Research Ideas and Outcomes 3: e12032
doi: 10.3897/rio.3.e12032

Small Grant Proposal

Data Intensive Science

Nicolas Schmelling [‡]

[‡] Heinrich Heine University, Düsseldorf, Germany

Corresponding author: Nicolas Schmelling (nicolasschmelling@gmail.com)

Reviewable v1

Received: 31 Jan 2017 | Published: 01 Feb 2017

Citation: Schmelling N (2017) Data Intensive Science. Research Ideas and Outcomes 3: e12032.
https://doi.org/10.3897/rio.3.e12032

Abstract

A proposal to create a full-semester zero-entry level course about the responsible handling of research data and the associated analyses, storage, and sharing. The syllabus will comprise open science workflows, the creation of data management plans, as well as the addressing issues about reproducibility and data sharing in science. The course and all its materials will be licensed under CC-BY or if possible under CC-0.

Keywords

OER, Learning, Open Data, Data Management, Open Science, Reproducibility, Data Analysis, Python, R

Problem Statement

The scientific community is facing a reproducibility crisis. The majority of research findings are not or only partly reproducible. One reason for this is the insufficient description of the statistical analysis and the incorrect storage of underlying research data and their inaccessibility. Here, we propose to address the problem by teaching correct data storage and sharing in open access repositories and the comprehensive description of data production and data analysis pipelines. Even though these tools are already available to the scientific community a lack of awareness about these tools as well as knowledge about the correct use hinders the progress of eliminating this crisis.

Proposed Project

The proposed course "Data Intensive Science" consists of two main parts, a series of workshops and an accompanying lecture series. This course provides learning materials and personal instructions to students and researches of all areas of science. The resulting course materials will be published under CC-BY or CC-0 license for reuse and will be structured in a way that each workshop can be taught as a single workshop. The project will reuse as much open and well tested materials as possible and only refine parts or create new ones if no suitable alternative is available in order to build upon universally applicable materials. The goal of the course is to teach students and researches a responsible way of working with research data and introduce them to open science workflows to accelerate scientific progress.

Lecture

The lecture will introduce the participants to concepts and workflows of open science, the principles of data and code storage and sharing, as well as the concepts of statistics thinking and analyses. The lecture will be held in a blended learning environment. Students are asked to prepare for the lecture with provided videos or open access reading materials, which will be accessible through a wiki hosted on GitHub, the Open Science Framework, or Wikiversity. The lecture will then be either a group discussion or flipped classroom, where students will step into the role of the lecturer. Using mobile apps as electronic voting systems throughout the lecture will help to identify misunderstood areas, where additional materials are needed. Online quizzes will help the students to recapitulate the lecture content and will also provide a resource for the lecturer about problems and misunderstandings. Depending on the topic, it is intended to invite external experts as lecturers that can expand the materials with some practically relevant information.

Workshop

The workshops will provide hands-on materials and instructions about data and code sharing as well as storage, reproducibility of statistical analyses as well as introductions to programming and data analysis in Python and R. The workshop series of this course starts with an introduction to version control using git and GitHub, as well as the platforms Zenodo, figshare, and the Open Science Framework. In the following, the students learn how to analyze their data with Python and R and how to create reproducible analyses using Jupyter Notebooks and RMarkdown notebooks. These notebooks allow for a fully reproducible pipeline and an interactive reuse of the analysis code. At the end of the practical part, the students are asked to perform data analyses on a given data set and to produce a reproducible pipeline. The practical part will be supported by screencasts to repeat the materials. Furthermore, in order to promote self-paced learning, additional materials are provided to the students. This will encourage students who cope well with the basic material to go beyond and explore more advanced topics. On the other side, this opens up space for the instructor to work more closely with the students that are having problems with some of the materials.

Implementation

For the implementation of the project, the schedule was split up into four sections. The **first phase** of the project implementation is concerned with the **collection of existing open materials** that fits into the course syllabus. Afterwards, in the **second phase**, these **materials** will be **refined** and structured such that they fit best into the course. If adequate material is missing, **new materials**, which fill the gaps, will be **created**. In the **third phase, video tutorials and screencasts** for the workshops and **online quizzes**, which support the lecture series, will be **created**. In the last, **fourth phase**, all these materials and additional tools will be combined into one **learning environment**.

Timeline

A period of ten months is expected for the preparation and collection of the materials for this course. The breakdown of stages and roles of different actors is outlined in Table 1.

Table 1.

Tentative Schedule for Implementation

Duration	Lead Instructor	Assistants
4 Months	Collection of existing open materials	
2 Months	Refining materials and creating additional materials	Video Tutorials
1 Month	Questions Pools	Video Tutorials
2 Months	Video Tutorials & Wiki	
1 Month	Setting up of the learning environment	Wiki

Budget

The details of the budget are outlined in Table 2. It is to be noted that the salaries of the two assistants are based on a normal salary for an undergraduate assistant working five hours on average per week.

Table 2.

Preliminary Budget

Item	Description	Cost (in EUR)
Personnel	Assistant 1	**2,684.89** (268.5 per month)

	Assistant 2	**2,684.89** (268.5 per month)
Total Cost		**5,369.78**

Acknowledgements

Special thanks go to Stephanie Dittebrand and Daniel Mietchen for their stellar support in creating this proposal.

Funding program

The proposal was submitted to the eLearning Support Fund at Heinrich Heine University Düsseldorf in January 2017.