Exercise 1: Information Coding & Data Structures

Exercise 1 for the lecture ‘Foundations of Data Science’

Prof. Dr. Karsten Donnay, Assistant: Philipp Kling

This session covers

- General data science process
- Introduction to git
- Introduction to our working case
- Data import in R

University of Zurich[UZH]

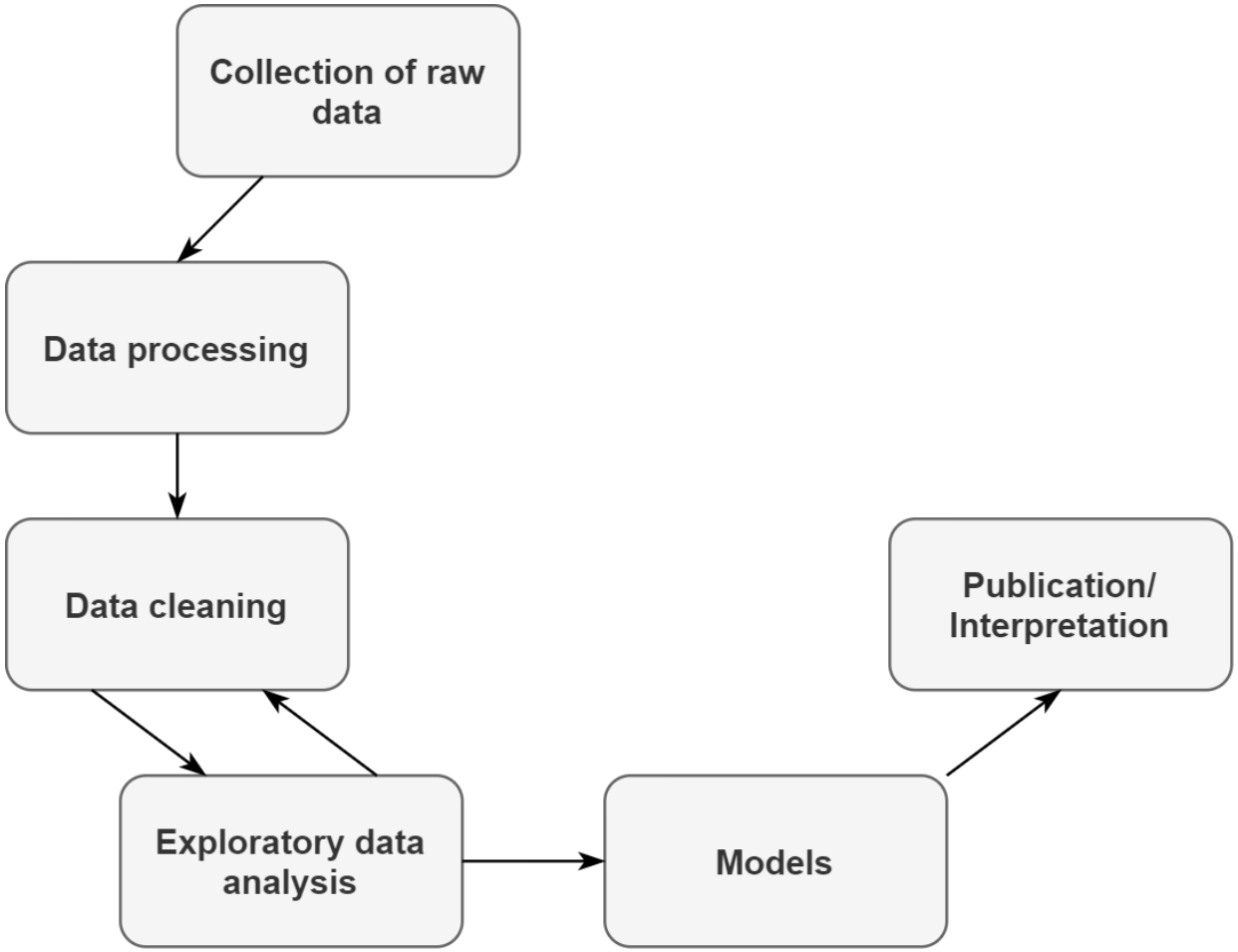
Department of Political Science

Data science process

Data science process

"Make sense of new and/or large data and communicate insight"

- Access innovative and **large data** resources.
- **Process data** to make it machine readable.
- Use statistical methods or machine learning to **detect structure** in the data.
- **Provide** meaningful **insights.**



Data science process

What is good science?

- **peer review**, **transparency** and **replicability** (Apart from other criteria)
  - Karl Popper (1934): “non-reproducible single occurrences are of no significance to science”.
  - Emphasizes the need for **publication** of employed **methods**, **documentation** of the **data collection** and cleaning process, and the **provision of datasets**.

Data science process

Why is reproducibility in data science difficult?

- Available **resources** (e.g. computing power, storage)
- Data on the Internet often **in flux** (e.g. websites change, Tweets get deleted…)
- **Permission** to use data (e.g. Facebook data)

- **Git** is one way to improve on one part of the reproducibility crisis: make method transparent and easily accessible.

University of Zurich[UZH]

Department of Political Science

Introduction to git

Git (Version control)

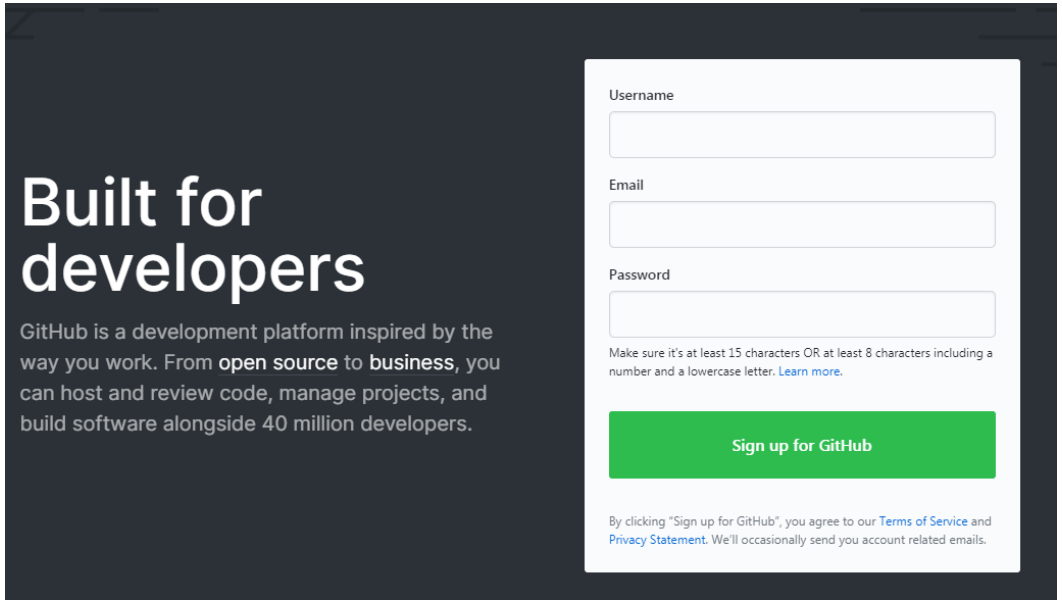
- The ‘Dropbox’ for programming
- Documents the different stages (versions) of files
- Makes it easy to track changes and restore previous versions.
- Enables controlled collaborations with others
- Eases the publication of work and increases transparency

Popular platform: Github (alternatively: bitbucket)

- Go to https://github.com/
- Select your role (student) and the purpose of usage and confirm the email.
- Create your first repository

  e.g. `github.com/philippklinguzh/datascience`

- Download and install git
  - Use notepad++ to edit (https://notepad-plus-plus.org/downloads)
  - Enable 3rd party software
  - OpenSSL library
  - Checkout Windows-style
  - Use MinTTY
  - Enable file system caching and enable Git credential manager



Version control with git

- Two possibilities:
  - download existing data from a repository
  - Initialize repository from data from your computer

- Let's add some data to your repository
  - Create a folder where you want to have your repository saved (`C:\Philipp\Documents\git`)
  - Open git bash/console
  - Navigate to the folder

    ```
    cd C:/Philipp/Documents/git
    ```

    (make sure to change the \ to /)

Version control with git

```
git clone https://github.com/username/name_of_repository.git
```

- Create a file in a folder of your choice (*test.txt*)
- Store your account information

```
git config --global user.email "YOUR GITHUB EMAIL"
git config --global user.name "YOUR GITHUB NAME"
```

- Check the status of all files in this repository

```
git status
```

- Add the file to the current list of files to be committed and check status again

```
git add test.txt
git status
```

Version control with git

- Commit your changes and add a message/description to the commit

```
git commit -m "Initial upload"
```

- Upload (“push”) your local changes to the repository

```
git push
```

- Inspect your changes by visiting your repository in the web browser
- Open your *test.txt* file and insert some text, then save it
- Check the status of all files in this repository again. You should see now that *test.txt* has changed.
- Repeat the previous steps

```
git add test.txt
git commit –m “Added some text to the test.txt”
git push
```

Version control with git

Version control with GitHub

– You can apply for educational discount (GitHub Pro for free):

https://education.github.com/

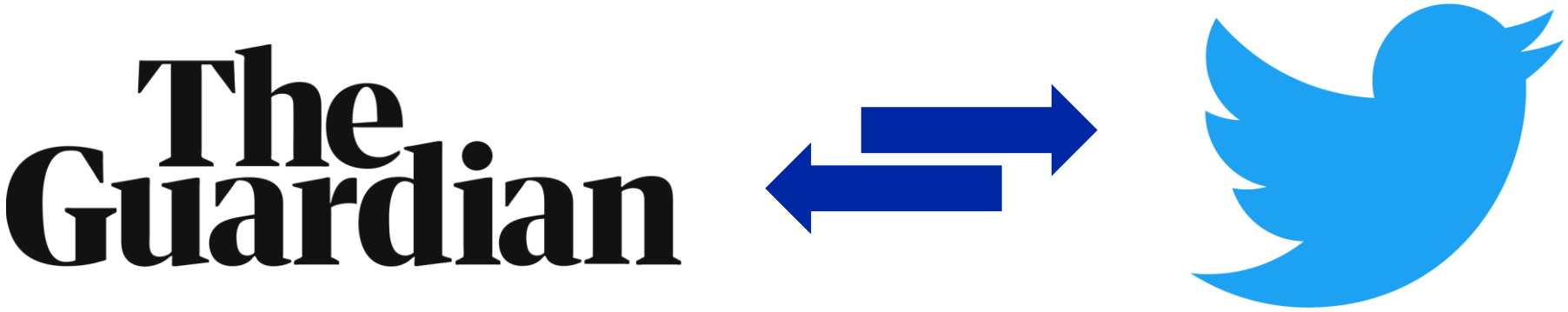
Introduction to our working case

Introduction to our working case

- While we talk about the case, you may already install the required packages for today's exercise.
  - `git clone https://github.com/css-zurich/fds-2020-exercise.git`
  - Open the Rmarkdown file "ex1/exercise1.Rmd" in Rstudio
  - Run the first lines of code

Introduction to our working case

- Hypothetical use case: How do characteristics of news articles relate to reactions on social media?
- Our two examples: The Guardian for news and Twitter for social media data



Introduction to our working case

- **Goal**: combine news data with social media data
- There are already **established packages** in **R** that retrieve data from these platforms. However, we will use these platforms to **build some applications from scratch** and demonstrate core concepts of data science using R.
- Keep in mind: before starting to build your own application, **do some research on existing work**. Often there are already established ways that work efficiently.
- After the five exercises you will be able to…
  - …manage and **process data efficiently**.
  - …**manipulate text** into formats that you can work with.
  - …**read data** from **websites** into R.
  - …retrieve data from application programming interfaces (**APIs**)

Introduction to our working case

Why use the **Internet** to collect data?

- Has a plethora of **useful data sources**:
  - Government publishes data (e.g. speeches, voting…)
  - Social media data to analyze human communication
  - News data for public discourse and attention to events
  - User interactions with e.g. products (Amazon reviews), Films (IMDB)…
- Why is this **relevant**?
  - Re-evaluation of existing research with new data
  - Enables entirely new research questions
  - Cost and time efficient
  - Theoretically easily reproducible

Introduction to our working case

Why use the **R**?

- **Free** and open source
- Excels in **data visualization** and application of statistical methods
- Also: can be used to collect data on the Internet
- **Beginner friendly** for people with no programming background

➡ Can be used at **every stage** of our **workflow** (no need to switch programs)!

University of Zurich[UZH]

Department of Political Science

Data import in R

Data import in R

- You can save R objects (e.g. a dataframe) in in .Rda-files
- You use “load()” to import an .Rda file into R.

- Most datasets will not be prepared for R (e.g. .csv-files, Excel files, etc.) and we will learn in the next exercise more about the ways to recognize data formats and how to import them.

Outlook

- There are many packages suitable to load specific types of data into R:
  - **jsonlite**: for JSON data
  - **xml2**: for XML data
  - **readr**: for Text data
  - **haven**: for SPSS, SAS, Stata files
  - **readxl**: for Microsoft excel files (.xls or .xlsx)
  - **DBI**: for connections to data bases
  - **httr**: to retrieve data from APIs
  - **rvest**: to retrieve data from websites/html