# Data Analysis and Model Classification
# Miniproject 1: Description and Assignment

Fumiaki Iwane Ping-Keng Jao Bastien Orset Julien Rechenmann
Ricardo Chavarriaga José del R. Millán

## Objective

The first miniproject will make you explore the realm of supervised machine learning problems. You will be provided real-world dataset (descriptions below) which you will use throughout the whole miniproject. You are ultimately requested to familiarize yourself with the data (statistics, distributions, visualization), apply a suitable feature selection method, establish a classifier which should be able to generalize to new data from the same problem, and properly estimate the performance of your classifier.You will be provided with guide sheets over the course of the next weeks which help you with the implementation of the course concepts for your mini-projects. They are supposed to give you the structure necessary to be able to **search and find (theoretical and practical) information on your own** for most of your questions. The "hands on" sections on the guide sheets provides you with tasks to implement and questions to answer. Although none of these tasks are mandatory, we strongly advise you to work through the entire guide sheet week by week, as they will help you progress through the miniproject and will prepare you well for the exam.

## Dataset descriptions

Event Related Potentials (ERPs) are electrical brain waves generated as reaction to certain external events. Here, participants watched a cursor moving towards an indicated target. In 30% of the time the cursor moved the opposite way unexpectedly, eliciting an error-related ERP in the participant. This dataset contains 597 samples (observed cursor movements) for one participant and 2048 continuous-value features corresponding to the signal amplitude for sixteen EEG channels at different time points. For each sample the labels (the types of the cursor movement) are provided.

The dataset is given in the form of an .mat file (MATLAB file format). **To load it, you will have first to register on Kaggle and join the competition** : `https://www.kaggle.com/t/6787641bdfc9452b8325add185f4420f`. Then, you can find the dataset on Kaggle: `https://www.kaggle.com/c/damc/data`. The dataset contains two variables:

**features** is a matrix of the structure *samples* × *features*. You will use this subset for doing all your analysis and computations (feature selection, cross validation, classifier construction).

**labels** is a column vector where every element is the label to the corresponding sample (column vector) in the features matrix.

## Nomenclature

Throughout this course (and beyond) there is a certain nomenclature linked to data analysis. The main reason for using a fixed vocabulary is to be able to communicate one's problems and findings without uncertainty or further demand for definition or explanation. Therefore, **please use this exact nomenclature in your reports**! Do not be worried about "ugly" word repetitions. We rather prefer to know exactly what you mean, than come closer to a novelesque reading experience (which will not happen anyway :P).

**A sample** is an observation of an investigated group or phenomenon. When multiple variables are measured for the same observation, the resulting datapoint aka sample will be multidimensional, and thus be represented as a vector.
*Example: in the ERP dataset each sample represents an EEG waveform generated after a discrete single movement of a cursor. It is represented as a vector with 2048 dimensions.*

**A feature** is any combination or transformation of the measured variables.
*Examples of features from a variable vector $[a, b, c, d]$: variable $a$. The ratio of variable $b$ to variable $c$. The Fourier transform of variable $d$.*

**A dataset** is a collection of observations (samples) from a certain study or experiment. The ensemble of the collected data can then be represented as a 2-D matrix of size $n \times m$, where $n$ is the number of samples and $m$ the number of features extracted for each sample.
*Example: the ERP dataset is given as a 2D matrix:* $597\,\textit{samples} \times 2048\,\textit{features}$.

**A class** is a population (or experimental condition) from which data is drawn.
*Example: in the ERP dataset, the two classes are "correct" and "erroneous" movement of the cursor.*

**A label** is the class-identifier of each sample in the dataset.
*Example: in the ERP dataset, the label "1" corresponds to "correct movement" and "0" to "erroneous movement".*

# Report

At the end of the first miniproject you are requested to hand in a report, motivating the choices you made and discussing the results you obtained from your implementation. This report should cover **all aspects** from **dataset exploration** over **feature selection** to **classification**, as well as the role of **(nested) cross-validation** in your analysis. Please stick to the following structure:

**Introduction** In 5 sentences, state the research question and the goal of the report.

**Methods** Describe step by step and explain in details what you did exactly, including

- where in the the processing pipeline do you apply method X or Y? Before the cross-validation? Inside the cross-validation?
- on which (part of a) dataset you are applying method X or Y (Training set? Folds of training set? Testing set? Validation set?)
- which parameters you choose to fix (and to what value)? And which parameters you choose to optimize (using cross-validation)?
- the rationale behind all your choices (why method Y and not X or Z?).

**Results** This section is uniquely reserved to show the results you obtained. Of course, it is not necessary to put in the outcome of every single thing you have tried for this miniproject. So choose well what data to show and what not to show.

You should be creative and pragmatic about how to display your data. Add mean and standard deviation to your plots when possible. You want to show for example how your train/validation error evolves as you include more features, but you don't want to show 10 times the same plot with different parameters (e.g. one for LDA for feature selection method A, one for LDA for feature selection method B, one for QDA for feature selection method A, one for ...). Pick only the interesting ones that illustrate your choice!

**Discussion** Use this chapter to briefly sum up your results and discuss about following points:

- What are the particularities of your dataset?
- Which methods do you recommend for it and why?
- Anything especially difficult or cool you did, and why?
- Any points where your methods did not suffice, and you would be in need of more advanced concepts?

**Note:** Your report **must not exceed 6 pages** (excluding a titlepage)! Every additional page will not be considered for grading.

# Report writing: how to's

- There are different ways how you can present your results: you could just write them in the continuous text, make a table, or visualize them in a plot.
- Graphs are an extremely powerful tool; they can make it easier to grasp results and their implications. But to unleash their full potential, the reader has to understand **everything** the graph is showing. Therefore, make sure your graphs **always** contain
  - a title (`title`)
  - axis labels for all axes (`xlabel/ylabel`)
  - meaningful axis units. You can specify them manually, if necessary, with `axis`
  - a legend, to show which color / linestyle corresponds to what data (`legend`)
  - a caption in the report, describing what the graph shows, and why this is important.

  Below you can find an example of how to present a figure so to maximize content and readability. (The data presented here is not related to any miniproject.). Please be also careful to present your figure in the report in a way that it can be observed by the assistant. You can easily change the fontsize of your figure on Matlab.

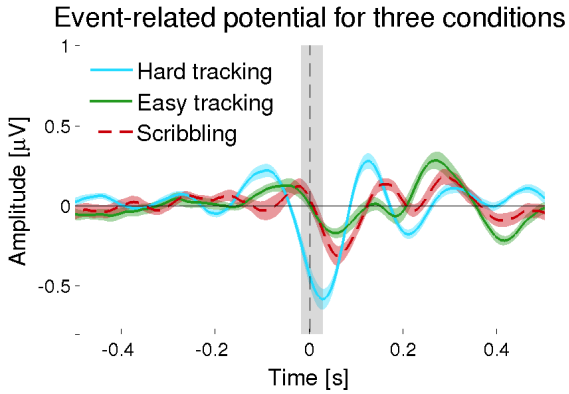


**Figure 1:** The event-related potential (ERP) is modulated by the difficulty of the task. In hard tracking conditions subjects elicit on average higher amplitude ERPs. The thick lines are the subject means, and the shaded areas indicate the standard error of the means. The gray-shaded bar marks the time of significant difference between the hard and both other tasks ($p = 0.0084$).

- Be aware that a result is not significant until tested for significance (e.g. by using a t-test)! So if you state that something is significant, **always** put the test statistic (e.g. p-value and chosen level of significance).
- Use exact language wherever possible! Describe dynamics and uncertainties always in a scientific way and **with numbers**.

  Examples from previous years:
  - Since there is some scattering in the y axis, we could call it 'moderate' negative correlation. - *Yes you can, but you shouldn't. Why not just report the exact value of the correlation?*
  - If I repeat the function more than 10 times the confusion matrix varies only slightly and the clustering result is always the same, it is always centered on -1.5. - *It would be so much more helpful, if you just report the mean and standard deviation of the elements across the 10 confusion matrices.*

– As a result, the resulting centroids are not the same, but within the margin of error. - *There is no such thing as a standard "margin of error". Please simply state the mean and standard deviation.*

## Submission

The **deadline** for the report submission is **November 12th, 2018 at 23:59**. Make sure your report is in .pdf format and name it according to following convention: `Miniproject1_Group`⟨*groupnumber*⟩`.pdf`. The group number is sent out per e-mail. Please also attach any MATLAB code used to obtain the results in a .zip file. The code will not be graded and is looked into only in cases where results or plots seem dubious to find the reason (bug in the code or conceptual error).

Your finished report has to be uploaded to the moodle **only by one person per group**. The submission function will be activated approximately one week before the deadline.

## Kaggle Competition

Before submitting anything, please register your team name here: `https://www.kaggle.com/c/damc/team`. Please follow this convention to name your team: ⟨`team_name`⟩`_Group_`⟨`groupnumber`⟩.

Several times during the project, you will be asked to test your model on unseen data (199 samples × 2048 features). You are provided with this unlabelled testing dataset along with seen training data. First train your model on the training data then predict the classes of the unseen dataset. You should a vector of 199 predicted labels (made of 0 and 1). Use the function `labelToCSV.m` (available on `https://www.kaggle.com/c/damc/data`) to convert your vector in a format that is readable by Kaggle's interpreter: `labelToCSV(predictedLabels, 'predictedLabels.csv', '.')`. Upload your prediction file on Kaggle: `https://www.kaggle.com/c/damc/submit`.