

# Predicting the potential of professional soccer players

Ruben Vroonen

Tom Decroos

Jan Van Haaren

Jesse Davis

MLSA17 @ ECML/PKDD17

18/09/2017

# Predicting the potential of professional soccer players

Ruben Vroonen

Tom Decroos

Jan Van Haaren

Jesse Davis

MLSA17 @ ECML/PKDD17

18/09/2017

# Meet Bob, a young professional soccer player



Bob

Age: 19

Year: 2017

# Bob has a set of skill ratings



Bob

Age: 19

Year: 2017

Attacking: 75/100  
Defending: 67/100  
Stamina: 50/100  
Intelligence: 72/100

# Meet Bob from the future



Bob  
Age: 19  
Year: 2017



Bob  
Age: 21  
Year: 2019

Attacking: 75/100  
Defending: 67/100  
Stamina: 50/100  
Intelligence: 72/100

# What are his skill ratings?



Bob  
Age: 19  
Year: 2017

Attacking: 75/100  
Defending: 67/100  
Stamina: 50/100  
Intelligence: 72/100



Bob  
Age: 21  
Year: 2019

Attacking: ?/100  
Defending: ?/100  
Stamina: ?/100  
Intelligence: ?/100

# Overview

## Related Work

PECOTA and CARMELO

## Data

SoFIFA.com ratings

## APROPOS

Our approach for predicting players' potential

## Experiments

Evaluating the predictive accuracy

# Overview

## Related Work

PECOTA and CARMELO

## Data

SoFIFA.com ratings

## APROPOS

Our approach for predicting players' potential

## Experiments

Evaluating the predictive accuracy



Similar systems have already been deployed in baseball (1) and basketball (2)

**(1) PECOTA**

*Player Empirical Comparison Analysis Test Algorithm*

Nearest neighbors analysis on player statistics using Bill James's similarity scores

**(2) CARMELO**

*Career-Arc Regression Model Estimator with Local Optimization*

Nearest neighbors analysis on Wins Above Replacement (WAR) using simple similarity score

# Overview

## Related Work

PECOTA and CARMELO

## Data

SoFIFA.com ratings

## APROPOS

Our approach for predicting players' potential

## Experiments

Evaluating the predictive accuracy

# A player card from SoFIFA.com contains 24 skill ratings for a specific player and age

Players Teams Leagues Squads Discussions Search Player ... Language / 语言 Login Old layout

FIFA FIFA 17 Dec 8, 2016 In Game In Real Life Change Log Comments Related Squads Calculator Compare (0)

## Static information

Lionel Messi (ID: 158023)  
 RW Age 29 (Jun 24, 1987) 5'7" 159lbs

Overall rating **93** Potential **93** Value € 89M Wage € 340K

Preferred Foot Left International Reputation 5★ Weak Foot 4★ Skill Moves 4★ Work rate Medium / Medium Body Type Messi Real face Yes

FC Barcelona **85**★★★★★ Position RW Jersey number 10 Joined Jul 1, 2004 Contract valid until 2018

Argentina **83**★★★★★ Position RW Jersey number 10

My Shortlist (418) Like (405) Dislike (149) Jump to Version ▾

## Position ratings

POSITION	OVA	IR
RS ST LS	88	+3
RW LW	91	+2
RF CF LF	92	+1
RAM CAM LAM	92	+1
RCM CM LCM	84	+3
RM LM	90	+3
RDM CDM LDM	59	+3
RCB CB LCB	45	+3
RB LB	57	+3
RWB LWB	62	+3

## Dynamic information

Attacking	Skill	Movement	Power
77 Crossing	97 Dribbling	92 Acceleration	85 Shot Power
95 Finishing	89 Curve	87 Sprint Speed	68 Jumping
71 Heading Accuracy	90 Free Kick Accuracy	90 Agility	74 Stamina
88 Short Passing	87 Long Passing	95 Reactions	59 Strength
85 Volleys	95 Ball Control	95 Balance	88 Long Shots
Mentality	Defending	Goalkeeping	Traits
48 Aggression	13 Marking	6 GK Diving	Finesse Shot
22 Interceptions	28 Standing Tackle	11 GK Handling	Long Shot Taker
93 Positioning	26 Sliding Tackle	15 GK Kicking	One Club Player
90 Vision		14 GK Positioning	Specialities
74 Penalties		8 GK Reflexes	Dribbler
94 Composure			Free Kick Specialist
			Acrobat
			Clinical Finisher

# The data

## Competitions:

England, France, Germany, Italy and Spain

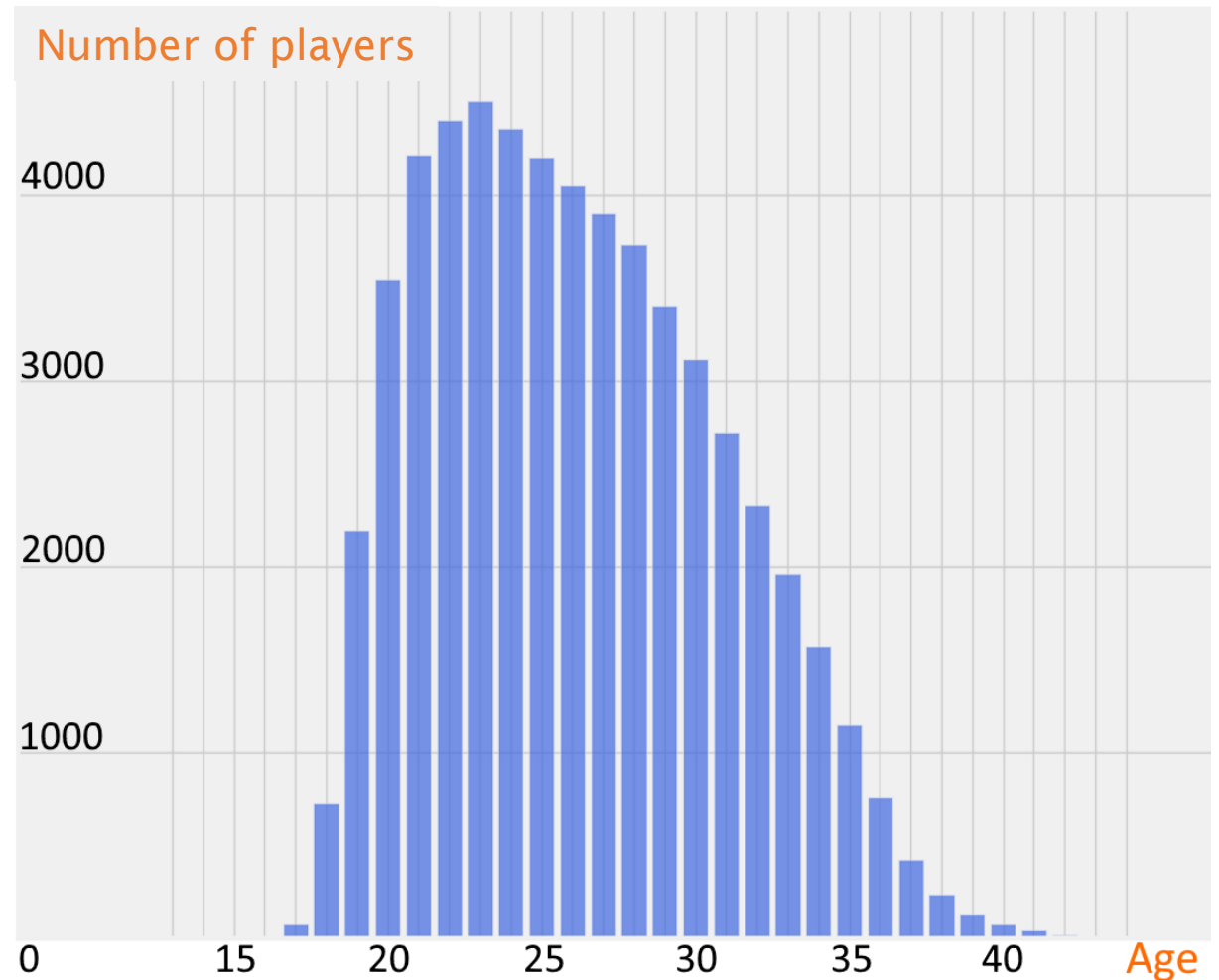
## Stats:

- 10,000 players
- 57,000 player cards
- Data from 2007-2017

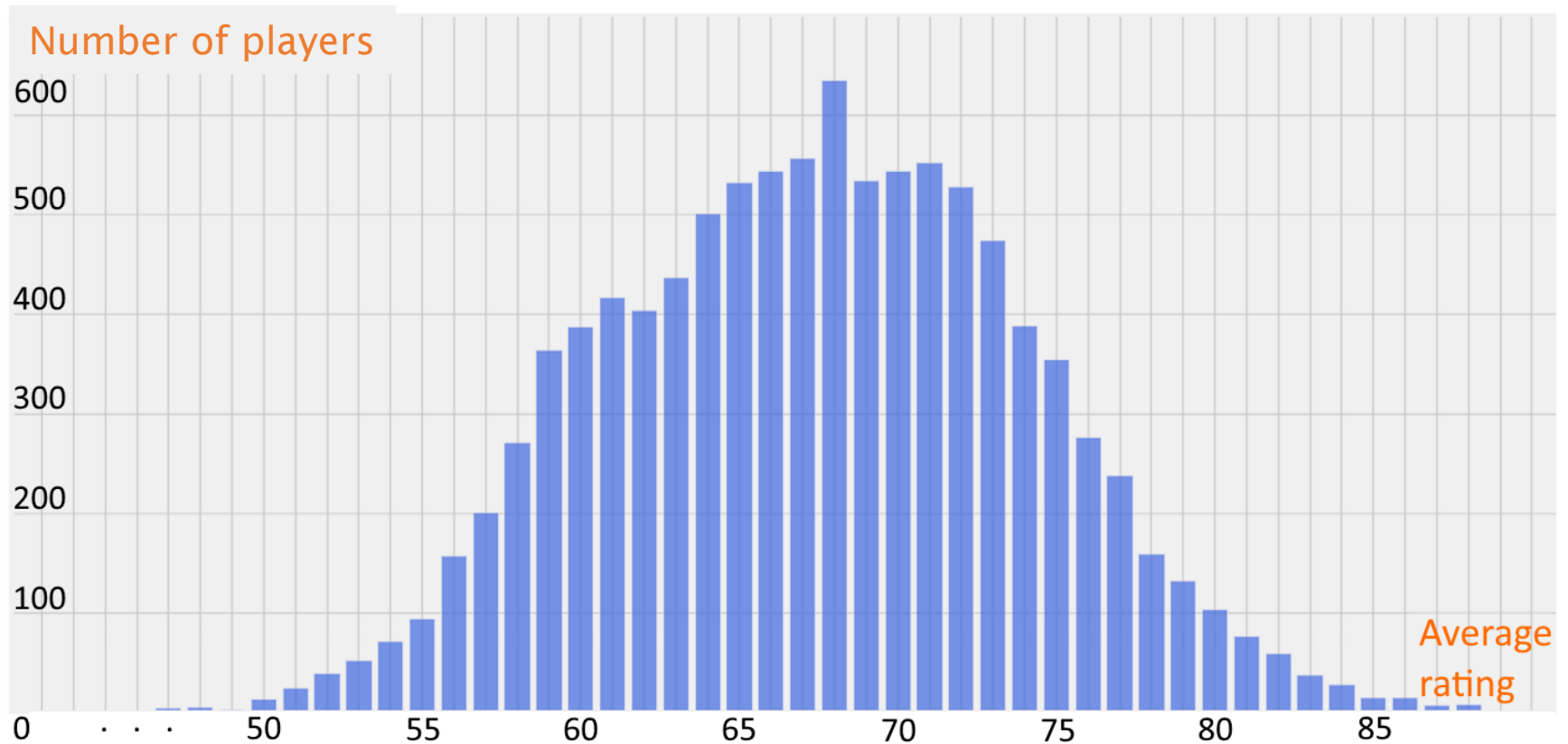
## Preprocessing challenges:

- Incorrect or missing age
- Position of substitute players

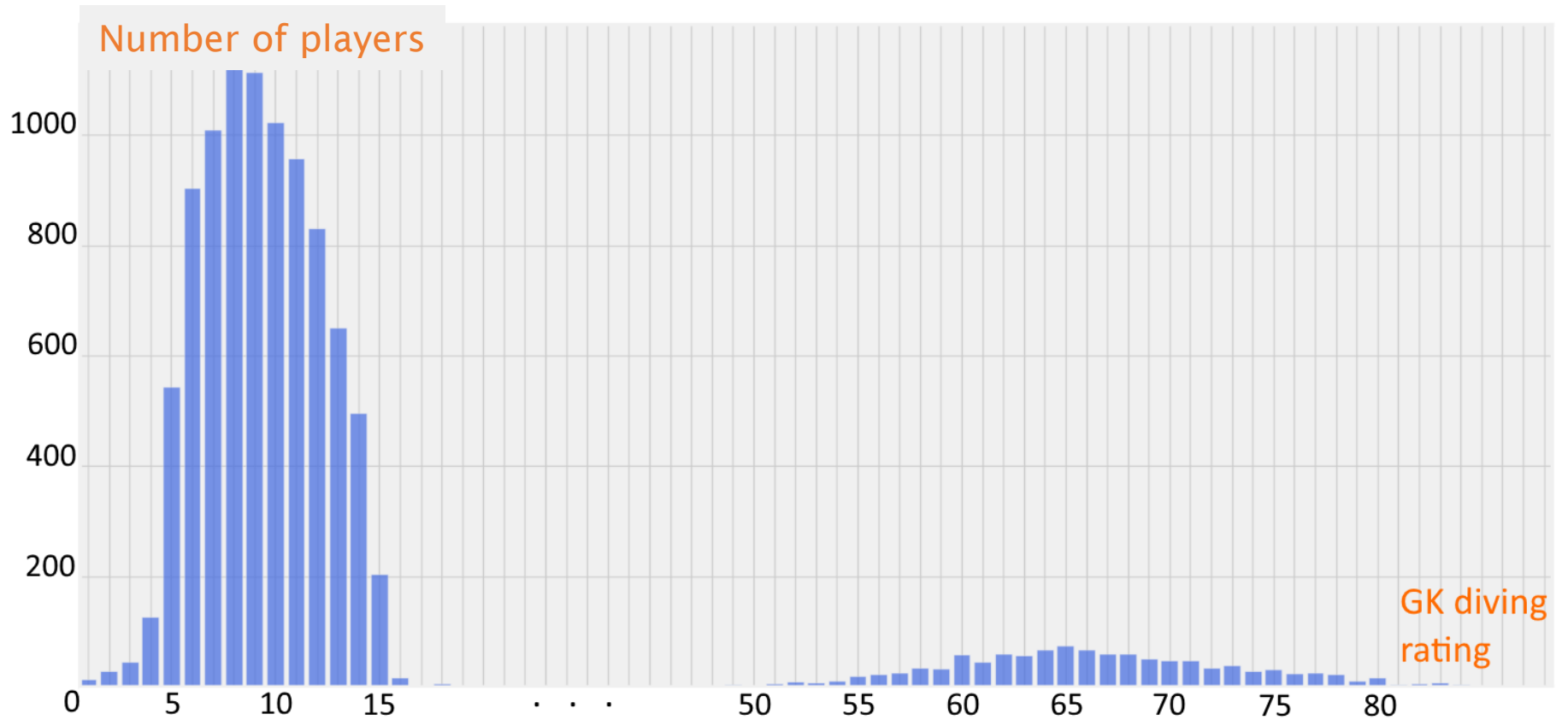
The most interesting categories (young and old players) have the least available data



# Most skill ratings follow a normal distribution...



... except goalkeeping skills



# Overview

## Related Work

PECOTA and CARMELO

## Data

SoFIFA.com ratings

## **APROPOS**

Our approach for predicting players' potential

## Experiments

Evaluating the predictive accuracy



Reminder: our task is to predict the skill ratings of future Bob



Bob  
Age: 19  
Year: 2017

Attacking: 75/100  
Defending: 67/100  
Stamina: 50/100  
Intelligence: 72/100



Bob  
Age: 21  
Year: 2019

Attacking: ?/100  
Defending: ?/100  
Stamina: ?/100  
Intelligence: ?/100

# APROPOS follows a nearest neighbors approach

Given:

- a player  $p$  and his current age  $a_1$
- a future age  $a_2$
- a database of players  $D$

Then:

1. Search players in  $D$  that are similar to  $p$  at age  $a_1$  and have data available for age  $a_2$ .
2. Predict the rating of  $p$  at age  $a_2$  by combining the ratings of similar players at age  $a_2$ .

# APROPOS follows a nearest neighbors approach

Given:

- a player  $p$  and his current age  $a_1$
- a future age  $a_2$
- a database of players  $D$

Then:

1. Search players in  $D$  that are **similar to  $p$  at age  $a_1$**  and have data available for age  $a_2$ .
2. Predict the rating of  $p$  at age  $a_2$  by combining the ratings of similar players at age  $a_2$ .

# APROPOS follows a nearest neighbors approach

Given:

- a player  $p$  and his current age  $a_1$
- a future age  $a_2$
- a database of players  $D$

Then:

**Similarity score**

1. Search players in  $D$  that are **similar to  $p$  at age  $a_1$**  and have data available for age  $a_2$ .
2. Predict the rating of  $p$  at age  $a_2$  by combining the ratings of similar players at age  $a_2$ .

APROPOS follows  
a nearest neighbors approach

Given:

- a player  $p$  and his current age  $a_1$
- a future age  $a_2$
- a database of players  $D$

Then:

**Similarity score**

1. Search players in  $D$  that are **similar to  $p$  at age  $a_1$**  and have data available for age  $a_2$ .
2. Predict the rating of  $p$  at age  $a_2$  by **combining the ratings** of similar players at age  $a_2$ .

APROPOS follows  
a nearest neighbors approach

Given:

- a player  $p$  and his current age  $a_1$
- a future age  $a_2$
- a database of players  $D$

Then:

**Similarity score**

1. Search players in  $D$  that are similar to  $p$  at age  $a_1$  and have data **Prediction method**
2. Predict the rating of  $p$  at age  $a_2$  by combining the ratings of similar players at age  $a_2$ .

APROPOS follows  
a nearest neighbors approach

Given:

- a player  $p$  and his current age  $a_1$
- a future age  $a_2$
- a database of players  $D$

Then:

**Similarity score**

**Absolute**

**Evolutional**

1. Search players in  $D$  that are  
**similar to  $p$  at age  $a_1$**  and  
have data

**Prediction method**

**Absolute**

2. Predict the rating of  $p$  at age  $a_2$  by  
**combining the ratings**  
of similar players at age  $a_2$ .

**Evolutional**

The **absolute** similarity score expresses the difference between skill **ratings**

	Bob			Alice		
Age	17	18	19	17	18	19
Dribbling score	68	72	78	81	82	85



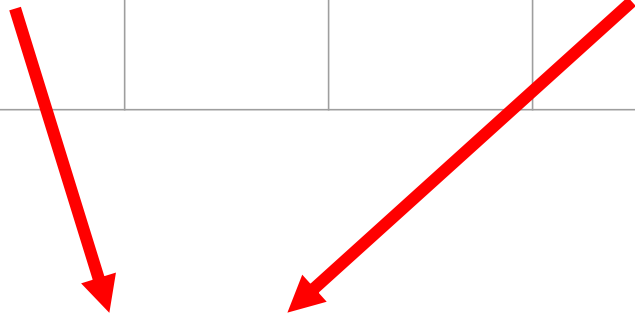
The **absolute** similarity score expresses the difference between skill **ratings**

	Bob			Alice		
Age	17	18	19	17	18	19
Dribbling score	68	72	78	81	82	85

$$f(\text{Bob}, \text{Alice}) = \sqrt{\hspace{20em}}$$

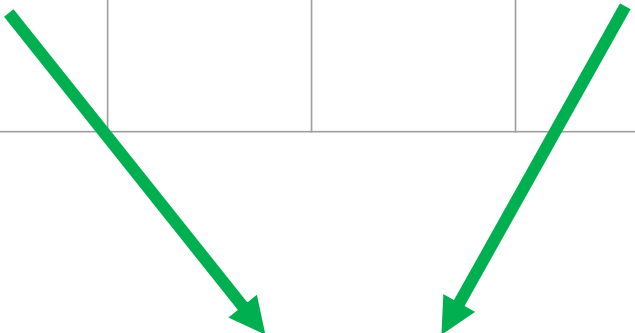
The **absolute** similarity score expresses the difference between skill **ratings**

	Bob			Alice		
Age	17	18	19	17	18	19
Dribbling score	68	72	78	81	82	85


$$f(\text{Bob}, \text{Alice}) = \sqrt{(68 - 81)^2}$$

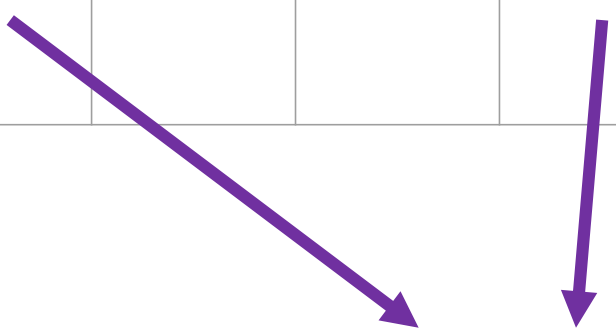
The **absolute** similarity score expresses the difference between skill **ratings**

	Bob			Alice		
Age	17	18	19	17	18	19
Dribbling score	68	72	78	81	82	85


$$f(\text{Bob}, \text{Alice}) = \sqrt{(68 - 81)^2 + (72 - 81)^2}$$

The **absolute** similarity score expresses the difference between skill **ratings**

	Bob			Alice		
Age	17	18	19	17	18	19
Dribbling score	68	72	78	81	82	85







$$f(\text{Bob}, \text{Alice}) = \sqrt{(68 - 81)^2 + (72 - 81)^2 + (78 - 85)^2}$$

The **evolutional** similarity score expresses the difference between skill **evolution**

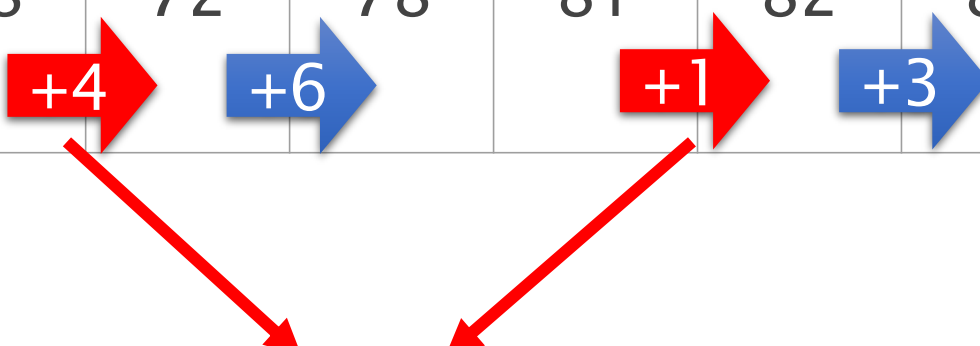
	Bob			Alice		
Age	17	18	19	17	18	19
Dribbling score	68	72	78	81	82	85

The **evolutional** similarity score expresses the difference between skill **evolution**

	Bob			Alice		
Age	17	18	19	17	18	19
Dribbling score	68	72	78	81	82	85
						

The **evolutional** similarity score expresses the difference between skill **evolution**

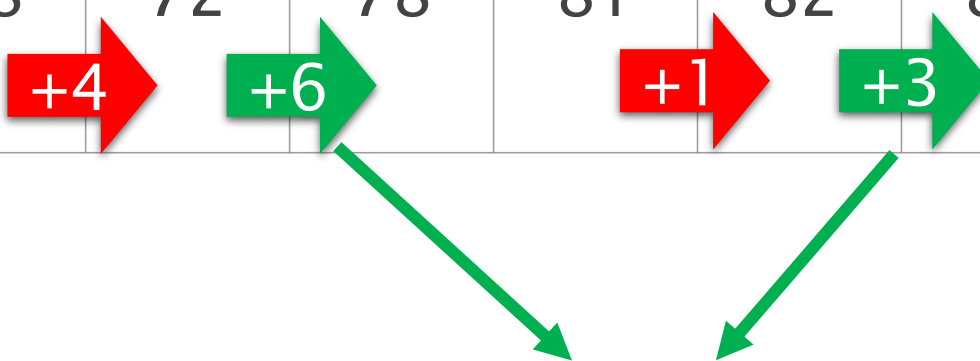
	Bob			Alice		
Age	17	18	19	17	18	19
Dribbling score	68	72	78	81	82	85



$$f(\text{Bob}, \text{Alice}) = \sqrt{(4 - 1)^2}$$

The **evolutional** similarity score expresses the difference between skill **evolution**

	Bob			Alice		
Age	17	18	19	17	18	19
Dribbling score	68	72	78	81	82	85



$$f(\text{Bob}, \text{Alice}) = \sqrt{(4 - 1)^2 + (1 - 3)^2}$$



The similarity score between players is computed as the average over all skills

$$\mathit{sim}(p, p') = \frac{\sum_{v \in V} \mathit{sim}_v(p, p')}{|V|}$$

The similarity score between players is computed as the average over all skills

**Total similarity  
between 2 players**

$$sim(p, p') = \frac{\sum_{v \in V} sim_v(p, p')}{|V|}$$

The similarity score between players is computed as the average over all skills

Total similarity  
between 2 players

Normalized similarity  
per skill (e.g. dribbling)

$$sim(p, p') = \frac{\sum_{v \in V} sim_v(p, p')}{|V|}$$

The similarity score between players is computed as the average over all skills

Total similarity  
between 2 players

Normalized similarity  
per skill (e.g. dribbling)

$$sim(p, p') = \frac{\sum_{v \in V} sim_v(p, p')}{|V|}$$

Total number of  
skills (=24)

APROPOS follows  
a nearest neighbors approach

Given:

- a player  $p$  and his current age  $a_1$
- a future age  $a_2$
- a database of players  $D$

Then:

**Similarity score**

**Absolute**

**Evolutional**

1. Search players in  $D$  that are  
**similar to  $p$  at age  $a_1$**  and  
have data

**Prediction method**

**Absolute**

2. Predict the rating of  $p$  at age  $a_2$  by  
**combining the player ratings** at age  $a_2$ .

**Evolutional**

We want to predict Bob's dribbling rating at age 21

	Bob	
Age	19	21
Dribbling	78	?

Alice is a similar player to Bob  
for whom we have historical data

	Bob		Alice	
Age	19	21	19	21
Dribbling	78	?	85	86

$$\begin{aligned} \text{Sim}(\text{Bob}, \text{Alice}) \\ = 0.7 \end{aligned}$$

Eve is also a similar player to Bob for whom we have historical data

	Bob		Alice		Eve	
Age	19	21	19	21	19	21
Dribbling	78	?	85	86	64	75

$$\begin{array}{ll} \textit{Sim}(\textit{Bob}, \textit{Alice}) & \textit{Sim}(\textit{Bob}, \textit{Eve}) \\ = 0.7 & = 0.8 \end{array}$$



The **absolute** prediction method uses the skill **ratings** of similar players

	Bob		Alice		Eve	
Age	19	21	19	21	19	21
Dribbling	78	?	85	86	64	75

$$\begin{array}{ll} \textit{Sim}(\textit{Bob}, \textit{Alice}) & \textit{Sim}(\textit{Bob}, \textit{Eve}) \\ = 0.7 & = 0.8 \end{array}$$

The **absolute** prediction method uses the skill **ratings** of similar players

	Bob		Alice		Eve	
Age	19	21	19	21	19	21
Dribbling	78	?	85	86	64	75

$$\begin{array}{ll}
 \textit{Sim}(\textit{Bob}, \textit{Alice}) & \textit{Sim}(\textit{Bob}, \textit{Eve}) \\
 = 0.7 & = 0.8
 \end{array}$$

*Dribbling prediction* = \_\_\_\_\_

The **absolute** prediction method uses the skill **ratings** of similar players

	Bob		Alice		Eve	
Age	19	21	19	21	19	21
Dribbling	78	?	85	86	64	75

$$\begin{aligned} \text{Sim}(\text{Bob}, \text{Alice}) &= 0.7 \\ \text{Sim}(\text{Bob}, \text{Eve}) &= 0.8 \end{aligned}$$

$$\text{Dribbling prediction} = \frac{0.7 * 86 + 0.8 * 75}{0.7 + 0.8} = 80$$



The **evolutional** prediction method uses the skill **evolutions** of similar players

	Bob		Alice		Eve	
Age	19	21	19	21	19	21
Dribbling	78	?	85	86	64	75

$$\begin{array}{ll} \textit{Sim}(\textit{Bob}, \textit{Alice}) & \textit{Sim}(\textit{Bob}, \textit{Eve}) \\ = 0.7 & = 0.8 \end{array}$$



The **evolutional** prediction method uses the skill **evolutions** of similar players

	Bob		Alice		Eve	
Age	19	21	19	21	19	21
Dribbling	78	?	85	86	64	75

$$\begin{array}{l} \textit{Sim}(\textit{Bob}, \textit{Alice}) \\ = 0.7 \end{array} \quad \begin{array}{l} \textit{Sim}(\textit{Bob}, \textit{Eve}) \\ = 0.8 \end{array}$$

The **evolutional** prediction method uses the skill **evolutions** of similar players



	Bob		Alice		Eve	
Age	19	21	19	21	19	21
Dribbling	78	?	85	86	64	75
						

$$\begin{array}{ll}
 \textit{Sim}(\textit{Bob}, \textit{Alice}) & \textit{Sim}(\textit{Bob}, \textit{Eve}) \\
 = 0.7 & = 0.8
 \end{array}$$

$$\textit{Dribbling prediction} = 78 + \text{—————}$$

The **evolutional** prediction method uses the skill **evolutions** of similar players

	Bob		Alice		Eve	
Age	19	21	19	21	19	21
Dribbling	78	?	85	86	64	75

$$\begin{aligned}
 \text{Sim}(\text{Bob}, \text{Alice}) &= 0.7 & \text{Sim}(\text{Bob}, \text{Eve}) &= 0.8
 \end{aligned}$$

$$\text{Dribbling prediction} = 78 + \frac{0.7 * 1 + 0.8 * 11}{0.7 + 0.8} = 84$$

# Overview

## Related Work

PECOTA and CARMELO

## Data

SoFIFA.com ratings

## APROPOS

Our approach for predicting players' potential

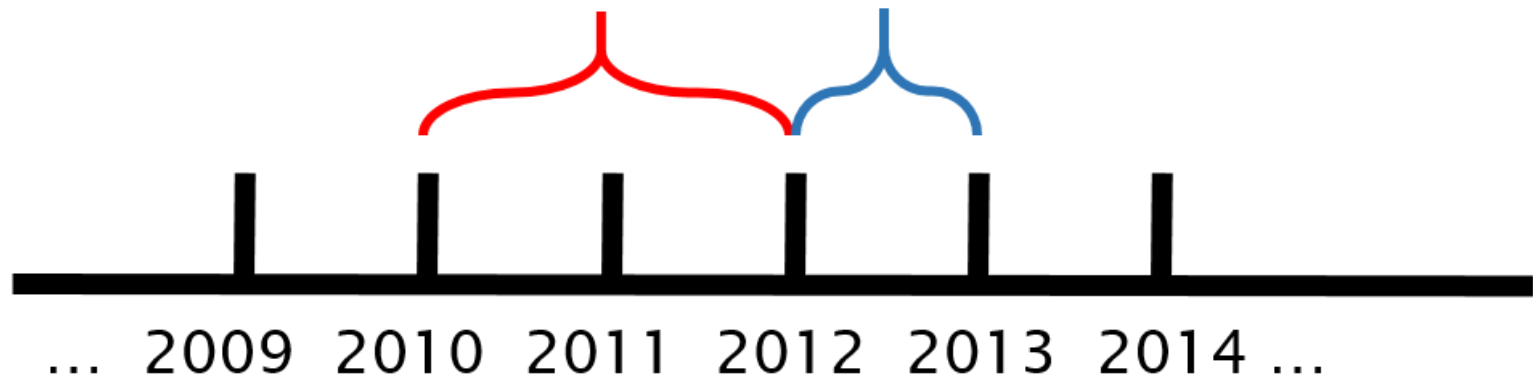
## Experiments

Evaluating the predictive accuracy



We predict skill ratings  
for 1000 players in 2012

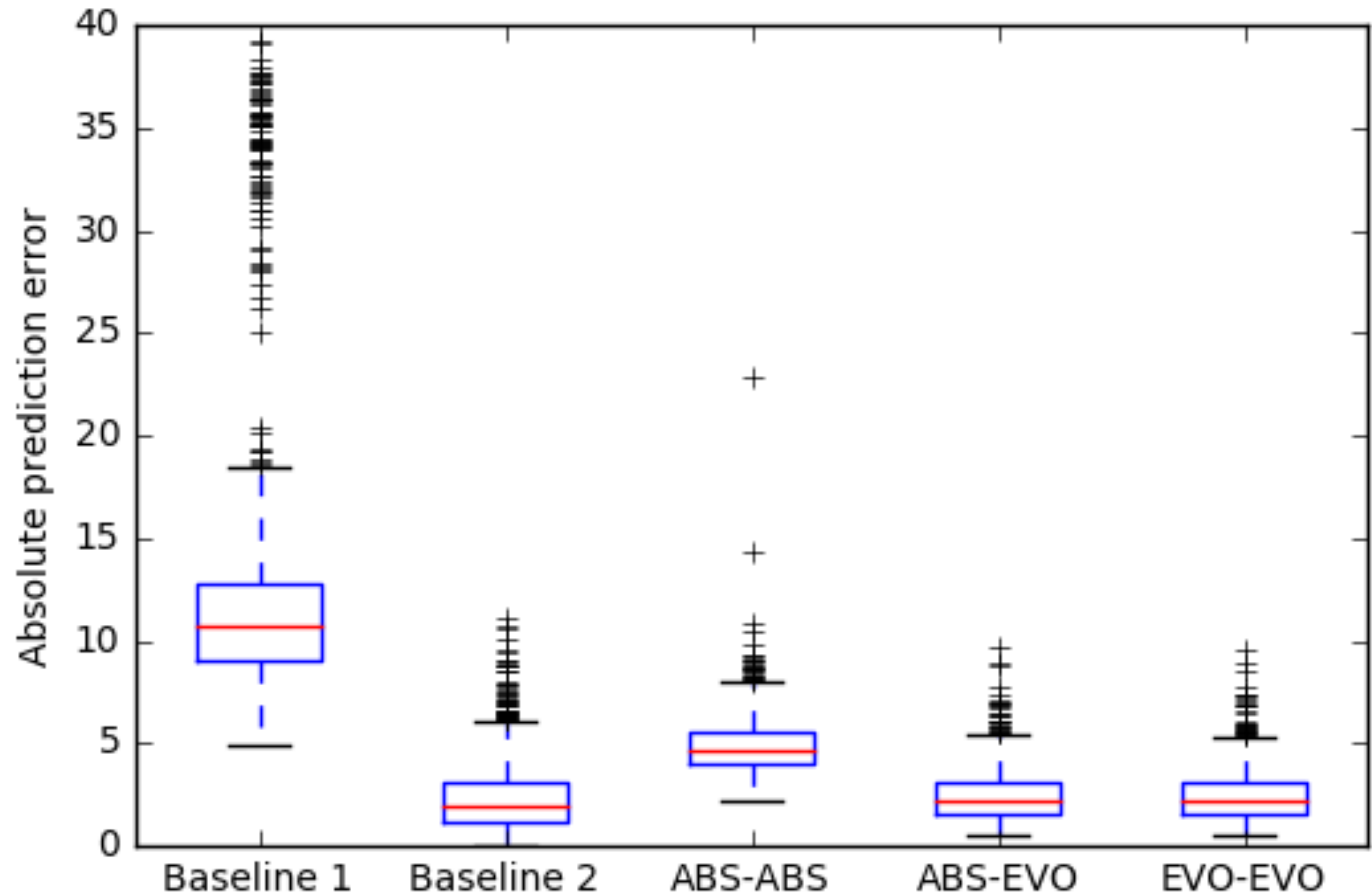
Similarity period = 3 years  
Prediction period = 1 year



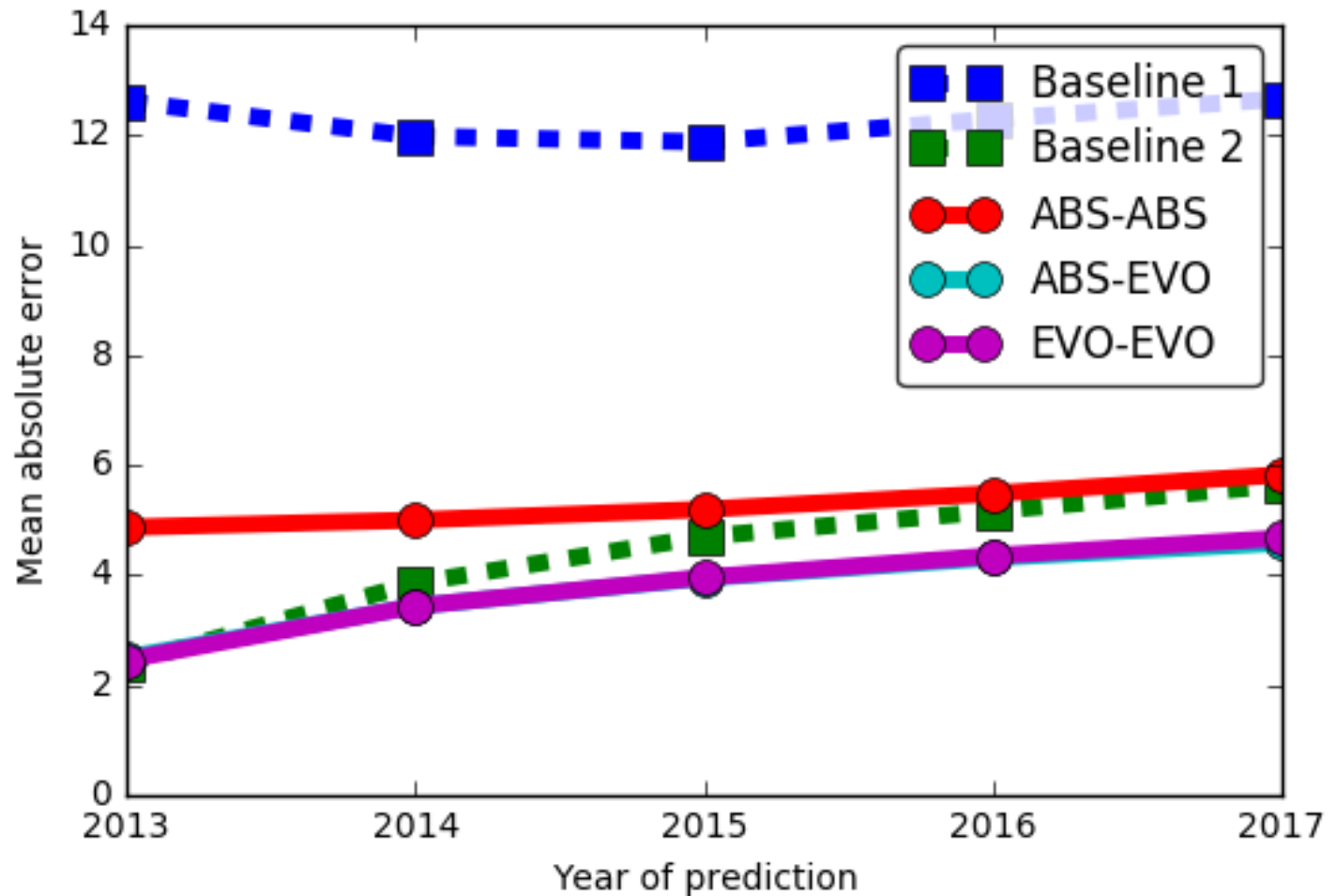
We compare 2 baseline models  
against 3 instances of APROPOS

1. Baseline 1: average skill rating of age group
2. Baseline 2: current skill rating as prediction
3. ABS-ABS
4. ABS-EVO
5. EVO-EVO

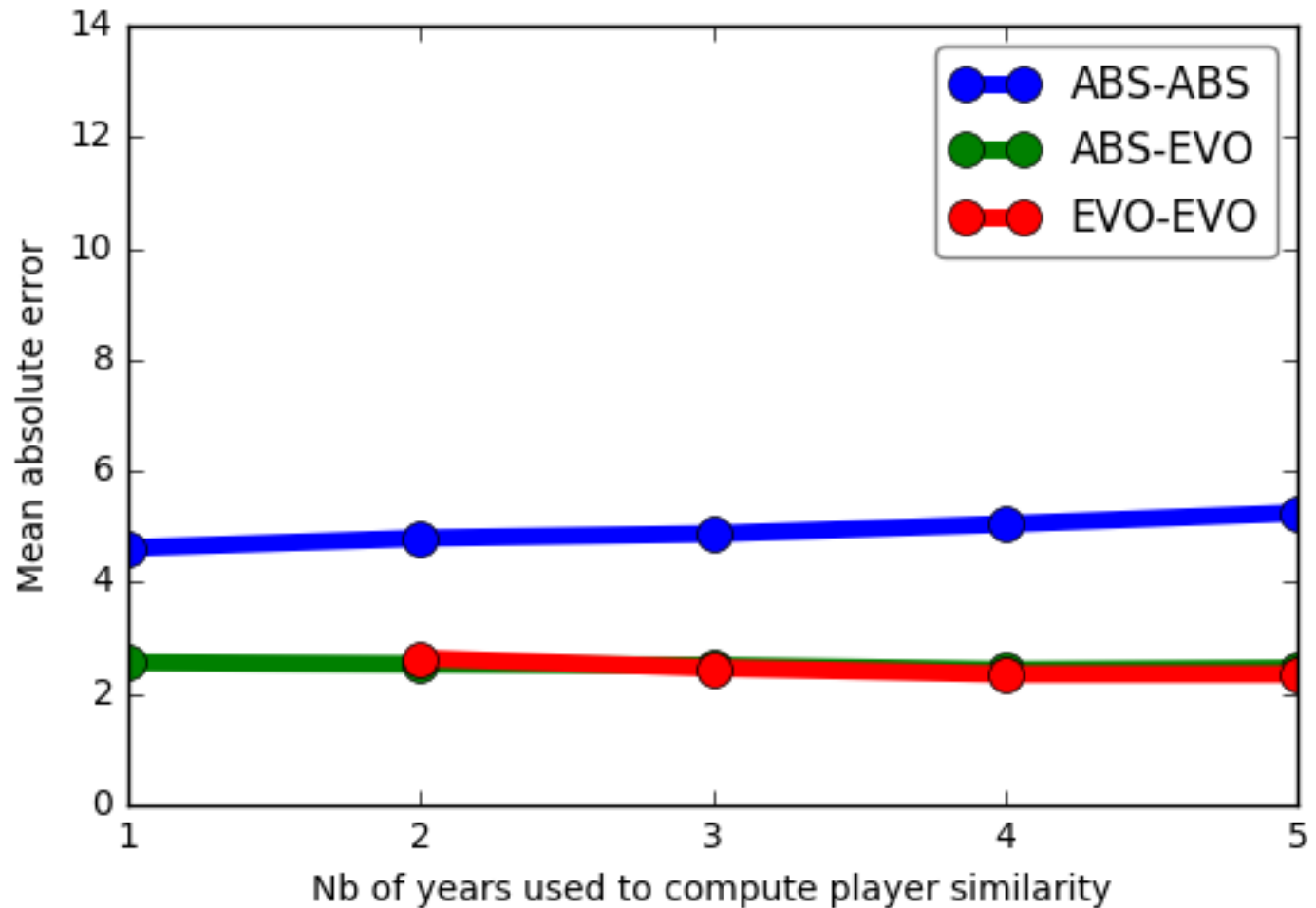
APROPOS performs better than baseline 1 and roughly equal to baseline 2.



# APROPOS beats Baseline 2 when predicting farther in the future



The nb of years used to compute player similarity has little effect on performance



# Conclusion

Predicting the potential of professional soccer players is an interesting task.

APROPOS solves this task using a nearest neighbors approach.

The best results are obtained by combining player-specific info with population-based info.