# Identifying Extreme Representatives in Sports. An Archetypal Approach

Daniel Fernández

Universitat Politècnica de Catalunya - BarcelonaTech, Spain

Workshop in Advances in Sports Analytics
Statistical Methods, Data Science, and Research Practices
Università degli Studi di Brescia
July 8th, 2026

# Team work





**Marica Manisera**

**Paola Zuccolotto**

**Daniel Fernández**

**Jordi Cortés**

**Martí Casals**

Big&Open Data Innovation Laboratory (BODaI-Lab)

University of Brescia, Italy

Institute for Research and Innovation in Health (IRIS)

Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Spain

National Institute of Physical Education of Catalonia (INEFC)

Universitat de Barcelona, Spain

# Outline

# Motivation

# Motivation example

**NBA Player Statistics Season 2024-25**

**$n = 501$ players**

**$p = 17$ variables**

| Var | Description |
|---|---|
| **MIN** | **Total minutes played** |
| **PTS** | **Points per game** |
| **REB** | **Rebounds per game** |
| AST | Assists per game |
| FGA | Field goal attempts |
| FG% | Field goal efficiency |
| P3A | 3-point attempts |
| P3% | 3-point efficiency |
| FTA | Free throw attempts |
| FT% | Free throw efficiency |
| STL | Steals per game |
| BLK | Blocks per game |
| TOV | Turnovers per game |
| PF | Personal fouls |

Head of the NBA dataset (some variables)

| Player | Team | PTS | FG% | P3% | REB | AST | BLK | MIN |
|---|---|---|---|---|---|---|---|---|
| A.J. Lawson | TOR | 22.5 | 42.1 | 32.7 | 8.2 | 3.0 | 0.6 | 486 |
| Aaron Gordon | DEN | 24.4 | 53.1 | 43.6 | 8.1 | 5.3 | 0.5 | 1447 |
| Aaron Holiday | HOU | 20.1 | 43.7 | 39.8 | 4.6 | 4.9 | 0.7 | 792 |
| Aaron Nesmith | IND | 22.8 | 50.7 | 43.1 | 7.5 | 2.3 | 0.7 | 1122 |
| Aaron Wiggins | OKC | 24.7 | 48.8 | 38.3 | 8.0 | 3.6 | 0.5 | 1744 |
| Adam Flagler | OKC | 15.4 | 26.0 | 19.4 | 6.4 | 2.8 | 0.7 | 203 |

# Motivation example

**NBA Player Statistics Season 2024-25**

**$n = 501$ players**

**$p = 17$ variables**

| Var | Description |
|---|---|
| **MIN** | **Total minutes played** |
| **PTS** | **Points per game** |
| **REB** | **Rebounds per game** |
| AST | Assists per game |
| FGA | Field goal attempts |
| FG% | Field goal efficiency |
| P3A | 3-point attempts |
| P3% | 3-point efficiency |
| FTA | Free throw attempts |
| FT% | Free throw efficiency |
| STL | Steals per game |
| BLK | Blocks per game |
| TOV | Turnovers per game |
| PF | Personal fouls |

Head of the NBA dataset (some variables)

| Player | Team | PTS | FG% | P3% | REB | AST | BLK | MIN |
|---|---|---|---|---|---|---|---|---|
| A.J. Lawson | TOR | 22.5 | 42.1 | 32.7 | 8.2 | 3.0 | 0.6 | 486 |
| Aaron Gordon | DEN | 24.4 | 53.1 | 43.6 | 8.1 | 5.3 | 0.5 | 1447 |
| Aaron Holiday | HOU | 20.1 | 43.7 | 39.8 | 4.6 | 4.9 | 0.7 | 792 |
| Aaron Nesmith | IND | 22.8 | 50.7 | 43.1 | 7.5 | 2.3 | 0.7 | 1122 |
| Aaron Wiggins | OKC | 24.7 | 48.8 | 38.3 | 8.0 | 3.6 | 0.5 | 1744 |
| Adam Flagler | OKC | 15.4 | 26.0 | 19.4 | 6.4 | 2.8 | 0.7 | 203 |

- Can we found players who are **extreme** profiles based on their performance? Archetypoid analysis (**Standard ADA**).

# Motivation example

**NBA Player Statistics Season 2024-25**

**$n = 501$ players**

**$p = 17$ variables**

| Var | Description |
|---|---|
| **MIN** | **Total minutes played** |
| **PTS** | **Points per game** |
| **REB** | **Rebounds per game** |
| AST | Assists per game |
| FGA | Field goal attempts |
| FG% | Field goal efficiency |
| P3A | 3-point attempts |
| P3% | 3-point efficiency |
| FTA | Free throw attempts |
| FT% | Free throw efficiency |
| STL | Steals per game |
| BLK | Blocks per game |
| TOV | Turnovers per game |
| PF | Personal fouls |

Head of the NBA dataset (some variables)

| Player | Team | PTS | FG% | P3% | REB | AST | BLK | MIN |
|---|---|---|---|---|---|---|---|---|
| A.J. Lawson | TOR | 22.5 | 42.1 | 32.7 | 8.2 | 3.0 | 0.6 | 486 |
| Aaron Gordon | DEN | 24.4 | 53.1 | 43.6 | 8.1 | 5.3 | 0.5 | 1447 |
| Aaron Holiday | HOU | 20.1 | 43.7 | 39.8 | 4.6 | 4.9 | 0.7 | 792 |
| Aaron Nesmith | IND | 22.8 | 50.7 | 43.1 | 7.5 | 2.3 | 0.7 | 1122 |
| Aaron Wiggins | OKC | 24.7 | 48.8 | 38.3 | 8.0 | 3.6 | 0.5 | 1744 |
| Adam Flagler | OKC | 15.4 | 26.0 | 19.4 | 6.4 | 2.8 | 0.7 | 203 |

- Can we found players who are **extreme** profiles based on their performance? Archetypoid analysis (**Standard ADA**).
- Are they players determined by ADA **representative**? e.g. do they have play enough minutes? Our proposal: **Penalized ADA**.

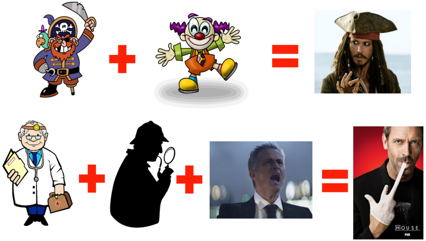
# Background: AA & ADA

# What are the Archetypes?

- Archetype (Wikipedia): from Greek:
  - *archē*: “beginning”, “origin”.
  - *tupos*: “pattern”, “model”, or “type”.
  - **Original pattern from which copies are made.**

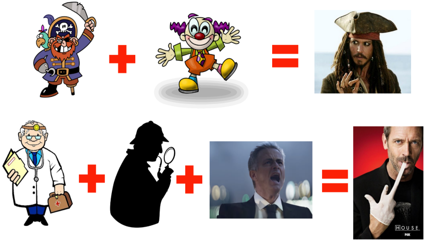
# What are the Archetypes?

- Archetype (Wikipedia): from Greek:
  - *archē*: "beginning", "origin".
  - *tupos*: "pattern", "model", or "type".
  - **Original pattern from which copies are made.**
- Archetypes in everyday language:
  - Jack Sparrow: 40% pirate and 60% clown.
  - Dr. House: 20% doctor, 30% detective, and 50% bad temper.



# What are the Archetypes?

- Archetype (Wikipedia): from Greek:
  - *archē*: "beginning", "origin".
  - *tupos*: "pattern", "model", or "type".
  - **Original pattern from which copies are made.**
- Archetypes in everyday language:
  - Jack Sparrow: 40% pirate and 60% clown.
  - Dr. House: 20% doctor, 30% detective, and 50% bad temper.




- In Statistics, the **concept of archetypes** is the same as in common life.

# Archetype Analysis (AA). Definition

- **Objective** (Cutler and Breiman, 1994): to find a few, not necessarily observed, extremal cases or pure types (the archetypes) such that:
  - all the observations are **obtained by convex combinations of the archetypes**, and
  - all the archetypes are convex combinations of the observations.

# Archetype Analysis (AA). Definition

- **Objective** (Cutler and Breiman, 1994): to find a few, not necessarily observed, extremal cases or pure types (the archetypes) such that:
  - all the observations are **obtained by convex combinations of the archetypes**, and
  - all the archetypes are convex combinations of the observations.
- **Solution** (Cutler and Breiman, 1994): Alternating minimizing algorithm and implemented in R by Eugster and Leisch (2009).

# Archetype Analysis (AA). Example

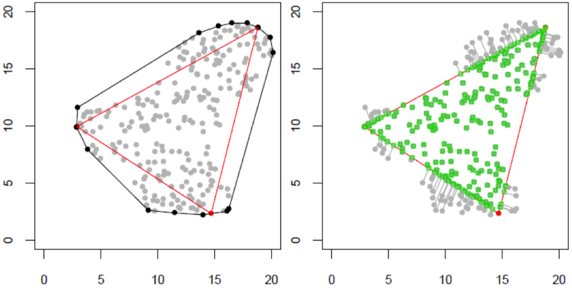



Figure 2: Visualization of three archetypes.

Suppose we want to find $k = 3$ archetypes:

1. Start with 3 random points.

# Archetype Analysis (AA). Example

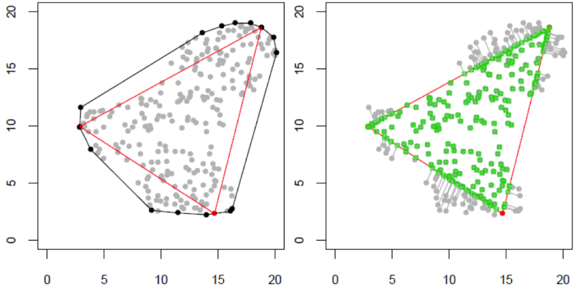



Figure 2: Visualization of three archetypes.

Suppose we want to find $k = 3$ archetypes:

1. Start with 3 random points.
2. The algorithm iteratively updates these points to identify archetypes.

# Archetype Analysis (AA). Example

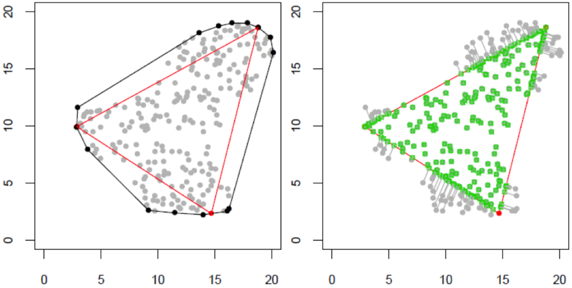



Figure 2: Visualization of three archetypes.

Suppose we want to find $k = 3$ archetypes:

1. Start with 3 random points.
2. The algorithm iteratively updates these points to identify archetypes.
3. In the end, they are the 3 corners of the data.

# Archetype Analysis (AA). Example

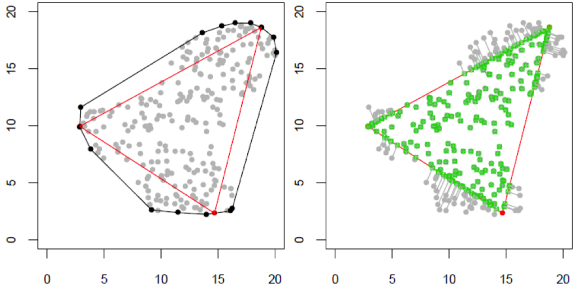


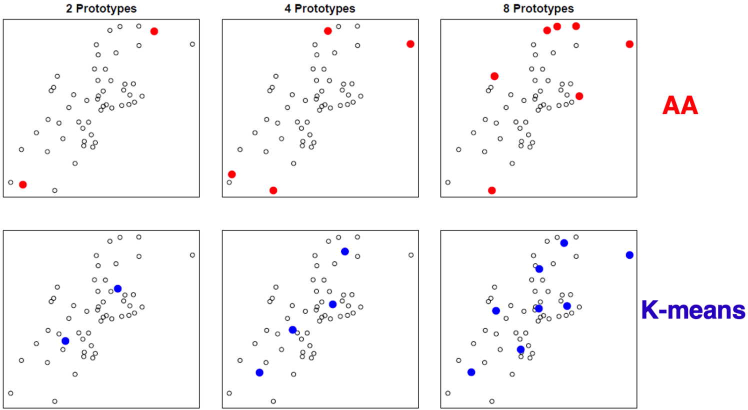

Figure 2: Visualization of three archetypes.

Suppose we want to find $k = 3$ archetypes:

1. Start with 3 random points.
2. The algorithm iteratively updates these points to identify archetypes.
3. In the end, they are the 3 corners of the data.
4. The **observations** are approximated by convex combinations of the archetypes.

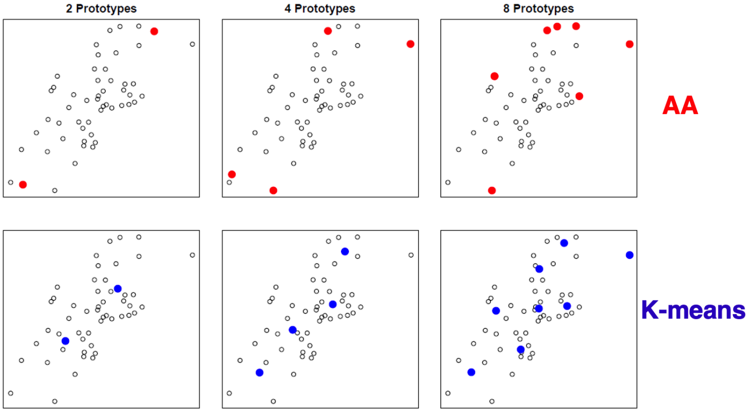
# Archetype Analysis (AA). Example

Archetypes **are more interpretable** than the centroids of k-means.



# Archetype Analysis (AA). Example

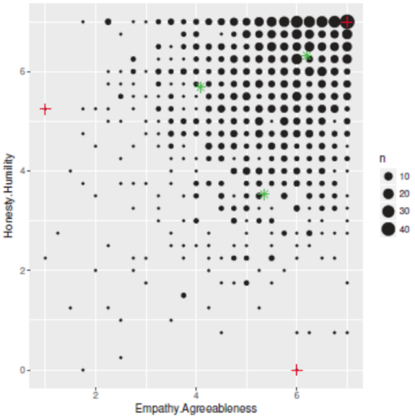
Archetypes **are more interpretable** than the centroids of k-means.




For humans, it is **easier to understand opposites than centroids**: they provide **an intuitive understanding** of the results (Vinúe *et al.*, 2015; Thurau *et al.*, 2012).

# AA. Interpretation

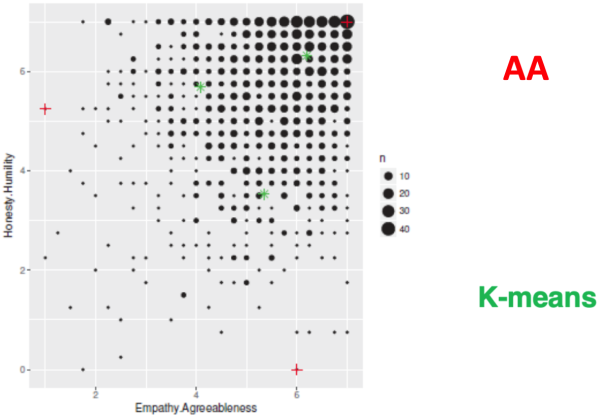
For example, consider two variables: empathy and honesty.




**AA**

**K-means**

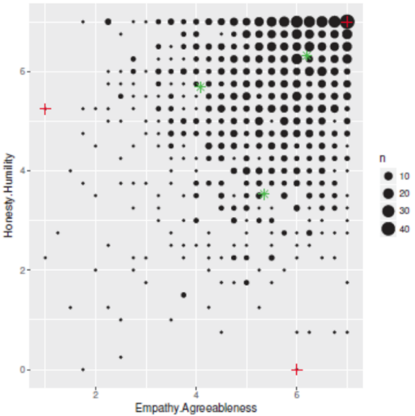
For example, consider two variables: empathy and honesty.




- **AA**: 3 archetypes:
  - 1) Maximum empathy and honesty
  - 2) Maximum empathy and zero honesty
  - 3) Zero empathy and high honesty

# AA. Interpretation

For example, consider two variables: empathy and honesty.




AA

K-means

- **AA**: 3 archetypes:
  - 1) Maximum empathy and honesty
  - 2) Maximum empathy and zero honesty
  - 3) Zero empathy and high honesty
- **k-means**:
  - Centroids are in the **middle of the cloud**, not far apart. This method attempts to find **separate groups** in **continuum** data.

# AA for multivariate data

- Let $X$ be an $n \times m$ matrix with $n$ observations and $m$ variables.
- The objective of AA is to find the matrix $Z$ of $k$ $m$-dimensional archetypes.
- AA computes two matrices $\alpha$ and $\beta$ which **minimize the residual sum of squares (RSS)**:

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2,$$

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $j = 1, \ldots, k$.

The elbow criterion is used to select the number of archetypes.

# Archetypoid Analysis (ADA)

- **A**rchetypoid **A**nalysis (ADA) is a **specific case** of Archetype Analysis (AA).

# Archetypoid Analysis (ADA)

- **A**rchetypoid **A**nalysis (ADA) is a **specific case** of Archetype Analysis (AA).
- **Objective** (Vinúe, Epifanio, & Alemany, 2015): To find a few observed, **extremal cases or pure types (the archetypoids)** such that:
  - All observations are approximated by convex combinations of the archetypoids (**as before**), and
  - All archetypoids are **real observations**.

# ADA for multivariate data

- The objective of ADA is to find the matrix $Z$ of $k$ $m$-dimensional archetypes.
- AA computes two matrices $\alpha$ and $\beta$ which **minimize the residual sum of squares (RSS)**:

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2,$$

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \in \{0, 1\}$ and $j = 1, \ldots, k.$ $\longleftarrow$


- **ADA solution** (Vinué, Epifanio, & Alemany, 2015): algorithm based on PAM.

# Summary of main developments in AA and ADA.

1. **Research**
   - **Functional AA (FAA) and functional ADA (FADA)** (Epifanio, 2016): to find k archetype functions (mixture of the data) or to find k functions of the sample (archetypoids).
   - **Outliers** (Vinúe & Epifanio (2021); Cabero, Epifanio *et al.* , (2021)): Detecting outliers with AA and ADA.
   - **Binary data** (Cabero & Epifanio, 2020).
   - **Ordinal data** (Fernández, Epifanio, & McMillan 2022).
   - **Biarchetype analysis** (Alcacer, Epifanio, & Gual-Arnau, 2024): Extends AA to simultaneously identify archetypes of both observations and features.

# Summary of main developments in AA and ADA.

1. **Research**
   - **Functional AA (FAA) and functional ADA (FADA)** (Epifanio, 2016): to find k archetype functions (mixture of the data) or to find k functions of the sample (archetypoids).
   - **Outliers** (Vinúe & Epifanio (2021); Cabero, Epifanio *et al.* , (2021)): Detecting outliers with AA and ADA.
   - **Binary data** (Cabero & Epifanio, 2020).
   - **Ordinal data** (Fernández, Epifanio, & McMillan 2022).
   - **Biarchetype analysis** (Alcacer, Epifanio, & Gual-Arnau, 2024): Extends AA to simultaneously identify archetypes of both observations and features.

2. **R packages**:
   - archetypes (Eugster and Leisch, 2009)
   - adamethods (Vinué & Epifanio, 2021)
   - Anthropometry (Vinué and Epifanio, 2021)

# Illustration: ATP data

INTERNATIONAL JOURNAL OF SPORTS PHYSIOLOGY AND PERFORMANCE

**Identifying Extreme Representative Tennis Players and Match External Load in Male Grand Slam**

| | |
|---|---|
| Journal: | *International Journal of Sports Physiology and Performance* |
| Manuscript ID | IJSPP.2024-0532 |
| Manuscript Type: | Original Investigation |
| Date Submitted by the Author: | 09-Dec-2024 |
| Complete List of Authors: | Brich, Quim; Institut Nacional d'Educació Física de Catalunya,<br>Casals, Martí; National Institute of Physical Education of Catalonia; Universitat de Vic - Universitat Central de Catalunya, Sport and Physical Studies Centre (CEEAF)<br>Cortés, Jordi; Universitat Politècnica de Catalunya, Department of Statistics and Operations Research<br>Fernández, Daniel; Universitat Politecnica de Catalunya, Department of Statistics and Operations Research<br>Baiget, Ernest; National Institute of Physical Education of Catalonia, |
| Keywords: | Workload, Multivariate Data, Archetypoid Analysis, Unsupervised Learning, Professional Players |
| | |

# ATP data: Dataset Description

- **Source**: Point-level data from 2017 Grand Slam men's singles (**only one year**).
  - Retrieved from Jeff Sackmann's GitHub: `github.com/JeffSackmann/tennis_slam_pointbypoint`.
  - Excluded: Walkover matches or those featuring players who retired.

# ATP data: Dataset Description

- **Source**: Point-level data from 2017 Grand Slam men's singles (**only one year**).
  - Retrieved from Jeff Sackmann's GitHub: `github.com/JeffSackmann/tennis_slam_pointbypoint`.
  - Excluded: Walkover matches or those featuring players who retired.
  - **Aggregated data by player** in 2017 (normalized by game and minute) available at: `github.com/jordicortes40/clustering_tennis`.
  - The profiles created are based on only one year, meaning they do not reflect the players' overall profiles, but rather **their state during that specific year**.

# ATP data: Dataset Description

- **Source**: Point-level data from 2017 Grand Slam men's singles (**only one year**).
  - Retrieved from Jeff Sackmann's GitHub: `github.com/JeffSackmann/tennis_slam_pointbypoint`.
  - Excluded: Walkover matches or those featuring players who retired.
  - **Aggregated data by player** in 2017 (normalized by game and minute) available at: `github.com/jordicortes40/clustering_tennis`.
  - The profiles created are based on only one year, meaning they do not reflect the players' overall profiles, but rather **their state during that specific year**.
- **Data Overview**: 282 matches involving 151 male players. All matches include:
  - **General match info**: surface type, match stats.
  - **Performance metrics** (categorized as):
    - **Volume**: points played, shot count, distance.
    - **Intensity**: hitting frequency, speed.
    - **Efficiency**: serve success rate, movement speed.

# ATP data: Dataset Description

- **Source**: Point-level data from 2017 Grand Slam men's singles (**only one year**).
  - Retrieved from Jeff Sackmann's GitHub: `github.com/JeffSackmann/tennis_slam_pointbypoint`.
  - Excluded: Walkover matches or those featuring players who retired.
  - **Aggregated data by player** in 2017 (normalized by game and minute) available at: `github.com/jordicortes40/clustering_tennis`.
  - The profiles created are based on only one year, meaning they do not reflect the players' overall profiles, but rather **their state during that specific year**.

- **Data Overview**: 282 matches involving 151 male players. All matches include:
  - **General match info**: surface type, match stats.
  - **Performance metrics** (categorized as):
    - **Volume**: points played, shot count, distance.
    - **Intensity**: hitting frequency, speed.
    - **Efficiency**: serve success rate, movement speed.

- **Objective**: **Player Profiling** $\Rightarrow$ to identify **real extreme** player profiles (**ADA**).

# ATP data: ADA results (players)

- Each player who participated in a Grand Slam in 2017 is represented as a mixture of the four archetypoids .

**Player representatives**

# ATP data: ADA results (players)

- Each player who participated in a Grand Slam in 2017 is represented as a mixture of the four archetypoids .

**Player representatives**

- **Cluster 1 (19.1%)**: Most physically demanding archetype; **moderate-high volume and high intensity**, with low serve velocity. *Representative: Darian King*

# ATP data: ADA results (players)

- Each player who participated in a Grand Slam in 2017 is represented as a mixture of the four archetypoids .

**Player representatives**

- **Cluster 2 (25.7%)**: Least physically demanding profile; **moderate-low volume and low intensity**, with high serve velocity. *Representative: Sam Groth*

# ATP data: ADA results (players)

- Each player who participated in a Grand Slam in 2017 is represented as a mixture of the four archetypoids .

**Player representatives**

- **Cluster 3 (34.2%)**: **Low volume, moderate-high intensity** archetype; players with efficient, high-impact performance. *Representative: Luca Vanni*

# ATP data: ADA results (players)

- Each player who participated in a Grand Slam in 2017 is represented as a mixture of the four archetypoids .

**Player representatives**

- **Cluster 4 (21.1%)**: Highest volumes with **moderate-low intensity**; endurance-based style with steady output.
  *Representative: Janko Tipsarevic*

## ATP data: ADA results (players)

# Darian King


Sam Groth, Luca Vanni, and Janko Tipsarevic

# Proposed method: Penalized ADA

# Penalized ADA

- **Main idea**: Modify the **objective function**, explicitly controlling the selection via regularization (weights).

# Penalized ADA

- **Main idea**: Modify the **objective function**, explicitly controlling the selection via regularization (weights).

$$RSS = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j \right\|^2 = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{\ell=1}^{n} \beta_{j\ell} \mathbf{x}_\ell \right\|^2 + \lambda \sum_{\mathbf{j=1}}^{\mathbf{k}} \sum_{\ell=1}^{\mathbf{n}} \beta_{\mathbf{j}\ell} \mathbf{w}_\ell,$$

# Penalized ADA

- **Main idea**: Modify the **objective function**, explicitly controlling the selection via regularization (weights).

$$RSS = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j \right\|^2 = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{\ell=1}^{n} \beta_{j\ell} \mathbf{x}_\ell \right\|^2 + \lambda \sum_{\mathbf{j=1}}^{\mathbf{k}} \sum_{\ell=\mathbf{1}}^{\mathbf{n}} \beta_{\mathbf{j}\ell} \mathbf{w}_\ell,$$

  under the constraints

  1. $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and $\sum_{\ell=1}^{n} \beta_{j\ell} = 1$ with $\beta_{j\ell} \in \{0, 1\}$ and $j = 1, \ldots, k$.
  2. The **penalty weights** are defined as:
     - $w_\ell^0 = 1/(\mathrm{minutes}_\ell + \epsilon)$, where $\epsilon$ is a small constant to avoid division by zero,
     - $w_\ell = w_\ell^0 / \bar{w}^0$, where $\bar{w}^0$ is the mean of all the $w_\ell^0$,
     - $\lambda \geq 0$ controls the strength of the penalty; $\lambda = 0$ represents no penalization.
- **Interpretation**: High-minute players $\rightarrow$ low contribution to the penalty

# Application: NBA data set

# Motivation example

**NBA Player Statistics Season 2024-25**

***n* = 501 players**

***p* = 17 variables**

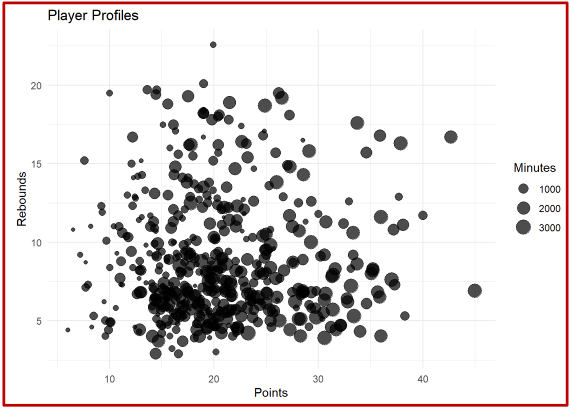
| Var | Description |
|---|---|
| **MIN** | **Total minutes played** |
| **PTS** | **Points per game** |
| **REB** | **Rebounds per game** |
| AST | Assists per game |
| FGA | Field goal attempts |
| FG% | Field goal efficiency |
| P3A | 3-point attempts |
| P3% | 3-point efficiency |
| FTA | Free throw attempts |
| FT% | Free throw efficiency |
| STL | Steals per game |
| BLK | Blocks per game |
| TOV | Turnovers per game |
| PF | Personal fouls |

Head of the NBA dataset (some variables)

| Player | Team | PTS | FG% | P3% | REB | AST | BLK | MIN |
|---|---|---|---|---|---|---|---|---|
| A.J. Lawson | TOR | 22.5 | 42.1 | 32.7 | 8.2 | 3.0 | 0.6 | 486 |
| Aaron Gordon | DEN | 24.4 | 53.1 | 43.6 | 8.1 | 5.3 | 0.5 | 1447 |
| Aaron Holiday | HOU | 20.1 | 43.7 | 39.8 | 4.6 | 4.9 | 0.7 | 792 |
| Aaron Nesmith | IND | 22.8 | 50.7 | 43.1 | 7.5 | 2.3 | 0.7 | 1122 |
| Aaron Wiggins | OKC | 24.7 | 48.8 | 38.3 | 8.0 | 3.6 | 0.5 | 1744 |
| Adam Flagler | OKC | 15.4 | 26.0 | 19.4 | 6.4 | 2.8 | 0.7 | 203 |

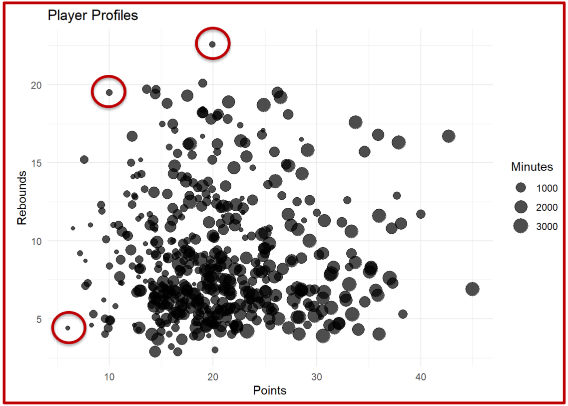
- Can we found players who are **extreme** profiles based on their performance? Archetypoid analysis (**Standard ADA**).
- Are they players determined by ADA **representative**? e.g. do they have play enough minutes? Our proposal: **Penalized ADA**.

# ADA results (players)




- The variables **Points** & **Rebounds** are our features of interest.
- The variable **Minutes** serves to compute weights.

# Standard ADA




- The variables **Points** & **Rebounds** are our features of interest.
- The variable **Minutes** serves to compute weights.
- Note that some players with very **low minutes** are located at the **extremes** of the data cloud $\Rightarrow$ Archetypoids in the **standard ADA**.

# ADA results: Standard vs. Penalized (I)

- Example: $k = 3$ archetypoids, 20 starting points, & $\lambda = 1$ as the penalty strength.

| | **Standard** | **Penalized** |
|---|---|---|
| RSS | 0.027 | 0.159 |
| Avg. minutes | 988.97 | 2421.67 |
| Computation time | 27.95 | 41.11 |

- **RSS** is higher in the penalized version $\Rightarrow$ adding the penalty selects **less extreme** but more **representative** archetypoids.
- **Average minutes** nearly triples in the penalized version (989 $\rightarrow$ 2422 min) $\Rightarrow$ confirms the penalty favors **high-minute** players.

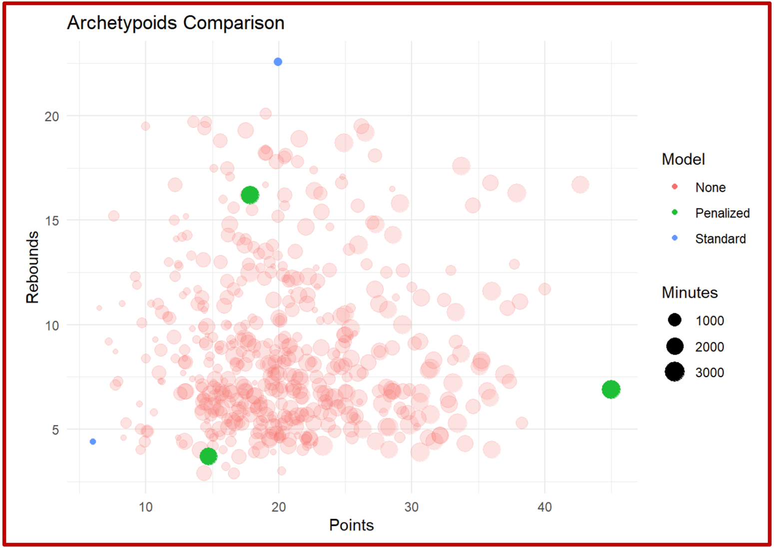
# ADA results: Standard vs. Penalized (II)

**Selected archetypoids**

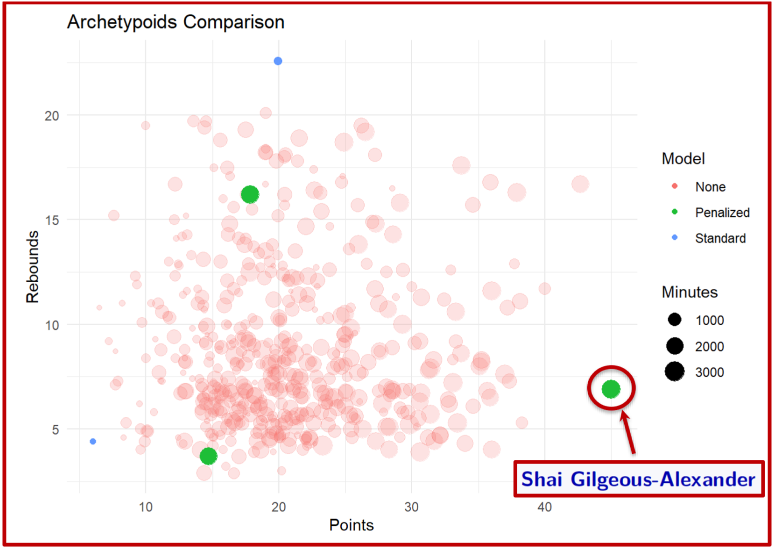
| Player | PTS | REB | MIN | Model |
|---|---|---|---|---|
| Joe Ingles | 6.0 | 4.4 | 114.1 | Standard |
| Oscar Tshiebwe | 19.9 | 22.6 | 255.2 | Standard |
| **Shai Gilgeous-Alexander** | 45.0 | 6.9 | 2597.6 | **Both** |
| Rudy Gobert | 17.8 | 16.2 | 2388.4 | Penalized |
| Kentavious Caldwell-Pope | 14.7 | 3.7 | 2279.0 | Penalized |

- Standard ADA picks players with **very low minutes** (Joe Ingles: 114 min, Oscar Tshiebwe: 255 min), which may be **misleading** in practice.
- **Shai Gilgeous-Alexander** is selected in **both** versions: he is simultaneously **extreme and representative**.
- $\lambda \geq 0$ controls the trade-off: higher $\lambda$ favors **representativeness**, lower $\lambda$ favors **extremeness**.

# ADA results: Standard vs. Penalized (III)



# ADA results: Standard vs. Penalized (III)



# Penalized ADA results (players)

- Each NBA player is represented as a mixture of the 3 archetypoids.
  Example Cluster 3:

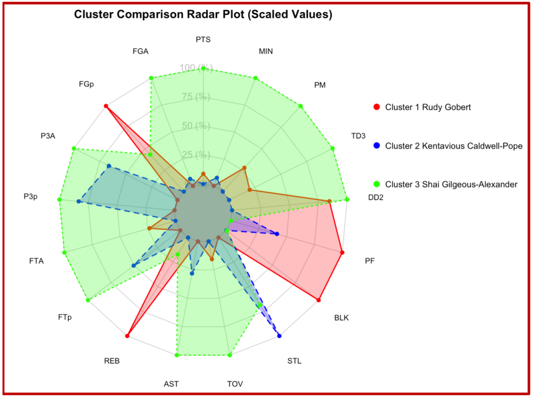
**Cluster 3 (N=58; 11.6%)**


***Representative: Shai Gilgeous-Alexander***

- Point guard - Shooting guard.
- 4-time NBA All-Star.
- 3-time All-NBA First Team member.
- 2-time NBA MVP.

# Penalized ADA results (players)

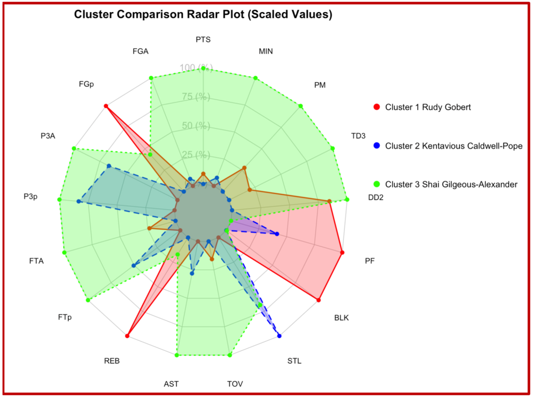
## Cluster 3 (N=58; 11.6%). Representative: Shai Gilgeous-Alexander




- **Cluster 3** players are **elite, high-usage franchise stars who dominate offensively**.

# Penalized ADA results (players)

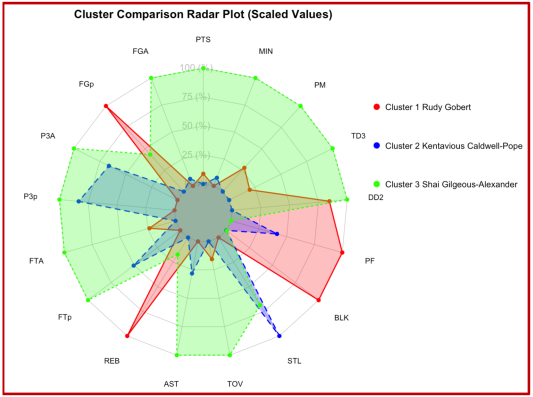
## Cluster 3 (N=58; 11.6%). Representative: Shai Gilgeous-Alexander




- **Cluster 3** players are **elite, high-usage franchise stars who dominate offensively**.
- **Leading all clusters** in scoring (32.9 pts), assists (7.22), minutes (1,857), and free throw attempts (7.57).

# Penalized ADA results (players)

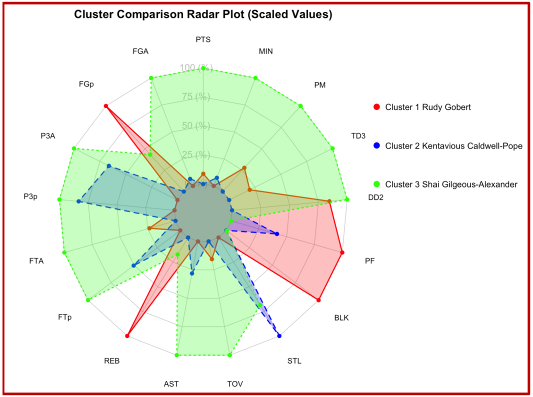
## Cluster 3 (N=58; 11.6%). Representative: Shai Gilgeous-Alexander




- **Cluster 3** players are **elite, high-usage franchise stars who dominate offensively**.
- **Leading all clusters** in scoring (32.9 pts), assists (7.22), minutes (1,857), and free throw attempts (7.57).
- They are the only cluster with a positive plus/minus (+1.37), confirming their **winning impact**.

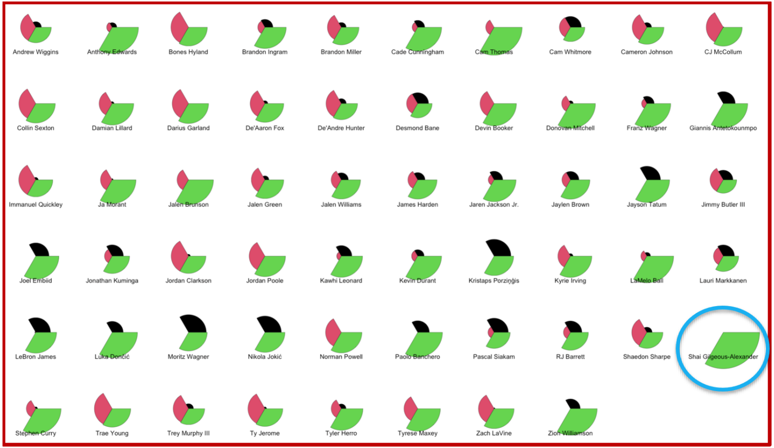
# Penalized ADA results (players)

## Cluster 3 (N=58; 11.6%). Representative: Shai Gilgeous-Alexander




- **Cluster 3** players are **elite, high-usage franchise stars who dominate offensively**.
- **Leading all clusters** in scoring (32.9 pts), assists (7.22), minutes (1,857), and free throw attempts (7.57).
- They are the only cluster with a positive plus/minus (+1.37), confirming their **winning impact**.
- Though their heavy playmaking responsibility comes at the cost of the **highest turnover rate** (3.77).

## Shai Gilgeous-Alexander (green archetypoid)



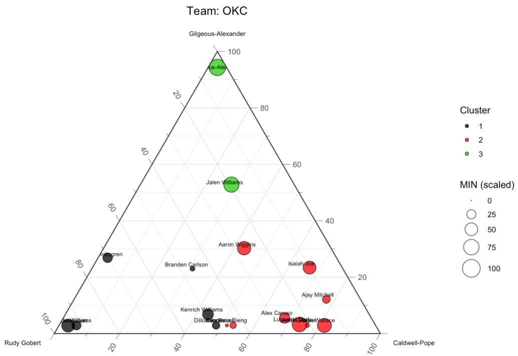

Rudy Gobert, and Kentavious Caldwell-Pope

# Penalized ADA results (players)

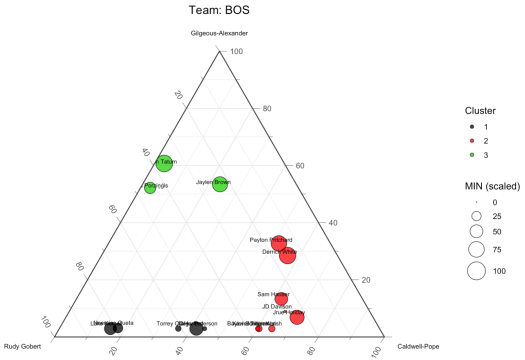
## Shai Gilgeous-Alexander (green archetypoid)


Rudy Gobert, and Kentavious Caldwell-Pope

# Penalized ADA results (teams) - playing style




OKC's simplex is dominated by two star players pulling the team **toward the extremes (ADAs)**.




Boston's players are spread more broadly through the **interior of the triangle,** consistent with a deeper rotation and a more collectively **balanced playing style**.

# Summary

# Summary

- We introduced a **penalized ADA** framework that incorporates player **representativeness** into the selection process.

# Summary

- We introduced a **penalized ADA** framework that incorporates player **representativeness** into the selection process.
- Standard ADA may select **outliers** with few minutes played, leading to **misleading** prototypes.
- The penalized version tends to select players with **higher minutes**, improving **interpretability** and practical relevance.

# Summary

- We introduced a **penalized ADA** framework that incorporates player **representativeness** into the selection process.
- Standard ADA may select **outliers** with few minutes played, leading to **misleading** prototypes.
- The penalized version tends to select players with **higher minutes**, improving **interpretability** and practical relevance.
- In **sports analytics**, where real-world relevance is essential, penalized ADA provides a more **actionable** solution than standard ADA.

# Next Steps

### 1 Sensitivity Analysis in $\lambda$

How the penalty **parameter $\lambda$** affects the selection of archetypoids.

### 2 Simulation Study

Simulation experiments to evaluate the method under **known ground truth conditions**.

### 3 Radial Approach

**Alternative ways** of incorporating representativeness via a radial approach for selecting archetypoids.

### 4 R Package

**R package** implementing the proposed methodology.

# Main References of this Talk

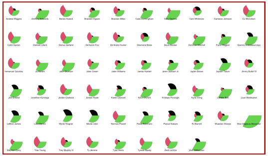

- Cutler, A., and Breiman, L. (1994). Archetypal analysis. *Technometrics*, 36(4), 338-347.
- Eugster, M., and Leisch, F. (2009). From spider-man to hero-archetypal analysis in R. *Journal of Statistical Software*, 30(8), 1–23.
- Fernández, D., Epifanio, I., and McMillan, L. F. (2021). Archetypal analysis for ordinal data. *Information Sciences*, 579, 281-292.
- Vinúe, G., Epifanio, I., and Alemany, S. (2015). Archetypoids: A new approach to define representative archetypal data. *Computational Statistics & Data Analysis*, 87, 102-115.
- Vinúe, G., and Epifanio, I. (2021). Robust archetypoids for anomaly detection in big functional data. *Advances in Data Analysis and Classification*, 15, 437-462.

# The End

**Acknowledgments**:

- MICIU & FEDER [PID2023-148033OB-C21] (Spain)
- Departament de Recerca i Universitats de la Generalitat de Catalunya [2021 SGR 01421] (Spain)
- NZ Government funding the Marsden Fund [E11772-5342] (New Zealand)

# Grazie Mille!!!



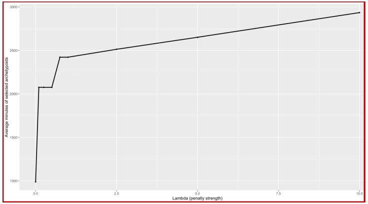
# Extra

# Influence of the Penalty ($\lambda$)

- For fixed $k = 3$, we compute the **average minutes** of the selected archetypoids across a range of $\lambda$ values.
- As $\lambda$ increases:
  1. **Avg. minutes** $\uparrow \rightarrow$ more representative players.
  2. **RSS** $\uparrow \rightarrow$ less extreme archetypoids.
  3. **Ratio** (min/RSS) $\uparrow \rightarrow$ better balance.
- Large $\lambda$ favors **high-minute** players at the cost of a higher RSS.

Results for varying $\lambda$ ($k = 3$)

| $\lambda$ | Avg. Min | RSS | Ratio |
|---|---|---|---|
| 0.00 | 1238 | 8.69 | 142 |
| 0.10 | 1589 | 8.72 | 182 |
| 0.25 | 1589 | 8.72 | 182 |
| 0.50 | 1589 | 8.72 | 182 |
| 0.75 | 1589 | 8.72 | 182 |
| 1.00 | 1589 | 8.72 | 182 |
| 2.50 | 2063 | 8.92 | 231 |
| 5.00 | 2241 | 9.06 | 247 |
| 10.00 | 2603 | 9.92 | 262 |



# Influence of the Penalty ($\lambda$)

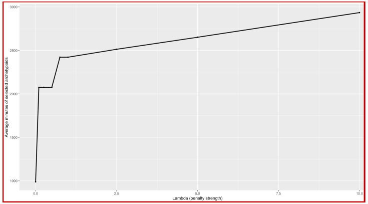
- For fixed $k = 3$, we compute the **average minutes** of the selected archetypoids across a range of $\lambda$ values.

- As $\lambda$ increases:
  1. **Avg. minutes** $\uparrow \rightarrow$ more representative players.
  2. **RSS** $\uparrow \rightarrow$ less extreme archetypoids.
  3. **Ratio** (min/RSS) $\uparrow \rightarrow$ better balance.

- Large $\lambda$ favors **high-minute** players at the cost of a higher RSS.

Results for varying $\lambda$ ($k = 3$)

| $\lambda$ | Avg. Min | RSS | Ratio |
|---|---|---|---|
| 0.00 | 1238 | 8.69 | 142 |
| 0.10 | 1589 | 8.72 | 182 |
| 0.25 | 1589 | 8.72 | 182 |
| 0.50 | 1589 | 8.72 | 182 |
| 0.75 | 1589 | 8.72 | 182 |
| 1.00 | 1589 | 8.72 | 182 |
| 2.50 | 2063 | 8.92 | 231 |
| 5.00 | 2241 | 9.06 | 247 |
| 10.00 | 2603 | 9.92 | 262 |



# Interpretation

- As $\lambda$ increases, the model favors players with **higher minutes**, leading to more **representative** archetypoids.
- This illustrates a key **trade-off**:
  1. **Low** $\lambda \rightarrow$ more **extreme** but less reliable archetypoids.
  2. **High** $\lambda \rightarrow$ more **representative** but less extreme.
- A well-chosen $\lambda$ **balances** both aspects.
- The penalized version tends to select players with **higher minutes**, improving **interpretability** and practical relevance.
- In **sports analytics**, where real-world relevance is essential, penalized ADA provides a more **actionable** solution than standard ADA.
- Standard ADA may select **outliers** with few minutes played, leading to **misleading** prototypes.

# Conclusions. Penalized ADA

- We introduced a **penalized Archetypoid Analysis** framework that incorporates player reliability into the selection process.
- The penalized version tends to select players with **higher minutes**, improving **interpretability** and practical relevance.

# Conclusions. Penalized ADA

- We introduced a **penalized Archetypoid Analysis** framework that incorporates player reliability into the selection process.
- The penalized version tends to select players with **higher minutes**, improving **interpretability** and practical relevance.

- In **sports analytics**, where real-world relevance is essential, penalized ADA provides a more **actionable** solution than standard ADA.

# Conclusions. Penalized ADA

- We introduced a **penalized Archetypoid Analysis** framework that incorporates player reliability into the selection process.
- The penalized version tends to select players with **higher minutes**, improving **interpretability** and practical relevance.
- In **sports analytics**, where real-world relevance is essential, penalized ADA provides a more **actionable** solution than standard ADA.
- Standard ADA may select **outliers** with few minutes played, leading to **misleading** prototypes.