**A Needle in a Data Haystack**

* 1. **Writeup**
* Project Title: Sports – Stats & Facts
* Team member info:
* Lior Cohen, [lior.cohen8@mail.huji.ac.il](mailto:lior.cohen8@mail.huji.ac.il), lior\_13
* Ben
* Ori Broda, [ori.broda@mail.huji.ac.il](mailto:ori.broda@mail.huji.ac.il), orib
* Link to project repository on github: <https://github.com/jjgold012/data-science>
* Problem description: There are massive-unused data about sports that can be used for betting, or just for fun…
* Data: We found many data on sport types in .csv and .xls formats.

Those files were found in different betting websites and forums that share them, such as bet365.

Each excel contains data on many matches occurred over years/seasons (we got from about 4 years of data in hockey to about 40 years in football) and betting odds from betting websites. The main attributes we addressed are: match date, the league/tournament, names of competing teams, score of each team, the winner and the betting odds.

The different sport types we collected and the amount of records are:

soccer – 113157

football – 35795 for american and 1508 for australian

basketball – 11021

hockey – 6848

rugby – 2563

tennis – 41187 for men and 24705 for women

cricket – 194

Summing up for a total of 236978 records.

Total amount of space used for repository: 304MB

* We took all the excel files we downloaded and inserted into SQL tables, with scripts we wrote, the columns we thought are interesting the most which are mentioned above, and created a table for each sport type. All can be found in the git folder.

After doing so we were able to perform any SQL queries we want and display it in many ways. We decided to display the queries as facts and as graphs.

Every detail can be observed using an easy-to-use software we developed, which can be installed and executed (hopefully) in your computer.

* Experiments:

We should make some experiments to evaluate our performance (we believe in you Ben ☺ )

* Future work: It is possible to extend the tables we built and add a column of the referee and reveal dark secrets about them, like their favorite team that barely lose while the referee in charge. Also, we can decide, by the months matched where played, what is the season with the highest chance to predict scores and make a lot of money. Can look for the temperature in that day and check if the weather affect game result. Another possibility for future work is to expend the data to other sport types, like cycling, horse racing, F1 and the Olympics.
* Conclusion: We managed to collect various data and establish a decent sports DB. Using the data we were able to obtain valuable and solid information about teams/players over the course of more than a decade. We believe that it is possible to predict future game results at some high probability based on this data, and maybe even help the betting sites to be more accurate (for example, we saw in the 'underdog' chart (see below) that the amount of wrong predictions is roughly around 25% to 30% for the majority of sports).