**HPI Course Overview**

**Task:**

* Look up HPI courses and their course related data based on different parameters (e.g. ECTS Points, teaching form, etc.)
* Combine, visualize and integrate these into a “Database” and enable batch-downloading of course materials (e.g. Slides)

**Dataset:**

* The Dataset is scraped from the **HPI websites**
* It contains:
  + Course name
  + Teaching form
  + Enrolment deadline and type
  + Course language
  + Maximum participants
  + Weekly hours
  + ETCS and creditability
  + Grading
  + Course website of the chair
  + Pdf files linked on course site

**Method:**

* Data preparation tools:
  + **Python** to crawl the HPI websites
  + Combination of **XPATH** and **Regex** to extract useful information from different sites into a **Pandas Dataframe**
  + see Script: HPI\_course\_Overview\_Web\_Crawlin.py
  + result in coursesExort.csv
  + **Googles OpenRefine** to improve data quality and clear up some inconsistencies
  + result in coursesExort-refined.csv