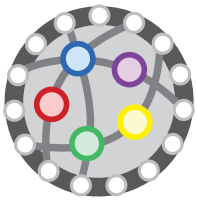

Before Starting

Get acquainted with the Command Line Interface. Search for a tutorial on command line interfaces in the OS you'll be using.

Installing Software

We will use R and Python in this course.

1. Install R from https://cran.rstudio.com/ (or see the package manager for your OS and distribution) and RStudio from https://www.rstudio.com/.
2. Install Anaconda from https://www.continuum.io/downloads and follow the instructions on https://docs.continuum.io/anaconda/r_language on how to install R packages in Anaconda.
3. Follow the instructions and links on https://docs.docker.com/engine/installation/ to install Docker. Run Docker and install the Jupyter Notebook Data Science Stack from https://github.com/jupyter/docker-stacks/tree/master/datascience-notebook .

Remember to update the conda packages after installing Anaconda (and whenever necessary):

```
conda update -all
```

Different versions of RStudio and Anaconda **may cause library compatibility issues**, make sure you have the latest versions and no older libraries installed.

Creating GitHub account and repository

1. Create a GitHub account at https://github.com/ .
2. Read the Hello World guide: https://guides.github.com/activities/hello-world/ .
3. Create a public repository with a README. Use a name related to this course!
4. Learn MarkDown, e.g. at https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet .
5. Edit that README to tell us about your project.
6. Send the link to that repository to the instructors (rafael.santos@inpe.br ***and*** gilberto.queiroz@inpe.br).

Do I really need to code?

Let's try this basic example. Get the file `defpatterns.missing.csv` from the course website. You can open it in any spreadsheet software.

The file contains a dataset with 1472 observations of 18 variables, with some position/ID variables, some metrics and classes for deforestation patterns. The name for each variable is in the first line of the file, and metrics' names start with c_. Variable "padrao" describes the class for the observation.

Some of the data is missing, though. Try to figure out how to answer the following questions with a spreadsheet program:

1. How many different values there are for the column "padrao"?
2. What is the minimum value for a given column, e.g. "c_PSMetric"?

3. What is the minimum non-zero value for a given column, e.g. "c_PSMetric"?
4. How many values are missing from each column?
5. How many observations are complete, i.e. there is no missing value for its variables?
6. How many complete observations there are for each different value of "padrao"?

It is possible to do these exercises with spreadsheets or other GUI-based analysis tool, but that would require a lot of work, and if you need to process different datasets with similar measurements you would need to redo most of the work. We will see that this is relatively simple to do in R, with the advantage of easy reproducibility.

To-Do: Define your project!

This is very important. Your grade on this course will be given based on the development and implementation of a project. Ideally this project should have something to do with your MsC or PhD research – you can use what you'll learn in this course to advance your specific research. Ask your advisor about a dataset related to your research or an approach that may be interesting to explore.

At a first moment you will need to be able to answer the following questions:

1. What is the science problem I need to solve?
2. What is the data that is available for this problem? Is it simple to get hold of this data? What is the volume of the data? Where is it? What is its format? (Hint: most of these questions were listed on the first lecture, and then some).
3. What do I think I will achieve by analyzing this data? What are the expected results?

If you can't think about a project, contact the instructors.