

First Year Project, Spring 2021

# Lecture 1: Introduction, Logistics, Project 1

Instructor: Michael Szell

Feb 5, 2021



# Today I introduce you to the course and to Project 1

## Course Logistics



## Project 1 Data set first exploration

1	Accident_Index,Location_Easting_NGR,Location_Northing_NGR,Longitude,Latitude,Police_Force,Accident_Severity,Number_of_Vehicles,Number_of_Casualties,Date,Day_of_Week,Time,Local_Authority_(District),Local_Authority_(Highway),1st_Road_Class,1st_Road_Number,Road_Type,Speed_Limit,Junction_Control,Junction_Priority,2nd_Road_Class,2nd_Road_Number,Pedestrian_Crossing_Human_Control,Pedestrian_Crossing_Physical_Facilities,Light_Conditions,Weather_Conditions,Road_Surface_Conditions,Special_Conditions_at_Site,Carriageway_Hazards,Urban_or_Rural_Area,Did_Police_Officer_Attend_Scene_of_Accident,LSOA_of_Accident_Location
2	2819010128300, 526210, 180407, -0.153842, 51.500057, 1, 3, 2, 3, 18/02/2019, 2, 17:50, 1, E09000033, 3, 4202, 1, 30, 1, 2, 3, 4202, 0, 5, 1, 1, 1, 0, 0, 1, 3, E01004762
3	2819010152270, 538219, 172463, -0.127949, 51.436208, 1, 3, 2, 1, 15/01/2019, 3, 21:45, 9, E09000022, 3, 23, 2, 30, 0, -1, -1, 0, -1, -1, 4, 1, 1, 0, 0, 1, 3, E01003117
4	2819010155191, 530222, 182543, -0.124193, 51.526795, 1, 3, 2, 1, 01/01/2019, 3, 01:50, 2, E09000027, 4, 504, 6, 30, 3, 4, 6, 0, 0, 4, 1, 1, 0, 0, 1, 1, E01000943
5	2819010155192, 525531, 184605, -0.191044, 51.546387, 1, 2, 1, 1, 01/01/2019, 3, 01:20, 2, E09000087, 4, 510, 6, 28, 3, 4, 4, 510, 0, 0, 4, 1, 2, 2, 0, 1, 1, E01000923
6	2819010155194, 524920, 184004, -0.209064, 51.541121, 1, 3, 2, 2, 01/01/2019, 3, 00:40, 28, E09000005, 3, 4083, 6, 30, 5, 4, 6, 0, 0, 0, 4, 1, 2, 0, 1, 1, E01000254
7	2819010155195, 520188, 185261, 0.020461, 51.548879, 1, 3, 2, 3, 01/01/2019, 3, 02:45, 17, E09000025, 5, 0, 3, 38, 0, -1, -1, 0, 0, 0, 4, 1, 1, 0, 0, 1, 1, E01003544
8	2819010155196, 532424, 164086, -0.099071, 51.367605, 1, 3, 1, 1, 01/01/2019, 3, 01:35, 20, E09000008, 3, 235, 5, 30, 5, 2, 4, 271, 0, 5, 4, 1, 1, 0, 0, 1, 1, E01001843
9	2819010155198, 532773, 178468, -0.088078, 51.489500, 1, 3, 3, 5, 01/01/2019, 3, 02:10, 8, E09000028, 6, 0, 6, 28, 3, 4, 6, 0, 0, 0, 4, 1, 1, 0, 0, 1, 1, E01000392
10	2819010155206, 540535, 188113, 0.141857, 51.572326, 1, 3, 2, 1, 01/01/2019, 3, 01:15, 16, E09000082, 3, 1112, 6, 30, 5, 2, 3, 110, 0, 5, 4, 1, 1, 0, 0, 1, 1, E01000031
11	2819010155207, 522267, 188185, -0.243769, 51.390529, 1, 3, 2, 1, 01/01/2019, 3, 04:30, 22, E09000024, 4, 282, 7, 30, 1, 4, 3, 3, 0, 4, 1, 1, 0, 0, 1, 1, E01003469
12	2819010155209, 543649, 186237, 0.070730, 51.556734, 1, 3, 1, 1, 01/01/2019, 3, 01:15, 14, E09000026, 3, 110, 3, 20, 3, 2, 6, 0, 0, 4, 1, 1, 0, 0, 1, 1, E01003689
13	2819010155210, 537356, 183448, -0.021065, 51.533230, 1, 3, 1, 1, 01/01/2019, 3, 02:00, 5, E09000019, 3, 105, 3, 20, 9, 3, 6, 0, 0, 4, 1, 1, 0, 0, 1, 1, E01003276
14	2819010155216, 532724, 185103, -0.007162, 51.549218, 1, 3, 1, 1, 01/01/2019, 3, 02:45, 3, E09000019, 3, 105, 3, 20, 9, 3, 6, 0, 0, 4, 1, 1, 0, 0, 1, 1, E01002776
15	2819010155217, 543616, 188274, 0.070277, 51.557875, 1, 2, 2, 1, 01/01/2019, 3, 04:10, 14, E09000026, 3, 1118, 3, 40, 0, -1, 1, 0, 0, 5, 4, 2, 1, 0, 0, 1, 1, E01003728
16	2819010155220, 527868, 179868, 0.178009, 51.496210, 1, 3, 2, 1, 01/01/2019, 3, 00:20, 12, E09000022, 3, 4, 30, 0, -1, 1, 0, 0, 0, 4, 1, 1, 0, 0, 1, 1, E01002821
17	2819010155221, 538487, 180538, -0.007064, 51.506832, 1, 2, 1, 1, 01/01/2019, 3, 05:55, 5, E09000030, 6, 0, 6, 28, 5, 4, 5, 0, -1, 0, 0, 0, 4, 1, 1, 0, 0, 1, 1, E01003270
18	2819010155225, 512750, 180199, -0.376691, 51.509401, 1, 3, 2, 1, 01/01/2019, 3, 00:50, 27, E09000089, 3, 3085, 6, 30, 3, 4, 6, 0, 0, 4, 1, 1, 0, 0, 1, 1, E01001351
19	2819010155226, 511549, 182802, -0.393167, 51.533115, 1, 3, 2, 1, 01/01/2019, 3, 02:45, 27, E09000089, 6, 0, 6, 30, 0, -1, -1, 0, 0, 0, 7, 1, 1, 0, 0, 1, 1, E01001331
20	2819010155232, 531510, 167115, -0.111369, 51.387849, 1, 2, 2, 1, 01/01/2019, 3, 05:54, 20, E09000008, 3, 235, 5, 30, 3, 4, 5, 0, 0, 0, 7, 8, 1, 0, 0, 1, 3, E01001181
21	2819010155234, 542152, 185138, 0.048712, 51.547165, 1, 3, 2, 1, 01/01/2019, 3, 07:30, 17, E09000025, 3, 117, 3, 30, 7, 2, 6, 0, 0, 0, 4, 1, 1, 0, 0, 1, 1, E01003534
22	2819010155242, 515593, 177263, -0.336784, 51.482519, 1, 3, 2, 1, 01/01/2019, 3, 09:41, 25, E09000018, 5, 0, 3, 30, 7, 2, 3, 4, 0, 5, 1, 0, 0, 0, 1, 1, E01002675
23	2819010155254, 534128, 189077, 0.065175, 51.559268, 1, 3, 2, 1, 01/01/2019, 3, 14:18, 31, E09000014, 6, 0, 6, 20, 3, 4, 6, 0, 0, 0, 1, 1, 0, 0, 1, 1, E01002408
24	2819010155254, 531638, 180408, -0.104600, 51.507028, 1, 3, 1, 1, 01/01/2019, 3, 15:20, 8, E09000020, 3, 3209, 5, 30, 5, 2, 3, 201, 2, 4, 1, 1, 1, 0, 0, 1, 1, E01003934
25	2819010155256, 528616, 177428, -0.149195, 51.481194, 1, 3, 2, 1, 01/01/2019, 3, 07:30, 10, E09000032, 3, 3216, 6, 30, 0, -1, -1, 0, 0, 1, 1, 0, 0, 1, 1, E01003909
26	2819010155257, 595349, 183292, -0.482399, 51.537901, 1, 3, 2, 1, 01/01/2019, 3, 11:20, 26, E09000017, 3, 408, 5, 30, 5, 2, 5, 0, 0, 5, 1, 1, 0, 0, 1, 1, E010033724
27	2819010155263, 538072, 171167, -0.002610, 51.422484, 1, 2, 2, 1, 01/01/2019, 3, 13:48, 7, E09000023, 3, 21, 6, 30, 3, 4, 6, 0, 0, 0, 1, 1, 0, 0, 1, 1, E010003243
28	2819010155276, 519377, 184736, 0.514896, 51.508896, 1, 2, 2, 1, 01/01/2019, 3, 16:39, 28, E09000005, 5, 0, 6, 28, 7, 0, 0, 0, 1, 1, E01000528
29	2819010155282, 537568, 182299, -0.418574, 51.528263, 1, 2, 2, 1, 01/01/2019, 3, 17:37, 5, E09000030, 5, 0, 6, 28, 0, -1, -1, 0, 1, 4, 4, 1, 1, 0, 0, 1, 1, E01004239
30	2819010155284, 534595, 181757, -0.061492, 51.538706, 1, 3, 2, 1, 01/01/2019, 3, 06:15, 5, E09000030, 3, 11, 5, 20, 9, -1, 6, 0, 0, 1, 1, 0, 0, 1, 1, E01004322
31	2819010155294, 526033, 177533, -0.186335, 51.482719, 1, 3, 2, 1, 01/01/2019, 3, 17:51, 12, E09000082, 3, 304, 5, 30, 3, 4, 6, 0, 0, 1, 1, 0, 0, 1, 1, E01004232
32	2819010155297, 523656, 175151, -0.221362, 51.461835, 1, 3, 1, 1, 01/01/2019, 3, 12:00, 10, E09000032, 3, 205, 6, 30, 0, -1, -1, 0, 0, 0, 1, 1, 0, 0, 1, 1, E01004505

## Project 1 Motivation



# Example of new skills/tools you will learn in Project 1

# MaskedArray for data cleaning

# Pearson's test of independence

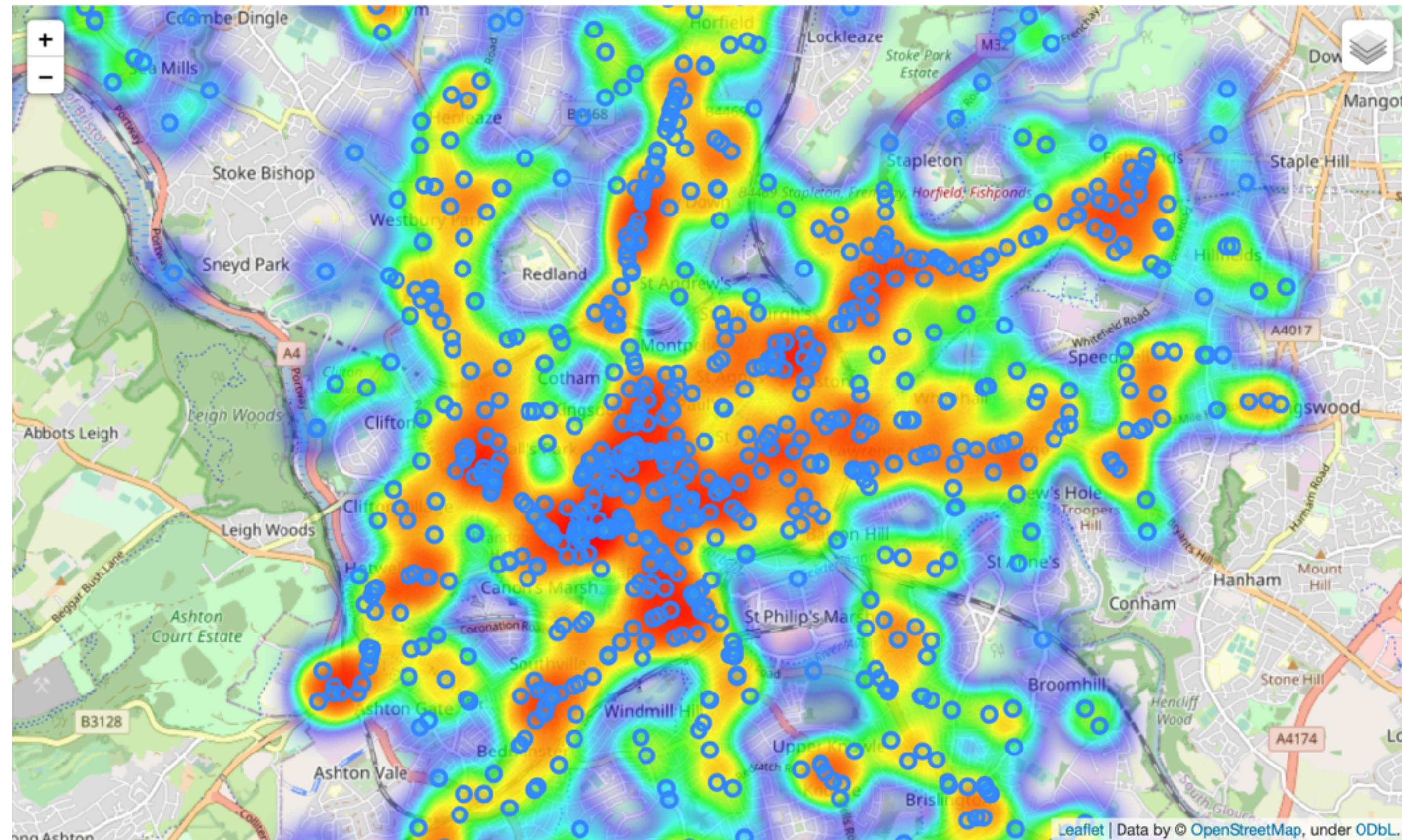
$\chi^2$

The figure consists of three vertically aligned dot plots. The y-axis for all three is labeled 'Age of Casualty' and ranges from 0 to 80. The x-axis categories are 'Pedestrian', 'Cyclist', and 'Car occupant'. Each category has a distinct color: blue for Pedestrian, orange for Cyclist, and green for Car occupant. The data points are represented by small dots. In the Pedestrian plot, there is a dense cluster of points between ages 10 and 50, with a few outliers up to 85. The Cyclist plot shows a similar distribution but with fewer points overall. The Car occupant plot has the highest density of points, particularly between ages 20 and 40, with a significant number of points clustered around 30.

# Categorical scatterplots

# shapely + folium

## for spatial filtering and map visualization



# Why this course?

You need **hands-on** experience (lectures and tutorials are not enough)

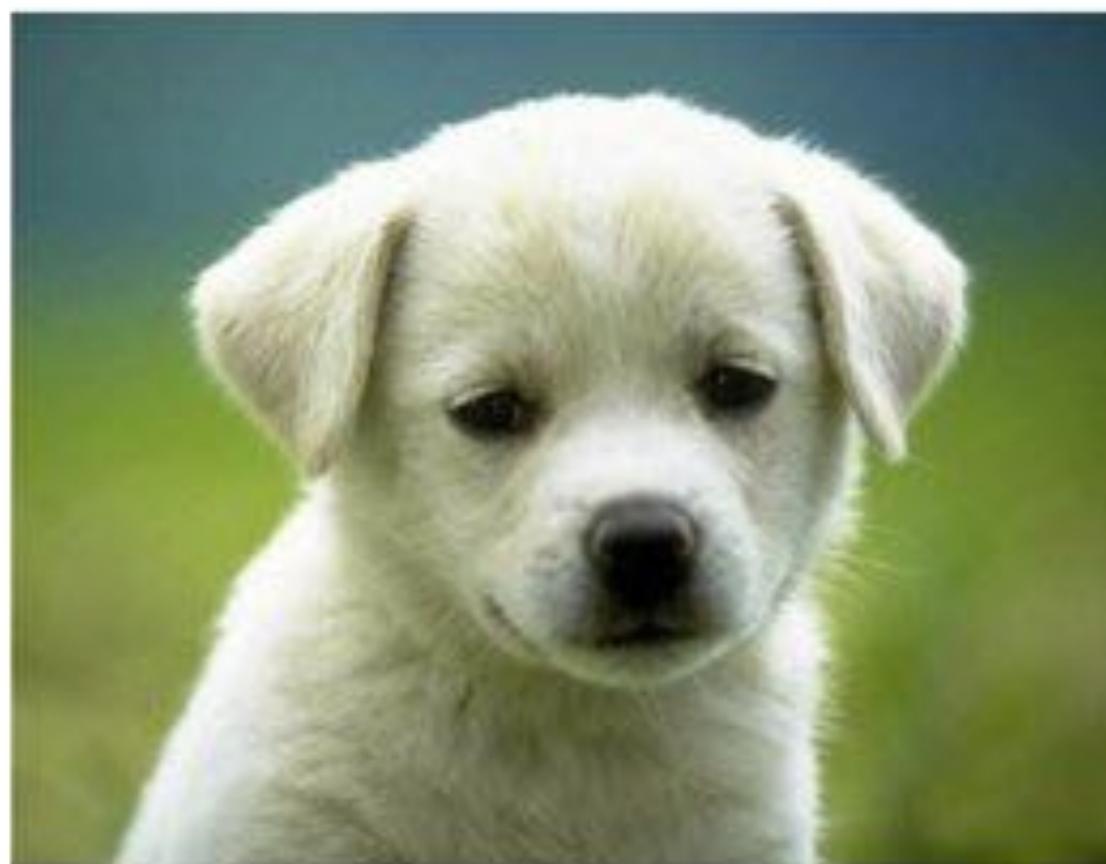
You need to work on **\*real-world\*** problems with **\*real-world\*** data

# Why this course?

You need **hands-on** experience (lectures and tutorials are not enough)

You need to work on **\*real-world\*** problems with **\*real-world\*** data

Data sets  
in tutorials



Data sets in  
the wild



# Why this course?

You need **hands-on** experience (lectures and tutorials are not enough)

You need to work on **\*real-world\*** problems with **\*real-world\*** data

You need to experience **the whole data science pipeline**:

- Problem formulation
- Data analysis
- Communication of results

Prepare you for **group work** with **RANDOM** people



After you have taken this course, you will be able to:

Develop code with a professional version control system (git)

Identify and discuss issues in a real data science project

After you have taken this course, you will be able to:

Develop code with a professional version control system (git)

Identify and discuss issues in a real data science project

Implement the whole data science pipeline:

- Problem formulation
- Data analysis
- Communication of results

After you have taken this course, you will be able to:

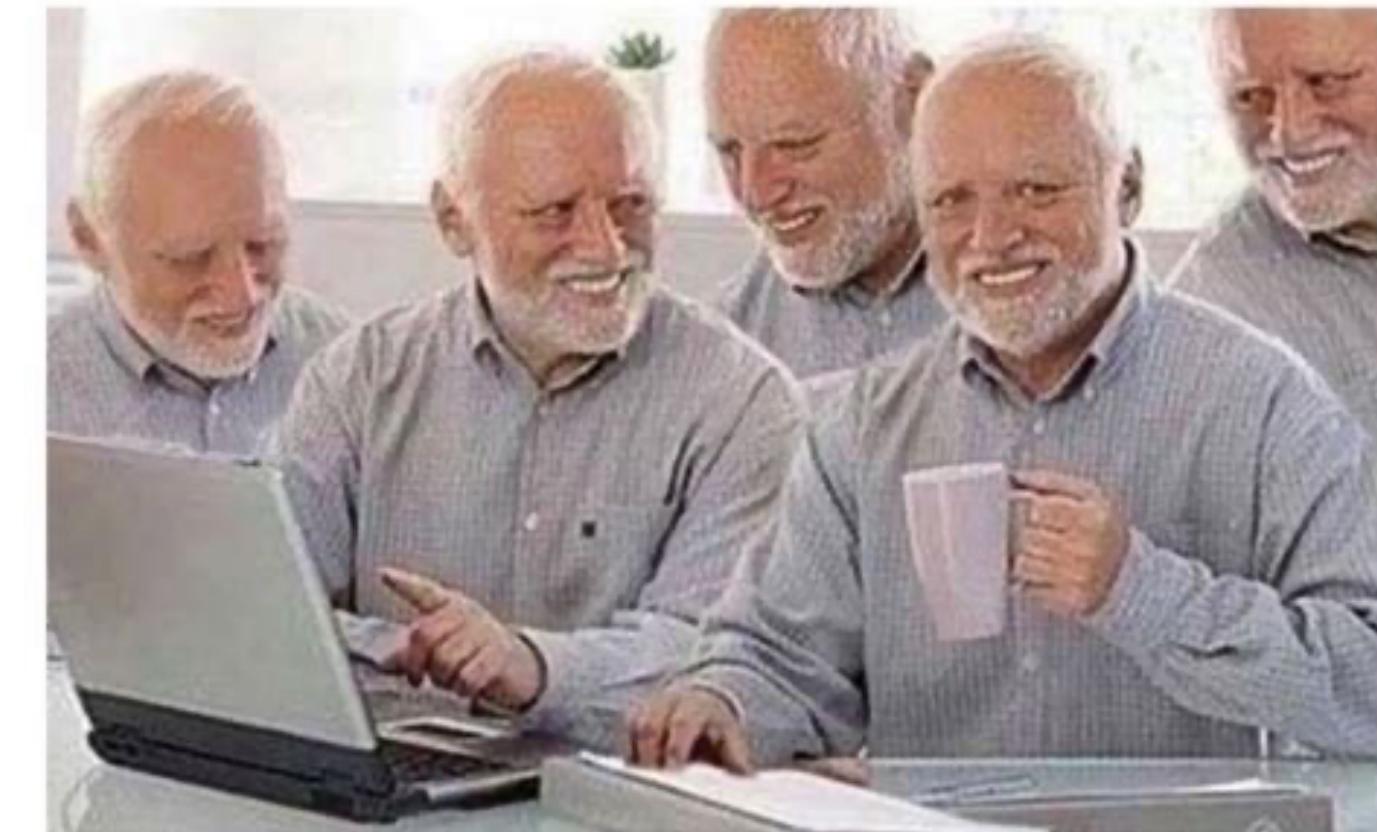
Develop code with a professional version control system (git)

Identify and discuss issues in a real data science project

Implement the whole data science pipeline:

- Problem formulation
- Data analysis
- Communication of results

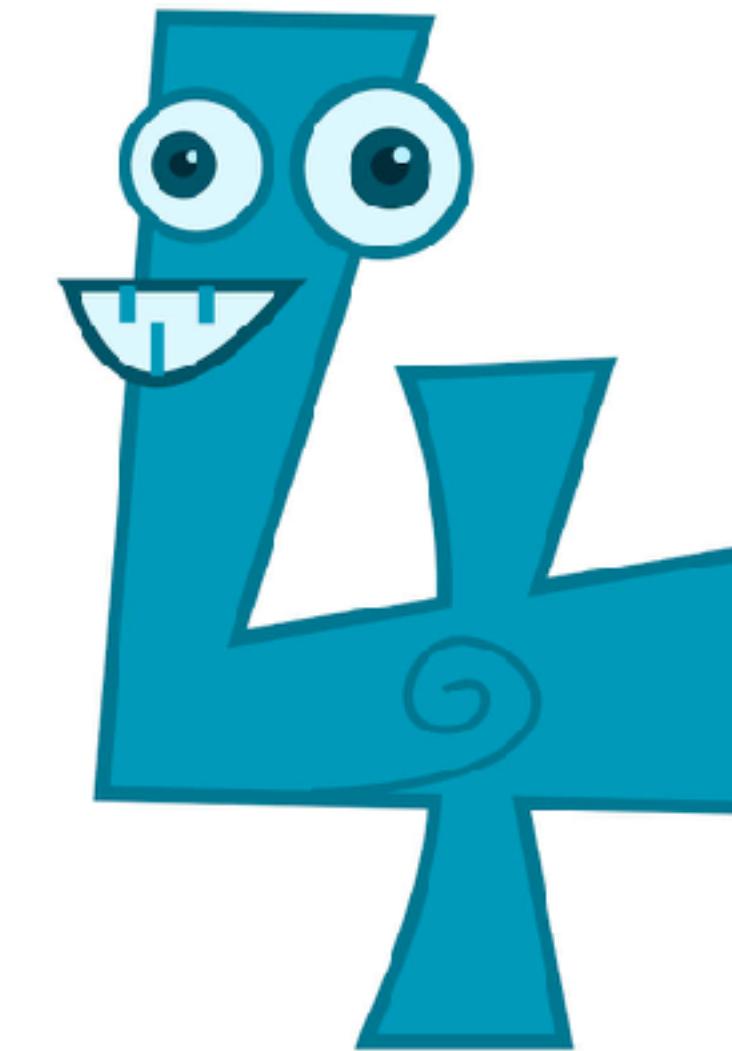
When you're doing a group project  
but you end up doing all the work



Work with RANDOM people, be able to handle teamwork issues

# What is going to happen?

You work on 4 data science projects from 4 diverse fields in groups of 5 over 4 months



4 instructors, 4 TAs, 4 hand-ins, 2 exams

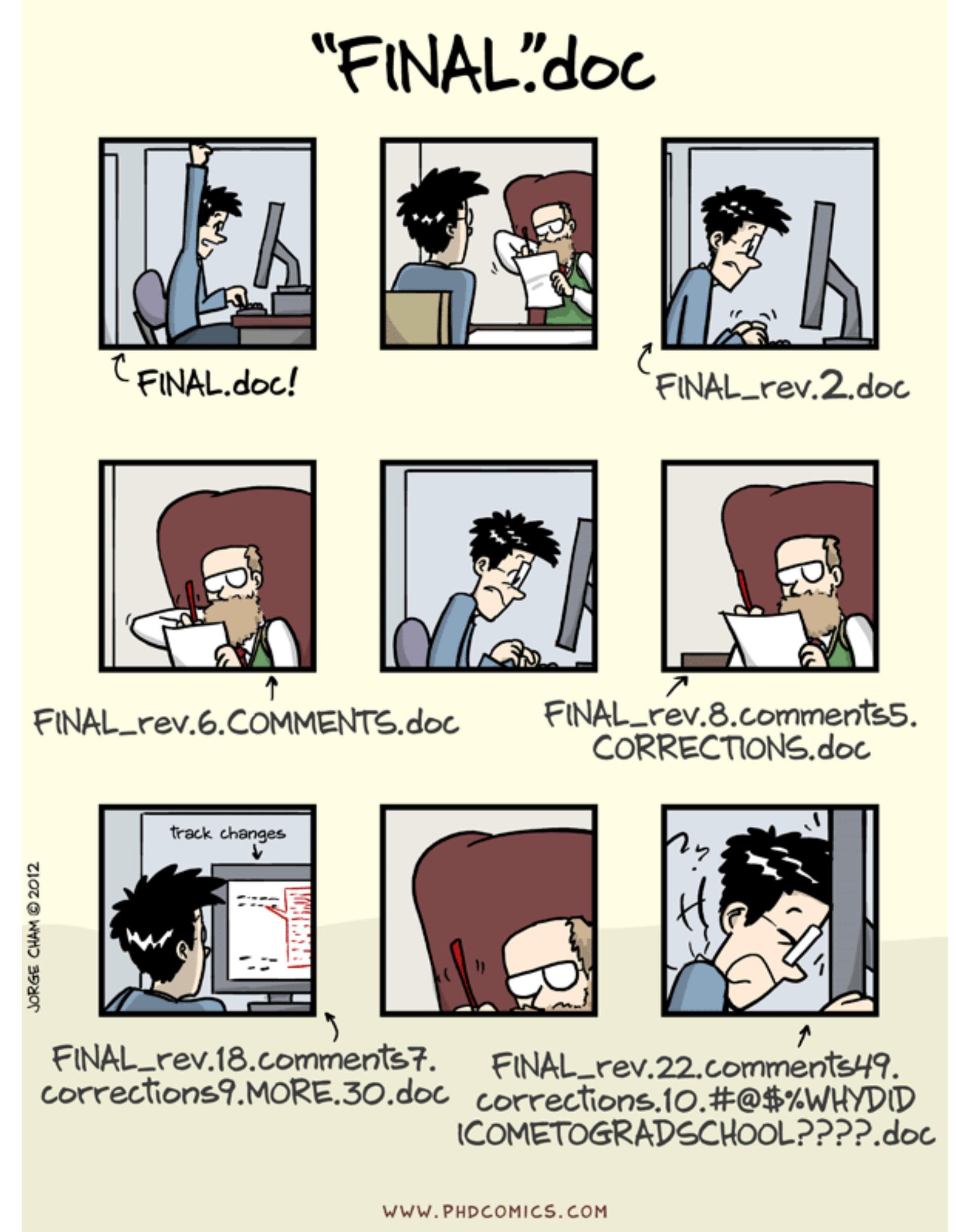
# Questions

# Get acquainted with git / GitHub

Git is a distributed version-control system for tracking changes in any set of files, originally designed to coordinate group work.



Similar systems: CSV, SVN, Mercurial



# Get acquainted with git / GitHub

GitHub, Inc. is a subsidiary of Microsoft which provides hosting for software development and version control using Git

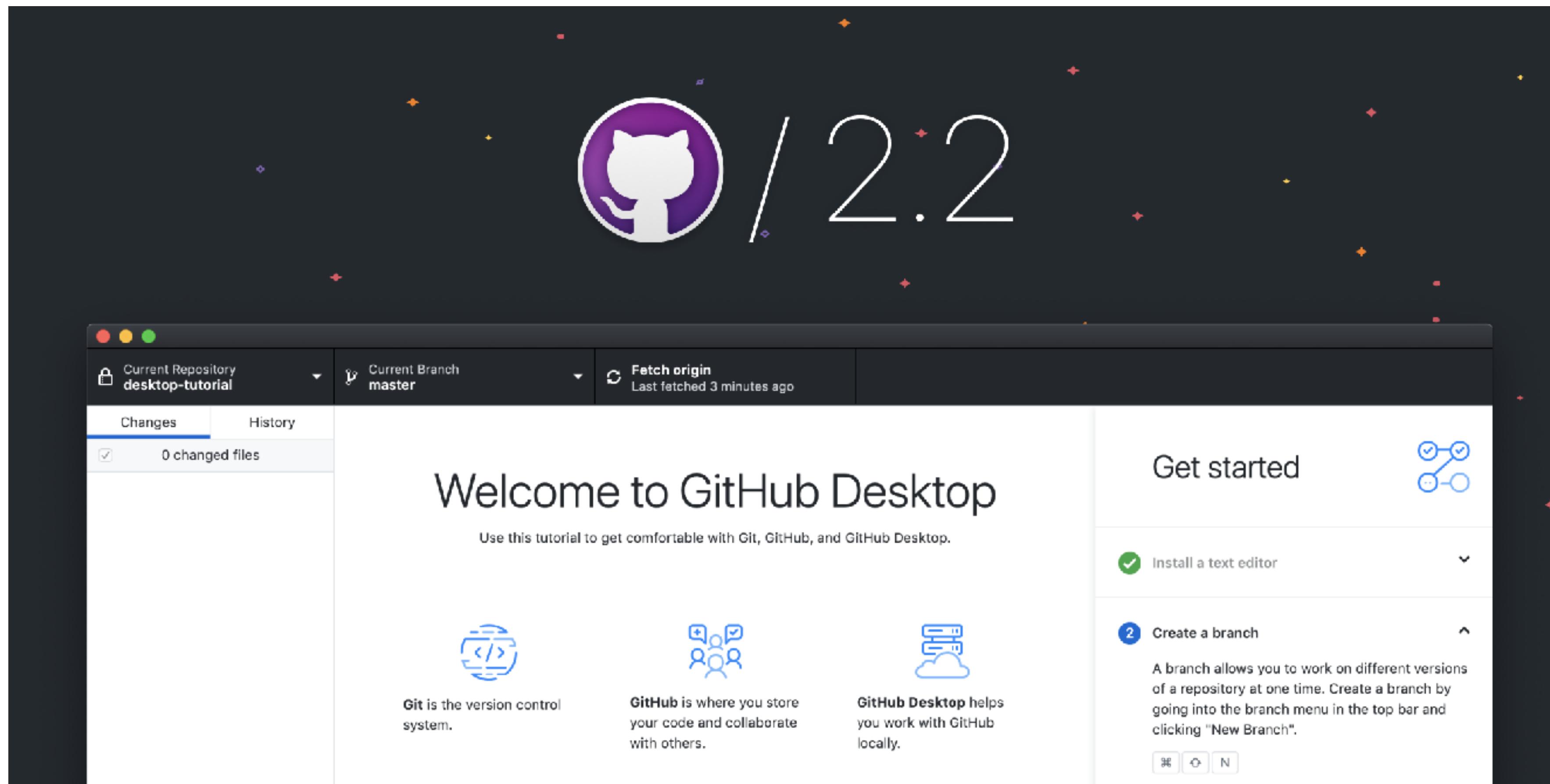


GitLab is a good alternative



# Get acquainted with git / GitHub

GitHub Desktop is a Graphical User Interface (GUI) for Github



**Let's get started with the project!**

# Before we delve into the project, you MUST get organized!

```
├── LICENSE
├── Makefile      <- Makefile with commands like `make data` or `make train`
├── README.md     <- The top-level README for developers using this project.
└── data
    ├── external   <- Data from third party sources.
    ├── interim    <- Intermediate data that has been transformed.
    ├── processed   <- The final, canonical data sets for modeling.
    └── raw         <- The original, immutable data dump.

── docs          <- A default Sphinx project; see sphinx-doc.org for details

── models        <- Trained and serialized models, model predictions, or model summaries

── notebooks     <- Jupyter notebooks. Naming convention is a number (for ordering),
                    the creator's initials, and a short '-' delimited description, e.g.
                    `1.0-jqp-initial-data-exploration`.

── references    <- Data dictionaries, manuals, and all other explanatory materials.

── reports       <- Generated analysis as HTML, PDF, LaTeX, etc.
    └── figures    <- Generated graphics and figures to be used in reporting

── requirements.txt <- The requirements file for reproducing the analysis environment, e.g.
                      generated with `pip freeze > requirements.txt`

── src
    ├── __init__.py  <- Source code for use in this project.
    │               <- Makes src a Python module
    ├── data         <- Scripts to download or generate data
    │   └── make_dataset.py
    ├── features     <- Scripts to turn raw data into features for modeling
    │   └── build_features.py
    ├── models       <- Scripts to train models and then use trained models to make
                        predictions
    │   ├── predict_model.py
    │   └── train_model.py
    └── visualization <- Scripts to create exploratory and results oriented visualizations
        └── visualize.py

── tox.ini        <- tox file with settings for running tox; see tox.readthedocs.io
```

**Folder structure is not a science, but based on years of experience**

**Cookiecutter Data Science**

<https://github.com/drivendata/cookiecutter-data-science>

# Before we delve into the project, you MUST get organized!

```
├── LICENSE  
├── Makefile      <- Makefile with commands like `make data` or `make train`  
├── README.md     <- The top-level README for developers using this project.  
└── data           
    ├── external    <- Data from third party sources.  
    ├── interim     <- Intermediate data that has been transformed.  
    ├── processed   <- The final, canonical data sets for modeling.  
    └── raw         <- The original, immutable data dump.  
  
── docs          <- A default Sphinx project; see sphinx-doc.org for details  
  
── models        <- Trained and serialized models, model predictions, or model summaries  
  
── notebooks     <- Jupyter notebooks. Naming convention is a number (for ordering),  
                  the creator's initials, and a short '-' delimited description, e.g.  
                  `1.0-jqp-initial-data-exploration`.  
  
── references    <- Data dictionaries, manuals, and all other explanatory materials.  
  
── reports       <- Generated analysis as HTML, PDF, LaTeX, etc.  
    └── figures     <- Generated graphics and figures to be used in reporting  
  
── requirements.txt <- The requirements file for reproducing the analysis environment, e.g.  
                      generated with `pip freeze > requirements.txt`  
  
── src           <- Source code for use in this project.  
    ├── __init__.py  <- Makes src a Python module  
    └── data         <- Scripts to download or generate data  
        └── make_dataset.py  
  
    ├── features     <- Scripts to turn raw data into features for modeling  
        └── build_features.py  
  
    ├── models       <- Scripts to train models and then use trained models to make  
                      predictions  
        └── predict_model.py  
        └── train_model.py  
  
    └── visualization <- Scripts to create exploratory and results oriented visualizations  
        └── visualize.py  
  
tox.ini      <- tox file with settings for running tox; see tox.readthedocs.io
```

## MUST haves:

- README.md
- Data: raw, interim, processed
- Notebooks / Code
- References
- Reports / Figures

Cookiecutter Data Science

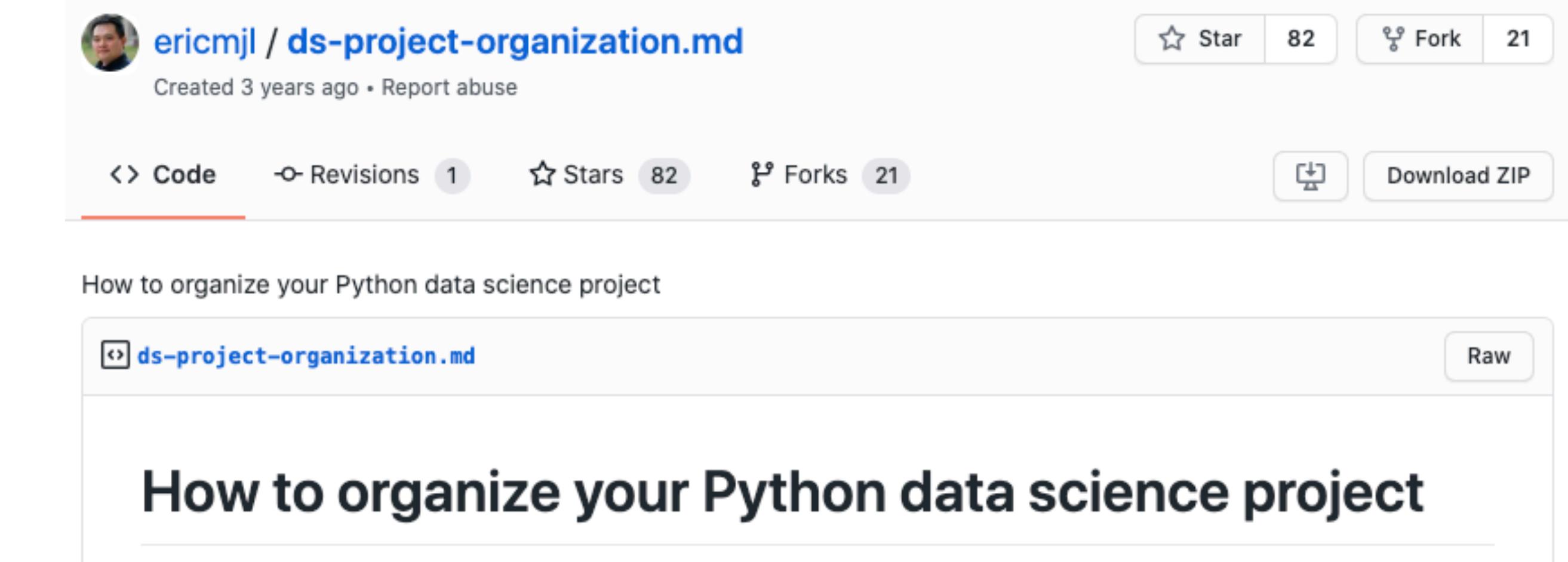
<https://github.com/drivendata/cookiecutter-data-science>

# Before we delve into the project, you MUST get organized!

## How to Create a Professional Github Data Science Repository

Good practices in repository structure, documenting jupyter notebooks, and writing an informative README

 Ahilan Srivishnumohan Jul 25, 2020 · 7 min read \*



The screenshot shows a GitHub gist page for a file named `ds-project-organization.md`. The page title is "How to organize your Python data science project". The file content is displayed in a monospaced font, featuring the same title at the top. Below the title, there is a section with the heading "How to organize your Python data science project". The GitHub interface includes standard navigation and statistics: "Code", "Revisions 1", "Stars 82", "Forks 21", "Star" (82), "Fork" (21), "Download ZIP", and a "Raw" link.

<https://gist.github.com/ericmjl/27e50331f24db3e8f957d1fe7bbbe510>

<https://towardsdatascience.com/how-to-create-a-professional-github-data-science-repository-84e9607644a2>

# Data is usually NOT stored on Github

Github is for code and work documents, not data.

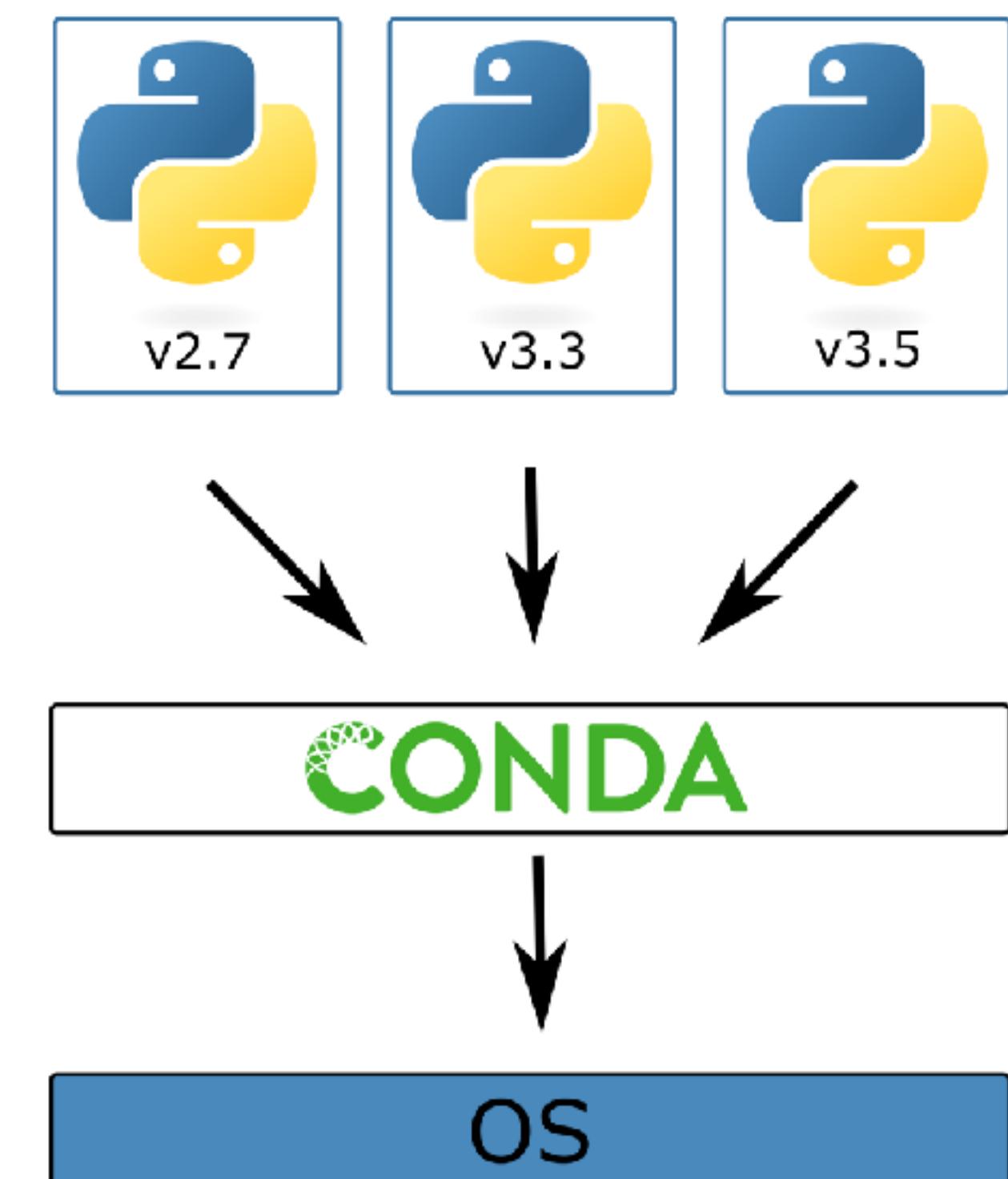
There are hard limitations: 100 MB files, 2 GB push, ~5 GB repo

If you have hundreds of MB data, add the data folder to .gitignore

# Advanced organization (not for this course)

In Data Sci, we use a **virtual environment** (venv) for each project

Why?



# Advanced organization (not for this course)

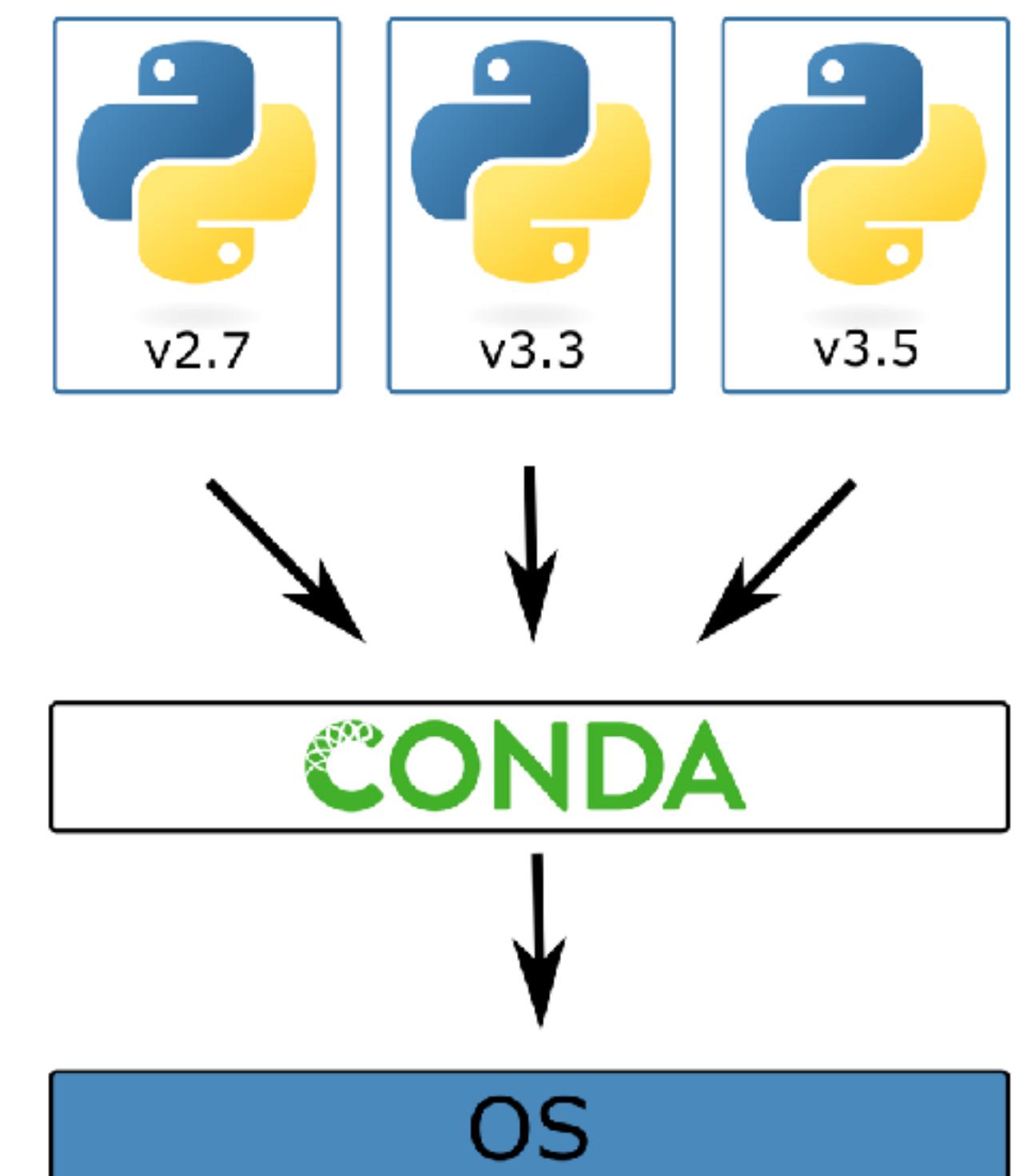
In Data Sci, we use a **virtual environment** (venv) for each project

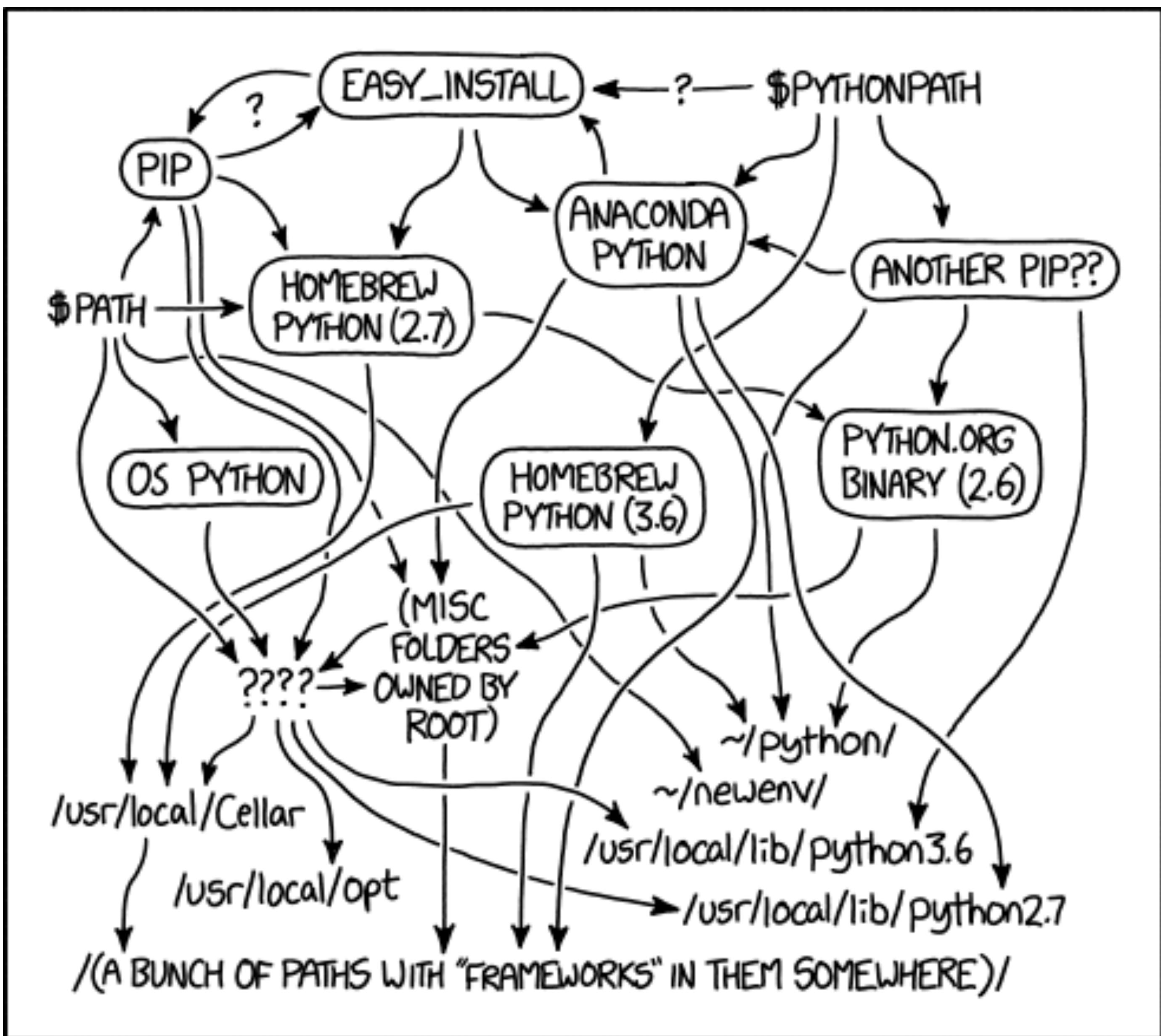
Why?

1) Reproducibility:

- Python changes all the time
- Python packages change all the time
- requirements.txt differ between projects

2) To not mess up your Python install





MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED  
THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

# Advanced organization (not for this course)

```
requirements.txt      *
1  matplotlib>=3.3.3
2  numpy>=1.19.4
3  pandas>=1.0.3
4  pyproj>=2.6.1.post1
5  geojson>=2.5.0
6  shapely>=1.7.0
7  csv>=1.0
8  networkx>=2.5
9  igraph>=0.8.3
10 fiona>=1.8.18
11 osmnx==0.16.2
12 geopandas>=0.8.1
13 tqdm>=4.55.0
14 haversine>=2.3.0
```

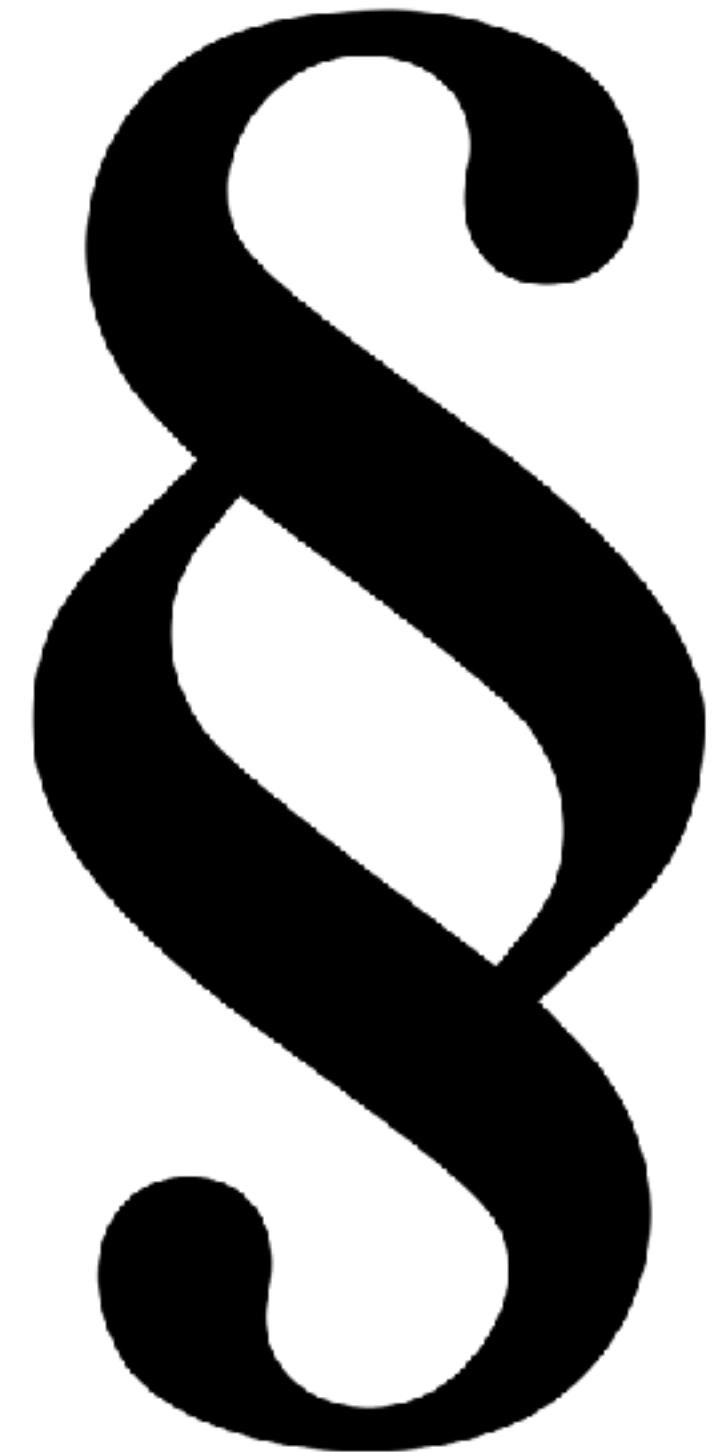
```
pip freeze > requirements.txt
```

# Advanced organization (not for this course)

## LICENSE

If you publish your project, the license is the legal document telling others how they can use it.

No serious company will ever risk building on your code if you don't have a license.



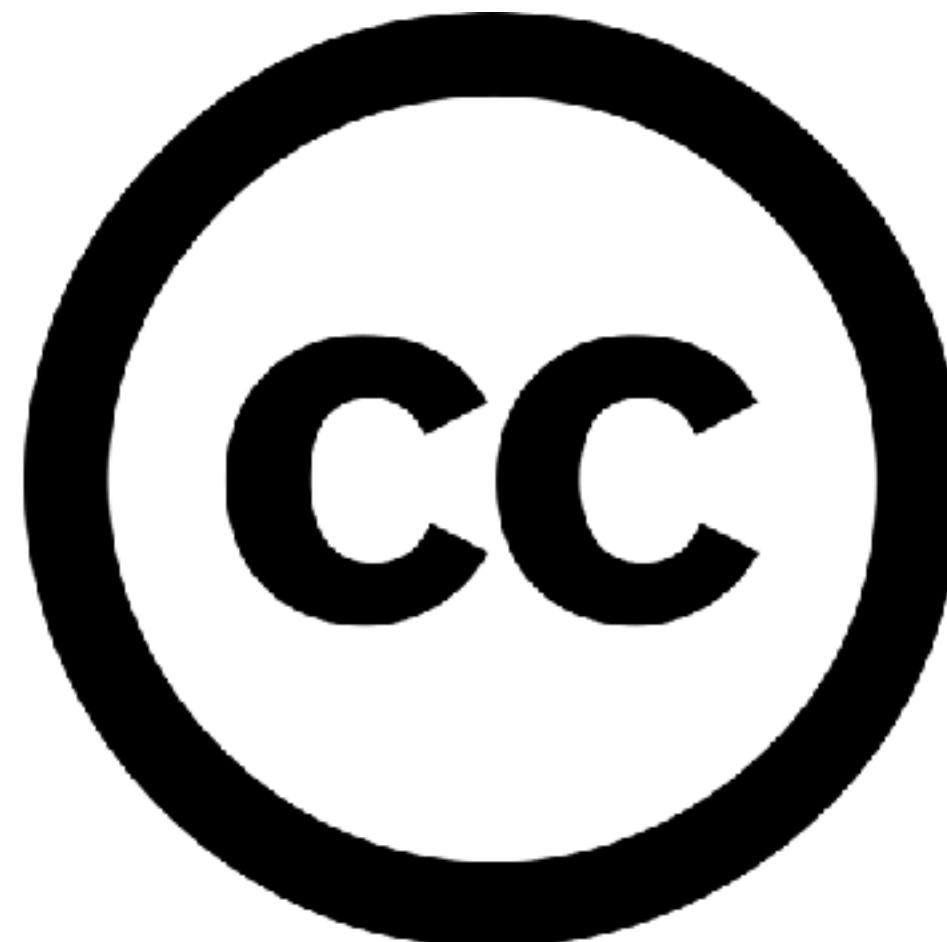
# Advanced organization (not for this course)

## LICENSE

If you publish your project, the license is the legal document telling others how they can use it.

Choose an established license,  
don't write it yourself.

Consider copyleft and creative commons to protect freedoms.



<https://creativecommons.org/>

<https://www.gnu.org/philosophy/free-sw.en.html>

<https://softwareengineering.stackexchange.com/questions/47028/how-could-we-rewrite-the-no-evil-license-to-make-it-free>

# Advanced organization (not for this course)

## DOCS

If you want others to use and build on your code, you need excellent documentation. It makes or breaks its success.

Expect investing >20% of the whole project time on docs.

**Let's get started with the project!**  
**(for real)**

# Data set: Road collisions in the UK in 2019

data.gov.uk | Find open data

**BETA** This is a new service – your [feedback](#) will help us to improve it

[Home](#) > Department for Transport > Road Safety Data

## Road Safety Data

**Published by:** Department for Transport

**Last updated:** 08 January 2021

**Topic:** Transport

**Licence:** [Open Government Licence](#)

### Summary

[Road Safety Statistics releases](#)



3 Data Tables  
Metadata

# How many road deaths since 2000?

<https://www.mentimeter.com/s/a3b93a55f54d81cf41095bd34a3a6fd5/4117bf082c85>



## Overview: Urban / Spatial Data Science

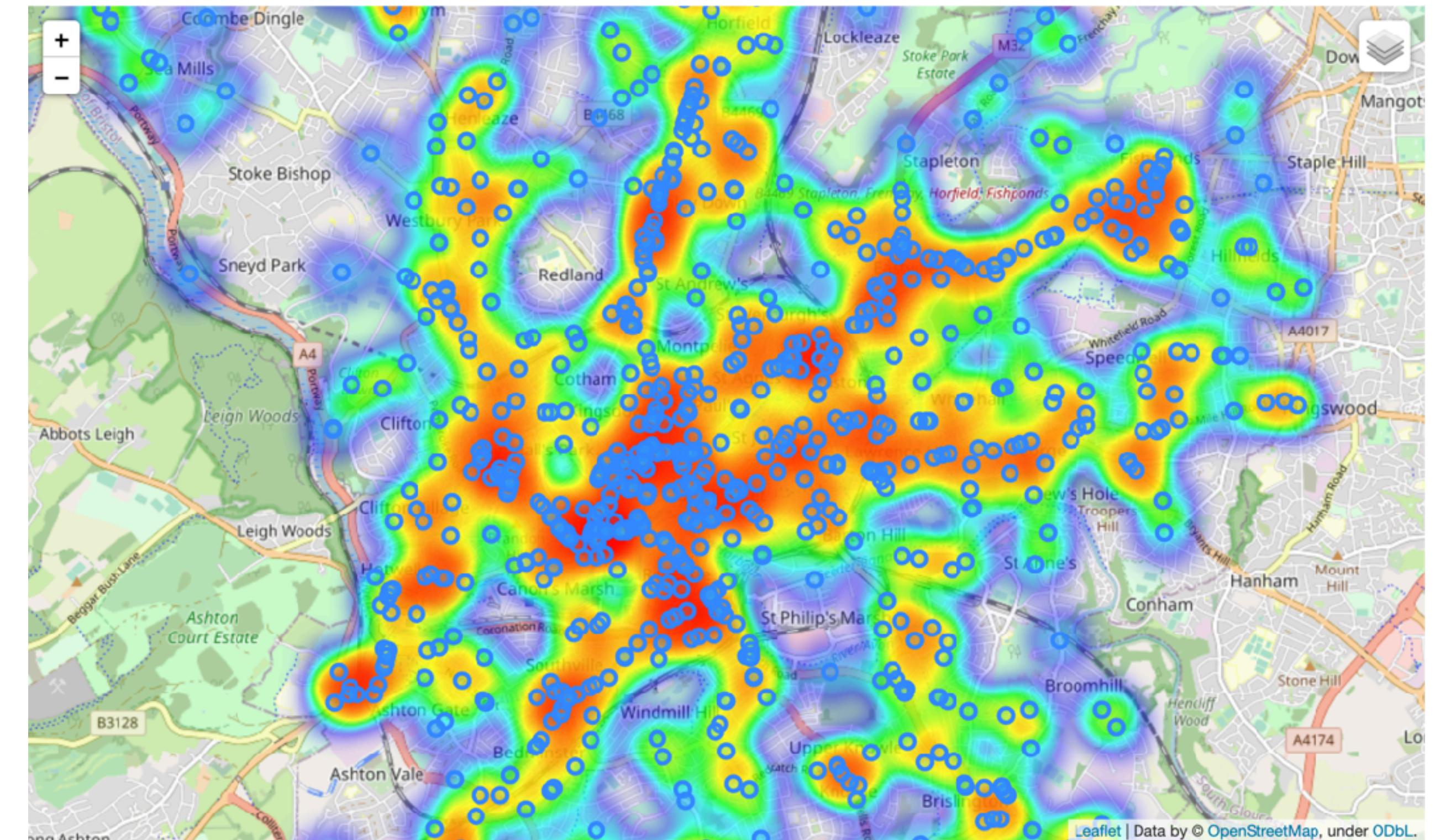
In this project, you will complete tasks similar to data scientists working for a department of transportation or a city government, to inform city leaders about traffic fatalities and injuries, and give insights for urban/transport planning. You will explore a data set of all recorded road collisions in Great Britain in the year 2019 from the UK Department of Transport with details about the circumstances of personal injury.

The major parts of the project are:

- Exploring and transforming the data, making numerical and visual reports
- Connecting data tables (accidents, vehicles, casualties)
- Reporting possible statistical associations by filtering for a variety of attributes
- Visualizing the data on a map
- Involving self-downloaded/scraped external data sets in the analysis

# Project Tasks

- 0) Data filtering and cleaning**
- 1) Single variable analysis**
- 2) Associations**
- 3) Map visualization**
- 4) Open question**



Details: [project01\\_description.pdf](#)

# Project Hand-in

**Hand-in 1) a report, 2) your code (Jupyter), 3) the git log.**

In your report you analyze the data set answering specific research questions.

Your group's **github repo fyp2021p01gXX** must document your whole process.

**All Python packages allowed, including pandas. Comment** your code assuming the reader has no knowledge of such extra packages.

Details: [project01\\_description.pdf](#)

# Hand-ins: What is important

**Argue for your choices of data/analysis, explain your decisions/results**

**is much more important than:**

Size of data set, number of results/analyses/methods

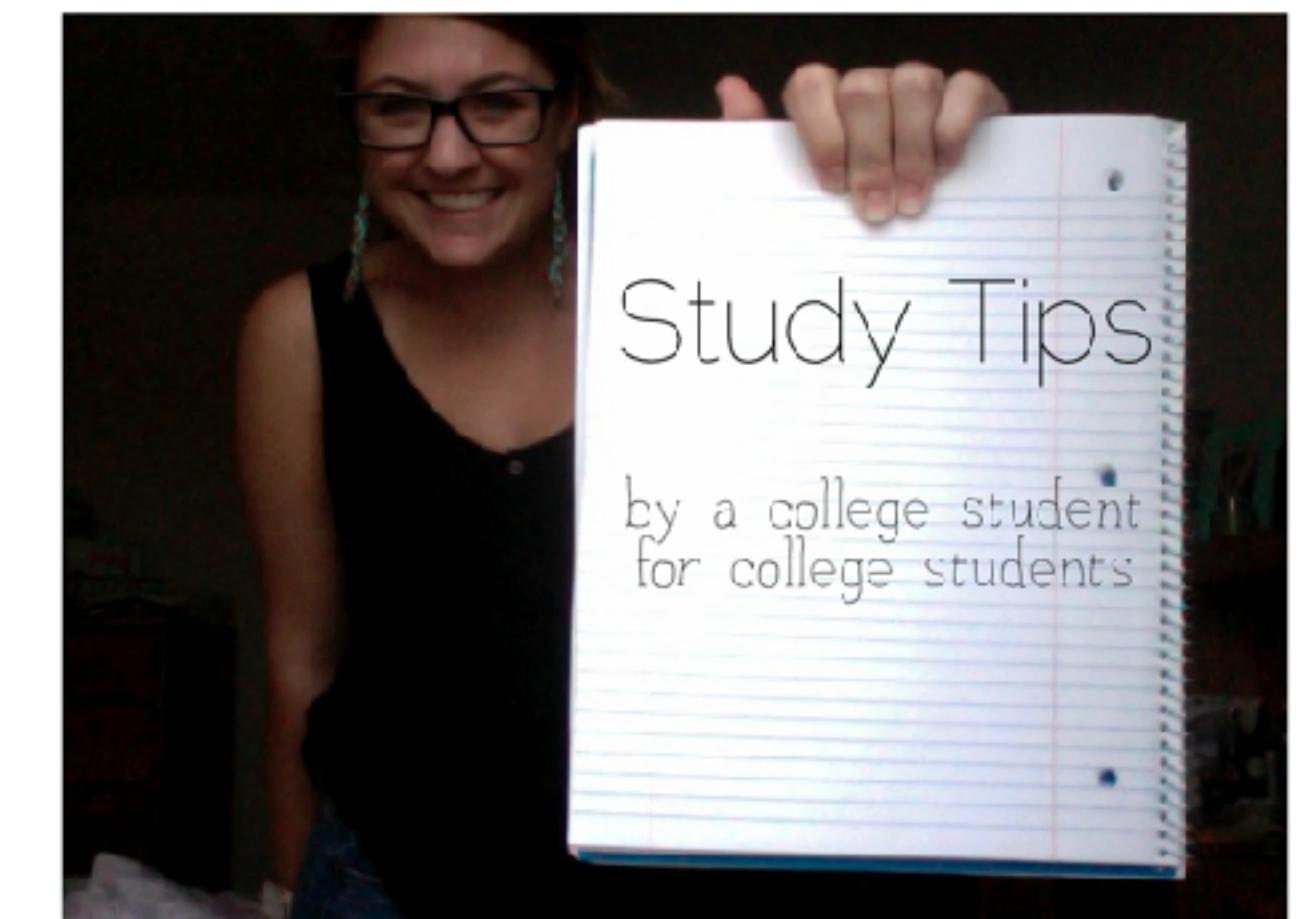
10 min: Explore the data (in any way you like)

Then: give first impressions

# Jupyter

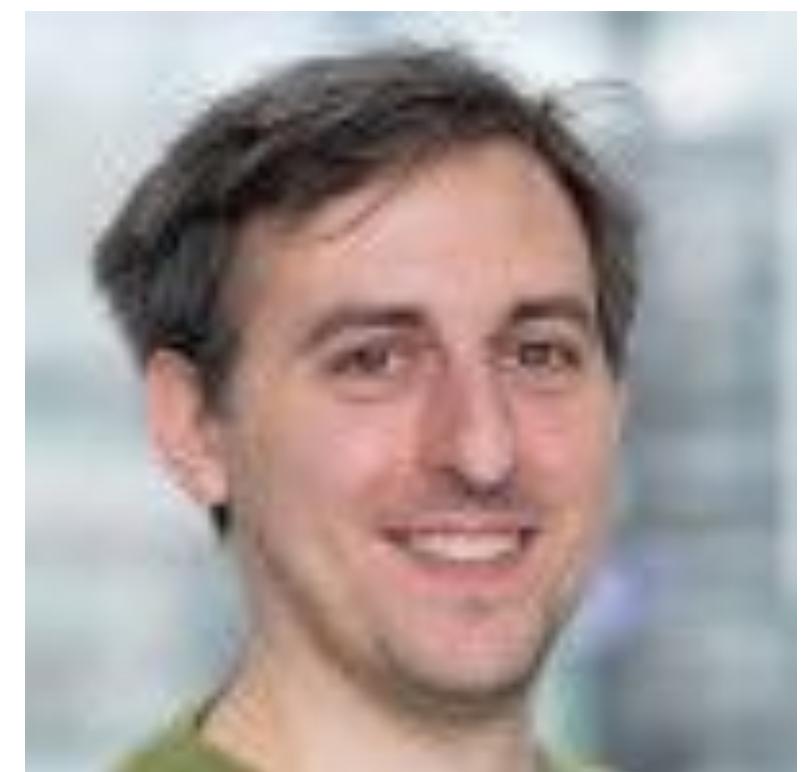
# Tips that past students give you

- “Don’t spend too long on literature in the beginning”
- “Make a plan, especially for time away from campus”
- “Start coding early”
- “Don’t be afraid to try things out”
- “You can do it even if nobody has programmed before!”



# Tips I give you

- “Group meet 2+ times / week, starting this week”
- “Work on the project every week (1-2 tasks/week)”
- “Start with a "shitty first draft" report >1w before deadline”
- “Google/Stackoverflow is your friend”
- “Ask on LearnIT! There are no stupid questions”



# Sources and further materials for today's class

<https://creativecommons.org/>

<https://www.gnu.org/philosophy/free-sw.en.html>

<https://softwareengineering.stackexchange.com/questions/47028/how-could-we-rewrite-the-no-evil-license-to-make-it-free>

<https://data.gov.uk/dataset/road-accidents-safety-data>

<https://gist.github.com/ericmjl/27e50331f24db3e8f957d1fe7bbbe510>

<https://towardsdatascience.com/how-to-create-a-professional-github-data-science-repository-84e9607644a2>

<https://github.com/drivendata/cookiecutter-data-science>

<https://deparkes.co.uk/2016/10/21/anaconda-python-environments/>

<https://www.oreilly.com/library/view/python-for-data/9781449323592/ch04.html>