

# Introduction to the course and final project specifications

Silvio Peroni

[silvio.peroni@unibo.it](mailto:silvio.peroni@unibo.it) – <https://orcid.org/0000-0003-0530-4305> – [@essepuntato](https://www.instagram.com/essepuntato)

Data Science (A.Y. 2021/2022)

Second Cycle Degree in Digital Humanities and Digital Knowledge

Alma Mater Studiorum - Università di Bologna

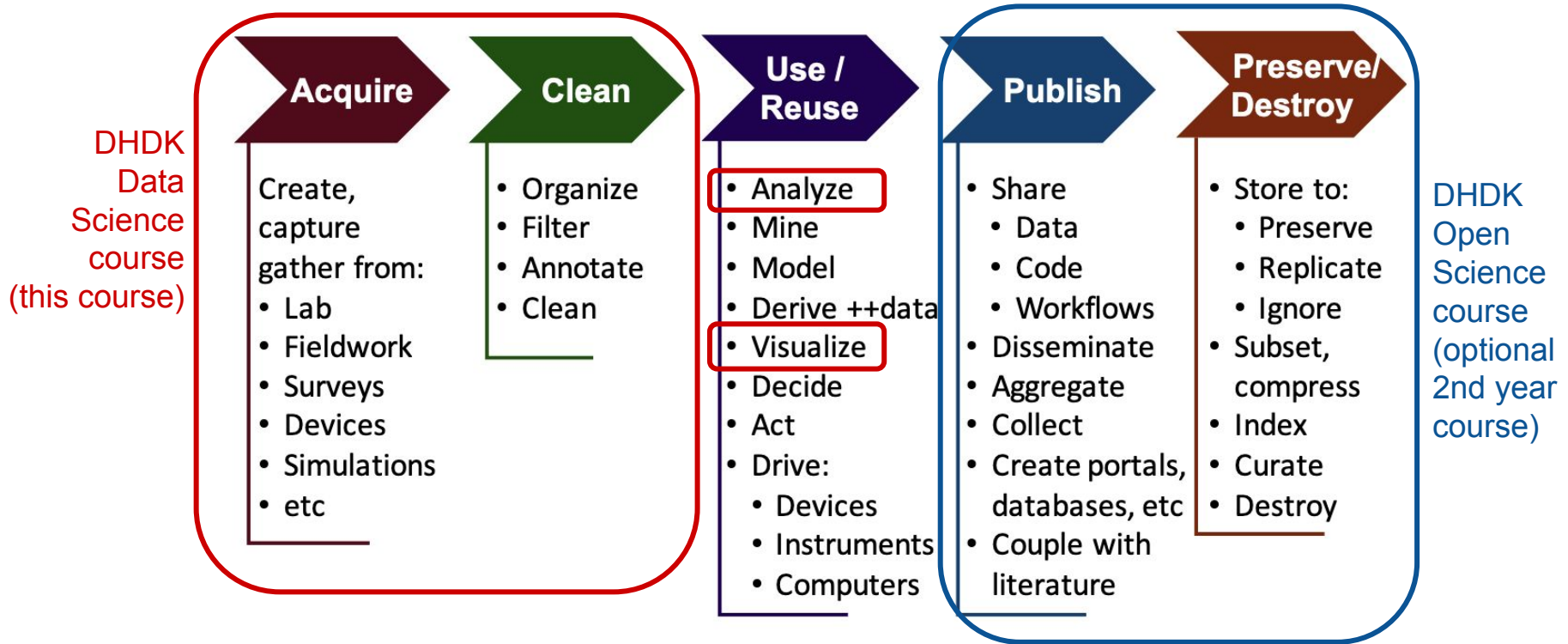


A preliminary introduction to  
the basic tools that  
enable you to start a  
**Data Science**  
project

(and get the final score for  
the integrated course

“Computational Management of Data”)

# Data Science life cycle (and what we will address)



# It is a practical course



## Computer, your best friend



# Rules

Let's see the rules of this course:

- There are **no text books** provided: if you have a doubt, google it
- Material that will be introduced in the course can be found on the [GitHub repository](#) of the course
- If you cannot attend, do not attend (even if attendance is **highly recommended**: you can ask questions)
- At least six exam sessions per academic year
- Max score one can obtain is 32 (out of 30)

# Course material

All the material (including slides) is available in the GitHub repository of the course at <https://github.com/comp-data/2021-2022/>

Have you found a mistake in the material? Please write me an email at [silvio.peroni@unibo.it](mailto:silvio.peroni@unibo.it)

Some (partial) suggested (external) material (for Python and data management with Python):

- Peroni, S. (2020). Computational Thinking and Programming book. <https://comp-think.github.io>
- Tagliaferri, L. (2018). How To Code in Python. ISBN: 978-0999773017.  
<https://www.digitalocean.com/community/books/digitalocean-ebook-how-to-code-in-python>
- Walsh, M. (2021). Introduction to Cultural Analytics & Python.  
<https://melaniewalsh.github.io/Intro-Cultural-Analytics/>
- The Programming Historian. ISSN: 2397-2068. <https://programminghistorian.org/en/>
- Dombrowski, Q. (2020). Jupyter notebooks for digital humanities.  
<https://github.com/quinnanya/dh-jupyter>

# Communications

We use a Slack discussion group for communicating with each other

[Slack](#) is a communication platform offering many [IRC-style](#) features, including persistent chat rooms (channels) organized by topic, private groups, and direct messaging, and can be used in any device (it has been developed for mobile and desktop devices)

Invitation link to join the workspace:

[https://join.slack.com/t/essepuntato-uni/shared\\_invite/zt-11x0yju5e-mIRqjP\\_7dcew~yFDJ4G0vg](https://join.slack.com/t/essepuntato-uni/shared_invite/zt-11x0yju5e-mIRqjP_7dcew~yFDJ4G0vg)

Once you entered in the workspace following the previous link, you will be added to the private channel dedicated to the course (“data-science-21-22”) by me (just poke me if I forgot to add you after one day, OK?)

Please, do a Slack account as soon as possible (e.g. during the break) since I will close the invitation link in at the end of this week – and we will start to use it since Wednesday

# Interacting

We will use the [GitHub repository of the course](#) for a series of activities, such as exercises and raising issues

All questions about coding to ask to either me or your fellows must be asked as new issues

Thus, please, create a new account on GitHub: <https://github.com/join>

Once logged into GitHub, go to the issue page of the course repository, and then feel free to add and respond to issues there

A very introductory guide to GitHub can be find [online](#)



# Course organisation

This introductory lecture + a series of

- (highly-practical) theoretical lectures (7 lectures, 14 hours), where I provide a theoretical introduction about the specific topic – computer not necessary but you can bring it with you of course
- hands-on sessions (7 lectures, 14 hours), where I run a laboratory activity session based on existing tools that enable the experimentation with the topics introduced in the theoretical lectures – computer mandatory

In case you do not have a laptop, there is the possibility to use a remote service using the computers in Aula Affreschi (please contact me in the break if you are in this situation)

# Course calendar

All lectures 12:30 - 14:30, **lectures in red**, **hands-on in blue**

02/02/2022	What is a datum and how it can be represented computationally
04/02/2022	Data formats and methods for storing data in Python
07/02/2022	Introduction to data modelling
09/02/2022	Implementation of data models via Python classes
11/02/2022	Processing and querying the data
14/02/2022	Introduction to Pandas
16/02/2022	Database Management Systems

18/02/2022	Configuring and populating a relational database
21/02/2022	SQL, a query language for relational databases
23/02/2022	Configuring and populating a graph database
25/02/2022	SPARQL, a query language for RDF databases
28/02/2022	Interacting with databases using Pandas
02/03/2022	Describing and visualising data
04/03/2022	Descriptive statistics and graphs about data using Pandas

# Exam

The exam consists of

1. the implementation of a project – yes, you have to write a software, and the specifications will be introduced today
2. an oral colloquium on the project implemented, for assessing the contribution of each student

It will be assigned maximum 16 points for the correctness of the project – I will run a series of tests aiming at assessing all the code developed

The points above are assigned to all the students that have worked to the project

Other 16 points (maximum) will be assigned to each individual student as result of the oral colloquium

# Groups

You have to organise in groups

- 3-4 people (not less, not more)
- You have to choose a name for the group (please, be creative)

Before the next lecture, I will post a link on Slack to subscribe your group

In working on the project, I would strongly suggest to split the work, so as to have a reasonable amount of code to implement

# Registering final grade

As you know, this course is just a module of an Integrated Course (I. C.)

- Computational Management of Data (Integrated Course)
  - Module 1: Computational Thinking and Programming (first semester course, done)
  - Module 2: Data Science (this course!)

You can register only the final grade of the Integrated Course, which is computed as the average of the final scores of the two modules (max. 30L each module)

# Evaluation of the course

The last lectures of the course, you will be asked to fill-up a questionnaire on the organisation of the course and related stuff - it is anonymous, of course

Please, do it carefully and honestly, since it is one of the most important inputs I will have to understand what can be improved in the next year course

This is the very first year for this course, and thus it is crucial to have your feedback in order to understand how to improve it for the next year

# JupyterLab

[JupyterLab](#) is a web-based interactive development environment for notebooks, code, and data

You need to have Python installed to running Jupyter – you can find instruction for installing Python in the [first laboratory lecture](#) of the Computational Thinking and Programming course

To install JupyterLab, you can use the command `pip` as follows:

```
pip install jupyterlab
```

It will be used for all the hands-on sessions, thus be sure to have it available on your computer, and the website includes a very good documentation, while a lot of tutorials are [available online](#)

The project



# Overall goal of the project

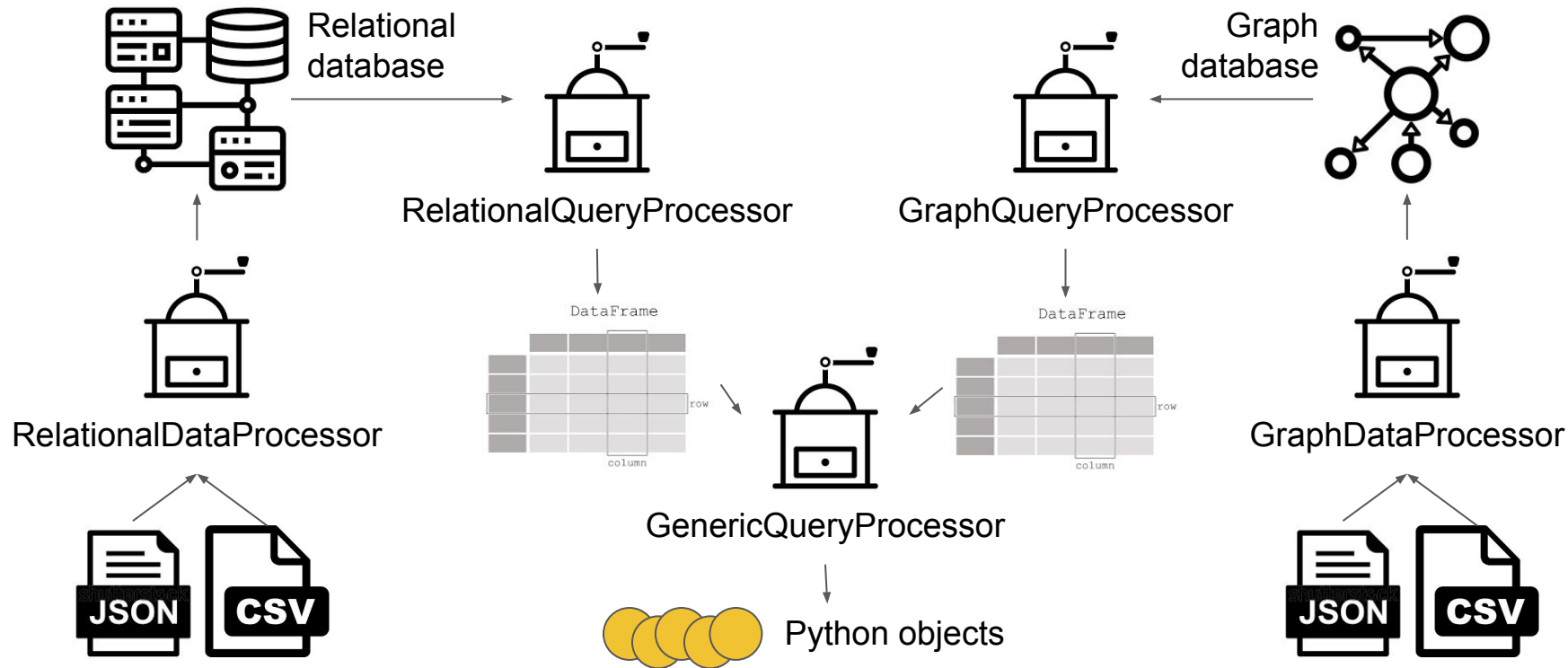
Develop a software that enables one

1. to process data stored in different formats and to upload them into two distinct databases
2. to query these databases simultaneously according to predefined operations

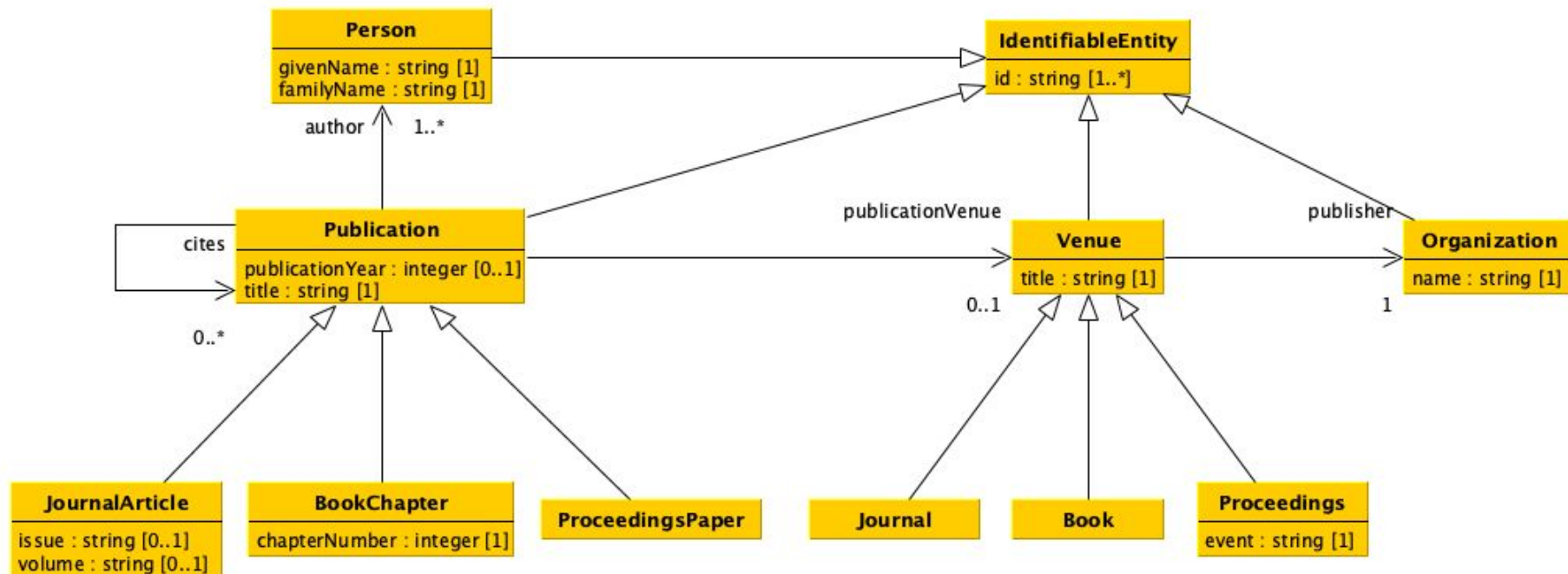
The software must be accompanied by a document (i.e. a Jupyter notebook) describing the data to process (their main characteristics and possible issues) and how the software has been organised (name of the files, where have been defined the various Python classes, etc.)

Information about the project can be found in the [GitHub repository](#) of the course (and will be updated in due course)

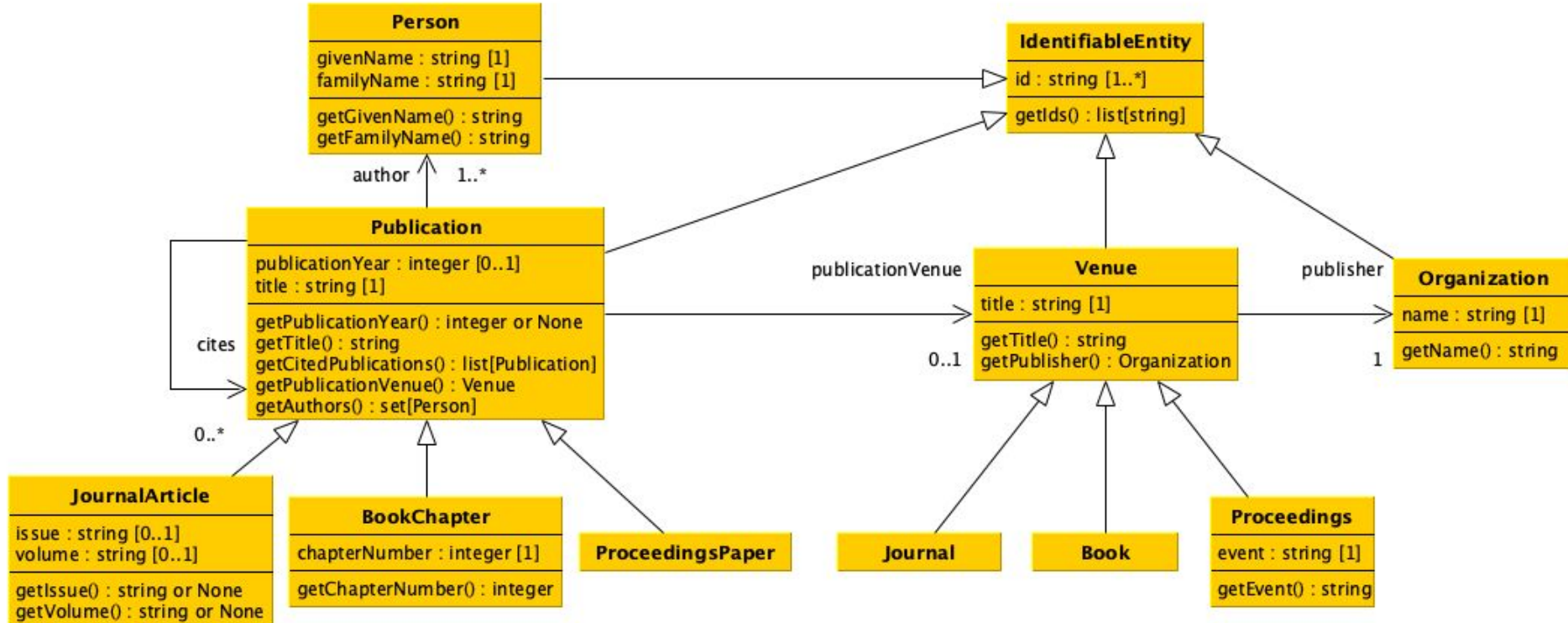
# Workflow



# Data model



# Software: UML diagram (1/2)



# Software: UML diagram (2/2)



# Submission of the project

You have to provide all Python files implementing your project, by sharing them in some way (e.g. via OneDrive)

You have to send all the files one week before the exam session you want to take

The schedule of the first four exam sessions are as follows:

- 16 May 2022, 11:00-16:00
- 20 June 2022, 11:00-16:00
- 15 July 2022, 11:00-16:00
- 5 September 2022, 11:00-16:00

# Suggestion: use the test-driven development

You already know what is the test-driven development (TDD) – to refresh your memory, you can look at [this slide](#) (and scream...)

I strongly suggest you to systematically adopt this development technique for developing your code

In addition, there is at least a Python library available which has been entirely developed to facilitate the creation of tests, i.e. [unittest](#)

Online, you can find several documents describing how to create tests in Python, such as [one at the Real Python website](#)

# End

## Introduction to the course and final project specifications

Silvio Peroni

[silvio.peroni@unibo.it](mailto:silvio.peroni@unibo.it) – <https://orcid.org/0000-0003-0530-4305> – [@essepuntato](https://twitter.com/essepuntato)

Data Science (A.Y. 2021/2022)

Second Cycle Degree in Digital Humanities and Digital Knowledge

Alma Mater Studiorum - Università di Bologna

