LEUPHANA
UNIVERSITÄT LÜNEBURG

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Feature Extraction and Aggregation for Predicting the Euro 2016

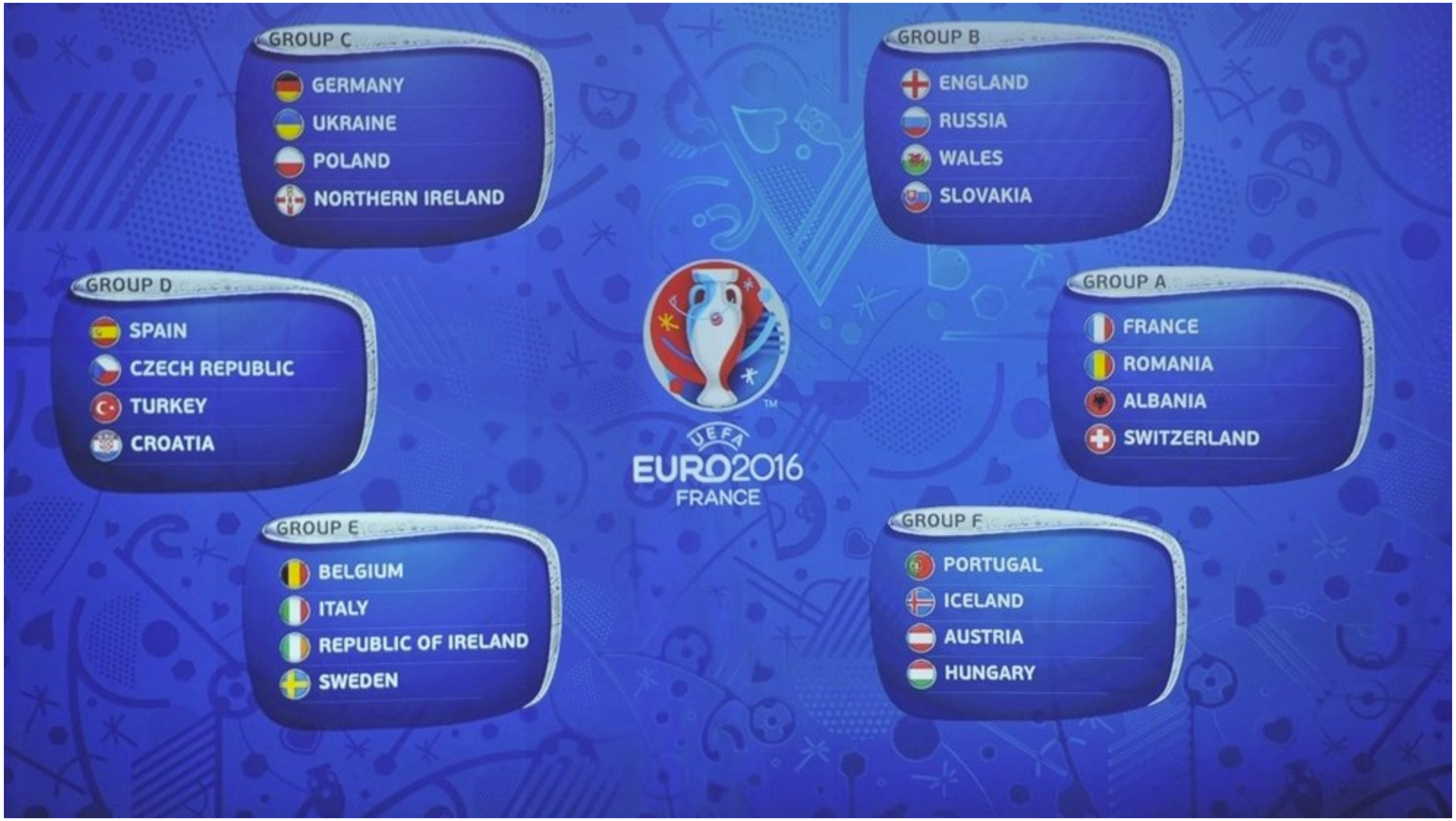
Maryam Tavakol
Hamid Zafartavanaelmi, and Ulf Brefeld

Riva del Garda, Sep 19, 2016

Agenda

- Introduction
- Feature Extraction
- Prediction & Learning
- Performance Analysis
- Summary

Introduction



GROUP C
- GERMANY
- UKRAINE
- POLAND
- NORTHERN IRELAND

GROUP B
- ENGLAND
- RUSSIA
- WALES
- SLOVAKIA

GROUP D
- SPAIN
- CZECH REPUBLIC
- TURKEY
- CROATIA

GROUP A
- FRANCE
- ROMANIA
- ALBANIA
- SWITZERLAND

GROUP E
- BELGIUM
- ITALY
- REPUBLIC OF IRELAND
- SWEDEN

GROUP F
- PORTUGAL
- ICELAND
- AUSTRIA
- HUNGARY

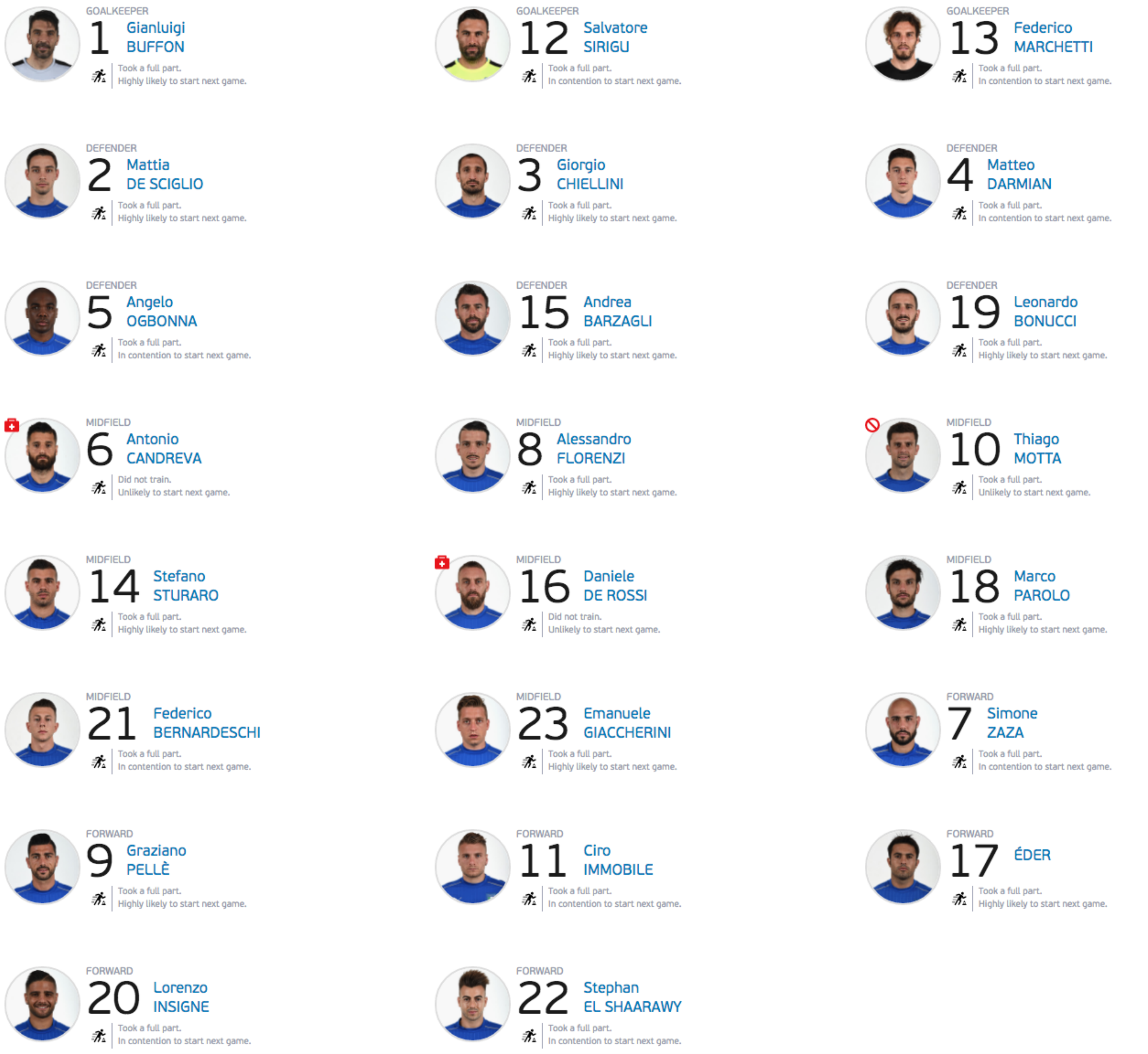
Feature Extraction

- Based on available data from the past tournaments
- General country data
  - FIFA ranking, FIFA points, UEFA ranking, etc.
  - Normalising features using ***min*** and ***max*** rescaling —keep the order

Feature Extraction

- Player specific data
    - Market value, age, num of matches/goals, etc.
    - Obtaining the current squads
    - Goal/play ratio —host advantage for France
    - Averaging for all players of a team
    - Normalising features using ***min*** and ***max*** rescaling

Add a New Feature



GOALKEEPER
1 Gianluigi
BUFFON
Took a full part.
Highly likely to start next game.

GOALKEEPER
12 Salvatore
SIRIGU
Took a full part.
In contention to start next game.

GOALKEEPER
13 Federico
MARCHETTI
Took a full part.
In contention to start next game.

DEFENDER
2 Mattia
DE SCIGLIO
Took a full part.
Highly likely to start next game.

DEFENDER
3 Giorgio
CHIELLINI
Took a full part.
Highly likely to start next game.

DEFENDER
4 Matteo
DARMIAN
Took a full part.
In contention to start next game.

DEFENDER
5 Angelo
OGBONNA
Took a full part.
In contention to start next game.

DEFENDER
15 Andrea
BARZAGLI
Took a full part.
Highly likely to start next game.

DEFENDER
19 Leonardo
BONUCCI
Took a full part.
Highly likely to start next game.

MIDFIELD
6 Antonio
CANDREVA
Did not train.
Unlikely to start next game.

MIDFIELD
8 Alessandro
FLORENZI
Took a full part.
Highly likely to start next game.

MIDFIELD
10 Thiago
MOTTA
Took a full part.
Unlikely to start next game.

MIDFIELD
14 Stefano
STURARO
Took a full part.
Highly likely to start next game.

MIDFIELD
16 Daniele
DE ROSSI
Did not train.
Unlikely to start next game.

MIDFIELD
18 Marco
PAROLO
Took a full part.
Highly likely to start next game.

MIDFIELD
21 Federico
BERNARDESCHI
Took a full part.
In contention to start next game.

MIDFIELD
23 Emanuele
GIACCHERINI
Took a full part.
Highly likely to start next game.

FORWARD
7 Simone
ZAZA
Took a full part.
In contention to start next game.

FORWARD
9 Graziano
PELLÈ
Took a full part.
Highly likely to start next game.

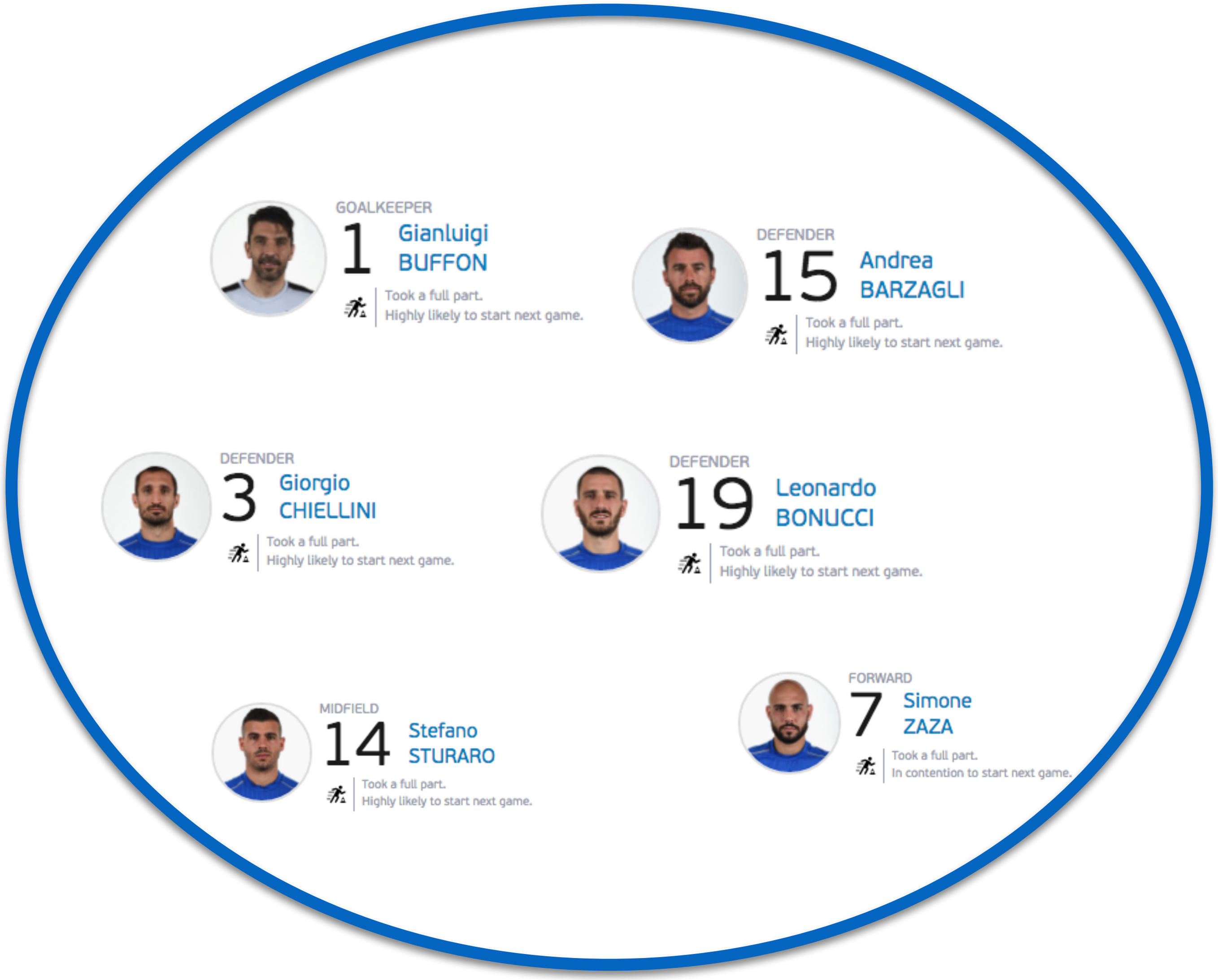
FORWARD
11 Ciro
IMMOBILE
Took a full part.
In contention to start next game.

FORWARD
17 ÉDER
Took a full part.
Highly likely to start next game.

FORWARD
20 Lorenzo
INSIGNE
Took a full part.
In contention to start next game.

FORWARD
22 Stephan
EL SHAARAWY
Took a full part.
In contention to start next game.

Club Division



GOALKEEPER
1 Gianluigi
BUFFON
Took a full part.
Highly likely to start next game.

DEFENDER
15 Andrea
BARZAGLI
Took a full part.
Highly likely to start next game.

DEFENDER
3 Giorgio
CHIELLINI
Took a full part.
Highly likely to start next game.

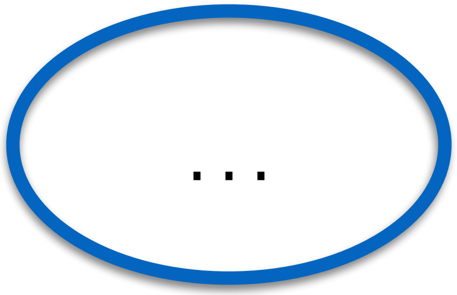
DEFENDER
19 Leonardo
BONUCCI
Took a full part.
Highly likely to start next game.

MIDFIELD
14 Stefano
STURARO
Took a full part.
Highly likely to start next game.

FORWARD
7 Simone
ZAZA
Took a full part.
In contention to start next game.

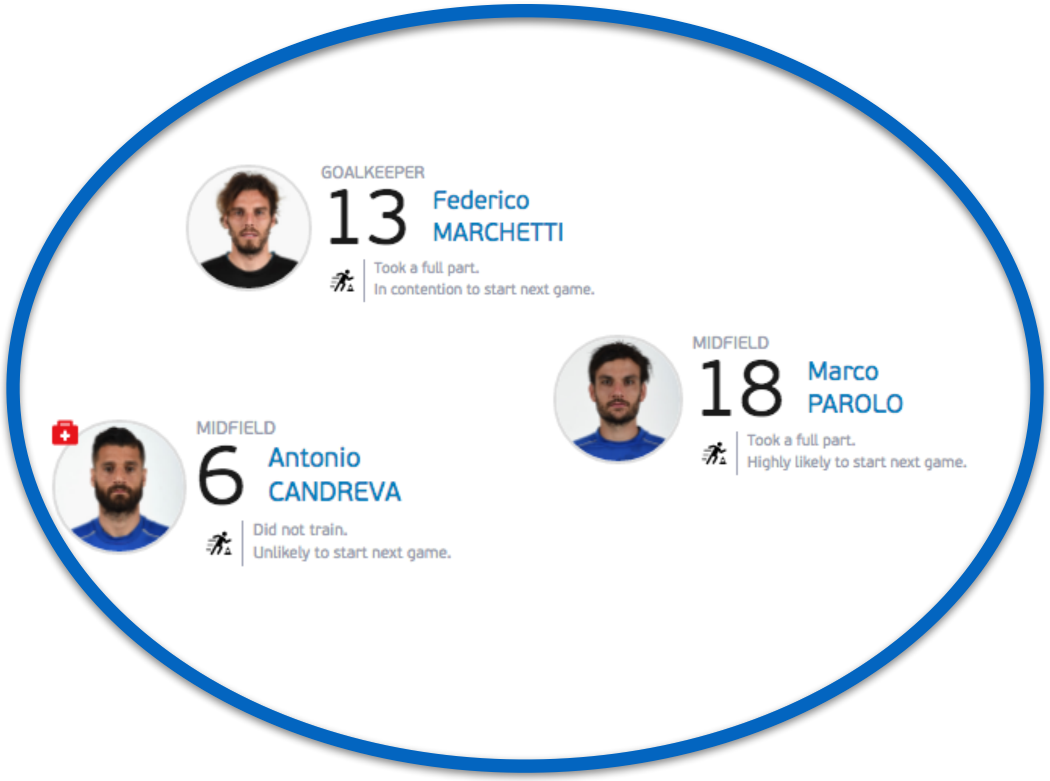
Juventus

Club rank = 2



Lazio

Club rank = 212



GOALKEEPER
13 Federico
MARCHETTI
Took a full part.
In contention to start next game.

MIDFIELD
18 Marco
PAROLO
Took a full part.
Highly likely to start next game.

MIDFIELD
6 Antonio
CANDREVA
Did not train.
Unlikely to start next game.

Team-Club Harmony

Country	Num of Players	Club	Club Rank
Spain	5	Barcelona	1
Italy	6	Juventus	2
France	2	Juventus	2
Germany	5	Bayern Munich	4
Belgium	3	Liverpool	42
Poland	3	Legia	52
Portugal	4	Sporting CP	179
Wales	3	Crystal Palace	0*
Iceland	2	Hammarby	0*

(Normalised Club rank) x (num of players)

Prediction

- A score per country is defined as a weighted sum of features, i.e., linear function

$$s_i = \boldsymbol{\theta}_i^\top \mathbf{x}_i$$

- The probabilities are computed based on obtained scores

Prediction

Win probability for team *i*

```
if $s_i \geq s_j$:
    $P_{w_i} = \frac{s_i}{(s_i+s_j)}$
    $P_{w_j} = (1 - P_{w_i}) * s_j = P_{l_i}$
else:
    $P_{w_j} = \frac{s_j}{(s_i+s_j)}$
    $P_{w_i} = (1 - P_{w_j}) * s_i = P_{l_j}$
$P_d = 1 - P_{w_i} - P_{w_j}$
```

Probability of draw

Lose probability for team *j*

Learning

- Capture the outcome probabilities from the head to head record of pair of countries
  - **Germany** vs. **France**: 27 times
  - 10 win for **Germany**, 12 for **France** and 5 draw

$$p_{w_G} = \frac{10}{27}, p_{w_F} = \frac{12}{27}, p_d = \frac{5}{27}$$

Learning

- Converting probabilities to scores
- Obtaining parameters from the closed form solution of ridge regression problem

$$\hat{\boldsymbol{\theta}} = (X^\top X + I)^{-1} X^\top \hat{\mathbf{s}}$$

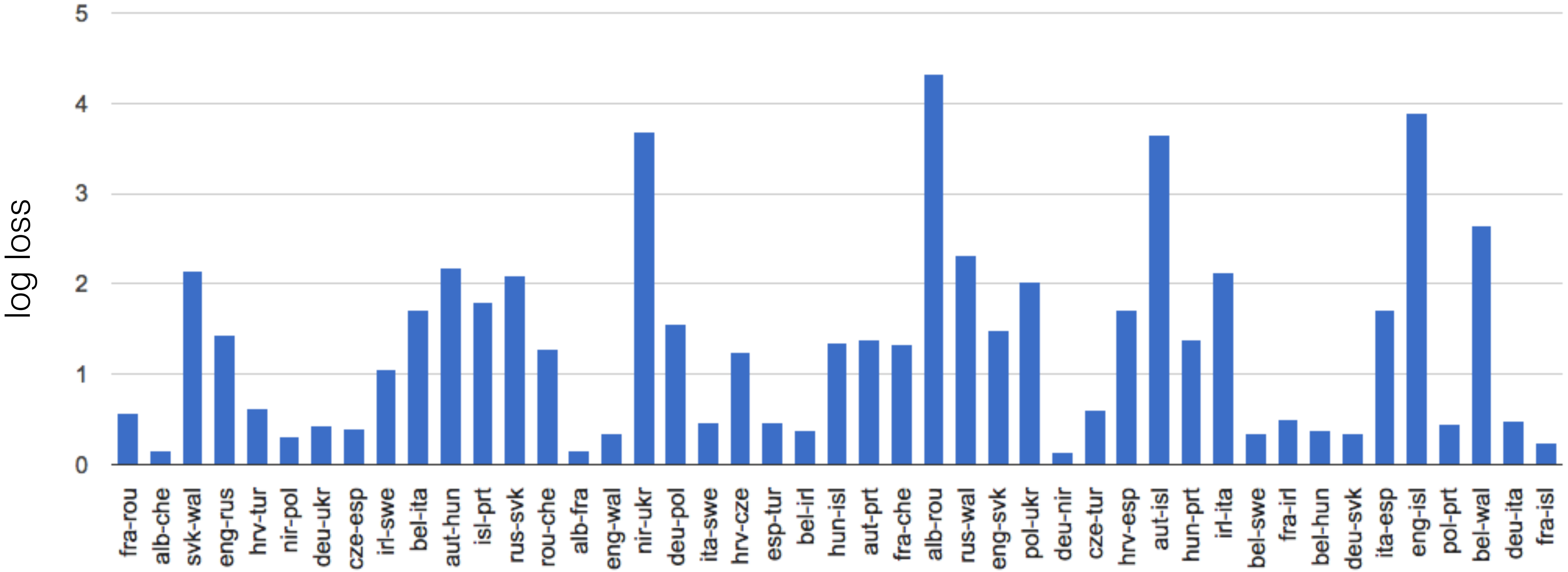
Performance Analysis

- Compare prediction results to actual tournament outcome
  - Until Quarter-Final (QF)
- Evaluation by multi class logarithmic loss

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * log(p_{ij})$$

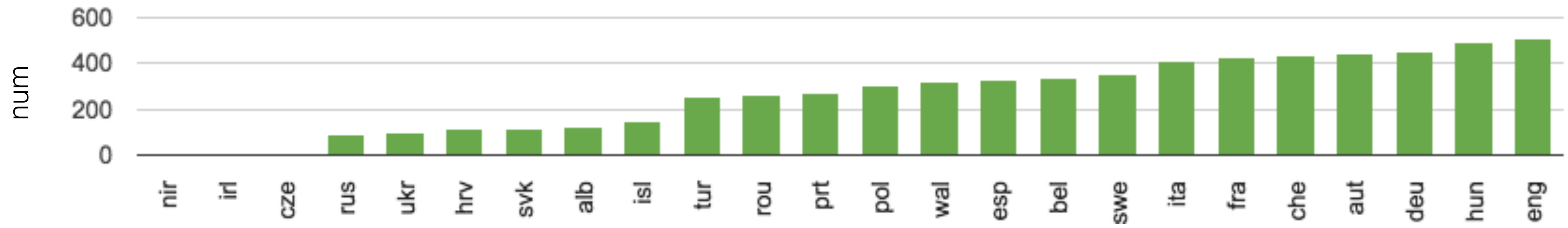
Overal Performance

- Error of prediction for 45 matches before QF
  - Average error: **1.3187**

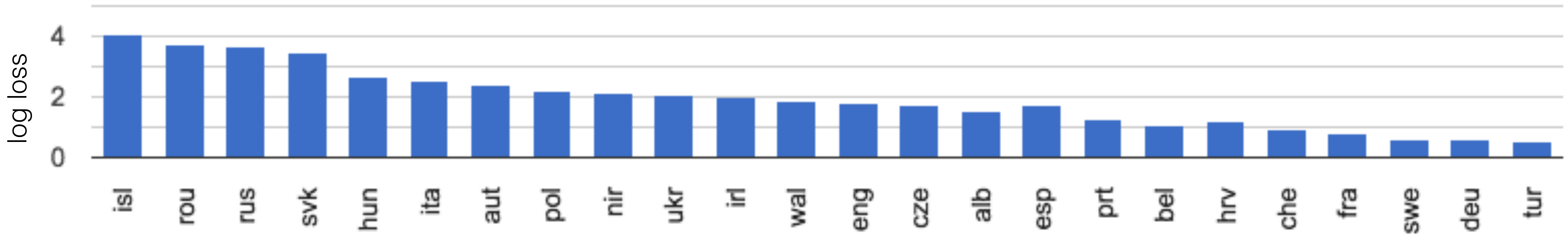


Insufficient Data

- Relation of performance with amount of historical data



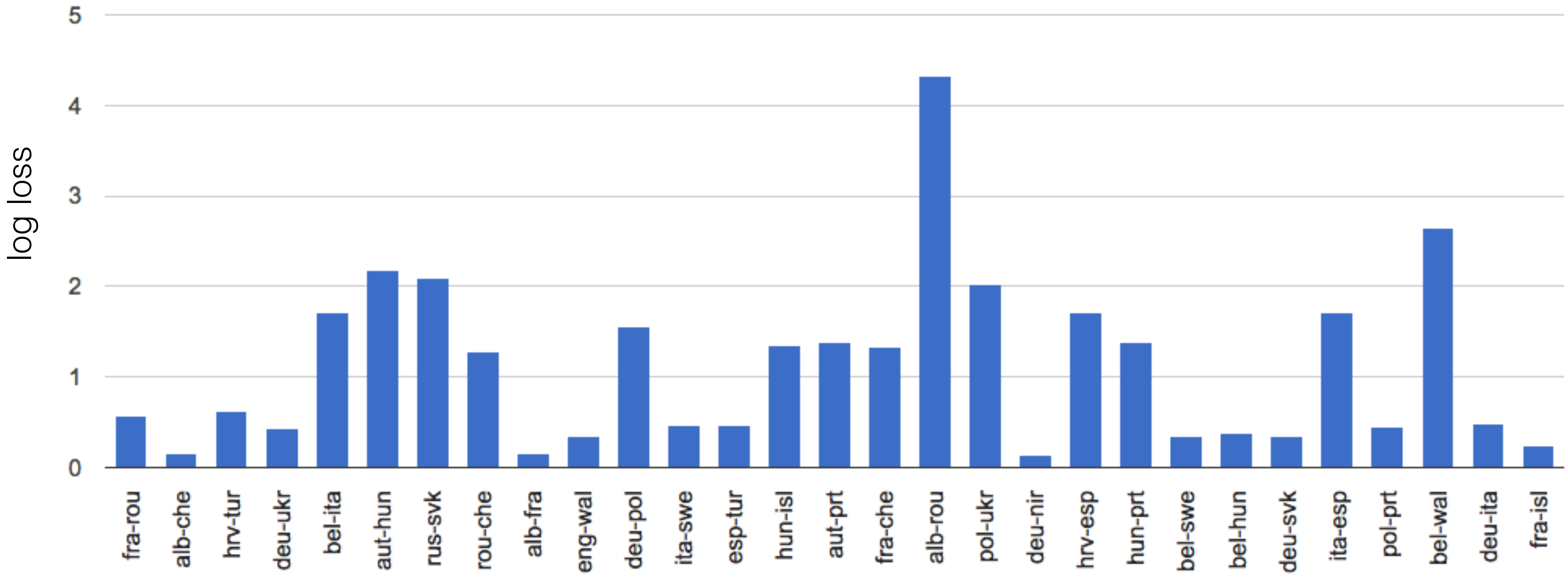
Num of historical data



Error per country

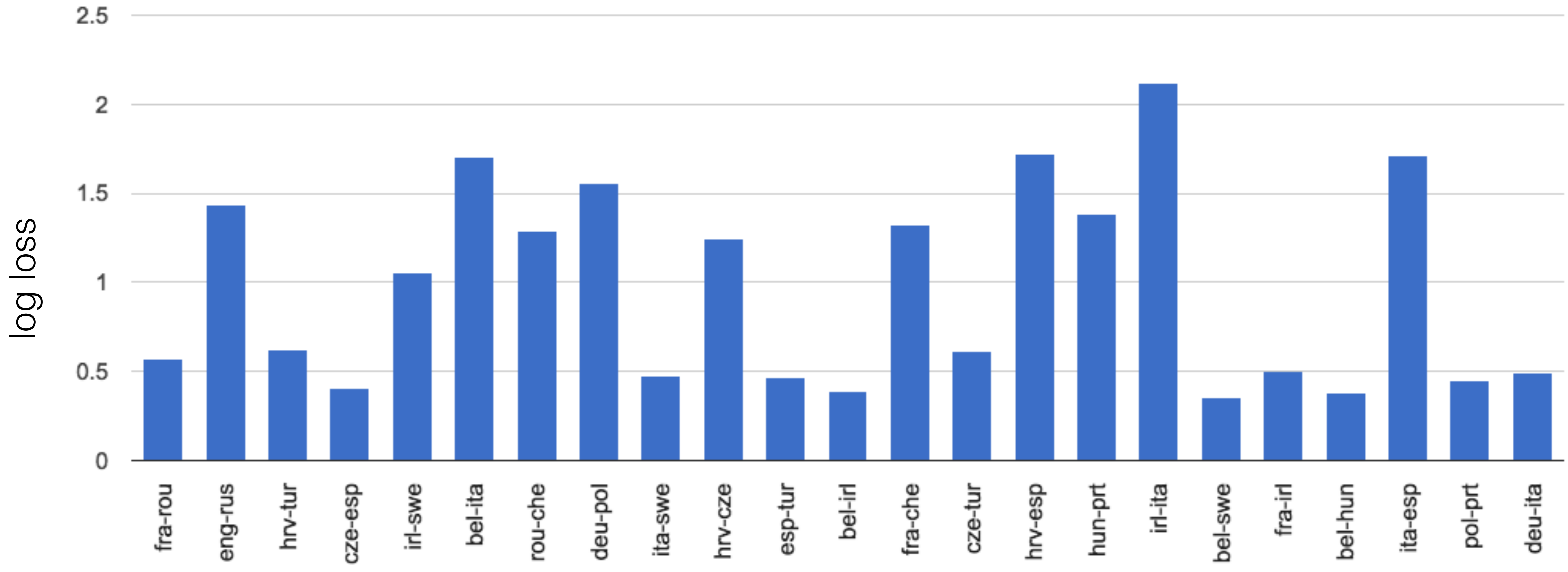
Sufficient Data

- Reduction of error from **1.3187** to **1.1129** for teams with more than 4 historical records



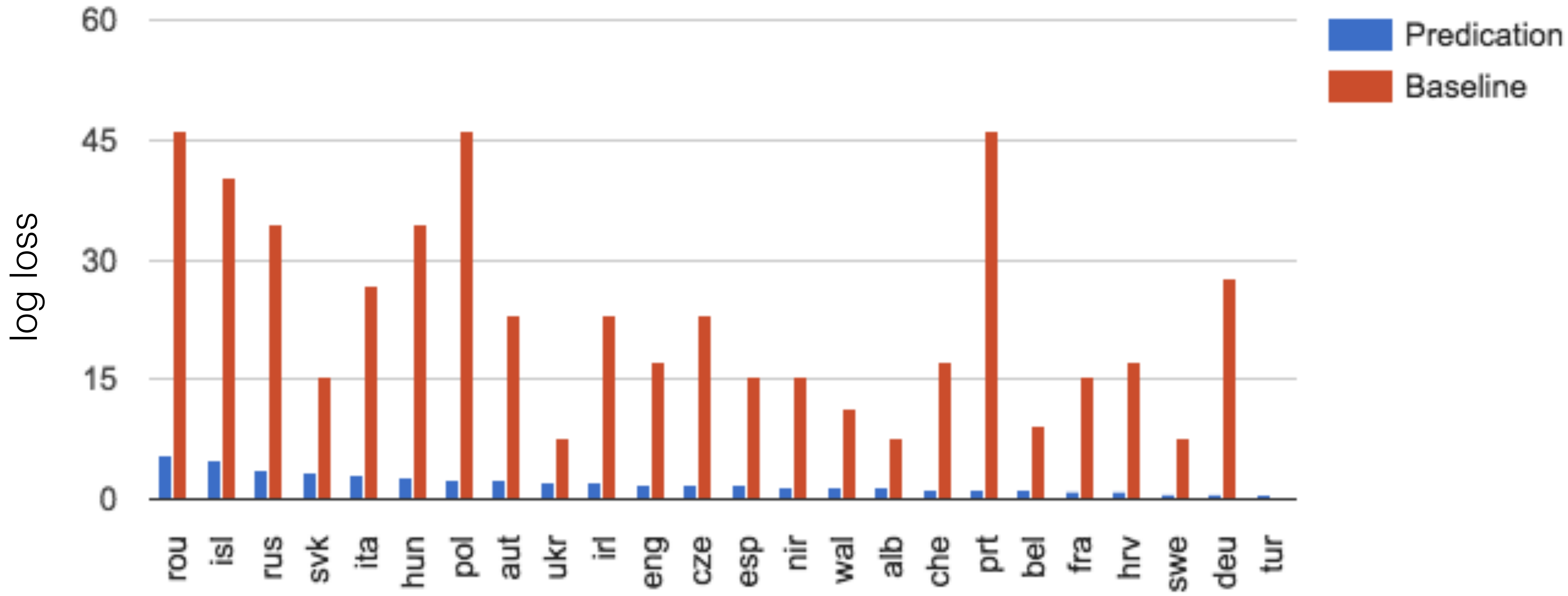
Role of Past Euros

- Eliminating teams with less than 2 appearance in past Euro cups, error: **0.9680**



Baseline

- Compare to a simple baseline (based on FIFA ranking only)



Summary

- Collecting data
- Feature extracting/cleaning
- New feature: team-club harmony
- Learn a linear model
- Effect of historical data on the performance

Thanks for your attention

Questions?

Email: tavakol@leuphana.de