

Sports Analytics with R

UseR! Conference 2017

Stephanie Kovalchik



About Me

- ISEAL Research Fellow at Victoria University
- Principal Data Scientist for Tennis Australia's Game Insight Group
- Tennis Blog: on-the-t.com
- @StatsOnTheT



Tutorial Resources

- Course package: [deuce](#)
- Course tutorial: [UseR Sport Tutorial](#)
- Contact: s.a.kovalchik@gmail.com

What It's Like to Be a Sports Statistician

- Finding sports data
- Web scraping sports data
- Data wrangling
- Exploratory data analysis
- Predictive modelling
- Blogging



Finding Sports Data

Popular Types of Sports Data

- Box Scores
- Performance Metrics
- Tracking Data
- Wearable Data

Box Scores

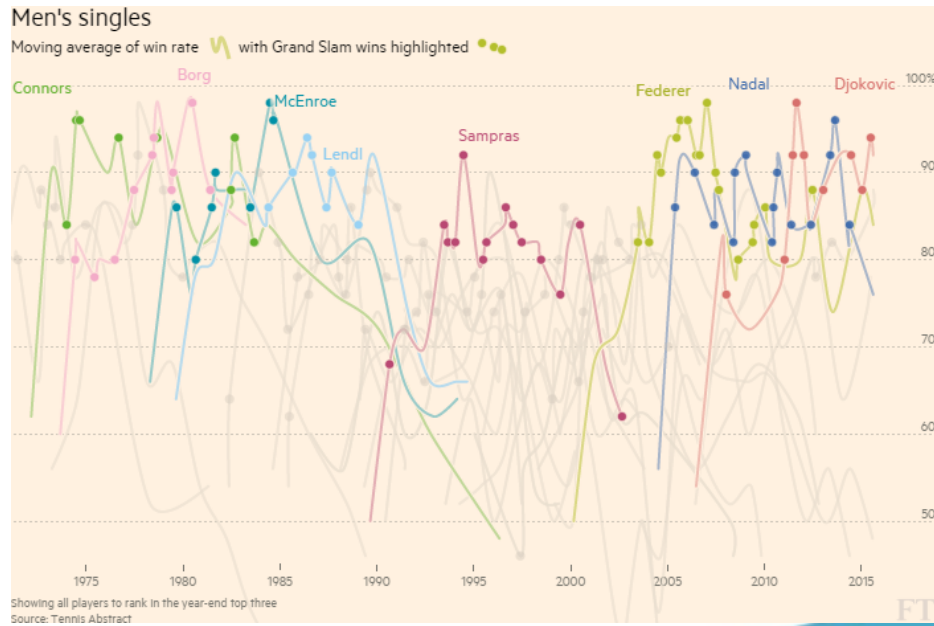
A 'box score' is a generic term for the summary statistics reported about a single sporting event.

BOSTON.							ATHLETIC.						
T.	R.	l.b.	P.O.	A.	E.		T.	R.	l.b.	P.O.	A.	E.	
G. Wright, s.s.	6	4	4	1	5	2	Force, s. s. . .	5	1	2	1	3	2
Leonard, 2b.	6	3	3	4	4	3	Eggler, c. f. . .	5	3	3	0	0	0
O'Rourke, 1b	6	2	3	9	0	1	Fisler, r. f. . .	5	0	1	2	0	0
Murnan, l. f.	6	1	0	3	1	0	Meyerle, 3db	5	1	2	2	3	3
Schafer, 2d b	6	3	3	3	1	2	Sutton, 1st b.	5	1	2	10	0	0
McGinley, c. f	6	0	0	0	0	1	Coons, c. . . .	5	1	0	1	1	3
Manning, r. f.	6	0	2	2	0	0	Hall, l. f. . . .	5	1	3	5	0	0
Morrill, c. . .	6	2	2	4	1	2	Fowser, 2d b.	6	1	2	6	7	5
Josephs, p. .	5	4	4	1	1	2	Knight, p. . .	5	2	2	0	1	2
Totals. . . .	53	19	21	27	13	13	Totals. . . .	46	11	17	27	15	15
Boston. . . .	9	1	3	3	4	1	0	2	5	19			
Athletic. . . .	1	0	0	0	3	3	2	2	0	11			
Runs earned—Boston, 4; Athletic, 5. Home-run—Hall, 1.													
Total bases on hits—Boson, 22; Athletic, 20. First base by													
errors—Boston, 8; Athletic, 5. Umpire, George White of													
Lowell, Mass. Time 2h. 47m.													

[1] Baseball box score from 1876.

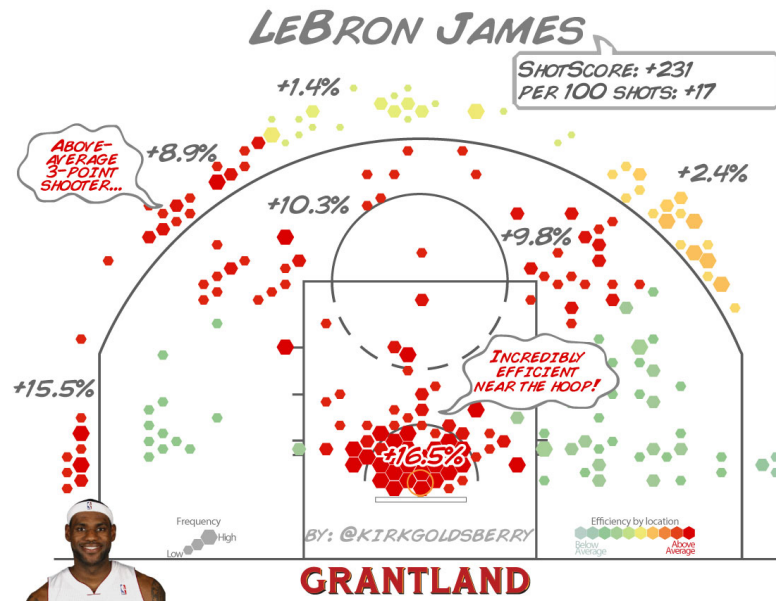
Performance Metrics

Performance metrics are usually repeated measures of team or players that are derived from historical box scores.



Tracking Data

'Tracking data' is spatio-temporal data of objects in a sporting event.



Wearables

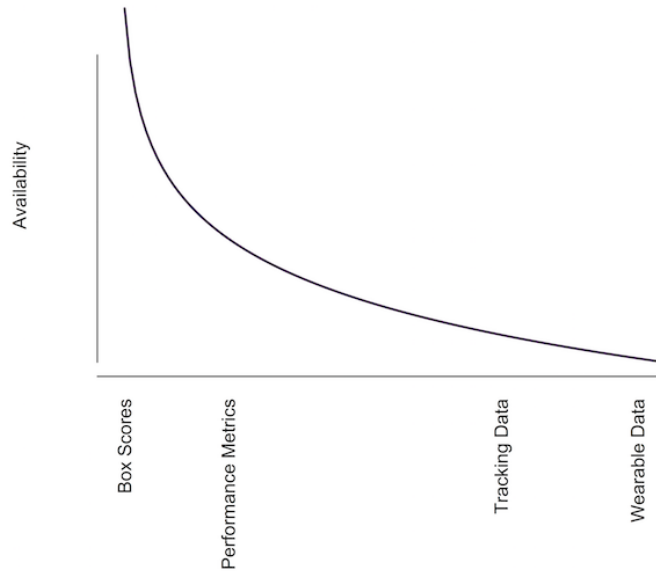
- 'Wearables' are technologies that collect data about athletes when worn.
- Biometrics and movement are two common types of data collected.



[1] VU Biomechanics Lab

Data Availability Curve

Sadly, not all types of sports data are equally available though the steepness of the curve depends on the particular sport.



Category	URL
Multiple	www.databaseSports.com www.opensourcesports.com http://www.espn.com.au/blog/statsinfo http://www.masseyratings.com/data.php
American Football	http://www.bballsports.com/ http://www.cfbstats.com/blog/college-football-data/ http://www.databaseFootball.com/ http://www.nfl.com/stats/player http://www.pro-football-reference.com http://stats.washingtonpost.com/cbk/teams.asp
Baseball	http://www.baseball-almanac.com/teamstats/statmaster.php http://www.baseballamerica.com/statistics/ http://www.baseballheatmaps.com/graph/distanceleader.php http://www.baseballmusings.com/cgi-bin/DayByDayDatabase.py http://www.baseballprospectus.com/sortable/ http://www.baseball-reference.com http://baseballguru.com/battingdatabase.html http://www.bballsports.com/ http://www.databasebaseball.com http://www.dougstats.com/ http://www.fangraphs.com/ http://www.hittrackeronline.com/index.php http://www.baseball-links.com/ http://www.retrosheet.org http://www.sabr.org/ http://www.seanlahman.com/ http://www.statcorner.com/ http://www.thebaseballcube.com/
Basketball	http://www.82games.com/ http://www.apbr.org/ http://www.basketball-reference.com/ http://www.basketballvalue.com/ http://www.bballsports.com/ http://www.nba.com/statistics/index.html http://www.databasebasketball.com/ http://www.dougstats.com/ http://www.finalfour.net/ http://www.hoopdata.com/ http://www.kenpom.com/ http://www.ncaa.org/ http://stats.washingtonpost.com/nba/

Category	URL
Hockey	https://puckalytics.com http://www.bballsports.com/ http://www.databasehockey.com/ http://stats.hockeyanalysis.com/ http://www.hockey-reference.com/ http://www.hockeydb.com/ http://www.nhlpa.com/stats/league-wide http://www.nhl.com/ice/statshome.htm http://www.quanthockey.com/ http://stats.washingtonpost.com/nhl/teams.asp
Soccer	http://www.football-data.co.uk/ http://www.football-data.co.uk/data.php http://soccerway.com http://www.squawka.com/ http://www.statto.com/football/stats http://www.transfermarkt.com/ http://www.whoscored.com/ http://www.worldcup.com/
Other	http://www.uci.ch/ http://www.espnricinfo.com/ http://www.databasegolf.com/ http://www.tennisabstract.com/ http://www.databaseSports.com/olympics http://www.rugbydata.com/ http://www.afl.com.au/stats http://www.volleyball.org/
Category	http://www.databaseping.com/ http://games.crossfit.com/leaderboard

Case Studies

Navigating sports sites and finding where data lives can sometimes be tricky. Let's get some practice by inspecting the following sites. For each case, determine where the main data lives, what type is there, and in what format.

1. <http://www.hockeydb.com/>
2. <http://www.football-data.co.uk/data.php>

Answer: Hockey Database

These data are under the 'Statistics' page and primarily exist as HTML tables.

Welcome to hockeydb.com, the internet's largest repository of hockey data!

Build the Perfect API
Learn how to plan, design, build, manage and share the perfect API.
mulesoft.com/Build_Perfect_API

TRENDING


5 Minutes Ago

1. Jonathan Drouin
2. Jonas Brodin
3. Brendan Warren
4. Nick Cousins
5. Mikhail Sergachev
6. Ryan Strome
7. Sam Reinhart
8. Ryan Nugent-Hopkins
9. Sergei Samsonov
10. Chris Bourque

1 Hour Ago

1. Jonathan Drouin
2. Nick Cousins
3. Brendan Warren
4. Alex Galchenyuk
5. Jonas Brodin
6. Merrick Madsen
7. Mikhail Sergachev
8. Simon Despres
9. Dion Phaneuf
10. Derek Stepan


HOCKEY STATISTICS



The standings and player statistics for nearly every professional hockey player to play — ever!

[» 2016-17 Morning Report](#)
[» Advanced Player Search](#)
[» Team Search](#)
[» Standings & Rosters](#)
[» Last Year's Players](#)
[» All-Time Records](#)
[» NHL Attendance](#)


NHL PLAYER LISTS



Interesting and different ways to look at NHL players.

[» Alphabetical Player List](#)
[» One Game Wonders](#)
[» Playoff-only Players](#)
[» More Playoff Players](#)
[» Most Teams](#)
[» Pack Your Bag Club](#)


TRADING CARDS



Checklists of NHL and minor league trading cards.


[» Player Search](#)
[» Set Lists](#)
[» Sets by Season](#)
[» Info & Rumors](#)
[» Minor Card Forum](#)

DRAFT PICKS



Lists and performance of players selected by the NHL and WHA.

[» Draft by Year](#)
[» Draft by Team](#)
[» Draft by Source Team](#)
[» Top Players by Pick #](#)

 Find us on Facebook
Find us on Google+

Answer: Football UK

A repository of betting data on soccer matches is available under the 'Historical Data' and these data are available as CSV files.

football-data.co.uk



bet365 JOIN NOW
Terms & Conditions Apply

Gamble Responsibly 18+

Network Sites

Data Updated: 11th June 2017

SHARE

Follow

HomeFree BetsLivescoresBetting AdviceBetting SystemOddsTipsCasinoPokerTennisBooks

bet365£200 Free

William HILL£30 Free

Sky BET£20 Free

Paddy Power£30 Free

BETVICTOR£40 Free

Ladbrokes£50 Free

BETFRED£60 Free

betfair£30 Free

Boylesports£25 Free

Paddy Power

McGregor 40/1

Mayweather 20/1

FREEBIES & PROMOS

£250,000 Super 6

bet365 Promotions

William Hill Promotions

Paddy Power Offers

£200 in Free Bets

£60 in Free Bets

£50 in Free Bets

£40 in Free Bets

£30 in Free Bets

£25 in Free Bets

Historical Football Results and Betting Odds Data

23 seasons results | 16 seasons betting odds | 16 seasons match stats

All FREE!!! Access Data via Country Links

In addition to the new [Livescore](#), [Tables](#) and [Statistics](#) service Football-Data continues to provide the football punter with computer-ready football results, match statistics and betting odds data for use with spreadsheet applications, to help with the development and analysis of football betting systems. What's more since July 2007 this data is now **FREE**. In doing so Football-Data takes the time out of recompiling pages and pages of results data and past betting odds found on a number of football results and [odds comparison](#) websites.

Forecasting profit: England 200W02



Acca Five Insurance

Now includes all sports

Min odds 1/5 per selection. Max free bet £50. Restrictions & T&Cs apply.

Join now

William HILL

SITE RESOURCES

Livescores

Historical Data

Betting Advice

Football News

Free Bets

Odds Comparison

Rating Systems

Football Ratings

In-Play Betting

Wisdom of Crowds

Tips Community

Forum

Betting Articles

Betting Problems

15 / 19

Packages

Sport	Package	What it Does
General	stattleshipR	Data capture API for multiple sports
	SportsAnalytics	Sports datasets and WEB import functions
	odds.converter	Convert odds to probabilities
Soccer	engsoccerdata	Repository of soccer datasets from 1891 to present
Baseball	pitchRx	Access to MLB Gameday data including pitchFX
	Lahman	Sean Lahman baseball database
Tennis	deuce	Repository of multiple tennis datasets
Cricket	cricketr!	Functions for analyzing ESPN Cricinfo stats

More on deuce

More on deuce

- deuce is a package I created to make it easy to access large historical data on tennis matches

More on deuce

- deuce is a package I created to make it easy to access large historical data on tennis matches
- It combines data from multiple sites:
 - [flashscore.com](#)
 - [ATP Tour](#)
 - [WTA Tour](#)
 - [Tennis Abstract](#)

More on deuce

- deuce is a package I created to make it easy to access large historical data on tennis matches
- It combines data from multiple sites:
 - [flashscore.com](#)
 - [ATP Tour](#)
 - [WTA Tour](#)
 - [Tennis Abstract](#)
- Some of the data you can obtain from deuce includes:
 - Match results from Open Era (1968) to the present
 - Historical rankings
 - Player demographics
 - Point-level data for multiple years at Grand Slams
 - Shot level data from the Match Charting Project

Installing deuce

You can install deuce using devtools. It may take several minutes because of the size of the datasets being transferred.

```
library(devtools)  
devtools::install_github("skoval/deuce")
```

Using Documentation

One of the best ways to get familiar with what a package does is to look at the package index. Try the following:

```
library(deuce)  
help(package = "deuce")
```

How would you learn about the contents of `atp_matches`?

Using Documentation

One of the best ways to get familiar with what a package does is to look at the package index. Try the following:

```
library(deuce)  
  
help(package = "deuce")
```

How would you learn about the contents of `atp_matches`?

```
help("atp_matches", package = "deuce")
```