

# Sports Strategy Analytics using Player Clustering & Probabilistic Model Checking

Capstone Presentation

Prepared by: AARON LOK HIN CHAN (A0117352M)

Advised by: Professor JIN SONG DONG

December 2022



# OUTLINE

## INTRODUCTION

- BACKGROUND & RELATED WORKS
- OBJECTIVE

## METHODOLOGY

- DATA SOURCE
- FEATURE SELECTION
- CLUSTERING
- EVALUATION

## EXPERIMENT RESULTS

## FUTURE WORKS

## TECHNICAL RESEARCH CONTRIBUTIONS



Background

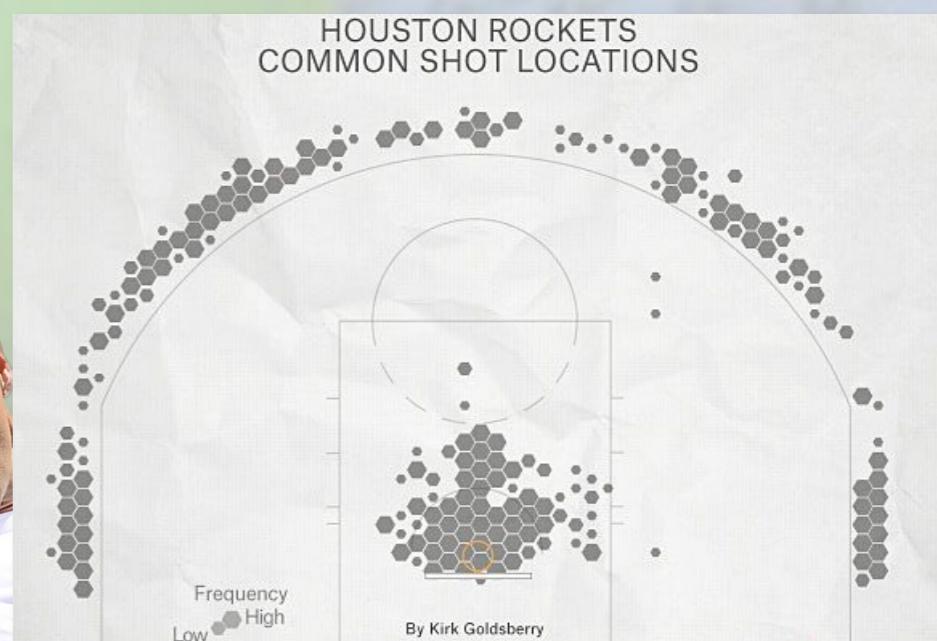
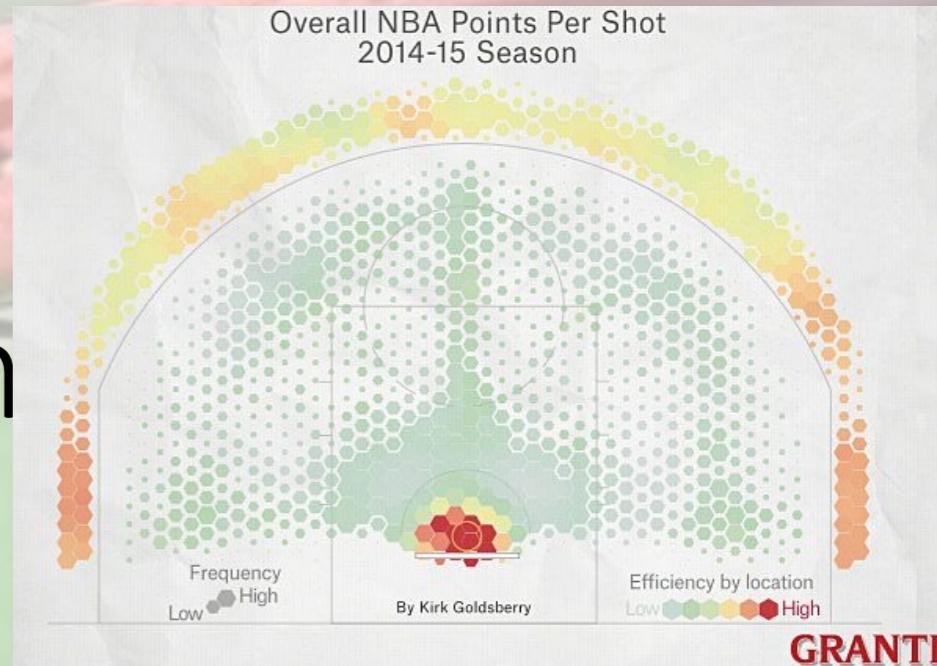
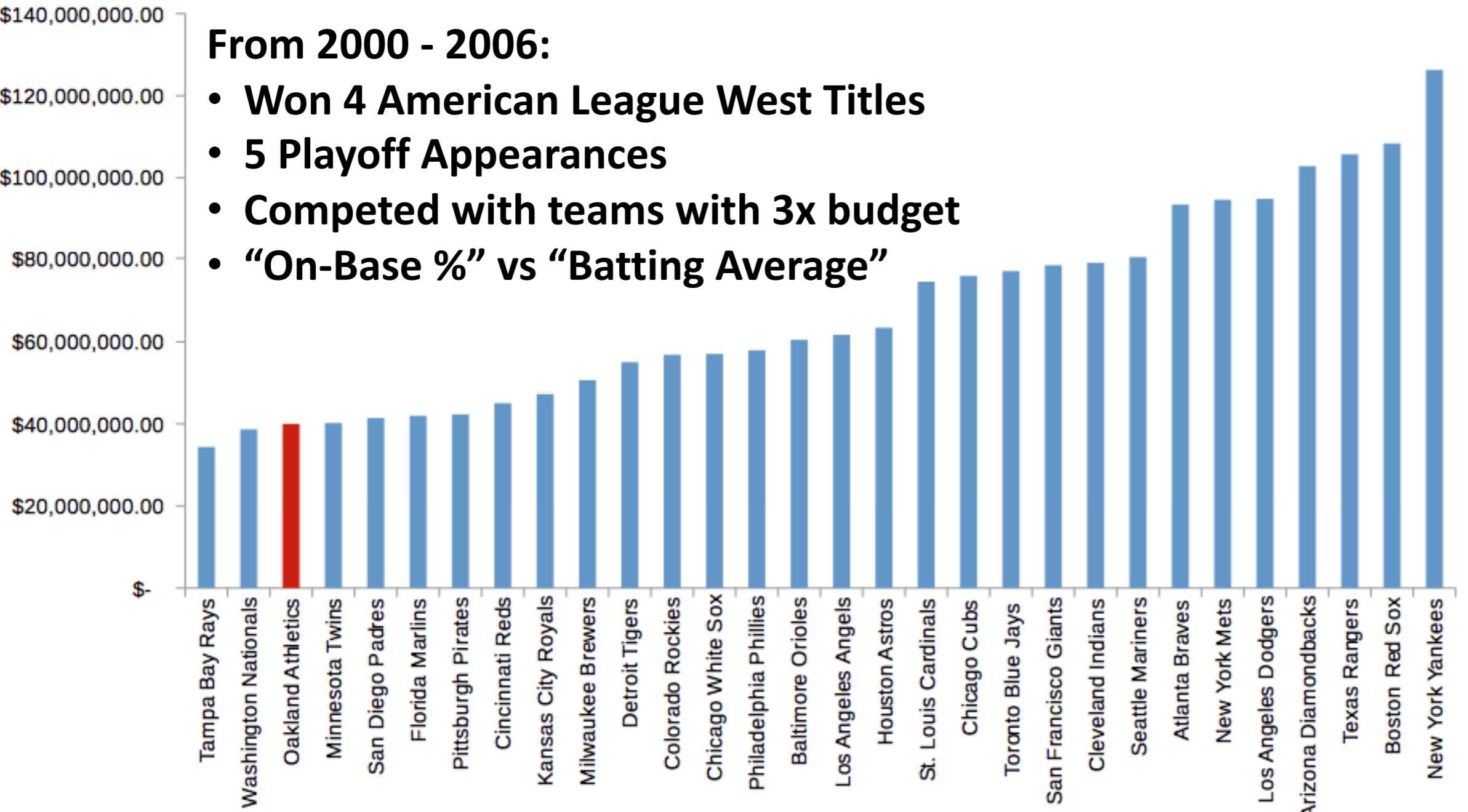
# DATA



## Moneyball Year (2002) MLB Team Salaries

From 2000 - 2006:

- Won 4 American League West Titles
- 5 Playoff Appearances
- Competed with teams with 3x budget
- “On-Base %” vs “Batting Average”



## RELATED WORKS



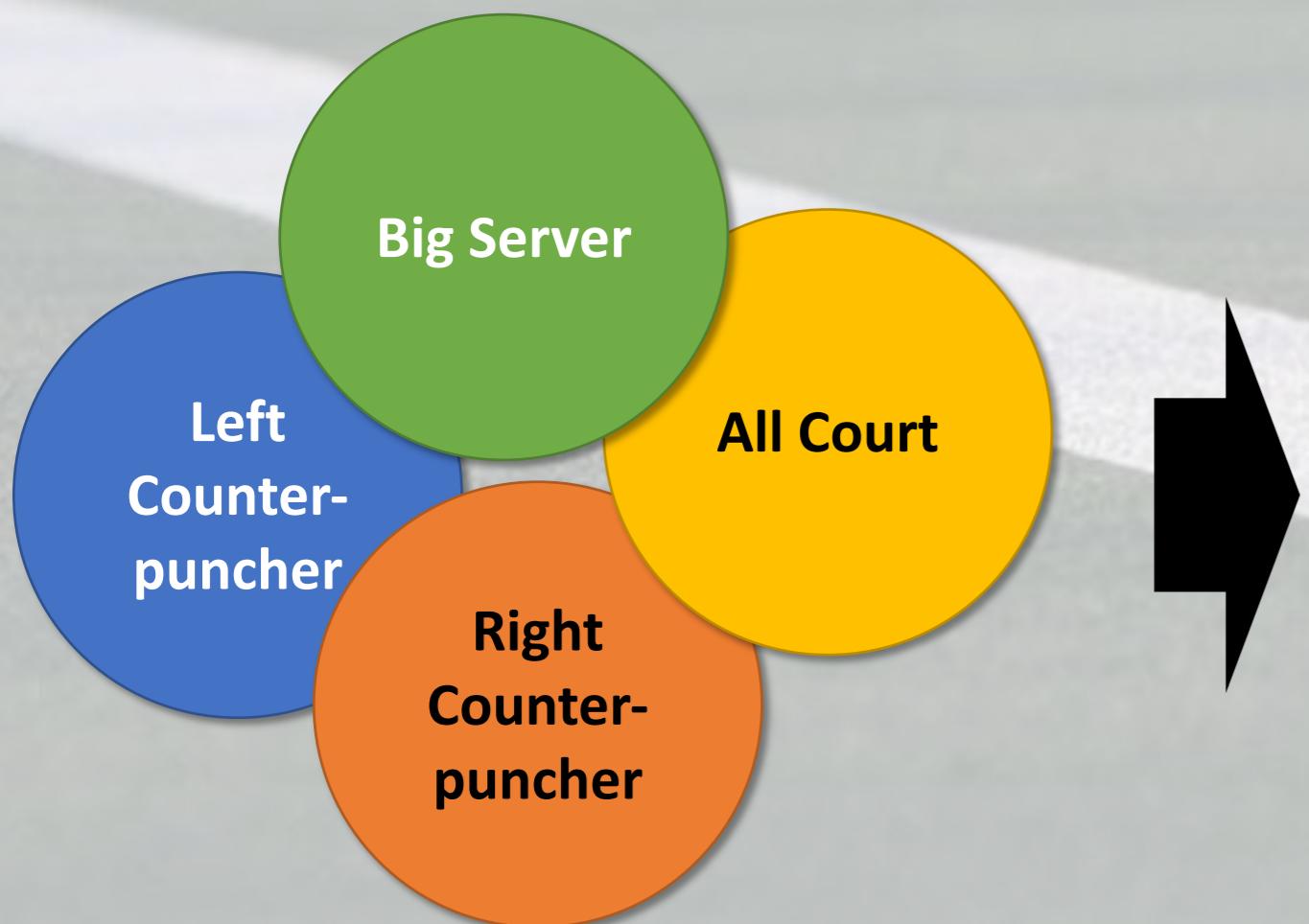
# Study #1: Sorting Strokes - Classifying Tennis Players Based on Style

(by Harvard Sports Analysis Collective)



# HSAC

- Cluster into play styles, and impact on win probability.



## Differences:

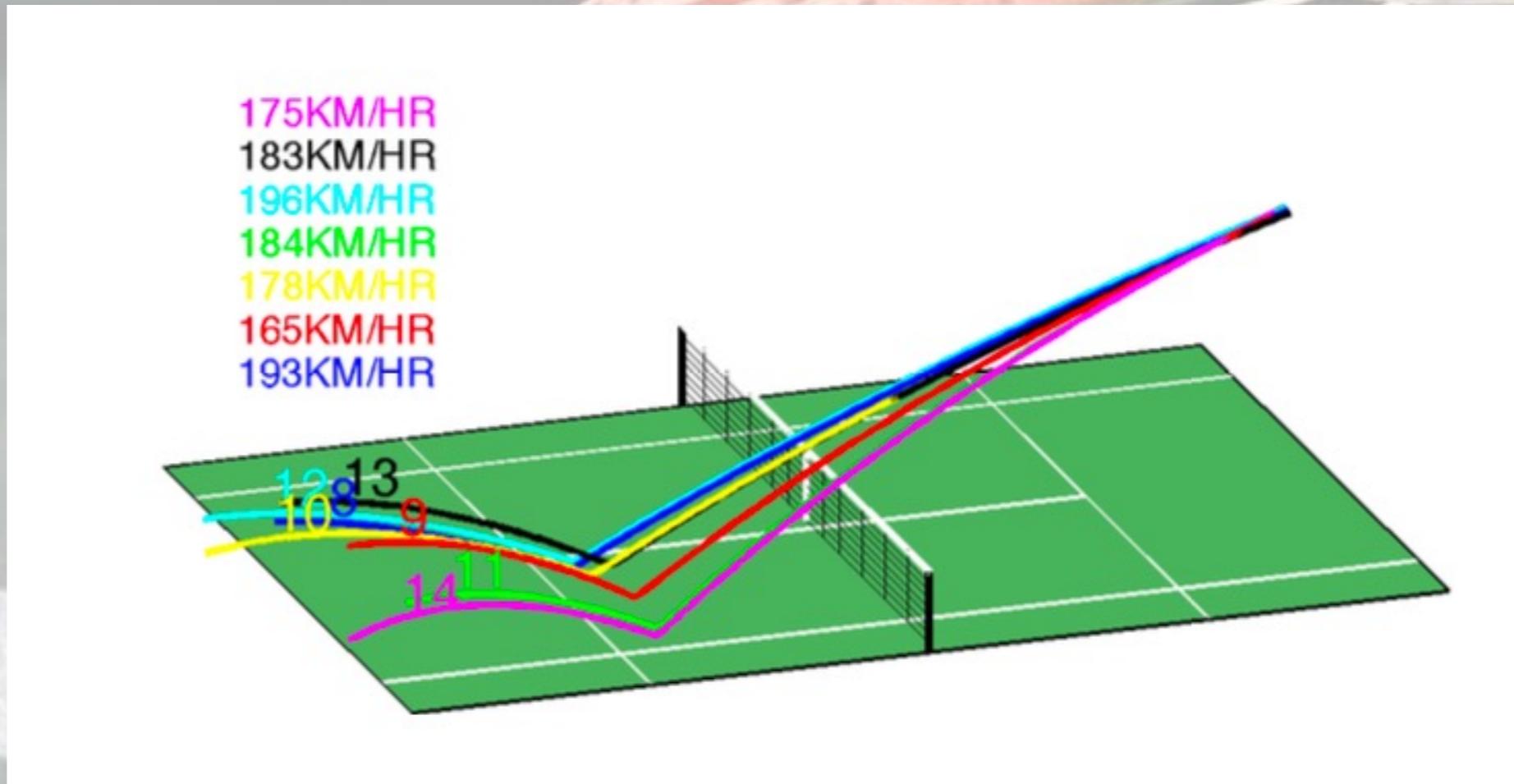
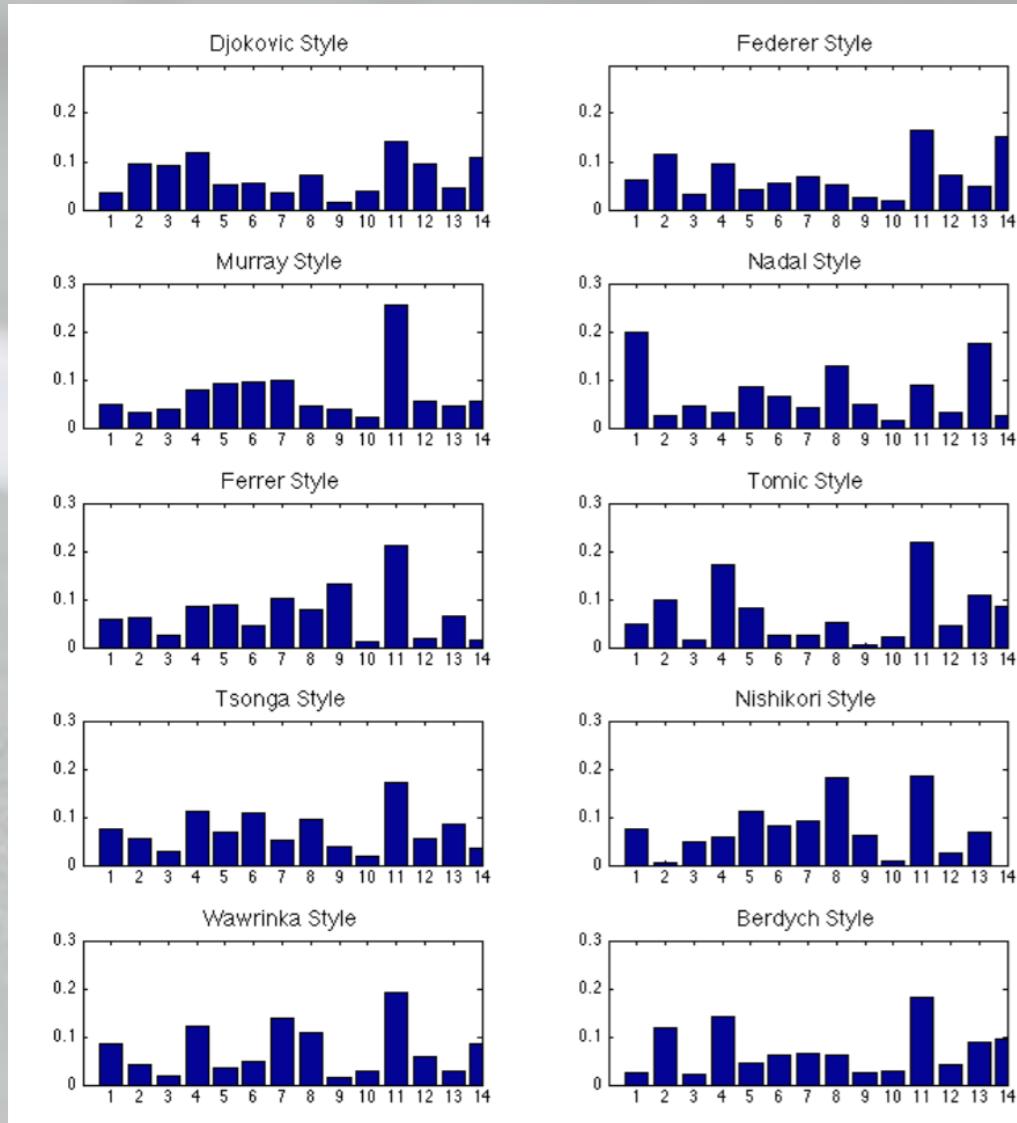
- High-level data only
- Lack of quantitative evaluation – take results as is
- Uses historical match-level data to determine winning likelihood

Elo Adv. (P1)	Elo Disadv. (P2)	Multiplied Effect on Win Odds Elo Adv.
Left Counter-puncher	Big Server	1.95
Right Counter-puncher	Left Counter-puncher	1.21
Right Counter-puncher	All-court	0.84
All-court	Big Server	0.4
Big Server	Right Counter-puncher	1.4

# Study #2: Predicting Serves in Tennis using Style Priors

(by Wei et al.)

- Used Hawk-eye data and “style priors” to predict **serve location**.



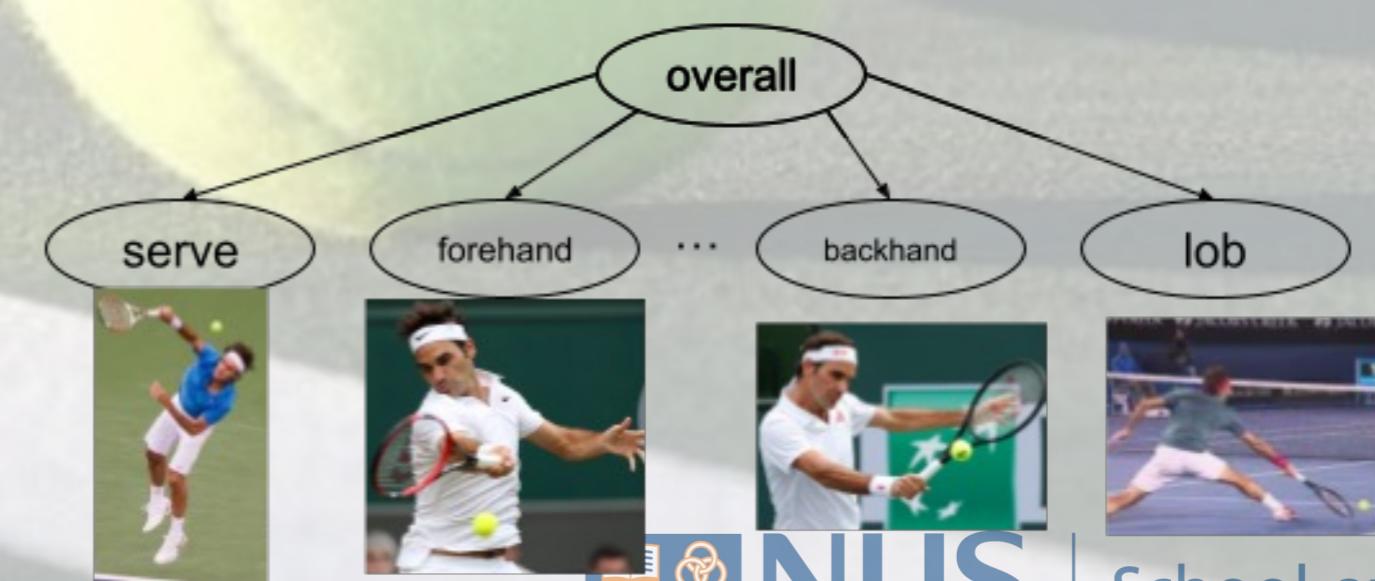
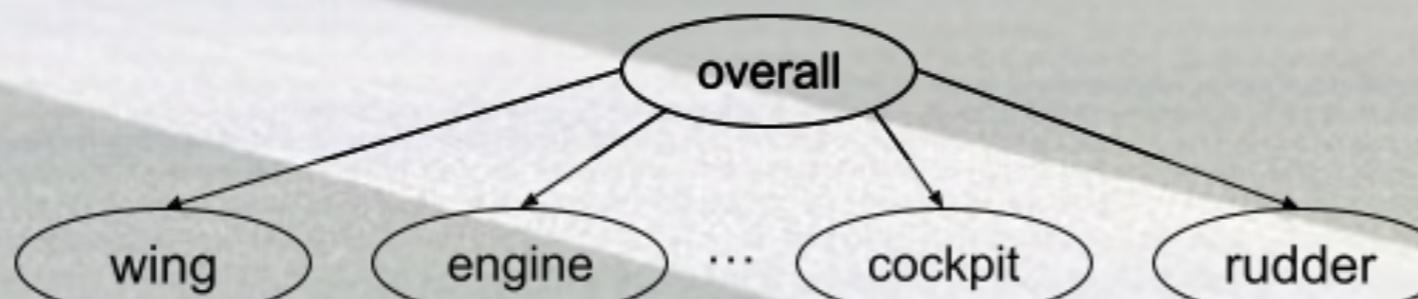
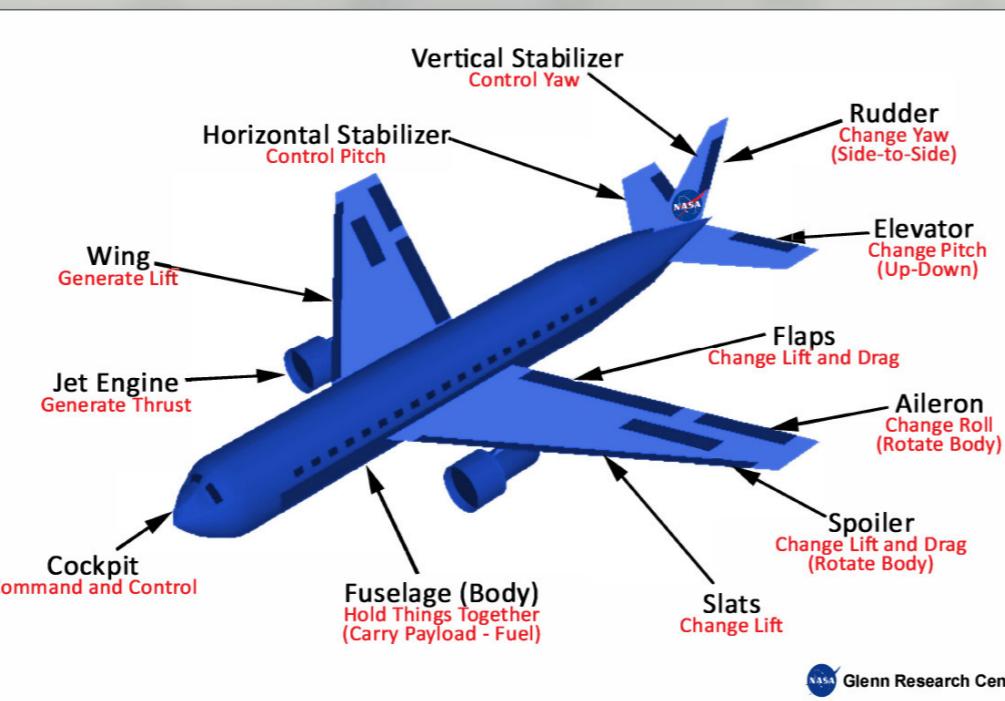
## Differences:

- Hawk-eye data
- Focuses on 1 part of the game

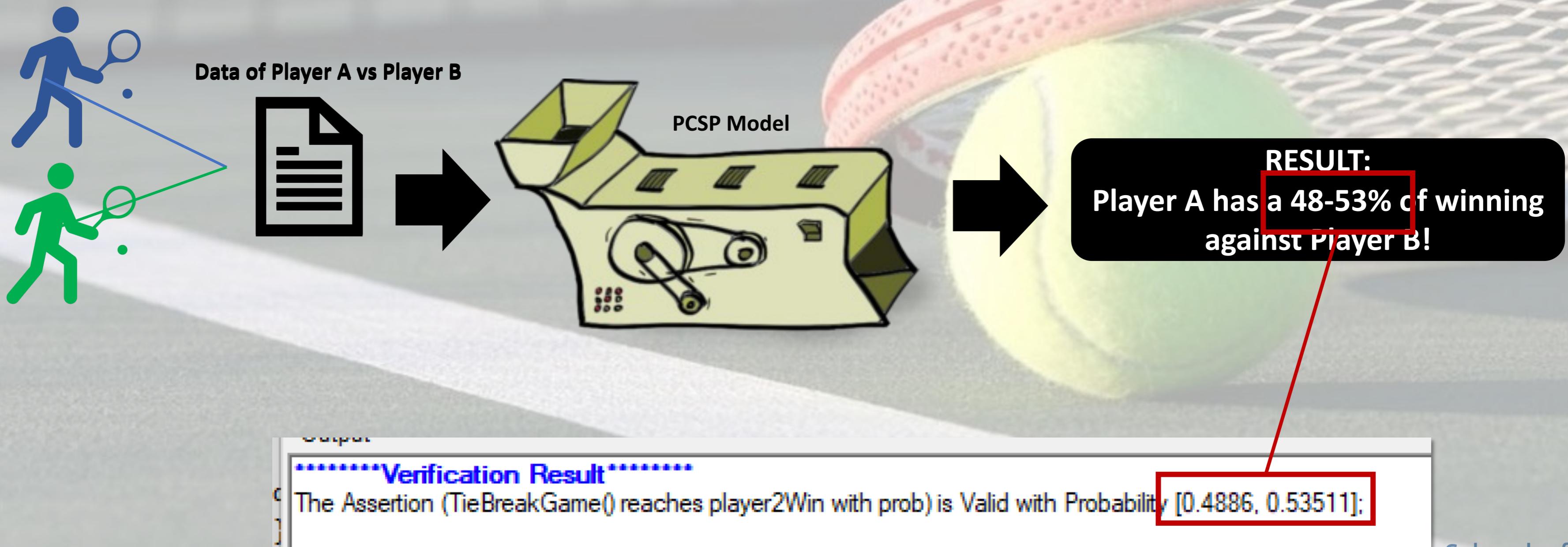
A photograph of two tennis players, Rafael Nadal and Roger Federer, smiling and laughing together. They are both wearing white headbands and blue shirts. Nadal is on the left, holding a tennis racket, and Federer is on the right. The background is dark.

# PCSP Model & PAT Model Checker

# What is PCSP & PAT Model Checker?

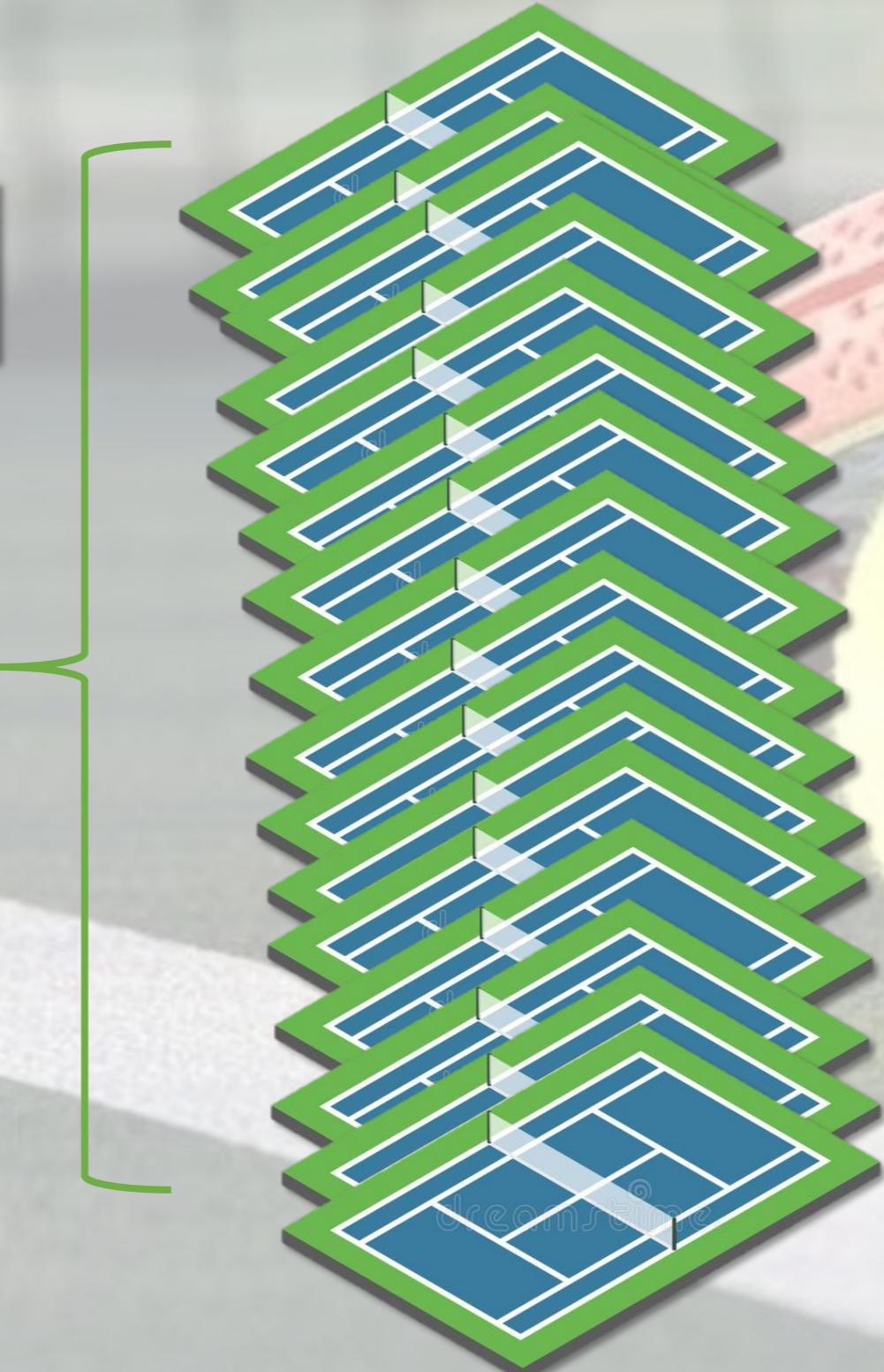


# PCSP & PAT Model Checker



**35 matches recorded**  
(~35-45 between each of the **BIG 3**)

Federer



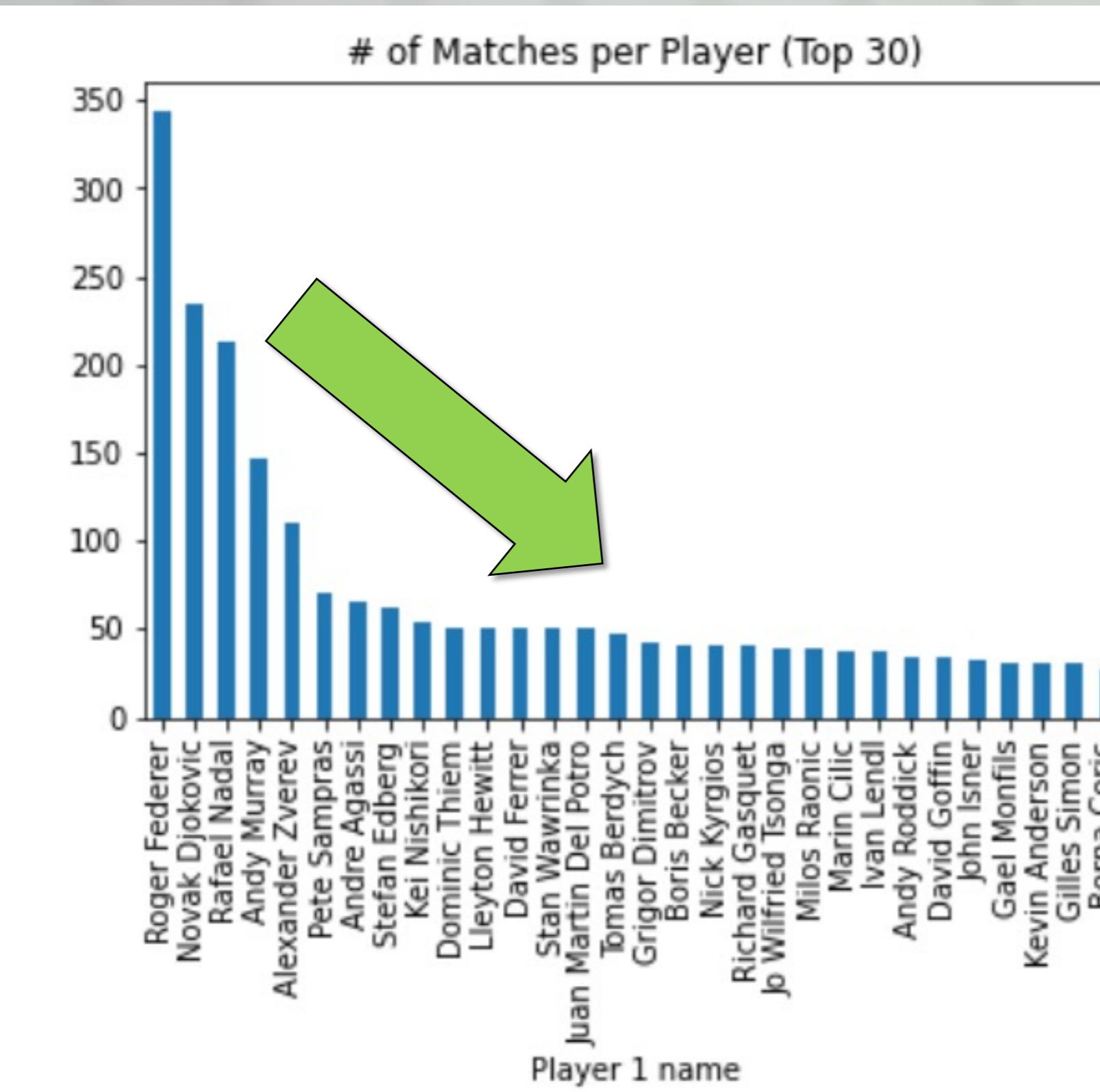
Nadal

of  
ating

# CHALLENGES

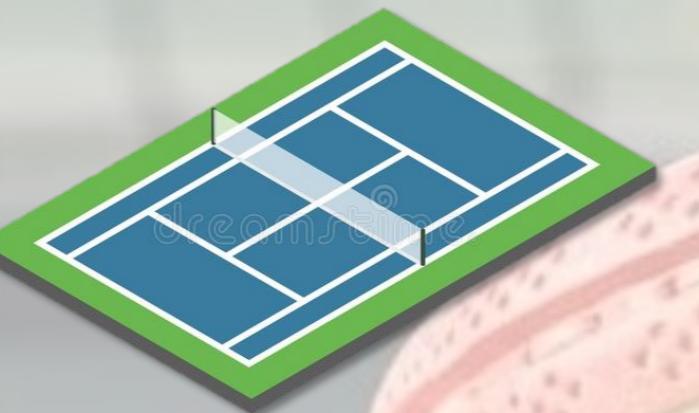


# CHALLENGES



**1 total match recorded**

Milos Raonic  
Ranked #18  
(2018)



I'VE NEVER EVEN  
PLAYED AGAINST  
THESE GUYS!

Nick Kyrgios  
Ranked #35  
(2018)



# OBJECTIVE

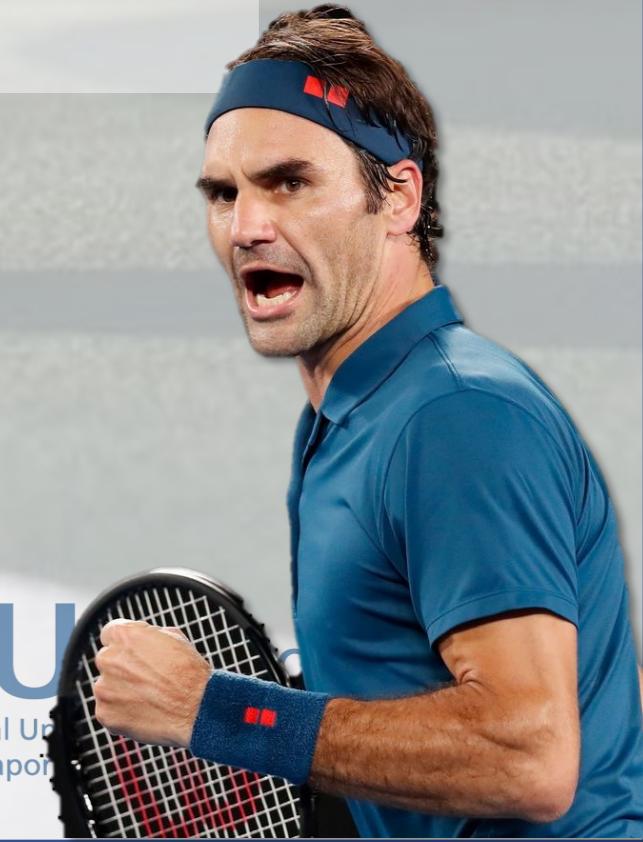
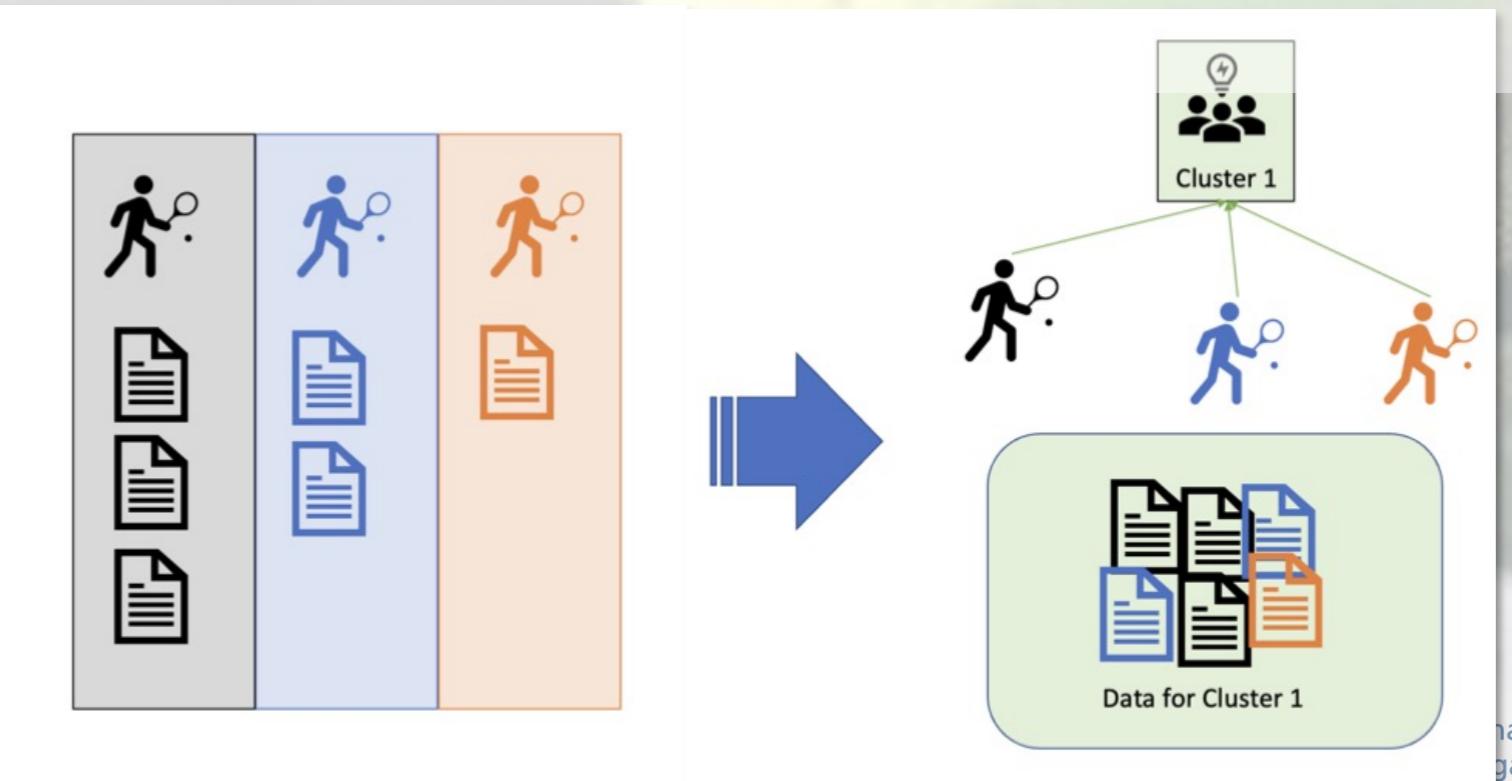


# OBJECTIVE

Group together similar tennis players into **clusters**, which will enable us to model their interactions **even if they never played with one another**, so long as similar players in their clusters have.

## Benefits:

- Increase data availability and access of PCSP functionality for players outside of top players
- Allow players to use “clustered” data in PCSP model
- Many more implications!



# APPROACH

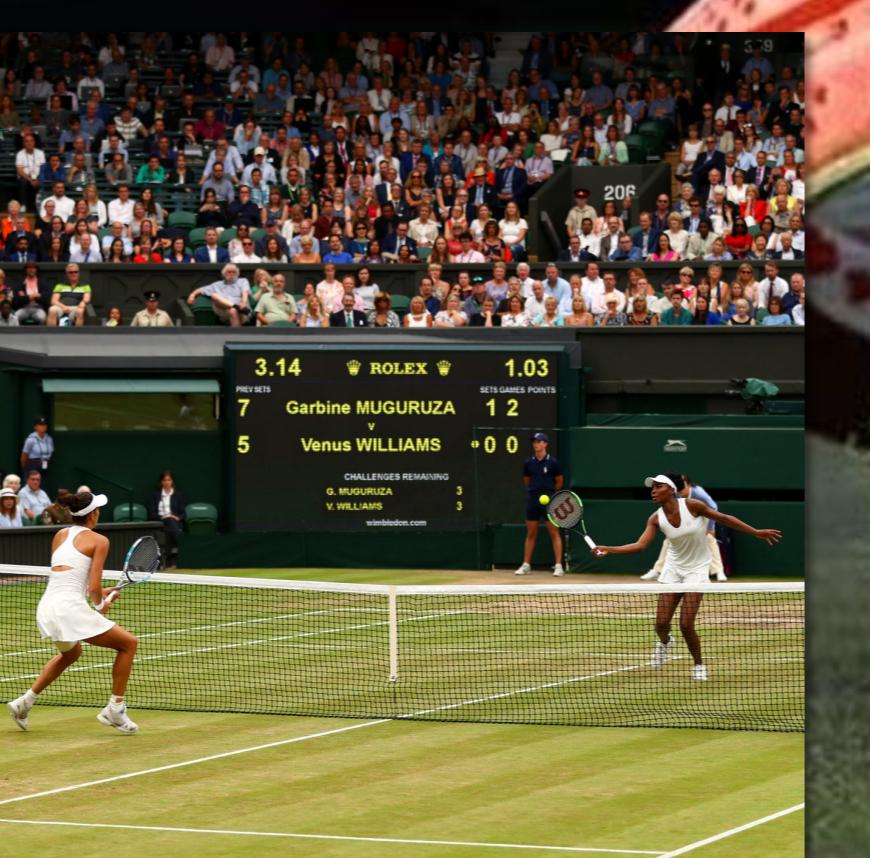
## Game Components



Serve

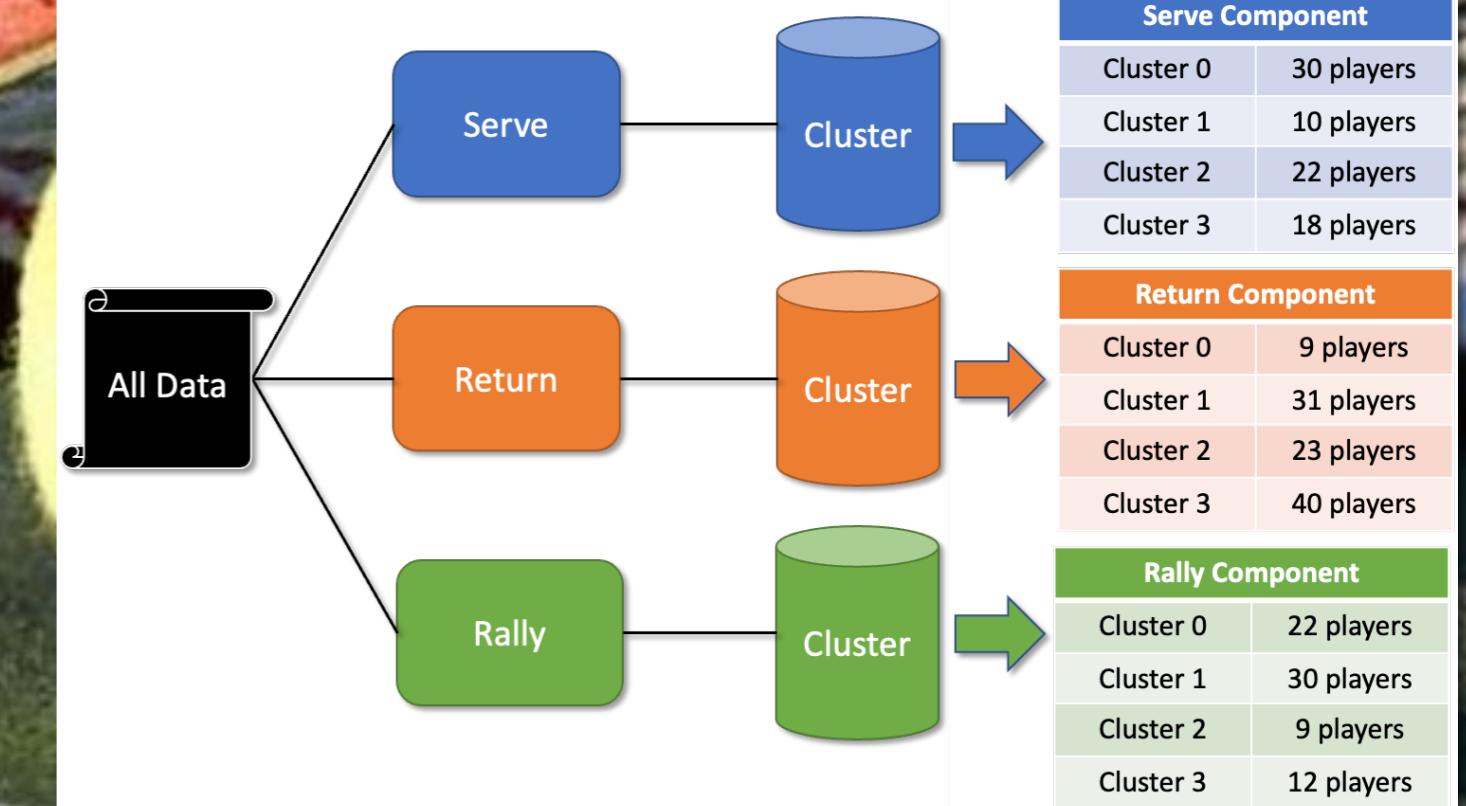


Return



Rally

## Game Components

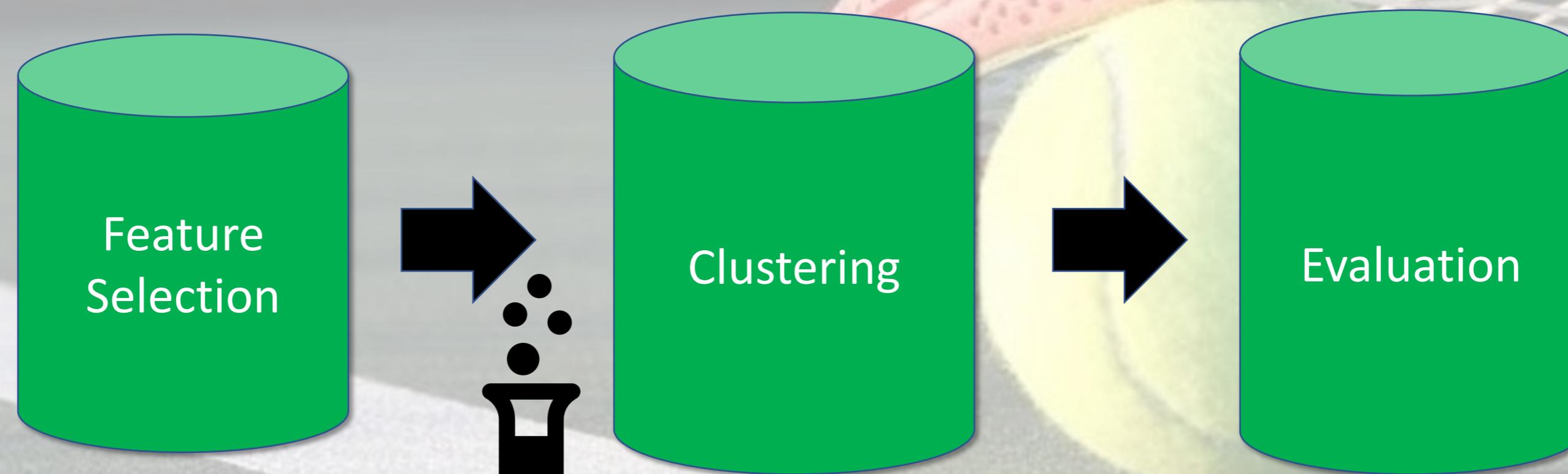


# The Process...

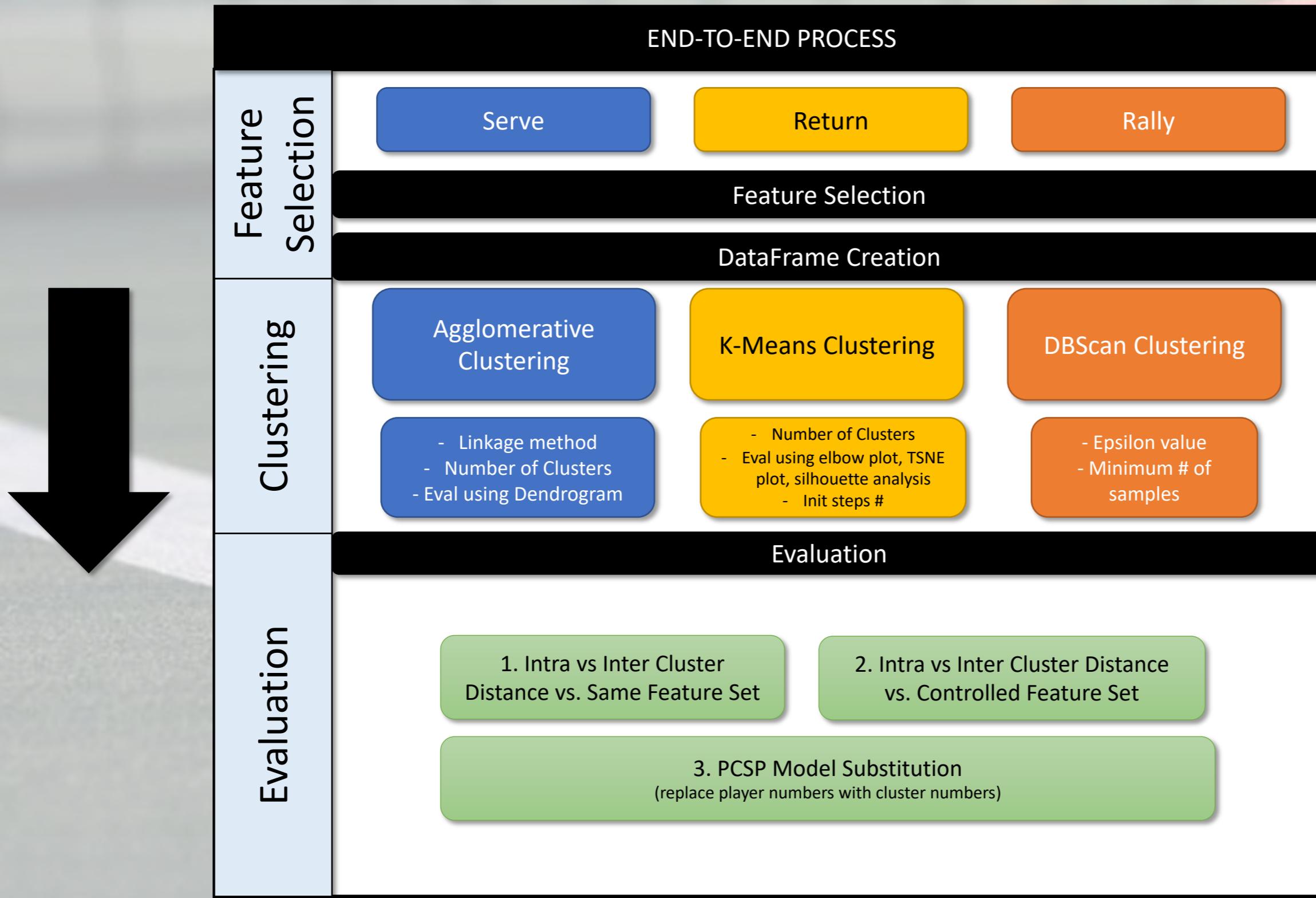
Let's hope this doesn't blow up..



# METHODOLOGY



# METHODOLOGY



# DATA SOURCE

## 1. High-level data: Career Summary Statistics

### Men's Reports

#### [ATP Stats Leaderboard](#)

The ATP top 50, sortable by almost 60 stats.

#### Match Charting Project Leaderboards:

Last 52: [Serve](#) | [Return](#) | [Rally](#) | [Tactics](#)

Career: [Serve](#) | [Return](#) | [Rally](#) | [Tactics](#)

#### [ATP H2H Matrix](#)

Head-to-head records of the current top 15.

#### [The Best ATP Players Who Haven't...](#)

...won titles, reached main draws, etc.

#### [ATP Lottery Matches](#)

...and equally befuddling results from the [Challenger tour](#)

FedEx ATP RANKINGS AS OF — 02.08.2021		
1	NOVAK DJOKOVIC	12,113
2	DANIIL MEDVEDEV	10,220
3	RAFAEL NADAL	8,270
4	STEFANOS TSITSIPAS	8,000
5	ALEXANDER ZVEREV	7,340
6	DOMINIC THIEM	7,095

Service Game	Feature Definition
Matches	Matches logged by the Match Charting Project
Unret%	Percent of serves that were unreturned
<=3 W%	Percent of service points won on either the serve or second shot
RiP W%	Percent of points won when return was put in play
SvImpact	Serve Impact: Advanced stat estimating how many service points were won due to the serve
1st: Unret%	Percent of first serves that were unreturned
1st: <=3 W%	Percent of first serve points won on either the serve or second shot
1st: RiP W%	Percentage of first serve points won when return was put in play
1st: SvImpact	Serve Impact: Advanced stat estimating how many first serve points were won due to the serve
D Wide%	Percent of deuce-court first serves that were hit wide
A Wide%	Percent of ad-court first serves that were hit wide
BP Wide%	Percent of (ad-court) break point first serves that were hit wide
2nd: Unret%	Percent of second serves that were unreturned
2nd: <=3 W%	Percent of second serve points won on either the serve or second shot
2nd: RiP W%	Percentage of second serve points won when return was put in play
1st: D Wide%	Percent of deuce-court second serves that were hit wide
1st: A Wide%	Percent of ad-court second serves that were hit wide
1st: BP Wide%	Percent of (ad-court) break point second serves that were hit wide
2ndAgg	2nd Serve Aggression Score (0 is average, higher = more aggressive)

Player	Matches	Unret%	<=3 W%	RiP W%	SvImpact	1st: Unret%	<=3 W%	RiP W%	SvImpact	D Wide%	A Wide%	BP Wide%	2nd: Unret%	<=3 W%	RiP W%	D Wide%	A Wide%	BP Wide%	2ndAgg
Ivo Karlovic	35	48.0%	63.2%	50.9%	55.3%	59.0%	74.8%	57.0%	70.1%	42.9%	56.9%	57.5%	31.6%	48.0%	42.7%	26.4%	30.2%	14.3%	126
Maxime Cressy	26	43.8%	59.2%	55.8%	52.1%	51.0%	68.5%	58.1%	64.3%	55.4%	59.9%	74.2%	43.6%	59.5%	51.4%	47.0%	55.3%	56.0%	245
John Isner	84	45.5%	59.2%	48.5%	52.1%	54.1%	67.4%	51.2%	63.5%	50.1%	45.9%	57.2%	26.7%	42.7%	44.1%	22.4%	52.1%	68.5%	34
Reilly Opelka	53	46.3%	58.8%	48.6%	51.7%	57.5%	69.2%	50.6%	65.6%	50.0%	44.1%	44.8%	26.9%	42.2%	46.2%	33.3%	53.9%	82.7%	33
Goran Ivanisevic	44	42.1%	53.9%	49.8%	47.4%	58.9%	72.1%	55.4%	68.6%	38.1%	55.0%	56.8%	26.3%	38.3%	45.4%	16.9%	51.7%	65.6%	119
Milos Raonic	78	39.7%	54.0%	51.6%	47.1%	52.0%	68.1%	57.3%	63.8%	53.6%	50.2%	45.5%	21.4%	34.1%	45.0%	28.1%	54.4%	71.8%	76
Richard Krajicek	20	40.2%	53.6%	47.8%	46.8%	54.9%	69.8%	56.7%	66.5%	45.7%	39.9%	27.8%	23.9%	37.0%	39.9%	12.4%	49.8%	70.0%	60
Nick Kyrgios	111	40.5%	51.8%	51.0%	46.4%	51.0%	63.2%	54.2%	59.8%	43.3%	40.9%	57.0%	23.0%	33.7%	46.4%	31.1%	50.5%	43.3%	103
Gilles Muller	20	37.3%	54.3%	53.1%	45.9%	49.3%	69.5%	60.4%	63.2%	45.5%	46.6%	40.0%	19.8%	33.2%	44.5%	24.2%	30.4%	23.5%	75
Sam Querrey	34	39.7%	51.2%	48.5%	45.6%	54.0%	67.4%	52.8%	63.7%	47.3%	47.0%	51.7%	20.0%	30.0%	44.1%	36.4%	47.4%	41.8%	77

# DATA SOURCE

## 2. Low-level data: Shot-by-Shot data

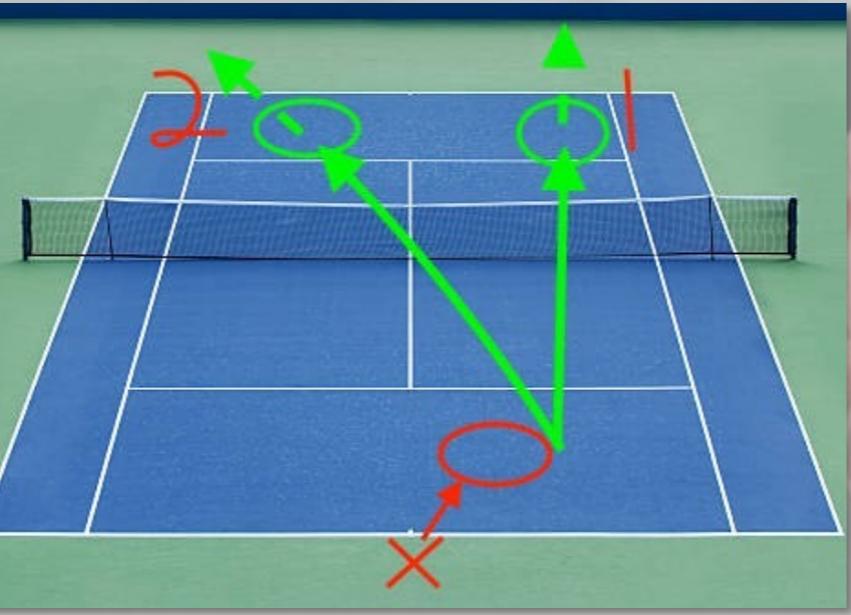
### Summary:

- 3 million+ rows (shots)
- 800+ players
- 8000+ matches
- Around 400 shots/match
- 59 columns
  - 11 features = “shot attributes”

### Tennis Abstract (GitHub)

Short	Player 1 name	Player 2 name	Player 1 handness	Player 2 handness	Player 1 points	Player 2 points	Player 1 games	Player 2 games	Player 1 sets	Player 2 sets	...	Prev prev shot From which court	Prev prev shot Prev Shot	Prev prev shot Direction	Prev prev shot To which court	Prev prev shot Depth	Prev prev shot Touched Net?	Prev prev Hit at what depth
0	Novak Djokovic	Alexander Zverev	RH	RH	0	0	0	0	0	0	...	99	99	99	99	99	99	99
1	Alexander Zverev	Novak Djokovic	RH	RH	0	0	0	0	0	0	...	99	99	99	99	99	99	99
2	Novak Djokovic	Alexander Zverev	RH	RH	0	0	0	0	0	0	...	1	99	6	2	99	2	
3	Alexander Zverev	Novak Djokovic	RH	RH	0	0	0	0	0	0	...	2	22	2	2	2	2	
4	Novak Djokovic	Alexander Zverev	RH	RH	0	0	0	0	0	0	...	2	1	7	1	99	2	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3306473	Ken Rosewall	Rod Laver	RH	LH	3	4	3	5	2	2	...	99	99	99	99	99	99	
3306474	Rod Laver	Ken Rosewall	LH	RH	4	3	5	3	2	2	...	99	99	99	99	99	99	
3306475	Ken Rosewall	Rod Laver	RH	LH	3	4	3	5	2	2	...	3	99	5	3	99	2	
3306476	Rod Laver	Ken Rosewall	LH	RH	4	3	5	3	2	2	...	3	22	2	2	2	2	
3306477	Ken Rosewall	Rod Laver	RH	LH	3	4	3	5	2	2	...	2	13	2	2	99	2	

3306478 rows x 45 columns



Field Name	Feature Definition
Player 1 name	Player's name
Player 2 name	Opponent's name
Player 1 handness	Dominant hand of player (RH for Right-hand, LH for Left-hand)
Player 2 handness	Dominant hand of opponent
Date	Date of match
Tournament Name	Name of tournament
Shot Type	Type of shot (1=1st serve, 2=2nd serve, 3=return, 4=rally)
From which court	Court location of when ball is hit by Player 1 (1=deuce court, 2=middle court, 3=ad court, 99=unknown)
Shot	Shot description (1~20 are forehand shots, 21~40 are backhand shots, 41 is trick shot, 99=unknown), detailed codes for all shots are:
Direction (serve)	Serve direction (1=deuce court, 2=middle court, 3=ad court, 4=serve to wide, 5=serve to body, 6=server to T, 99=unknown)
To which court	Shot direction (1=deuce court, 2=middle court, 3=ad court, 99=unknown)
Depth	(1=shallow, 2=deep, 3=very deep, 99=unknown)
Touched Net?	(1=yes touched net, 2=not)
Hit at what depth?	(1=at net, 2=at baseline, 99=unknown)
Approach shot?	(1=yes, 2=no)
Shot outcome	Outcome of shot (1=ace, 2=fault, 3=forced error, 4=unforced error, 5=-winner, 6=service winner, 7=no outcome)
Fault type	Fault type, if Shot Type = serve (1=(net), 2=(wide), 3=(long), 4=(wide and long), 5=(foot fault), 6=(shank), 7=no fault, 99=other)

# EXPLORATORY DATA ANALYSIS

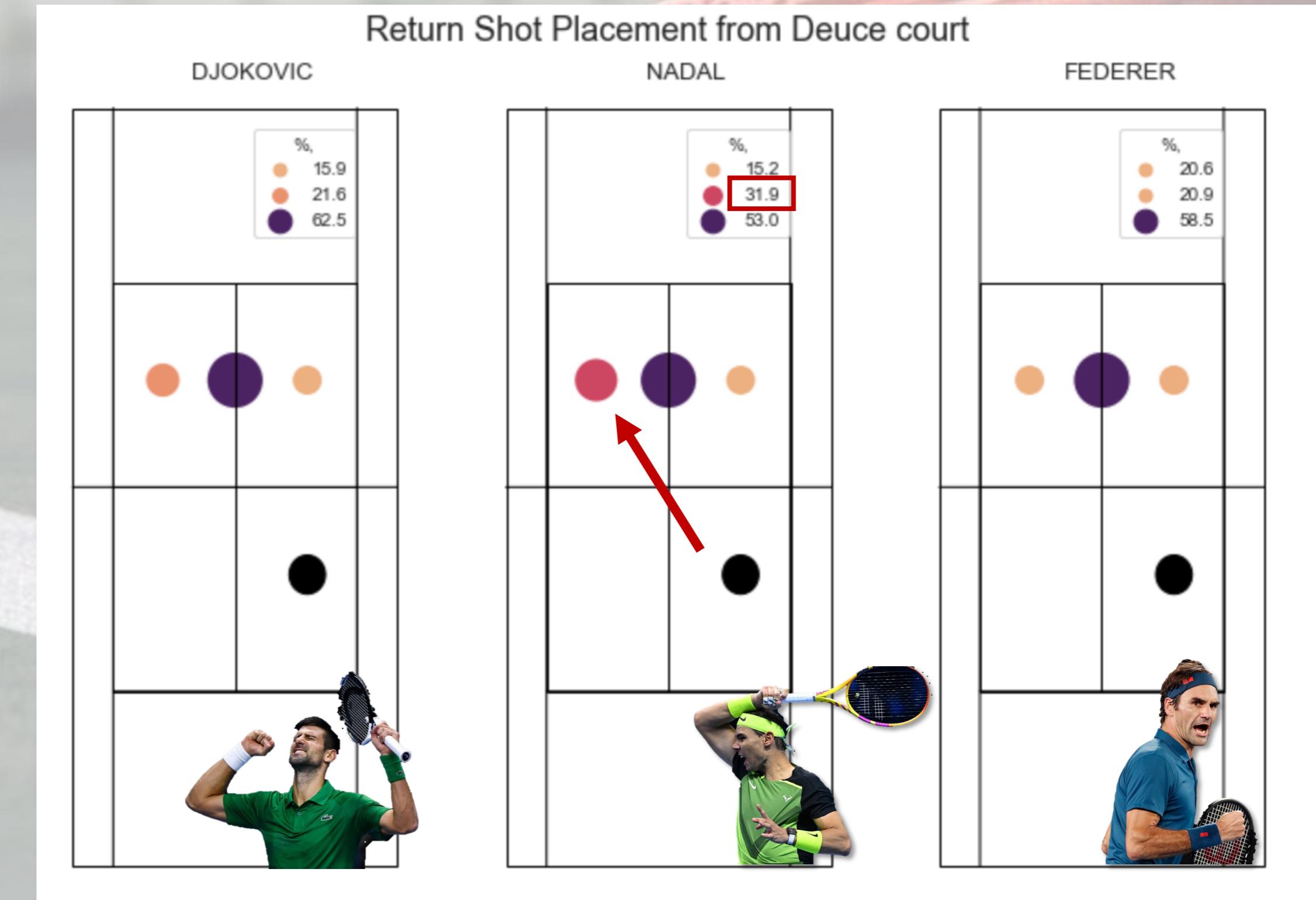
Djokovic			
court	count	percent	
mid	3594	62.5	
deuce	1242	21.6	
ad	912	15.9	

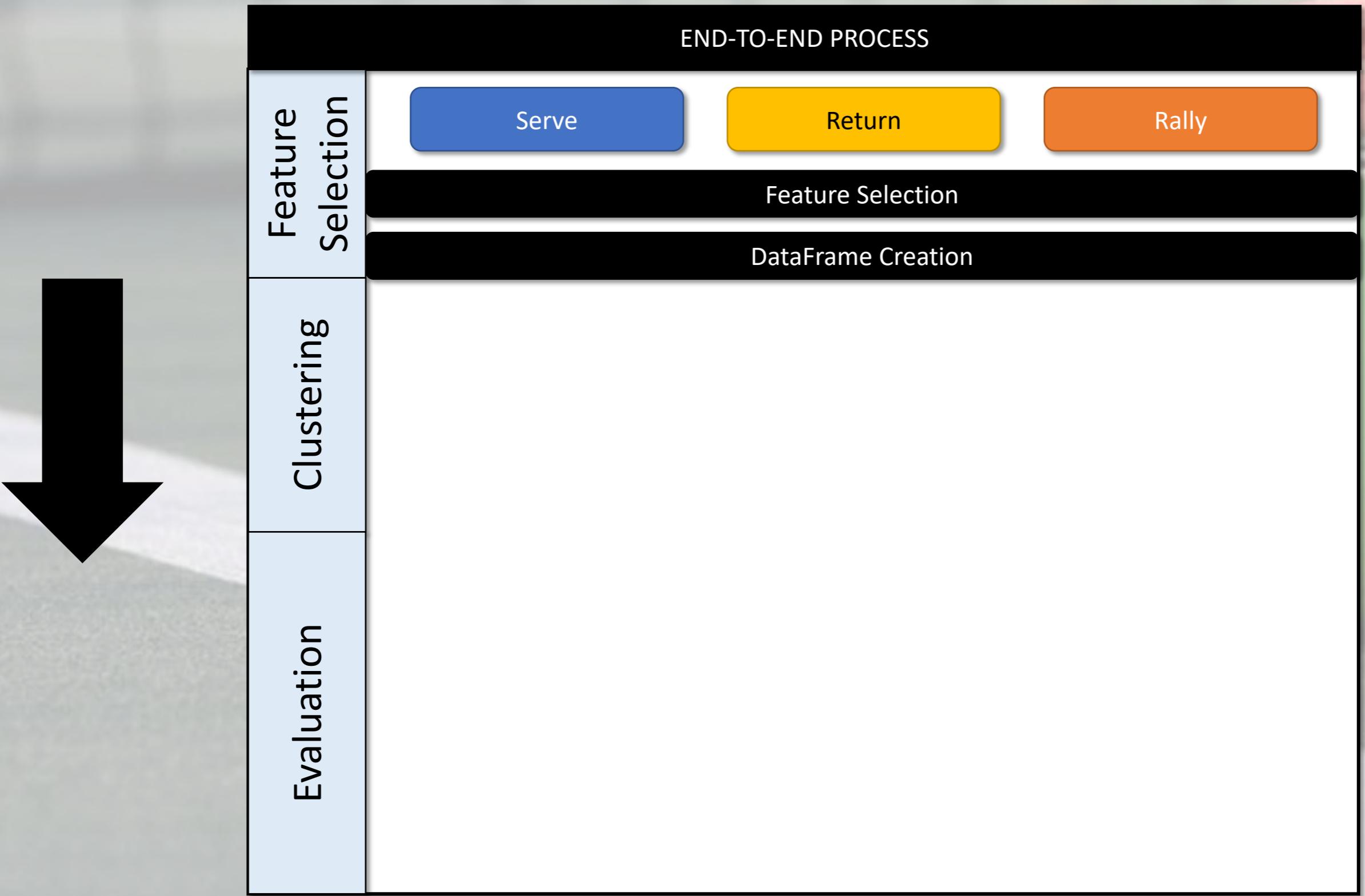
Nadal			
court	count	percent	
mid	3363	53.0	
deuce	2023	31.9	
ad	963	15.2	

Federer			
court	count	percent	
mid	4678	58.5	
ad	1668	20.9	
deuce	1647	20.6	



# PHASE 1: FEATURE SELECTION



## FEATURE SELECTION

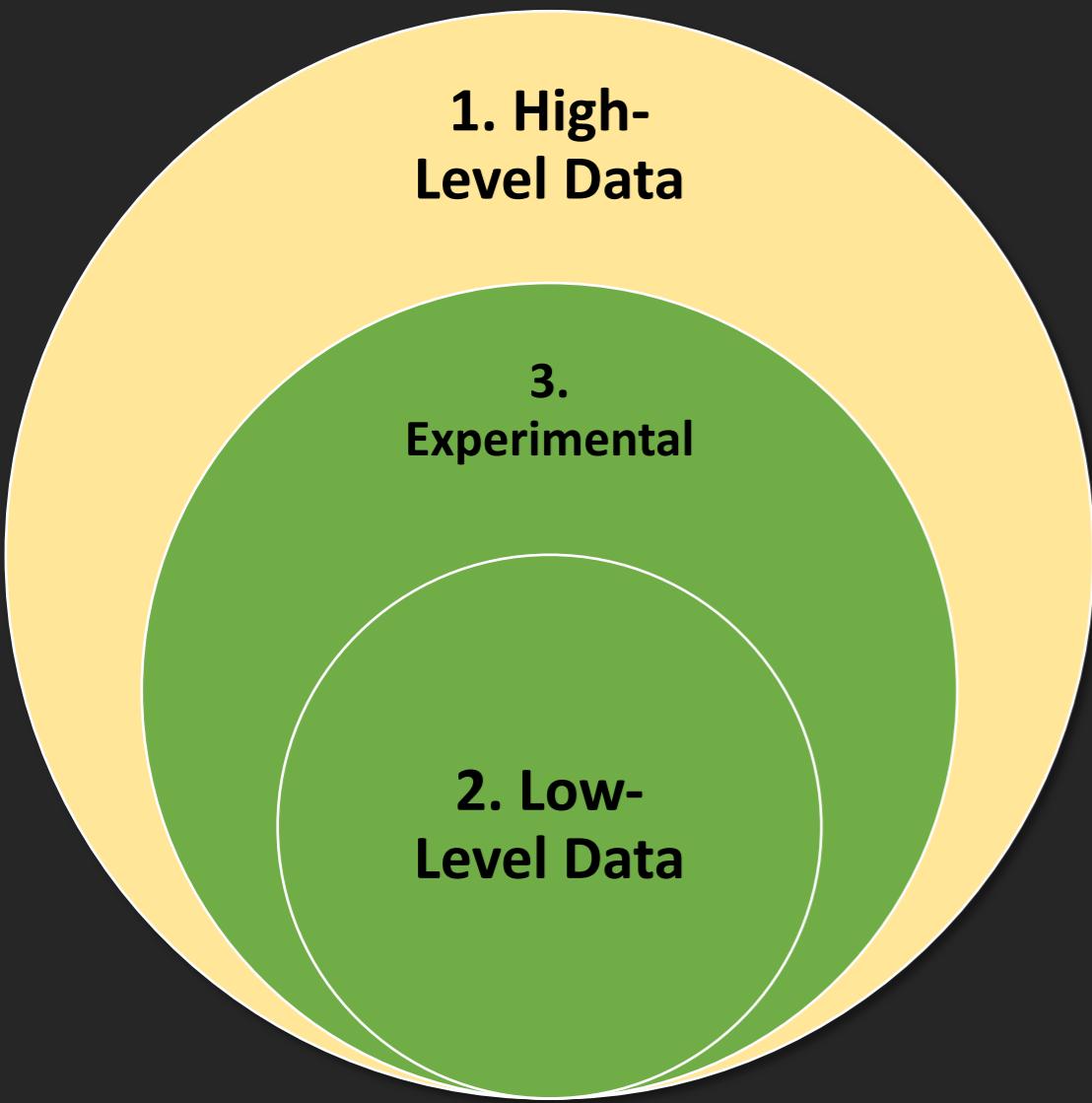
A scene from Disney's Alice in Wonderland. Alice, in her blue dress, stands in a dark, winding path. Above her, the Cheshire Cat is perched on a branch, smiling mischievously. The text is overlaid on the image, with Alice's speech in a black box and the Cheshire Cat's response in another.

ALICE: WOULD YOU TELL ME,  
PLEASE, WHICH FEATURES TO CHOOSE?

K-Means

CLUSTERING ALGORITHM: THAT DEPENDS A GOOD  
DEAL ON WHAT PARTS OF THE  
GAME YOU WANT TO REPRESENT

# FEATURE SELECTION



## 1. High-Level Summary Data

Service Game	Feature Definition
Matches	Matches logged by the Match Charting Project
Unret%	Percent of serves that were unreturned
<=3 W%	Percent of service points won on either the serve or second shot
RiP W%	Percent of points won when return was put in play
SvImpact	Serve Impact: Advanced stat estimating how many service points were won due to the serve
1st: Unret%	Percent of first serves that were unreturned
1st: <=3 W%	Percent of first serve points won on either the serve or second shot
1st: RiP W%	Percentage of first serve points won when return was put in play
1st: SvImpact	Serve Impact: Advanced stat estimating how many first serve points were won due to the serve
D Wide%	Percent of deuce-court first serves that were hit wide
A Wide%	Percent of ad-court first serves that were hit wide
BP Wide%	Percent of (ad-court) break point first serves that were hit wide
2nd: Unret%	Percent of second serves that were unreturned
2nd: <=3 W%	Percent of second serve points won on either the serve or second shot
2nd: RiP W%	Percentage of second serve points won when return was put in play
1st: D Wide%	Percent of deuce-court second serves that were hit wide
1st: A Wide%	Percent of ad-court second serves that were hit wide
1st: BP Wide%	Percent of (ad-court) break point second serves that were hit wide
2ndAgg	2nd Serve Aggression Score (0 is average, higher = more aggressive)

# FEATURE SELECTION

## PCSP Functions

1. High-Level Data

3. Experimental

2. Low-Level Data

```
// deuce stroke is when player1 hit position is 1
Ply1_de_stroke = pcase{
  13: FH_Crosscourt { ball = 6} -> Ply2_de_stroke
  13: FH_Downline { ball = 4} -> Ply2_ad_stroke
  13: FH_DownMid { ball = 5} -> Ply2_mid_stroke
  13: BH_InsideIn {ball = 4} -> Ply2_ad_stroke
  12: BH_InsideOut { ball = 6} -> Ply2_de_stroke
  12: BH_DownMid { ball = 5} -> Ply2_mid_stroke
  12: FH_Error { ball = 9} -> {nscore++; if (nscore == 7) {won = player2}
    else { turn = (turn+1)%4 }
    } -> NextPt
  12: BH_Error { ball = 9} -> {nscore++; if (nscore == 7) {won = player2}
    else { turn = (turn+1)%4 }
    } -> NextPt
}.
```

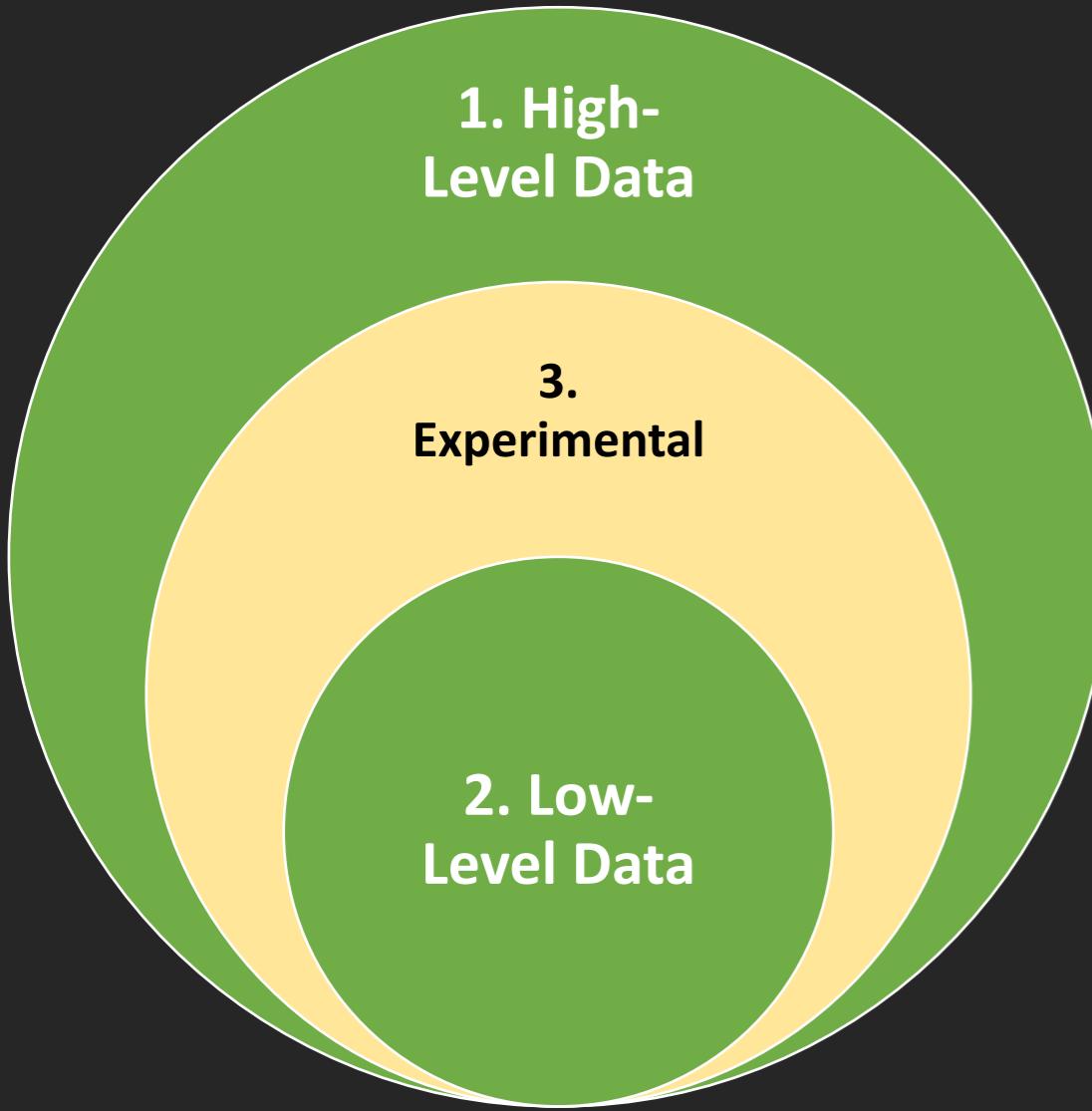
```
// mid stroke is when player1 hit position is 2
Ply1_mid_stroke = pcase{
  13: FH_InsideOut{ball = 4} -> Ply2_ad_stroke
  13: FH_Crosscourt { ball = 6} -> Ply2_de_stroke
  13: FH_DownMid { ball = 5 } -> Ply2_mid_stroke
  13: BH_InsideOut { ball = 6} -> Ply2_de_stroke
  12: BH_DownMid { ball = 5} -> Ply2_mid_stroke
  12: FH_Crosscourt { ball = 4} -> Ply2_ad_stroke
  12: FH_Error { ball = 9} -> {nscore++; if (nscore == 7) {won = player2}
    else { turn = (turn+1)%4 }
    } -> NextPt
  12: BH_Error { ball = 9} -> {nscore++; if (nscore == 7) {won = player2}
    else { turn = (turn+1)%4 }
    } -> NextPt
}.
```

```
// ad stroke is when player1 hit position is 3
Ply1_ad_stroke = pcase{
  13: BH_Crosscourt { ball = 4} -> Ply2_ad_stroke
  13: BH_Downline { ball = 6} -> Ply2_de_stroke
  13: BH_DownMid { ball = 5 } -> Ply2_mid_stroke
  13: FH_InsideOut { ball = 4} -> Ply2_ad_stroke
  12: FH_InsideIn { ball = 6} -> Ply2_de_stroke
  12: FH_DownMid { ball = 5 } -> Ply2_mid_stroke
  12: FH_Error { ball = 9} -> {nscore++; if (nscore == 7) {won = player2}
    else { turn = (turn+1)%4 }
    } -> NextPt
  12: BH_Error { ball = 9} -> {nscore++; if (nscore == 7) {won = player2}
    else { turn = (turn+1)%4 }
    } -> NextPt
}.
```

## 2. Low-Level Shot % Data

0: From_Court	rally_outcome	
middle court	BH_Error	0.080589
	BH_cross_court	0.170786
	BH_down_mid	0.097845
	BH_inside_out	0.094301
	FH_Error	0.093835
	FH_cross_court	0.195411
	FH_down_mid	0.079657
	FH_inside_out	0.187576
deuce court	BH_Error	0.002019
	BH_down_mid	0.001863
	BH_inside_in	0.003416
	BH_inside_out	0.001708
	FH_Error	0.261180
	FH_cross_court	0.318944
	FH_down_line	0.260404
	FH down mid	0.150466
ad court	BH_Error	0.193075
	BH_cross_court	0.354506
	BH_down_line	0.185508
	BH_down_mid	0.168769
	FH_Error	0.015746
	FH_down_mid	0.010166
	FH_inside_in	0.027669
	FH inside out	0.044562

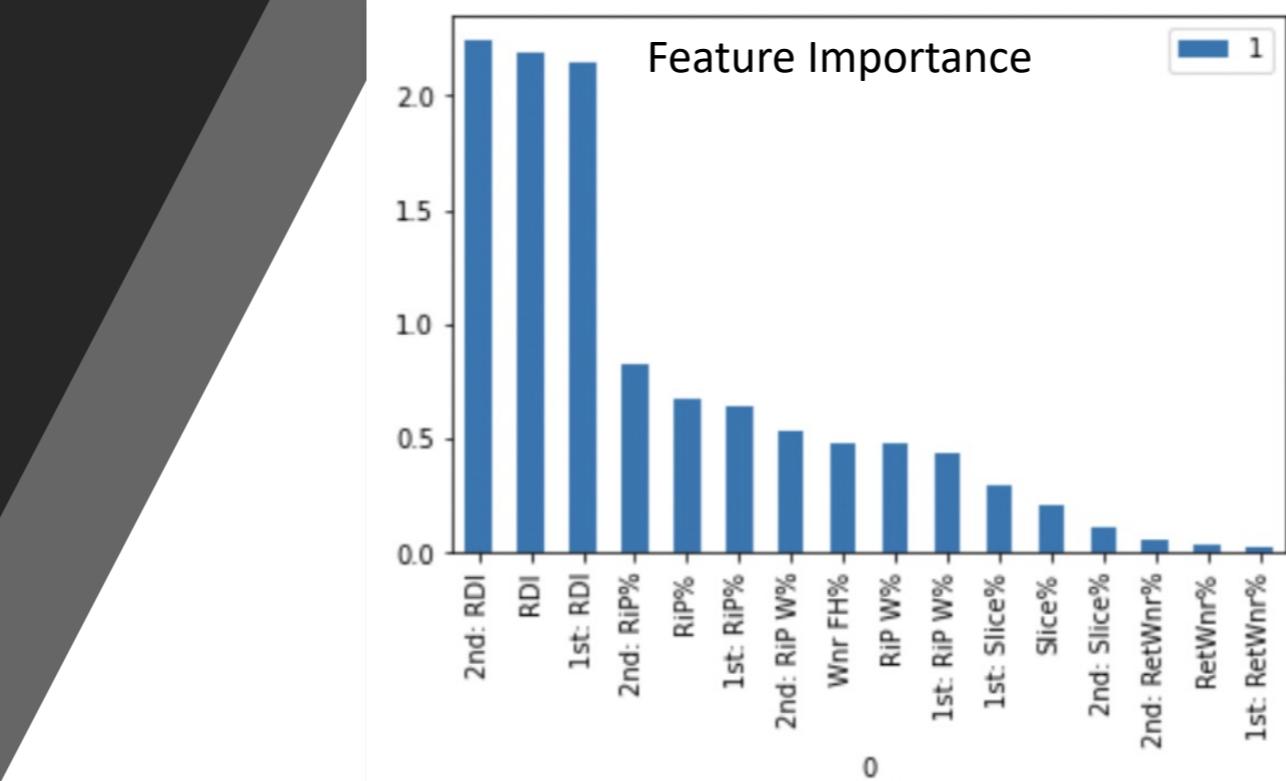
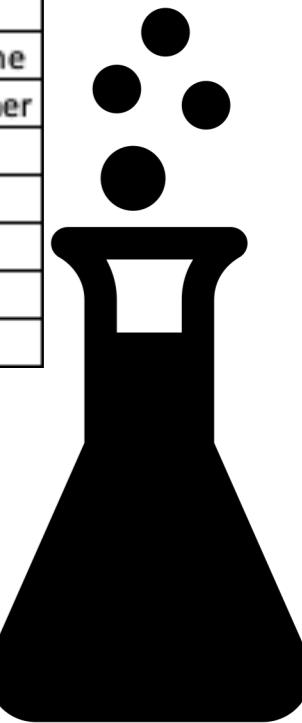
# FEATURE SELECTION



## 3. Experimental

Feature Combinations

Shot Type	Shot Description	From Court	Direction	Outcome
1st serve	Forehand	Deuce court	Deuce court	Fault
2nd serve	Backhand	Ad court	Middle court	No outcome
Return	Forehand Volley	Middle court	Ad court	Serve winner
Rally	Slice			
	Smash			
	Backhand lob			
	Forehand drop shot			
	...			

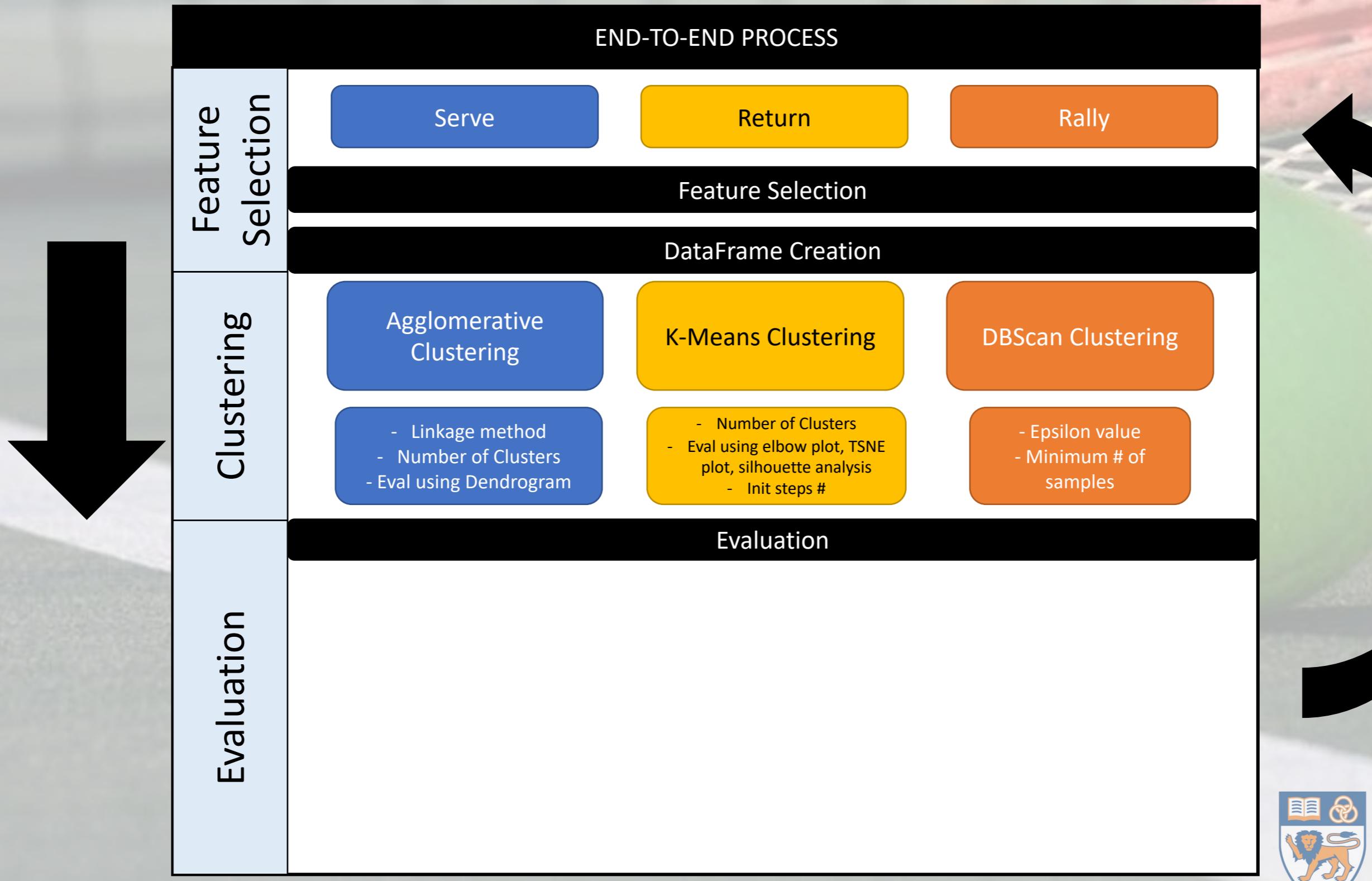


Feature Variance

feature: 0 Mean difference: 1.69 Total Variance: 459.464  
feature: 1 Mean difference: 0.879 Total Variance: 449.829  
feature: 2 Mean difference: 66.213 Total Variance: 154932.184  
feature: 3 Mean difference: 2.076 Total Variance: 2731.953

feature	percentage
0	middle court-BH_Error
1	middle court-BH_cross_court
2	middle court-BH_down_mid
3	middle court-BH_inside_out
4	middle court-FH_Error
...	...
19	ad court-BH_down_mid
20	ad court-FH_Error
21	ad court-FH_down_mid
22	ad court-FH_inside_in
23	ad court-FH_inside_out

# PHASE 2: CLUSTERING



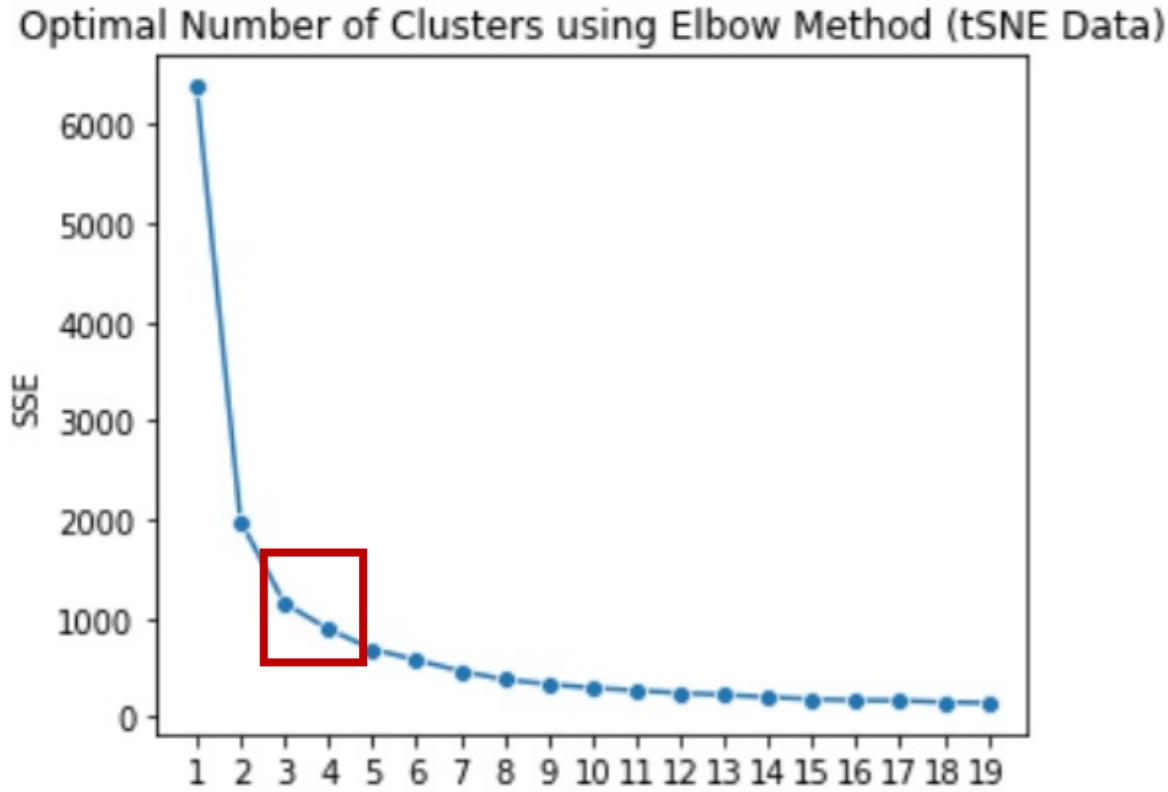
# K-Means Clustering

1. Decide on # of clusters (k)
2. Randomly assign centroid to each cluster
3. Calculate distance of each observation (player's feature) to each of the k centroids
4. Assign each observation to closest centroid
5. Find new location of centroid (mean of all observations in cluster)
6. Repeat steps 3-5 until centroids do not change position

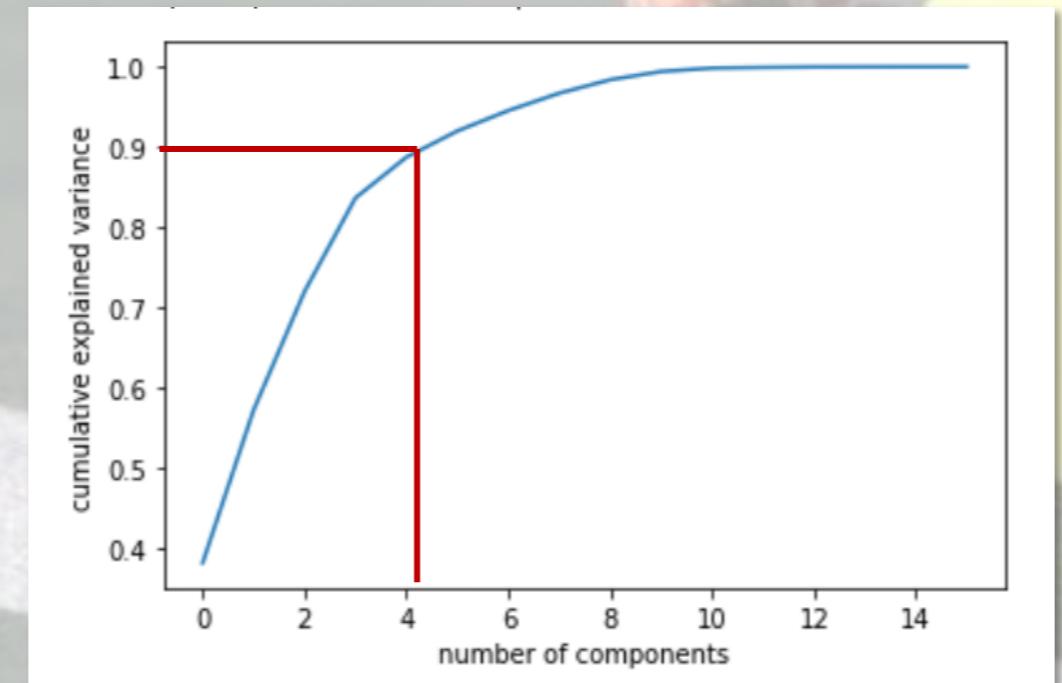
\*K-Means minimizes **Intra-cluster Distance**

# K-Means Clustering: Considerations

## Elbow Method:

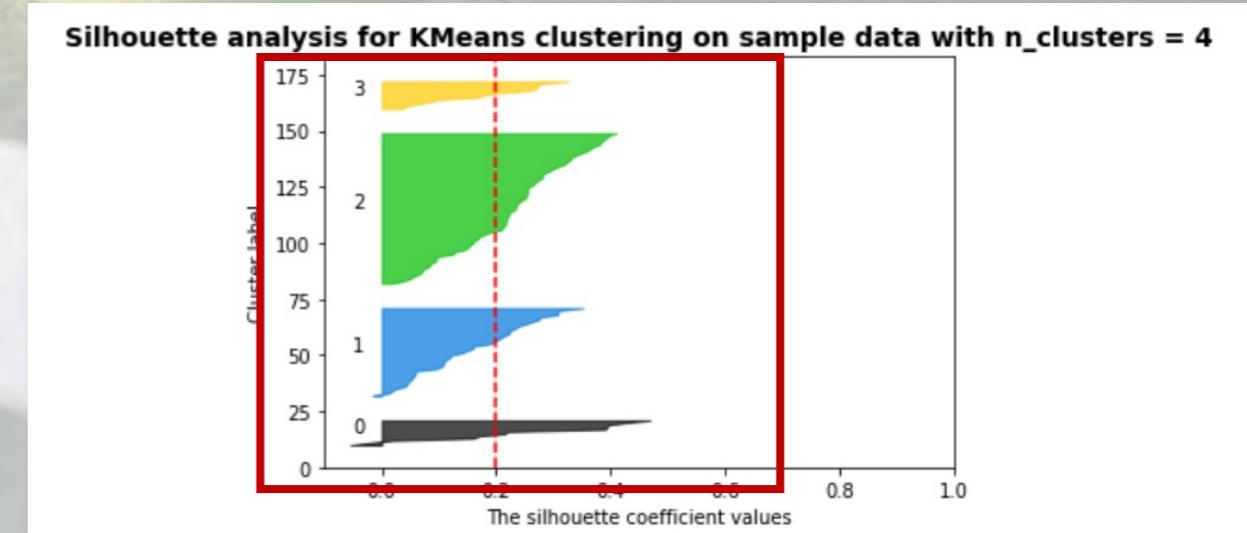
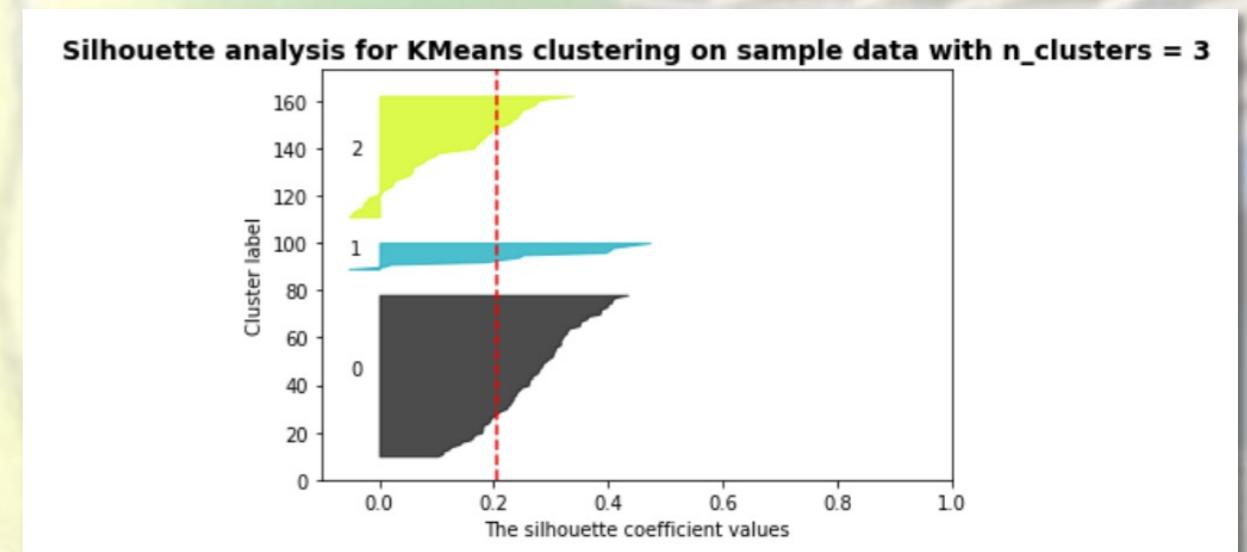
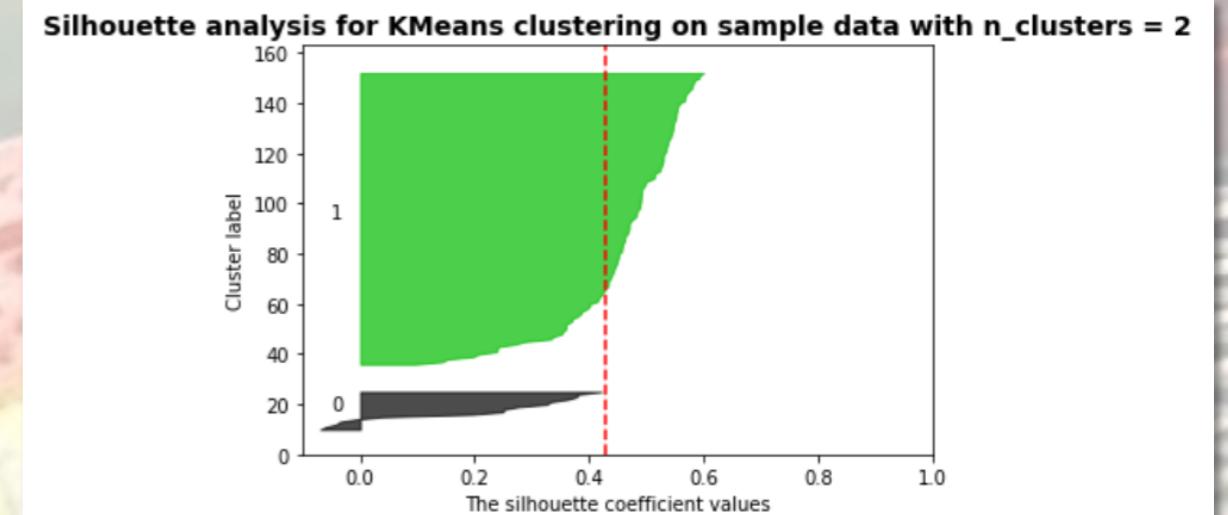


## PCA Analysis:



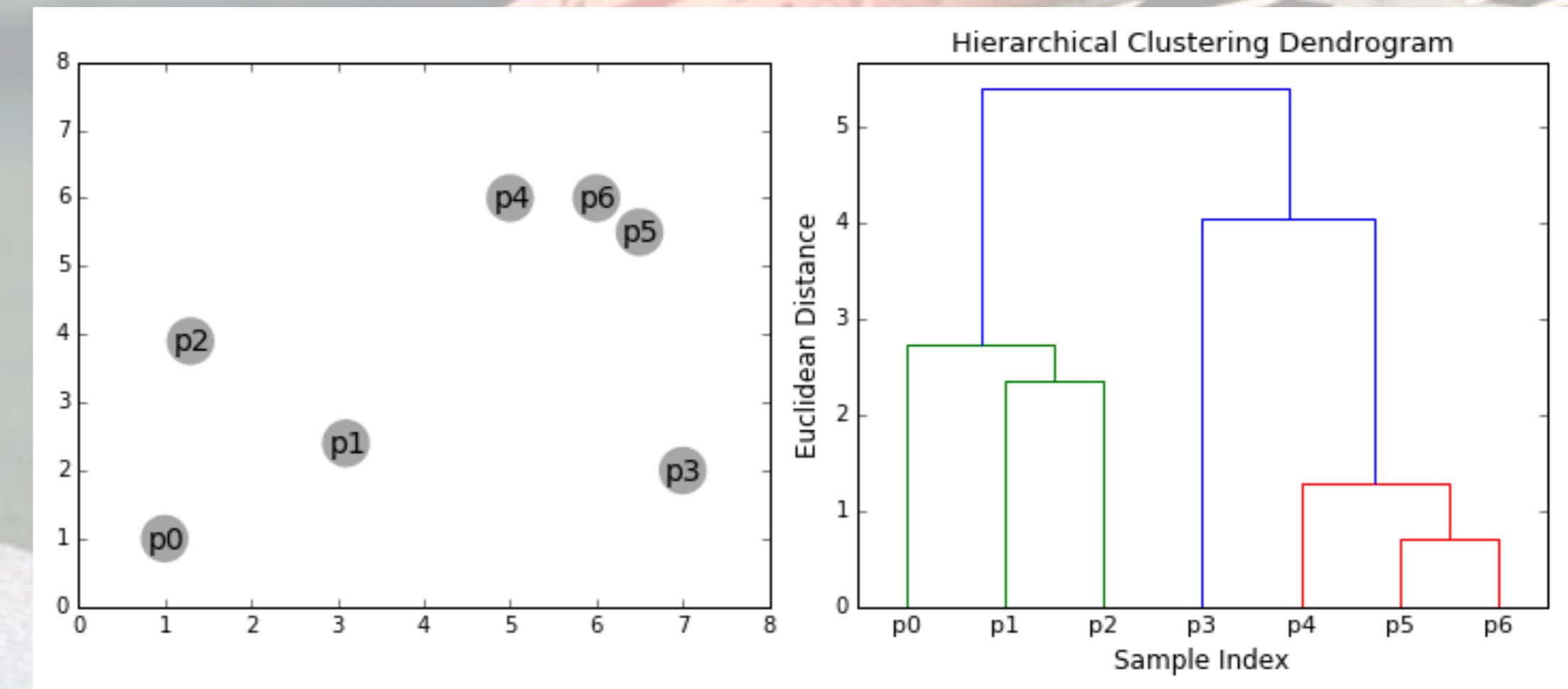
	PC1	PC2	PC3	PC4	PC5	PC6
Standard Deviation	2.468874	1.746458	1.538812	1.367278	0.901955	0.723117
Proportion of Variance	0.380959	0.190632	0.147996	0.116841	0.050845	0.032681
Cumulative Proportion	0.380959	0.571591	0.719587	0.836428	0.887273	0.919954

## Silhouette Analysis:



# Agglomerative Clustering

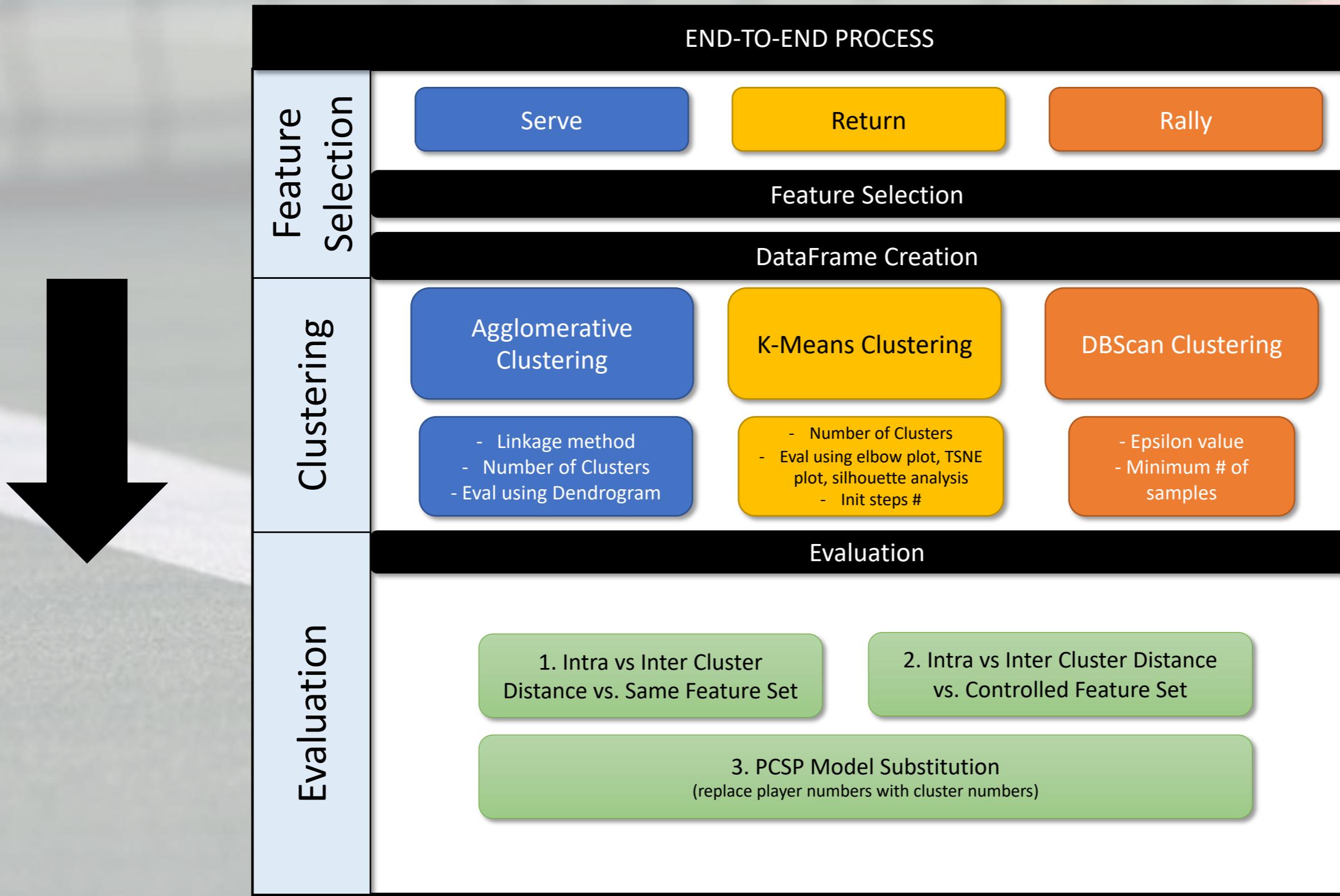
1. Determine distance measurement (eg. Euclidean distance)
2. Determine the linkage criteria to merge clusters (eg. Ward, Complete, Single)
3. Every data point starts as its own cluster
4. Continually merge the nearest clusters
5. Repeat the process until every data point become one cluster



## Considerations:

- Where to “cut” the dendrogram
- Linkage method: Ward

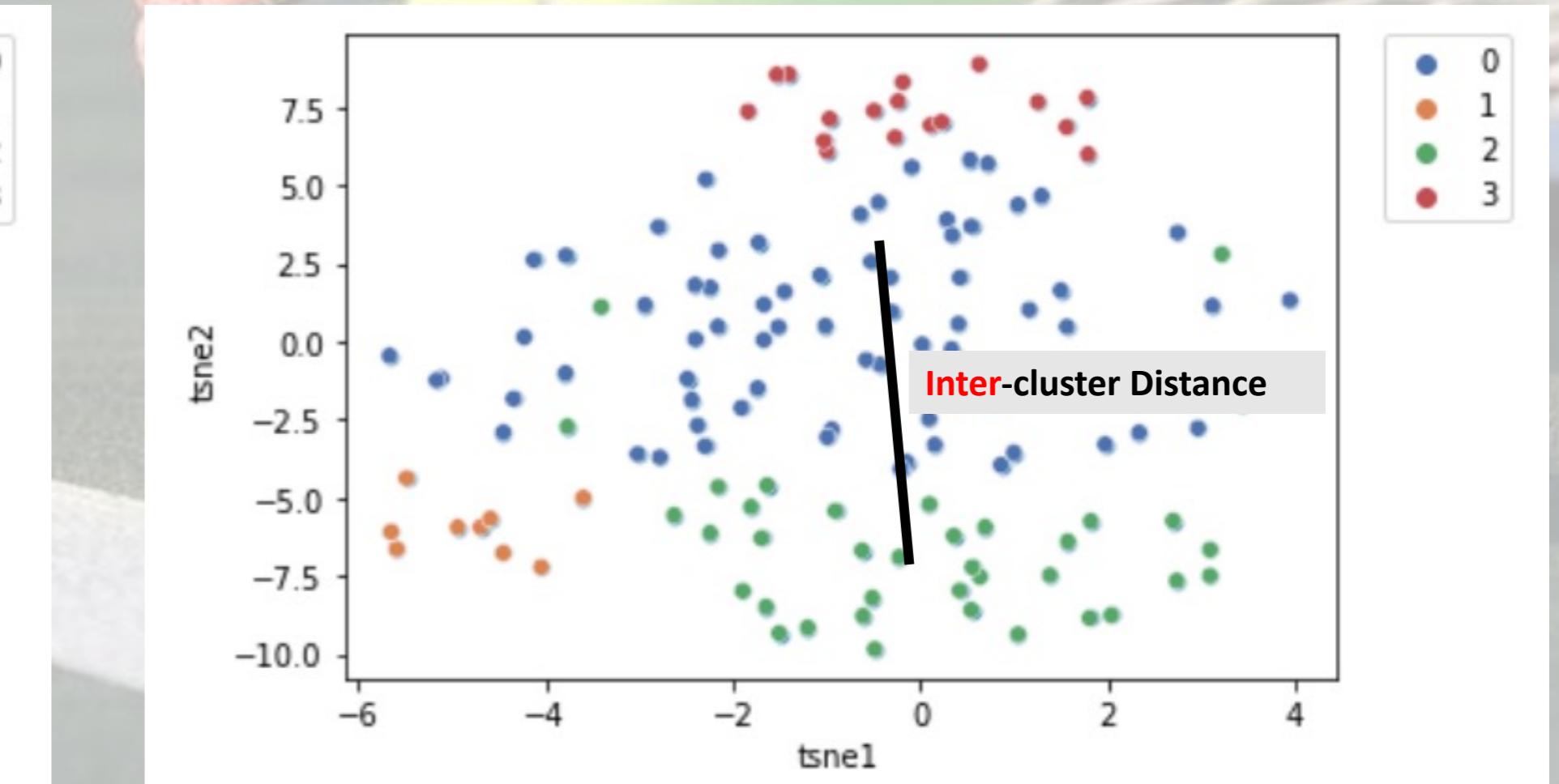
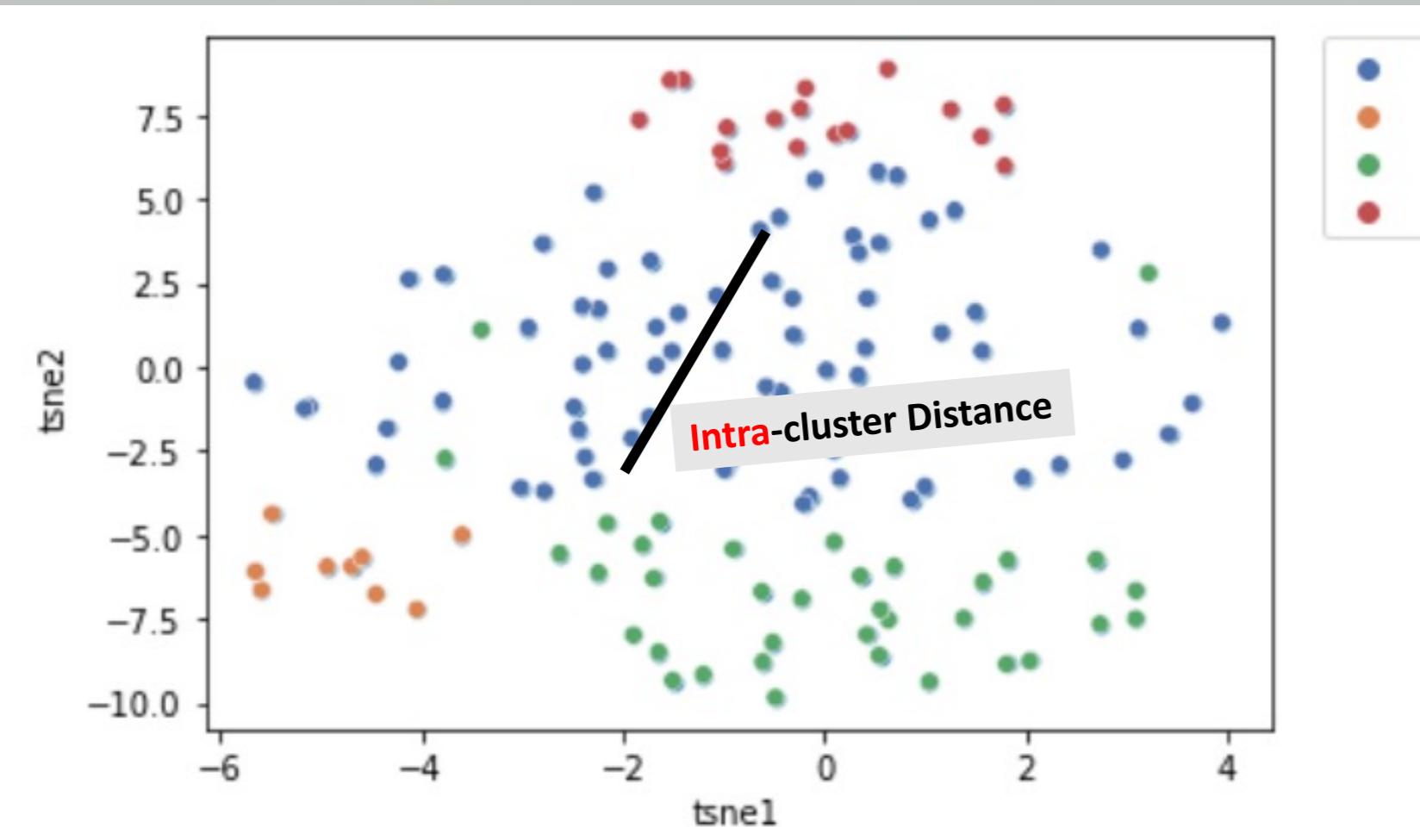
# PHASE 3: EVALUATION



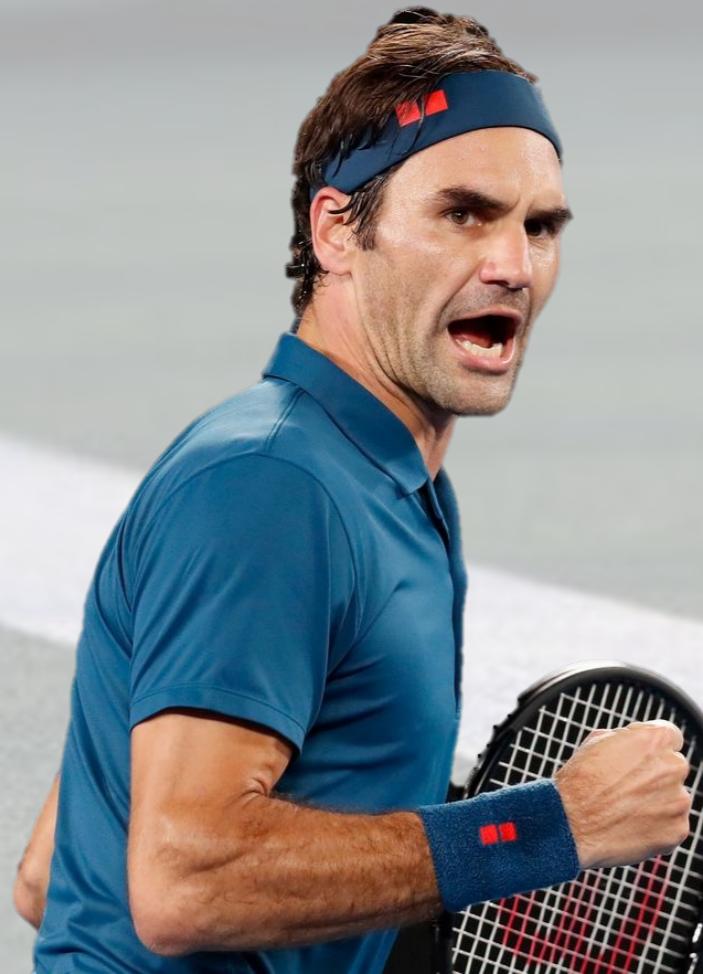
# EVALUATION #1 & #2

## Metric: Intra-Cluster Distance & Inter-Cluster Distance

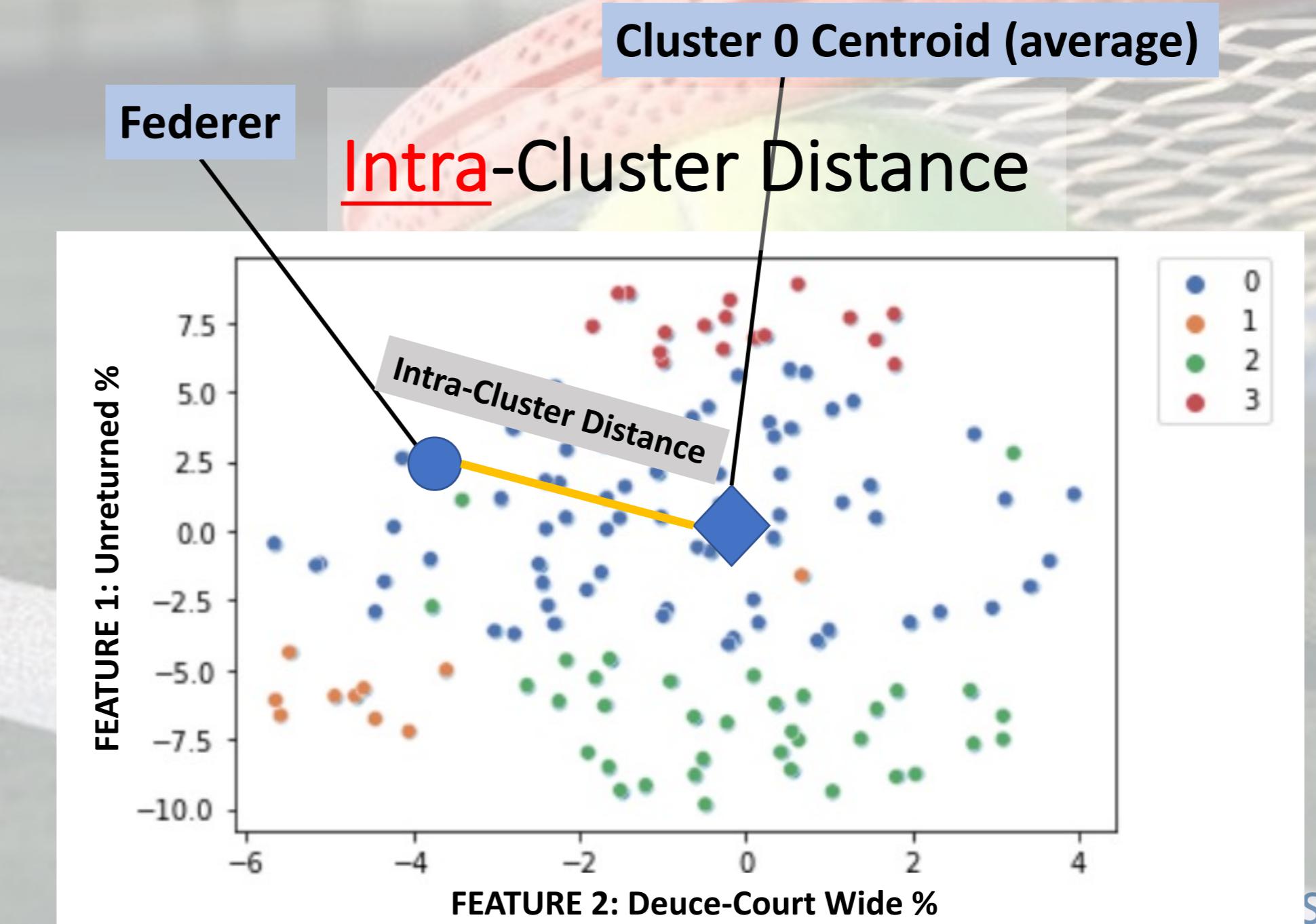
- #1: Using Same Feature Set
- #2: Using Different Feature Set for Evaluation



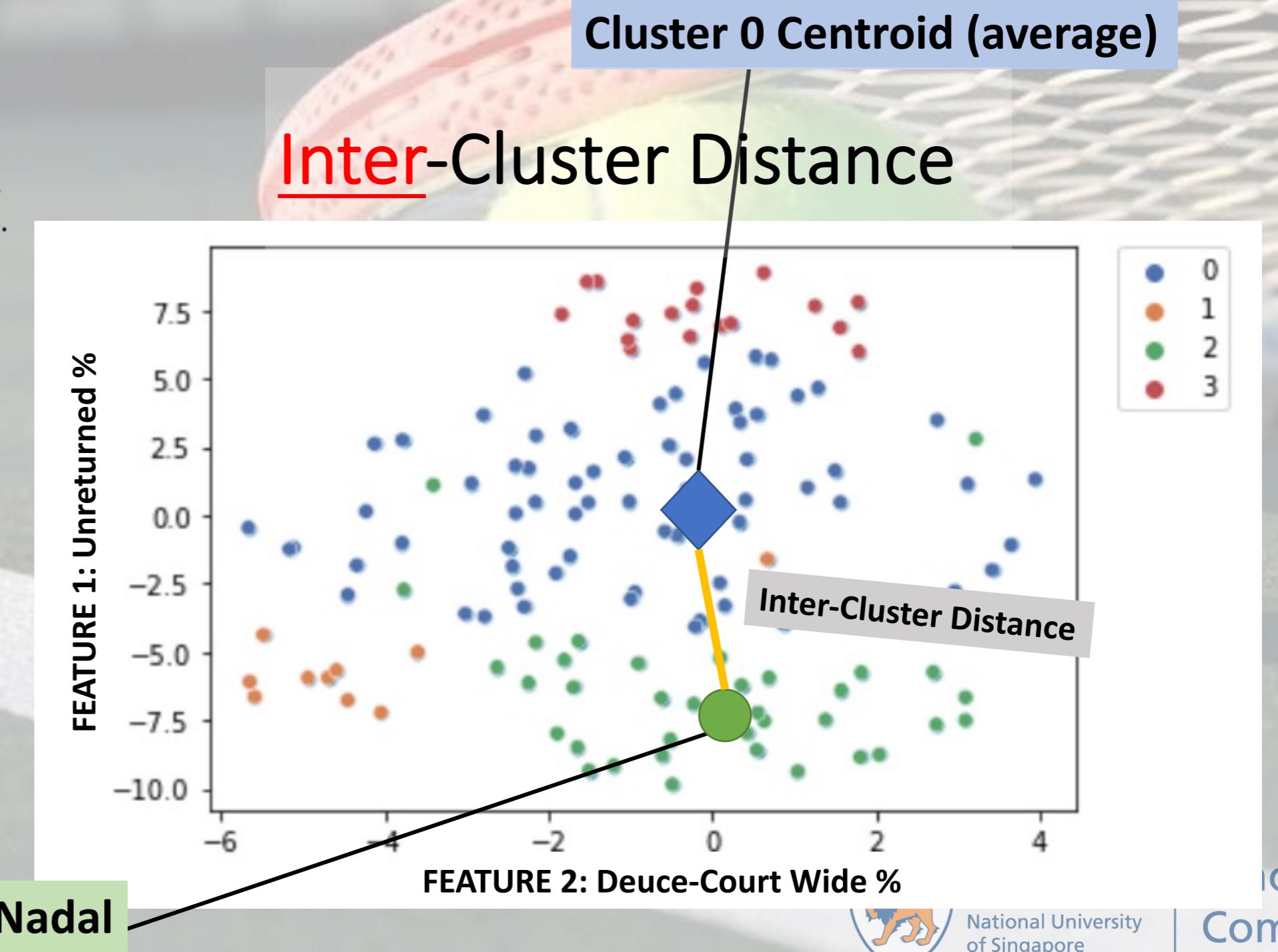
# Evaluation #1: Intra-Inter Cluster Distance using Same Feature Set



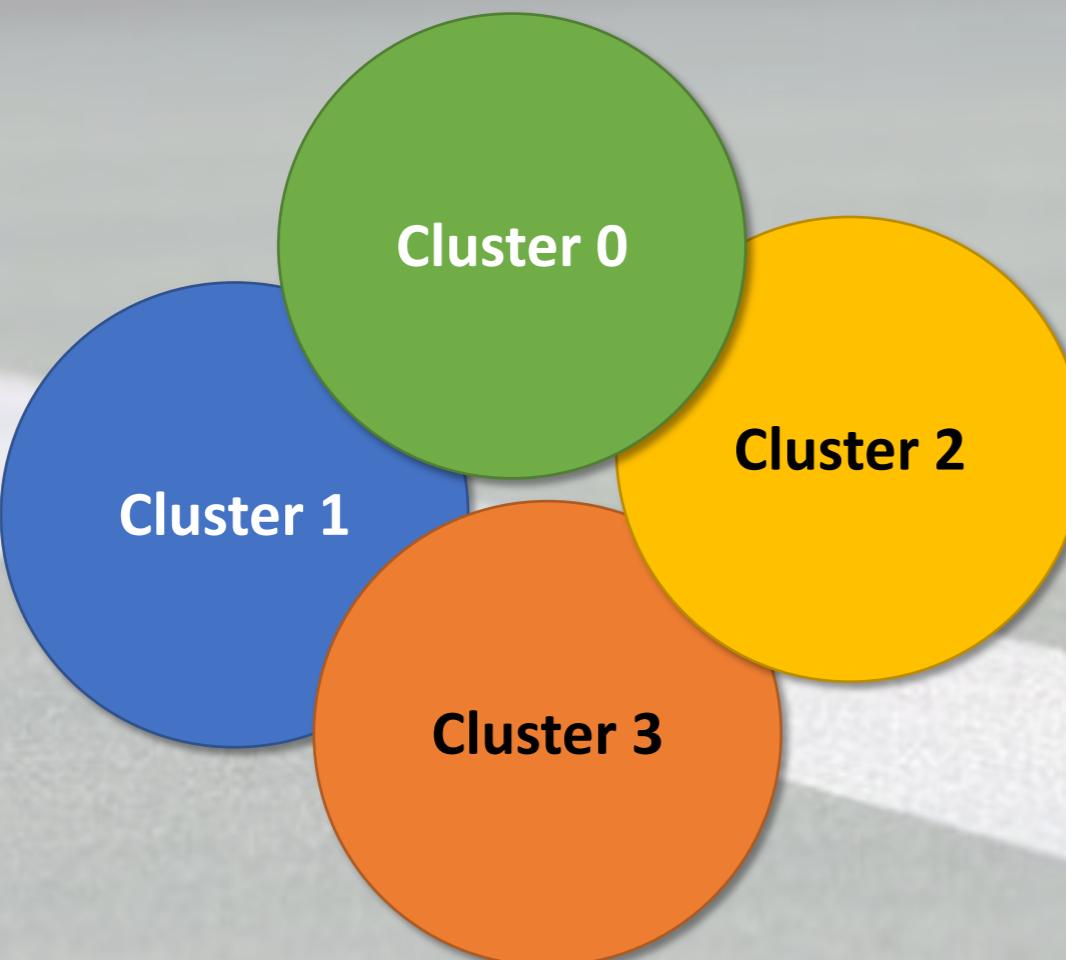
Roger Federer (Serve Cluster 0)



# Evaluation #1: Intra-Inter Cluster Distance using Same Feature Set



# Evaluation #1: Intra-Inter Cluster Distance using Same Feature Set



---Evaluating Cluster 0---  
Size of Cluster: 70  
Intra-cluster Distance Avg: 10.06  
Inter-cluster Distance Avg: 27.78  
% Difference (Inter - Intra) for Cluster 0: 176.10%

---Evaluating Cluster 1---  
Size of Cluster: 10  
Intra-cluster Distance Avg: 10.14  
Inter-cluster Distance Avg: 41.94  
% Difference (Inter - Intra) for Cluster 1: 313.73%

---Evaluating Cluster 2---  
Size of Cluster: 36  
Intra-cluster Distance Avg: 9.80  
Inter-cluster Distance Avg: 32.95  
% Difference (Inter - Intra) for Cluster 2: 236.37%

---Evaluating Cluster 3---  
Size of Cluster: 17  
Intra-cluster Distance Avg: 10.68  
Inter-cluster Distance Avg: 52.43  
% Difference (Inter - Intra) for Cluster 3: 391.07%

Overall Average of % Difference (Inter-Intra): 279.32%

# Evaluation #2: Intra-Inter Cluster Distance using Controlled (Evaluation) Feature Set

## CLUSTER FEATURE SET

Rally Game	Rally Definition
<b>Matches</b>	Matches logged by the Match Charting Project
<b>RallyLen</b>	Average rally length
<b>RLen-Serve</b>	Average rally length on serve points
<b>RLen-Return</b>	Average rally length on return points
<b>1-3 W%</b>	Win percentage on points between 1 and 3 shots
<b>4-6 W%</b>	Win percentage on points between 4 and 6 shots
<b>7-9 W%</b>	Win percentage on points between 7 and 9 shots
<b>10+ W%</b>	Win percentage on points of at least 10 shots
<b>FH/GS</b>	Forehands per groundstroke
<b>BH Slice%</b>	Slices per backhand groundstroke
<b>FHP/Match</b>	Forehand Potency per match (higher is better)
<b>FHP/100</b>	Forehand Potency per 100 forehands
<b>BHP/Match</b>	Backhand Potency per match (higher is better)
<b>BHP/100</b>	Backhand Potency per 100 backhands

## EVALUATION FEATURE SET

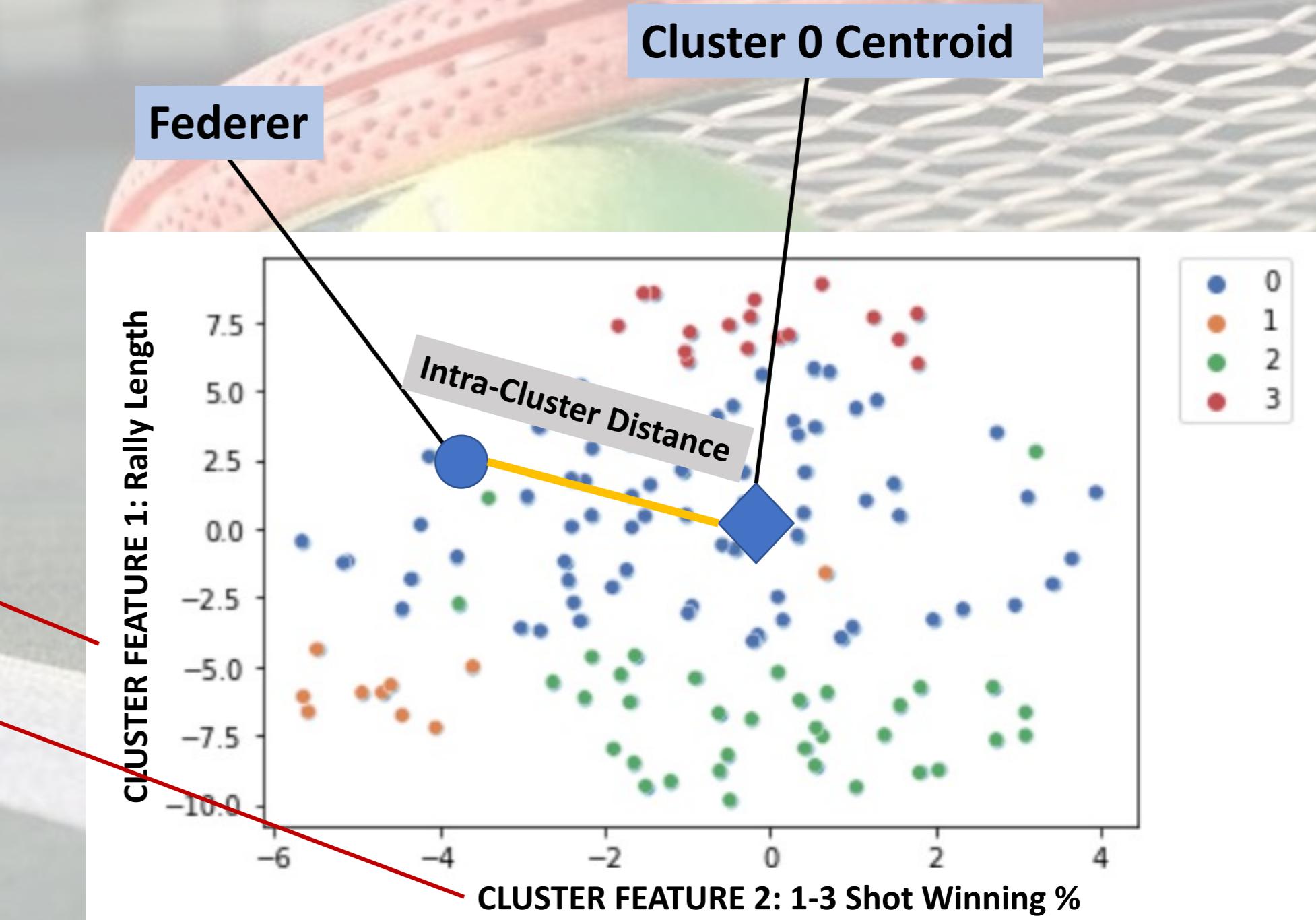
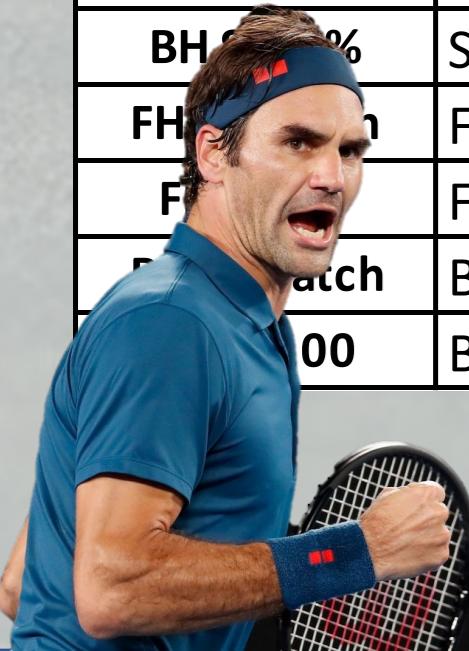
	feature	percentage
0	middle court-BH_Error	0.044010
1	middle court-BH_cross_court	0.120941
2	middle court-BH_down_mid	0.109487
3	middle court-BH_inside_out	0.063157
4	middle court-FH_Error	0.080256
...	...	...
19	ad court-BH_down_mid	0.189106
20	ad court-FH_Error	0.032507
21	ad court-FH_down_mid	0.031118
22	ad court-FH_inside_in	0.054525
23	ad court-FH_inside_out	0.102736



# Evaluation #2: Intra-Inter Cluster Distance using Controlled (Evaluation) Feature Set

## During Clustering:

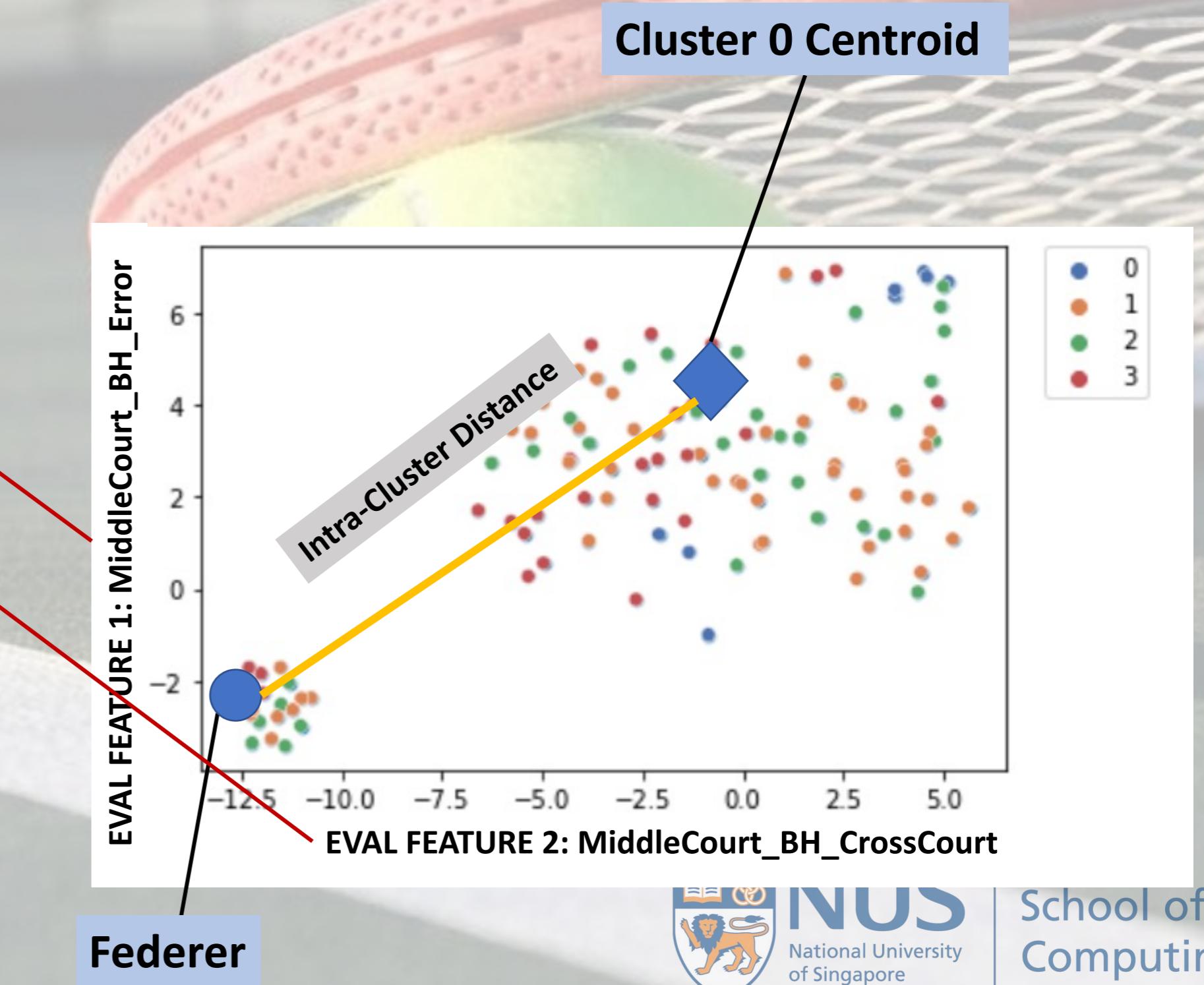
Rally Game	Rally Definition
Matches	Matches logged by the Match Charting Project
RallyLen	Average rally length
RLen-Serve	Average rally length on serve points
RLen-Return	Average rally length on return points
1-3 W%	Win percentage on points between 1 and 3 shots
4-6 W%	Win percentage on points between 4 and 6 shots
7-9 W%	Win percentage on points between 7 and 9 shots
10+ W%	Win percentage on points of at least 10 shots
FH/GS	Forehands per groundstroke
BH Slices %	Slices per backhand groundstroke
FH Potency	Forehand Potency per match (higher is better)
FH Potency 100	Forehand Potency per 100 forehands
BH Potency	Backhand Potency per match (higher is better)
BH Potency 100	Backhand Potency per 100 backhands



# Evaluation #2: Intra-Inter Cluster Distance using Controlled (Evaluation) Feature Set

During Evaluating:

	feature	percentage
0	middle court-BH_Error	0.044010
1	middle court-BH_cross_court	0.120941
2	middle court-BH_down_mid	0.109487
3	middle court-BH_inside_out	0.063157
4	middle court-FH_Error	0.080256
...	...	...
19	ad court-BH_down_mid	0.189106
20	ad court-FH_Error	0.032507
21	ad court-FH_down_mid	0.031118
22	ad court-FH_inside_in	0.054525
	ad court-FH_inside_out	0.102736



## 2. Intra-inter Cluster Distance using Controlled Feature Set

Cluster Results:

---Evaluating Cluster 0---

Size of Cluster: 57

Intra-cluster Distance Avg: 28.14

Inter-cluster Distance Avg: 46.32

% Difference (Inter - Intra) for Cluster 0: 64.59%

-----

---Evaluating Cluster 1---

Size of Cluster: 7

Intra-cluster Distance Avg: 24.95

Inter-cluster Distance Avg: 55.82

% Difference (Inter - Intra) for Cluster 1: 123.70%

-----

---Evaluating Cluster 2---

Size of Cluster: 33

Intra-cluster Distance Avg: 27.58

Inter-cluster Distance Avg: 49.83

% Difference (Inter - Intra) for Cluster 2: 80.68%

-----

---Evaluating Cluster 3---

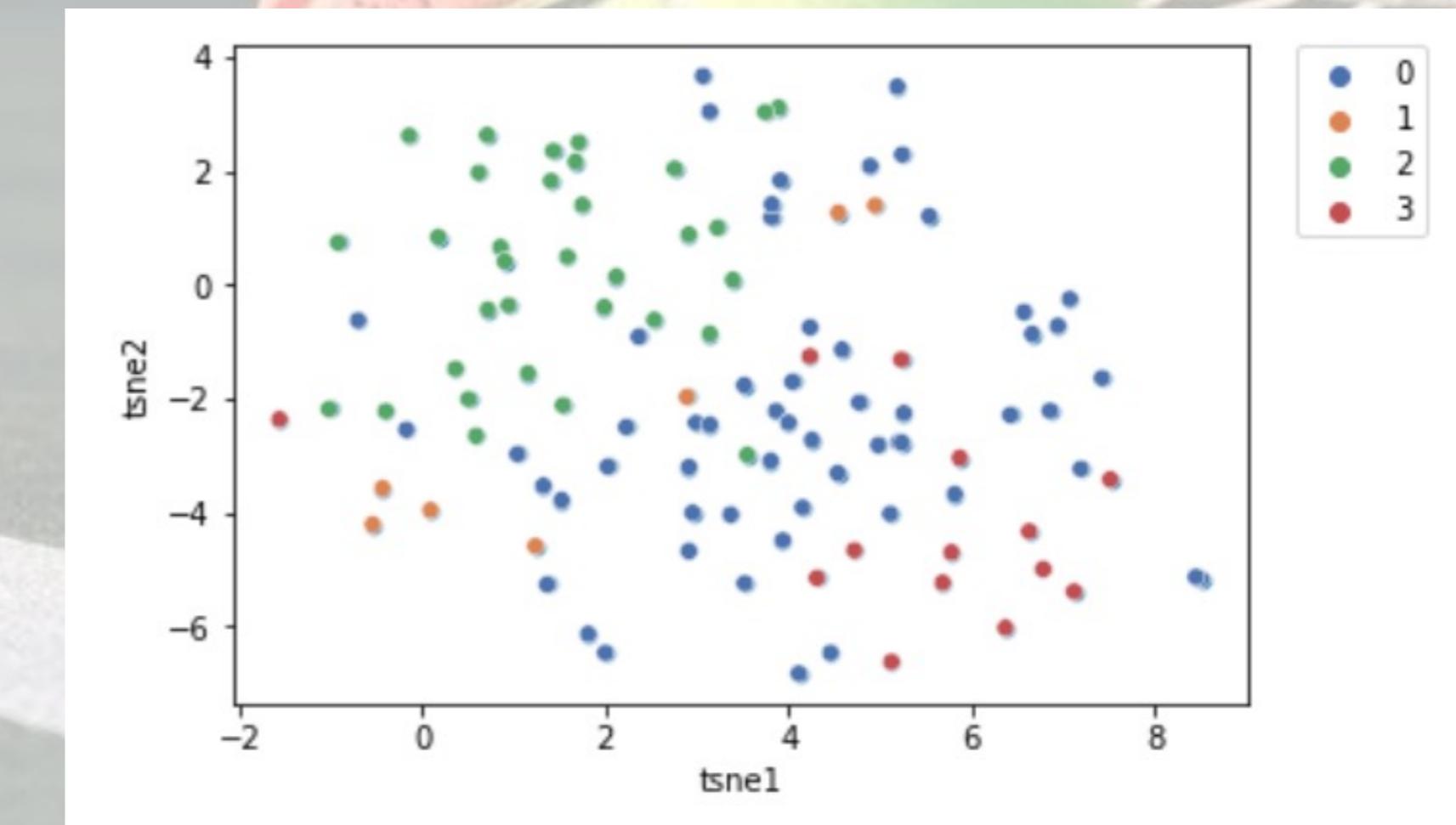
Size of Cluster: 14

Intra-cluster Distance Avg: 45.72

Inter-cluster Distance Avg: 54.84

% Difference (Inter - Intra) for Cluster 3: 19.93%

Overall Average of % Difference (Inter-Intra): 72.22%



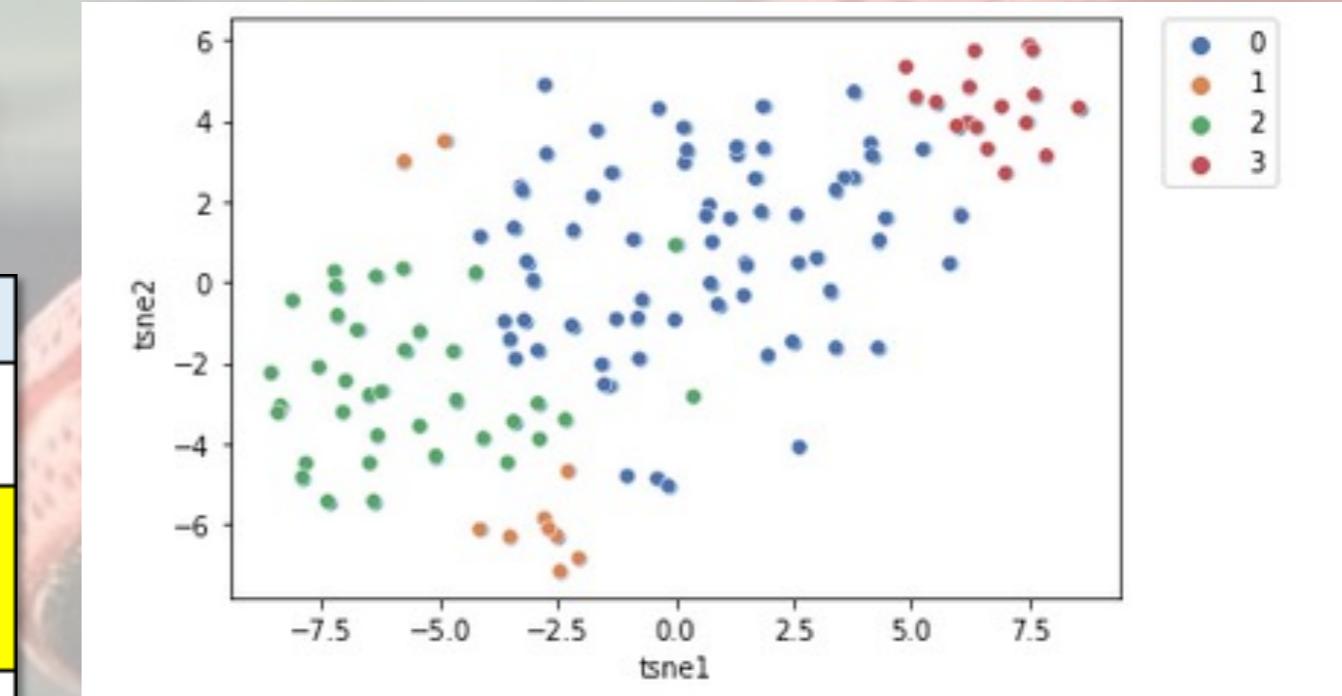
# EXPERIMENT RESULTS



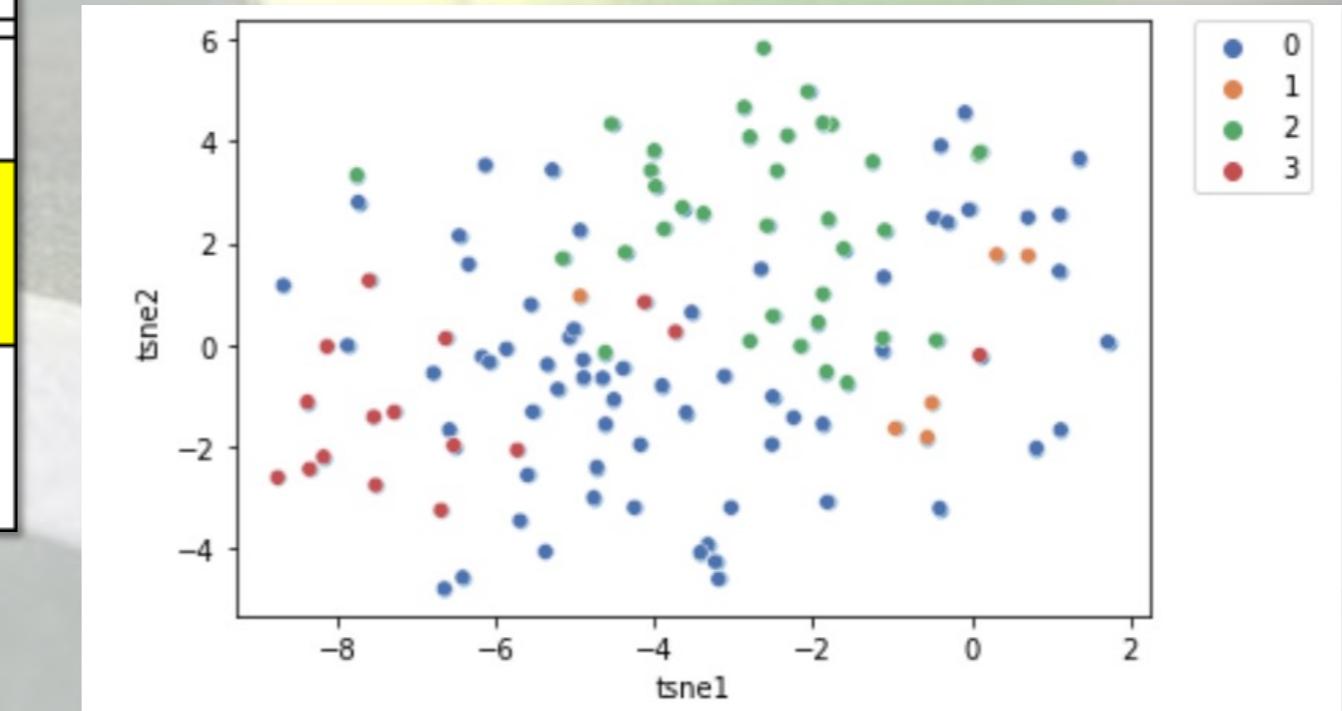
# RESULTS: SERVE GAME

Serve Cluster			
Evaluation #1	Cluster Method	Cluster Feature Set	Evaluation Feature Set
	Overall Average of % Difference (Inter-Intra)		
Evaluation #1	Agglomerative, Ward linkage, 4 clusters	19 summary-level features	19 summary-level features
	279.32%		
Evaluation #2	Kmeans, init=100, 4 clusters	19 summary-level features	19 summary-level features
	242.79%		
Serve Cluster			
Evaluation #2	Cluster Method	Cluster Feature Set	Evaluation Feature Set
	Overall Average of % Difference (Inter-Intra)		
Evaluation #2	Agglomerative, Ward linkage, 4 clusters	19 summary-level features	36 PCSP serve function features
	72.22%		
Evaluation #2	Kmeans, init=100, 4 clusters	19 summary-level features	36 PCSP serve function features
	54.72%		

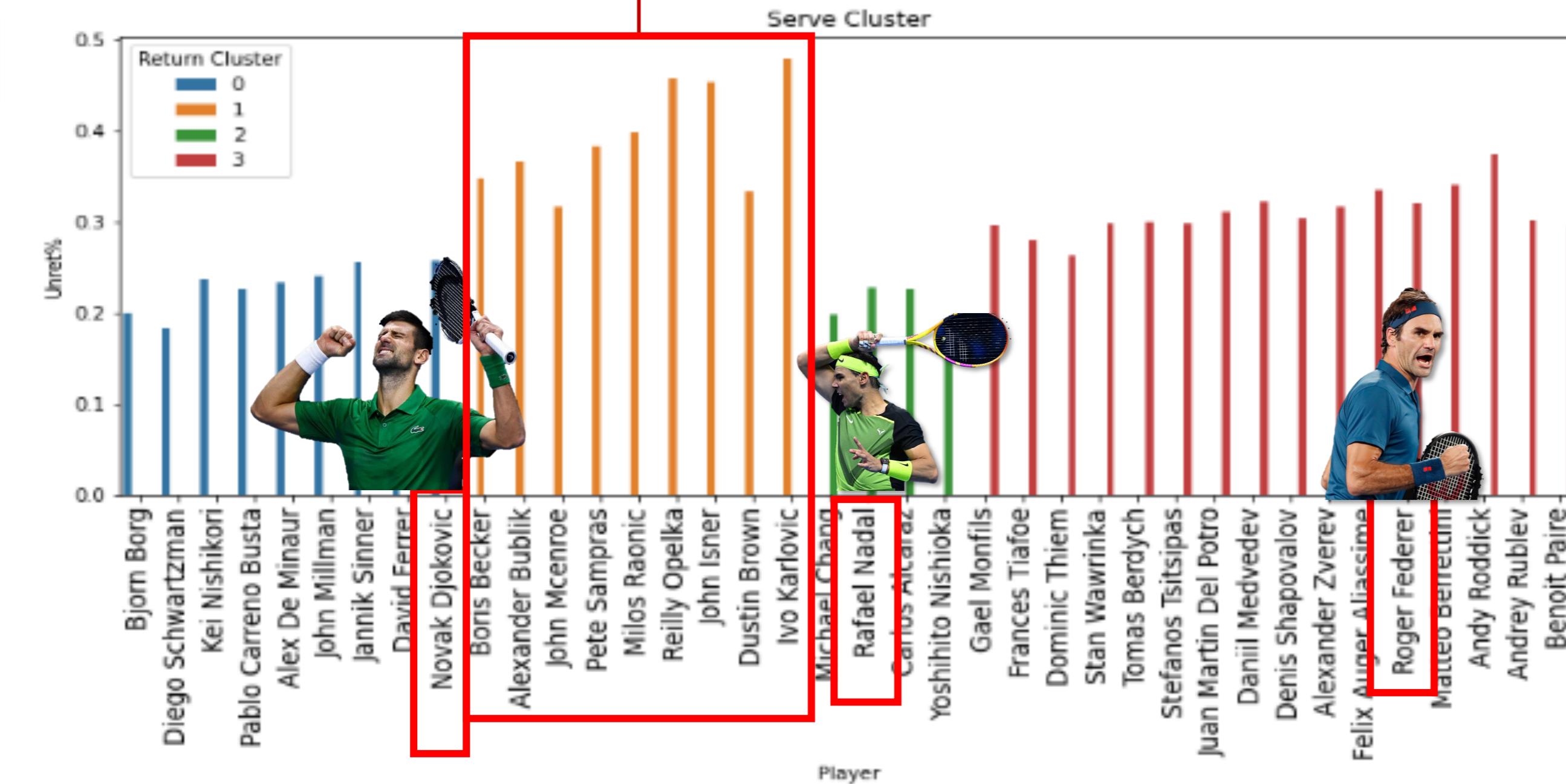
Eval #1 using Agglo



Eval #2 using Agglo



# RESULTS: SERVE GAME

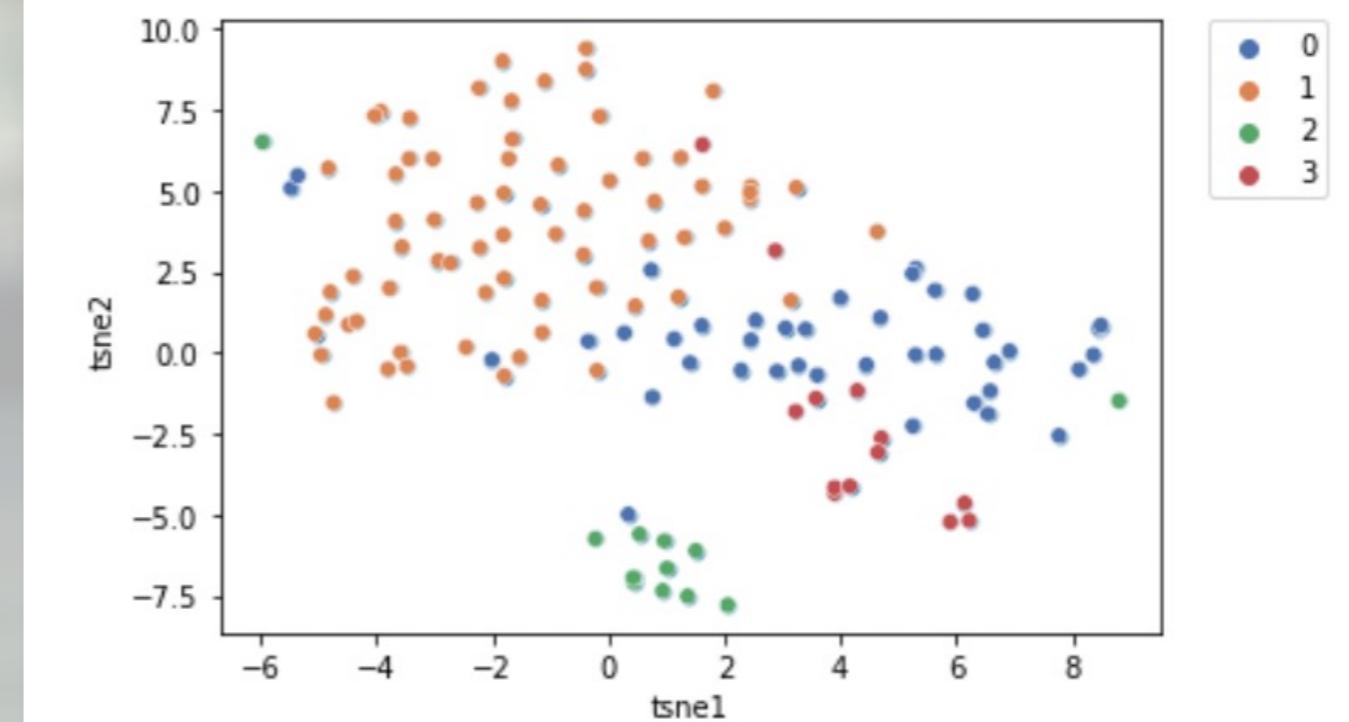


Cluster 3: BIG SERVERS

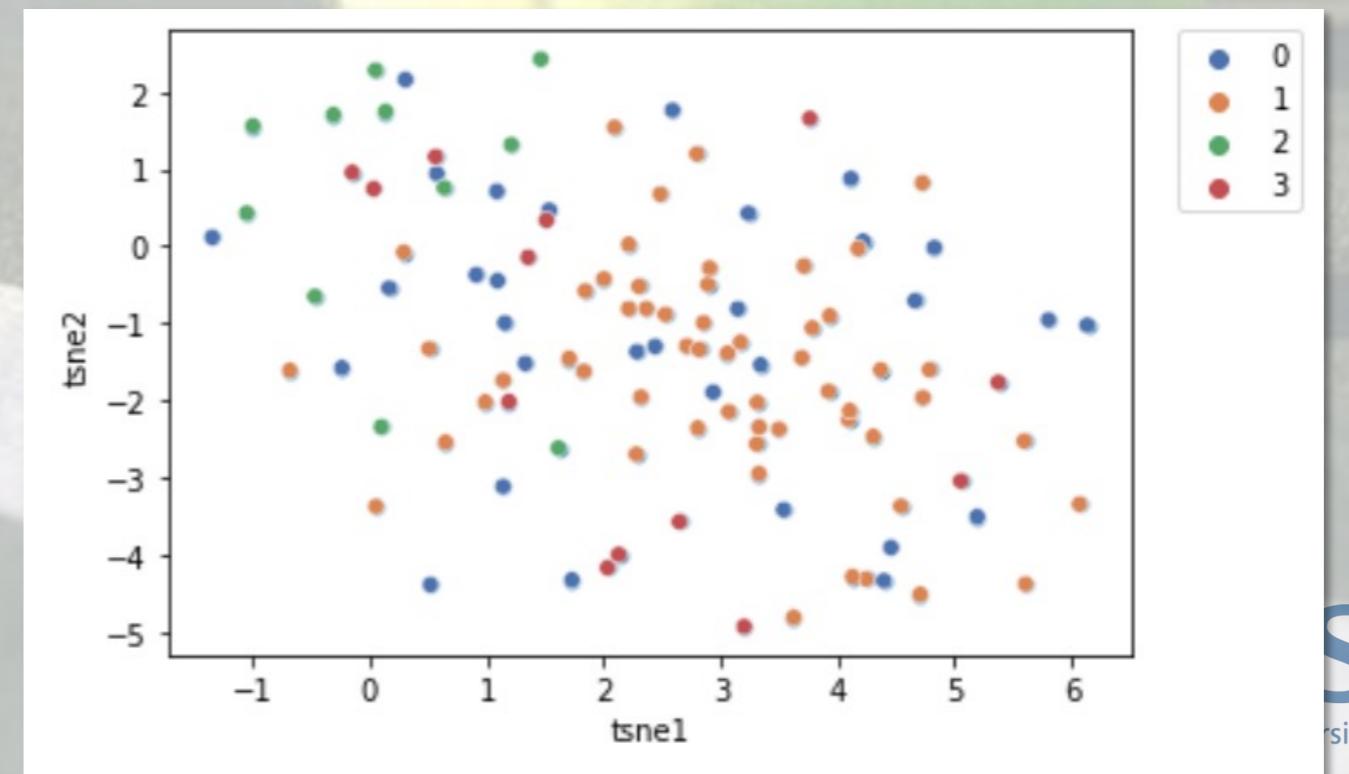
# RESULTS: RETURN GAME

Return Cluster			
Evaluation #1	Cluster Method	Cluster Feature Set	Evaluation Feature Set
	Overall Average of % Difference (Inter-Intra)		
Evaluation #1	Agglomerative, Ward linkage, 4 clusters	16 summary-level features	16 summary-level features
	Kmeans, init=100, 4 clusters	16 summary-level features	16 summary-level features
Evaluation #2	Cluster Method	Cluster Feature Set	Evaluation Feature Set
	Agglomerative, Ward linkage, 4 clusters	16 summary-level features	72 PCSP return function features
Evaluation #2	Kmeans, init=100, 4 clusters	16 summary-level features	72 PCSP return function features
			254.57%

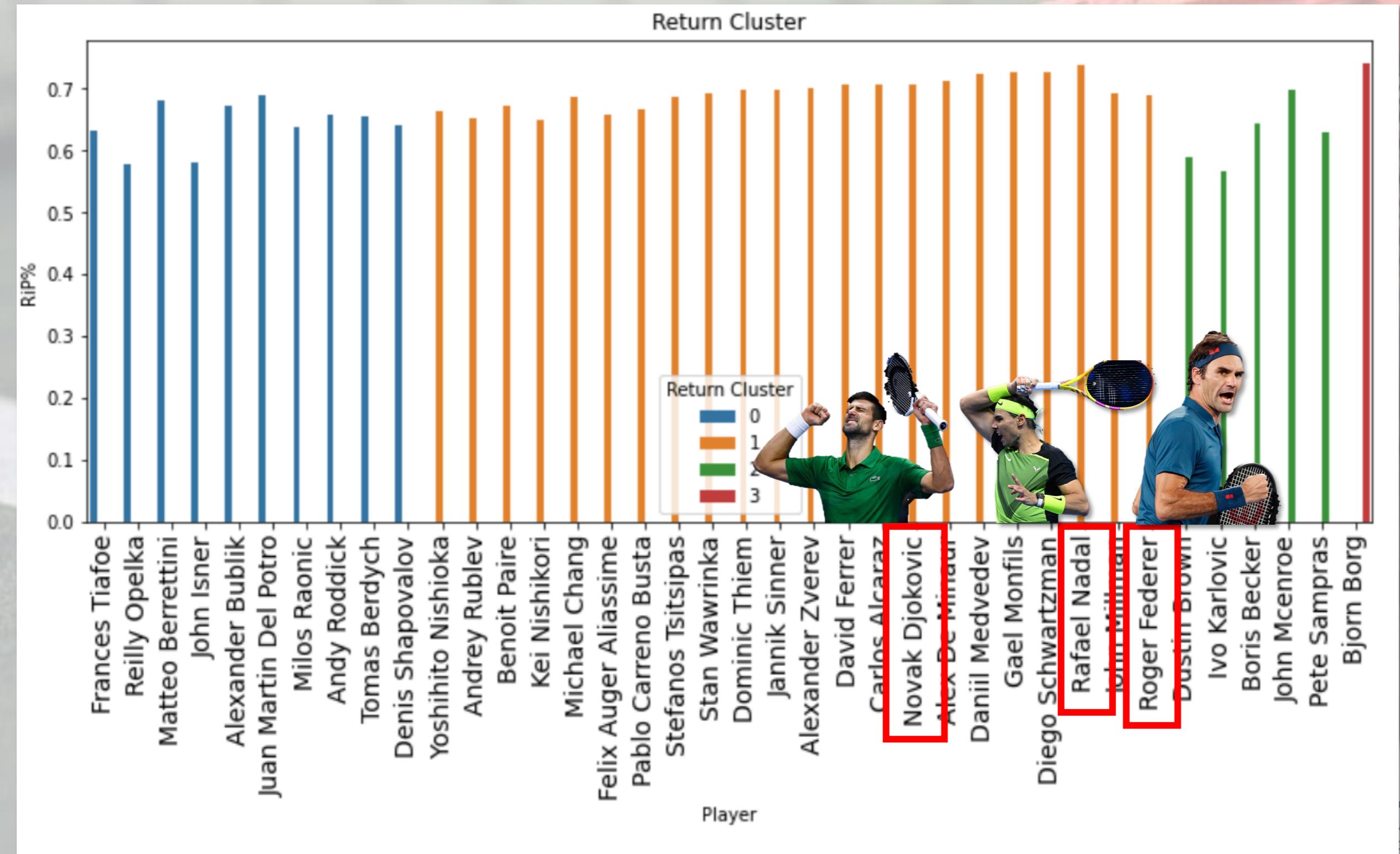
Eval #1 using K-Means



Eval #2 using K-Means



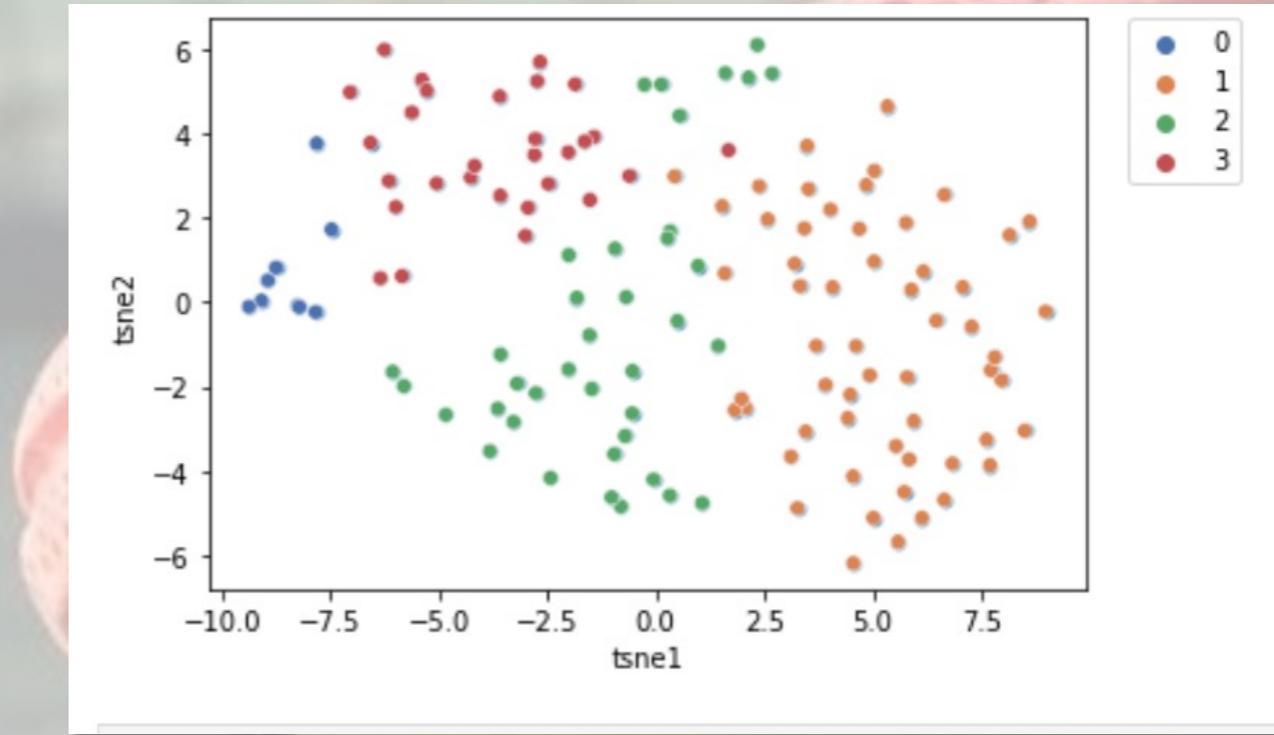
# RESULTS: RETURN GAME



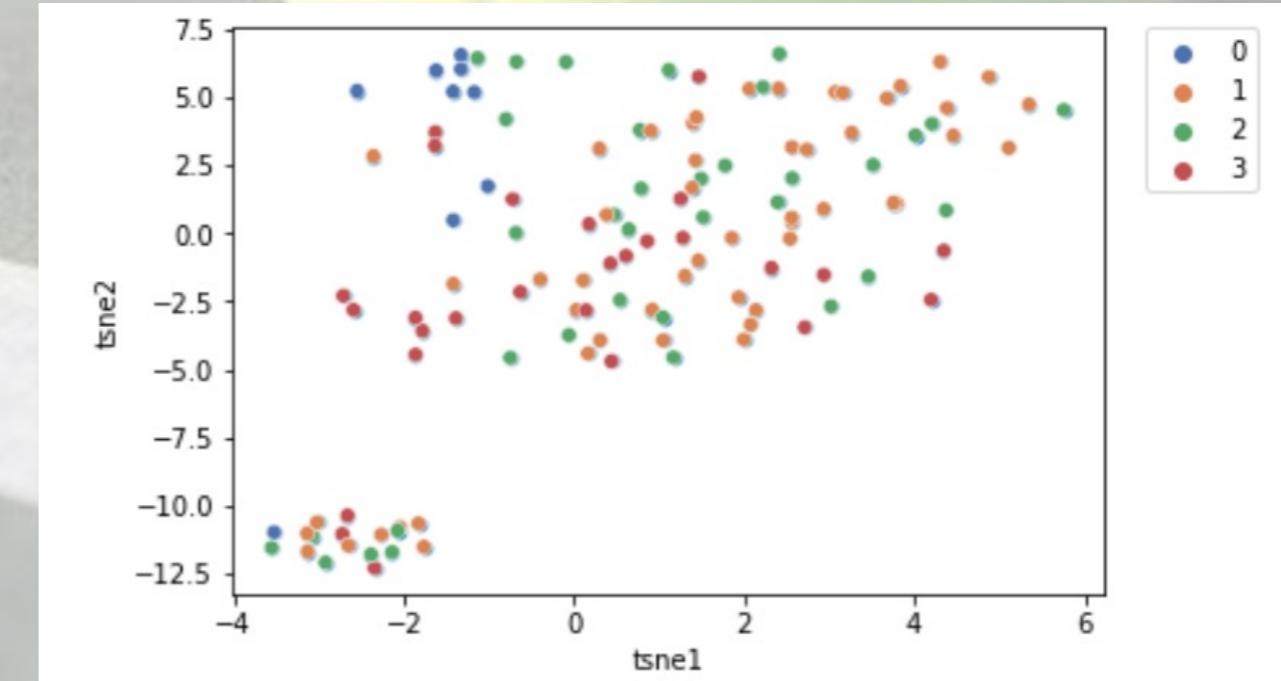
# RESULTS: RALLY GAME

Rally Cluster				
Evaluation #1	Cluster Method	Cluster Feature Set	Evaluation Feature Set	Overall Average of % Difference (Inter-Intra)
	Agglomerative, Ward linkage, 4 clusters	13 summary-level features	13 summary-level features	173.18%
Evaluation #2	Kmeans, init=100, 4 clusters	13 summary-level features	13 summary-level features	186.19%
	Cluster Method	Cluster Feature Set	Evaluation Feature Set	Overall Average of % Difference (Inter-Intra)
	Agglomerative, Ward linkage, 4 clusters	13 summary-level features	24 PCSP rally function features	34.66%
	Kmeans, init=100, 4 clusters	13 summary-level features	24 PCSP rally function features	31.90%

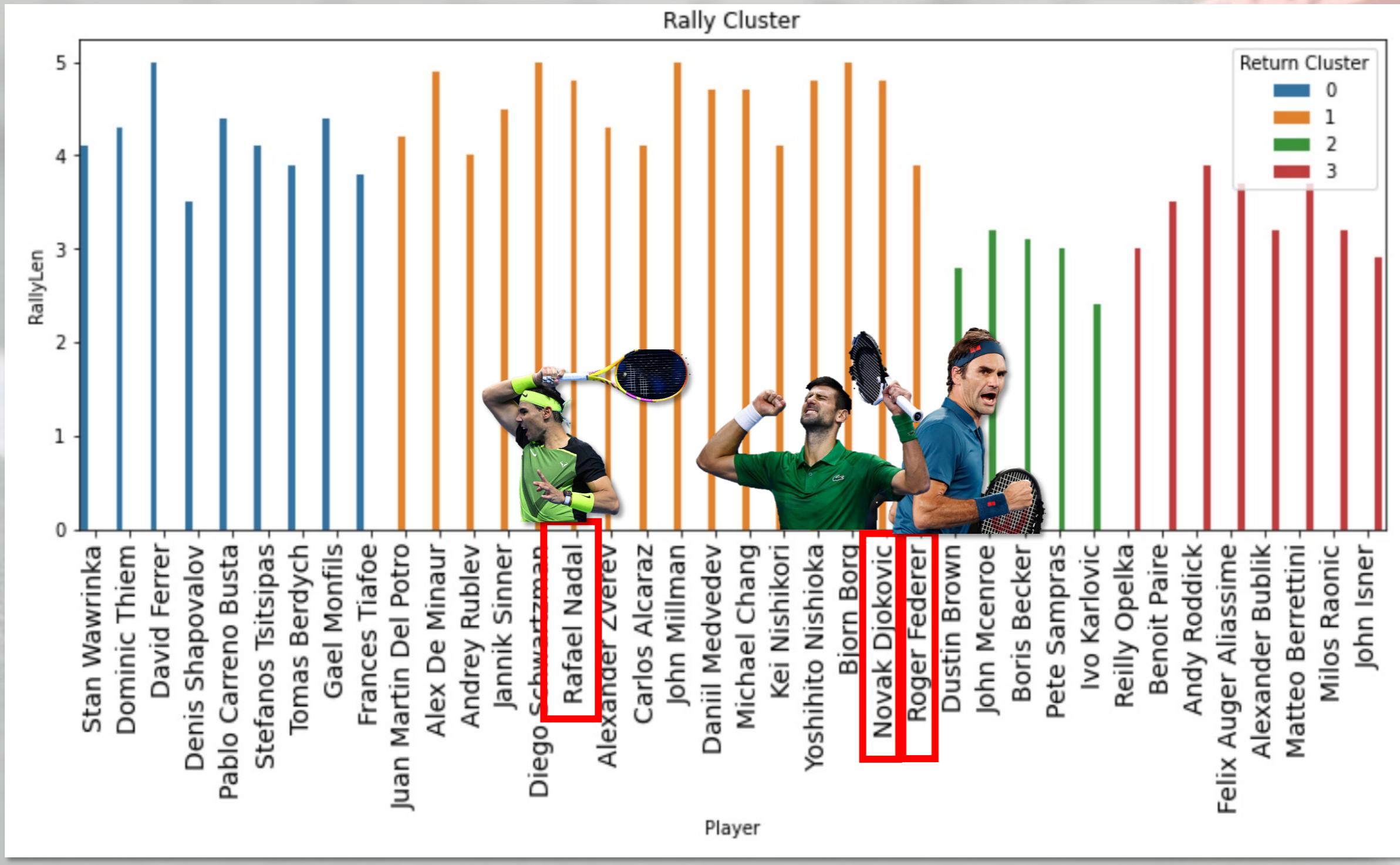
Eval #1 using K-Means



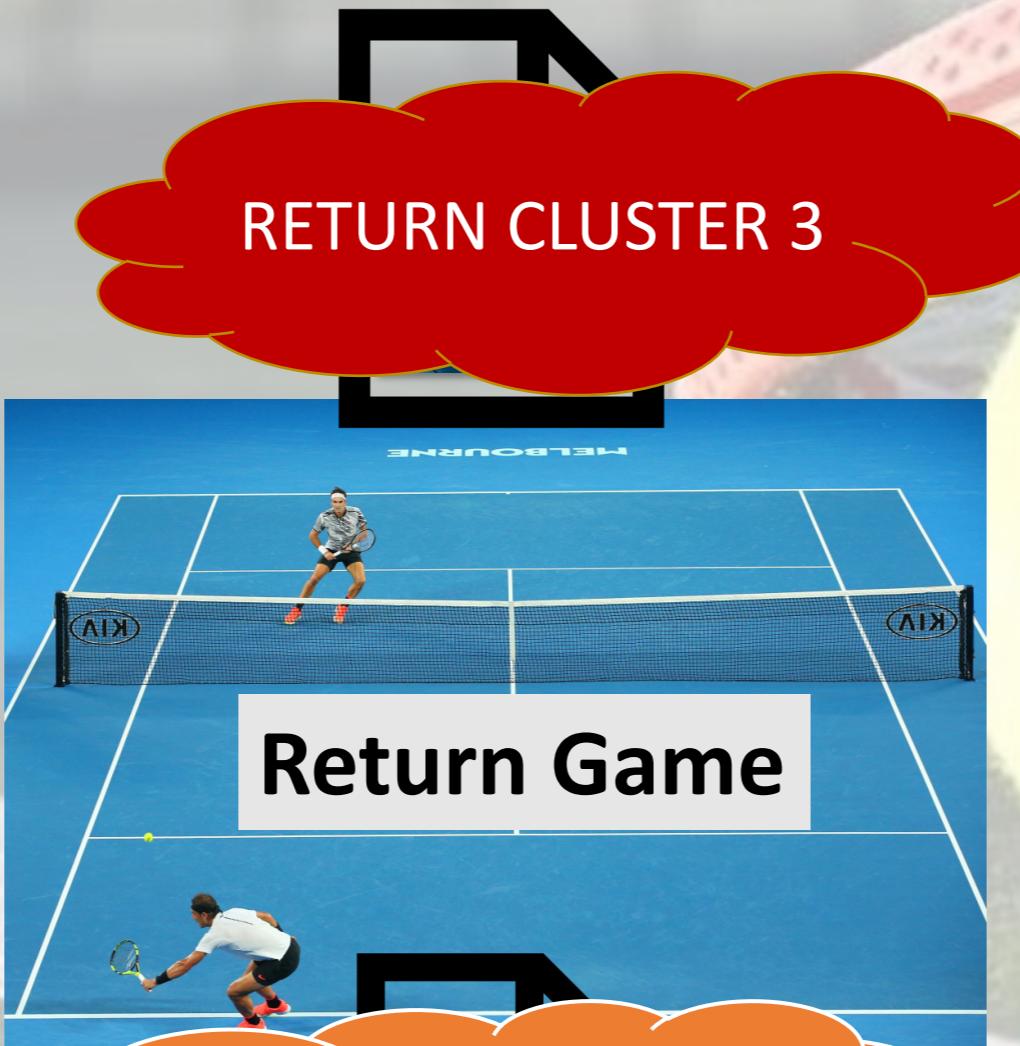
Eval #2 using K-Means



# RESULTS: RALLY GAME



# Evaluation #3: PCSP Data Replacement



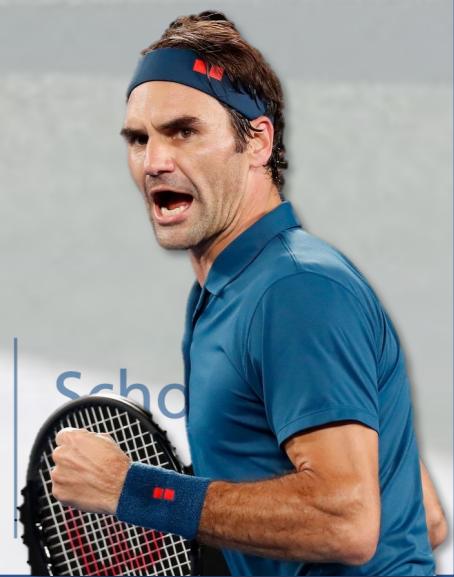
# Evaluation #3: PCSP Data Replacement

```
- De_Ply1Serve = pcase {                                     // all probability is ba:  
    231: ServeT_in{ball= 6} -> Ply2_BackHandR // T will have opponent  
    119: ServeT_err{ball=9} -> De_Ply1Serve_2nd // Federer tries to serve  
    337: ServeWide_in{ball =6} -> Ply2_ForeHandR  
    173: ServeWide_err{ball=9} -> De_Ply1Serve_2nd  
    92: ServeBody_in{ball=6} -> (Ply2_BackHandR [] Ply2_ForeHandR)  
    48: ServeBody_err{ball=9} -> De_Ply1Serve_2nd};  
  
- De_Ply1Serve = pcase {  
    30: ServeT_in{ball= 6} -> Ply2_BackHandR  
    17: ServeT_err{ball=9} -> De_Ply1Serve_2nd  
    29: ServeWide_in{ball =6} -> Ply2_ForeHandR  
    14: ServeWide_err{ball=9} -> De_Ply1Serve_2nd  
    7: ServeBody_in{ball=6} -> (Ply2_BackHandR [] Ply2_ForeHandR)  
    2: ServeBody_err{ball=9} -> De_Ply1Serve_2nd};
```



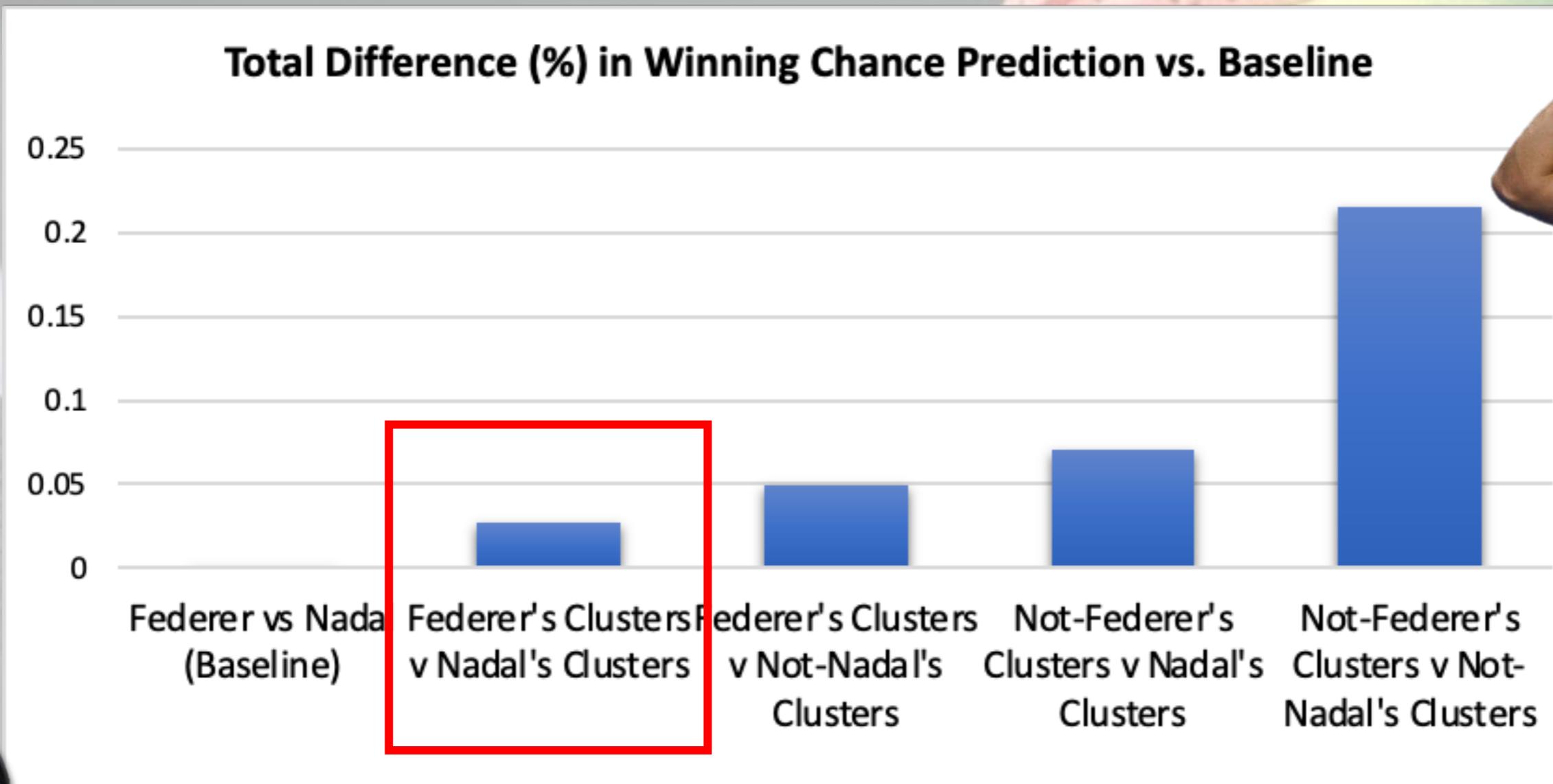
# RESULTS: Evaluation #3 - PCSP Data Replacement

#	Test Case	Game components				Results				Difference from Baseline			
		Ply1	Ply2	Serve	Return	Rally	Ply1 prob	Ply1 2 prob	Ply1 Avg	Ply2 Avg	Ply1 Difference	Ply2 Difference	Total Difference
0	Baseline	Federer	Nadal	Fed vs Nadal	Fed vs Nadal	Fed vs Nadal	[48.084, 50.94]	[49.059, 51.915]	0.49512	0.50487	-	-	-
1	Cluster v Cluster	Federer Cluster	Nadal Cluster	Fed's cluster (3) vs Nadal's cluster (2)	Fed's cluster (1) vs Nadal's cluster (1)	Fed's cluster (0) vs Nadal's cluster (1)	[0.46736, 0.5087];	[0.4913, 0.53263];	0.4880	0.5120	1.43%	-1.41%	2.67%
2	Cluster v Non-Cluster	Federer Cluster	Non-Nadal Cluster	Fed cluster (3) vs Isner cluster (1)	Fed cluster (1) vs Isner cluster (0)	Fed cluster (0) vs Medvedev cluster (2)	[0.48341, 0.53201];	[0.46798, 0.51658];	0.5077	0.4923	-2.54%	2.49%	4.91%
3	Non-Cluster v Cluster	Non-Federer Cluster	Nadal Cluster	John Isner cluster (1) vs Nadal cluster (2)	John Isner cluster (0) vs Nadal cluster (1)	Medvedev cluster (2) vs Nadal cluster (1)	[0.45157, 0.50242];	[0.49757, 0.54842];	0.4770	0.5230	3.66%	-3.59%	7.07%
4	Non-Cluster v Non-Cluster	Non-Federer Cluster	Non-Nadal Cluster	John Isner cluster (1) vs Mats Wilander cluster (0)	John Isner cluster (0) vs Karlovic cluster (2)	Medvedev cluster (2) vs Sampras cluster (3)	[0.51802, 0.58273];	[0.41726, 0.48197];	0.55038	0.44962	-11.16%	10.94%	21.55%



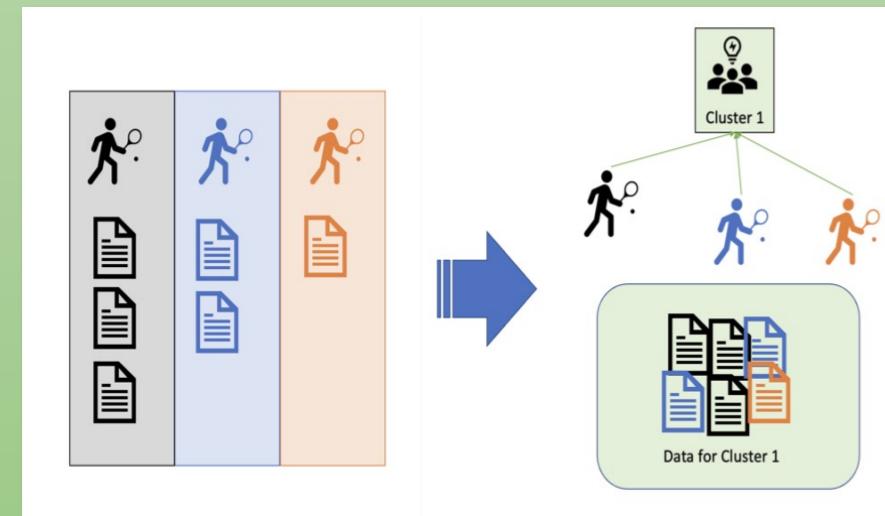
# CONCLUSION

Based on these results, clusters are a good representation of players themselves!

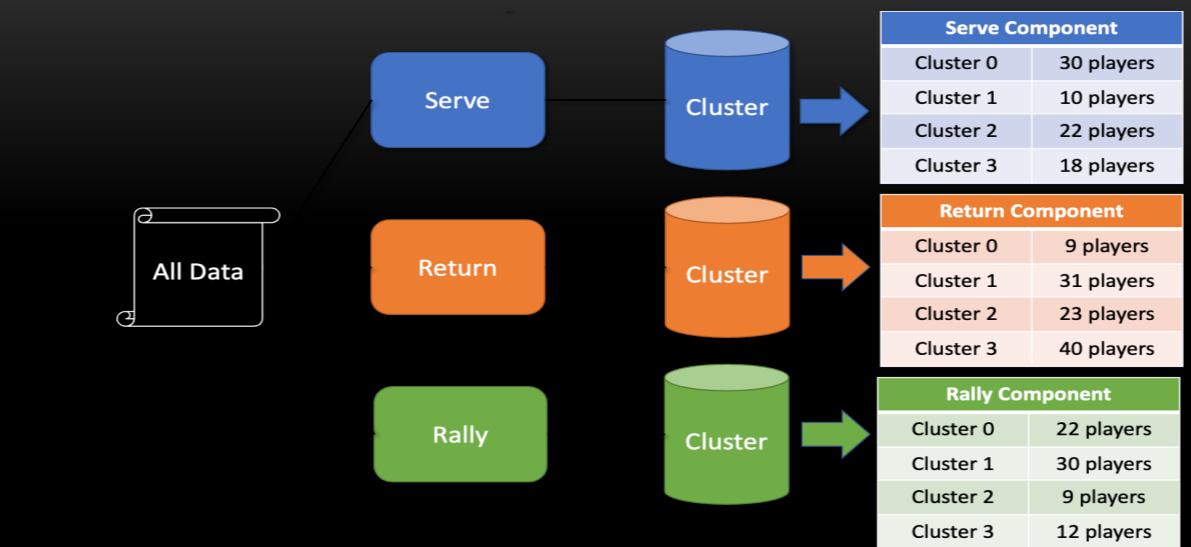


# TECHNICAL RESEARCH CONTRIBUTIONS

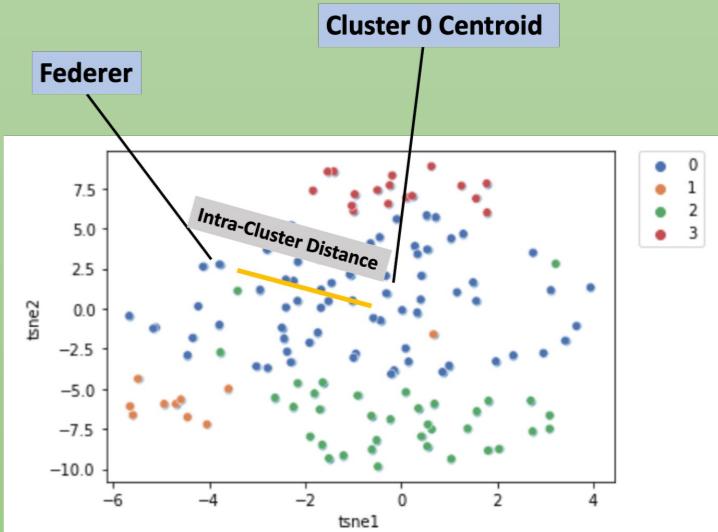
## 1. Data Generalization via Clustering



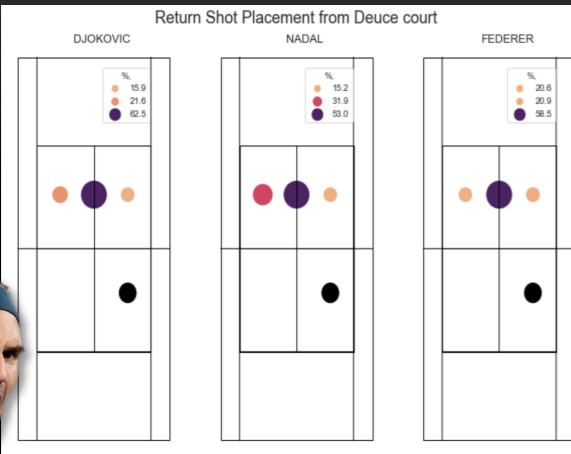
## 2. Player Clustering via Game Components



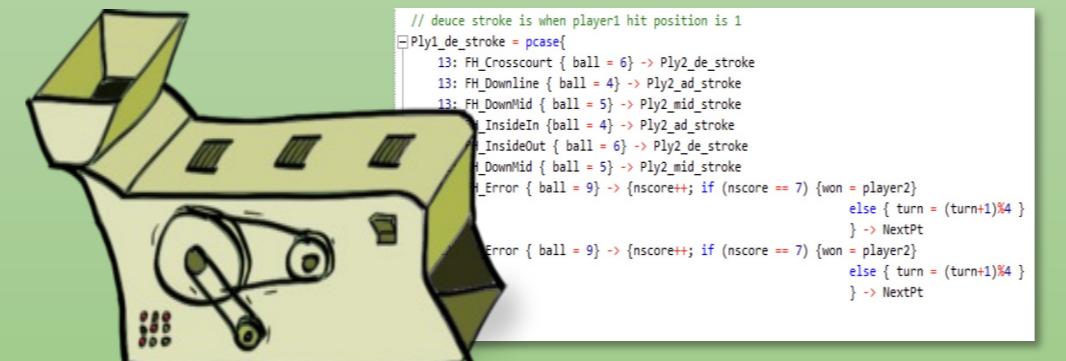
## 3. 3-Part Evaluation Process



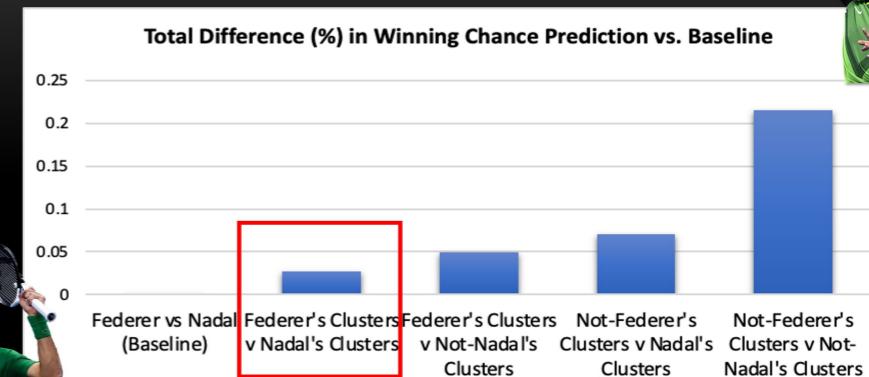
## 4. Interactive Shot Location Preference Chart



## 5. Automated PCSP Cluster Data Populator Script



## 6. Experiment Results



# FUTURE WORKS

## 1. Systematically Test Combo's:

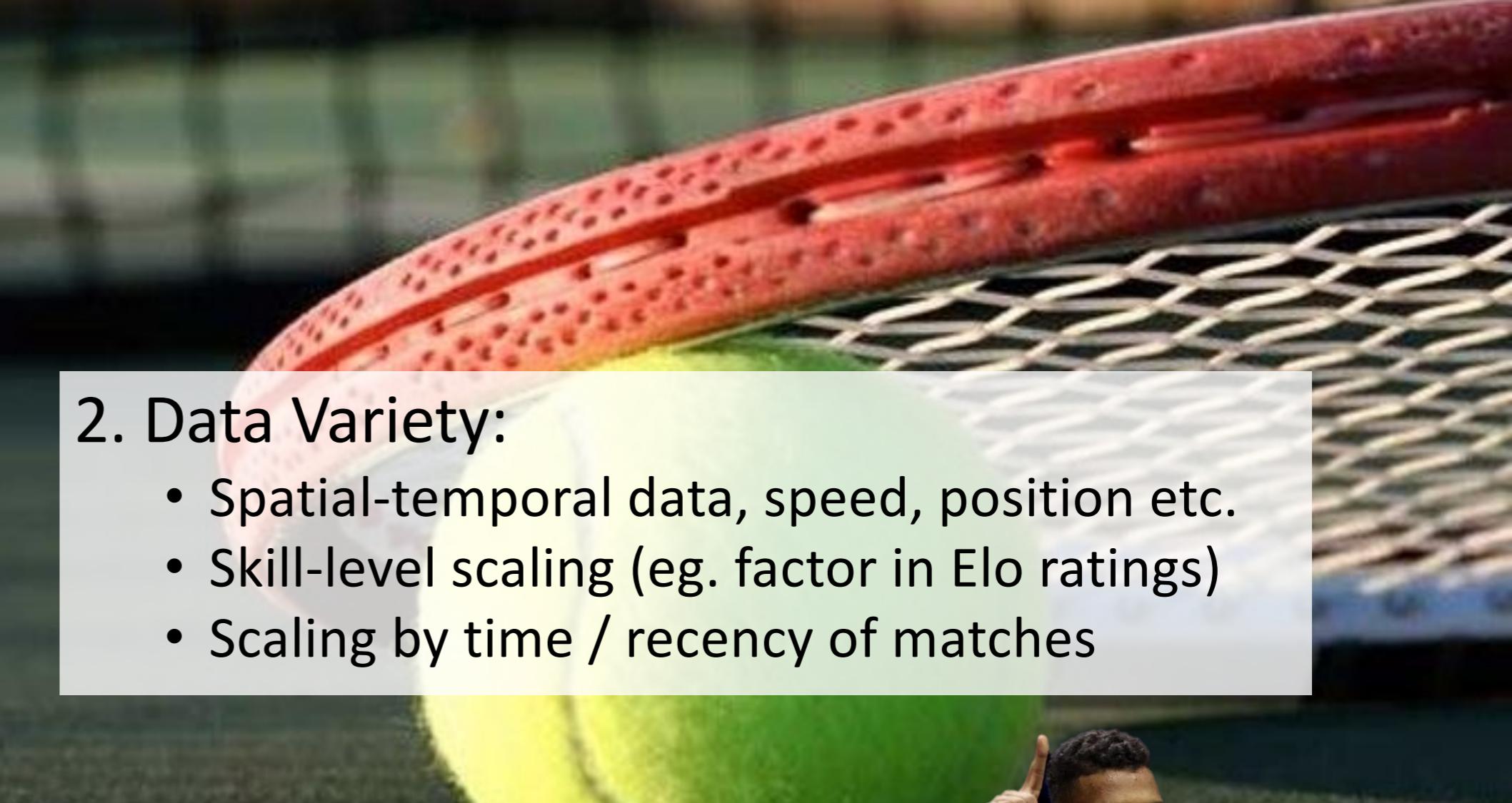
- Feature set combinations
- # of clusters, cluster-size
- # of game components
- More players for PCSP-replacement
- Etc..

## 2. Data Variety:

- Spatial-temporal data, speed, position etc.
- Skill-level scaling (eg. factor in Elo ratings)
- Scaling by time / recency of matches

## 3. Play Styles:

- Sequence patterns - eg. Serve-and-volley likelihood
- Isolating left-handed players



# Lessons Learned

Let's hope this doesn't blow up..



- Data modeling is a very empirical process
- No “one size fits all” solution

# QUESTIONS

