# Participants instructions for the empirical evaluation

In this experiment we are going to evaluate the usability and feasibility of DescribeML and the domain-specific language which is based on. To do so, the only previous work we expect from participants, as you, is to have a look at this short video (It's 4 minutes) presenting the tool.

If you have already seen the video, then you can proceed with the experiments.

Installation of the tool:

To install the tool, you need an instance of Visual Studio Code in your PC.

Then go to the extension tab, and type: DescribeML.
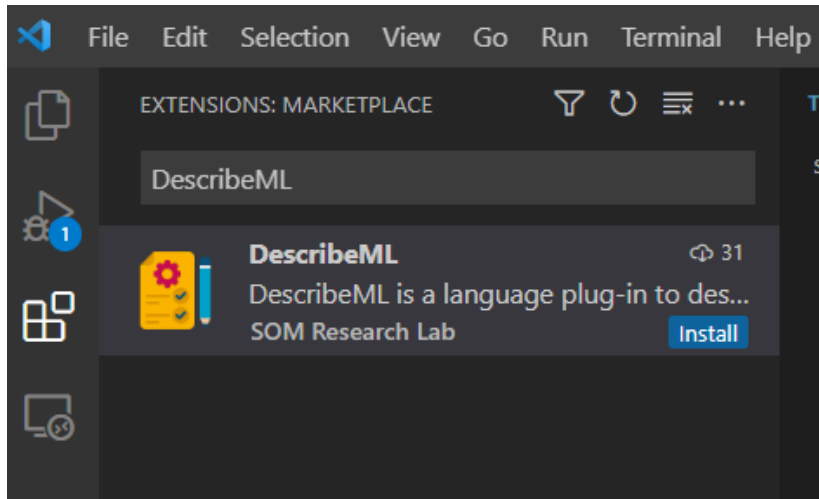
Click install, and that's it!

Exercise 1: Reading a dataset description

For this exercise open the file **Melanoma.descml** sent together with these instructions with your Visual Studio Code. You will see the description shown below (*With syntax colors*).


```
Dataset: Melanoma_Classification_Dataset
    Metadata:
        Title: "2020 SIIM-ISIC Melanoma Classification Challenge Dataset"
        Unique-identifier: SIC_Melanoma_Classification_Challenge_Dataset
        Version: v0001
        Dates:
            Release Date: 10-08-20
        Citation:
            Title: "A patient-centric dataset of images and metadata for identifying melanomas using clinical context"
            Authors: "Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M.,
            Year: 2021
            Journal: "Sci Data 8, 34"
            DOI: 10.1038/s41597-021-00815-z
        Description:
            Purposes: "The 2020 SIIM-ISIC Melanoma Classification challenge dataset..."
            Tasks:    [classification, image-classification]
            Gaps:     "As the leading healthcare organization for informatics in medical imaging..."
            Areas:    HealthCare
            Tags:     Images Melanoma
        Applications:
            Past Uses: "Yes. The 2020 SIIM-ISIC Melanoma Classification challenge dataset
            was created for the purpose of conducting a machine learning competition to
            identify melanoma in lesion images. The competition was launched on DATE on
            the Kaggle platform and ran through DATE. You can learn more about the competition
            and results here: LINK"
            Recommended:
                "Identify melanoma in lesion images"
                "Predict incidence of melanoma in a population"
            Non-recommended: "Due to low population prevalence and challenges with access
            to care in different populations, the images gathered for large datasets such
            as this for AI classification have a strong tendency to under-represent darker
```

This is a full description of the ISIC Melanoma dataset used in famous Kaggle competitions to detect melanoma over the skin of a patient.

The goal of this experiment is to evaluate the readability of the DSL by the final dataset users and by data science engineers. To do so, you are asked to answer the questions **only of pages 1 and 2** of the following form about the dataset by finding the relevant information in the .descml file.

Please, do not answer questions of page 3, until the next exercise is finished

Exercise 2: Building a dataset description

> **Record your screen:** *For this exercise,* please record your screen during the experiments
>
> To record your screen, open the VSCode studio;
>
> *For Windows:*
>
> - type WINDOWS + G to open the XBOX record app.
> - Then click at the record button.
> - When you finish the exercise, type WINDOWS + G and click at the stop button.
> - The video will be in the VIDEOS default folder of windows.
>
> Use the Screenshot toolbar
>
> *For MACOS:*

Once you answered the question of section 2 of the form, you can start the next exercise. In this exercise you need to describe a simple dataset published in Kaggle. You can access to the dataset repository following this link. Looking at the dataset website, the goal is to use the tool to express the following points:

- Provide the release date of the dataset
- Which license have the dataset
- Describe the attribute "genre"
- Provide the "Mode" of the attribute "genre"
- Provide a description of the gathering process

To do so, **open the file videogames.descML.** From there you can start filling the different terms. Type CTRL + Space to get autocompletion help from the IDE.

The time limit to do the exercise is 30 minutes. Once you finish the exercise, please stop the recording video (WINDOWS + G, or SHIFT + COMMAND + 5) and send the video to jginermi@uoc.edu.

Please fill the question of the **last section of the** form to finalize the experiment.

If you have doubts you can take a look to the Melanoma dataset (presented before), and the Reference Language Guide that contains some examples.