Ontologies

In-class practicum

1. Using GitHub for code and project managements
   1. Create a GitHub account
   2. Create a new repository and choose Apache 2.0 license
   3. Upload the random\_normal.py code and resulting output on GitHub. There are several options to do this:
      1. Web Interface – we will be using this one
      2. GitHub Desktop
      3. Learning git
   4. Create a ReadMe describing the code.
2. Reporting on your data science project

Problem statement: The El Niño Southern Oscillation is a climate phenomenon characterized by shifting wind patterns and sea surface temperatures in the tropical Pacific. It is the leading source of interannual climate variability on Earth. In this project, we are interested in exploring whether changes in sea surface temperature in Nino3.4 region (taken as the average anomaly over 5N-5S, 170W-120W, referred to as the Nino3.4 index) causes changes in precipitation, as recorded by the Standard Precipitation Index, in other parts of the world. We will be working with the following data: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview> . Since the data is observational, we will be using [Granger Causality](https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.grangercausalitytests.html) to establish causality. Time series X “Granger-causes” time series Y if past values of X can be used to predict future values of above and beyond the information contained in past values of Y. Granger causality can be posed as an hypothesis test:

Null hypothesis: X does NOT Granger causes Y

Alternate Hypothesis: X Granger causes Y

Pre-processing steps:

* Sea surface temperature data: Calculate Nino 3.4
* Precipitation data: Calculate PDSI (which can be done for each grid cell separately).

Analysis:

* Granger causality, X : sea surface temperature, Y: PDSI

1. Sketch the workflow for the analysis, including the pre-processing steps.
2. Which steps can be run in parallel? Explain why you can/cannot do so.
3. Using PROV, describe the provenance of your record.
   1. What entities are involved?
   2. What activities are involved?
   3. Which agents are involved?
   4. How do they relate?
4. Describe in details how would you share:
   1. The data
   2. The software
   3. The workflow