Course 1: A Practical Introduction to Data Gathering Methods with a focus on Web Scraping and Data Access via APIs

Format/Length: 2 half days (4 hours per session)

Brief Description: This short course will offer a very practical introduction to data gathering geared at social scientists and survey researchers. This course begins with an overview of web scraping discussing some basic technical jargon, types of web data and various methods for scraping. The course also includes a discussion and illustration of Application Programming Interfaces (APIs) use for gathering web data when they are available. Some websites are designed to be easily accessible by web crawlers or scraping algorithms while others require much more advanced, custom programming. And some web data can be accessed using an API that is provided by the website. In this course we will illustrate how participants can discern these differences as well as presenting several motivating examples of the various ways web scraped data can be used throughout a study's lifecycle from design to calibration to analysis. We provide an extensive introduction to a suite of freeware programs that allow virtually syntax free, but customizable, web scraping capabilities. We contrast this type of gathered data access to APIs for some websites like Zillow or Twitter and discuss pros and cons of using web scraping or APIs to gather this type of web data. The course concludes with specific focus on the import.io tool where we demonstrate its capabilities and provide several, hands-on practical examples for participants to begin scraping several websites of increasing complexity. We will also illustrate API calls in R for Zillow, the Census and others as time permits.

Outline:

- Overview of Web Scraping
  - **What is Web Scraping?**
  - **Why should survey and social science researchers care about it?**
  - **Legal Issues with Web Scraping…**
  - **Examples of how modern social science researchers are and have used web scraping.**
- A Detailed Introduction of Web Scraping
  - **A brief Overview of How websites work**
  - **Common Web Scraping Tasks for Education Researchers**
  - **Web Scraping Cycle**
  - **Components of Chrome's Page Inspection Tool**
  - **Technologies for Modifying/Cleaning Scraped Data**
- Web Scraping Software/Tools Available
  - **Google Chrome Extension**
  - **Excel**

- **Import.io**
- **R and Python**
- **Web scraping Resources**

- A Brief Overview of Web Scraping with Import.io
- Detailed Examples of Web Scraping
  - **Scraping a Basic Table from Wikipedia or other Similar Source**
  - **Scraping Zillow and School Webpages**
  - **Scraping Lists that span Multiple pages**
  - **Scraping More Advanced Webpages including Internet Resources**
    - **Using XPath and RegExes**
- Discussion and illustration of APIs for Zillow, Census and Other Sources
- Hands-On Web Scraping Lab Session Including 4 Specific Examples (2-2.5 hours)