The Journal of Open Source Software

Hacking Health Covid-19: An Online Data Science Course

Thierry Warin[1, 2]

1 Professor, HEC Montreal (Montreal, 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. In this context, I decided to transform my on-site course on data science into an online course about data science with Covid-19 as a case study. This initiative just wanted to be a positive use of our time and resources to help communities around us or the world potentially. The course is about teaching data science using the R language while using Covid-19 as a case study. 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 was called "Hacking Health Covid-19" for it had similarities with hackathons and also one main difference: it was similar to traditional hackathons since it required a strong involvement to create a viable contribution, and it was also different in the sense that it spanned over six weeks.

In this course, 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 was offered at the graduate level. A thorough set of learning elements was created (e.g. R modules, R data packages, a computing platform, a communication platform) for this course to provide a comprehensive online environment for students.

Statement of Need

This course is about teaching data science leveraging Covid-19 data, 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).

During this course, students were provided access to a platform dedicated to research in data science. This course was a great way to teach the R language in the data science context while offering a hands-on perspective on a significant current issue: Covid-19. The question was to know how data scientists and aspiring data scientists could participate in the world's efforts against Covid-19, for instance, by providing a database of scientific references in epidemiology or finding new ways to inform policymakers and develop more efficient public 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. Furthermore, although visualizations were an essential element for our course, we could augment them with some data science perspectives, such as sentiment analysis, or predictive modelling.

In terms of development principles, by providing a scientific perspective, we meant to propose contributions that could be validated by others and shedding a slightly different light on the pandemic. A Github repo had been created to host the code and also served 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 can:

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

Moreover, open science principles were presented. Our main objective in this course was to provide a scientific perspective to the world conversation and open up our contributions so that other people could leverage them. We taught and promoted open science principles, such as this simple equation: Open data + open code + open research = reproducible research. For this initiative, we wanted to be as open as possible to be at the highest level possible of reproducibility. The need to provide reliable sources of information and 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 could learn how to disseminate their results more widely in a reproducible spirit.

Content and Instructional Design

The course developed over six weeks from April to May 2020. Formal online meetings were organized weekly for teams to present their ideas and their achievements. To participate, students needed to form teams of 3 people. 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 could have different levels. Teams had access to online peer learning sessions on our communication platform and our virtual campus. They were able to train themselves on the fundamentals of R on top of receiving synchronous training.

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

Access to cutting-edge technology. Students were granted access to our research-oriented data science platform to contribute to the conversation about Covid-19. The goal was to contribute to the conversation with science at its core. This course 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 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 promote 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 60,000 references and 22 metadata), called EpiBibR (https://github.com/warint/EpiBibR), and (2) a second package, called shapeR (https://github.com/warint/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 exciting feature of the course is that it was a jump-in jump-out system where teams could enter the competition at any point in time during this period and could submit their module whenever they wanted during this period and at the latest May 18th, 2020. It offered some necessary agility to transition from a course in a physical 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 is 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 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 many 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, Antoine, Debien, Anne-Sophie Gill, Catherine Feuillet and Martin Paquette at the SKEMA Global Lab in AI (Montreal, Canada). 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