Welcome to Week 1 of Getting and Cleaning Data! This course focuses on preparing you for collecting and cleaning data for downstream analysis and sharing.

One of the major components of a data scientist's job is to collect and clean data. Whether at a small organization or a major enterprise, the first step in using data is getting, cleaning and understanding the data. In this course, we focus on R packages and a few outside tools that can be used to collect data from a variety of sources, from Excel files to databases like MySQL. We will also cover a variety of formats including JSON, XML, and flat files (.csv, .txt).

The emphasis of this course is on creating tidy data sets that can be used in downstream analyses. Once you have mastered the material in this course you will be ready to learn about the techniques for exploring, analyzing, and summarizing data offered through our courses track or other Statistics, Data Science, or Machine Learning MOOCs.

One important note is that as part of this class you will be required to use the Github account you set up in The Data Scientist's Toolbox. Github is a tool for collaborative code sharing and editing. During this course and other courses in the track you will be submitting links to files you publicly place in your Github account as part of peer evaluation. If you are concerned about preserving your anonymity you should set up an anonymous Github account and be careful not to include any information you do not want made available to peer evaluators.

The recommended background for this course are the courses The Data Scientist's Toolbox and R programming. It will be challenging to take this class concurrently with those classes. You may have to read ahead in previous classes to get the relevant background for this class. For a complete set of course dependencies in the Data Science Specialization please see the course dependency chart.

Please see the course syllabus for information about the Quizzes, the Course Project, and grading. Don't forget to say hi on the discussion forums. The community developed around these courses is one of the best places to learn and the best things about taking a MOOC!

Jeff Leek and the Data Science Track Team