Sports Data Analytics

Joost N. Kok, Leiden Centre of Data Science

Universiteit Leiden
The Netherlands

Leiden University Medical Center

LCDS
LEIDEN CENTRE OF DATA SCIENCE

Dutch Research Agenda

- Value of Sports
- Performance
- More people, more active, more often

Why data analytics?

the next step in pursuit of medals and talent promotion

Canadian Tire Data Analysts To Help Put Athletes On The Podium

Canadian Tire Corporation's deep experience in data analytics extends to sports by helping Own the Podium identify future Olympic and Paralympic medallists.

Through a three-year partnership with Own the Podium, the team is building on Canadian Tire's decades-long history in predictive modeling. Canadian Tire's data analysts are using various data from global sports competitions, dating back to the 1930s, to provide new insights and next-generation predictive modelling to identify which athletes, given their current performance, are most likely to medal in future events. The information will also be used to help coaches and athletes refine their training programs and identify opportunities on the path to the podium.

EIS Director of Performance Solutions provides fascinating insight into the future of sport science

"By 2045, we may well have reached the age of 'information doping'. Right now, the sports world is clamouring over the competitive advantages potentially offered by data analytics. In 30 years the growth of artificial intelligence (AI) may be dominating information doping.

"A huge number of sports are underpinned by decision-making. The difference between winners and losers is the decisions they make. As AI advances there is a very real possibility that the ability of the human brain to detect stimuli, process them and produce a solution may be exceeded by machines."

Why data analytics?

fueling recreational sports, fraud detection and rehabilitation

MIT Technology Review

sas THE POWER TO KNOW.

MLS

A new scientific approach

the unknown known

Why now?

- ✓ More data than ever: physiological, spatial-temporal, image, sensor, etc.
- ✓ Scientific progress in data science: machine learning/statistics, HCI, etc.
- ✓ Emerging data management: FAIR, semantic web, integration etc.
- ✓ Increasing realtime computer power, downsizing of hardware, etc.

The prospect of unveiling the “Unknown Known”

- ✓ Data integration from multiple sources, multi format, unstructured data
- ✓ From single sample and population average to personalized outcome
- ✓ Pattern recognition, hypothesis setting and predictive modelling
- ✓ User interaction & context relevant feedback / representation

Sport Data Center

- Five research lines centered around Sports Data Analytics:

1. Elite Sports
2. Recreational Sports
3. Adaptive Sports & Rehabilitation
4. Sport Policy & Economics
5. Fraud & Risks

TU Delft

AISS Amsterdam Institute of Sport Science

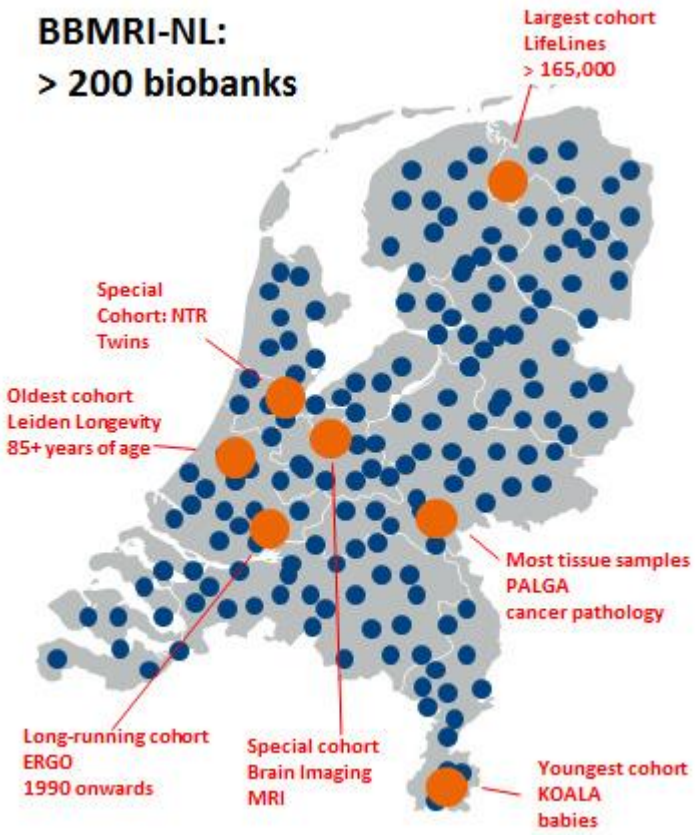
Universiteit Leiden

LUMC Leiden University Medical Center

SDC (Sport Data Center)

- Data Scientists in core
- Surrounded by sport domain knowledge
- Data & Sport field labs
- Sports data
- Open Network

Elite Sports



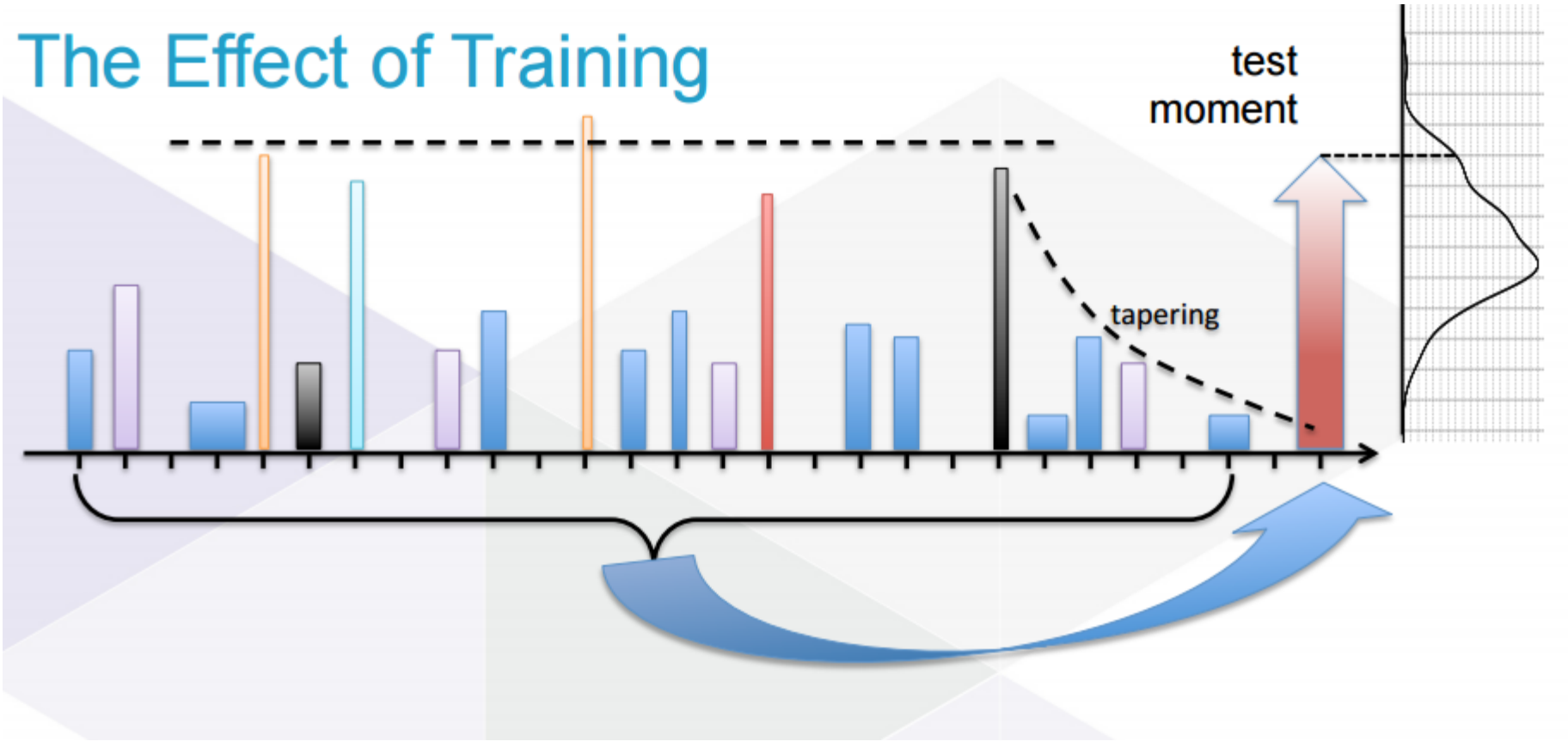
Transfer of Knowledge

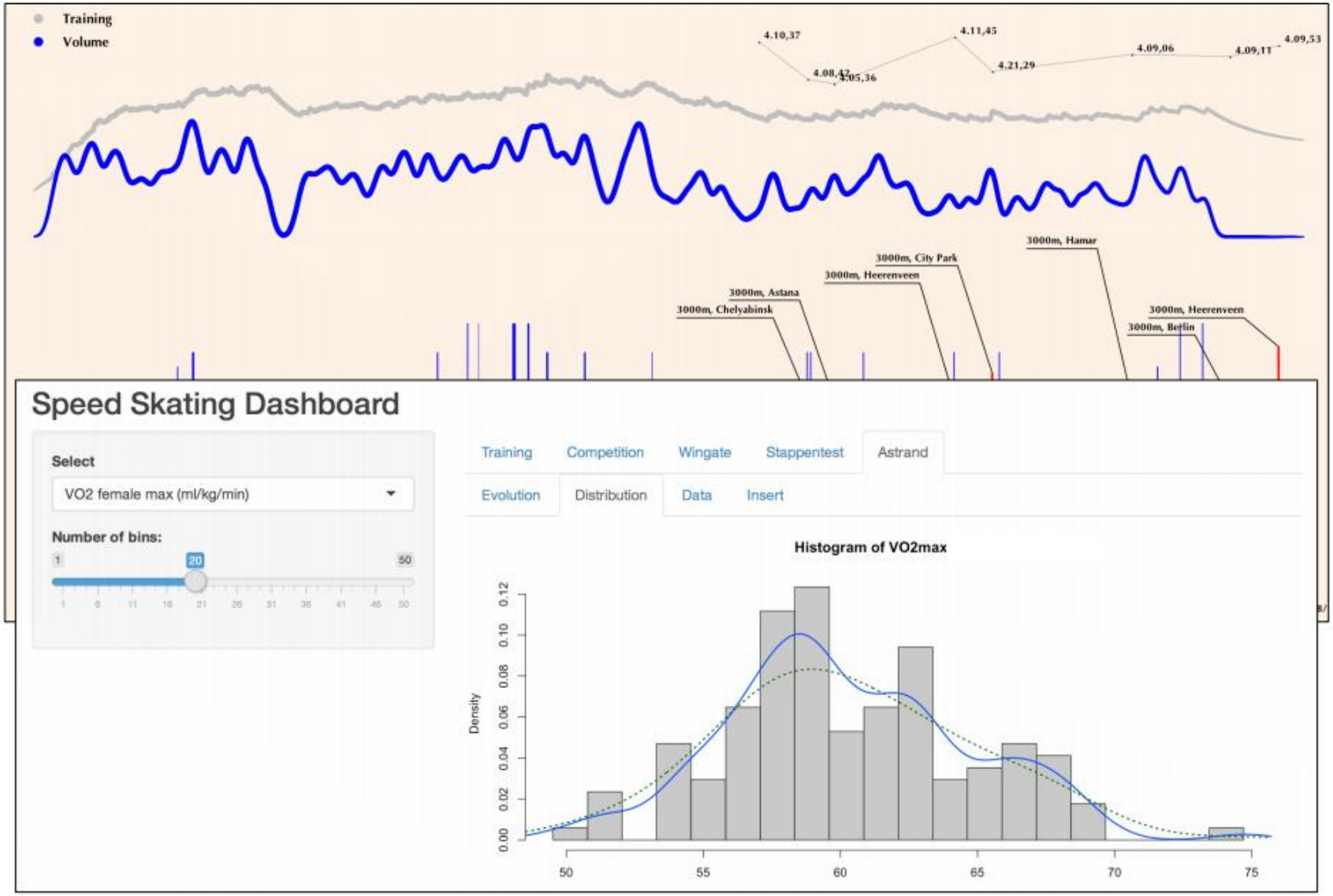
Historical Training Data

- 15 years of data collected
- Some 40 athletes, currently nine: seven men, two women
- Some 30 Olympic medals + numerous championships
- Daily training details
  - Morning and afternoon training
  - Six days per week
  - Training type, intensity (subjective), duration, load
- Roughly bi-weekly physical test, aerobic, anaerobic
- Competition data
  - Corrected for track-differences

Work by Arno Knobbe

The Effect of Training




Training
Volume
4.10,37
4.08,42
4.05,36
4.11,45
4.21,29
4.09,06
4.09,11
4.09,53
3000m, Chelyabinsk
3000m, Astana
3000m, Heerenveen
3000m, City Park
3000m, Hamar
3000m, Heerenveen
3000m, Berlin
Speed Skating Dashboard
Select
VO2 female max (ml/kg/min)
Number of bins:
1
20
50
Training
Competition
Wingate
Stappentest
Astrand
Evolution
Distribution
Data
Insert
Histogram of VO2max
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
50
55
60
65
70
75

Using sensors to measure playground dynamics

18 January 2016

How do playground interactions contribute to children's social competence? Developmental psychologists Carolien Rieffe (Leiden University) and Guida Veiga (University of Évora, Portugal) joined forces with the Leiden Institute of Advanced Computer Science to investigate this. A paper on their study is currently in press.

Academic staff

Carolien Rieffe
Professor by Special Appointment

Joost Kok
professor

Arno Knobbe
postdoc / guest

Organisation

Science

Leiden Centre of Data Science

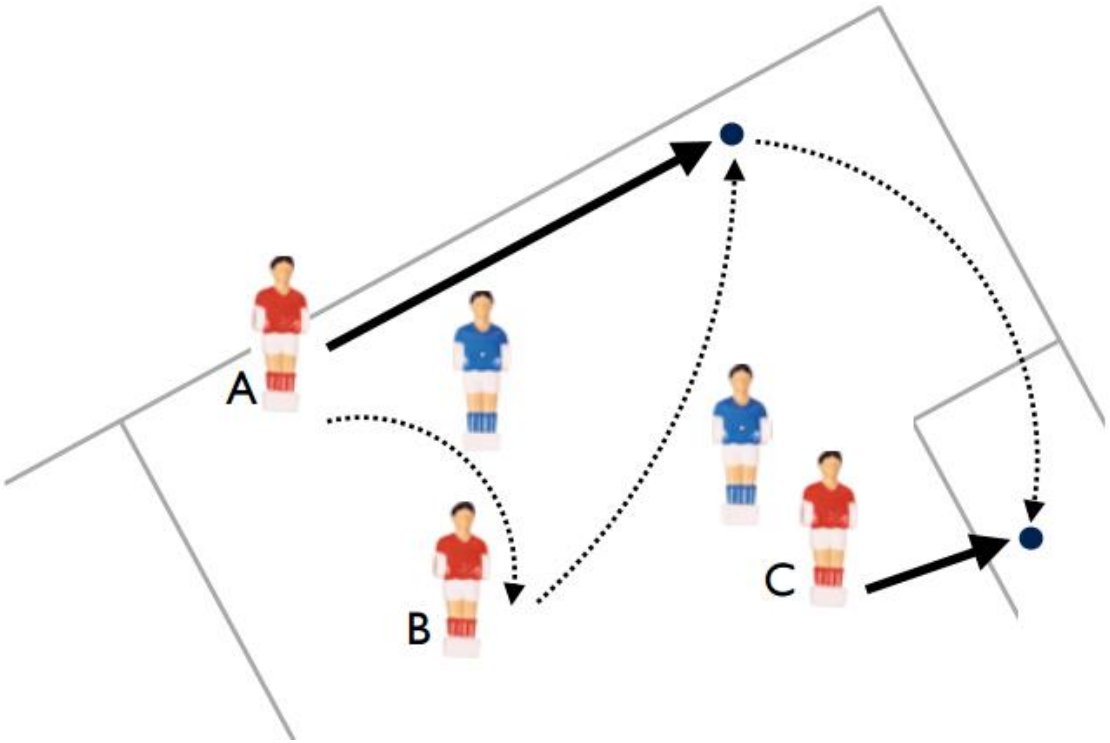
Leiden Institute of Advanced Computer Science (LIACS)

Social and Behavioural Sciences

Psychology

Developmental and Educational Psychology

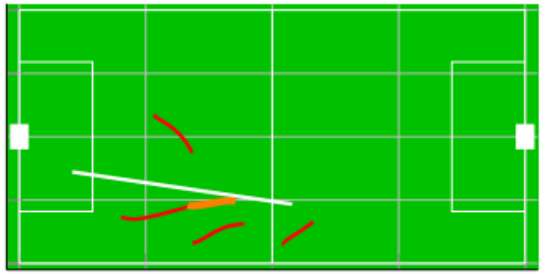
- Pattern = "interesting" event
- E.g., A plays 1-2 with B and crosses to C



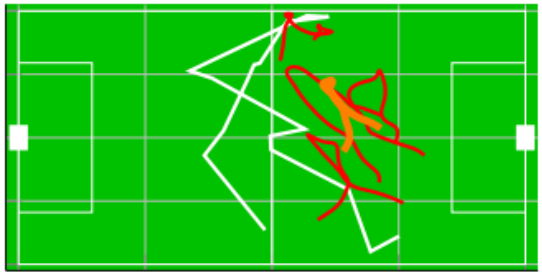
Slides by Ulf Brefeld

- Individual level
- Group level
- Team level

- 4 defence players
  → game initiations



- 4 offence players
  → scoring opportunities



Different Scales

- Analyse opponent tactics
- Detect strengths/weaknesses in strategy
- Automatic game plans
- Serious games / training
- Player scouting
- Improved media coverage

Policy & Economics


15:17
John
Delete
From 0 to 21 years
Recommended area Choose...
Record details
Growth references Netherlands
Gender Boy
Date of birth 9-17-2013
Gestational age 40 wks 2 days
Height of father 183 cm (-0.14 SDS, P44.6)
Height of mother 178 cm (1.15 SDS, P87.4)
Target height 186.47 cm (0.37 SDS, P64.5)
Measurements
9-17-2015 (2.000 yrs)
110 cm / 23 kg / BMI: 19.0
9-17-2014 (1.000 yrs)
Height (cm) →
200
180
160
140
120
100
80
60
0
2
4
6
8
10
12
14
16
18
20
Age (years) →
Vijfde Landelijke Groeistudie, www.tno.nl/groei
Growth Analyser

SPORT DATA CENTER | MAP | ATHLETES | COUNTRIES | PREDICTIONS | ABOUT

Choose discipline(s):

Horizontal Bar Men

Choose Athlete(s):

Epke Zonderland

- Emre Arabacioglu
- Endriadi
- Enkhmunkh Munkhjargal
- Enrico Pozzo
- Enrique Navarro
- Enzo Bernardoni
- Epke Zonderland

Rating History | Rating by Date | Age | potential winners | athletes overview | Info | About

Personal information for Epke Zonderland

Name: Epke Zonderland

Discipline: Horizontal Bar

Category: Men

Age: 29.5

Country: Netherlands

ELO: 1651.31

Rank: 2

SPORT DATA CENTER | MAP | ATHLETES | COUNTRIES | PREDICTIONS | ABOUT

Choose discipline:

Horizontal Bar Men

Top Athletes Now

How many:

None 3 5 10

Top Athletes Ever

How many:

None 3 5 10

Other Athletes

Choose Athlete(s):

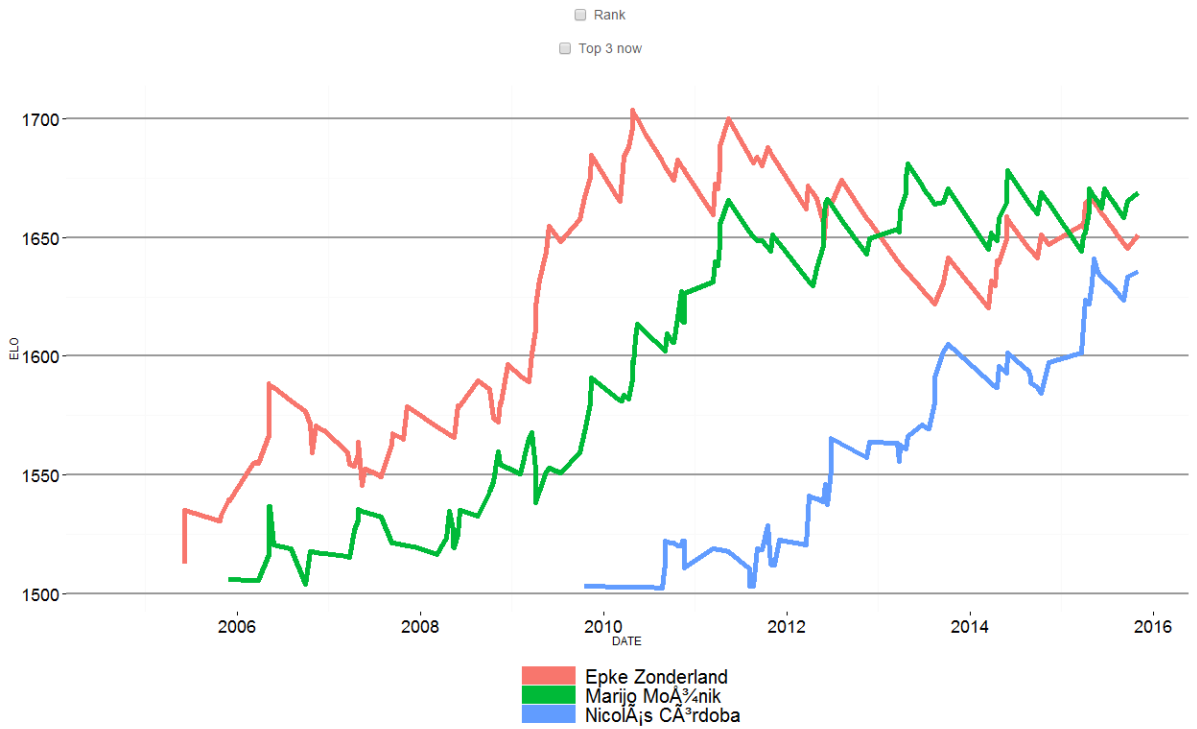
Please pick one or more athletes

- Adam Rzepa
- Adam Wong
- Adan Santos
- Adelin Kotrong
- Adham Al-Sqour
- Adickxon Trejo
- Aditya Rana

Rating History | Rating by Date | Age | potential winners | athletes overview | Info | About

Rank

Top 3 now



SPORT DATA CENTER | MAP | ATHLETES | COUNTRIES | PREDICTIONS | ABOUT

choose discipline:

Horizontal Bar Men

Top Countries Now

How many:

None 3 5 10

Top Countries Ever

How many:

None 3 5 10

Other Countries

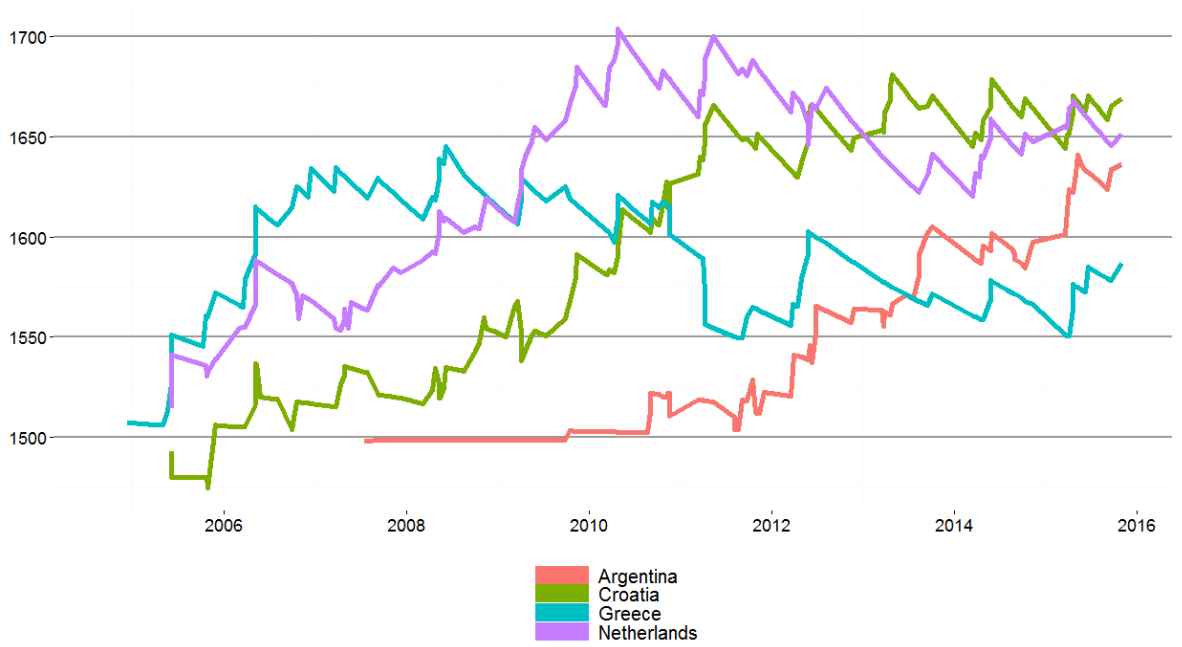
Choose country:

Greece

Slovenia
Slovakia
Finland
Bulgaria
Venezuela
Austria
Albania
Algeria

History per Discipline | History per Discipline by Date | Disciplines vs | Info Countries | About

vs top 3 now



Fraud & Risks


Sun Sport
EIGHT YEARS AFTER DENYING OUR STORY, DJOKOVIC ADMITS:
I WAS offered £110k to
GUIDO'S SPL-ASH LANDING
TENNIS FIX SCANDAL

Data: about 2G (2.000.000.000) records

Year	General Practitioner	Dentists	Pharmacy	Mental Health Care	Fysiotherapy	Hospitals
2009				1.165.353		
2010	262.584.340	69.297.896	191.744.461	1.218.992	55.575.780	16.412.981
2011	304.654.670	68.449.999	208.515.505	1.251.854	57.068.264	17.150.880
2012	313.926.643	51.934.447	219.200.187		53.549.109	15.407.850
Totaal	**881.165.653**	**189.682.342**	**619.460.153**	**3.636.199**	**166.193.153**	**48.971.711**

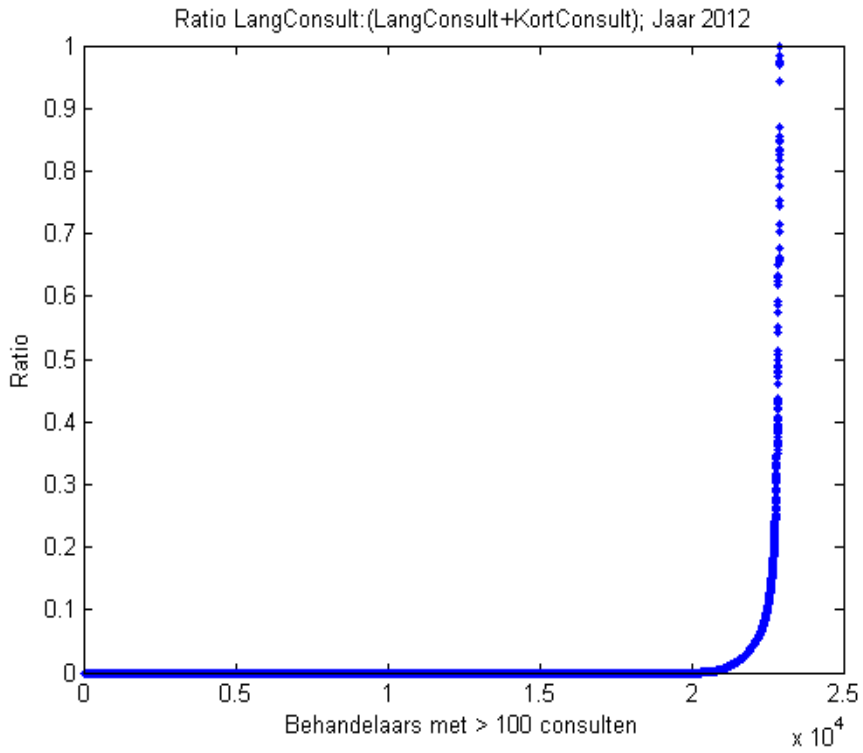
Methodology

- Formulate tests, with help of domain experts:
  - Hard rules
  - Soft rules
- Translate results of tests into estimates of losses

Soft Rules

- Certain things are allowed, but when they happen too frequently, then they become suspected

- The 99% threshold:
  - the last percentile is suspected
  - the loss is estimated as the extra cost "above" 99th percentile



	Total Amount Meuro	Hard Rules Meuro	Soft Rules Meuro	Hard and Soft Rules Meuro	Percentage %
GP	2619	15,4	6,2	21,6	0,8
Dental Care	2180	0,7	1,0	1,7	0,1
Pharmacy	5280	10,5	0,9	11,4	0,2
Mental Health	3980	4,2	-	4,2	0,2
Physio-therapy	1446	0,6	11,1	11,7	0,8
Hospitals	16676	11,9	54,7	66,6	0,4
Total	32181	43,3	73,9	117,2	0,4

Fraud and Risks

- 100% Control
  - Check all data (instead of sampling)
  - Powerful when different kinds of data are combined (from molecular level to sport events

Scientific Challenges

- Match the ability to gather data by the ability to process it in a safe and meaningful way
- Compare, evaluate and integrate different methods from Data Science
- Develop standards in techniques and methodology
- Use data to gain real life competitive advantage
- Challenge existing training methods

TU Delft

AIS Amsterdam Institute of Sport Science

Universiteit Leiden

LUMC Leiden University Medical Center

Make Data Science meaningful to end-users!!

SDC (Sport Data Center)

- Five research lines centered around Sports Data Analytics:

1. Elite Sports
2. Recreational Sports
3. Adaptive Sports & Rehabiliation
4. Policy & Economics
5. Fraud & Risks

TU Delft

Amsterdam Institute of Sport Science

Universiteit Leiden

LUMC Leiden University Medical Center