MONASH University

MONASH BUSINESS SCHOOL

Accessing data and developing analysis software, to study game strategies, with examples from tennis and soccer

Di Cook
Professor of Business Analytics
Econometrics and Business Statistics

Outline

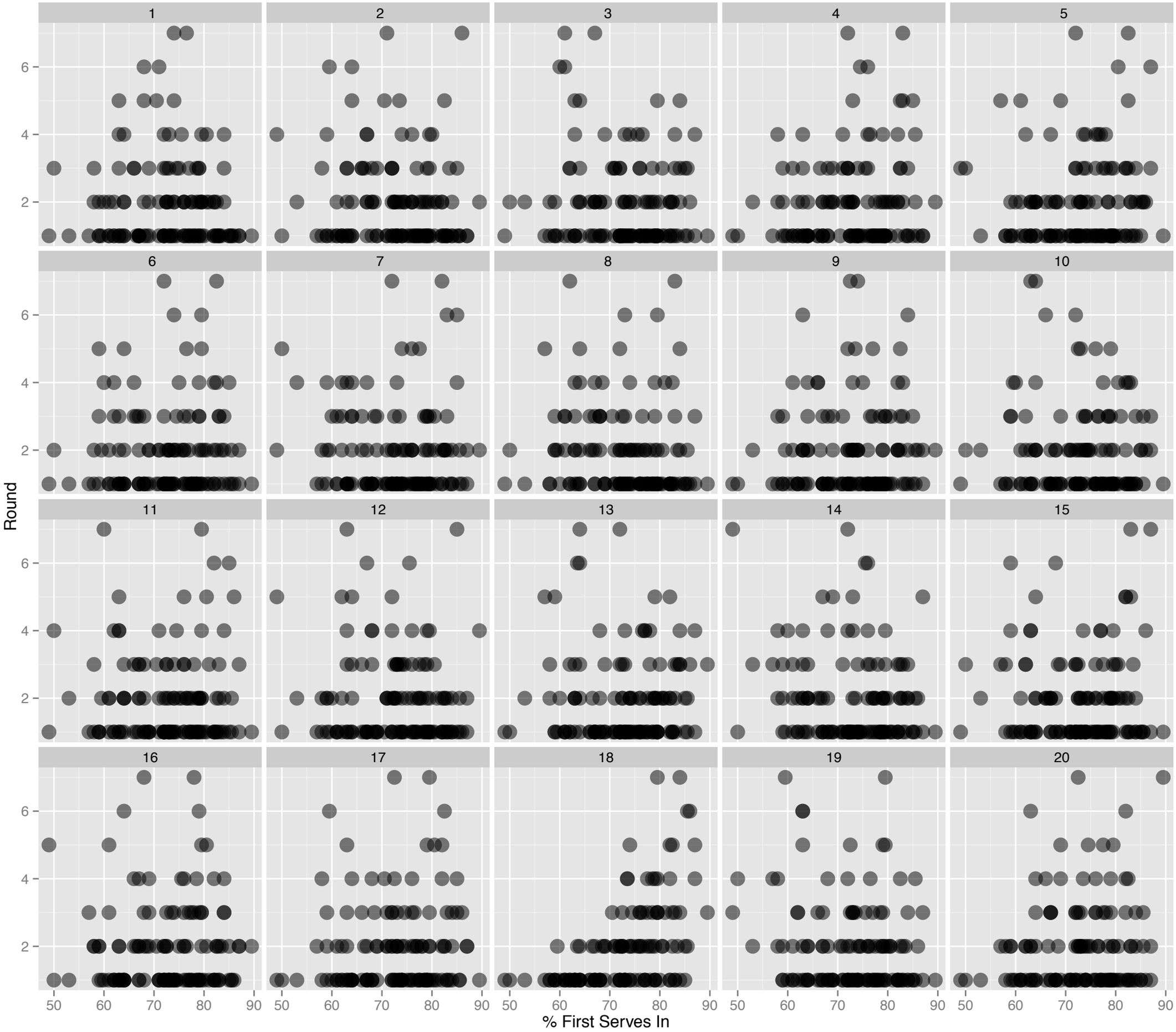
- Is Kygrios a future No 1? Inference with graphics (2014 analysis)
- Iowa high school soccer: public databases, web apps, quantitative rankings

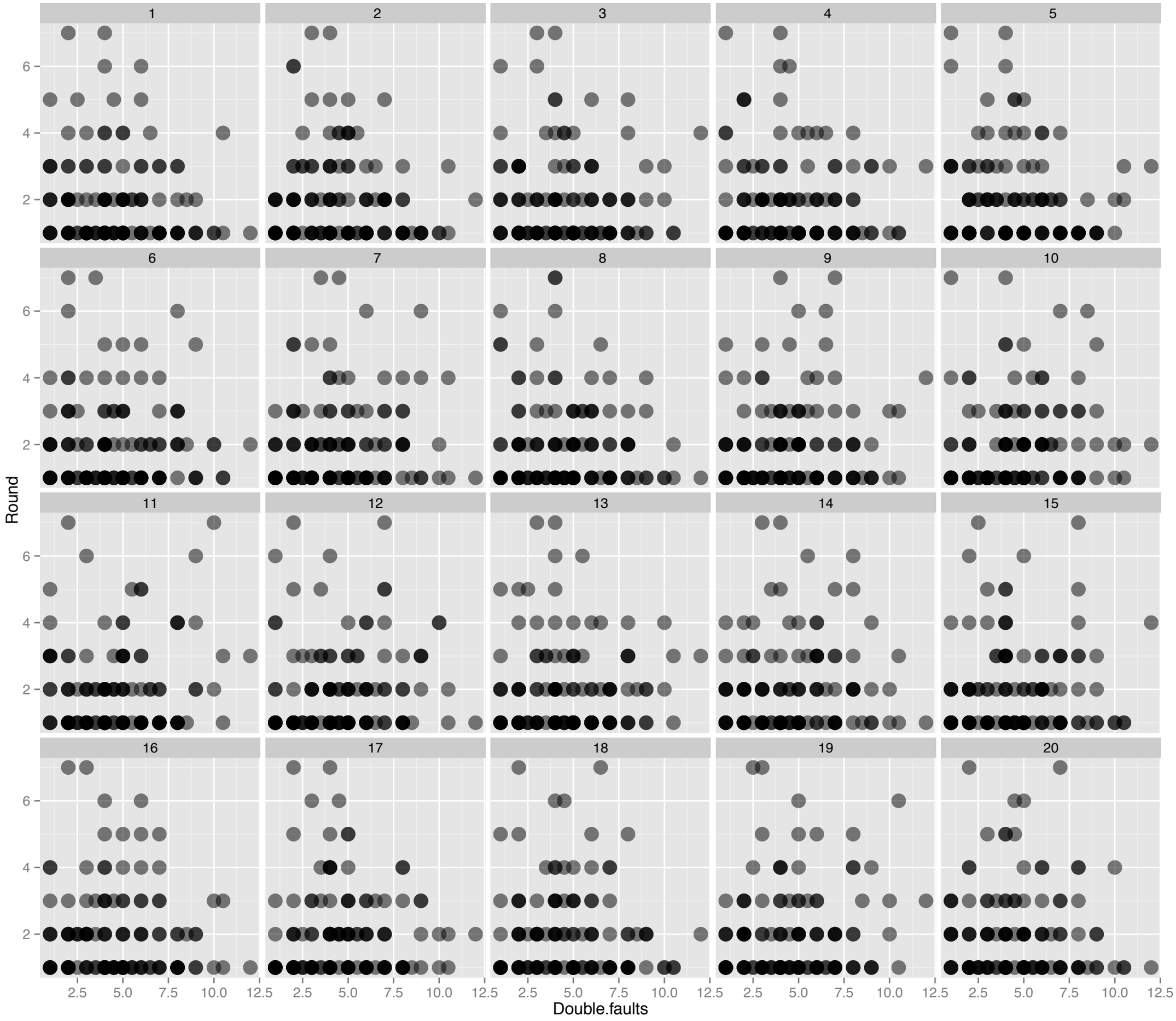
tennis

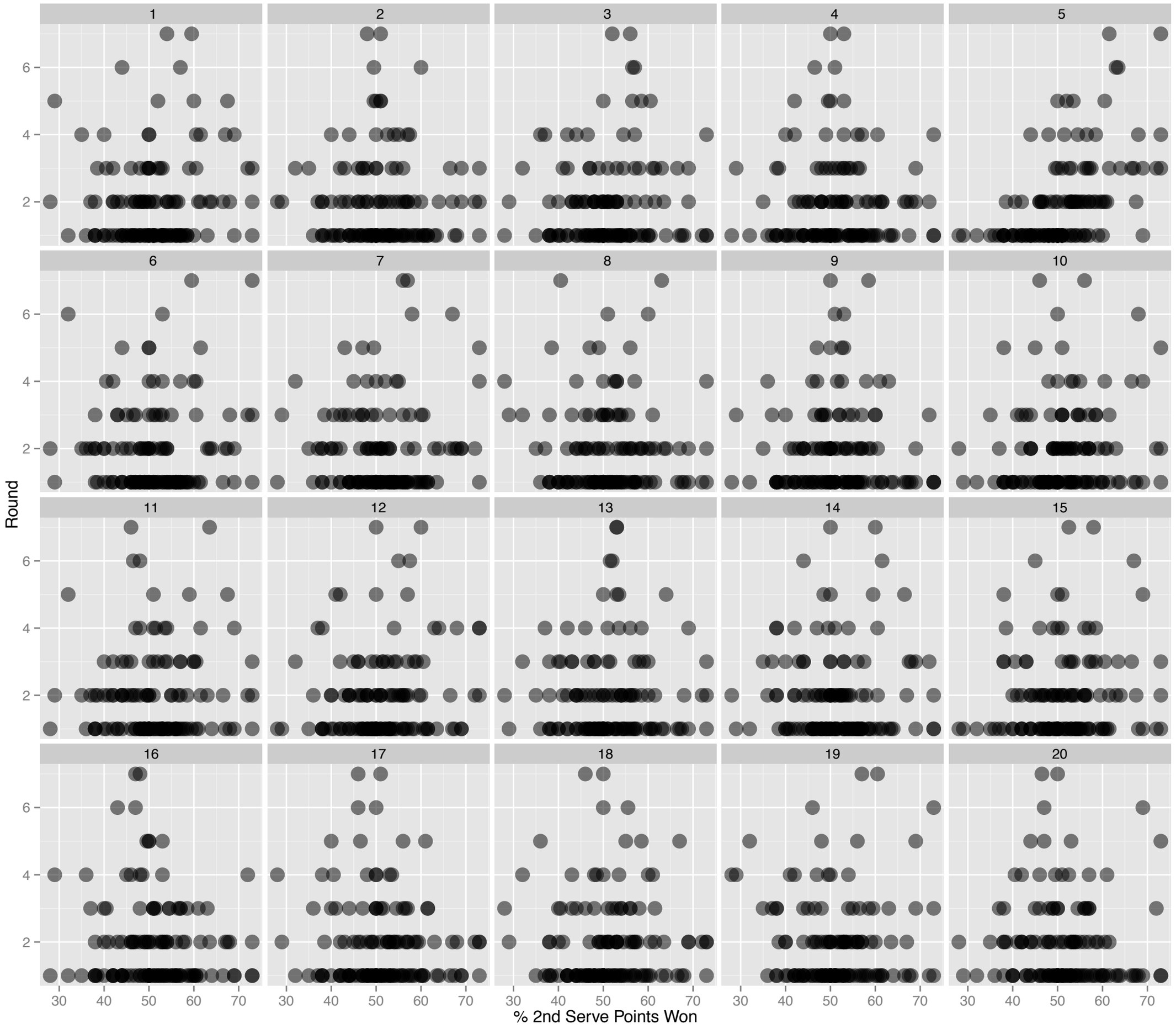
- Nick Kyrgios burst onto the international tennis on July 1, 2014 by beating world number 1, Rafael Nadal.
- Went to the Wimbledon web site, scraped the data from the matches
- Plots of aggregated statistics from first two rounds against round reached
- How good is Kyrgios relative to other players

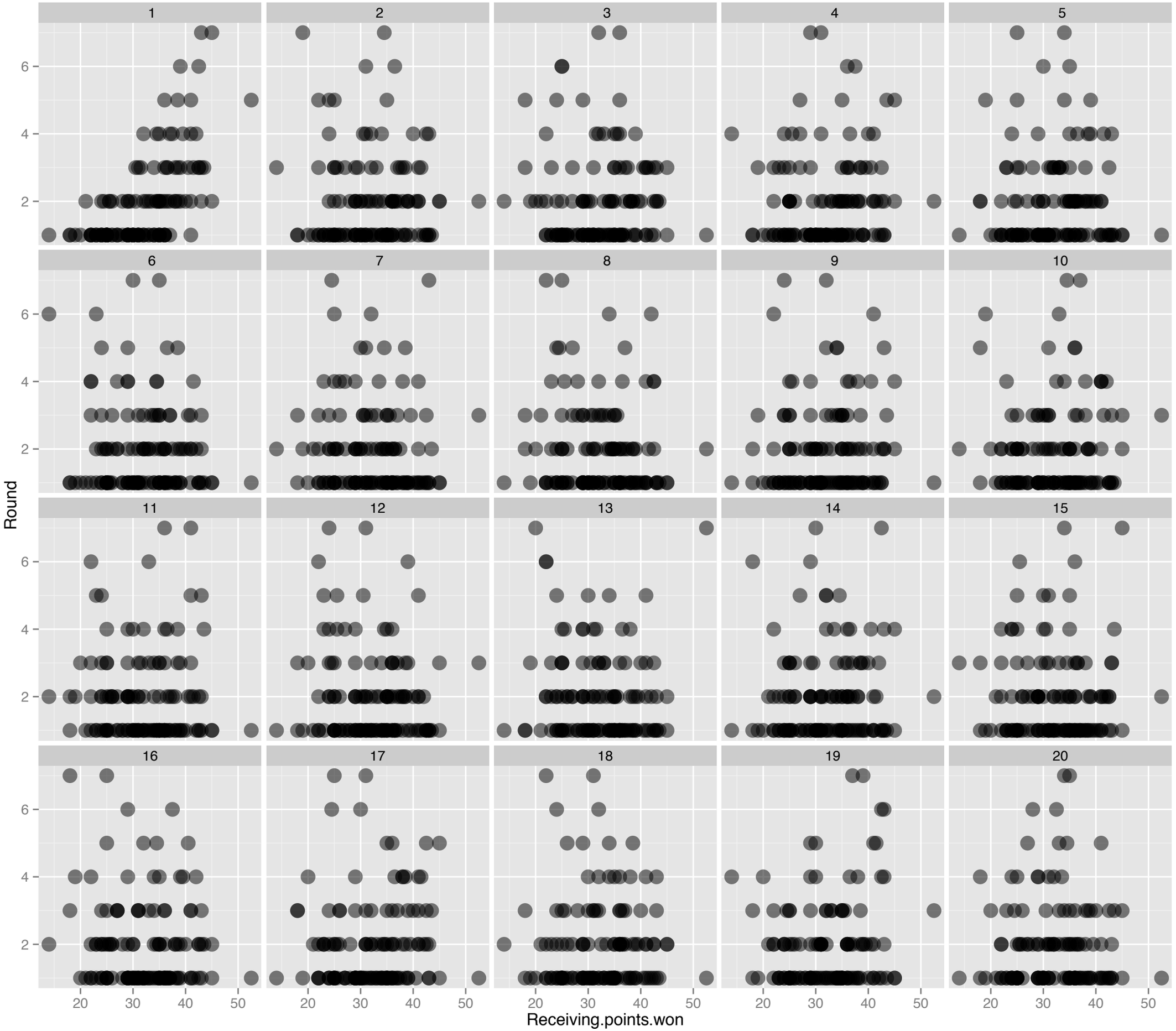
Rather than show the plot of the data alone, I'm going to show you the data plot embedded in a page of plots of null data (almost the same data) where there is NO ASSOCIATION.

YOUR TASK: Pick the plot that is most different from the others.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Round
% First Serves In
2
4
6
50
60
70
80
90


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Round
Double.faults
2
4
6
2.5
5.0
7.5
10.0
12.5

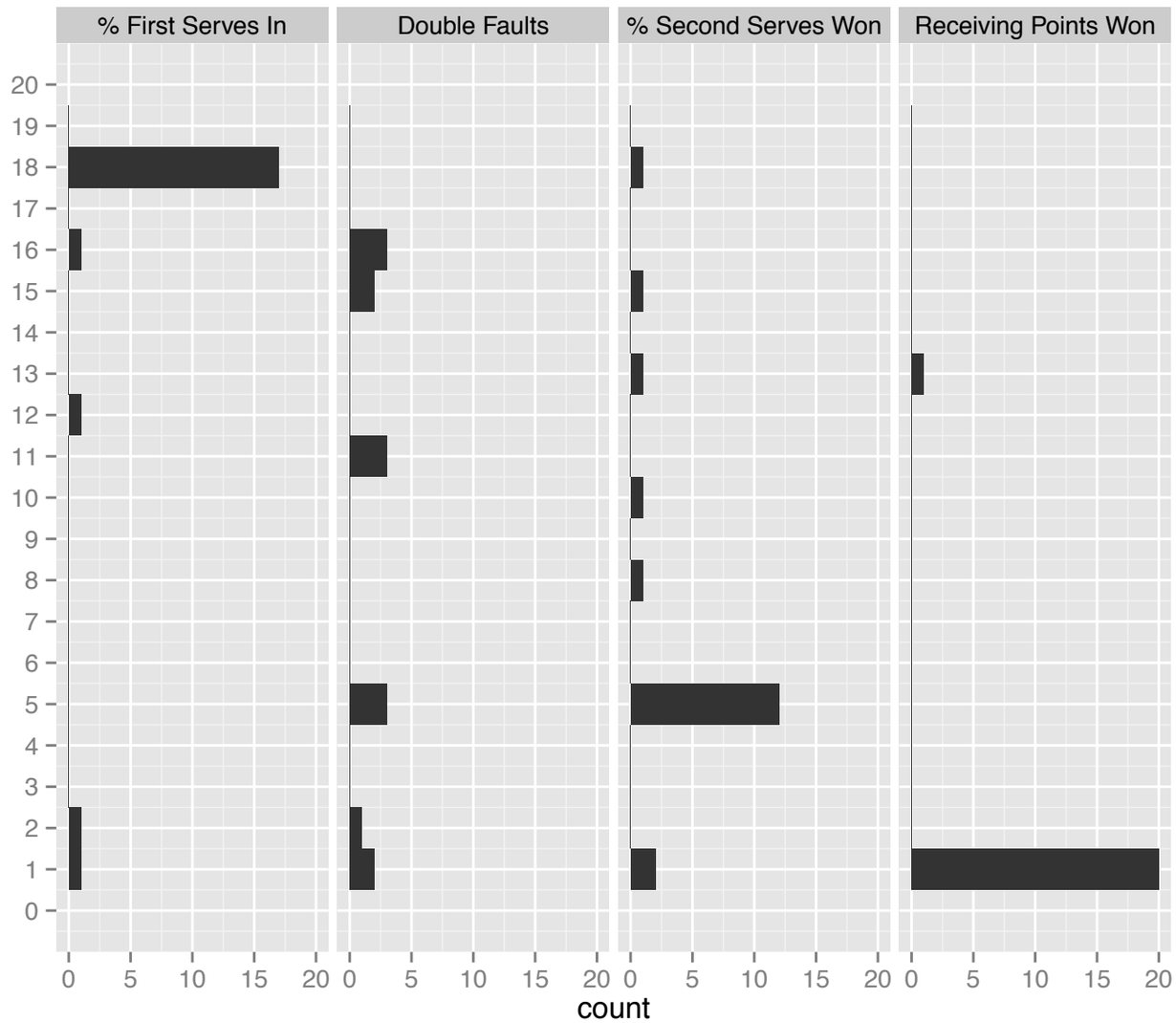

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
2
4
6
30
40
50
60
70
Round
% 2nd Serve Points Won


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Round
Receiving.points.won
2
4
6
20
30
40
50

Data: 18, 5, 5, 1

What do we learn? Advancing in tournament is related to % first serves in, %second serves won, receiving points won, but not so much the double faults.

Results of crowd-sourcing



How does it work?

- NULL PLOT: Association between variables is broken by permutation

%serves in	round
58	2
65	6
47	3
61	5

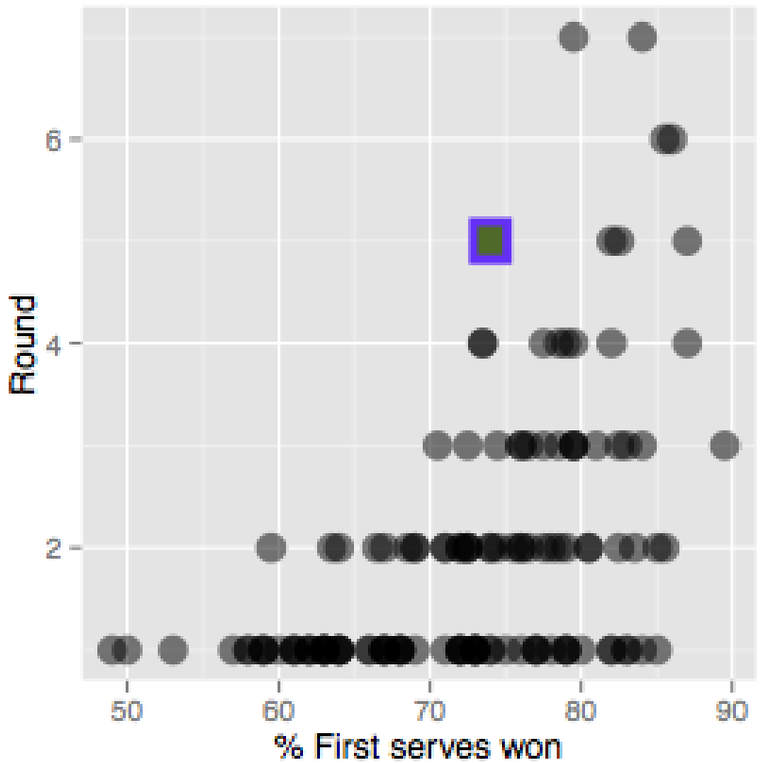
Permute round

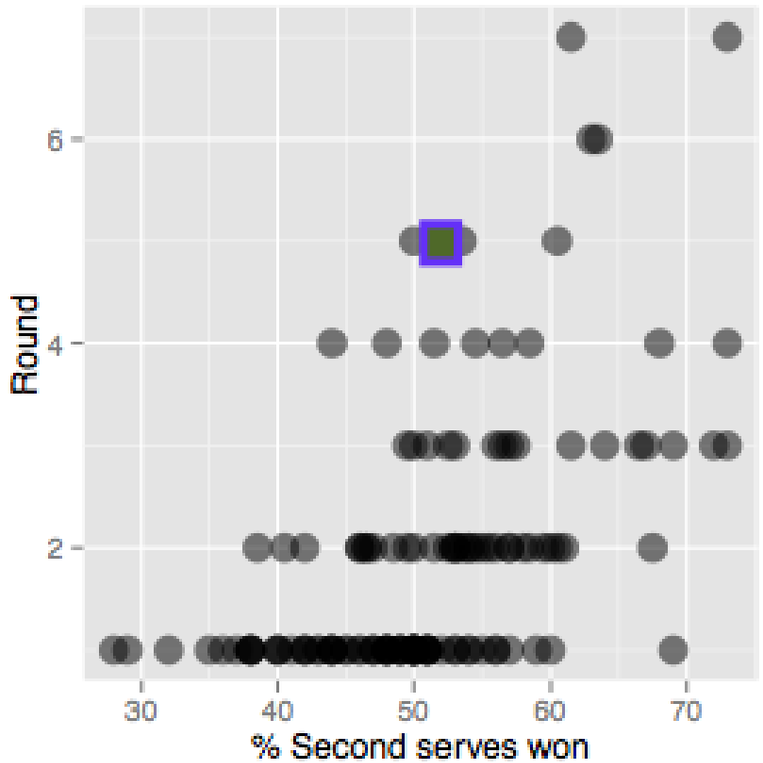
%serves in	round
58	**6**
65	**5**
47	**2**
61	**3**

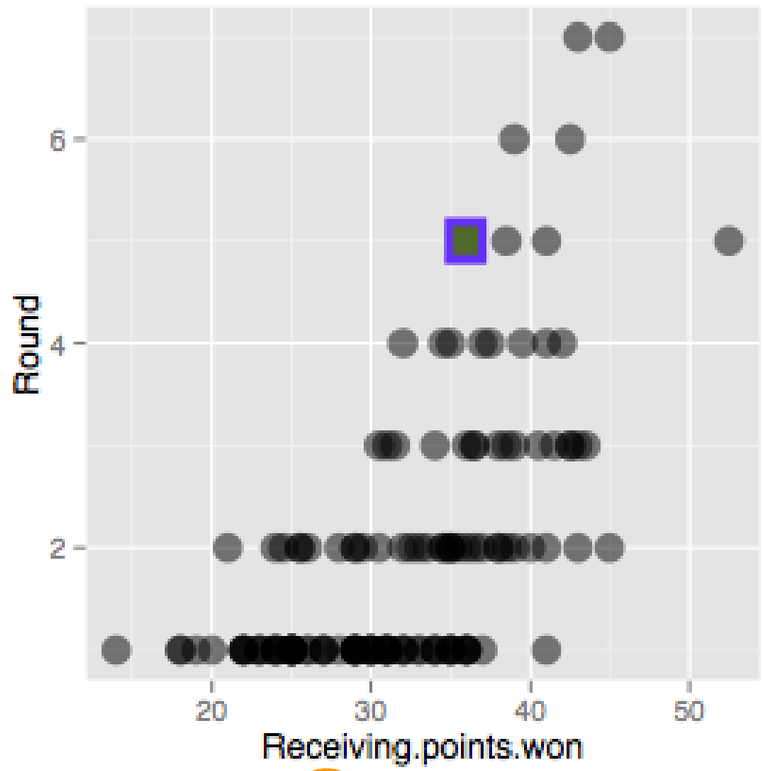
Why?

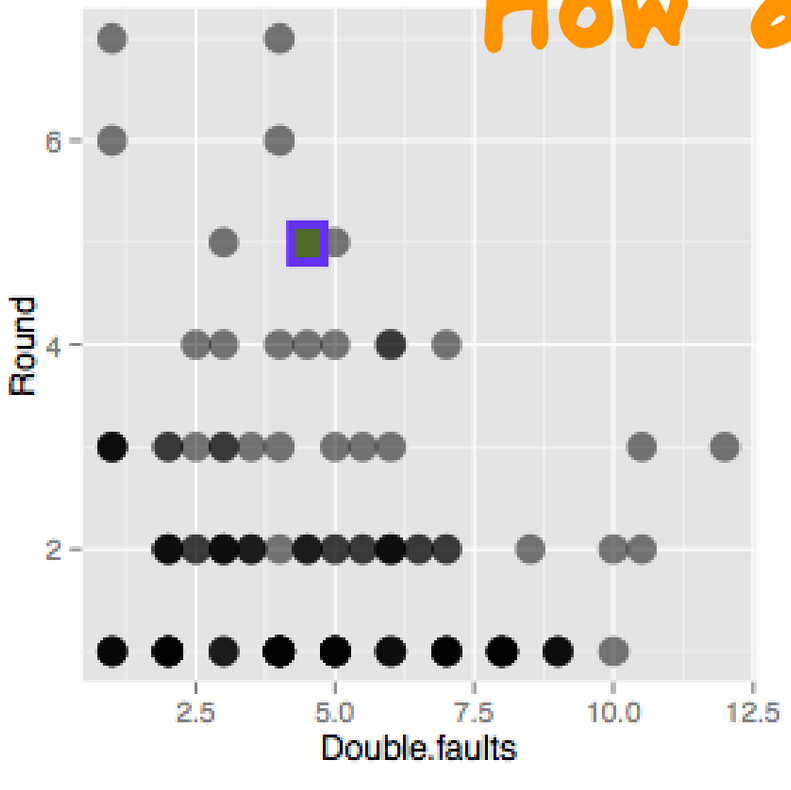
- It allows broader application of statistics
- Plots are more (can be) complex statistical summary
- Lineup procedure enables us to learn whether "what we think we see is really there"

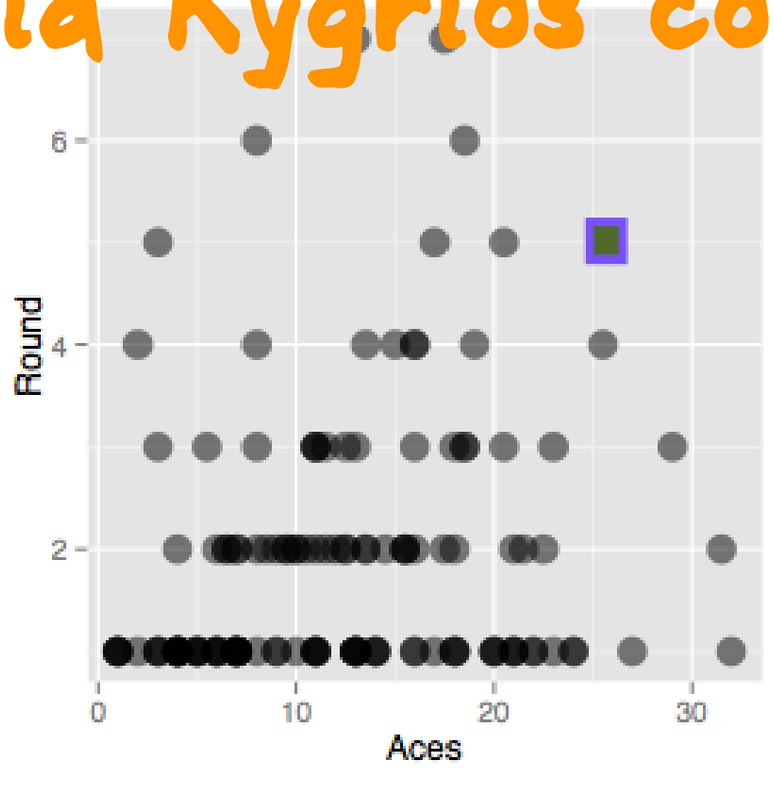
How did Kygrios compare?

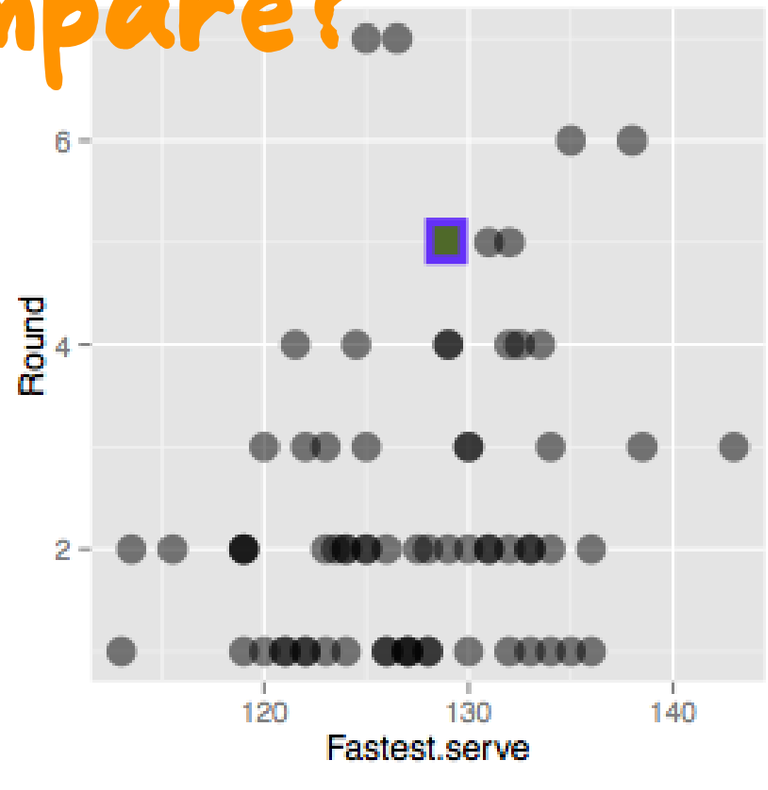












Soccer

- Teams in Iowa need to enter their stats each week with a web app, and these is made public

http://quikstatsiowa.com

- The motivation: to rank Iowa high school soccer teams in a way that is free of bias and to create a "fair" seeding algorithm to be used in Substate tournaments.

SOCCER

Season: 2016-17 2015-16 **2014-15** 2013-14 2012-13 2011-12 2010-11

Team Stats

Team: All Teams | Class: | Conference:

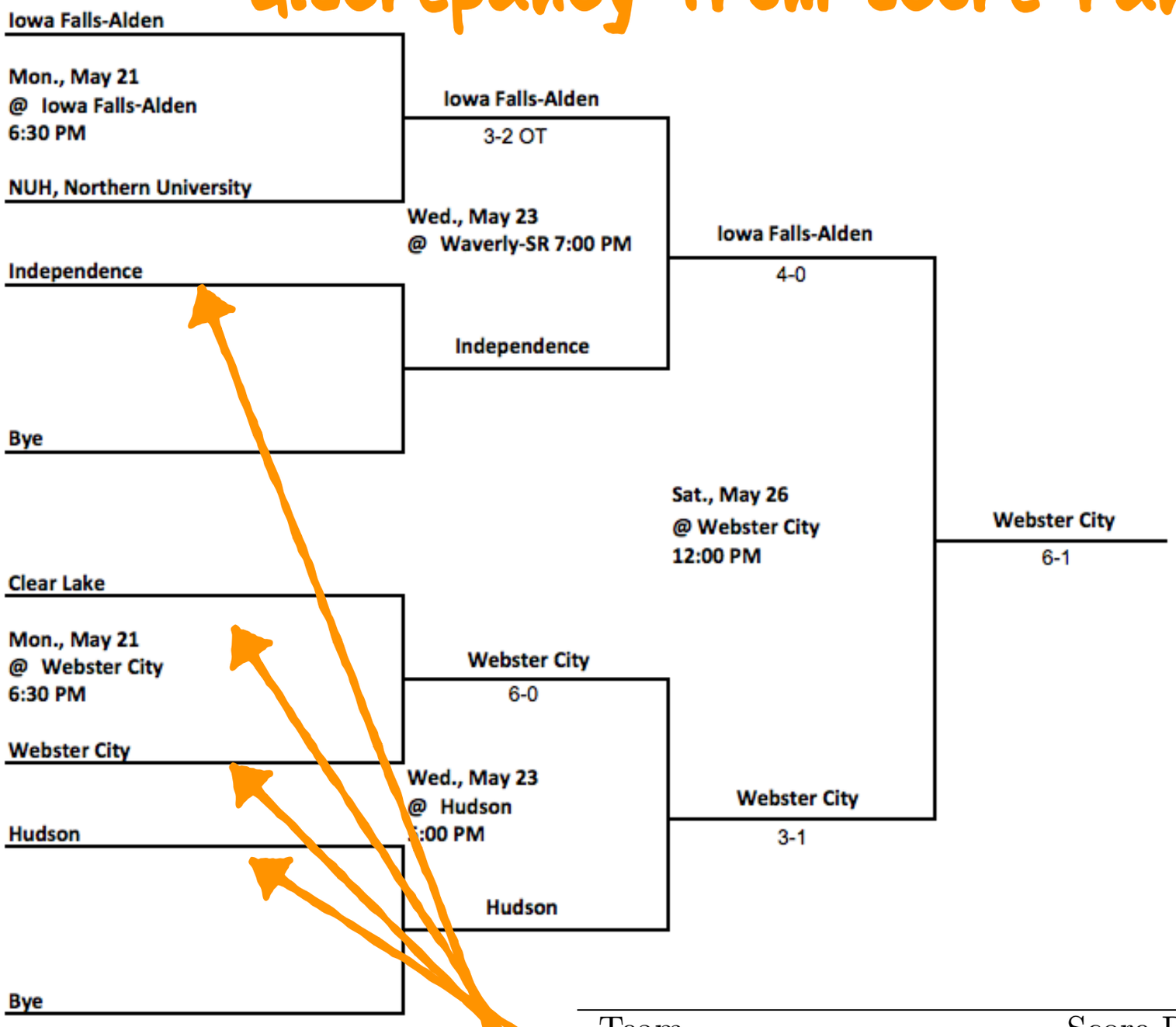
State Soccer Qualifiers: | Load List

Team	Record	GP	G	A	P	Sh	Sh %	SOG	SOG %	CK	PKM	PKA	GA	GA Avg	S	S %
Beckman, Dyersville	18 - 0 - 0	18	86	62	234	400	0.215	161	0.534	130	0	0	7	0.439	48	0.873
Waverly-Shell Rock	16 - 0 - 0	16	61	54	176	279	0.219	207	0.295	43	2	0	7	0.500	50	0.877
Johnston	21 - 1 - 0	22	79	53	211	529	0.149	265	0.298	118	4	4	10	0.476	108	0.915
Ankeny Centennial	20 - 1 - 0	21	88	71	247	0	0.000	0	0.000	118	5	5	14	0.685	31	0.689
Central Clinton, DeWitt	17 - 1 - 0	18	68	61	197	286	0.238	180	0.378	81	5	6	15	0.857	47	0.758
Bishop Heelan, Sioux City	15 - 1 - 0	16	49	38	136	236	0.208	89	0.551	72	6	10	9	0.500	42	0.824
Storm Lake	13 - 1 - 0	14	85	50	220	182	0.467	193	0.440	46	2	2	4	0.400	32	0.889
Sioux City, West	20 - 2 - 0	22	115	78	308	350	0.329	274	0.420	107	22	9	24	1.110	91	0.791
Xavier, Cedar Rapids	17 - 2 - 0	19	78	78	234	401	0.195	205	0.380	157	6	7	15	0.771	70	0.824
Danville	15 - 2 - 0	17	87	76	250	367	0.237	197	0.442	125	2	3	14	0.815	79	0.849
Underwood	14 - 2 - 0	16	75	69	219	470	0.160	295	0.254	60	10	12	17	1.126	167	0.908
Linn-Mar, Marion	20 - 3 - 0	23	80	83	243	466	0.172	237	0.338	116	3	4	17	0.790	36	0.679

Process

- During season, teams play a sample of other teams in the state, negotiated by their athletic director.
- Play can be across bracket (1,2,3)
- Partial rankings, convert to full rankings incorporating level and results on opponents, used Colley method
- Lots of DATA CLEANING and pre-processing, of web-scraped data!

A 2A substate draw with particularly large discrepancy from score ranking results



Team	Score Rank	Suggested Seed	Suggested Match
Hudson	26	1	ByeA
Webster City	40	2	ByeB
Clear Lake	110	3	B1
Iowa Falls-Alden	115	4	A1
Cedar Falls, Northern Univ	127	5	A2
Independence	139	6	B2

Iowa Soccer Rankings

Choose a Dataset

Boys 2015 Data

Download Data

Boys Data
Boys Substate Data
Girls Data

Links

quikstats
IAHSAA

App by: Danny Bero, Dr. Dianne Cook

Overall Ranking | Matchups | Team Statistics | 2015 Pre-Season | Substate Groups | Help

Show 10 entries Search:

Team	Divison	Record	Colley	Score	Colley Rank	Score Rank
MOC-Floyd Valley	1A	12-1	3.224940	5.841548	97	30
Beckman, Dyersville	1A	6-0	3.132916	5.330535	111	38
Danville	1A	9-0	2.641215	5.154636	142	42
Clear Lake	1A	9-3	3.849672	5.001997	51	43
Regina, Iowa City	1A	7-4	3.166786	4.707187	109	47
Des Moines Christian	1A	9-3	3.128105	4.416338	114	52
Gilbert	1A	12-3	3.170473	4.371476	107	53
Gladbrook-Reinbeck	1A	10-2	3.038049	3.979906	122	60
West Liberty	1A	7-6	3.129262	3.977112	113	61
Albia	1A	6-5	3.626371	3.943519	65	62
Team	1A	Record	Colley	Score	Colley Rank	Score Rank

Showing 1 to 10 of 53 entries (filtered from 149 total entries) Previous 1 2 3 4 5 6 Next

~/soccer-stats/rankings_app - Shiny

http://127.0.0.1:7795 | Open in Browser | Publish

Iowa Soccer Rankings

Choose a Dataset

Boys 2015 Data

Download Data

Boys Data
Boys Substate Data
Girls Data

Links

quikstats
IAHSAA

App by: Danny Bero, Dr. Dianne Cook

Overall Ranking | Matchups | Team Statistics | 2015 Pre-Season | Substate Groups | Help

Team 1

Gilbert

Team 2

Des Moines Christian

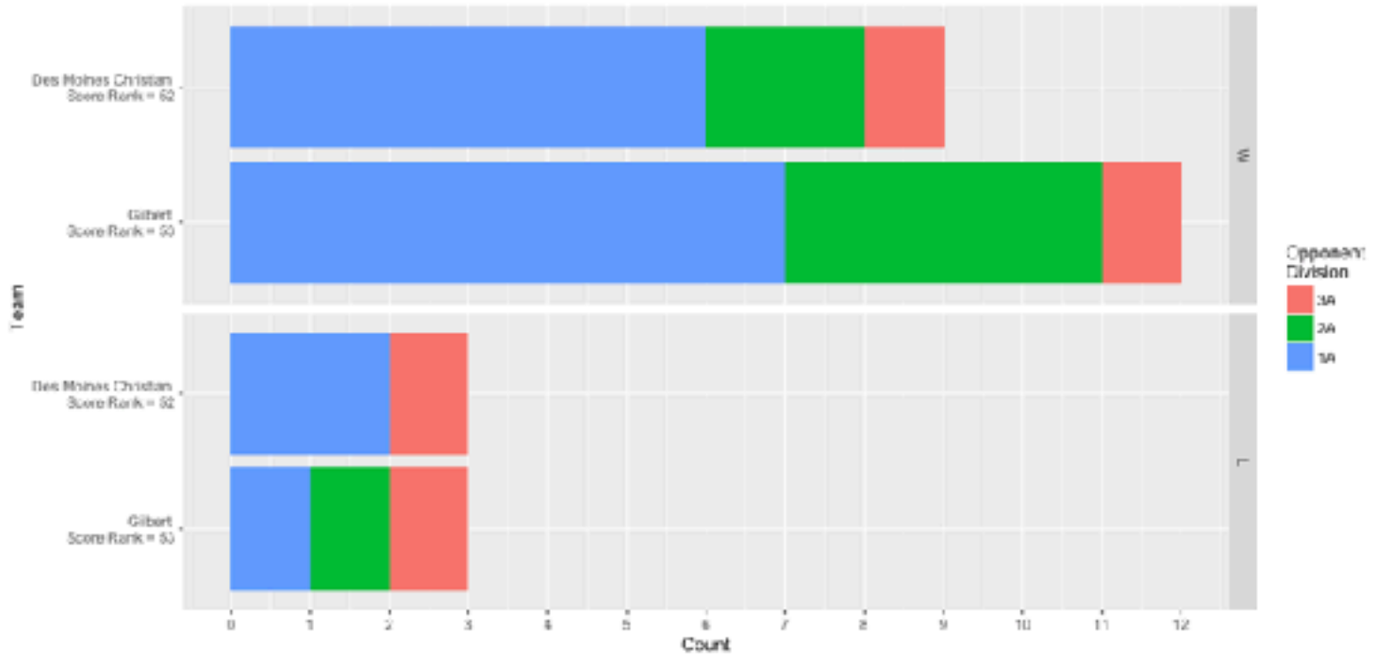
Matchup Predictions for Week 8, 2015

Predicted Score Differential	Win Probability
-0.04	0.49

Season Matchups

Week	Date	Team	Opponent	Score
5	2015-04-30	Gilbert	Des Moines Christian	1-0

Season Comparison



Division Comparison

~/soccer-stats/rankings_app - Shiny

http://127.0.0.1:7795 | Open in Browser | Publish

Iowa Soccer Rankings

Choose a Dataset

Boys 2015 Data

Download Data

Boys Data
Boys Substate Data
Girls Data

Links

quikstats
IAHSAA

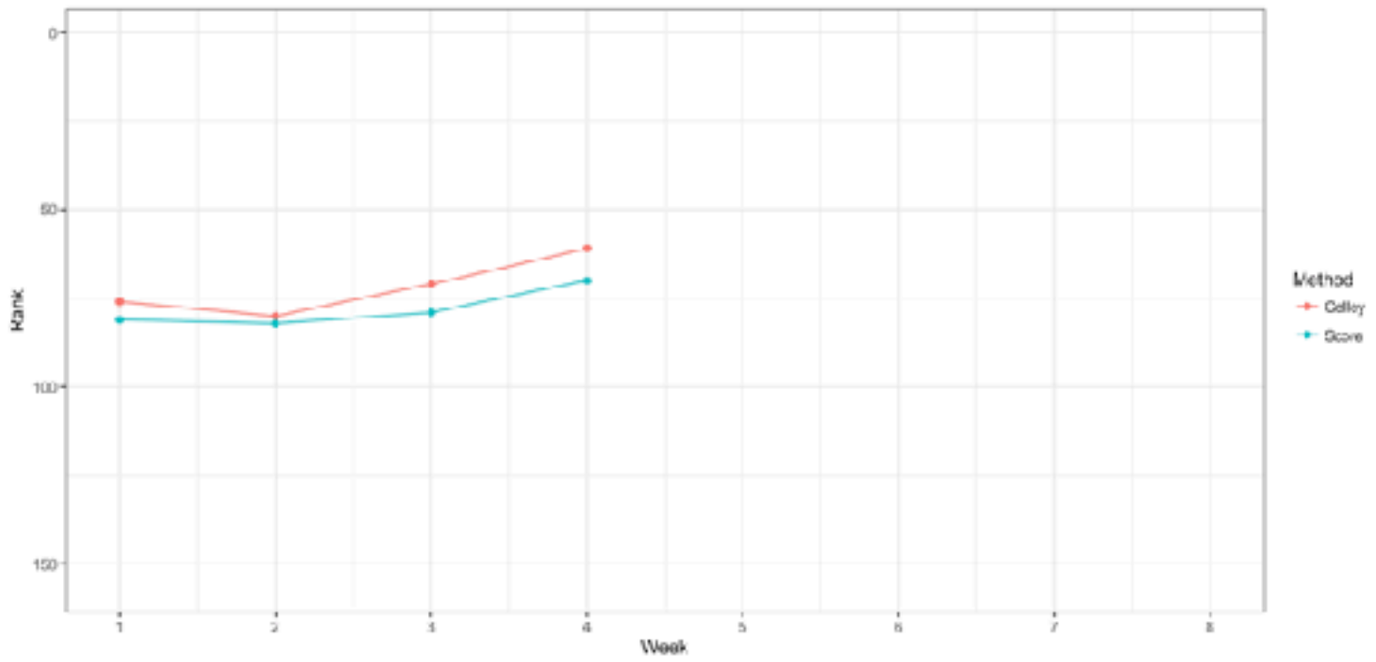
App by: Danny Bero, Dr. Dianne Cook

Overall Ranking | Matchups | Team Statistics | 2015 Pre-Season | Substate Groups | Help

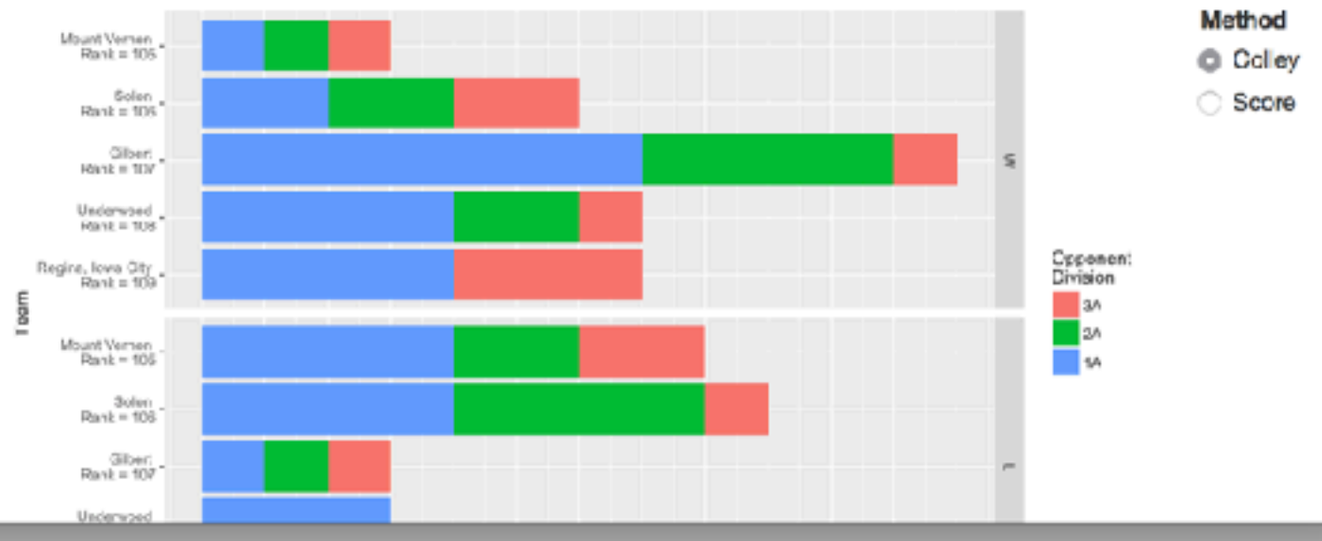
Select a Team

Gilbert

Weekly Rankings



Nearest Neighbors: Season Comparison Plot



Why?

- More objective ranking process
- Tool for coaches and players to see how their team is performing relative to other teams
- Data can be used in statistics classes

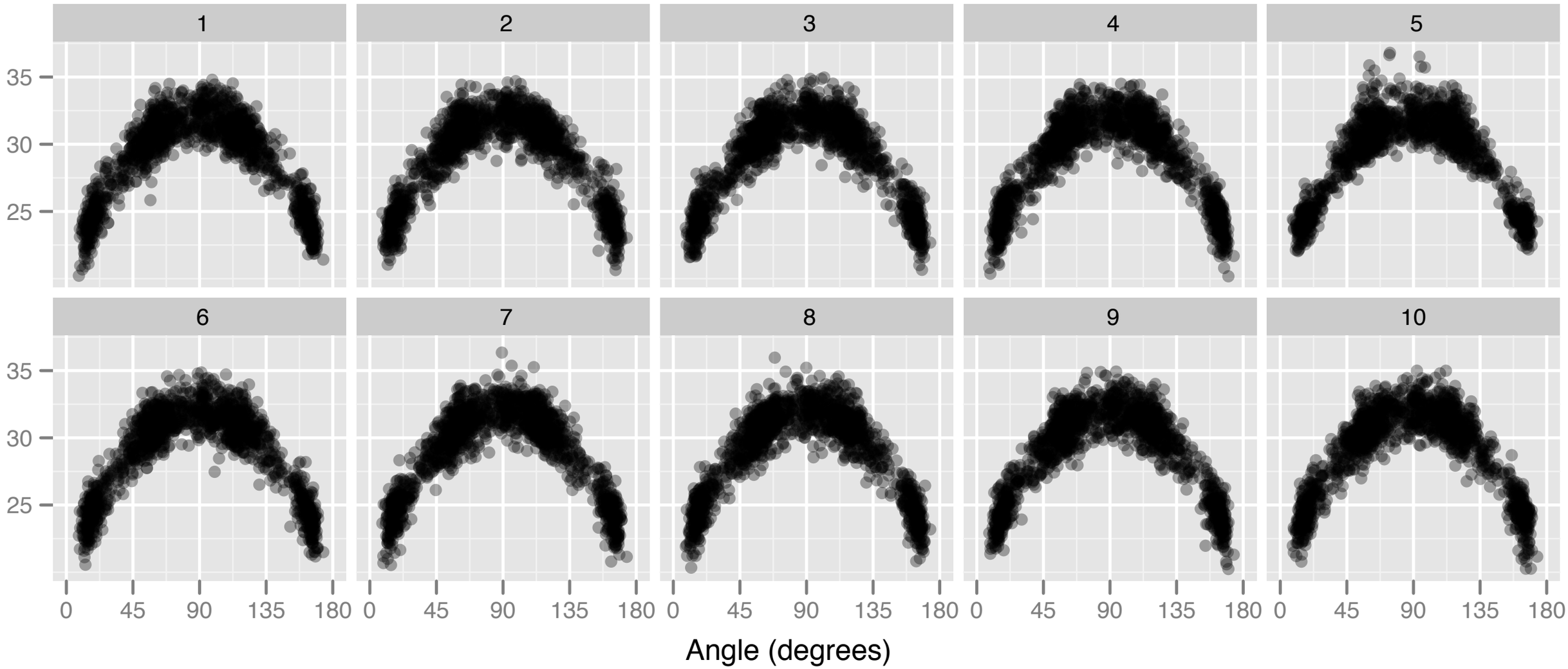
Key questions

- Beyond traditional statistical inference, supporting data exploration
- (Public) databases, gender equity
- Better visualisation of sports data, incorporating uncertainty, describing usual patterns
- Engaging students in statistics
- All analysis is done with open data and open source software

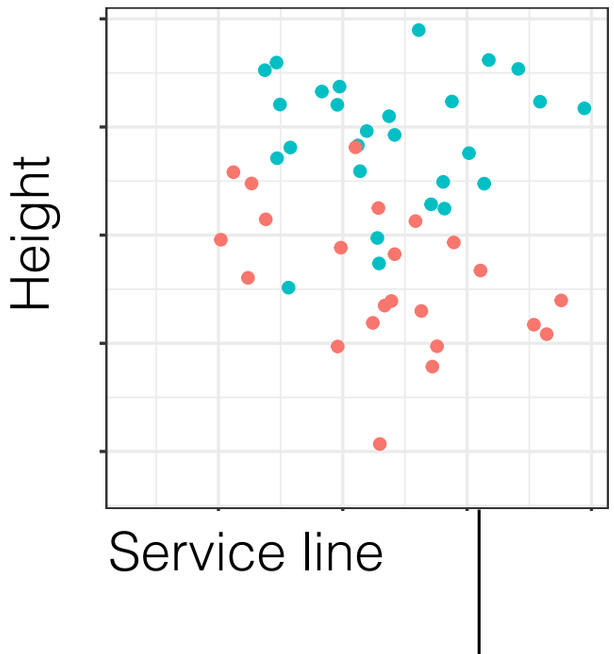
Tennis

- Research with Tennis Australia, Dr Stephanie Kovalchik
- Monash students: Steph Kobakian, Alwin Wang, Braden Churcher, Madeline Jom
- Visualisation of tennis serve
- Classifying different serve styles
- Face and emotion detection

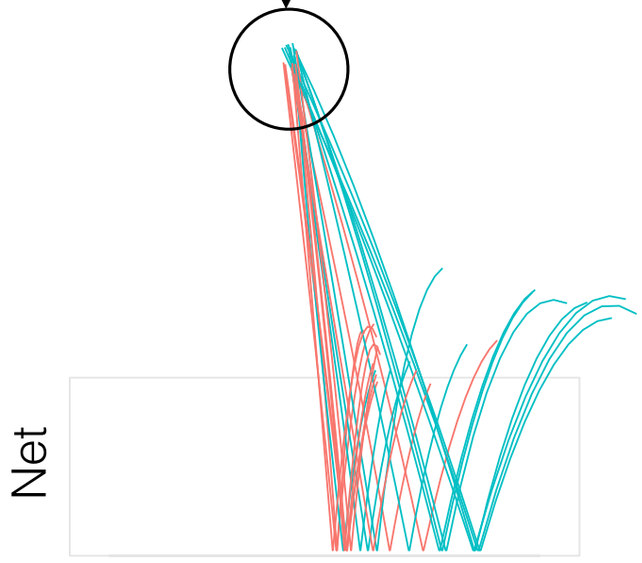
Basketball



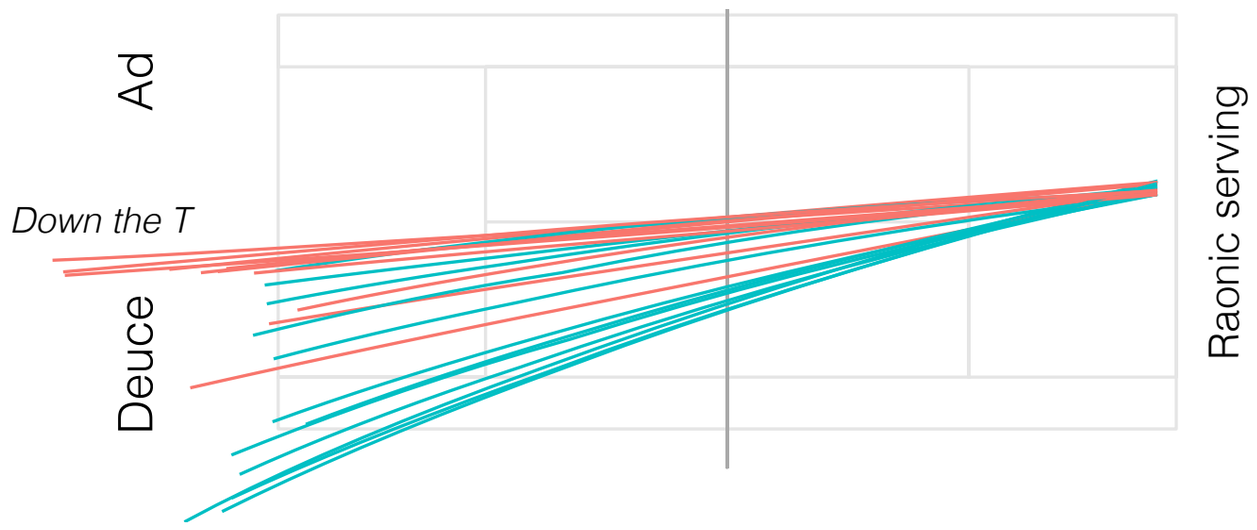
Raonic's hit position



Fifty deuce side serves from Raonic at Australian Open 2016. A different toss height tends to produce a different serve. The down the T serves (red) are hit from a lower height.

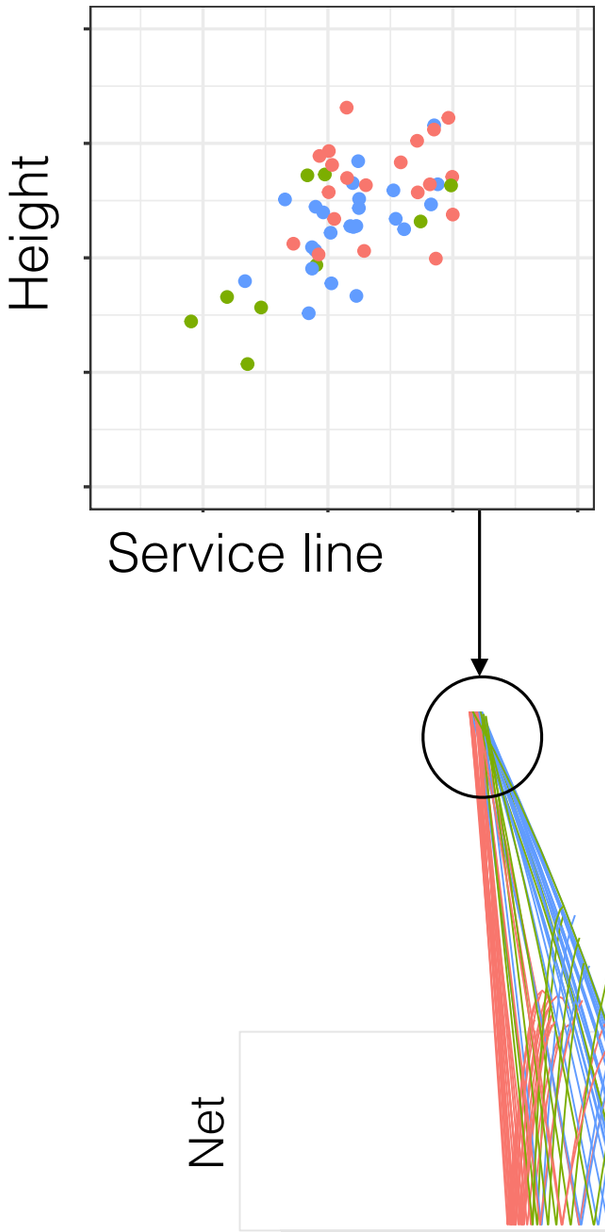


Receiver's view



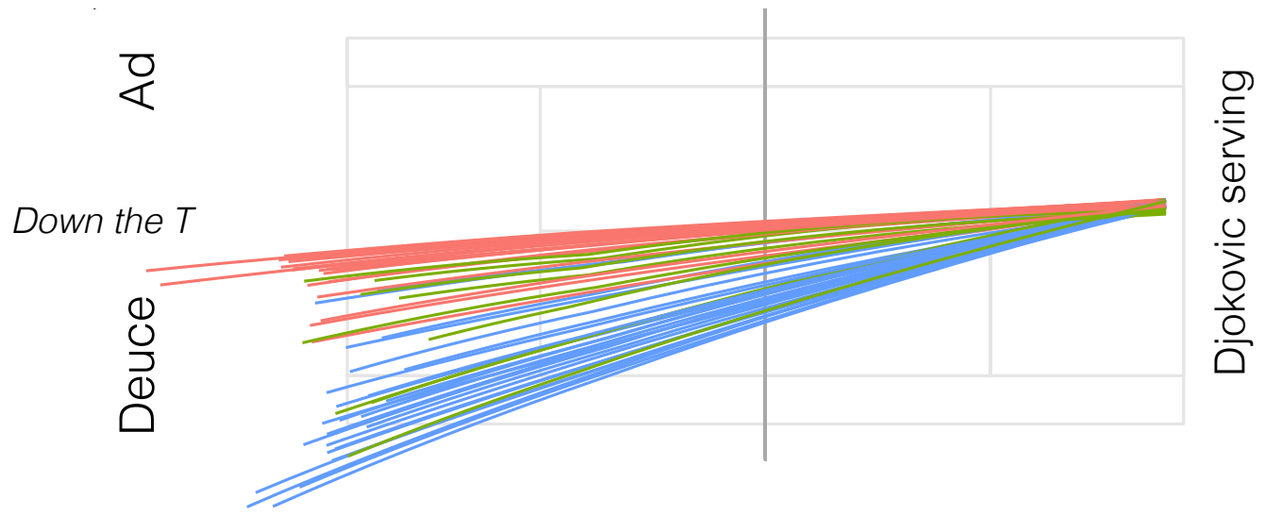
Bird's eye view of the court

Djokovic's hit position



Fifty-three deuce side serves from Djokovic at Australian Open 2016. Almost identical hit position leads to different type of serve.

Receiver's view



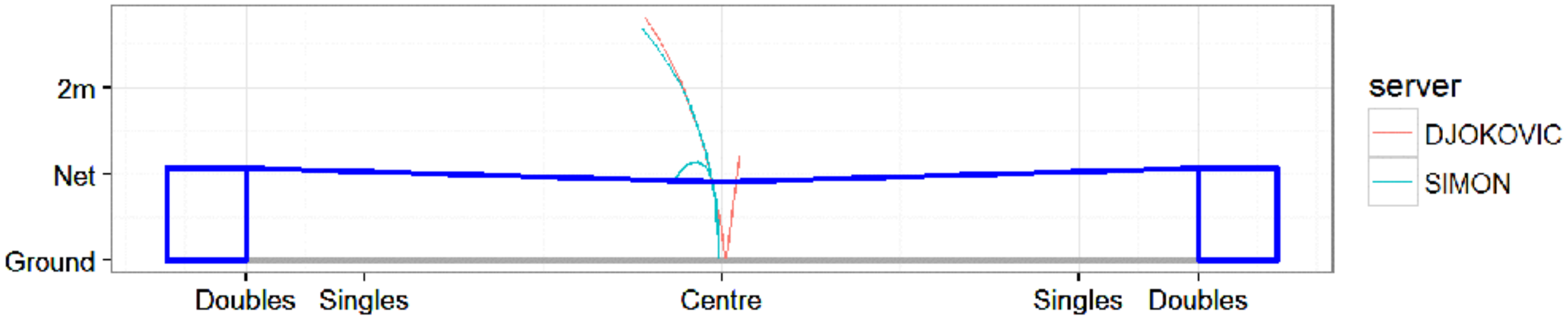
Bird's eye view of the court

Computing spin

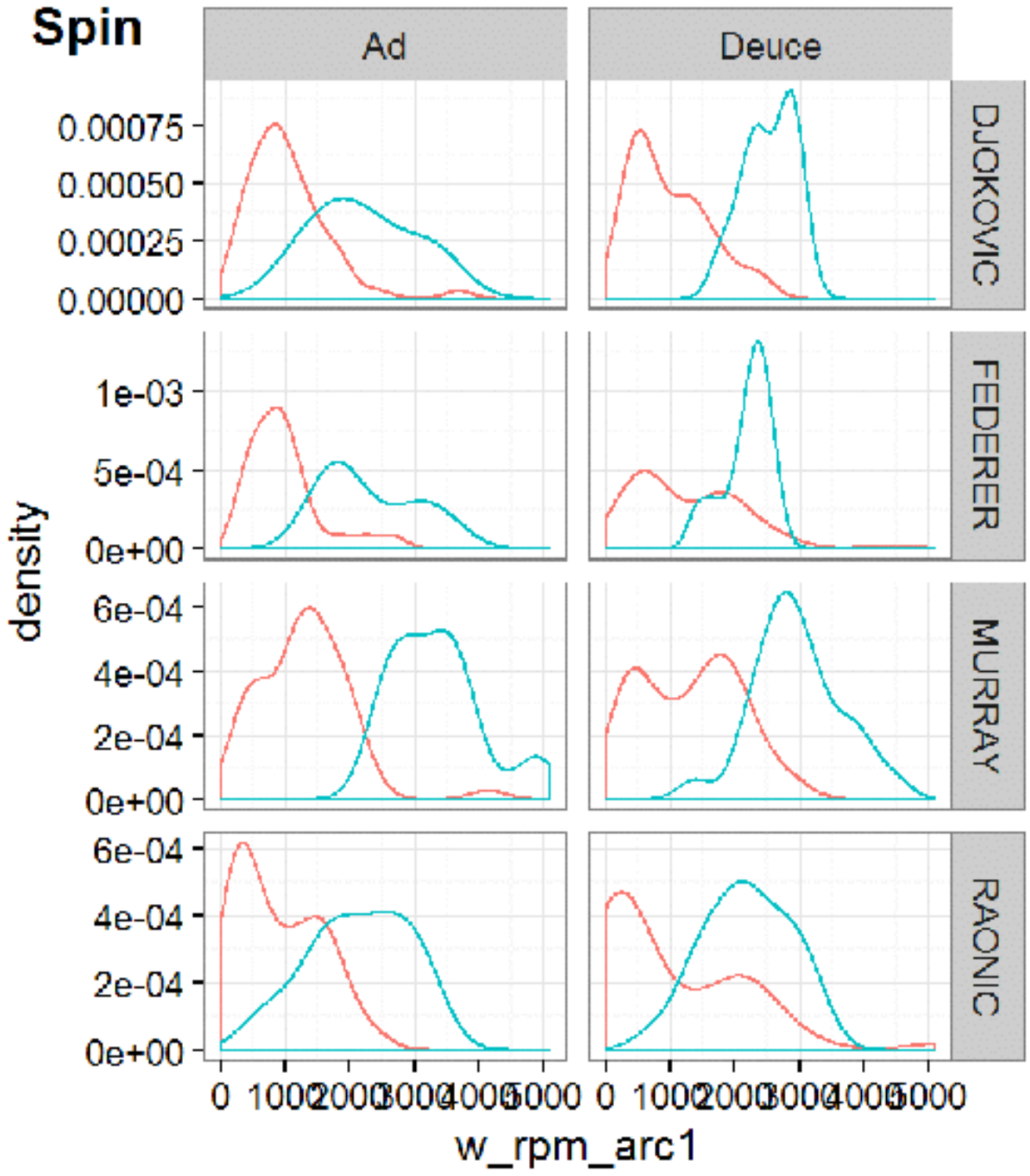
Prediction: $$\vec{a}_{\mathrm{p}} = \frac{\vec{F}_{\mathrm{net}}}{m} = \frac{\vec{L}}{m} + \frac{\vec{D}}{m} + \vec{g}$$

Lift Force: $$\vec{L} = |L| \cdot \frac{\vec{W} \times \vec{V}}{\left|\vec{W} \times \vec{V}\right|} \qquad |L| = A\frac{\rho}{2}V^2\left(\frac{1}{2 + \frac{V}{V_s}}\right)$$

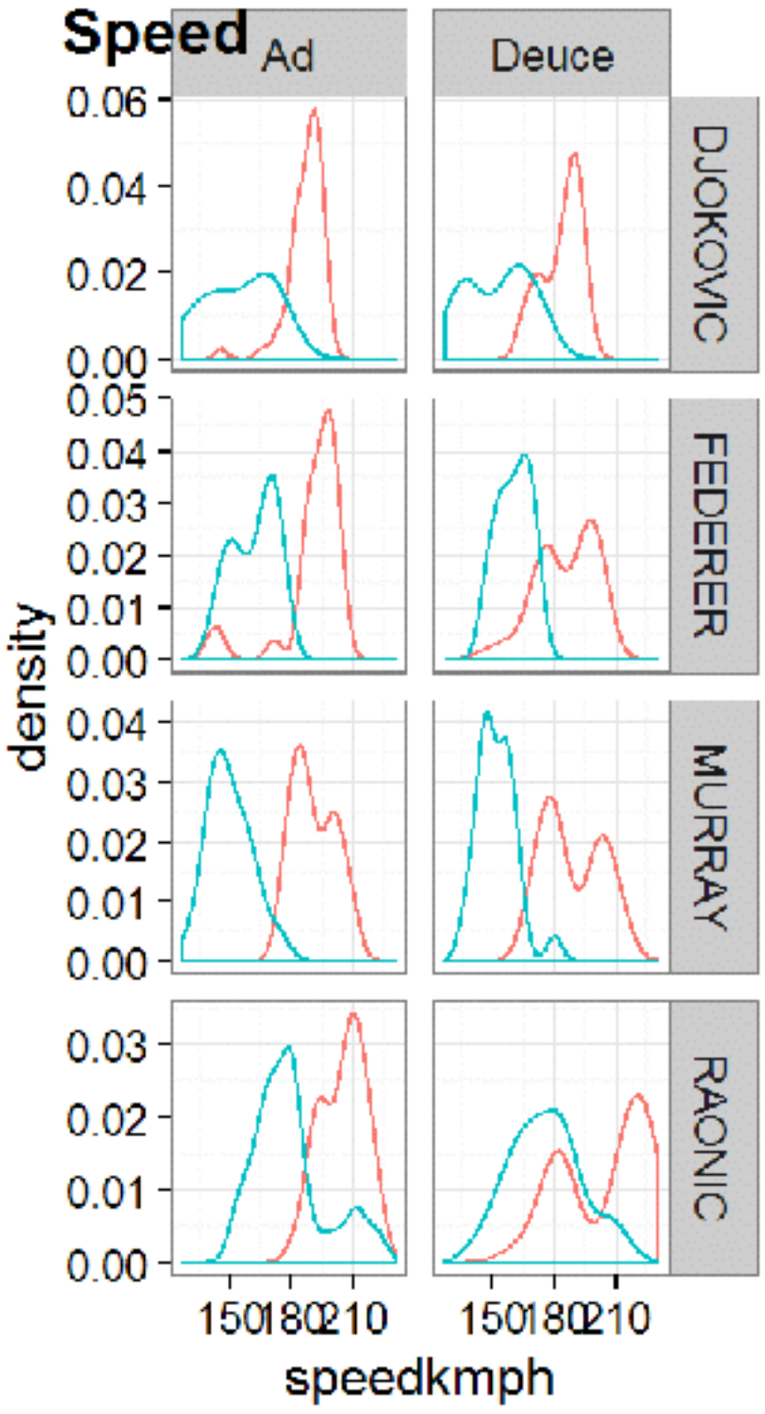
Drag Force: $$\vec{D} = -|D| \cdot \frac{\vec{V}}{|V|} \qquad |D| = A\frac{\rho}{2}V^2\left(0.55 + \frac{1}{\left(22.5 + 4.5\left(\frac{V}{V_S}\right)^{2.5}\right)^{0.4}}\right)$$

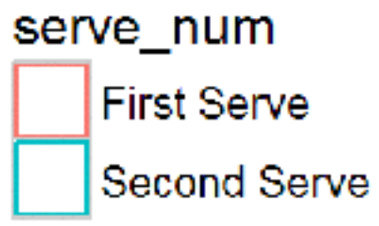


Spin

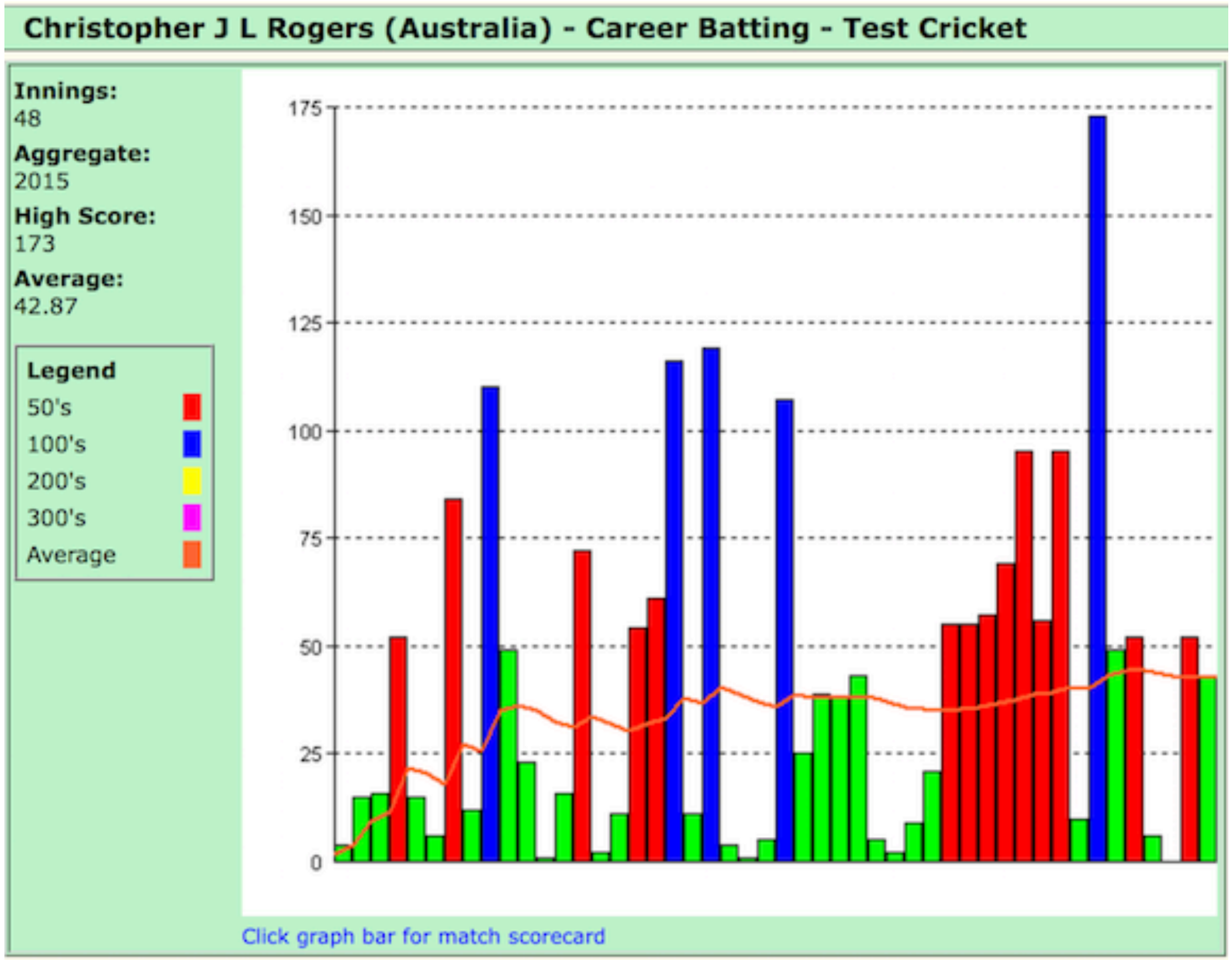

Ad
Deuce
DJOKOVIC
FEDERER
MURRAY
RAONIC
density
0.00075
0.00050
0.00025
0.00000
1e-03
5e-04
0e+00
6e-04
4e-04
2e-04
0e+00
0 1000 2000 3000 4000 5000
w_rpm_arc1

Speed


Ad
Deuce
DJOKOVIC
FEDERER
MURRAY
RAONIC
density
0.06
0.04
0.02
0.00
0.05
0.04
0.03
0.02
0.01
0.00
150 180 210
speedkmph


serve_num
First Serve
Second Serve

Cricket plots

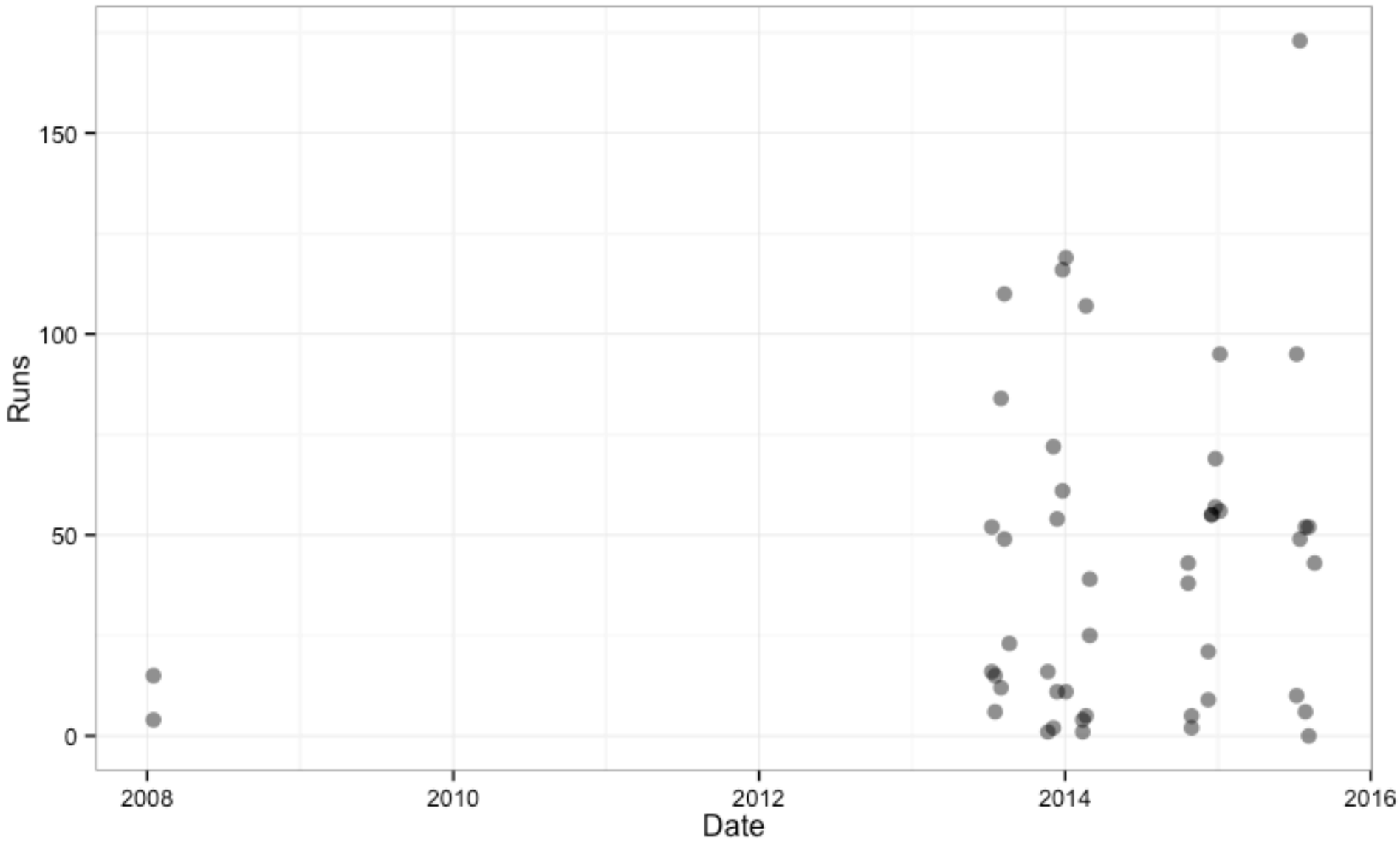


Analysis

- What is the purpose?
  - ✓ Is there a temporal trend in his scores?
  - ✓ Do he suffer from slumps and highs?
- Colour selection unfortunate
- Bar charts take space away from variance perception - 0's lost as focus is on the highs
- Time is not real time

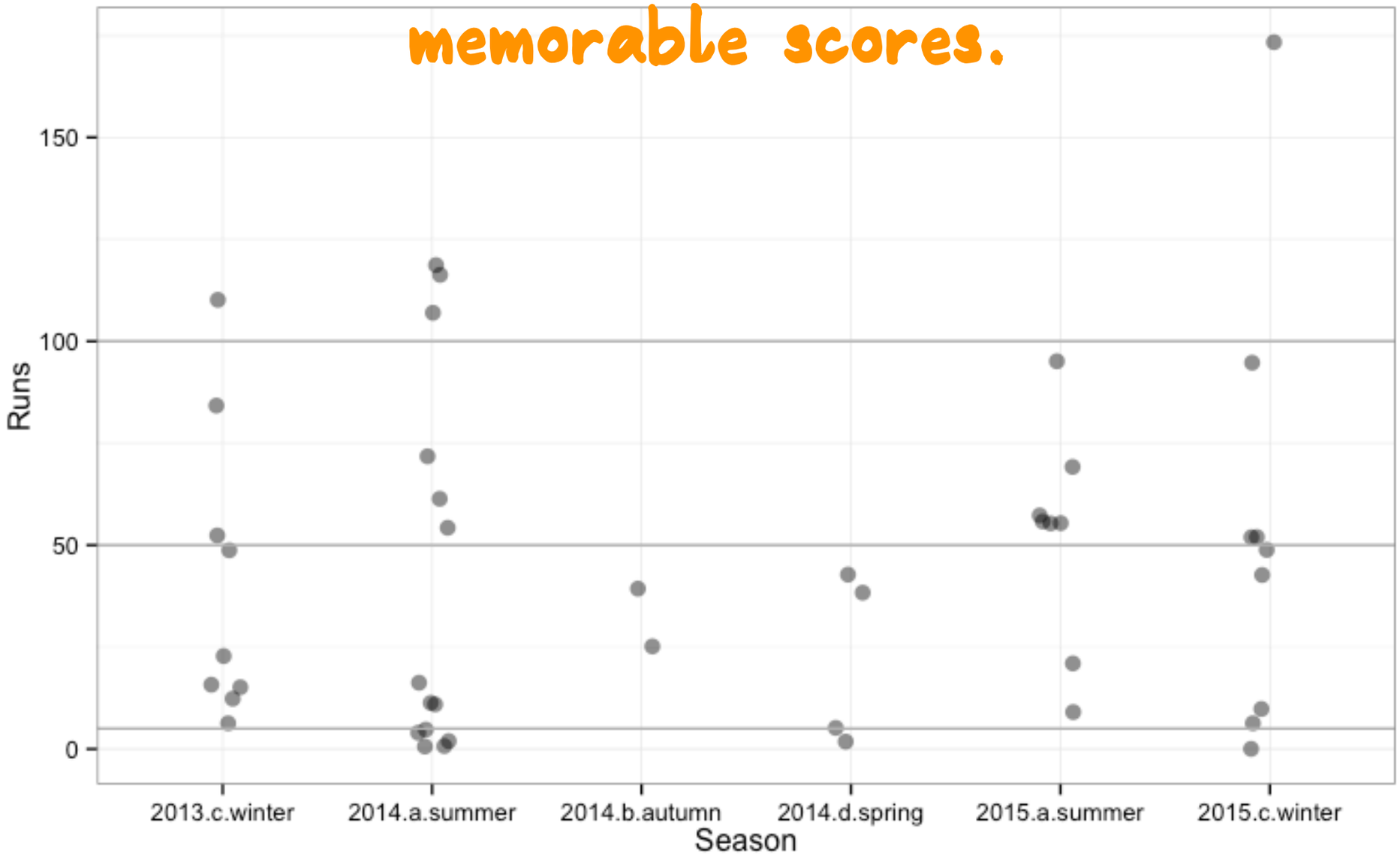
Temporal trend

His first scores were back in 2008, then none until 2013



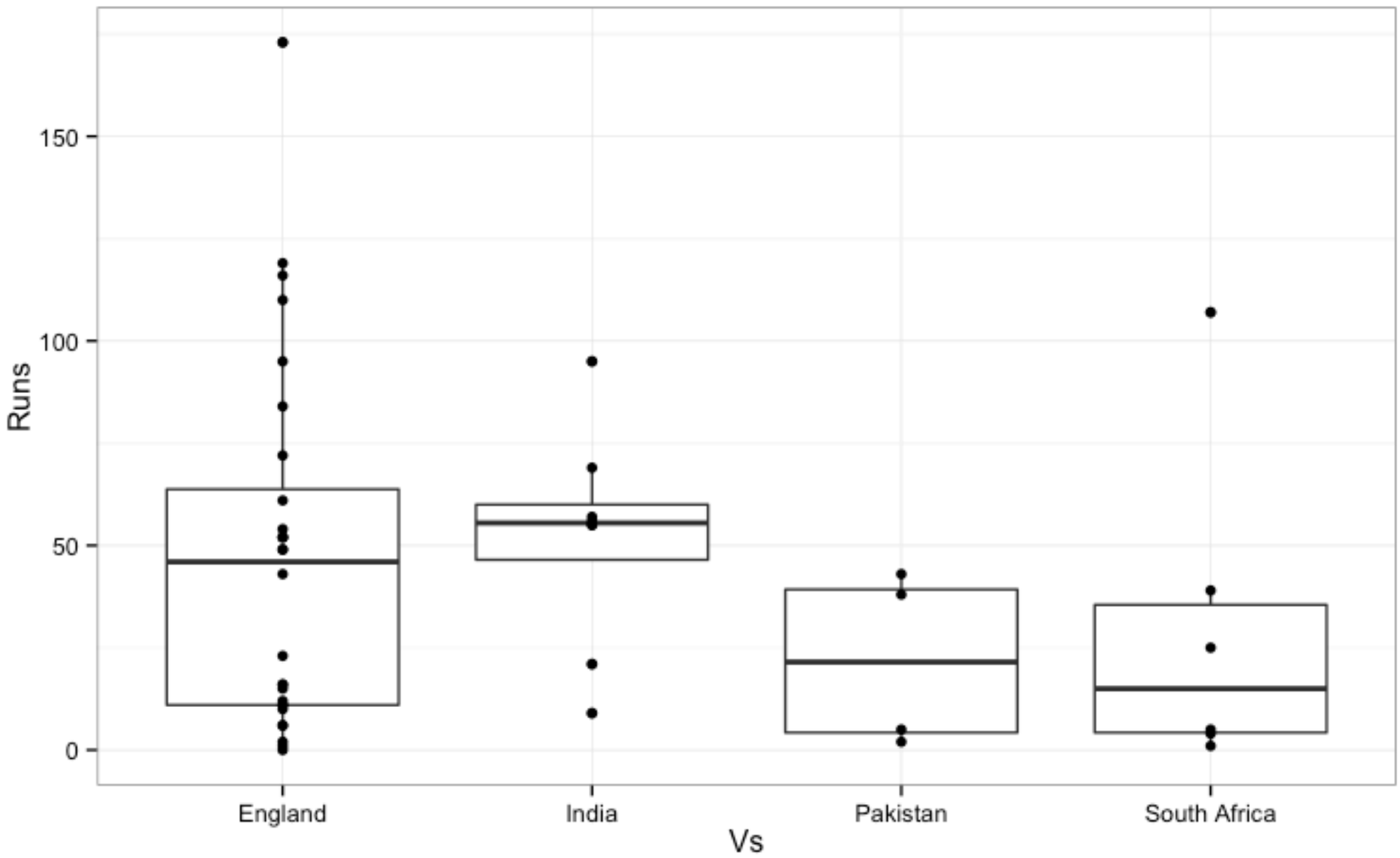
Temporal trend

Focus on the >2013 scores. Make a new variable season, to reflect cricket time. Guide lines for memorable scores.



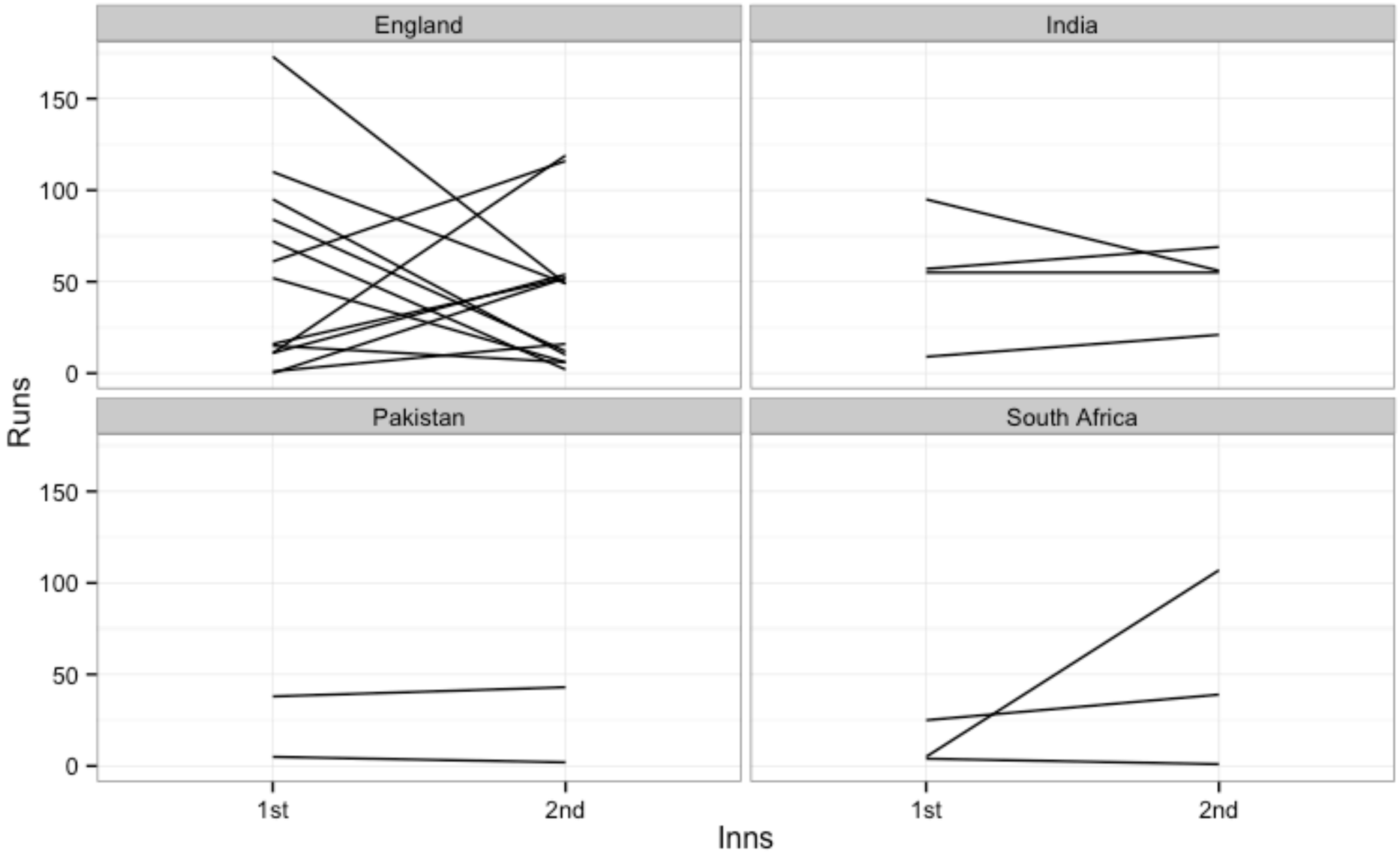
Performance against different foes

Side-by-side boxplots by opponent, plus dotplots. Top scores were mostly against England

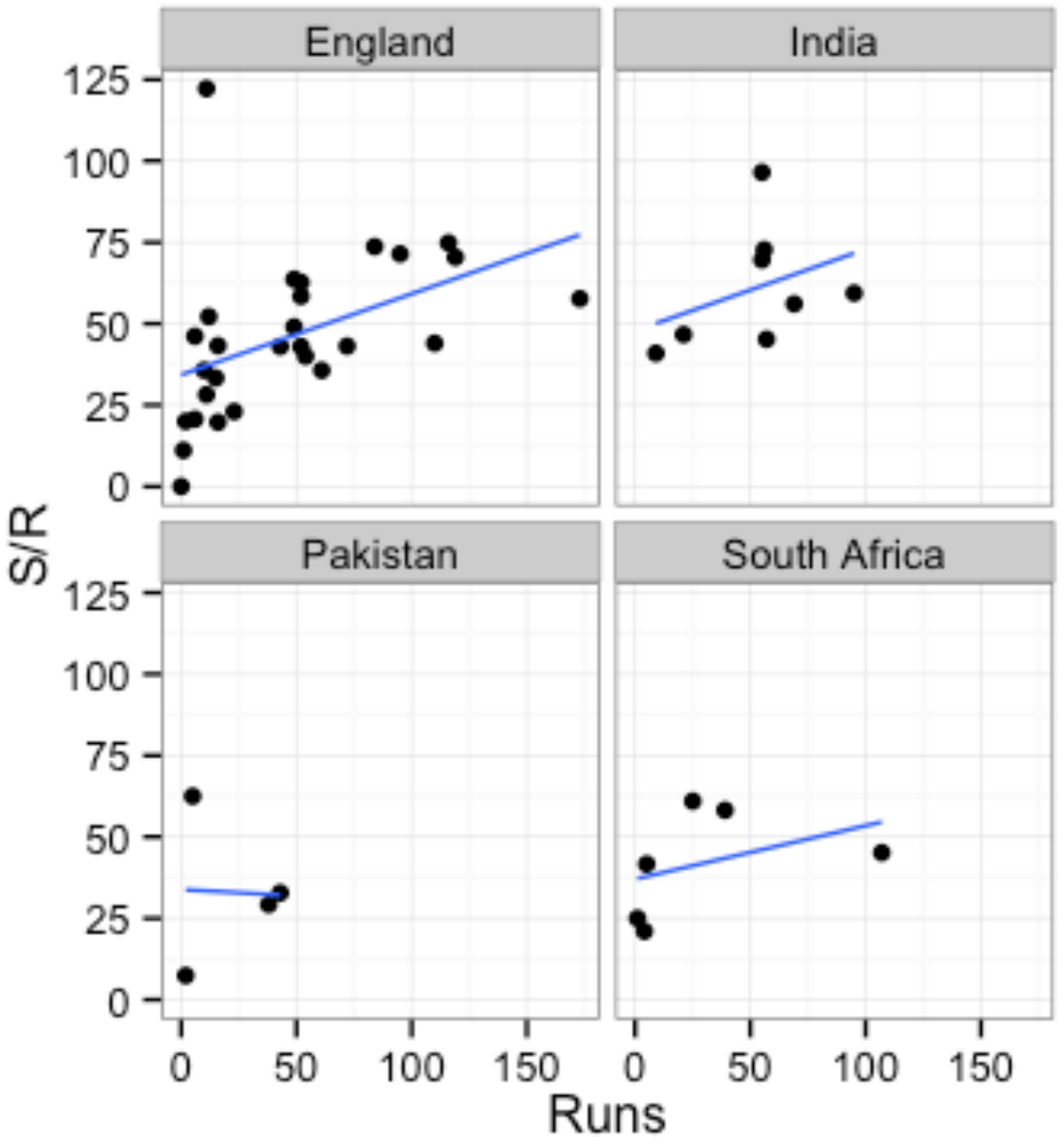


First vs second innings

Lines not level. Tendency may be to be "good"-"bad" or "bad"-"good"



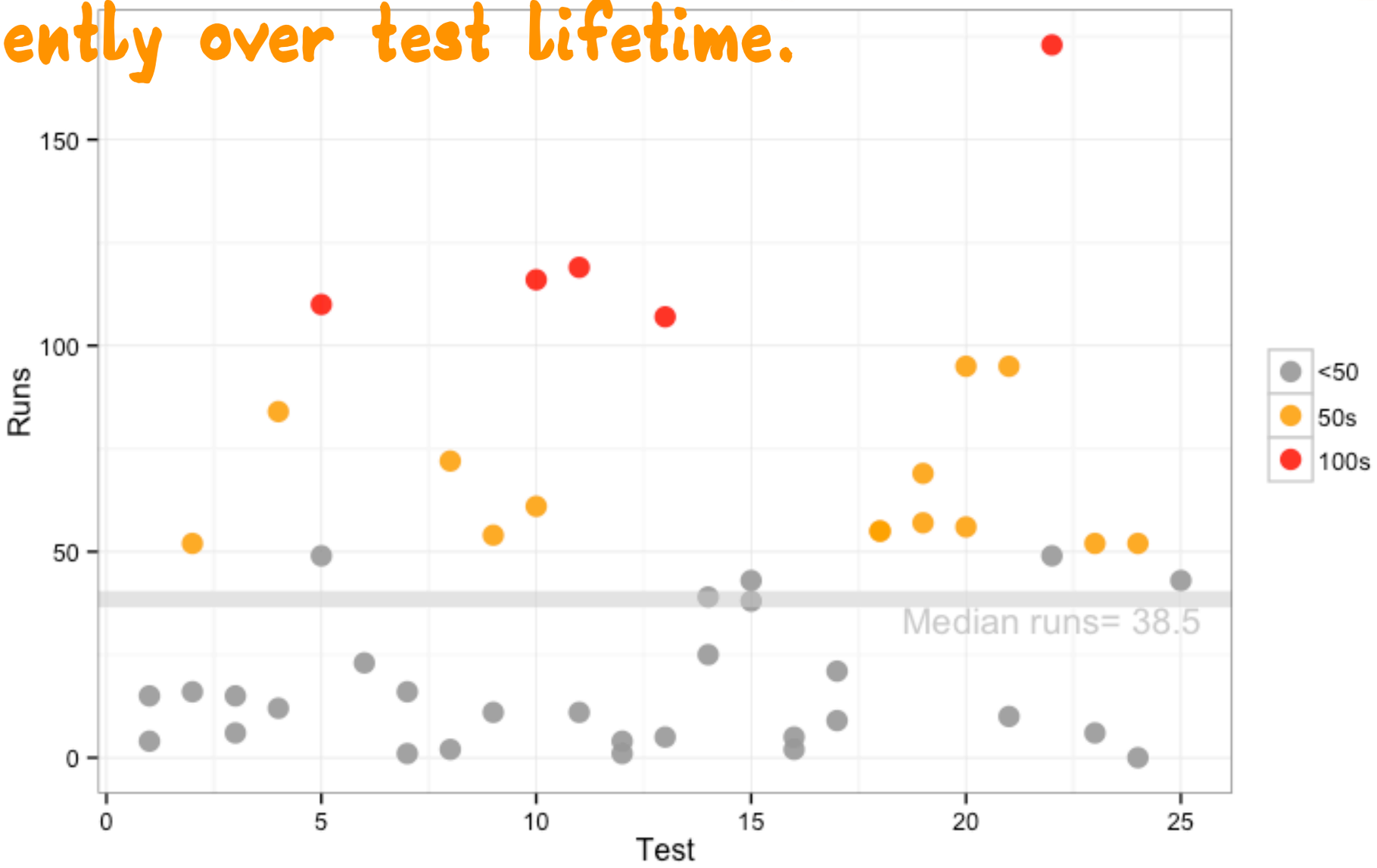
Run rate and scores



Strike rate by runs.
Strike rate higher when he scored more runs.

First vs second innings

Test used as time variable. Guideline showing his median score. Benchmark totals (50, 100) mapped to sequential colour scale. Pretty variable scoring, consistently over test lifetime.



Workflow tools

- R Notebook: Code and document together, hooks to data
- github: Collaborative research, analysis and writing
- slack: Group organisation