# Information Visualization lab

## First practical work

### *Introduction*

Imagine you are rector of UPC and wonder how the faculty is faring in terms of research. Your university has a lot of departments, researchers, and research groups. And you have no idea how many articles are produced every year, just a rough approximation. It would be interesting to see if the COVID-19 affected the production, for good or bad…

Although the UPC has a tool called Futur (https://futur.upc.edu) where anybody can search for a researcher or a department, the trove of information that it is contained by the repository might answer some of the questions above. Therefore, the goal of this project is to create a set of visualizations using Python and altair that answers the following questions:

- How have the different departments performed along the last years?
- How have the different research groups performed along the last years?
- Is there any correlation between department size and production?
- Do any departments produce more papers per person?
- How big is each area of research (the journal area of research in the dataset) in UPC?
- Is there an intersection between departments and areas of research?
- How is the publications growth of each research group compared to the average at UPC?
- How departments (in size) compare according to the number of authors that have had publications in the period of the dataset?

Questions in blue are only intended for groups of 3.

You may add extra questions (such as the ones related to the COVID-19). Some of the questions can be answered directly from the data. Some others require data derivation. Most of the data processing and derivation can be carried out using Open Refine. Some derivations/calculations can be done interactively, but precalculation is typically always better.

The visualization must be a multi-view visualization. Think carefully the design. Test and redesign.

### *Data*

You will receive a file in text format that has 137695 entries. Each entry has the following structure:

*Publication date; id_article; title article; ISSN journal; name of the journal; research area of the journal; id_author; surname1; surname2; name; department name; research group*

Some (or most of the) articles have multiple authors. Thus, several entries will appear. But article identifiers are unique. Authors also have a unique id that may be useful. Note that the format is not very common.

### *Data confidentiality*

The data we provide for this project can exclusively be used for the project of this course. It cannot be shared, published, or distributed in any way. It also cannot be used for any other purpose. Please be careful and

respectful when handling this dataset. By downloading the dataset, you agree that you agree to these terms. The data can be downloaded from this link: https://mydisk.cs.upc.edu/s/9LoxgrLtPK4GmTz

## *Data processing*

You can process the data using either Open Refine, or another tool. You can also process the data programmatically. You need to deliver the clean dataset.

Independently of the cleaning tool and process, you must describe your cleaning steps in your Google Colab document. If these are using pandas, for example, include the code in the document. We must be able to reproduce the steps and go from the raw data to the clean version.

## *Design and implementation*

For the visualization, we need you to describe the design process also in the Google Colab document. This means that you may include all the steps that led you to the final visualization. You can remove (or group) some steps in the final document if you think it is better. But we need to see the design process, we want to understand how did you reach to the final visualization.

Before you start coding anything, you need to think on what visualizations will be provided. Note that the user needs to be able to answer the questions above with a single visualization, that will include multiple views.

Consider all sorts of charts that might be useful: line charts, bar charts, heat maps, treemaps… Some views will contain several variables, so use visual cues, proper palettes to ensure they are understood properly.

## *Delivery instructions*

The work can be implemented in pairs or individually. You have to provide the clean data. You have to describe the cleaning procedure, so that we can generate the clean data from the raw data following your steps. This description must go in the Colab document.

You must include a step-by-step description on **how to solve tasks**. These can go in the Collab document. For example, one might have:

- Question 1: How have the different departments performed along the last years?
- Answer to Q1 could be: "In chart C1 you can see a line chart with the department's publications along the last years. To avoid clutter we did this and that… And by checking *something that is clear enough*, we can see that the behavior is…"

The delivery must consist on a single ZIP file with a name that includes the authors, that contains the datasets (raw and clean), the Colab file(s) (*ipnyb*) and optional extra documents if required. The Colab file must be named after the names of the authors. Treat the Colab document as a report, include titles, boldfaces, etc., to make it easier to read.

The deadline for the delivery of this lab project is the 18$^{th}$ of November.

## *Important remarks*

The final grade will take into account the number of variables included in the visualizations (these may include new calculated variables, such as averages, maxima, minima, etc.)(e.g., number of accidents, victim age, gender, weather condition…). Additionally, we will value the number of non-trivial tasks (adequately described in the documentation) that can be properly solved with your visualization tool. In this sense, adding

other data sources if suitable (e.g., you could add information regarding when the COVID-19 period started, or the number of people in a department…). You can get interesting data in Futur, if needed.

Don't leave the project for the last day or do the minimum amount of work. In case of doubt, ask us whether the current work is enough or needs more effort.