

Data Acquisition and Management

The background of the slide features a close-up of a hand holding a black pen, pointing towards a digital screen. The screen displays a complex financial chart, specifically a candlestick chart with red and white bars, overlaid with various blue and red trend lines and moving averages. The overall image has a blue tint.

Web Data

Scraping and APIs

Overview

Finding the Data We Want

Acquiring the Data

Processing the Data for Analysis

Storing the Data

web scraping

html

less structured . . .

web APIs

**xml
json**

. . . more structured

Web APIs	Web Scraping
<ul style="list-style-type: none"> • Safer to Work With (APIs provide implicit contracts between providers and consumers) 	<ul style="list-style-type: none"> • More Dangerous to Work With (Any changes to structure of web page may break accessing routine)
<ul style="list-style-type: none"> • Easier to Code – requires ability to work with XML and/or json. Advanced skills: DOM (document object model), web services; web protocols (https, ftp) 	<ul style="list-style-type: none"> • Harder to Code – requires additional skills, including knowledge of HTML (start with: tables, paragraphs, forms), CSS selectors, XPATH expressions, and regular expressions; web crawling; browser-specific developer tools
<ul style="list-style-type: none"> • Data more likely to have been scrubbed 	<ul style="list-style-type: none"> • Data less likely to have been scrubbed
	<ul style="list-style-type: none"> • Web Pages may contain data that's not included in API
	<ul style="list-style-type: none"> • Web Page data may be more current or more frequently updated
	<ul style="list-style-type: none"> • Web APIs may not exist or may be restricted (legal and financial issues here)

Best-of-Class R Packages for Web Data

RCurl

XML

rjson

selectr

ROAuth

httr

rvest

stringr