ENGAGING EVERYONE WITH OPEN DATA SCIENCE

Kimmo Vehkalahti
University of Helsinki, Finland
kimmo.vehkalahti@helsinki.fi

Teaching of statistics should focus more on practical data science, with a special emphasis on data wrangling: Preparing the data for the analyses, looking at the data via clever visualizations, and learning the principles and practices of open science and reproducible research. The statistics curriculum should be updated and the term "data science" used as a synonym to statistics. In all possible fields, there is a huge need to have more data scientists. To engage everyone with "open data science" (open data, open science, and data science), we have created a new course, where students from all levels and fields work together and share their ideas with openly available data sets and freely available state-of-the-art software tools, such as RStudio, R Markdown, and GitHub. The new course has been quite successful in engaging extremely heterogeneous groups of students to challenge themselves to a "next level" by learning new skills of open data science.

INTRODUCTION

The ongoing "Data revolution" (Kitchin, 2014) sets more requirements for the researchers on all fields of science. One could say (without exaggerating too much) that *we should all be data scientists*. Indeed, the term "data science" is a good synonym to statistics. This is also a question of a brand or an image. Although statisticians themselves do know that statistics is a collection of important skills that almost everyone needs, the typical image of statistics (and statisticians) is far from that. In addition, as everyone knows, the word "statistics" itself is a bit problematic (at least in English), as it refers both to the field or discipline but also to the numerical summaries of data sets, which might easily give a completely wrong idea of the field.

Interestingly, in Finnish language (spoken only by less than 6 million people in the world), the discipline is called "*tilastotiede*", where "*tiede*" means science and "*tilasto*" may refer either to a statistic or – to data. Hence, "data science" is something that we could have used for a long time (at least in Finnish), but of course it is not so easy to cha(lle)nge traditional terms. Perhaps statistics (as a term) worked well decades ago, but the world has changed rapidly, and we should react by updating old terms. Data are everywhere, and science is important, so "data science" makes up a simple, compact, and useful term.

Luckily, computer science started using the term "data science" quite recently and it has spread everywhere, also to increasing number of applied fields and even to everyday usage. Statisticians should also use the term to refer to their field and consider using "data scientist" to describe themselves as professionals of working with data. The choice of terms is not meaningless, and it will have consequences, for example, in the choices related to studies and careers, because the image difference is so clear: "statistics" sounds boring, while "data science" sounds cool.

In all possible fields, there is a huge need to have more data scientists. To engage everyone with "open data science" (open data, open science, and data science), we have created a new course, where students from all levels and fields work together and share their ideas with openly available data sets and free software tools, such as RStudio, R Markdown, and GitHub. The new course has been successful in engaging extremely heterogeneous groups of students to challenge themselves to a next level by learning new state-of-the-art skills of open data science.

OPEN DATA SCIENCE

We have developed a new course, entitled, "Introduction to Open Data Science", to respond to a serious need we have observed all around, but especially in the social sciences and humanities. The amount of available data is continuously increasing, and instead of (or at least in addition of) traditional methods, skills like algorithmic thinking, programming, and sharing data and code are becoming necessary. The primary targets of our course are doctoral students of those fields, but it is open and accessible for anyone interested. Indeed, students from all levels (doctoral / master's / bachelor's) and all possible fields (e.g., humanities, social, behavioral, agricultural, medical, or natural sciences) have participated on the three courses organized in 2017 and 2018.

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. iase-web.org/Conference_Proceedings.php

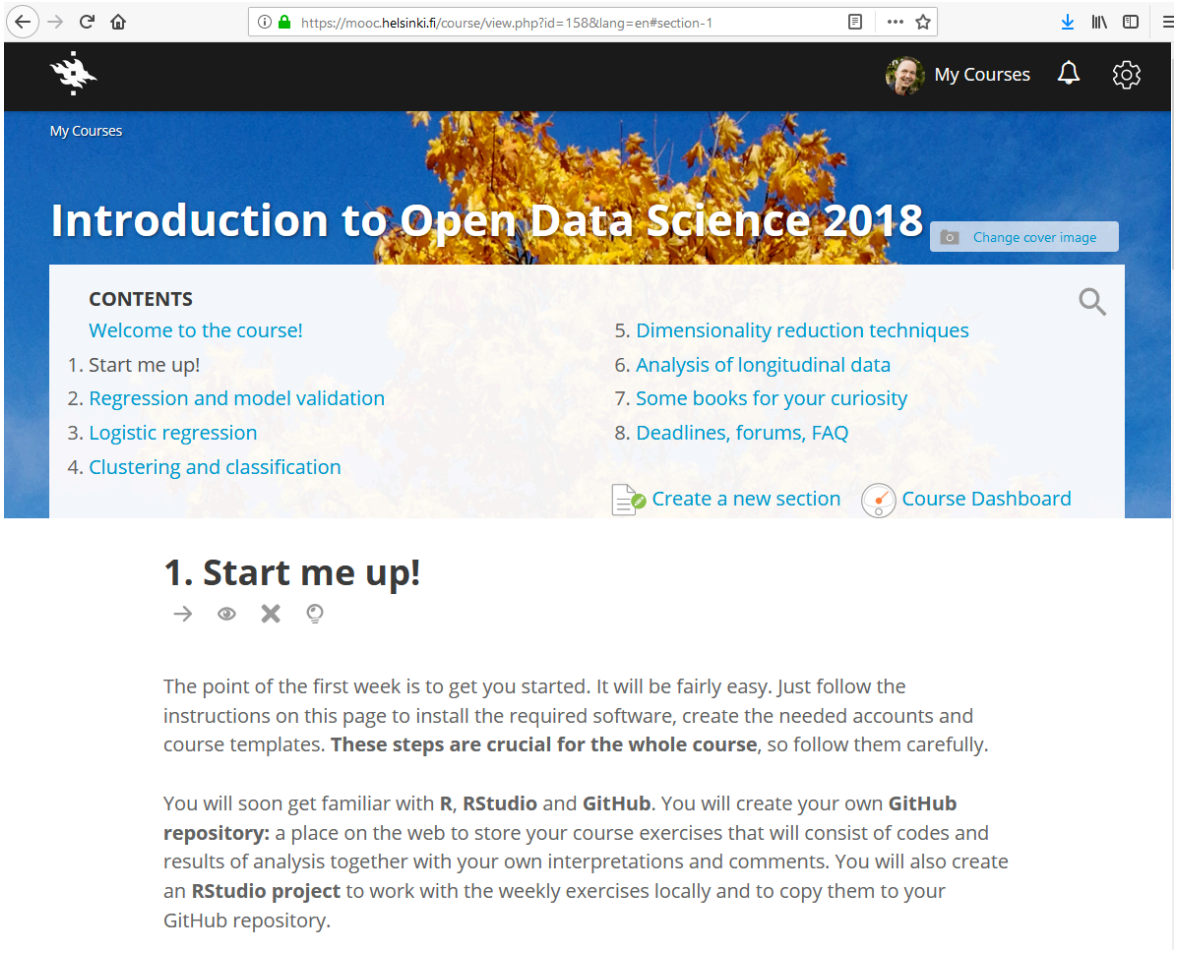

Figure 1. View of the course on the Moodlerooms platform

Already in the planning phase we wrote some motivational sentences that would describe the characteristics of the course: "*We must discover the patterns hidden behind numbers in matrices and arrays. We are not afraid of coding, recoding, programming, or modeling. We want to visualize, analyze, interpret, understand, and communicate. These are the core themes of Open Data Science (Open Data – Open Science – Data Science). And this course is THE course for learning these skills.*" The general learning objectives were stated as follows: "*After completing this course you will understand the principles and advantages of using open research tools with open data and understand the possibilities of reproducible research. You will know how to use RStudio, R Markdown, and GitHub for these tasks, and know how to learn more of these open software tools. You will also know how to apply certain statistical methods of data science, that is, data-driven statistics.*"

In more detail, the course consists of seven weeks, where the first week is quite crucial. It is used for introducing the course in general, but also presenting and installing the software tools to be learned and the learning platform (see Figure 1) with its peer-review tools to be used during the subsequent weeks. We expect the participants to bring their own laptop computers, install the required software, and create an account on GitHub (www.github.com) in the beginning of the course. There is only one face-to-face meeting each week, a combination of a short lecture and a workshop, where the lecturer and the teaching assistants give advice with any technical details encountered by the students. However, the course can be studied completely from distance. There is a strict weekly schedule for the submissions and the peer-reviews of the assignments. There are no exams, as the grades are based on the points given by peers and checked by the teachers.

After the first week, there are five quite dense working weeks guiding the students in various techniques of data wrangling and statistical analysis with selected, openly available data sets. The first theme is called "Regression and model validation". It is an easy start from basic data management and wrangling, making scatter plots and other graphs, creating simple and multiple regression models, and

doing some explorative regression diagnostics. Although there are no pre-requirements, many students have faced regression analysis in some form before. It is also a smooth way to get to grips with the R language (R Core Team, 2019), RStudio (www.rstudio.com), R Markdown (Xie, Allaire & Grolemund, 2018) and GitHub.

The second theme is "Logistic regression", which provides with a brief introduction in the world of generalized linear models in a form of typical regression models for binary outcomes. It includes training and testing a predictive model with the means of cross-validation. Again, a few data wrangling techniques and suitable visualizations are learned.

The third theme underlines that instead of a systematic introduction to statistics or statistical methods, the point of this course is to practice working with data and the software tools. The name of the theme is "Clustering and classification", and it focuses on k-means clustering and discriminant analysis. The next theme continues with these core methods of data science under the title "Dimensionality reduction techniques", and it includes the basics of principal components analysis and multiple correspondence analysis (Greenacre & Blasius, 2006).

The fifth and the last theme of the course, namely, "Analysis of longitudinal data", returns to the topic areas of statistical modeling. After the data wrangling part that introduces the wide and

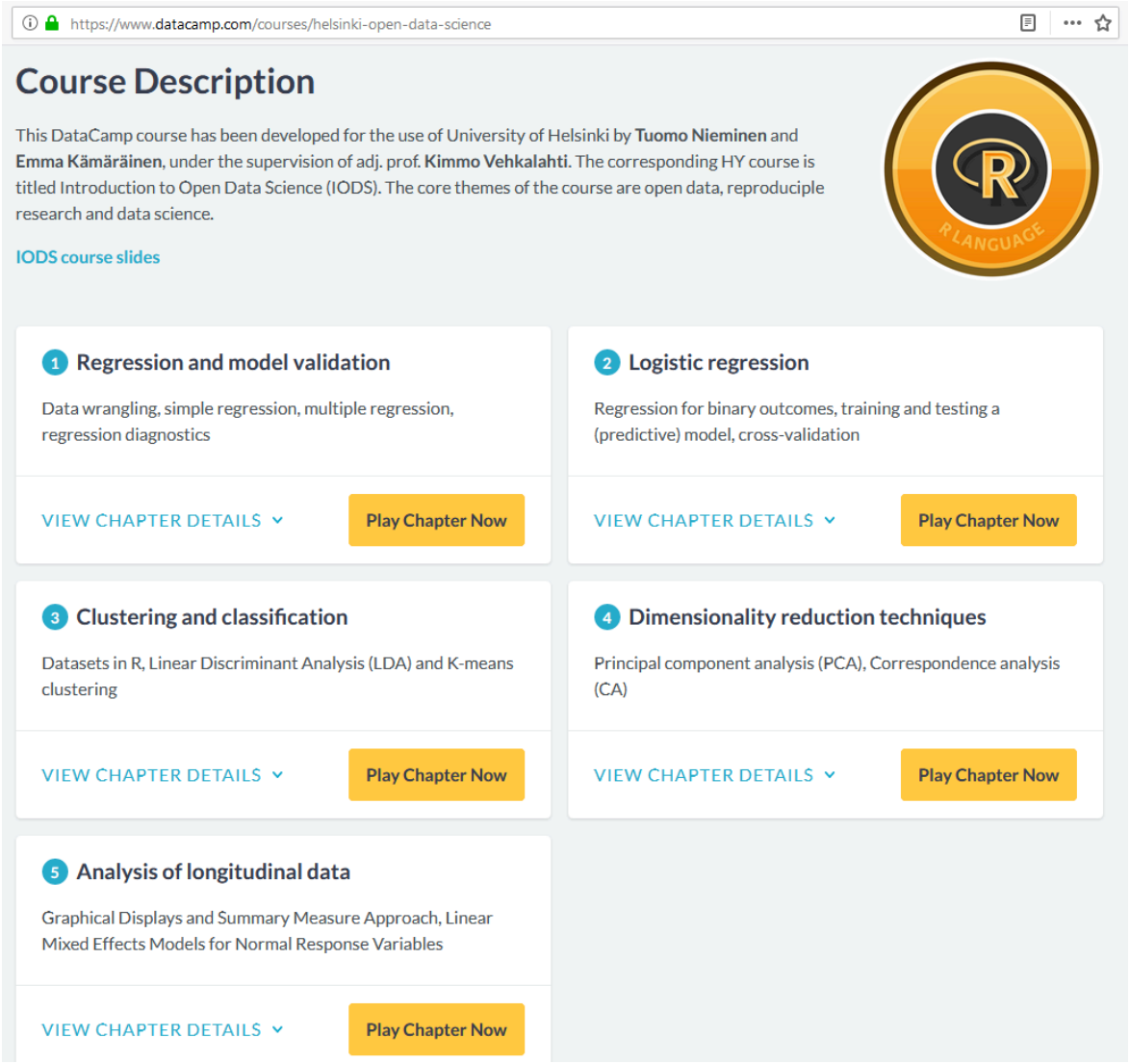

Figure 2. View of the course on the DataCamp platform

long forms of data sets and how to convert them from a form to another, it goes through typical graphical displays and simple methods of summary measures. As a final topic of the course, it then

gives the basics of the linear mixed effects models in the context of longitudinal research settings that are becoming more and more typical in practice.

Originally, in the place of the last theme, there was a final assignment that gathered together the course topics in a small research report, but it was later replaced with a new working week, because it became too tedious and time-consuming for the teachers to assess and grade the individual reports. With the weekly peer-review that is now applied six times during the course (every student evaluates the assignments of three other students, selected automatically by the learning platform) makes the course better scalable, as in the future, the number of the students may increase significantly, while the teacher resources will remain about the same. So far, the number of students has been quite moderate (about 100 to 150), but now it should be possible to offer it for many times larger audiences without serious organizing problems.

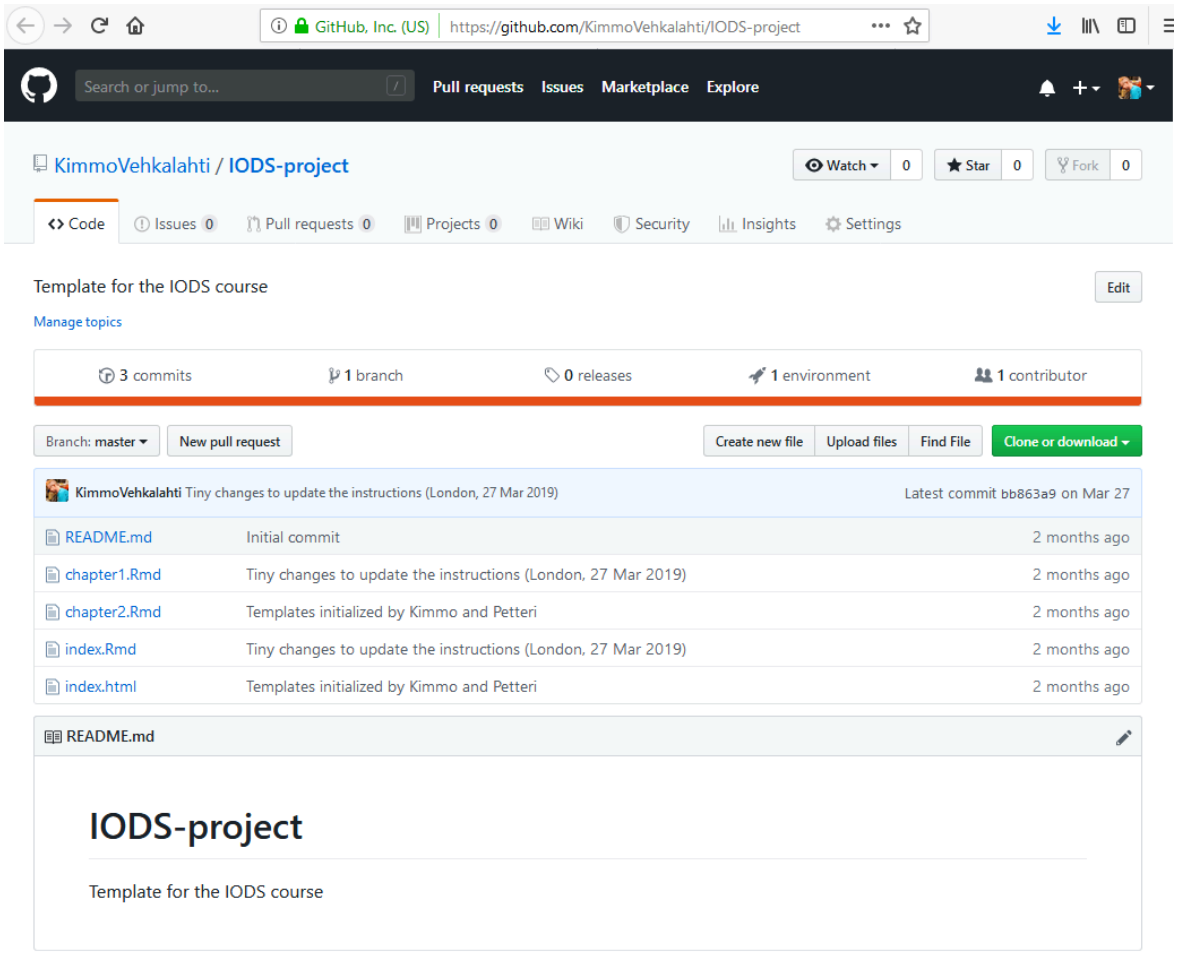

Figure 3. View of the course template on GitHub

Supporting the activities on the primary learning platform of the course (see Figure 1) there is another platform on the DataCamp (www.datacamp.com), where we have implemented a sub-course called “Helsinki Open Data Science” (see Figure 2). Its role is to provide a set (a chapter in DataCamp vocabulary) of interactive exercises related to each of the five themes explained above and teach the basics of the R programming. In principle, the DataCamp gives points or scores of completing those exercises, but those scores are not used as a part of the course grade (originally, they used to be until the technical link between DataCamp and Moodle environments was disconnected by DataCamp). Instead, the course points and grades are earned by peer-reviewing the assignments that follow the flow of the DataCamp exercises, use the data sets introduced in the exercises, and are completed by the students with their own computers using RStudio software. Typically, the students will first learn the necessary R codes on DataCamp and then copy-paste and modify the codes in RStudio, producing their report in an R Markdown template that is initiated on the first week of the course (see Figure 3).

In addition to the online materials on various platforms, including short videos done by the teaching assistants and the free-form online discussions (Q&A etc.), the course utilizes a recently published textbook of applied multivariate data analysis by Vehkalahti & Everitt (2019) as the background material supporting the learning of the statistical methods. The book has a GitHub repository that includes all the data sets of the book and the R codes for its examples and exercises. They provide additional material for the students to continue learning after the course.

RESULTS

The course has been a success story from the very beginning. This is nicely reflected in the following excerpts from anonymous student feedback:

"*I really enjoyed this course, to be honest this is the best course that I had in Helsinki. Combining both DataCamp and Rstudio exercise was amazing idea, it helped me alot. Even though I have been using R since couple of years but during this course I learned more sophisticated ways of programming.*"

"*The course was really interesting and hands-on approach worked well. Datacamp exercises were well organised. Need for this kind of applied statistical (data science) courses where you're needed to clean your dataset and then use correct statistical methods is in high demand. You can get a feel that you're learning something actually useful for real life. Learning Github has been really huge benefit.*"

"*Now I feel that this was the best course, in which I have ever been participating, because: 1. The time schedule was very flexible, and I had an opportunity to work according my time schedule. 2. I like very much the interactive DataCamp exercises part, where I have got an idea how to start, and only after that to continue with the RStudio exercises. 3. The video lectures helped me a lot because this suits a lot to my way of studying: first to read and watch theoretical lectures, and after that continuing with the practical exercises part. 4. I like the idea for peer reviews, because in this way we had an opportunity to compare (to some extension) what we have done with the work of the others.*"

"*I must admit that I attended the course to see how you've done the arrangements, what technologies you've chosen and to get some ideas how to run similar courses on my own. Learn from the best is the word ☺ I've used R since 2000. I still learned new things on the course!*"

"*First of all I want to thank you all about this course which has been the funniest and most interesting ever. This was my first touch to R, GitHub and Slack. I never thought that I would get this excited about something, but I did. I noticed that the R environment is an endless world and its not as difficult as I thought at first. I will definitely continue to learn codes and statistics.*"

DISCUSSION

There is a huge need to have more data scientists. Teaching of statistics should focus more on data science, with a special emphasis on data wrangling. The statistics curriculum should be updated and the term "data science" used as a synonym to statistics. Our new course gives an excellent example of how to engage students to learn skills of reproducible, open data science.

REFERENCES

Greenacre, M., & Blasius, J., eds. (2006). *Multiple Correspondence Analysis and Related Methods*. Boca Raton, FL: Chapman and Hall/CRC.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. London: SAGE.

R Core Team (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. *https://www.R-project.org/*

Vehkalahti, K., & Everitt, B. S. (2019). *Multivariate Analysis for the Behavioral Sciences*, Second edition. Boca Raton, FL: Chapman and Hall/CRC. *https://github.com/KimmoVehkalahti/MABS*

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Boca Raton, FL: Chapman and Hall/CRC.