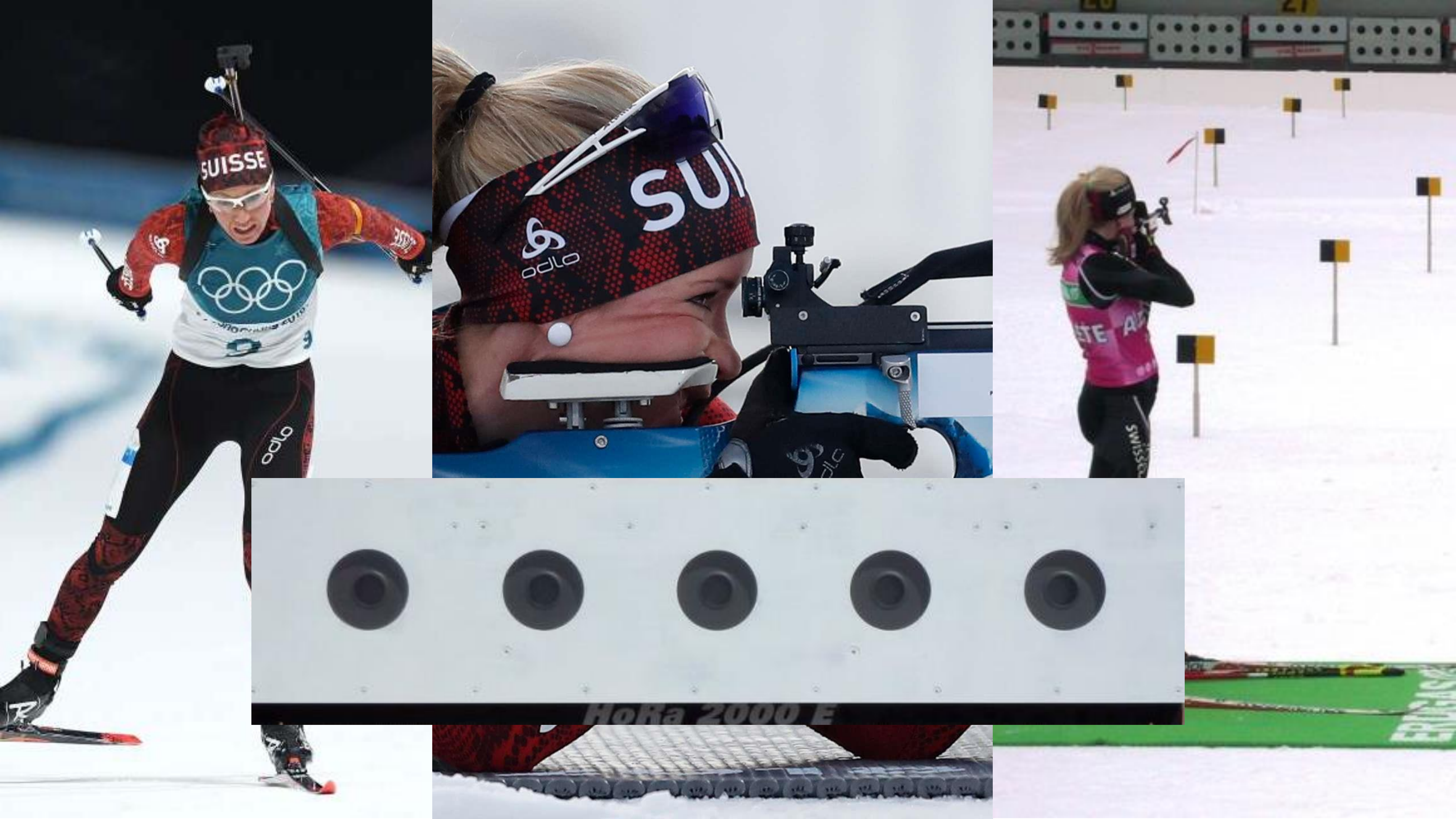
für Simon
Cologna
Thanx for the support

Predicting biathlon shooting performance using machine learning

Thomas Maier[1], Daniel Meister[2], Severin Trösch[1], Jon Peter Wehrlin[1]

[1]*Eidgenössische Hochschule für Sport Magglingen EHSM*
[2]*Datahouse AG*


SUISSE
SUI
odlo
HoRa 2000 E
SWISS

Introduction

- Shooting is crucial for end ranking (~50%)
  (Luchsinger et al. 2017)

- Influence of fatigue and biomechanical parameters
  (Hoffmann et al. 1992; Sattlecker et al. 2017)

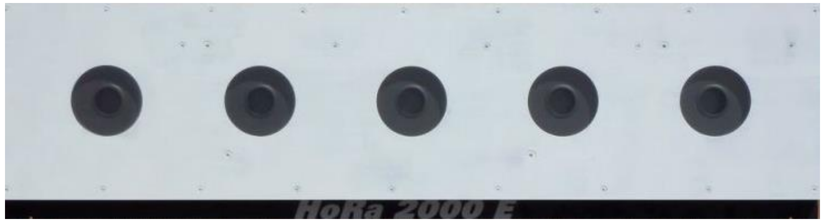
- Shooting mode, athlete level, variation in performance
  (Luchsinger et al. 2017; Skattebo & Losnegard 2017)

- **How predictable are individual shots?**

Data

- World Cup, World Championships und Olympic Games (only single athlete categories)

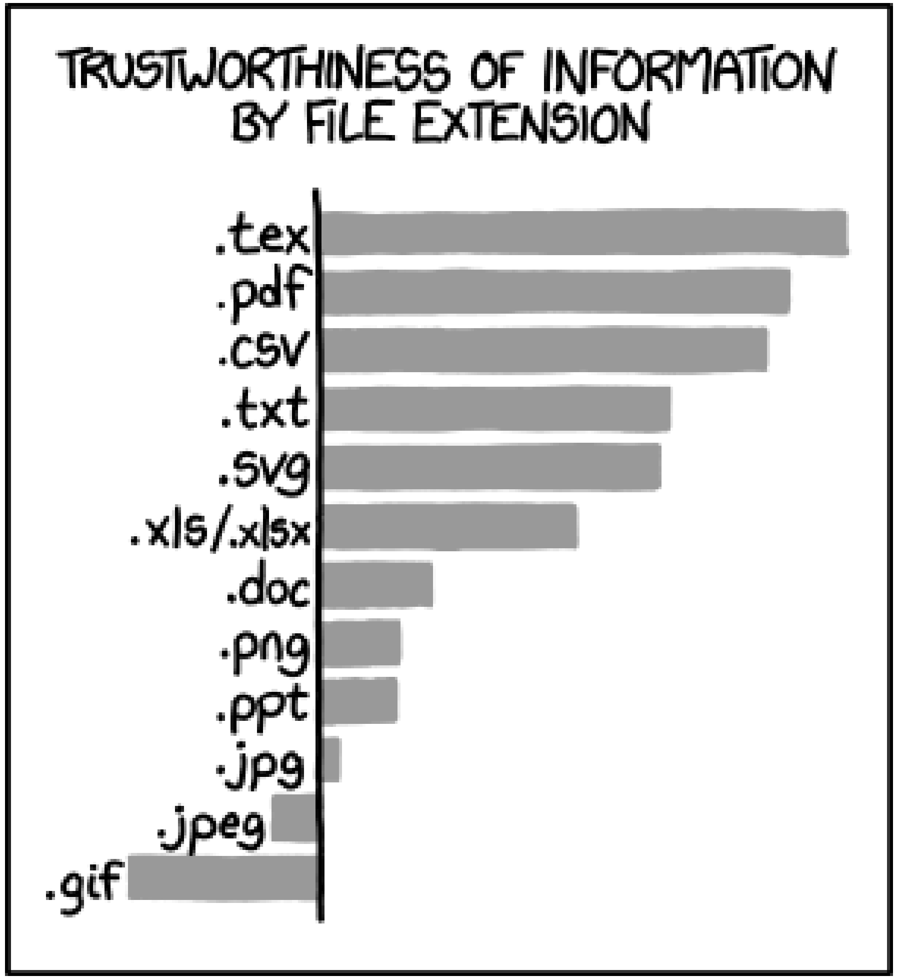
- From HoRa, supplier of target system



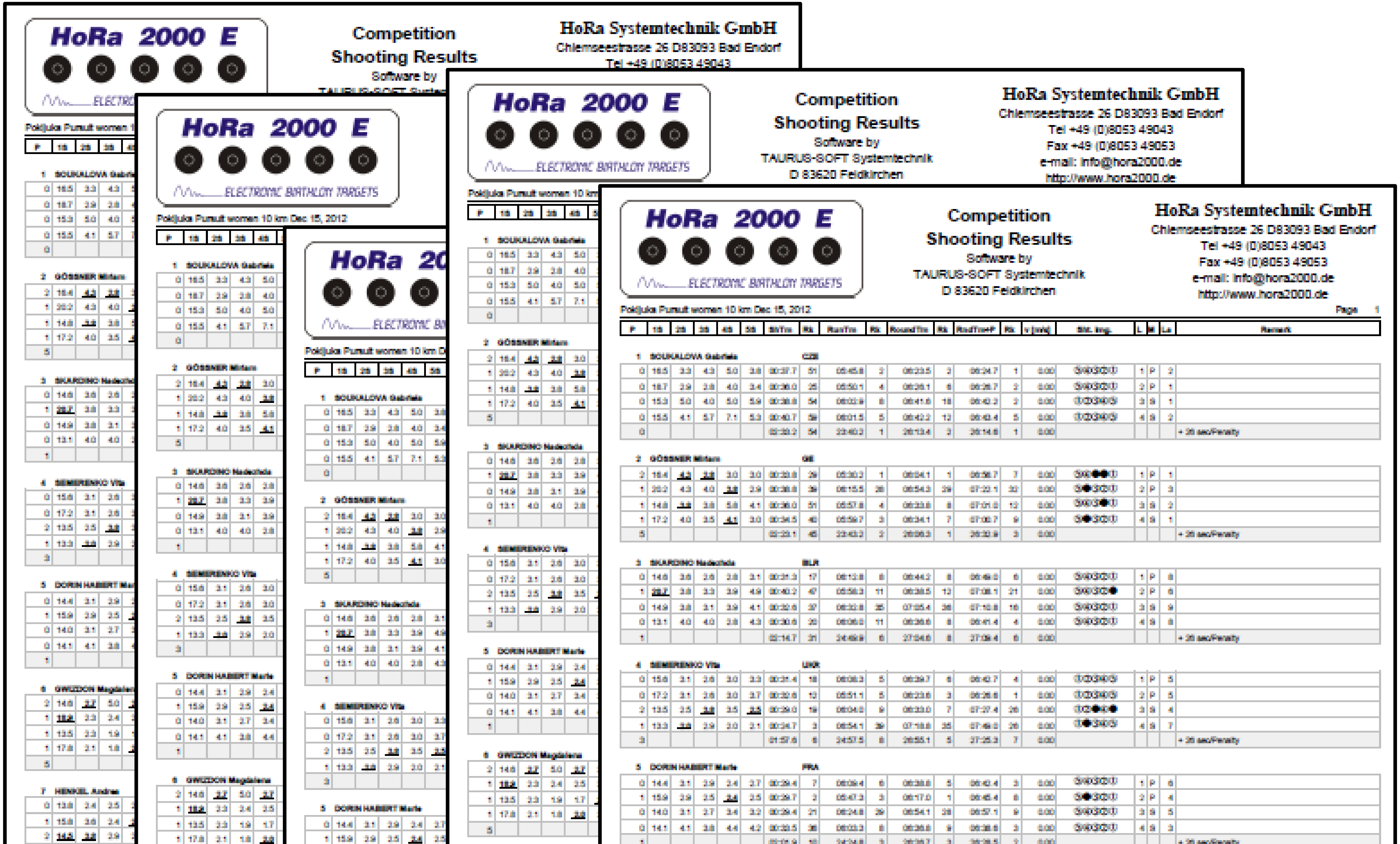
- **Training data**: 2012/13 – 2015/16 → **Test data**: 2016/17

  Total of 152’640 shots

Data … as PDF



xkcd



HoRa 2000 E

ELECTRONIC BIATHLON TARGETS

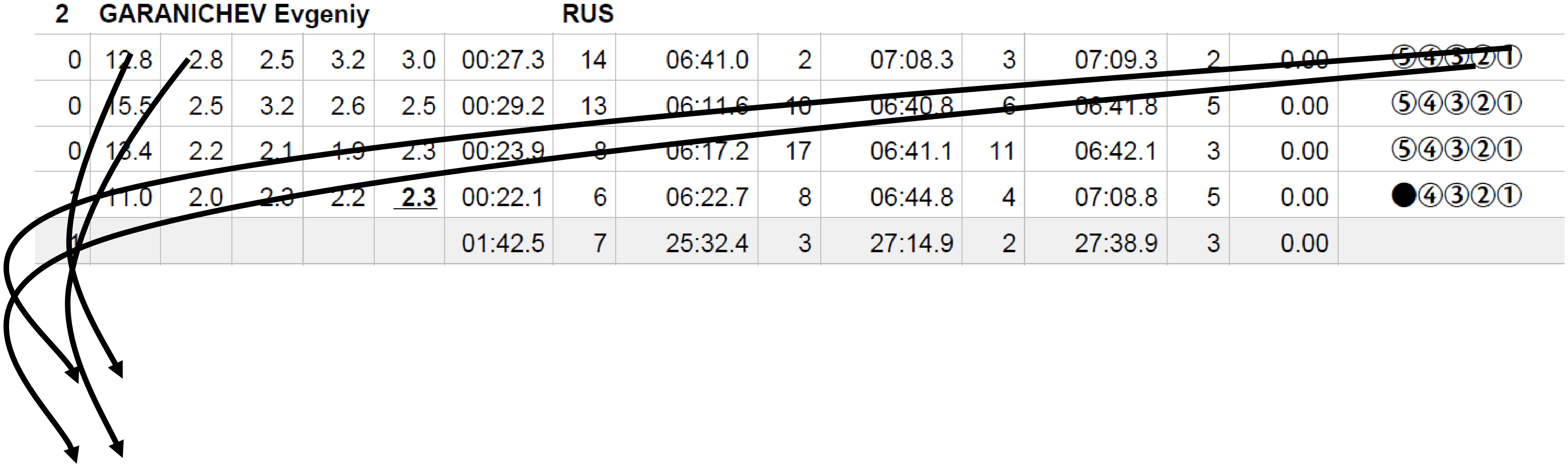
Competition
Shooting Results
Software by
TAURUS-SOFT Systemtechnik
D 83620 Feldkirchen

HoRa Systemtechnik GmbH
Chiemseestrasse 26 D83093 Bad Endorf
Tel +49 (0)8053 49043
Fax +49 (0)8053 49053
e-mail: info@hora2000.de
http://www.hora2000.de

Pokljuka Pursuit women 10 km Dec 15, 2012 Page 1

P	1S	2S	3S	4S	5S	ShTm	Rk	RunTm	Rk	RoundTm	Rk	RndTm+P	Rk	v [m/s]	Sht. img.	L	M	La	Remark
1 SOUKALOVA Gabriela							CZE												
0	16.5	3.3	4.3	5.0	3.8	00:37.7	51	05:45.8	2	06:23.5	2	06:24.7	1	0.00	○○○○○	1	P	2	
0	18.7	2.9	2.8	4.0	3.4	00:36.0	25	05:50.1	4	06:26.1	6	06:26.7	2	0.00	○○○○○	2	P	1	
0	15.3	5.0	4.0	5.0	5.9	00:38.8	54	06:02.9	8	06:41.6	18	06:42.2	2	0.00	○○○○○	3	S	1	
0	15.5	4.1	5.7	7.1	5.3	00:40.7	59	06:01.5	5	06:42.2	13	06:43.4	5	0.00	○○○○○	4	S	2	
0						02:33.2	54	23:40.2	1	26:13.4	2	26:14.8	1	0.00					+ 20 sec/Penalty
2 GÖSSNER Miriam							GER												
2	16.4	4.3	3.8	3.0	3.0	00:33.8	29	05:30.3	1	06:04.1	1	06:56.7	7	0.00	○○●●○	1	P	1	
1	20.2	4.3	4.0	3.8	2.9	00:38.8	39	06:15.5	28	06:54.3	29	07:22.1	32	0.00	○●○○○	2	P	3	
1	14.8	3.8	3.8	5.8	4.1	00:36.0	51	05:57.8	4	06:33.8	8	07:01.8	13	0.00	○○○●○	3	S	2	
1	17.2	4.0	3.5	4.1	3.0	00:34.5	40	05:59.7	3	06:34.1	7	07:00.7	9	0.00	○●○○○	4	S	1	
5						02:23.1	45	23:43.2	2	26:06.3	1	26:32.8	3	0.00					+ 20 sec/Penalty
3 SKARDINO Nadezhda							BLR												
0	14.6	3.6	2.6	2.8	3.1	00:31.3	17	06:12.8	8	06:44.2	8	06:49.0	6	0.00	○○○○○	1	P	8	
1	20.7	3.8	3.3	3.9	4.9	00:40.2	47	05:58.3	11	06:38.5	12	07:08.1	21	0.00	○○○○●	2	P	6	
0	14.9	3.8	3.1	3.9	4.1	00:32.6	37	06:32.8	35	07:05.4	36	07:10.8	16	0.00	○○○○○	3	S	9	
0	13.1	4.0	4.0	2.8	4.3	00:30.6	20	06:06.0	11	06:36.6	8	06:41.4	4	0.00	○○○○○	4	S	8	
1						02:14.7	31	24:49.9	6	27:04.6	8	27:09.4	6	0.00					+ 20 sec/Penalty
4 SEMERENKO Vita							UKR												
0	15.6	3.1	2.6	3.0	3.3	00:31.4	18	06:08.3	5	06:39.7	6	06:42.7	4	0.00	○○○○○	1	P	5	
0	17.2	3.1	2.6	3.0	3.7	00:32.6	13	05:51.1	5	06:23.6	3	06:26.6	1	0.00	○○○○○	2	P	5	
2	13.5	2.5	3.8	3.5	3.5	00:29.0	19	06:04.0	9	06:33.0	7	07:27.4	26	0.00	○○●○●	3	S	4	
1	13.3	3.8	2.9	3.0	2.1	00:24.7	3	06:54.1	39	07:18.8	35	07:49.0	26	0.00	○●○○○	4	S	7	
3						01:57.6	6	24:57.5	8	26:55.1	5	27:25.3	7	0.00					+ 20 sec/Penalty
5 DORIN HABERT Marie							FRA												
0	14.4	3.1	2.9	2.4	2.7	00:29.4	7	06:09.4	6	06:38.8	5	06:42.4	3	0.00	○○○○○	1	P	6	
1	15.9	2.9	2.5	2.4	2.5	00:29.7	2	05:47.3	3	06:17.0	1	06:45.4	8	0.00	○●○○○	2	P	4	
0	14.0	3.1	2.7	3.4	3.2	00:29.4	21	06:24.8	29	06:54.1	28	06:57.1	9	0.00	○○○○○	3	S	5	
0	14.1	4.1	3.8	4.4	4.2	00:30.5	38	06:03.3	8	06:36.8	9	06:38.8	3	0.00	○○○○○	4	S	3	
1						02:01.9	10	24:24.8	3	26:26.7	3	26:26.5	2	0.00					+ 20 sec/Penalty

Tidy data

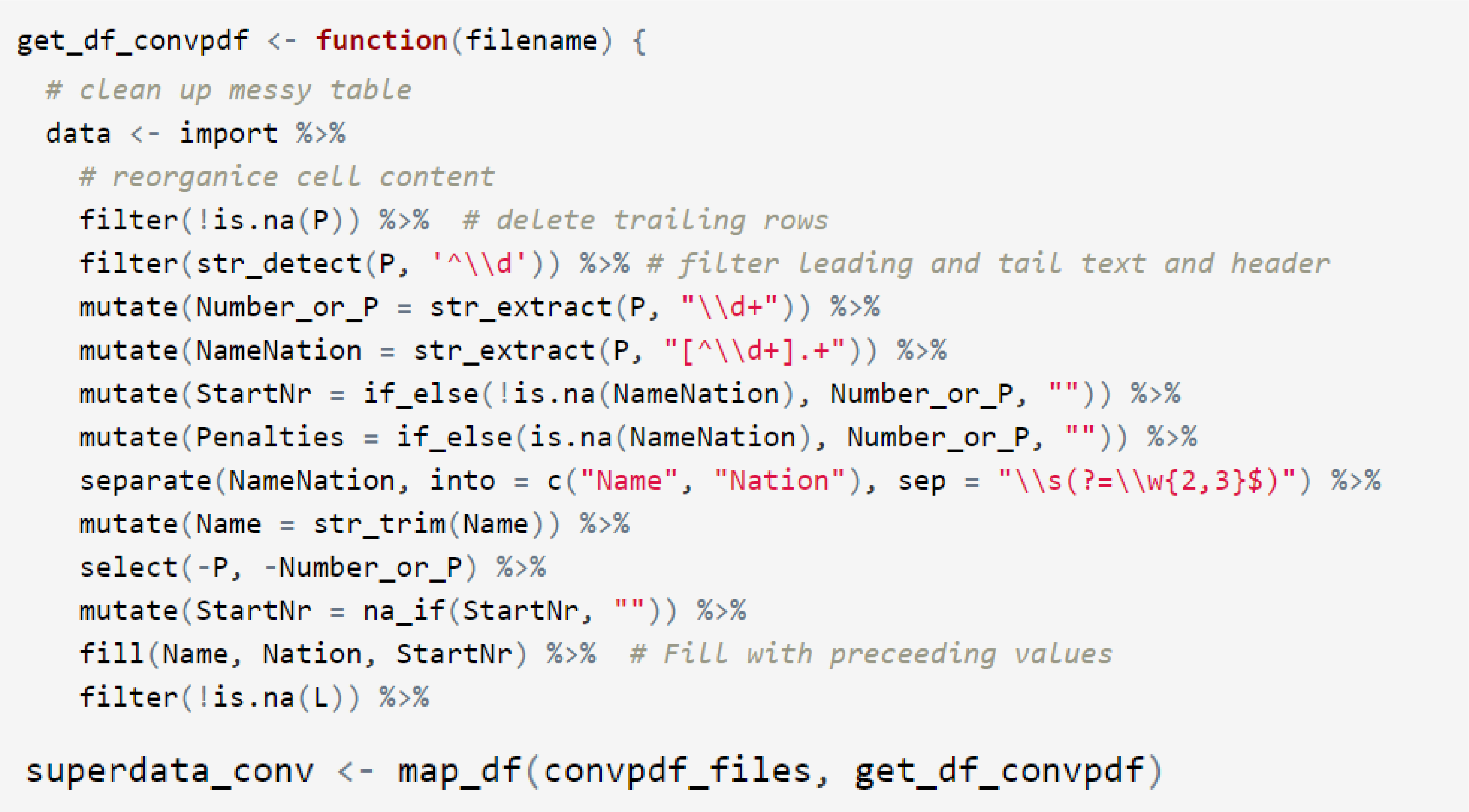


2 GARANICHEV Evgeniy RUS

0	12.8	2.8	2.5	3.2	3.0	00:27.3	14	06:41.0	2	07:08.3	3	07:09.3	2	0.00	⑤④③②①
0	15.5	2.5	3.2	2.6	2.5	00:29.2	13	06:11.6	10	06:40.8	6	06:41.8	5	0.00	⑤④③②①
0	13.4	2.2	2.1	1.9	2.3	00:23.9	8	06:17.2	17	06:41.1	11	06:42.1	3	0.00	⑤④③②①
1	11.0	2.0	2.3	2.2	**2.3**	00:22.1	6	06:22.7	8	06:44.8	4	07:08.8	5	0.00	●④③②①
1						01:42.5	7	25:32.4	3	27:14.9	2	27:38.9	3	0.00	

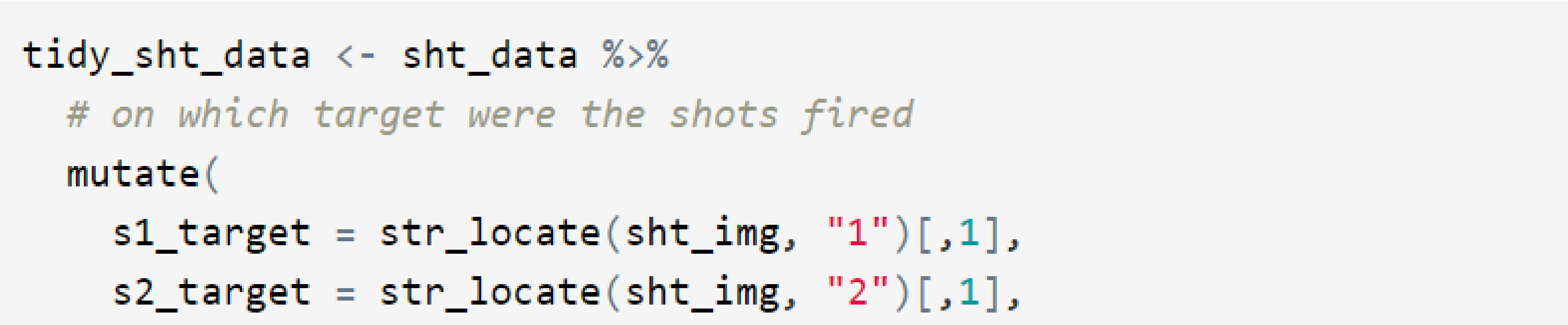
One row for each shot

Reorganise data with dplyr

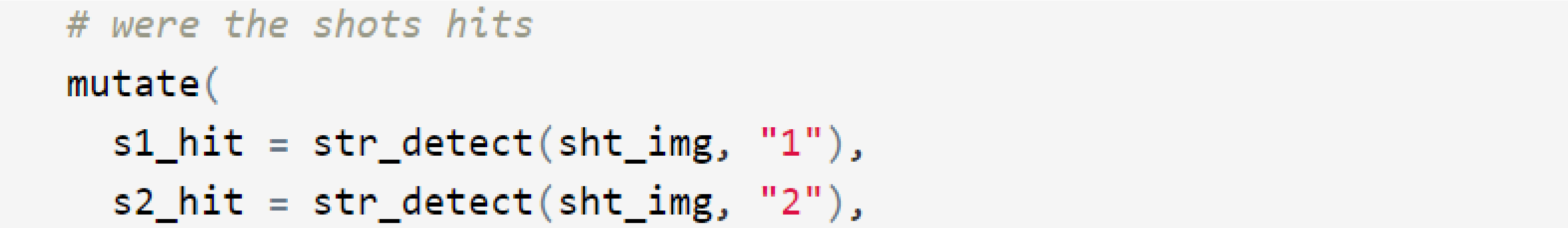


```
get_df_convpdf <- function(filename) {
  # clean up messy table
  data <- import %>%
    # reorganice cell content
    filter(!is.na(P)) %>%  # delete trailing rows
    filter(str_detect(P, '^\\d')) %>% # filter leading and tail text and header
    mutate(Number_or_P = str_extract(P, "\\d+")) %>%
    mutate(NameNation = str_extract(P, "[^\\d+].+")) %>%
    mutate(StartNr = if_else(!is.na(NameNation), Number_or_P, "")) %>%
    mutate(Penalties = if_else(is.na(NameNation), Number_or_P, "")) %>%
    separate(NameNation, into = c("Name", "Nation"), sep = "\\s(?=\\w{2,3}$)") %>%
    mutate(Name = str_trim(Name)) %>%
    select(-P, -Number_or_P) %>%
    mutate(StartNr = na_if(StartNr, "")) %>%
    fill(Name, Nation, StartNr) %>%  # Fill with preceeding values
    filter(!is.na(L)) %>%

superdata_conv <- map_df(convpdf_files, get_df_convpdf)
```

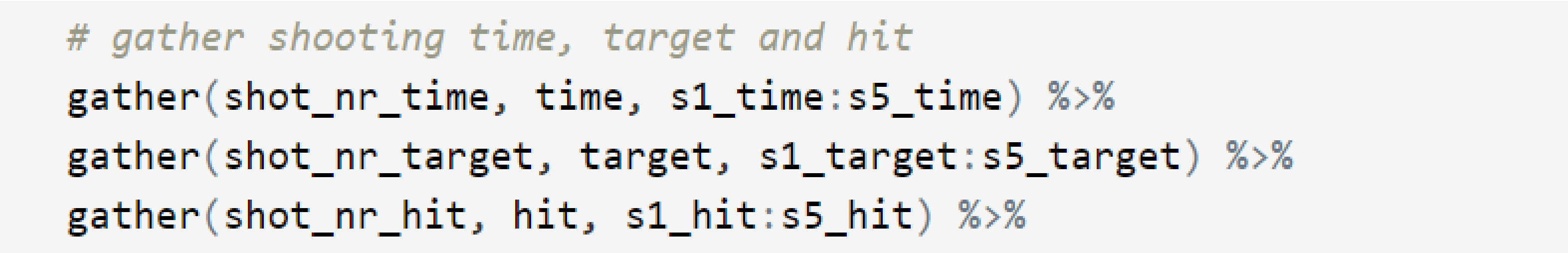
Gather data



```
tidy_sht_data <- sht_data %>%
  # on which target were the shots fired
  mutate(
    s1_target = str_locate(sht_img, "1")[,1],
    s2_target = str_locate(sht_img, "2")[,1],
```

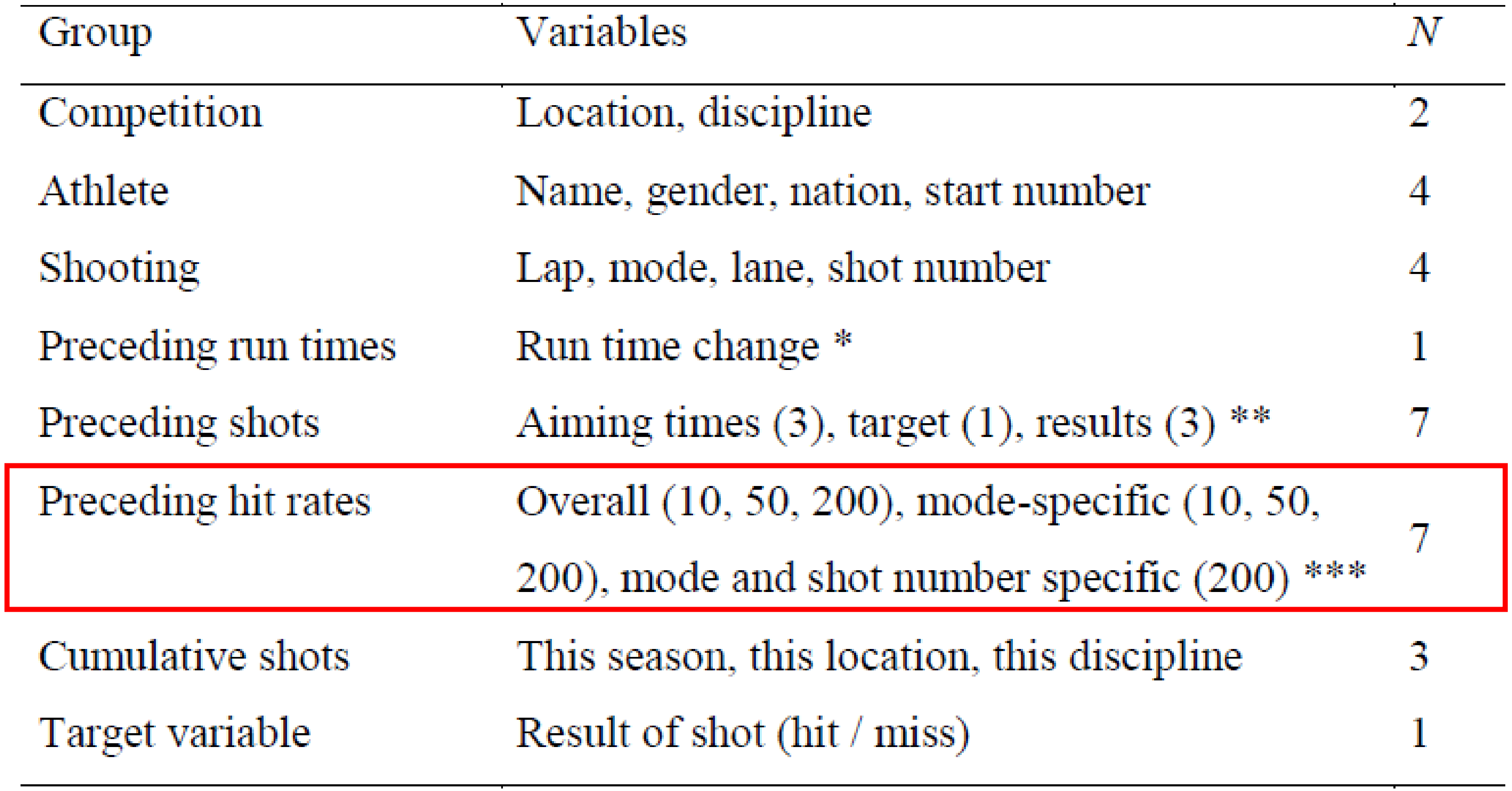


```
  # were the shots hits
  mutate(
    s1_hit = str_detect(sht_img, "1"),
    s2_hit = str_detect(sht_img, "2"),
```



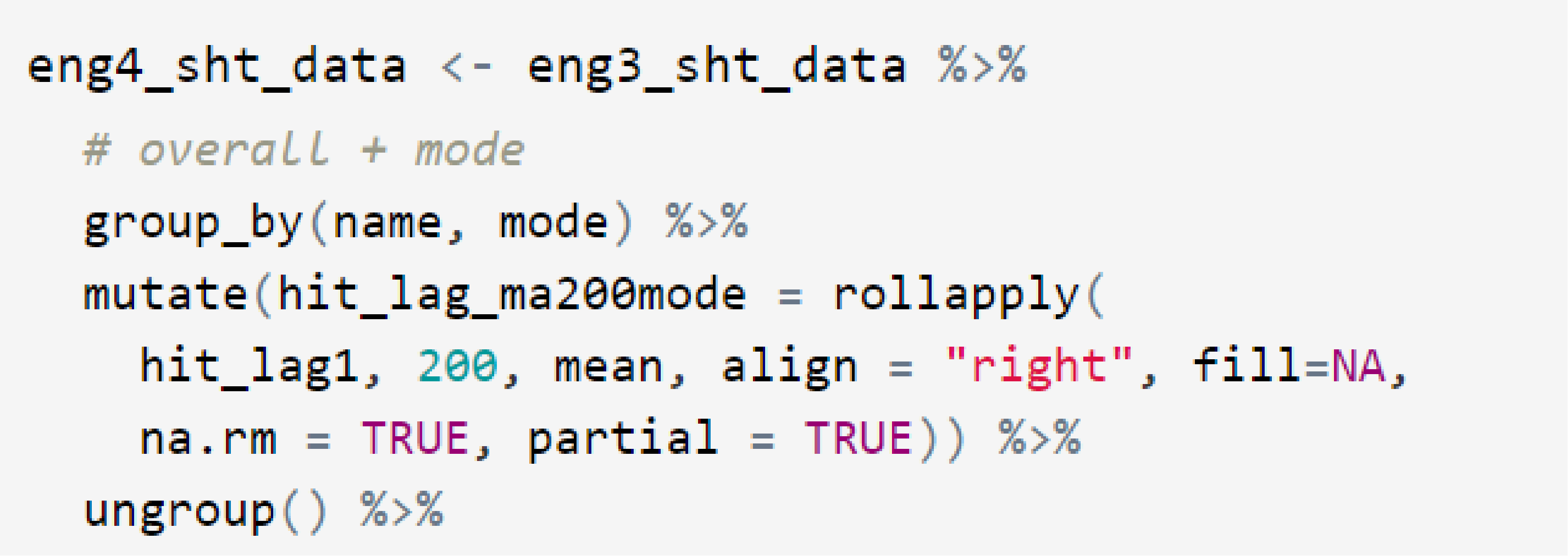
```
  # gather shooting time, target and hit
  gather(shot_nr_time, time, s1_time:s5_time) %>%
  gather(shot_nr_target, target, s1_target:s5_target) %>%
  gather(shot_nr_hit, hit, s1_hit:s5_hit) %>%
```

Feature Engineering (29 Variables)



Group	Variables	*N*
Competition	Location, discipline	2
Athlete	Name, gender, nation, start number	4
Shooting	Lap, mode, lane, shot number	4
Preceding run times	Run time change *	1
Preceding shots	Aiming times (3), target (1), results (3) **	7
Preceding hit rates	Overall (10, 50, 200), mode-specific (10, 50, 200), mode and shot number specific (200) ***	7
Cumulative shots	This season, this location, this discipline	3
Target variable	Result of shot (hit / miss)	1

Rolling functions with zoo



```
eng4_sht_data <- eng3_sht_data %>%
  # overall + mode
  group_by(name, mode) %>%
  mutate(hit_lag_ma200mode = rollapply(
    hit_lag1, 200, mean, align = "right", fill=NA,
    na.rm = TRUE, partial = TRUE)) %>%
  ungroup() %>%
```

Analysis

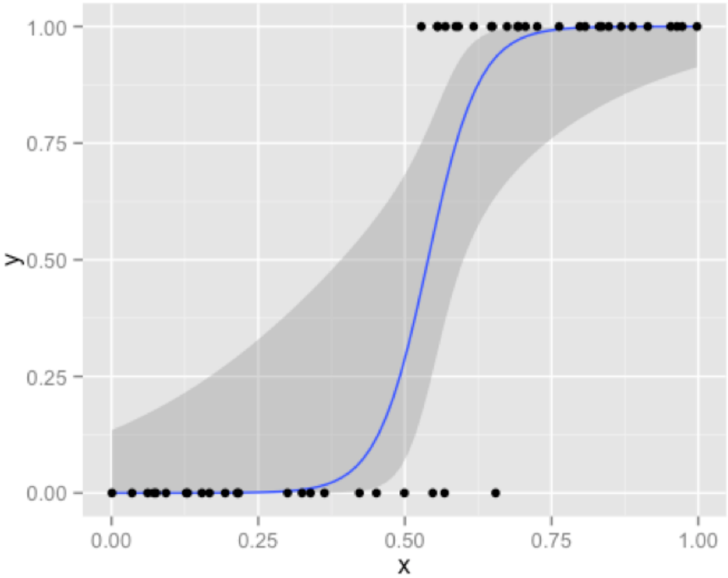
Exploratory Data Analysis

- 95% Confidence limits
- Pearson Correlations
- Chi-squared- / Mann-Whitney-U-Tests

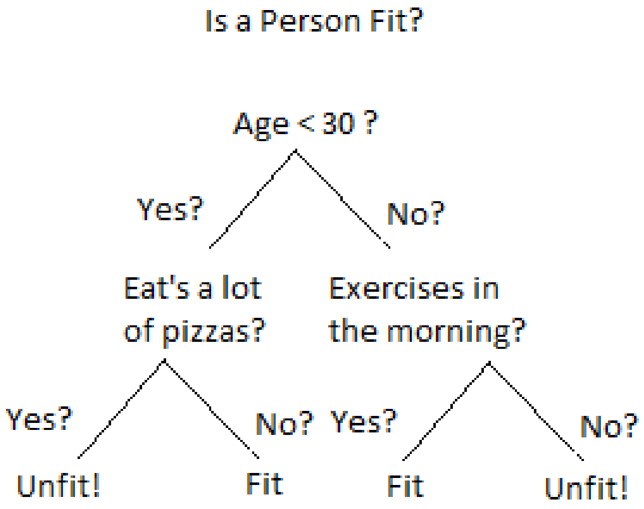
Machine Learning

- **LogReg**: logistic regression using only 1 input-variable
- **XGB**: extreme gradient boosting with trees
- **NNet**: artifical neural network

LogReg



XGB



Is a Person Fit?

Age < 30 ?

Yes?

No?

Eat's a lot of pizzas?

Exercises in the morning?

Yes?

No?

Yes?

No?

Unfit!

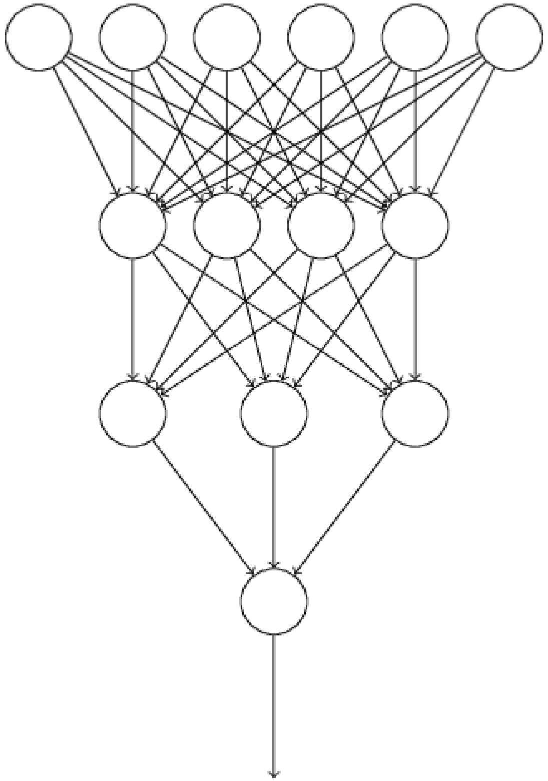
Fit

Fit

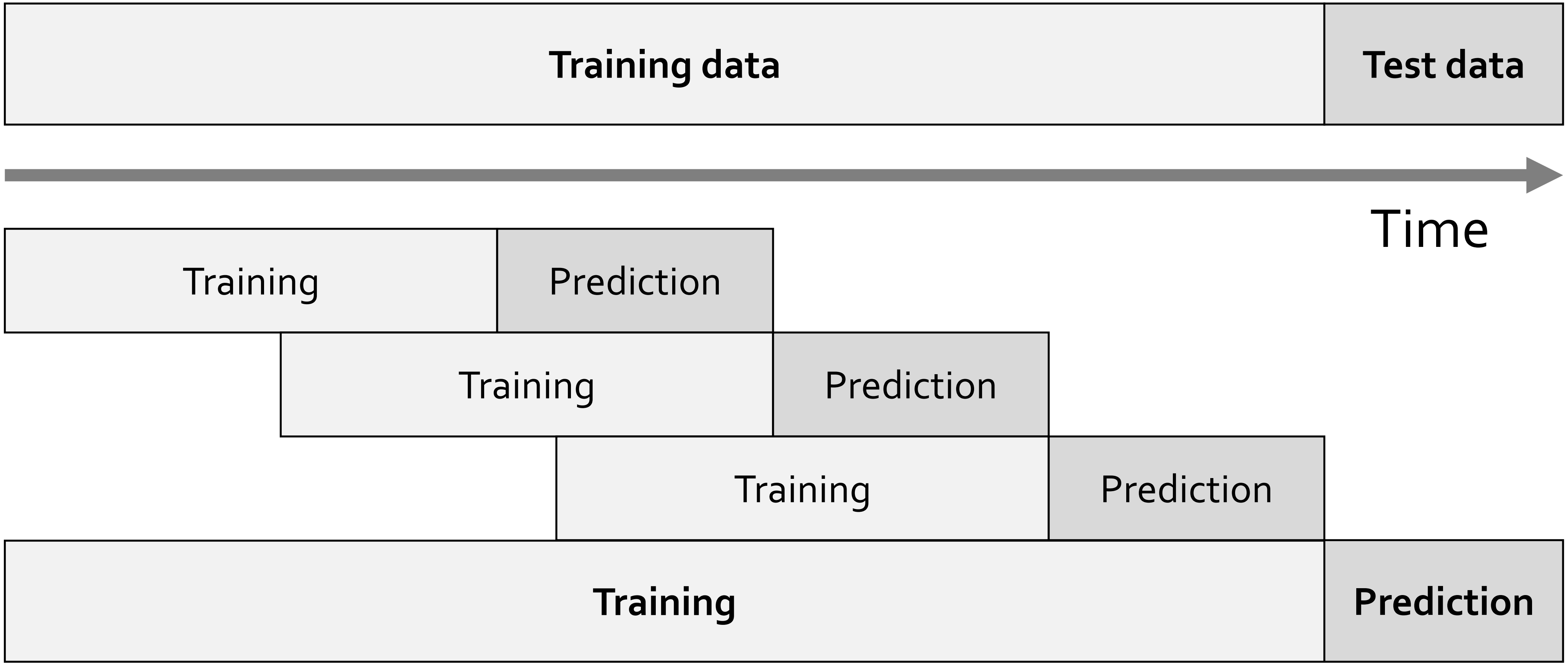
Unfit!

Sequential trees to fit errors of previous trees

NNet



Time sliced cross-validation



Training data **Test data**

Time

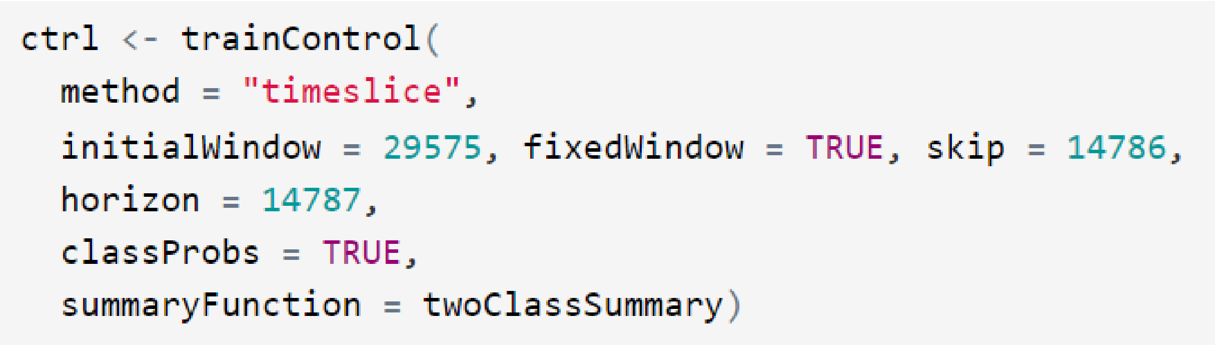
Training Prediction

Training Prediction

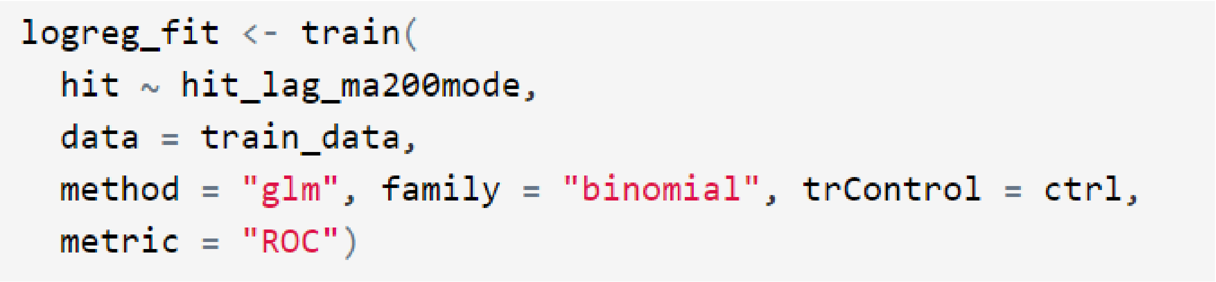
Training Prediction

Training **Prediction**

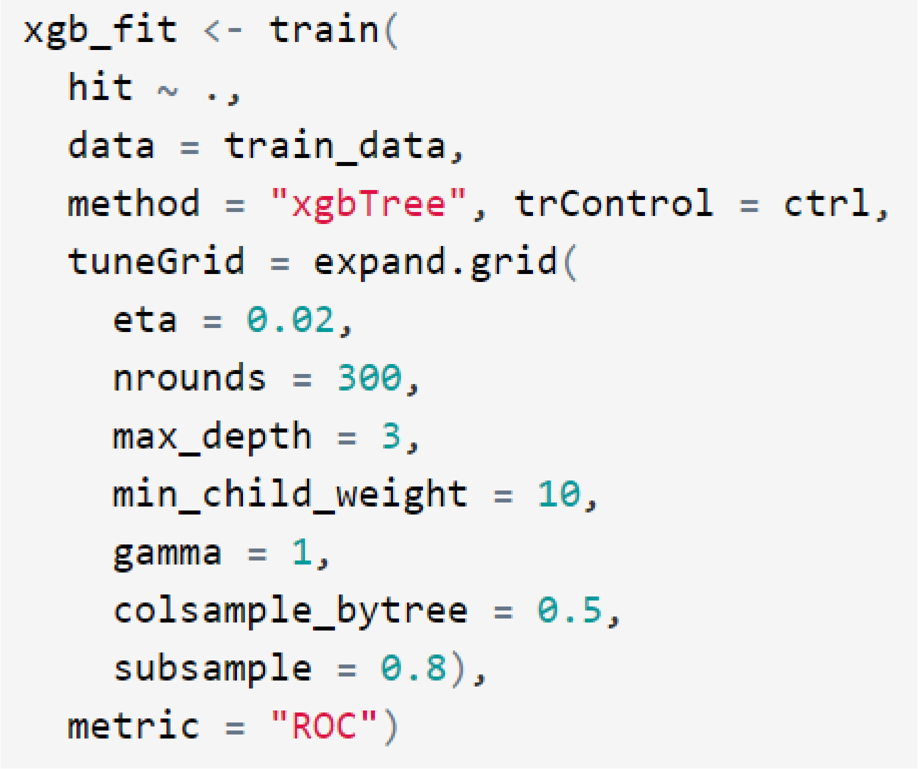
Caret – ML model wrapper



```r
ctrl <- trainControl(
  method = "timeslice",
  initialWindow = 29575, fixedWindow = TRUE, skip = 14786,
  horizon = 14787,
  classProbs = TRUE,
  summaryFunction = twoClassSummary)
```



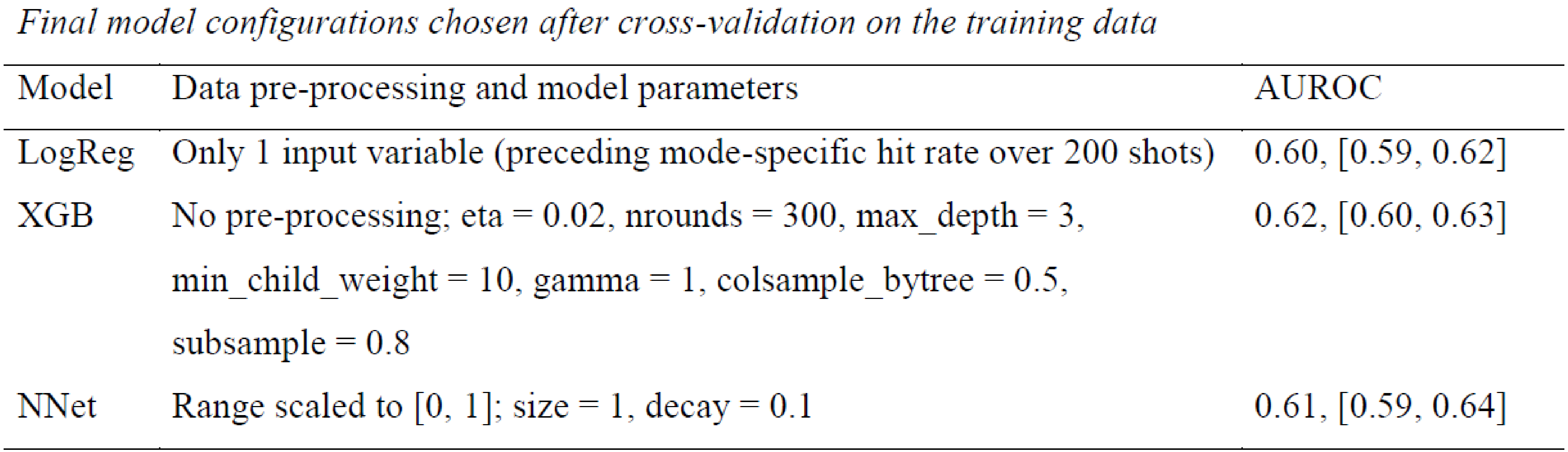
```r
logreg_fit <- train(
  hit ~ hit_lag_ma200mode,
  data = train_data,
  method = "glm", family = "binomial", trControl = ctrl,
  metric = "ROC")
```



```r
xgb_fit <- train(
  hit ~ .,
  data = train_data,
  method = "xgbTree", trControl = ctrl,
  tuneGrid = expand.grid(
    eta = 0.02,
    nrounds = 300,
    max_depth = 3,
    min_child_weight = 10,
    gamma = 1,
    colsample_bytree = 0.5,
    subsample = 0.8),
  metric = "ROC")
```

IBU WORLD CHAMPIONSHIPS BIATHLON 2016
PRESENTED BY
HOLMENKOLLEN
IBU
DKB
e-on
2
1
Holmenkollen
3

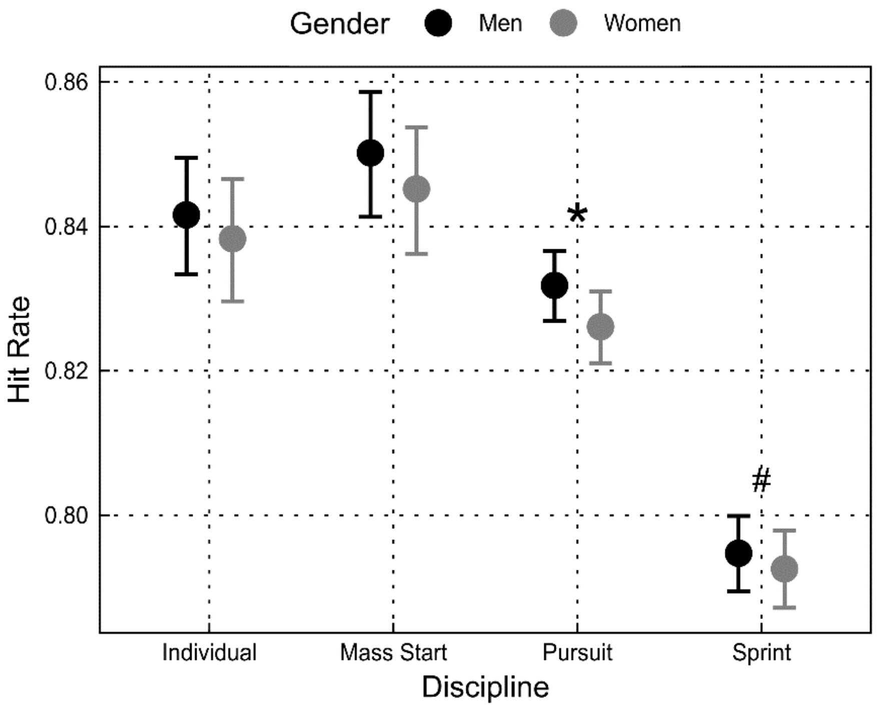
Final model configurations

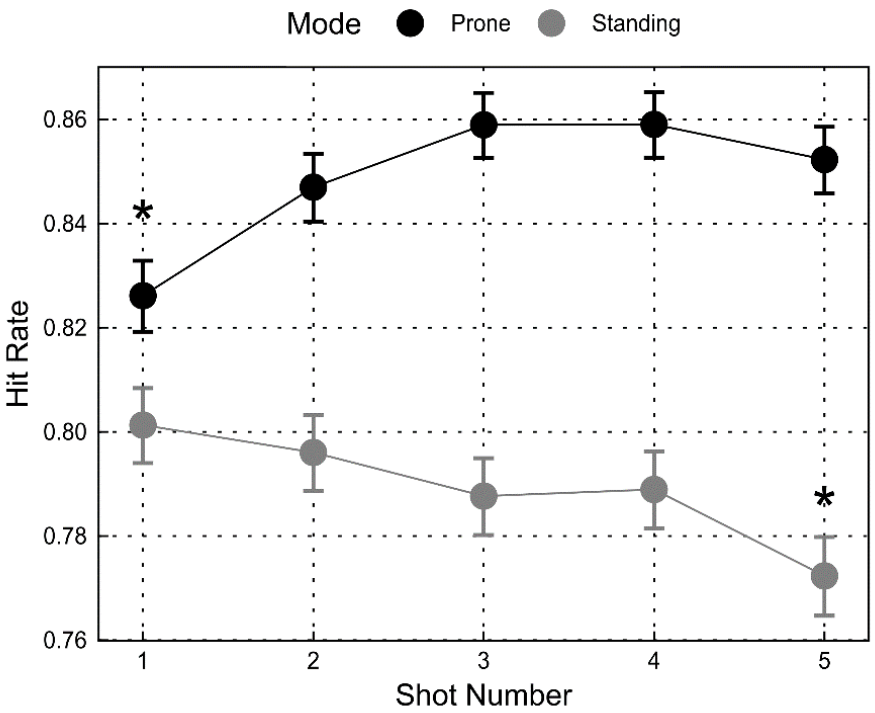


Final model configurations chosen after cross-validation on the training data

Model	Data pre-processing and model parameters	AUROC
LogReg	Only 1 input variable (preceding mode-specific hit rate over 200 shots)	0.60, [0.59, 0.62]
XGB	No pre-processing; eta = 0.02, nrounds = 300, max_depth = 3, min_child_weight = 10, gamma = 1, colsample_bytree = 0.5, subsample = 0.8	0.62, [0.60, 0.63]
NNet	Range scaled to [0, 1]; size = 1, decay = 0.1	0.61, [0.59, 0.64]

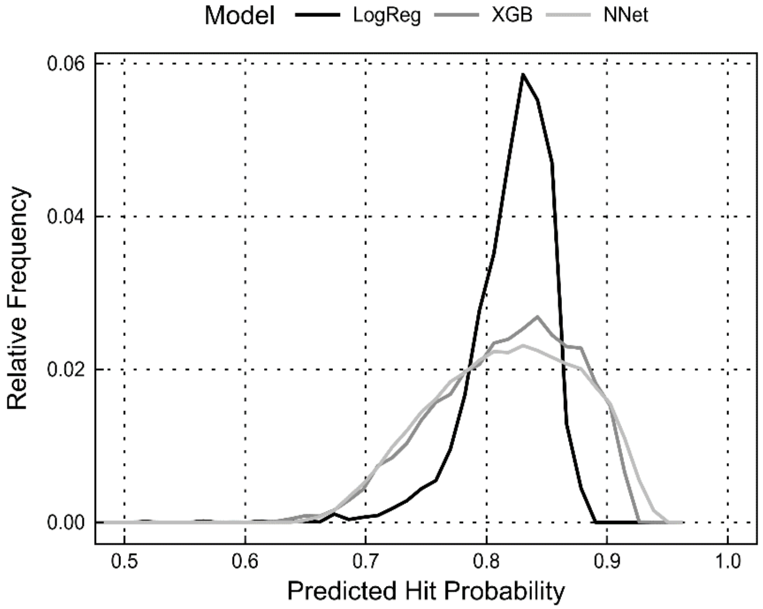
Results – Exploratory Analysis

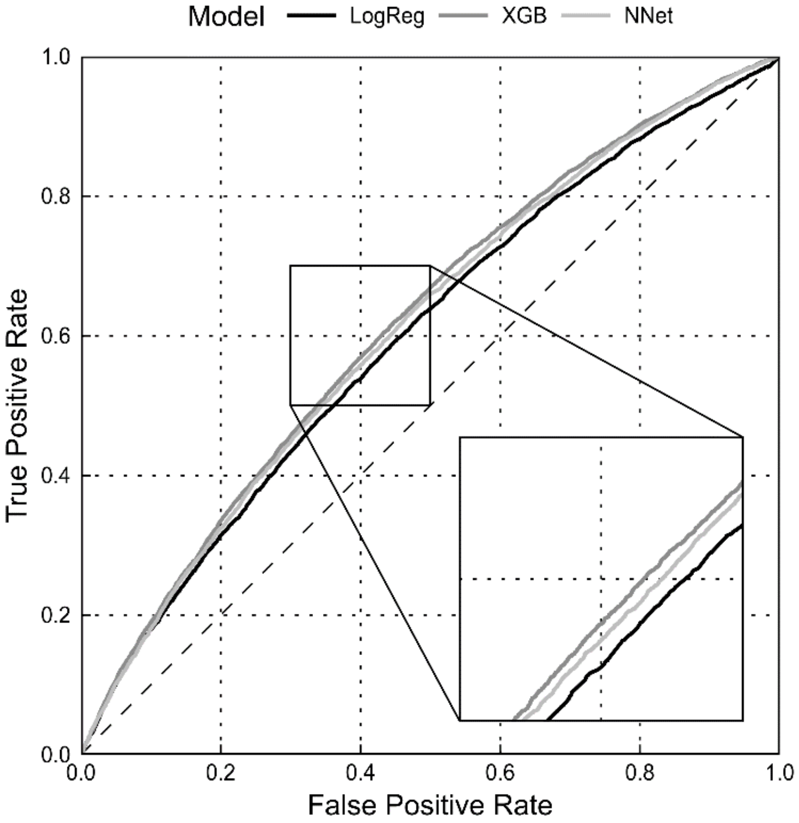




Hit rate varies between: **Athletes** > disciplines > shooting modes > shot number

Results – ML Models





All models show low predictive power

Complex models show about the same performance as LogReg

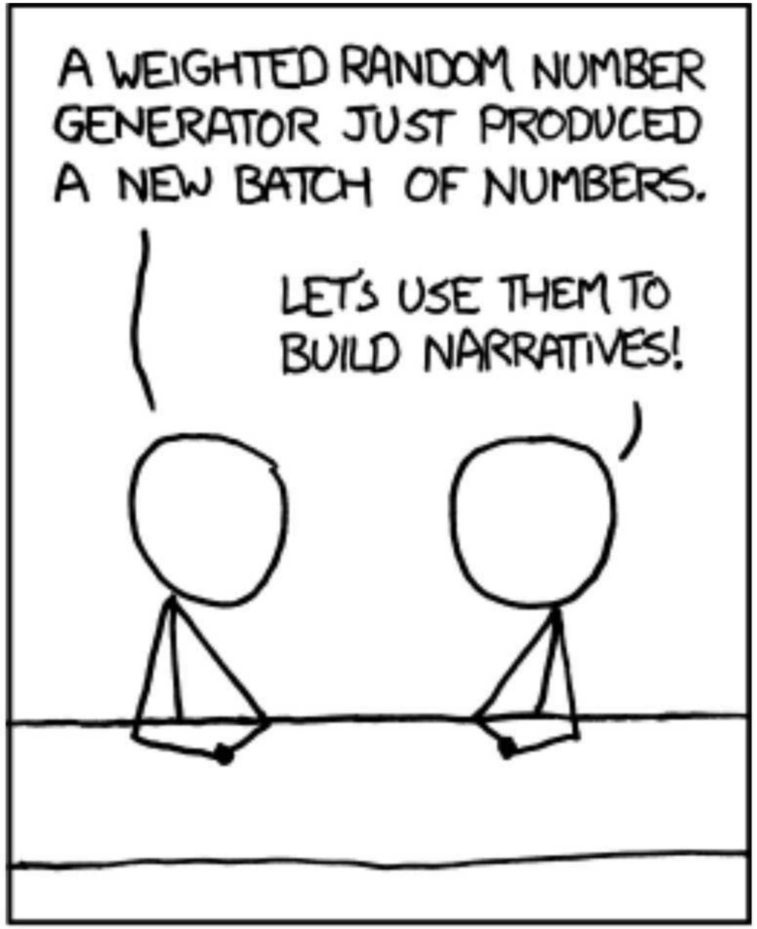
Discussion

- Largest differences in hit rates between athletes

- **Individual preceding mode-specific hit rate holds almost all predictive information**

- Individual shots can be modelled as Bernoulli trial
  → explains observed variation

- High random influence in competition results (± 1-2 hits / competition)


A WEIGHTED RANDOM NUMBER GENERATOR JUST PRODUCED A NEW BATCH OF NUMBERS.
LET'S USE THEM TO BUILD NARRATIVES!

ALL SPORTS COMMENTARY

xkcd

Selina was really concentrated today, so she was able to access her true potential. She is a professional athlete!

A Swiss coach

Irene was losing her confidence midway where she started to think too much, the pressure was too high on the last two shots.

Another Swiss coach

The hot hand [in basketball] is a massive and widespread cognitive illusion.

Daniel Kahneman

2
22 LAURA DAHLMEIER GER 14:17.1
0
CZE 14:41.6
2 DORIN HABERT FRA 15:21.4
3 FROLINA KOR 15:33.3

Hit Prob. = 84%

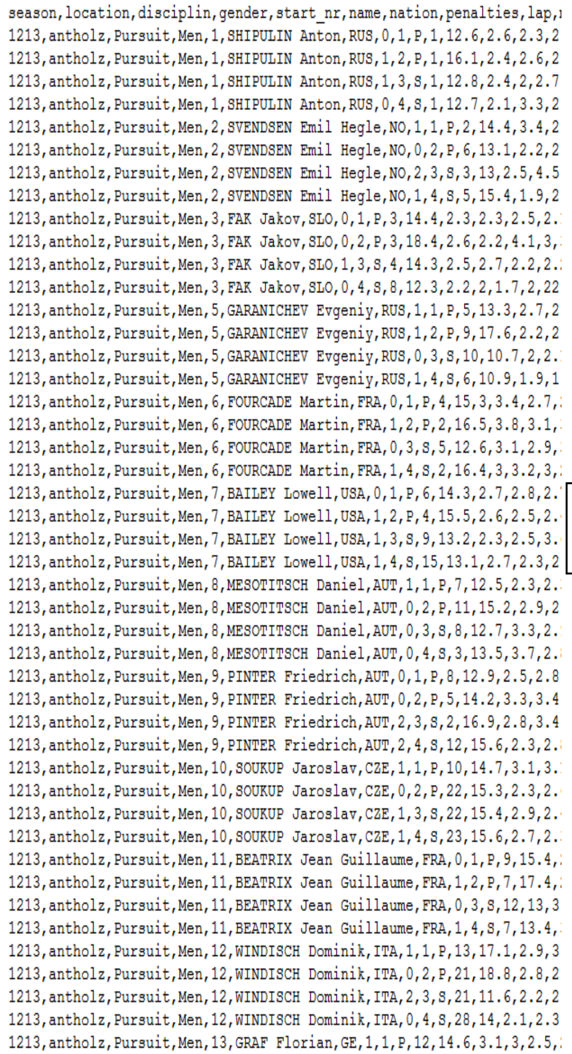
Final thoughts…

- Not everyone understands probabilities / randomness
- Not everyone is interested in the complexity of your models

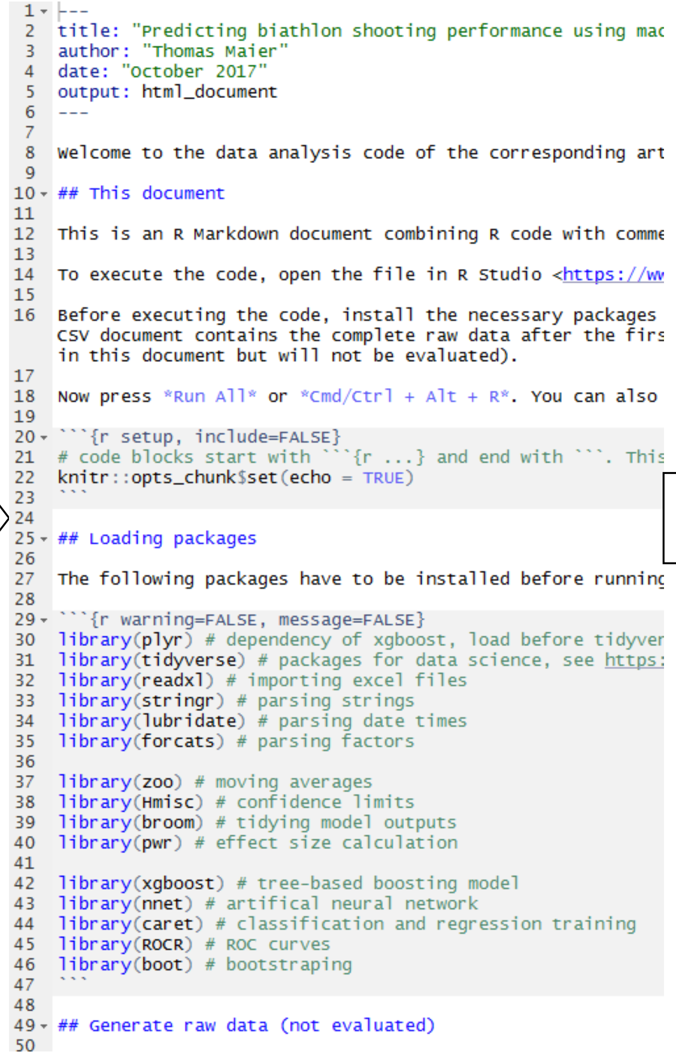
- Coaches / customers / executives / the public …

… are interested in stories and specific instructions

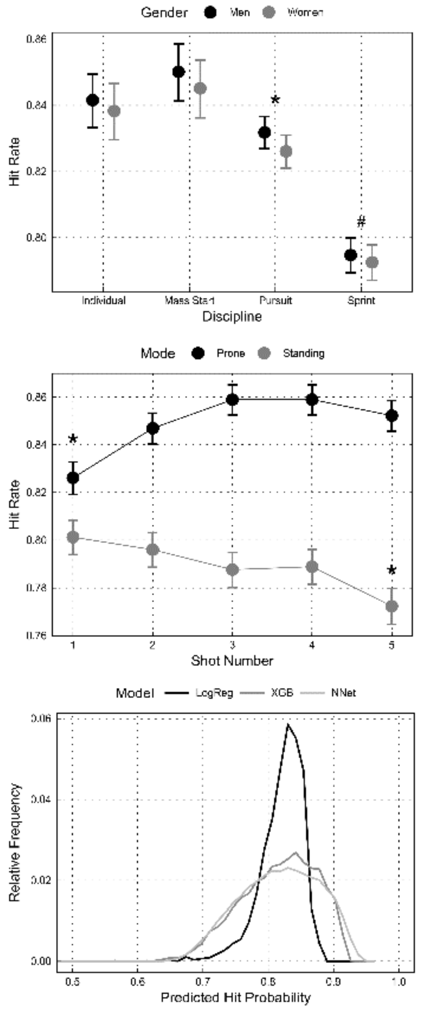
Thomas Maier
Senior Data Scientist
Datahouse AG
Alte Börse - Zürich
044 289 92 63
thomas.maier@datahouse.ch



```
season,location,disciplin,gender,start_nr,name,nation,penalties,lap,
1213,antholz,Pursuit,Men,1,SHIPULIN Anton,RUS,0,1,P,1,12.6,2.6,2.3,2
1213,antholz,Pursuit,Men,1,SHIPULIN Anton,RUS,1,2,P,1,16.1,2.4,2.6,2
1213,antholz,Pursuit,Men,1,SHIPULIN Anton,RUS,1,3,S,1,12.8,2.4,2,2.7
1213,antholz,Pursuit,Men,1,SHIPULIN Anton,RUS,0,4,S,1,12.7,2.1,3.3,2
1213,antholz,Pursuit,Men,2,SVENDSEN Emil Hegle,NO,1,1,P,2,14.4,3.4,2
1213,antholz,Pursuit,Men,2,SVENDSEN Emil Hegle,NO,0,2,P,6,13.1,2.2,2
1213,antholz,Pursuit,Men,2,SVENDSEN Emil Hegle,NO,2,3,S,3,13,2.5,4.5
1213,antholz,Pursuit,Men,2,SVENDSEN Emil Hegle,NO,1,4,S,5,15.4,1.9,2
1213,antholz,Pursuit,Men,3,FAK Jakov,SLO,0,1,P,3,14.4,2.3,2.3,2.5,2.
1213,antholz,Pursuit,Men,3,FAK Jakov,SLO,0,2,P,3,18.4,2.6,2.2,4.1,3,
1213,antholz,Pursuit,Men,3,FAK Jakov,SLO,1,3,S,4,14.3,2.5,2.7,2.2,2.
1213,antholz,Pursuit,Men,3,FAK Jakov,SLO,0,4,S,8,12.3,2.2,2,1.7,2,22
1213,antholz,Pursuit,Men,5,GARANICHEV Evgeniy,RUS,1,1,P,5,13.3,2.7,2
1213,antholz,Pursuit,Men,5,GARANICHEV Evgeniy,RUS,1,2,P,9,17.6,2.2,2
1213,antholz,Pursuit,Men,5,GARANICHEV Evgeniy,RUS,0,3,S,10,10.7,2,2.
1213,antholz,Pursuit,Men,5,GARANICHEV Evgeniy,RUS,1,4,S,6,10.9,1.9,1
1213,antholz,Pursuit,Men,6,FOURCADE Martin,FRA,0,1,P,4,15,3,3.4,2.7,
1213,antholz,Pursuit,Men,6,FOURCADE Martin,FRA,1,2,P,2,16.5,3.8,3.1,
1213,antholz,Pursuit,Men,6,FOURCADE Martin,FRA,0,3,S,5,12.6,3.1,2.9,
1213,antholz,Pursuit,Men,6,FOURCADE Martin,FRA,1,4,S,2,16.4,3,3.2,3,
1213,antholz,Pursuit,Men,7,BAILEY Lowell,USA,0,1,P,6,14.3,2.7,2.8,2.
1213,antholz,Pursuit,Men,7,BAILEY Lowell,USA,1,2,P,4,15.5,2.6,2.5,2.
1213,antholz,Pursuit,Men,7,BAILEY Lowell,USA,1,3,S,9,13.2,2.3,2.5,3.
1213,antholz,Pursuit,Men,7,BAILEY Lowell,USA,1,4,S,15,13.1,2.7,2.3,2
1213,antholz,Pursuit,Men,8,MESOTITSCH Daniel,AUT,1,1,P,7,12.5,2.3,2.
1213,antholz,Pursuit,Men,8,MESOTITSCH Daniel,AUT,0,2,P,11,15.2,2.9,2
1213,antholz,Pursuit,Men,8,MESOTITSCH Daniel,AUT,0,3,S,8,12.7,3.3,2.
1213,antholz,Pursuit,Men,8,MESOTITSCH Daniel,AUT,0,4,S,3,13.5,3.7,2.
1213,antholz,Pursuit,Men,9,PINTER Friedrich,AUT,0,1,P,8,12.9,2.5,2.8
1213,antholz,Pursuit,Men,9,PINTER Friedrich,AUT,0,2,P,5,14.2,3.3,3.4
1213,antholz,Pursuit,Men,9,PINTER Friedrich,AUT,2,3,S,2,16.9,2.8,3.4
1213,antholz,Pursuit,Men,9,PINTER Friedrich,AUT,2,4,S,12,15.6,2.3,2.
1213,antholz,Pursuit,Men,10,SOUKUP Jaroslav,CZE,1,1,P,10,14.7,3.1,3.
1213,antholz,Pursuit,Men,10,SOUKUP Jaroslav,CZE,0,2,P,22,15.3,2.3,2.
1213,antholz,Pursuit,Men,10,SOUKUP Jaroslav,CZE,1,3,S,22,15.4,2.9,2.
1213,antholz,Pursuit,Men,10,SOUKUP Jaroslav,CZE,1,4,S,23,15.6,2.7,2.
1213,antholz,Pursuit,Men,11,BEATRIX Jean Guillaume,FRA,0,1,P,9,15.4,
1213,antholz,Pursuit,Men,11,BEATRIX Jean Guillaume,FRA,1,2,P,7,17.4,
1213,antholz,Pursuit,Men,11,BEATRIX Jean Guillaume,FRA,0,3,S,12,13,3
1213,antholz,Pursuit,Men,11,BEATRIX Jean Guillaume,FRA,1,4,S,7,13.4,
1213,antholz,Pursuit,Men,12,WINDISCH Dominik,ITA,1,1,P,13,17.1,2.9,3
1213,antholz,Pursuit,Men,12,WINDISCH Dominik,ITA,0,2,P,21,18.8,2.8,2
1213,antholz,Pursuit,Men,12,WINDISCH Dominik,ITA,2,3,S,21,11.6,2.2,2
1213,antholz,Pursuit,Men,12,WINDISCH Dominik,ITA,0,4,S,28,14,2.1,2.3
1213,antholz,Pursuit,Men,13,GRAF Florian,GE,1,1,P,12,14.6,3.1,3,2.5,
```



~~~
---
title: "Predicting biathlon shooting performance using mac
author: "Thomas Maier"
date: "October 2017"
output: html_document
---

Welcome to the data analysis code of the corresponding art

## This document

This is an R Markdown document combining R code with comme

To execute the code, open the file in R Studio <https://ww

Before executing the code, install the necessary packages
CSV document contains the complete raw data after the firs
in this document but will not be evaluated).

Now press *Run All* or *Cmd/Ctrl + Alt + R*. You can also

```{r setup, include=FALSE}
# code blocks start with ```{r ...} and end with ```. This
knitr::opts_chunk$set(echo = TRUE)
```

## Loading packages

The following packages have to be installed before running

```{r warning=FALSE, message=FALSE}
library(plyr) # dependency of xgboost, load before tidyver
library(tidyverse) # packages for data science, see https:
library(readxl) # importing excel files
library(stringr) # parsing strings
library(lubridate) # parsing date times
library(forcats) # parsing factors

library(zoo) # moving averages
library(Hmisc) # confidence limits
library(broom) # tidying model outputs
library(pwr) # effect size calculation

library(xgboost) # tree-based boosting model
library(nnet) # artifical neural network
library(caret) # classification and regression training
library(ROCR) # ROC curves
library(boot) # bootstraping
```

## Generate raw data (not evaluated)

~~~


Gender ● Men ● Women
Hit Rate
*
#
Individual
Mass Start
Pursuit
Sprint
Discipline
Mode ● Prone ● Standing
Hit Rate
*
*
Shot Number
Model LogReg XGB NNet
Relative Frequency
Predicted Hit Probability