

Soccer Analytics

when Data Science takes the field



Charles Reep (1904 - 2002)



Valeri Lobanovskyi
(1939 - 2002)

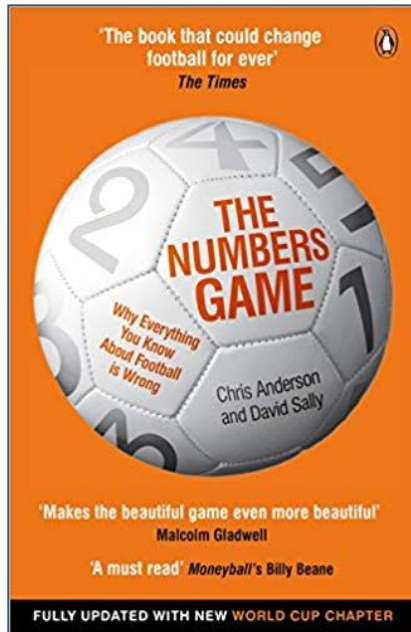




“The greatest mistake of my life”
Sir Alex Ferguson



The Maldini principle:
the better a defender
the fewer the tackles



The numbers game: why everything you know about football is wrong

C. Anderson, D. Sally

Soccer analytics: Unravelling the complexity of “the beautiful game”

Luke Bornn, Dan Cervone, Javier Fernandez

First published: 29 May 2018 | <https://doi.org/10.1111/j.1740-9713.2018.01146.x>

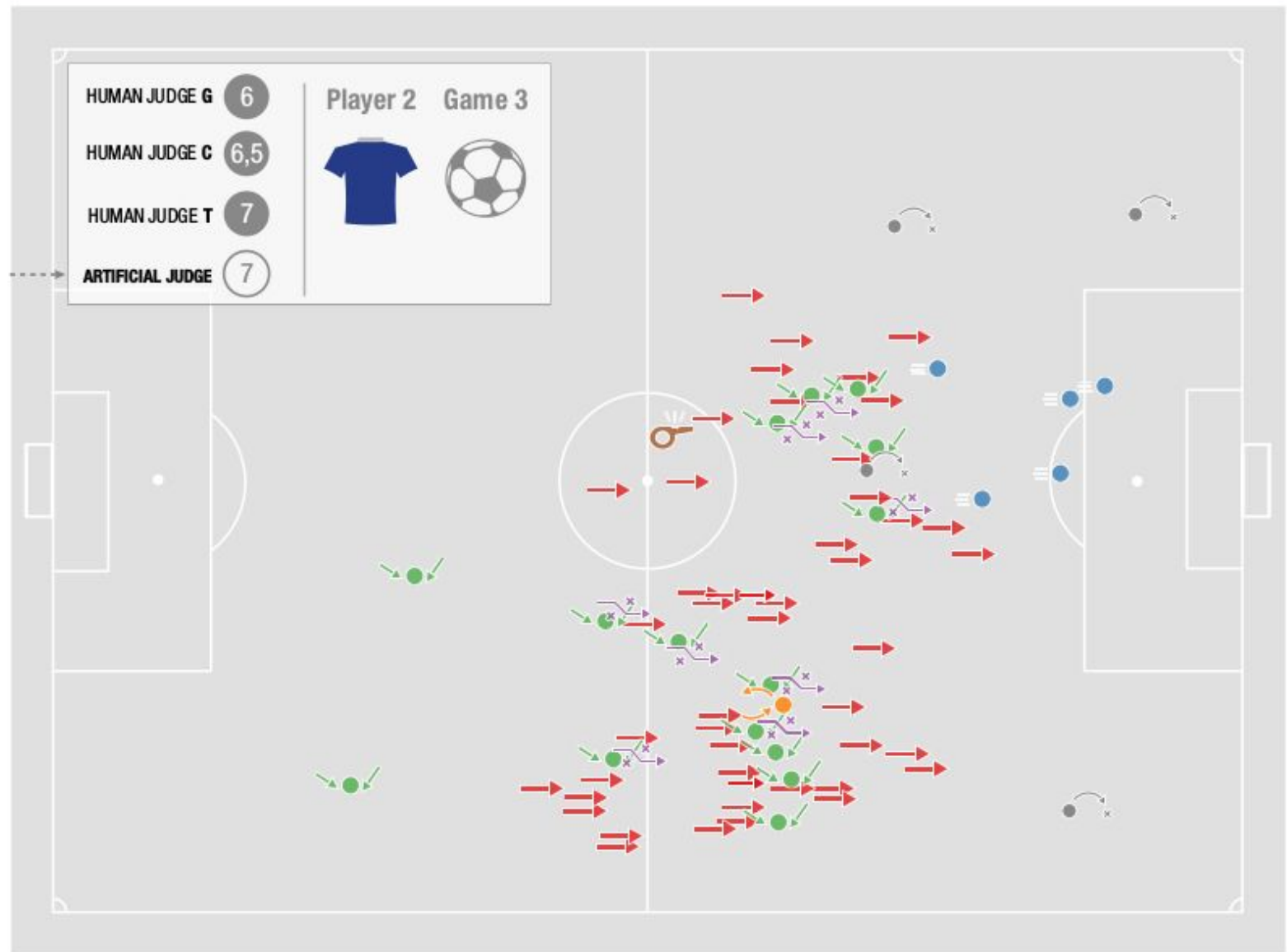
Quando il computer scese in campo

Il colonnello Lobanovski arrivò alla Dinamo Kiev nel 1973: per prima cosa chiese un computer e uno statistico. Da allora il calcio non ha potuto fare a meno di Big Data

28/06/2015

Soccer-logs

- events involving the ball occurring during a game
- player, team, position, time, outcome
- semi-automatic collection



Soccer-logs collection system

test save reset

LAZ 0-0 INT 00:11 1T

Commento
R. Trevisani - D. Adani

clinicapaideia.it PAIDEIA clinicapaideia.it clinicapaideia.it clinicapaideia.it
WORK HARD. PLAY HARDER. macron macron macron macron

87 7 11 37 1 8 9 10 17 77 25 44 33

Saves Shots Through Passes

GoalKeeper Leaving Line Key Passes 1VS1 & dribbling

Cross Accelerations

Defending Duels Smart Passes

Cornet Kick

100% +

04:10 04:20 04:30 04:40 04:50 05:00 05:10 05:20 05:30 05:40

Lazio

Internazionale

77 M. Brozović

Teams


```
{ 'eventName': 8,
  'eventSec': 8.221464,
  'id': 217097515,
  'matchId': 2576132,
  'matchPeriod': '1H',
  'playerId': 8306,
  'positions': [{ 'x': 42, 'y': 14 }, { 'x': 74, 'y':
33}],
  'subEventName': 83,
  'tags': [{ 'id': 1801 }],
  'teamId': 3158 }
```

pass

identifiers

high pass

accurate

1700 events
per match
(in average)

```
In [ ]: from bokah.widgets import CustomUI, ColumnDataSource, ModelView, RangeSlider, Select, ColorPicker, ColorPicker
from bokah.plotting import Figure, output_sandbox, show
from bkplot.io import push_sandbox

import numpy as np
import json
from sklearn import preprocessing
from bokah.layouts import widgetbox
from bokah.widgets import ColumnDataSource
from bokah.widgets.widgets import DataTable, DataFrame, TableColumn

output_sandbox()

source = ColumnDataSource(data=dict(x=[], y=[], width=[], height=[]))

> calibart = CustomWidget(dict(source=source), code='')

box_select = ModelView(callback=callback)

> p = Figure(plot_width=800, w

> view = Select('x', w

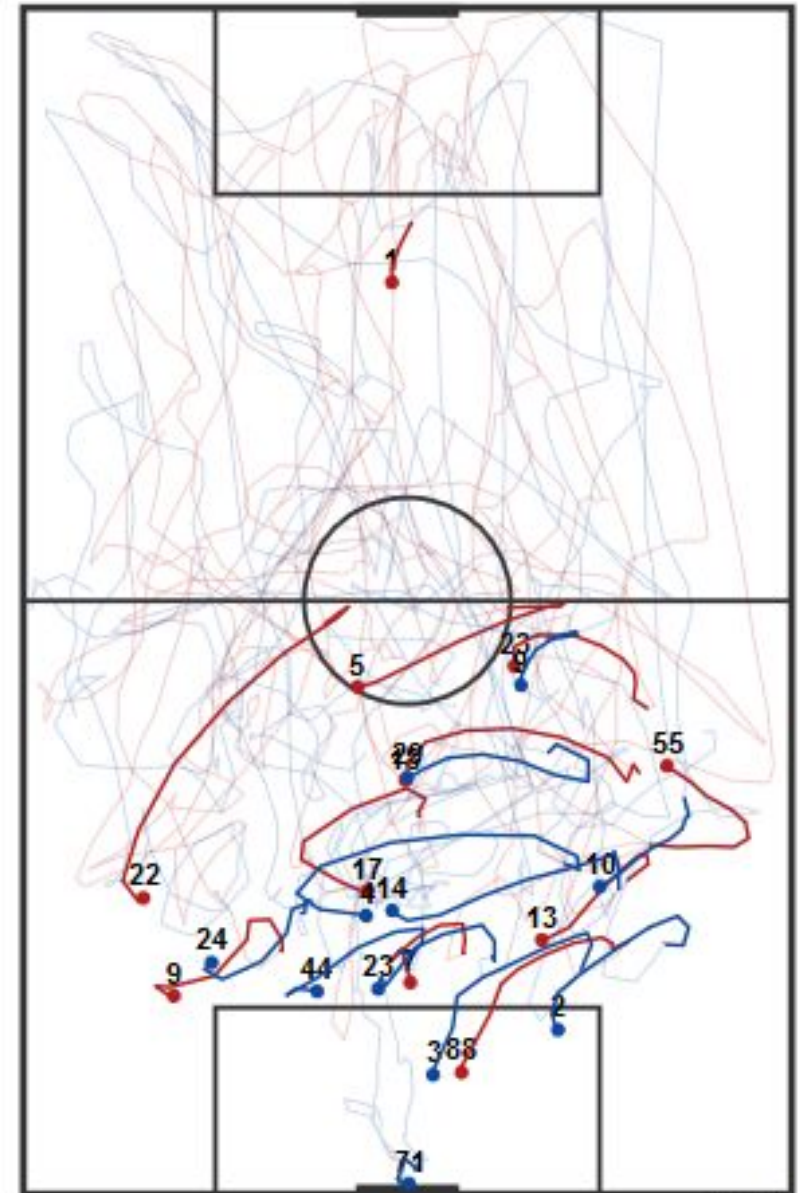
p.add_glyph(source, view, selection_glyph='x', size={'x': 100})

from bokah.widgets import CustomUI

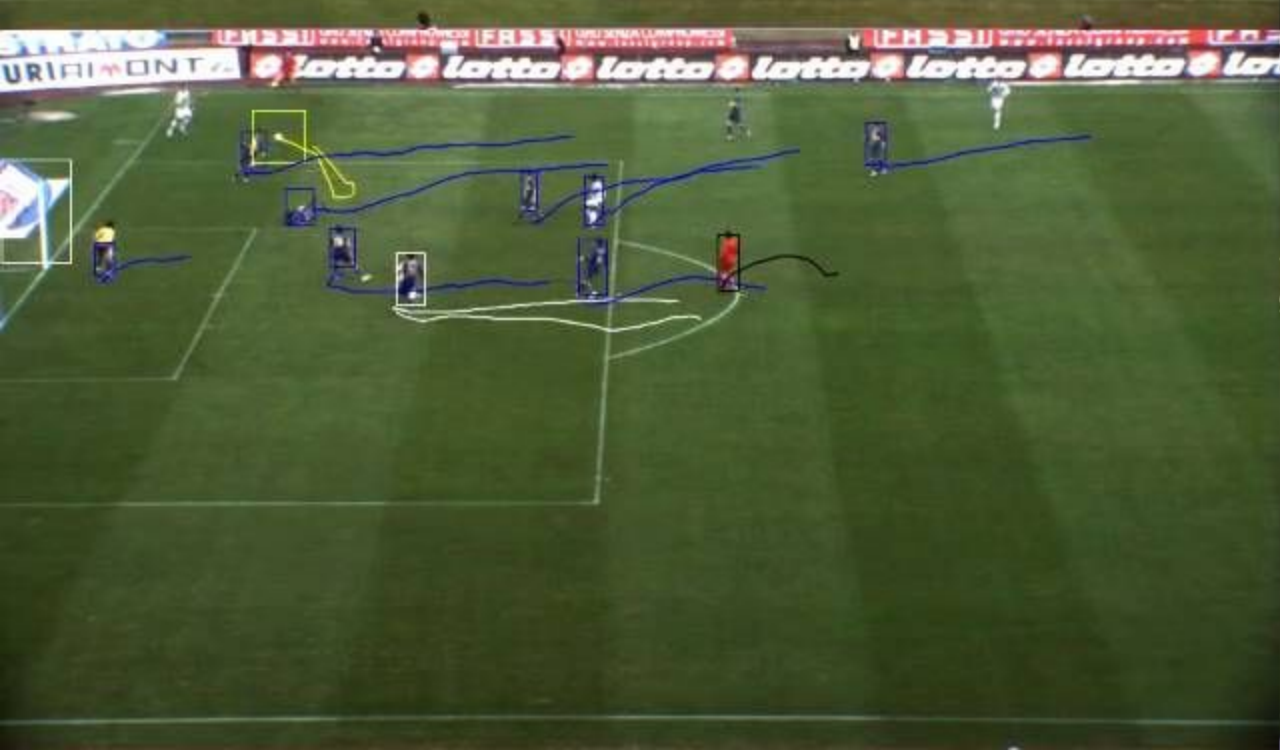
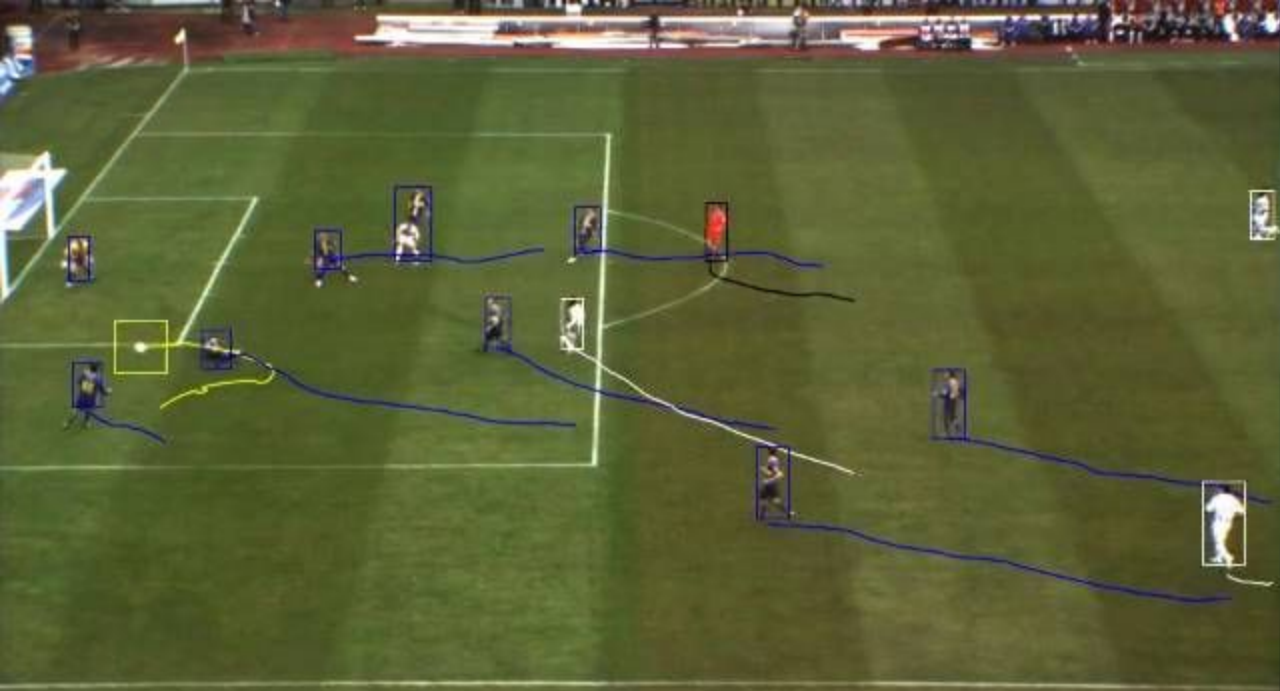
> def on(EventCallback(view, old, new):
```

Video tracking

- video-cameras are installed in stadiums
- each player identified
- trajectory of the player is inferred



01:28



Soccer Atlas



GPS devices
track training
sessions



Prediction is better than cure

using AI to predict injuries of
soccer players

📍 Sochi, Russia



2



1



FIFA WORLD CUP
RUSSIA 2018



📍 Sochi, Russia



2

1



FIFA WORLD CUP
RUSSIA 2018





188M €

in Spain

16%

days absence

Economic costs estimation of soccer injuries in first and second spanish division professional teams

http://bit.ly/cost_injuries_soccer



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



Number of injuries that players had occurred before each training session

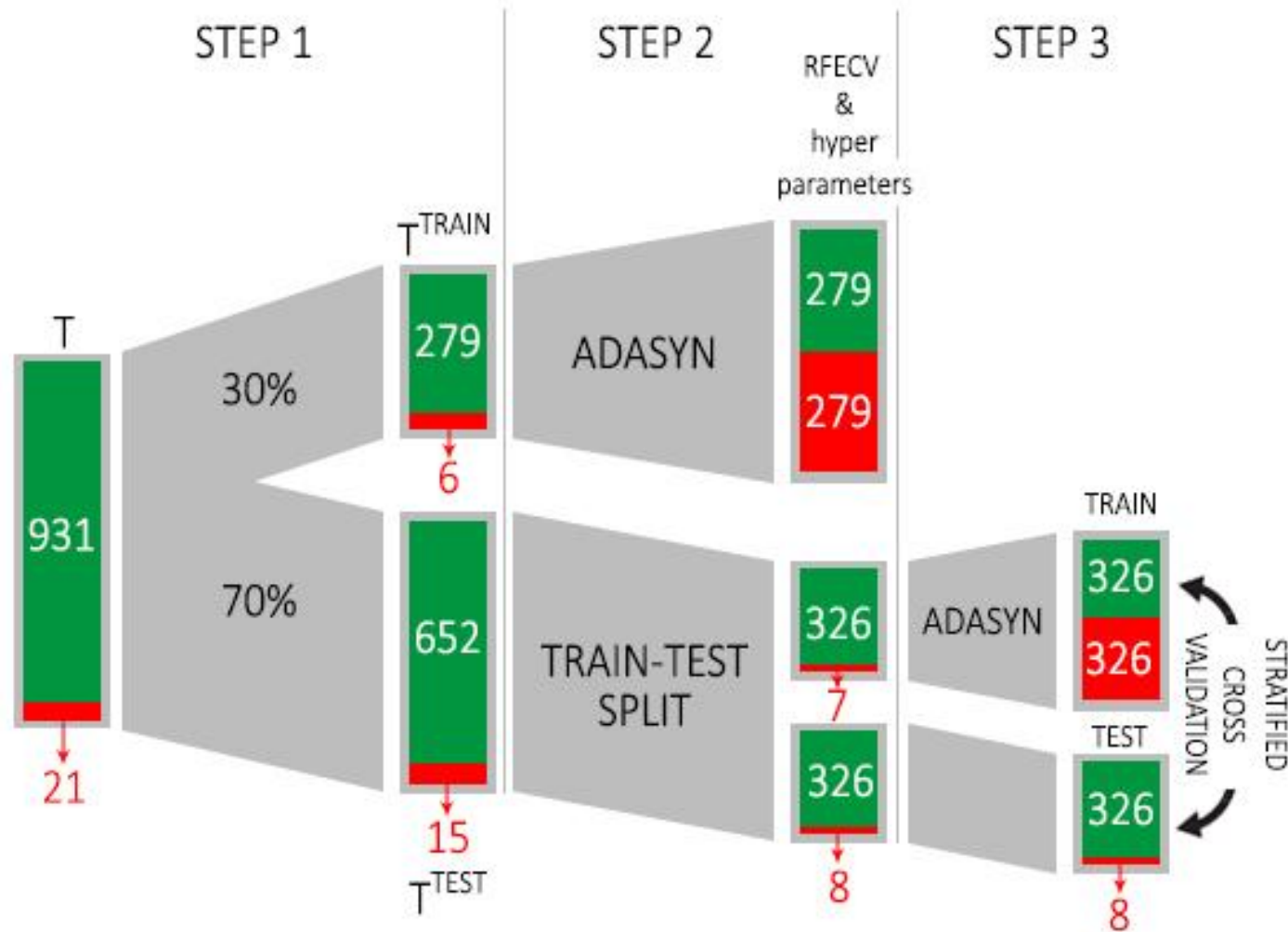
a classification problem

	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

injury examples
are very **rare**

(just 2% of the
examples)

Re-balancing the dataset



ADASYN:
a technique to rebalance
the dataset

It generates synthetic
examples of the
minority class

State of the art

ACWR =

acute workload (7 days)

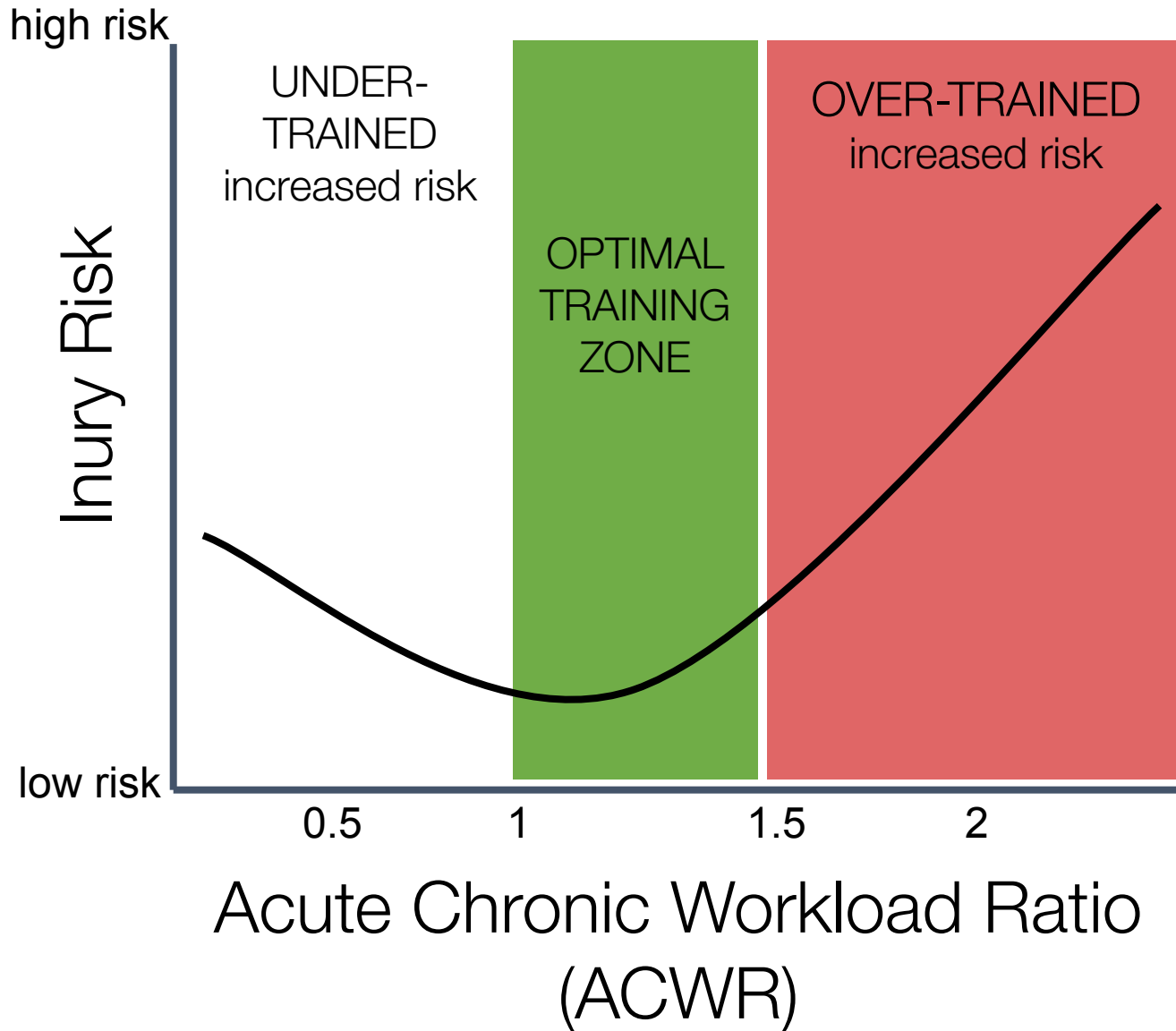
chronic workload (28 days)

Pro:

- simple to compute
- high recall

Cons:

- monodimensional
- low precision
- many false alarms



high recall

> 90

low precision

< 4%

	class	prec	rec	F1	AUC
ACWR	0	1.00	0.43	0.60	0.67
	1	0.04	0.91	0.07	
Null model	0	0.98	0.98	0.98	0.51
	1	0.06	0.05	0.05	

high recall

> 90

low precision

< 4%

Will get injured?

Yes | No

High Speed Running

Accelerations

Age, role, weight

Dynamic stress load

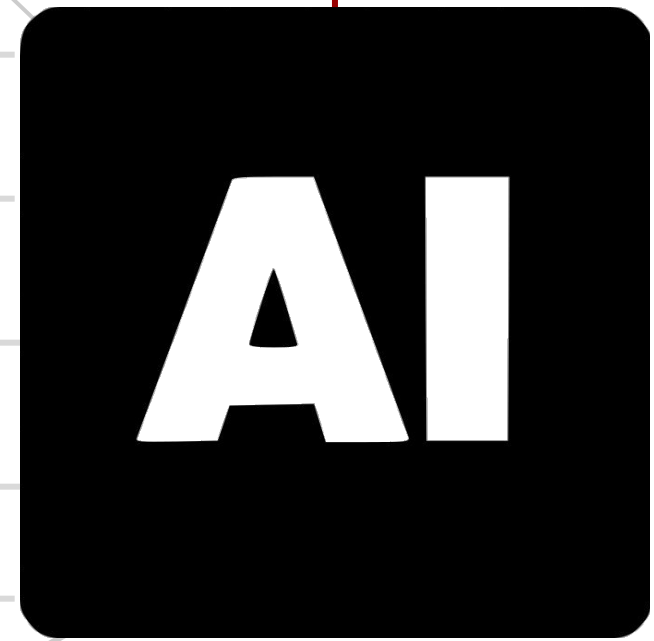
Fatigue index

Metabolic Distance

Decelerations

High Metabolic Load D.

Total distance



Injury Forecaster

infer the relation between workload variables and injury likelihood

High Speed Running

Accelerations

Age, role, weight

Dynamic stress load

Fatigue index

Metabolic Distance

Decelerations

High Metabolic Load D.

Total distance



classify a session as injury or non-injury

Will get injured?

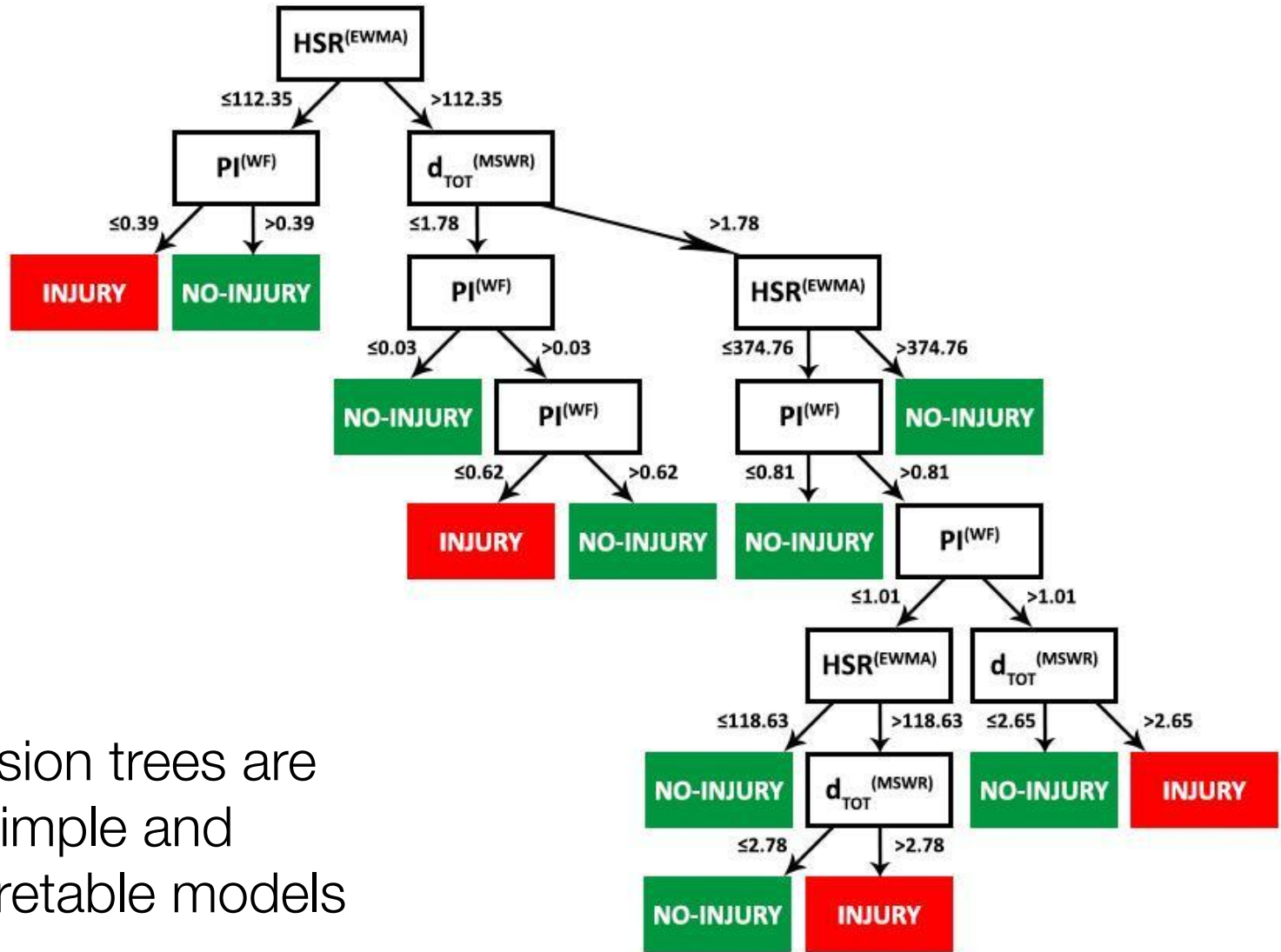
Yes

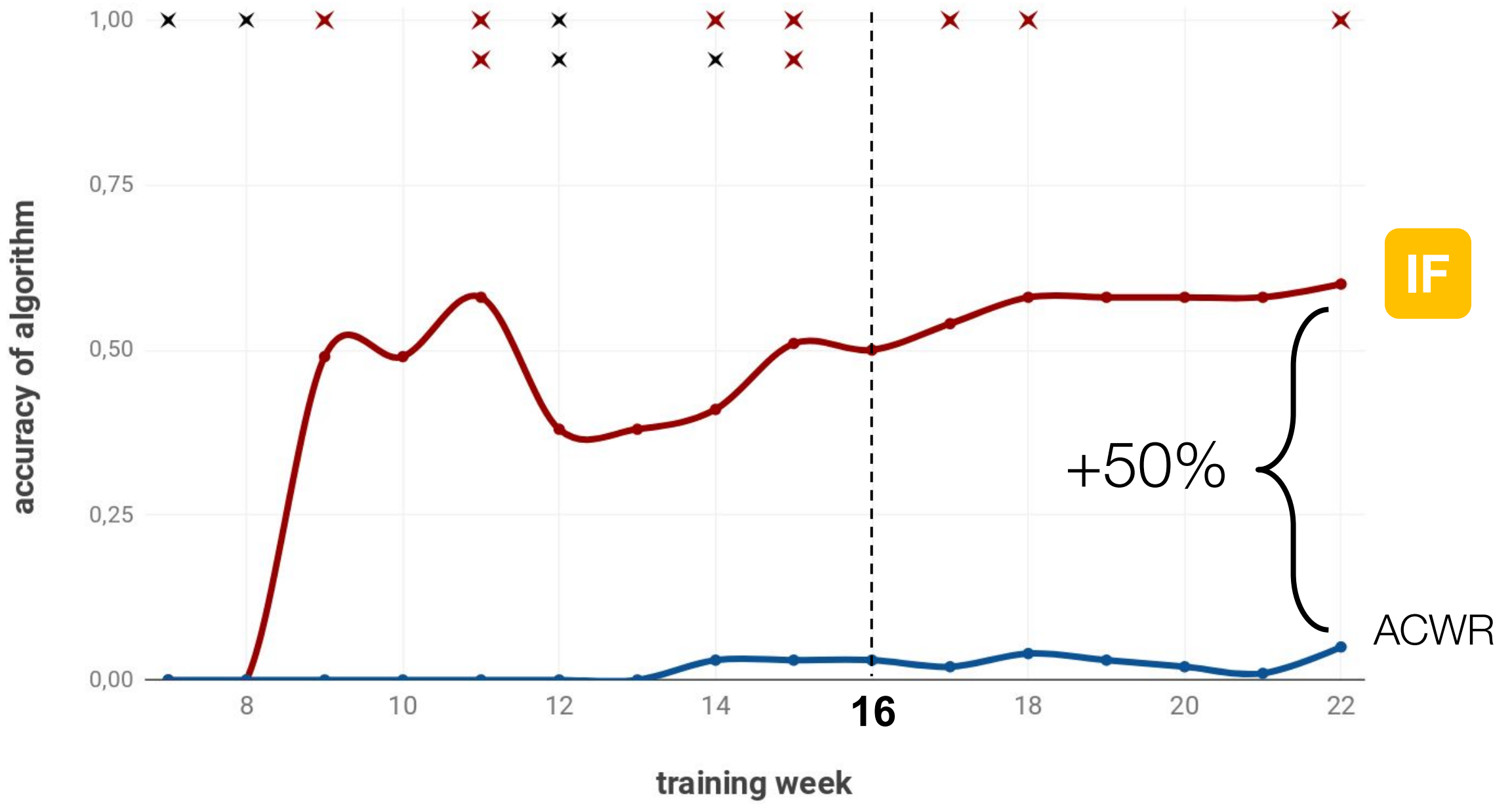
No

How IF is made?



Decision trees are simple and interpretable models



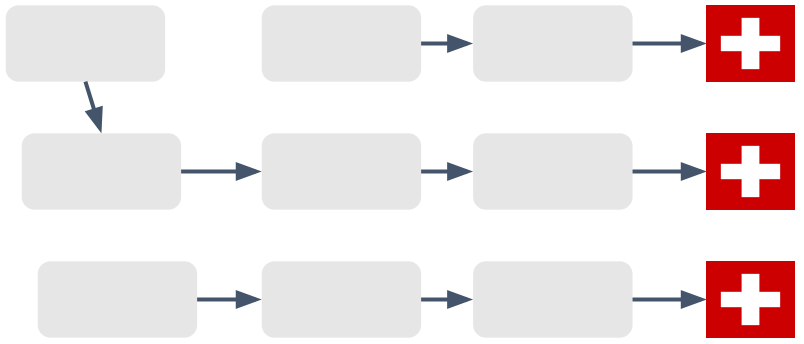
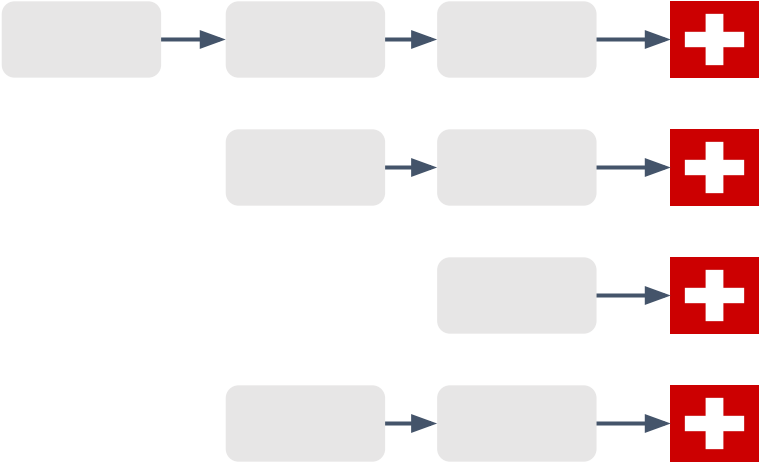


How IF is made?

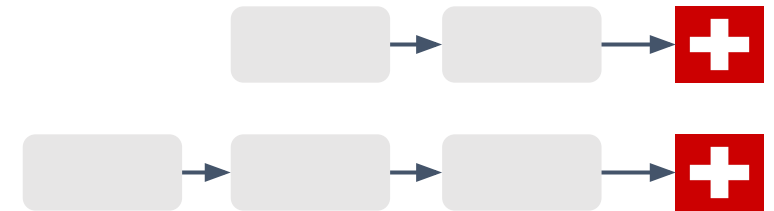
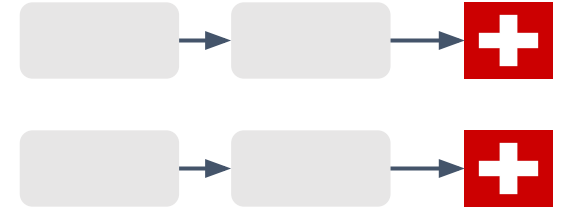


Specialized IF

JUVENTUS



Personalized IF



EL PAÍS

Un algoritmo para saber cuándo se va a lesionar un jugador

Los especialistas afirman que las soluciones tecnológicas para evitar daños en los atletas profesionales están todavía en fase embrionaria

NewScientist

DAILY NEWS 3 August 2018

Football teams secretly using AI to predict injuries before they occur

Stop agli infortuni e mercato al top Con due algoritmi cambia il futuro

● Cnr e Università di Pisa hanno creato due sistemi di intelligenza artificiale

La Gazzetta dello Sport

WIRED

L'intelligenza artificiale aiuta i calciatori a evitare gli infortuni

Un algoritmo calcola con una precisione del 50% il rischio per il singolo calciatore di farsi male nell'allenamento successivo. Le stime attuali hanno una precisione solo del 4%. Alla ricerca partecipano Università di Pisa, Cnr e Università di Milano, nonché alcuni calciatori

Il Sole
24 ORE

SPORT E DATA ANALYTICS

Infortuni previsti con precisione: arriva la manutenzione predittiva del calciatore



Effective injury forecasting in soccer with
GPS training data and machine learning

http://bit.ly/plosone_injury

Phases of the project

1. **Motivate your proposal:** find material demonstrating the importance of your proposal;
2. **State of the art:** search for existing solutions
3. **Define:** formalize your problem in terms of predictive task
4. **Extract information** extract meaningful features
5. **Implement:** realize your solution using the most suitable technique
6. **Evaluate:** evaluate the quality of your solution
7. **Interpret:** interpret your model to extract new knowledge