# Assignment 2: validation theory

Mark van der Loo and Edwin de Jonge

useR!2021

You do not need to use R in this excersise. It is aimed at thinking in a systematic way about data processing and data validation.

## Exercise 1: the online job vacancy case

You work in a project aimed at estimating which type of skills are required on the job market. To this end, a large collection of job vacancy descriptions are obtained by webscraping.

From the scraped data, the job title, function description and job requirements sections are extracted and stored in a file. The date of scraping and original url are stored too. All text is converted to `UTF-8` where necessary.

Next, the function description sections are searched for skills such as 'programming', 'presenting' , and so on. The list of skills is obtained from ESCO in `UTF-8` csv format. This yields a data set where each row corresponds to a single vacancy, and each column is a skill. The entries are 0 when the skill is not mentioned, and 1 if it is.

Using a pre-trained machine learning model, the job vacancies are classified into different categories such as IT, health care, education, and so on. Finally, summary statistics over the classes are computed and reported in with nice-looking tables and visualisations in a dashboard.

Design and describe the input data, the output data and one or two important data stages in between. For each stage write down some quality demands, include things like

- Technical format (HTML, csv, in-database, other?)
- Technical quality demands (data type, encoding, ...)
- Content checks (variable ranges, code lists (= allowed categories), comparison with other results, ...)

For this exercise it is better to get a rough sketch of the whole process than to have only one step with many details.

## Exercise 2

Determine the complexity level of the following data validation checks.

a. Checking that someone under the age of 15 is not employed.
b. There are not more than 20% missing in a column
c. The answer to Question 7 in a questionnaire is in {`"Yes"`, `"No"`}
d. The ratio between variables `Height` and `Weight` of an individual record does not exceed twice the median ratio accross the whole data set.