Technische Universität Berlin

# World Cup Prediction

Milad Abbaszadeh & Malek Trabelsi

Faculty IV | Data Science Project | 13.02.2019

# Agenda

1. Problem Statement
2. Challenges
3. Datasets
4. Experiments
5. Results & Interpretations
6. Demo
7. Conclusion & Future Work

# Problem Statement

**Predict the outcome of a match given previous data**

# Challenges

- Modelling the Dataset
- Data Collection (only FIFA ?)
- Data Integration (Different datasets)
- Model: Classification VS Regression
- Feature Engineering

## Challenges

Names from one dataset as **reference**

DR Congo → Congo

Northern Ireland → Ireland

Dominican Republic → Dominica

England → United Kingdom

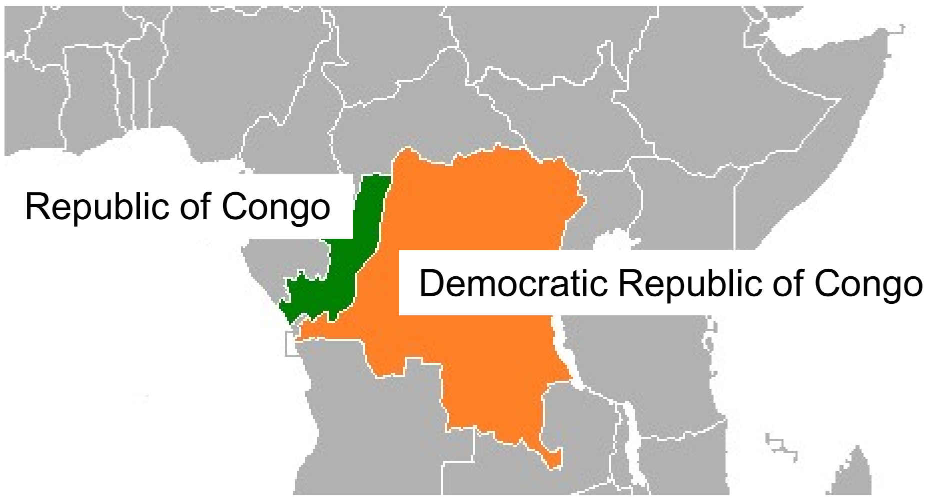
# Challenges

**DR Congo → Congo**

Northern Ireland → Ireland

Dominican Republic → Dominica

England → United Kingdom




https://fr.wikipedia.org/wiki/Congo

## Challenges

DR Congo → Congo

**Northern Ireland → Ireland**

Dominican Republic → Dominica

England → United Kingdom

| | Team1 | Team2 | Score1 | Score2 | Date/Time |
|---|---|---|---|---|---|
| 50 | Ireland | Northern Ireland | 0 | 1 | 29.05.1999/00:00 |
| 4137 | Ireland | Northern Ireland | 5 | 0 | 24.05.2011/20:45 |
| 8633 | Ireland | Northern Ireland | 0 | 0 | 15.11.2018/20:45 |

# Challenges

DR Congo → Congo

Northern Ireland → Ireland

**Dominican Republic → Dominica**

England → United Kingdom


https://www.quora.com/Whats-the-difference-between-Dominica-and-the-Dominican-Republic

## Challenges

DR Congo → Congo

Northern Ireland → Ireland

Dominican Republic → Dominica

**England → United Kingdom**


https://www.pinterest.com/pin/380343131002611548/

# Challenges

DR Congo → Congo

Northern Ireland → Ireland

Dominican Republic → Dominica

England → United Kingdom

**Serbia and Montenegro → Yugoslavia**

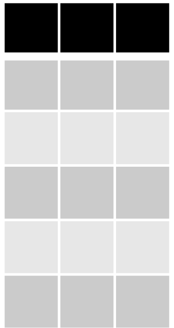
# Datasets

wide dataset → **Features**

data from Friendly & World Cup matches

from 1994

9071 * 40



long dataset → **Observations**

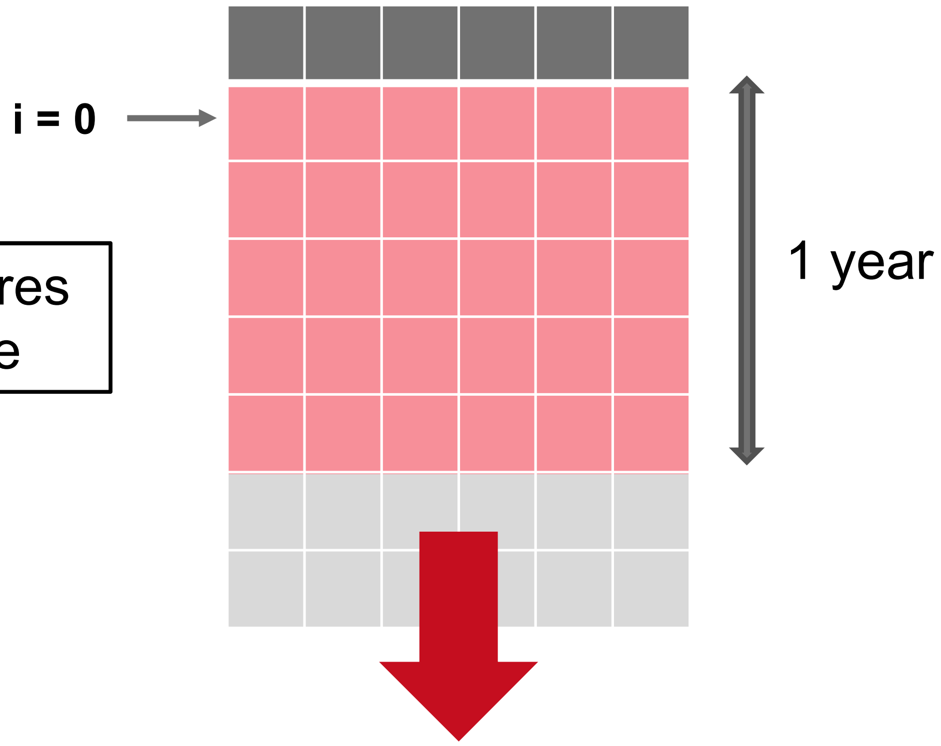
data from many **International tournaments**
(Friendly, World Cup, African Cup of Nations,
World Cup Quanlifications …)

From 1872

38900 * 30

# Datasets

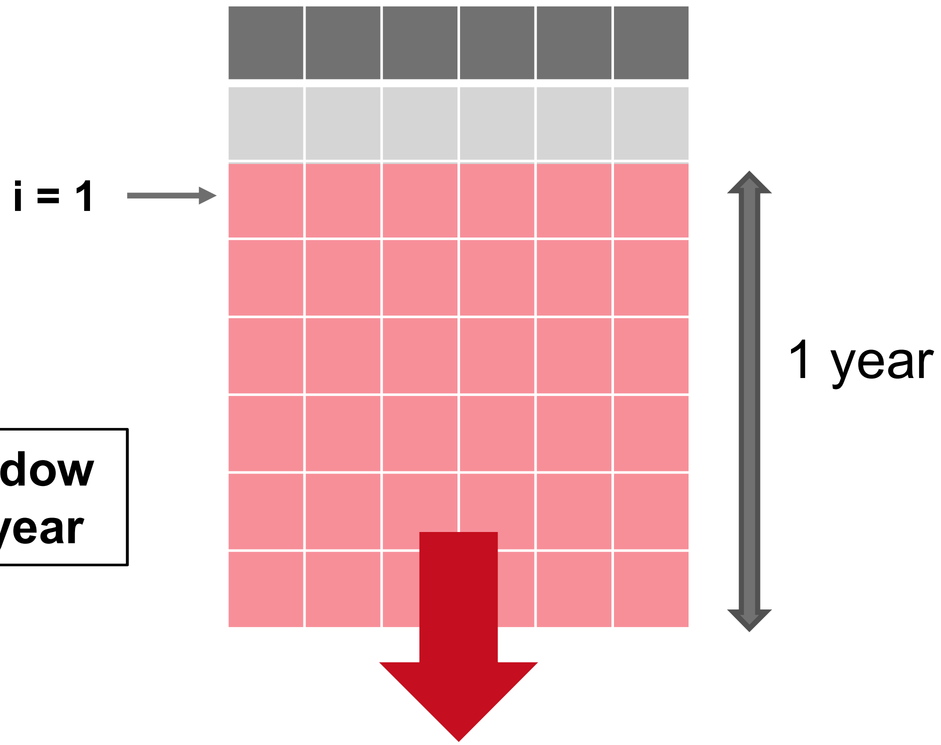
Teams + Scores
+ Date/Time




**nb matches**
**nb & ratio goals**
**nb & ratio wins / losses / draws**
**nb & ratio win / losses / draws against opponent**

Technische Universität Berlin

# Datasets




**Sliding Window of width 1 year**

**nb matches**
**nb & ratio goals**
**nb & ratio wins / losses / draws**
**nb & ratio win / losses / draws against opponent**

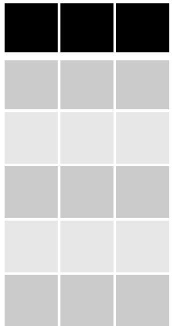
# Datasets

wide dataset → **Features**

data from Friendly & World Cup matches

from 1994

**+ FIFA Score, FIFA Ranking, Population, Surface, Density**

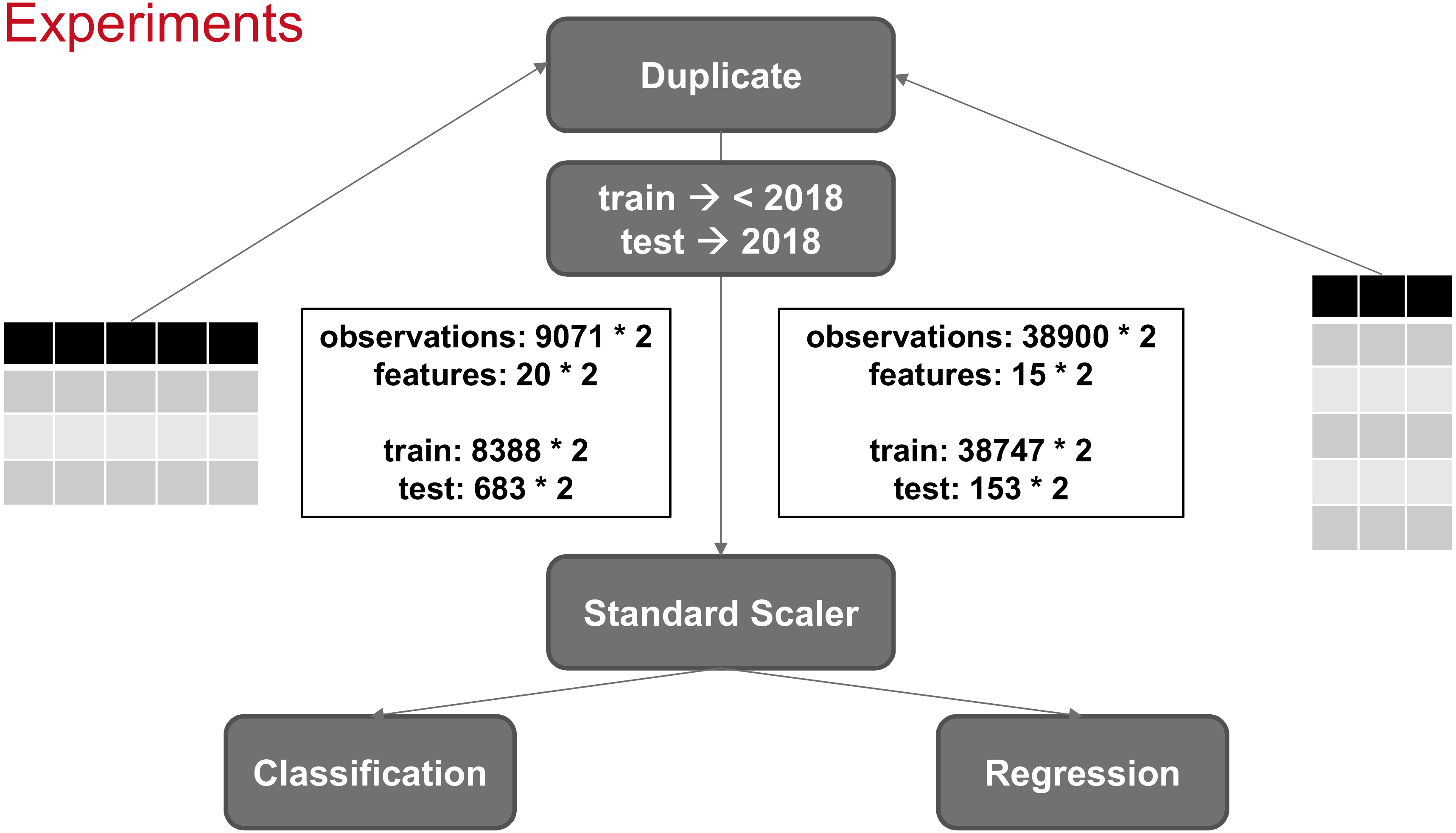


long dataset → **Observations**

data from many **International tournaments** (Friendly, World Cup, African Cup of Nations, World Cup Quanlifications …)

From 1872

# Experiments



# Results (Classification)

| Classifier | Accuracy Score |
|---|---|
| Dummy Classifier | 36,82 % |
| **Random Forest** | **48,17 %** |
| Bernoulli NB | 46,92 % |
| Extra Trees | 41,58 % |
| KNN | 39 % |
| MLP | 44,66 % |
| **Nearest Centroid** | **48,61 %** |
| **Ridge Classifier** | **48,76 %** |
| **SVC** | **49,04 %** |

| Classifier | Accuracy Score |
|---|---|
| Dummy Classifier | 36,27 % |
| Random Forest | 36,6 % |
| Bernoulli NB | 41,5 % |
| Extra Trees | 38,56 % |
| KNN | 34,64 % |
| **MLP** | **45,42 %** |
| Nearest Centroid | 39,22 % |
| **Ridge Classifier** | **41,18 %** |
| **SVC** | **42,48 %** |

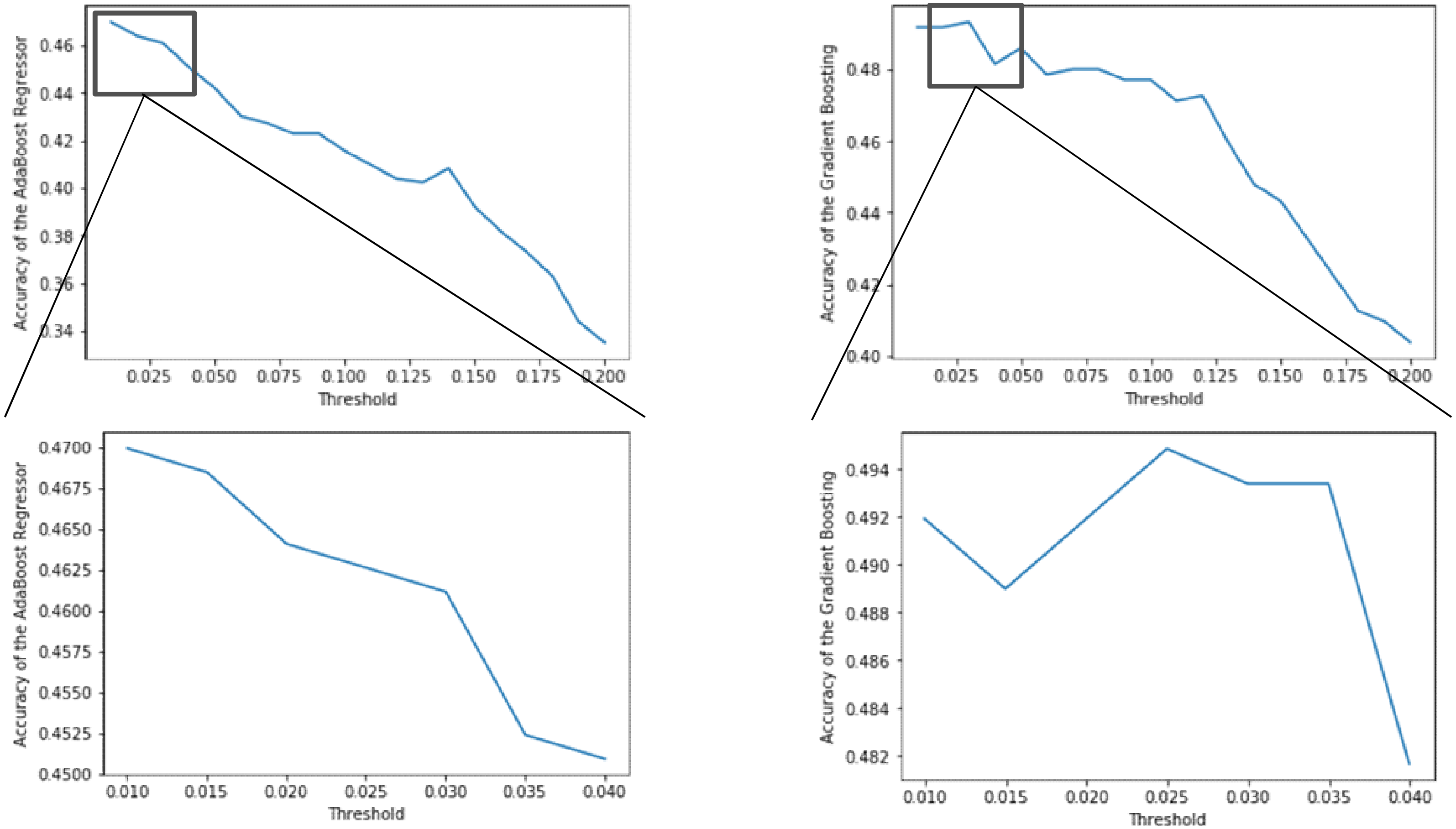
# Results (Regression)

| Regressor | Accuracy Score | Accuracy (threshold = 0,03) |
|---|---|---|
| MLP Regressor | 45,82 % | 46,41 % |
| **Gradient Boosting** | **48,76 %** | **49,34 %** |
| Random Forest | 45,24 % | 42,17 % |
| AdaBoost | 48,02 % | 43,19 % |
| Bagging Regressor | 44,22 % | 45,39 % |
| **Transformed Target** | **48,90 %** | **48,02 %** |

# Results (Regression)

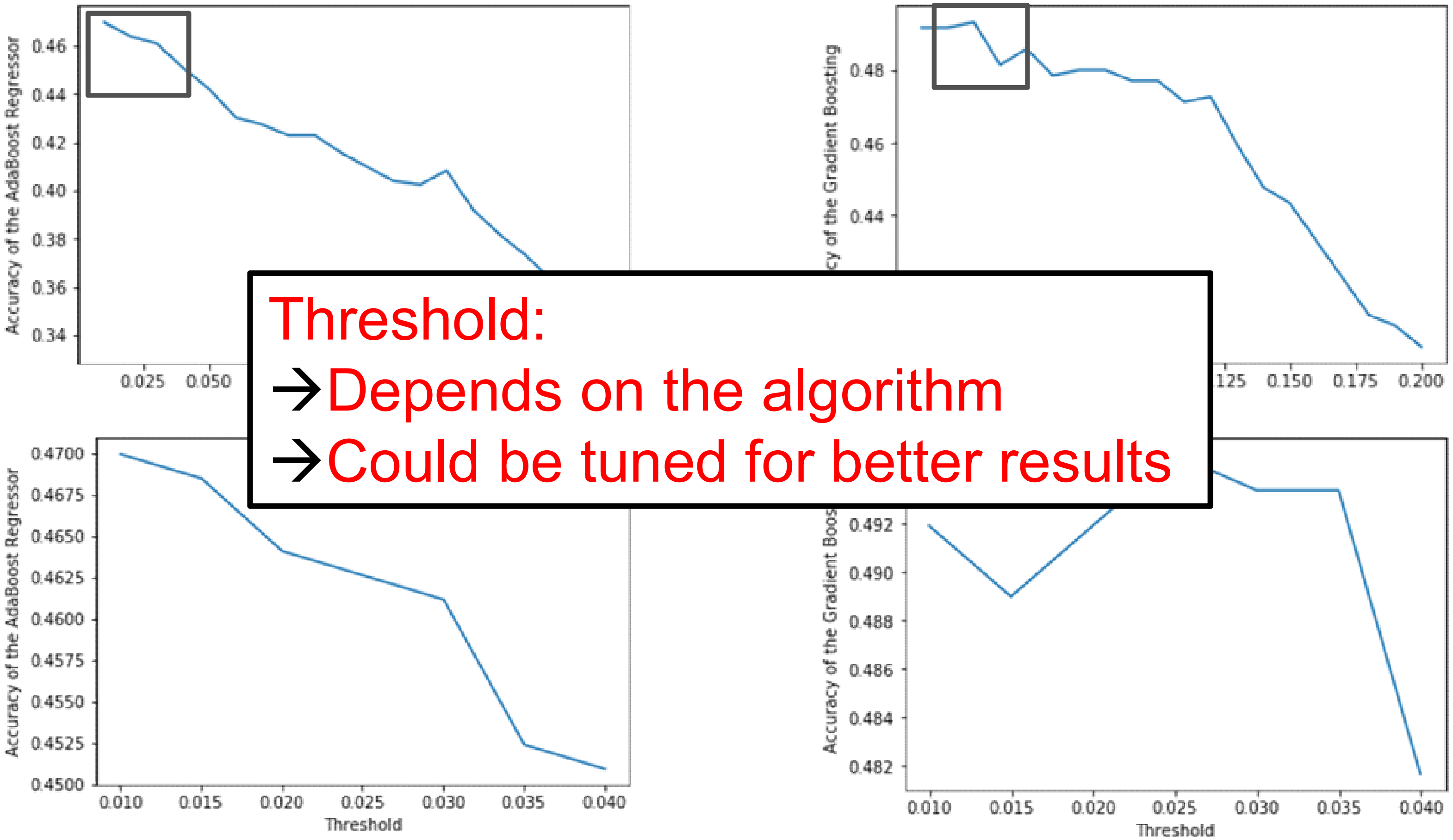
| Regressor | Accuracy Score | Accuracy (threshold = 0,05) |
|---|---|---|
| **MLP Regressor** | **43,79 %** | **44,44 %** |
| **Gradient Boosting** | **41,83 %** | **43,14 %** |
| Random Forest | 37,91 % | 34,64 % |
| AdaBoost | 39,87 % | 39,87 % |
| Bagging Regressor | 39,22 % | 37,25 % |
| **Transformed Target** | **41,18 %** | **41,38 %** |

# Results (Regression)



Technische Universität Berlin

# Results (Regression)



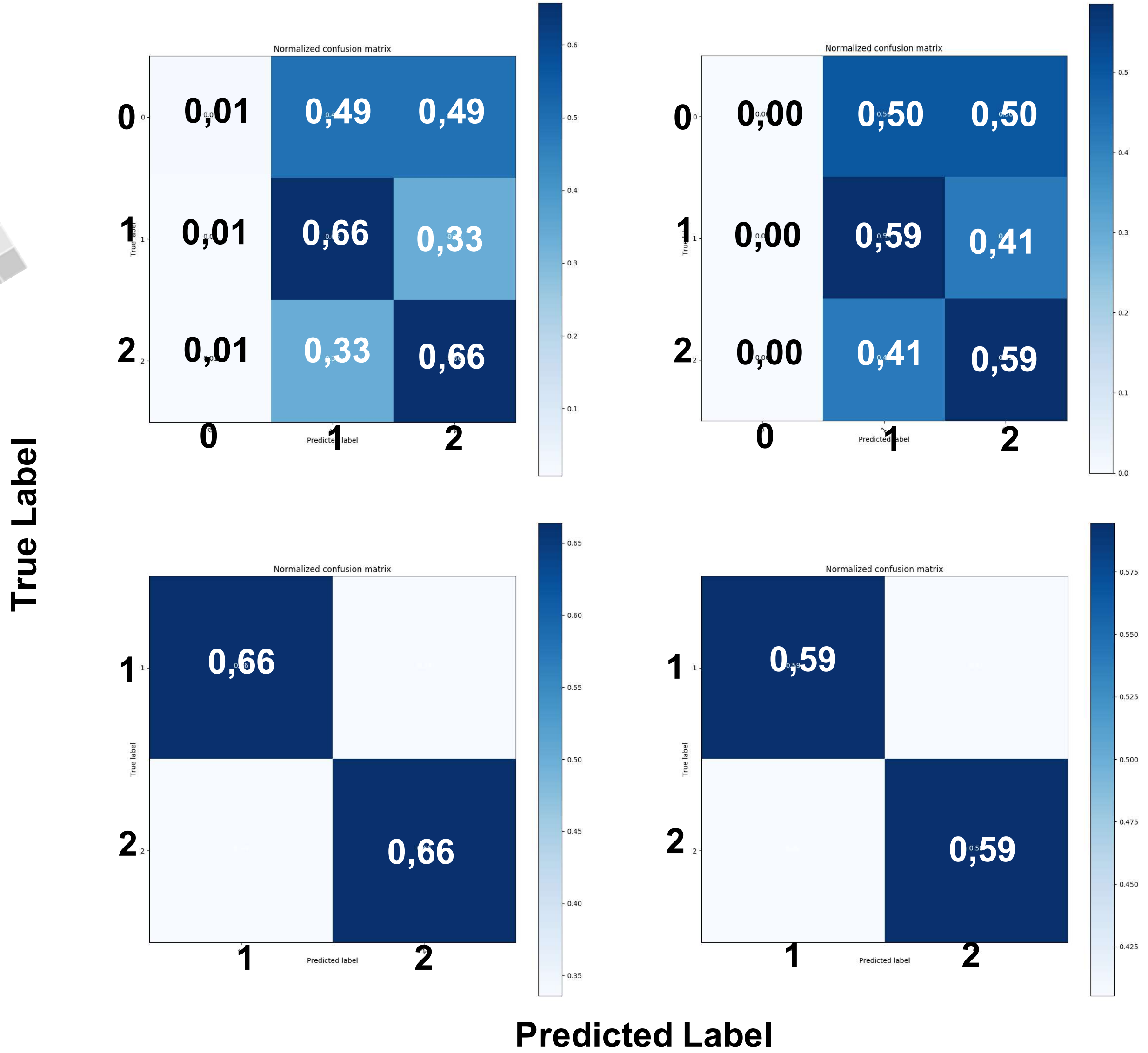

Threshold:

→Depends on the algorithm

→Could be tuned for better results

| Model | Class | Precision | Recall | F1-Score |
|---|---|---|---|---|
| SVC | 0 | 0.38 | 0.02 | 0.03 |
| | 1 | 0.49 | 0.66 | 0.56 |
| | 2 | 0.49 | 0.66 | 0.56 |
| Random Forest | 0 | 0.31 | 0.04 | 0.07 |
| | 1 | 0.49 | 0.64 | 0.55 |
| | 2 | 0.49 | 0.64 | 0.55 |
| Nearest Centroid | 0 | 0.33 | 0.26 | 0.29 |
| | 1 | 0.53 | 0.57 | 0.55 |
| | 2 | 0.53 | 0.57 | 0.55 |
| Ridge Classifier | 0 | 0.00 | 0.00 | 0.00 |
| | 1 | 0.49 | 0.66 | 0.56 |
| | 2 | 0.49 | 0.66 | 0.56 |

Technische Universität Berlin




**SVC Classification**

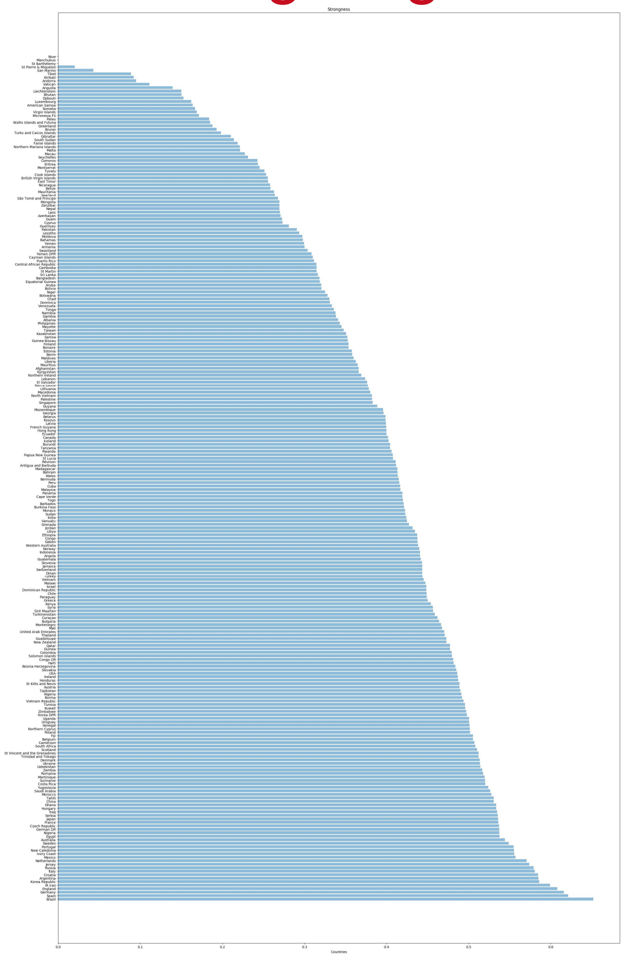


# Investigating the Results

| Regressor | Accuracy Score | threshold = 0,03 | Accuracy (2 classes) |
|---|---|---|---|
| MLP Regressor | 45,82 % | 46,41 % | 62,62 % |
| **Gradient Boosting** | **48,76 %** | **49,34 %** | **65,20 %** |
| Random Forest | 45,24 % | 42,17 % | 63,22 % |
| AdaBoost | 48,02 % | 43,19 % | 54,67 % |
| Bagging Regressor | 44,22 % | 45,39 % | 62,82 % |
| **Transformed Target** | **48,90 %** | **48,02 %** | **66,60 %** |

→ Draw matches are usually more difficult to predict (even for humans) !!

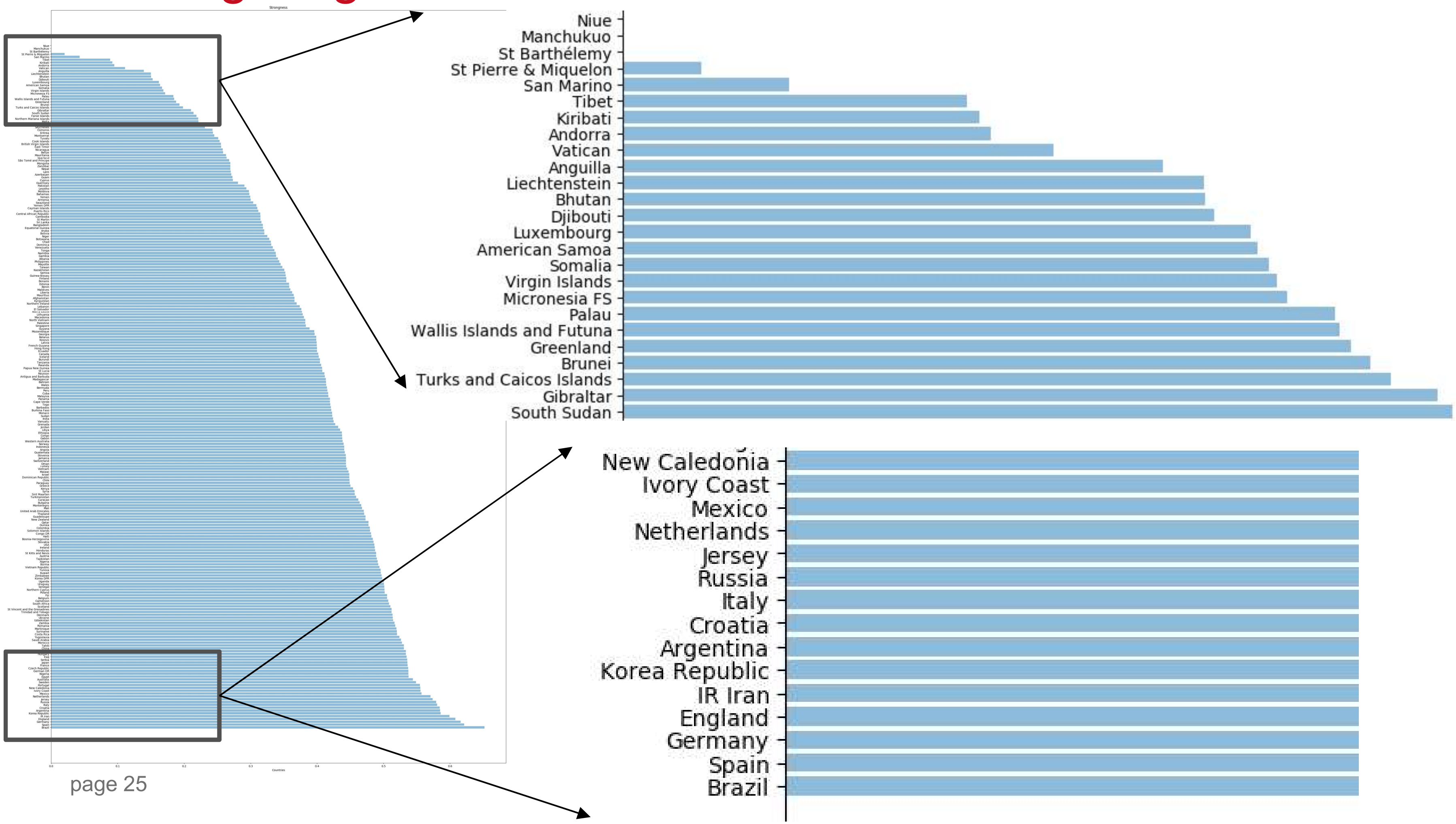
# Investigating the Results



strength (team$_i$) =

$$\text{average} \; \frac{score(team_i)}{score(team_i) + score(opponent_i)}$$

# Investigating the Results



# Demo

DEMO

Technische Universität Berlin

# Future Work

| | |
|---|---|
| Features > Observations | • Focus on adding more features rather than more data |
| Regression > Classification | • Allows us to model the strength of a team, rather than only the winner<br>• 6-1 VS 2-1 → 0,86 VS 0,67 |
| Model | • Tune the threshold, accordingly to the model |
| Draw Matches | • Online Learning<br>• User implication |



# Resources

https://www.flashscore.com/football/world/friendly-international/archive/

https://www.fifa.com/

https://ourworldindata.org/

http://en.fifaranking.net/ranking/

https://data.worldbank.org/

http://projectbritain.com/population.html

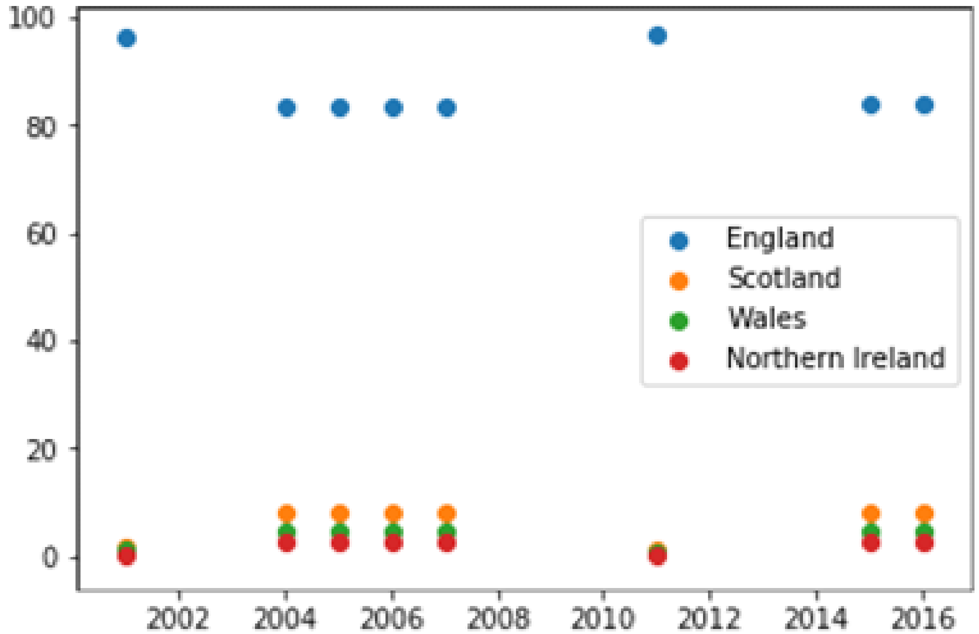
http://www.worldometers.info/world-population

# ☺ Thank you for Your Attention ☺

# Questions ?

Technische Universität Berlin

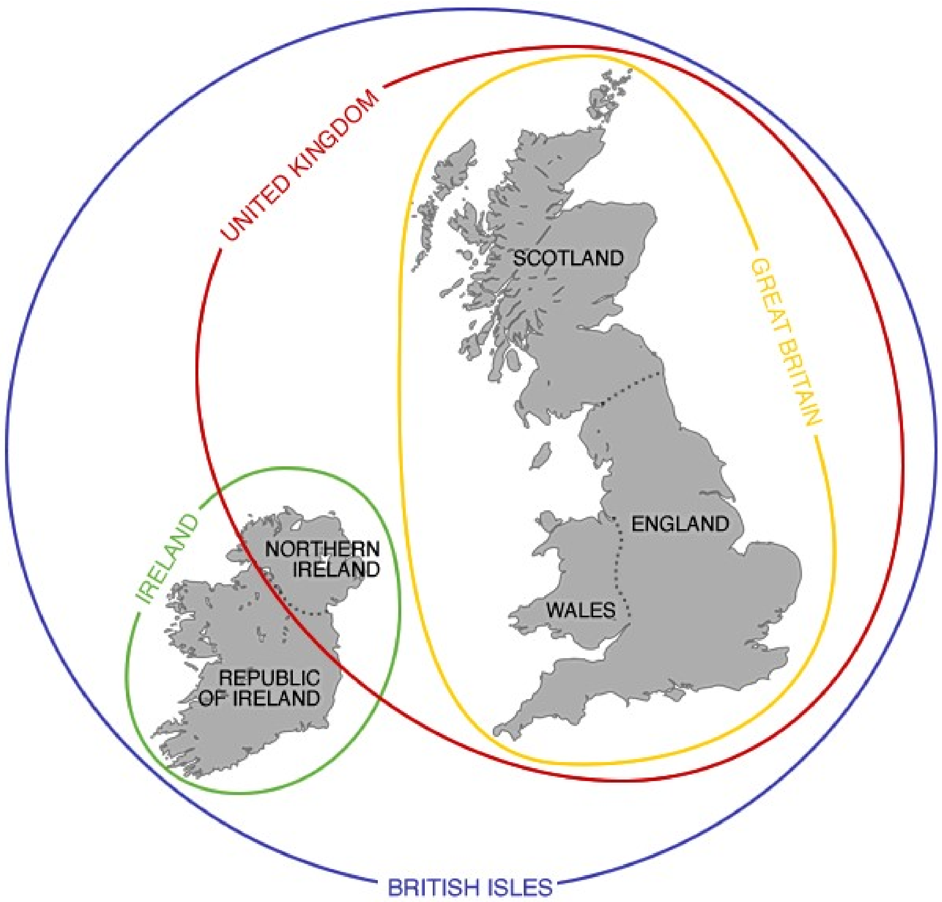
# Challenges





England: 84%
Scotland: 8.3%
Wales: 4.8%
Northern Ireland: 2.9%

# Challenges


UNITED KINGDOM
GREAT BRITAIN
IRELAND
SCOTLAND
ENGLAND
WALES
NORTHERN IRELAND
REPUBLIC OF IRELAND
BRITISH ISLES