MLPM Project

Dr. Svetlin Penkov

Fall 2019

Problem Requirements

- Supervised learning problem
- Real data
  - Lots of open source datasets
- Sequential / Temporal aspect of the data

Choosing a problem

- Recommended domain - human activity recognition
- Problem of your choice
  - Make sure it meets the requirements
  - Make sure you discuss it with me before starting any work on it
    - During tutorials
    - linkedin.com/in/svpenkov/
    - sv.penkov@gmail.com
- Everyone should have chosen a problem and discussed it with me by 7th Jan 2020

Human Activity Recognition

WISDM Smartphone and Smartwatch Activity and Biometrics Dataset

WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set

Download: Data Folder, Data Set Description

Abstract: Contains accelerometer and gyroscope time-series sensor data collected from a smartphone and smartwatch as 51 test subjects perform 18 activities for 3 minutes each.

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	15630426	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	6	**Date Donated**	2019-10-06
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	28648

Source:

Dr. Gary Weiss, gaweiss '@' fordham.edu, Computer and Information Sciences Department, Fordham University.

SSAB Dataset

+

THE 18 ACTIVITIES REPRESENTED IN DATA SET

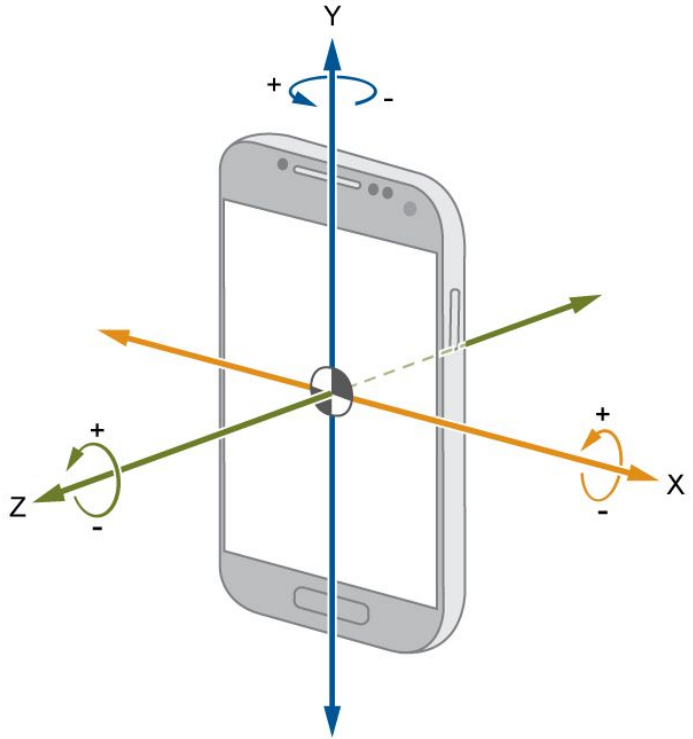
Activity	Code
Walking	A
Jogging	B
Stairs	C
Sitting	D
Standing	E
Typing	F
Brushing Teeth	G
Eating Soup	H
Eating Chips	I
Eating Pasta	J
Drinking from Cup	K
Eating Sandwich	L
Kicking (Soccer Ball)	M
Playing Catch w/Tennis Ball	O
Dribblinlg (Basketball)	P
Writing	Q
Clapping	R
Folding Clothes	S

SSAB Dataset Raw

Accelerometer: linear acceleration
Gyroscope: angular velocity

TABLE 1
SUMMARY INFORMATION FOR THE DATASETS

Number of subjects	51
Number of activities	18
Minutes collected per activity	3
Sensor polling rate	20Hz
Smartphone used	Google Nexus 5/5x or Samsung Galaxy S5
Smartwatch used	LG G Watch
Number raw measurements	15,630,426



SSAB Dataset Features

- Proposed dataset features for 10s windows
- Feel free to explore them or propose your own features

TABLE 5
LAYOUT OF ARFF HEADER FILE

Line #	Attribute Name	Attribute Type or Values
3	ACTIVITY	{A,B,C,D,E,F,G,H,I,J,K,L,M,O,P,Q,R,S}
4-13	X{0-9}	numeric
14-23	Y{0-9}	numeric
24-33	Z{0-9}	numeric
34-36	{X,Y,Z}AVG	numeric
37-39	{X,Y,Z}PEAK	numeric
40-42	{X,Y,Z}ABSOLDEV	numeric
43-45	{X,Y,Z}STANDDEV	numeric
46-48	**{X,Y,Z}VAR***	numeric
49-61	**XMFCC{0-12}***	numeric
62-77	**YMFCC{0-12}***	numeric
75-87	**ZMFCC{0-12}***	numeric
88-90	**{XY, XZ, YZ}COS***	numeric
91-93	**{XY, XZ, YZ}COR***	numeric
94	RESULTANT	numeric
95	**class***	{16XX}

* The value for class is the subject identifier for the file and is a single value between 1600 and 1650.

SSAB Dataset Open Problems

- Classify activity types (online or offline)
- Classify activities for a given person with high accuracy
- Classify user from activity pattern (probably the hardest)
- Semi-supervised learning (supervised + unsupervised):
    - Any of the above with as few labels as possible

Solution Requirements

- Feature learning or selection e.g.:
  - Clustering
  - Dimensionality reduction
  - Sparse priors
  - Proxy linear learners
- Supervised learning approach
- Model selection
  - At least 1 baseline model

Outcome Requirements

- Detailed specification of the proposed model:
  - Likelihood, prior distributions
  - Identify free parameters and hyperparameters
  - Describe learning and inference procedures
- Implementation of the proposed model
  - Language of your choice
  - Specialised ML packages - allowed, but talk to me
- Experimental results
  - Plots and tables
  - Well chosen metrics
  - Critical analysis

Purpose of the Project

- Design and implement a probabilistic ML model
- Solve a real problem
  - Open
  - Hard
  - Hairy
- No right or wrong solution, but...
- Objectively better or worse

Project Outcome

- 4-6 page report (printed A4) with a standard paper layout
  - Introduction
  - Related Work
  - Methods
  - Results
  - Discussion & Conclusion
- 5-10 min presentation
- Expected due date 11 Feb 2020 (may be changed)
- Ultimately you have to convince me (and the others) that you have come up with a good model for your problem.
- We will use the remaining 2 tutorials for Q&A regarding the project

Guidelines: DOs

- Make sure your data is clean (i.e. filter and remove outliers)
- Start from the simplest possible idea you have and iteratively build on
- Be creative and try out your ideas! (there is no right answer...)
- Explore various methods as long as you understand the probabilistic models behind them
- Prioritise thoughtful design and thorough analysis over
  - Complex models
  - Marginal improvements in the final metrics
- Split your dataset in
  - 70% training + 30% test
  - 50% training + 20% validation + 30% test

Guidelines: DON’Ts

- Avoid deep learning style end-to-end architectures
  - You can use neural networks if they have a very specific role e.g. (learning an approximate distribution with variational inference)
  - You should clearly demonstrate that there are no other simpler and reasonable choices
- Don’t build overly complicated models
  - Model selection should be part of your iterative process
- Don’t just *“design”* a model because somebody else has implemented it
- Don’t just use code for methods you do not understand

Example: Classification based on Multiple Temporal Sequences

Cognitive Control Loop

Eye-tracking Glasses

Intention inference

ERP Detection & Classification

Emotiv EPOC EEG Headset

Robot

Observe

Act

Perceive

Working Environment

Act

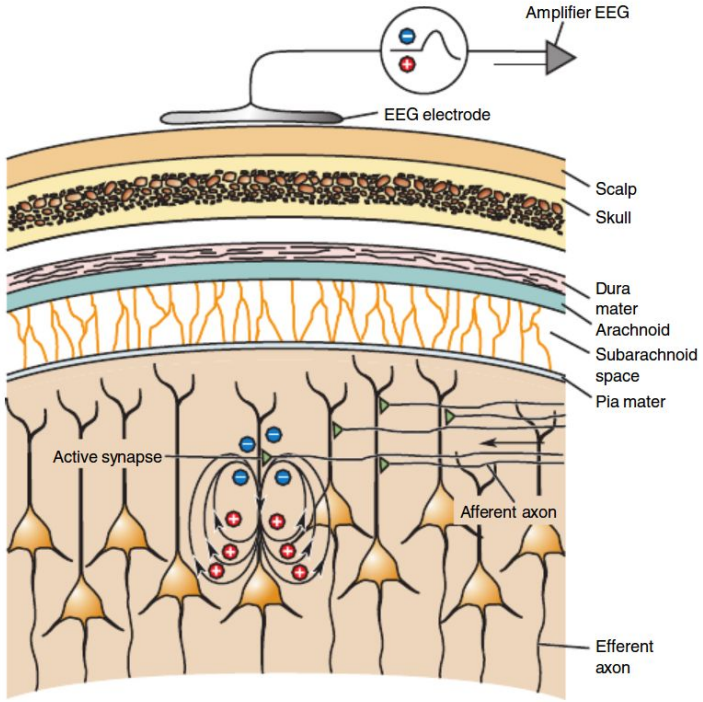
Perceive

Human

EEG & Event Related Potentials

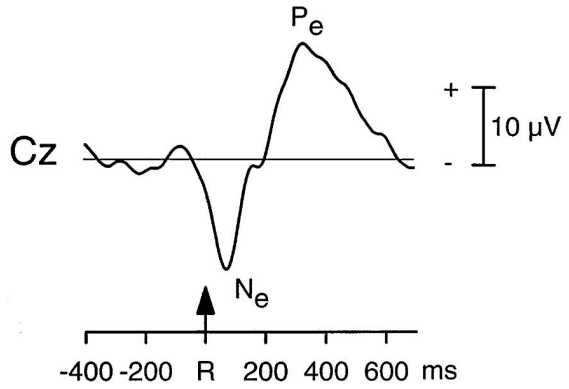
Spatiotemporal patterns of brain activity corresponding to:

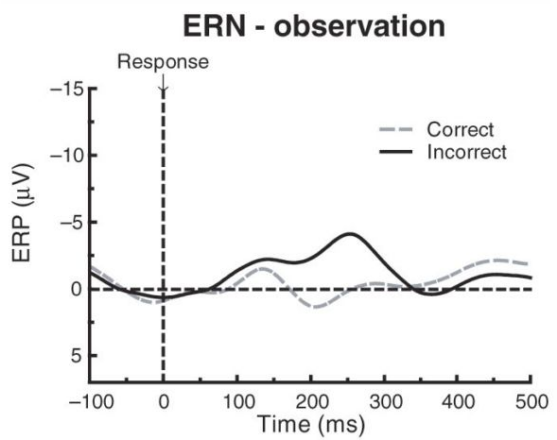
- Anticipation of action or a stimulus
- Violation of the expected sensory input
- Perception of error



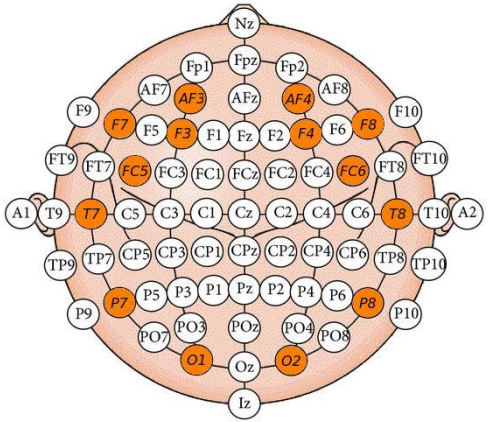
Error Related Potentials (ErrP)

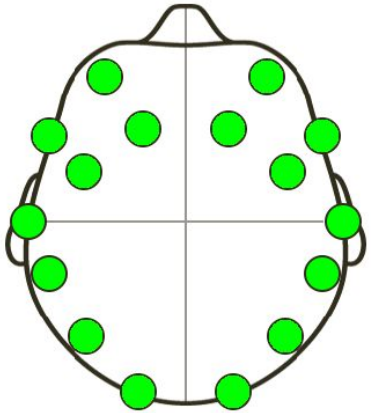
- Response ErrP
- Feedback ErrP
- Observation ErrP
- Interaction ErrPs

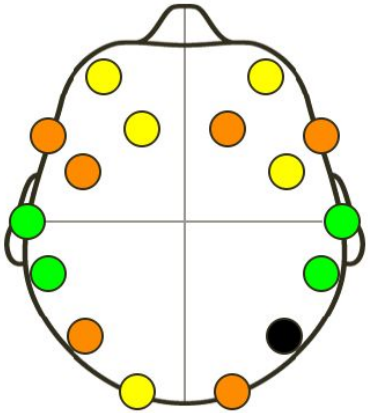




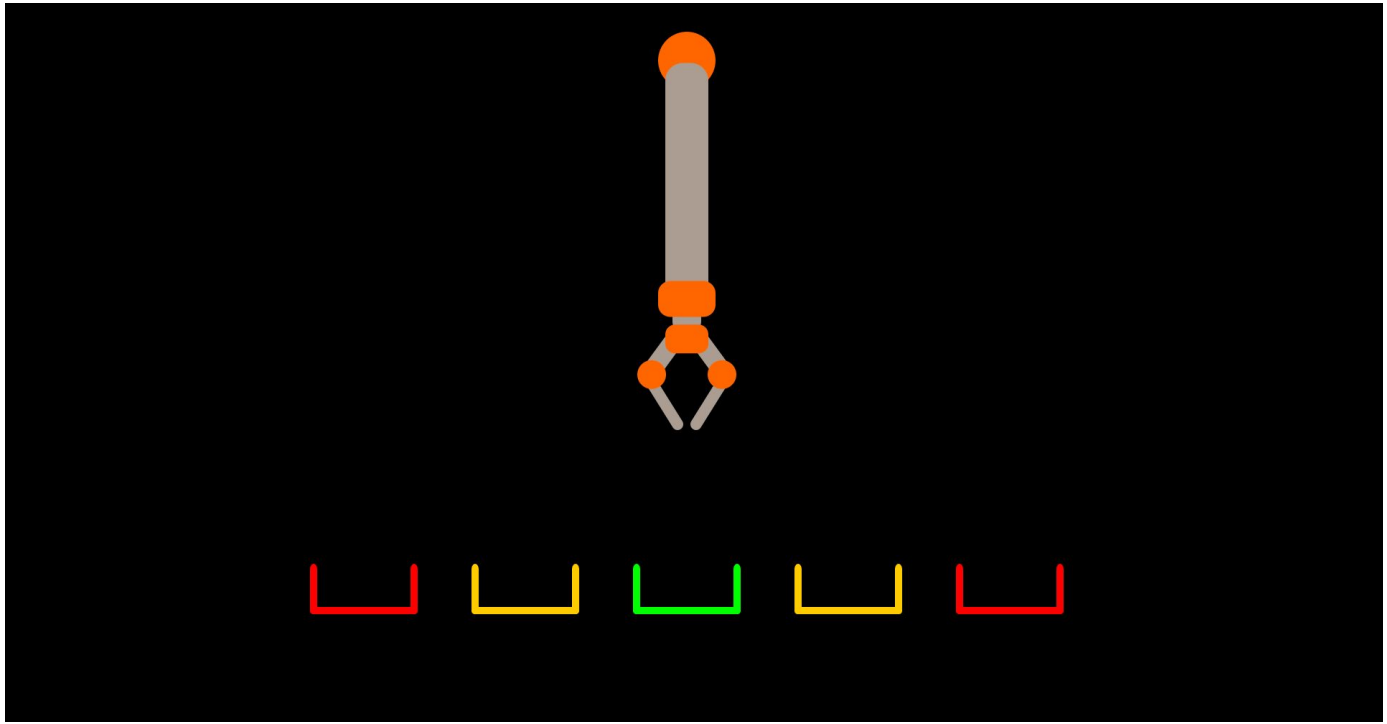
Emotive EEG Headset



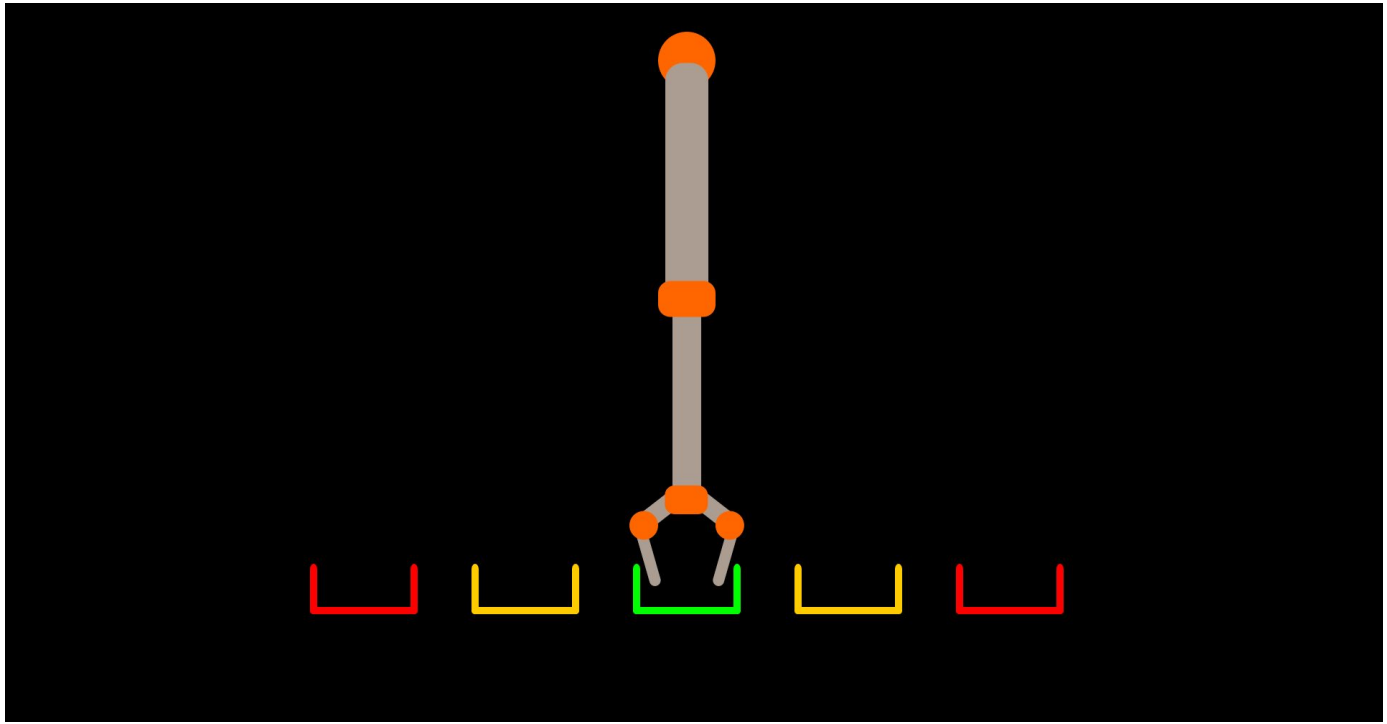




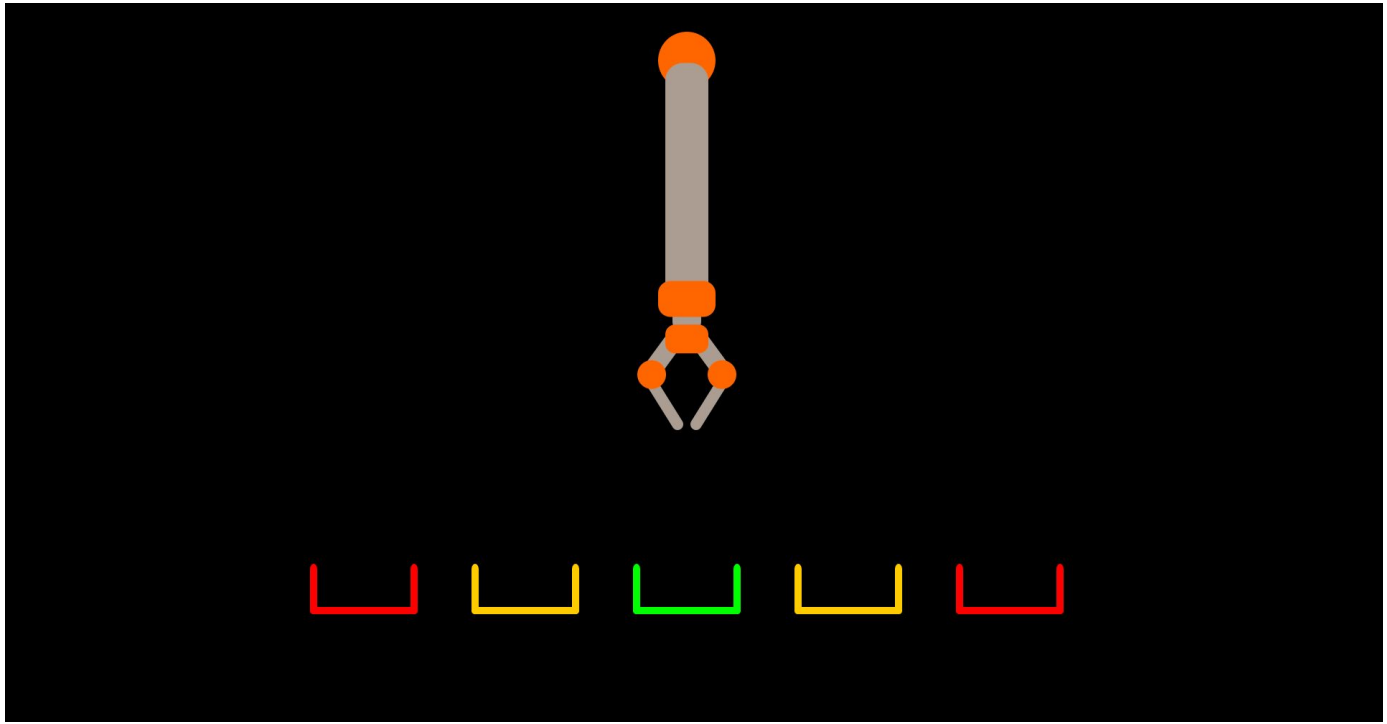
Experiment



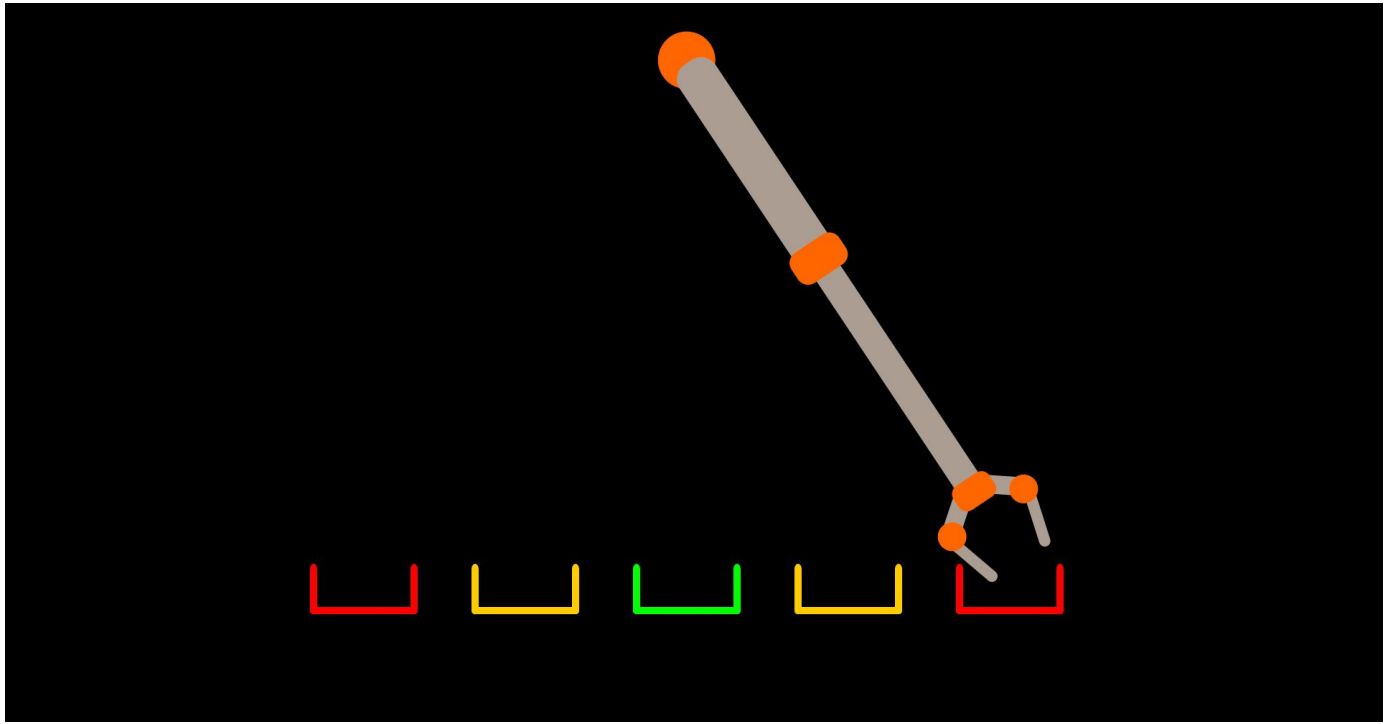
Experiment



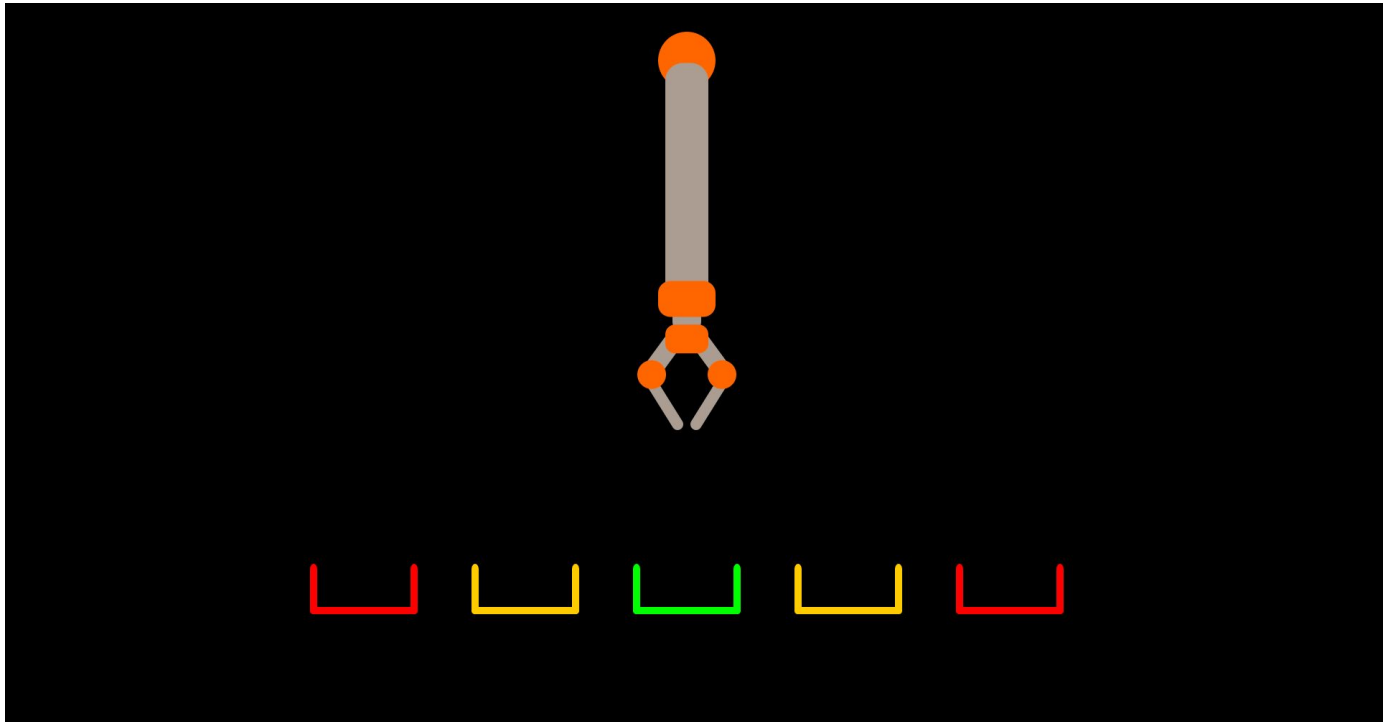
Experiment



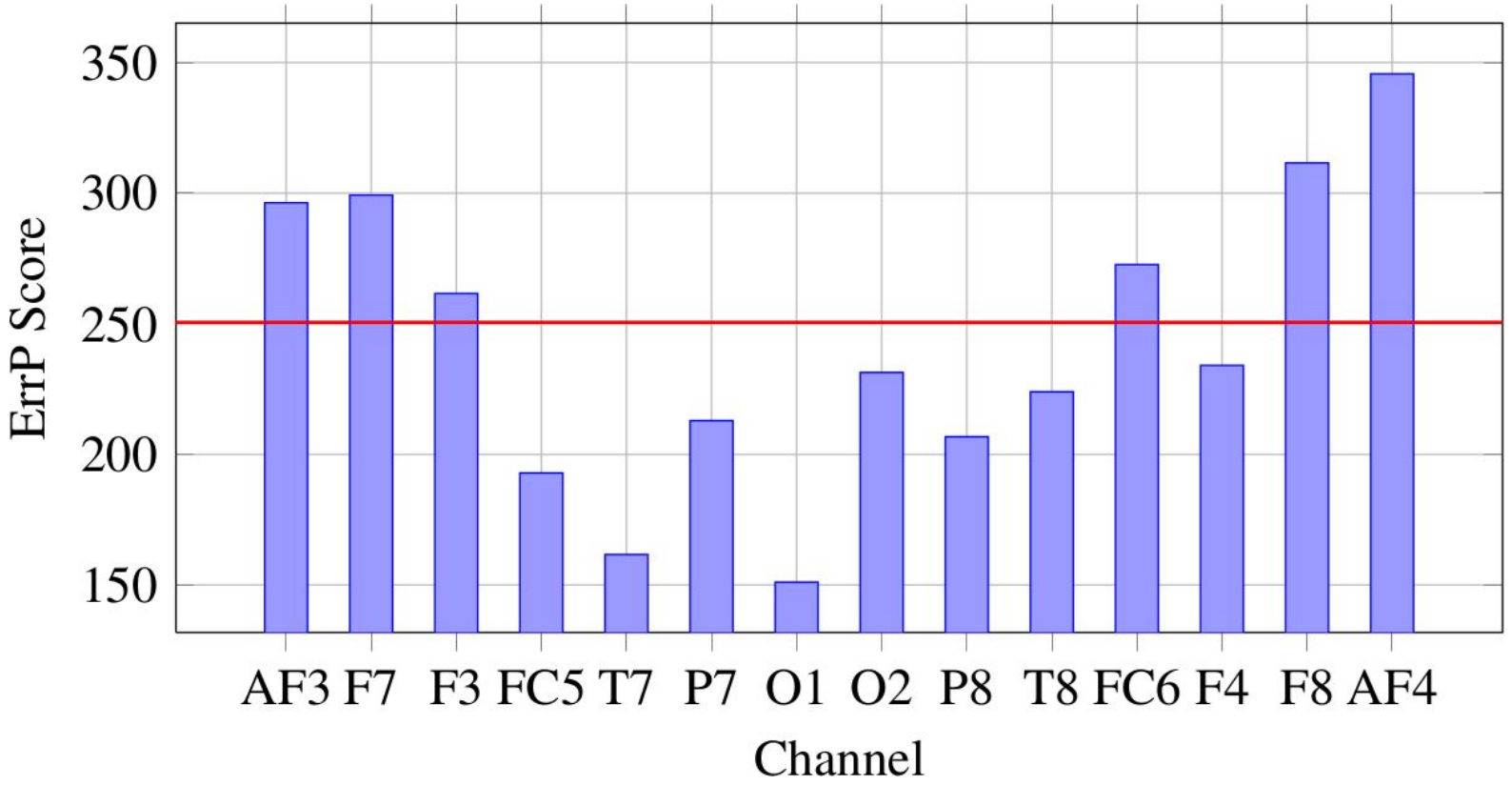
Experiment



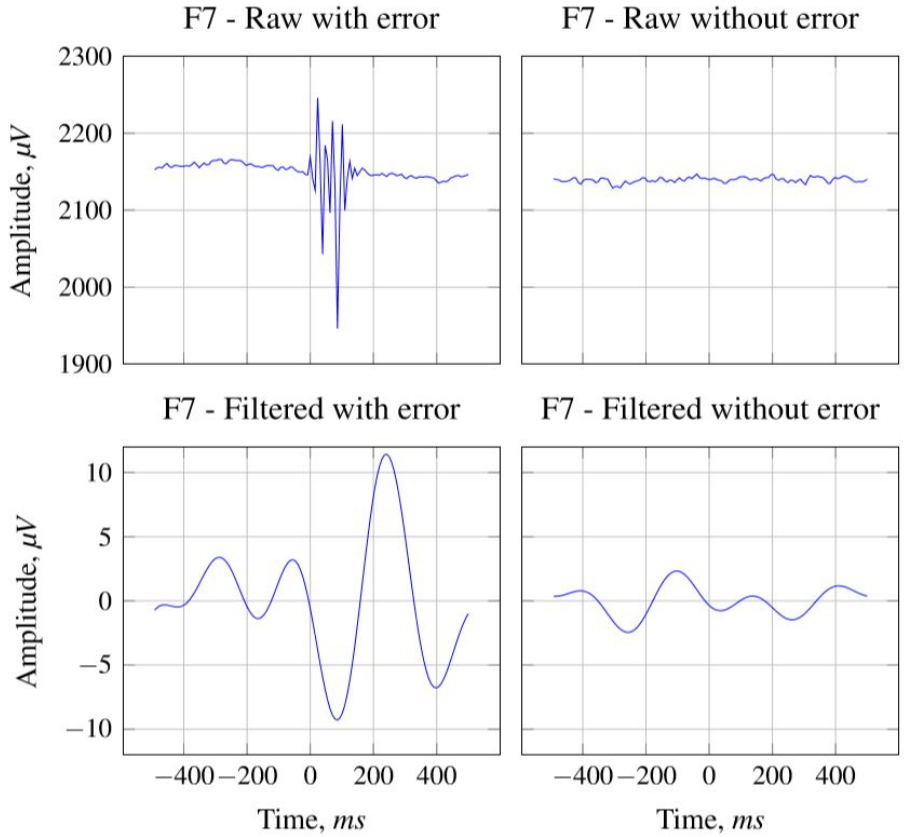
Experiment



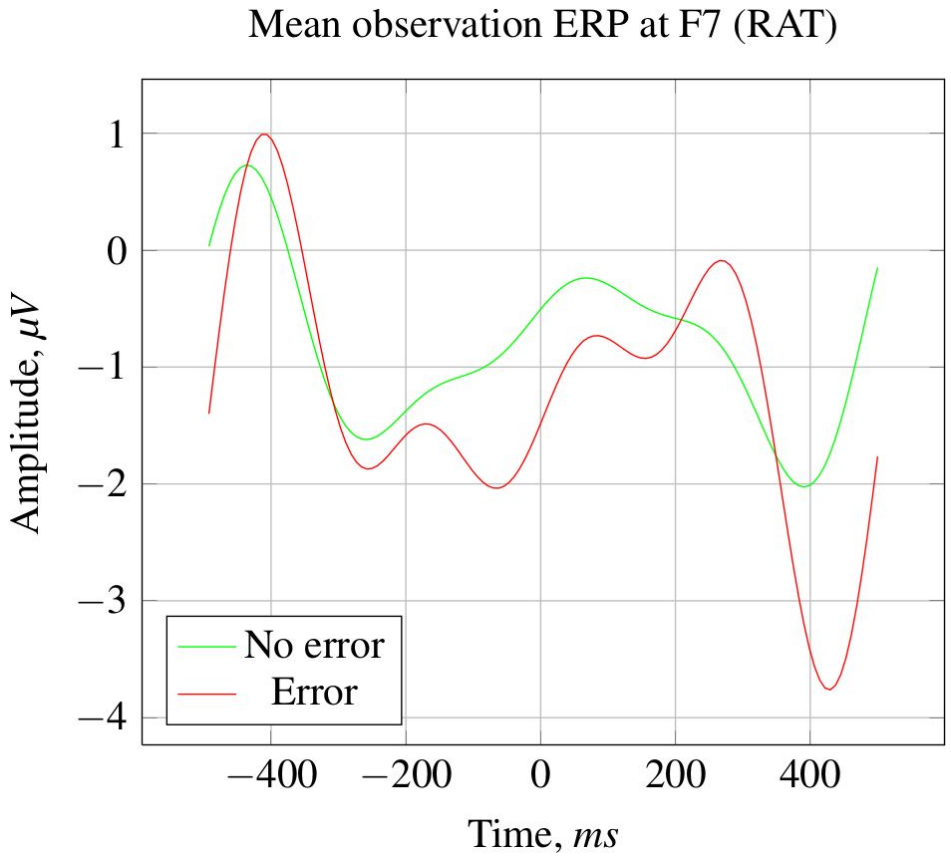
Channel Selection (Robust Discriminant Power)



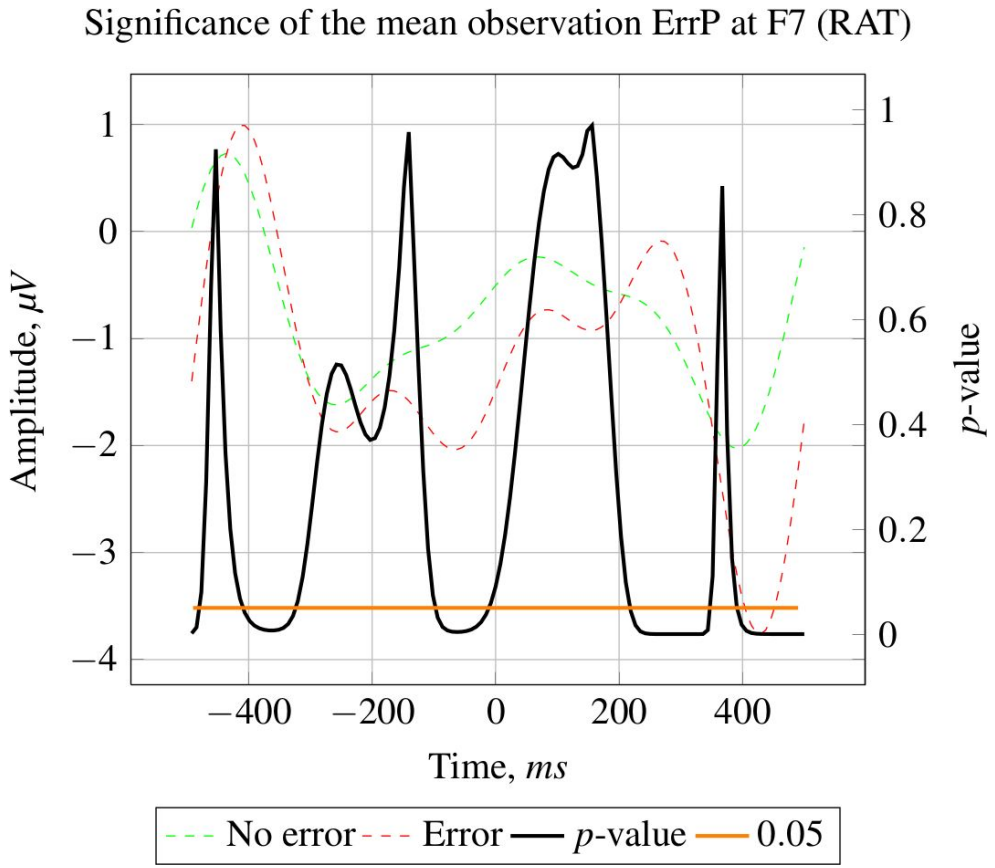
Data Cleaning



Mean ERP

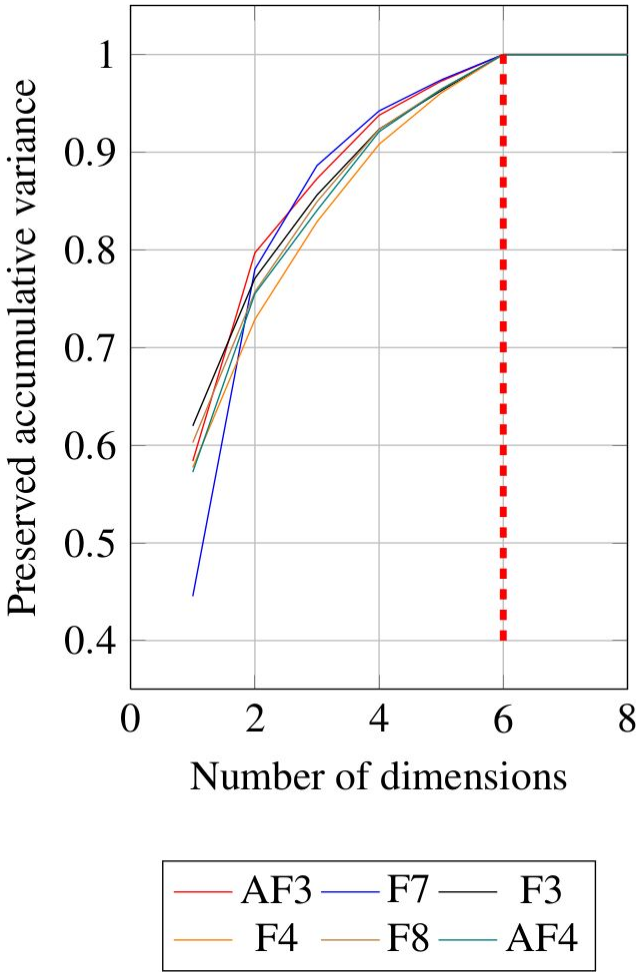

Mean observation ERP at F7 (RAT)
Amplitude, μV
1
0
−1
−2
−3
−4
−400
−200
0
200
400
Time, ms
No error
Error

ANOVA


Significance of the mean observation ErrP at F7 (RAT)
Amplitude, μV
1
0
−1
−2
−3
−4
−400
−200
0
200
400
Time, ms
1
0.8
0.6
0.4
0.2
0
p-value
No error
Error
p-value
0.05

Dimensionality Reduction

- Analyse 1 sec segments
- 1 segment = 128 samples
- 128-dimensional data points
- Most dimensions are correlated!



Machine Learning

Classifier	***Best Channel***	***Accuracy (%)***	***F1 Score***
Logistic Regression	*AF4*	***69.66***	*0.5172*
Mixture Of Gaussians	*AF4*	*57.44*	***0.5297***
Hidden Markov Model	*F3*	*51.20*	*0.3973*