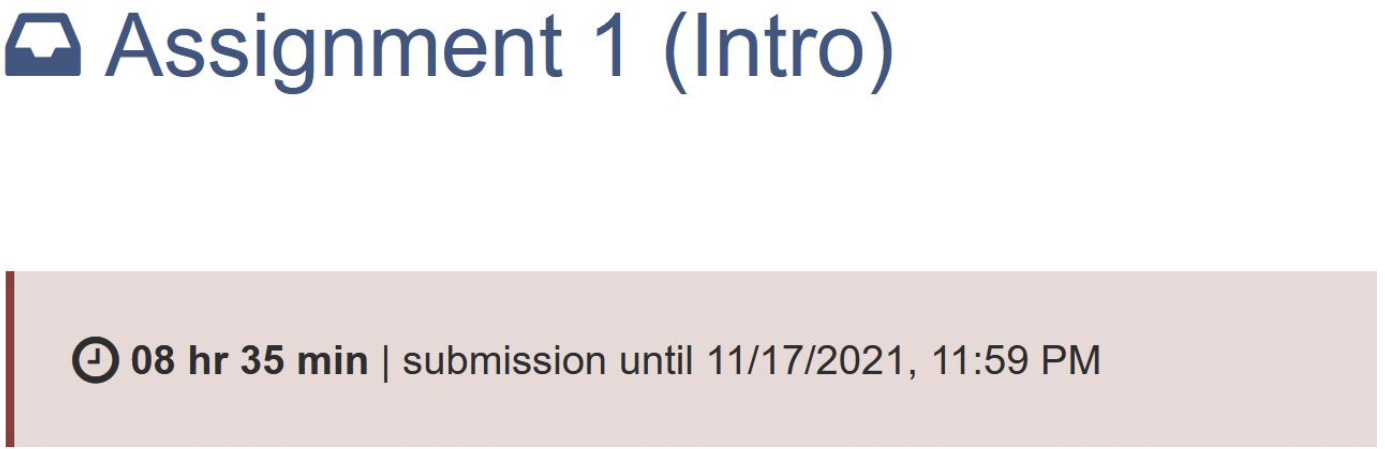
Q&A

Data Collection
Prof. Dr. Ralf Lämmel & **Johannes Härtel**
(johanneshaertel@uni-koblenz.de)

Deadline

- We have a small deadline extension for assignment 1.



Reference Solution

- We will try to implement an additional feedback loop, adding interesting solutions provided by you; or common misconceptions into the presentation of the reference solution.
- Hence, the reference solution needs to be provided with a **short delay (0-3 days)** to the submission deadline of an assignment.

Do the Research Questions in Assignment 1 and 2 need to be the related?

- No!

Are there any restrictions on APIs or Wrappers I can use in Assignment 2?

- No!

Some lessens (I) learned (over the past years) on data collection

- If you crawl heavy, you may need to make the program sleep in between requests (regardless of rate limits).
- In some cases, request are only handled if pretending to be a browser. Adding something like the following to the request header works: `'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_5) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/50.0.2661.102 Safari/537.36'`
- Using (OO) **wrappers** around original APIs is often just an annoying indirection. Maintainers of (OO) wrappers are people too (thery code might be a mess)!

Some lessens (I) learned (over the past years) on data collection

- Often the basic solution is the best (request + text, JSON or XML processing).
- Be fault resistant. Dump your data directly and no maintain in any data structure.
- Do **NOT OPTIMIZE**, unless rough estimates on crawling times (computed by you manually) forces you to.
- Know the bottlenecks; threading might not be the problem (but frequently “threading” turns out to be a problem).