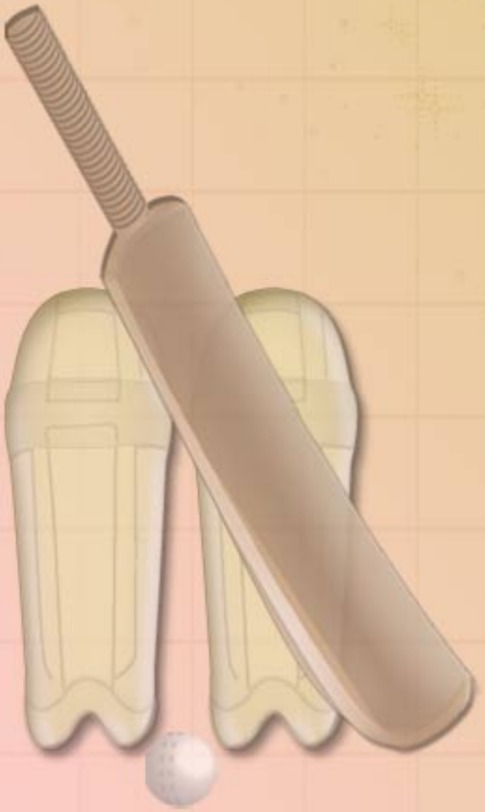
# Data Mining Lab Course
# Cricket Dataset



Week 4 Presentation:

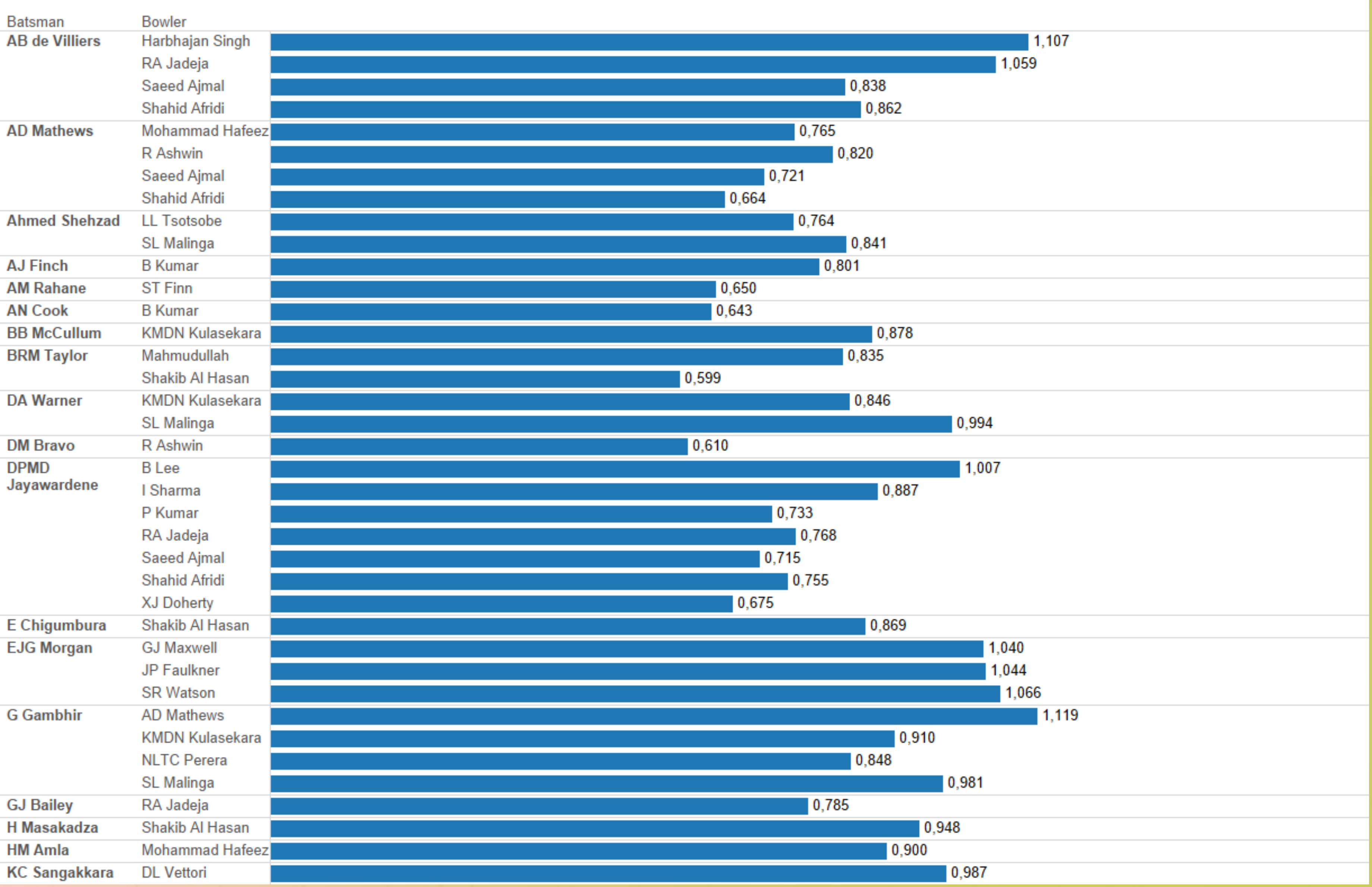
Avinash Mishra
Matti Lorenzen
Raveekiat Singhaphandu

# Main Achievements

- Performance of a Batsman against a particular bowler.
- Parsing of Players profile from cricinfo website.
- Setting up the Performance parameter for players.
- Players performance with a particular nonstriker batsman.
- Detailed and easy description.
- Updated last week Wiki to be more descriptive.

- Feedback for the Mashable Online Popularity Data Set

# Strike Rate of a Batsman against a particular bowler



| Batsman | Bowler | |
|---|---|---|
| AB de Villiers | Harbhajan Singh | 1,107 |
| | RA Jadeja | 1,059 |
| | Saeed Ajmal | 0,838 |
| | Shahid Afridi | 0,862 |
| AD Mathews | Mohammad Hafeez | 0,765 |
| | R Ashwin | 0,820 |
| | Saeed Ajmal | 0,721 |
| | Shahid Afridi | 0,664 |
| Ahmed Shehzad | LL Tsotsobe | 0,764 |
| | SL Malinga | 0,841 |
| AJ Finch | B Kumar | 0,801 |
| AM Rahane | ST Finn | 0,650 |
| AN Cook | B Kumar | 0,643 |
| BB McCullum | KMDN Kulasekara | 0,878 |
| BRM Taylor | Mahmudullah | 0,835 |
| | Shakib Al Hasan | 0,599 |
| DA Warner | KMDN Kulasekara | 0,846 |
| | SL Malinga | 0,994 |
| DM Bravo | R Ashwin | 0,610 |
| DPMD Jayawardene | B Lee | 1,007 |
| | I Sharma | 0,887 |
| | P Kumar | 0,733 |
| | RA Jadeja | 0,768 |
| | Saeed Ajmal | 0,715 |
| | Shahid Afridi | 0,755 |
| | XJ Doherty | 0,675 |
| E Chigumbura | Shakib Al Hasan | 0,869 |
| EJG Morgan | GJ Maxwell | 1,040 |
| | JP Faulkner | 1,044 |
| | SR Watson | 1,066 |
| G Gambhir | AD Mathews | 1,119 |
| | KMDN Kulasekara | 0,910 |
| | NLTC Perera | 0,848 |
| | SL Malinga | 0,981 |
| GJ Bailey | RA Jadeja | 0,785 |
| H Masakadza | Shakib Al Hasan | 0,948 |
| HM Amla | Mohammad Hafeez | 0,900 |
| KC Sangakkara | DL Vettori | 0,987 |

# Players profile

- We parsed the profile of each player from espncricinfo.com.
- Some attributes:
    - Playing_role, Batting_style, Bowling_style, Batting/Bowling Average
    - SR: Strike rate

BUT WHY??

- More detailed analysis of player performance.
- Players performance comparison during 10 year time frame.

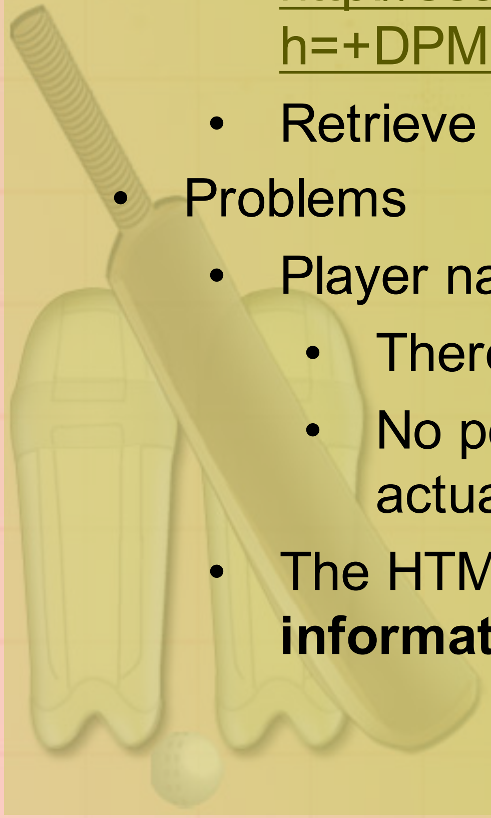
Problem :

- Different name of a player in different dataset.
- IN PlayerInfoAll table
    - Name: Mahela Jayawardene
    - Full Name: Denagamage Proboth Mahela de Silva Jayawardene
- In Innings table
    - Batsman: DPMD Jayawardene

# How do we solve it?(1)

First method:

- We found that ESPNcricinfo have the search bar which we can simply build URL request
  - For DPMD Jayawardene
    - http://search.espncricinfo.com/ci/content/site/search.html?search=+DPMD%20+Jayawardene;type=player
    - Retrieve all the link result to player profile
  - Problems
    - Player name are more common than we think
      - There are **five** Amjad Ali from Pakistan
      - No possible way to easily distinguished them without actually look through the data
    - The HTML page for each player **does not contain any information about the abbreviation**



# How do we solve it?(2)

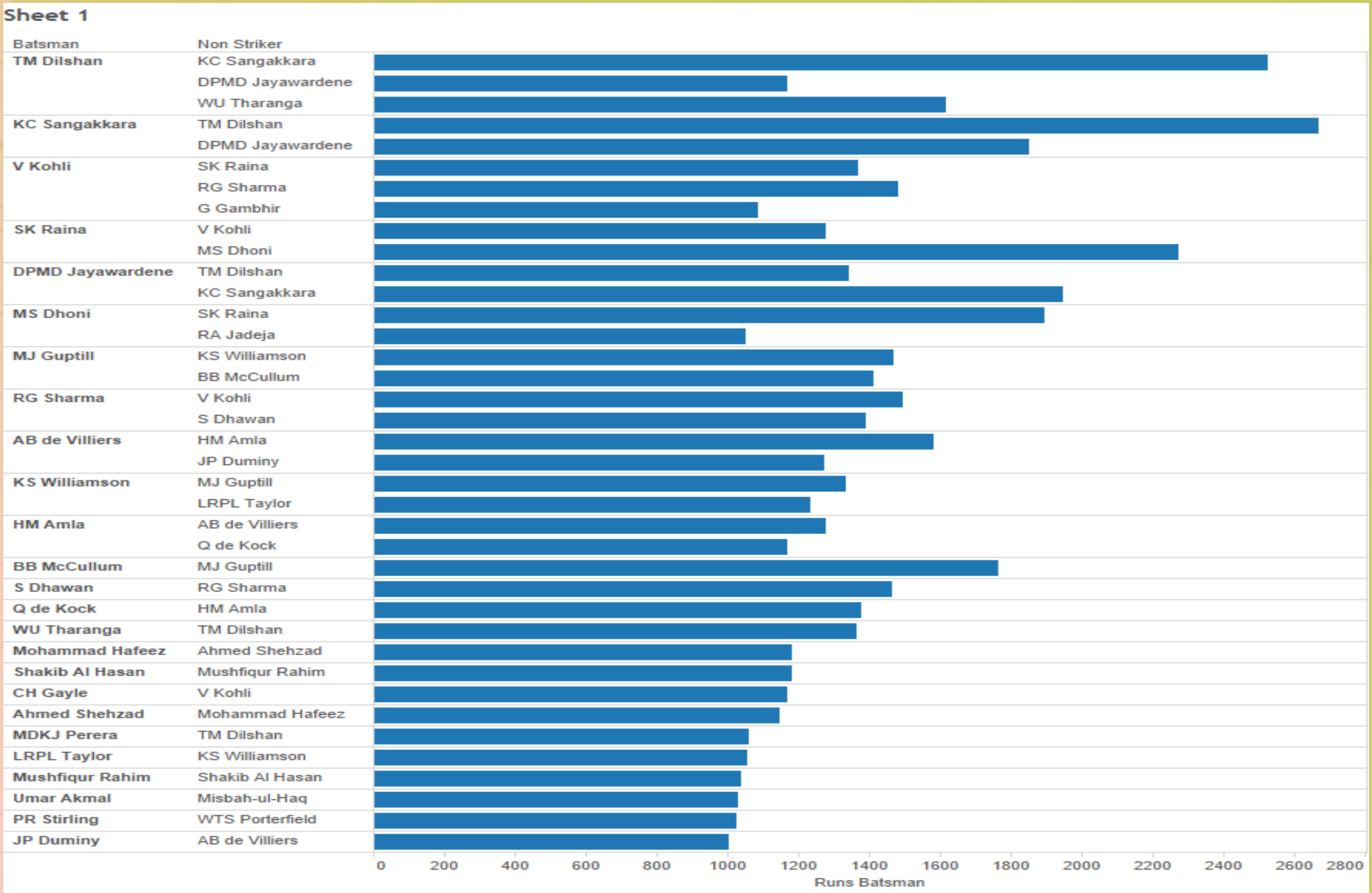
Second method:

- Again, we found that ESPNcricinfo have the list currently active team member and also using **abbreviation name ☺** for each player
- Now we list all the player abbreviation name and their link
- Download those link and save in HTML format
- Parsed interested data in to CSV
- Finally we can link this recently acquired data with the innings table

# Performance Parameters

| ' | ODI | T20 |
|---|---|---|
| Batting Average | 37 | 28.02 |
| Batting Strike rate | 77.84 | 119.27 |
| Bowling Average | 27.4 | 30 |
| Bowling Strike rate | 31.5 | 23.5 |
| Bowling Economy rate | 5 | 6.5 |

# Batsman and nonstriker batsman



Sum of Runs Batsman for each Non Striker broken down by Batsman. The view is filtered on sum of Runs Batsman, which ranges from 1,000 to 12,294.

# Next Task…..

- Player performance against a type of bowler/batsman.
- Batsman performance with type of batsman at non striker end.
- Players performance in this time duration as compare to whole career.
- Player performance grouping

# Feedback for Team 4 :
# Mashable Online Popularity Data Set

- No longer using original dataset due to errors discussed last week
- Extraction of statistical features, categorized number of shares → 3 groups
- Analysis of frequent words, channels usually published in
- In the presentation: „Most Valuable Authors are Publishing Their Most Articles In The Same Channels”
  In the wiki: “Against our intuition, each author is publishing in all channels and does not have a single channel of expertise.”
- Very well structured, a clear plan, the wiki is looking very nice.
- Out of interest – did you notify the author of the paper regarding the faults in the dataset?