The Journal of Open Source Software

Hacking Health Covid-19: An Online Data Science Course-Project

Thierry Warin[1, 2]

1 Professor, HEC Montréal (Canada) **2** Principal Investigator, CIRANO (Montreal, Canada)

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

During the Covid-19 first quarantine period, lots of universities closed their facilities and offered online courses. Ours was subject to the same predicament. In this context, our data science laboratory has decided to organize a course-project about Covid-19. This initiative just wanted to be a positive use of our time and resources to potentially help communities around us or around the world. The course-project is about teaching data science using the R language. The question is to define the perimeter of data science: from unstructured data to mapping, through Natural Language Processing (NLP) and predictive modelling.

The course-project was called "Hacking Health Covid-19" for it had similarities with hackathons and also one main difference: it is similar since it requires a strong involvement on a short period of time to create a viable contribution, and it is also different in the sense that it spans over a 6-week period.

In this course-project, the goal for students was to create a module based on the R language that contributes scientifically to the data conversation about Covid-19. This course-project was offered at the graduate level.

Statement of Need

This course-project is about teaching data science, and mostly how data science can be useful to students. It is therefore, in essence, a course about the definition of data science. Computer science has taken a prominent role in the advancement of knowledge creation when it comes to data. This is not new. Already in 1966, Peter Naur coined the term "datalogy" based on his understanding of the role of computing power in data analysis (Naur, 1966). Data science is broader than computer science and multidisciplinary. If we were to propose the three data science skillsets, we could propose: database management, statistics and machine learning, and distributed and parallel systems (Jones, 2015). We could also add other domains. In short, data science still seems to be more of an umbrella of disciplines (Meng, 2019) around the value extraction from data, "data science is an umbrella term to describe the entire complex and multistep processes used to extract value from data" (Wing, 2019).

We are fortunate to have access to a platform dedicated to research in data science. This course-project is a great way to teach the R language in the context of data science, while offering a hands-on perspective on a major current issue: Covid-19. The question is to know how data scientists and aspiring data scientists can participate to the world efforts against Covid-19, for instance by providing a database of scientific references in epidemiology, or by finding new ways to inform policymakers and develop more efficient pulic responses in these tough times.

We are starting to see a lot of online initiatives about Covid-19, and notably about data

visualizations. Even if they are numerous, they tend to be very close to each other. And although visualizations are important for our own course-project, we can augment them with some data science perspectives, for instance in terms of sentiment analysis, or predictive modelling.

In terms of development principles, by providing a scientific perspective, we mean to propose contributions that can be validated by others, and shedding a slightly different light to the pandemic. A Github repo has been created to host the code and serves also as a data warehouse.

Learning objectives

The following learning objectives are inspired by other learning modules from the Journal of Open Source Education (Arribas-Bel, 2019).

Upon completion, students are able to:

- demonstrate advanced concepts in data science,
- use the tools to import, manaipulate and analyse data about Covid-19,
- critically evaluate the suitability of a specific technique,
- and select a number of analytical techniques in R to turn data into relevant information.

Moreover, open science principles were presented. Our main objective in this course-project is to provide a scientific perspective to the world conversation and to open up our contributions so that other people can leverage them. We taught and promoted open science principles, such as this simple equation: Open data + open code + open research = reproducible research. Indeed, for this initiative, we wanted to be as open as possible in order to be at the highest level possible of reproducibility. The need to provide reliable sources of information as well as validated data to form better decisions has become crucial. By building up a workflow of integrated tools such as data, code and methods, our students can learn how to disseminate their results more widely in a reproducible spirit.

Content and Instructional Design

The course developed over a 6 week period from April to May 2020. Formal online meetings were organized on a weekly basis for teams to present their ideas and their achievements. To participate, students needed to form teams of 3 people. In order to cover the whole range from web apps development to statistics and machine learning, the r language was used. Coding knowledge in R is thus advisable. Now, team members can have different levels. Teams will have access to online peer learning sessions on our communication platform and on our virtual campus. They will be able to train themselves on the fundamentals of R.

The main portal for the course-project is www.warin.ca/covid-19-hackathon/. On this portal, students could find all the supported material for their contribution to this course-project. For this course-project to be implementable, we realized we needed to provide (1) technological tools and (2) tailored content.

Access to cutting-edge technology. The Lab's team decided to devote some resources to open its platform, both in terms of technology and human resources, to contribute to the world conversation about Covid-19. The goal is to make a contribution to the world conversation with science at its core. This course-project relied on our laboratory's technological platform. Students were given access to our servers for the computing power, as well as our online communication platform based on an open source software

(rocket.Chat), they were provided access to servers equipped with RStudio Server, as well as a server with a Nextcloud installation to share their documents and work collaboratively. As aforementioned, as much as we promoted open science, we also relied exclusively on open source software.

Access to teaching material in the R language. A whole curriculum - called "coding school" - was also created to train them in teams in the R language (see Table 1).

Table 1. Coding School Topics

Coding School		
R Nanocourse 1. Reproducible Research	R Nanocourse 6. simple Linear Regression	PDF Text Extraction
R Nanocourse 2. Data Import and Graphics	R Nanocourse 7. Multiple Linear Regression	Structural Topic Model
R Nanocourse 3. Data Wrangling	Mapping, Spatial Analysis and Econometrics in R	API: newsAPI
R Nanocourse 4. dynamic Documents	Flexdashboards with RMarkdown	Collecting Twitter Data
R Nanocourse 5. Descriptive Statistics	Dash and Plotly in R	API: shapeR

To simplify the process, we also created two packages for this project: (1) a first package about scientific references in epidemiology and coronaviruses more specifically (around 32,000 references and 22 metadata), called EpiBibR, and (2) a second package, called shapeR, to facilitate the use of shapefiles, should they want to do a mapping exercise.

We proposed 12 examples of modules covering the wider perimeter of data science (see Table 2).

Table 2. Potential Modules

Potential Modules		
Data Visualization	Data Warehouse	News Collection & Analysis
Predictive Modelling	Social Media Collection	Mapping
Bibliometrics	Topic Modelling	Covid-19 & International Flows
Covid-19 & Finance	Covid-19 & Public Policies	Covid-19 & Ethics

In each module, students would find a series of examples as well as the code from our virtual campus to replicate the principles behind these examples (see here for an example).

An interesting feature of the course-project is that it was a jump in - jump out system where teams could enter in the competition at any point of time during this period and could submit their module whenever they wanted during this period and at the latest on May 18th, 2020. It offered some necessary agility to help transition from a course in a physcial classroom to an online course.

Experience of Use

The materials in this computational resource have been used, updated and maintained for over five years.

As aforementioned, we made our research platform available. It's a combination of cutting-edge tools that we use in our research. This platform includes a communication platform,

a virtual campus where they will have access to a lot of the code we are using in research, and to a computing platform where they will be able to integrate their R code.

We have also developed a unique learning environment to support this Hacking Health. We developed a lot of resources to help the students build their module (data, examples of potential modules, etc.) but also many tutorials to help them in the design and coding of their module.

Acknowledgements

The author would like to thank the Center for Interuniversity Research and Analysis of Organizations (CIRANO, Montreal) for its support, as well as Thibault Senegas, Marine Leroi and Martin Paquette. The usual caveats apply.

References

Arribas-Bel, D. (2019). A course on Geographic Data Science. *Journal of Open Source Education*, *2*(16), 42. https://doi.org/10.21105/jose.00042

Jones, T. (2015). The Identity of Statistics in Data Science. Retrieved April 30, 2020, from Amstat News website: https://magazine.amstat.org/blog/2015/11/01/statnews2015/

Meng, X.-L. (2019). Data Science: An Artificial Ecosystem. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.ba20f892

Naur, P. (1966). The science of datalogy (Forum). *Comm. ACM*, *9*(7), 485. Retrieved from https://scholar.google.com/scholar?hl=fr&as_sdt=0%2C5&q=+Naur%2C+P.+The+science+of+datalogy+%28Forum%29.+Comm.+ACM+9%2C+7+%28July+1966%29&btnG=

Wing, J. M. (2019). The Data Life Cycle. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.e26845b4