Actions Speak Louder than Goals: Valuing Player Actions in Soccer

Tom Decroos, Lotte Bransen, Jan Van Haaren, Jesse Davis

KDD 2019

August 7, 2019

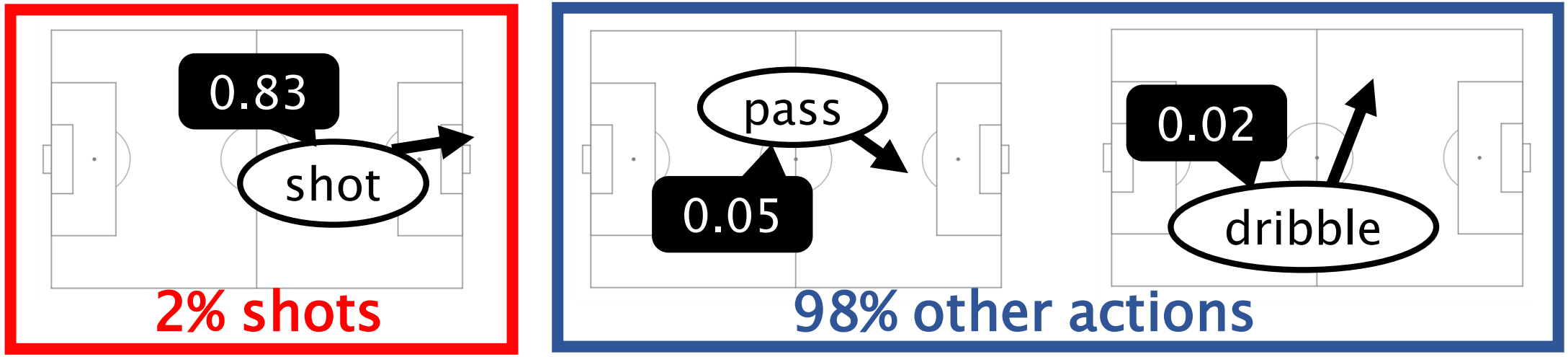
KU LEUVEN

DTAI
Declarative Languages & Artificial Intelligence

SCISPORTS

nano4SPORTS
TECHNOLOGY FOR AN ACTIVE LIFESTYLE

Key soccer analytics task: Valuing on-the-ball actions



Problem: Existing soccer stats value only a single type of action

Our contribution: A framework that values ALL on-the-ball actions

Challenge 1: Real-world action sequences are messy

- Missing or unrecorded actions

- Useless events

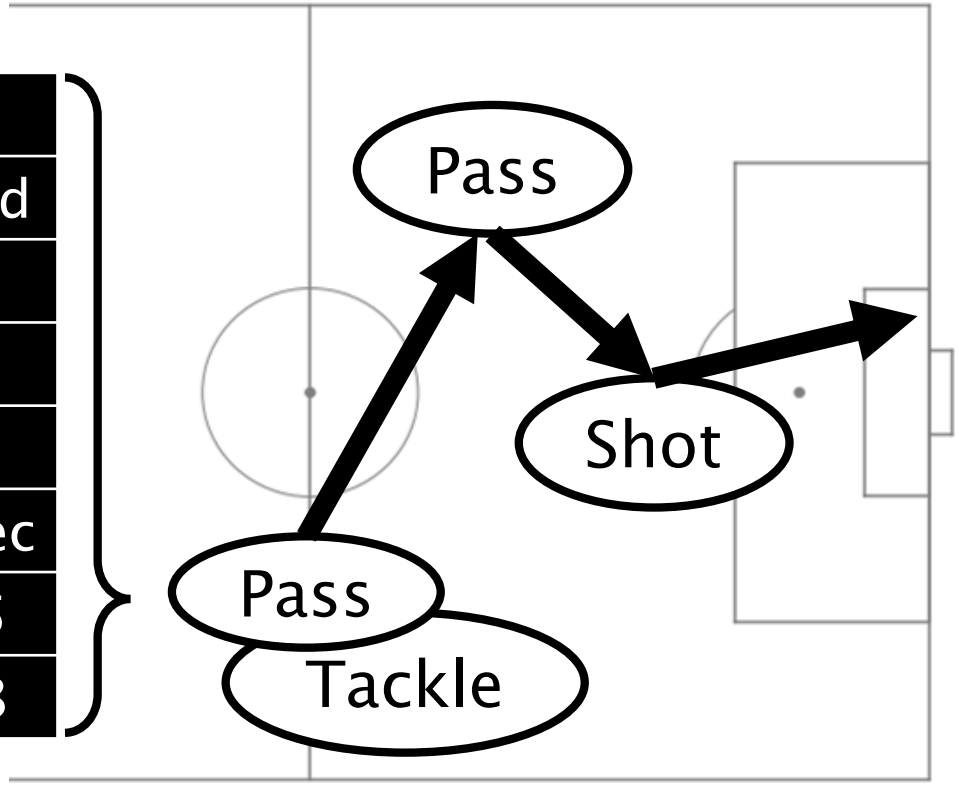
- Vendor-specific terminology

opta wyscout STATSBOMB STATS

- Optional extra information

Contribution 1: SPADL is a unified and simple language for describing on-the-ball player actions

Type:	Pass
Player:	Eden Hazard
Team:	Chelsea
Result:	Success
Bodypart:	Foot
Time:	12min 36sec
Start:	x=53 y=15
End:	x=74 y=48



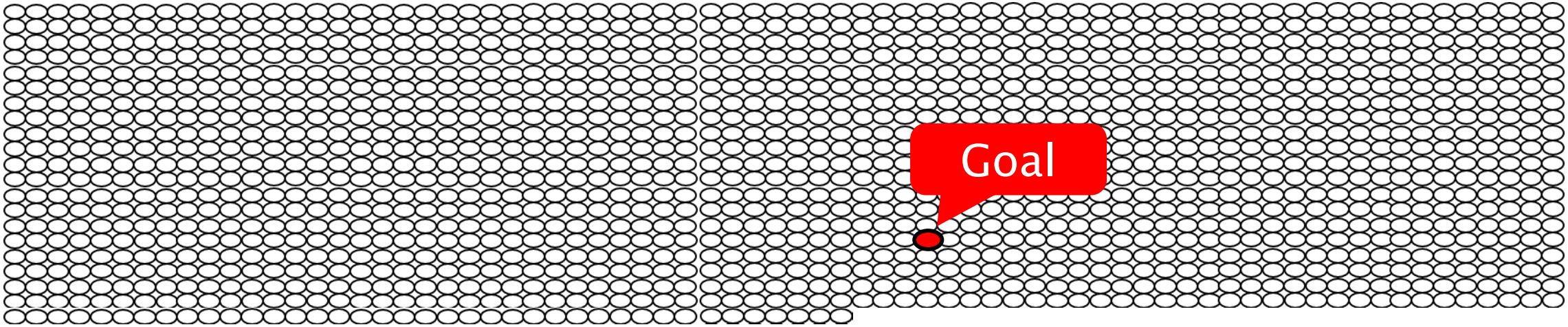
Converters available at: https://github.com/ML-KULeuven/socceraction/

KU LEUVEN DTAI SCISPORTS opta wyscout STATSBOMB

Challenge 2: Most actions do not affect the score

A soccer game contains ±1600 actions

Most common final score: 1 – 0



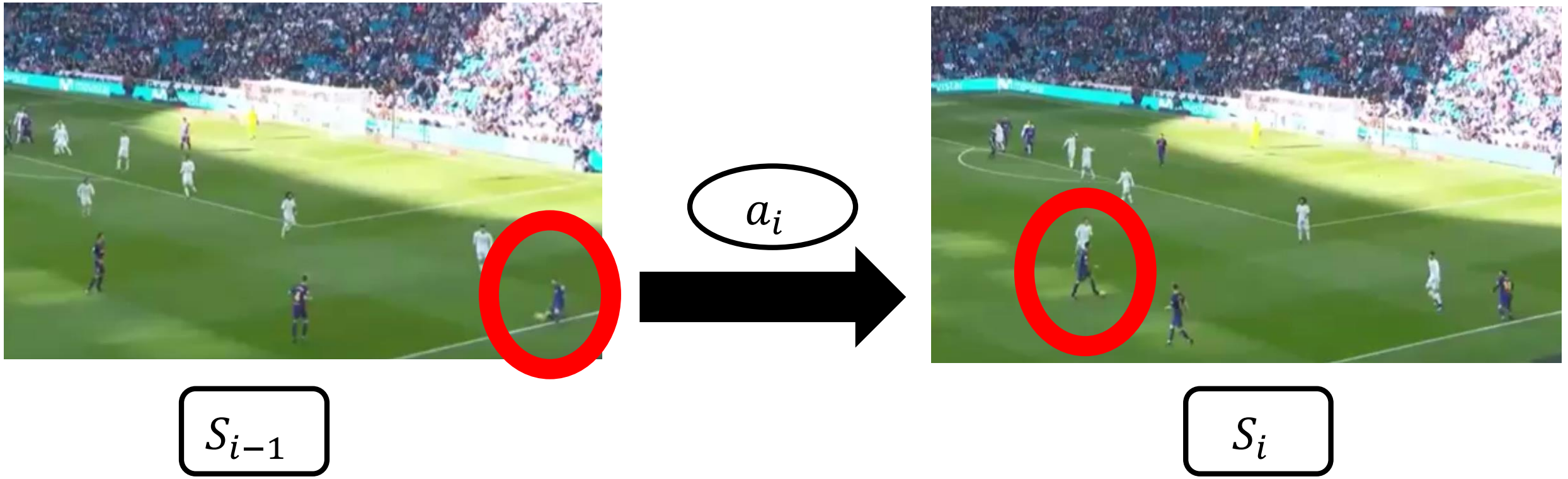
What is the value of an action?

How good is a player?

High-level questions with no objective ground truth

Example action: Pass from Messi to Busquets

Action a_i moves the game from state S_{i-1} to state S_i



Contribution 2: The VAEP framework values an action by its expected impact on the score

Intuition: a good action a_i for team T
(1) Increases the short-term probability of team T scoring and/or
(2) Decreases the short-term probability of team T conceding

$$\text{VAEP value}(a_i) = \Delta P_{scores}(a_i) - \Delta P_{concedes}(a_i)$$

$$\Delta P_{scores}(a_i) = P_{scores}(S_i, T) - P_{scores}(S_{i-1}, T)$$
$$\Delta P_{concedes}(a_i) = P_{concedes}(S_i, T) - P_{concedes}(S_{i-1}, T)$$

Transformation from subjective task to objective ML task: estimating P

Our ML task: Estimate $P_{scores}(S_i, T)$ and $P_{\text{concedes}}(S_i, T)$

X: Features

Y: Labels

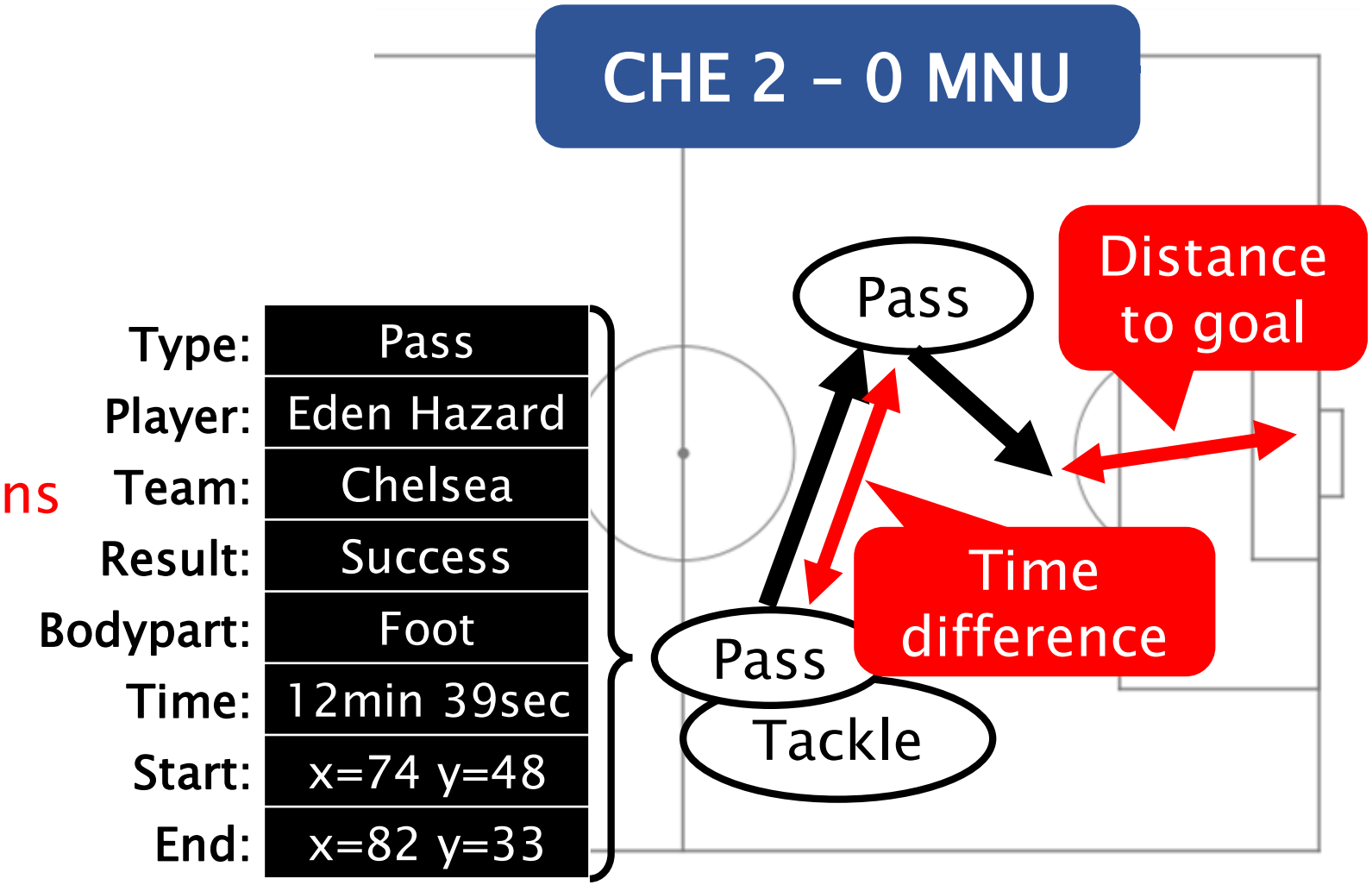
F: Probabilistic classifier

X: Features that describe game state S_i

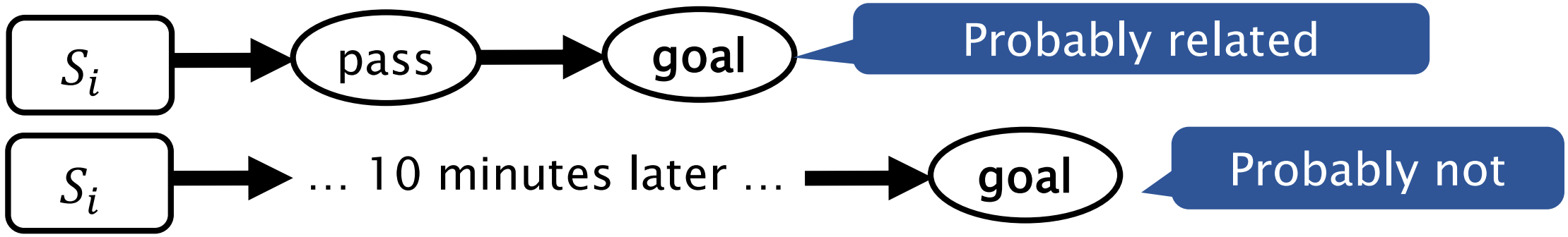
a) **Simple features**
- Action type
- Result
- …

b) **Complex features**
- Distance to goal
- Time between actions
- …

c) **Context features**
- Goal difference (e.g., +2, -1)



Y: Labels that capture S_i's limited temporal influence

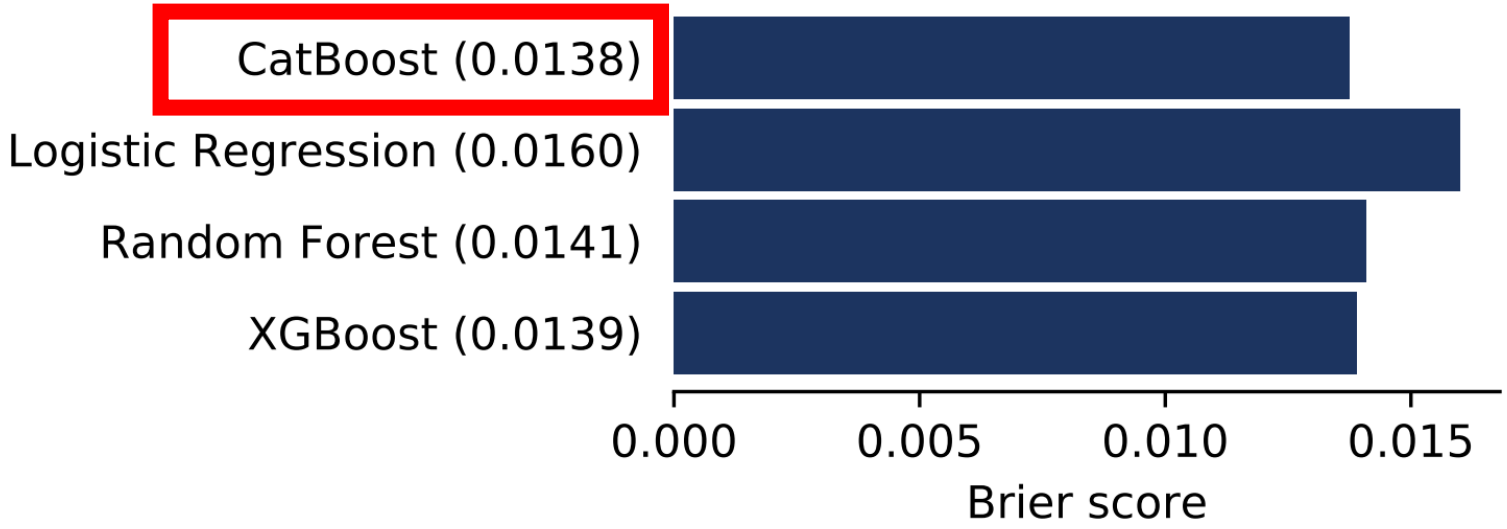


$$Y_{scores}(S_i, T) = \begin{cases} \mathbf{1} \text{ if team } T \text{ scores in the next 10 actions} \\ \mathbf{0} \text{ otherwise} \end{cases}$$

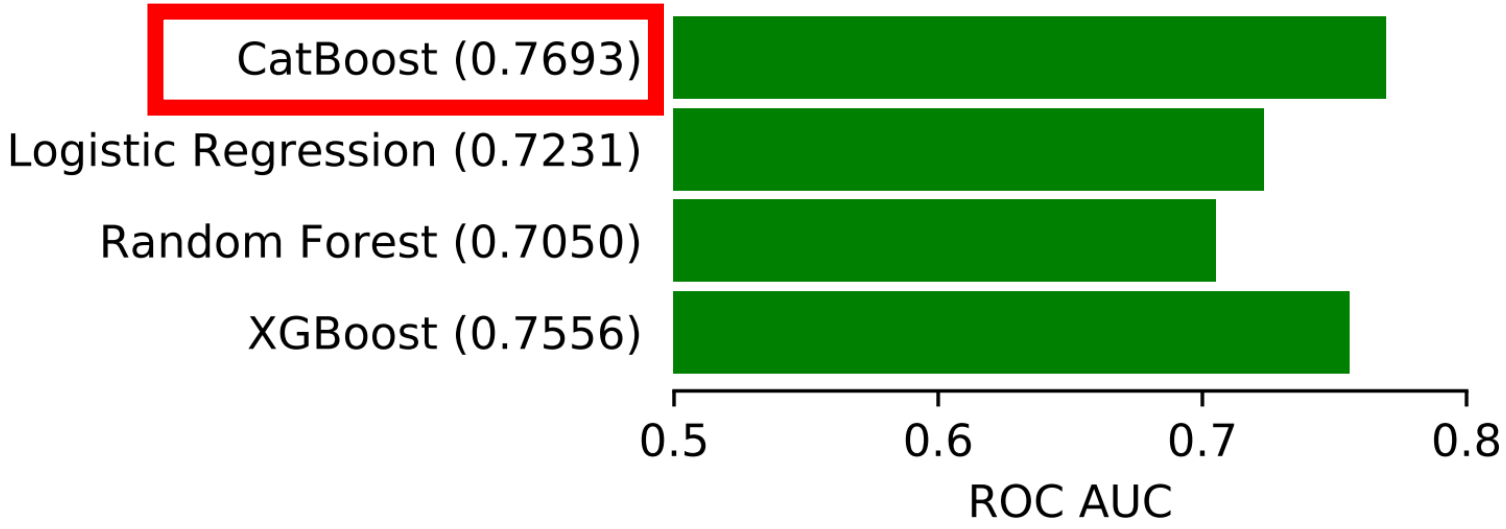
$$Y_{concedes}(S_i, T) = \begin{cases} \mathbf{1} \text{ if team } T \text{ concedes in the next 10 actions} \\ \mathbf{0} \text{ otherwise} \end{cases}$$

F: Probabilistic classifier

Brier score
- accuracy
- calibration



ROC AUC
works well for
unbalanced data sets



Our ML task: Estimate $P_{scores}(S_i, T)$ and $P_{\text{concedes}}(S_i, T)$

X: Features

Simple features + Complex features + Context features

Y: Labels

1 if team T scores/concedes in the next 10 actions

F: Probabilistic classifier

CatBoost

Our soccer analytics task: Value on-the-ball actions

X: Features

Simple features + Complex features + Context features

Y: Labels

1 if team T scores/concedes in the next 10 actions

F: Probabilistic classifier

CatBoost

Formula:

$$\text{VAEP value}(a_i) = \Delta P_{scores}(a_i) - \Delta P_{concedes}(a_i)$$

Intuitive illustration of VAEP values:
Barcelona's 3-0 goal vs Real Madrid (Dec 23, 2017)

Applications in player scouting

Rating players

Identifying top players

Comparing playing styles

The big question

Our soccer data

7

European competitions

Premier League, La Liga, Eredivisie, …

5

seasons

2012/13 – 2017/18

11,565

games

14,427,803

actions

Our soccer data

7

European competitions

Premier League, La Liga, Eredivisie, …

5

seasons

2012/13 – 2017/18

Test data

11,565

games

14,427,803

actions

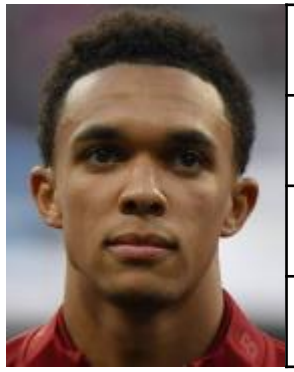
Rating players on expected score impact

Romelu Lukaku
Striker at Manchester United

2869 minutes
966 actions
16 goals
7 assists

VS

Trent Alexander-Arnold
Defender at Liverpool



1575 minutes
1528 actions
1 goal
2 assists

Naive metric: goals + assists per 90 minutes

0.72

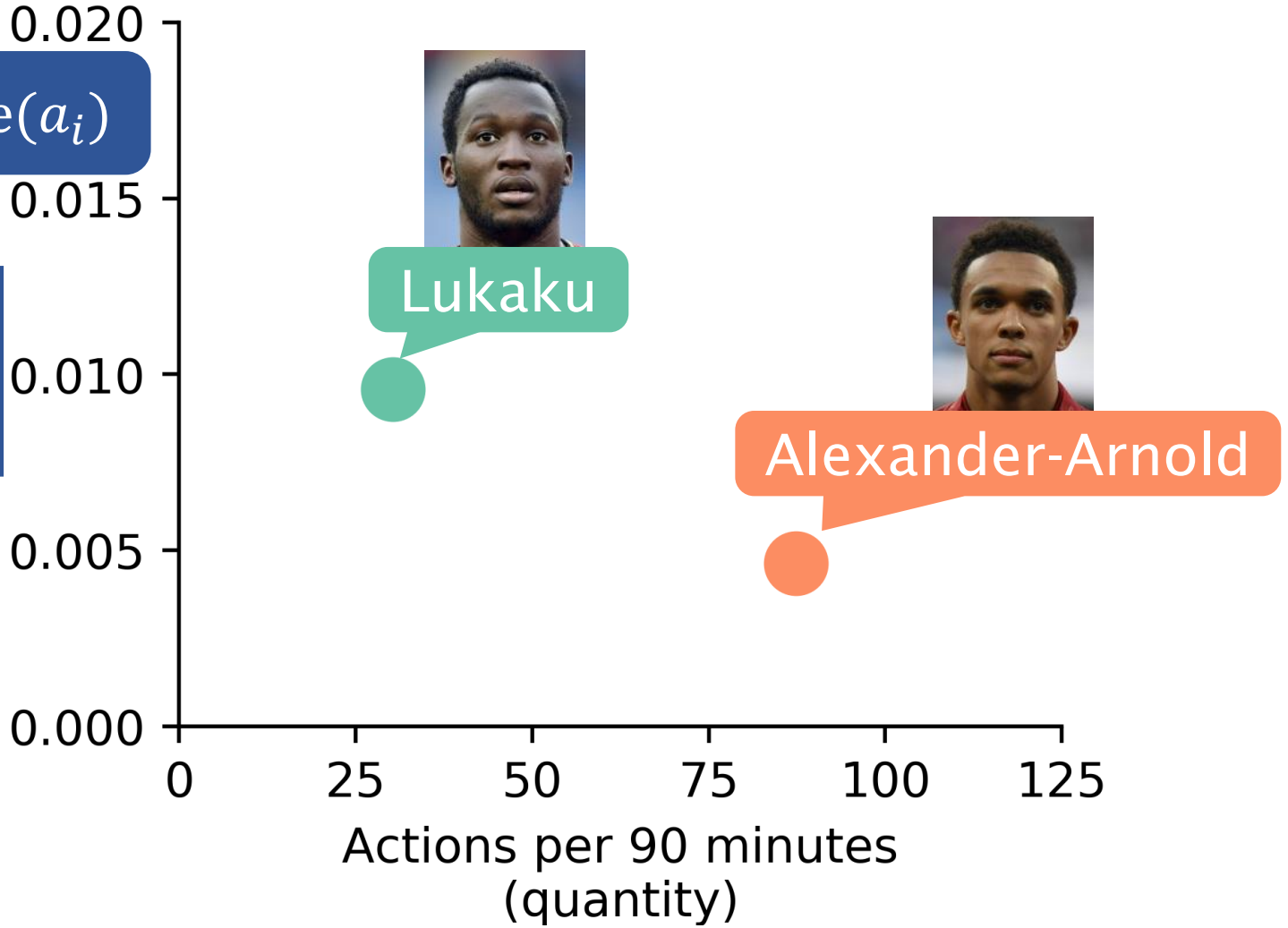
0.11

Let's rate players on ALL their actions instead

Rating players on expected score impact

$= \frac{1}{n} \sum_i^n \text{VAEP Value}(a_i)$

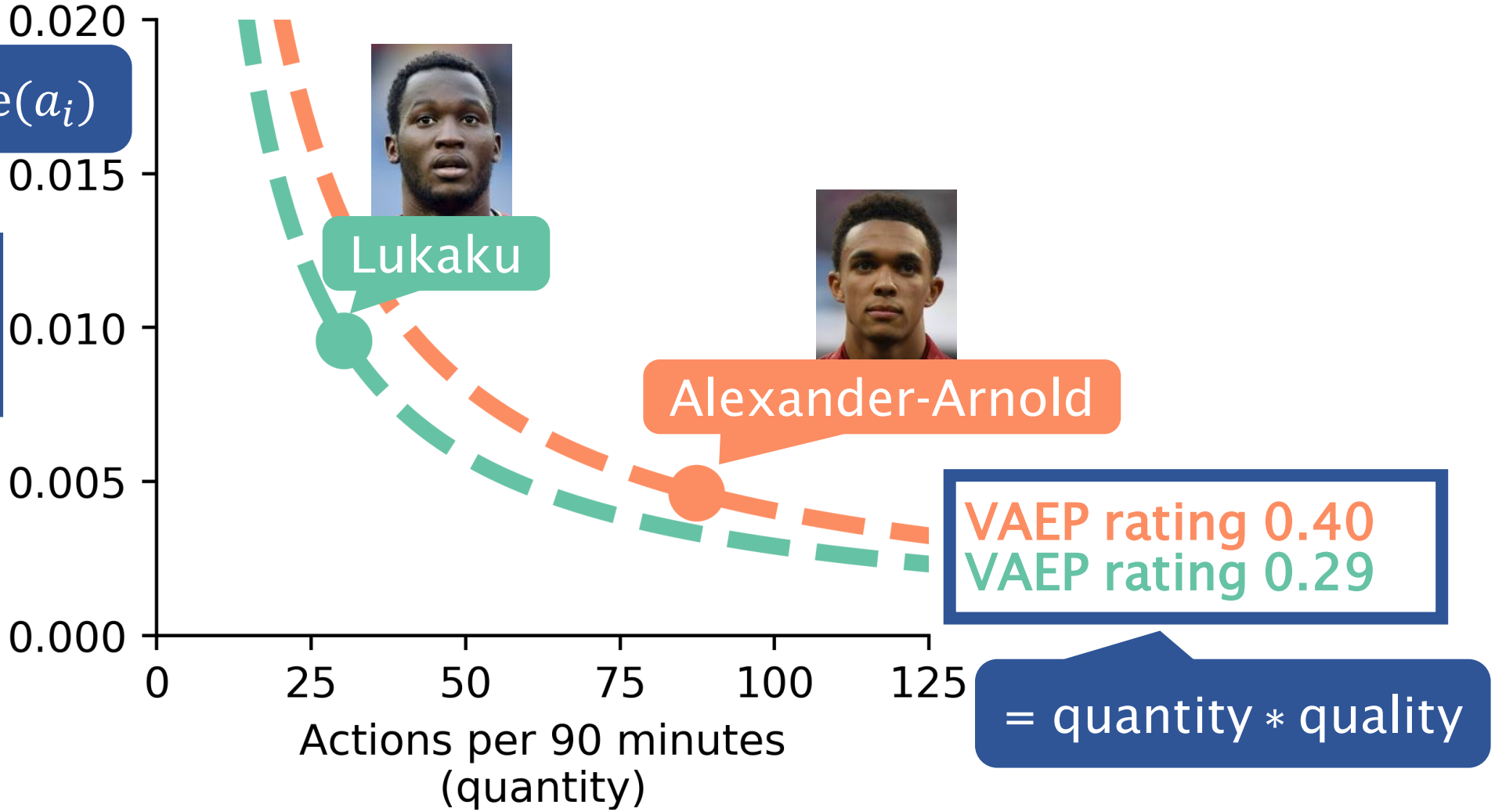
Average action value (quality)



Rating players on expected score impact

$= \frac{1}{n} \sum_i^n \text{VAEP Value}(a_i)$

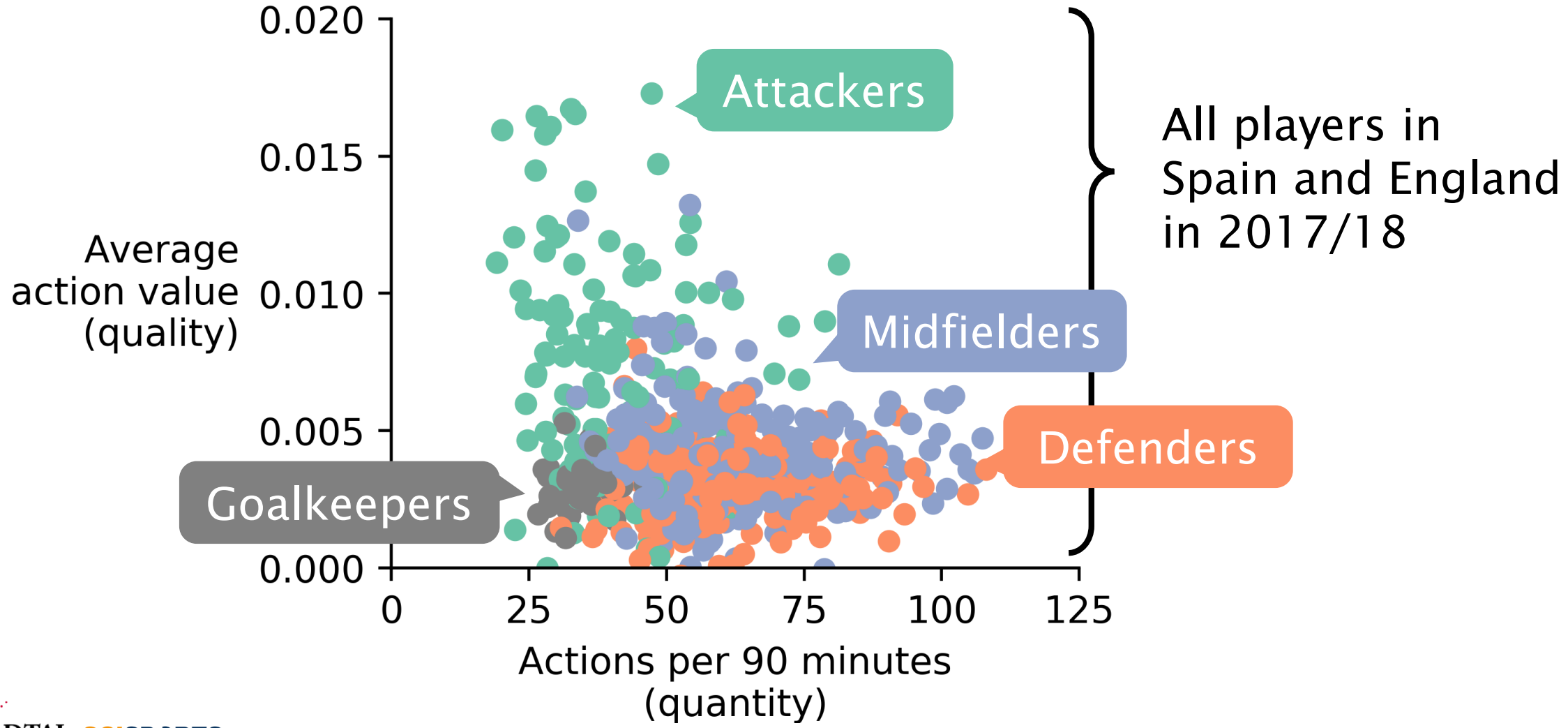
Average action value (quality)



= quantity ∗ quality

KU LEUVEN DTAI SCISPORTS

Rating players on expected score impact



Top-5 players in the 2017/18 Premier League

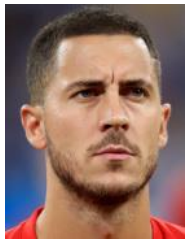
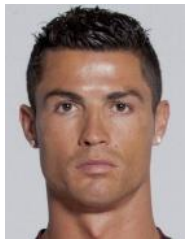
Rank	Player	VAEP rating	Price
1	Philippe Coutinho	0.90	€ 140m
2	Mohammed Salah	0.82	€ 150m
3	Kevin De Bruyne	0.64	€ 150m
4	Eden Hazard	0.64	€ 150m
5	Riyad Mahrez	0.63	€ 60m

Top-5 U21 players in the 2017/18 Dutch League

Rank	Player	Team	Age	VAEP rating	June 2018
1	David Neres	Ajax	21	0.62	€ 20m
2	Mason Mount	Vitesse	19	0.62	€ 4m
3	Frenkie de Jong	Ajax	20	0.50	€ 7m
4	Steven Bergwijn	PSV	20	0.49	€ 12m
5	Donny van de Beek	Ajax	21	0.47	€ 14m

Top-5 U21 players in the 2017/18 Dutch League

Rank	Player	Team	Age	VAEP rating	June 2018	June 2019	Price delta
1	David Neres	Ajax	21	0.62	€ 20m	€ 45m	+ €25m
2	Mason Mount	Vitesse	19	0.62	€ 4m	€ 12m	+ €8m
3	Frenkie de Jong	Ajax	20	0.50	€ 7m	€ 85m	+ €78m
4	Steven Bergwijn	PSV	20	0.49	€ 12m	€ 35m	+ €23m
5	Donny van de Beek	Ajax	21	0.47	€ 14m	€40m	+ €26m

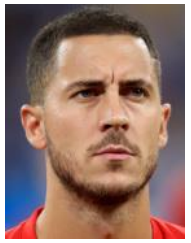
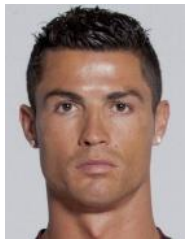
Can Hazard  replace Ronaldo  ?

VAEP rating 0.64

VAEP rating 0.61

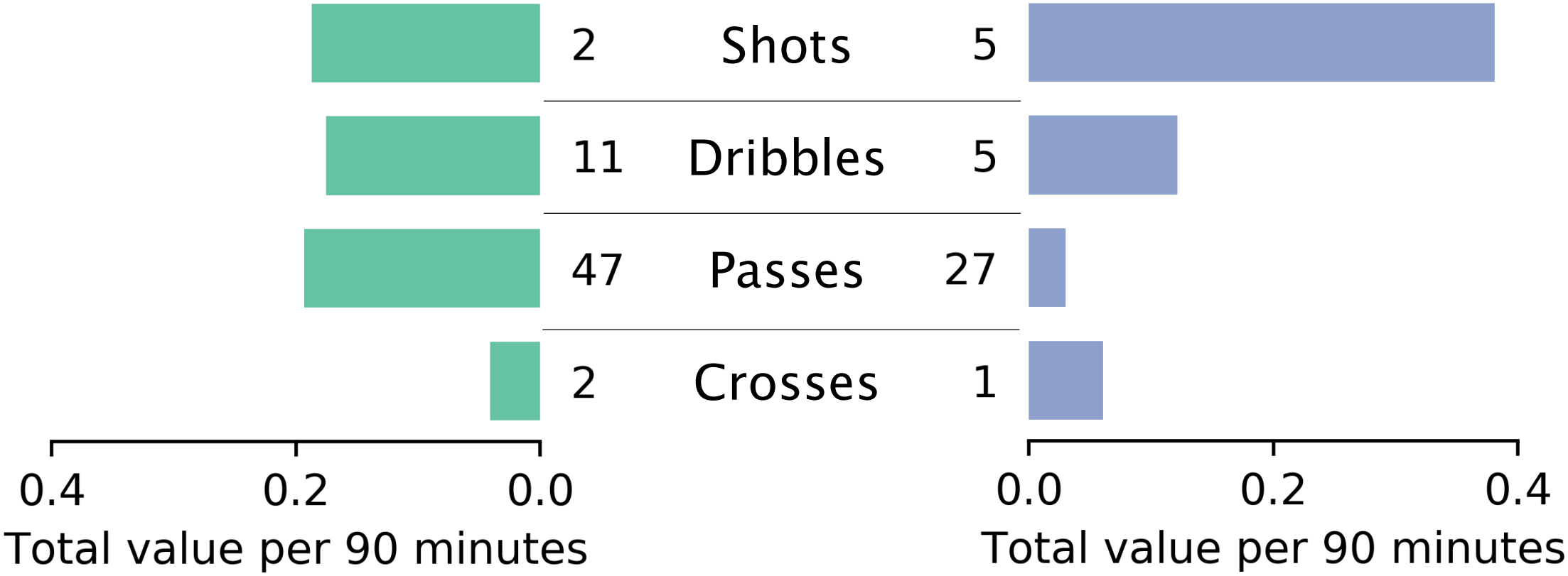
2	Shots	5
11	Dribbles	5
47	Passes	27
2	Crosses	1

Actions per 90 minutes

Can Hazard  replace Ronaldo  ?

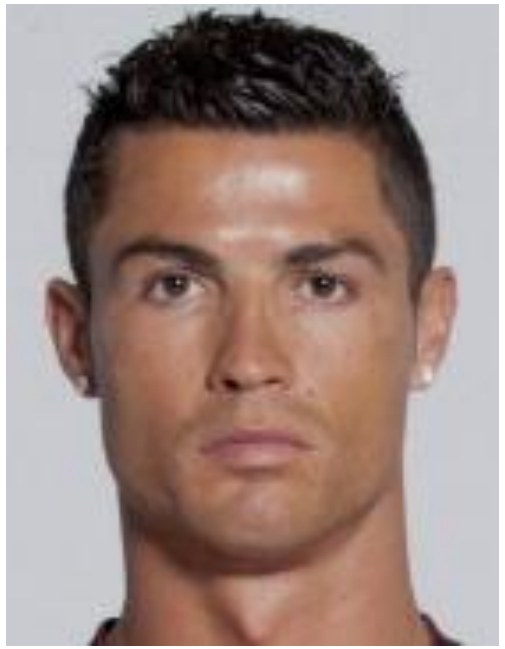
VAEP rating 0.64

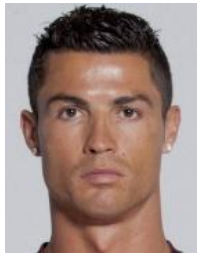
VAEP rating 0.61

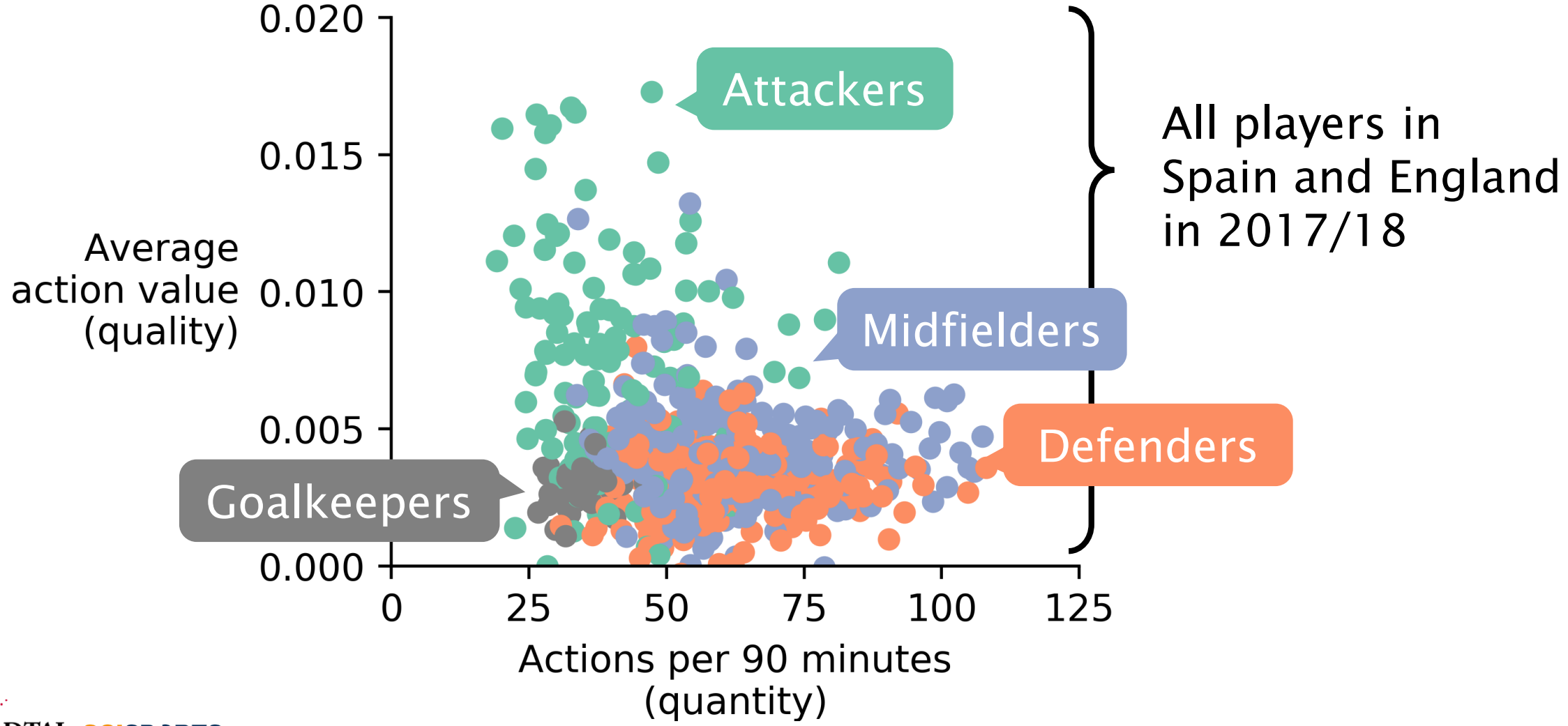


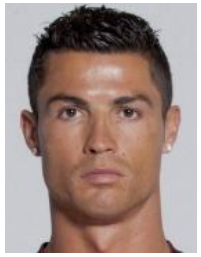
KU LEUVEN DTAI SCISPORTS

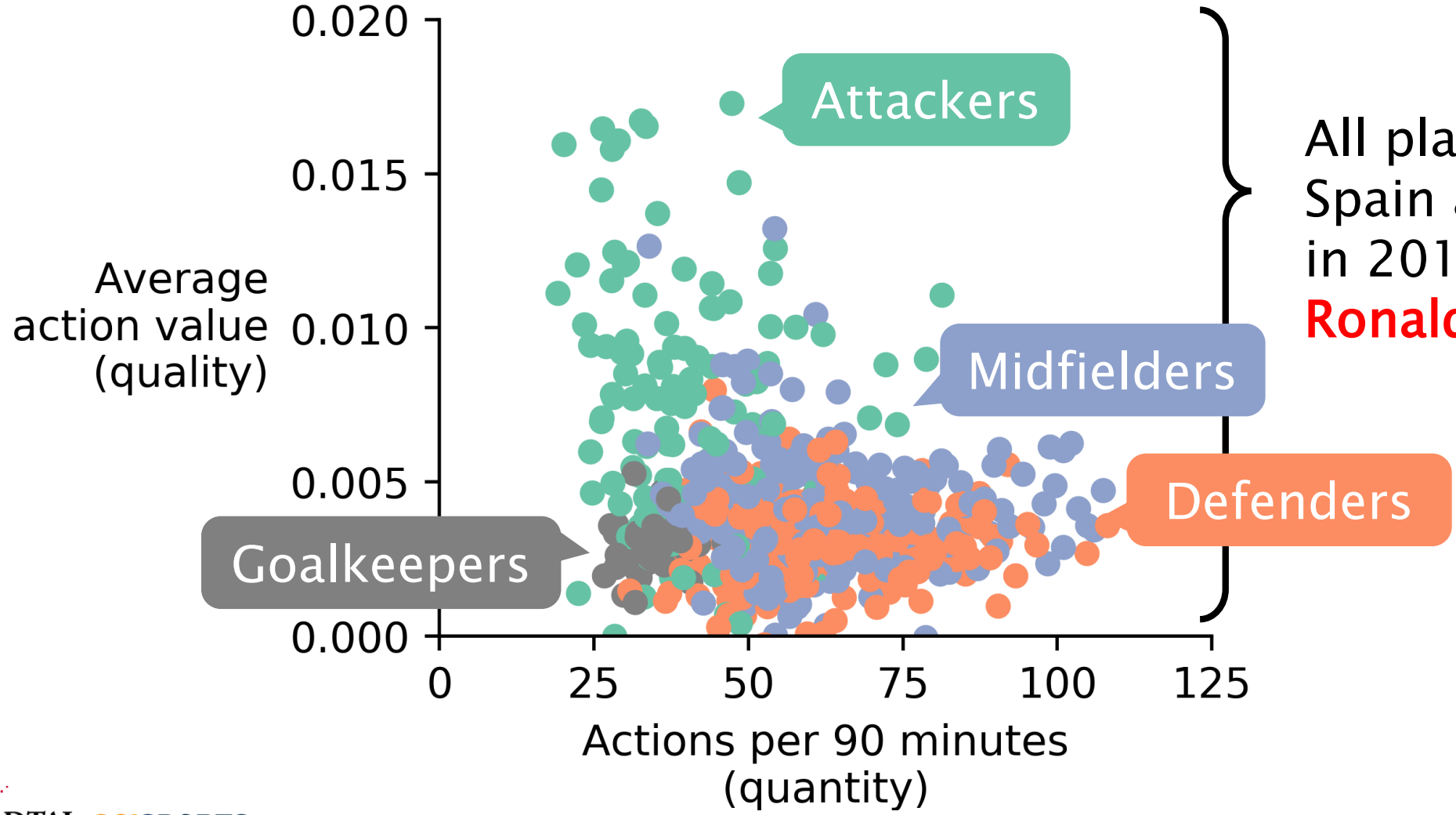
The big question



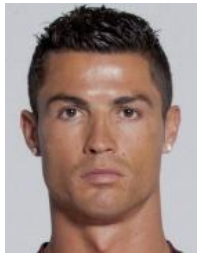
Who is better, Ronaldo  or Messi  ?

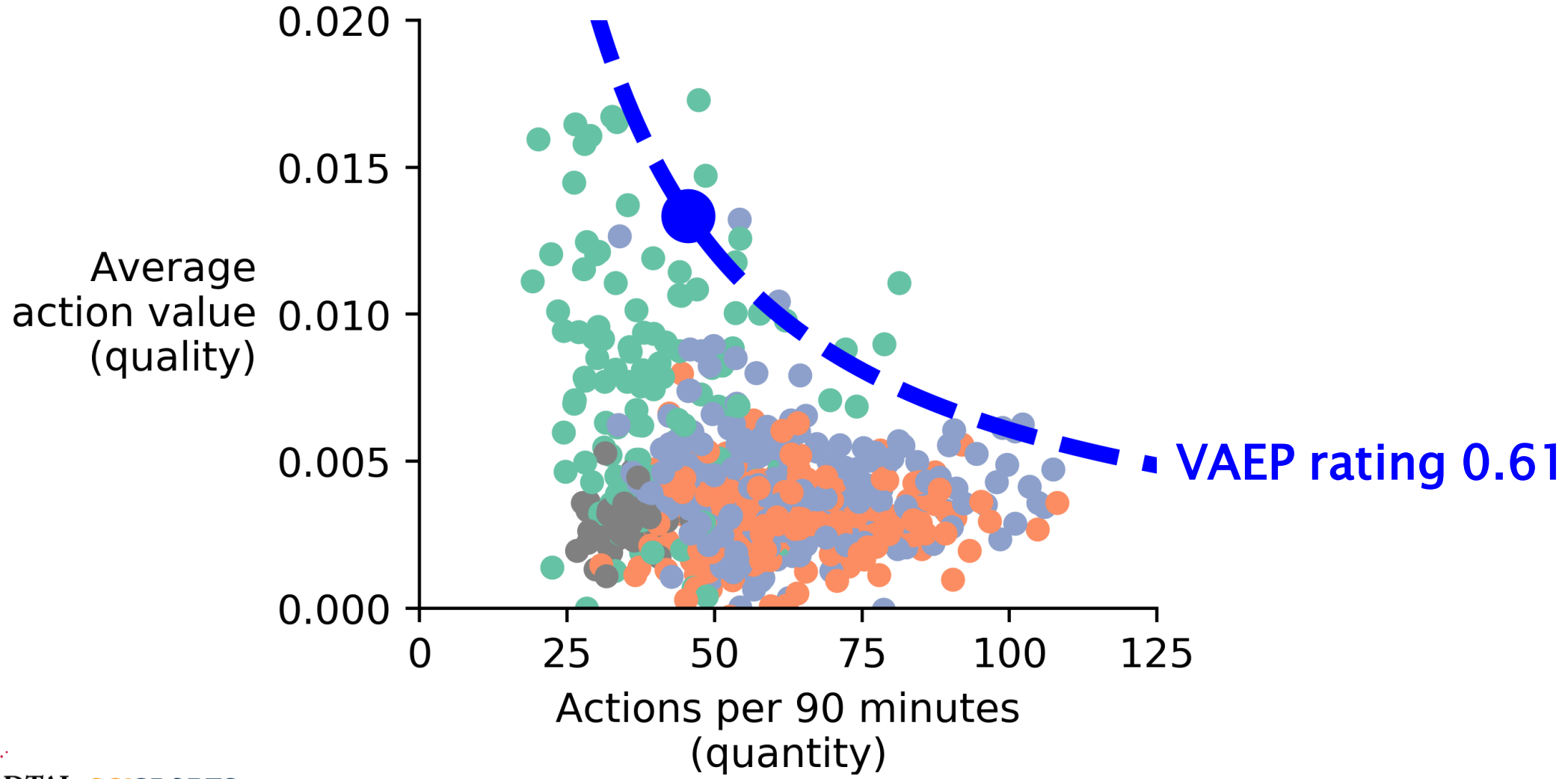


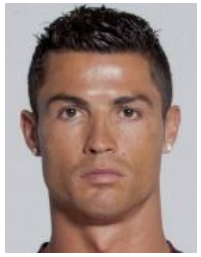
Who is better, Ronaldo  or Messi  ?

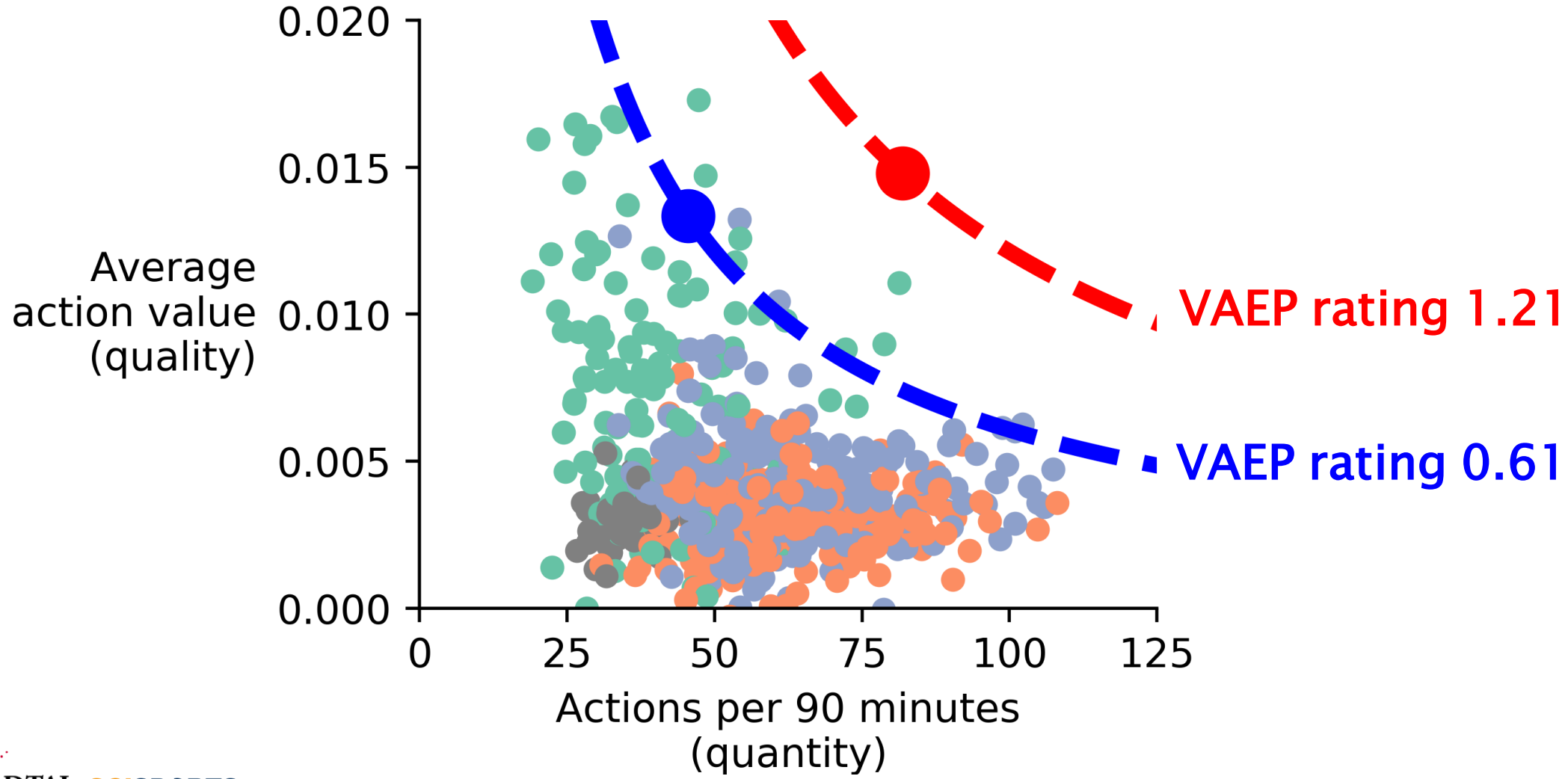


All players in Spain and England in 2017/18 **except Ronaldo and Messi**

Who is better, Ronaldo  or Messi  ?



Who is better, Ronaldo  or Messi  ?

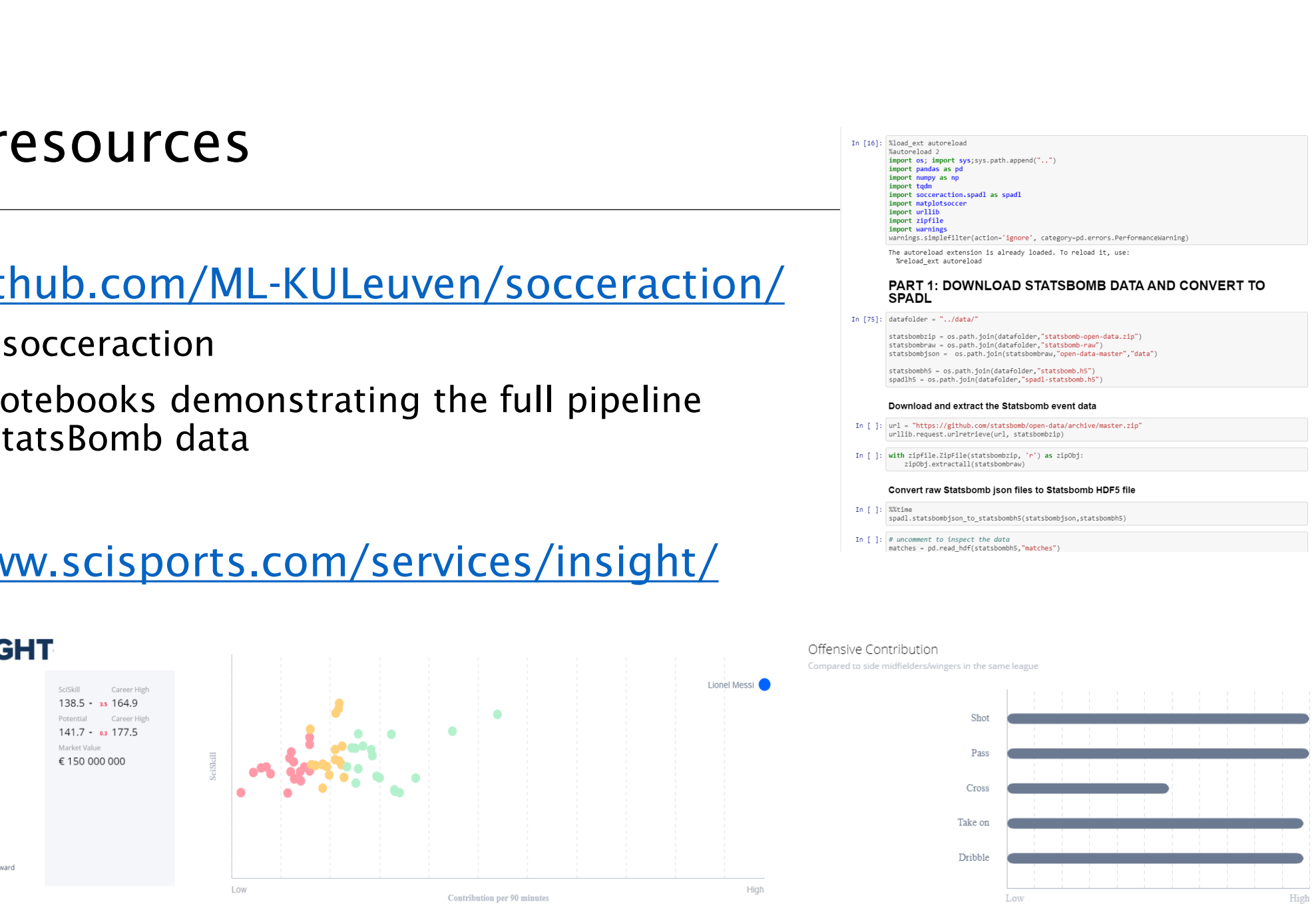


Online resources

https://github.com/ML-KULeuven/socceraction/

- pip install socceraction
- Example notebooks demonstrating the full pipeline with free StatsBomb data

https://www.scisports.com/services/insight/



KU LEUVEN DTAI SCISPORTS

Concluding thoughts

Challenges:

- Real-world soccer data != UCI data sets
- Often no ground truth available

Valuing all on-the-ball player actions:

- Captures information ignored by existing soccer stats
- Has many use cases, e.g., player scouting

Messi > Ronaldo ☺

Contributions

Authors


Tom


Lotte


Jan


Jesse

1. **SPADL**: a unified and simple language for soccer actions
2. **VAEP**: a framework to assign values to ALL actions in soccer
3. Use cases relevant for scouting
4. Code + notebooks: https://github.com/ML-KULeuven/socceraction/

KU LEUVEN

DTAI
Declarative Languages & Artificial Intelligence

SCISPORTS

nano4SPORTS
TECHNOLOGY FOR AN ACTIVE LIFESTYLE