# Week 2 (25.10 - 31.10.2019)

## Task 1

Write a python script which takes a list of PubChem CIDs and returns three files: a list of structures in SMILES format, a list in SDF format, and a table (CSV or TSV) containing properties of the compounds. You can use the library pubchempy and you might find rdkit useful as well for converting between file formats.

- Went through some basic API's from the pubchempy library to know more about it.
- There are API's to convert return files in SDF and CSV formats given the CIDS in the pubchempy library called download().
- To get the file in SMILES format, the MolToSmiles() format converter from the rdkit library can be used.
- This task can be found in the getData.py script file. Please provide the list of cids as cmd arguments.

## Task 2

Perform an analysis of the attached data using Galaxy.

- We use the rdkit tool on Galaxy to generate molecular descriptors from the given dataset.
- The descriptors will act as the features on which the model will be trained.
- Some preprocessing is done on the molecular-descriptors.tabular file to select required columns and separate the labels from the feature columns.
- Finally, the data is split into train and test sets (80-20 ratio) and we use the ensemble tools on Galaxy to create a model (I have used a Random Forest Classifier).
- The preprocessing of the data is done in the preprocessing.py script.