A Statistical Analysis of Cricket One Day Internationals

Aishwarya Pratap Singh (1211162781) Ayushi Jain (1204841348)
Rahul Aakunuru (1209394573) Nikhil Lohia (1211168085)

Abstract—Cricket is a game played in over a hundred countries worldwide; One Day International (ODI) is one of the most common forms played. Cricket enthusiasts may be interested in an easy-to-use method to see how their favorite team has been performing throughout the years, while sports analysts may be interested in seeing the trends of different teams' performances throughout the years. In this paper, we examine the effectiveness of various visualization techniques for data generated from ODI matches between 2006 and 2017. We had created four visualizations through which users may explore team performance, player performance, and the relationships among the two. The visualizations are interactive. Users may sort and filter the data to see only the information that is meaningful to them. Our visualizations comprise a bar chart, a scatter plot, a histogram, and a stacked area chart. By combining the visualizations, we had formed a system to provide users better insight to the games. This system allows them to see different players' proficiency, their contributions to the success of their teams, and how those teams perform against one another. Finally, we present case studies to give evidence for the applicability and merits of our proposed visualization techniques.

Index Terms—Cricket, Performance, Visualization types, scatter-plot, histogram

1 Introduction

CRICKET is an extremely popular game around the world. To put this in perspective, there were an estimated 2.2 billion viewers watching the game during the ICC Cricket World Cup 2011. With cricket becoming extremely fast paced and increasingly competitive, performance analysis is steadily making its way into the teams dressing rooms. Data visualization and data analysis in cricket has increasingly become a critical part of the game. For instance, players can try to become technically better by analyzing their own performances. Second, teams can make plans for their upcoming matches by analyzing the performance of their competitors. Third, Selectors can analyze team performance to select appropriate teams. Finally, the International Cricket Council can analyze trends to maintain the level of the games taking place. Our project aims to further investigate many such metrics that may affect the performance of a team or the popularity of the game. We present our visualizations in a way that they tell a story to the end user and let them understand how the game has progressed over the past years, in addition to how the teams have gone about looking at their performance. We try to present a visual narrative that is "an account of a series of events, facts, etc., given in order and with the establishing of connections between them". [?]

Each game of cricket can generate a lot of data that can be difficult to comprehend and analyze without some sort of visual representation. Thus, we have created a system of four interactive data visualizations in which users can explore team and player performance and discover underlying trends in the data. To start, we discuss current related work that deals with analysis and information visualization of sports statistics from various sports like baseball and soccer. Second, we explain the source of our data, the characteristics of the data, the relevant parts of the data, and how we aggregated and cleaned the data to enable our visualizations. This is followed by a detailed description of our approach in which we explain our techniques their applicability. Third, we introduce our data visualization tool and describe its features. We explain how users can use the tool to get valuable insights from the data. Finally, we discuss a usability survey that was carried out to evaluate our approach and its effectiveness.

2 Related Works

With the increase in the use of technology in games like cricket, a lot of data is being collected. Exploring such data and finding patterns or insights will help analyze players performance and their effect on match outcomes. This indeed helps the teams to develop a strategy and tactics to win a game. Charles Perin at al. [?] has visualized soccer data in which they provide the design guidelines for different facet views and their layout in spacio-temporal flow. This will help experts to quickly look at the game and find the phase where the game can turn around.

Prasant Nair has built a predictive model for Cricket game [?], where he predict the outcome of the game based on ball-by-ball real time scores. But he has used visualization to detect the correlation and coordination between different variable. This didn't give any insights about the game or about the players. We went through bunch of visualizations and best practices used in visualization to effectively communicate the insights of game in general and cricket in particular

2.1 Baseline Bar Display for Scores and Margin

Andy Cox and John Stasko have created a baseline bar display in their sportsViz [?] system to visualize the scores in a baseball match and the margin by which a team has won or lost. In their system, they have functionality to sort the data by runs or margin. In cricket, we have similar statistics called runs and margin of runs by which the team can win or lose. We have used similar design to visualize the runs scored by a team. Since we have multiple visualizations in our systems, we allow user to select each bar and the other visualizations will be updated accordingly.

2.2 Scatter Plot:

Fengbo et al. designed and built an application called NBAViz [?] in which they visualize the performance of NBA players. In their visualization, they have used scatter plot to show the impact of players on the performance of team. For each player played in a game, they have plotted number of minutes that the player played vs number of points that the player scored. In our cricket visualization, we have similar metric to measure batting performance of team, which is the number of runs scored vs number of balls faced by a batsman. So, we have used this idea to visualize how a batsman is impacting the outcome of the game.

2.3 Brushing and plotting a Bar Graph

The Scatter plot tells how good of a batsman a team has, but it doesn't allow user to zoom in to area and filter the values. We have used Brushing and linking, which is a well-known technique in infoViz, to allow user to select and zoom an area. The principle of brushing and linking was used by Becker and Cleveland in their paper Brushing Scatter-plots [?]. Brushing is a process of selecting a portion of graph or plot. Often brushing is followed by linking to some other visualization in the system. We have used brushing to allow user to select a part of scatter plot and consider the players involved in that area. This brushing is linked to a bar graph in which we have a stacked bar for each player who are present in the brushed area and height of the bar represents the number of times an inning played by the player is present in the brushed area.

2.4 Stacked Area Graph

Along with team's performance and player's impact on team, it is important to look at the player's performance over the years. We could use histogram for each player where length of each bar in histogram is proportional to the runs scored by the player and show the histogram of different players in multiple views. This kind of visualization will give the player's performance over the years, but it will be difficult for the user to compare the performance between players. Theme-river [?] works better for these kind of visualizations. The width of the theme is consistent with the number of runs scored by a batsman and it connects the value with runs scored in before year with a smooth curve which will be visually pleasing. The horizontal flow represents the flow of time and the color currents within the river represents the player's data. But theme river doesn't give us the percentage of runs scored a player in total runs scored by top 10 players because theme-river expands on top and bottom part of x-axis. And hence we have choose Stacked area chart [?] which reduces the variation is graph making it easy to read and compare the values

3 Design Principles

The amount of information that can be represented on a unit area of screen is very limited and with increasing it data becomes harder to fetch relevant insights from it. Showing all the data at once can cause confusions and make the judgment even harder. The main goal of our design is to compare team vs team performance and how each player makes an impact on the outcome. To achieve this we visualize one team at a time and then add the option to compare it with another.

The process can be summarized as:

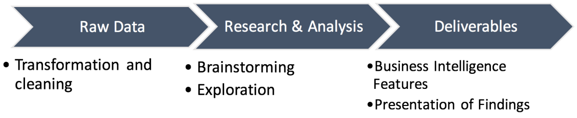

Raw Data
Research & Analysis
Deliverables
• Transformation and cleaning
• Brainstorming
• Exploration
•Business Intelligence Features
•Presentation of Findings

Fig. 1. Design process

We start with raw data obtained from Cricsheet [?] and transform it to a *csv* format. Since data loss is inevitable, we choose to abstract attributes explicitly based upon the target audience. Our visualization caters to 2 types of audiences

- A sports journalist who wants to analyze a teams performance
- A fan who wishes to discover trends of his/her favorite team.

Keeping it in mind, we discover the relevant insights they would be interested in. Some of them could be

- Analyze a players contribution to the teams success.
- Identify players in a certain bracket.
- Team vs Team metrics
- Compare the performance of players against different oppositions.

A good visualization is one which has interactivity and makes the comparison engaging. We present the viewer with data controls and let them uncover things on their own. Based upon these reasonings we follow 2 main design principles. Firstly, the user should be able to select the teams to compare. Secondly, the user should be able to interact with the visualization to get additional insights in his area of interest.

As per the guidelines of Kuchinsky et al. [?] multiple views are used when there is diversity in attributes and level of abstraction. It was also recommended to use Multiple views when different view brings out correlation or disparity or when we want to break down the complex data into multiple manageable chunks and provide insights. In our cricket data, we have attributes like player names, player country, runs scored etc. where the attributes are very diverse and we have different level of abstraction at game level and at player level. Although the dimension of data is not very high, we want to break it down to provide insights of game. And hence we have chosen to use multiple views instead of using single view and overwhelm the user.

Kuchinsky et al. [?] has suggested the way in which multiple views can be used. Our system is in consistent with the rules specified and we have placed multiples views close to each other to optimize time and space and views are kept consistence.

The multiple views which we have provided are not isolated, but they have relationship between them. These relationships are utilized by co-ordination. "Coordination is apparent foremost when user interaction comes into play". [?] Maximilian Scherr in his paper "Multiple and coordinated views in Information Visualization" [?] talks about the work of North and Schneidermans Snap-together visualization [?] where they talk about the way to give a user the ability to explore the data and coordination between the views.

According to this model, the system need to have a relational database and user is given opportunity to query the database and once he does that, the existing views should be updated. Coordination is done by bringing all the updated visualization together and creating a snap version of the system. Although our system doesn't have relational database at back end, we have created multiple files on which the user can query. Based on the user's query, we load the appropriate file and update the views of each visualization and present the snap of the system to user.

The outcome of the matches will be win or lose, which are categorical variable but can be thought of binary variables. For the variables which are binary in nature, we used binary color scheme which differs in the hue. [?]

4 System

4.1 Data

A single game of Cricket generates a huge amount of data. One game of two teams consists of three hundred deliveries divided into fifty overs each. Each delivery can result into scoring runs ranging from zero to six scored by eleven different batsmen. We can aggregate this data at many levels along this hierarchy and create several interactive visualizations. The user is served into exploring data with respect to a particular team against an opponent. For example, we can explore the insights when Australia plays against India. We leverage the ball by ball data obtained from Cricsheet [?]. From this data we extract the outcome of each game, teams involved, and the runs scored by each player over the years. This creates a large collection of data for several players and teams. For example, we get a data on over $30,000$ different innings played over the time. Furthermore, this data is annotated by attributes like balls faced and runs scored.

Limited by the scope of this project, we focus only on the runs scored by a team and a batsmen as it has been proved to affect the outcome of a game. '*Strike Rate*' is a major factor which determines the performance of the player and is defined as:

$$Strike\ Rate = \frac{Runs\ Scored}{Balls\ Faced}$$

4.2 Implementation

The entire dataset is obtained from Cricsheet [?] as a YAML file for each game. The YAML file contains the data in the following format and we have around 1300 files for the matches played.

```
info:
  city: Mumbai
  dates:
    - 2006-06-13
  match_type: ODI
  outcome:
    by:
      runs: 38
    winner: England
  overs: 50
  player_of_match:
    - ME Trescothick
  teams:
    - India
    - England
  toss:
    decision: bat
    winner: England
  umpires:
    - R Dill
    - DB Hair
  venue: 'Wankhede Stadium'
innings:
  - 1st innings:
      team: England
      deliveries:
        - 0.1:
            batsman: ME Trescothick
            bowler: Anil Kumble
            non_striker: EC Joyce
            runs:
              batsman: 0
              extras: 0
              total: 0
```

Each file is parsed and we create 3 types of csv files

- matchTable.csv

  ```
  match_id,team_1,team_2,
  team_1_runs,team_2_runs,
  outcomeBy,outcomeByValue,winner,date
  ```

- performanceTable.csv

  ```
  match_id,playerName,runsScored,
  ballsFaced,team,result,opposition
  ```

- streamTable.csv

  ```
  playerName,runsScored,date,
  team,opposition
  ```

These files are then ingested into the system to perform visualizations.

We performed several tests to analyze the performance and behavior of several of our design decisions and our prototype implementation. Our prototype system was written as a JAVASCRIPT application using D3 [**?**] and C3 [**?**] libraries and some helper libraries like ‘d3-queue’ [**?**] to resolve asynchronous issues. Our prototype was tested on Google Chrome running Apache Tomcat server in a machine with an Intel i5-3770k processor, 8GB of RAM, and an NVIDIA GTX 670 graphics card, all under Windows 7 64-bit. Our system work best for screen resolution 1280 x 800 and when chrome window is maximized.

4.3 Layout

Our system is comprised of four modules that visualize team and player performance. We have primarily used the number of runs scored as the metric to measure the performances of the players as well as the impact it has on the overall performance of the team. In our system, we provide the user with two drop-down menus, that list the names of the various cricket playing nations. These drop downs provide an effective way of filtering the results based on a team and the opposition it has played against. The first drop-down allows users to select a country for which they would like to view the team and player performances, the second, optional drop-down allows users to narrow results by selecting a single country that has played against the first country. By default the second drop down is set to show the results against all the oppositions.

4.4 Views

4.4.1 *Bar charts*

The first visualization contains two parallel bar charts (positioned one below another) which can be used to analyze a teams performance using their runs score. The first chart shows the number of runs scored by the team in a time-line fashion. The x-axis contains an ordinal scale with match dates while the y-axis contains the scores (runs). The second chart shows the margin by which the team has won or lost in each match. The length of the each bar in both charts represents the score or margin respectively. In case the team won the match while batting second, the margin is taken to be one. The color of each bar shows whether the match was won or lost.

The visualization also has a sorting feature which allows users to sort matches by date, runs score, or margin of runs scored. The sorting mechanism is useful and really effective in arriving at the conclusions such as, what is a safe score for a team if batting second, what is the probability that a team would successfully defend or chase a target set by the opposition.

Filtering based on the margin of runs helps us understand whether the team plays too many close games. This could be due to the fact that the team loses out in the critical points in the game.

Another valuable insight that can be derived by sorting the data based on the margin is to gaze whether the team's performance has been increasing, if the margins have been decreasing, or vice versa.

Using these sorting features along with the second drop down to select the opposition, can help us understand how the team has been performing against different oppositions. Fig **??** shows a bar chart for India

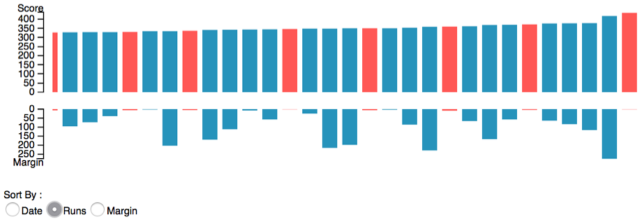


Fig. 2. Bar chart for Team India

The next two visualizations in our system combine to make a multiple linked view. Using the data visualization technique of brushing, we linked a scatter-plot with a histogram.

4.4.2 *Scatter-plot*

A scatter-plot is used to depict the relationship between the number of balls faced by a batsman in innings and the runs they scored. The x-axis contains the number of balls in an inning faced by the batsmen and the y-axis contains the number of runs scored. The relationship between these two values is known as a “strike rate” in cricket. Scoring a lot of runs in the game is as important as the number of balls faced in scoring those runs. Thus, the players who score a lot at an extremely fast rate are considered to be more valuable players when compared to players who make a lot of runs at a slower rate. Essentially, each point in the scatter-plot will represent an inning played by player in terms of their strike rate. Fig **??** shows the scatter-plot for each innings played by the players of Team India.

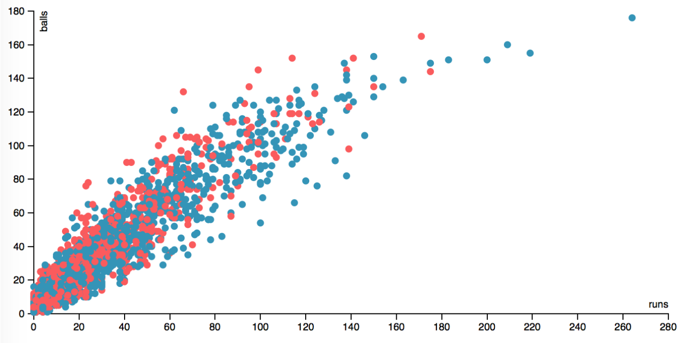


Fig. 3. Scatter-plot for Team India

The points in the scatter-plot are color-coded; the color of a point represents the outcome of the match in which the inning (i.e. the point of interest) was played. Blue is used to represent a win while red is used to represent a loss. For example, a point depicting the strike rate of Player A will be blue if the team had won the particular match. Color-coding the scatter-plot help us find an underlying relationship between a players performance and the outcome of the match.

On each point in the scatter-plot, we show a tool-tip which gives additional information such as the corresponding player name, the players runs scored, and the number of balls faced by the player.

The scatter-plot also has an option to select a range of points (brushing). When a set of points is selected in the scatter-plot, a dataset is generated with all the points in the selected area and passed to a histogram. that would visualize a stacked bar graph for each player for whom a point is found in the selection. The stacked bar will represent the number of points in the selected area for a lost cause vs winning cause for a player. Fig **??** shows the brushing feature in the scatter-plot.

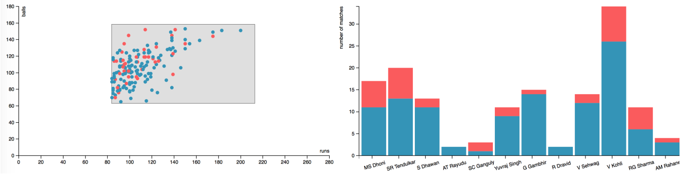


Fig. 4. Brushing in scatter plot

4.4.3 Stacked Bar

The stacked bar chart is designed to visualize player frequencies that will be retrieved from the dataset mentioned above. For example, if two points are selected in the scatter-plot and both correspond with strike rates of player “X”, the histogram will contain a bar for player “X” with a height of two. This visualization is aimed at understanding how a particular player’s performance affects the success of his team. A stacked bar in the chart represents the count of innings found in the selection made in the data visualized in the scatter plot. Each bar will be stacked based on the number of innings that were played in a lost cause against the number of innings played where the team won the game.

‘Finishers’ is a term often used to describe a player who has the ability to stay ‘not out’ till the end of the game and thus himself lead the team to a win. These players often have to face the pressure situation in a game. Fig **??** shows the stacked bar chart for Team India.

When used with the opposition filter, this visualization also helps us identify how certain players play against different oppositions.

4.4.4 Stacked Area Chart

Stacked area chart is an extremely effective visualization in visualizing the performance of the team based on the sum of the number of runs, the players of the team, have been scoring each year. Stacked area chart also provides an effective means of analyzing the contribution made by a key batsman in the total number of runs scored. This visualization is useful in understanding when the team has peaked and coinciding this with the important tournaments played over the years gives us an understanding of the results of a lot them. Fig **??** shows a stacked area chart for performance of Team India over the years.

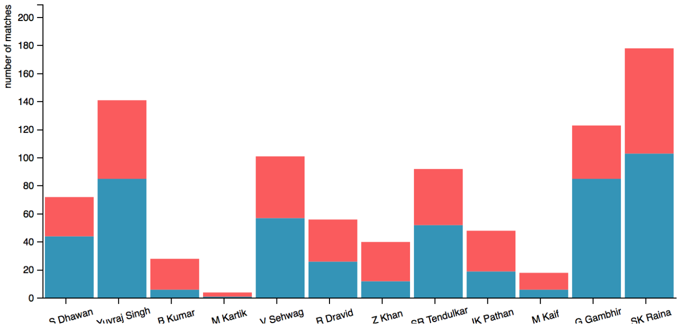


Fig. 5. Stacked Bar Chart

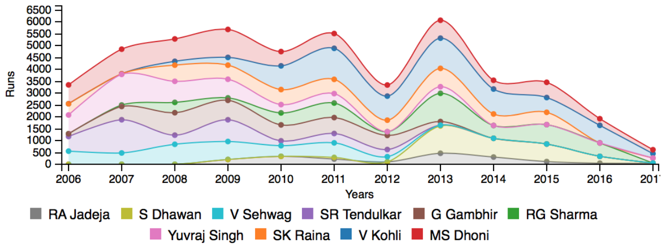
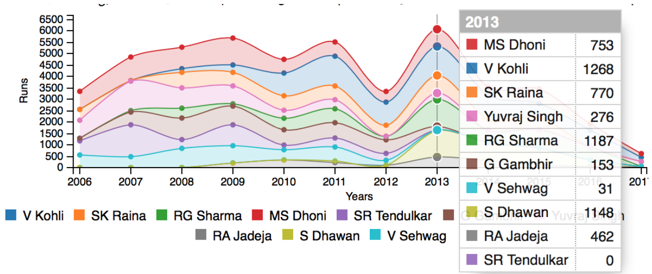


Fig. 6. Stacked Area Chart for Team India

Not only this, the visualization also helps us understand the effect that the retirement of the important players has on a team.

This visualization also helps analyze how the careers of individual players have evolved over the years. From the debuts and retirements of the players to injuries, that led to a temporary lapse in the performance of a player, can all be visualized using this visualization. Fig **??** shows the run details by each player of Team India in the year 2013.



2013	
🟥 MS Dhoni	753
🟦 V Kohli	1268
🟧 SK Raina	770
🟪 Yuvraj Singh	276
🟩 RG Sharma	1187
🟫 G Gambhir	153
🟦 V Sehwag	31
🟨 S Dhawan	1148
⬛ RA Jadeja	462
🟪 SR Tendulkar	0

Fig. 7. Details on demand for Team India in 2013

5 Results

As we have seen so far, this visualization is aimed at analyzing the performance of the teams over the last 10 years. The visualization also tries to identify how the performance of certain players affects the overall team performance more than the others. While analyzing this data, we consider the performance of the batsmen of a team as the key indicator of their performances. Specifically, we use the metric "Strike Rate" to quantify the performances of the players.
Using this visualization, we have been able to

1) Explain why, or how the team performed in a specific way at a period
2) Derive at facts that have thus far not been popular in the cricket fraternity.

We present below, a case study of some specific scenarios where the visualization has helped us achieve the results mentioned above.

5.1 What has been a defend-able total for a team?

By using the simple sort by runs feature in the bar graph representing the runs made by team in each of their matches, we can derive at a very interesting conclusion. When we attempt to understand this metric for different teams, we see that most of the teams do not have an evident score beyond which they win a hundred percent of their matches. The only exception to this is South Africa. We can clearly see that South Africa wins all their games when they score beyond 300. It is also interesting to see how Australia, often considered to be the best team in the world has lost the match they scored the highest in, along with several others where their scores were at the higher end of the spectrum. Even more interesting is the fact, that most of these high scoring games that Australia has lost are against South Africa.

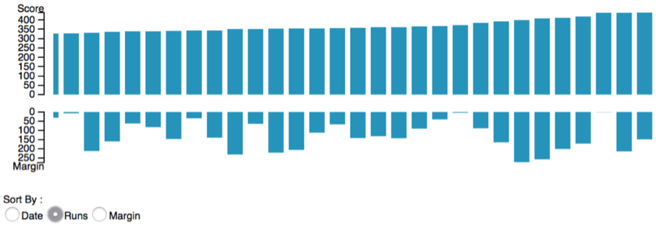


Fig. 8. South Africa against all

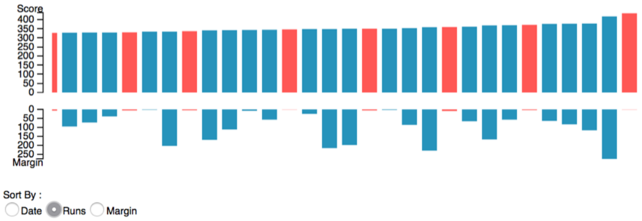


Fig. 9. Australia against all

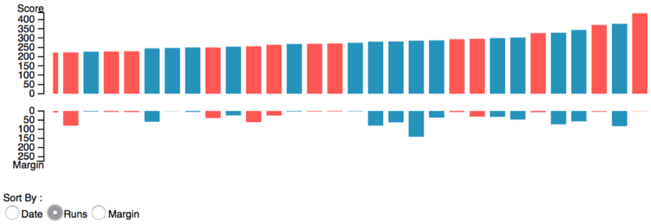


Fig. 10. Australia against South Africa

5.2 What have been the causes for some of the exceptionally good and exceptionally bad periods of performance of a team?

This case presents itself as a unique study of how the omission of some of the important players of the teams may explain why the team had a slump in their performance at a particular period. For instance, in the year 2016, when three of the most prolific batsmen in the Sri Lankan cricket team, *Jayawardene*, *Sangakara* and *Dilshan*, retired, their performance as team went down to an all time low in the last 10 years.

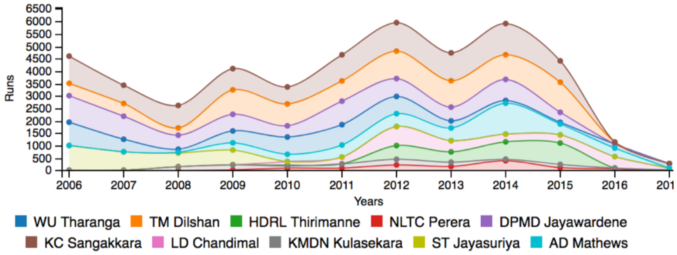


Fig. 11. Performance of Sri Lanka over the years

A similar situation happened with India in the year 2012 when *Sachin Tendulkar* retired while *Virendra Sehwag* and *Yuvraj Singh* were left out of the team due to inconsistent performances and injury issues respectively. Though this may be a self-explaining scenario, what is important to see is what happened after this slump in the performances. It is interesting to see that India emerged out even better in the following year with 2013 being their best year in terms of performance in the last 10 years, 'Sri Lanka' is still struggling to find their footing since. This fact sheds a lot of light on the planning done by the team managements on phasing out the players of the team as well keeping a strong bench strength ready at all times. Bench Strength is a term used to describe the how good a set of backup players does a team have.

5.3 How were the champions made?

This case helps explain why certain teams have performed exceptionally well in some tournaments. India who were the winners of the International Cricket World Cup in the year 2011 and the winners of the ICC Champions trophy in 2013 were at the top of their performance in both those years. They however could not defend their title of the world champions in 2015 due to an indifferent performance by the

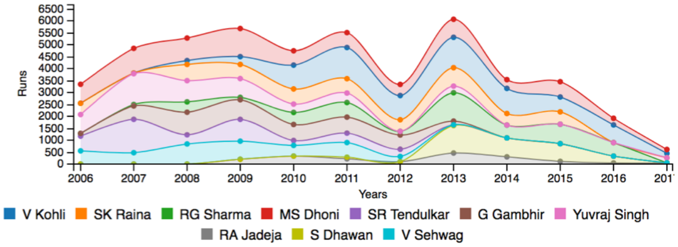


Fig. 12. Performance of India over the years

team. On the other hand, *New Zealand* had an exceptionally good year as a team in 2015 which resulted in them being one of the finalists in the World Cup. This was the first time ever that New Zealand had reached the finals in the tournament.

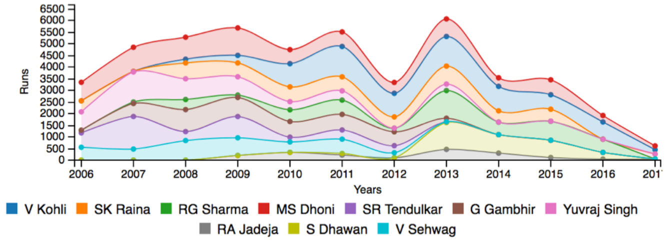


Fig. 13. Performance of India over the years

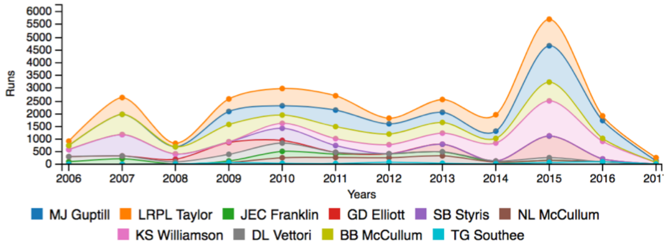
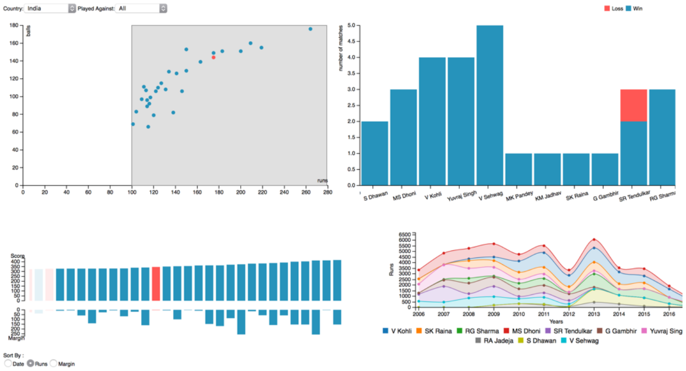


Fig. 14. Performance of New Zealand over the years

5.4 Who are the real match winners in the team?

Selecting all the top scoring matches of the team India we can view the innings played by their batsmen in the scatter-plot. On brushing the area, selecting the scores of 100 or more, we can view how many times each batsman has played an innings that lies in that area in the stacked bar chart. *Sachin Tendulkar*, often considered as the 'God of Cricket' in India, does his reputation no good as he is the only one to have played big innings in a lost cause for India. We can also clearly see how *Virender Sehwag* appears to be the one to contribute maximum towards the success of the team. Another important find that we came across is to identify the real finishers for the teams. Finishers are the players who usually play towards the end of the innings in high pressure situations and often are the ones that take the team through. *MS Dhoni* is often tagged to be the best finisher in the history of the game. We wanted to contrast his performances against some of the other players who may have played a similar part but may not have received their due credit in the process.


Country: India Played Against: All

Sort By : ○ Date ● Runs ○ Margin

Fig. 15. Performance of Indian players at a glance

To select the finishers, we select the area with the range of runs between $40 - 90$ and the range of balls played in the range of $0 - 60$. This window was chosen as this range of runs is the minimum to have a significant effect on the game and the range of balls played lets us select players who have scored at a rate of more than a run a ball.

We came across a startling discovery to find that another player from the same team, *SK Raina*, has in fact had as many number of innings in the matches that India has gone on to win. The important thing to note is that his win loss ratio is far lesser compared to the 'all time finisher' *Mr. MS Dhoni*.

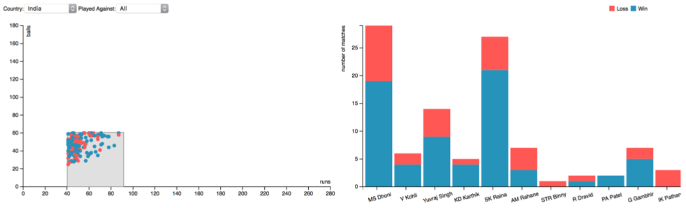
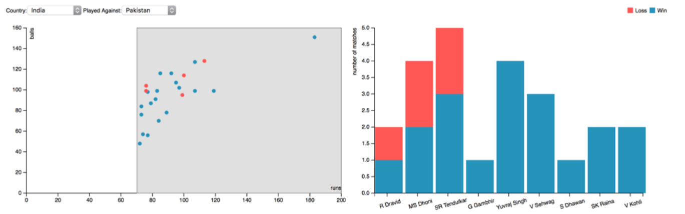

Country: India Played Against: All

Fig. 16. Performance of MS Dhoni and SK Raina

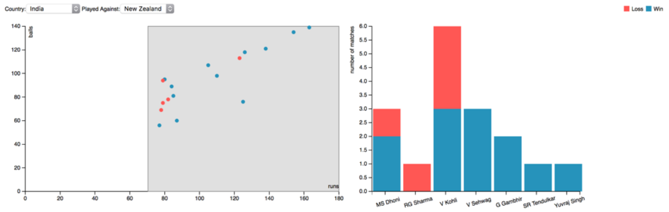
5.5 How do players fare against different oppositions

In this scenario, we tried to find out how different players have performed against different oppositions. We have made some interesting discoveries here as well. When selecting all the matches played between India and Pakistan, we see how different players of India have performed against this opposition. *Sachin Tendulkar* though again emerges as the player with the maximum number of scores in the range of 70 or more has a $3 : 2$ win-loss ratio. On the other hand, *Yuvraj Singh* has had 4 such innings with India winning all these games. When we change the opposition to 'New Zealand', *Yuvraj Singh's* contribution comes down to just one successful inning. He also has a dismal record against 'Australia'. Interestingly, *Rohit Sharma* appears to have done well in terms of his performances against 'Australia', but those innings have only led to a win-loss ratio of $4 : 3$.



Country: India Played Against: Pakistan

Fig. 17. India vs Pakistan



Country: India Played Against: New Zealand

Fig. 18. India vs New Zealand

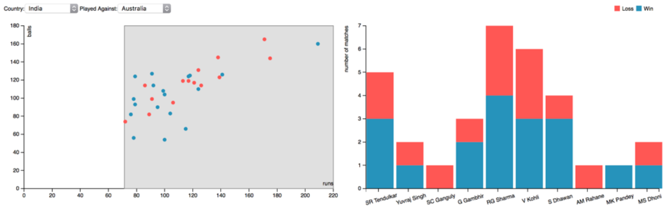
There are a few more metrics that go on to decide the outcome of a game, but it is interesting to see how with just the batting metric we can explain so many trends and results with respect to the teams. It can be an interesting endeavor to correlate our findings with some of the other metrics in the game.

6 Conclusion

We have presented a system that analyzes a teams match plan by following Ben Shneiderman's information-seeking mantra [?]. Namely, we proposed analysis and visualization techniques that follow the steps "overview first" (Scatter Plot), "zoom and filter" (Brushing and sorting bar chart), and "detail-on-demand" (Histogram and stacked area chart). We focused on the fixed parameter of strike rate of a player. However, there still exists a challenge to extract information and gain insights from bowling statistics of a team. Some teams are known for their exceptional spinners or for players who can conceive all the wickets in a game. Such players can heavily impact the outcome of a game. In the future, we plan to integrate bowling information in our visualization system.

Interactivity among the visualizations is a crucial part of our prototype. We integrated methods to analyze players performances over time. However, we may need to search for similar player behavior on more abstract levels. We may want to know how players adapt their behavior when playing with others on their teams; player partnership may be a very interesting area of analysis.

Aside from player strike rates, other important metrics to consider could be team behavior and team strategy. For example, a team could play offensive under one captain but defensive under another one. Thus, we may want to know how players perform under these kinds of conditions. Although we dont currently have a defined approach to visualize strategies/tactics, doing a sentiment analysis could be a potential first step toward this problem.



Country: India Played Against: Australia

Fig. 19. India vs Australia