Data Acquisition and Management

Web Data

Scraping and APIs

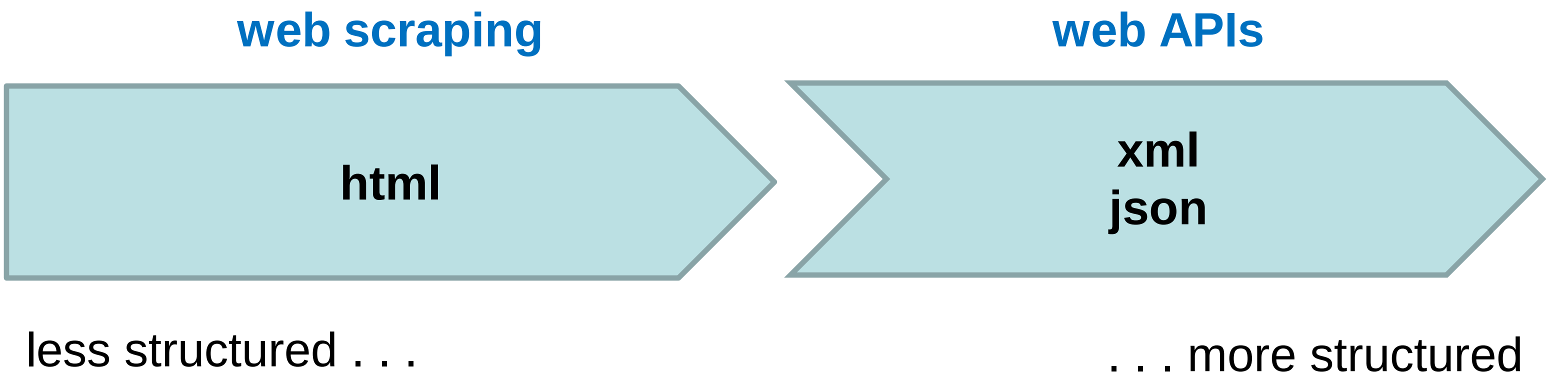
Overview

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
• **Safer** to Work With (APIs provide implicit contracts between providers and consumers)	• More Dangerous to Work With (Any changes to structure of web page may break accessing routine)
• **Easier** to Code – requires ability to work with XML and/or json. Advanced skills: DOM (document object model), web services; web protocols (https, ftp)	• 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
• Data more likely to have been scrubbed	• Data less likely to have been scrubbed
	• Web Pages may contain data that's not included in API
	• Web Page data may be more current or more frequently updated
	• 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