Introduction & overview

17.03.2020, Data Science (SpSe 2022): T1-1

Prof. Dr. Claudius Gräbner-Radkowitsch
Europa-University Flensburg, Department of Pluralist Economics
www.claudius-graebner.com | @ClaudiusGraebner | claudius@claudius-graebner.com

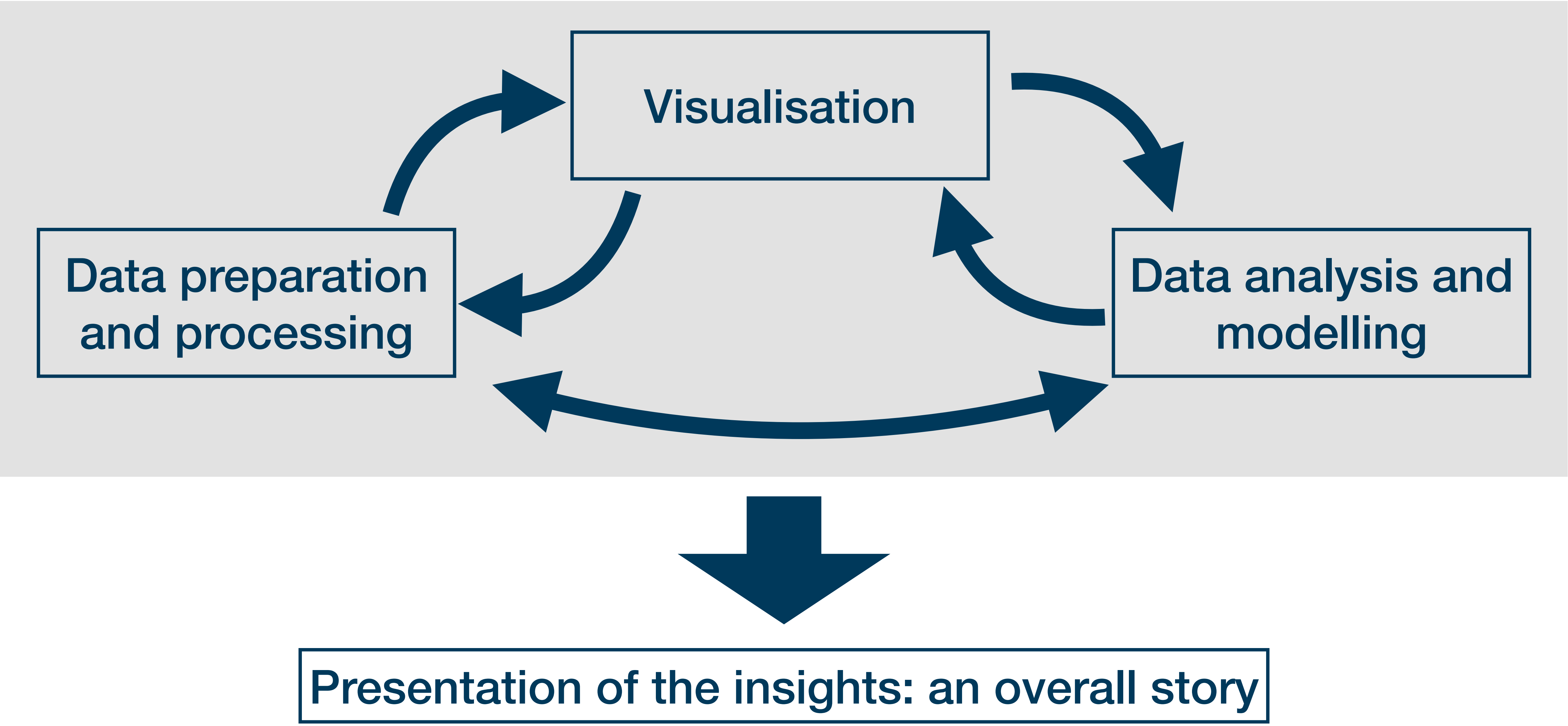
Europa-Universität
Flensburg

Europa-Universität
Flensburg
International Institute of Management
and Economic Education
Department of Pluralist Economics

Part I: Organization & outlook

Goal of the course

- In this course you will learn how to prepare, analyse, and present quantitative data using the software R → four key areas



Visualisation

Data preparation and processing

Data analysis and modelling

Presentation of the insights: an overall story

Why R?

- R allows you to conduct all steps of this data science pipeline within one consistent framework in a transparent and reproducible manner
- R is free, OS-independent and open source → inclusive, transparent, and vibrant tool
- For statistical analysis, R is among the most widely used and demanded programming languages
- R is demanded in almost every industry
- Learning R makes it easier to learn other widely used programming languages
- There is a great and friendly R Community

> “The days of commercial statistical languages and packages such as SAS, Stata and SPSS are over”
>
> Paul Jansen, CEO of Tiobe Software

#	RedMonk	TIOBE	PYPL
1	JavaScript	Python	Python
2	Python	C	Java
3	Java	Java	JavaScript
4	PHP	C++	C/C++
5	C#	C#	C#
6	C++	Visual Basic	PHP
7	CSS	JavaScript	**R**
8	TypeScript	PHP	Objective C
9	Ruby	Assembly	Swift
10	C	SQL	TypeScript
11	Swift	Go	Matlab
12	**R**	Swift	Kotlin
13	Objective C	**R**	Go
14	Shell	Matlab	Ruby
15	Scala	Delphi	VBA

What you will be able to do

- Read in data sets from various sources
- Prepare ‘messy’ data and produce ‘tidy’ data
- Create illustrative visualisations on a publication-ready level

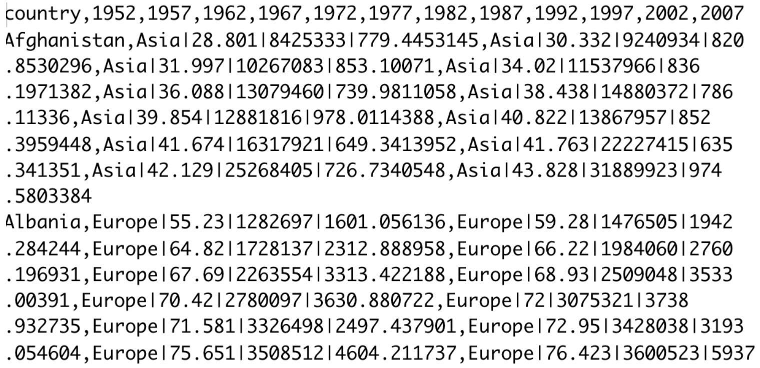
EUROPEAN COMMISSION

OECD

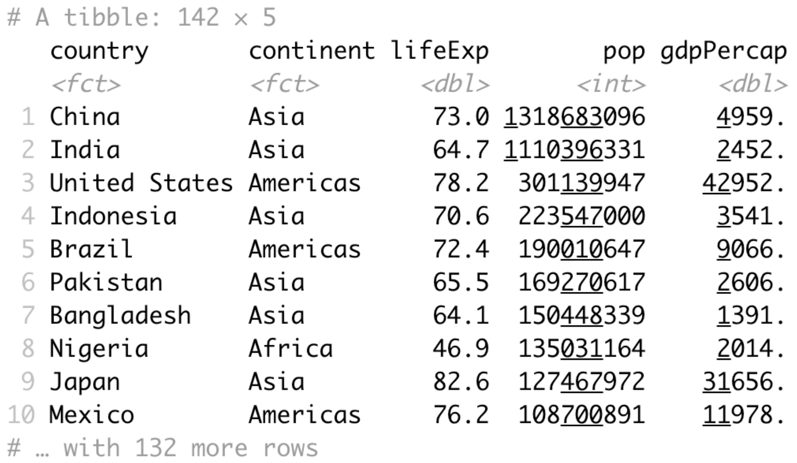
THE WORLD BANK

INTERNATIONAL MONETARY FUND

DESTATIS
Statistisches Bundesamt

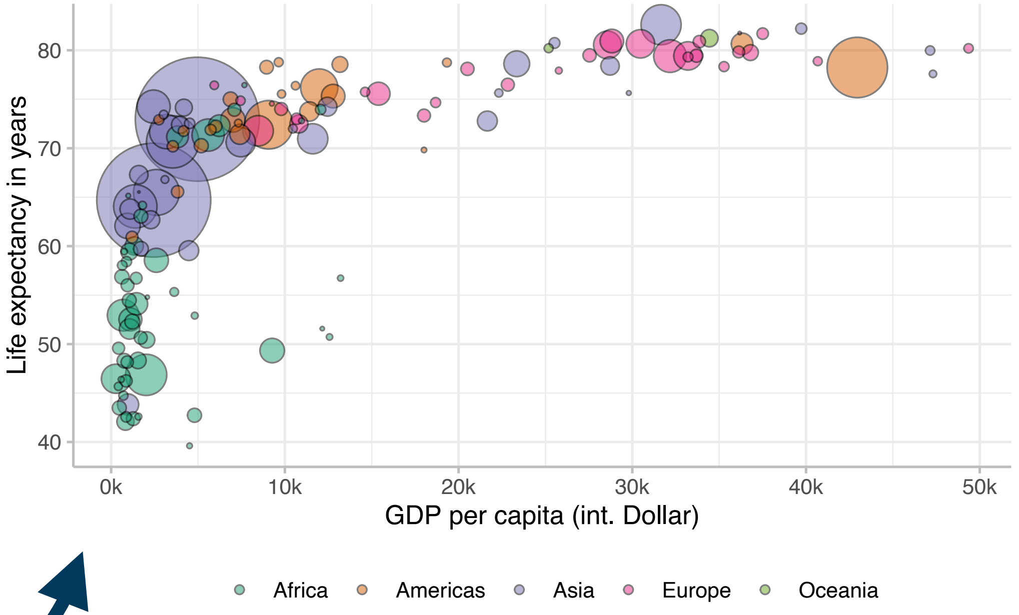


```
country,1952,1957,1962,1967,1972,1977,1982,1987,1992,1997,2002,2007
Afghanistan,Asia|28.801|8425333|779.4453145,Asia|30.332|9240934|820
.8530296,Asia|31.997|10267083|853.10071,Asia|34.02|11537966|836
.1971382,Asia|36.088|13079460|739.9811058,Asia|38.438|14880372|786
.11336,Asia|39.854|12881816|978.0114388,Asia|40.822|13867957|852
.3959448,Asia|41.674|16317921|649.3413952,Asia|41.763|22227415|635
.341351,Asia|42.129|25268405|726.7340548,Asia|43.828|31889923|974
.5803384
Albania,Europe|55.23|1282697|1601.056136,Europe|59.28|1476505|1942
.284244,Europe|64.82|1728137|2312.888958,Europe|66.22|1984060|2760
.196931,Europe|67.69|2263554|3313.422188,Europe|68.93|2509048|3533
.00391,Europe|70.42|2780097|3630.880722,Europe|72|3075321|3738
.932735,Europe|71.581|3326498|2497.437901,Europe|72.95|3428038|3193
.054604,Europe|75.651|3508512|4604.211737,Europe|76.423|3600523|5937
```



```
# A tibble: 142 × 5
   country       continent lifeExp        pop gdpPercap
   <fct>         <fct>       <dbl>      <int>     <dbl>
 1 China         Asia         73.0 1318683096     4959.
 2 India         Asia         64.7 1110396331     2452.
 3 United States Americas     78.2  301139947    42952.
 4 Indonesia     Asia         70.6  223547000     3541.
 5 Brazil        Americas     72.4  190010647     9066.
 6 Pakistan      Asia         65.5  169270617     2606.
 7 Bangladesh    Asia         64.1  150448339     1391.
 8 Nigeria       Africa       46.9  135031164     2014.
 9 Japan         Asia         82.6  127467972    31656.
10 Mexico        Americas     76.2  108700891    11978.
# … with 132 more rows
```

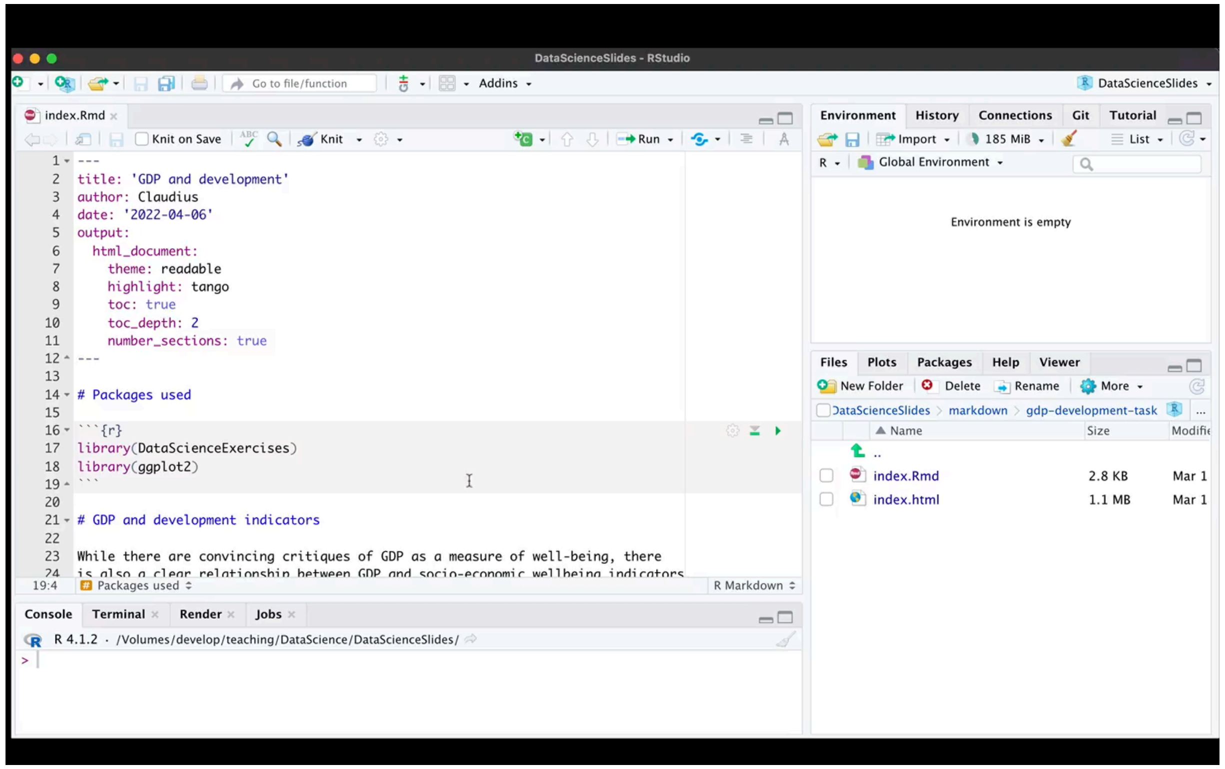
Life expectancy and income per capita



Note: size of bubbles represents population. Data: Gapminder

What you will be able to do

- Identify hidden patterns in data and make predictions using a variety of modelling techniques
- Write reproducible research reports in Markdown
- Publish visually appealing reports on the web via Netlify
- Reflect upon the potentials and limits of quantitative data analysis

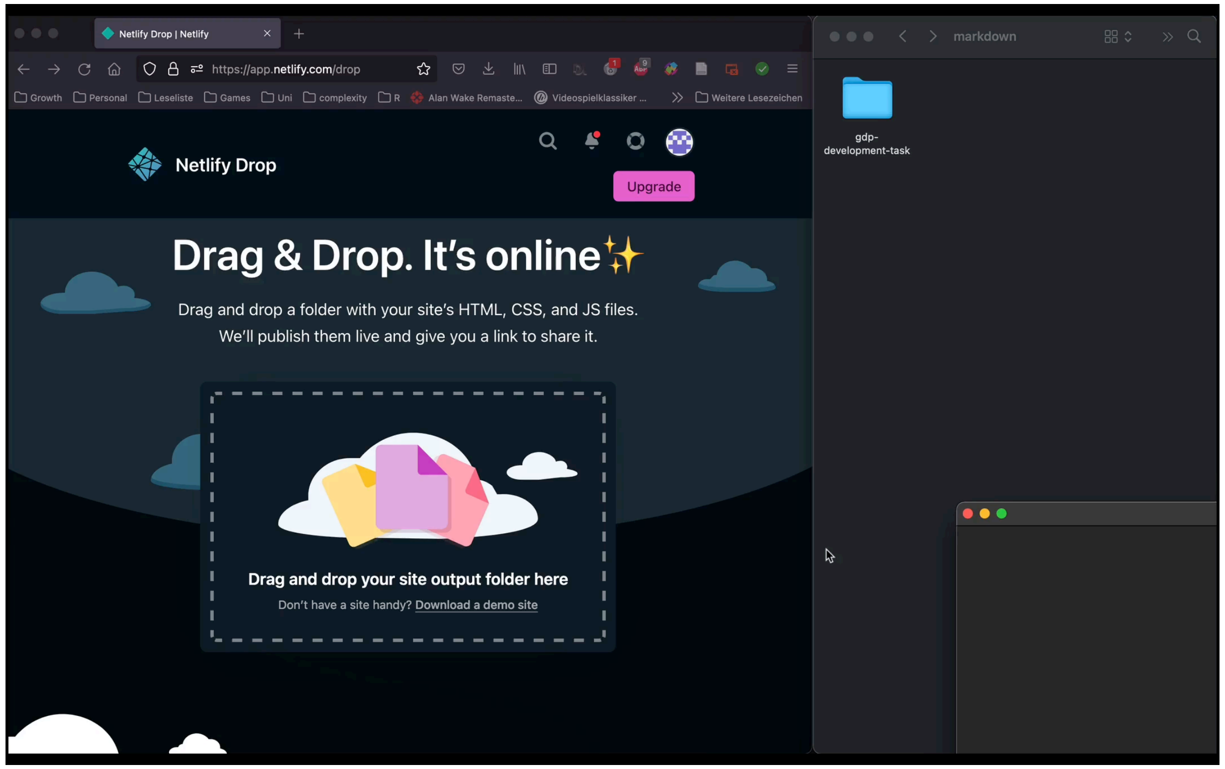


~~~
---
title: 'GDP and development'
author: Claudius
date: '2022-04-06'
output:
  html_document:
    theme: readable
    highlight: tango
    toc: true
    toc_depth: 2
    number_sections: true
---

# Packages used

```{r}
library(DataScienceExercises)
library(ggplot2)
```

# GDP and development indicators

While there are convincing critiques of GDP as a measure of well-being, there
is also a clear relationship between GDP and socio-economic wellbeing indicators
~~~

```
R 4.1.2 · /Volumes/develop/teaching/DataScience/DataScienceSlides/
>
```

Environment is empty

Name	Size	Modifie
..		
index.Rmd	2.8 KB	Mar 1
index.html	1.1 MB	Mar 1



Drag & Drop. It's online ✨

Drag and drop a folder with your site's HTML, CSS, and JS files. We'll publish them live and give you a link to share it.

Drag and drop your site output folder here

Don't have a site handy? Download a demo site

The road to our goal

- This is the first time I am teaching this particular course at the EUF → our outline is tentative and subject to change
- We will regularly consult three open source and free textbooks
- I will provide you with practical exercises, which I recommend you to complete every week
  - Work together, find study groups
  - Use the Moodle forum for questions
  - Try to follow the course constantly

- Ask questions and **provide feedback**
  - There will be *very short* feedback forms for each session, the results will be presented at the beginning of the next week

Organization of the lectures

- There will be no strict separation of theoretical lectures and practical labs
- Each session comprises aspects of both → always bring your laptops👩🏼‍💻👨🏾‍💻
  - Several session will feature group work
- Questions – about the exercises or any other practical challenges – should always be posted online in the Moodle forum
  - Questions should most of all be answered by other students → solving each others' problems helps tremendously for understanding
  - The forum ensures that answers to questions are (i) recorded and (ii) available to everybody
  - Particularly intriguing questions can be discussed in the beginning of a session

Logistics

- There is one weekly and one bi-weekly on-site session
- The course material as such will be made available via a course webpage
  - Written in R → easier for me to maintain than via Moodle ;)
  - Makes material publicly available
- Discussion and announcements are organised via Moodle
  - Most important: the forum for our questions and the announcements
- For the dates of all sessions please consult the course outline
  - There will be changes during the semester!
- How the exams can take place will be decided in due course

Examination

- Upon successful completion, this course is worth 5 CP
    - Corresponds to 150 working hours, about 35 being lecture time
- Your overall grade comprises of…
    - A mid-term exam during the middle of the semester (50%)
    - A final exam at the end of the semester (50%)
- You will need to analyse artificial data sets, write reproducible reports, and answer content questions:
    - Includes data preparation, visualisation and analysis
    - Open book character is meant to mimic the practical application of the tools

Summary: our ‘learning agreement’

The goal

You learn to be confident in using R when turning raw data into a comprehensible story. This includes **importing**, **transforming**, **modelling**, and **visualising** data, and to **communicate** the overall results.

You will also learn to critically reflect on data scientific practices and products, produced by yourself and others.

What I offer

I provide **slides, example codes, tutorials, and exercises**, which are tailored to your learning needs. I will give my best to facilitate an **amicable working environment**, and answer questions in class and via Moodle.
I seek your **feedback** and implement it, when feasible.

What I expect

I expect you to **attend** classes regularly, to be **honest** about what you did not understand, to **support each other** through Moodle and in class, that you do the **homework** and **exercises** such that you keep up with the course, and that you make use of the **feedback** tools.

Summary: our ‘learning agreement’

- Why do I expect these activities from you?
  - Learning a programming language is a consecutive activity: you miss basics in the beginning → you’ll quickly become frustrated and get lost
  - This is a demanding course: catching up later on what you missed earlier will be difficult
  - Learning a programming language works mainly through practice and *doing* → practical exercises have a *huge* benefit
  - Learning a programming language is *difficult* and at times *frustrating* → we need an amicable environment and must support each other
  - Few things have a bigger learning effect than helping others with their problems

Learning a programming language can be a lot of fun and really brings you forward – if we do this together as a team💪🏽

Open questions?

Time to introduce ourselves…