

What Really Makes You Move?

Identifying relationships between physical activity and health through applying machine learning techniques on high frequency accelerometer data and survey data

Seyit Höcük – Data Scientist@CentERdata

Joris Mulder – Pradeep Kumar – Natalia Kieruj

May 21, 2019

NPSO conference, Rotterdam





CentERdata core activities

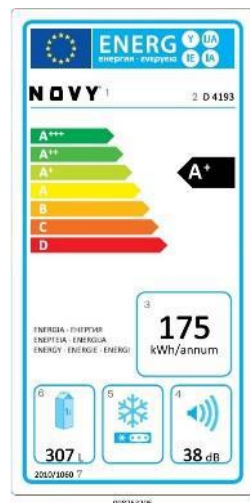
Survey research

Online surveys
Experiments
Data Dissemination
Hostings



Consumer research

Consumer and
behavioral research



Data Science

Machine learning
Deep learning
Text analytics
Data maturity
Data visualisation



ICT

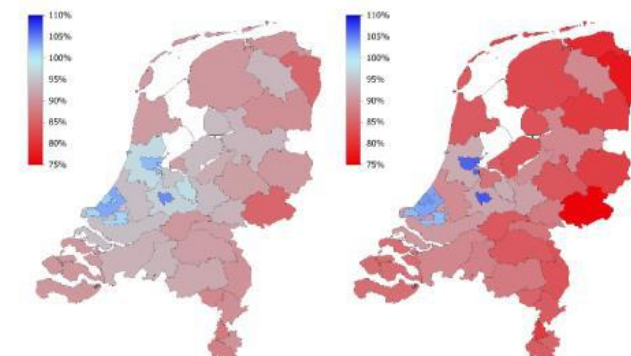
Software and
App development
Dashboards



Quantitative analysis

Policy and behavioral research

Figuur 12: Groei/krimp van het aantal leerlingen in het vo naar rpa in 2022 (links) en 2027 (rechts) ten opzichte van 2017.





Accelerometer study

- Data collected in 2013 in the LISS panel
- LISS: Online panel of 4,500 households, comprising 7,000 individuals (age 16+)
- Relationship health & physical activity

GENEActiv UK

- 1011 panel members participated
- 8 days, day and night





What

- 1 - Detect specific activity patterns from accelerometer data
- 2 - Find the relationship between health and physical activity



Activity patterns

Sedentary	(sleeping, sitting)
Low	(driving, commuting)
Moderate	(walking)
High	(cycling, tooth brushing)
Vigorous	(jogging, exercising)



What

- 1 - Detect specific activity patterns from accelerometer data
- 2 - Find the relationship between health and physical activity

First study, where no Machine Learning/AI was involved is published in Journal of Epidemiology and Community Health, 2018:



OPEN ACCESS

What they say and what they do: comparing physical activity across the USA, England and the Netherlands

Arie Kapteyn,¹ James Banks,² Mark Hamer,³ James P Smith,⁴ Andrew Steptoe,⁵
Arthur van Soest,⁶ Annemarie Koster,⁷ Saw Htay Wah¹



How

Recognizing activity patterns with machine learning

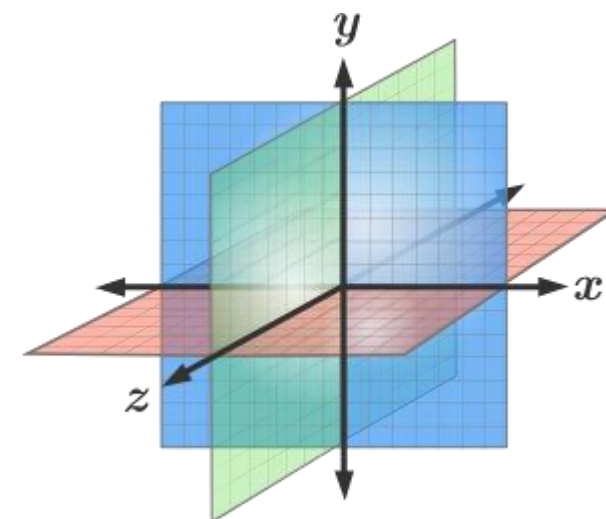
Accelerometer device wearable as watch water-proof
→ can be worn at all times

Measures:

- Acceleration X-Y-Z axis (± 8 g)
- Skin temperature (0-70 °C)
- Light intensity (0-5000 lux)
- 60 Hz, 8 days, 24/7

about 3.3 GB for each participant

852 usable respondents = 3 TB raw data (csv)





Why

The advantages of a wearable device

Self-report misses daily activities and sedentary behavior

Unaffected by cognitive and social desirable bias

Impartial toward cultural and socio-economic differences

Non-invasive of privacy and cheaper than other means

Allows harmonizing across respondents, countries, and studies



Training a Machine

Supervised machine learning on labeled data



Method: model building pipeline

Data Cleaning & pre-processing

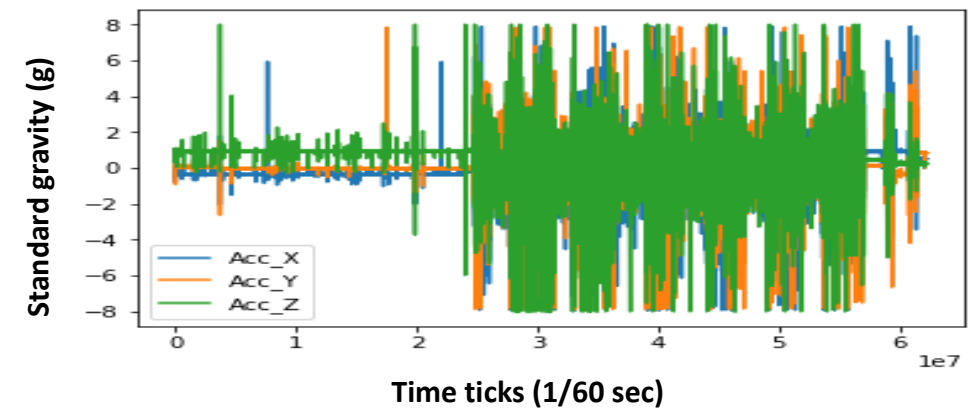
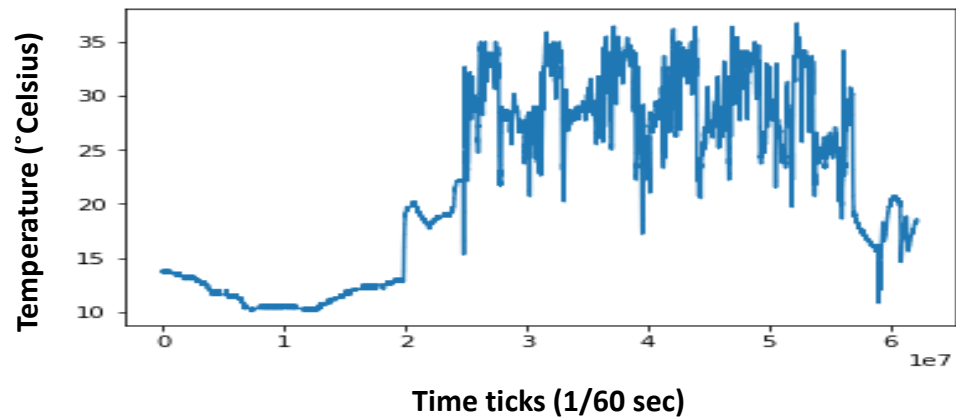
- Removal non-wear time
- Removal of high frequency (frequency higher than 15 Hz)
- Data with wear time less than 7 days discarded

Feature Engineering

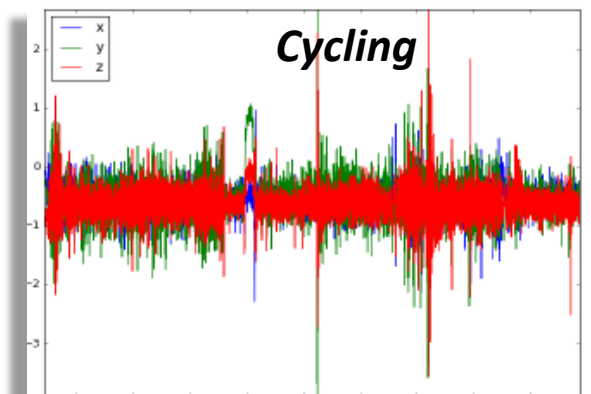
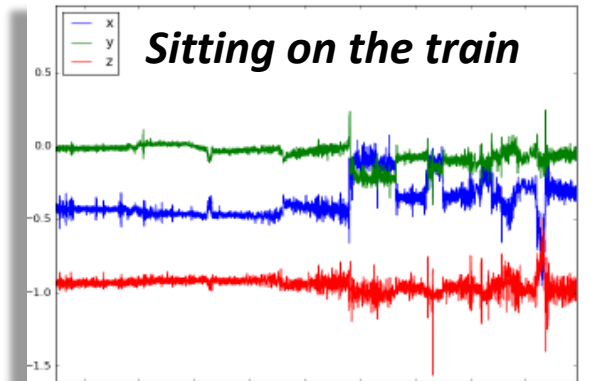
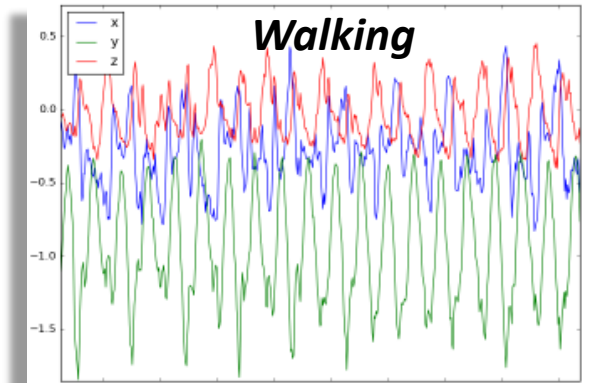
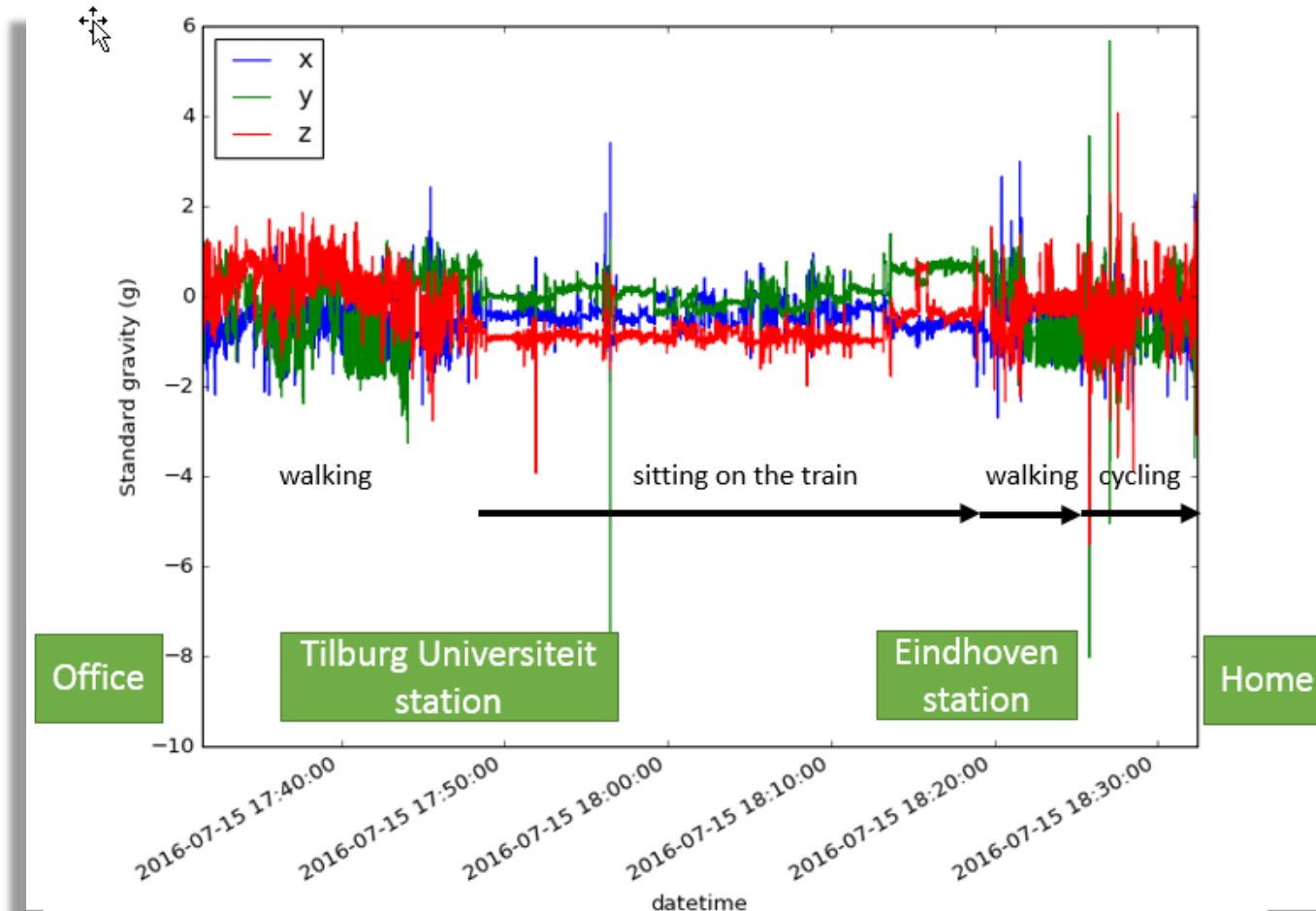
- **Time domain:** X, Y, Z, temperature, mean, median, standard deviation, RMS, percentile distribution
- **Frequency domain:** Fourier transform, dominant frequency selection, power of signal, wavelets

Model building & validation

- Optimizing the epoch time
- Preparing balanced dataset
- Train/test splitting of 80%/20%
- Training and validation of the model (SVM, RF, and LR model)



Method: data labeling



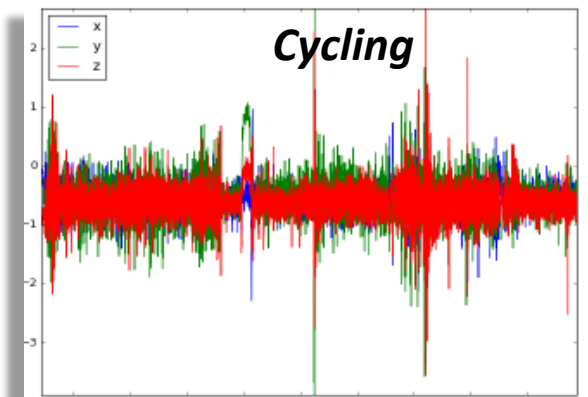
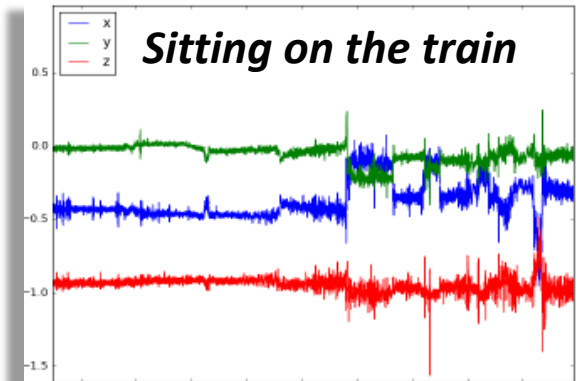
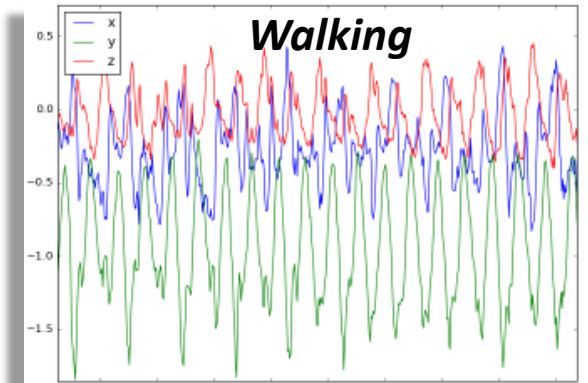
Labelling the data

Method: data labeling

Categorize activities now,
specifically identify in the future

Activity patterns

Sedentary	(sleeping, sitting)
Low	(driving, commuting)
Moderate	(walking)
High	(cycling, tooth brushing)
Vigorous	(jogging, exercising)





Method: applying ML

Model and epoch selection

Trained 3 machine learning models
(with different epoch lengths: from 5 to 10 seconds)

Model	5 seconds	6 seconds	7 seconds	8 seconds	9 seconds	10 seconds
SVM testing	93.7	93.6	93.7	93.5	95	95.1
LR testing	91	91.6	92.6	92.5	93.7	93.5
RF testing	94.5	94.6	95.2	95.1	97.8	98.1

SVM = Support Vector Machine

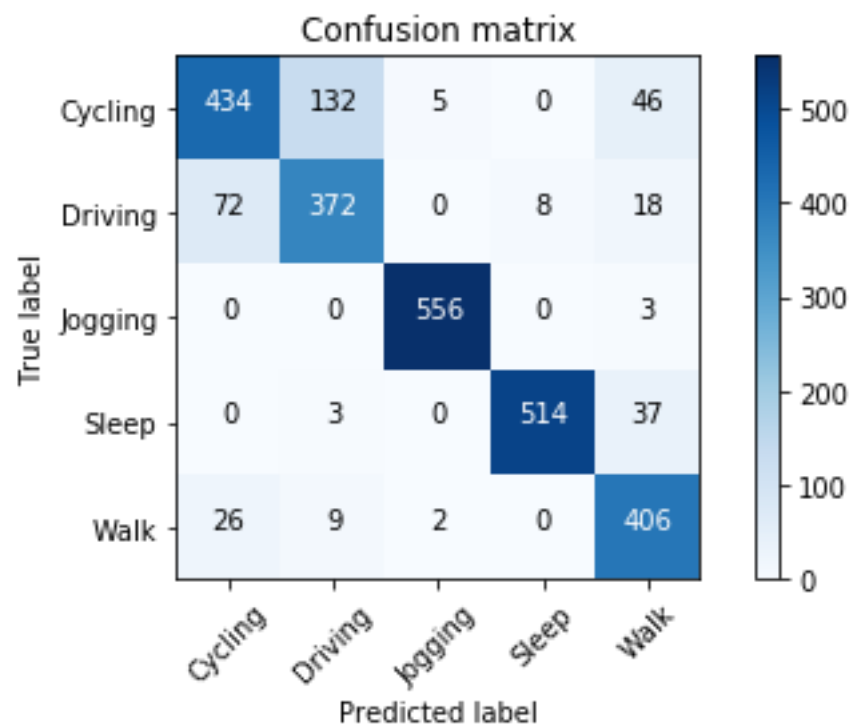
LR = Logistic Regression

RF = Random Forest



Method: applying ML

Model and epoch selection



Activity	Precision	Recall	F ₁ -score
Cycling	0.82	0.70	0.76
Driving	0.72	0.79	0.75
Jogging	0.99	0.99	0.99
Sleep	0.98	0.93	0.96
Walk	0.80	0.92	0.85
Avg/total	0.87	0.86	0.86

Precision = True positives / (True positives + False positives)

Recall = True positives / (True positives + False negatives)



Results
























Applying models to collected data



Predicting Activities

Raw data

About 650 participant files

 800228.csv	09-Feb-19 02:52	Microsoft Excel C...	7,296 KB
 800402.csv	12-Feb-19 04:51	Microsoft Excel C...	6,516 KB
 800674.csv	09-Feb-19 17:44	Microsoft Excel C...	7,378 KB
 800881.csv	11-Feb-19 08:36	Microsoft Excel C...	7,373 KB
 800894.csv	09-Feb-19 11:22	Microsoft Excel C...	7,468 KB
 800930.csv	09-Feb-19 11:30	Microsoft Excel C...	6,970 KB
 801004.csv	11-Feb-19 16:57	Microsoft Excel C...	6,742 KB
 801141.csv	08-Feb-19 13:17	Microsoft Excel C...	7,204 KB
 801144.csv	10-Feb-19 16:18	Microsoft Excel C...	6,397 KB
 801269.csv	10-Feb-19 16:26	Microsoft Excel C...	7,237 KB
 801327.csv	10-Feb-19 16:34	Microsoft Excel C...	7,055 KB
 801394.csv	09-Feb-19 23:53	Microsoft Excel C...	6,968 KB
 801556.csv	09-Feb-19 03:01	Microsoft Excel C...	7,220 KB
 801736.csv	09-Feb-19 11:38	Microsoft Excel C...	7,344 KB
 801776.csv	08-Feb-19 13:25	Microsoft Excel C...	6,459 KB
 801849.csv	10-Feb-19 00:01	Microsoft Excel C...	7,278 KB
 802011.csv	10-Feb-19 16:42	Microsoft Excel C...	6,933 KB
 802340.csv	08-Feb-19 13:33	Microsoft Excel C...	6,938 KB
 802395.csv	09-Feb-19 11:47	Microsoft Excel C...	7,400 KB
 802398.csv	11-Feb-19 08:44	Microsoft Excel C...	7,232 KB
 802819.csv	08-Feb-19 13:41	Microsoft Excel C...	7,315 KB
 802859.csv	10-Feb-19 16:50	Microsoft Excel C...	6,978 KB
 802920.csv	11-Feb-19 08:51	Microsoft Excel C...	6,648 KB

About 120,000 epochs per file →

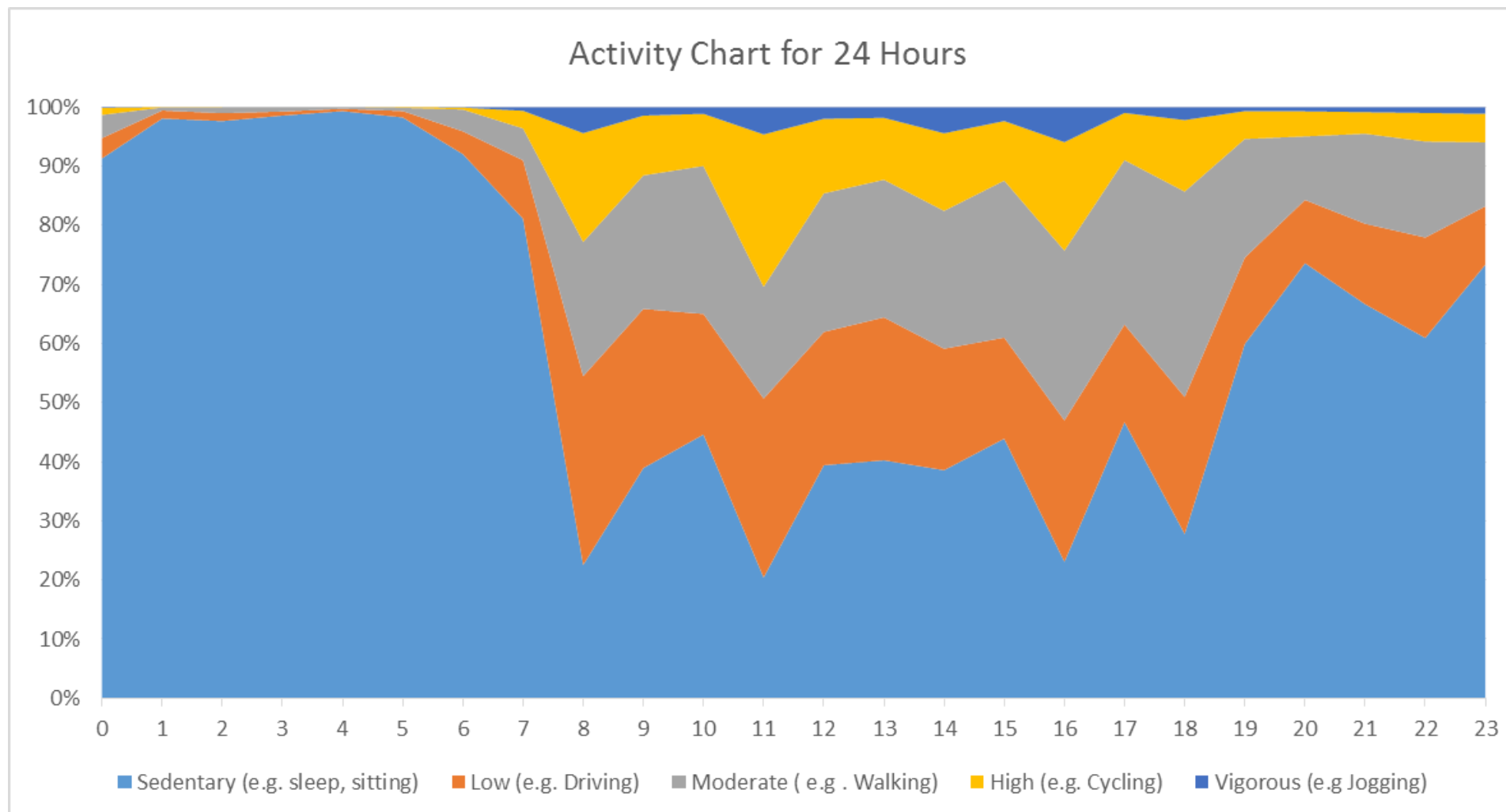
	Starting Time	End Time	Epoch time (s)	Predicted Activity
0	2013-03-30 09:13:57:084	2013-03-30 09:14:03:044	6.0	sleep
1	2013-03-30 09:14:03:044	2013-03-30 09:14:09:004	6.0	sleep
2	2013-03-30 09:14:09:004	2013-03-30 09:14:14:964	6.0	sleep
3	2013-03-30 09:14:14:964	2013-03-30 09:14:21:124	6.0	sleep
4	2013-03-30 09:14:21:124	2013-03-30 09:14:27:084	6.0	sleep
5	2013-03-30 09:14:27:084	2013-03-30 09:14:33:044	6.0	sleep
6	2013-03-30 09:14:33:044	2013-03-30 09:14:39:004	6.0	sleep
7	2013-03-30 09:14:39:004	2013-03-30 09:14:44:964	6.0	sleep
8	2013-03-30 09:14:44:964	2013-03-30 09:14:51:124	6.0	sleep
9	2013-03-30 09:14:51:124	2013-03-30 09:14:57:084	6.0	downstairs
10	2013-03-30 09:14:57:084	2013-03-30 09:15:03:044	6.0	downstairs
11	2013-03-30 09:15:03:044	2013-03-30 09:15:09:004	6.0	sleep
12	2013-03-30 09:15:09:004	2013-03-30 09:15:14:964	6.0	sleep
13	2013-03-30 09:15:14:964	2013-03-30 09:15:21:124	6.0	sleep
14	2013-03-30 09:15:21:124	2013-03-30 09:15:27:084	6.0	sleep
15	2013-03-30 09:15:27:084	2013-03-30 09:15:33:044	6.0	sleep
16	2013-03-30 09:15:33:044	2013-03-30 09:15:39:004	6.0	downstairs
17	2013-03-30 09:15:39:004	2013-03-30 09:15:44:964	6.0	brush
18	2013-03-30 09:15:44:964	2013-03-30 09:15:51:124	6.0	brush
19	2013-03-30 09:15:51:124	2013-03-30 09:15:57:084	6.0	downstairs
20	2013-03-30 09:15:57:084	2013-03-30 09:16:03:044	6.0	downstairs
21	2013-03-30 09:16:03:044	2013-03-30 09:16:09:004	6.0	downstairs
22	2013-03-30 09:16:09:004	2013-03-30 09:16:14:964	6.0	sleep
23	2013-03-30 09:16:14:964	2013-03-30 09:16:21:124	6.0	sleep
24	2013-03-30 09:16:21:124	2013-03-30 09:16:27:084	6.0	downstairs
25	2013-03-30 09:16:27:084	2013-03-30 09:16:33:044	6.0	downstairs
26	2013-03-30 09:16:33:044	2013-03-30 09:16:39:004	6.0	downstairs
27	2013-03-30 09:16:39:004	2013-03-30 09:16:44:964	6.0	downstairs
28	2013-03-30 09:16:44:964	2013-03-30 09:16:51:124	6.0	downstairs
29	2013-03-30 09:16:51:124	2013-03-30 09:16:57:084	6.0	downstairs
30	2013-03-30 09:16:57:084	2013-03-30 09:17:03:044	6.0	downstairs
31	2013-03-30 09:17:03:044	2013-03-30 09:17:09:004	6.0	downstairs
32	2013-03-30 09:17:09:004	2013-03-30 09:17:14:964	6.0	downstairs
33	2013-03-30 09:17:14:964	2013-03-30 09:17:21:124	6.0	downstairs
34	2013-03-30 09:17:21:124	2013-03-30 09:17:27:084	6.0	downstairs
35	2013-03-30 09:17:27:084	2013-03-30 09:17:33:044	6.0	downstairs
36	2013-03-30 09:17:33:044	2013-03-30 09:17:39:004	6.0	downstairs



Activity patterns

Probability diagram for activity patterns

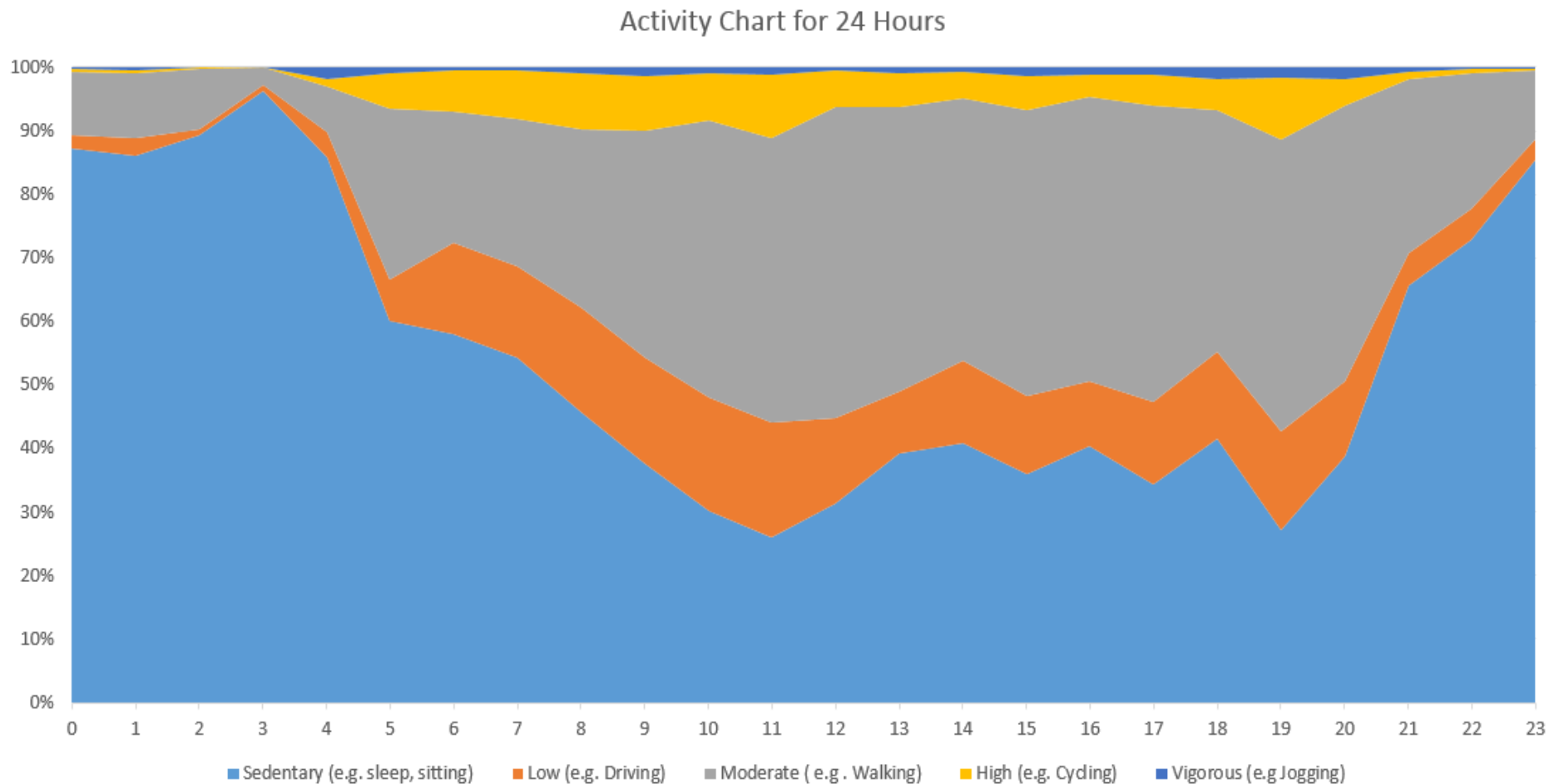
Example: Person A



Activity patterns

Probability diagram for activity patterns

Example: Person B





Statistical results

Perceived health versus activity

Perceived Health	count	Activity type [mean (% time)]				
		Vigorous	High	Moderate	Low	Sedentary
Poor	11	0.27%	2.38%	13.13%	7.35%	76.87%
Moderate	107	0.30%	2.50%	15.65%	7.57%	71.40%
Good	380	0.46%	2.80%	15.70%	8.48%	69.98%
Very good	140	0.53%	3.06%	16.17%	8.67%	68.70%
Excellent	30	0.46%	3.37%	17.03%	7.80%	68.76%

Red = lowest

Blue = highest



Who is who?



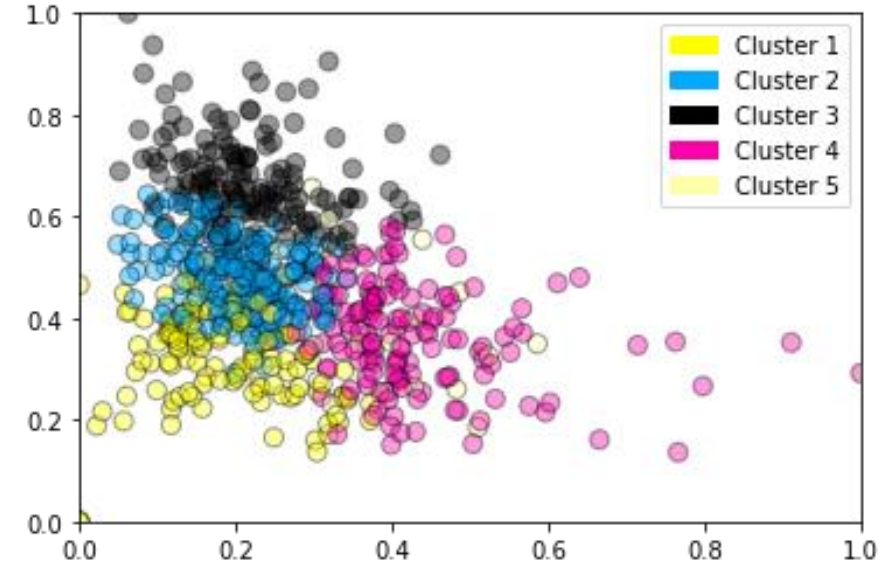
Can we identify what kind of people?



Clustering (K-modes)

Variables used for clustering

- Background variables
 - e.g., age, gender, handedness
- Socio-economic variables
 - e.g., income, domestic situation
- Self-report data (LISS core questionnaire Health)
 - e.g., drinking, eating, smoking, perceived health, functional disabilities
- Personality (LISS core questionnaire Personality)
 - Big five personality traits
- Objective activities (Accelerometer)
 - e.g., sleep, activity (sedentary, low, medium, high, vigorous)





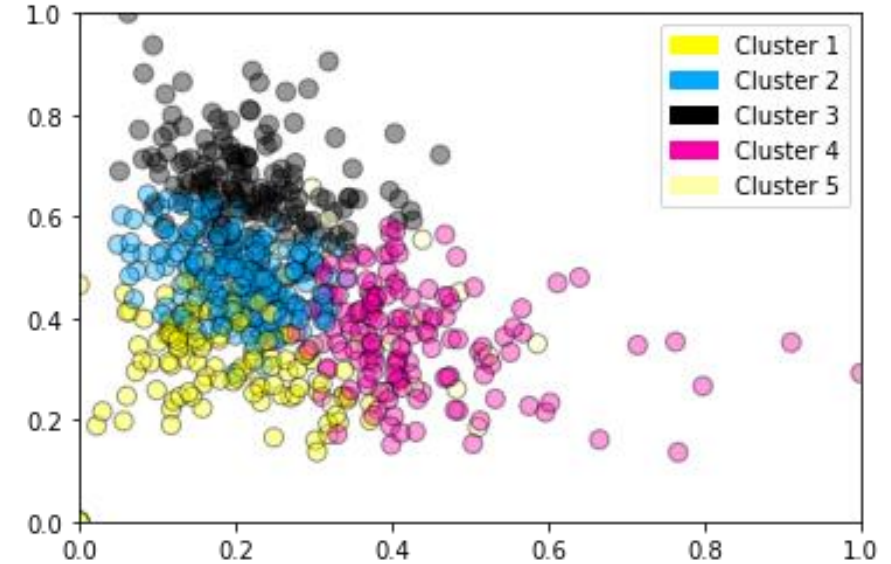
Clustering (K-modes)

Variables selection/importance

Variables selected by deviation from mean

Top variables for the clusters

- Perceived level of depression
- Family situation (children)
- Level of physical activity
- Smoking (yes/no)
- Alcohol consumption
- Education level
- Economic status
- Big five personality traits





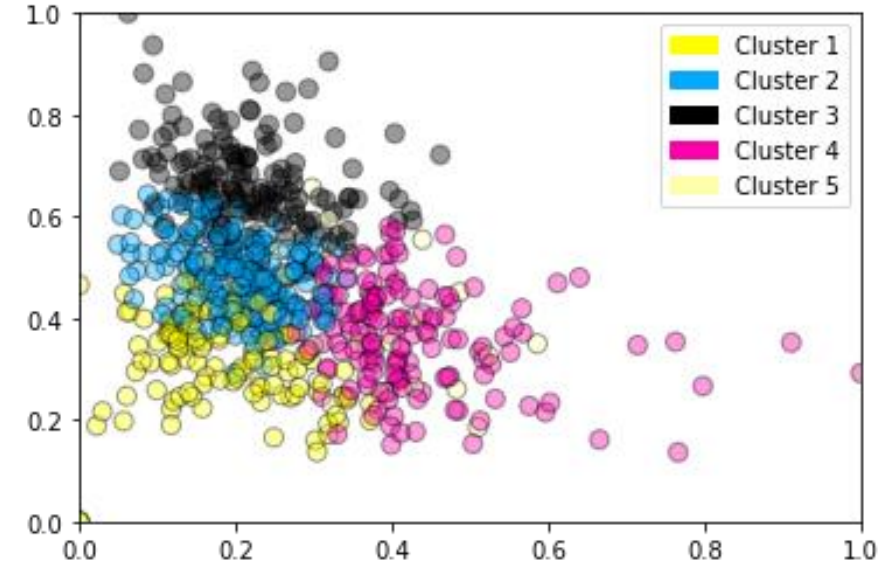
Clustering (K-modes)

Variables selection/importance

Variables selected by deviation from mean

Top variables for the clusters

- Perceived level of depression
- Family situation (children)
- Level of physical activity
- Smoking (yes/no)
- Alcohol consumption
- Education level
- Economic status
- Big five personality traits



Considered but less important, e.g.,

- Eating habit
- BMI
- Age
- Happiness



Clustering Results

Red = stands out negatively
Green = stands out positively

Cluster/ Persona	Income	Depression	Number of kids at home	Activity	Education	Smoking	Alcohol consumption	Characteristic personality (Big 5)	Perceived health
1 Bob	Average	Average	Much Higher	Higher	Average	Higher	Average	Conscientiousness	Average
2 Stella	Lower	Much Higher	Higher	Lower	Lower	Much Higher	Much Lower	Neuroticism	Below average
3 Adam	Higher	Lower	Higher	Average	Much Higher	Average	Average	Openness	Above average
4 Olivia	Average	Higher	Much Lower	Much Higher	Much Lower	Lower	Much Higher	Extraversion	Average
5 Kyle	Higher	Average	Much Lower	Much Lower	Average	Much Lower	Higher	Neuroticism	Below average

Clustering of panel members based on activity, health, and personality.
Validation of clustering using perceived health.



Results

Personas

From Clusters to Personas



Bob

Mister Average
A common conscientious guy that takes care of himself and his family



Stella

The Compulsive
A tense person with functional problems who tends to deal with issues the wrong way



Adam

The Techie
An educated person with decent income who considers his health important



Olivia

The Activist
An initiative taker who isn't afraid to express herself, but this sometimes leads to bad decisions



Kyle

The Player
The lone working guy who likes to go out, have fun and values wealth over health



Conclusions and future research

- ❑ We can detect activity patterns under free living conditions using machine learning with high precision
- ❑ We find relation between activity and perceived health
- ❑ **Next:** Collect more training data from diverse population and backgrounds (generalizing predictions)
- ❑ **Next:** Use deep learning for improved precision/accuracy
- ❑ **Next:** Detect more specific activities (vs. low, medium, high)
- ❑ **Next:** Patterns and relations CBS microdata & other LISS data



Computing power + access to register data

Data stored at
Statistics
Netherlands

The diagram illustrates the architecture of the LISS data system, showing the flow of data between various components:

- User:** Represented by a person icon at a computer, connected to the Internet and the Odissei Data facility.
- Internet:** The primary communication channel for the User and the CBS component.
- CBS (Central Business System):** Consists of a **DSC** (Data Storage Component) and a **CBS** (Central Business System) database. The CBS database is connected to the Internet via a **VPN** (Virtual Private Network).
- Odissei Data facility:** A central hub for combining and analyzing data. It contains a **CBS** database and a **User** database. It is connected to the Internet via a **VPN**.
- SURFsara:** A component for computation on a supercomputer. It contains an **ODF** (Open Data Facility) and a **User** database. It is connected to the Odissei Data facility via a **VPN**.
- Data Flow:**
 - User data:** Flows from the User to the Odissei Data facility and then to the SURFsara.
 - LISS data:** Flows from the CBS to the Odissei Data facility and then to the SURFsara.
 - Remote access:** A green box highlights the interaction between the CBS and the Odissei Data facility, indicating a secure connection via a **VPN**.
 - Computation on supercomputer:** The Odissei Data facility sends data to the SURFsara for processing.



Open Data Infrastructure for Social Science and Economic Innovations



Thank you! Any questions?

