



(<https://www.ire.org/>)

SEARCH

Enter search terms...

SEARCH

NICAR

NICAR ([HTTPS://WWW.IRE.ORG/NICAR](https://www.ire.org/nicar))

ABOUT ([HTTPS://WWW.IRE.ORG/NICAR/ABOUT](https://www.ire.org/nicar/about))

DATABASE LIBRARY ([HTTPS://WWW.IRE.ORG/NICAR/DATABASE-LIBRARY](https://www.ire.org/nicar/database-library))

Practice Datasets (<https://www.ire.org/nicar/database-library/practice-datasets>)

Terms and Conditions (<https://www.ire.org/nicar/database-library/terms-and-conditions>)


Stanford Review Of Tools (<https://www.ire.org/nicar/database-library/stanford-review-of-tools>)

ANALYSIS JOBS ([HTTPS://WWW.IRE.ORG/NICAR/ANALYSIS-JOBS](https://www.ire.org/nicar/analysis-jobs))

TOOL LIBRARY ([HTTPS://WWW.IRE.ORG/NICAR/TOOL-LIBRARY](https://www.ire.org/nicar/tool-library))

 LOGIN ([HTTPS://WWW.IRE.ORG/WP-LOGIN.PHP](https://www.ire.org/wp-login.php))

Not a member? Join (</membersonly/join/register>) ([Why?](#)
(</membership>).

Subscribe to the IRE Radio Podcast  (<https://itunes.apple.com/us/podcast/ire-radio-podcast/id900544465?mt=2>)

NICAR ([HTTPS://WWW.IRE.ORG/NICAR](https://www.ire.org/nicar)) » DATABASE LIBRARY ([HTTPS://WWW.IRE.ORG/NICAR/DATABASE-LIBRARY](https://www.ire.org/nicar/database-library)) » STANFORD REVIEW OF TOOLS ([HTTPS://WWW.IRE.ORG/NICAR/DATABASE-LIBRARY/STANFORD-REVIEW-OF-TOOLS](https://www.ire.org/nicar/database-library/stanford-review-of-tools))

Written by Tobin Asher, compiled by Arun Chaganty, Nathaniel Okun, Deger Turan, Cayman Simpson and Max Bodoia

Data Wrangling and Analysis

BASIC TOOLS

Excel and Google Sheets

- Description: A user-friendly tool for data wrangling and basic data analysis.
- Website: <https://drive.google.com> (<https://drive.google.com/>)
- When to Use: Quick and basic data analysis.
- Case Study: For small tabular datasets, some groups first used a spreadsheet to look at the data. Google Sheets allowed groups to share access to a dataset.
- Ease of Use: Beginner
- Cost: Excel is paid, Google is free
- Works Well With: Trifacta

- **Strengths:** They are effective for exploring and doing summaries (i.e. summing columns) for small datasets. They make it easy to quickly skim spreadsheets to become familiar with what data is available. They require no programming experience.
- **Weaknesses:** They are generally only effective for relatively small datasets (less than 10MB). Best used for basic data analysis.

SQL

- **Description:** Structured Query Language (SQL) is a programming language designed to query and analyze data in relational databases.
- **Website:** Various software programs based in SQL, such as MySQL, PostgreSQL, and SQLite. SQL is used in a variety of other tools, such as CartoDB for example.
- **When to Use:** Initial data exploration and analysis, particularly in concert with other applications.
- **Case Study:** One group used SQL as part of their CartoDB mapping application. CartoDB stores data in a PostGIS database, and the CartoDB.js API allows users to modify the SQL queries used in their CartoDB applications. In this case, the group used this functionality to create interactive filters for the data shown on their map.
- **Ease of Use:** Beginner, with some instruction.
- **Cost:** Paid and free versions
- **Works Well With:** Useful in doing many kinds of data journalism and analysis, particularly for joining different data sets.
- **Strengths:** It is effective for exploring data. It makes it easy to quickly analyze data and helps in becoming familiar with the data and determining next steps in data-cleaning.
- **Weaknesses:** Can have a higher learning curve for journalists unfamiliar with using data analysis techniques.

CLEANING

OpenRefine

- **Description:** OpenRefine is a tool that runs on a local machine and allows a user to filter and transform data.
- **Website:** <http://openrefine.org/> (<http://openrefine.org/>)
- **When to Use:** To clean medium to large amounts of messy tabular data.
- **Case Study:** For one investigation, a team performed a data analysis on inspection reports, containing about 40,000 – 60,000 observations. A number of fields were unstandardized. For example, the suggested timeline for fixes to had numerous variations: “3 months”, “within 3 months”, “less than 3 months”, as well as spelling mistakes. OpenRefine made it easy to see these different variations and cluster and normalize them with the “text facet” feature. The “numeric” and “date” facets allowed the team to identify outliers and other problems with the data.
- **Ease of Use:** Beginner
- **Cost:** Open-source/free
- **Strengths:** One can use Python to write scripts for it, allowing for a wide-range of data manipulations. Has a fairly useful set of pre-programmed transformations.
- **Weaknesses:** It can require a some programming knowledge for tasks that go beyond the abilities of the pre-programmed operations.

Trifacta

- **Description:** Trifacta is a data wrangling tool with a very intuitive user interface to simplify the process for non-technical users. The user interface can be used for simple data analysis, such as understanding each variable’s distribution. It is built on Hadoop and can handle large datasets.
- **Website:** <http://www.trifacta.com/> (<http://www.trifacta.com/>)
- **When to Use:** When a dataset needs cleaning or wrangling, or for quick analysis of variable distributions.
- **Ease of Use:** Beginner
- **Cost:** Paid program
- **Works Well With:** R, any tool that can input tabular data

- **Strengths:** It is an effective formatting tool that saves development time in structuring data. Its biggest, underrated strength is developing scripts on random samples of the data — it provides a snapshot of what the distribution of columns looks like. It abstracts and provides a user-friendly interface for basic regular expressions. It is good at predicting, without user-input, what kinds of transformations might be necessary to clean a dataset. It does a good job of autocompleting commands. Prior familiarity with SQL might help, but isn't necessary.
- **Weaknesses:** Input data needs to be sufficiently regular, for example rows with a uniform number of columns.

GENERAL PURPOSE AND SCRIPTING TOOLS

Python

- **Description:** Python is a powerful tool for data-wrangling, preprocessing and data analysis. It is a general programming language, and its utility comes from its packages. The Pandas package, in particular, is useful for data manipulation and analysis.
- **Website:** <https://www.python.org/>
- **When to Use:** Great for data wrangling and cleaning. When data needs partial work that can't be done cell-wise — date conversion, partial summing or aggregation. It is a good fall-back if other tools do not work.
- **Ease of Use:** Intermediate
- **Cost:** Open-source/free
- **Works Well With:** Tableau
- **Strengths:** It is versatile. If there is not a specialized tool for a given task, Python can usually be successfully used. Python makes it easy to modify the raw data contained in a CSV file, as opposed to other programming languages in which a user can modify data only after it has been read into a new data table or other programming object. Its programming syntax is easy to read and understand.
- **Weaknesses:** It requires a moderate level of programming knowledge.

R and RStudio

- **Description:** The combination of R and Rstudio are extremely effective tools for any kind of data manipulation and analysis. It uses a command-line interface, but with only a few commands even beginners can do meaningful work such as viewing, manipulating, and analyzing their data. Its strength comes from the thousands for data analysis packages written for it. Particularly useful are the dplyr and tidyr packages for manipulation and the ggplot2 package for visualization. RStudio is an integrated development environment for R with a script editor, an interactive console for running code, and windows with variable information, graphical output, and documentation. Overall, R and RStudio are probably the most effective data wrangling and analysis tools with their variety of packages available and R's flexible and compact language.
- **Website:** <http://www.r-project.org/>
- **When to Use:** Wrangling data, such joining data sets, and for a wide range of analysis and visualization.
- **Ease of Use:** Beginner to advanced: simple functionality is easily learned; advanced programming and packages can have a steep learning curve.
- **Cost:** Open-source/free
- **Works Well With:** Trifacta
- **Strengths:** Its functionality extends beyond basic tasks (reading in a table and printing out rows/columns) to almost any kind of analysis. Due to its popularity among the data analysis community, there are a variety of packages available for different tasks, as well as a large body of tutorials/help forums available online. It is useful in quickly generating graphs and computing statistics about data with only a few simple lines of code. Its statistical capabilities proved useful in detecting patterns in the data.
- **Weaknesses:** Using its more advanced programming features requires moderate programming knowledge, include functional programming to efficiently process large datasets.

Summary and Comparisons

- Excel, Google Sheets, and R are great tools for basic data analysis. Each allows you to quickly view your data, do summaries, and look for patterns. R's dplyr package and SQL can both filter, select, summarize, and join datasets.

SQL is suitable when the data are stored in a database. R can handle data stored in databases and in files.

- Python is a general purpose programming language designed for ease of use. R is special purpose programming language design for working with data. Python has packages in many domains. R has extensive, high-quality packages for manipulating, visualizing, analyzing, and modeling data.
- Trifacta makes wrangling data easier for non-technical users with its intuitive user interface and by automatically suggesting cleaning steps and transformations. Some programming knowledge is helpful but not essential.
- OpenRefine and Trifacta are both useful tools to clean data. The Trifacta interface guesses good defaults and transformations, while OpenRefine is very manually driven. That said, OpenRefine might be a better option with smaller datasets because it lets you view all the data and execute transformations better. Trifacta saves a transformation script, which is very useful when redoing a transformation on a dataset.

DOCUMENT ANALYSIS

Overview

- Description: Overview processes PDF files, either uploaded from a local drive or imported directly from DocumentCloud, and finds common words and phrases that appear in multiple documents. It is a great tool for finding basic patterns across documents.
- Website: <https://www.overviewproject.org/> (<https://www.overviewproject.org/>)
- When to Use: Document sorting, management
- Case Study: One group used Overview to manage the documents provided by regulatory agencies for easy and shared access.
- Ease of Use: Beginner
- Cost: Free/open-source
- Works Well With: DocumentCloud
- Strengths: Automatic category detection is quite easy to use and makes Overview well-suited for initial stages of data analysis when one may not be sure what patterns may exist in the data. Links with DocumentCloud for easy parsing of PDF documents.
- Weaknesses: Overview automatic tagging mechanism was largely ineffective and the manual tagging could be mostly duplicated by using Google Docs. There are some problems when searching for phrases rather than individual words, most notably when those phrases involve 'stop words' (common words such as 'a' and 'the')

DocumentCloud

- Description: DocumentCloud is a great platform to store, annotate, analyze, and read primary source documents. One feature of DocumentCloud is the ability to use dates in documents to generate a timeline to aid in analysis.
- Website: <https://www.documentcloud.org/> (<https://www.documentcloud.org/>)
- When to Use: PDF to text conversion, entity extraction
- Case Study: One group had to read through dozens of press releases and large reports in PDF form. They used DocumentCloud's entity extraction tool to find potential contacts.
- Ease of Use: Beginner
- Cost: Open-source/Free (with journalism organizational account)
- Works Well With: Overview
- Strengths: DocumentCloud automatically generates a text version of all documents through OCR processing. Having raw text allows users to take advantage of various other tools for parsing raw text. Entity extraction. This tool is built into DocumentCloud and can identify persons, names, contact details, and other relevant information in each document. Document hosting. DocumentCloud allows you to host documents publically and privately and is a worthwhile document repository and method of publishing documents.
- Weaknesses: The timeline tool was inaccurate and lacked enough detail to supplant human analysis. The PDF to text conversion failed on documents that were improperly (not completely straight) scanned. Unfortunately, documents were scanned in this manner a surprising amount of the time.

Google OCR

- **Description:** Google has a straightforward web-based OCR tool that requires no previous knowledge, training or programming knowledge. **How to use:** Upload document to Google Drive, right click it and open with Google Docs. It will automatically OCR your pdf.
- **Website:** N/A
- **When to Use:** When documents are saved in PDF format.
- **Case Study:** One team took annual statistical reports in PDF format and used Google OCR to convert them to text.
- **Ease of Use:** Beginner
- **Cost:** Free
- **Works Well With:** PDFs
- **Strengths:** Good for converting PDFs to text for analysis.
- **Weaknesses:** Spotty performance on table. Maximum 10MB per file. Only performs OCR on first 10 pages.

Summary and Comparisons

- DocumentCloud was generally effective at converting PDFs to text. It had trouble when PDFs were scanned even slightly incorrectly but worked very well on most other documents. The fact that you can host documents and share them with other members of your team is also quite helpful. DocumentCloud is useful when needing to search through and annotate documents or publish them through an embed code. DocumentCloud might be the best choice with large sets of documents that you want to keep together and publish.
- Google OCR was useful for its OCR function that is very easy to use for specific documents.

Data Visualization and Mapping

GENERAL PURPOSE

Tableau

- **Description:** Tableau is a powerful, customizable, and easy-to-use tool that should be a part of every data visualization tool box. Great for basic charting and exploring your data quickly
- **Website:** <http://www.tableau.com/>
- **When to Use:** Quick analyses and visualizations of data.
- **Case Study:** Several groups used Tableau for exploratory data analysis and to prototype visualizations before finalizing them using D3 and CartoDB. One can make graphs in Tableau significantly quicker than by using D3 (though the graphs are less customizable), making it an excellent tool for prototyping.
- **Ease of Use:** Beginner/Intermediate
- **Cost:** Free version for journalists
- **Works Well With:** Trifacta and data in a wide range of data types or stored in databases
- **Strengths:** Most of its tools are easy-to-use, quick-to-learn, and intuitive. It can create professional level visualizations and requires no programming knowledge.
- **Weaknesses:** Hard to do visualizations other than the subset it is designed for. More advanced features have a significant learning curve. For example, parameterization and graph coloring require knowledge of their Excel-like formula building.

R – ggplot2

- **Description:** A plotting system for R, capable of producing an extensive array of high-quality graphics. Can be used for both exploratory data analysis and for data visualization presentations.
- **Website:** <http://ggplot2.org>
- **When to Use:** For R users, ggplot2 is an excellent choice for data visualization.
- **Case Study:** One of the groups used ggplot2 to make sense of a large volume of housing data, for example plotting eviction trends over time.
- **Ease of Use:** Intermediate
- **Cost:** Free
- **Works Well With:** Any data that can be read into R

- Strengths: It is easy to use for simple plots, but gives users extensive customizability.
- Weaknesses: It does require programming knowledge. To make the best use, it is helpful to learn the Grammar of Graphics it uses (the gg in its name).

D3

- Description: D3 is a JavaScript library for data visualizations.
- Website: <http://d3js.org/>
- When to Use: When designing intricate, highly customized, and interactive visualizations.
- Case Study: One group wanted to develop an interactive timeline-tree to represent Twitter discourse over time. D3 was the only tool that could provide both hover and automation simultaneously.
- Ease of Use: Advanced
- Cost: Open-source/free
- Works Well With: Any structured data format
- Strengths: It is highly customizable and creates beautiful, interactive graphics.

There are several extensive libraries of publicly available examples that can be used to develop visualizations, eg. <http://bl.ocks.org/mbostock>

- Weaknesses: It has an unintuitive syntax and precipitous learning curve. Strong programming expertise is required to successfully use D3. Its developmental cycle is long. Therefore, it is best used only after concrete details of the final visualizations have been established.

GEOSPATIAL DATA

CartoDB

- Description: CartoDB is an online tool designed for analyzing and visualizing map data.
- Website: <https://cartodb.com/>
- When to Use: When mapping data.
- Case Study: One group used CartoDB to create a tool which would allow parents to find and compare daycares near them in California. They modified the HTML, CSS and JavaScript of the map to create a highly customized final product.
- Ease of Use: Beginner-intermediate
- Cost: Paid and free version
- Works Well With: Google Drive, automatically syncing data
- Strengths: It easily allows the user to plot a variety of geospatial data on a map and customize their appearance (color, shape, size). For basic functions, it allows the user to carry out tasks through a graphical interface-based wizard, requiring no programming ability. While there are limitations to the basic interface, for more nuanced settings, one may modify the HTML, CSS, SQL queries, and JavaScript to create highly customizable maps.
- Weaknesses: A moderate amount of programming knowledge is required to modify the HTML, CSS, SQL queries, and JavaScript. Limits the amount of geocoding allowed per account, so users with large datasets must use another tool to geocode.

GeoPy

- Description: GeoPy is a Python library and geocode is an R function that make it easy to use the Google Geocoding API to geocode addresses.
- Website: <https://geopy.readthedocs.org/en/1.10.0/>; <http://cran.r-project.org/web/packages/ggmap/ggmap.pdf>
- When to Use: Geocoding
- Case Study: One group used GeoPy / Google Geocoding API to generate latitudes/longitudes for ~20k addresses in California.
- Ease of Use: Both are straightforward, requiring only basic coding knowledge
- Cost: Free
- Works Well With: Each works well with their respective languages

- Strengths: Both make writing the code to interface with Google's API easier and shorter than writing a script from scratch.
- Weaknesses: Google's API limits the number of geocoded addresses to 2,500 per 24-hour period per IP address for non-paying customers. Both can geocode incorrectly even when the input address is well-formatted. It is important to check for obvious outliers.

TIMELINES

TimelineJS

- Description: TimelineJS is a tool that allows users to easily create interactive timelines and embed them in web pages.
- Website: <http://timeline.knightlab.com/>
- When to Use: Building a timeline
- Case Study: One group used TimelineJS to create timelines they later used in their final report. Building the timeline was easy and intuitive even for team members with no programming experience. A similar timeline created by Time Magazine can be viewed [here](#).
- Ease of Use: Beginner
- Cost: Open-source/free
- Works Well With: Google Spreadsheets
- Strengths: It requires no programming knowledge and is extremely easy to use The timeline looks extremely professional with little effort. It is linked up to Google Sheets, which means that all one has to do to create a timeline is fill out the spreadsheet template provided. It has an active help desk for any questions one might have.
- Weaknesses: The template provided is very difficult to customize.

NETWORKS

Gephi

- Description: Gephi is an interactive graph data analysis and visualization tool. It is geared towards network analysis, and contains powerful layout algorithms.
- Website: <http://gephi.github.io/>
- When to Use: When mapping network data
- Case Study: One group experimented with Gephi for visualizing tweet flow. Since number of tweeters was small, and the group wished to have greater control of the interactivity, they ultimately choose to use D3.
- Ease of Use: Intermediate
- Cost:
- Works Well With: Data stored in a graph format
- Strengths: Great tool for more graph data analysis and visualization It can be scaled to large graphs While it has extensive features that are harder to configure, can be used minimally to develop quick analysis
- Weaknesses: It is difficult to customize visualizations. The tool is not designed for data visualization but for analysis. Once the graph data is uploaded, it is very hard to change nodes, edges or attributes, such as size or descriptive text to accompany the data nodes. For a more interactive network building and analysis process, the user expects to not only delete nodes but change their values. Gephi displays a spreadsheet containing all the information, but it is not modifiable. In terms of visualizations, it is not possible to select multiple nodes to rearrange. Visualizations can be designed either using algorithms or manually per node.

Summary and Comparisons

- The above visualization tools are most suited for the following tasks:
- Tableau: Exploratory data analysis and visualization prototyping
- R: ggplot2, exploratory data analysis and visualizations for presentations
- CartoDB: Map visualizations
- D3: Custom designed, interactive visualizations
- Gephi: Graph analysis

- TimelineJS: Timelines
- If deciding between Tableau and D3, Tableau is better for prototyping during the design process and D3 is better for the final product. That said, D3 has a high learning curve, and Tableau is able to create professional level visualizations. If a journalist doesn't have experience with D3, Tableau is likely the better choice.

Data Acquisition

TWITTER DATA

Twitter Archiving Google Spreadsheets (TAGS)

- Description: TAGS is a Google spreadsheet-based program that allows for easy collection of Twitter data based on hashtags.
- Website: <https://tags.hawksey.info/> (<https://tags.hawksey.info/>)
- When to Use: Collecting tweets given a set of hashtags.
- Case Study: After running an algorithm with Python to collect hashtags of interest, we fed the resulting hashtags to TAGS and collected more than 15,000 tweets in a fully automated process.
- Ease of Use: Beginner
- Cost: Free
- Works Well With: Data analysis programs such as Trifacta (see below) or R that can import its CSV results.
- Strengths: It is easy to use and requires no programming knowledge, an understanding of Google Drive is sufficient. When fully automated, can collect 6000 tweets in an hour, and can be further configured to collect more without duplicates. It produces a well-formatted CSV file with detailed information about the tweets, such as location and language.
- Weaknesses: It does not support complicated collection queries. It does not automatically collect public information about the tweeter.

Scraping from Websites

Organizations often publish information online through web pages rather than publish their own internal database. In these situations, it can be necessary to scrape data from the website. Unfortunately, the nature of the problem also often requires custom solutions: there is no one-size-fits-all solution. For simple lists of files to be downloaded, a tool such as DownThemAll (a Firefox add-on) can work. Other times, programs such as import.io can help for straightforward scraping jobs.

However, if data needs to be extracted from particular parts of the web pages themselves, a scripting solution is often the only option. Here, are two approaches to scrape data from websites using Python.

Python – Scrapy

- Python: Python is a popular programming language accessible by all major operating systems.
- Description: Scrapy is a package that automatically crawls websites and scrapes data matching a given pattern.
- Website: <http://scrapy.org/> (<http://scrapy.org/>)
- When to Use: Scrapy is useful for scraping data from an entire website by following all the links. It provides a programming abstraction that requires one to only specify the information to extract from each web page (i.e. table headings, etc.) and which links to follow next. It can be too large of a hammer for scraping a single web page.
- Case Study: For one story, the team needed to extract audit reports. The list was paginated: only 20 reports could be accessed on single page and to get to the next page, the students had to follow a link. Normally, this would be perfect for Scrapy. However, the website had an undocumented protocol that made it difficult to follow the links with Scrapy. So the team hand-coded a scraper using the Python requests package.
- Ease of Use: Intermediate – requires a decent knowledge of Python and programming to use.
- Cost: Open-source/free
- Strengths: It is easy to scrape large websites with many pages.

- Weaknesses: Sometimes websites do strange things to prevent scraping. In these cases, the Scrapy abstraction fails and it doesn't serve the stated purpose.

Python – requests + lxml etree

- Description: Requests is a package that makes it easy to query websites, and lxml.etree is a DOM tree parser
- Website: <http://docs.python-requests.org/en/latest/> (<http://docs.python-requests.org/en/latest/>); <http://lxml.de/> (<http://lxml.de/>)
- When to Use: This approach is very low-level and requires a good understanding of how the web works, in particular what GET and POST requests are. We would recommend using it if one needs to access a single web page or if one has to work around a website designed to prevent people from scraping it.
- Case Study: As mentioned in the Scrapy case study, one team needed to scrape links and text from a paginated website. The website had a particular protocol which required a hidden field on the website to be used as an argument to link to the next page. Once the team scraped the web page, they extracted the hidden field using the lxml.etree package.
- Ease of Use: Advanced – requires a good understanding of Python, programming, and how the Internet and web pages work.
- Cost: Open-source/free
- Strengths: Its uses are highly customizable and it is easy to use.
- Weaknesses: Needs one to implement all the logic.

	DATA AND RESOURCES	NEWS AND INFORMATION	OUR ORGANIZATION	GET INVOLVED
(https://www.ire.org/)	Database Library (https://www.ire.org/nicar/database-library)	IRE News (https://www.ire.org/nicar/database-library/category/ire-news)	About (https://www.ire.org/about)	Join (https://www.ire.org/membership)
141 Neff Annex Missouri School of Journalism Columbia, MO 65211	Story Library (https://www.ire.org/resources/center/stories)	Extra Extra (https://www.ire.org/nicar/database-library/category/extra-extra)	History (https://www.ire.org/about/history)	Donate (https://www.ire.org/donate)
573-882-2042 (tel:5738822042)	Tipsheets (https://www.ire.org/resources/center/tipsheets)	Uplink (https://www.ire.org/nicar/database-library/category/uplink)	Board of Directors (https://www.ire.org/about/board-of-directors)	Contact (https://www.ire.org/about/contact)
info@ire.org (mailto:info@ire.org)	Listserve (https://www.ire.org/resources/center/listserve)	IRE Journal (https://www.ire.org/nicar/database-library/category/ire-journal)	Legal Documents (https://www.ire.org/about/legal-documents)	
Staff Directory (https://www.ire.org/about/office-directory/)		Member News (https://www.ire.org/publications/ire-journal/member-news)		
Advertise With Us (https://www.ire.org/publications/advertise-us/)		Job Postings (https://www.ire.org/jobs)		
Privacy Policy (https://www.ire.org/privacy-policy/)				