The MLEnd Deception Dataset

*This year we are going to create the **MLEnd Deception Dataset**, a collection of truthful and deceptive stories narrated by individuals as their own experience, in English and in their native language.*

We hope that while working on the creation of this dataset, you will learn fundamental aspects of Machine Learning and most importantly, that you will have fun.

Background

We are going to create new dataset, the MLEnd Deception Dataset, consisting of labelled audio recordings of storytelling. Each audio recording will consist of a story describing a personal experience. Each story might be a true recount or a false recount.

How will you contribute to the dataset? By recording yourself telling such stories. Audio recordings will be **anonymised** (i.e. the identity of the storyteller will not be disclosed) and the dataset will be made public in the future for anyone around the world to play with.

Overview of the task

During the creation of the dataset, you will be asked to:

- Think of 6 stories, such that
  - 3 are true recounts (2 in English, 1 in Native).
  - 3 are false recounts (2 in English, 1 in Native).
- Record 6 audio files, one per story.
- Upload the audio files, together with additional labels.
- Fill in an optional demographic questionnaire.

Table 1: Audio Files

Language	True Stories	Deceptive Stories
English	2	2
Native	1	1
Sub-total	3	3
Total stories	**6**	

Contributing to the MLEnd Deception Dataset is worth 4 marks.

Please read, understand and follow the instructions in this guide.

If you feel that you are not able to record yourself or have any concerns about this activity, please let us know and we will discuss alternative arrangements.

Guide: Instructions in details

This guide will describe:

1. Choosing your stories
2. Language
3. Structure of the story
4. Recording a story
5. How to use Audacity to record a story
6. Naming the audio files and labelling
7. Submitting your samples
8. Demographic questionnaire
9. Final dataset
10. Dos and Don'ts

1. Choosing your stories

You will be recording yourself narrating **6 stories**, 3 corresponding to real-life experiences and 3 deceptive (i.e. imagined or false) stories. To choose a story, you need to know what type of story you can choose from. In this section we explain what kind of stories you can choose. This will also be helpful to make deceptive or imaginative stories.

1.1 Type of story

We believe that you have plenty of experiences and adventures that you could choose from. However, to keep this task simple and consistent, we would like you to choose stories corresponding to a **time when you visited a place or a person**. Your story should **narrate what happened** and **what you did.** You could also **describe how you felt**.

As an example of a story about a time when you visited a place, you could choose a time when you visited a memorable place and start your narration as:

"When I visited Italy in 2016, I landed in Rome and …".

As an example of a story about a time when you visited a person, you could choose a cherished individual, and your story could start as:

"In 2010 I visited my grandmother. I asked my her …".

1.2 Emotion of the story

Your actions and incidents are an important aspect of your story. You can also add your feelings and emotions to your story. The stories you tell don't have to be fun or happy memories only. The emotion of your stories could range from happy to sad, fun and excitement to anger, or it can be calm or contained.

1.3 Length of story

For the dataset to be useful, we need to record stories that are **at least 2 minutes long**. This corresponds to approximately 300 words. Whilst a story can be longer if you like, to keep it consistent, the duration should never exceed **4 minutes**. Your audio file should **never exceed 50MB**.

1.4 Appropriate language and context

We believe you are mature enough to understand what kind of language and context you could use. However, we would like to remind you about the wording and the context you choose for the story. ***Be mindful about the wording, phrases and context of your stories. They should not be abusive, discriminatory or hurtful to any individual or group****.* Please avoid topics of controversial nature.

1.5 Deceptive stories

A deceptive story should be narrated **as if it was real**. In other words, **it should sound truthful and real, as if you were recounting a true story.** While telling a deceptive story you should try to be as convincing as possible.

One of the ways for succeeding at telling a deceptive story is **to avoid saying something very improbable** or something that is obviously a lie. Examples of stories that you should avoid are a story about a time "when I landed on the moon" or "when I visited my great-great-great-grandfather back in 1754 using a time machine". These stories might be funny but not very useful for the purpose of this dataset.

Examples of good deceptive stories include the following.

If you have never been to USA, one deceptive story could be about a time *"when I visited Chicago, USA, back in 2019. I was roaming the streets, when I found an Italian restaurant. …".*

If you have never met your grandfather, you can tell a story of a time *"last year, I visited my grandfather and I asked him about my father's childhood. I loved hearing about how my father grew-up and how similar he was to me, when he was young…"*

Deceptive stories, like real ones, should have a minimum duration of 2 mins (maximum 4 mins), be about a time when you visited a place or a person, and use of appropriate language and context.

2. Language

Out of the 6 stories, 4 stories should be in English and other 2 should be in your native language. If your native language is English, then all the stories should be in English. There should be an equal number of true and deceptive stories in each language (See Table 1).

For example, if your native language is Spanish, you will record 2 true stories in English, 1 true story in Spanish, 2 deceptive stories in English and 1 deceptive story in Spanish.

3. Structure of a story

All the stories **MUST start with the following English sentence:**

> *"My name is _____ _____, my QM ID is _________ and this is my story."*

We will remove this sentence before making the dataset available to ensure anonymity.

After this sentence, you should start your story. For example:

"When I visited …". [English/Native Language]

For example, your first story could be as follow:

[In English]
"My name is Patrick Smith, my QM ID is 123456789, and this is my story. Back in 2010 I visited my grandmother. My grandmother loved it when I visited her, she used to make special sweets for me …"

[In Spanish – if your native language is Spanish]
"Mi nombre es Patrick Smith, mi QM ID es 123456789 y esta es mi historia. En 2010 visité a mi abuela. A mi abuela le encantaba que le visitara, solía hacerme un dulce especial…"

4. Recording a story

To record a story, you will need three materials:

1) Audacity – the audio recording software,
2) Yourself – the storyteller, and
3) A partner – the listener.

Once you have chosen a story and you have installed Audacity (see Section 5), find a partner. Your partner is just a passive listener. To record the story, find a quiet place with no background noise (avoid public spaces). Once you are all set-up, narrate your story to your partner. Your partner MUST NOT speak nor interrupt you while your story is being recorded.

Once you finish recording a story, we recommend that you **DO NOT listen to your recording**. **No one likes their own voice!** This is where your partner will come in handy. Your partner will reassure you that your story and recording is fine. Remember that **recordings will be anonymised**.

We know that you will be tempted to listen to your recording and might want to record yourself again. Please remember that these recordings are **NOT meant for a TV, MOVIE or COMMERCIAL audition**. Do not try to perfect your tone, pitch, grammar or flow of the story, just be natural, as if you were telling your story to your partner. There is no need to take multiple attempts to submit the perfect story.

If you are a partner, please reassure the storyteller that the recording is fine. **Having a partner will save you a lot of your time**!

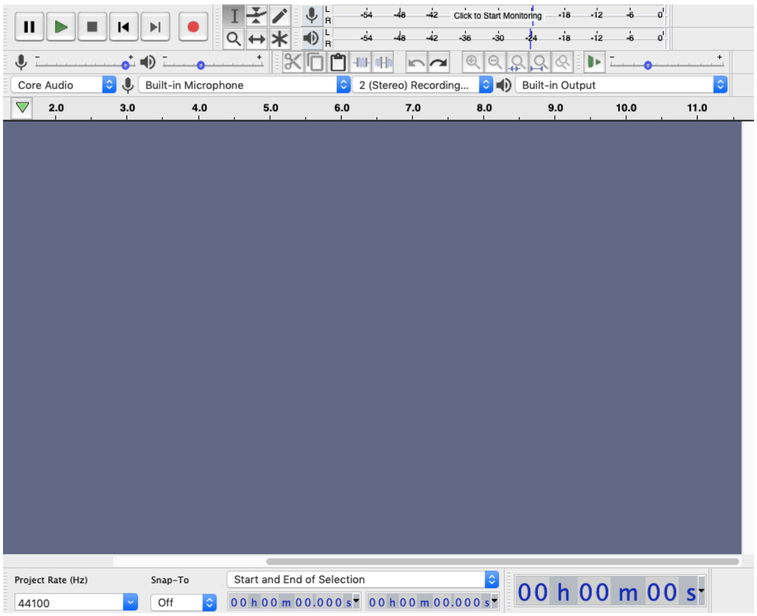
Tip: If you like, you can write down a few points of the story on a piece of a paper (flash cards), to remind you, if you forget. But **DO NOT write down the full story and read it word-by-word while recording**, as this will defeat the purpose of the data collection.

5. How to use Audacity to record a story

Audacity is a free, open-source, cross-platform audio software. You can download it from:

https://www.audacityteam.org/

The GUI looks like this (or similar to this, depending on your platform):



Core Audio

Built-in Microphone

2 (Stereo) Recording...

Built-in Output

Project Rate (Hz) 44100

Snap-To Off

Start and End of Selection

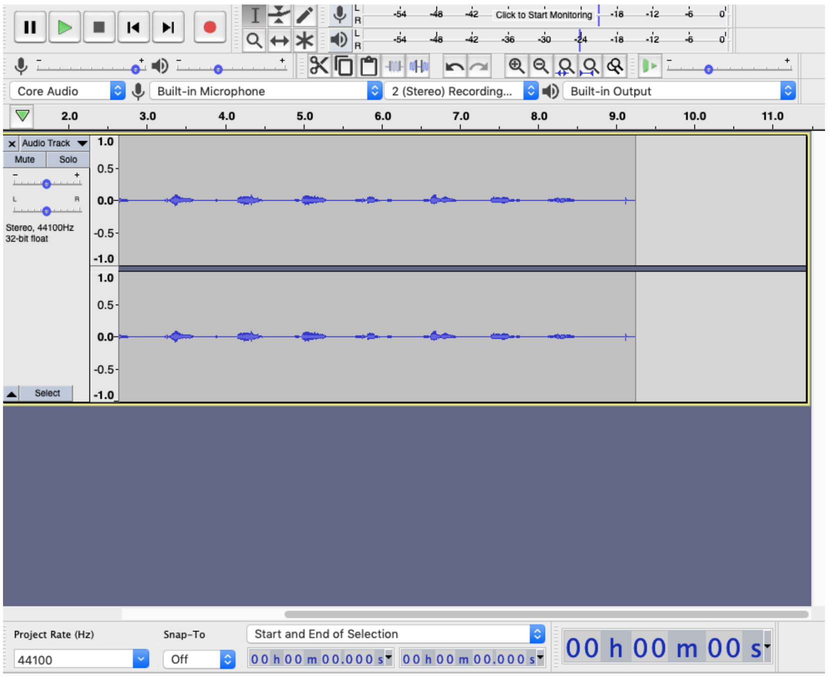
00 h 00 m 00.000 s 00 h 00 m 00.000 s

00 h 00 m 00 s

Do not change the default settings. Simply proceed as follows for every recording:

- Press the record button.
- Start narrating your story.
- Press the stop button, when story is finished (give 2-3 sec extra).

That's it, you've recorded yourself. After recording, you will see something like this:



Project Rate (Hz): 44100

Snap-To: Off

Start and End of Selection: 00 h 00 m 00.000 s 00 h 00 m 00.000 s

00 h 00 m 00 s

The next step will be to export your audio as a WAV audio file. Proceed as follows:

- File -> Export -> Export as WAV.
- Keep default encoding (Signed 16-bit PCM) and save. Name your file as described in Section 6.

Remember that **the audio file size should never exceed 50MB**. If the duration of your story does not exceed 4 minutes, the audio file will meet this restriction.

6. Naming the audio files and labelling

We expect 6 different stories:

- 3 instances of true stories
- 3 instances of deceptive stories

The names of each file should be:

- story_1_[Language]_[Truth/Lie]
- story_2_[Language]_[Truth/Lie]

  ..

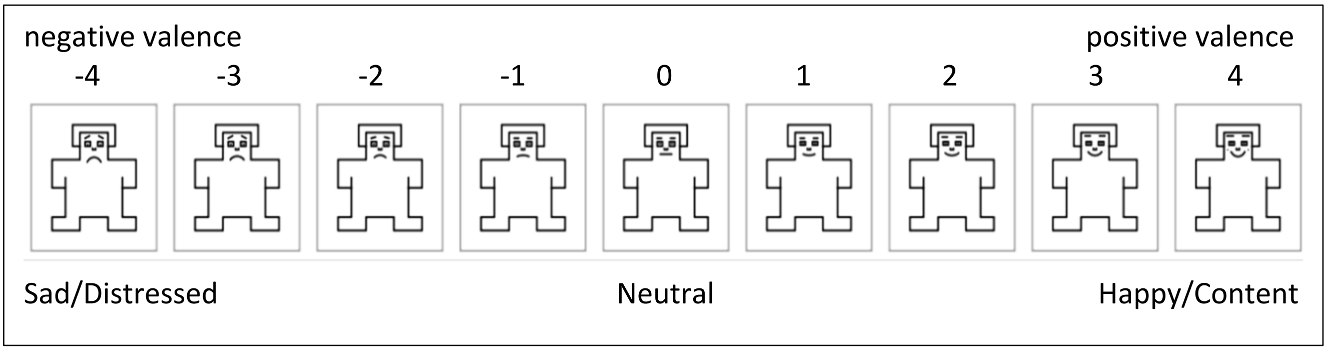
- story_6_[Language]_[Truth/Lie]

If your native language is **NOT** English	If your native language is English
story_1_English_Truth.wav story_2_English_Truth.wav story_3_English_Lie.wav story_4_English_Lie.wav story_5_Spanish_Truth.wav story_6_Spanish_Lie.wav	story_1_English_Truth.wav story_2_English_Truth.wav story_3_English_Truth.wav story_4_English_Lie.wav story_5_English_Lie.wav story_6_English_Lie.wav

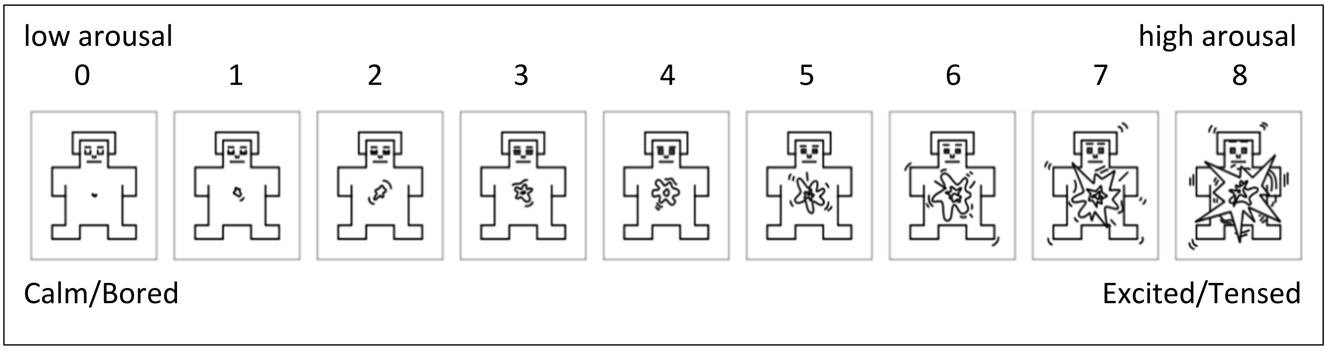
At the end, you will have produced **6 stories, 3 truthful** and **3 deceptive**.

We would also like you to label each story to describe its emotion content. We will be using the so-called Valence and Arousal Rating system.

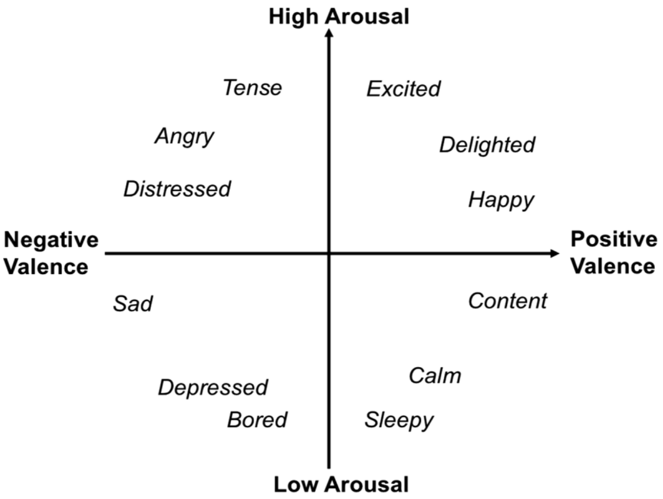
Valence: Valence is term used to describe **how positive or negative an emotion is**. We will use the rating as described in the following figure:



Arousal: Arousal is the term that describes **level of energy or intensity of an emotion**. We will use the rating as described in the following figure:



The Valence and Arousal values can be used to quantify our emotions, as shown in the figure below:



For example, a happy story would have a valence of 4 and arousal of 5, whereas a sad story would have a valence of -4 and arousal as 3.

7. Submitting your samples

We expect **6 separate submissions**, **one per story**. You will find the submission link on QM+, where you will be able to upload each audio file and label it.

Please make sure you submit **all the files and labels**, and strictly **follow the instructions** in this document to receive the full 4 marks.

8. Demographic questionnaire

There is a small demographic questionnaire that we would like you to complete and will allow us to potentially enrich the dataset. It includes questions such as your age, sex, and the approximate coordinates of the place (town, region, etc) where you feel most identified with. If you have any concerns about this questionnaire, please let us know.

9. Final dataset

The **MLEnd Deception Dataset** will consist of a collection of items with the following attributes:

- **Short audio file**.
- **Story**: categorical attribute, 2 values: Truthful/Lie
- **Valence**: continues attribute (*-4 to 4*)
- **Arousal**: continues attribute (*0 to 8*)
- **Language**: categorical attribute, English or other
- **Storyteller Anonymised ID**. The anonymised ID is simply number that allows us to distinguish one storyteller from another. It is NOT your QM ID.
- **Demographic data** about the storyteller.

We would like to make the dataset public (e.g. GitHub, Kaggle). No other data will be shared. If you have any concern, please let us know.

10. Dos and Don'ts

Please follow the instructions properly to record the story.

Here is a short summary:

- Your story must be about "visiting a place or a person"
- Your story should start with the following sentence in English: *"My name is ___, my QM ID __ , and this is my story"*.
- Recordings should be:
  - at least 2 minutes long, no longer than 4 minutes.
  - recorded using Audacitiy using its default settings in '.wav' format.
  - recorded in silence with no background noise (avoid recording it in public spaces).
- Do not worry about producing a perfect narration. Do not try multiple times.
- Do not listened to your recordings: no one likes their own voice.
- Use a partner as a listener: they will reassure you and save you time.
- Deceptive stories should be convincing, avoid obvious lies.
- Use appropriate language and context.
- Properly label each recording with valence and arousal.