# A Practical Guide to Big Data Research in Psychology

Eric Evan Chen

University of California, Irvine

Sean P. Wojcik

Upworthy

## Author Footnotes

Eric Evan Chen, Department of Psychology and Social Behavior, University of California, Irvine; Sean P. Wojcik, Upworthy.

This article includes discussion of data and methods from previously published work, and these are noted and referenced as such. The other data and methods have not been previously presented or published. We thank Peter Ditto, Clint McKenna, and James Chen for their assistance in the preparation of this article. We thank Lisa Harlow and two anonymous reviewers for their helpful and insightful comments and suggestions.

Correspondence concerning this article should be addressed to Eric Evan Chen, Department of Psychology & Social Behavior, University of California, Irvine, Irvine, CA 92697. E-mail: chene5@uci.edu

Abstract

The massive volume of data that now covers a wide variety of human behaviors offers researchers in psychology an unprecedented opportunity to conduct innovative theory- and data-driven field research. This article is a practical guide to conducting big data research, covering data management, acquisition, processing, and analytics (including key supervised and unsupervised learning data mining methods). It is accompanied by walkthrough tutorials on data acquisition, text analysis with latent Dirichlet allocation topic modeling, and classification with support vector machines. Big data practitioners in academia, industry, and the community have built a comprehensive base of tools and knowledge that makes big data research accessible to researchers in a broad range of fields. However, big data research does require knowledge of software programming and a different analytical mindset. For those willing to acquire the requisite skills, innovative analyses of unexpected or previously untapped data sources can offer fresh ways to develop, test, and extend theories. When conducted with care and respect, big data research can become an essential complement to traditional research.

*Keywords:* big data, application programming interfaces, text analysis, data analytics, data mining, machine learning

# A Practical Guide to Big Data Research in Psychology

Big data is about the real world, captured in an array of formats, from medical records to texts of speeches to photographs. The massive volume of data and the wide-ranging variety of behaviors captured in those data provide researchers in psychology the opportunity to conduct a quantitative form of naturalistic field research. The volume of the data generated and collected is orders of magnitude larger than has ever been the case, across a broad variety of human behaviors, and at a deep granularity (Lin, 2015). This "datafication" of everyday life is a product of both the extent to which people now use digital and online technologies (e.g., digital and social media), as well as the extent to which digital technologies have replaced old ways of life (e.g., digital storage of medical records) (Mayer-Schönberger & Cukier, 2013). The Internet—and social media in particular—is a powerful force now, and will likely only increase its role in our daily lives. According to the Pew Research Center (2015), 92% of U.S. teens aged 13 to 17 reported going online daily for social media and other uses, with 56% going online several times a day, and 24% going online "almost constantly."

The goal of this article is to introduce key big data acquisition and analytics concepts and provide a framework for developing behavioral science research projects based on these concepts. This article is geared toward those who have little or no prior big data research experience but who have some technical knowledge and, ideally, some programming experience. Because of the vast range of data acquisition, management, and analytics techniques, it is far beyond the scope of any single article to describe all of these big data techniques in full detail. Thus, this article focuses on conveying essential underlying concepts and on the practical aspects of carrying out big data research, including software engineering concerns related to acquiring and analyzing big data.

This article follows four major steps of a typical big data research project: data management planning, data acquisition, data processing, and data analytics. It is accompanied by three walkthrough tutorials in the Supplemental Materials: one on data acquisition and two on data analysis. In order to gain hands-on experience with and more fully understand the techniques and software discussed here, working through these tutorials is strongly recommended. This article begins by describing big data characteristics and outlines some of the challenges that a big data research project might present for psychologists. It then discusses aspects of data sources and data handling, including management, acquisition, and processing. Finally, it discusses two sets of commonly used big data analytic techniques: 1) text and multimedia analytics and 2) data mining and machine learning.

## Big Data, and Big Data for Psychologists

Big data was first characterized with three V's: Volume, Velocity, and Variety (Laney, 2001). *Volume* refers to the sheer scale of the data. There is no specific numerical threshold above which data becomes Big. Rather, data becomes Big when the scale exceeds the computing capacities of generally available hardware and software. *Velocity* refers to the speed with which the data are generated and the speed of the analytics process required to meet the demands. If we think of the pace of traditional data collection as akin to drinking from a water fountain, then Big Data collection is akin to drinking from a firehose, as Twitter's real-time, full capacity streaming interface is often referred to (https://dev.twitter.com/streaming/firehose, Twitter, n.d.-a). *Variety* refers to the many forms that Big Data can take—including structured numeric data, text documents, audio, video, and social media. Examples of this include image files, property values data, crime reports, and death certificates. Each of the potential forms that big data can take requires its own method of acquiring, managing, and analyzing. A fourth V, *veracity*, is often

added, though it is more of a challenge than a defining feature. Veracity refers to the challenge of handling the ambiguities and varying quality of big data. Often, Big Data (such as that from social network sites) are not generated for the purpose of analysis. Other data may not be entirely accurate. Raw, unstructured data must be translated and structured to prepare them for analyses. Throughout this process, opportunities abound for inaccuracies to be introduced or exacerbated.

As many have pointed out (e.g., Gandomi & Haider, 2015), there are a variety of definitions of Big Data, with some adding several more V's. A wide variety of types of data and analytical approaches are often subsumed under the rubric of Big Data. The particular perspective that we offer here is one useful for psychologists.

**Big *Data***

What data qualifies as big data, and what can we do with it? On a broad conceptual level, the "data" of Big Data is broken up into *unstructured data:* raw, unstructured digital information, such as text or images; and traditional, *structured data*: data that can be represented in a spreadsheet or statistics program. A hallmark of Big Data is that data are often *unstructured*. Unstructured "data" is not data in the sense of data as a collection of measurements. Rather, unstructured data can be thought of as pure raw digital information, such as text documents. In between these two types of data, conceptually, is *semi-structured data*. These data are similar to annotated data—elements of the data are tagged with information about their meaning. Although semi-structured data is an important type of Big Data, its use is often somewhat more specialized. For example, unstructured data has uses in areas such as information retrieval and in processing raw texts for analysis by computational linguists. This paper will focus on turning unstructured data into structured data, as this is likely to be more useful to psychologists.

A typical example of unstructured data would be a collection of tweets from Twitter. Raw tweet information from Twitter includes many pieces of information, such as the text of the tweet, number of followers of the tweeter, and whether it was a retweet. See Figure 1 for an abridged example of tweet data. From raw information, we can then take measurements, and analyze these in many different ways. For example, counts of positive and negative emotion words in the texts of the tweets could be measured, and these counts could be compared to the number of followers a user has.

**Small Data**

One of the most important uses of big data is to generate small data (Berman, 2013). Small data refers to data that do not have volume, velocity, and variety issues. This is the sort of data that most psychologists would be used to: data that can be represented in a spreadsheet and on a single computer. Using big data as Big Data—acquiring, managing, and analyzing massive volumes of a variety of data at high velocity—involves a significant degree of technical engineering expertise. Thus, we believe that it will be most useful to psychologists for this article to focus on the small data generation aspect of big data.

Admittedly, if we are moving back and forth between big data and small data, the line becomes fuzzy. In practice, whether a specific dataset actually qualifies as Big Data is often less important than whether a big data acquisition, management, and/or analytics technique could be usefully applied to that dataset. From this perspective, it is thus crucial to be able to identify and deploy relevant techniques to different data, as appropriate.

**Engineering a Big Data Research Project**

"Good programmers know what to write. Great ones know what to rewrite (and reuse)" (Raymond, 2000, p. 2).

Although computing skills are necessary for big data research, expert-level abilities are generally not required, in part because of the availability of pre-existing software libraries that implement advanced techniques. Capitalizing on collective effort is a fundamental aspect of big data (Lin, 2015), and wherever relevant, we point out software that can be used to implement a particular technique.

The general programming language Python and the statistical programming environment R—both freely available—are widely used in big data work. Python is a popular general-programming language, with a large community of active users who develop and make available a wide range of tools. R is popular in data analytics, and is already widely used in psychology (Culpepper & Aguinis, 2011).

In addition to the usefulness of these languages' core programming and analytical functionalities, these functionalities can be augmented by adding on new packages. These supplemental packages implement useful new features (e.g., additional graphing features, structural equation modeling) that may not otherwise be available in the core language package. The main repository of packages for Python is PyPI (https://pypi.python.org/pypi) and for R is CRAN (https://cran.r-project.org/). Also, GitHub (https://github.com/) hosts packages for a number of languages. Most of the specific acquisition and analytical techniques discussed in this article are available as free software packages. In addition, there are also general toolkits for broad processing tasks. For natural language processing (NLP) in Python, this includes the Natural Language Toolkit (NLTK: http://www.nltk.org/), and in R, the tm package (https://cran.r-project.org/web/packages/tm/index.html). For computer vision (including facial recognition) work in Python, opencv (http://opencv.org/) is widely used.

The Text Analysis, Crawling, and Interpretation Tool (TACIT: http://tacit.usc.edu/ Dehghani et al., 2015) may be of particular interest to psychologists interested in conducting text analyses. It implements many of the text acquisition, management, and analysis tasks described in this paper with an easy to use interface.

**Additional Resources**

Big data projects can face a wide variety of challenges and standards and implementations evolve rapidly. Much, if not most, of the practical knowledge exists outside of the traditional media forms of journals and books. This knowledge is online, in forms such as official documentation, blog posts, and technical forums. Resources such as these offer instruction and advice on tasks of varying specificity, addressed to readers of varying levels of skill and knowledge. Software repositories (e.g., CRAN, PyPI, GitHub) are vital resources. Online education platforms such as Codecademy (https://www.codecademy.com/), Coursera (https://www.coursera.org/), DataCamp (https://www.datacamp.com/), and edX (https://www.edx.org/) offer courses that cover a wide range of big data-related topics and technical skills. For problem-solving specific issues, blog posts and forum discussions often cover frequently encountered issues. The Supplemental Materials includes a summary of key resources mentioned in this article.

One of the most important skills to develop is the ability to search for and evaluate the online technical information needed to solve those challenges. Many people encounter similar problems, and there is a vibrant community of developers and users who share their collective knowledge and experience. Many online resources serve as open forums for collective problem-solving and as repositories for the solutions reached. Stack Overflow (http://stackoverflow.com/) is one widely used forum. Because search engines provide access to resources of many different

types, it is essential to be able to succinctly characterize problems that arise, and may also involve having some idea of the potential solutions to the problem.

It is also crucial to note that the *velocity* of big data could also refer to the speed at which protocols and formats—particularly proprietary protocols and formats—and software and hardware can change, and often with little, if any, warning. It is thus essential to be well-versed in the underlying principles, to be sufficiently technically proficient, to be familiar with the structure of the data, and to constantly watch out for changes in how a data source operates and how their data are structured. Please be aware that the specific details in the examples given and in the software and resources discussed in this article may have changed since this article was written. Thus, in the discussion below, we have sought to convey underlying principles while also providing the concrete, technical details necessary to understand these principles.

## Data Management Planning

Before beginning a big data project, one must plan how to manage the data. One key decision is whether the data will be stored using the native file system of the computer's operating system (OS) or in a specialized database storage system. In making this decision, some important questions include: What is the expected size of the data? How should the data be organized? How will the data be accessed for analysis? How will the data be distributed? Although many big data practitioners use specialized database storage systems, this will usually be excessive for most psychologists, and would thus be an unnecessary technical challenge. Given this, we recommend using the OS's file system. Nevertheless, databases are an important aspect of big data and psychologists interested in big data should at least be aware of them. We discuss both approaches below.

### Native OS File Systems

Using the native OS file system is suitable for most research projects aiming to generate small data from big data, given the right planning. The main issues that arise stem from the challenges of large file sizes. For example, for the FAT 32 file system, which is commonly used for portable storage devices such as flash drives, the maximum file size is 4 GB. Collecting data from streaming interfaces such as Twitter's can easily exceed this. Many software packages that collect data from such interfaces simply copy the data directly from the data source to a single file. Although modern versions of Windows and Mac OS use file systems that support much larger maximum file sizes, transferring files between computers using different operating systems can still present problems.

**Managing large data files.** Because storage space is, in general, much more plentiful than memory space, we can typically acquire and store much more than we can analyze. Because most statistics packages such as SPSS or R read the entire dataset into memory, this means that the amount of memory will need to be at least the size of the dataset, plus overhead space (Danneman & Heimann, 2014). Unfortunately, computers available to consumers often do not have sufficient memory for this.

However, this challenge can often be overcome with awareness and preparation. For example, through a combination of judicious sampling from the data source (e.g., De Choudhury et al., 2010) and power analysis, psychologists can acquire manageably-sized datasets. If a large volume of data has been acquired in the form of excessively large files, software is available that can shrink and/or split these large files into smaller pieces, and each of these smaller files can be handled separately, as appropriate. With structured data, many statistical analysis programs allow random sampling from large datasets.

**Interoperability.** Another consideration in choosing how to manage the data is the requirements of the planned analysis software. An advantage of using the native file system for unstructured and structured data is that statistics and other analytics programs typically access data through the OS file system. Thus, storing data in the file system from acquisition to analysis can streamline the entire process.

## Databases

Traditional database systems are designed to optimize the storage, management, and access of structured data. In many ways, a traditional database is analogous to a searchable spreadsheet. One of the main strengths of databases is that they can be queried to search and retrieve data in an almost limitless variety of ways. If search and cross-reference functionalities are truly necessary for a research project, storing the data in a database may be worthwhile. Such cases would include projects for which a collection of data will be preserved for future studies that each handle and analyze the data or subsets of the data in very different ways.

Different database technologies use their own languages to communicate between the database management system and the users of the data. A good starting point for understanding this process is through the Structured Query Language (SQL). This is one of the most common ways of interfacing with databases that are structured based on specified relations between the data. MySQL is a popular, free database implementation of SQL (https://www.mysql.com/). There are also many commercial implementations of SQL.

With the advent of big data, new database management systems have been developed, in part to better handle unstructured data. Some of these database technologies, grouped under the moniker NoSQL, go beyond the use of SQL and relational structures. NoSQL can be interpreted as both “Non-SQL” as well as “Not Only SQL.” Some widely-used NoSQL databases include

MongoDB (https://www.mongodb.org/), Apache's Cassandra (http://cassandra.apache.org/), and Redis (http://redis.io/). Learning how to deploy databases is a useful skill, and free online courses and other resources on databases are available. But managing databases can become challenging—the study of database technologies is its own subarea of computer science, and entire companies specialize in providing database management expertise.

## Data Acquisition

### Data Sources

Big data research in psychology has been closely associated with social media. However, this is only one type of big data. The digitization of modern life cuts across a broad range of areas. For example, an international effort has been working to systematize legislation in a machine-readable format (Sartor, 2011), and the U.S. government has initiated a policy of data availability across many government agencies (Project Open Data; https://www.data.gov). This policy has produced numerous ways to access these data sources, and useful tools to work with these data.

Some important categories of data sites are archival repositories, traditional media sites, and social network sites. Archival repositories include those of entities such as the U.S. Government (e.g., data.gov), as well as general collections, such as Amazon Web Services' Public Data Sets repository (AWS, n.d.: http://aws.amazon.com/public-data-sets/) which includes the 1000 Genomes Project, with full genomic sequences for 1,700 individuals, and the Common Crawl Corpus, with data from over five billion web pages. The University of California, Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/) hosts a large number of useful public datasets. Traditional media sites include the online presences of outlets such as CNN or the International Herald Tribune.

Social network sites have been defined as "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system" (boyd & Ellison, 2008, p. 211). Sites differ in how public the interactions of its users are. Twitter, for example, provides their users a public platform for communicating with as broad an audience as possible. Other sites provide more selective interactions, be it more personal (e.g., Facebook) or more professional (e.g., LinkedIn). In general, for social network sites like Facebook and LinkedIn, new social network connection requests must be explicitly approved by the user. For social media sites like Twitter, however, users' communications are public by default (Twitter, n.d.-b) and anyone can subscribe to these communications (Twitter, n.d.-c). This distinction between social sites is also reflected in the accessibility of the sites' data, both to developers (Facebook, n.d.; LinkedIn, n.d.; Twitter, n.d.-d) and to researchers (Schroepfer, 2014). If Twitter is like one giant public park, then Facebook and LinkedIn are like collections of individual homes and businesses.

**Acquiring Data**

There are several important ways in which data can be acquired from data sources, depending on the data source. Some data sources essentially give away their data, and provide interfaces that are designed to make it as easy as possible to access their data. These sources include social media sites, many of which aim to provide a broad public forum, and government agency data repositories. Other data sources have accessible data—a site's webpages, for example—but do not provide special interfaces to acquire their data in bulk. Acquiring data from these two particular types of sources will be discussed below.

In addition to these two approaches, data sources may share their data directly. Google provides their ngrams data, for example, for the public to download (Google, 2015: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html). This is not new or specific to big data. The Inter-university Consortium for Political and Social Research (ICPSR, 2015: https://www.icpsr.umich.edu/icpsrweb/landing.jsp), for example, has long maintained datasets for public use. Nevertheless it is important to take note of these sources of data because combining datasets from various sources is common in big data.

**Application Programming Interfaces (APIs).** For many sources, making their data as broadly and easily accessible as possible is a core part of their organizational mandate. To achieve this accessibility, they provide mechanisms known as *APIs*. An API, in this context, is a set of protocols and technologies that provide a systematic way for users to request and receive data. The typical process is straightforward: the user requests a particular piece of data, and the data source's server automatically sends the users that piece of data. The U.S. Government Printing Office (GPO), Reddit, Twitter, and other sources of big data typically provide documentation explaining the specific parameters that can be used to refine each source's API calls.

***Accessing APIs.*** Some APIs are open access and relatively unconstrained. Others place restrictions on how they are used. Sites like Twitter, for example, require users of its API to register as a developer. For Twitter and many other social network sites, this registration is free and open to most users. Typically, in this registration process, the site provides identifiers for each developer, and these identifiers must be supplied every time the API is accessed. Among other things, it provides a way for the site to regulate access to its API. It can limit the speed at which the user accesses data, or even block the user altogether.

***Example: Requesting congressional speeches via the GPO API.*** Suppose a researcher is interested in comparing the speech of Democratic and Republican politicians. The following example demonstrates how to acquire U.S. Congressional speech text data from the GPO (n.d.: https://github.com/usgpo) through its public API. This API follows the same format as other APIs, and it is free and open access with relatively few restrictions on how the data are accessed. Registration is not required. For further documentation, see also: https://github.com/usgpo/link-service and https://api.fdsys.gov/. For this example, we are interested in the Daily Digest, in the Congressional Record, for January 15, 2014. The Daily Digest is the daily summary of the proceedings and debates of the U.S. Congress.

```
http://api.fdsys.gov/link?collection=crec&section=dailydigest&publishdate=2014-01-15
```

There are three parts to this request.

The first portion:

```
http://api.fdsys.gov/
```

is the address of the GPO server, and specifies that we will connect to it using the HTTP protocol (the standard protocol for web communication).

The second portion:

```
link?
```

is the GPO API's specified command for indicating that we are requesting data.

The third portion:

```
collection=crec&section=dailydigest&publishdate=2014-01-15
```

is made up of the parameters that specify what data are being requested. These parameters are specified by the API's documentation. The parameters are separated by a `&` sign:

```
collection=crec
section=dailydigest
publishdate=2014-01-15
```

`collection=` Indicates the requested collection. `crec` refers to the Congressional Record – Daily. Some other collections that the GPO offers include: Congressional Documents (`cdoc`), Compilation of Presidential Documents (`cpd`), and Public and Private Laws (`plaw`).

`section=` Indicates the requested section. `dailydigest` indicates that you are requesting data from the Daily Digest. Because each of the collections is organized in different ways, you will need to know how they are set up. The Congressional Record is broken down into the Daily Digest (`dailydigest`), Senate Proceedings (`senate`), House Proceedings (`house`), and Extensions to the record (`extensions`).

`publishdate=` Indicates the publication date of the requested document. `2014-01-15` refers to January 15, 2014.

See Appendix A for Python code to automatically generate requests for acquiring Daily Digest data.

***Other APIs.*** This general API format is used by many interfaces. The following command:

```
http://www.reddit.com/search.json?q=psychology&limit=10&sort=new
```

for the internet site Reddit (`http://www.reddit.com/`), the self-described "front page of the internet," searches (`search.json?)` for the word "psychology" (`q=psychology`) and returns the ten (`limit=10`) newest (`sort=new`) listings.

For Twitter, this command:

```
https://twitter.com/search?q=psychology%20since%3A2016-01-01%20until%3A2016-01-02
```

searches Twitter's timeline for all Tweets matching the term "psychology" for January 1, 2016. In this command, `%20` represents space and `%3A` represents colon. This is known as URL encoding or percent encoding (Internet Engineering Task Force, 2005). For further explanation, please see the section in the Supplemental Materials on URL encoding.

For popular APIs, there are often existing software packages that implement these functionalities. The TACIT tool, for example, (Dehghani et al., 2015) provides access to several data sources, including the U.S. Congress and Supreme Court, Twitter, and Reddit. For accessing GPO data, the Capital-Words software toolkit (source code at: https://github.com/sunlightlabs/Capitol-Words) from the Sunlight Foundation implements a broad range of functionality (Sunlight Foundation, 2015). This organization also offers its own API through which users can acquire and analyze government documents.

Tutorial 1, in the accompanying Supplemental Materials, walks through the acquisition of data from an API. This tutorial acquires a large batch of real data from the GPO. These data will then be analyzed in Tutorial 2.

**Pulling data.** For data sources that do not have APIs, data can sometimes still be acquired. With websites, for example, when a site's server transmits data to users to be displayed, acquisition tools can capture these data. When a user navigates to a website, his or her web browser sends a request to the site's servers for a page (e.g., the front page of Amazon.com). In response to this request, the server then sends the data comprising the webpage to the browser. Software can be written that mimics web browsers. Web servers send requested data to these software programs as if they were filling requests from a web browser. Because webpage data is in a standardized format (HTML: Hypertext Markup Language), software can be written that then interprets the data from the server and extracts the desired pieces of information.

This type of automated retrieval should be used with caution. Although some sites will allow you to "scrape" data in this way, provided that you do so at a reasonable rate of acquisition, it is often expressly forbidden in sites' Terms of Service. Automated data acquisition in big data research can consume a disproportionately large amount of a data source's finite pool of computing resources. Even when the data collection process by researchers goes unnoticed by end users of a site (as is often the case with research on social network sites, for example), for the site itself, this process always uses some amount of resources. The processors that serve the data requests and the network connections that provide the data both have finite capacity. Each request for data takes time for a processor to complete. The network connections through which the data flow are like pipes—only a finite volume can pass through at any given time. In fact, one common cyber-attack, known as a denial of service (DoS) attack, involves flooding a server with so many requests that actual users cannot access the server (Internet Engineering Task Force, 2006). Automated data collection can generate many requests for data in a short amount of time that can appear similar to a denial of service attack. Even if the collection process does not overload a server, sites often have measures to protect against access patterns that resemble an attack, including safeguards to block the perceived attacker.

One way to access data respectfully is to pause between requests. An automated process can generate request after request, each following the other on the order of milliseconds. Pausing appropriately is also important if there are communication errors or if a server is momentarily overloaded. It is consistent with internet standards to use linear and/or exponential back-off (i.e., waiting) times (e.g., Internet Engineering Task Force, 2009). The sample acquisition code given in Tutorial 1 provides an example implementation of a back-off algorithm.

Ethical big data research should consider the needs and requests of those who maintain data sources, in addition to those of their users. One should use care when acquiring data, and for relatively smaller data acquisition tasks, these data can be collected manually.

### Additional Resources

There is a wide range of online and offline resources that cover various aspects of data acquisition. Danneman and Heimann (2014) cover several aspects of acquiring and analyzing social media data, and R. Russell (2013) covers the acquisition and analysis of data from a variety of social network sites, in Python. Among the online resources, there are also forums specific to particular data sources, such as Twitter (https://twittercommunity.com/).

## Data Processing

Once collected, the data must be cleaned and prepared for analysis. This data processing phase is the bridge phase between data acquisition and data analysis. It could also be thought of as the data *pre*-processing phase, for those who prefer to think of analysis as the processing of data. Much newly-acquired big data will be raw and unstructured, and cleaning, extracting, and processing these raw, unstructured data is a fundamental challenge (Gandomi & Haider, 2015; Labrinidis & Jagadish, 2012). This task depends closely on the idiosyncrasies of the data sources and the planned analyses, and on the particular requirements of the analysis software to be used.

### Example: Extracting Text from HTML Files

For text data, one common task is to remove extraneous elements that were a result of the data acquisition process. Text content is often embedded in files such as HTML webpages. When this is the case, HTML tags must be stripped. (Note that this is similar to the HTML parsing aspect of scraping.)

To illustrate, we can download an HTML document from the GPO with the following API command:

```
http://api.fdsys.gov/link?collection=chrg&congress=105&jacketid=48-707&link-type=html
```

A web browser would automatically format the page, based on the HTML formatting. However, looking at the raw source of the retrieved document (this is the entirety of the data sent from the server to a web browser), we see the second line is the title displayed by the web browser, along with the HTML tag to specify how it is to be formatted:

```
<title> - GOVERNMENT PROGRAMS AND OVERSIGHT OF THE SMALL BUSINESS COMMITTEE AND THE SUBCOMMITTEE ON BENEFITS OF THE COMMITTEE ON VETERANS' AFFAIRS</title>
```

The symbols `<title>` and `</title>` are HTML tags to mark that the text in between them is the title of the document. This is the title shown by the browser. There are many other types of tags that each serve different purposes. For example, the tags `<p >` and `</p>` indicate that the text in between them is a paragraph of text.

Processing these kinds of files involves writing programs that use the tags to automatically recognize and extract particular pieces of information from HTML files. For example, if we are interested in extracting the title from the previous file, the following Python code is a (simplified) way to automatically extract the title:

```
if line.startswith("<title>") and line.endswith("</title>"):
   title = line.replace("<title>", "").replace("</title>", "")
```

The first line of code detects whether a line of raw input starts and ends with the title open (`<title>`) and close (`</title>`) tags. The second line of code then removes the tags by replacing them with nothing (i.e., an empty string: `""`).

One of the challenges of this approach is that the format of the raw HTML files must be analyzed before writing the processing code in order to determine how the relevant information is encoded within each HTML document. Often, the data of interest are buried within an HTML

document, and programs must implement the logic to automatically detect where these data are and extract them. In practice, extracting even the title from an HTML file can be more complicated than what is given in this code snippet. Furthermore, web pages are updated constantly, and are often the end result of processing by JavaScript, Flash, PHP, and/or Ruby, to name just a few web technologies. This can make the data of interest a moving target.

Unstructured data are often very messy, and processing these data can require a great deal of effort. This data processing phase is often the most time-consuming part of a big data project. The key to this bridge phase is to focus on the beginning and the end points. Where did the raw data come from? How are the unstructured data formatted? What information is needed? What is the target format? In some cases, such as with text analysis, the goal of processing is not to structure the data, but rather to render the data into the specific format required by the next step of the process.

The processing phase may also need to handle missing data. Not all analysis software packages handle missing data in a robust manner—some may simply not function. In such cases, missing data must be handled separately (e.g., imputed). Unfortunately, some software packages do not document how they handle missing data.

**Additional Resources**

There is a wide variety of specific forms of data that will be handled in this phase and the most useful resources will address each form's specific challenges. While there are some books (e.g., McCallum, 2012) that cover a selection of topics, data processing work often requires knowledge of specific foundational standards. For text processing, for example, depending on the software used, it is important to be at least familiar with the various standards by which text is encoded in digital formats (e.g., for Python version 2:

https://docs.python.org/2/howto/unicode.html). For processing HTML and other data from the web, familiarity with web standards is important (e.g., https://www.w3.org/standards/).

## Data Analytics

As argued above, big data research can ultimately involve the analysis of datasets that are not particularly big at all. Although big data analyses are generally automated, for smaller datasets in particular, they need not be (Lewis, Zamith, & Hermida, 2013). For many analyses in psychology, human coding remains the gold standard (Iliev, Deghani, & Sagi, 2014). Integrating big data research into psychology should not be held back by the lack of an automated version of a particular analytical technique. One can fruitfully apply a non-big data analytical technique to data acquired from a big data source. Conversely, one can also fruitfully apply a big data analytical technique to data acquired from a non-big data source.

Big data analytics techniques—especially machine learning-based big data analytics techniques—are, in some ways, in direct opposition to the hypothesis-driven, hypothesis pre-registration approach to scientific research. Hypothesis-driven research is a top-down structured approach; it starts with a hypothesis aiming to make a conclusive decision. Big data analytics typically involves exploratory analysis, with a bottom-up speculative approach into ideas for hypotheses.

Many major data analytics techniques have been implemented in freely available software. We highly recommend using this software, and we point out examples throughout the next section. Generally, these are written by experts in computer science and/or statistics, and are widely used in published studies. There is no need to reinvent the wheel, much less reinvent the car.

The following sections will cover four types of big data analytics techniques: 1) text analytics, 2) multimedia analytics, 3) supervised learning, and 4) unsupervised learning. The last two types of techniques constitute the two major types of data mining and machine learning analytics techniques.

## Text Analytics

Text analytics is widely used in many disciplines, including computer science, computational linguistics, and the humanities (Indurkhya & Damerau, 2010; see Iliev et al., 2014 for a review from a psychological research perspective). The analysis of language has a long history in psychology (Pennebaker, Mehl, & Niederhoffer, 2003), making these techniques highly relevant to psychologists interested in big data research. The following brief overview discusses three different approaches: user-defined dictionaries, feature extraction, and word co-occurrences. The user-defined dictionary approach (e.g., word counting analyses) is already widely used in psychology. The feature extraction and word co-occurrences approaches are based on algorithmically recognizing patterns within texts and are particularly useful given the unstructured nature of big data. These latter two approaches offer a rich representation of the words and ideas in text data.

**User-defined dictionaries.** This popular approach includes word count-based analyses (Iliev et al., 2014; Liu, 2010). This approach analyzes the frequency of a given set of words (i.e., the dictionary) that are conceptually related to the attributes of interest. These dictionaries can be based on prior research and development, or generated for a specific project. Commercial software, such as the popular Linguistic Inquiry Word Count (LIWC: http://www.liwc.net/), and free NLP software such as NLTK (http://www.nltk.org/) offer this functionality.

**Feature extraction.** This bottom-up approach starts with texts that differ in some known way and can identify the features (e.g., particular words) that most accurately distinguish the texts from each other (Iliev et al., 2014).

Diermeier and colleagues (2012) used the support vector machine (SVM) feature extraction approach to identify the terms most indicative of liberal and conservative ideology, using U.S. Senate speeches. They used the SVM$^{Light}$ software implementation (Joachims, 1999), though the LIBSVM implementation (Chang & Lin, 2011) is also popular with researchers. Both of these libraries, as with many toolkits, were developed with the express purpose of making it as easy possible for users to analyze their data. (For example, the developers of LIBSVM provide a script for running it called "easy.py"). In addition, the text analysis software TACIT (Dehghani et al., 2015) and most machine learning packages offer SVM functionality. SVM will be discussed further in the section below on data mining and machine learning.

**Word co-occurrence analyses.** This set of analyses focuses on how words are used together, within and between texts (Iliev et al., 2014). Two common approaches are latent semantic analysis (LSA) and latent Dirichlet allocation (LDA; unfortunately, this is also the acronym for linear discriminant analysis). Both of these techniques learn underlying, meaningful relationships between words by analyzing *corpora* (collections of texts). In some ways, this type of computational learning is analogous to the manner in which children learn new concepts through intuitive inductive reasoning (Landauer & Dumais, 1997).

Topic modeling is one of the most important classes of word co-occurrence analyses. This approach analyzes the words of a set of text documents to discover common themes (topics) across documents. This allows researchers to examine the connections between these themes and

how they may change over time (Steyvers & Griffiths, 2007). LSA and LDA are two widely-used probabilistic approaches.

***Latent semantic analysis.*** This family of analyses assumes that words in texts are not randomly distributed. Words with similar meanings are more likely to appear in the same texts than are words that have less similar meanings (Landauer & Dumais, 1997). This approach traditionally relies on singular value decomposition, a form of factor analysis (Landauer, Foltz, & Laham, 1998), to reduce the number of dimensions. The gensim software package (Řehůřek & Sojka, 2010) is one implementation, and the Python machine learning package scikit-learn (Pedregosa et al., 2011) also has the functionality to perform LSA.

LSA has roots in information retrieval, where search algorithms aim to identify documents (e.g., webpages) that are similar in meaning, but may not even contain the full search term. This type of analysis generates quantitative measures of semantic similarity between terms. For example, the terms "self-perception theory" or "differential susceptibility" would each be considered to be semantically related to the general concept of "psychology;" and they would be relatively unrelated to the concept of "football."

***Latent Dirichlet allocation.*** LDA-based topic modeling uses differences and similarities in word distributions in a group of texts to generate the main themes (i.e., topics) that arise across the texts (Blei, 2012; Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004). For example, for a set of scientific documents, these topics might be data analysis, evolutionary biology, and genetics. A set of topics is generated based on similarities in word usage within and between those topics. For each topic, each text is assigned a score capturing the text's degree of relatedness to that topic. Importantly, each text has a score for every topic—this captures the idea that a text can discuss more than one topic. In addition, these scores can be used to divide the

texts into subgroups. Texts can be grouped together based on their topic distributions. Some implementations include MALLET (McCallum, 2002), GibbsLDA++ (Phan & Nguyen, 2007), Stanford Topic Modeling Toolbox (http://nlp.stanford.edu/software/tmt/tmt-0.4/), gensim (Řehůřek & Sojka, 2010), a Matlab Topic Modeling Toolbox (http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm), and TACIT (Dehghani et al., 2015). Tutorial 2, in the accompanying Supplemental Materials, walks through an example of LDA topic modeling using MALLET. This tutorial uses real text data from the GPO, acquired in Tutorial 1.

An example of output from an LDA analysis using MALLET is shown in Figure 2. This output is from Tutorial 2. This analysis generated these topics from U.S. Senate Additional Statements documents, which are relatively free-form statements given in the U.S. Senate, such as tributes and congratulatory statements covering a broad range of topics. Each pair of lines represents a topic and gives the key words representative of that topic.

Topic modeling has begun to see use in various areas in psychological research. For example, in the area of family psychology, Atkins and colleagues (2012) used LDA topic modeling to analyze transcripts of therapy sessions and communication assessments from couples in a randomized trial comparing two therapy treatments, traditional behavioral couple therapy and integrative behavioral couple therapy (Christensen et al., 2004). Some of the topics discovered by the topic modeling algorithm included Family, Finances, Negative Emotional Content, and therapy process-related topics. For example, the algorithm established that the words “mom,” “mother,” “dad,” “sister,” and “brother” tended to occur together in transcript documents. The authors interpreted this to be a topic about family. Similarly, the algorithm

established that the words “money,” “dollars,” “buy,” and “hundred” tended to occur together. The authors interpreted this to be a topic about finances.

Because a subset of the therapy sessions in the study were coded for behavioral items, Atkins and colleagues were then able to fit a model to evaluate the association between the linguistic content of couples’ therapy sessions and behaviors within those therapy sessions (e.g., constructive problem-solving skills, blame, positive emotion). They were indeed able to establish a general concordance between external outcomes and the topics covered in therapy. For example, interestingly, the association between blame behavior and topics covered in a session was more subtle than simply negative emotionality. Rather, blame behavior was related to statements that, while not overtly critical, “functioned to create an emotional distance from a spouse” (p 826).

**Features: Imposing Structure on Unstructured Data.** Generating structured data from unstructured data involves selecting the right features. These are the measurements derived from the raw, unstructured data. These features are tied to the types of analyses that can be run. Text data, for example, could be given a categorical variable structure (e.g., the political party of the speaker in a speech transcript, the marital status of the couple in a therapy transcript) or a continuous variable data structure (e.g., percent usage of positive emotion words).

Vector representation and, more generally, matrix representation is a common way of representing features (Flach, 2012). In text analysis, documents are often represented as vectors representing word frequencies. In one type of representation, each element in the vector corresponds to the number of occurrences of a particular word. (This class of representation is known as *bag of words* because it does not preserve structural information such as word order.)

For example, suppose we are interested in using the words *gene*, *twin*, *implicit*, and *attitude* in an analysis of a corpus of psychology papers:

$$\begin{bmatrix} gene \\ twin \\ implicit \\ attitude \end{bmatrix}$$

A document (perhaps about behavioral genetics) in which the word *gene* occurs 10 times, the word *twin* occurs 20 times, the word *implicit* occurs zero times, and the word *attitude* occurs zero times would be represented as:

$$\begin{bmatrix} 10 \\ 20 \\ 0 \\ 0 \end{bmatrix}$$

A document (perhaps about implicit attitudes) in which the word *gene* occurs zero times, the word *twin* occurs zero times, the word *implicit* occurs 10 times, and the word *attitude* occurs 20 times would be represented as:

$$\begin{bmatrix} 0 \\ 0 \\ 10 \\ 20 \end{bmatrix}$$

Finally, a document about the heritability of implicit attitudes might have the following structure:

$$\begin{bmatrix} 10 \\ 10 \\ 10 \\ 10 \end{bmatrix}$$

Additionally, text feature representations are often weighted based on other factors. A common weighting is term frequency-inverse document frequency (TF-IDF: Spärck Jones, 1972) which gives less frequent, more specific words more importance than words that frequently occur across all documents (such as the word “the”).

In practice, the vocabulary of a corpus of millions of documents might comprise thousands of words. A typical text analytics machine learning algorithm might generate millions of vectors, each thousands of elements long, to represent the documents in a corpus. Again, note that many types of features can be represented using this kind of matrix form.

**Stop words.** In many text analyses, some words, known as *stop words*, are excluded from the analyses. Stop words are words that are considered to not carry important content for a particular analysis. Typical stop words include articles (e.g., a, an, the), pronouns, conjunctions, and numbers. Removing stop words can assist with file size and memory usage issues during analyses. Importantly, for some analyses, these words do not carry useful information, while for others they do (Grimmer & Stewart, 2013). For topic modeling analysis, for example, removing stop words aids in the interpretation of the resulting topics. The decision to remove certain words, and which words to remove, thus depends on each analysis and this decision should be reported. It is important to understand when stop words are used by an analytics tool and, if they are used, what the stop words are.

**Software packages.** A number of software packages implement these text analytics algorithms. For LSA and LDA, implementations include gensim (Řehůřek & Sojka, 2010). For LDA, implementations include MALLET (McCallum, 2002), the R packages lda (https://cran.r-project.org/web/packages/lda/index.html) and topicmodels (https://cran.r-project.org/web/packages/topicmodels/index.html), and TACIT (http://tacit.usc.edu/; Dehghani et al., 2015).

Also, several software packages provide broad frameworks offering common NLP functionality, such as word counting and text cleaning (e.g., punctuation removal). For R, the package tm (https://cran.r-project.org/web/packages/tm/index.html) offers a variety of text

mining functions. NLTK (http://www.nltk.org/) provides NLP functionality for Python. The Stanford CoreNLP toolkit (Manning et al., 2014) is a widely-used JAVA-based framework.

**Multimedia Analytics**

The analysis of multimedia, including audio, images, and video, is a relatively new area of research. Multimedia analytics covers a wide range of data formats from a variety of sources. Analysis of emotion content in music, for example, can involve using information in the acoustic signal, lyric analyses, and metadata such as user tags (H. H. Chen, 2011). Techniques for the analyses of video data include methods to detect novel events in video data using an algorithm derived from human eye tracking behavior (Kang, Aurangzeb Ahmad, Teredesai, & Gaborski, 2007), and can, for example, detect soccer goal shots based on a combined analysis of video and audio data (M. Chen, Chen, Shyu, & Zhang, 2007).

Wojcik and colleagues' (2015) Study 2 examined the emotion expression of members of the U.S. Congress. The goal of the study was to assess the association between expressions of positive emotion and political ideology in members of the U.S. Congress. Expressions of emotion in this study were assessed through automated linguistic analyses combined with facial emotion expression analyses of photographs from the Congressional Pictorial Directory of the members of the 113th U.S. Congress (United States Government Printing Office, 2013). The analyses of the photographs were conducted by a human coder, certified in the Facial Action Coding System (FACS: Ekman & Friesen, 1978).

**Data Mining and Machine Learning Overview**

Data mining involves finding associations and patterns in data in order to predict some useful outcome (Fernandez, 2010). In the sense that qualitative data analysis involves examining "bits of information in the data, and looking for similarities and differences within these bits to

categorize and label the data" (Walker & Myrick, 2006, p. 549), data mining bears some similarity to the "open coding" process (Hesse-Biber & Leavy, 2011). Data mining also proceeds from a fine-grained, bottom-up analysis of the data that aims to take into account as many features of the data as possible. These analyses can be used by researchers in psychology on both data from traditional studies and data acquired from big data sources.

One of the most notable uses of big data is to use these bottom-up analyses to build models that make predictions beyond the range of the collected data, e.g., recommend new movies, locate optimal product placements in stores (Mayer-Schönberger & Cukier, 2013). Data mining often relies on machine learning, which is essentially the same as pattern recognition (Bishop, 2006). Conceptually, machine learning involves "using the right features to build the right models that achieve the right tasks" (Flach, 2012, p. 12). At its core, the goal is to extract useful information from data.

As such, these techniques can often be used with datasets that are not big data-scale, and this applies to data from studies in all areas of psychology. Indeed, as will be noted below, many data mining and machine learning versions of techniques such as regression and cluster analysis are very similar to those already in wide use in psychology. Although not every dataset from a typical psychological study can be usefully analyzed with these techniques, for those that can, these techniques offer an approach that integrates many features of a dataset to yield new insights.

**Software packages.** There are many freely available software packages that implement a wide variety of data mining and machine learning techniques. The following are just a few of the packages. R (https://www.r-project.org/) is widely used for data mining analyses and many additional packages have been developed for these analyses (https://cran.r-project.org/). For

Python, scikit-learn (http://scikit-learn.org/stable/; Pedregosa et al., 2011) is widely used. The Waikato Environment for Knowledge Analysis (WEKA) is a standalone general machine learning package with a graphical user interface (http://www.cs.waikato.ac.nz/ml/weka/; Hall et al., 2009). TACIT (http://tacit.usc.edu/; Dehghani et al., 2015) implements several of these techniques for text analysis.

The following is an overview of several types of data mining analyses. Data mining techniques can be divided into supervised learning methods and unsupervised learning methods.

**Supervised Learning**

Supervised learning algorithms aim to characterize the associations between a set of feature (predictor) variables and a target variable by learning their relationships from the data. Each data sample is represented by a set of feature variables and a target outcome variable. Supervised learning involves training an algorithm (i.e., building a model) on a training dataset in which the outcome of the target variable is known. Knowing the answers, the algorithm can deduce a model that best fits the data to the answers. Typically, the goal of supervised learning is to build a model that will be used on future data to predict as yet unknown outcomes. For example, a model based on customers' existing movie viewing preferences can be used to predict how much a customer will enjoy viewing a new movie (e.g., the Netflix prize: http://www.netflixprize.com/).

These big data techniques are geared toward generating models that provide actionable information. In psychology, there are many potential applications of these algorithms, such as in the construction of complex models to predict adverse mental health outcomes or to identify at-risk youths.

**Classification.** Classification involves identifying the class (i.e., category) that a particular instance or observation belongs to, based on its features. Classification is typically a supervised learning task: the classes are specified prior to the training of the classifier. For example, along the lines of previous work (e.g., Atkins et al., 2012, Imel et al., 2015), based on linguistic features of transcripts of interactions of couples, each text could be classified by whether, after five years, the couple remained together. Common classification algorithms include classification trees and random forests, support vector machines (SVM: Cortes & Vapnik, 1995; Joachims, 1998), k-nearest neighbors, and logistic regression (Flach, 2012; Wu et al., 2008).

A simple example of supervised learning would be to give an algorithm that aims to classify flowers. This algorithm is given a training set of data about red roses and blue violets, with feature data about color, red or blue, and flower species, rose or violet. The goal is to build a model that, given a new flower, will predict whether the flower is a rose or a violet. A simple solution would be to build a model that predicts that all red objects are roses, and all blue objects are violets.

Tutorial 3, in the Supplemental Materials, builds a classifier using the support vector machine approach. This tutorial demonstrates how machine learning can be used on even simple datasets.

***Feature extraction.*** As noted above in the text analytics section, one use of classification techniques is to extract the features that the algorithm has found to be the most important for determining how to classify the instances. For example, in the classification of text documents, certain words may be most indicative of one speaker or another. Feature extraction is useful in

the selection of features for model building. The Recursive Feature Elimination variant of support vector machines (SVM-RFE: Guyon et al., 2002) is one important approach.

**Regression.** Regression analyses in data mining are essentially the same as regression analyses used in non-big data settings. In machine learning contexts, regression analysis aims to identify the set of feature variables that have informative relationships with the target variable. The goal is to build a model that can be used to predict the outcome of new data samples.

Big data regression models often include many predictors, many of which are uninformative. One method to handle these uninformative predictors is to use regularized regressions. This method helps make the results more interpretable by reducing the number of coefficients that are non-zero by shrinking the coefficients (Flach, 2012; Raschka, 2015). Less informative predictors are reduced to zero or essentially zero. This aids in model selection by eliminating uninformative variables. One approach to regularization uses ridge regression (Hoerl & Kennard, 1970). Another uses the least absolute shrinkage and selection operator (lasso) approach in which some coefficients are set to zero, and others are shrunk (Tibshirani, 1996).

**Model assessment with training data and test data.** When creating a predictive model, the goal is to use the model to predict the outcomes of future data. Yet how do we know how it will perform on these future, as yet unseen data? The answer is to "generate" unseen data from the existing dataset. Basically, a dataset is divided into two subsets: a training set and a test set. The test set is held out of the model building process, and is only used to validate the resulting model. During the validation phase, this test set serves as the unseen data on which to test the model. This allows for an unbiased estimate of the performance of the model. Split sample and k-fold cross validation are two commonly used approaches to assess the performance of a classification model (Raschka, 2015). Binary classification models, for example, are commonly

evaluated in terms of sensitivity (the proportion of correct identification of positive samples), specificity (the proportion of correct identification of negative samples), and accuracy (Hastie, Tibshirani, & Friedman, 2001). In addition, measures and statistical tests that are used in contexts outside of data mining for the assessment of the quality of a model (goodness-of-fit), and model comparison and model selection are also used in data mining. These include deviance, Akaike information criterion, and Bayesian information criterion (Hastie et al., 2001).

**Unsupervised Learning**

In general, unsupervised learning methods aim to build descriptive models to represent interesting structure or patterns in the data (Flach, 2012). In contrast to supervised learning, the data samples are not labeled with outcomes; no information about outcomes is given and there is no target variable. The variables are exploratory—results arise upwards from the data based on the analyses.

**Clustering.** Cluster analysis uses similarities and differences between samples to find hidden patterns and structures in the data (Jain, 2010; Jain et al., 1999). Samples that are similar to each other, based on particular attribute variables, are grouped together into clusters. Samples that are in different clusters should be less similar to each other than samples that are in the same cluster. For example, in analyses of social network users, group mining (user clustering) can identify users who share common attributes. These results can be used to make recommendations for new users to follow or befriend. Cluster analysis has a long history in psychology including Cattell's (1943) work in identifying clusters of personality traits. There are several popular clustering approaches used in data mining (Wu et al., 2008) including k-means (Lloyd, 1982) and hierarchical clustering (Hastie et al., 2001).

**Association rule learning.** Association rule learning aims to identify associations hidden in large datasets (Agrawal et al., 1993; Agrawal & Srikant, 1994). These associations contain an antecedent and a consequent. Following an example from Agrawal and colleagues (1993), take the following rule, referring to a hypothetical dataset of supermarket transactions:

{Bread, Butter} → {Milk} with confidence factor of 90%

This states that 90% of those who purchase bread and butter also purchase milk. This might find use in psychological research for those interested in using an unsupervised learning approach to exploratory data analysis. One possibility might be to mine a dataset for disorder comorbidities. The Apriori algorithm (Agrawal & Srikant, 1994) is widely used. Data mining software packages such as WEKA (http://www.cs.waikato.ac.nz/ml/weka/; Hall et al., 2009), as well as packages for R (including RWeka: https://cran.r-project.org/web/packages/RWeka/index.html), implement association rule analyses.

## Network Analysis

Network analysis provides a "precise formal definition to aspects of the political, economic, or social structural environment" (Wasserman & Faust, 1994, p. 3). A network provides a structure to describe complex interconnected relationships in terms of nodes and edges. For example, in the World Wide Web the nodes are the web pages and the edges are the hyperlinks; in a citation network the published papers are the nodes and the links are the references to previously published papers. A social network is constructed to study relationships between individuals, groups, or organizations (social entities). Social network analysis investigates structures of social entities to identify local and global patterns, and locate influential entities and interactions (Borgatti, Everett, & Johnson, 2013; Easley & Kleinberg, 2010; Kadushin, 2012).

This important class of analyses is, so far, less widely used in data mining than in fields such as sociology, which has a long history of social network analysis including with social network site data (e.g., Lewis et al., 2012). Work in psychology has also begun to examine network relations (e.g., Barberá, Jost, Nagler, Tucker, & Bonneau, 2015). UCINET is a widely-used network analysis software package (https://sites.google.com/site/ucinetsoftware/home; Borgatti, Everett, & Freeman, 2002) in network analysis. The RSiena package (https://www.stats.ox.ac.uk/~snijders/siena/) for R and Pajek (http://mrvar.fdv.uni-lj.si/pajek/; Batagelj & Mrvar, 2003) are also popular.

**Visualization**

Common graphing techniques such as histograms, box plots, and scatter plots are widely used in big data analytics. In addition, to assist with understanding high dimensionality data, matrix visualization techniques have also been used (Chen et al., 2007; Jacoby, 1998). Statistical packages, including R, and data mining and machine learning software include visualization packages, such as matplotlib for Python. In addition, for graphing networks, Gephi (https://gephi.org/), d3.js (https://d3js.org/), and graphviz (http://www.graphviz.org/) are commonly used software packages.

**Analyzing Big Data as Big Data with MapReduce**

One of the challenges in analyzing massive datasets is that it can quickly far outstrip the computing power of a single machine. The big data approach is to combine multiple machines with typical processors and memory. However, managing and synchronizing the operation of multiple machines running in parallel creates its own set of challenges. Addressing these problems, Google's MapReduce (Dean & Ghemawat, 2008) has been revolutionary. MapReduce provides a framework for parallel, distributed computing. Essentially, the programmer translates

a given analytical problem into a MapReduce job by identifying a way in which the problem can be broken up into a repeating, independent process (or set of such processes) that executes a core task of the problem. Breaking up the problem into independent processes allows multiple, independently-operating computers to carry out the analyses in parallel. The outputs of these processes are then combined in such a way that it represents a solution to the original problem. Some analyses in the statistical programming language R, for example, can be translated to run on a cluster of computers using the MapReduce framework. In order to do so, it must be possible for these analyses to be fit into the MapReduce paradigm.

This algorithm was published and made open to the community, allowing others to implement and make available MapReduce frameworks. Hadoop (http://hadoop.apache.org/), by the Apache Software Foundation, is a free implementation and has been widely adopted. The Hadoop framework includes a distributed filesystem, the Hadoop Distributed File System (HDFS), to store data across multiple machines; an implementation of the MapReduce algorithm, Hadoop MapReduce, to analyze these data using multiple computers; and Hadoop YARN, a scheduler and cluster resource management module. Hadoop can be linked to other software so that programs can take advantage of this distributed processing.

It is beyond the scope of this article to fully explain the MapReduce algorithm, describe how to translate analytical problems into a MapReduce job, deploy a Hadoop cluster, etc. But there are a number of resources available for learning about distributed computing in data science (e.g., Apache Software Foundation, 2016; Lescovec et al., 2014).

Distributed processing frameworks are complex and require the right expertise to manage, and for researchers analyzing small data generated from big data, it will not be needed. Adler (2012, p. id592042) provides this set of guidelines for deciding to consider using Hadoop:

> "1. You cannot solve the problem with one machine, even after shrinking your data or expanding the machine.
>
> 2. It's possible to formulate the problem as a Map/Reduce problem. Many, but not all, important problems fit into a Map/Reduce model.
>
> 3. You have the right expertise to run a Hadoop cluster. Companies that actively use Hadoop often have teams of people to manage Hadoop."

**Additional Resources**

There are further online and offline resources for both specific and general analytics issues. Some relevant resources will be particular to an analytical task or software package. For example, for text analysis, Iliev and colleagues' (2014) review provides an excellent overview of automated text analysis techniques, geared toward psychologists. NLTK is accompanied by extensive online resources, as well as a book (Natural Language Processing with Python, http://www.nltk.org/book/: Bird, Klein, & Loper, 2009) covering basic programming concepts and common NLP tasks.

For overall machine learning background—including both supervised and unsupervised learning, the following are some starting points. Flach (2012) discusses conceptual and mathematical aspects of machine learning within a holistic framework. Hastie and colleagues (2001) cover the statistics underlying many machine learning techniques. Raschka (2015) offers a step-by-step guide on how to conduct various machine learning analyses in Python.

## Discussion and Conclusion

Thinking about how to generate small data from big data allows us to realize that big data research in psychology is not simply or necessarily research with massive datasets. The focus can be on the *data* part of Big Data. A novel dataset based on real-world behavior can contribute

new insights and provide evidence that complements studies using traditional research methods. Nevertheless, big data research also presents many new challenges.

A checklist of key questions to consider at various stages throughout a big data project is given in Appendix B. These questions raise important conceptual, methodological, technical, and ethical issues that should be thought through, as necessary.

**Domain Knowledge**

Big data tends to be complex and messy, and there are considerable methodological challenges in big data analytics associated with this. One common challenge is that, often, there are a large number of variables. When the number of variables (high dimensional data) is large, there are issues of over-fitting, spurious clusters, spurious correlations, multiple testing (controlling for false positives), feature selection, sparseness, etc.

Another problem encountered in big data analytics is that since the sample size is huge, it is virtually impossible *not* to conclude statistical significance. This is an issue of statistical significance versus practical significance. Meehl (1990) and Cohen (1994) have argued that, essentially, every psychological phenomenon is related to every other phenomenon in some way—everything is part of a web of an unknown multitude of interconnected causal factors. A particular association may be unimportantly small, but with a large enough sample size, it can be statistically significant. In big data research, it is therefore crucial to be clear about what constitutes importance. One remedy is that, in addition to the p-value, its associated estimate (e.g., effect size or correlation coefficient) should be evaluated. At the planning stage, researchers should specify how to decide whether a finding is important and practically significant. Is there a threshold for expected effect sizes above which an effect size is important? Is there some amount of explained variance above which a set of variables offers important

explanatory or predictive power? Is there some level of accuracy above which a classifier becomes useful?

Building predictive models can run the risk of ignoring whether or not they have any explanatory power (Lin, 2015). If ice cream sales are useful in a model built to predict homicide rates, the naïve approach to predictive analytics would include them in the model. Thus, having a clear theoretical focus is essential. An arsenal of data science analytical techniques is available, but there is no single best algorithm—there are "no free lunches" (Wolpert, 1996; Wolpert, 2002), and subject knowledge is essential in evaluating what algorithm to apply. As with other forms of analyses, big data analytics complement rather than replace traditional research.

**Security and Privacy**

The behavioral sciences already have a strong foundation in securing data and protecting the privacy of participants. Issues of privacy, confidentiality, and anonymity are central principles of institutional review boards, which already frequently handle studies dealing with sensitive information such as criminal records and health records. Institutional review boards must play a central role in extending and adapting these principles to big data research. In understanding how to handle these issues, psychology is, in many ways, already far ahead of businesses and others who use big data. Nevertheless, the degree to which big data accesses and measures everyday behaviors is unprecedented and raises many serious new issues. These issues are also important given the movement in psychology toward archiving and distributing data (e.g., the Open Science Framework: https://osf.io/).

For example, protecting identities is not merely a matter of deleting or masking a person's name or other specific identifier. Given the size of big data, identities can also be reconstructed by combining pieces of information, each of which would not be enough to

identify a person but, combined, would allow individuals to be identified (Berman, 2013). When data are made open—which frequently occurs with big data, additional care must be taken to ensure that individuals cannot be identified in unexpected ways.

Psychological research overall has adopted a higher standard of conduct above that of the general public. Just because a set of data is publicly available does not automatically mean that it can be ethically used for research. The principle of informed consent suggests that researchers should take into consideration how people expect their data to be used. Going forward, such issues may need to be evaluated on a case-by-case basis.

**Critical Questions**

Unlike the direct observations that researchers in traditional naturalistic field studies make, big data observations are typically obtained via an entity besides the researcher. Medical records are recorded by the healthcare providers and shopping habits are recorded by the vendors. Online posts are made on a particular platform (e.g., Twitter, Instagram). These behaviors are mediated through the structure of the data source. The manifestations of a psychological process would look very different in Twitter's 140-characters-or-less format compared to Reddit's call-and-response conversational format. Much of the power of a platform comes from specific restrictions and affordances, restrictions and affordances that channel psychological processes in unique ways. In addition, sites often have editorial policies that restrict certain types of behaviors, such as those deemed offensive or illegal. Ensuring the veracity of big data requires understanding the idiosyncrasies of the data sources (Berman, 2013; Ruths & Pfeffer, 2014).

Big data research has bounds. Like much about big data, these bounds are, to date, fuzzy. The size of big data datasets may tempt us to overlook their limits (boyd and Crawford, 2012).

Access to new technology is often limited to higher socioeconomic classes. Data may be easily accessible, but the sources of those data may be unaware of just how accessible the data are and in what unexpected ways they might be used. boyd and Crawford (2012) provide a good entry point into some of these critical issues.

**Are you Equipped to do a Big Data Research Project?**

For most psychologists, the biggest challenge will be the software engineering requirements, and there may be hardware challenges as well. The analytical techniques are, conceptually, similar to—if not the same as—the analytical techniques already in wide use in traditional psychological research. However, interfacing with and handling data from highly variable real-world sources requires understanding the underlying principles on which they operate. Accessing a social media source through the internet, for example, may require delving into how internet protocols work. Being able to understand the underlying algorithms and implementations of one's tools is crucial to the internal validity of research (Ruths & Pfeffer, 2014; Vihinen, 2015). Thus, critical use requires being able to understand how algorithms and techniques are implemented.

**Conclusion**

Big data research offers insight into real-world behaviors, though taking on a big data project requires significant technical skills. Nevertheless, for those willing to take on such a commitment, innovative analyses of unexpected or previously untapped data sources can offer fresh ways to develop, test, and extend theories. The big data approach provides structure to and extracts analyzable features from raw information about people, and derives quantitative answers to substantive research questions. Despite the potential pitfalls of automated analytical techniques (Grimmer & Stewart, 2013; Iliev et al., 2014), the care required in effectively

operationalizing independent and dependent variables can prompt researchers to sharpen their thinking about their theoretical constructs, just as it can with traditional forms of field research (Paluck & Cialdini, 2014). When conducted with care and respect, big data research can become an essential complement to traditional research.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, *22*(2), 207-216.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, *1215*, 487-499.

Apache Software Foundation. (2016). Apache Hadoop documentation. Retrieved from: http://hadoop.apache.org/docs/current/

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology*, *26*(5), 816.

AWS. (n.d.). Public data sets on AWS. from http://aws.amazon.com/public-data-sets/

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right Is Online Political Communication More Than an Echo Chamber? *Psychological science*, *26*(10), 1531-1542.

Batagelj, V. & Mrvar, A. (2003). Pajek - Analysis and Visualization of Large Networks. In Juenger, M. & Mutzel, P. (Eds.), *Graph Drawing Software* (pp. 77-103). Berlin: Springer.

Berman, J. J. (2013). *Principles of big data: Preparing, sharing, and analyzing complex information*. Waltham, MA: Elsevier.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media, Inc. http://www.nltk.org/book/

Bishop, C. M. (2006) *Pattern recognition and machine learning.* New York City: Springer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84. doi: 10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of machine Learning Research, 3*, 993-1022.

Boe, B., Pedersen, A. D., & Mellor, T. Python Reddit API Wrapper. Retrieved from https://github.com/praw-dev/praw

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for Windows: Software for Social Network Analysis.* Harvard, MA: Analytic Technologies.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks.* Los Angeles: Sage Publications.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

boyd, d., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210-230.

Capozio, A. (2015). tumblR (Version 2). Retrieved from http://cran.r-project.org/web/packages/tumblR/index.html

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology, 38*(4), 476-506.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 27:21-27:27.

Chen, C., Hardle, W., & Unwin, A. (2008). *Handbook of data visualization.* Berlin: Springer-Verlag.

Chen, H. H. (2011). *Music emotion recognition*. Boca Raton, FL: CRC Press.

Chen, M., Chen, S.-C., Shyu, M.-L., & Zhang, C. (2007). Video event mining via multimodal content analysis and classification. In V. A. Petrushin & L. Khan (Eds.), *Multimedia data mining and knowledge discovery* (pp. 234-258). London: Springer London.

Christensen, A., Atkins, D. C., Berns, S., Wheeler, J., Baucom, D. H., & Simpson, L. E. (2004). Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. *Journal of Consulting and Clinical Psychology*, *72*(2), 176.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Culpepper, S. A., & Aguinis, H. (2011). R is for Revolution A Cutting-Edge, Free, Open Source Statistical Package. *Organizational Research Methods, 14*(4), 735-740.

Danneman, N., & Heimann, R. (2014). *Social media mining with R*. Birmingham: Packt Publishing.

Data, P. O. (n.d.). Project Open Data. from https://project-open-data.cio.gov/

data.gov. (n.d.). Search for a dataset. from http://catalog.data.gov/dataset

De Choudhury, M., Lin, Y.-R., Sundaram, H., Candan, K. S., Xie, L., & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media? *ICWSM, 10*, 34-41.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113.

Dehghani, M., Johnson, K., M., Garten, J., Balasubramanian, V., Singh, A., Shankar, Y., Rajkumar, A., Parmar, N. J., Hoover, J., Pulickal, L., & Boghrati, R.

Tacit: An Open-Source Text Analysis, Crawling and Interpretation Tool. Retrieved from SSRN: http://ssrn.com/abstract=2660651. Software available at: http://tacit.usc.edu/

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2012). Language and ideology in congress. *British Journal of Political Science, 42*(01), 31-55. doi: doi:10.1017/S0007123411000160

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). Social media update 2014: Pew Research Center.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and Markets: Reasoning about a highly connected world.* New York: Cambridge University Press.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: Investigator's guide*. Washington, D.C.: Consulting Psychologists Press.

Facebook. (n.d.). Graph API. from https://developers.facebook.com/docs/graph-api

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., & Martone, M. E. (2014). Big data from small data: Data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience, 17*, 1442-1447. doi: 10.1038/nn.3838

Fernandez, G. (2010). Data mining: A gentle introduction *Statistical Data Mining Using SAS Applications, Second Edition* (pp. 1-14). Boca Raton, FL: CRC Press.

Flach, P. (2012). *Machine learning: The art and science of algorithms that make sense of data.* Cambridge: Cambridge University Press.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144. doi: http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101,* 5228-5235.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*, 267-297. doi: 10.1093/pan/mps028

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning, 46,* 389-422.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, R., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*, 10-18.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer.

Hesse-Biber, S. N., & Leavy, P. L. (2011). *The practice of qualitative research* (2nd ed.). Los Angeles: Sage.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Iliev, R., Deghani, M., & Sagi, E. (2014). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition, 7*, 265-290. doi: doi:10.1017/langcog.2014.30

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy, 52*(1), 19.

Indurkhya, N., & Damerau, F. J. (2010). *Handbook of natural language processing* (2nd ed.): Chapman & Hall/CRC.

Internet Engineering Task Force. (2005). Uniform Resource Identifier (URI): Generic Syntax. Retrieved from https://tools.ietf.org/html/rfc3986

Internet Engineering Task Force. (2006). Internet denial-of-service considerations. Retrieved from https://tools.ietf.org/html/rfc4732

Internet Engineering Task Force. (2009). TCP congestion control. Retrieved from https://tools.ietf.org/html/rfc5681

Jacoby, W. G. (1998). *Statistical graphics for visualizing multivariate data.* Thousand Oaks, CA: Sage Publications.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR), 31*(3), 264-323.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings.* C. Nédellec and C. Rouveirol. Berlin, Heidelberg, Springer Berlin Heidelberg**:** 137-142.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press.

Kadushin, C. (2012). *Understanding social networks.* New York: Oxford University Press.

Kang, J. M., Aurangzeb Ahmad, M., Teredesai, A., & Gaborski, R. (2007). Cognitively motivated novelty detection in video data streams. In V. A. Petrushin & L. Khan (Eds.), *Multimedia data mining and knowledge discovery* (pp. 209-233). London: Springer London.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment, 5*(12), 2032-2033. doi: 10.14778/2367502.2367572

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240. doi: 10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284. doi: 10.1080/01638539809545028

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*, 70-73.

Lescovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets.* Cambridge, UK: Cambridge University Press.

Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences, 109*(1), 68-72.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media, 57*(1), 34-52. doi: 10.1080/08838151.2012.761702

Lin, J. (2015). On building better mousetraps and understanding the human condition: Reflections on big data in the social sciences. *The ANNALS of the American Academy of Political and Social Science, 659*, 33-47. doi: 10.1177/0002716215569174

LinkedIn. (n.d.). Documentation. from https://developer.linkedin.com/docs

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing* (2nd ed., pp. 627-666). Boca Raton, FL: CRC Press.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129-137.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *ACL (System Demonstrations)*. 55-60.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved from http://mallet.cs.umass.edu/

McCallum, Q. E. (2012). *Bad data handbook.* Sebastopol, CA: O'Reilly Media, Inc.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*(1), 195-244.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books, T., . . . Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176-182. doi: 10.1126/science.1199644

Natural Language Toolkit. Retrieved from http://www.nltk.org/

Office, U. S. G. P. (2013). *Congressional Pictorial Directory of the 113th Congress*. Washington, D.C.: United States Government Printing Office Retrieved from http://www.gpo.gov/fdsys/pkg/GPO-PICTDIR-113/pdf/GPO-PICTDIR-113.pdf.

Office, U. S. G. P. (n.d.). United States Government Printing Office API Documentation. from https://github.com/usgpo

opencv. Retrieved from http://opencv.org/

Paluck, E. L., & Cialdini, R. B. (2014). Field research methods. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 81-100). New York: Cambridge University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547-577.

Phan, X.-H., & Nguyen, C.-T. (2007). GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA).

Raschka, S. (2015). *Python machine learning.* Birmingham, UK: Packt Publishing.

Raymond, E. S. (2000). *The cathedral and the bazaar* Retrieved from http://www.catb.org/esr/writings/cathedral-bazaar/cathedral-bazaar/

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, 46-50.

Russell, M. A. (2013). *Mining the social web: Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more.* Sebastopol, CA: O'Reilly Media, Inc.

Ruths, D., & Pfeffer, J. (2014). Supplemental material for Social media for large studies of behavior. *Science, 346*(6213), 1063-1064. doi: 10.1126/science.346.6213.1063

Sartor, G. (2011). Introduction: ICT and legislation in the knowledge society *Legislative XML for the semantic web* (pp. 1-10). New York: Springer.

Schroepfer, M. (2014). Research at Facebook [Press release]. Retrieved from http://newsroom.fb.com/news/2014/10/research-at-facebook/

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11-21.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424-440.

Sunlight Foundation. (2015). Capital Words API. from https://sunlightlabs.github.io/Capitol-Words/index.html

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267-288.

Twitter. (n.d.-a). API Reference documents: Streaming. Retrieved from https://dev.twitter.com/streaming/firehose

Twitter. (n.d.-b). About public and protected Tweets. Retrieved from https://support.twitter.com/articles/14016-about-public-and-protected-tweets

Twitter. (n.d.-c). Approving or denying follower requests. Retrieved from https://support.twitter.com/groups/50-welcome-to-twitter/topics/204-the-basics/articles/20169376-approving-or-denying-follower-requests

Twitter. (n.d.-d). API Overview. Retrieved from https://dev.twitter.com/overview/api

Vihinen, M. (2015). No more hidden solutions in bioinformatics. *Nature, 521*, 261.

Walker, D., & Myrick, F. (2006). Grounded theory: An exploration of process and procedure. *Qualitative Health Research, 16*(4), 547-559. doi: 10.1177/1049732305285972

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). New York: Cambridge University Press.

Wojcik, S. P., Hovasapian, A., Graham, J., Motyl, M., & Ditto, P. H. (2015). Conservatives report, but liberals display, greater happiness. *Science, 347*(6227), 1243-1246. doi: 10.1126/science.1260817

Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In Soft Computing and Industry (pp. 25-42). Springer London.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67-82.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems, 14*(1), 1-37.

Figure 1. Example Tweet data structure (abridged). This tweet's text is: "The text of the tweet. #example"

```
{"favorited": false,
"text": "The text of the tweet. #example",
"possibly_sensitive": false,
"in_reply_to_status_id": null,
"user": {
     "follow_request_sent": null,
     "profile_use_background_image": true,
     "id": #1,
     "profile_image_url_https":
     "https://pbs.twimg.com/profile_images/",
     "profile_sidebar_fill_color": "",
     "followers_count": ,
     "statuses_count":,
     "description": "Description",
     "friends_count": ,
     "location": "User's description of location",
     "name": "",
     "lang": "en",
     "created_at": "Mon Jan 01 01:01:01 +0000 2009",
     "time_zone": "Eastern Time (US & Canada)",
},
"geo": null,
"id": ,
"favorite_count": 0,
"lang": "en",
"retweeted_status": {
     "text": " The text of the tweet. #example",
     "in_reply_to_status_id": null,
     …
     "user": {"follow_request_sent": null,
          "profile_use_background_image": true,
          "default_profile_image": false,
          "id": #2,
          …
          },
     "geo": null,
     …
},
"entities": {
     …
"retweet_count": 0,
"created_at": "",
}
```

Figure 2. Topic modeling output: the ten topics generated by MALLET (from Tutorial 2).

```
0	5	service air veterans general army force u.s war military served national officer guard medal nation honor staff country corps defense

1	5	service president years served public law tribute career leadership california community member mrs board leader colleagues state city worked judge

2	5	team state national coach world year president indiana david congratulate record years hard championship high great game proud league basketball

3	5	university college state president history research years year professor education west montana york great art work institute washington time center

4	5	county communities local work development community iowa million programs economic state people public farm arkansas opportunities services health opportunity program

5	5	united states people government national bill efforts federal country energy american america issues work water organization americans nation working ensure

6	5	life family time lives children man day people age great george father made passed friends home born long love world

7	5	community today anniversary city health care south center town dakota north president great years medical area recognize park join people

8	5	business years company alaska small family colorado year today president maine success industry work state home businesses idaho continued job

9	5	school students education high award program community year children young schools national teacher outstanding work teachers youth middle elementary excellence

<1000> LL/token: -8.92572
```

Appendix A. Python code for retrieving Congressional Daily Digest data for January 24, 2005 via the GPO API.

```
import datetime
import urllib2

# Specify the type of document we're interested in.
request_type = 'sadditional'

# Set the date to request.
requested_date = "2005-01-24"

# Construct the command.
command = "http://api.fdsys.gov/link?collection=crec&link-type=html"
command += "&type=" + request_type
command += "&publishdate=" + requested_date

print "Requesting data for", requested_date

# Set up the connection to the server.
response = urllib2.urlopen(command,
                           timeout=30)

# Read the server's response.
gpo_data = response.read()

file_name = "crec_" + request_type + "_" + requested_date + ".htm"
with open(file_name, 'w') as output_file:
    output_file.write(gpo_data)
```

Appendix B. Checklist of key questions that may be useful to consider at various stages of a big data research project.

**Project Development**

What is the basic research question?

What are the data sources relevant to this question?

What are the analytical techniques applicable to the types of data available from these sources?

Are the operationalizations sound?

How will the importance of an effect or association be assessed? For example, what is the target effect size or amount of variance to be explained?

What are the ethical considerations?

Concerning the people from whom the data would be collected, how will their consent be obtained, if at all?

How will the data and software be archived? GitHub (https://github.com/), the Open Science Framework (https://osf.io/), and The Dataverse Project (http://dataverse.org/) are some possible repositories.

To whom will the data be made available?

**Managing the data**

How much raw data will there be?

In what format will the raw data be stored?

What data format does the planned analysis program require?

Do the data need to be anonymized and/or protected?

Will the data be redistributed? To whom? How?

Is a database required?

**Acquiring the data**

How can the data be acquired ethically and respectfully? Can the data be obtained by simply asking for it?

What are each data source's requirements and expectations?

What is the format of the data as generated by the source?

**Processing the data**

What is the format of the stored raw data?

For unstructured data, what are the relevant digital encoding standards?

What format is required for the analyses?

**Analyzing the data**

Given the analysis plan, what software package is most suitable? Some considerations include robustness of implementation, cost, and ease of use.

What validation method will be used for the analyses?

Is the available computing hardware sufficient to run the analyses? If not, an approach such as sampling from the full dataset may be necessary.

Are the interpretations warranted?