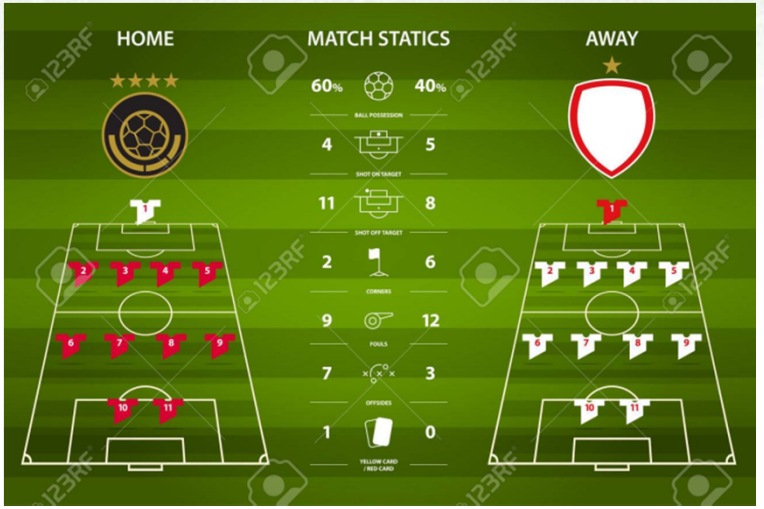
An overview of Football Analytics

Ioannis Ntzoufras

Athens University of Economics & Business

Introduction

Football/Soccer is the best sport for implementing Science/Statistics/Analytics



- Low number of events (so difficult to predict)
- High uncertainty (so difficult to predict)
- Very popular (because it is difficult to predict?)
- Very profitable (because it is difficult to predict?)
- High Financial Risk of investment (because passion becomes more important than numbers and science) – Professional Teams are usually acting as win-maximizers and not profit-maximizers

Main Topics Quantitative analysis of Football/Sports

- Prediction
- Player Evaluation & Performance analytics
- Physical Metrics of Players in training
- Inline game metrics with wearables
- Scheduling
- Sports Economics & Competitive Balance
- Other (Passing Network Analytics, Referee effects, Red card effect, Home effect, Corruption Analytics, Analysis of substitution times)

Prediction

- Offline (before the game)
- Inline (within the game)

Offline Prediction

Modeling of

- Game Scores
  - Poisson based models and extensions
  - Modeling the difference using the Skellam model
- Final outcome of a game (Win/Draw/Loss)
  - Multinomial regression model
  - Bradley Terry Model

Models for Scores

Models for Counts

- Simple Poisson Model (Maher, 1982; Lee, 1992; Dixon & Coles, 1997, Karlis and Ntzoufras, 2000)
- Bivariate Poisson Model (Karlis & Ntzoufras, 2003)
- Negative Binomial Model (see e.g. Ntzoufras 2009)
- Skellam Model for the goal difference (Karlis & Ntzoufras, 2009)
- Poisson-log-normal random effects model (not the best for football counts; see e.g. Ntzoufras 2009)

Models for Scores

Such models allow us not only to predict a single football game but also (simulation based results)

- Final League reproduction
- Estimate probabilities of wining a league, winning European tickets, or relegation.
- Estimate final rankings
- Estimate results under different scenarios/assumptions (by changing covariates i.e. conditions of the game)

Offline Prediction

Poisson Based models

- Vanilla model: home effect + teams attacking and defensive parameters
- Models with time evolved team parameters (time and form matters!)
- Additional covariates
  - Odds from betting teams (easily accessible – good covariates)
  - Team performance (ingame and before the game)
  - Information about events and formation (team strategy, formation, injuries etc.)
  - Economo-demographic variables (Stability, tradition, Budget, Player Value, Coach Value, Country of origin for European leagues)
  - Prior information (previous games between the teams)
  - Team form (e.g. performance in last 5 games)

Offline Prediction

The simple (vanilla) Poisson model

The model is expressed by

$$
\begin{aligned}
Y_{ij} &\sim \text{Poisson}(\lambda_{ik}) && \text{for } j = 1, 2 \\
\log(\lambda_{i1}) &= \mu + \text{home} + a_{\text{HT}_i} + d_{\text{AT}_i} \\
\log(\lambda_{i2}) &= \mu + a_{\text{AT}_i} + d_{\text{HT}_i} && \text{for } i = 1, 2, \ldots, n,
\end{aligned}
$$

where n = number of games, μ = constant parameter; home = home effect; HT_i and AT_i = home and away teams in i game; a_k and d_k = attacking and defensive effects–abilities of k team for $k = 1, 2, \ldots, K$; and K = number of teams in the data (here $K = 20$).

In full (balanced) round-robin leagues, the parameters can be easily calculated by considering averaged of scored/conceded goals for each team

Offline Prediction

Data for the simple (vanilla) model

- Observations
  - 2 × Number of games (N)
  - Each game will occupy two lines/observations (one for home team and one for away team)
- **Response Variable**: Goals scored by each team in each game
- **Covariates**
  - **Home effect**: Binary for home and away teams (1 for home teams and zero otherwise)
  - **Scoring team**: Categorical factor for the team scoring the number of goals (the corresponding coefficient will estimate the attacking ability of each team)
  - **Team accepting goals**: Categorical factor for the team receiving the number of goals (the corresponding coefficient will estimate the defensive ability of each team).

Offline Prediction

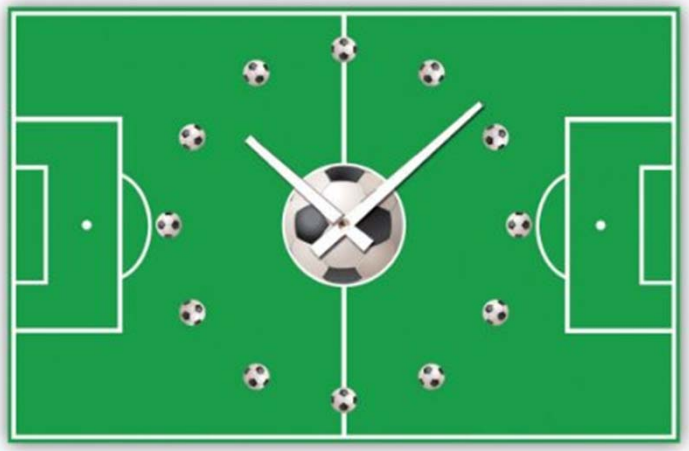
Important Assumptions

- Dependence/Independence of Goals of a game
- Time dependent attacking and defending parameters
- What about draw inflation?
- What about Over-dispesion?
- Shall we focus on modeling scores or outcomes (win/draw/loss)?

Checking the performance of the predictions

- Checking model fit and prediction using in-sample and out-of-sample measures

Prediction within the game



Modeling of

- Time to event (goal)
  - Survival analysis based models
    - Dixon & Robinson (1998, *RSSD*)
    - Nevo and Ritov (2013, *JQAS*)
    - Boshnakov, Kharrat, McHale (2017, *Int. J. Forecasting*)
    - Work in progress by our team
- Model the probability of event for short intervals (every 1 or 5 minutes)
  - Using Binomial mixed models for repeated measures

Player Evaluation

Aim

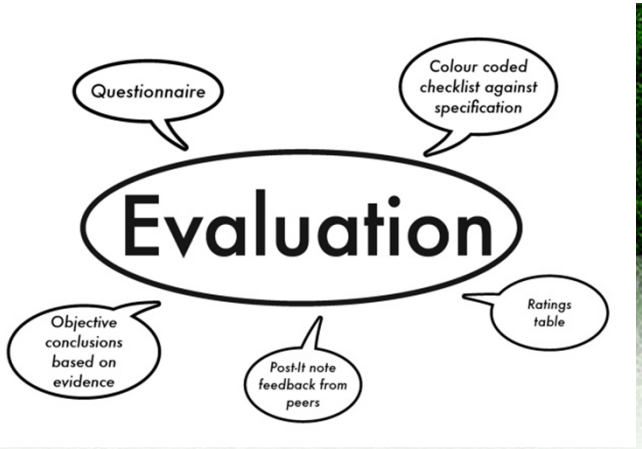
- Estimate the contribution of players in a team
- Rank, identify and reward best players
- Scouting – Early Identification of talents
- Estimate the future performance/value of a Player
- Help the manager to decide the best formation

Player Evaluation

Methods

- Simple approach with binary indicators
- Random effects
- Analysis based on Game Performance Indicators
- Expected Goals (xG) and Expected Assists (xA)
- Player Economic/Marketing Value and performance

Player Evaluation



Methods (2)

- Simple approach with indicators
  - Build a model with indicators whether a players was in the field
  - Binary indicators for players
  - Difficult to build a dataset. Each game should be splitted in multiple lines according to substitution times
- Analysis based on Game Performance Indicators
  - Build a model to identify the importance of each event in the game (goals, shots, steals, passes, speed, stamina, area covered etc.)
  - Use model indicators to build an index of players
  - McHale, Scarf & Folker (2012, *Interfaces*) building different indexes based on different response measures

Player Evaluation

Methods (3)

- Random effects
  - Use random effects to identify individual contribution
  - Goal Scoring: McHale & Szczepanski (2014, *JRSSA*)
  - Passing Skills: Szczepanski & McHale (2016, *JRSSA*)
- Player Economic/Marketing Value and performance
  - Saebo & Hvattum (2018, *Journal of Sports Analytics*): *Modelling the financial contribution of soccer players to their clubs*
  - Evaluating the efficiency of the association football transfer market using regression based player ratings (pre-print only)

Player Evaluation

Methods (4)

McHale, Scarf & Folker (2012, *Interfaces*)

building different indexes based on different response measures

Index ingredients:

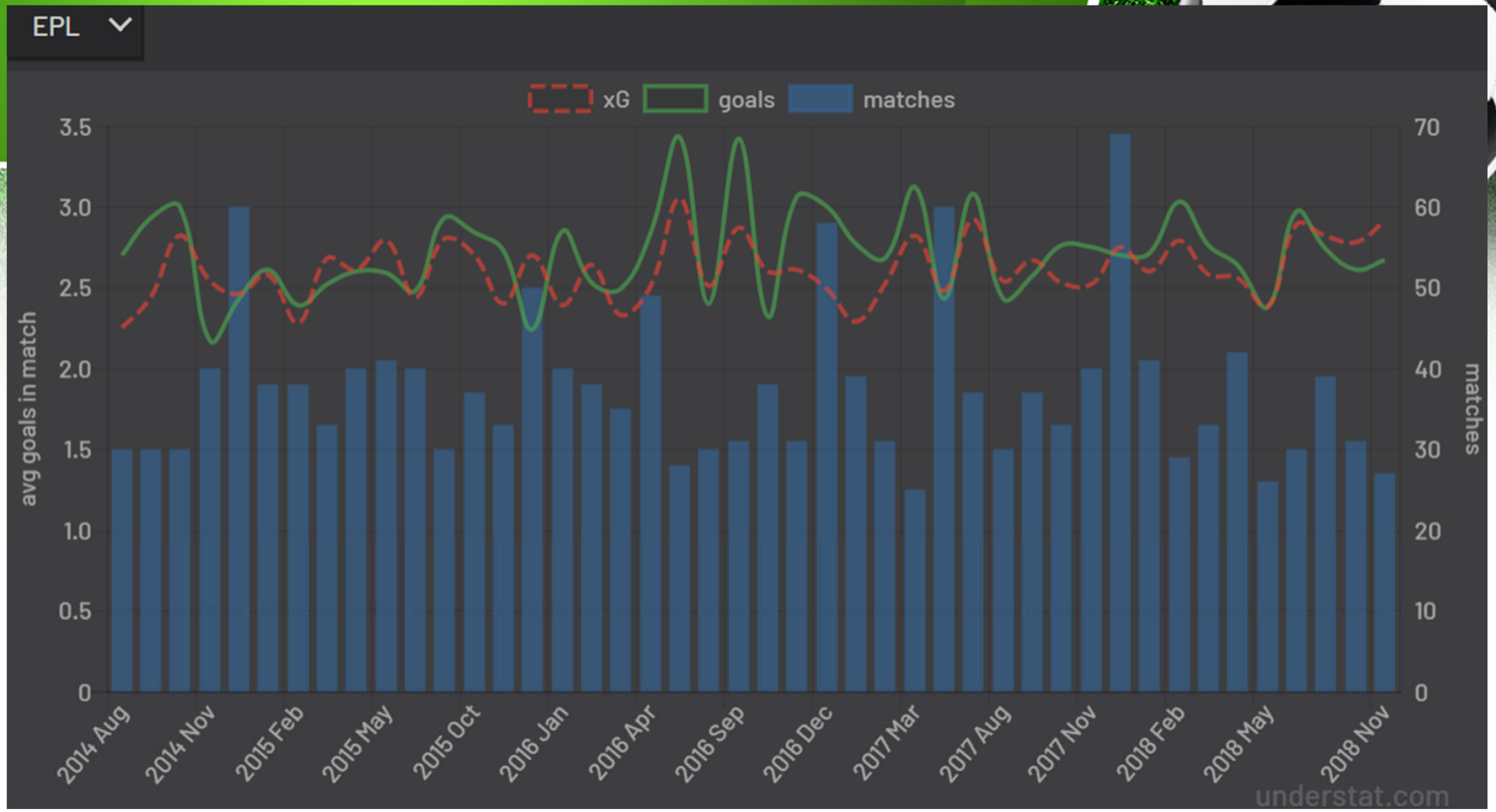
- Subindex 1: Modelling Match Outcome (model based with outcome probability)
- Subindex 2: Points-Sharing Index (time played by each players and points)
- Subindex 3: Appearance Index (time played by each players)
- Subindex 4: Goal-Scoring Index
- Subindex 5: Assists Index
- Subindex 6: Clean-Sheets Index

Player Evaluation

Methods (5)

Expected Goals (xG)

- We model every shot
- Response measure: is the probability of a shot resulting in a goal
- The sum of these probabilities will give the xG of a player and a team
- Similar for assists (xA)
- References:
  - https://www.optasports.com/services/analytics/advanced-metrics/
  - https://understat.com/

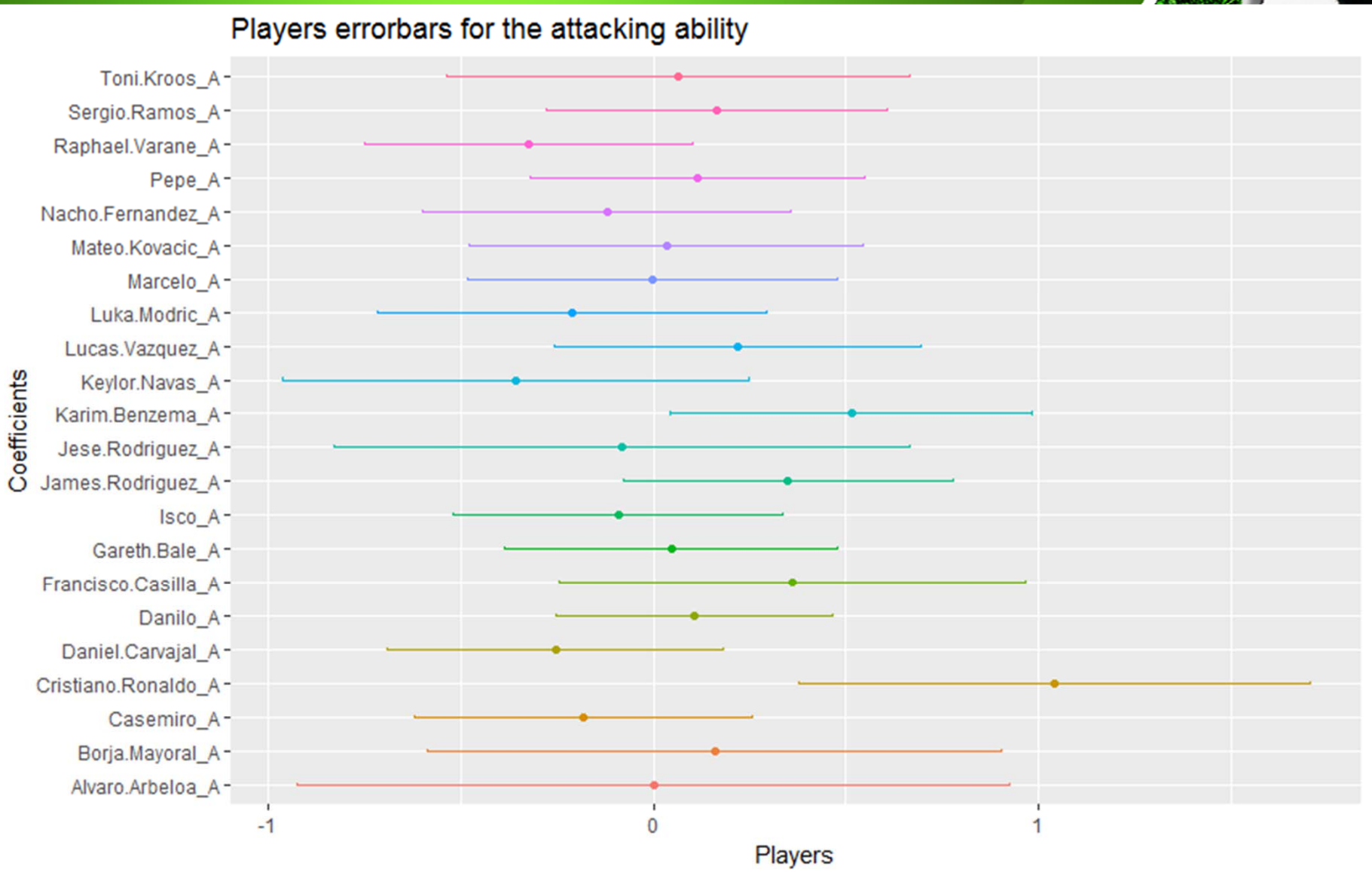


Expected Goals (xG): https://understat.com/

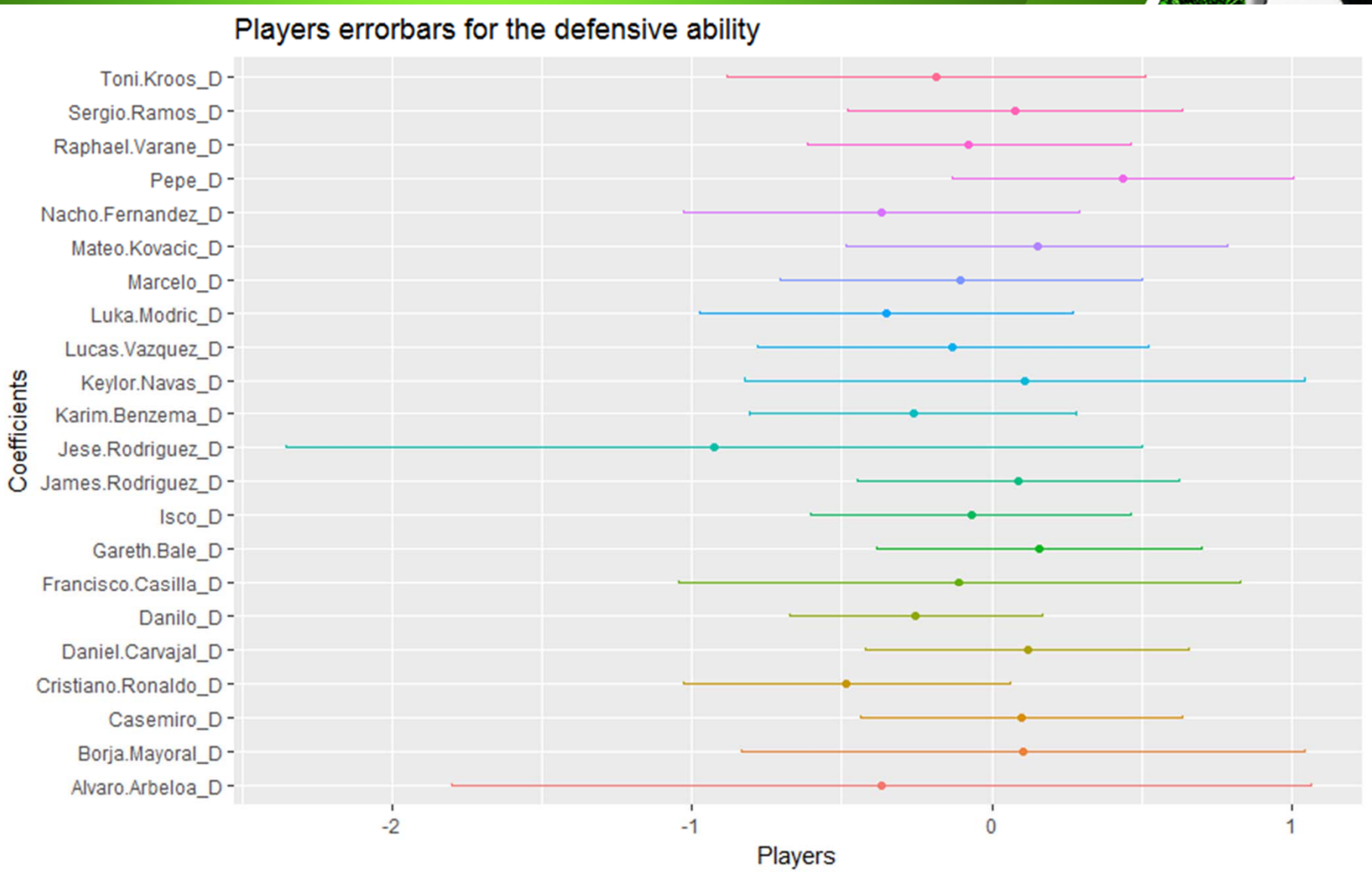
Player Evaluation

Example of the Simple approach with indicators

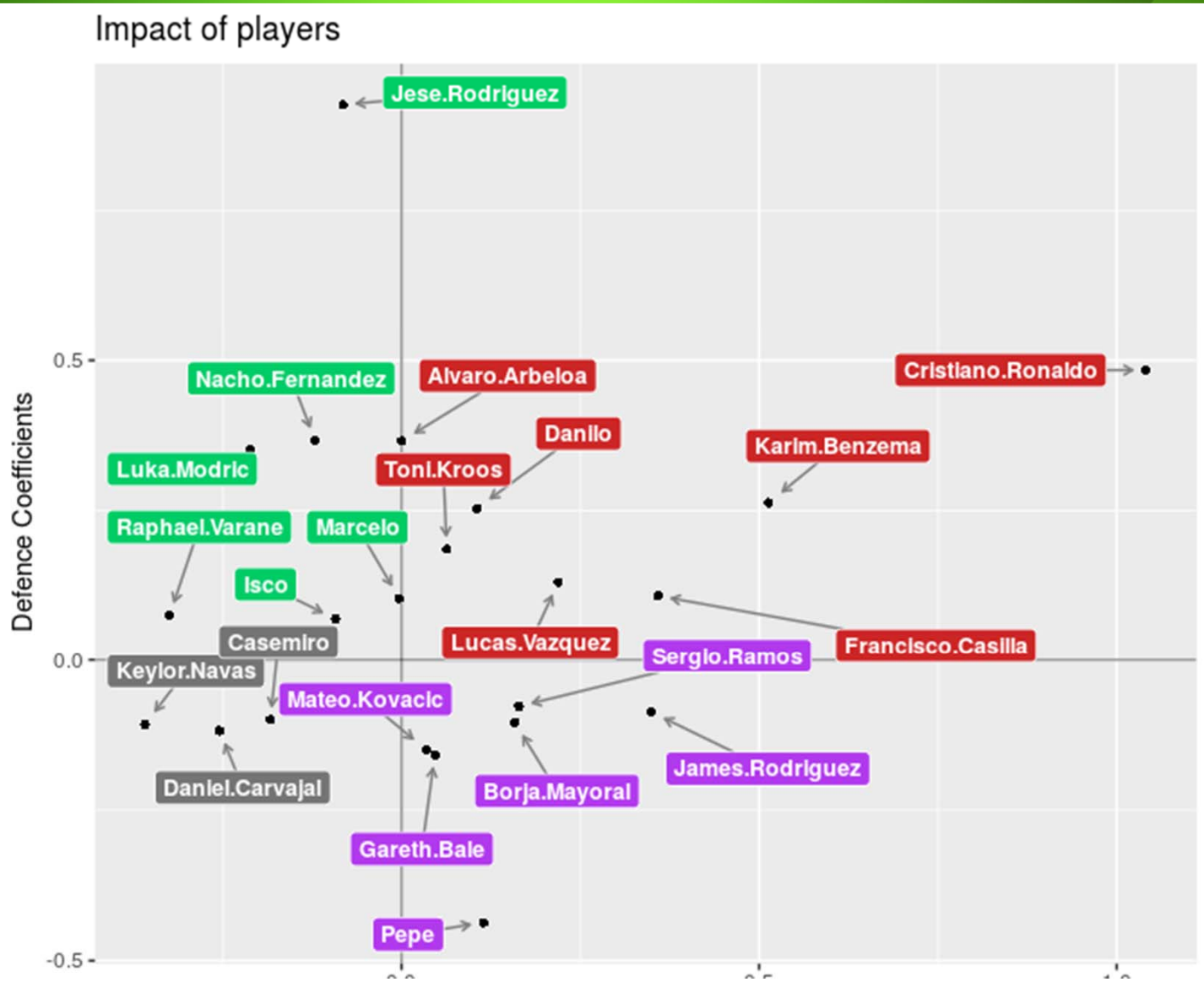
- 351 matches of the La Liga Season 2015/2016
- 954 goals (555 goals were scored by home teams, 399 conceded)
- 110 scored by Real Madrid, 34 conceded
- M.Sc. Thesis at AUEB by A. Mourtopallas


Players errorbars for the attacking ability
Toni.Kroos_A
Sergio.Ramos_A
Raphael.Varane_A
Pepe_A
Nacho.Fernandez_A
Mateo.Kovacic_A
Marcelo_A
Luka.Modric_A
Lucas.Vazquez_A
Keylor.Navas_A
Karim.Benzema_A
Jese.Rodriguez_A
James.Rodriguez_A
Isco_A
Gareth.Bale_A
Francisco.Casilla_A
Danilo_A
Daniel.Carvajal_A
Cristiano.Ronaldo_A
Casemiro_A
Borja.Mayoral_A
Alvaro.Arbeloa_A
Coefficients
-1
0
1
Players

Realmadrid


Players errorbars for the defensive ability
Toni.Kroos_D
Sergio.Ramos_D
Raphael.Varane_D
Pepe_D
Nacho.Fernandez_D
Mateo.Kovacic_D
Marcelo_D
Luka.Modric_D
Lucas.Vazquez_D
Keylor.Navas_D
Karim.Benzema_D
Jese.Rodriguez_D
James.Rodriguez_D
Isco_D
Gareth.Bale_D
Francisco.Casilla_D
Danilo_D
Daniel.Carvajal_D
Cristiano.Ronaldo_D
Casemiro_D
Borja.Mayoral_D
Alvaro.Arbeloa_D
Coefficients
-2
-1
0
1
Players

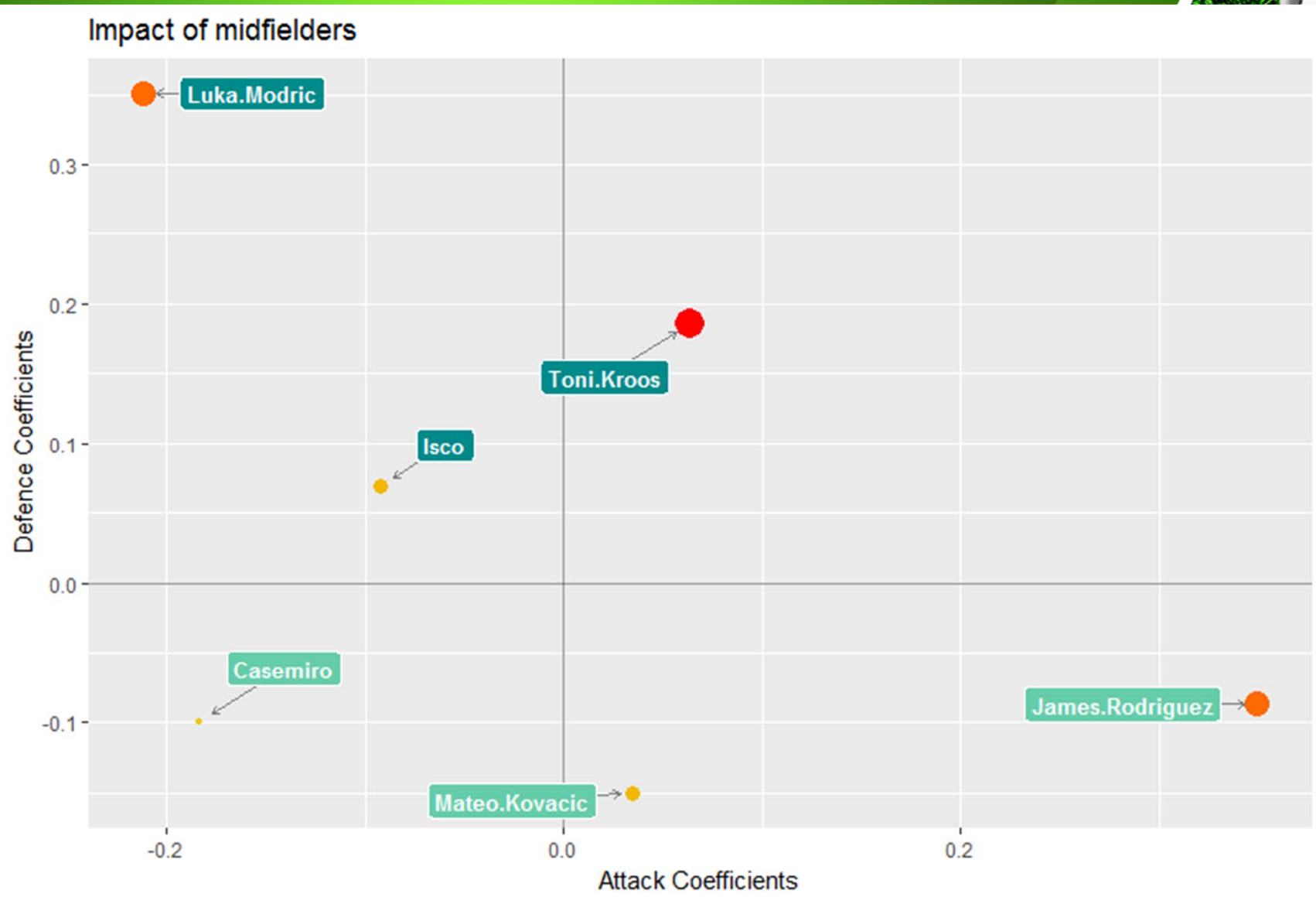
Realmadrid

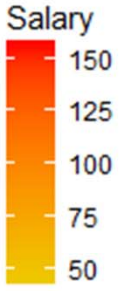

Impact of players
Defence Coefficients
0.5
0.0
-0.5
Jese.Rodriguez
Nacho.Fernandez
Alvaro.Arbeloa
Cristiano.Ronaldo
Danilo
Karim.Benzema
Luka.Modric
Toni.Kroos
Raphael.Varane
Marcelo
Isco
Casemiro
Lucas.Vazquez
Sergio.Ramos
Francisco.Casilla
Keylor.Navas
Mateo.Kovacic
James.Rodriguez
Daniel.Carvajal
Borja.Mayoral
Gareth.Bale
Pepe
Category
a Bad at both
a Better in attack
a Better in defence
a Good at both

Realmadrid

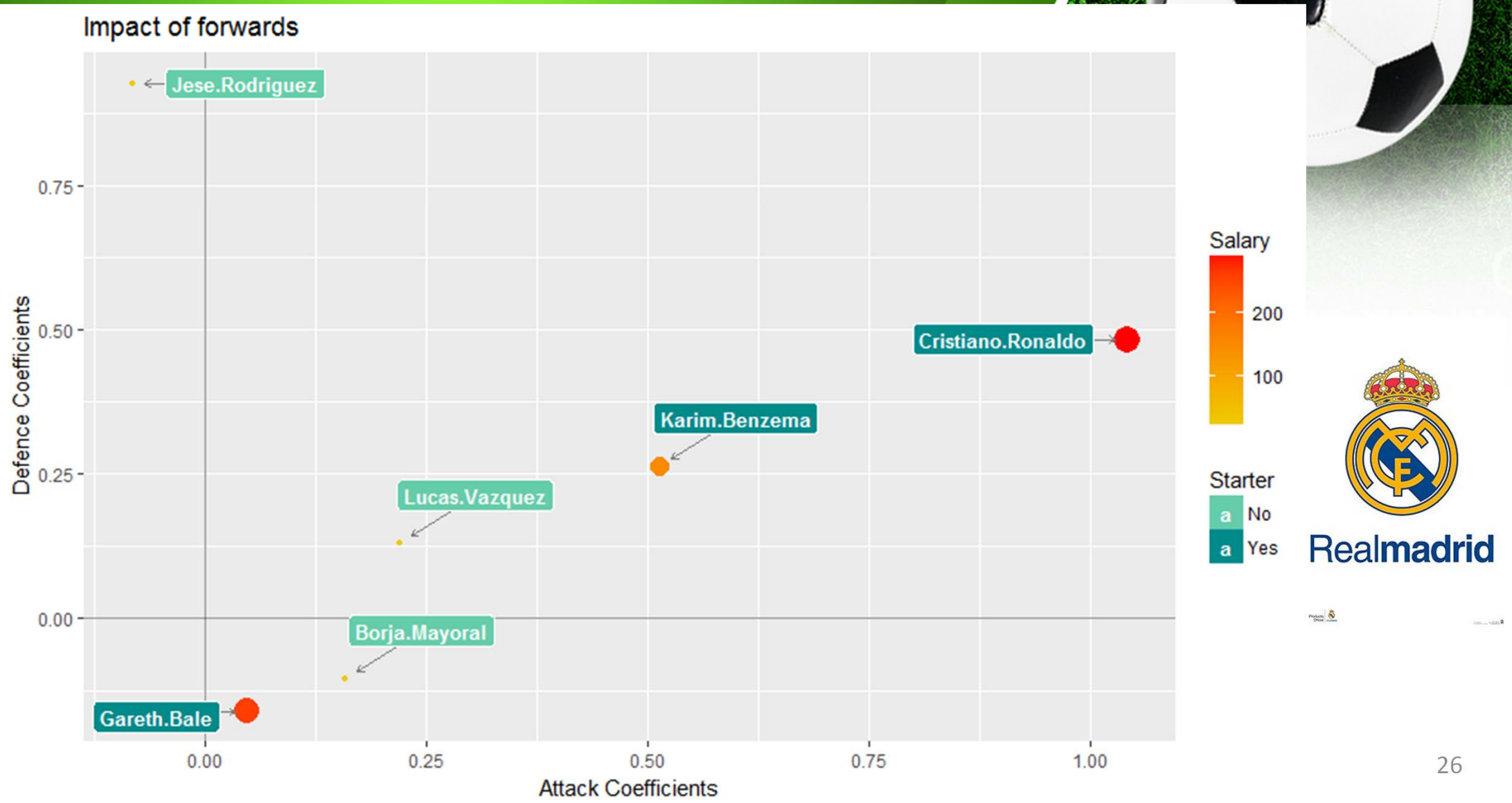

Impact of defenders
Defence Coefficients
Attack Coefficients
0.4
0.2
0.0
-0.2
-0.4
-0.3
-0.2
-0.1
0.0
0.1
Nacho.Fernandez
Alvaro.Arbeloa
Danilo
Marcelo
Raphael.Varane
Sergio.Ramos
Daniel.Carvajal
Pepe
Starter
No
Yes
Salary
140
120
100
80
60
40

Realmadrid


Impact of midfielders
Luka.Modric
Toni.Kroos
Isco
Casemiro
James.Rodriguez
Mateo.Kovacic
0.3
0.2
0.1
0.0
-0.1
-0.2
0.0
0.2
Defence Coefficients
Attack Coefficients


Salary
150
125
100
75
50
Starter
a No
a Yes

Realmadrid


Impact of forwards
Jese.Rodriguez
Cristiano.Ronaldo
Karim.Benzema
Lucas.Vazquez
Borja.Mayoral
Gareth.Bale
Defence Coefficients
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Attack Coefficients
Salary
200
100
Starter
No
Yes

Realmadrid

Real Madrid 2015/16
Player Evaluation

Conclusions

Cristiano Ronaldo is the key player of the team

Tony Kroos' impact is higher than we may presume

Nacho Fernandez improved since last season (very high def contribution)

Lucas Vasquez is a very promising player (contributed positively in both attack and defensive dimensions with low salary)

Gareth Bale performed less than expected (overprized)

Pepe $\Rightarrow$ low defensive contribution – high salary (overprized?)

Realmadrid

Metrics for physical improvement and training

Aim

- Improve the physical condition of athletes
- Focus on specific skills and measure them
- Avoid injuries
- Improves the team by optimizing allocated training time

Inline game metrics with wearables

The aim is to measure

- Movement of players in the game
- Speed and coverage
- Physical condition
- Physical and tactics performance

It helps

- Evaluate the performance of players and teams within a game
- The manager to decide formation and substitutions

League and Contest Scheduling

AIM

- Fair scheduling
- Eliminate bias due to the sequence of games
- Strengthen competitiveness (related with next slides)
- Incorporate constraints (incl. other sports, safety issues, other events, tv requirements etc.)

HOW?

- Using Operational Research and optimization methods
- Hybrid search methods
- Validate using simulation methods from Statistical models

Competitive Balance

- A balanced league increases the interest of the fans and improves the athletic product
- The notion of a balanced league is not yet very well defined
  - Equal Strength between all teams? or
  - Equal Strength between best teams (or the teams with the highest number of fans?)

Sports Economics & Competitive balance

What league do we want to see?

- All fans like the fact that a weaker team occasionally wins a game or a league
- May neutral fans follow the weakest team

e.g. Greece in Euro 2004

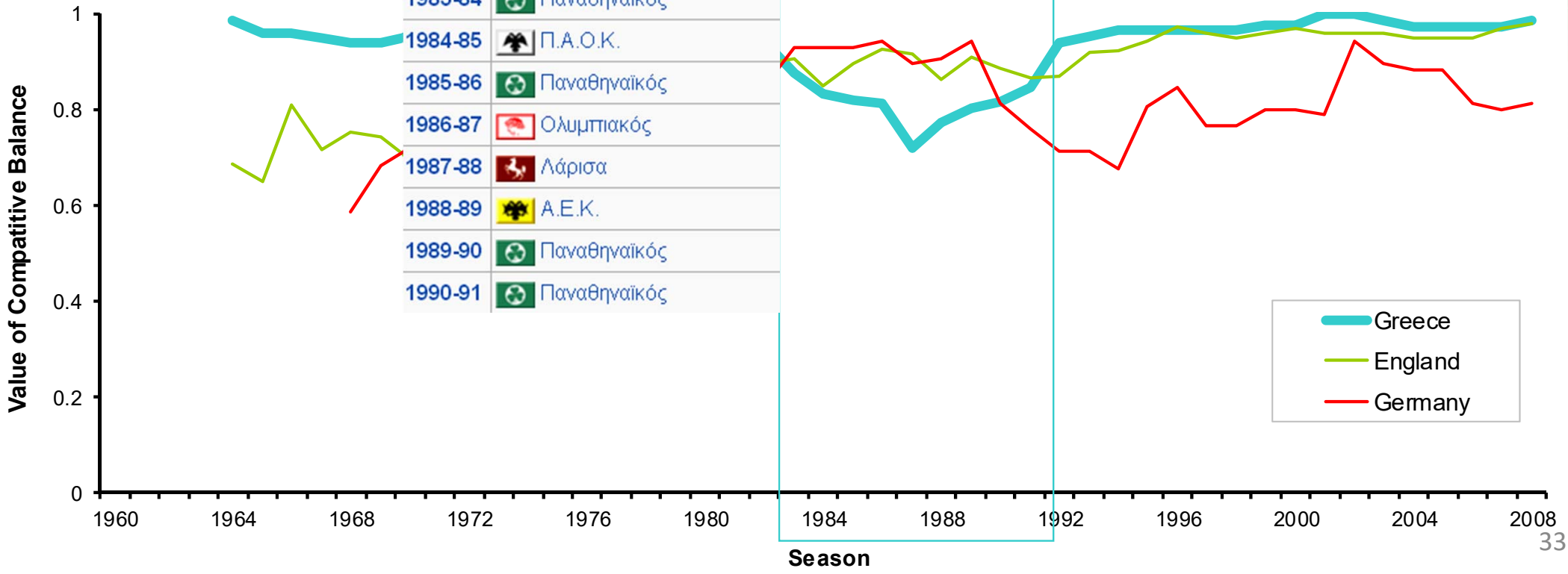
But

- They do not like their team to loose

- They like or they are willing to pay an expensive ticket to see a final with high ranked and expensive teams

e.g. Bayern-Barcelona

Moving Averages of lag five for DN_1 (Champion) from 1959-20



1983-84	Παναθηναϊκός
1984-85	Π.Α.Ο.Κ.
1985-86	Παναθηναϊκός
1986-87	Ολυμπιακός
1987-88	Λάρισα
1988-89	Α.Ε.Κ.
1989-90	Παναθηναϊκός
1990-91	Παναθηναϊκός

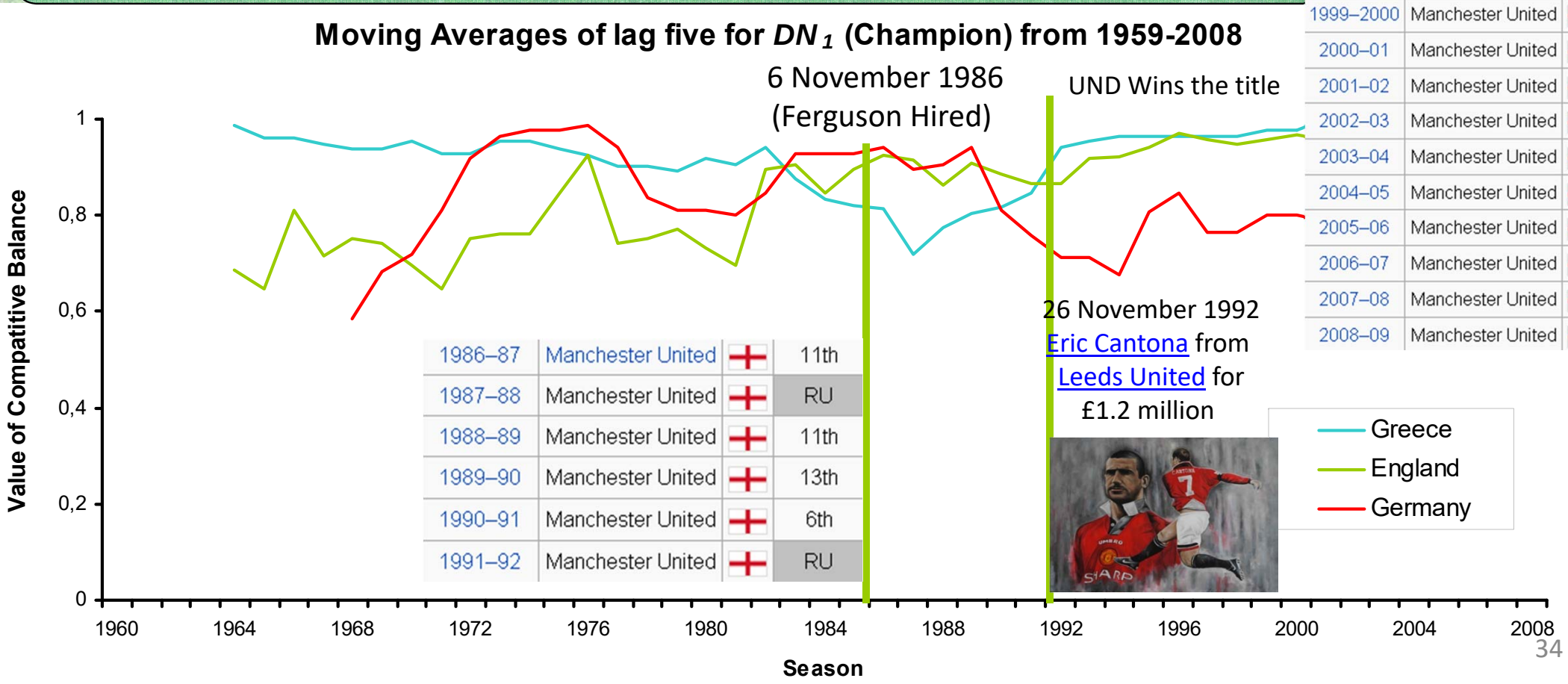
1991-92	Α.Ε.Κ.
1992-93	Α.Ε.Κ.
1993-94	Α.Ε.Κ.
1994-95	Παναθηναϊκός
1995-96	Παναθηναϊκός
1996-97	Ολυμπιακός
1997-98	Ολυμπιακός
1998-99	Ολυμπιακός
1999-00	Ολυμπιακός

2000-01	Ολυμπιακός
2001-02	Ολυμπιακός
2002-03	Ολυμπιακός
2003-04	Παναθηναϊκός
2004-05	Ολυμπιακός
2005-06	Ολυμπιακός
2006-07	Ολυμπιακός
2007-08	Ολυμπιακός
2008-09	Ολυμπιακός

ManU won 13 out of 17 leagues for the period 1992-2009 and it was not ranked in lower position than 3rd.

3 cases in England ⇒ promoted team ⇒ won the championship:

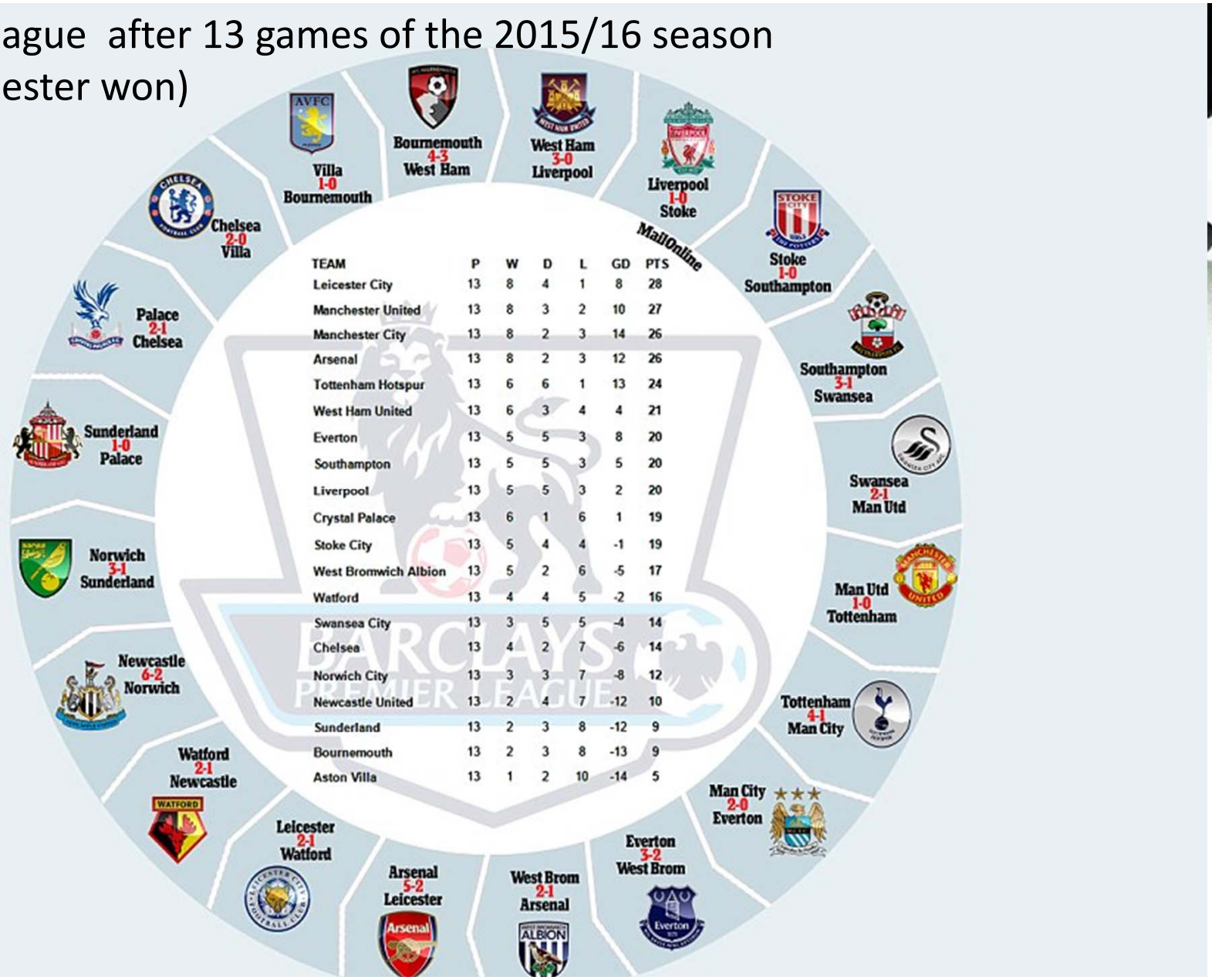
Ipswich (1961) & Nottingham (1997) & Leicester (2015-16 – not in the Figu



1986–87	Manchester United		11th
1987–88	Manchester United		RU
1988–89	Manchester United		11th
1989–90	Manchester United		13th
1990–91	Manchester United		6th
1991–92	Manchester United		RU

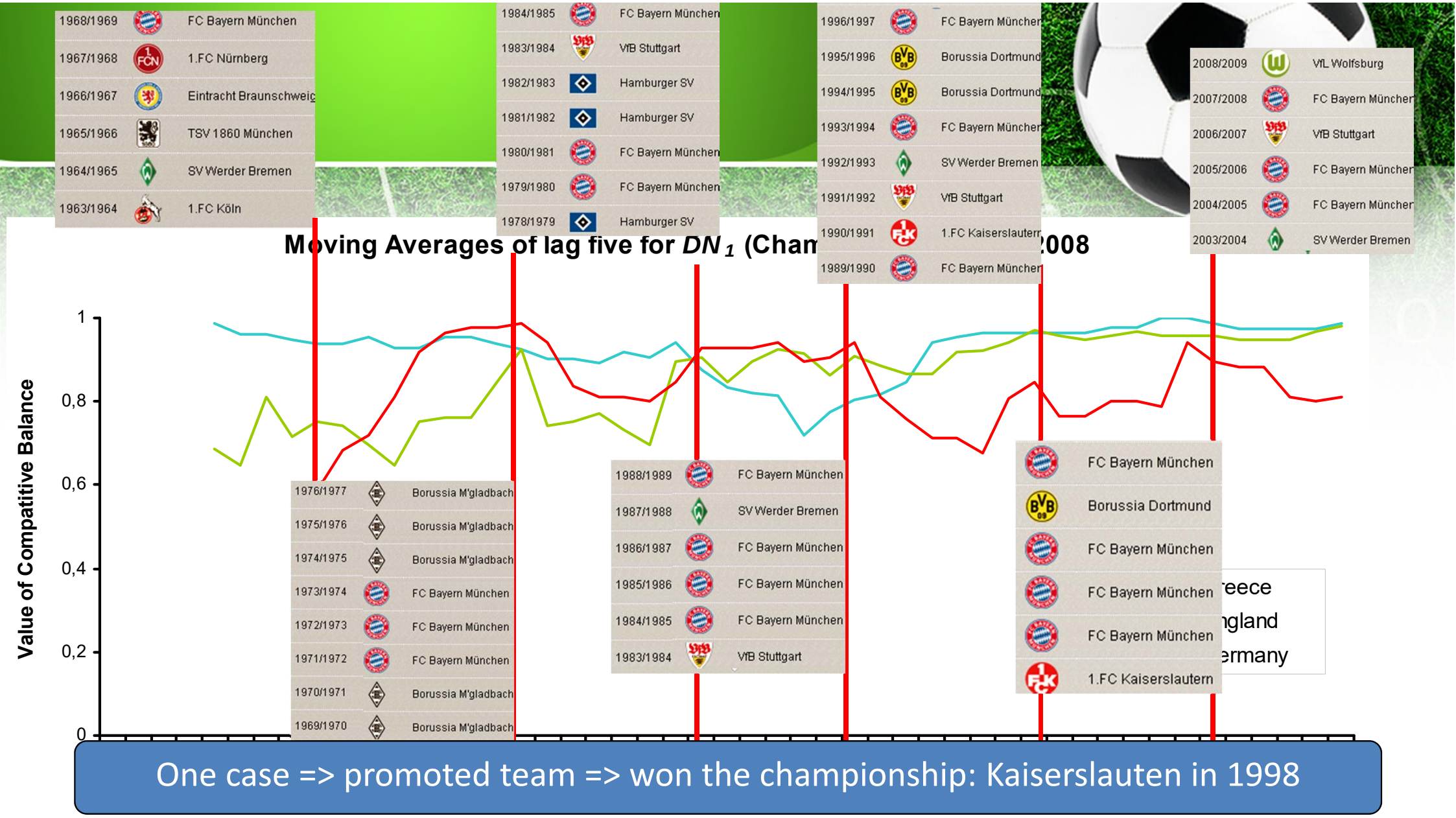
1992–93	Manchester United		W
1993–94	Manchester United		W
1994–95	Manchester United		RU
1995–96	Manchester United		W
1996–97	Manchester United		W
1997–98	Manchester United		RU
1998–99	Manchester United		W
1999–2000	Manchester United		W
2000–01	Manchester United		W
2001–02	Manchester United		3rd
2002–03	Manchester United		W
2003–04	Manchester United		3rd
2004–05	Manchester United		3rd
2005–06	Manchester United		RU
2006–07	Manchester United		W
2007–08	Manchester United		W
2008–09	Manchester United		W

Premier League after 13 games of the 2015/16 season (when Leicester won)



TEAM	P	W	D	L	GD	PTS
Leicester City	13	8	4	1	8	28
Manchester United	13	8	3	2	10	27
Manchester City	13	8	2	3	14	26
Arsenal	13	8	2	3	12	26
Tottenham Hotspur	13	6	6	1	13	24
West Ham United	13	6	3	4	4	21
Everton	13	5	5	3	8	20
Southampton	13	5	5	3	5	20
Liverpool	13	5	5	3	2	20
Crystal Palace	13	6	1	6	1	19
Stoke City	13	5	4	4	-1	19
West Bromwich Albion	13	5	2	6	-5	17
Watford	13	4	4	5	-2	16
Swansea City	13	3	5	5	-4	14
Chelsea	13	4	2	7	-6	14
Norwich City	13	3	3	7	-8	12
Newcastle United	13	2	4	7	-12	10
Sunderland	13	2	3	8	-12	9
Bournemouth	13	2	3	8	-13	9
Aston Villa	13	1	2	10	-14	5

Premier League after 13 games of the 2015/16 season (when Leicester won)



One case => promoted team => won the championship: Kaiserslauten in 1998

How to design Knockout Tournaments?

- Do we support the stronger or the weakest teams?

We do not wish to see

- **many** strong teams to be disqualified early
- Two weak or not popular teams in the final

We do wish to see

- **Some** strong teams to be disqualified early
- Some weak teams to qualify further against all odds

Sports Economics & Competitive balance

In round-robin contests (National leagues)?

- Do we support the stronger or weakest teams?
- Small or large leagues?
- Playoffs?
- Give more money to strong teams (reward) or to weak teams (balance)?
- What about promotion/relegation rules (refreshes the interest or just recycles bad teams?)

We do wish to see

- A large enough group of teams to be close and compete for the championship
- A large enough group of teams to be close and compete for European tickets

We do not wish to see

- A team having big margin of points from all the rest (so the champion is known early)
- Teams with low number of points so they are not competitive (early relegation)
- Teams with economic problems

Sports Economics & Competitive balance

For UEFA Champions League

- Does it need improvement?
- Not metrics to measure balance
- Big discussion of how to reward teams and share income
- Closed or Open League?
- How many teams from each National League/Country
- The current income share and reward system destroys the balance in National teams in second ranked leagues like Greece.

Concluding remarks

To conclude with

- **Prediction** is important for fans (in terms of betting) ⇒ increases profits of bet companies and interest for the sport product (in macro perspective)
- **Inline prediction** is important for fans (in terms of betting) ⇒ increases profits of bet companies and interest for the sport product (Media – TV, Radio, Internet).

Concluding remarks

- **Player performance and evaluation** $\Rightarrow$ Of main interest for: the fans (Player Ranking), Teams (Scouting, Future Performance and Value), Companies (Sponsoring), Players (A lot of money from all previous), Coaches/Managers (Selection of better players)

- **Physical Measurements** (Training and Games): It is related with player evaluation. Main value to help managers/coaches to improve their teams. In macro perspective also the teams financial position is also improving.
- **Scheduling and Competitive Balance**: More Fair and Balanced contests lead to better product and more profit.

THANK YOU

AUEB SPORTS ANALYTICS GROUP

INNOVATIVE RESEARCH IN ANALYTICS

NO Matter How Many Goals You Save People Always Remember The One You Miss.