Winnen met Data Science

Joost N. Kok, Leiden Centre of Data Science

Universiteit Leiden
The Netherlands

Overzicht

Data Science

Sport en Bewegen

Discover the world at Leiden University

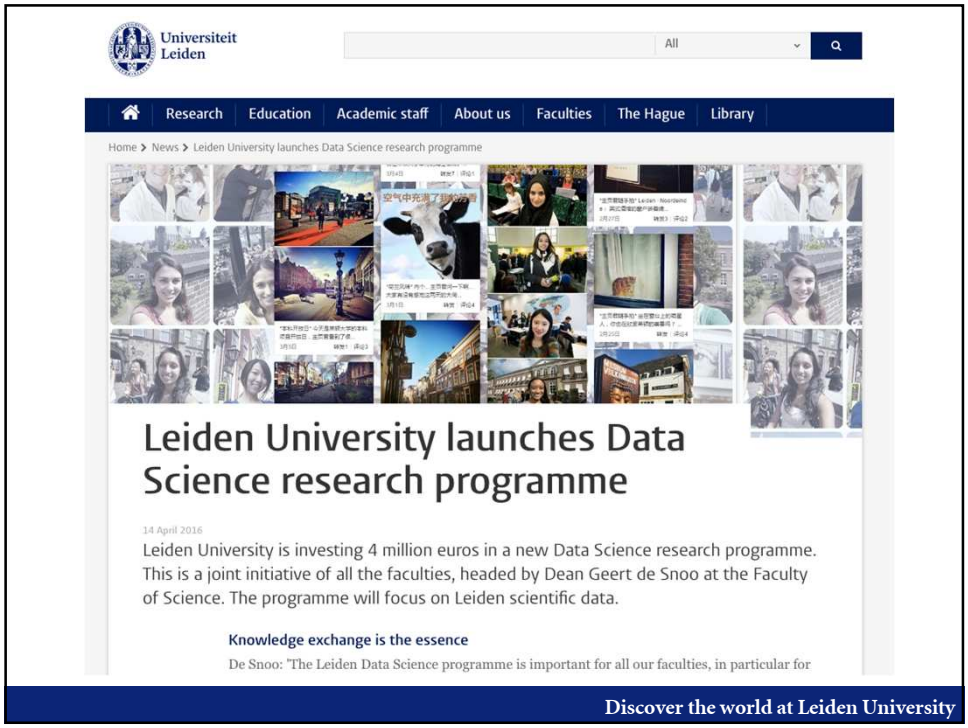
LIACS

- The Computer Science institute of Leiden University

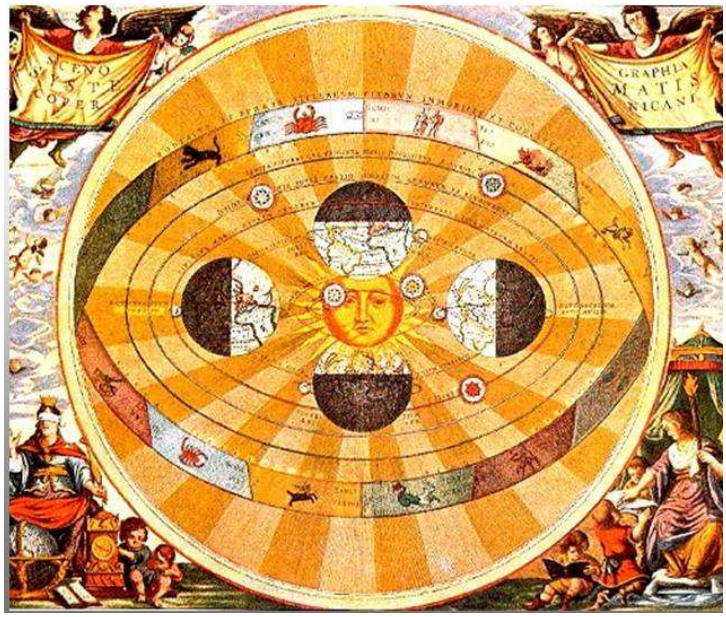
Discover the world at Leiden University

LCDS Profile

- **Data Science Dossier:** http://onderzoeksgebieden.leidenuniv.nl/data-science
- Focus op
  - Fundamenteel
  - Standaarden
  - Wetenschappelijke Data

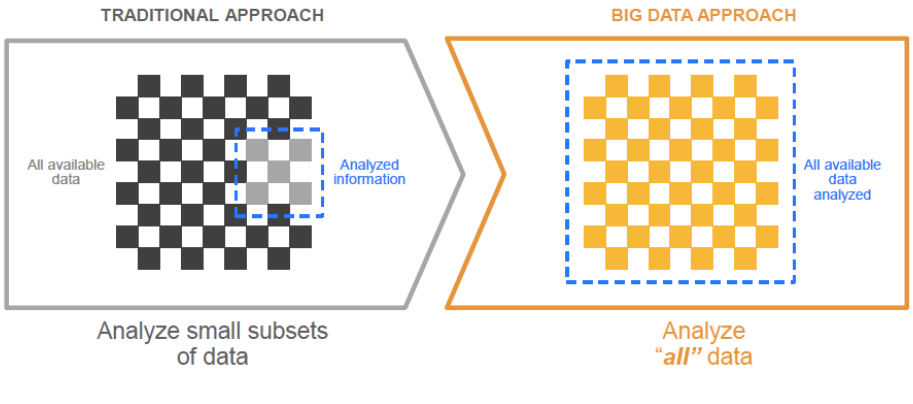

Universiteit Leiden
All
Research
Education
Academic staff
About us
Faculties
The Hague
Library
Home > News > Leiden University launches Data Science research programme
Leiden University launches Data Science research programme
14 April 2016
Leiden University is investing 4 million euros in a new Data Science research programme. This is a joint initiative of all the faculties, headed by Dean Geert de Snoo at the Faculty of Science. The programme will focus on Leiden scientific data.
Knowledge exchange is the essence
De Snoo: "The Leiden Data Science programme is important for all our faculties, in particular for
Discover the world at Leiden University

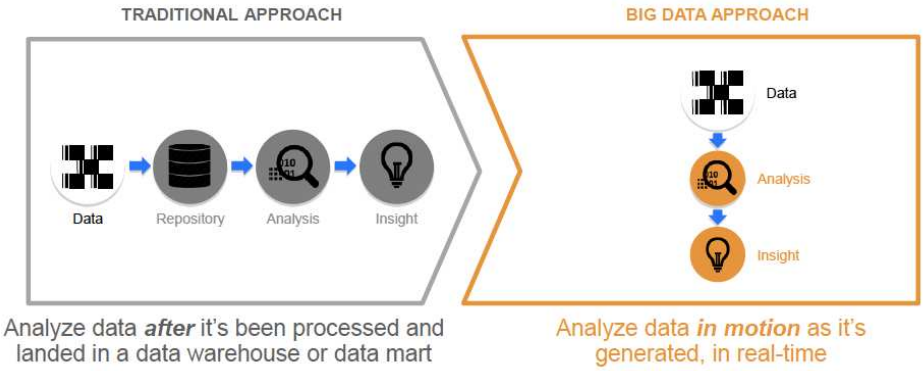

Discover the world at Leiden University

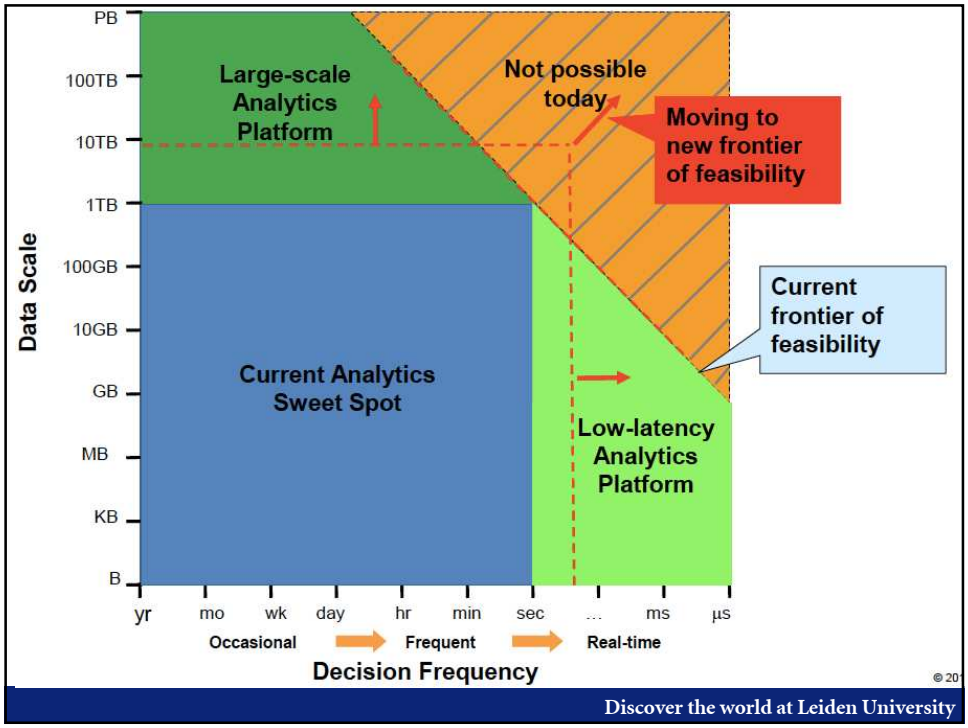


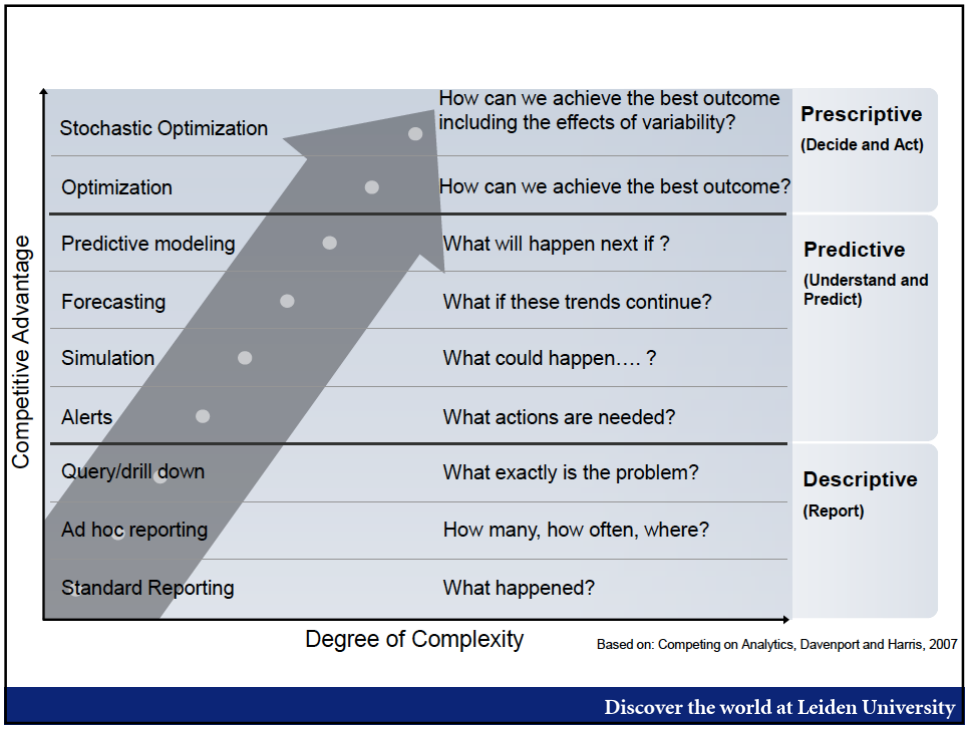
Volume	Velocity	Variety	Veracity
Data at Scale	**Data in Motion**	**Data in Many Forms**	**Data Uncertainty**
Terabytes to exabytes of existing data to process (e.g. CRM, ERP data, etc.)	Streaming data, milliseconds to seconds to respond (e.g. data from smart sensors, mobile device, etc.)	Structured, unstructured, text, multimedia (e.g. rational DB, images, free text, video, etc.)	Uncertainty due to data error, inconsistency & incompleteness, ambiguities, model approximations (e.g. manual errors, device errors, models errors, etc.)

Source: IBM GTO 2012


TRADITIONAL APPROACH
BIG DATA APPROACH
All available data
Analyzed information
All available data analyzed
Analyze small subsets of data
Analyze "all" data

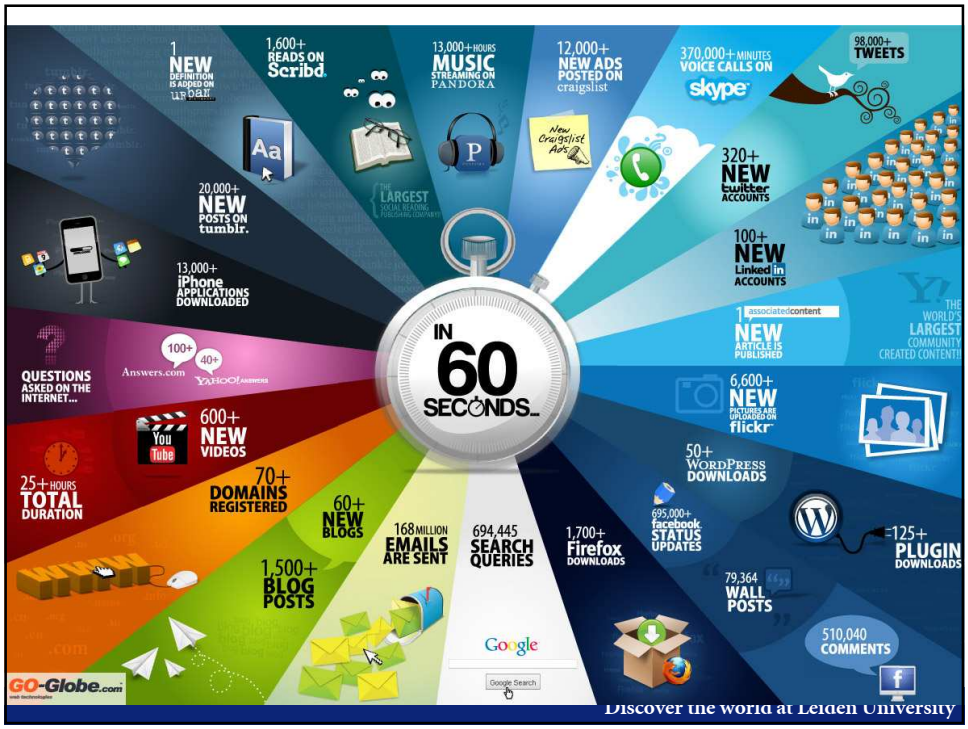

TRADITIONAL APPROACH
BIG DATA APPROACH
Data
Repository
Analysis
Insight
Data
Analysis
Insight
Analyze data after it's been processed and landed in a data warehouse or data mart
Analyze data in motion as it's generated, in real-time


PB
100TB
10TB
1TB
100GB
10GB
GB
MB
KB
B
Data Scale
Large-scale Analytics Platform
Not possible today
Moving to new frontier of feasibility
Current frontier of feasibility
Current Analytics Sweet Spot
Low-latency Analytics Platform
yr
mo
wk
day
hr
min
sec
…
ms
µs
Occasional
Frequent
Real-time
Decision Frequency


Competitive Advantage
Stochastic Optimization
How can we achieve the best outcome including the effects of variability?
Optimization
How can we achieve the best outcome?
Predictive modeling
What will happen next if ?
Forecasting
What if these trends continue?
Simulation
What could happen…. ?
Alerts
What actions are needed?
Query/drill down
What exactly is the problem?
Ad hoc reporting
How many, how often, where?
Standard Reporting
What happened?
Prescriptive
(Decide and Act)
Predictive
(Understand and Predict)
Descriptive
(Report)
Degree of Complexity
Based on: Competing on Analytics, Davenport and Harris, 2007

Data, data, data

- 90% of the world's data is created in the last two years
- 80% of the data is unstructured
- 1 trillion (10^12) connected devices generate 2.5 quintillion (10^18) bytes per day



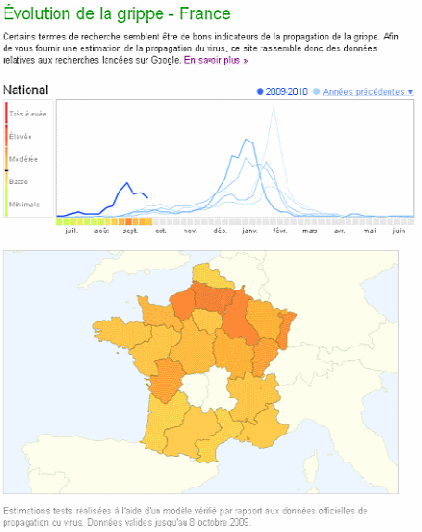
Value of Data

- http://tinyurl.com/valueofdata

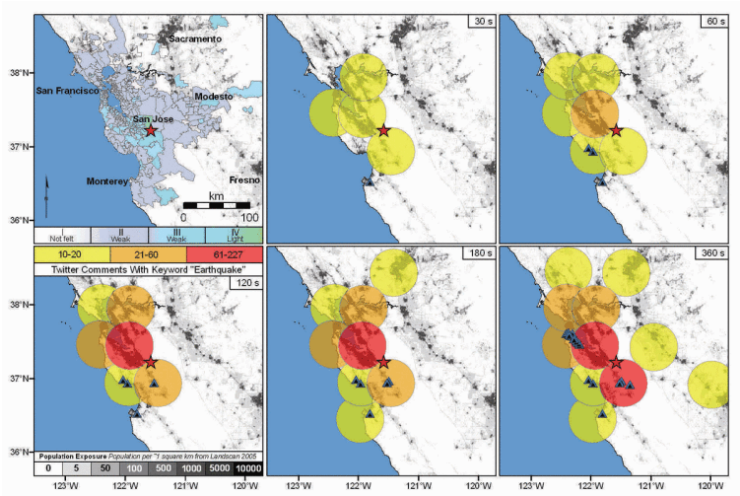
Computational Turn

- Correlaties volgen direct uit de data
- (Heel) veel data beschikbaar
- Grote vooruitgang in methoden
- Redeneer over de correlaties om de oorzakelijke verbanden te vinden

Google: verspreiding griep



Twitter: aardbeving



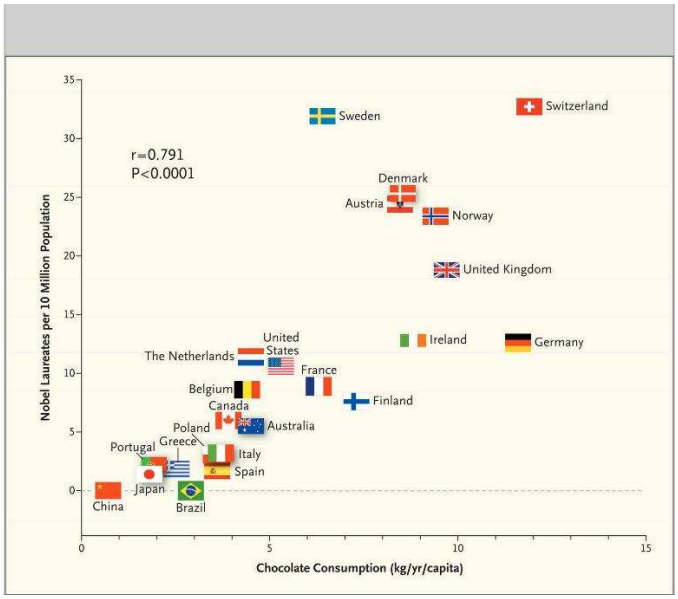
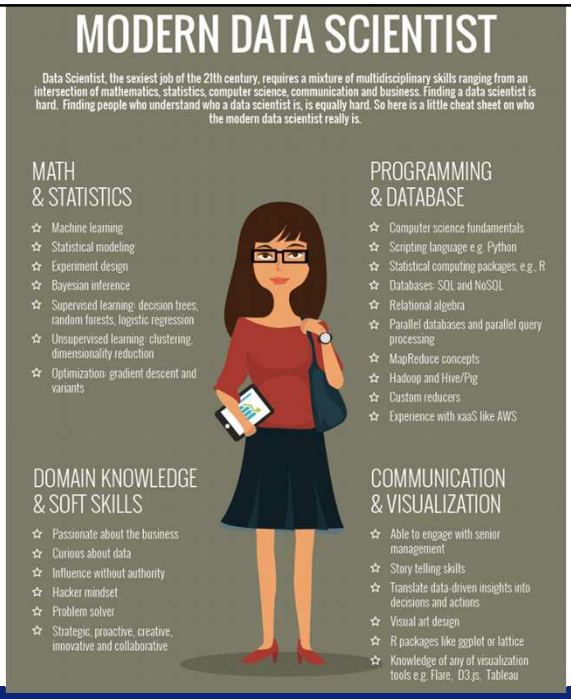


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Data Science

- Data geschikt maken
  - Managen, labelen, combineren, …
- Spelen met de data
  - Visualiseren, patroonherkenning, clusteren, ..
- Iets nuttigs mee doen
  - Classificeren, voorspellen, …



Data Science

- Data Science =
  Statistics + Computer Science + …

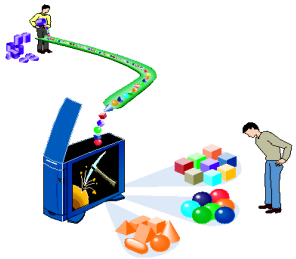
- More = Better ?
  - More data = more noise = more patterns
  - Methods have to be adapted
  - Data chain

Data Science

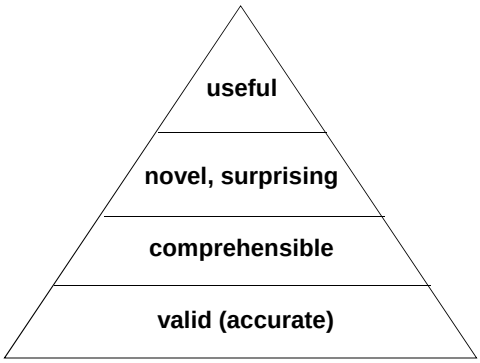
- Data geschikt maken
  - Managen, labelen, combineren,
- Spelen met de data
  - Visualiseren, patroonherkenning, clusteren,
- Iets nuttigs mee doen
  - Classificeren, voorspellen,

Data Mining definitions

- Secondary analysis of data
- Induction of understandable useful models and patterns from databases
- Algorithms for large quantities of data



- Data Mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data



Data Mining

- Data Mining is somewhat comparable to statistics (and often based on the latter), but takes it further in the sense that whereas
  - statistics aims more at validating given hypotheses,
  - in data mining often millions of potential patterns are generated and tested, in the hope of finding some that are potentially useful.

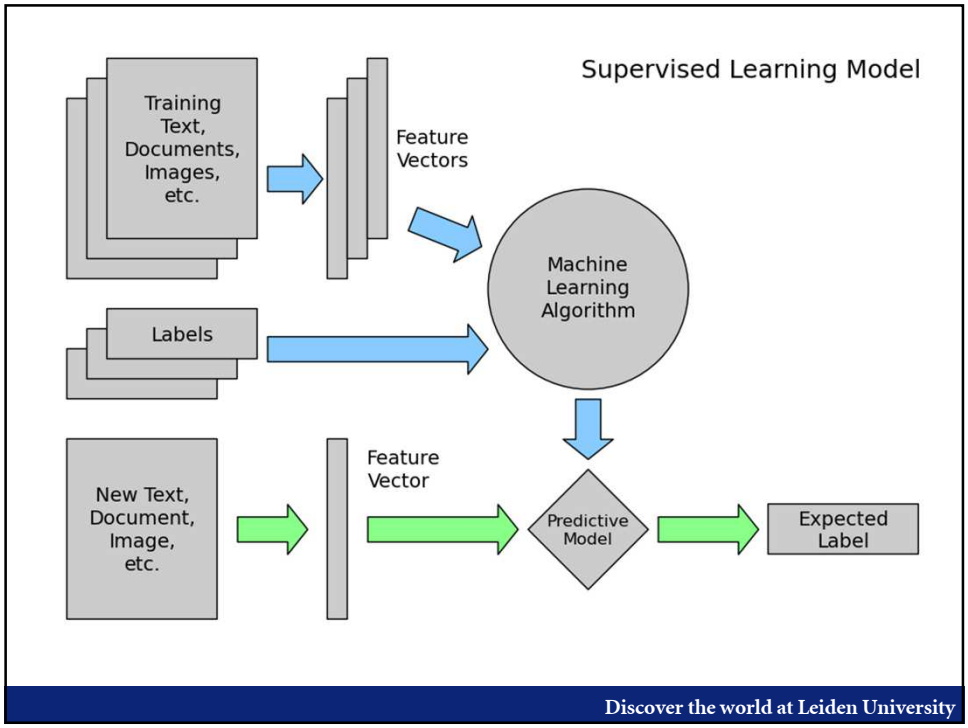
Meaningfulness of Patterns

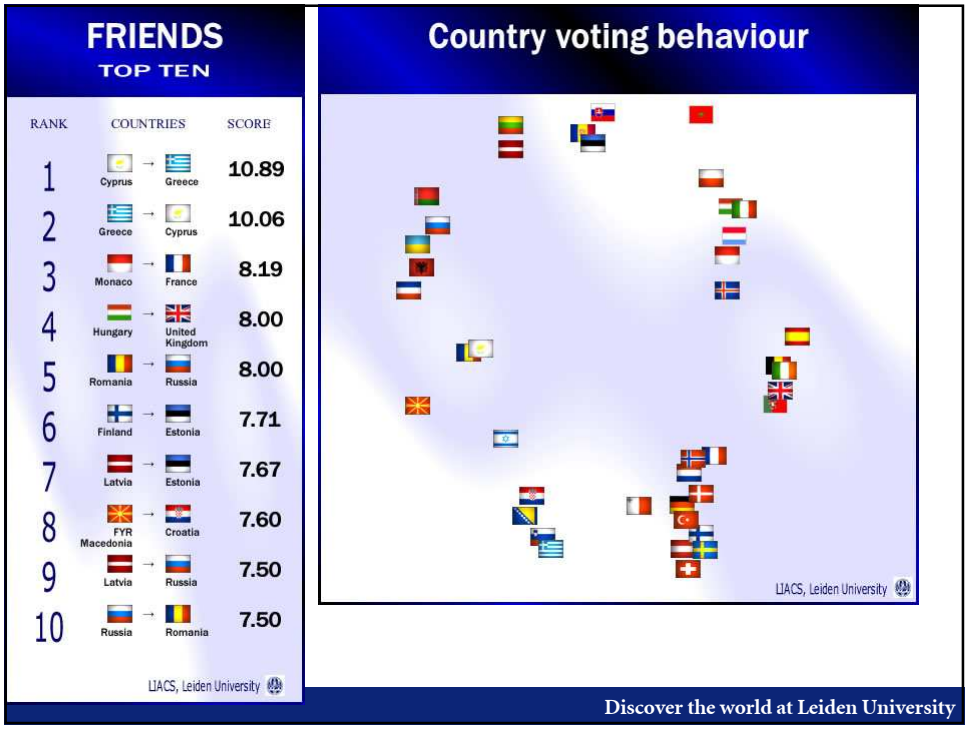
- A risk with Big-Data mining is that an analyst can "discover" patterns that are meaningless
- Bonferroni's principle

Meaningfulness of Patterns

Example:
•We want to find (unrelated) people who at least twice have stayed at the same hotel on the same day
- 10^9 people being tracked.
- 1000 days.
- Each person stays in a hotel 1% of the time (1 day out of 100)
- Hotels hold 100 people (so 10^5 hotels).
- If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?

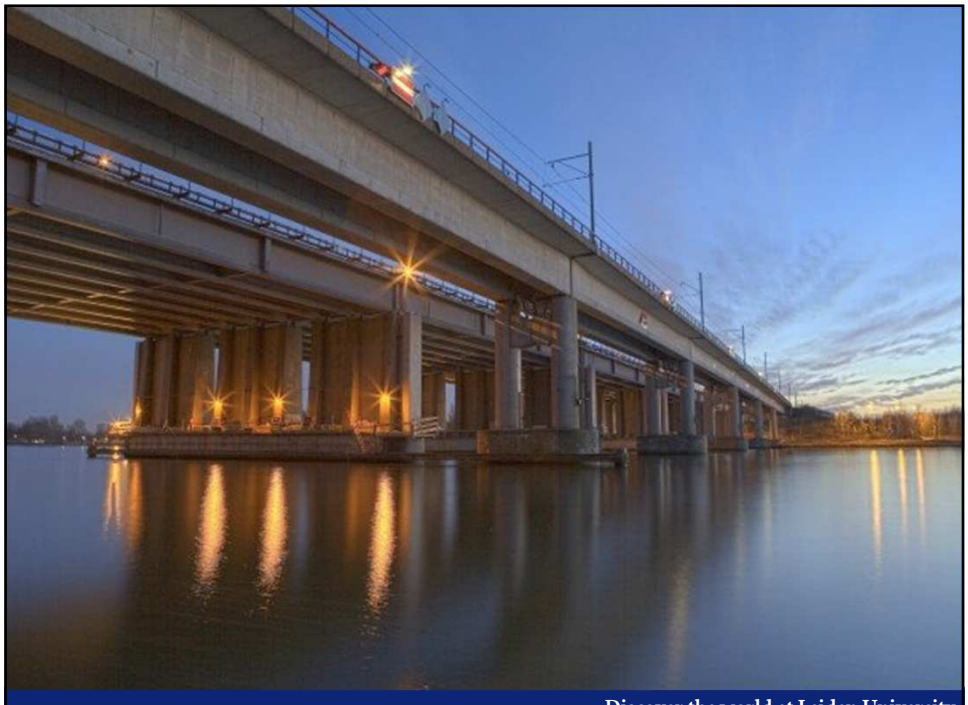
•Expected number of "suspicious" pairs of people:
- 250,000
- … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way


Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label
Discover the world at Leiden University


FRIENDS
TOP TEN
RANK
COUNTRIES
SCORE
1 Cyprus → Greece 10.89
2 Greece → Cyprus 10.06
3 Monaco → France 8.19
4 Hungary → United Kingdom 8.00
5 Romania → Russia 8.00
6 Finland → Estonia 7.71
7 Latvia → Estonia 7.67
8 FYR Macedonia → Croatia 7.60
9 Latvia → Russia 7.50
10 Russia → Romania 7.50
LIACS, Leiden University
Country voting behaviour
LIACS, Leiden University
Discover the world at Leiden University

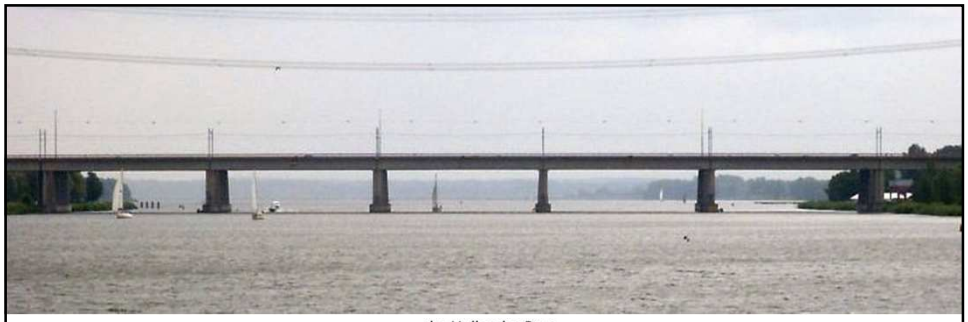
Monitoring

- Monitoring a Highway Bridge (Hollandse Brug)
- Health Care Data
- Sewers (SewerSense)
- School Children (Free Play)
- Animals (Oostvaardersplassen)
- Cohort Studies (Leiden 85+)
- Elderly people (Park Vossenberg)

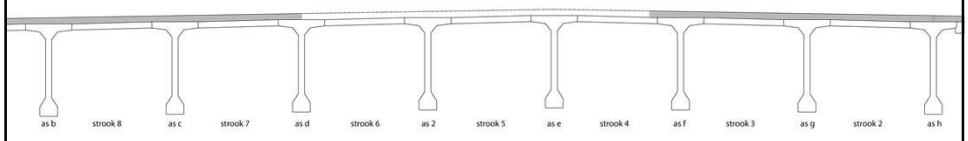


InfraWatch: Hollandse Brug
A6 between Amsterdam and Almere

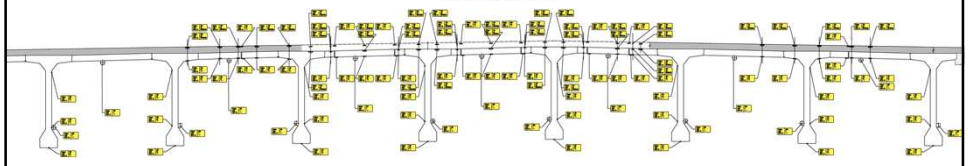
InfraWatch



the Hollandse Brug ...

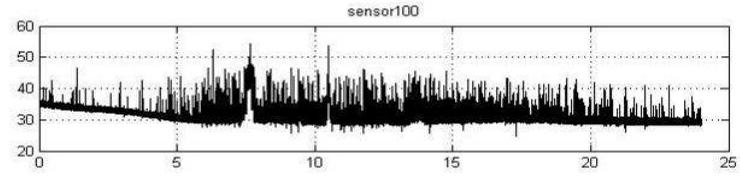


... and its sensors

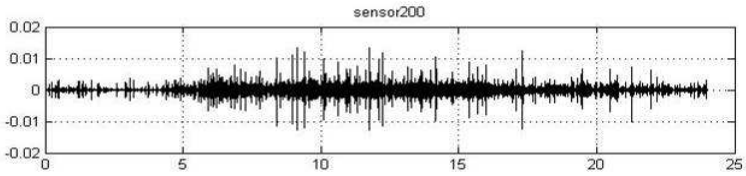


One day of data, multiple sensor types

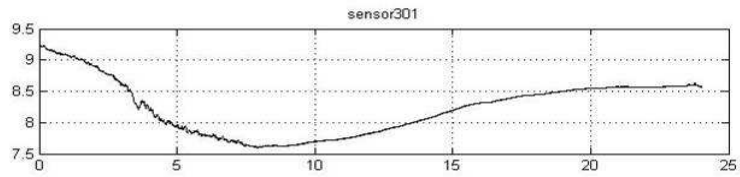
Strain



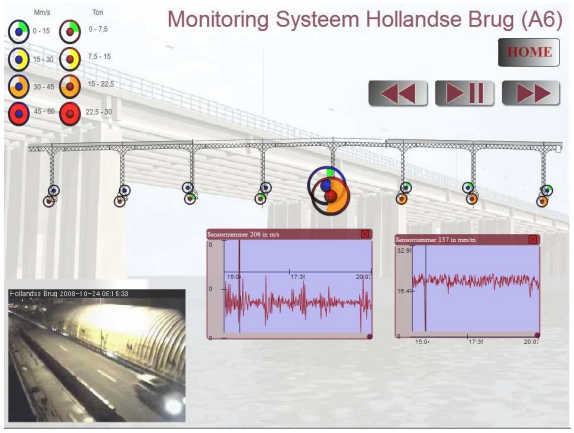
Vibration

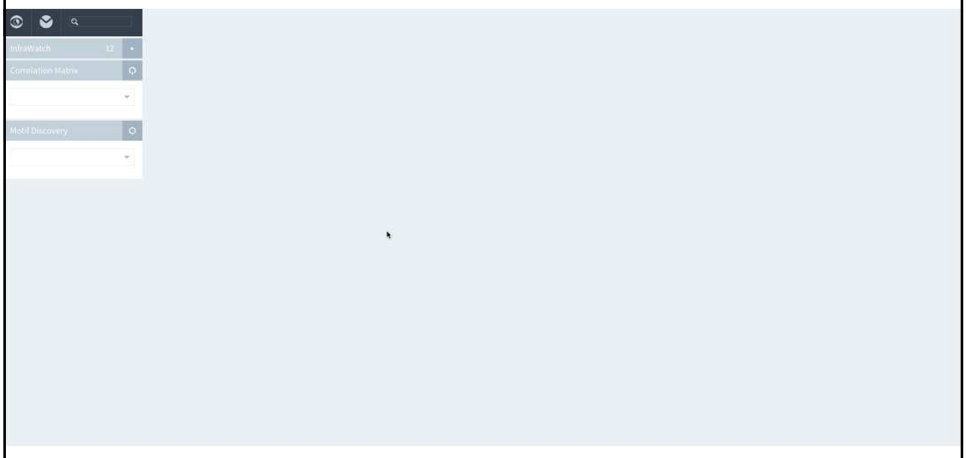


Temperature



Sensor Viewer

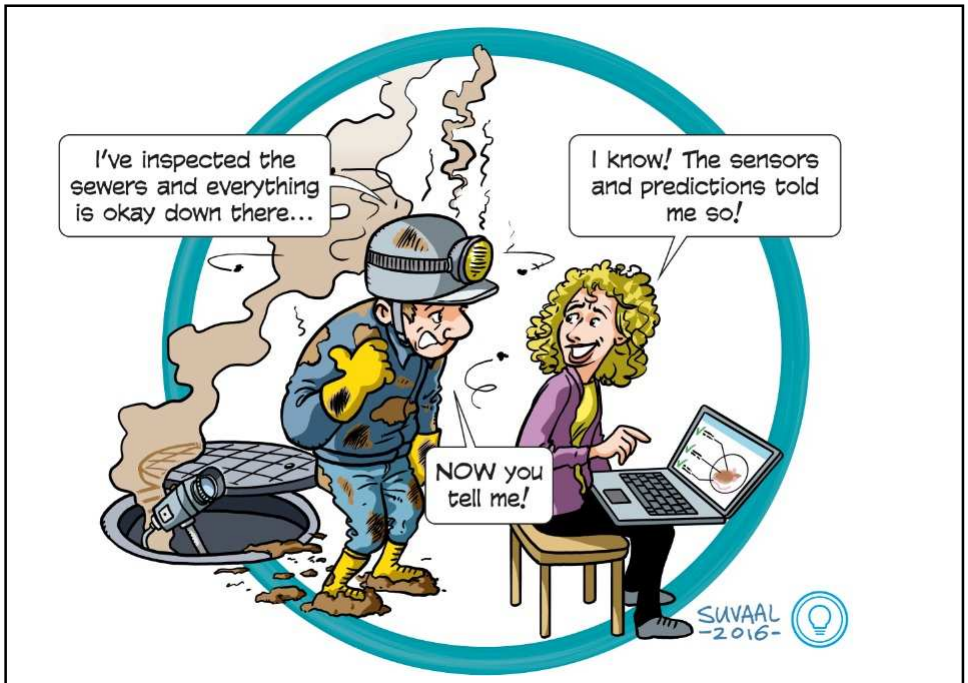




Using sensors to measure playground dynamics

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.



Data: about 2.000.000.000 records

Year	Family Doctors	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**

	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

Responsible Data Science

- [Fairness] Data Science without prejudice: How to avoid unfair conclusions even if they are true?
- [Accuracy] Data Science without guesswork: How to answer questions with a guaranteed level of accuracy?
- [Confidentiality] Data Science that ensures confidentiality: How to answer questions without revealing secrets?
- [Transparency] Data Science that provides transparency: How to clarify answers such that they become indisputable?

Ranking of differentially expressed genes

Gene	Score
$\text{gene}_{\sigma(1)}$	score 1
$\text{gene}_{\sigma(2)}$	score 2
$\text{gene}_{\sigma(3)}$	score 3
$\text{gene}_{\sigma(4)}$	score 4
.....	.....
$\text{gene}_{\sigma(100)}$	score 100
$\text{gene}_{\sigma(101)}$	score 101
.....	.....
$\text{gene}_{\sigma(9905)}$	score 9905

The genes are ordered in a ranked list (e.g. according to their differential expression between the classes).

The challenge is to extract meaning from this list, to describe subgroups with common properties.

.

Subgroup Discovery

- Population of individuals
- A property we are interested in
- Discover the subgroups of the population that are statistically "most interesting", i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest.

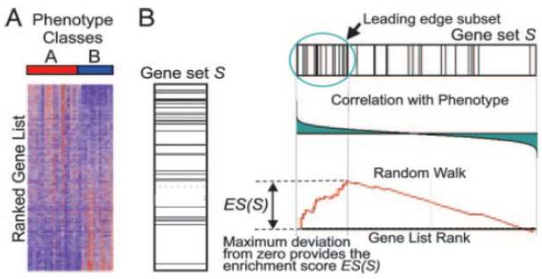
Subgroup Discovery

Discovery of gene **subgroups** which

- are "higher" in the ranked list
- can be compactly summarized using
  - knowledge (GO, ENTREZ, KEGG)
  - Interactions between genes
  - …

```
==========================================================================================


Rule 1
Distribution: [10, 0]
Score: 0.807160888001244
A participants: Adh1, Coq7, Cyp2c29, Cyp2c37, Cyp3a13, Cyp3a16, Hmox1, Nudt8, Ftmt, Dpyd]

All genes in the subgroup
        have the following properties:
                cellular_component(cytoplasmic part),
                molecular_function(transition metal ion binding),
                KEGG_pathway(Metabolism of cofactors and vitamins),


==========================================================================================


Rule 2
Distribution: [10, 0]
Score: 0.797726796885391
A participants: Aipl1, Cdk5, Cdkn1b, Fxn, Glrb, Itpr3, Mecp2, Myo7a, Sod1, Sod2]

All genes in the subgroup
        have the following properties:
                biological_process(sensory perception),
                cellular_component(intracellular membrane-bounded organelle),


==========================================================================================


Rule 3
Distribution: [11, 0]
Score: 0.771767724124038
A participants: Capg, Capza3, Cdkn1b, Mecp2, Mapt, Pax5, Prox1, Scin, Tmod3, Noc2l, Trim54]

All genes in the subgroup
        have the following properties:
                biological_process(regulation of organelle organization),
                cellular_component(intracellular organelle),


==========================================================================================
```


Enrichment Score



Data rond Sport en Bewegen
Sport Data Valley

Data Science and Sports

- Apply data science in new application domains yielding new results and insights in those domains.
- At the same time, the application domain serves as a source of inspiration for new data science research.

Vraagstukken

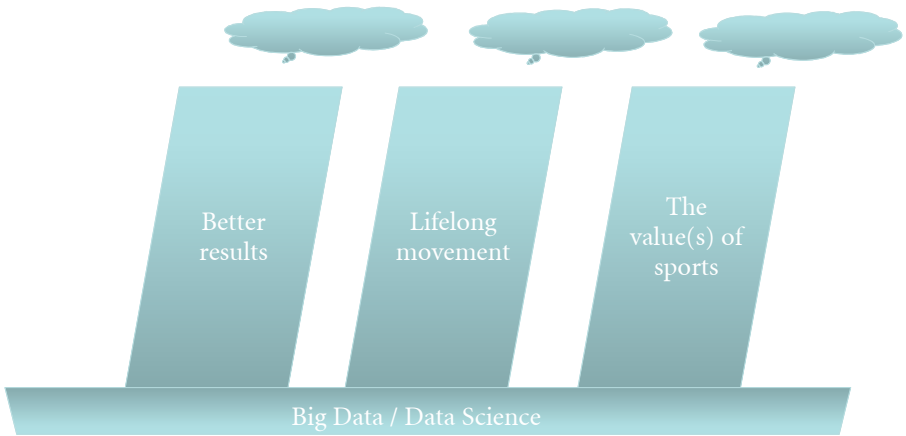
- Nederland staat niet in de top-10
- Kinderen zijn minder gezond
- Sociale cohesie neemt af
- Fraude en matchfixing
- Mensen bewegen te weinig
- Genezen door bewegen

Het kabinet verdubbelt de in 2016 afgesproken structurele intensivering voor de topsport van 10 miljoen euro per jaar naar 20 miljoen euro per jaar om meer kansen te bieden aan onze Olympische en Paralympische teams. Daarnaast komt er meer ruimte voor topsport talenten om onderwijs en topsport combineren.

Het kabinet trekt structureel 5 miljoen euro extra uit ter ondersteuning van de organisatie van sportevenementen in Nederland, waaronder EK's, WK's en multisportevenementen. Het initiatief voor de organisatie van een evenement ligt altijd bij de sport en haar partners.

Het kabinet zal het gesprek met de bonden aangaan over de handhaving van de openbare orde bij evenementen en de omgang met gedragingen tijdens evenementen en risicowedstrijden.

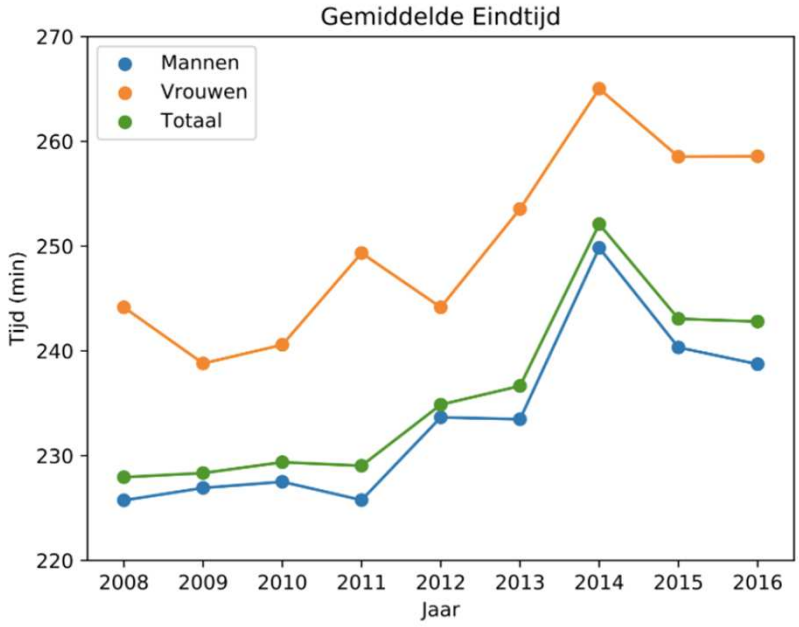
Het kabinet gaat steviger inzetten op de aanpak van dopinggebruik, matchfixing, corruptie en misbruik in de sport.

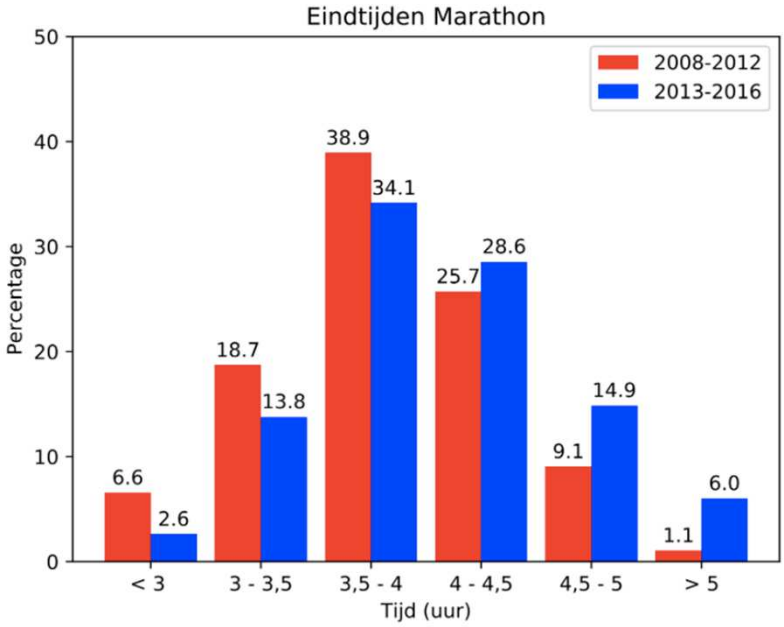


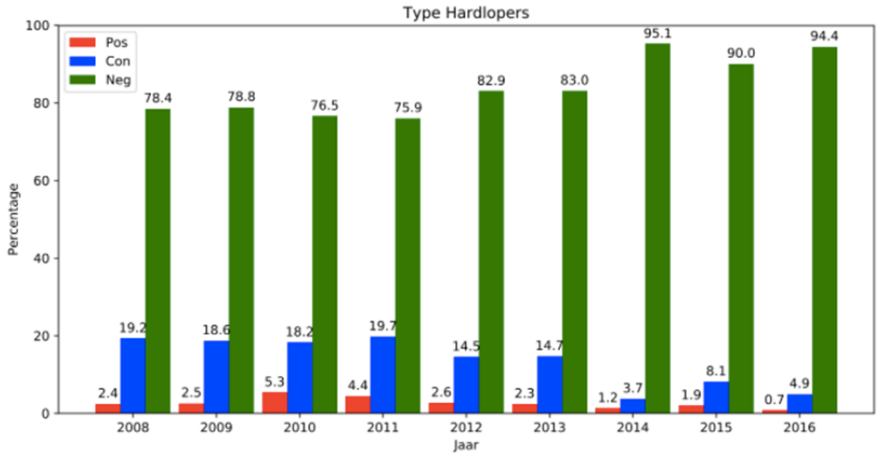
nationale wetenschaps agenda

Knowledge Agenda Sports and Movement

Tailored Advice


Gemiddelde Eindtijd
Mannen
Vrouwen
Totaal
Tijd (min)
270
260
250
240
230
220
2008
2009
2010
2011
2012
2013
2014
2015
2016
Jaar

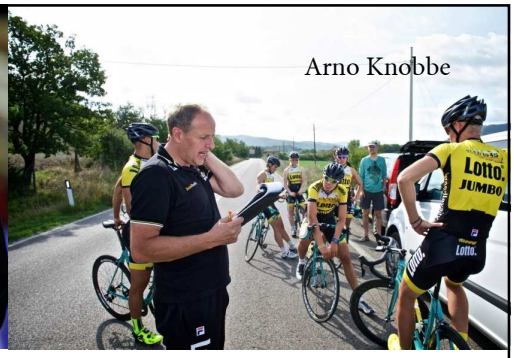

Eindtijden Marathon
2008-2012
2013-2016
Percentage
50
40
30
20
10
0
6.6
2.6
18.7
13.8
38.9
34.1
25.7
28.6
9.1
14.9
1.1
6.0
< 3
3 - 3,5
3,5 - 4
4 - 4,5
4,5 - 5
> 5
Tijd (uur)



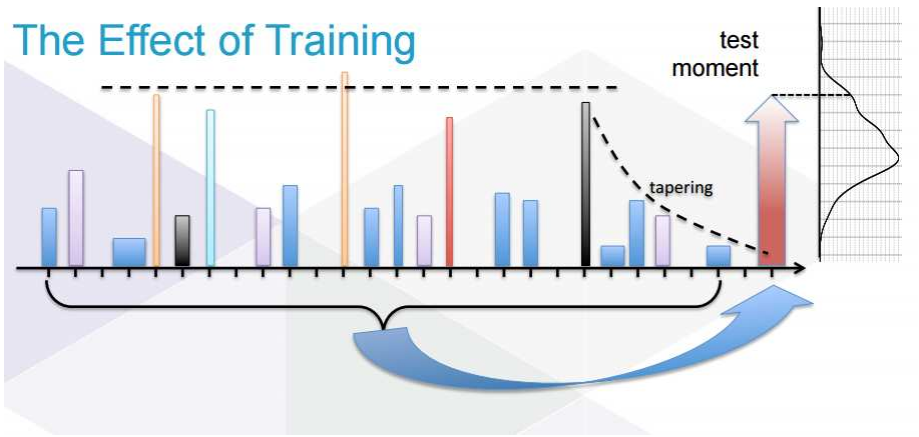
pos. : lopers die de tweede helft sneller lopen dan eerste helft.
con. : lopers die een constant tempo lopen.
neg. : lopers die de tweede helft langzamer lopen dan eerste helft.

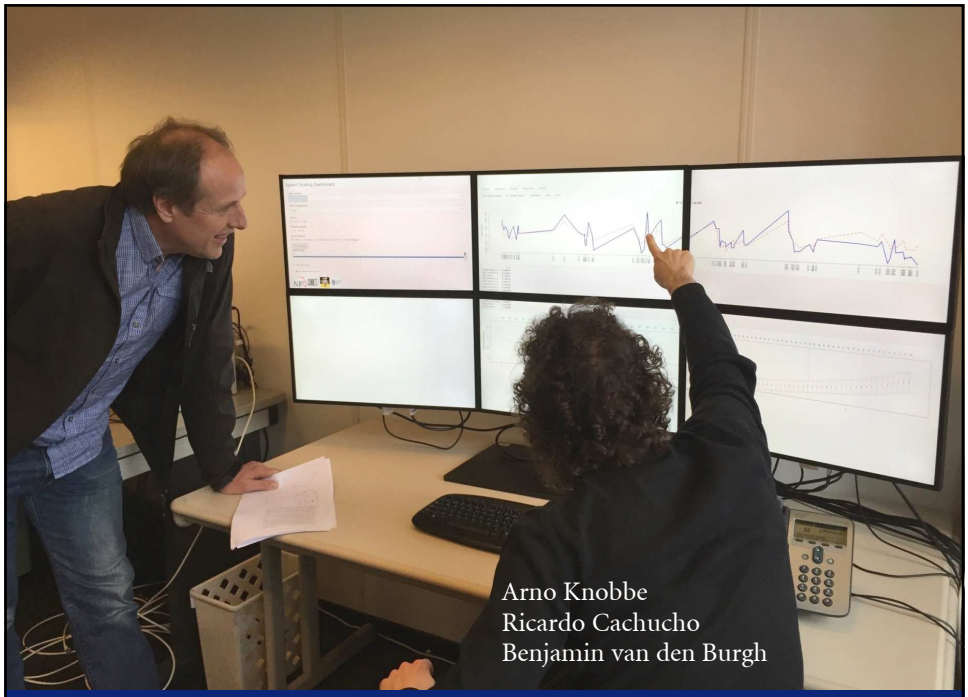
Loop Data

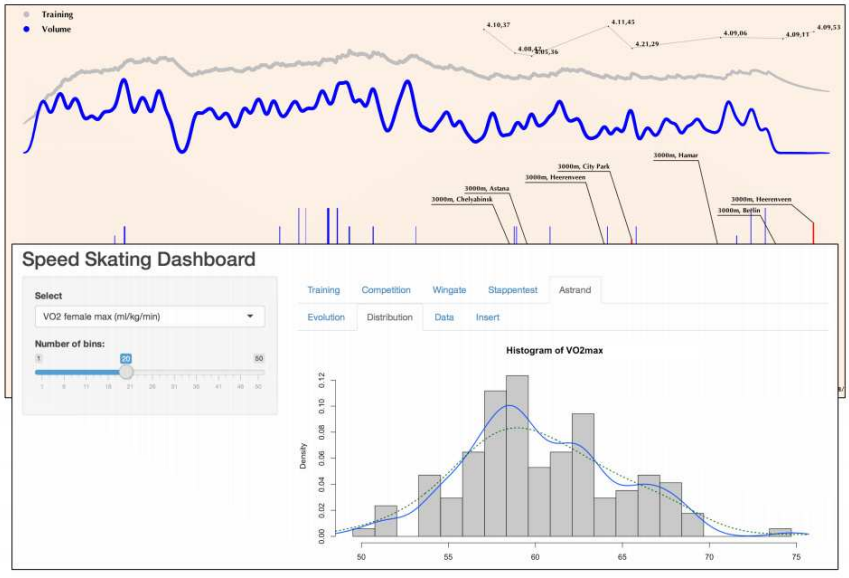
- Blessurepreventie
- Strategie



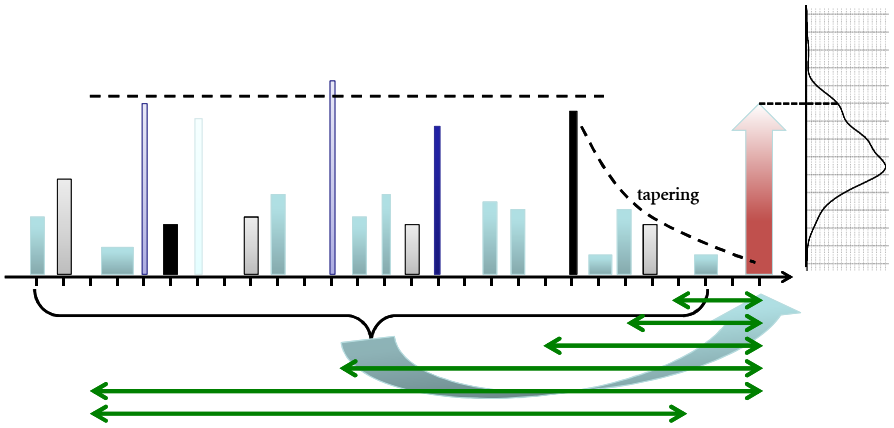
Training Structure




Arno Knobbe
Ricardo Cachucho
Benjamin van den Burgh


Training
Volume
4.10,37
4.11,45
4.08,45
4.05,36
4.21,29
4.09,06
4.09,11
4.09,53
3000m, Hamar
3000m, City Park
3000m, Heerenveen
3000m, Astana
3000m, Chelyabinsk
3000m, Heerenveen
3000m, Berlin
Speed Skating Dashboard
Select
VO2 female max (ml/kg/min)
Number of bins:
20
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

The Effect of Training



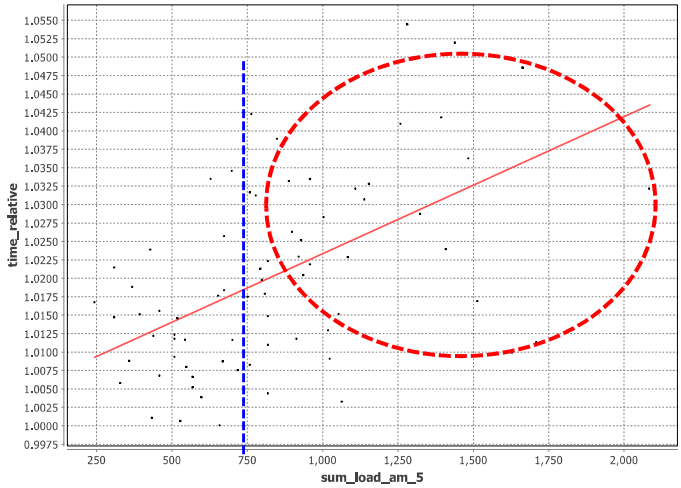
Training parameters

- Hundreds of training parameters that capture various aspects of periodization

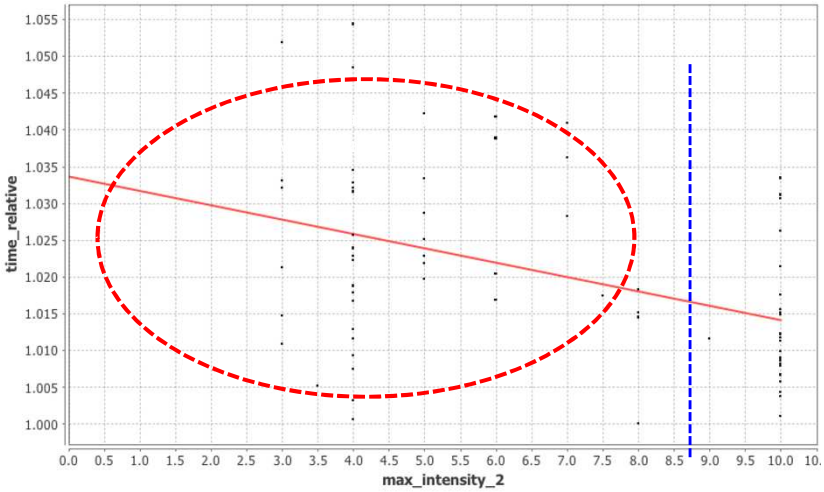
- Sum of duration over 14-day period
- Max of intensity over 2-day period
- Sum of duration over 21-day period, morning sessions
- Sum of duration over 21-day period, intensities 6, 7, 8, 9
- Maximum of load over 7-day period, cycling
- Count of cycling sessions over 4-day period

etcetera

Sum of loads over last 5 days, morning sessions



Maximum intensity 8 or higher in the two days prior to the race



Some Findings

- To increase *aerobic capacity*, make sure you
  - include at least one exercise longer than 3.5 hours
  - …over the period of 14 to 3 days before the test moment
  - avoid loads above 240 in the mornings, 2 days window

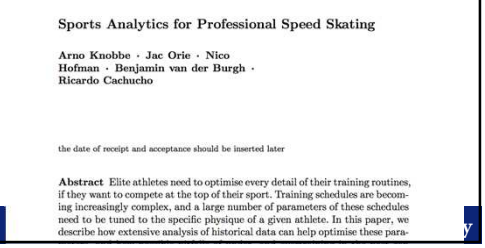
VO_2max will increase by 4%

- total time in intensity zone [1, 4] above 850 min/w, 21 days window
- average intensity above 3.8, 14 days window

VO_2max will increase by 11%

Results

- Longitudinal, detailed data, great potential
- Actionable results, taken up by coach
- No major revolution in training:
  - subtle tweaking of parameters
- Joint scientific publication
- Team funding



Sports Analytics for Professional Speed Skating

Arno Knobbe · Jac Orie · Nico Hofman · Benjamin van der Burgh · Ricardo Cachucho

the date of receipt and acceptance should be inserted later

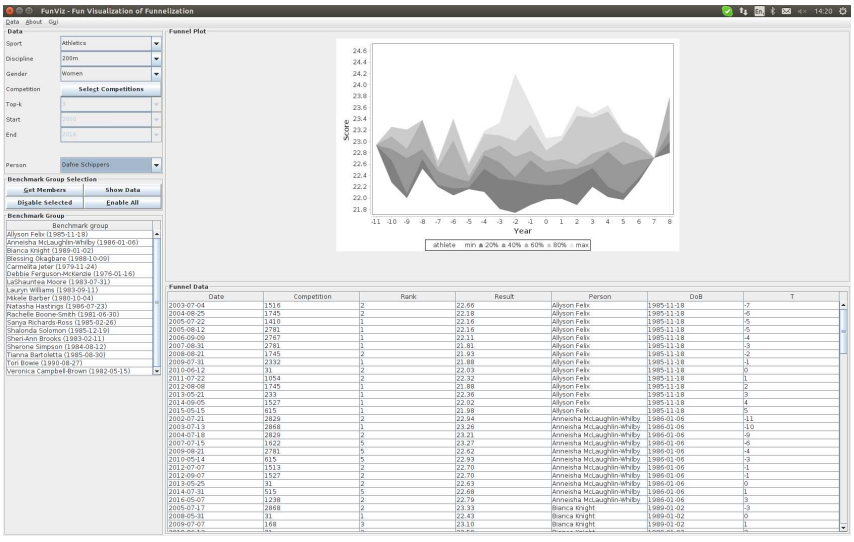
Abstract Elite athletes need to optimise every detail of their training routines, if they want to compete at the top of their sport. Training schedules are becoming increasingly complex, and a large number of parameters of these schedules need to be tuned to the specific physique of a given athlete. In this paper, we describe how extensive analysis of historical data can help optimise these para-

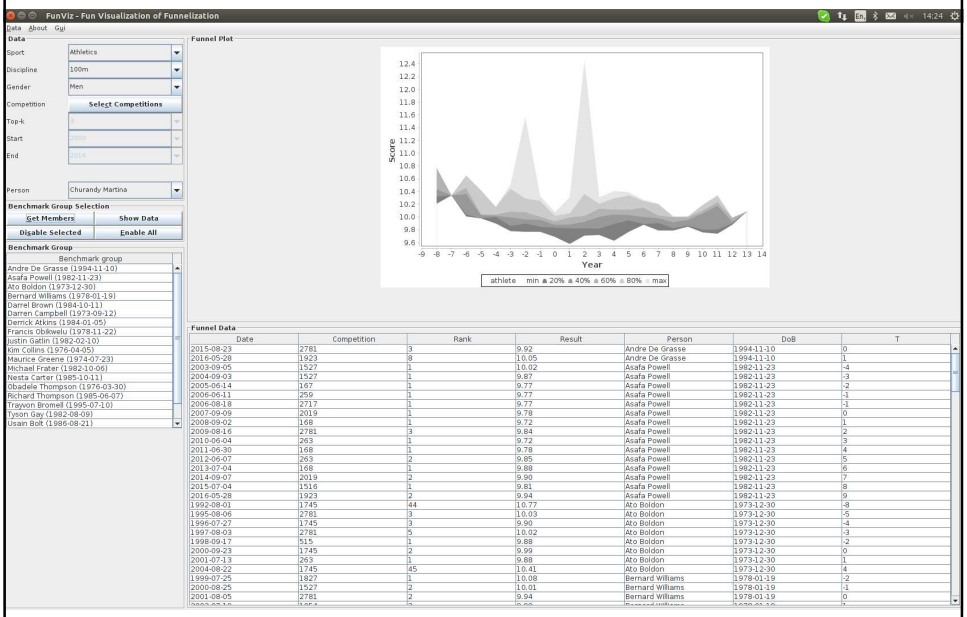

Investment Decisions
papendal
hotel • sport • congres
Marvin
Meeng
Discover the world at Leiden University

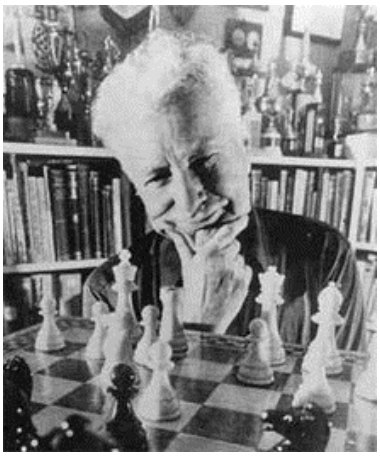

Discover the world at Leiden University

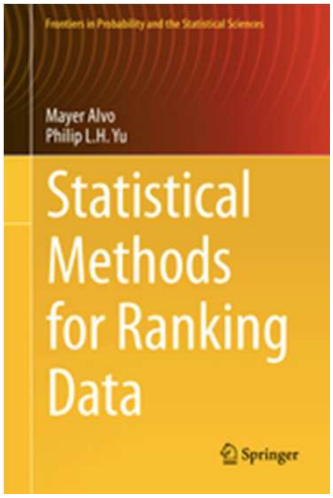
Funnels

Marvin Meeng



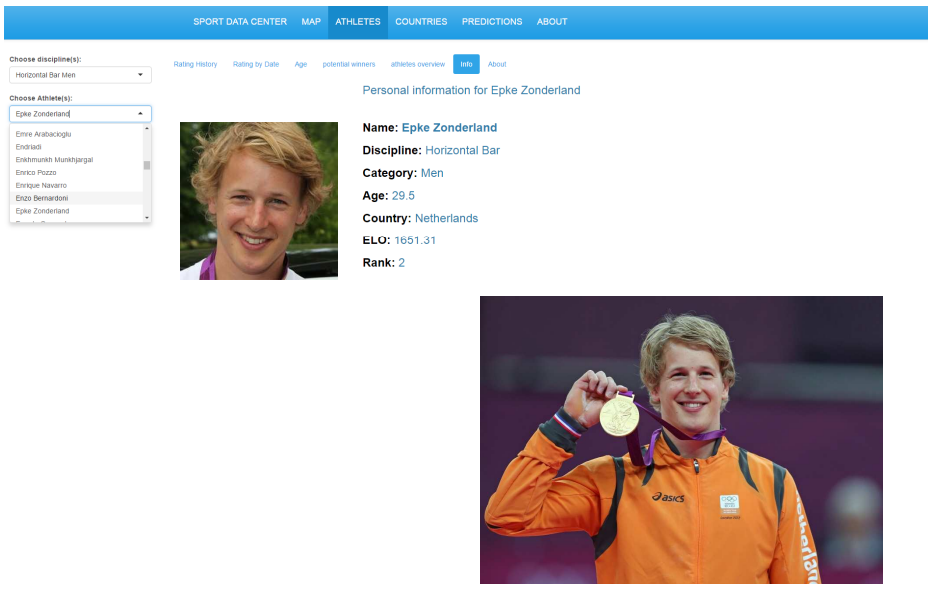




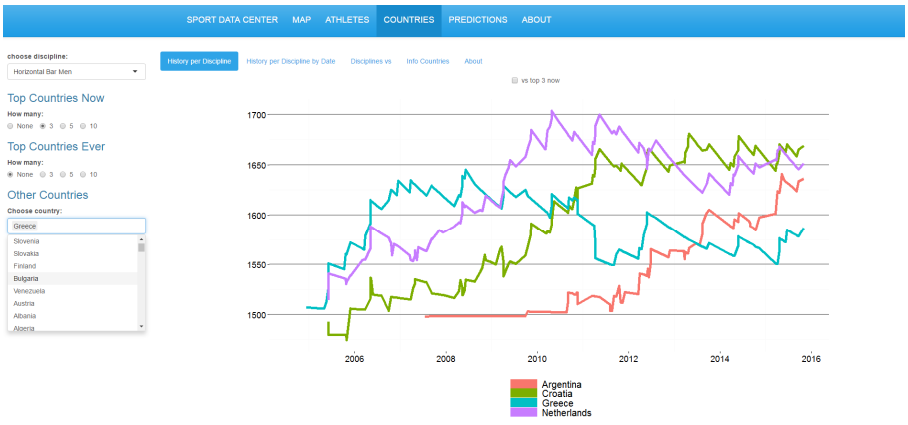


Discover the world at Leiden University


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)
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
1700
1650
1600
1550
1500
ELO
2006
2008
2010
2012
2014
2016
DATE
Epke Zonderland
Marijg MoÃ¾nik
NicolÃ¡s CÃ³rdoba


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
1700
1650
1600
1550
1500
2006
2008
2010
2012
2014
2016
Argentina
Croatia
Greece
Netherlands

Challenges

- Develop Elo-like ratings
  - Probabilities
  - Head-to-head including differences
  - Doubles
  - Teams

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

Research question

- Using an algorithmic approach that analyses possession switches during football matches, which variables have the highest influence on these moments?

Discover the world at Leiden University

Data

Available data

- Positional data
- Event data

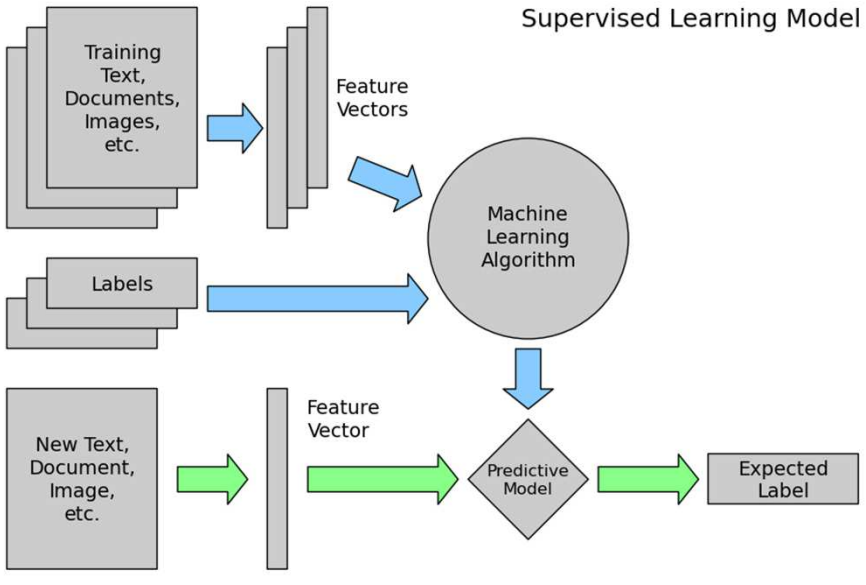
Timestamp (ms)	X	Y	Name
40,000	17.027	5.364	Home Team Player
40,000	-4.051	-3.378	Visiting Team Player 1
40,000	99,999	99,999	Visiting Team Player 2
40,000	11.094	-10.509	Referee
40,000	38.09	-7.985	Ball

Table 3.1: Frame 40,000 (first frame of 40th second).

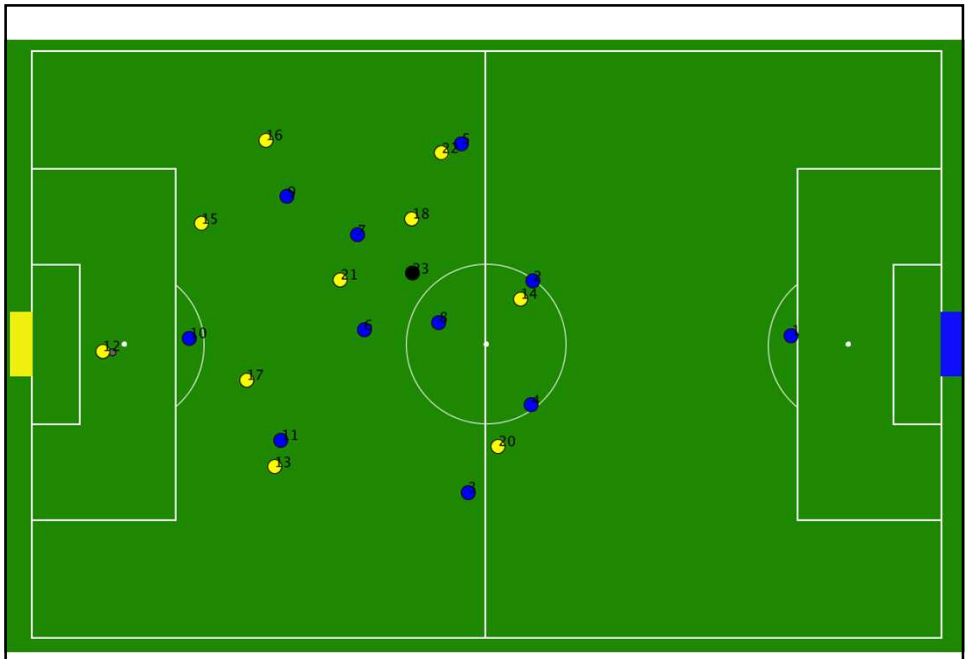
Time(ms)	Half	Category	Player	Team	Attribute	Definition	Match
7,132	1	attacking action	Player 8	Home Team	head — duel touched	isPossessionLoss — isDuelPart — isDuelAir — isAerial	Match 1
10,905	1	pass	Player 15	Visiting Team	right foot	isPassCompleted — isPassWide	Match 1
11,930	1	pass	Player 16	Visiting Team	right foot — direct	isPossessionLoss — isPassForward — isPassLong	Match 1
14,931	1	pass	Player 3	Home Team	direct — left foot	isPossessionGain — isPossessionGainInterception — isPassCompleted — isPassForward	Match 1
17,030	1	attacking action	Player 4	Home Team	duel touched	isDuelPart — isDuelWonByAttacker — isDuelStanding	Match 1

Table 3.2: Event data example.

Discover the world at Leiden University



Discover the world at Leiden University



Sum of distances to ball

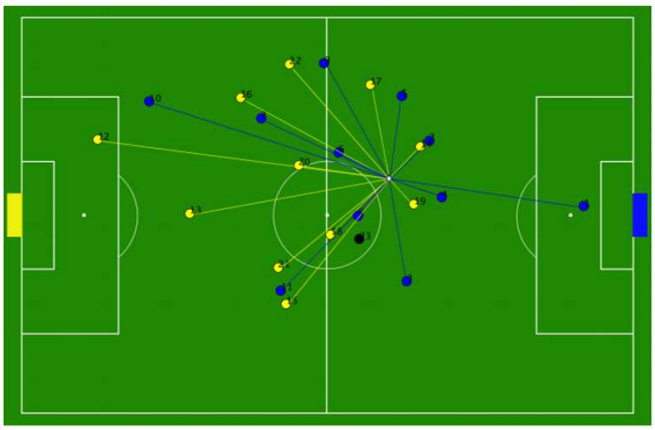


Figure 4.1: Distance to the ball.

Sum of Distances to closest opponent

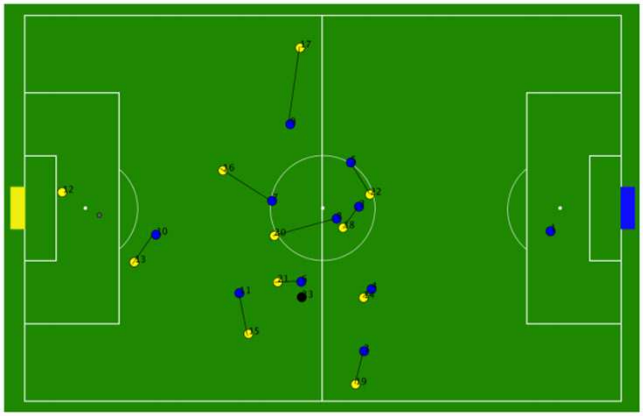
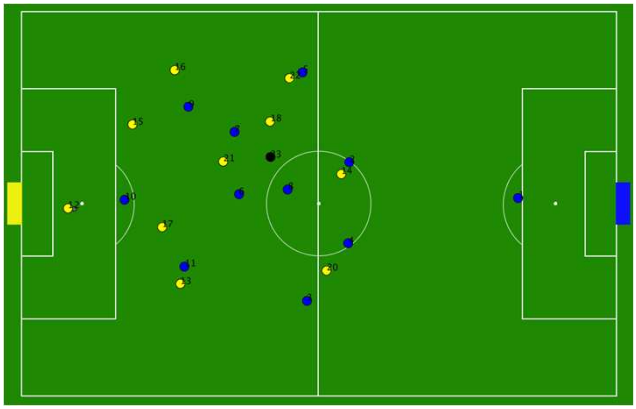


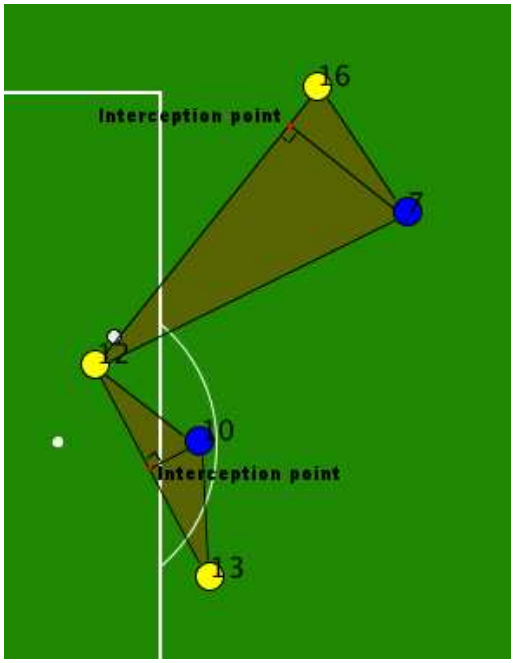
Figure 4.2: Optimal direct opponent.

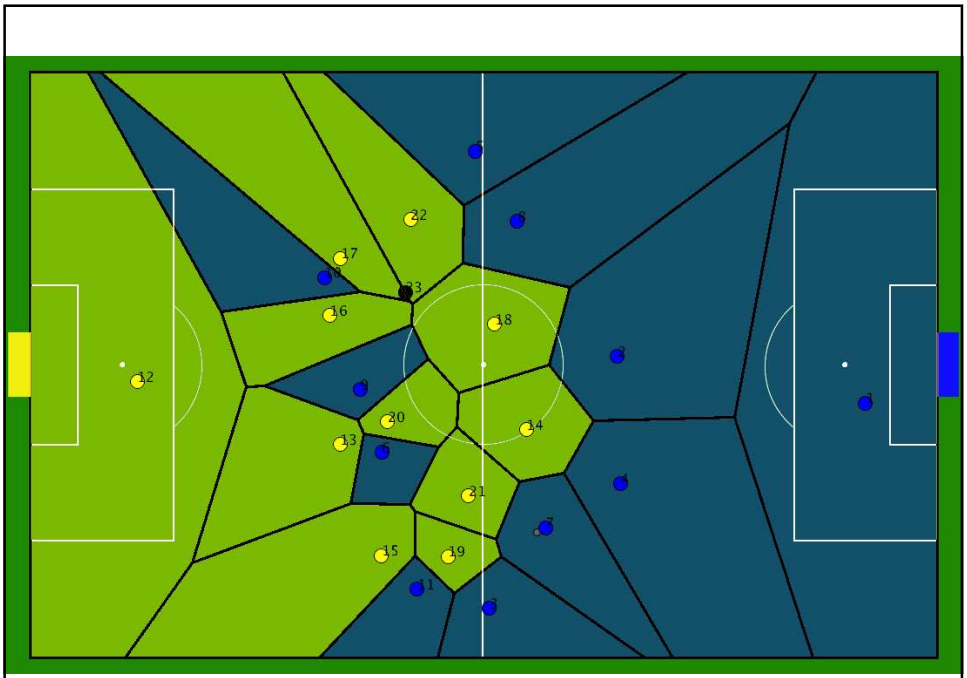
Counts

- # Players of own team within X meters
- # Players of opposite team within X meters



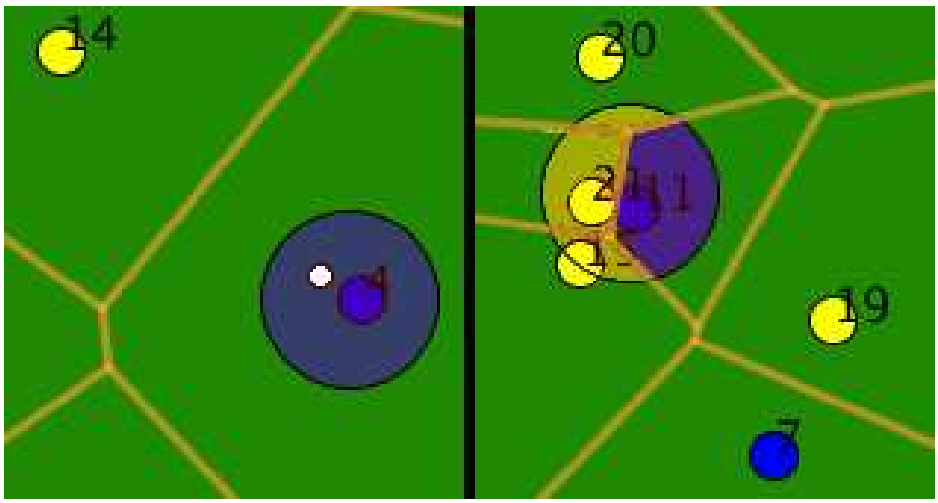
Discover the world at Leiden University


Interception point
16
7
12
10
Interception point
13


22
17
23
16
18
12
30
14
13
21
15
19
11

Voronoi Diagrams

- https://www.youtube.com/watch?v=7eCrHAv6sYY
- https://www.youtube.com/watch?v=Y5X1TvN9TpM
- https://www.youtube.com/watch?v=k2P9yWSMaXE



Features

Football knowledge
- Possession switch
- Pressure

Features
- Distances
- Positions
- Counts
- Ratios
- Surfaces
- Possible passes

Discover the world at Leiden University

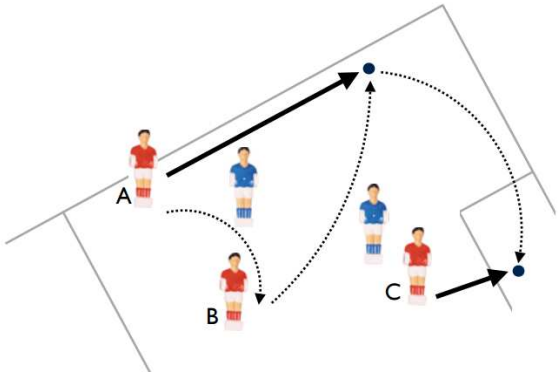
Subgroup Discovery

- Pattern

> *The distance between the ball and the closest 3 players of the defending team should be smaller or equal to 27 meters for a possession switch to occur*

Discover the world at Leiden University

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

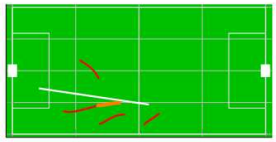


Slides by Ulf Brefeld

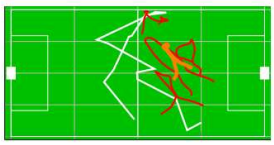
- Individual level
- Group level
- Team level

Different Scales

- 4 defence players
  → game initiations



- 4 offence players
  → scoring opportunities



Data is de nieuwe olie

Kennistransfer van
- extern naar intern
- topsport naar breedtesport
- sport naar sport
- expert naar groot publiek

Advies op maat
- van groep naar individu
- van lid naar klant

Data is the new oil. It's only useful when it's refined!
Jess Greenwood, Contagious

Fraud & Risks

SportDataCenter

Data rond Sport en Bewegen

SportDataCenter

- Datawetenschappers
- Sport Data Valley

- Domeinkennis essentieel
- Data & sport field labs

TU Delft · Amsterdam Institute of Sport Science · Universiteit Leiden · LUMC Leiden University Medical Center · The Hague University of Professional Education · TNO

SportDataCenter

Vijf onderzoekslijnen in sportdata analyse

1. Topsport
2. Breedtesport
3. Aangepast sporten
4. Economische waarde van sport
5. Fraude en risico's

Data Science and Sports

- Apply data science in new application domains yielding new results and insights in those domains.
- At the same time, the application domain serves as a source of inspiration for new data science research.

SPORT DATA Valley

SportDataCenter

Sportdata

- Data Scientists have no data
- Positive, easy to explain
- Quantified Self
- Testbed