DATA+AI SUMMIT 2022

Technical and Tactical Football Analysis

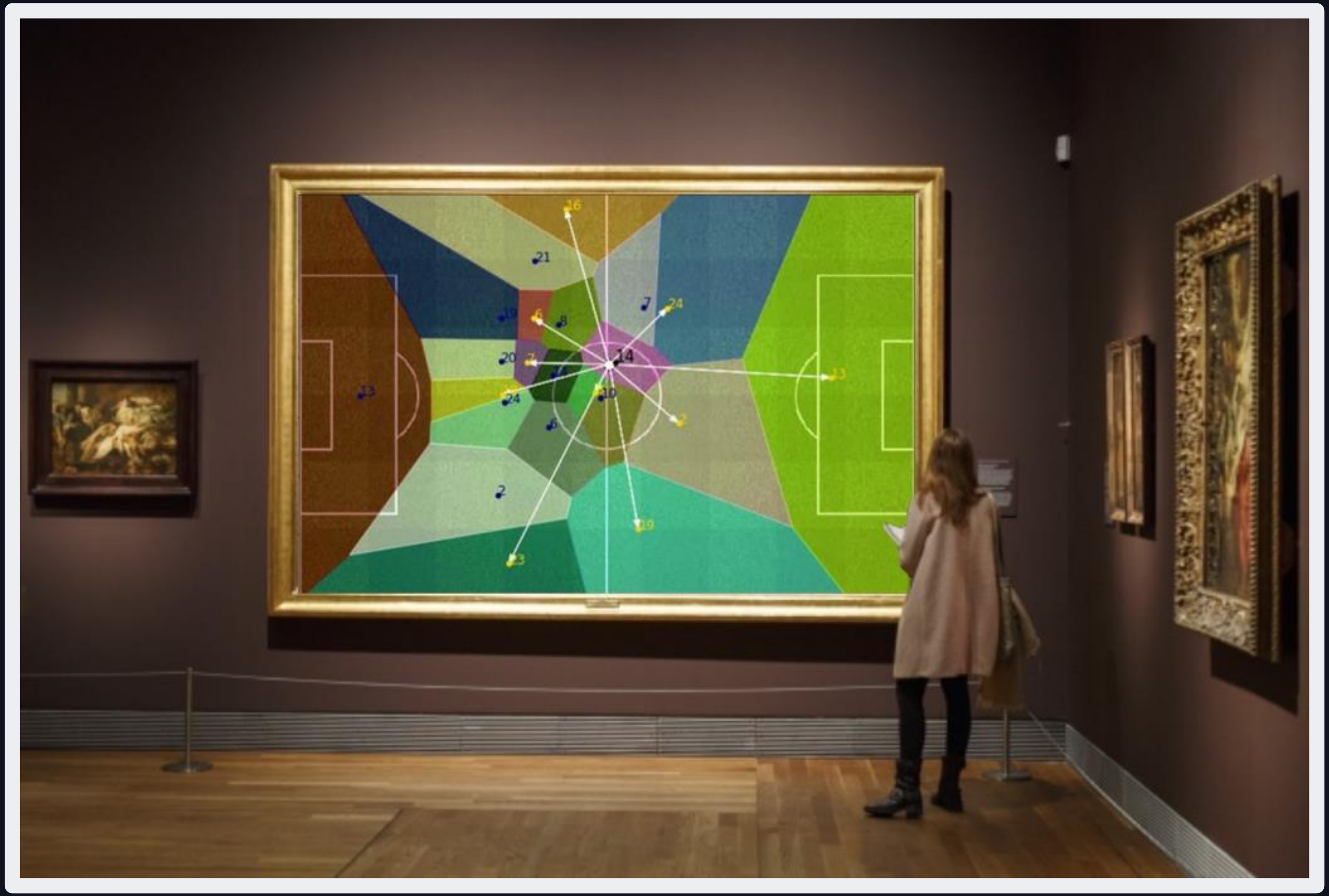
How LaLiga uses data to implement novel analytics and metrics in football

Rafael Zambrano
Head of Data Science, LaLiga Tech

ORGANIZED BY databricks

Introduction

- What do we have
- Data Preparation
- Use Cases
- Goal Probability



What do we have?

What do we have

Event Data

Fouls

Clearances

Goals

Penalty kicks

Passes

Saves

Cards

Injuries

Shots

Free kicks

Misses

Corners

What do we have

Event Data

What do we have

Event Data

```xml
<?xml version="1.0" encoding="UTF-8"?>
<Game id="943096" away_score="0" away_team_id="178" away_team_name="Barcelona"
competition_id="23" competition_name="Spanish La Liga" game_date="2018-03-31T19:45:00"
home_score="0" home_team_id="179" home_team_name="Sevilla" matchday="30"
period_1_start="2018-03-31T19:45:19" season_id="2017" season_name="Season 2017/2018">
        <Event id="1868950374" event_id="1" type_id="34" period_id="1" min="10" sec="4"
        team_id="179" outcome="1" x="0.0" y="0.0" timestamp="2018-03-31T18:57:54.122"
        last_modified="2018-03-31T18:57:54">
                <Q id="1534520812" qualifier_id="194" value="62991" />
                <Q id="2095178222" qualifier_id="130" value="8" />
                <Q id="1878619934" qualifier_id="197" value="260" />
        </Event>
        ...
```

What do we have

Tracking Data

Player and Ball coordinates sampled at 25Hz

What do we have

Tracking Data

3.6 million rows

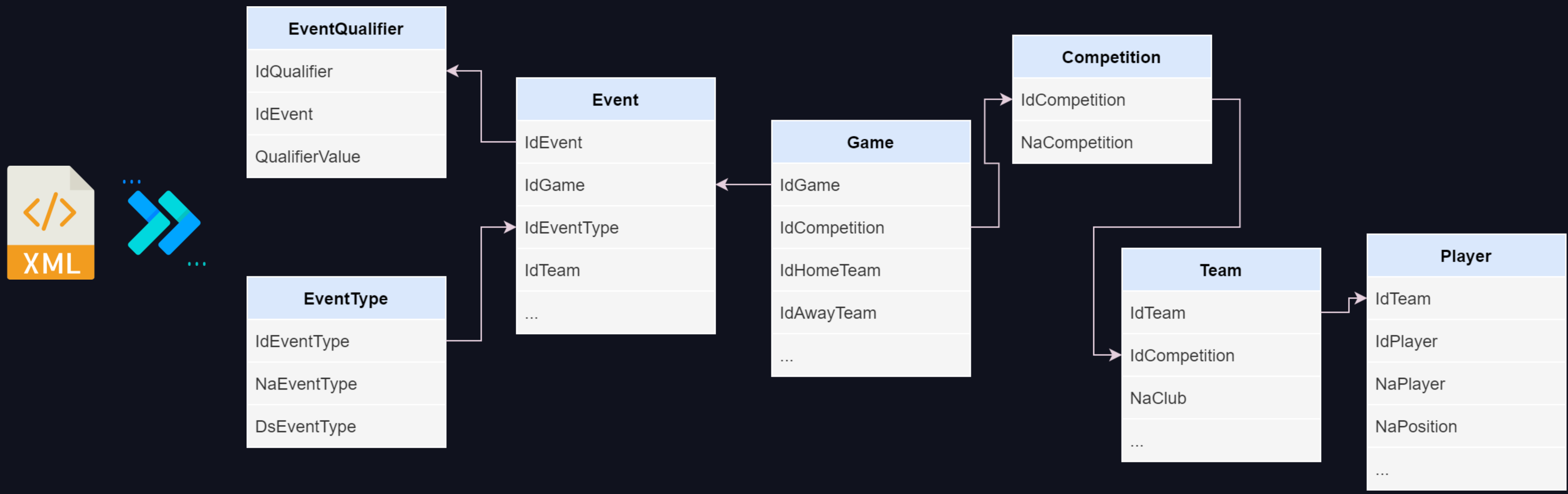

```
1891089:0,1,28,-1112,2770,0.4;1,2,8,834,403,0.2;0,3,3,-913,434,0.00;3,4,1,1643,-3447,0.00;1,5,18,633,-
693,0.22;0,6,4,-2159,-1070,0.00;1,7,11,-20,-924,0.3;1,8,20,1359,-1848,0.16;3,9,2,-
2199,3428,0.03;1,10,2,1421,1873,0.00;0,11,14,-14,-2541,0.28;0,12,20,40,-8,0.76;1,13,6,1495,-
876,0.2;1,14,21,18,956,0.05;1,15,25,4747,7,0.1;1,16,12,-22,1604,0.78;1,17,10,16,-
1639,0.09;0,18,25,57,-1169,1.56;0,19,2,-922,-2849,0.06;0,20,31,-3655,-5,0.64;0,21,27,-
31,2383,0.28;1,22,4,1682,223,0.00;0,23,17,-2219,436,0.00;0,24,8,-1144,-219,0.19;3,25,0,-333,924,0.07;-
1,26,-1,-38,-3483,0.00;4,27,-1,5565,4400,0.00;4,28,-1,5565,4400,0.00;4,29,-1,5565,4400,0.00;:-
83,13,30,0,A,Alive,SetAway;:

1891090:0,1,28,-1110,2769,0.4;1,2,8,833,404,0.13;0,3,3,-913,433,0.00;3,4,1,1642,-
3447,0.00;1,5,18,632,-695,0.3;0,6,4,-2159,-1069,0.00;1,7,11,-21,-925,0.34;1,8,20,1358,-
1850,0.19;3,9,2,-2199,3428,0.03;1,10,2,1421,1874,0.03;0,11,14,-14,-2541,0.28;0,12,20,35,-
10,0.84;1,13,6,1495,-877,0.15;1,14,21,13,952,0.09;1,15,25,4747,7,0.07;1,16,12,-
23,1605,0.66;1,17,10,16,-1638,0.03;0,18,25,55,-1170,1.52;0,19,2,-922,-2849,0.06;0,20,31,-3653,-
4,0.64;0,21,27,-31,2382,0.24;1,22,4,1682,223,0.00;0,23,17,-2219,436,0.00;0,24,8,-1144,-
220,0.19;3,25,0,-333,924,0.07;-1,26,-1,-38,-3483,0.00;4,27,-1,5565,4400,0.00;4,28,-
1,5565,4400,0.00;4,29,-1,5565,4400,0.00;:-120,1,30,0,A,Alive;:

1891091:0,1,28,-1107,2768,0.35;1,2,8,831,406,0.13;0,3,3,-913,433,0.00;3,4,1,1642,-
3448,0.00;1,5,18,630,-696,0.35;0,6,4,-2159,-1069,0.00;1,7,11,-21,-928,0.35;1,8,20,1357...
```


DAT

Data Preparation

Data Preparation

First Step: Transform files into tables



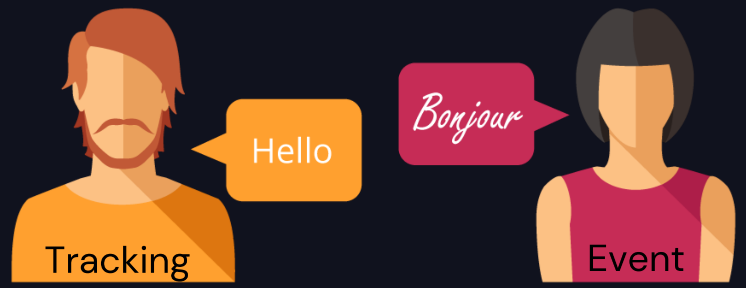
Data Preparation

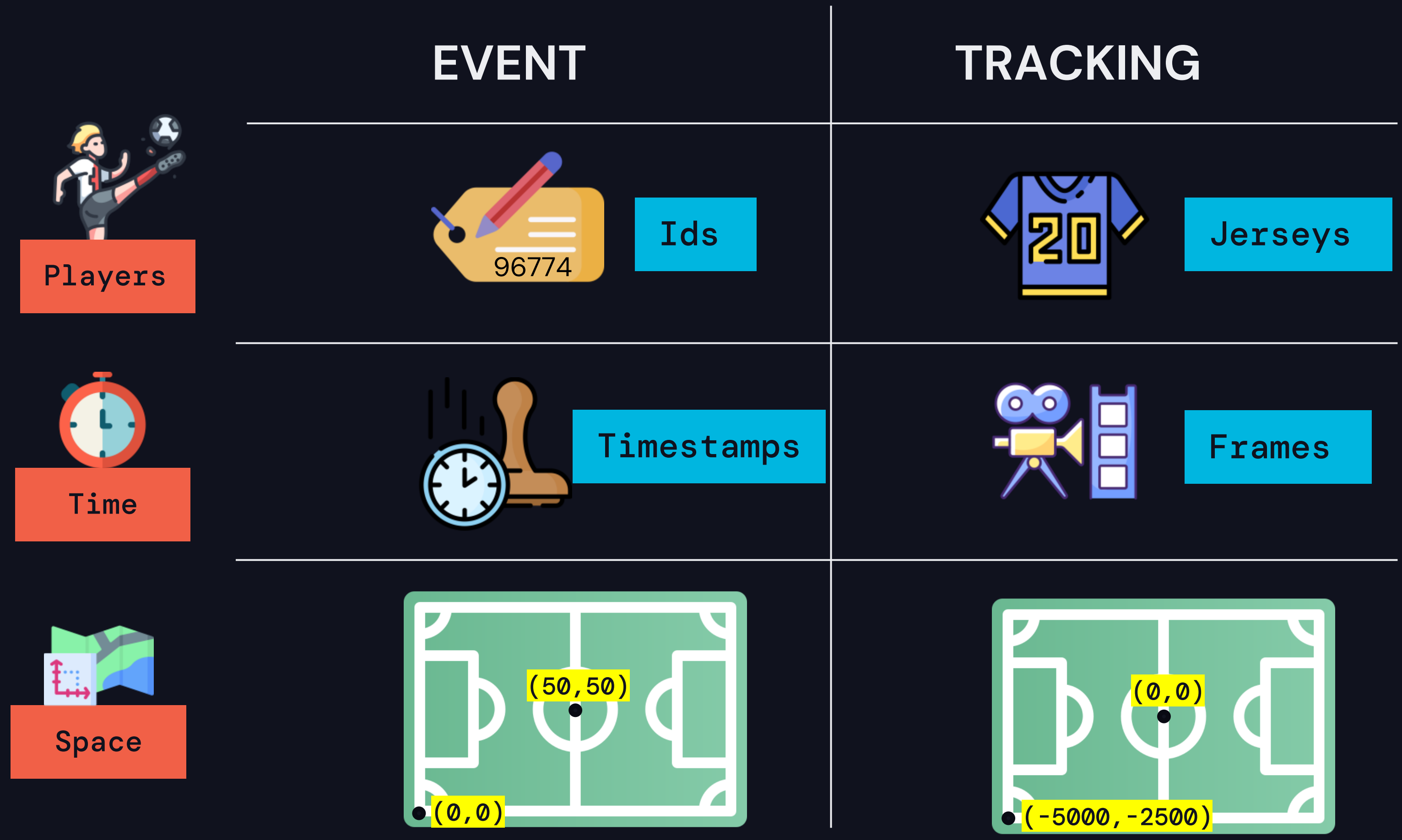
First Step: Transform files into tables

Tracking								
IdFrame	NuJersey	Speed	XPlayer	YPlayer	XBall	YBall	ZBall	IdGame
1891089	10	1.9	-1042	2617	-1022	241	0	12233
1891090	9	2.4	1922	2312	-1021	242	0	12233
...	...	...	...	...	...	...	...	...

Data Preparation

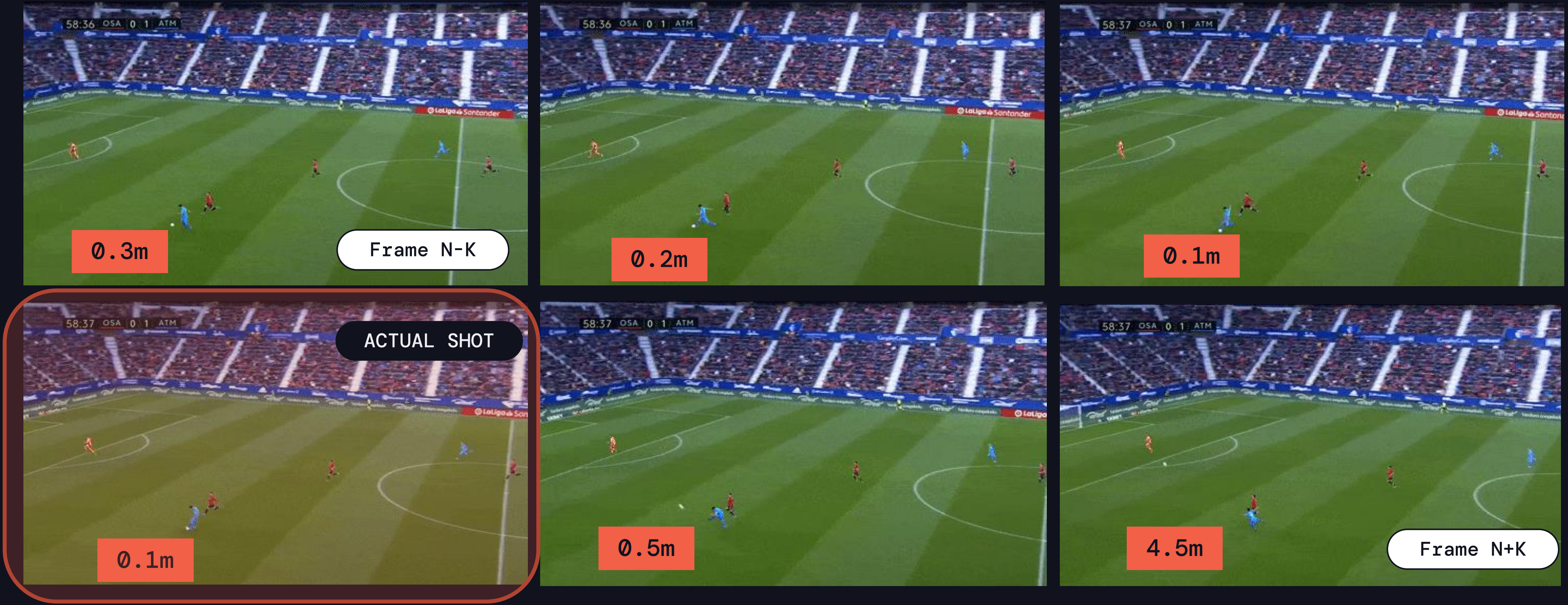
Then, combine event and tracking





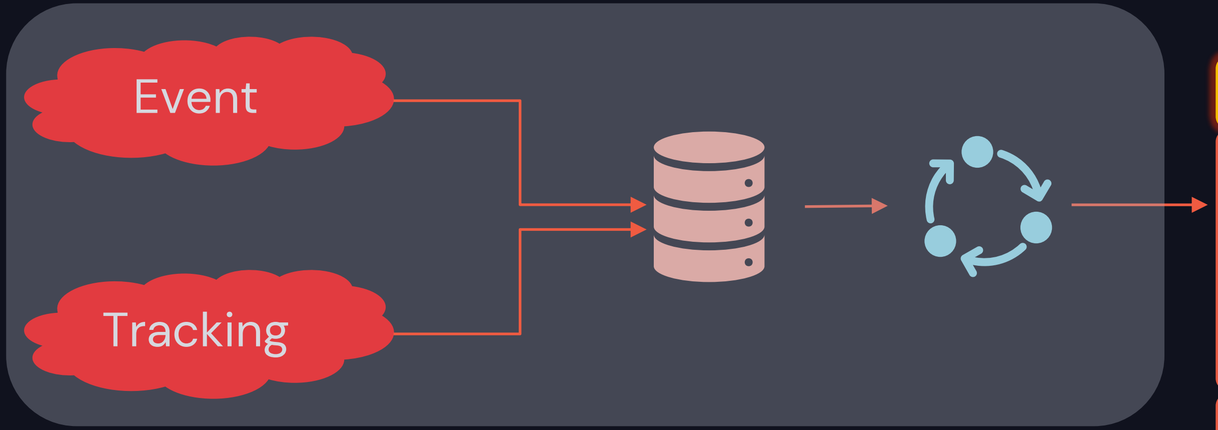
Data Preparation

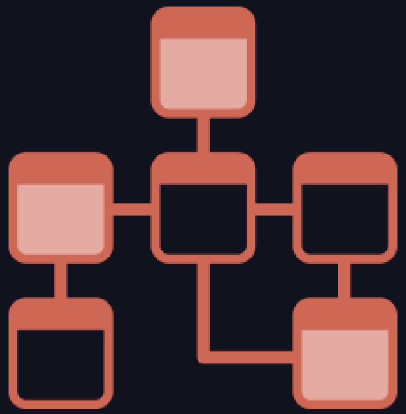
Frame correction

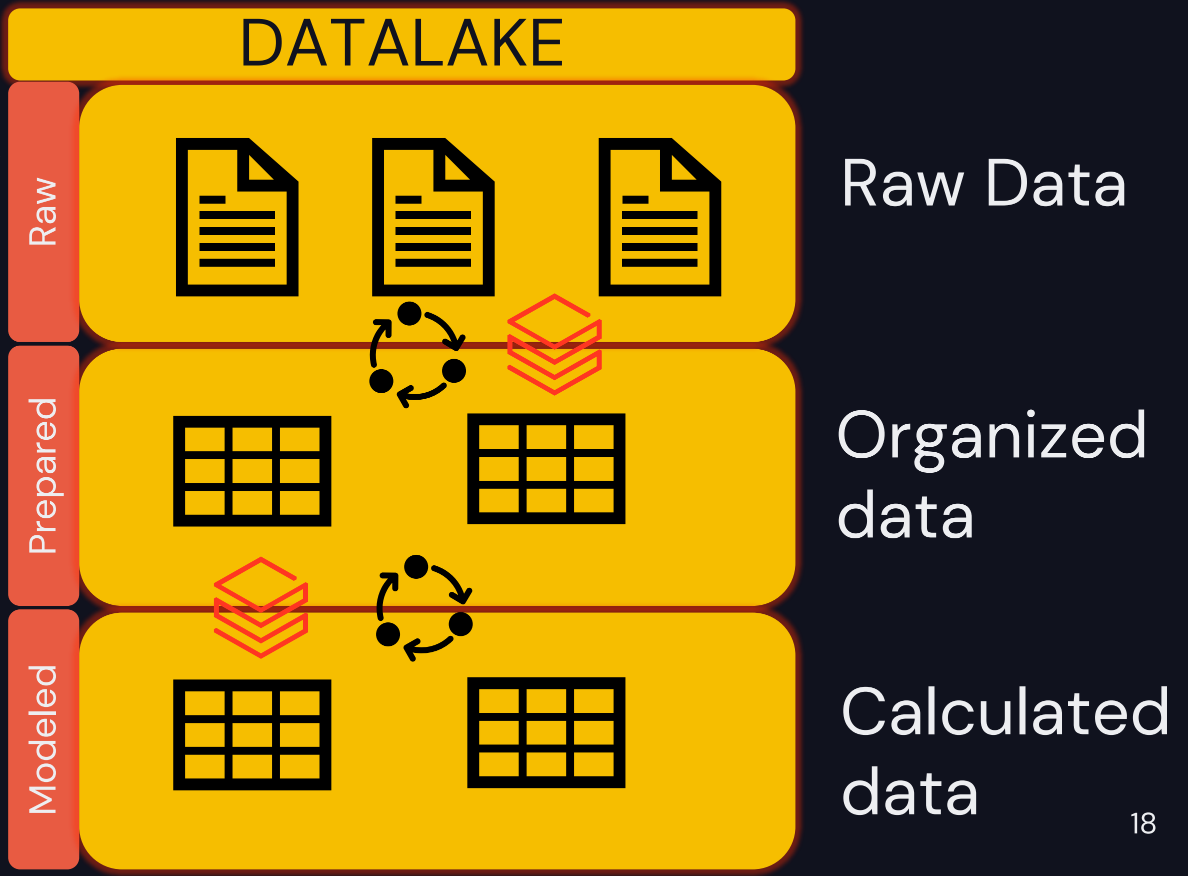


Data Preparation

Combine event and tracking into a data model





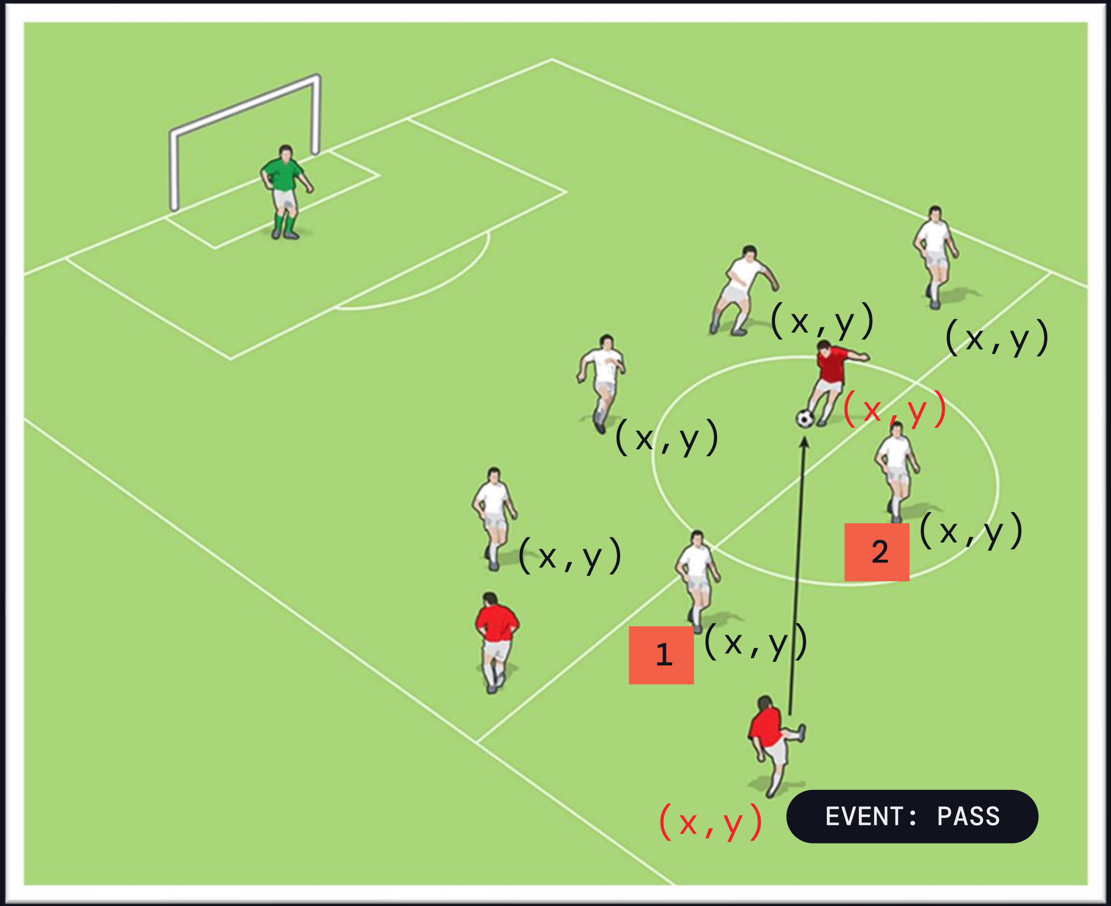


Use Cases

Use Cases

Combine event and tracking data for analyzing matches and developing new metrics

1. OPPONENTS PASSED BY A PASS

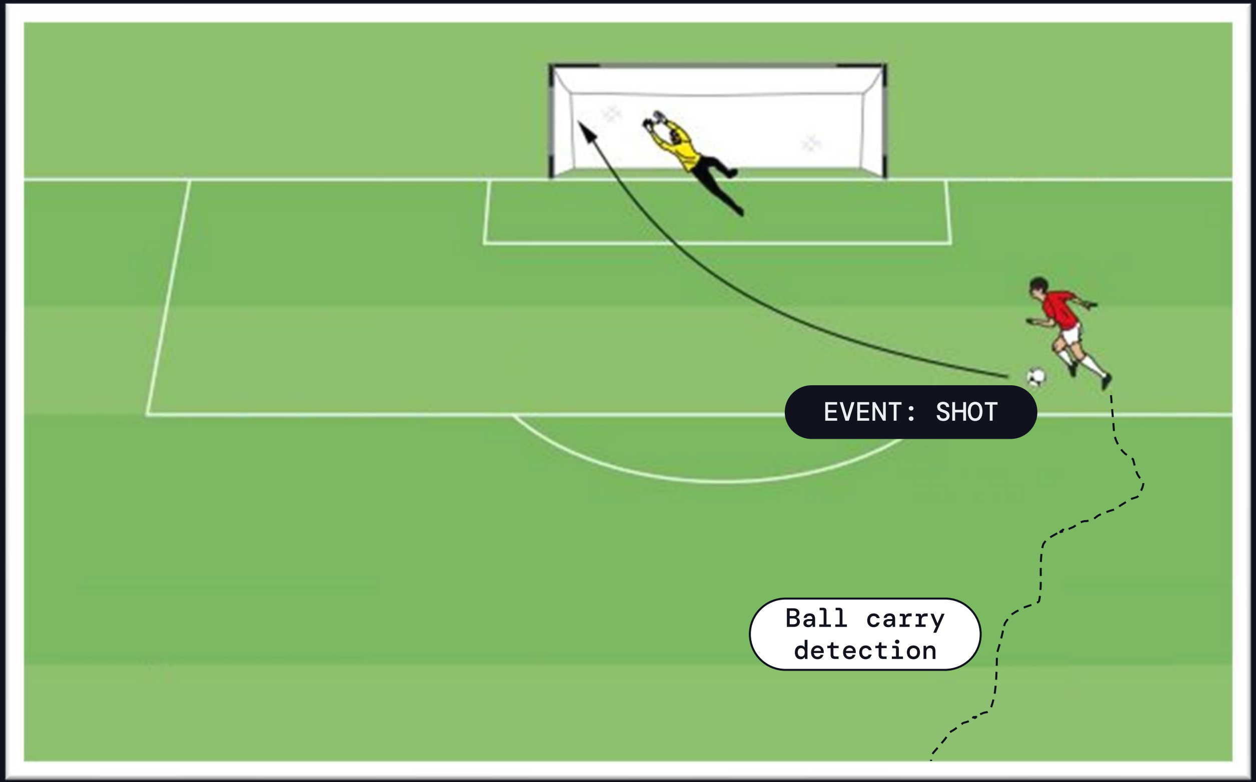


Use Cases

Combine event and tracking data for analyzing matches and developing new metrics

2. BALL CARRIES LEADING TO SHOT

A player carries the ball and finishes with a shot on target

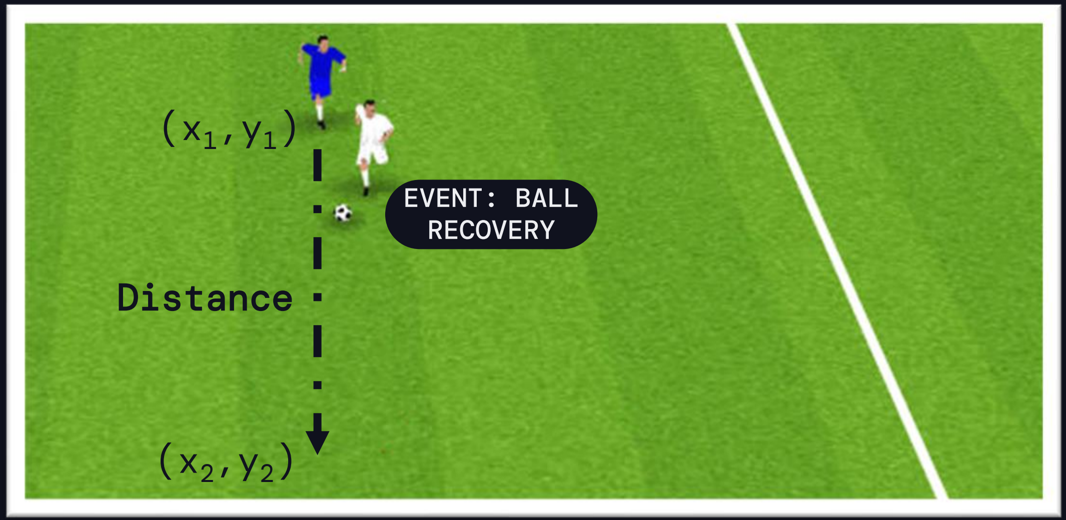


Use Cases

Combine event and tracking data for analyzing matches and developing new metrics

3. DISTANCE COVERED AFTER LOOSING POSSESSION

Distance covered during the 5 seconds after loosing possession

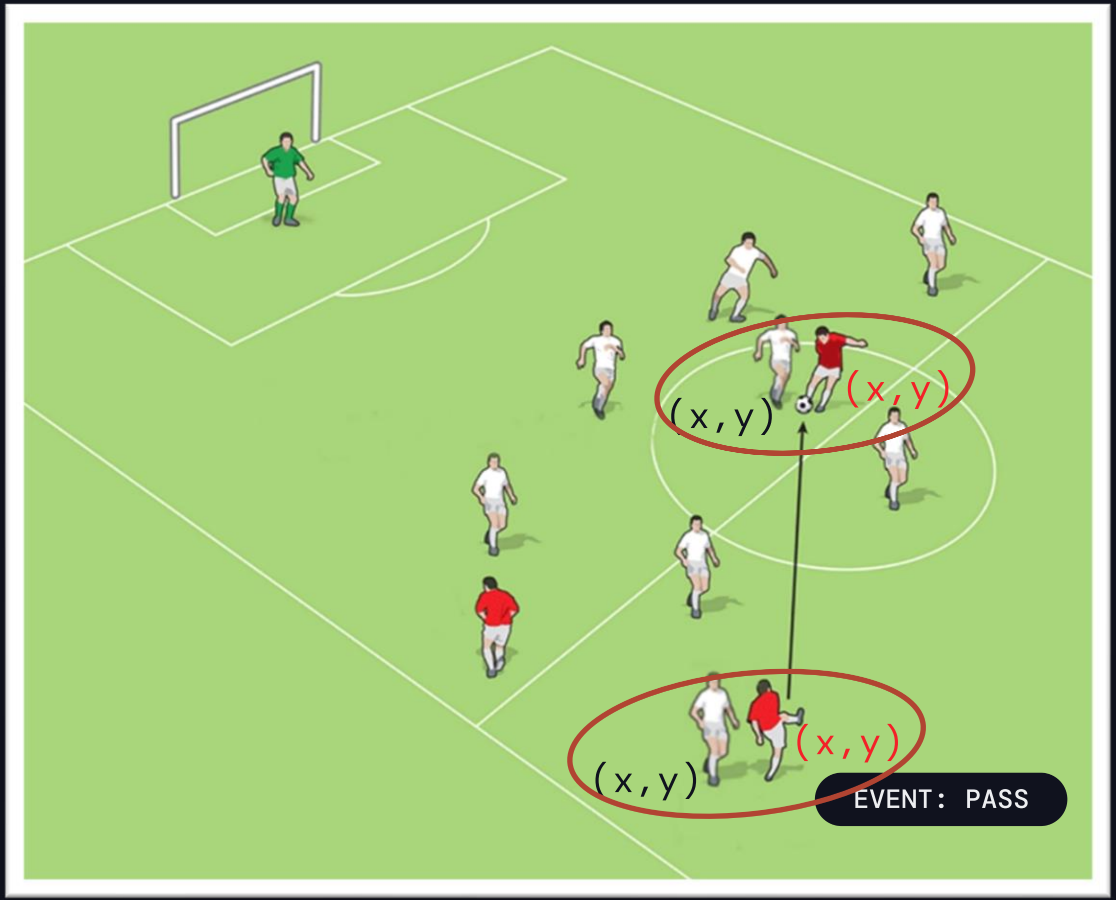


Use Cases

Combine event and tracking data for analyzing matches and developing new metrics

4. PASSES INTO DOUBLE PRESSURE

The player who gives the pass and the player who receives it are under pressure (less than 2 meters)

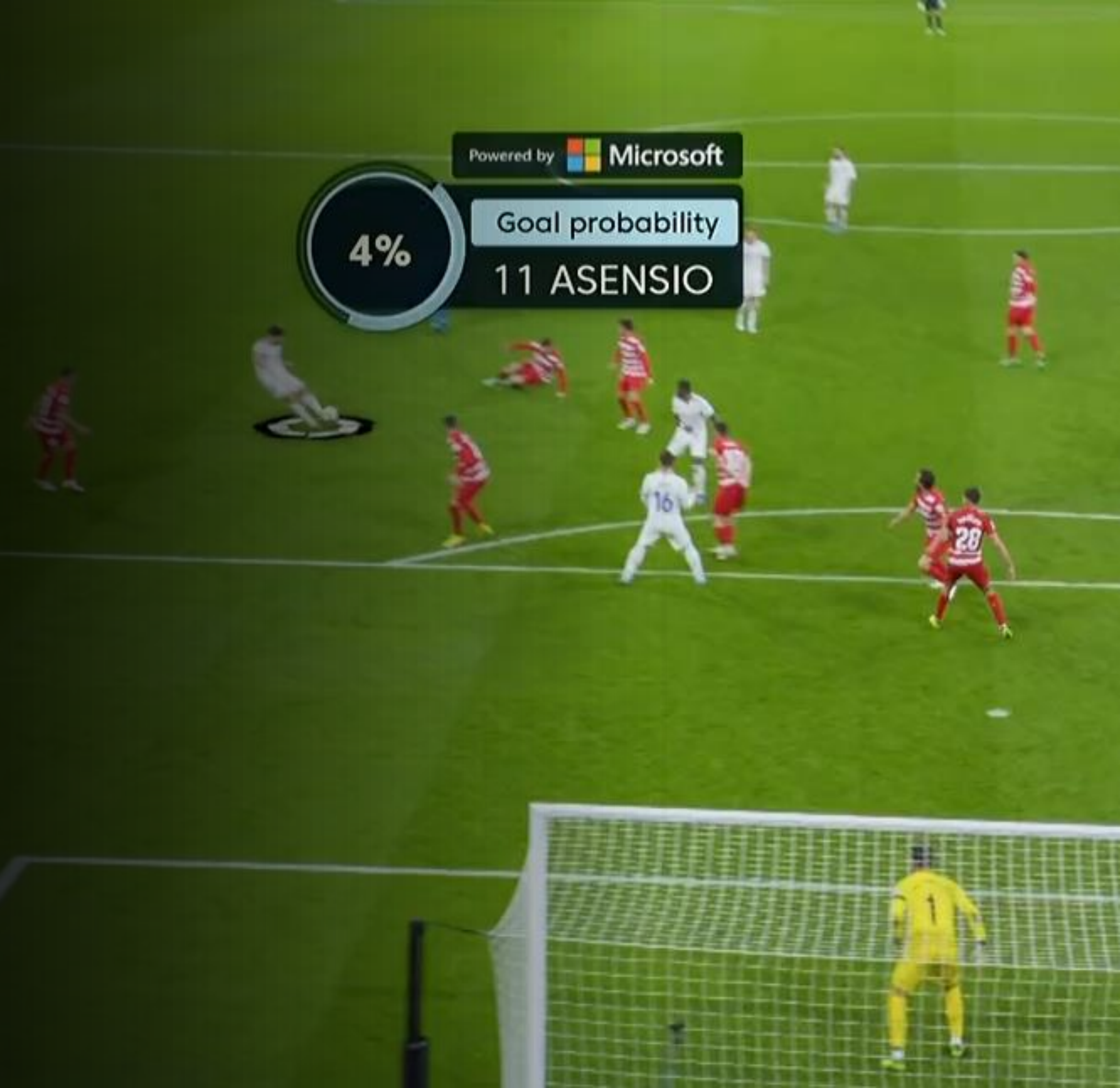


Goal Probability

What is it?

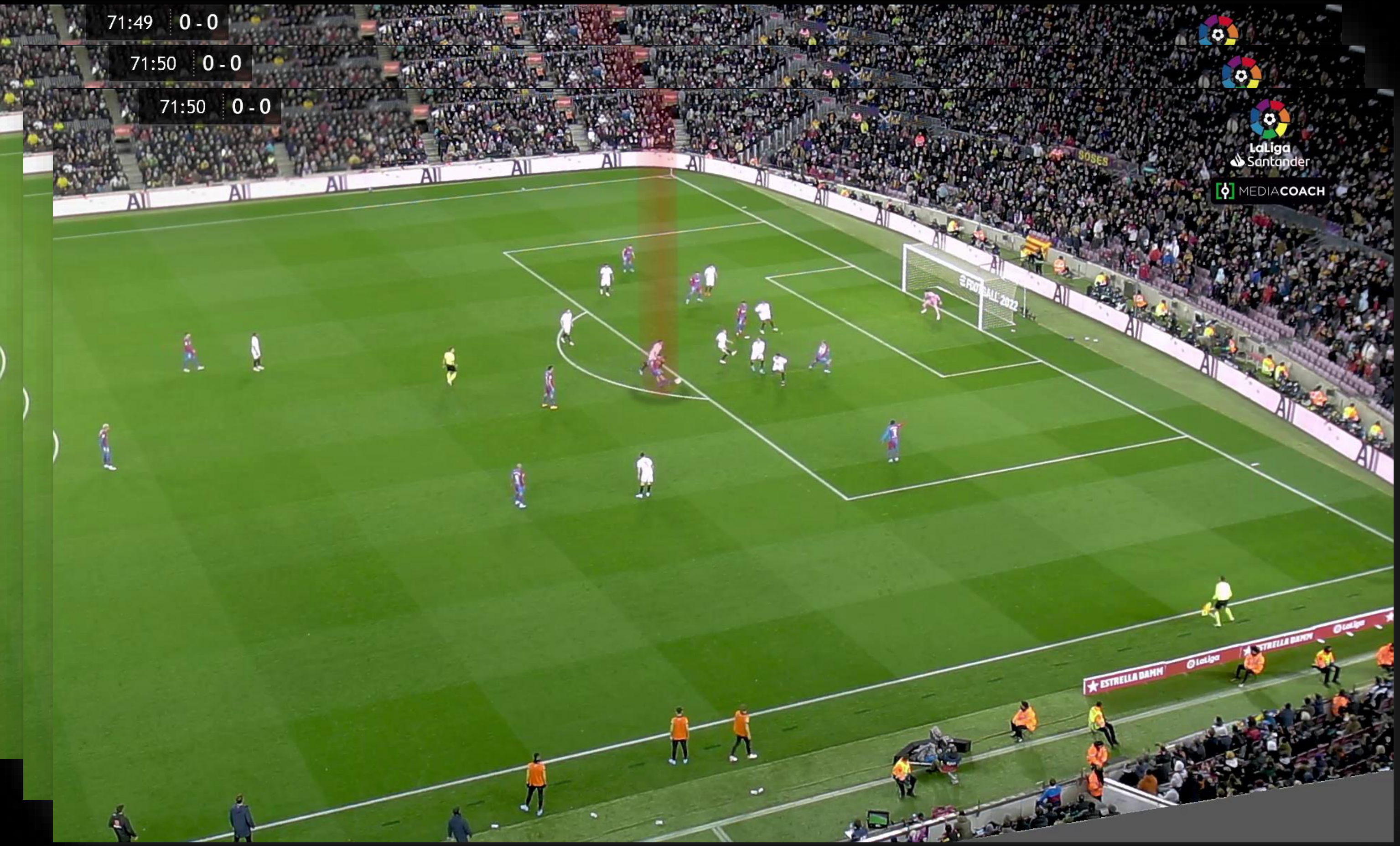
Goal Probability is a metric that evaluates the chances of a player scoring a goal based on his position and the position of his opponents.

It can be applied to any player, regardless of whether he is in possession of the ball or not. In this way, every player has an associate goal probability at each moment of the match



How is it generated?

1. We receive the position of the players and the ball 25 times per second in real time



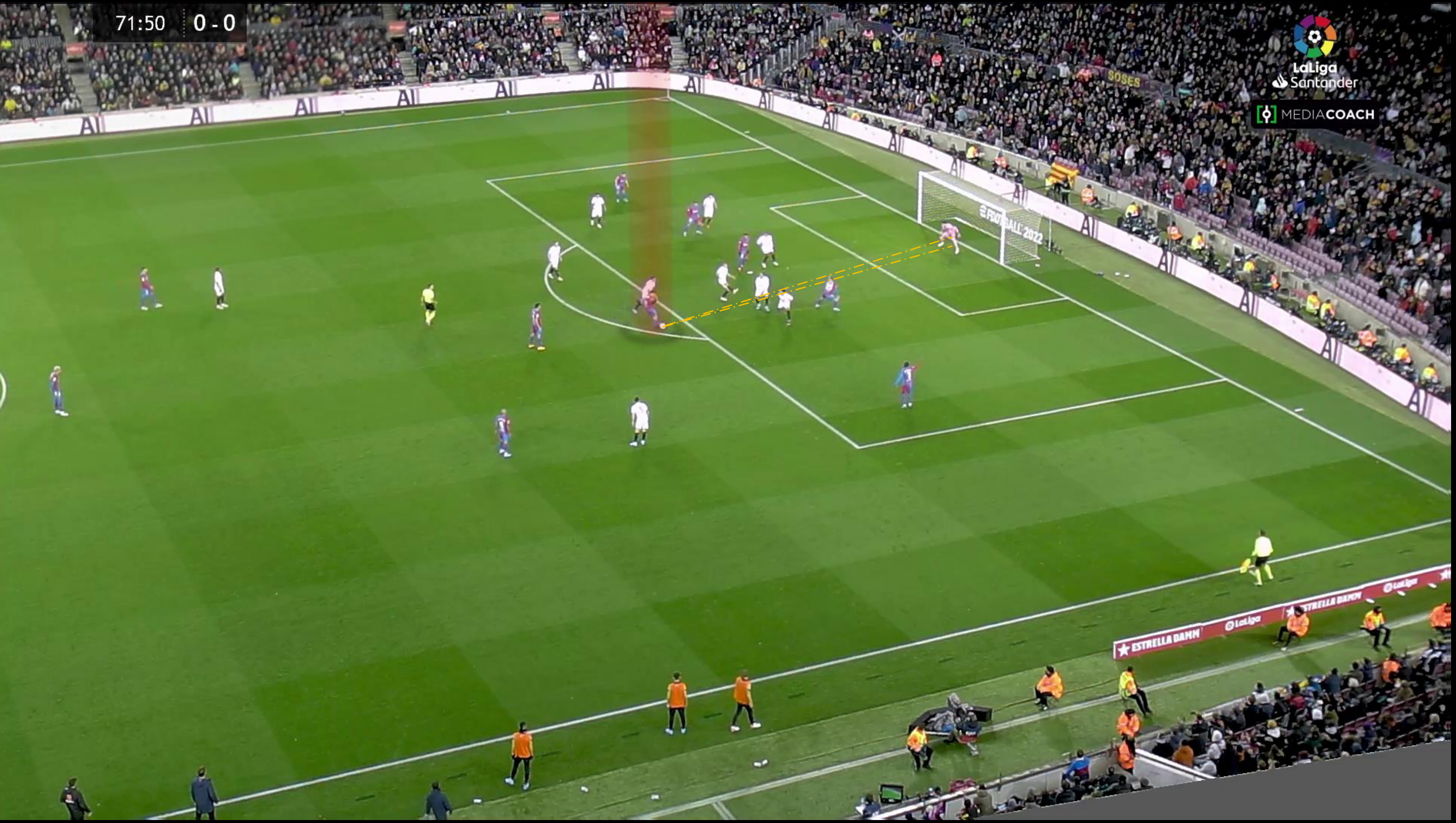
Frame = 2028092
Frame = 2028093
Frame = 2028094

Team Jersev X Y
Team Jersey X Y

Team	Jersey	X	Y
FCB	1	31.18	33.21
FCB	3	67.20	40.44
FCB	4	63.09	29.13
FCB	5	81.46	28.55
FCB	7	91.61	17.13
FCB	8	76.32	18.41
FCB	16	84.81	32.55
FCB	18	93.36	47.40
FCB	19	93.44	40.30
FCB	21	91.11	27.25
FCB	25	92.29	34.41
SEV	4	92.56	26.53
SEV	5	81.18	17.52
SEV	6	92.64	28.81
SEV	8	91.39	32.40
SEV	9	90.06	44.34
SEV	10	85.11	37.44
SEV	13	103.94	33.41
SEV	16	95.60	42.41
SEV	20	87.28	33.68
SEV	22	71.23	39.44
SEV	23	93.21	37.11

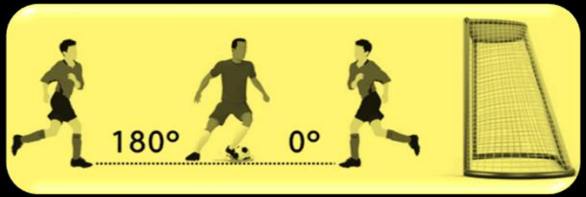
	X	Y
Ball	85.17	32.09

How is it generated?



2. This data is processed in order to calculate the **input variables** to the model

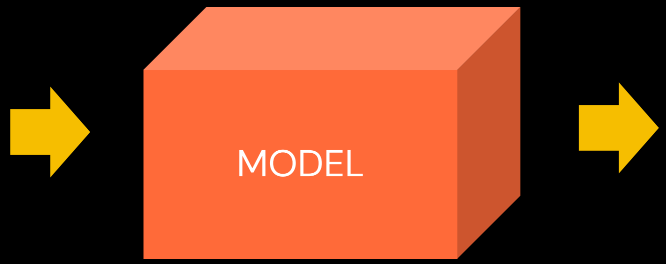
- Distance to goal 18.45
- Distance to goalkeeper 16.42
- Angle to goal 20.4
- Number of opponents in the cone of vision 2.5
- Distance to the nearest opponent 0.3
- Angle to the nearest opponent 25
- One-on-one 0
- Angle to goalkeeper 1.5
- Quality of the player 1



How is it generated?

3. The model returns the goal probability

- Distance to goal 18.45
- Distance to goalkeeper 16.42
- Angle to goal 20.4
- Number of opponents in the cone of vision 2.5
- Distance to the nearest opponent 0.3
- Angle to the nearest opponent 25
- One-on-one 0
- Angle to goalkeeper 1.5
- Quality of the player 1



Goal Probability = 3.4%

INPUT ⟶ MODEL ⟶ OUTPUT

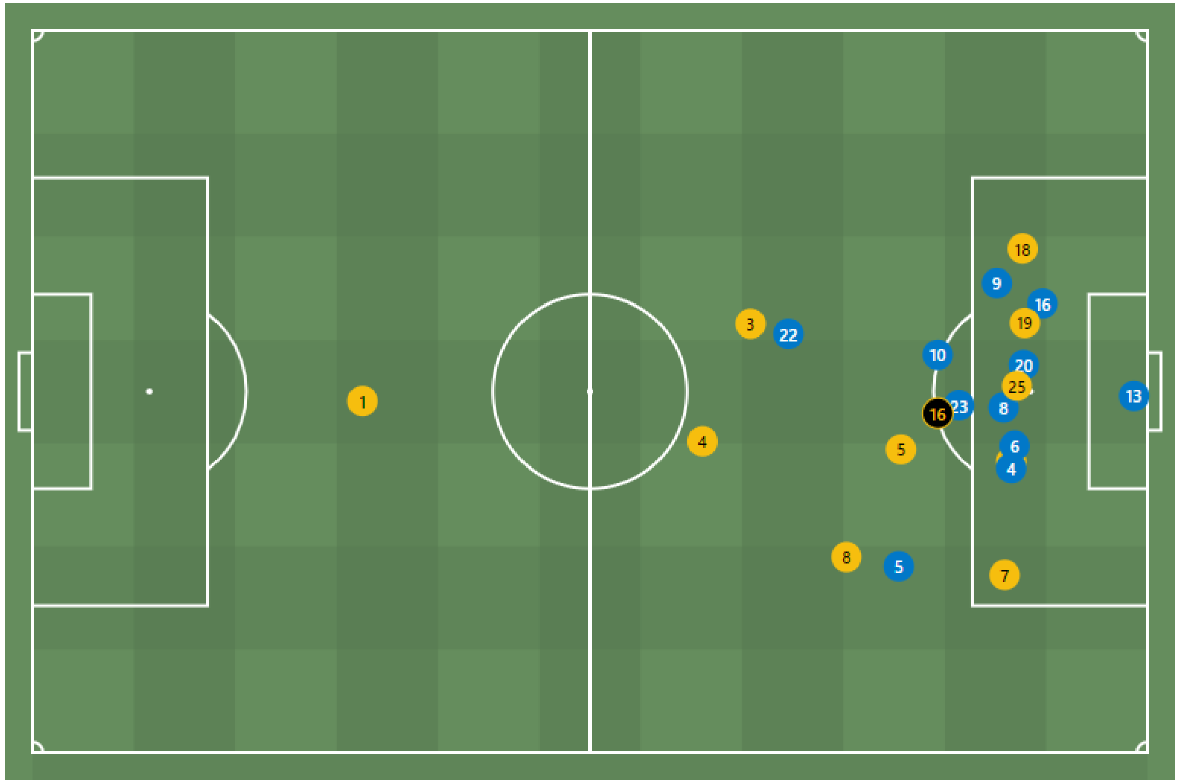
For every frame and every player

How is it generated?

4. We send the results to a web front-end via API

Selector de partido LaLiga Santander | Jornada 30 | Fútbol Club Barcelona - Sevilla Fútbol Club SAD

Selector de objetivo



Local		Visitante	
1	Ter Stegen	1	Dmitrovic
2	Dest	2	G. Montiel
3	Piqué	3	Augustinsson
4	R. Araujo	4	K. Rekik
5	Sergio	5	L. Ocampos
6	Riqui Puig	6	Gudelj
7	O. Dembélé	7	Suso
8	Dani Alves	8	Jordán
9	Memphis	9	Tecatito
10	Ansu Fati	10	I. Rakitic
11	Adama	11	Munir
12	Braithwaite	12	Rafa Mir
13	Neto	13	Bono
14	Nico	15	Y. En-Nesyri
15	Lenglet	16	J. Navas
16	**Pedri**	17	E. Lamela
17	De Jong	18	Delaney
18	Jordi Alba	19	Acuña
19	Ferran	20	Diego Carlos
20	S. Roberto	21	Oliver T.

HH:MM:SS (CET) < SEG < FRAME 22:32:03.23 delay (f) 00:00 > FRAME > SEG APLICAR

Goal Probability → xGoal

Player-Ball distance → Dist. al Balón

xGoal	4.3%	3.7%	3.8%	3.7%	3.8%	3.7%	3.7%	3.6%	3.5%	3.5%	3.5%	3.5%
Dist. al Balón	0.7m	0.7m	0.6m	0.5m	0.5m	0.4m	0.4m	0.3m	0.3m	0.3m	0.3m	0.2m
Frame	11	12	13	14	15	16	17	18	19	20	21	22

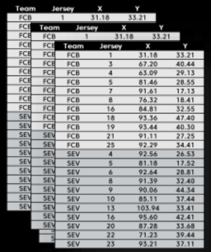
xGoal	3.4%	3.4%	3.4%	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%	2.9%	2.9%
Dist. al Balón	0.2m	1.3m	2.6m	4m	5.5m	6.1m	6.5m	6.9m	7.2m	7.5m	7.6m	7.7m
Frame	**23**	24	0	1	2	3	4	5	6	7	8	9

How is it generated?

5. Goal probability graphic is included in the broadcast

How is it generated?

1. We receive the position of the players and the ball 25 times per second in real time



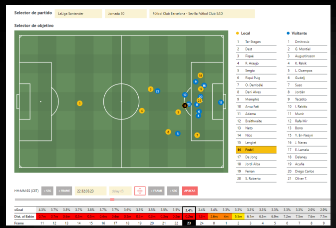
2. This data is processed in order to calculate the input variables to the model

18.45	16.42	20.4	2.5	0.3	25	0	1.5	1

3. The model returns the goal probability

3.4%

4. We send the results to a web front-end via API



5. Goal probability graphic is included in the broadcast

< 30 sec.

Amount of Data

3 million data for every game
30 million data for every matchday
High processing capacity

Multidisciplinary Team

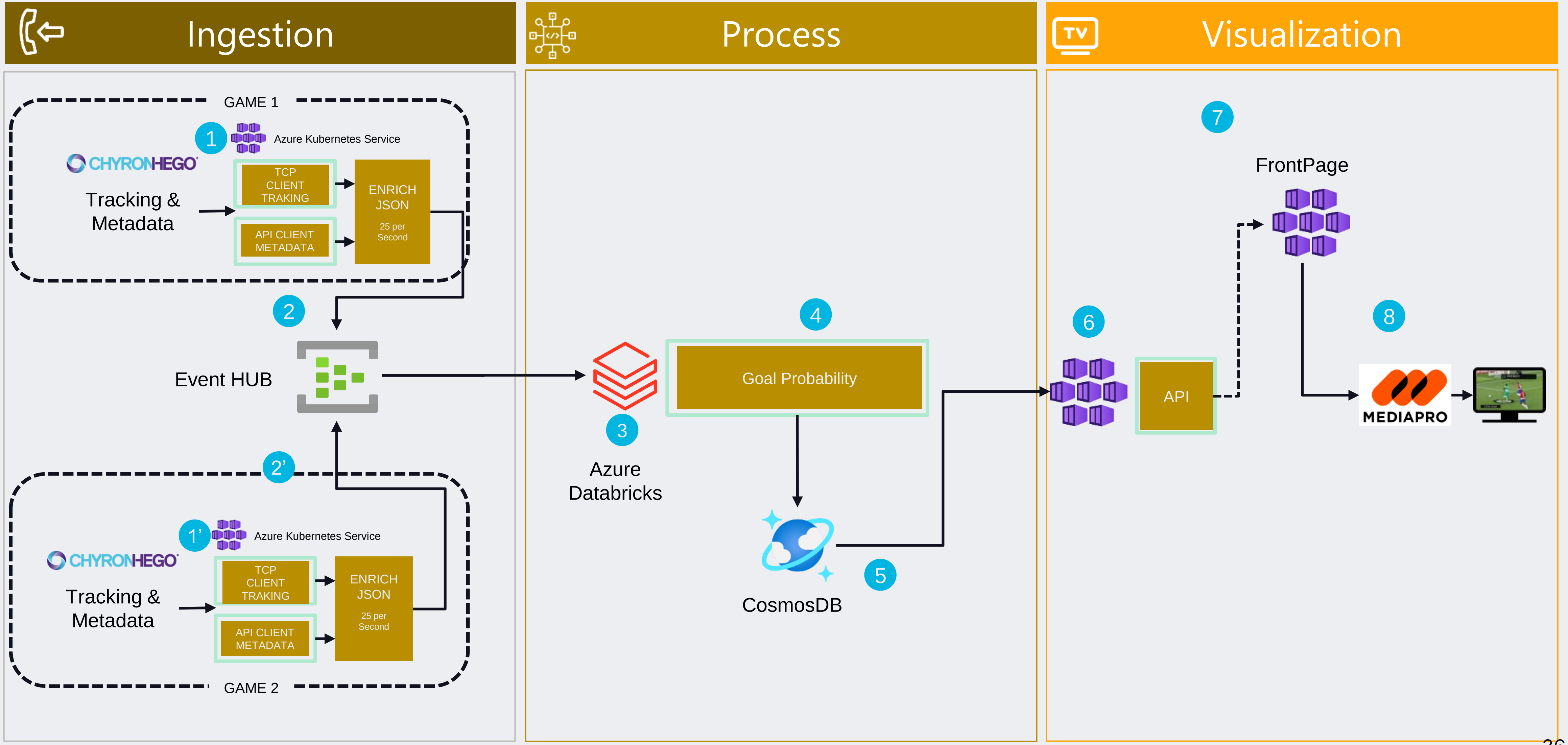
Data Scientists, Data Engineers, Data Arquitects, Audiovisual and Football experts make up the work team

Partnership

Partnership with Microsoft and its software tools helped develop the solution

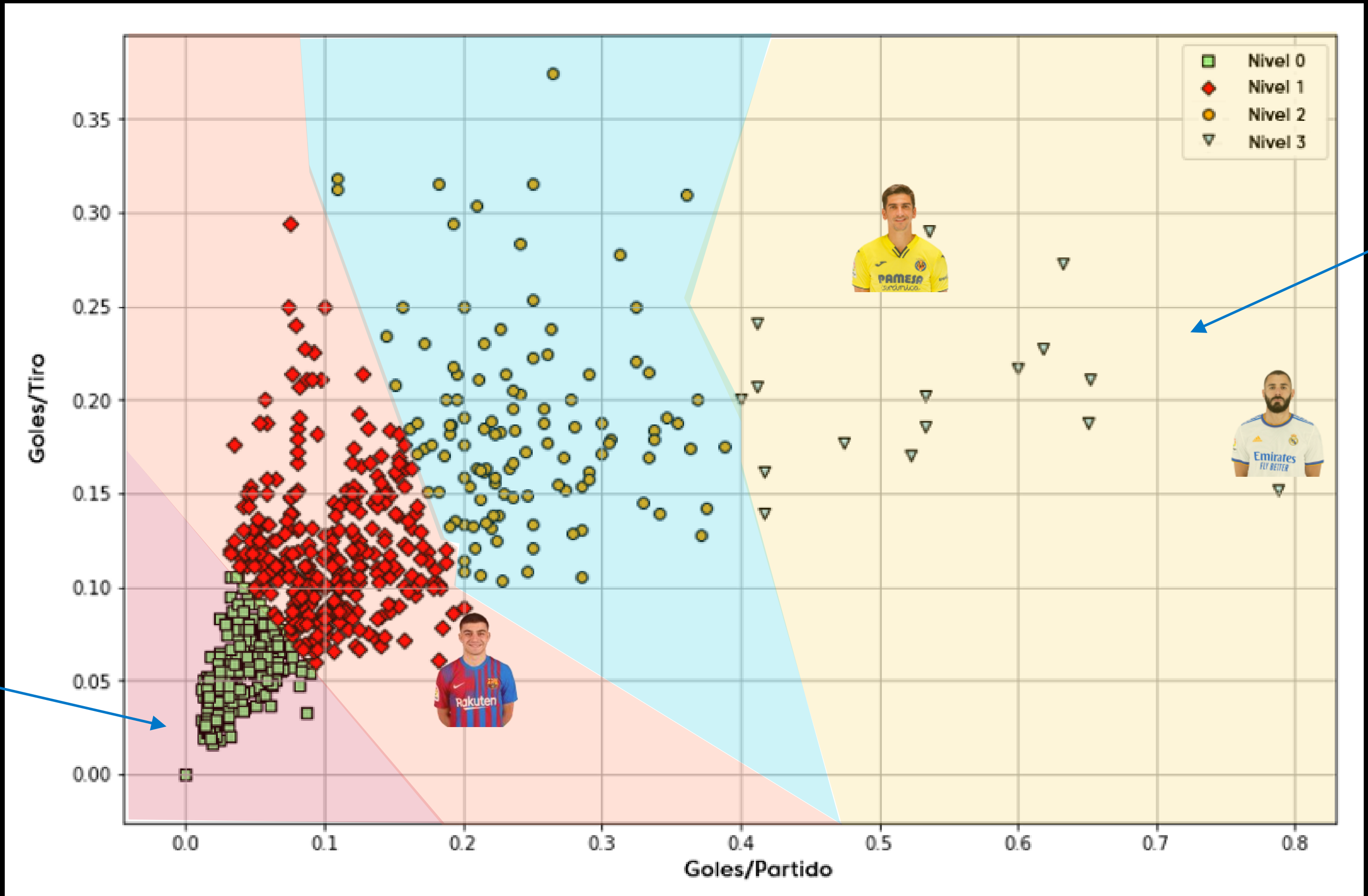
Microsoft

Technology’s Arquitecture



About the quality of the player

Players are grouped into 4 groups, based on the **number of goals per game** and **the number of goals per shot**



More likely to score

Less likely to score

About the model

>2.000 games

~20.000 shots

- We use historical data to collect examples of shots
- For every shot, we calculate and store the variables along with the output (goal/no-goal)

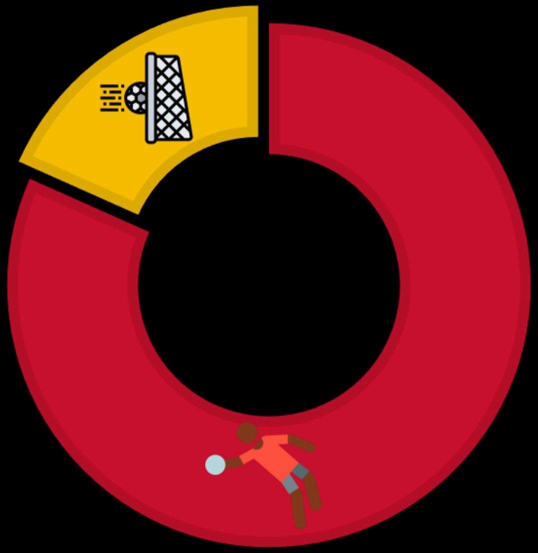
Tracking

Event

VARIABLES			OUTPUT
Distance to goal	Distance to GK	…	
4	2	…	1
5	6	…	0
5	4	…	0
25	6	…	1
12	4	…	0
…	…	…	…

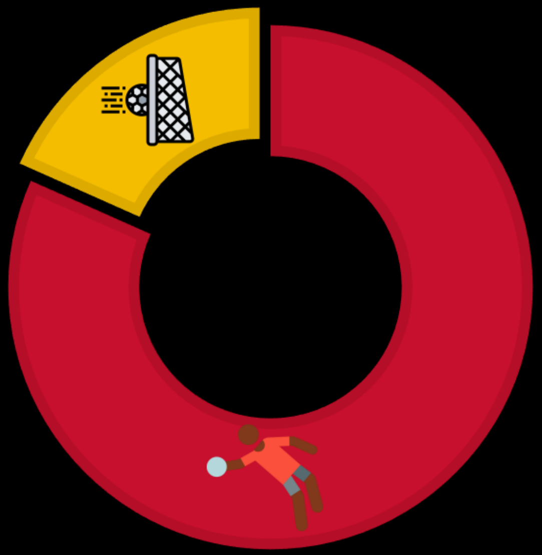
About the model

- Imbalanced dataset (~20% of shots on target are goals)
- Need for explainability



About the model

- Imbalanced dataset (~20% of shots on target are goals)



Goal distance	GK distance	Goal Angle	Opp. In cone	Dist. nearest	Angle nearest	One-on-one	GK Angle	Player Quality	Goal
5	6	30	0	1	10	0	10	1	1
15	10	20	2	3	55	0	15	3	1

Synthetic new sample

10	8	25	1	2	30	0	12	2	1

About the model

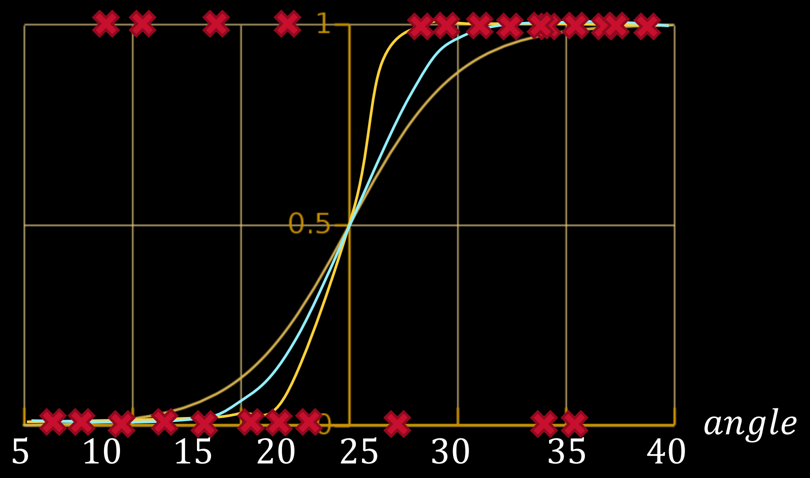
- **Logistic regression** is a traditional statistics technique that models the probabilities for classification problems

  Let's consider the angle to goal as the only explanatory variable for scoring

Sigmoid function

$$Probability = \theta(w_0 + w_1 \cdot angle)$$

Angle	Goal
5	1
35	0
21	0
10	1
...	...



We don't have access to the actual probability

We can only observe the occurrence of an event and try to infer that probability

About the model

$$\theta(s) = \frac{e^s}{1+e^s}$$

$$Probability = \theta(w_0 + w_1 \cdot angle)$$

Angle	Goal scored
5	1
35	0
21	0
10	1
…	…

$w_0 = 1.2$
$w_1 = -0.07$

Probability
70%
22%
43%
62%
…

$w_0 = 2.3$
$w_1 = -0.15$

Probability
82%
5%
3%
69%
…

Which one is better?

Maximum likelihood estimation

Select the parameters w_0 , w_1 that maximize this likelihood

$$\prod P(y|x) = 0.82 \times (1-0.05) \times (1-0.03) \times 0.69 = 0.52$$

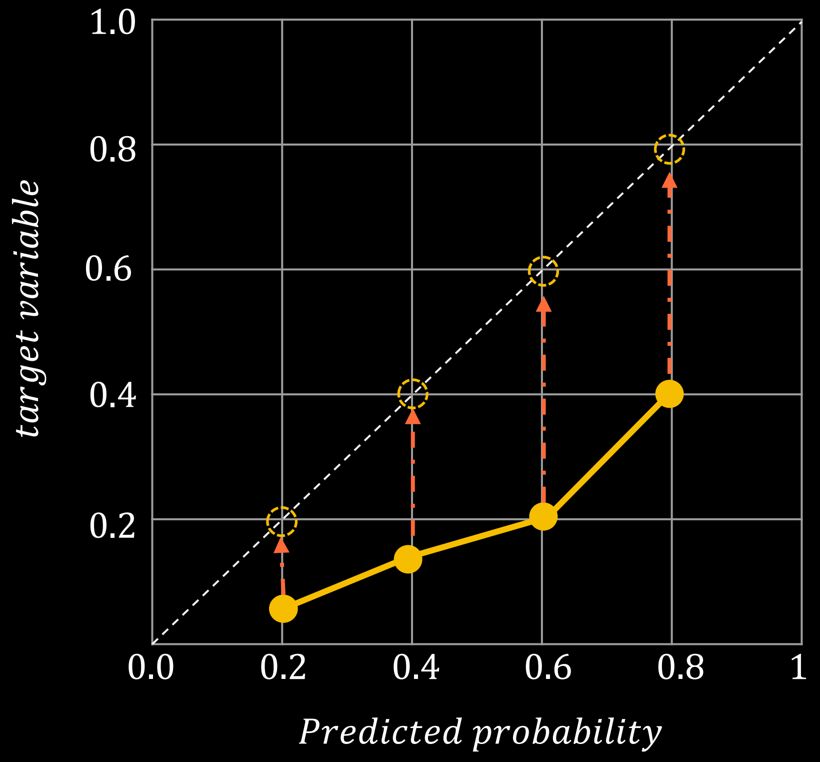
$$\prod P(y|x) = 0.7 \times (1-0.22) \times (1-0.43) \times 0.62 = 0.19$$

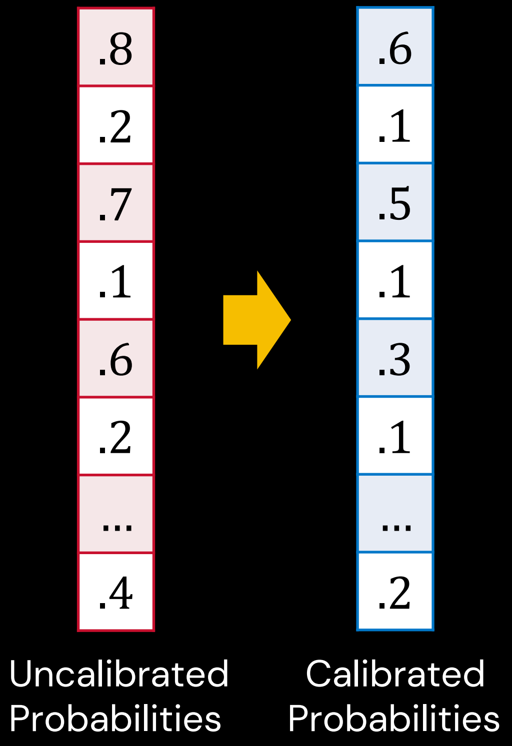
About the model

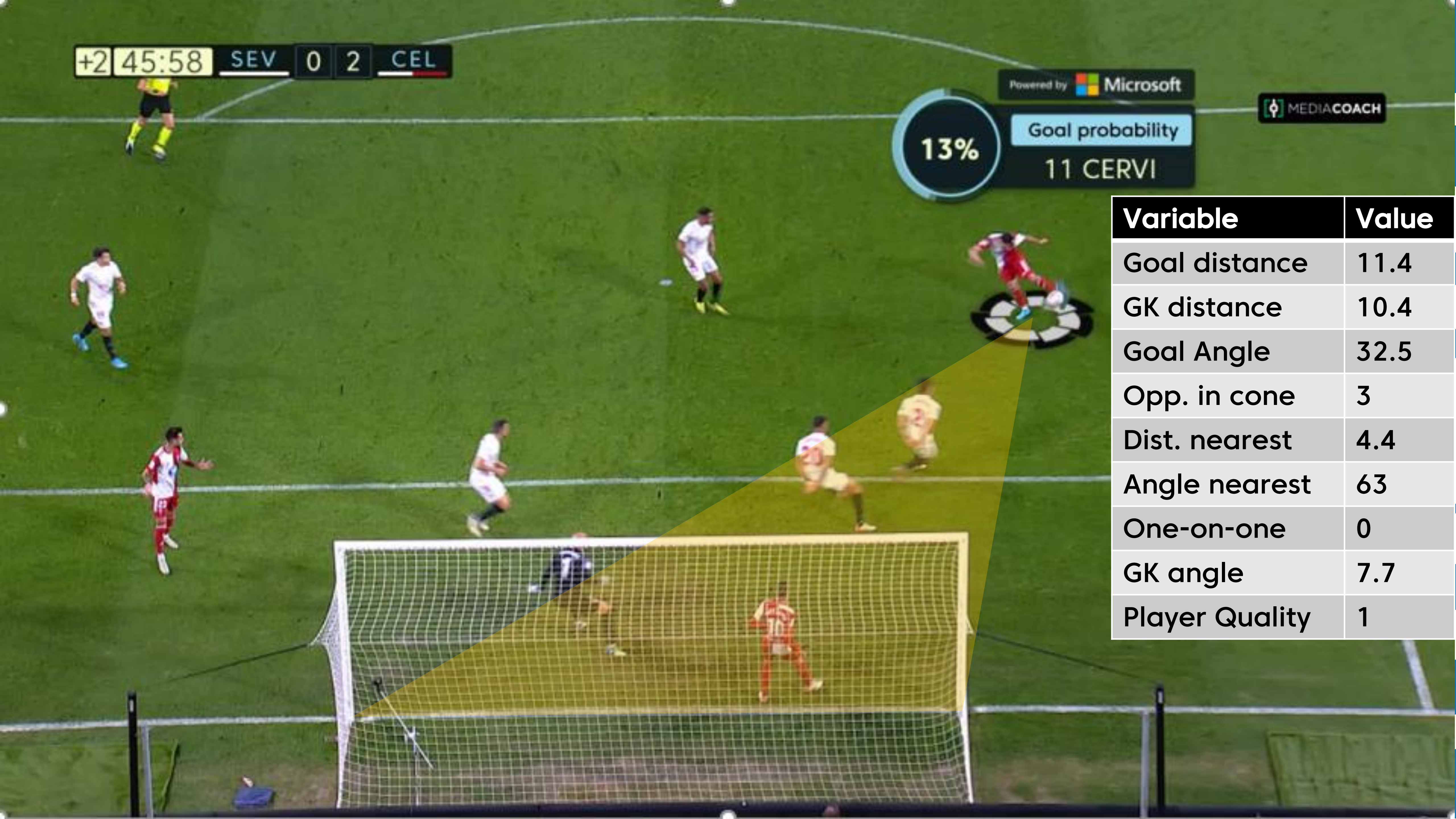
- Using all our variables:

$$Probability = \theta(w_0 + w_1 \cdot distance + w_2 \cdot angle + w_3 \cdot opp\ in\ cone + \cdots)$$

- However, this model doesn't give real probabilities, since we trained it with a balanced dataset

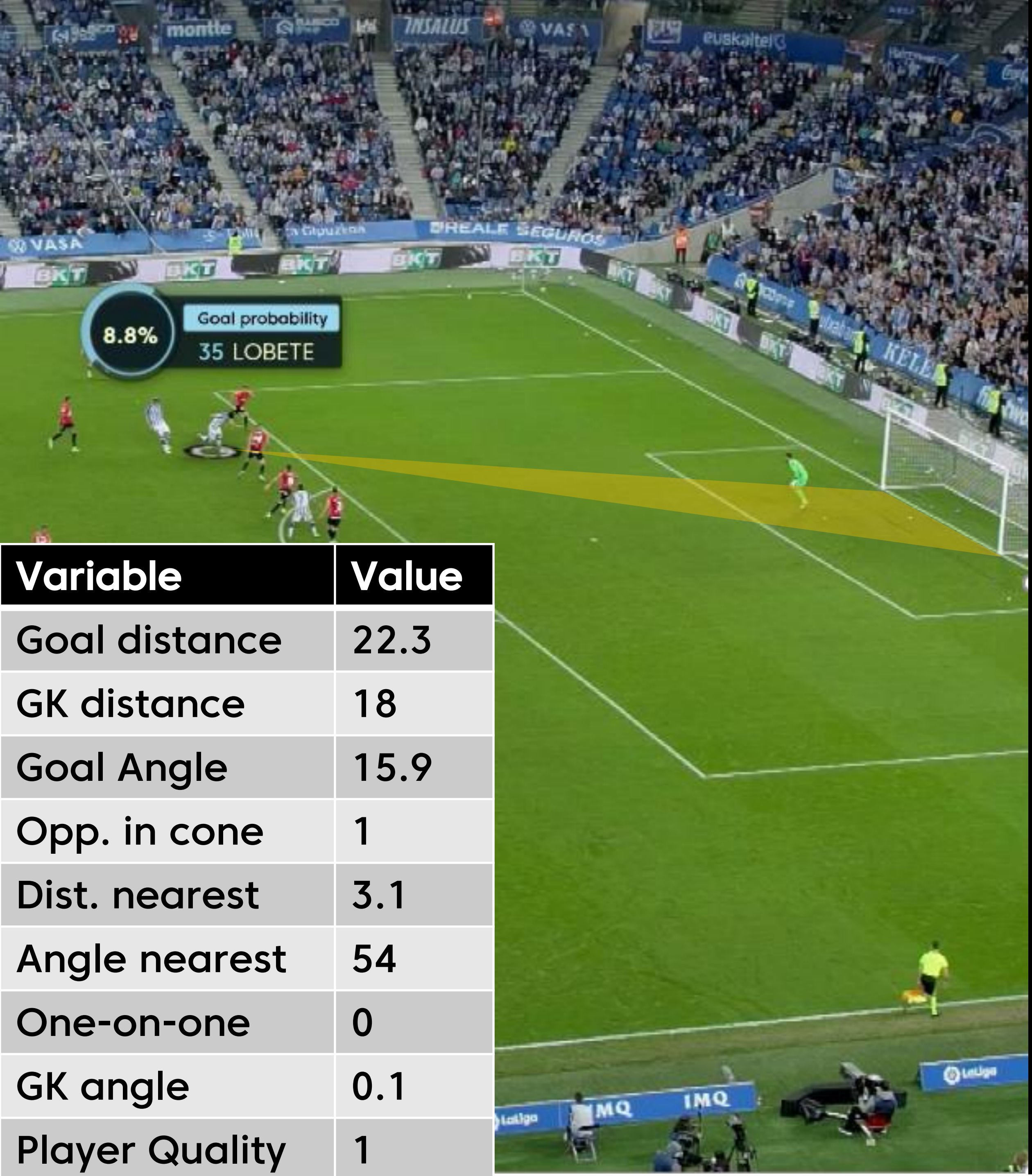






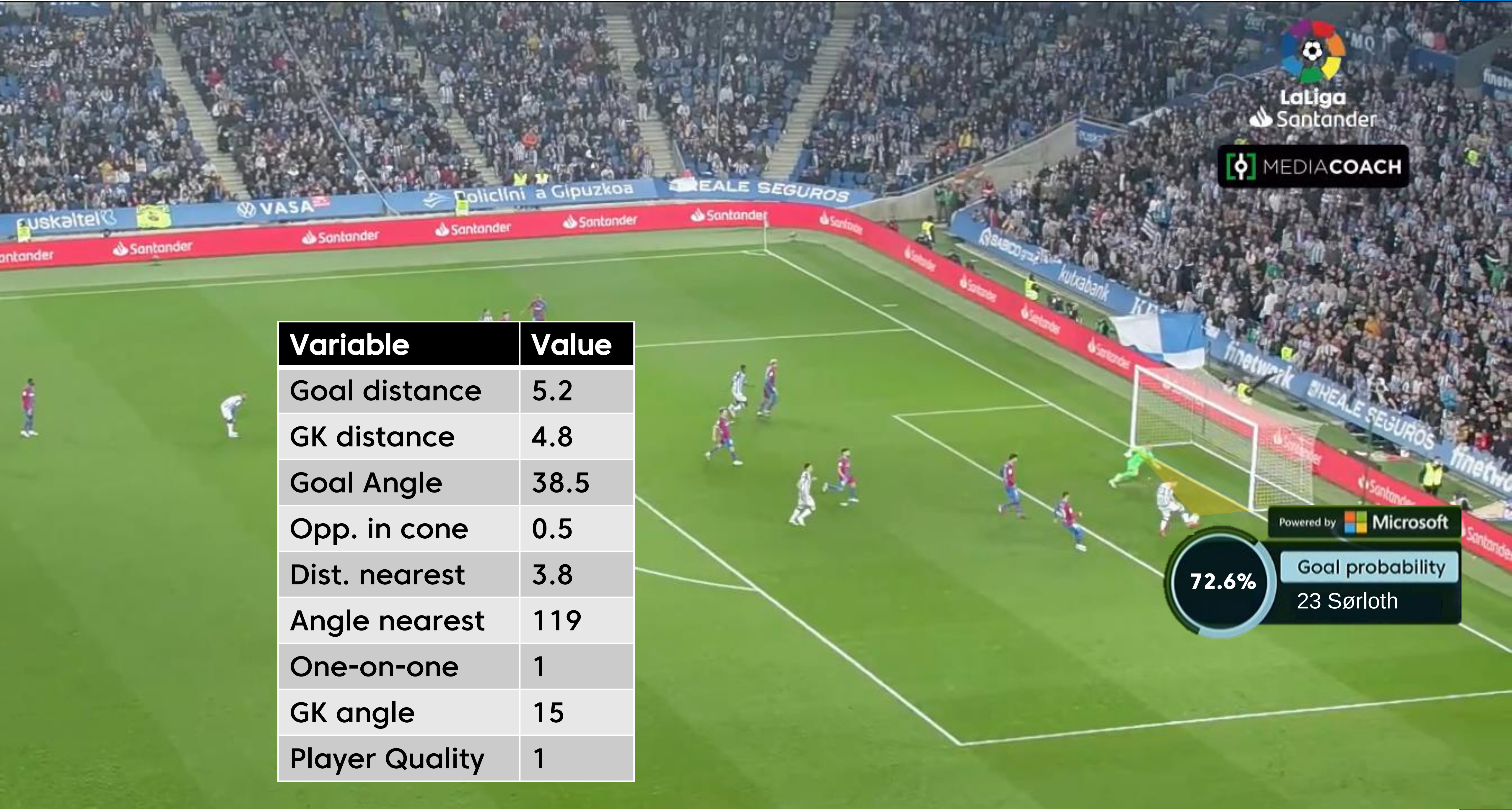
Variable	Value
Goal distance	11.4
GK distance	10.4
Goal Angle	32.5
Opp. in cone	3
Dist. nearest	4.4
Angle nearest	63
One-on-one	0
GK angle	7.7
Player Quality	1

Variable	Value
Goal distance	6.4
GK distance	5.2
Goal Angle	57.7
Opp. in cone	1.5
Dist. nearest	1.3
Angle nearest	138
One-on-one	0
GK angle	5.8
Player Quality	1



Variable	Value
Goal distance	22.3
GK distance	18
Goal Angle	15.9
Opp. in cone	1
Dist. nearest	3.1
Angle nearest	54
One-on-one	0
GK angle	0.1
Player Quality	1

Variable	Value
Goal distance	16.9
GK distance	11.9
Goal Angle	21.7
Opp. in cone	2
Dist. nearest	2.1
Angle nearest	25
One-on-one	0
GK angle	1.4
Player Quality	1



Variable	Value
Goal distance	5.2
GK distance	4.8
Goal Angle	38.5
Opp. in cone	0.5
Dist. nearest	3.8
Angle nearest	119
One-on-one	1
GK angle	15
Player Quality	1

LaLiga TECH

Empower the Sports & Entertainment sector with modular technology solutions that enable clients to address the challenges and opportunities of fan, competition and content management.

Enable data-driven decisions through a digital ecosystem and specific products

CONNECT | OPTIMIZE | CONTROL | EXECUTE

LaLigaTech.com

DATA+AI SUMMIT 2022

Thank you

Rafael Zambrano

Head of Data Science, LaLiga Tech