

# Mining the Play By Play

Alexandr Diaz-Papkovich  
Carleton University

[alexdiuz.ca](http://alexdiuz.ca)

@adp\_diaz

# Introduction

- Play By Play (“RTSS”) is now commonly used
- Can be used to evaluate teams or players
  - Very rich dataset, when accessed
- Well-established as a core resource by now
- Doesn't record penalty expirations or shift changes

# Implicit data

- Shifts, penalty expirations are not recorded

85	1	SH	12:00 8:00	FAC	TOR won Neu. Zone - TOR #11 MCCLEMENT vs MTL #48 BRIERE	11 37 4 51 34 C R D D G	48 51 67 76 79 31 C C L D D G
86	1	EV	12:17 7:43	HIT	TOR #28 ORR HIT MTL #67 PACIORETTY, Off. Zone	11 28 37 4 51 34 C R R D D G	48 51 67 76 79 31 C C L D D G
87	1	EV	12:43 7:17	MISS	MTL #15 PARROS, Wrist, Wide of Net, Off. Zone, 24 ft.	11 28 37 4 51 34 C R R D D G	15 8 32 24 55 31 R L L D D G

- Solution: Add them ourselves!
- Use RTSS penalty data and player TOI tables

```

20132014, 0001, TOR, MTL, 1, 85, FAC, SH, 12:00, 1, 5, 1, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0
20132014, 0001, TOR, MTL, 1, 86, SHFT, SH, 12:00, 1, 5, 1, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0
20132014, 0001, TOR, MTL, 1, 86, PEXP, EV, 12:03, 1, 6, 1, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0
20132014, 0001, TOR, MTL, 1, 86, SHFT, EV, 12:10, 1, 6, 1, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0
20132014, 0001, TOR, MTL, 1, 86, HIT, EV, 12:17, 1, 6, 1, 6, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0
20132014, 0001, TOR, MTL, 1, 87, MISS, EV, 12:43, 1, 6, 1, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0

```

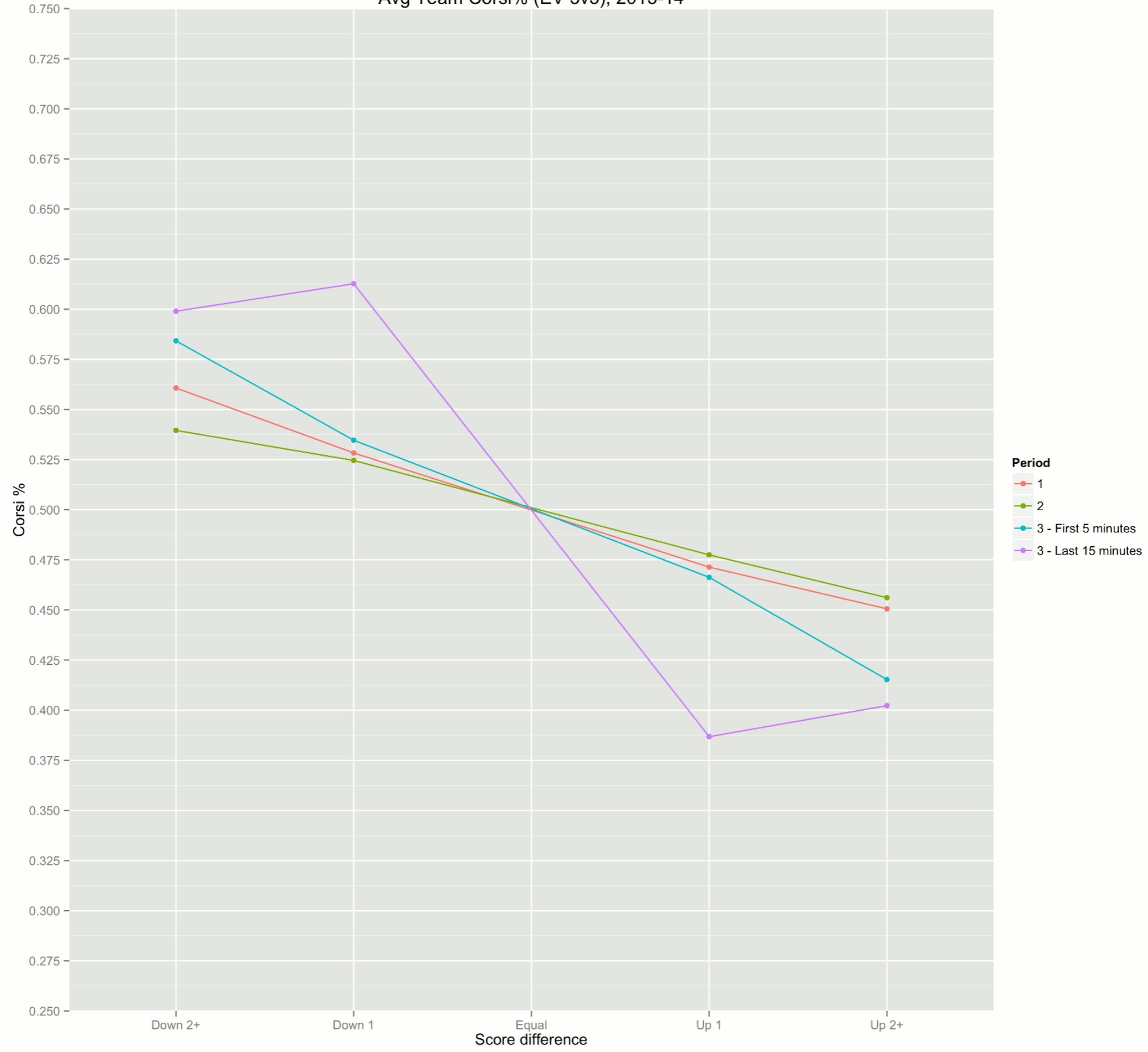
- Result: To-the-second accuracy for strength, score differential, and player presence

# Can we always get it right?

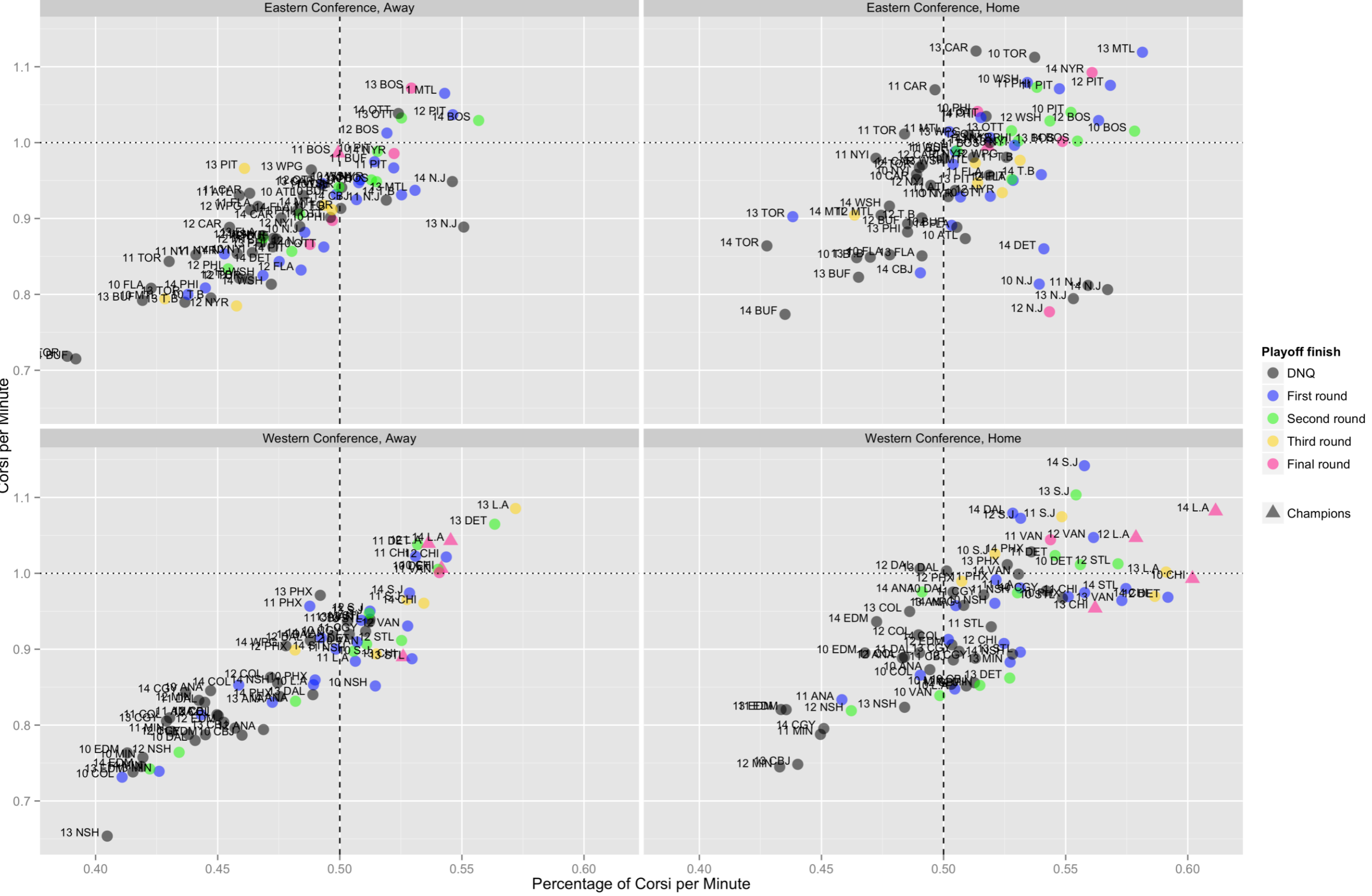
- Penalty info alone won't work
- Ambiguous cases exist

226	3	EV	2:48 17:12	PENL	PHI #37 ROSEHILL Misconduct (10 min)(10 min), Neu. Zone	17 18 82 51 77 25 C C R D D G	25 36 28 5 8 35 C C R D D G
227	3	EV	2:48 17:12	PENL	PHI #37 ROSEHILL Roughing(2 min) Served By: #19 HARTNELL, Neu. Zone Drawn By: FLA #82 KOPECKY	17 18 82 51 77 25 C C R D D G	25 36 28 5 8 35 C C R D D G
228	3	EV	2:48 17:12	PENL	PHI #37 ROSEHILL Roughing(2 min) Served By: #19 HARTNELL, Off. Zone Drawn By: FLA #19 UPSHALL	17 18 82 51 77 25 C C R D D G	25 36 28 5 8 35 C C R D D G
229	3	EV	2:48 17:12	PENL	FLA #82 KOPECKY Roughing(2 min), Neu. Zone Drawn By: PHI #36 RINALDO	17 18 82 51 77 25 C C R D D G	25 28 5 8 35 C R D D G
230	3	PP	2:48 17:12	FAC	FLA won Off. Zone - FLA #16 BARKOV vs PHI #28 GIROUX	17 18 51 77 25 C C D D G	25 28 5 8 35 C R D D G
231	3	SH	2:48 17:12	PENL	PHI #36 RINALDO Roughing(2 min), Neu. Zone Drawn By: FLA #82 KOPECKY	17 18 51 77 25 C C D D G	25 28 5 8 35 C R D D G

Avg Team Corsi% (EV 5v5), 2013-14



# Corsi Per Minute vs Corsi/minute Percentage (EV, 5v5, Tied)



# Association rule learning

- Data mining technique also known as market basket analysis
  - Cash registers record the items people purchase
  - Gigantic binary database
  - e.g. {BREAD, MILK} is a common combo
  - We create rules like {BREAD} => {MILK} to find common combinations

# Interest measures

- WARNING: Math.
- $\text{SUPP}(X)$  = % of transactions with itemset  $X$   
=  $\text{Pr}(X)$  = Probability  $X$  happens
- $\text{CONF}(X \Rightarrow Y)$  =  $\text{SUPP}(X, Y) / \text{SUPP}(X)$   
=  $\text{Pr}(Y|X)$  = Probability  $Y$  happens, Given  $X$



# More interest measures

- $LIFT(X \Rightarrow Y) = SUPP(X, Y) / [SUPP(X) * SUPP(Y)]$ 
  - Measure of independence
  - Farther from 1 means less independent
- $DOC(X \Rightarrow Y) = CONF(X \Rightarrow Y) - CONF(\neg X \Rightarrow Y)$   
 $= Pr(Y|X) - Pr(Y|\neg X)$   
 $= (\text{Probability } Y \text{ happens given } X \text{ has happened}) -$   
 $(\text{Probability } Y \text{ happens given } X \text{ has not happened})$

# Example transactions

BREAD	MILK	BEER	DIAPERS
1	1	0	0
1	0	1	1
1	1	0	1
0	0	1	1

- $\text{SUPP}(\text{BREAD}, \text{MILK}) = 0.5$
- $\text{CONF}(\text{DIAPERS} \Rightarrow \text{BEER}) = 0.5 / 0.75 = 0.667$
- $\text{LIFT}(\text{BEER} \Rightarrow \text{DIAPERS}) = 0.5 / (0.75 * 0.5) = 0.75$

# Example: 2013-14 Leafs

COLTON_ORR	JOHN_MICHAEL_LILES	JAMES_VAN_RIEMSDYK	SHOT_FOR	SHOT_AGAINST	MISS_FOR	MISS_AGAINST	BLOCK_FOR	BLOCK_AGAINST
0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0	0
0	0	1	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0

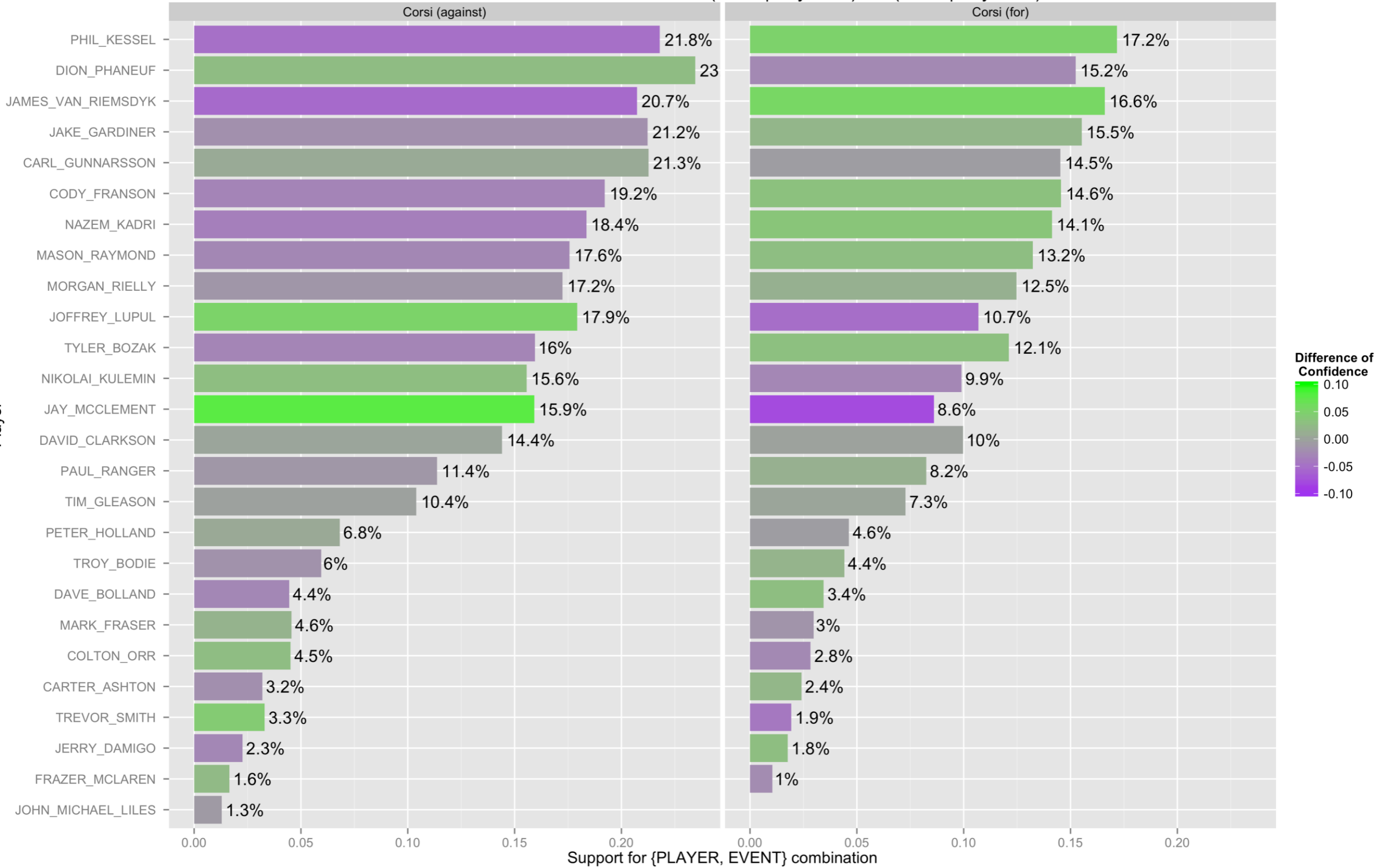
- Created a database of “transactions”
- Players on the left, events on the right
- For this example, only looking at Corsi events (EV, 5v5, tied)
- Allows us to see finer details, pick apart lines, gauge responsibility for events, etc

# Example: 2013-14 Leafs

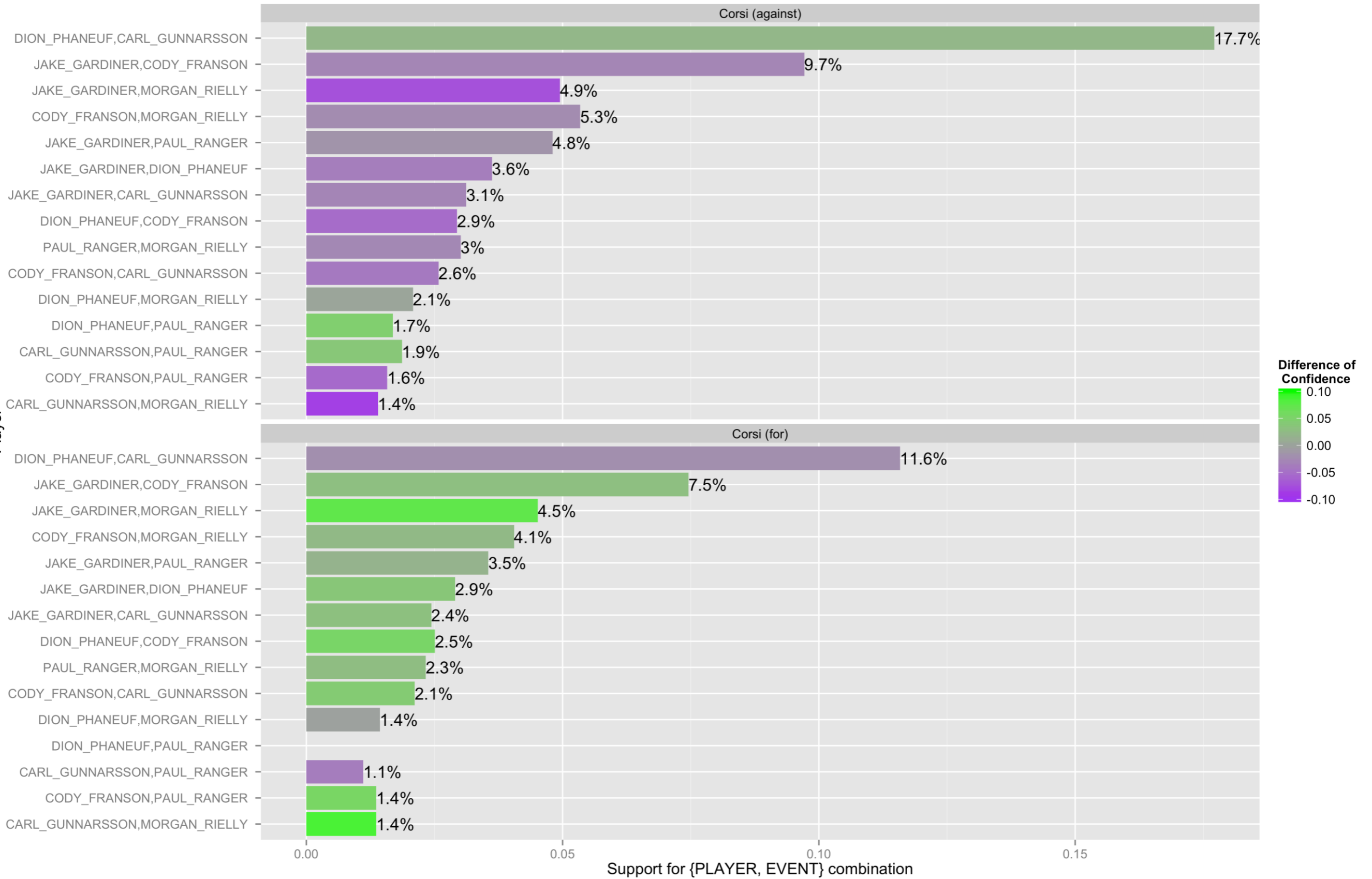
- Bar size is support for events
  - Bigger bar means that combination of players/events is more likely
- Colour is degree of confidence
  - More green means it is more likely that an event happened *because* of the player(s) on the ice
  - Purple is vice versa
  - Grey is less conclusive

# 2013-14 Toronto Maple Leafs, probability of Corsi events by player

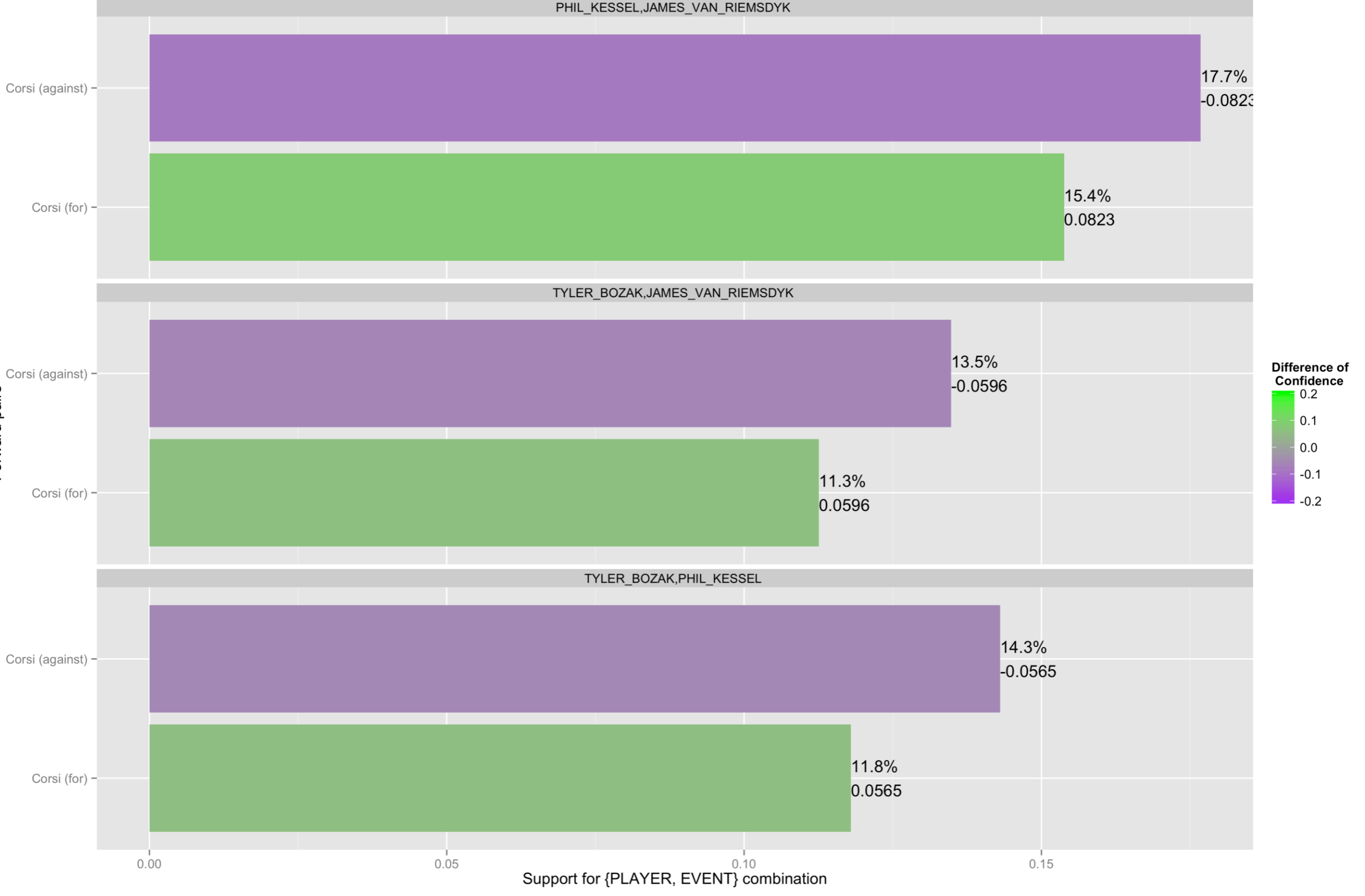
Difference of Confidence =  $\Pr(\text{Event}|\text{Player on}) - \Pr(\text{Event}|\text{Player off})$



### 2013-14 Toronto Maple Leafs defense pairs, probability of Corsi events

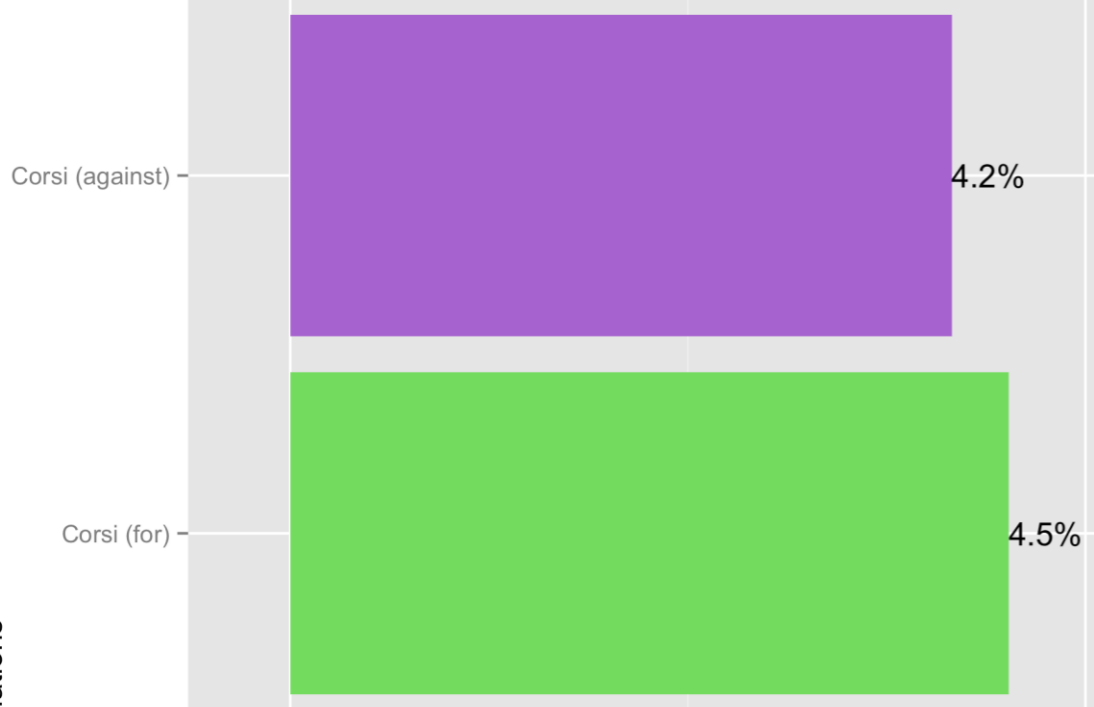


# 2013-14 Toronto Maple Leafs forward combinations, probability of Corsi events

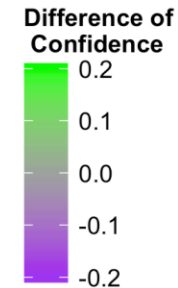


# 2013-14 Toronto Maple Leafs forward combinations, probability of Corsi events

PHIL\_KESSEL,NAZEM\_KADRI,JAMES\_VAN\_RIEMSDYK



TYLER\_BOZAK,PHIL\_KESSEL,JAMES\_VAN\_RIEMSDYK





# Possible conclusions

- Jay McClement was way out of his element
  - Very high probability of Corsi against when he was on the ice vs off the ice
- Phaneuf, Gunnarsson are possibly a weak combination
- Evidence that Gardiner, Rielly played very well together
- Bozak is the weakest member of the top line
  - Kadri would be an improvement

# Presentation conclusions

- Adding penalty times and shift changes gives greater accuracy at team- and player-level
- Association rules give insights on individual and combined player performances
- Goal is to combine the two using enriched RTSS datasets
- Especially since my thesis deadline is approaching

# Questions?

- [alexandr.diazpapkovich@carleton.ca](mailto:alexandr.diazpapkovich@carleton.ca)
- [alexdiаз.ca](http://alexdiаз.ca)
- @adp\_diaz