

LAS 6292: AUTOMATED DATA EXTRACTION

updated: 2021-07-09

Pre-Class Preparation (Instructor):

Send in an email to students:

- content of any pre-class emails.

Bring to Class:

- Snacks
- Flip charts and markers
- Dry write markers
- Tent cards for student names

Objectives and Competencies:

- OCR

Pre-class Preparation (Students):

Readings

Online Lectures:

- if any link here

Class Outline

Topic 1 Overview: Optical Character Recognition (OCR) (10 min)

1. OCR: google, website, R
2. Data from Published Figs
3. Data Packages Govt Data via API
4. text analysis

1.Video Primer: [What is OCR?](#)

2. Different OCR Tools - text and data from pdfs into csv,txt, etc.

- Google Drive - Video Primer: [OCR with Google Drive](#)
 - [pdf to excel](#)
 - [PDFTables](#) will convert PDF to .csv, and has an [API](#) so you can do your conversions in bulk with R. You can do ~25 pages free; large numbers are reasonably priced.
 - R tools: `pdfreader` and `tabulizer`
 - [written tutorial 1](#)
 - [written tutorial 2](#)
 - [written tutorial 3](#)
 - [Video Tutorial 1](#)
 - [Video Tutorial 2](#)
 - Detailed [Blog Post / Tutorial](#)
 - [Mathpix Snip](#) digitizes handwritten or printed text, and copies outputs to the clipboard that can be pasted into LaTeX editors like Overleaf, Markdown editors like Typora, Microsoft Word, and more.
3. Extracting Tables from images using R package `magick`.
- Detailed [Blog Post / Tutorial](#)
4. Extracting Data from Published Figures
- Ankit Rohagni's [Web Plot Digitizer](#)
 - WPD [Video Tutorial](#)
 - WPD Tutorial [Blog Post](#)
 - Alternative 1: Extracting data from images using R package `magick`
 - Alternative 2: [GetData](#) extracts data automatically from scanned images (~\$30).
 - Alternative 3: r package `digitize` will extract data from scatterplots within the R environment. [This article](#) will walk you through the process.
5. Text Mining
- [Text Mining with R](#) by Julia Silge and David Robinson
 - [gutenbergr](#): Download and Process Public Domain Works from [Project Gutenberg](#). Tutorial can be found [here](#)
6. Government & Related Data
- [Data.gov](#) (the open data portal of the US Government) and [Using Data.gov APIs in R](#)
 - the [rOpengov Project](#)
 - [Open Fiscal Data Package](#)
 - [educationdata](#): Retrieve data from the Urban Institute's Education Data API as a `data.frame` for easy analysis. See [also here](#)
 - a [huge list](#) of data sources for social scientists available with R tools *accessing [World bank Data](#) with R
7. Web Scraping

- Library Carpentry Lesson Webscraping <https://librarycarpentry.org/lc-webscraping/>
- Start Here: [Introduction to webscraping](#)
- Video: [Scraping WebData in R with rvest](#)
- Video: [Practical Introduction to Web Scraping using R](#)
- Very nice [written tutorial](#)...
-and another one, this time from the [UC Business Analytics R Programming Guide](#)
- [scraping HTML text](#) and [scraping HTML tables](#)
- [SelectorGadget](#) is useful to id CSS selectors

Topic 2 Overview: Topic (10 min)

Image is from: Pereira, Thales Augusto Zamberlan. (2018). Poor Man's Crop? Slavery in Brazilian Cotton Regions (1800-1850). Estudos Econômicos (São Paulo), 48(4), 623-655. <https://doi.org/10.1590/0101-41614843tzp>

Intro text

Breakout & Return Results

(we did this in-class together as a live-coding-type-exercise): using * [Web Plot Digitizer](#), we extracted data df from several figures that differ in quality and content.

Breakout (15 min): topic of breakout

Returning results & Take-home message (35 min) summary of results

2. **Take-home message:** message.

additional text

anything before the break? (10 min)

if so, describe here

Break (10 min)

Free Time

There are 30 min remaining that can be used to —

Tools & Resources

Collecting Social Media Data

1. Scraping [Twitter Data with R](#) or with [Tweetsets](#)

Cell Phone Data

1. Exploratory analyses [Part 1](#) and [Part 2](#)

Image Analysis

1. Pennekamp, F. and Schtickzelle, N. (2013), Implementing image analysis in laboratory-based experimental systems for ecology and evolution: a hands-on guide. *Methods Ecol Evol*, 4: 483-492. <https://doi.org/10.1111/2041-210X.12036>
2. How to build your own image recognition app with R! [Part 1](#) and [Part 2](#)

Wearable Devices

1. Izmailova, E.S., Wagner, J.A. and Perakslis, E.D. (2018), Wearable Devices in Clinical Trials: Hype and Hypothesis. *Clin. Pharmacol. Ther.*, 104: 42-52. <https://doi.org/10.1002/cpt.966>
2. Loncar-Turukalo T, Zdravevski E, Machado da Silva J, Chouvarda I, Trajkovic V. Literature on Wearable Technology for Connected Health: Scoping Review of Research Trends, Advances, and Barriers *J Med Internet Res* 2019;21(9):e14017 [doi: 10.2196/14017](https://doi.org/10.2196/14017)

Automated Data Collection

1. [Automated Data Collection \(ADC\) Basics](#)

Education Data

1. `edbuildr`: import [EdBuild's master dataset](#) of school district finance, student demographics, and community economic indicators for every school district in the United States.

Sources

1. Tafti A.P., Baghaie A., Assefi M., Arabnia H.R., Yu Z., Peissig P. (2016) OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym. In: Bebis G. et al. (eds) *Advances in Visual Computing*. ISVC 2016. *Lecture Notes in Computer Science*, vol 10072. Springer, Cham. https://doi.org/10.1007/978-3-319-50835-1_66
2. Correia, R.A., Ladle, R., Jarić, I., Malhado, A.C.M., Mittermeier, J.C., Roll, U., Soriano-Redondo, A., Veríssimo, D., Fink, C., Hausmann, A., Guedes-Santos, J., Vardi, R. and Di Minin, E. (2021), Digital data sources and methods for conservation culturomics. *Conservation Biology*, 35: 398-411. <https://doi.org/10.1111/cobi.13706>

[Building R and Stata packages for the Education Data Portal](#)