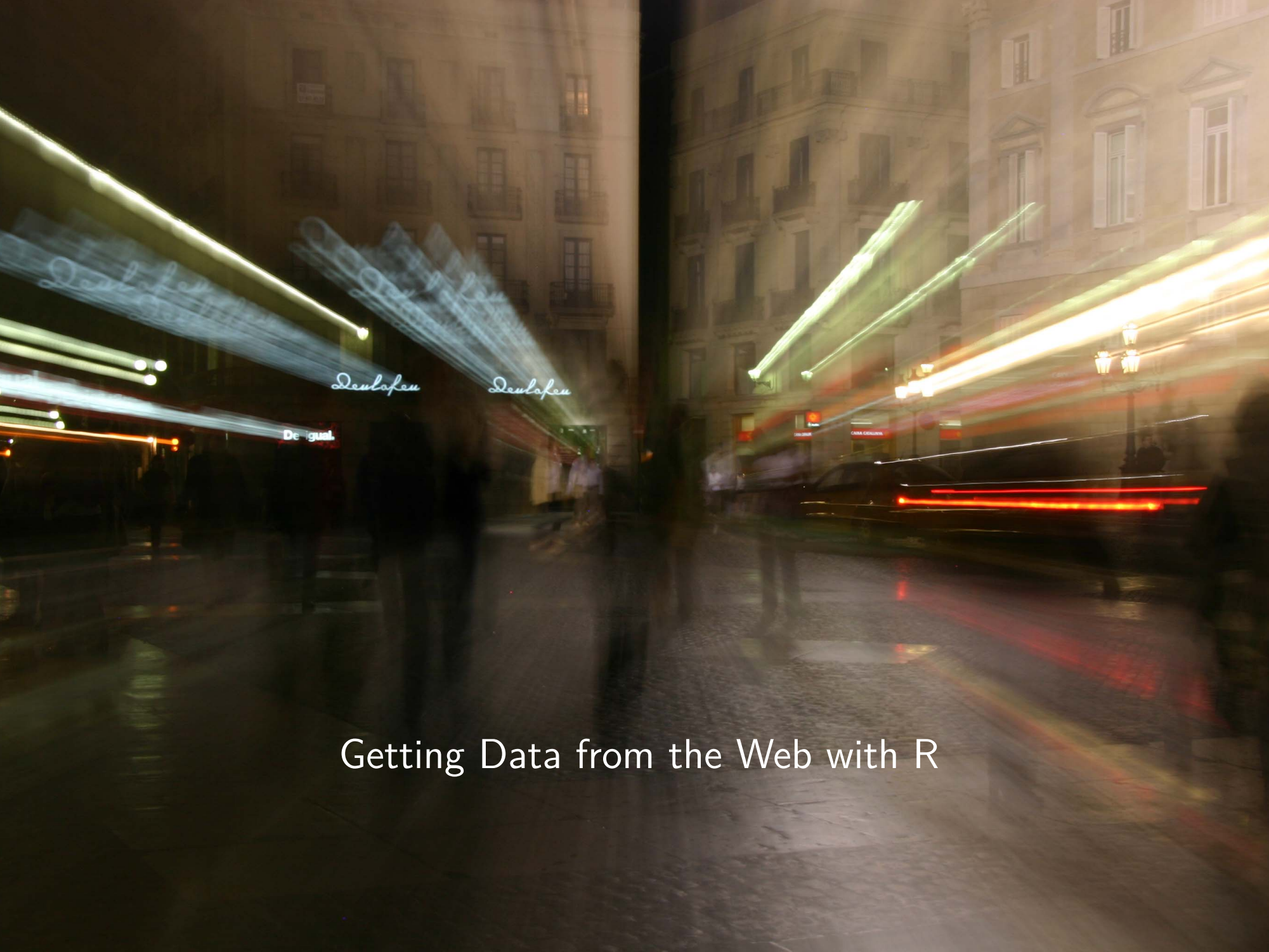
Getting Data from the Web with R

Part 1: Introduction

Gaston **S**anchez

April-May 2014


Getting Data from the Web with R

Readme

License:

You are free to:

Share — copy and redistribute the material

Adapt — rebuild and transform the material

Under the following conditions:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.

NonCommercial — You may not use this work for commercial purposes.

Share Alike — If you remix, transform, or build upon this work, you must distribute your contributions under the same license to this one.

Lectures Menu

Slide Decks

1. **Introduction**
2. Reading files from the Web
3. Basics of XML and HTML
4. Parsing XML / HTML content
5. Handling JSON data
6. HTTP Basics and the RCurl Package
7. Getting data via Web Forms
8. Getting data via Web APIs

About these lectures

Goal

My goal is **to give you an introduction** to some of the tools in R for getting data from the Web.

I don't pretend to cover everything nor going very deep. I just want to show you an overview of various Web Data scenarios you can handle with R.

Preliminaries

Requirements

Must have:

- Some experience working with R
- Some knowledge of HTML
- An insatiable curiosity for learning new things

Nice to have:

- Knowledge about data storage formats
- Some programming experience
- Knowledge on how the Web works

Software

You'll need:

- R (preferably the last version)
  http://cran.r-project.org/
- RStudio (highly recommended)
  https://www.rstudio.com/
- Text Editor
  (eg vim, emacs, TextWrangler, notepad, sublime text)
- Web Browser
  (eg Chrome, Safari, Firefox, Internet Explorer, Opera)
- and a good internet connection!

In my case ...

Software I used for these slides:

- R version 3.1.0 (2014-04-10) – "Spring Dance"
- Platform: x86_64-apple-darwin10.8.0 (64-bit)
- IDE: RStudio Version 0.98.501
- Text Editor: TextWrangler
- Web Browser: Google Chrome Version 34.0.1847.131
- Operating System: OS-X Version 10.8.5

Resources

Some R Books

- XML and Web Technologies for Data Sciences with R
  by Deb Nolan and Duncan Temple Lang
- Introduction to Data Technologies
  by Duncan Murdoch
- Data Manipulation with R
  by Phil Spector
- more references in each slide deck

Resources

Web Scraping with R

- Web scraping for the humanities and social sciences (by Rolf Fredheim and Aiora Zabala)

  http://quantifyingmemory.blogspot.co.uk/2014/02/web-scraping-basics.html

- Web Scraping with R (by Xian Nan)

  http://cos.name/wp-content/uploads/2013/05/Web-Scraping-with-R-XiaoNan.pdf

- R-bloggers posts on *Web Scraping*

  http://www.r-bloggers.com/?s=web+scraping

Some R Packages

Package	Description
`RCurl`	R interface to the `libcurl` library for making general HTTP requests
`RHTMLForms`	Tools to process Web/HTML forms
`XML`	Tools for parsing XML and HTML documents and working with structured data from the Web
`RJSONIO`	Functions for handling JSON data
`jsonlite`	Functions for handling JSON data
`rjson`	Functions for handling JSON data
`ROAuth`	Interface for authentication via OAuth 1.0
`SSOAP`	Use SOAP protocol to retrieve data

CRAN Task View: *Web Technologies and Services*

`http://cran.r-project.org/web/views/WebTechnologies.html`


The Web

VIP Questions

Very Important Preliminary Questions

The Data that you want:

1. Where is it located?
2. How accessible is it?
3. What is its structure / format?

VIP Questions

Location of Data

- Do you know the location (URL) beforehand?

  Or do you have to figure it out?
- Is it in one single specific place?
  (eg one HTML table, one file in the Web)
- Is it in one website but spread across several pages?
  (eg several HTML tables at different pages)
- Is it spread across several websites?
  (eg multiple pieces of information in various sites)
- Is it in one or several databases?

VIP Questions

Accessibility of Data

- Do you have free direct immediate access to data?
- Do you need to fill a Web Form?
- Do you need to use a Web API?
- Do you require username, password, authentication?
- Do you need to use a specifc transfer protocol?
- Do you need to use a specifc type/method of request?

VIP Questions

Format / Structure of Data

- Is it plain text?
- Is it in tabular (spreadsheet-like) form?
- Is it in HTML?
- Is it in some XML-dialect?
- Is it in JSON format?
- Other formats: binary, images, maps, etc?

Glossary

Some Acronyms

- **WWW** World Wide Web
- **W3C** World Wide Web Consortium
- **URL** Uniform Resource Locator
- **HTTP** HyperText Transfer Protocol
- **XML** Extensible Markup Language
- **HTML** HyperText Markup Language
- **JSON** JavaScript Object Notation