

Welcome - Read on GitHub

I'm glad that you decided to take Developing Data Products, part of the Data Science Specialization from Johns Hopkins Biostatistics!

A data product is the production output from a statistical analysis. Data products automate complex analysis tasks or use technology to expand the utility of a data informed model, algorithm or inference. This course covers the basics of creating data products using Shiny, R packages, and interactive graphics. This course focuses on the statistical fundamentals of creating a data product that can be used to tell a story about data to a mass audience.

You will learn how to communicate using statistics and statistical products. Emphasis will be paid to communicating uncertainty in statistical results. You will learn how to create simple Shiny web applications and R packages for their data products. In addition, we'll cover reproducible presentations and interactive graphics.

We believe that the key word in Data Science is "science". Our specialization is focused on providing you with three things: (1) an introduction to the key ideas behind working with data in a scientific way that will produce new and reproducible insight, (2) an introduction to the tools that will allow you to execute on a data analytic strategy, from raw data in a database to a completed report with interactive graphics, and (3) on giving you plenty of hands on practice so you can learn the techniques for yourself. This course represents the final cog in a data science application, creating an end-usable data product.

We are excited about the opportunity to attempt to scale Data Science education. We intend for the courses to be self-contained, fast-paced, and interactive.

Some Basics

A couple of first week housekeeping items. First, make sure that you've had R Programming and the Data Scientist's Toolbox. Reproducible Research would be helpful, but is not mandatory. At a minimum you must know: very basic git, basic R and very basic knitr.

An important aspect of this class is to peruse the materials in the github repository. All of the most up to date material can be found here: https://github.com/DataScienceSpecialization/Developing_Data_Products

You should clone this repository as your first step in this class and make sure to fetch updates periodically. (Please send pull requests too!) It is one of the most essential components of the Specialization that you start to use Git frequently. We're practicing what we preach as well by using the tools in the series to create the series, especially git.

You can clone the whole repo with (http)

- git clone https://github.com/DataScienceSpecialization/Developing_Data_Products.git
- or (ssh)
- git clone git@github.com:DataScienceSpecialization/Developing_Data_Products.git

The lectures are in the index.Rmd lecture files. In this class, we'll cover how to create these sorts of slides. You will see all of the R code to recreate the lectures. Going through the R code is the best way to familiarize yourself with the lecture materials.

The lecture material for this class is largely front-loaded. This is because the latter time of the class is devoted to developing your data application. Thus the class should be doable in about a 1 month's time or maybe less. Though make sure you're keeping up with the classes at the beginning so that you have some space in your schedule later on for app development!

If you'd like to keep up with the instructors I'm @bcaffo on twitter, Roger is @rdpeng and Jeff is @jtleek. The Department of Biostat here is @jhubiostat.

https://datasciencespecialization.github.io/Developing_Data_Products/welcome.html