Machine Learning for Sensory Data

MARK HOOGENDOORN

VRIJE UNIVERSITEIT AMSTERDAM
DEPARTMENT OF COMPUTER SCIENCE

VU VRIJE UNIVERSITEIT AMSTERDAM
Faculty of Science

OUTLINE

- Health and sensor
- Machine learning: what is it?
- Machine learning for sensory data
  - With example of a sensory dataset
- A case studies we have done:
  - Mood prediction
- Conclusions

HEALTH AND SENSORY DATA

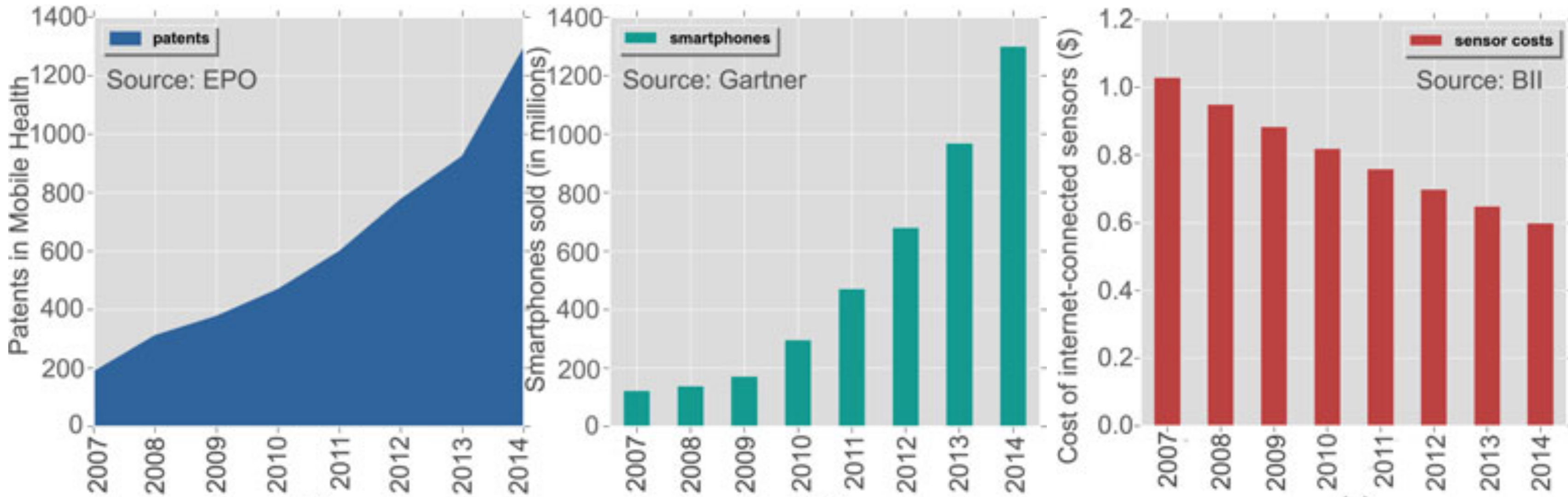
- Huge increase is seen in data collected about health and wellbeing
- Health data collected in various ways:
  - By medical staff (electronic medical records)
  - By smart device sensors (wearables, e.g. smart phone, smart watch)
  - By the people (prompts on your mobile phone, social media, etc.)
- On top, smart devices can easily be used as a mechanism for health interventions
  - Providing supporting messages
  - Providing exercises
  - ......

ELECTRONIC MEDICAL RECORDS

- Some data about Electronic Medical Records[1]:
  - Only in the US (in 2010) EMR's contained 150 exabytes of data
  - Prospected to become yottabytes (10^{24}) in the very near future
  - Large part of the data is *unstructured* (some say 80%), think of free text notes, medical images, etc.

MOBILE HEALTH

- Mobile health[1]:



- Over 100,000 mobile apps in the iTunes store alone (!)[2]

MACHINE LEARNING FOR HEALTH

- Traditional methods from the medical/health domain:
  - Cannot handle the huge amounts of data
  - Are hypothesis driven and cannot find new unexpected results
  - Cannot cope with unstructured data
  - Do not allow for tailoring therapies towards individuals (*personalization*)
- Machine learning can help here!
- My research is devoted to development of machine learning techniques for:
  - Predictive modeling for health (predicting health states)
  - Personalization for health (tailoring interventions)

MACHINE LEARNING

- Machine learning:
  - "Machine learning is to automatically identify patterns from data"
  - "A computer program is said to learn from experience E with respect to some class of tasks T and performance P, if its performance at tasks in T improves with E." (Mitchell)
- Data Mining (DM) is the whole process from data to insights (including machine learning as a step)

MACHINE LEARNING TASKS

- What are machine learning tasks?
  - **Supervised learning** is the machine learning task of inferring a function from a set of labeled training data
  - In **unsupervised learning**, there is no target measure (or label), and the goal is to describe the associations and patterns among the attributes
  - **Reinforcement learning** tries to find optimal actions in a given situation so as to maximize a numerical reward that does not immediately come with the action but later in time.

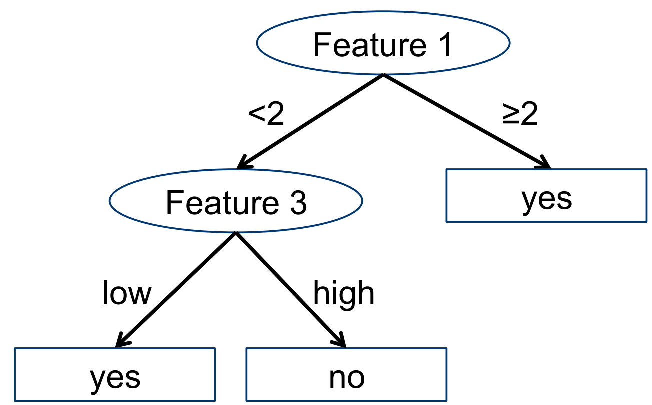
SUPERVISED LEARNING

- Supervised learning approaches learn using a labeled dataset:
  - “Regular” machine learning algorithms:

features

instances

Patient	Feature 1	…………	Feature p	Target
1	2		low	yes
2	6		medium	no
N	4		medium	yes



WHAT STEPS ARE NEEDED TO APPLY MACHINE LEARNING SUCCESSFULLY[11]?

Table 1 Side-by-side comparison of the major existing KDDM models

Model	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*	Generic model
Area	Academic	Industrial	Academic	Industrial	Academic	N/A
No of steps	9	5	8	6	6	6
Refs	(Fayyad *et al.*, 1996d)	(Cabena *et al.*, 1998)	(Anand & Buchner, 1998)	(Shearer, 2000)	(Cios *et al.*, 2000)	N/A
Steps	1 Developing and Understanding of the Application Domain	1 Business Objectives Determination	1 Human Resource Identification	1 Business Understanding	1 Understanding the Problem Domain	1 Application Domain Understanding
			2 Problem Specification			
	2 Creating a Target Data Set	2 Data Preparation	3 Data Prospecting	2 Data Understanding	2 Understanding the Data	2 Data Understanding
			4 Domain Knowledge Elicitation			
	3 Data Cleaning and Preprocessing		5 Methodology Identification	3 Data Preparation	3 Preparation of the Data	3 Data Preparation and Identification of DM Technology
	4 Data Reduction and Projection		6 Data Preprocessing			
	5 Choosing the DM Task					
	6 Choosing the DM Algorithm					
	7 DM	3 DM	7 Pattern Discovery	4 Modeling	4 DM	4 DM
	8 Interpreting Mined Patterns	4 Domain Knowledge Elicitation	8 Knowledge Post-processing	5 Evaluation	5 Evaluation of the Discovered Knowledge	5 Evaluation
	9 Consolidating Discovered Knowledge	5 Assimilation of Knowledge		6 Deployment	6 Using the Discovered Knowledge	6 Knowledge Consolidation and Deployment

WHICH STEP TAKES MOST TIME?

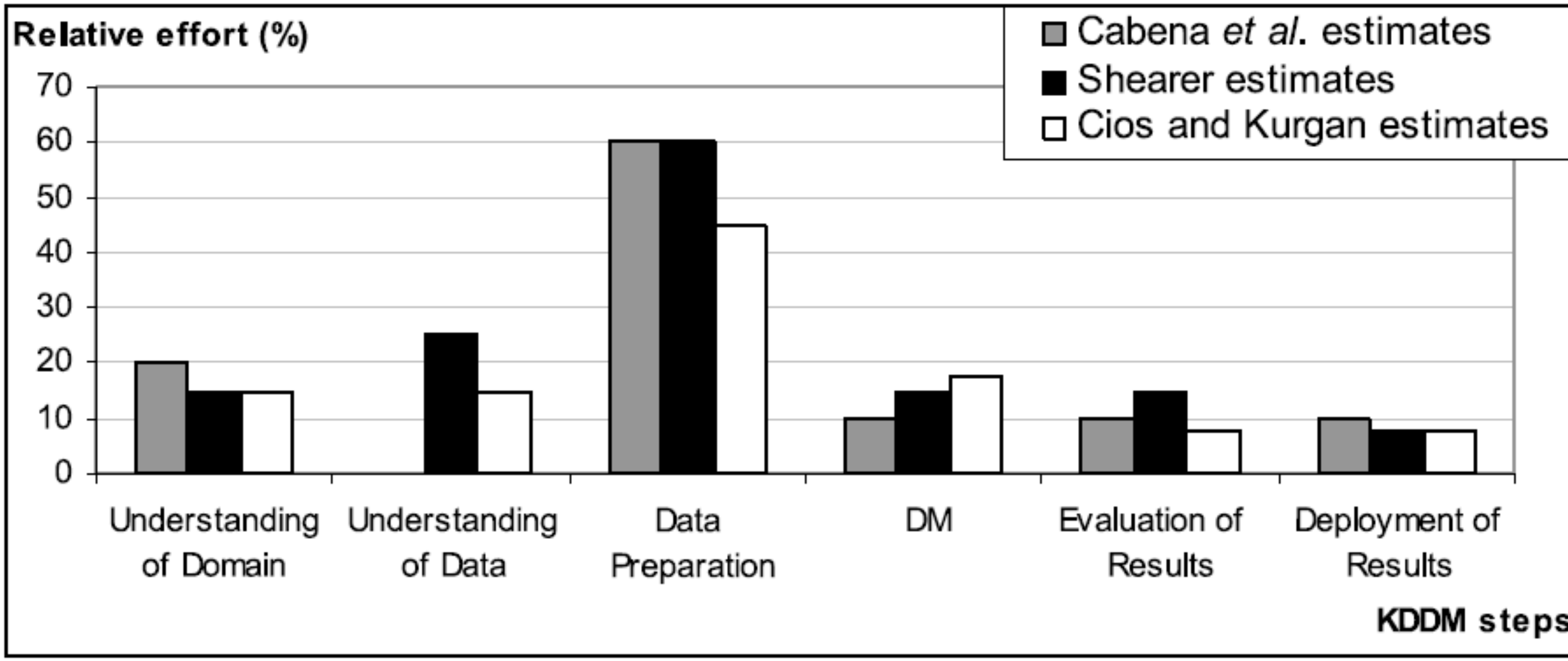
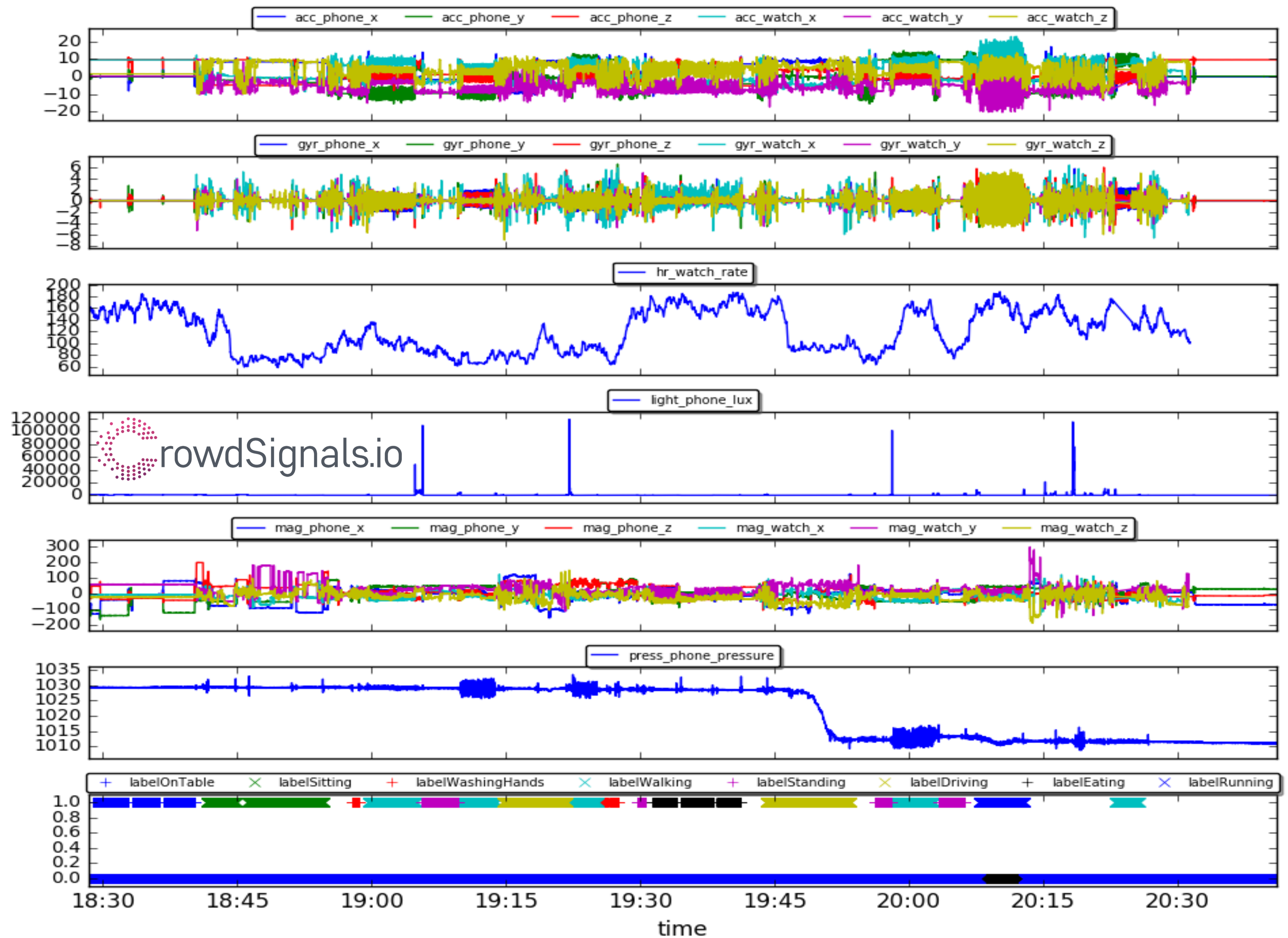


Figure 1 Relative effort spent on specific steps in the KDDM process

EXAMPLE WEARABLES DATASET

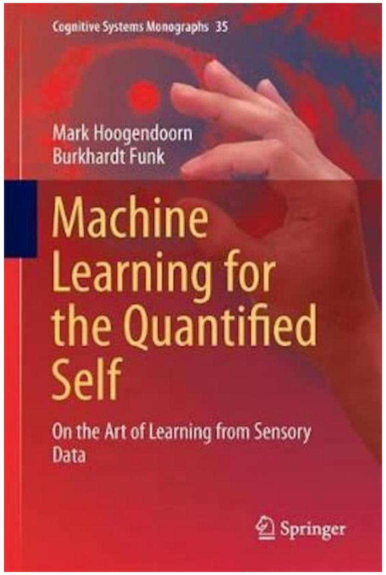
- Let us look at an example dataset from some wearables

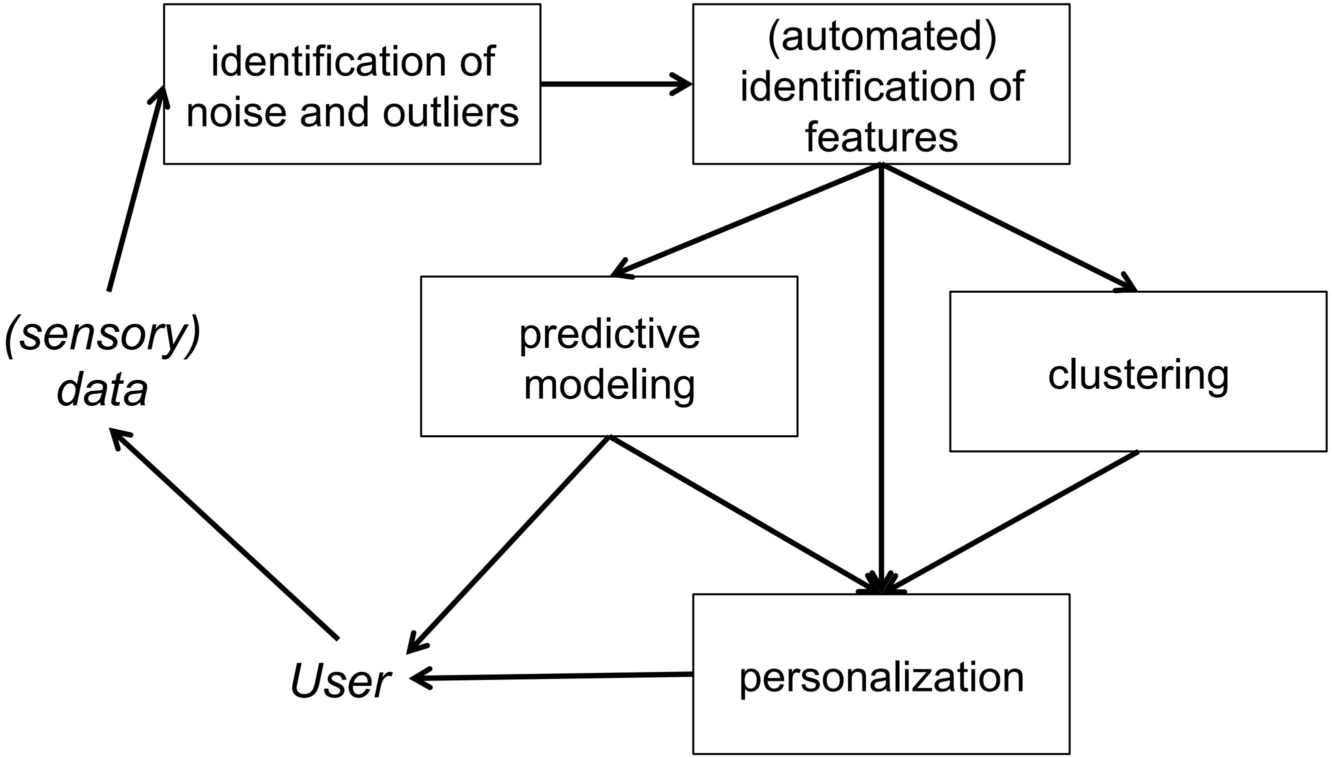


acc_phone_x
acc_phone_y
acc_phone_z
acc_watch_x
acc_watch_y
acc_watch_z
gyr_phone_x
gyr_phone_y
gyr_phone_z
gyr_watch_x
gyr_watch_y
gyr_watch_z
hr_watch_rate
light_phone_lux
CrowdSignals.io
mag_phone_x
mag_phone_y
mag_phone_z
mag_watch_x
mag_watch_y
mag_watch_z
press_phone_pressure
labelOnTable
labelSitting
labelWashingHands
labelWalking
labelStanding
labelDriving
labelEating
labelRunning
18:30
18:45
19:00
19:15
19:30
19:45
20:00
20:15
20:30
time

HOW DO WE LEARN FROM THIS DATA?

- We consider the following loop[3]:





NOISE AND OUTLIER REMOVAL

- What is an outlier?
  - An outlier is an observation point that is distant from other observations
- Causes?
  - Measurement error (a person with a heart rate of 400)
  - Variability (a person trying to push his limits with a heart rate of 190)

NOISE AND OUTLIER REMOVAL

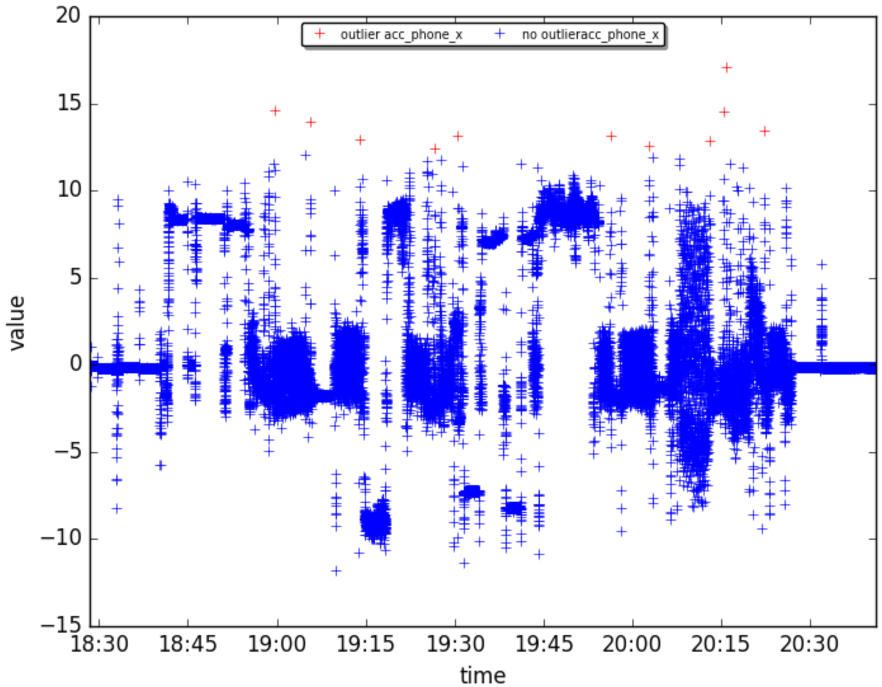
- Difference between measurement and variability outlier?
  - Former generated by another mechanism
- How to remove?
  - Domain knowledge (heart rate cannot be over 220)
  - Without domain knowledge (machine learning focus)
- Have to be **cautious** as you do now want to remove valuable information

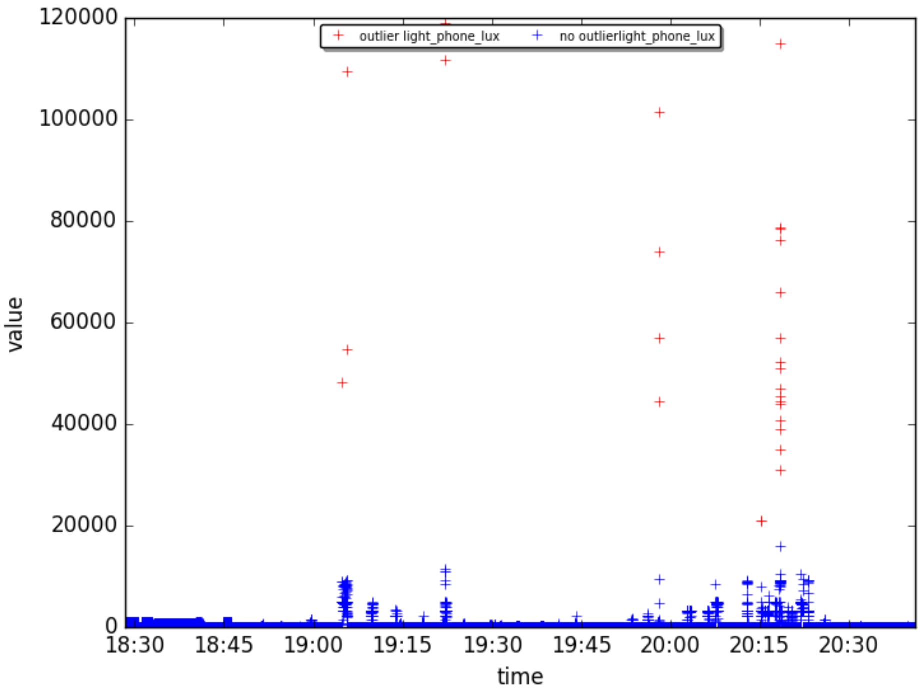
NOISE AND OUTLIER REMOVAL

- Lot of different approaches:
  - Distribution based (we assume a certain distribution of the data)
    - Chauvenet's criterion, mixture models, …
  - Distance based (we only look at the distance between data points)
    - Simple distance based, local outlier factor, …

NOISE AND OUTLIER REMOVAL

- Example outcome:





IDENTIFICATION OF FEATURES

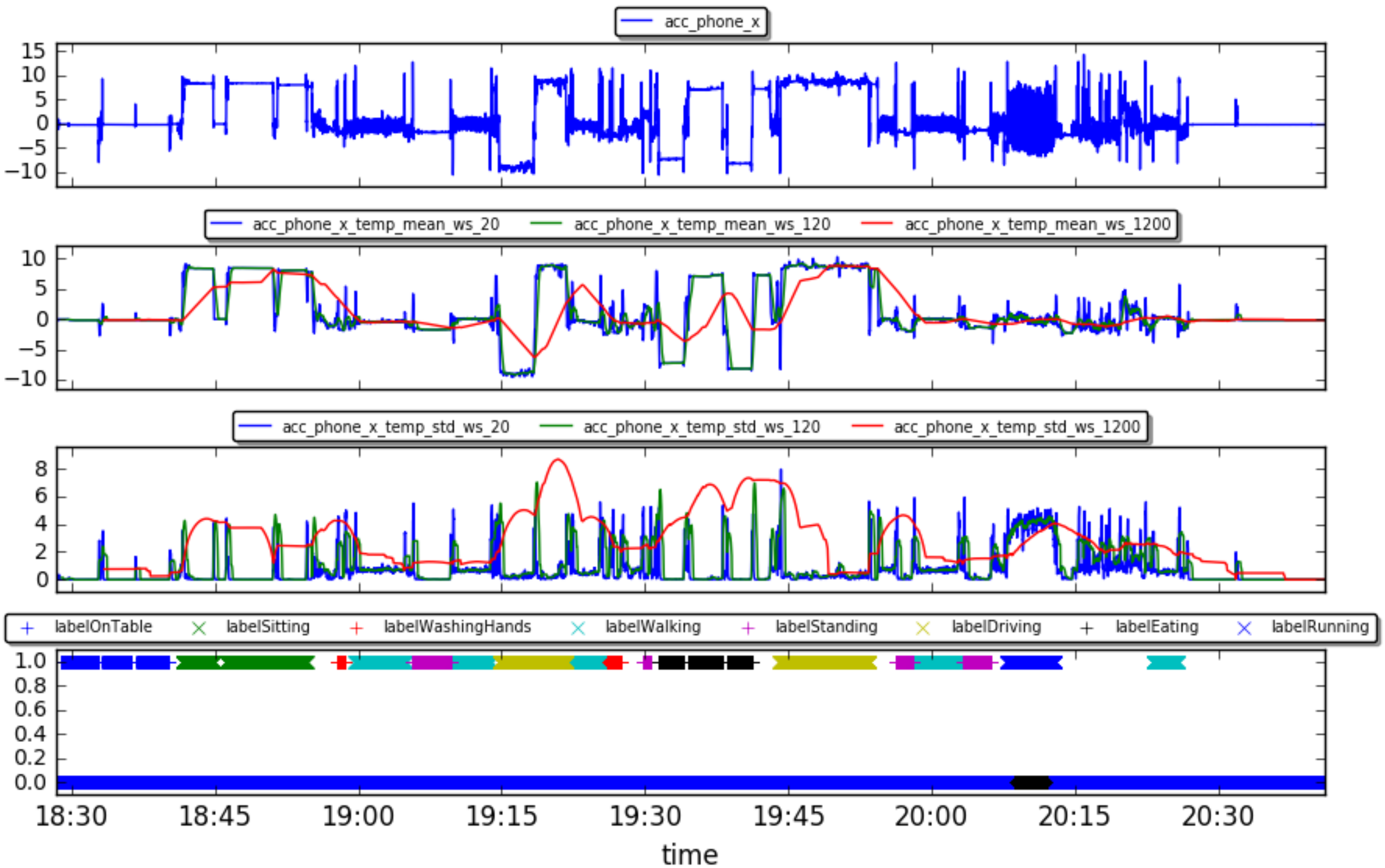
- Example dataset we have now:

Acc. x	Acc. y	Acc. z	Activity
-0.34	+3.45	-3.33	walking
+3.12	+3.14	+3.14	walking
-0.34	+0.34	0.45	walking
+6.54	+0.45	-5.43	running

- Would we able to learn properly?
  - Nope, we need better features/variables!

IDENTIFICATION OF FEATURES

- We need to learn based on patterns over time
- E.g. take the mean or SD over the last x time points points



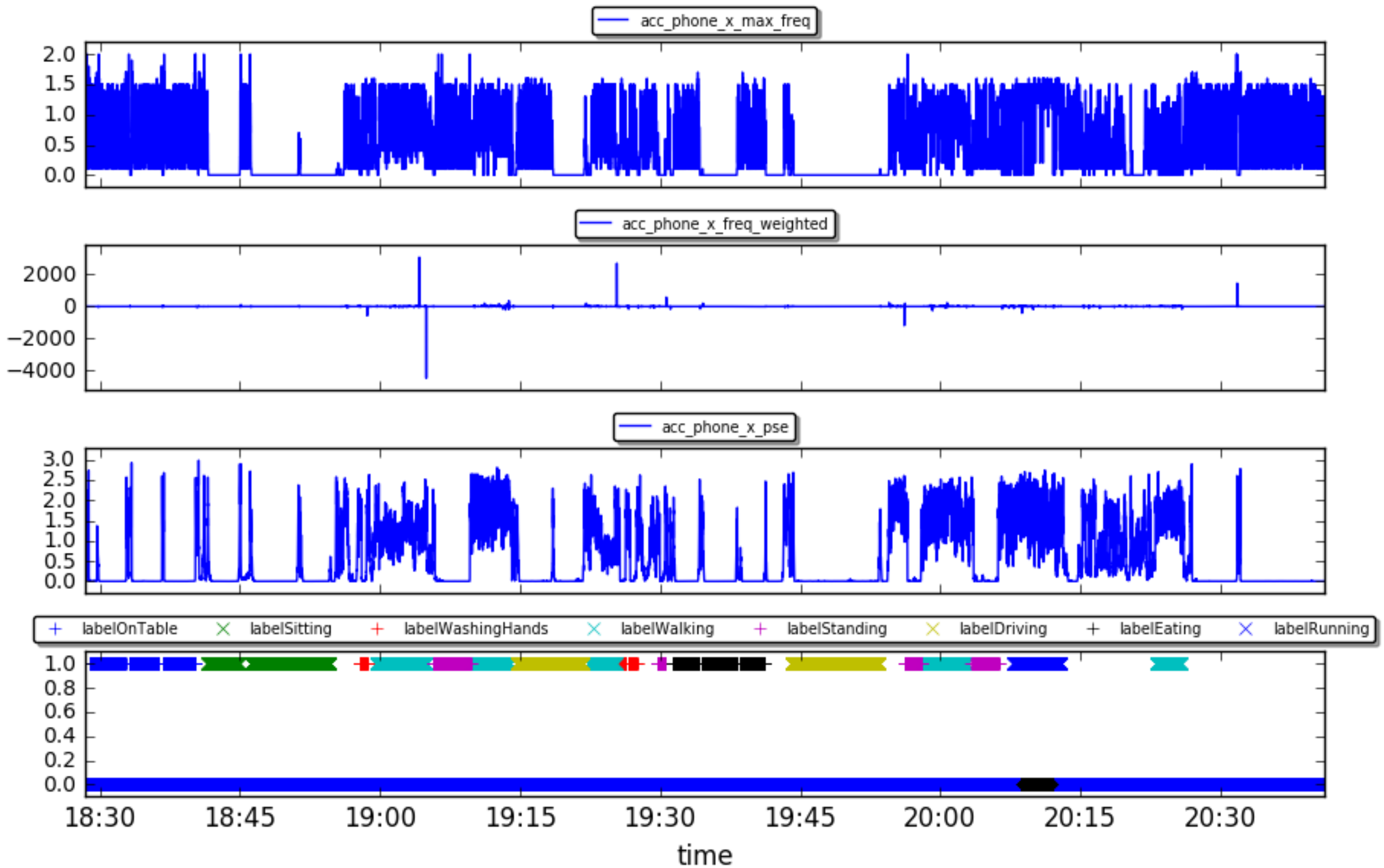
IDENTIFICATION OF FEATURES

- New dataset

Acc. x mean	Acc. y mean	Acc. z mean	Activity
Mean previous 3 + current time point	Mean previous 3 + current time point	Mean previous 3 + current time point	running

IDENTIFICATION OF FEATURES

- We need to learn based on patterns over time
- We can also look at the frequency of the signal (Fourier tr.)



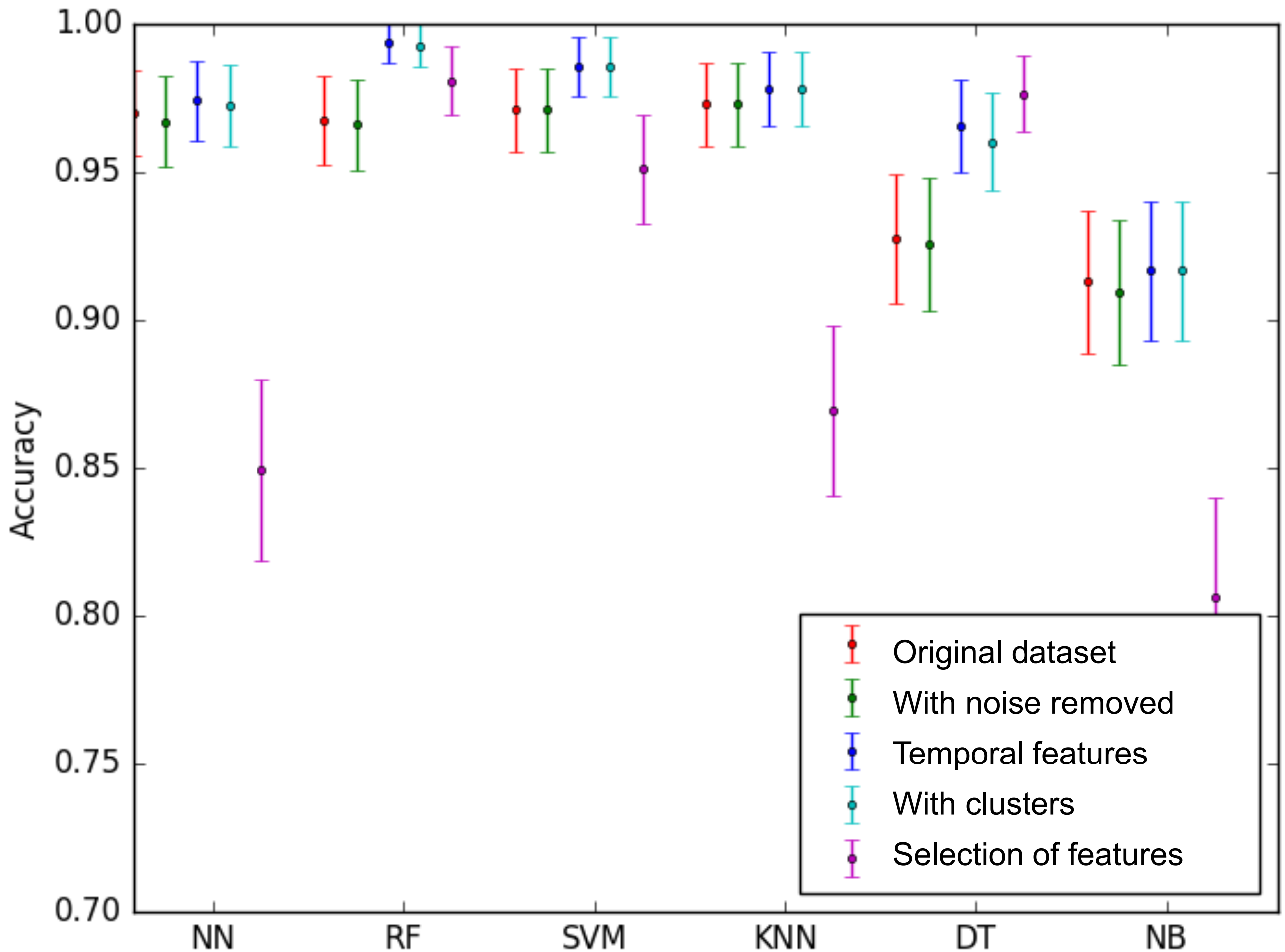
VU

LEARNING FROM THIS DATA

- Let's apply some machine learning algorithms
- We are going to learn how to predict the activity based on all other sensory data
- We are going to set aside part of the dataset as an independent test set
- How accurate would we be able to predict this?
  - < 50%
  - 50 – 75%
  - 75 – 90%
  - > 90%

LEARNING FROM THIS DATA

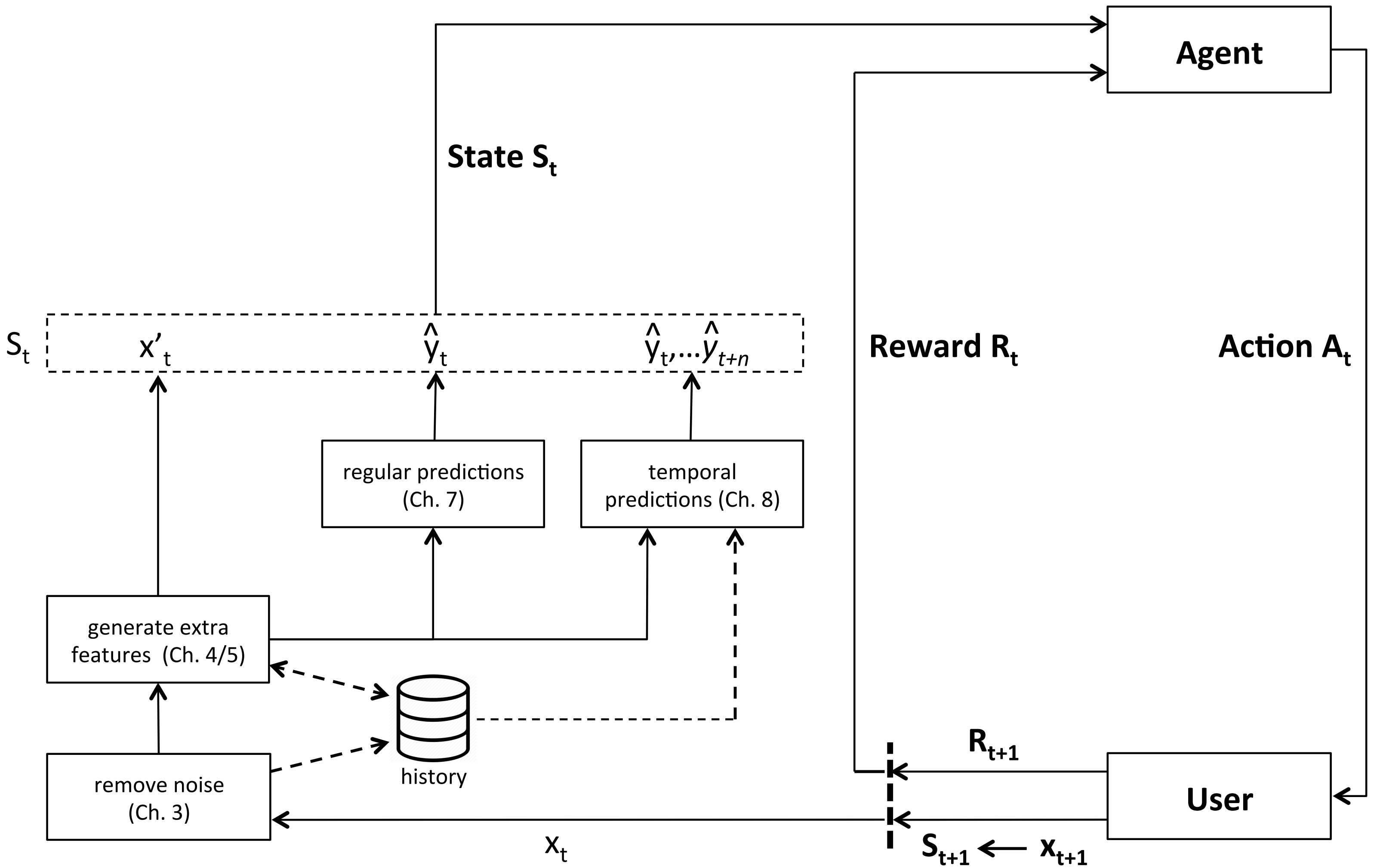
- We try different algorithms (no time for details, sorry)



LEARNING FROM THIS DATA

- Is is always this good?
- No, definitely not:
  - Dataset of a single person
  - Limited time
  - Limited examples of activities
- Still, an accuracy above 90% is typical for recognizing activity, also for more challenging cases
- More and more algorithms do not require the identification of features, but learn these temporal features themselves
  - E.g. LSTM (for Long Short Term Memory Networks)

PERSONALIZATION



CASE STUDY: DEPRESSION

Types of locations
- Social
- Sports

System usage

Larger questionnaires

Ecological Momentary Assessment (EMA)
- Mood
- Sleep quality
- Social interactions
- ……

Activities

ict4depression

E-COMPARED

Mobile intervention with various therapeutic modules
(cognitive behavior therapy, activity scheduling, exercise therapy, etc.)

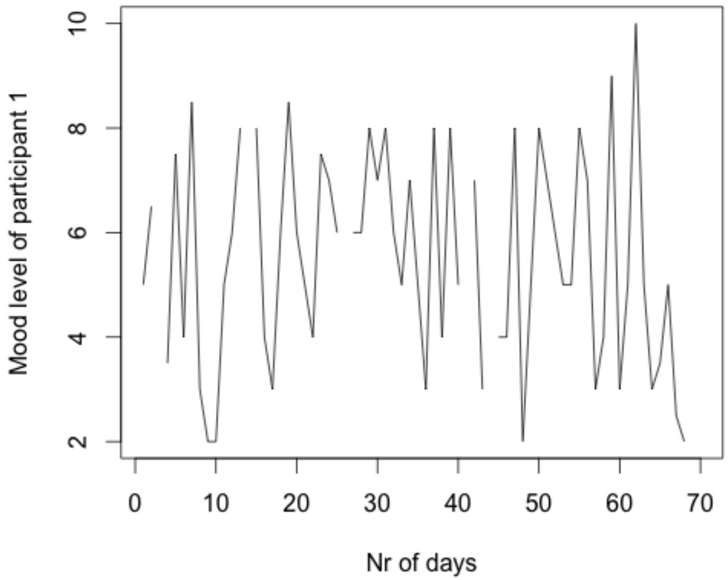
VU

CASE STUDY: DEPRESSION

- We have collected the following data:

Abbreviation	EMA question
Mood	How is your mood right now?
Worry	How much do you worry about things at the moment?
Self-Esteem	How did you sleep tonight?
Sleep	How much have you enjoyed the day?s activities?
Activities done	How good do you feel about yourself right now?
Enjoyed activities	To what extent have you carried out enjoyable activities today?
Social contact	How much have you been involved in social interactions today?

- 49 patients, over 70 days
- Want to predict mood per individual:



CASE STUDY: DEPRESSION

- We want to apply the same approach as has been explained before (summarize the data over days)
- What window size is optimal? We try to find this per measurement
- Learn per patient, within the windows we take the mean, standard deviation, and the trend
- We perform the following experiments:

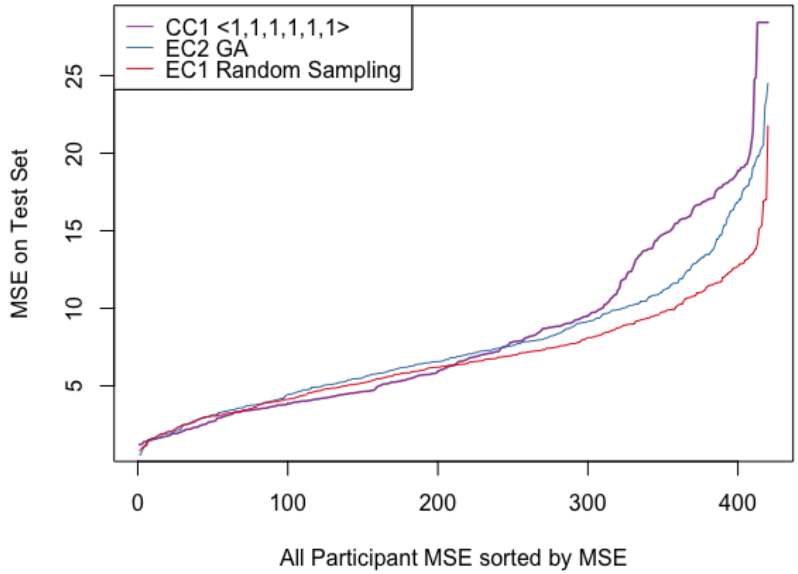
Condition	Aggregation Intervals	Solutions per Participant
CC1	$< 1, 1, 1, 1, 1, 1 >$	Fixed
CC2	$< 2, 2, 2, 2, 2, 2 >$	Fixed
CC3	$< 3, 3, 3, 3, 3, 3 >$	Fixed
CC4	$< 4, 4, 4, 4, 4, 4 >$	Fixed
EC3	Random Sampling	Best of 300
EC4	Genetic Algorithm	Best of 300 (10 gen * 30 pop)

CASE STUDY: DEPRESSION

- Results (MSE):

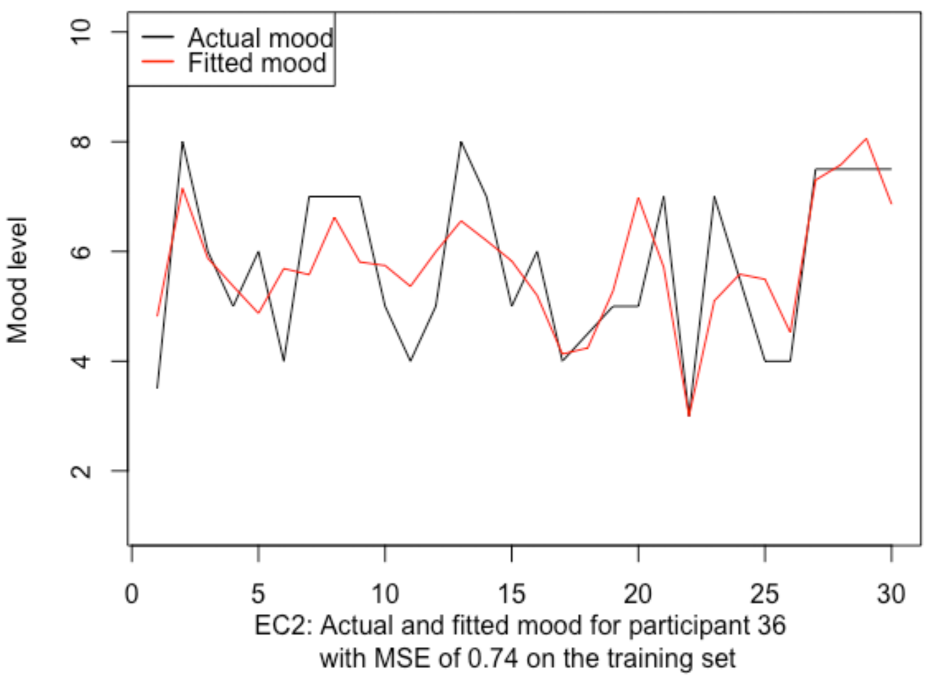
Condition	Training Set		Validation Set		Test Set	
	MSE	SD	MSE	SD	MSE	SD
CC1	1.30	0.05	6.09	0.73	8.18	1.09
CC2	0.91	0.05	7.25	1.29	9.82	1.87
CC3	0.85	0.04	8.98	1.17	9.75	1.80
CC4	0.78	0.04	10.17	1.36	10.79	2.07
EC1	0.94	0.05	2.22	0.32	6.68	1.40
EC2	0.93	0.05	2.15	0.31	7.60	1.38

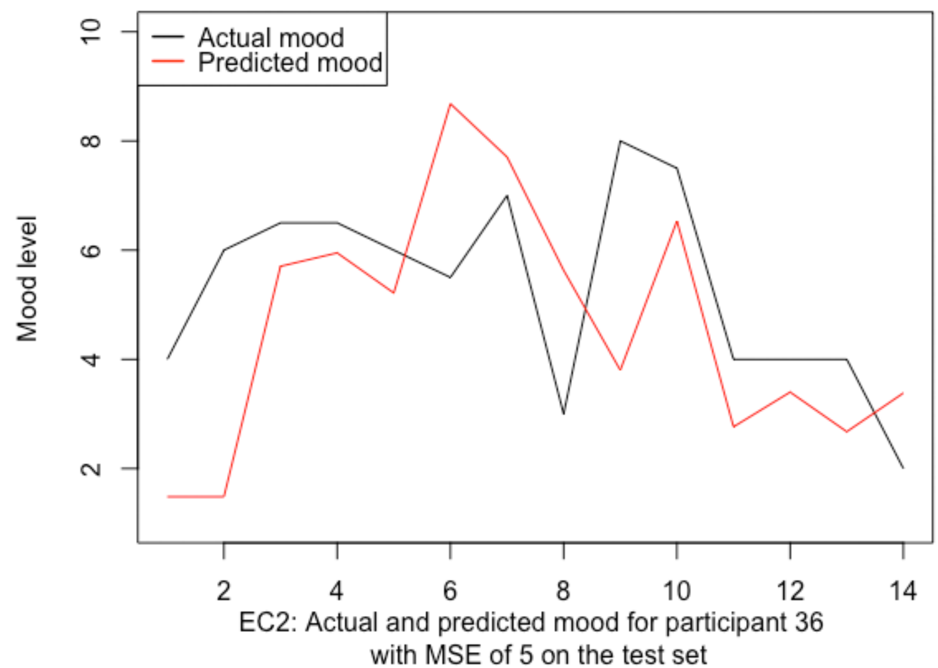
- Becomes significantly better



CASE STUDY: DEPRESSION

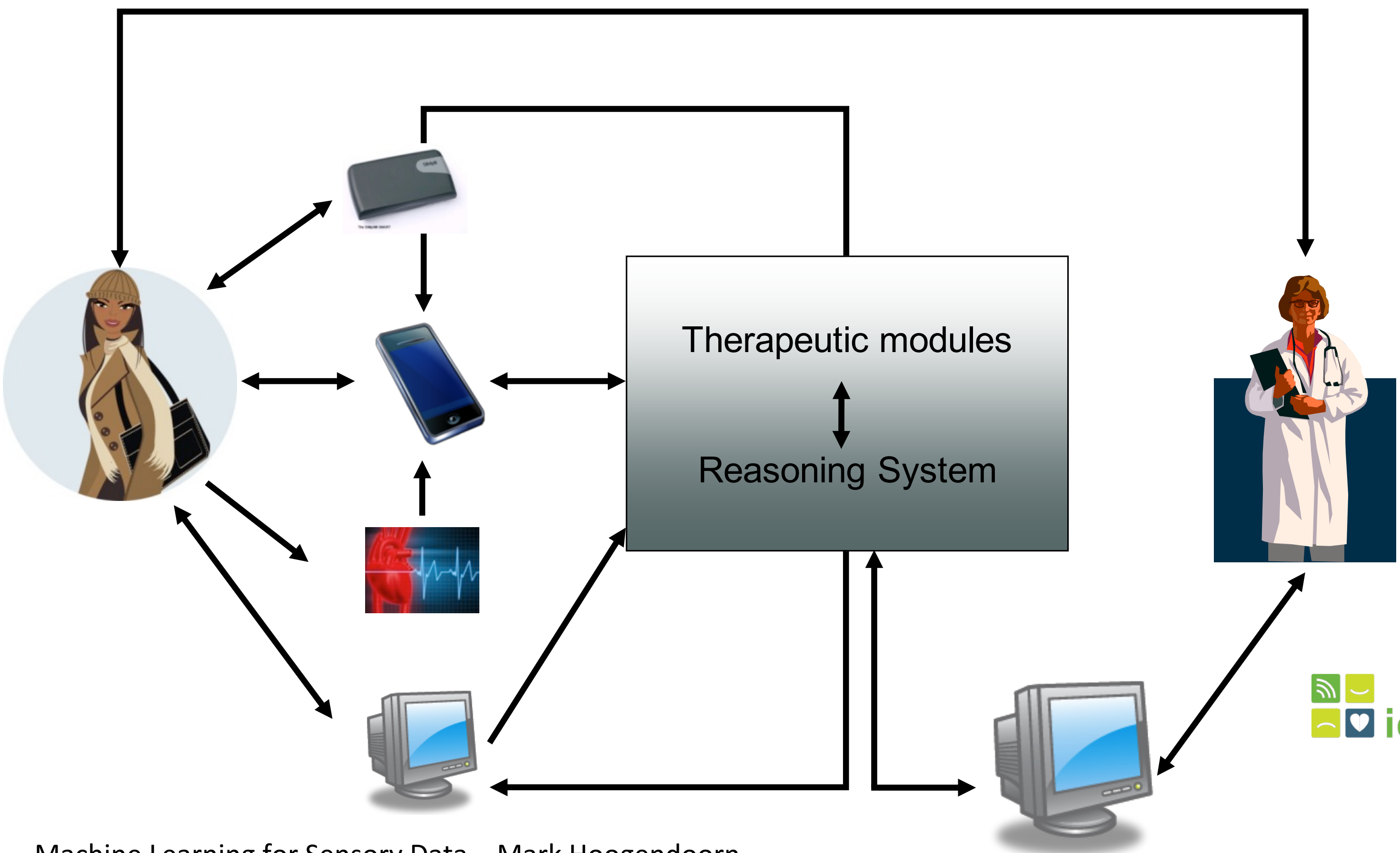
- Example patient:





INTEGRATE PREDICTIONS IN SYSTEM

- We have developed an approach for depression (knowledge driven personalization):

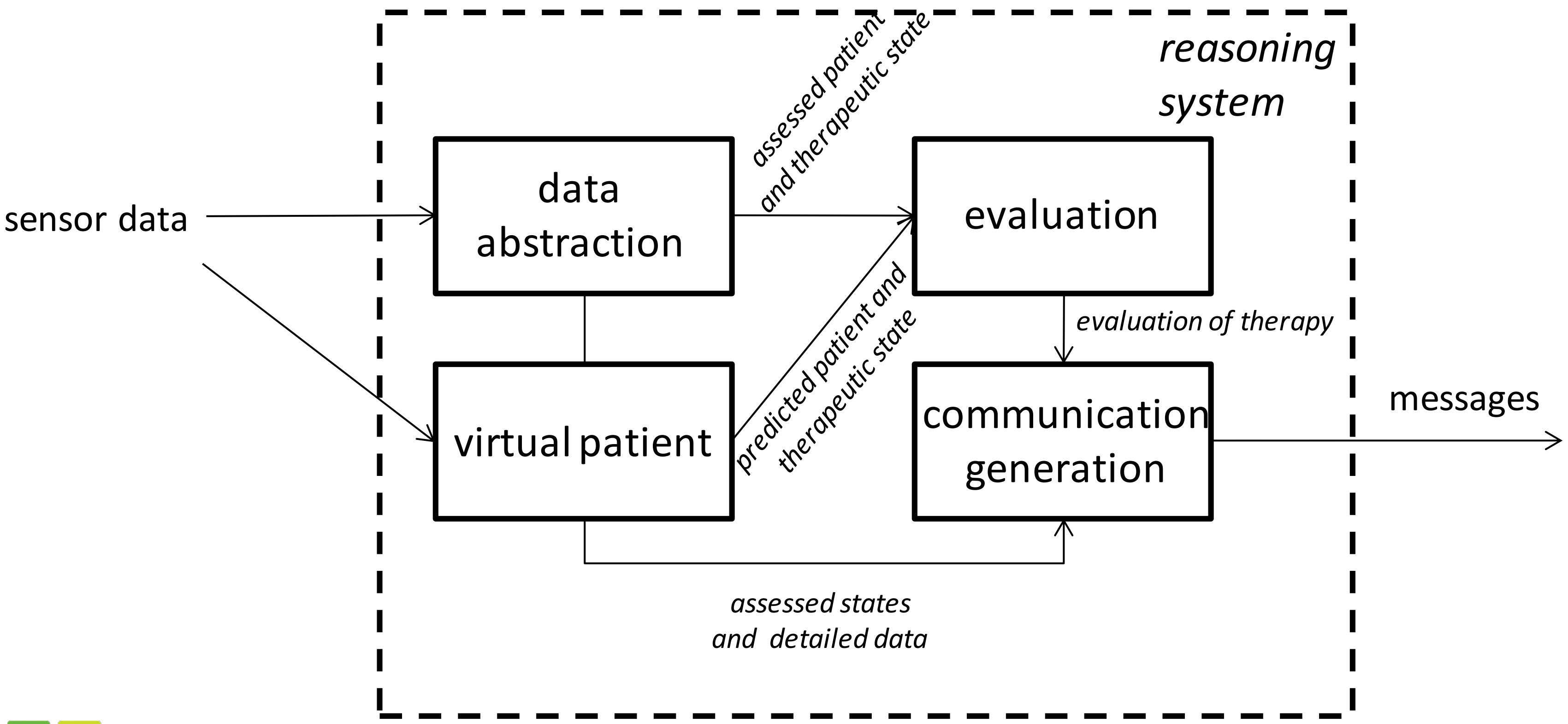


ict4depression

VU

WE PROVIDE TAILORED FEEDBACK

- Reasoning module:



ict4depression

PERSONALIZATION/FEEDBACK

- Communication generation: send appropriate feedback
- Types of feedback:
  - Feedback to the patient
    - Information permanently available via the website
    - Direct information via mobile phone
    - Reminders for therapeutic activities
    - Motivational messages (based on predictive models)
    - Weekly feedback about progress (based on predictive models)
    - Therapeutic advice (based on predictive models)
  - Feedback to GP
    - Permanently available via website
    - Suggestions about therapy change (based on predictive models)
    - Warning in case of very low mood levels

ict4depression

VU

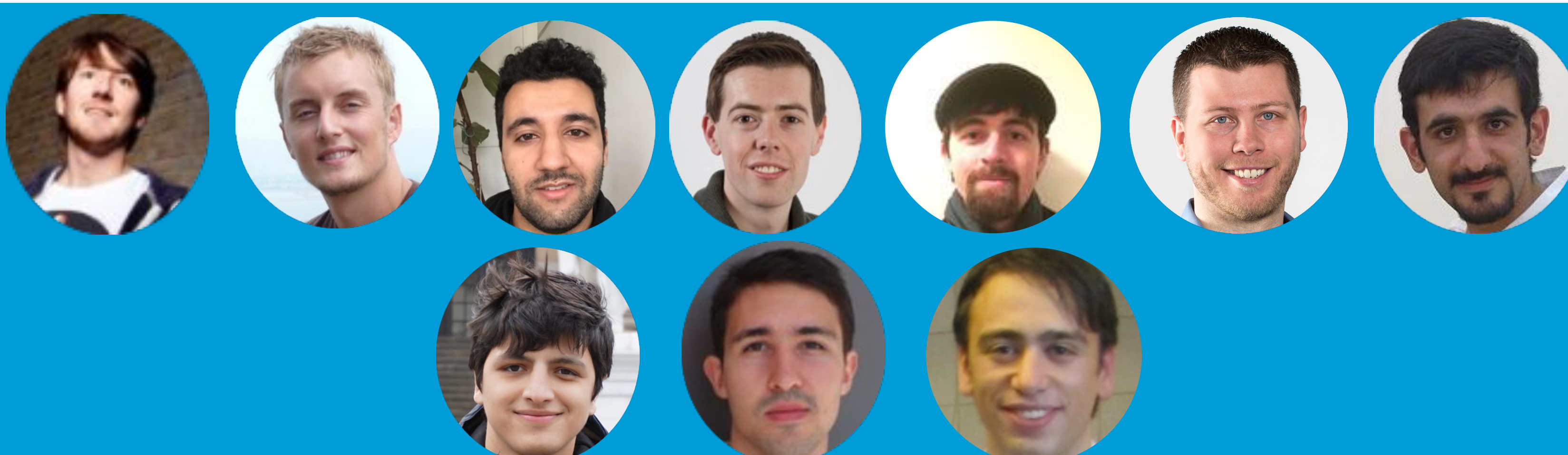
SUMMARY

- Machine Learning for Health poses a lot of challenges to generate accurate models
- Wearables pose interesting challenges to create accurate predictive models
- Feature engineering is crucial for some case
- Personalization is the next step
- Lots of developments seen now that focus on end-to-end learning, but: lack of insight

THANKS (NOT COMPLETE)

Former PhD students/Junior researchers

Current PhD students/PostDocs



Other collaborators for presented research (not exhaustive)

m.hoogendoorn@vu.nl