SOBIGDATA.it

ITALIAN RESEARCH INFRASTRUCTURE

AI in sports applications: when training data is actually training data

Pisa, 6-05-2025

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
SOBIGDATA.it ITALIAN RESEARCH INFRASTRUCTURE
Consiglio Nazionale delle Ricerche

Why sports? And why football, in particular?

- Sports are founded on data
- Data collection is challenging (and boring…)
- Football in particular is a complex game, with plenty of hidden features and a financial-like context

Main applications

Computer Vision

- Tracking players / data collection
- Biomechanics / pose estimation

Time Series Analysis

- Health monitoring
- Performance analysis
- Injury prevention

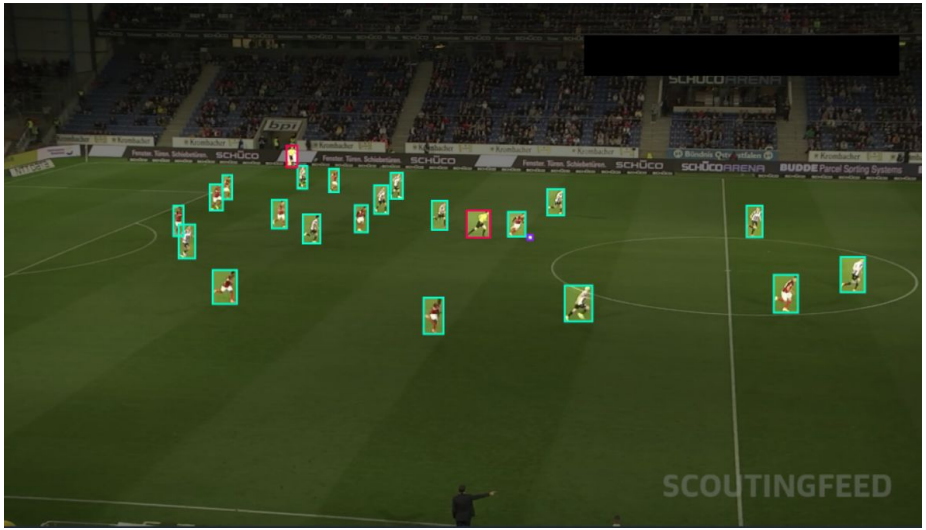
Machine learning

- Athletes profiling
- Ranking
- Game analysis

Computer Vision in Sports: Revolutionizing Data Collection

Key applications:

- Player tracking
- Motion analysis



Player Tracking & Pose Estimation

Techniques:

- Keypoint detection (OpenPose, MediaPipe)
- Optical flow for movement analysis

Use cases:

- Real-time player positioning (e.g., soccer, basketball)
- Speed and acceleration metrics

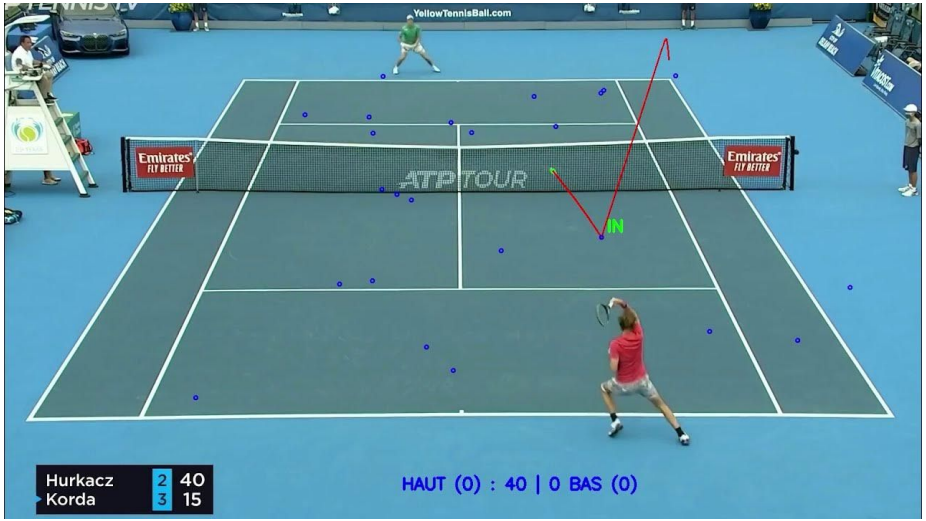
Ball & Object Tracking

Methods:

- YOLO (You Only Look Once) for fast ball detection
- Multi-camera triangulation for 3D tracking

Applications:

- Tennis serve speed analysis
- Baseball pitch trajectory tracking



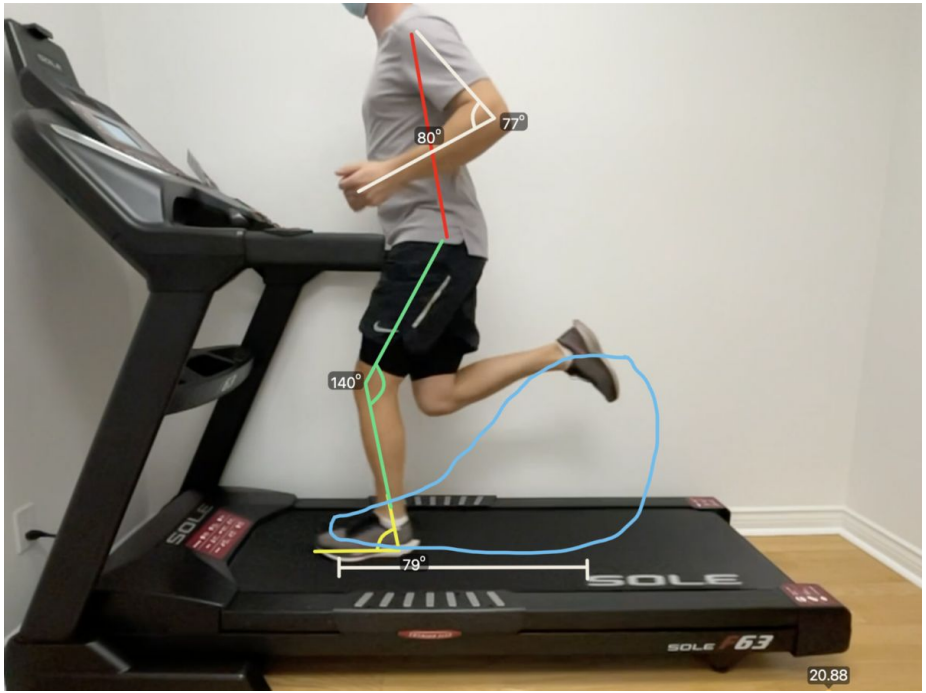
Motion & Biomechanics Analysis

Deep learning models analyze:

- Running gait (for injury prevention)
- Golf swing mechanics

Tools:

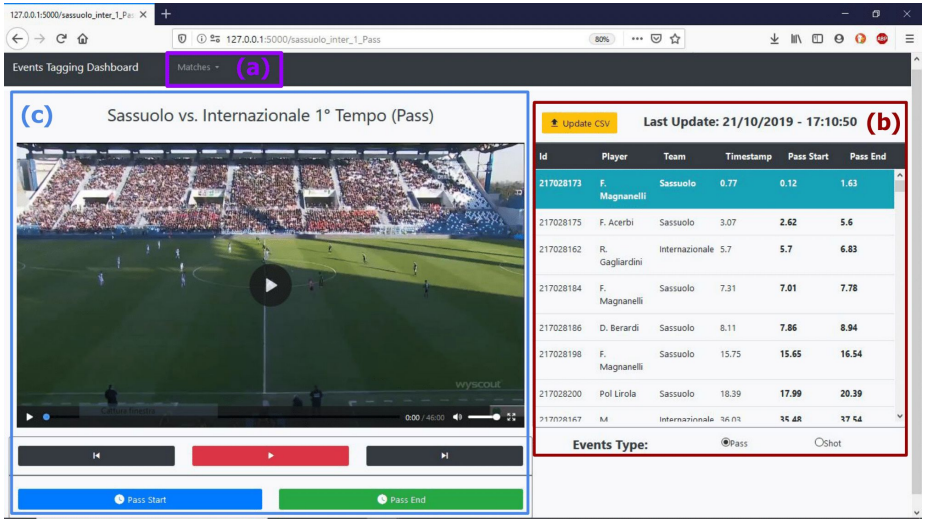
- OpenCV + TensorFlow for joint angle estimation



Automated Event Detection

AI classifies actions:

- Goals, fouls, tackles (using CNN + LSTM networks)
- Highlight generation for broadcasts



Reference:

- Automatic pass annotation from soccer video streams based on object detection and lstm

  D Sorano, et. al. - Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), 2021
  https://github.com/DaniloSorano/PassNet

Resources

Papers

- A survey on soccer player detection and tracking with videos https://doi.org/10.1007/s00371-024-03367-6
- SoccerNet-Tracking: Multiple Object Tracking Dataset and Benchmark in Soccer Videos [Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022]

Tracking tutorial:

- Roboflow - Football AI Tutorial: From Basics to Advanced Stats with Python https://www.youtube.com/watch?v=aBVGKoNZQUw

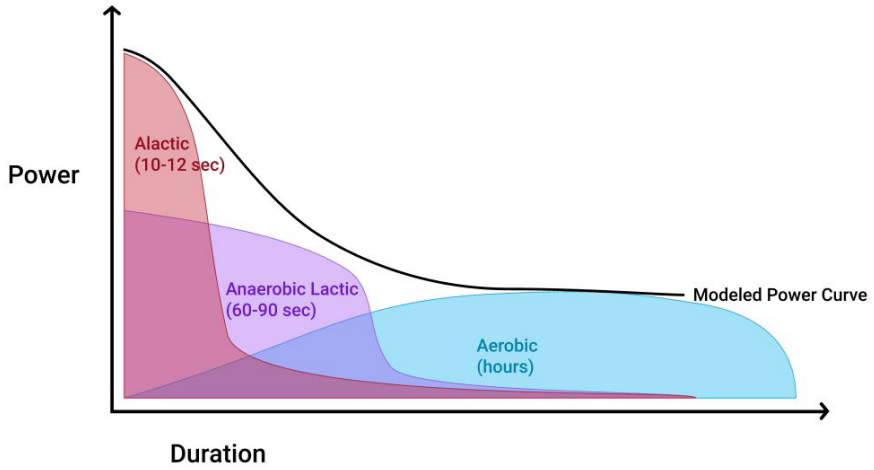
Time Series in Sports

Definition:

- Time series = Sequential data points collected over time (e.g., heart rate, speed, GPS coordinates).

- **Why it matters:**

- Tracks trends in performance.
- Identifies fatigue and/or injury risks.
- Joint time series to track many metrics over time

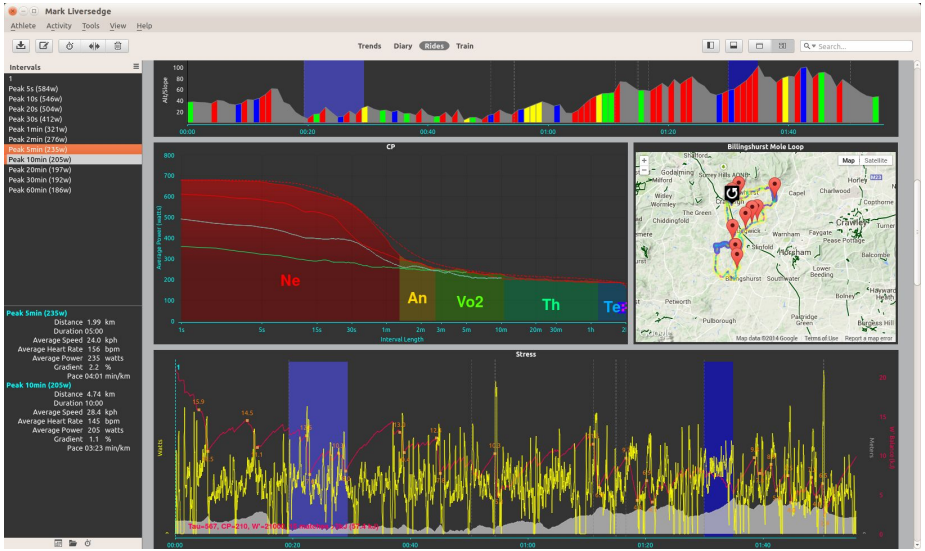


Key Metrics in Performance Monitoring

Examples of time-dependent data:

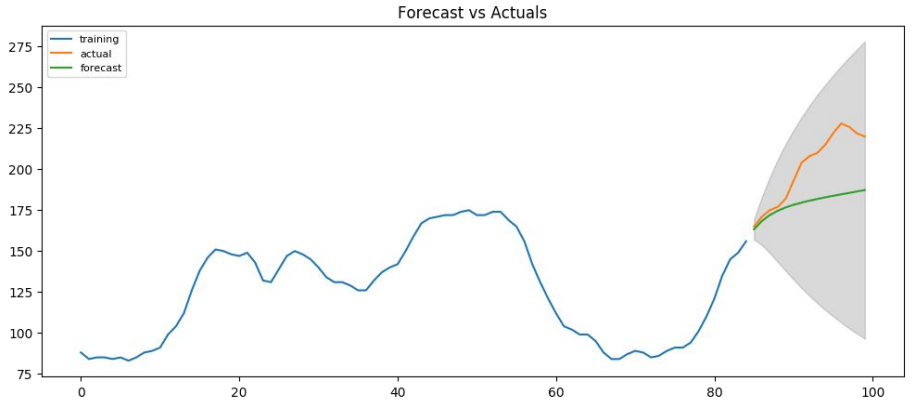
- Running speed/distance (GPS tracking).
- Jump height (in basketball/volleyball).
- Heart rate variability (HRV).

Tools:
Wearables (Catapult, WHOOP), optical tracking (Hawk-Eye).



Time Series Techniques for Performance Analysis

- Moving averages (smoothing noisy data).
- Fourier transforms (detecting periodicity, e.g., fatigue cycles).
- Models such as ARIMA,Recurrent Neural Network, N-BEATS, etc (predicting future performance trends).



Wearables & IoT Data Collection

Devices:

- **GPS vests, smart insoles, EMG sensors.**

Time series challenges:

- **Noise filtering.**
- **Sensor fusion (combining GPS + accelerometer data).**

Injury Risk Prediction

Just two stats for a single season of a single professional competition:

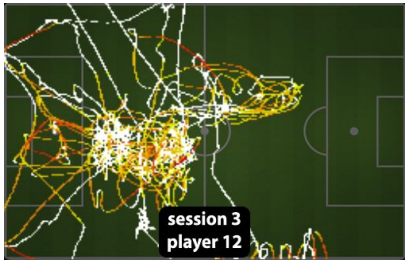
- 16.23% of season absence
- Total cost for teams: 188,058,072 €

The first ML approach to assess injury risk

Dataset

- 26 players
- 6 central backs
- 4 full backs
- 7 middlefields
- 8 wingers
- 2 strikers

- 23 weeks
- GPS portable (STATSports Viper)



Effective injury prediction in professional soccer with GPS data and machine learning

Alessio Rossi · Luca Pappalardo · Paolo Cintia · F. Marcello Iaia · Javier Fernández · Daniel Medina

Features

Training features (GPS)

Total Distance
High Speed Running (>19.8 km/h)
Metabolic Distance (>20W/kg)
High Metabolic Load Distance (>25.5 W/Kg)
High Metabolic Load Distance Per Minute
Explosive Distance (>25 W/kg <19.8 Km/h)
Accelerations >2m/s2
Accelerations >3m/s2
Decelerations >2m/s2
Decelerations >3m/s2
Dynamic Stress Load (>2g) Fatigue Index
(Dynamic Stress Load/Speed Intensity)

Players' features

Age
Height
Weight
Role
Previous injuries

State of the art – ACWR

Very low	Low	Moderate	High	Very High
<0.49	*0.50-0.99*	*1.00-1.49*	*1.50-1.99*	*>2.00*

Monodimensional method relying on this formula:

$$ACWR = \frac{\text{acute workload (7 days)}}{\text{chronic workload (28 days)}}$$

What if we build a predictor through ACWR?

		ACWR			
	class	prec	rec	F1	AUC
$C_{d_{TOT}}$	0	0.99	0.44	0.61	0.65
	1	0.03	0.86	0.06	
$C_{d_{HSR}}$	0	0.99	0.37	0.54	0.57
	1	0.03	0.76	0.05	
$C_{d_{MET}}$	0	0.99	0.43	0.60	0.59
	1	0.03	0.76	0.06	
$C_{d_{HML}}$	0	0.99	0.43	0.60	0.60
	1	0.03	0.76	0.06	
$C_{d_{HML/m}}$	0	0.99	0.39	0.56	0.60
	1	0.03	0.81	0.06	
$C_{d_{EXP}}$	0	1.00	0.43	0.60	0.67
	1	0.04	0.91	0.07	
C_{Acc_2}	0	0.99	0.47	0.64	0.64
	1	0.03	0.80	0.06	
C_{Acc_3}	0	0.99	0.45	0.64	0.58
	1	0.03	0.71	0.06	
C_{Dec_2}	0	0.99	0.46	0.63	0.66
	1	0.04	0.86	0.07	
C_{Dec_3}	0	0.99	0.46	0.63	0.66
	1	0.04	0.86	0.07	
C_{DSL}	0	1.00	0.42	0.60	0.66
	1	0.03	0.90	0.07	
C_{FI}	0	0.98	0.47	0.64	0.55
	1	0.03	0.62	0.05	

B_1	0	0.98	0.98	0.98	0.51
	1	0.06	0.05	0.05	
B_2	0	0.98	1.00	0.99	0.50
	1	0.00	0.00	0.00	
B_3	0	0.00	0.00	0.00	0.50
	1	0.02	1.00	0.04	
B_4	0	0.98	0.77	0.86	0.60
	1	0.04	0.43	0.07	

high recall > 90%

low precision < 6%

Switching to a multi-dimensional approach

	d_{TOT}	d_{EXP}	...	ACC_3	label
s_1	4,018.19	426.42	...	16.99	**0**
s_2	3,465.81	326.41	...	16.91	**0**
s_3	3,227.15	256.85	...	18.25	**1**
	⋮	⋮	⋮	⋮	⋮
s_n	3,199.58	273.69	...	19.64	**1**

Feature set:

12 Daily

12 Acute

12 ACWR

12 MSWR

7 Contextual

Total = 55 features

Label = {0:No-injury; 1:Injury}

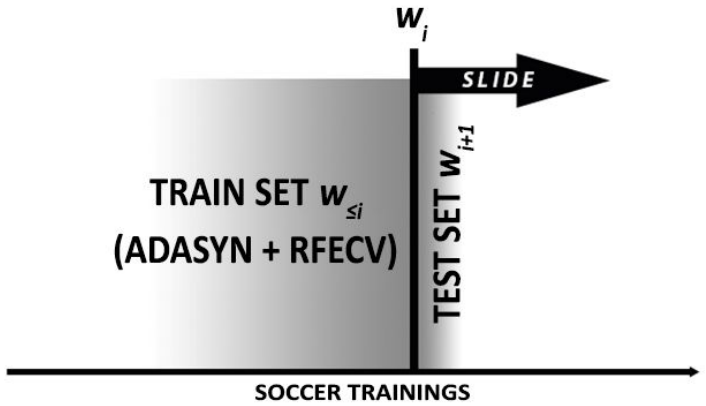
Adding time into context

TRAIN SET:

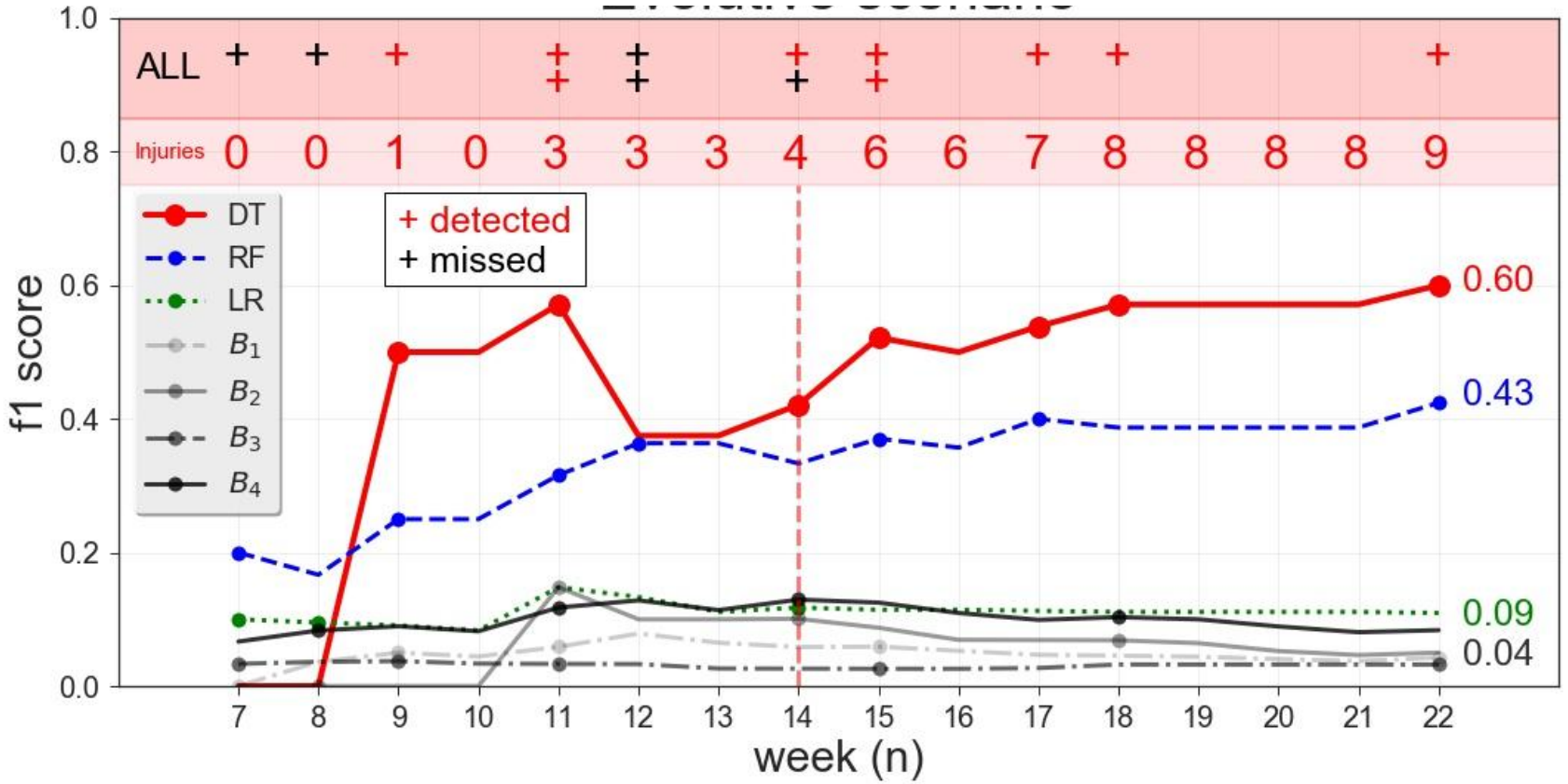
- Prediction start at w6 due low injury examples in first part of the season. The dimension of the test set increase as the season go by.
- ADASYN, RFECV and model fitting.

TEST SET:

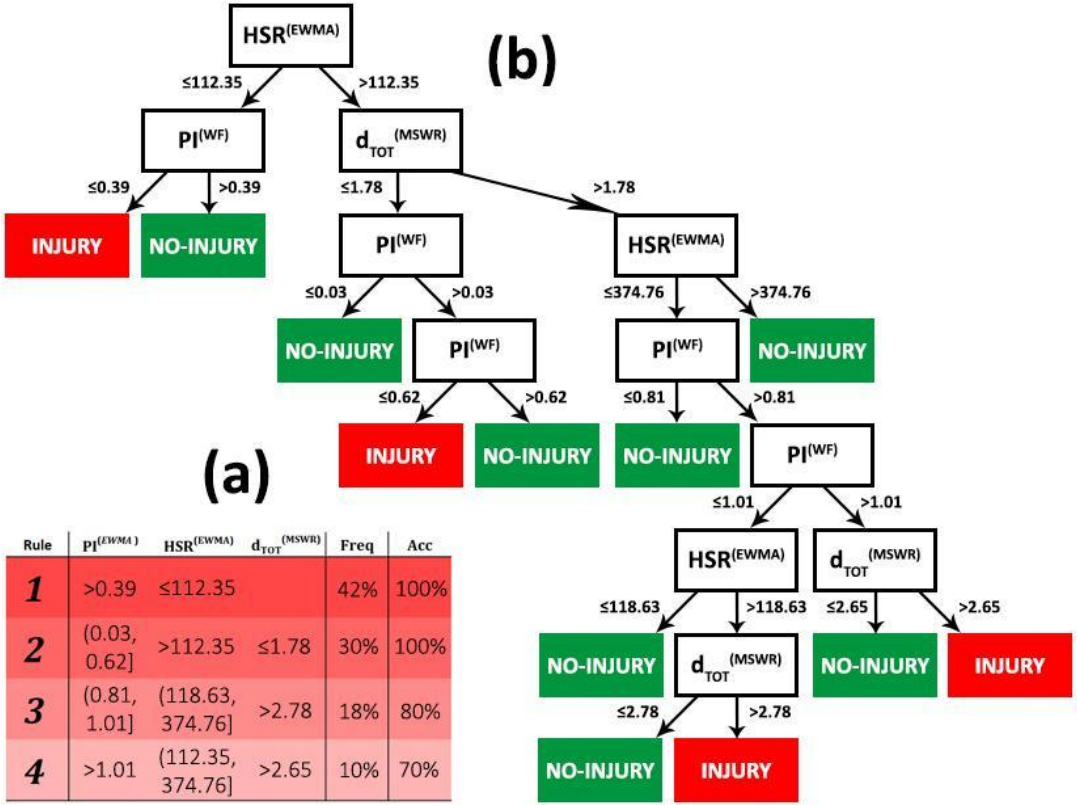
- Algorithms (i.e., DT, LR, RF) test
- Modelle assessment: Precision, Recall, F1



Digging into the results



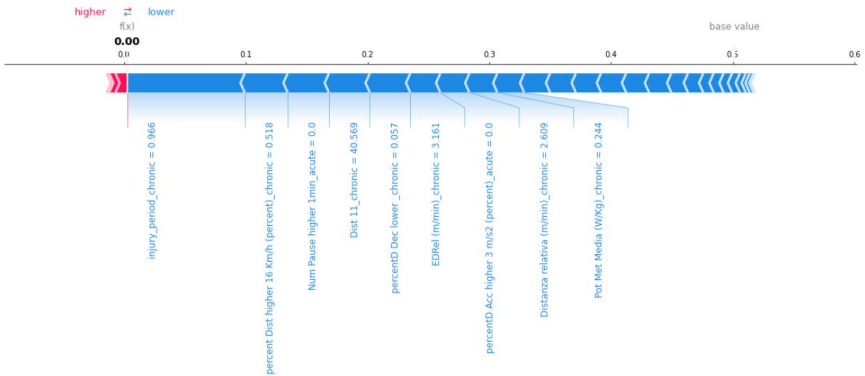
A plus when using decision trees: interpretation

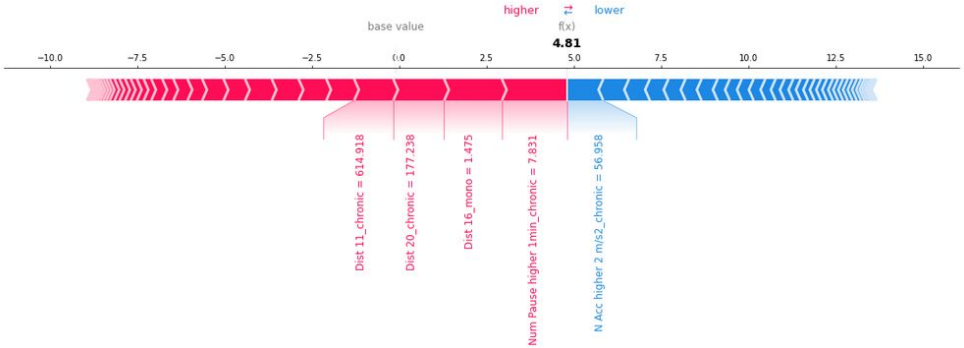


(a)

Rule	$PI^{(EWMA)}$	$HSR^{(EWMA)}$	$d_{TOT}^{(MSWR)}$	Freq	Acc
1	>0.39	≤112.35		42%	100%
2	(0.03, 0.62]	>112.35	≤1.78	30%	100%
3	(0.81, 1.01]	(118.63, 374.76]	>2.78	18%	80%
4	>1.01	(112.35, 374.76]	>2.65	10%	70%

Injury probability as load indicator





Injury probability < 50%

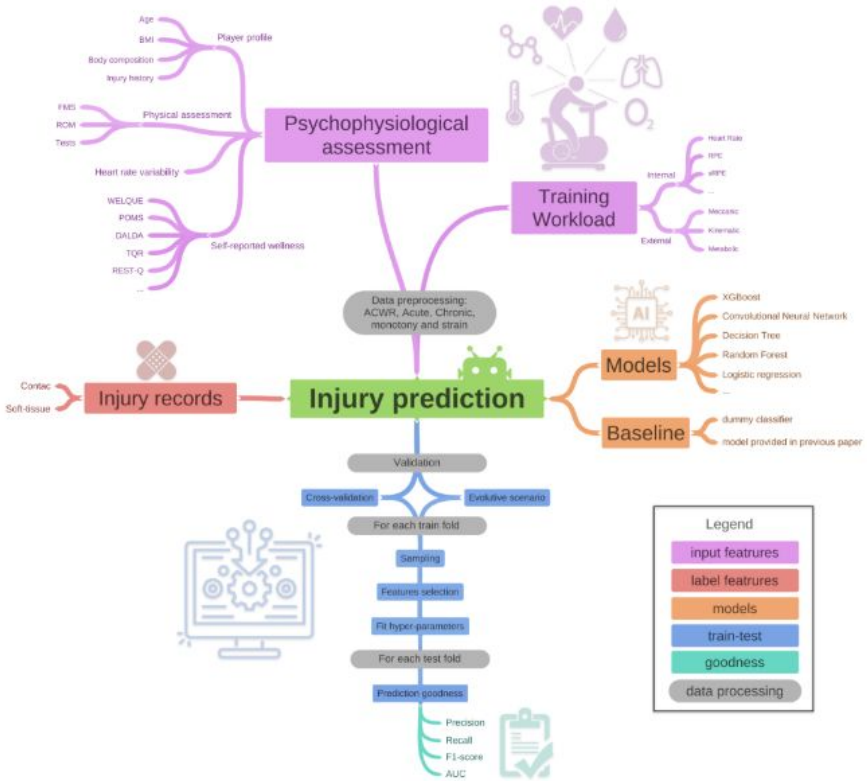
The player still got injured over the next 3 days

Injury probability > 60%

The player got injured over the next 3 days

Overview from the experiment

- From 6% to 94% precision
- Interpretable rules for coaches
- 14 weeks needed for training
- > 60% injuries detected



Machine learning use cases in football

Machine learning and data science in football (“technical” club staff):

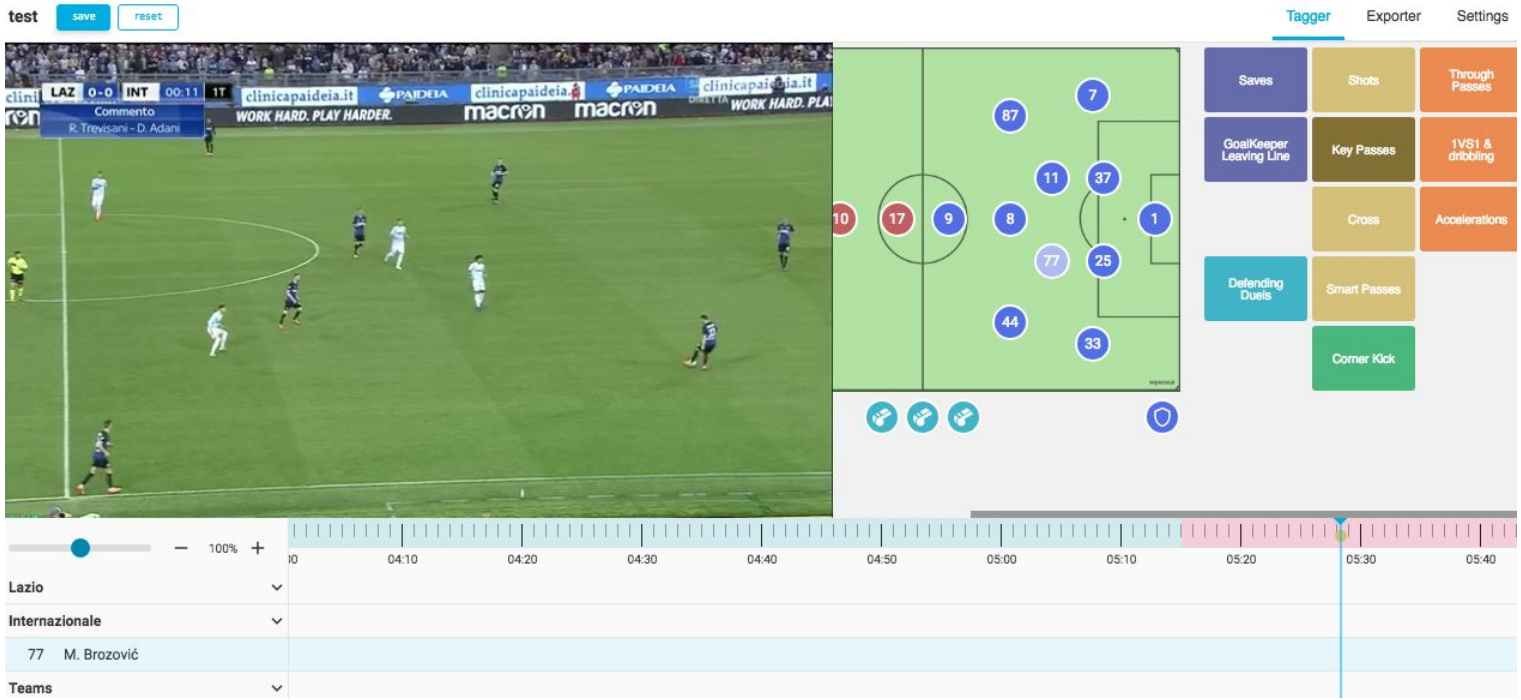
- Player scouting
- Match analysis
- Game style detection
- Tactical metrics development

A little deviation from a science perspective

- Football has been studied also by anthropologists, sociologists
- Many great minds from the past were passionate for football
- Still hard to explain why football and not any other sport

The unpredictable, in football, is still higher than the predictable. Even after AI taking over.

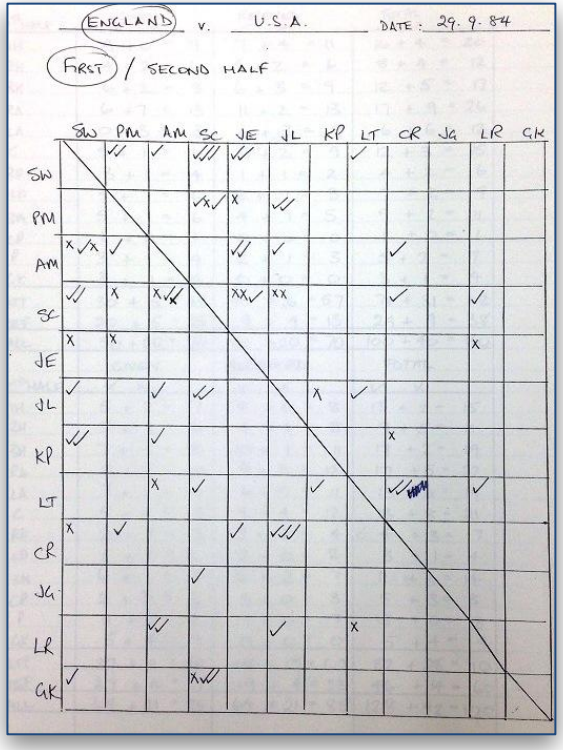
Match analysis: understanding the game

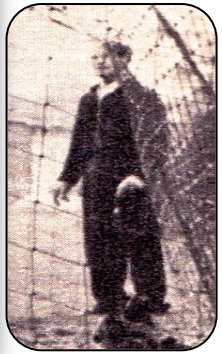


The revolution started here (2010s, video and related data started to become available in a structured way)

The first football data analyst

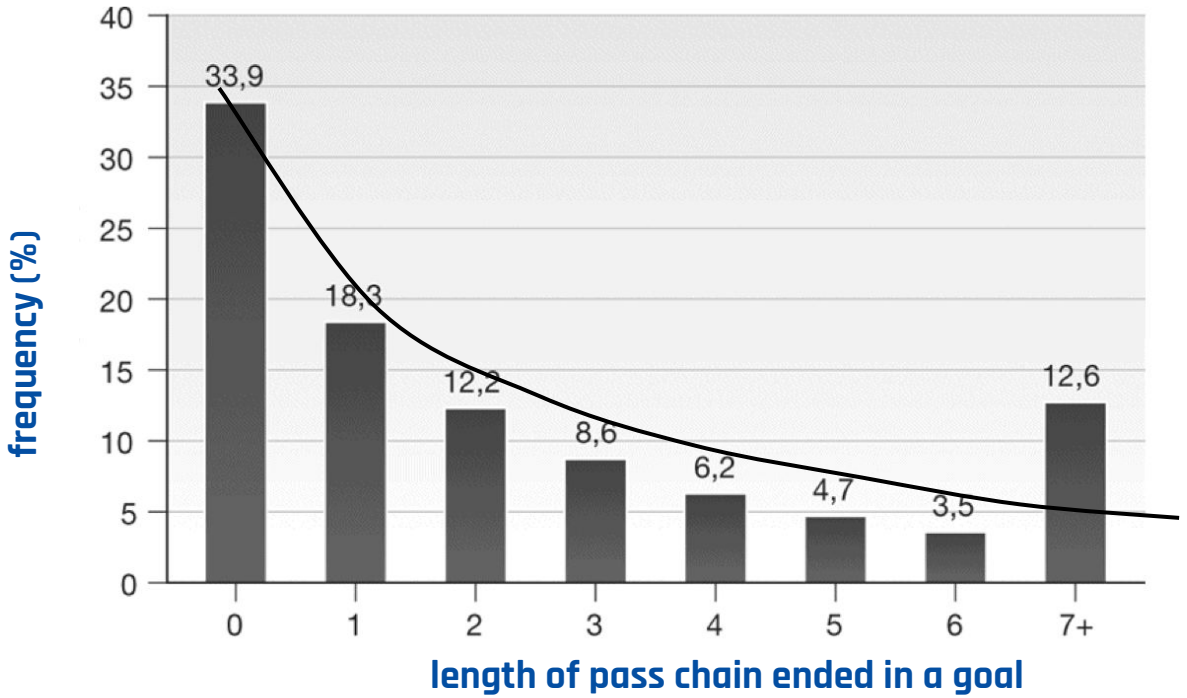
R. Pollard, Charles Reep (1904-2002):
pioneer of notational and performance analysis in football, Journal of Sports Sciences 20(10):853-855, 2002.





Not a good start

- England federation started relying on the long ball theory from Reep's studies
- The theory was not properly backtested



Not a good start

FiveThirtyEight

OCT. 27, 2016, AT 9:09 AM

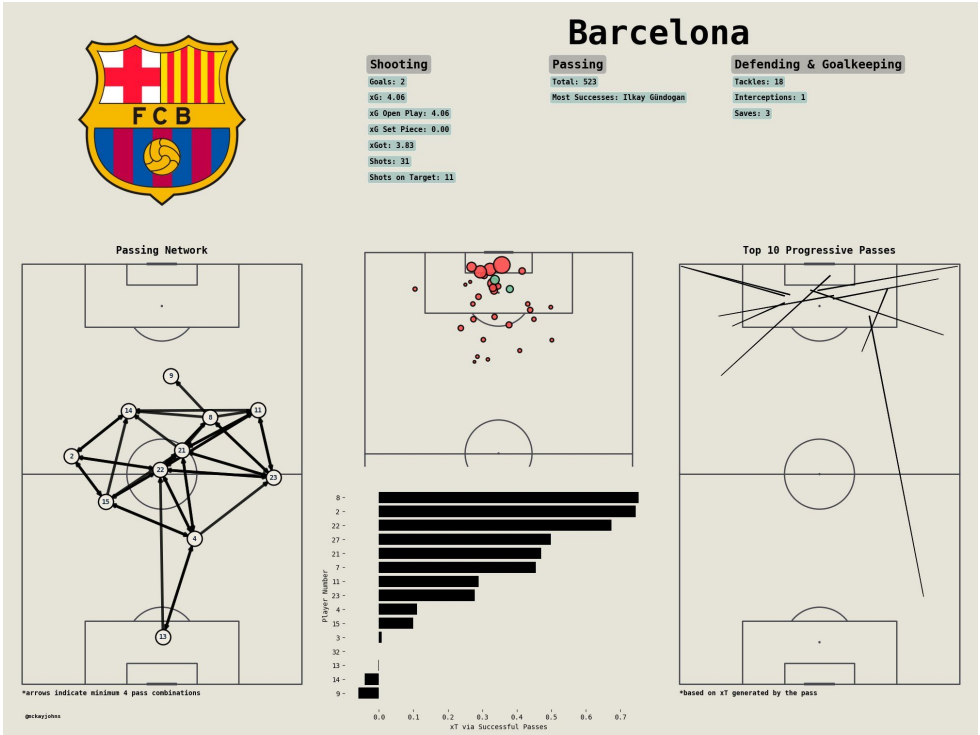
How One Man's Bad Math Helped Ruin Decades Of English Soccer

By Joe Sykes and Neil Paine

Data driven match report

- The introduction of data science in football has been really hyped
- Several metrics and approaches has not been properly validated, just dumped over coaches

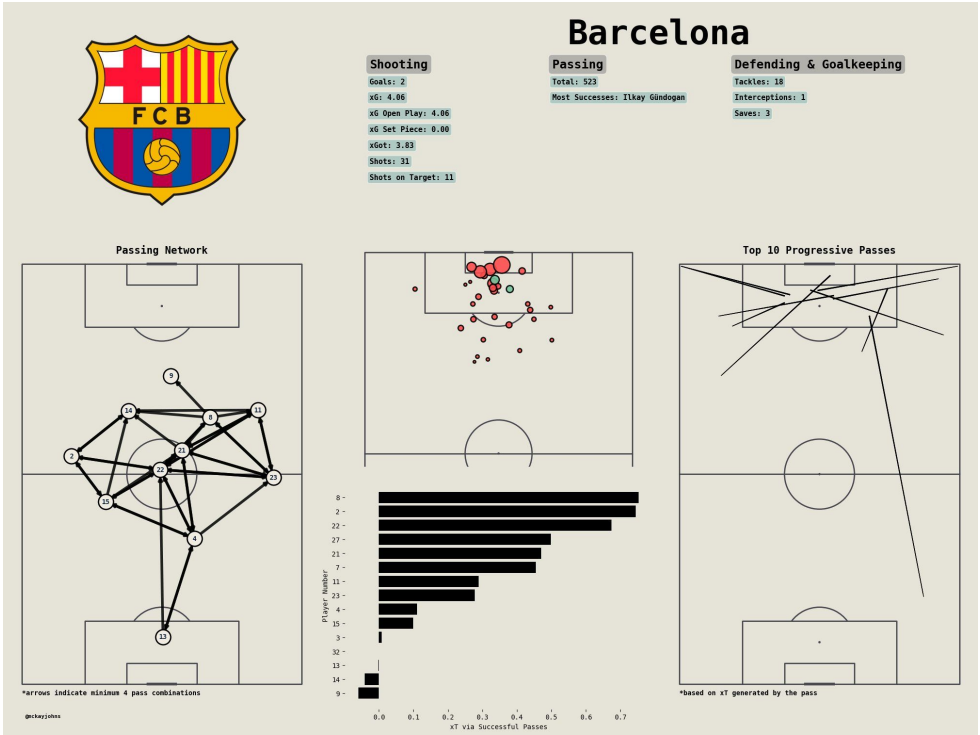
The trust of coaches in data was (is?) at the bare minimum



Data driven match report

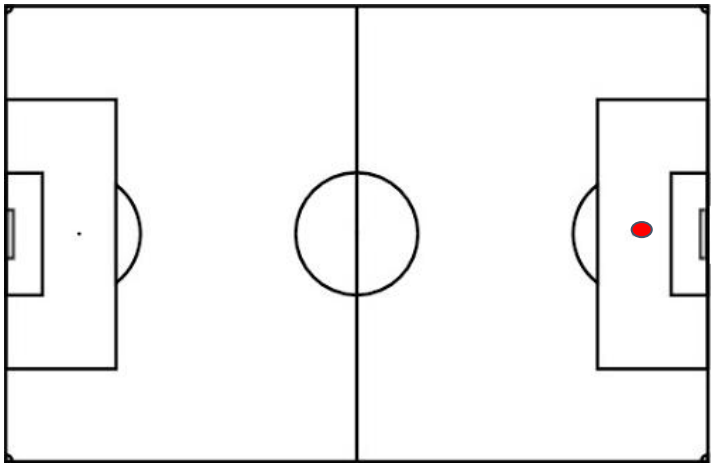
drawbacks:

- passing networks provide a wrong overview of ball possession
- summing xg of all the shots is a BIG statistics red flag
- xT is the only novelty from last years of research

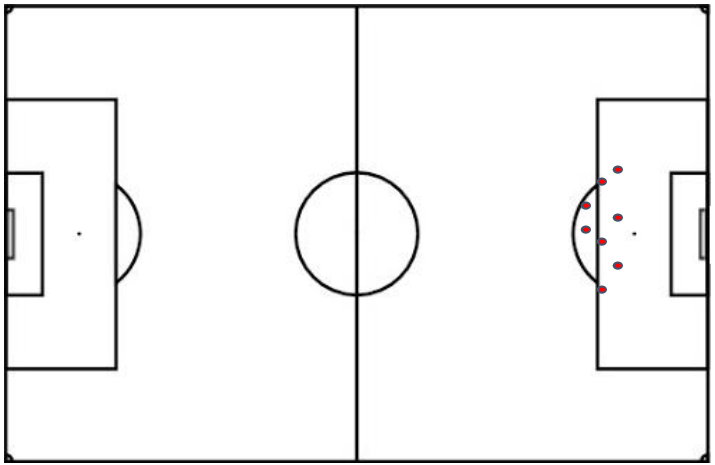


Is this a draw? (hint: probably not)

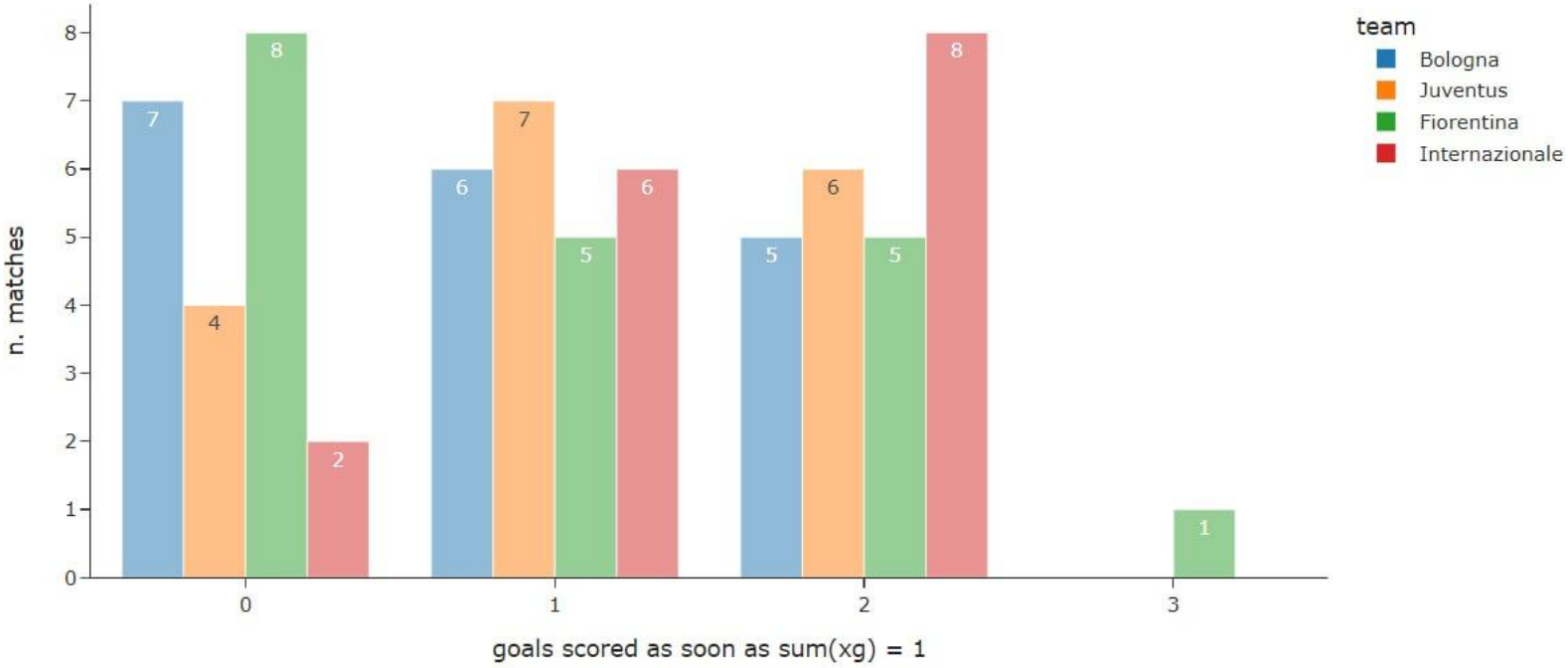
Xg = 0.8



Xg = 0.8

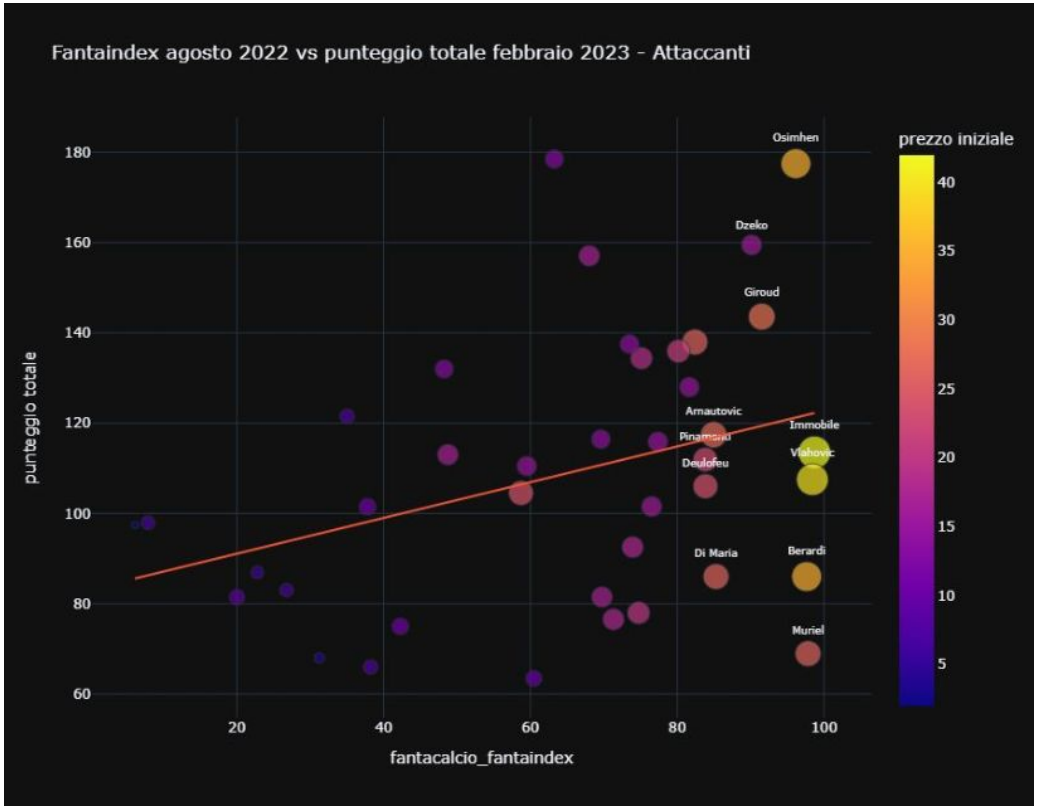


1 xg = 1 goal?


team
Bologna
Juventus
Fiorentina
Internazionale
n. matches
0
1
2
3
4
5
6
7
8
goals scored as soon as sum(xg) = 1
7
4
8
2
6
7
5
6
5
6
5
8
1

could you win fantasy football with data?

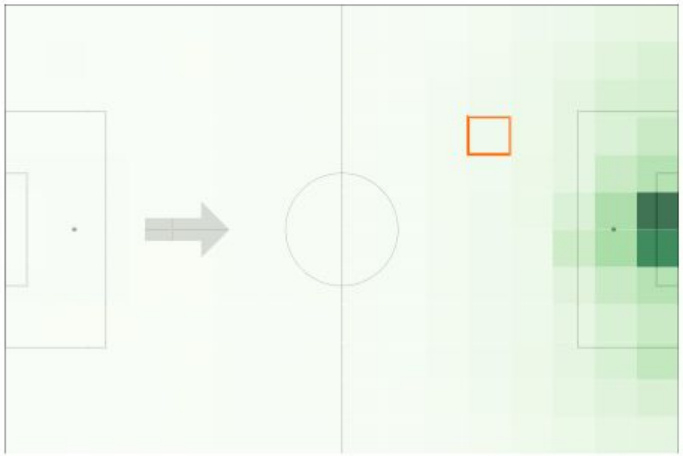
- Fantaindex was an AI metric sponsored by many media channels
- Claim: use this index to find the best players for next season
- Validation: not provided
- At the end of the season users have discovered that it did not work

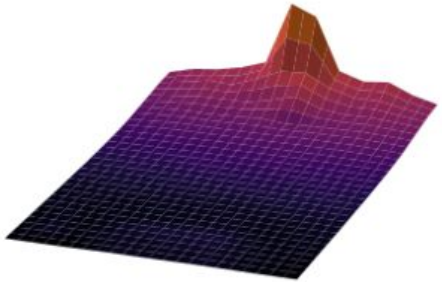


Getting back to real metrics: Expected Threat

Expected Threat (xT) = 0.034

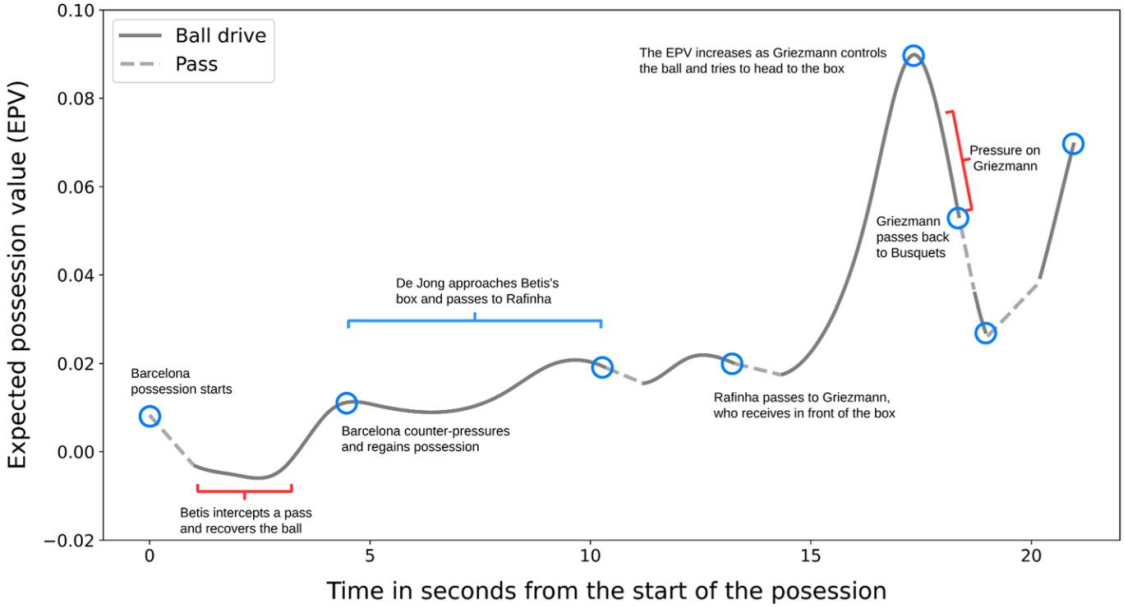
i.e. when the team has the ball in the highlighted zone, they will score in the next **5** actions **3.4%** of the time.





- **Scoring probabilities according to the stream of events on the ball**
- **Used to evaluate the quality of on-ball events performed by the players**

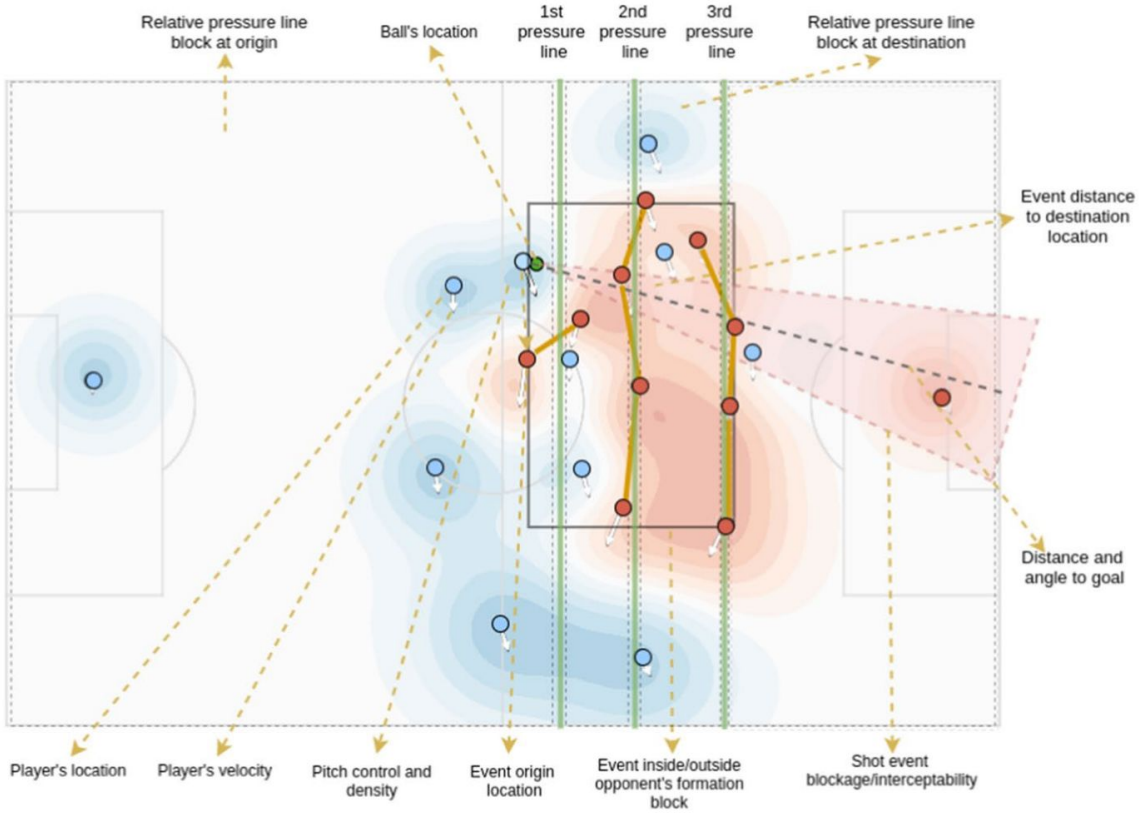
Expected Possession Value



Refinement of xt, relying on tracking data, i.e. the positions of every player at every instant of the game

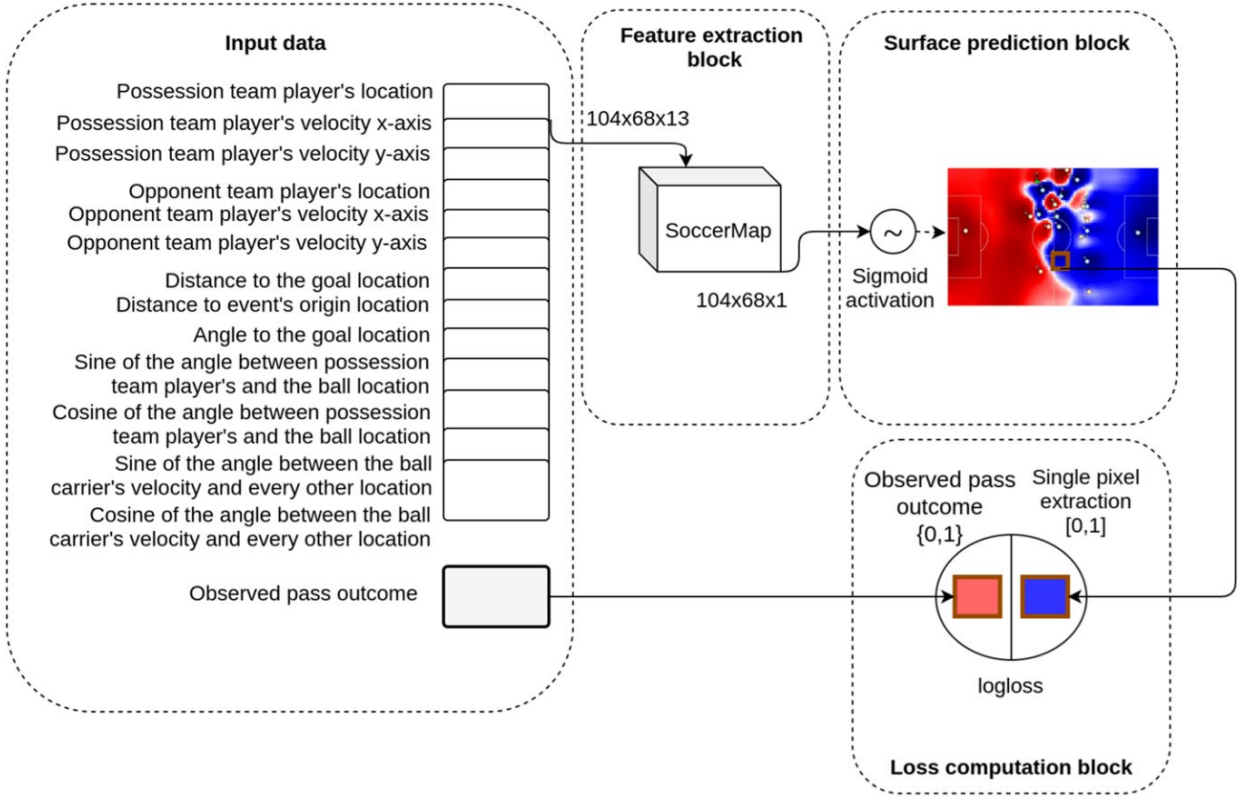
Expected possession value

- Spatial features
- Context features
- Value of the controlled space according to the “threat” you can pose to the oppponents



Expected possession value

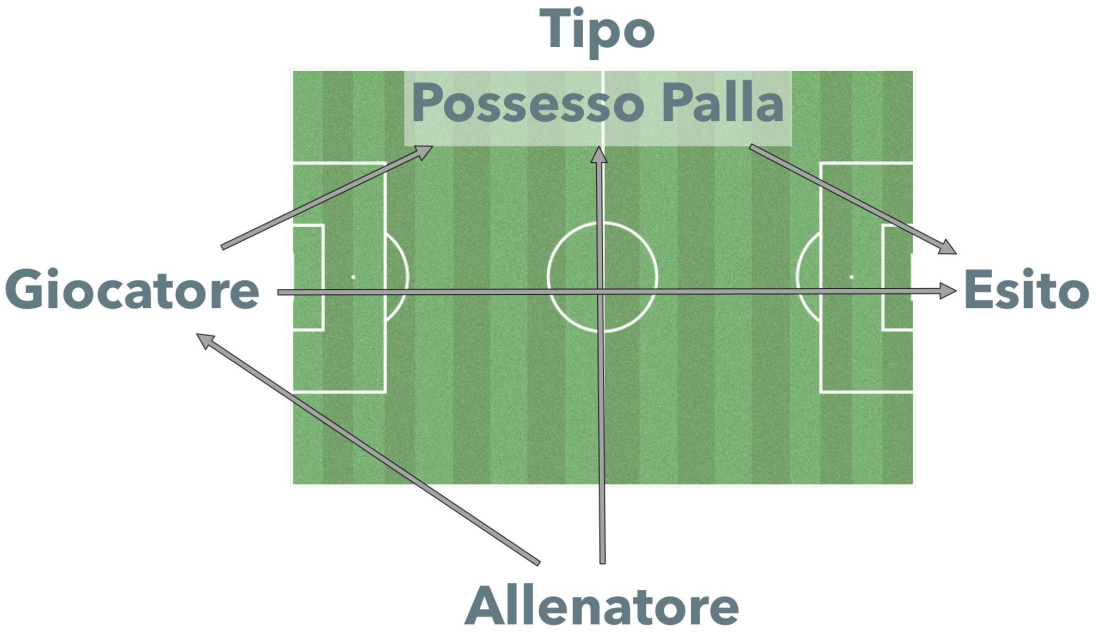
- **Task: predict who "owns" every pixel of the pitch**



Match analysis for scouting

- Adaptive scouting is a data analysis algorithm capable of simulating a player's adaptability to a playing style

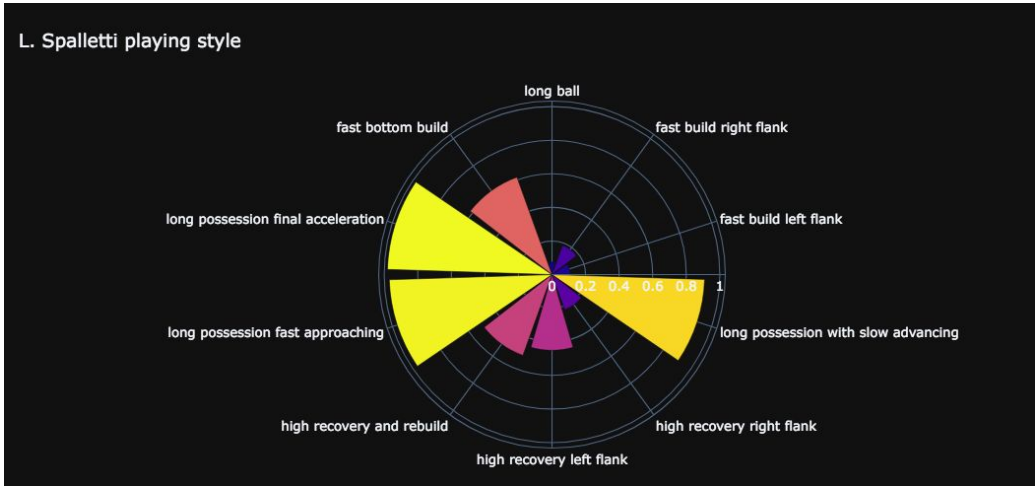
- The starting point is the collection of individual events recorded in each match (approximately 2000 per game). From these, the playing style of each coach/team is modeled, along with the effectiveness of each player based on the adopted style, and their potential adaptability to a different style.



Match analysis for scouting

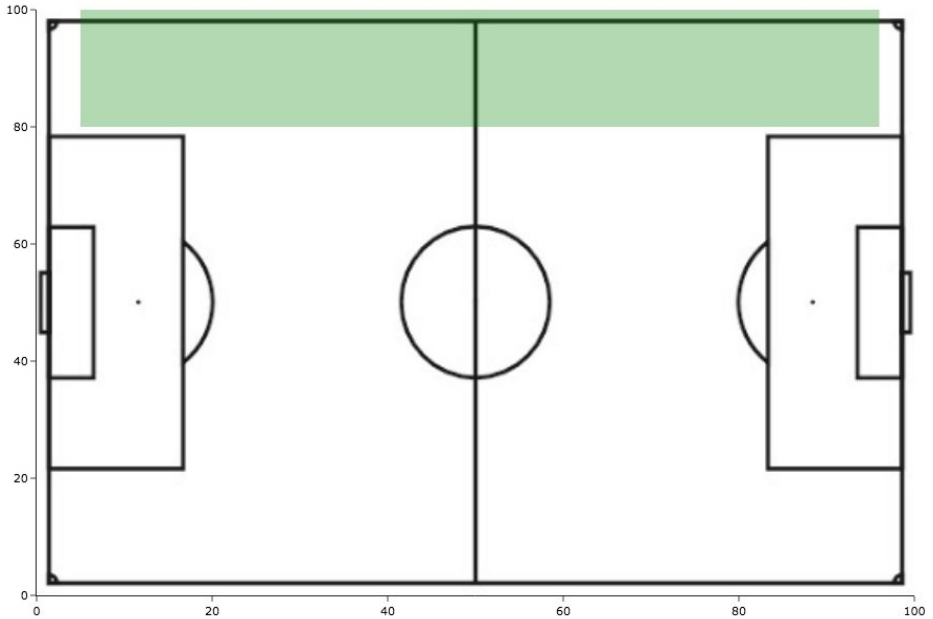
- First input: playing style. The analysis is performed by considering all ball possession sequences and their related characteristics (duration, speed of ball progression, starting point of possession, etc.)

- The graph shows Napoli's usage of different playing styles under Spalletti. A value above 0.5 for a given style indicates usage higher than the average among coaches in Europe's top leagues



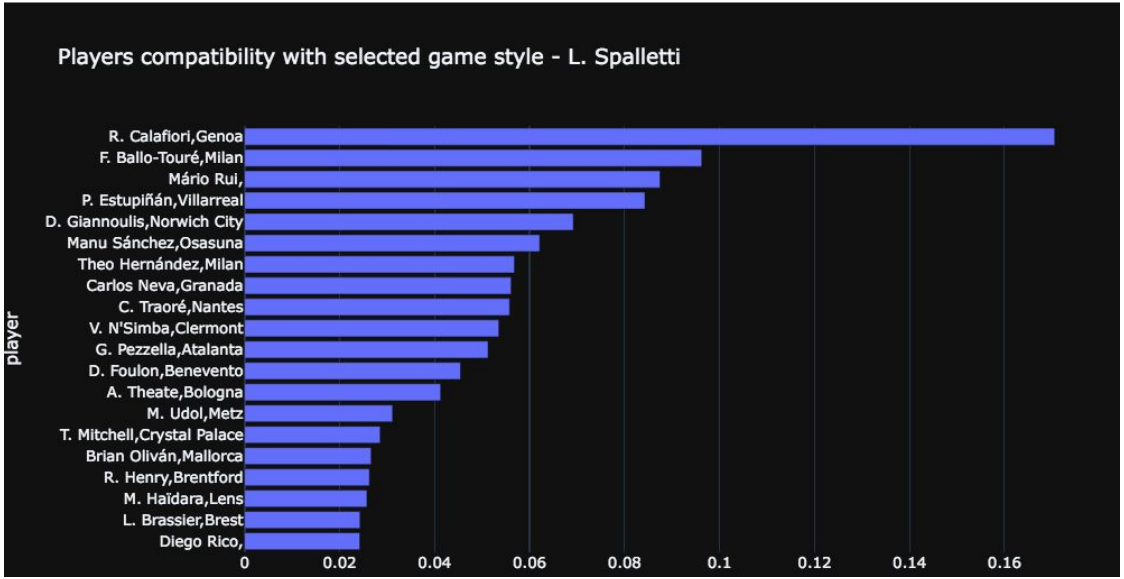
Overview from the experiment

Second input: a reference zone. The system identifies players who play the ball in the specified zone more than others**.**



Overview from the experiment

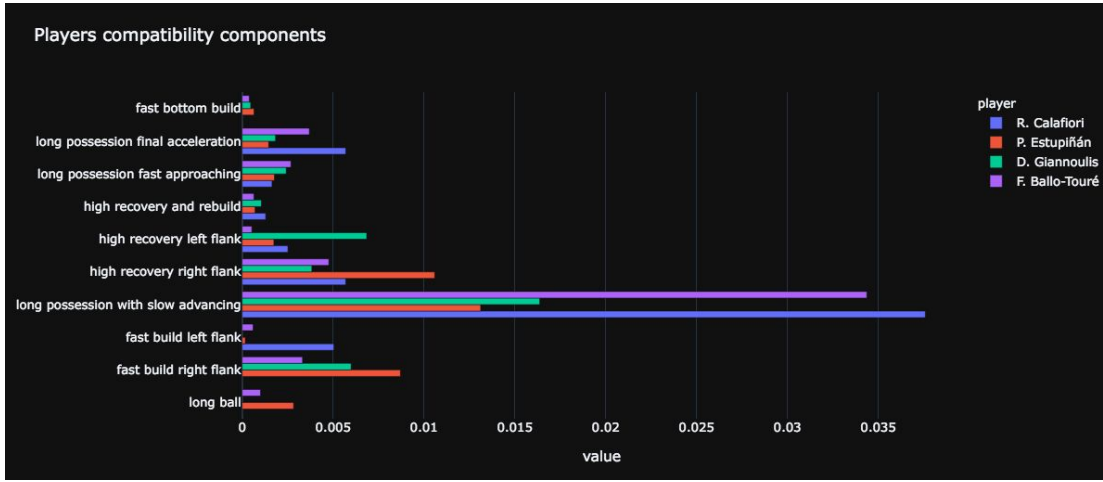
Using only the spatial filter, we selected players who primarily play the ball in the chosen zone. We then ranked them based on their 'compatibility' with Napoli's style of play



Match analysis for scouting

- The selected players show a higher compatibility with Spalletti's playing style. In the attacking moves they're involved in—particularly those that lead to goal-scoring opportunities—they demonstrate greater effectiveness based on the selected playing style

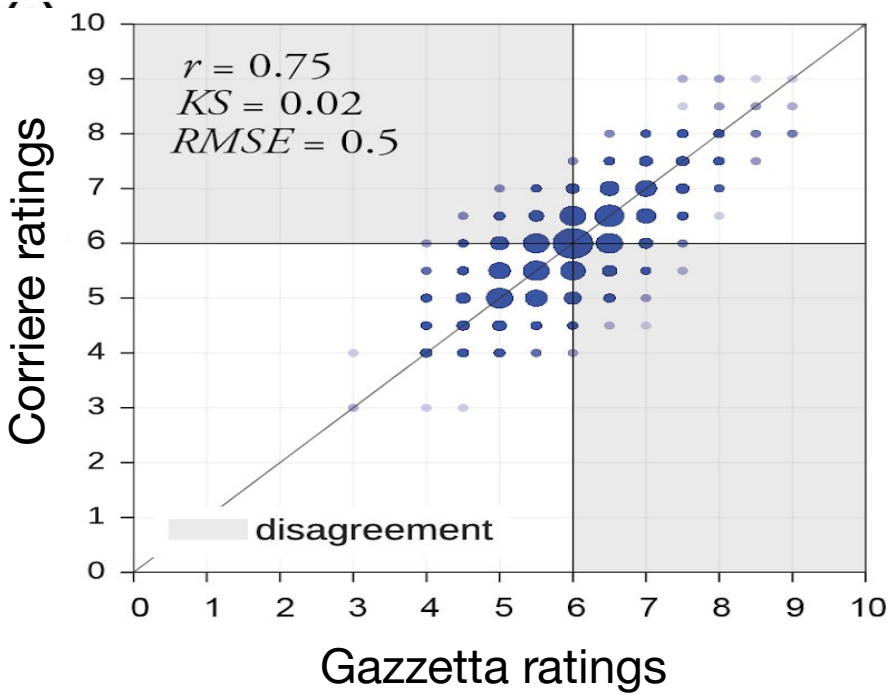
- We can therefore analyze the player's effectiveness within the reference coach's playing style for each type of ball possession



Player scouting

The AI journalist

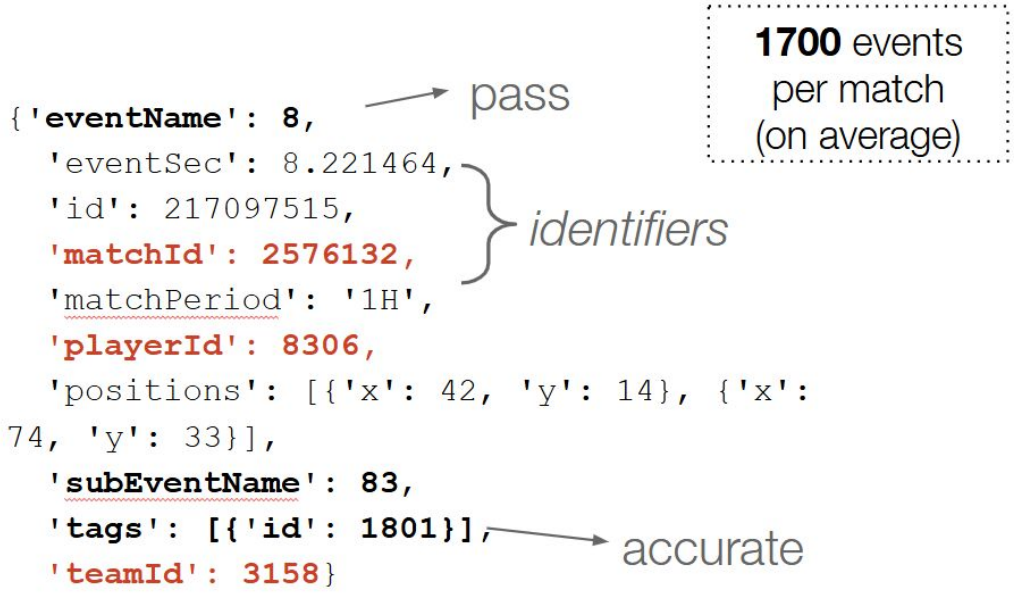
- Human judges are a bit noisy
- Using data to replicate their judgment process could be not worth it
- They use context that is not dependent from the sole player, not useful for real scouting



Playerank: evaluating performance and ranking players

Dataset:

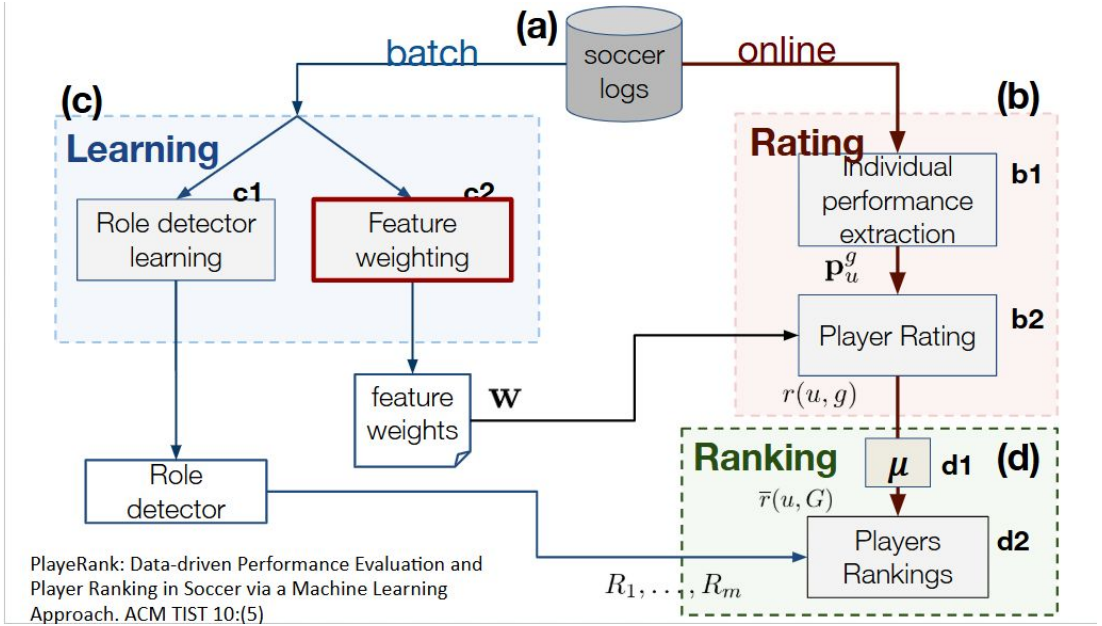
- 5 seasons
- 18 competitions
- 30M events
- 20K matches
- 21K players



Playerank: evaluating performance and ranking players

The challenge: evaluate the individual performance in a team performance

The approach: find the relationship between performance data and game result, then associate the value of the performance metric according to player contribution on achieving such performance metric value



Dataset creation

Team performance vector and corresponding game result

team1

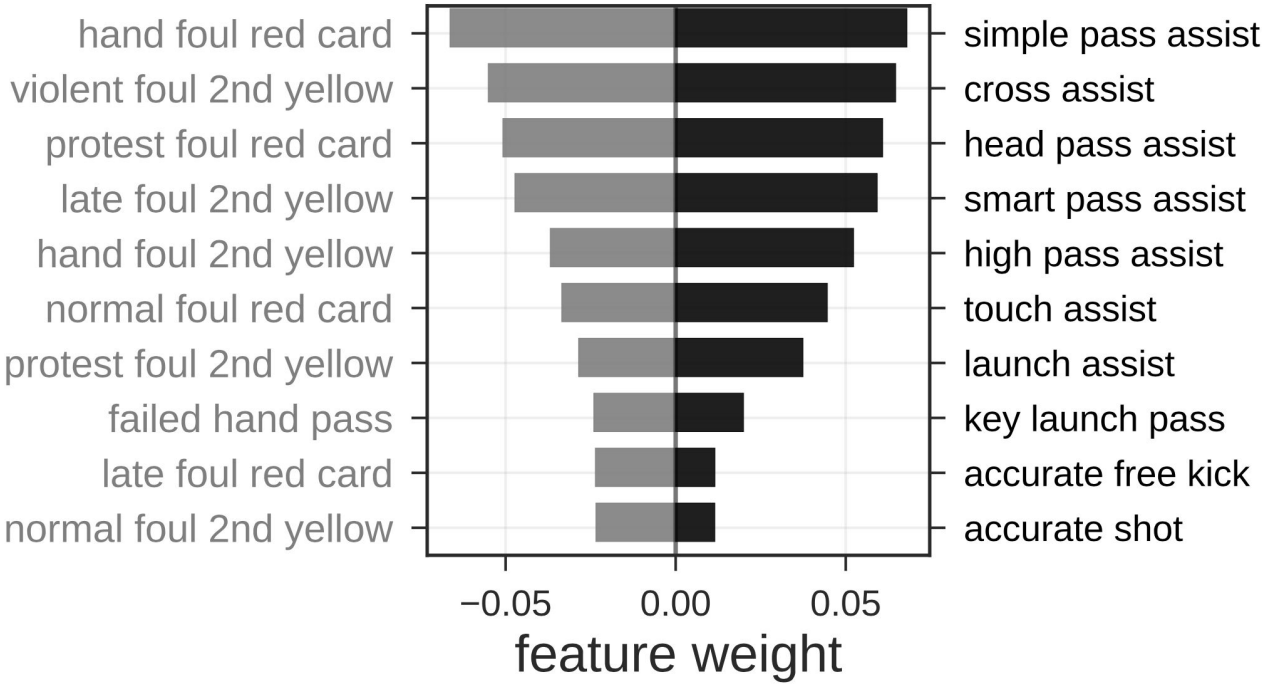
passes	duels	shots	discipline	...

team2

passes	duels	shots	discipline	...

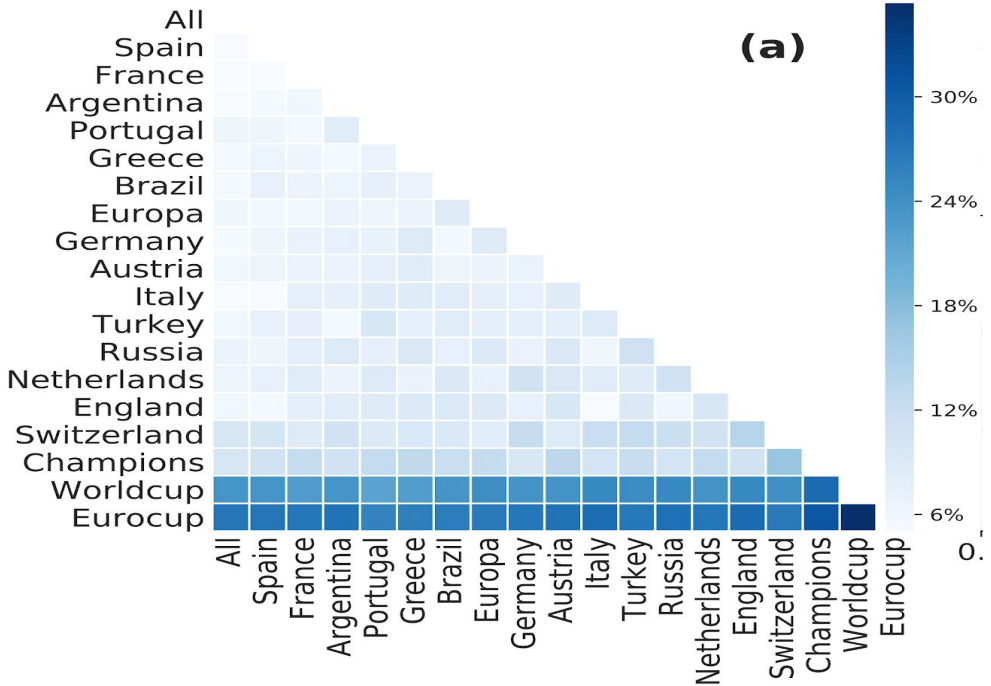
Features weight

The process: Extracting features weight from an SVM model created on the team performance dataset



A first validation

- Are those weights depending from the competition used to create the dataset?
- Are they dependent from time/season chosen?



Ranking the GOATs era

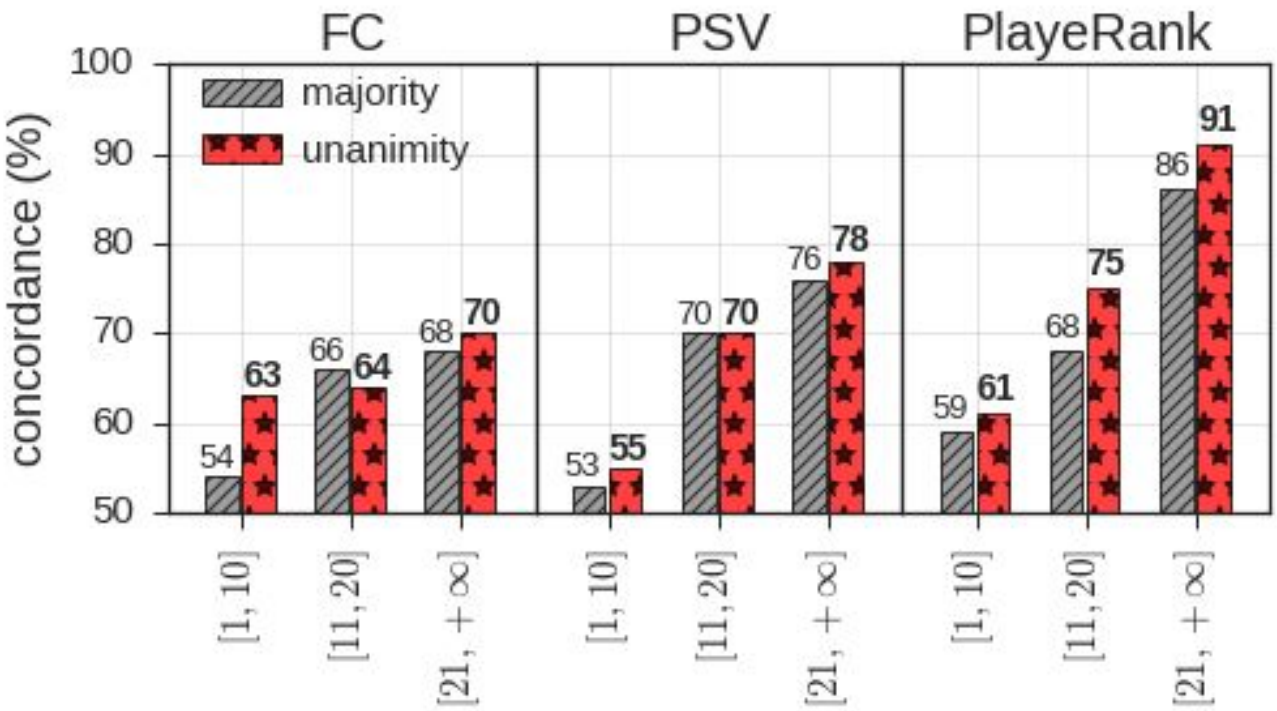
When ground truth is missing

algorithm	expert 1	expert 2	expert 3
			
			

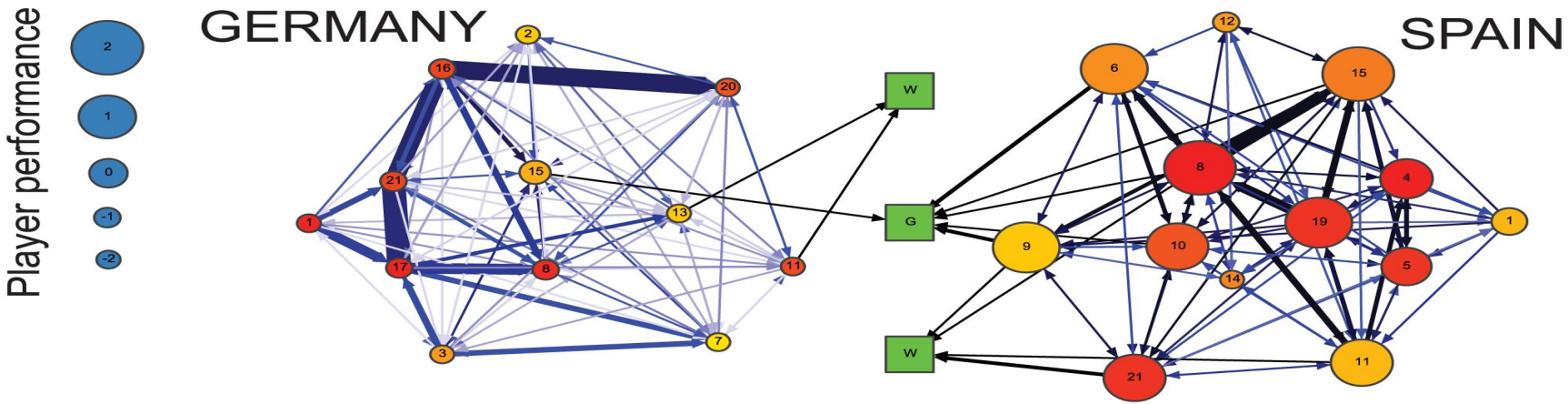
Football coaches asked to answer a survey
Results of the survey compared with playerank-based ranking

When ground truth is missing

The result: the longer the distance among two players across playerank-based ranking, the higher the agreement among football coaches



Flow centrality: network metrics for player ranking



Player contribution is measured according to his centrality in team's passes network

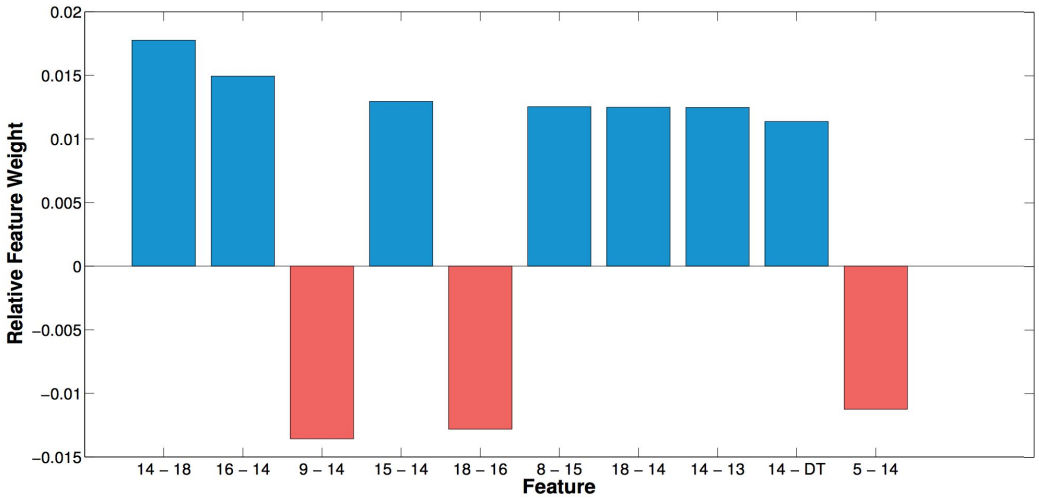
Validation: 8 of the 20 players in the list of the competition's best players (Fifa world cup 2010)
Duch et al. (2010) Quantifying the Performance of Individual Players in a Team Activity. PLoS ONE 5(6): e10937.

Pass Shot Value (PSV)

Evaluates the quality of a player's passes based on their likelihood of leading to a high-value shot.

Validation: correlation with assists and goals

Brooks et al. (2016) Developing a Data-Driven Player Ranking in Soccer using Predictive Model Weights, SIGKDD



VAEP: Valuing Actions by Estimating Probabilities

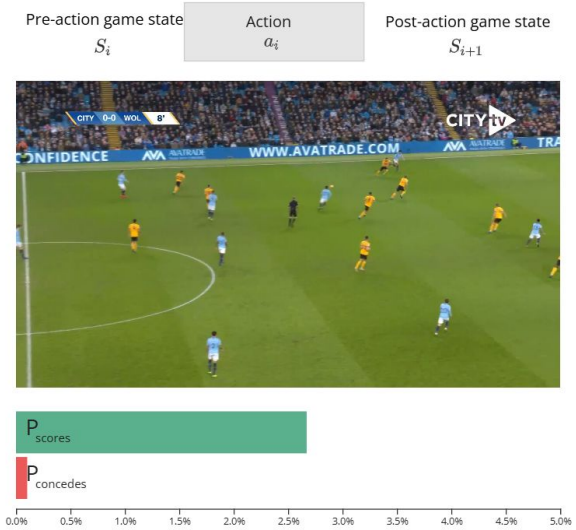
Computes the probability of scoring before and after every player's event on the ball. Delta between such probabilities is the value of the event

Validation: missing

DeCroos et. al., KDD 2019 - Valuing Actions Estimating Probabilities

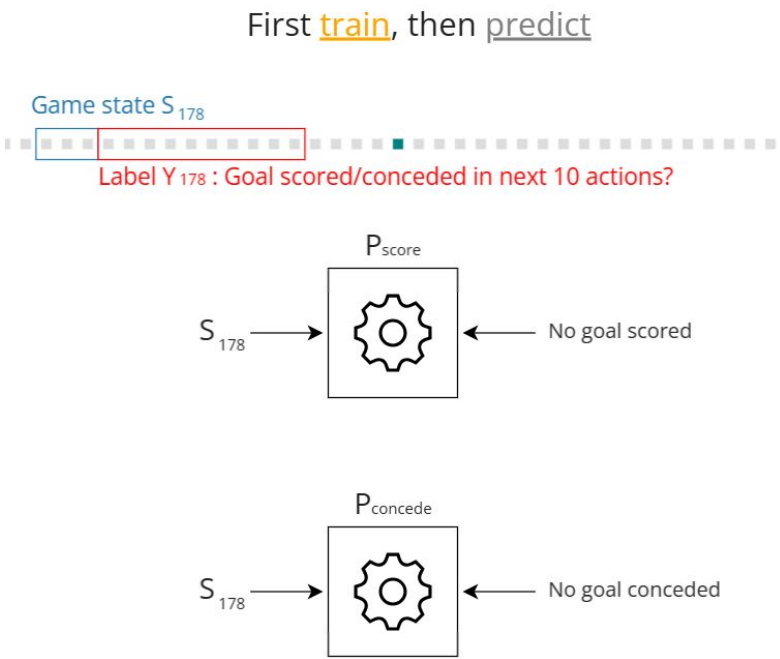
Good actions increase the probability of scoring and decrease the probability of conceding

Notice that the probability of scoring in the near future is much higher in the post-action state, compared to the pre-action state. At the same time, Manchester City's probability of conceding (slightly) decreases. Consequently, a natural way to assess the usefulness of an action is to assign a value to each game state. Then an action's usefulness is simply the difference between the scoring probabilities in the post-action game state and pre-action game state.



VAEP: Valuing Actions by Estimating Probabilities

Predicting the occurrence of a goal within next 10 events, given last 3 event as input



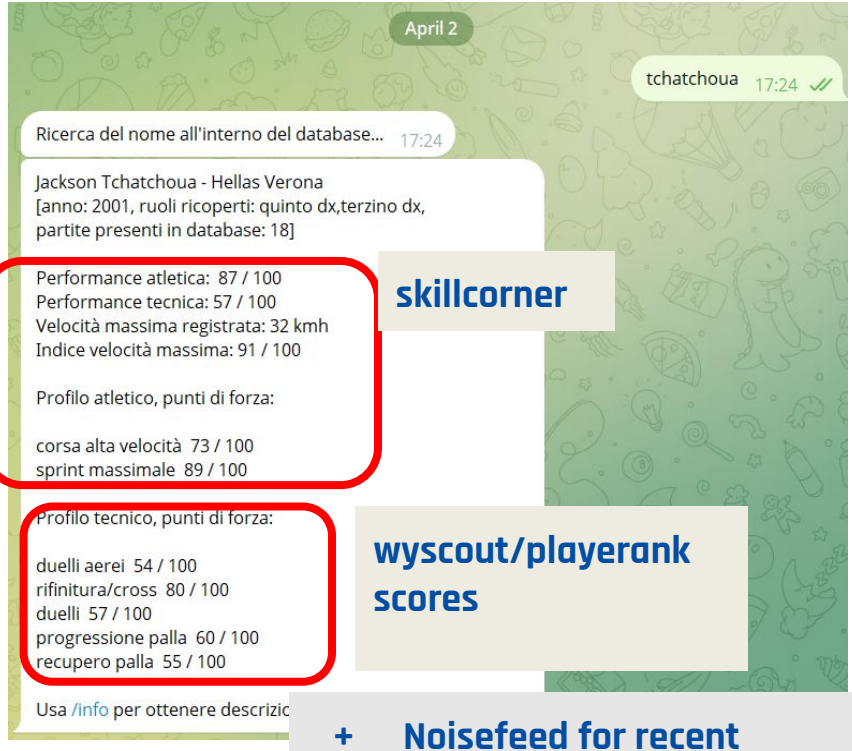
The player scouting dilemma

A real world scouting report in action

Scouting tools in use at Hella Verona FC

Features requested by coaches:

- **Extreme synthesis**
- **No plots**
- **Readable numbers**



Thanks!