October 10-11, 2018
Presented at the Bureau of Labor Statistics Conference Center

SIMON MUNZERT
Hertie School of Governance, Berlin, Germany

COURSE OBJECTIVES
By the end of the course, students will…

- have an overview of state-of-the-art research that draws on web-based data collection,
- have a basic knowledge of web technologies,
- be able to assess the feasibility of conducting scraping projects in diverse settings,
- be able to scrape information from static and dynamic websites as well as web APIs using R, and
- be able to tackle current research questions with original data in their own work.

WHO SHOULD ATTEND
Individuals in government, business, academia, and non-profit organizations who conduct data analytic work using the statistical software R. This course provides a condensed overview of web technologies and techniques to collect data from the web in an automated way. **Students are expected to be basically familiar with the statistical software R.** Besides base R, knowledge about the "tidyverse" packages, in particular, dplyr, magrittr, and stringr are of help. If you are familiar with R but have no experience in working with these packages, the best place to learn them is the primary reading "R for Data Science".

SUGGESTED READING

Simon Munzert, Christian Rubba, Peter Meißner, and Dominic Nyhuis, 2015: *Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining*. Chichester: John Wiley & Sons.

Wickham, Hadley, and Garrett Grolemund, 2016: *R for Data Science*. Sebastopol, CA: O'Reilly. (available for free at http://r4ds.had.co.nz/)

THE INSTRUCTOR

SIMON MUNZERT is Lecturer in Political Data Science at the Hertie School of Governance. He received his doctoral degree in Political Science from the University of Konstanz. His research interests include measuring and forecasting public opinion, political representation, and the use of new media in society. He is author of the textbook *Automated Data Collection with R* (Wiley). His research has been published in *Political Analysis, Journal of the Royal Statistical Society Series A, Political Science Research and Methods,* and *Social Science Computer Review.* Furthermore, he is an enthusiastic user of the statistical software R.

COURSE MATERIALS AND MEALS
Registrants will be provided with a course lecture notebook. JPSM group continental breakfasts, lunches and refreshments are included in the course fee.

DAILY CHECK-IN
Registrants must check-in with the JPSM onsite each day of the course.

WEDNESDAY: OCTOBER 11, 2018	
7:30--8:30	Registrant Check-in and Continental Breakfast
8:30--10:30	Introduction: Setup and First Step in Scraping the Web Using R
10:30--10:45	Morning Break
10:45--12:15	Scraping with Regular Expressions
12:15--1:15	Lunch
1:15--2:45	Scraping Static Webpages
2:45--3:00	Afternoon Break
3:00--4:30	Advanced Scraping of Static Webpages
4:30	Adjourn
THURSDAY: OCTOBER 12, 2018	
7:30--8:30	Registrant Check-in and Continental Breakfast
8:30--10:30	Scraping Dynamic Webpages
10:30--10:40	Morning Break
10:45--12:15	Tapping APIs
12:15--1:15	Lunch
1:15--2:45	Legal and Ethical Issues in Web Scraping
2:24—3:00	Afternoon Break
3:00—4:30	Scraping Workflow and Tricks of The Trade
4:30	Adjourn