AN INTRODUCTORY GUIDE TO DATA SCIENCE: THE TERMINOLOGICAL LANDSCAPE

Abhinav Yedla

MS, Computer Science, Iowa State University

Shawn Dorius

PhD, Sociology, Iowa State University

January 2016

AN INTRODUCTORY GUIDE TO DATA SCIENCE: THE TERMINOLOGICAL LANDSCAPE

ABSTRACT

The emerging field of data science has rapidly evolved into an extremely diverse field equipped with multi-disciplinary techniques to extract, analyze and classify structured and unstructured data. These methods offer researchers, policy analysts, and the lay public evidence-based insights into a tremendous range of human, organizational, and societal activities on a scale and scope that has rarely been possible with conventional scientific methods. At present, however, the multi-disciplinary nature of the data science space suffers a 'language' problem insofar as data scientists from different fields often use different terms to describe common methods and concepts. The aim of the present research is threefold. First, we report results of a literature review that identifies and defines the essential content domain of data science, with special focus on the classification of data collection techniques. Second, we establish a preliminary set of relationships among the most trafficked terms of data science to facilitate interdisciplinary communication among scientists from heterogeneous fields. And third, we develop a classification scheme of web-scraping methods based on their availability, the quality of the data procured by the method, the ease of data extraction, reproducibility, the technical skills required to leverage each method, and the types of data collected by each method.

Introduction

The emergence of web-accessible data repositories, databases of structured, semi-structured and unstructured data, and a substantial increase in user-generated, organic data has created exciting new opportunities for data scientists to collect and analyze larger volumes of data with greater efficiency than has heretofore been possible. Such advances, enabled by the proliferation of web and mobile technologies that have become ubiquitous in daily life for many people and by emergent technologies in storage, computing power, and data generating utilities, have converged over a very short time period to produce a wealth of unanalyzed, and therefore untapped, data. Based on early studies of organic, user-generated data sources, such data contain a wealth of scientific insights. As with any emergent methodology, research is needed to establish the reliability, validity, and generalizability of such data sources to known populations, to establish theoretically grounded measurement constructs, and develop formal methods for collecting, cleaning, and analyzing organic data.

The work ahead is substantial, but the opportunities to advance theory, develop new methods, and generate wholly new scientific data on a massive scale suggest that such efforts have the potential for significant returns to human capital and infrastructure investment. One of the challenges for data scientists working in this new area of research is that a number of fields that have historically worked in relative isolation from one another[30] now find themselves engaged in many of the same research activities but with little common language to facilitate collaborations on large-scale social and methodological problems. This sort of 'Babel' phenomenon is not new, but nonetheless impedes greater integration between social scientists, computer scientists, and software engineers, for example. The purpose of the present research

is to provide data science with some guidance in establishing a common language for the heterogeneous disciplines that find themselves data science bedfellows.

We make four contributions that we hope represent a first step toward greater integration between the scientific stakeholders engaged in data science research activities. First, we report results of an inventory we conducted to identify the key terms and research activities of data science (DS). The purpose of the inventory is to provide a baseline conceptual structure to the emerging field of DS. Second, we report results of bibliometric analysis that demonstrates a number of important trends in the language of DS. The terminology inventory, bibliometric analysis, and relational maps allow us to establish a topline review of the data science landscape that captures the central terms used to describe research activities in two of the three stages of the data pipeline (data collection and data curation). We organize DS terms into a conceptual model that establishes relationships among the terms. In the third contribution, we provide a more detailed overview of activities involving the first stage of the data pipeline, collection of data, including a summary and synthesis of the foundational data collection activities of DS. We give special attention to web scraping techniques, which we organize into a hierarchical tree comprised of the key data scraping techniques. In our fourth contribution, we develop a baseline rating system that we use to rank data collection methods according to five criteria: data quality, ease of use, reproducibility, technical skills, processing errors, and data types. Our purpose in ranking the various methods is not to suggest there is any one best method, but rather to provide a roadmap for those who are new to DS and who may lack one or more of the technical skills necessary to engage in the full range of data collection methods available to data

scientists. We hope such a ranking schematic will serve to lower the entry barrier to DS for researchers working in diverse fields of study.

The Data Pipeline

We begin with an overview of a conventional data pipeline, also referred to as a data flow, or data analysis pipeline[31], which we depict visually in Figure 1. The idea of the data pipeline will be new to many scientists, owing to the somewhat slower and more deliberate process of data generation that has characterized much of science practice prior to the 21st century. Take, for example, data collection in the social sciences. From a historical perspective, social scientific data have most commonly been generated by design (GROVES CITE). That is, scientists developed a data collection instrument (typically a survey instrument, interview protocol, or archive mining strategy) that was designed for either a single data collection, as in the case of a cross-sectional social survey, or for intermittent data collections, as in the case of panel studies. With designed data at their control, social scientists have had the luxury of also controlling the pace of data generation, which allowed researchers to carefully plan each phase of data collection, including survey development, sample selection, instrument pre-testing, often a protracted data collection of weeks, months or more, followed by an equally deliberative approach to data cleaning and analysis. Such an approach has been instrumental in reducing errors at each stage of the data pipeline[32] and such data has been essential to the development of empirical social research.

With the proliferation of organic data, and especially with continually produced data (e.g. Twitter and RSS feeds, user reviews, Reddit discussions), data generation never stops. The constant flow of organic data has required researchers to reorient their understanding of data from a slow, highly controlled endeavor to one in which

data is dynamic, large-scale, and more difficult to harness. Transitioning from static to dynamic data collection is made easier when data is conceptualized as flowing along a pipeline. In the data pipeline, data is captured and continually move from a state of limited structure and refinement to one in which data are structured, tidied, tagged with relevant metadata, and otherwise transformed into an easily analyzable format[33]. Such tidy formats facilitate the development of static reports, dynamic dashboards, and, in many cases real-time monitoring of key metrics and indicators.

For conceptual purposes, we reduced the DS scientific activity to three major activity research areas, including *data collection*, *data curation*, and *data analysis* (a reasonable case made be made for a fourth area, *data communication*). Together, these areas involve research activities as diverse as statistical inference, algorithm development, tool development, data blending, database management, and visualization, but at the core of each is data. With the rise of social media and the Internet of Things (IOT) a vast amount of raw information is continuously produced and stored in data warehouses, many of which are freely available to the research community. There is much to learn by mining such information. One aim of DS is to use such enterprise level data in creative ways, to leverage the unprecedented volumes of human generated data in pursuit of business value and scientific insights, and to analyze human behavior at a very large scale.

Organic and designed data is being generated in virtually every area society, from large-scale transactional enterprises (e.g. bank withdrawals, consumer purchases, air traffic logs) to globally integrated environmental sensors in the most remote corners of earth and the solar system. For the sake of brevity, we note only three such sources in Figure 1, including the internet of things (IOT), offline reports, and web data, though data of scientific value can be collected from a great many other

sources. Data scientists develop protocols, scripts, packages, programs, and applications to capture data from sundry sources, which we discuss in more detail below.

The next step in the data pipeline centers on the activities that organize, clean, structure, harmonize, and otherwise prepare the data for analysis. Once the data is collected and tidy, it is ready for the final stage in the data pipeline: analysis. Data analysis comes in as many forms as do the data and as such, analysis might be exploratory, descriptive, inferential, or predictive; it might be univariate, bivariate, or multivariate; and the analysis might involve numbers, text, images, documents, or videos. Conceptualizations of the data pipeline vary by fields of study, but the model illustrated in Figure 1 provides a general overview of a data pipeline that should be familiar to most data scientists.

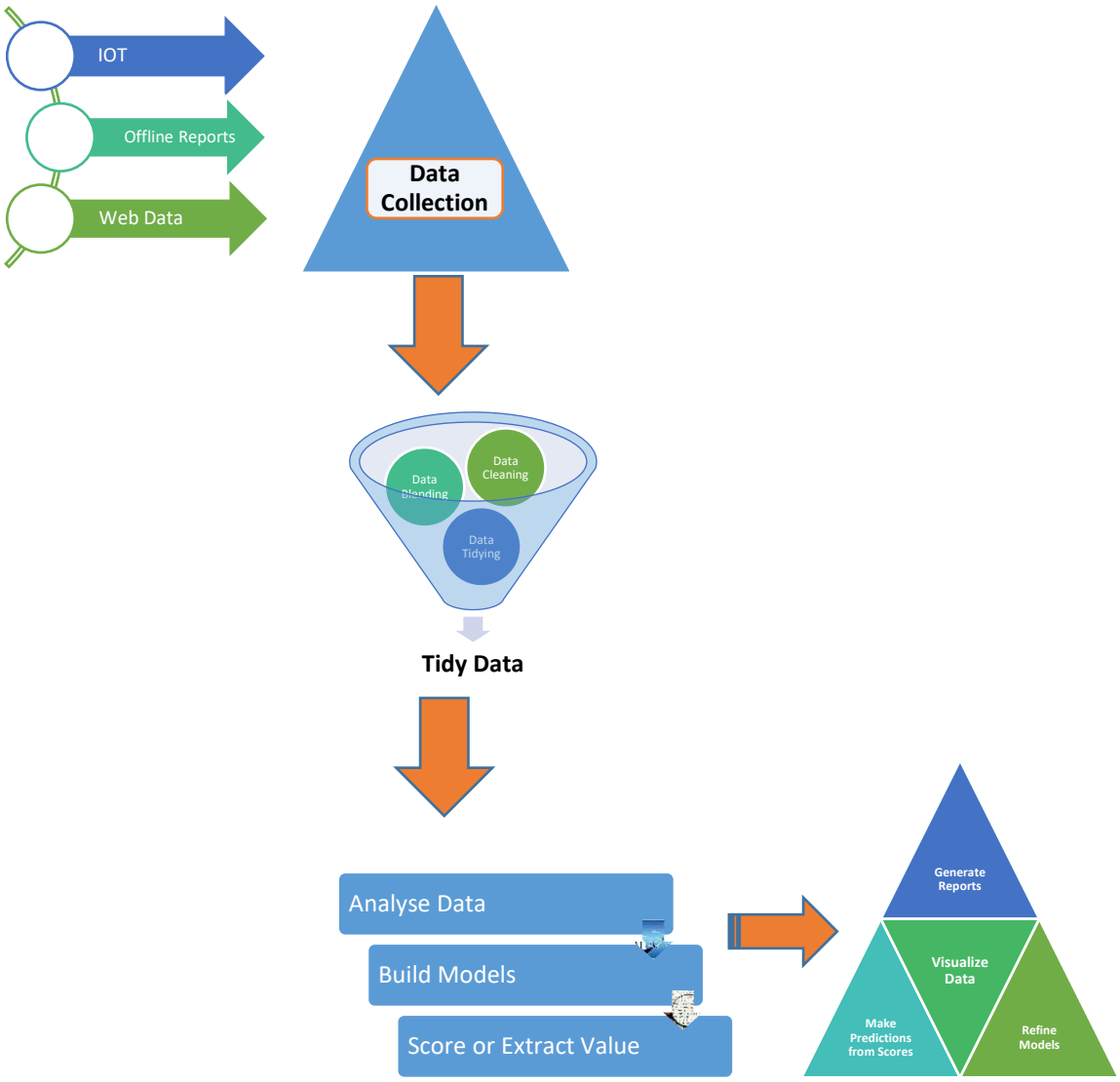

Figure 1: Illustration of a typical web-extraction data pipeline

In the next section, we will focus our attention on the first two stages of the data pipeline, including data collection and data cleaning/curation. We leave detailed discussion of the language of data analytics to future research.

Overview of Key Terms and Research Activities of Data Science

We now report results of a literature review that identifies salient DS terms. Our goal here is to provide a macroscopic view of the DS landscape and establish relations between terms that are sometimes synonymous, and other times indicators of

distinctive research activities. The term collection process involved a review of the DS literature contained in the foundational journals of DS and a wider review of the scientific and business literatures to develop a list of DS terms. We next reviewed the research articles to identify terms, which we then supplemented with additional terms found in online forums, pertinent message boards, published articles, and the like. These efforts yielded a core set of terms commonly used in DS to describe research activities involving the collection, cleaning, and curation of social scientific data.

We then used the list of terms to conduction a formal bibliometric analysis of the DS terminology landscape. Bibliometric analysis helped us to quantify terms according to most and least commonly used terms, and to identify patterns and trends in their usage, such as ‘vogue terms’ and those which have a long history of usage and clear, interdisciplinary meanings. Our bibliometric analysis was conducted using “Publish or Perish” program (PoP)[34], which scrapes citation data from the Google Scholar web database. PoP allows users to conduct bibliometric searches of specific terms, phrases or set of words and this can be done over specific date ranges and within targeted journals. The results were collected in a tidy data structure, with rows containing citations and columns containing citation metadata (authors, date, publisher, title, etc.). The metadata are useful for identifying the relative popularity and reach of particular articles, which, in turn, can be used to measure trends in the use of CSS terms. Following this process, we generated a citation database containing time-series data on a comprehensive set of DS terms, from which we were able to score terms according to common bibliometric statistics.

We report the list of terms in Figure 2 using a word cloud of most commonly used words in the DS literature, with terms scaled in size according the total number of citations (occurrences) in the literature. As can be seen, data collection, design data,

qualitative data, quantitative data, and primary data have, to date, been among the most commonly used in the DS literature. The list of terms shows that the most common terms (e.g. data collection, data visualization) are also the most general, though this is partly a function of the more extensive set of terms to describe the particular activities of DS.

Figure 2: Cumulative Number of Citations Containing Data Science Terms in Published Scientific Literature, 1980-2015

The centrality of these terms should not be interpreted to mean that all stages in the data pipeline have received equal attention in the scientific literature. In fact, 'data collection' has received far more attention than data cleaning or data analysis. This can be seen in Figure 3, which plots the usage of data collection, data cleaning, and data analysis—the three key DS research areas—based on the occurrence of each term in English language books. According to the Google Books Corpus[35], data collection is far more prevalent in books than either data analysis or data cleaning. This is not unsurprising, considering the relatively limited research attention to data processing[36].

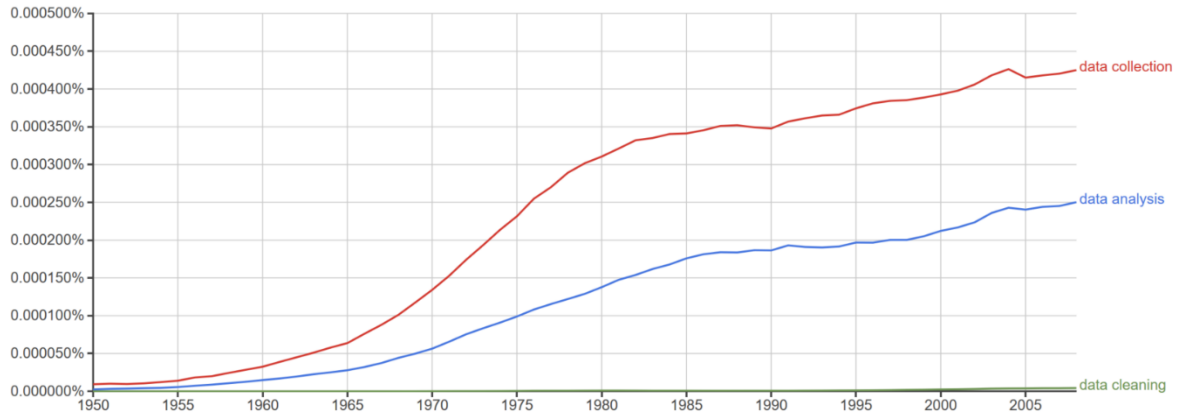

Figure 3: Annual Share of Occurrences of Data Science Terms in English Language Published Texts, 1950-2008

Term Definition

This section defines the terms that have, to date, been most prominently used in data science. All of the terms were collected from peer reviewed scientific journals, with occasional supplement from other online resources. This is not an exhaustive list of every term used to describe the foundational activities of data science, but they do represent many of the most commonly used terms. We organize terms according to their position in the data pipeline. We begin with

Data

Designed Data: Data is said to be 'designed' when it has been purposefully generated for business or scientific purposes. Designed data typically require less work to prepare for analysis and is often easily read and analyzed by commercial data analysis software. Examples of designed data include data collected by surveys, personal interviews, and focus groups.

Organic Data: Unlike designed data, organic data are typically generated as a byproduct (data exhaust) of user interactions with digital technologies. Such user interactions may be purposive, as in the case of consumers posting product reviews,

comments on message boards, or post to social media. Organic data also include ‘search’ data and data generated in the background of web and mobile devices, as in the case of GPS technologies that record and transmit location data to storage databases [22]. Organic data is considered ‘newer’ than designed data and the value of such data is often not clearly known to researchers at the time the data is generated.

Organic Search Results/Data: Organic search results are listings on search engine results pages that appear because of their relevance to the search terms. Organic search data stand in contrast to non-organic data generated by advertisements, as in the case of pay-per-click-advertising [13].

Quantitative data: Quantitative data is information about quantities, or information that is measured or summarized with numbers. Quantitative data are most typically associated with statistical analysis.

Qualitative data: Qualitative data is information about qualities, or information that cannot easily be described on understood with numbers [4]. Historically, qualitative data have been associated with text, often collected using conventional qualitative data collection techniques common in the social sciences, such as ethnographic field notes, personal interviews, and focus group transcripts. More recently, what constitutes ‘qualitative’ data has expanded to include images, videos, and other forms of non-numeric data. Such an enlargement of the understanding of what constitutes scientific ‘data’ has expanded the scope of scientific inquiry and made possible many exciting new discovery of human behavior and motivation.

Tidy Data: A dataset is said to be tidy if it satisfies the following conditions: 1) observations are in rows, 2) variables are in columns, and 3) data are self-contained in a single dataset. Tidy data is the primary format for analyzing data in spreadsheets and conventional statistical programs (e.g. Stata, SAS, SPSS) [14].

Course Data; Dirty Data: Course, or dirty, data refers to data that has not undergone a structuration process in preparation for analysis. Such data might be missing metadata, might be unstructured, containing missing or erroneous values (e.g. out of bounds values), or contain a mismatch between the data type and the data itself (e.g. integer data stored as string data).

Primary Data: Primary data is a term used most extensively among computational social scientists to refer to data that was designed and collected by the researchers for the express purpose of answering a specific research question(s). The term 'primary data' has historically been associated with survey, interview, and other observational data. Such data can be reused for other scientific purposes, but when this occurs, the data are typically then classified as 'secondary' data.

Secondary Data: Secondary data refers to data that was collected/generated for one purpose, but used for another purpose. Organic, user generated data is typically referred to as secondary data, but this occurs less frequently than designed data, where 'secondary data' is more commonly used.

Structured Data: Structured data refer to data that have been organized into a logical format that is readable by statistical program, algorithms, spreadsheets, scripts, and related software applications. In their most 'structured' format, data are organized into rows and columns and may containing quantitative or qualitative data.

Unstructured Data: Unstructured data is often associated with dirty, or course, data. Unstructured data may be found in reports, on websites, and in other formats in which the data are not readily analyzed by conventional methods due to their not being organized in a structure, standardized format.

Internet Data Collection

Data Extraction: Is the act or process of retrieving data out of original data sources for further data processing, analysis, or storage (see also data migration). Data extraction can involve structured or unstructured data [18].

Data Scraping: Is a data extraction technique in which a computer program extracts data from human-readable output, typically originating from another computer program [23].

Report Mining: Is the extraction of data from human readable computer reports. While report mining can, under limited conditions, be carried out manually, it most often involves automation [1].

Screen Scraping: Screen scraping is typically associated with the programmatic collection of visual data from a web source via a internet browser, instead of parsing data, as in Web scraping [1].

Web Scraping; Web Harvesting; Web Data Extraction: Is a computer software technique that extracts information from websites [26]. Data that is scraped/harvested/extracted can be quantitative or qualitative, and can be structured or unstructured.

Data Munging; Data Wrangling: Data munging and data wrangling are synonymous terms that refer to the collection, cleaning, and storage of disparate data. Munging/wrangling often involves the process of enhancing the ‘structure’ of data by organizing it into rows and columns, and enhancing data value with metadata. Such activities can be manual or semi-automated [11].

Fetching Data: Fetching data involves querying databases and extracting data in a structured format, typically involving the retrieval of data sequentially from rows. The term is more broadly used as a synonym for ‘getting data’, or ‘collecting data’.

Data Curation

Data Curation: Is a term used to refer to data management activities involving a) the organization and integration of data collected from various sources, annotation of the data, and publication and presentation of the data such that the value of the data is maintained over time, and the data remains available for reuse and preservation. Data curation includes "all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data" [1].

Data Cleansing; Data Cleaning; Data Scrubbing: These terms refer to procedures that detect and correct corrupt data from a record set, table, or database. The terms, which are primarily used in the context of spreadsheets and databases, are inclusive of the tasks of identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting dirty data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing with computer scripts [1].

Data Parsing: Literally meaning ‘parting data’, parsing data involves the splitting of data into two or more parts. Parsing is most commonly associated with string data, but quantitative data can also be parsed. By way of example, a string variable may contain a date in the format DDMMYY. For analytic purposes, it is often useful to analyze annual trends and test for seasonal effects. To do either, researchers would create three new variables, including day (DD), month (MM), and year (YY). Month data is used to measure seasonal effects and annual data is used to measures trends.

Data Blending: Typically conceptualized as a three-step process of data acquisition, data integration/fusion, and data cleaning. The key idea behind data blending is the joining of two or more datasets for the purposes of enhanced data insights.

Data Fusion: Is the process by which two or more data sets are integrated into a single one. The goal of data fusion is to bring together data from different sources and format them into a consistent, accurate, and useful representation of real-world phenomenon [21].

Data Harmonization: The process by which data from different sources, scales, and formats is transformed into a common scale for the purposes of making direct comparisons. Data harmonization is typically conceptualized and operationalized at the level of individual variables, rather than at the level of whole datasets. Data harmonization can involve either machine analytics or human coding. It combines both business-side uses of data as well as IT best practices for data quality [7].

Data Integration: Involves combining data residing in different sources and providing users with a unified view of these data. This process becomes significant in a variety of situations, which include both commercial (when two similar companies need to merge their databases) and scientific (combining research results from different bioinformatics repositories, for example) domains. Data integration appears with increasing frequency as the volume of data and the need to share it increases. Data integration has become the focus of extensive theoretical work, and numerous open problems remain unsolved [2].

Data Migration: Is the process of transferring data between storage types, formats, or computer systems. It is a key consideration for any system implementation, upgrade, or consolidation [19].

Data Democratization: The notion of making data available directly to users, as opposed to privatized data models in which third parties deliver data, often in the form of economic transactions. Data democratization is part of the larger, open data movement that seeks to make data a public, rather than a private, good [6].

Data Integrity: An overall assessment of the accuracy, completeness, timeliness, and validity of the data [6].

Data Management: Data management is the development and execution of architectures, policies, practices and procedures to manage the information lifecycle needs of an enterprise in an effective manner [17].

Data Replication: The process of sharing information to ensure consistency between redundant sources.

Data Virtualization: Data virtualization is an umbrella term used to describe any approach to data management that allows an application to retrieve and manipulate data without requiring technical details about the data, such as how it is formatted or where it is physically located [10].

Data Profiling: The process of collecting statistics and information about data in an existing source [25].

Term Usage Trends in Peer-reviewed Scientific Literature and Relationships Among Terms

We next consider citation trends from 1980 to 2015 for the foundational DS terms, which we graph in Figures 4 and 5. As noted above, two terms stand out: data collection and data extraction are by far the most frequently used terms. Both terms were in usage prior to 2000, though more so for data collection than data extraction.

Since about 2000 there has been steady growth in the use of many other DS terms. Screen scraping and internet data collection are two such terms that have seen increased usage in the scientific literature since the late 1990s. This increase in citations can be attributed to the rise of the internet, ease of access to online data storage, increase in storage capacities and the emergence of cloud storage. The internet of things (IoT) has produced vast quantities of structured and unstructured data that is housed and spread across the internet and such data can be extracted for analytic purposes. The declines observed from 2010 should not be interpreted as a decline in the salience or importance of such terms. Instead, the recent downward trend in the citation of articles containing DS terms is related to the nature of scientific citations, which accumulate over time. Newly published articles have, on average, few citations than older articles. In general, it takes time for a newly published research paper to have an impact, as measured by citations.

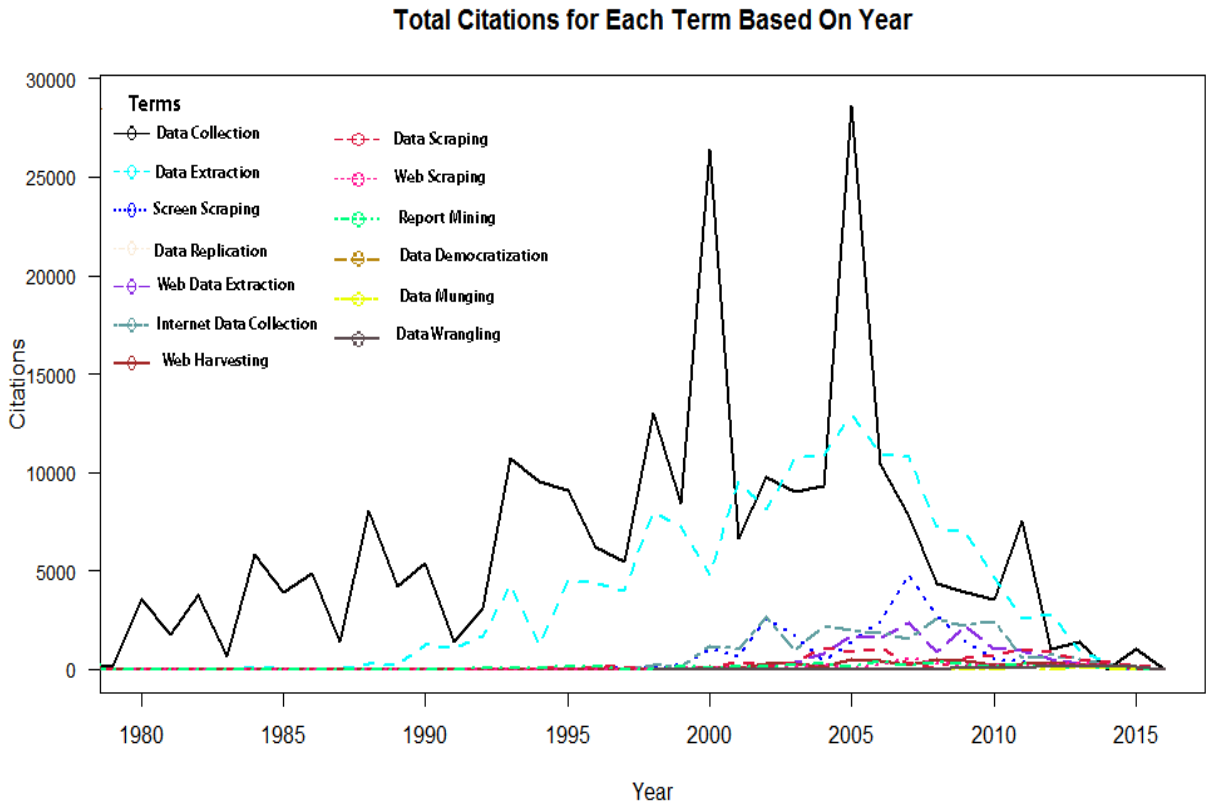

Figure 4: Total Citations for Each Data Collection Terms Based on Year

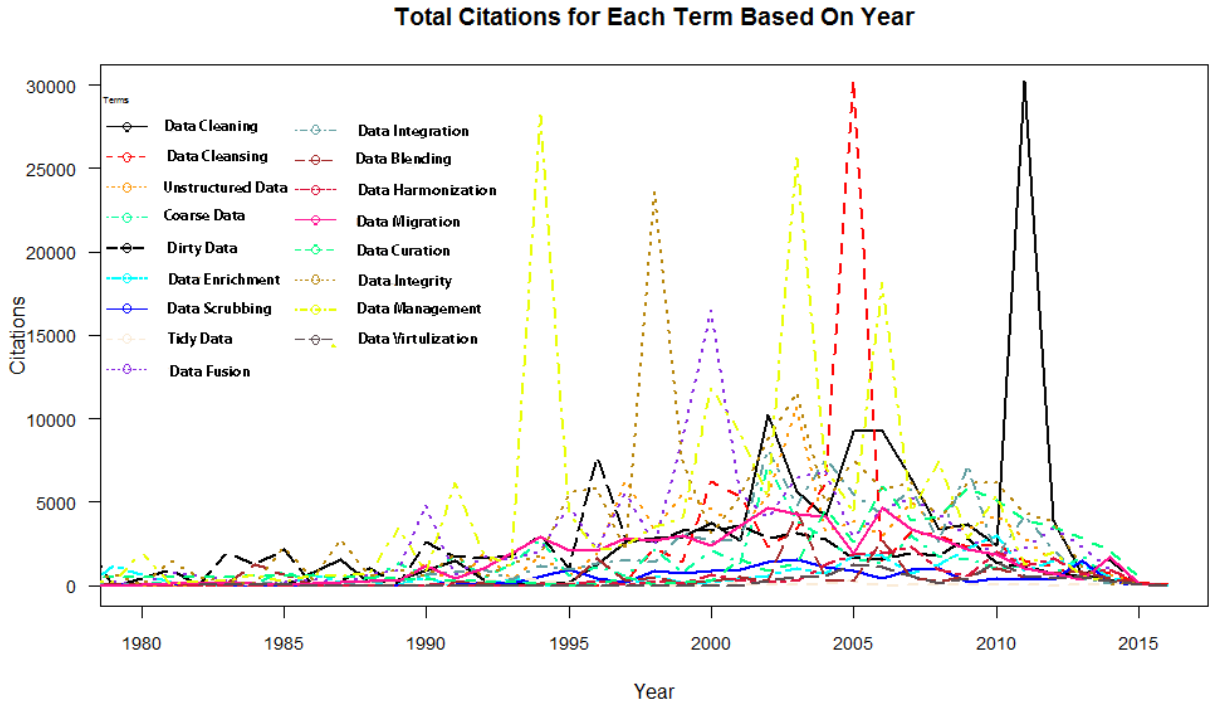

Figure 5: Total Citations for Each Data Curation Terms Based on Year

Figure 6 represents a conceptual overview of the DS terms landscape, with a focus on *Data Collection* and *Data Curation*. We use a hierarchical relations schema that relates terms as parent or child. Terms in green represents parent terms, pink represents first order child terms, and red identifies second order child terms. We first present DS terms in unorganized space to visually illustrate the 'Babel' phenomenon, after which, we establish linkages and associations between terms.

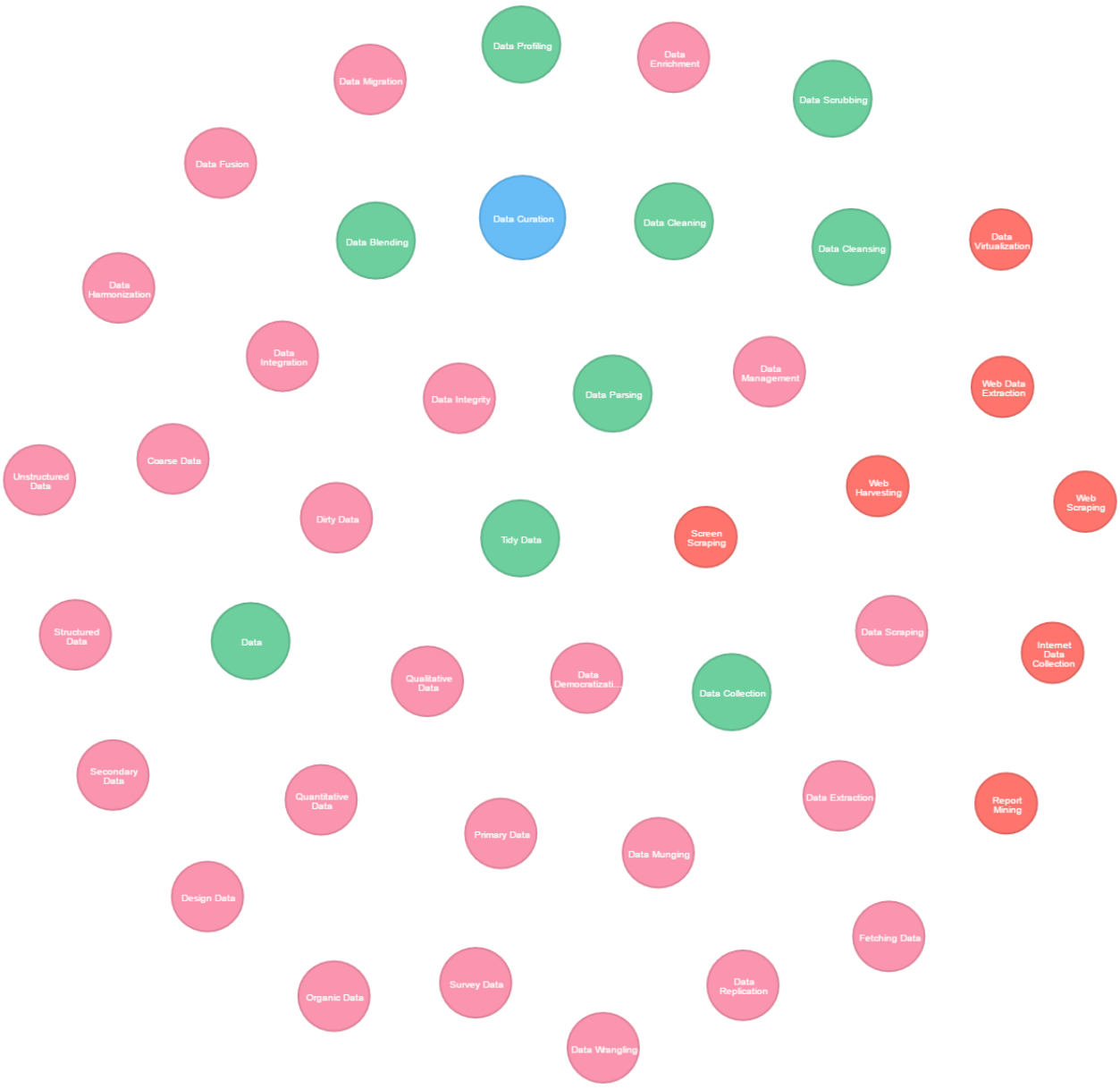

Figure 6: Glossary of Data Science Terms

Figure 7 identifies relationships between DS terms. We divide the terms into three groups, including: Data, Data Collection, and Data Curation. As in the prior graph, parent terms are in green, terms in pink are child terms, and terms in red are second level children. With this color coding scheme, terms in pink are children of terms in green and terms in red are children of terms in pink.

Figure 7 shows that there are a number of synonymous data collection terms, especially with regard to data scraping activities. Data scraping and data extraction involved the same kinds of research activities and are therefore considered synonyms. Web harvesting, web scraping, internet data collection, and web data extraction are synonyms and effectively describe highly similar activities focused on the digital and automated collection of data using computers and internet browsers. Two terms that were not easily categorized as distinctly 'data collection' activities include data wrangling and data munging. Both wrangling and munging involve the collection of data, but both terms are also commonly used to describe additional activities that are farther down the data pipeline, including data integration, blending, fusion and data tidying. For heuristic purposes, we categorized both munging and wrangling as primarily data collection activities.

Within the Data Curation hub of terms, we show that *data blending* has five child terms, including data fusion and integration, which are synonyms, and data migration, integrity, and harmonization. Data cleaning and its synonyms, tidy data, data scrubbing, and data cleansing, are involved with converting course, or dirty data into tidy, or cleaned data.

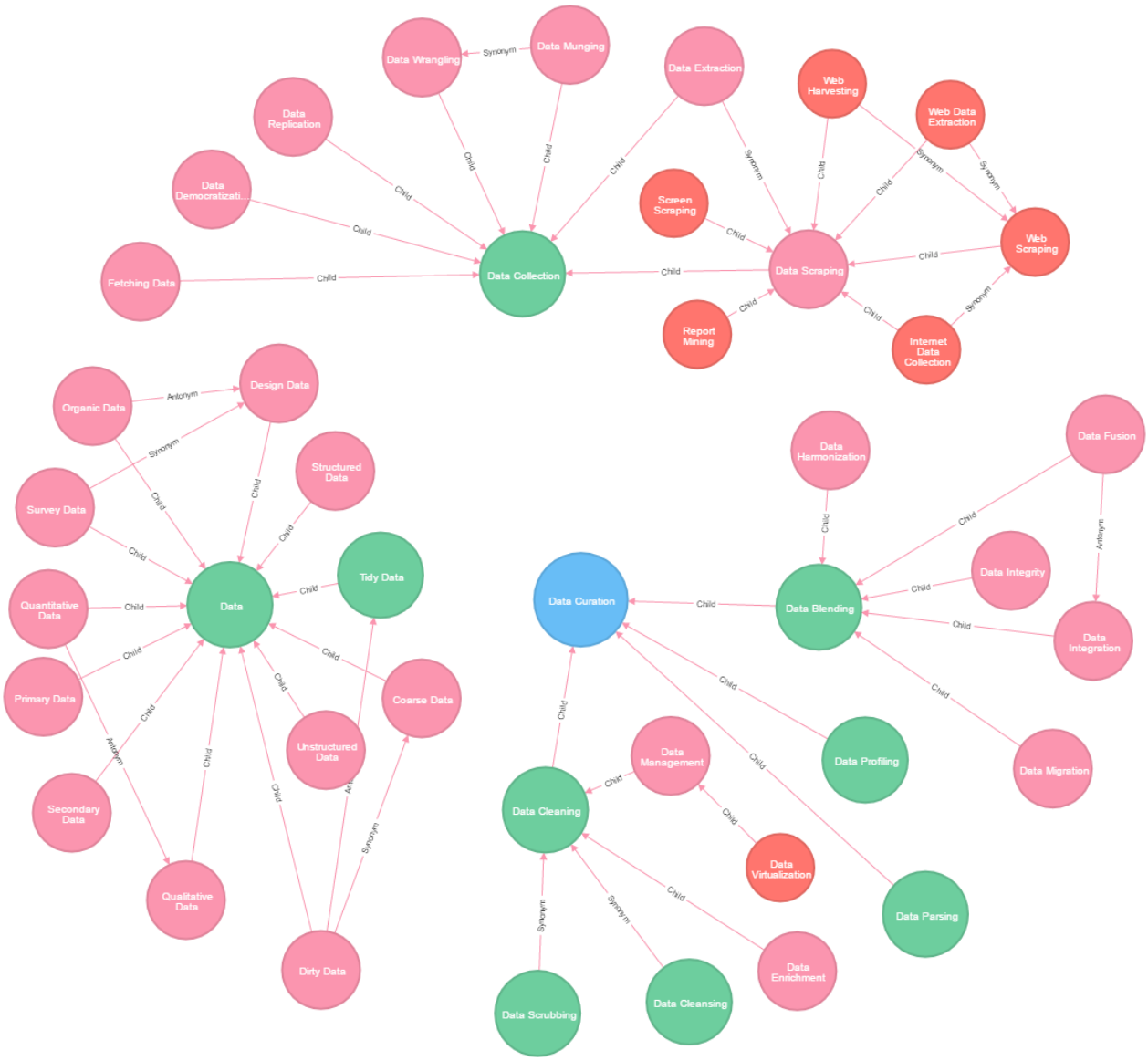

Figure 7: Current data science terminological (conceptual) landscape

In the next section, we discuss methods and techniques involving the collection of data using the internet, which we refer to generally as web data scraping techniques.

WEB DATA SCRAPING TECHNIQUES

Web data scraping or, more simply, web scraping, is the extraction of data from websites or web sources. Web scraping can be done by the software that directly access the World Wide Web using the HTTP (Hypertext Transfer Protocol) or through a web browser. Broadly speaking, web data can be classified as enterprise level data or social web level data. In this section, we give primary attention to social web data. At the level of the Social Web, data is most commonly extracted from social media and social networking sites. Web data extraction techniques allow researchers to gather a large amount of the structured data that is inceasingly generated and disseminated by web, social media and online social networks. This continuously generated human data offer unprecedented opportunities to analyze human attitudes, beliefs, values and behavior (ABVB) at a very large scale. Not only has organic data enabled the study of ABVBs at larger scales than previously possible, but the greater level of anonymity afforded by the internet means that measures of ABVBs extracted from the ‘exhaust’ of data generated by human interaction on the internet are less likely to suffer from social desirability bias[37]. The trick, then, for social scientists who have traditionally collected measures of ABVB directly from known respondents using surveys and interviews, is to identify analogous tools for collecting similar measurements contained in Web.

Web scraping is a method for collecting human data on the web and can be seen as similar, in its general approach and goals, to survey data collection. Web

scraping techniques can be broadly divided into four areas, each of which is discussed below: Copy & Paste, Structured Data Extraction, API's or RSS, and Screen Scraping.

Hierarchical Tree of Data Scraping Techniques

Having briefly discussed data scraping, it is important to stress that choosing the correct data scraping technique for scientific data collection is an extremely important step. The correct choice saves the researcher time and resources, ensures the science is reproducible, and that processing, coverage, and measurement errors are minimized (Groves 2010 Biemer 2010). It is also useful to remind readers that while we discuss web scraping methods as distinct data collection methods, it is often the case that a single data collection will require researchers to use multiple techniques to extract the required data.

We begin with a conceptual overview of data scraping techniques. Figure 8 visualizes data scraping activities in a tree structure and includes a summary description of each method. We follow with a more detailed description of each method. Figure 8 shows that there are two main *Data Scraping* activity areas: Web scraping and report mining. Web scraping can be further reduced to four distinct methods, including copy & paste, structured data extraction, APIs/RSS feeds, and screen scraping. Screen scraping can be further reduced to four methods, including browser tools, bot programs, computer programs, and regular expressions. We discussion each term, in turn, below.

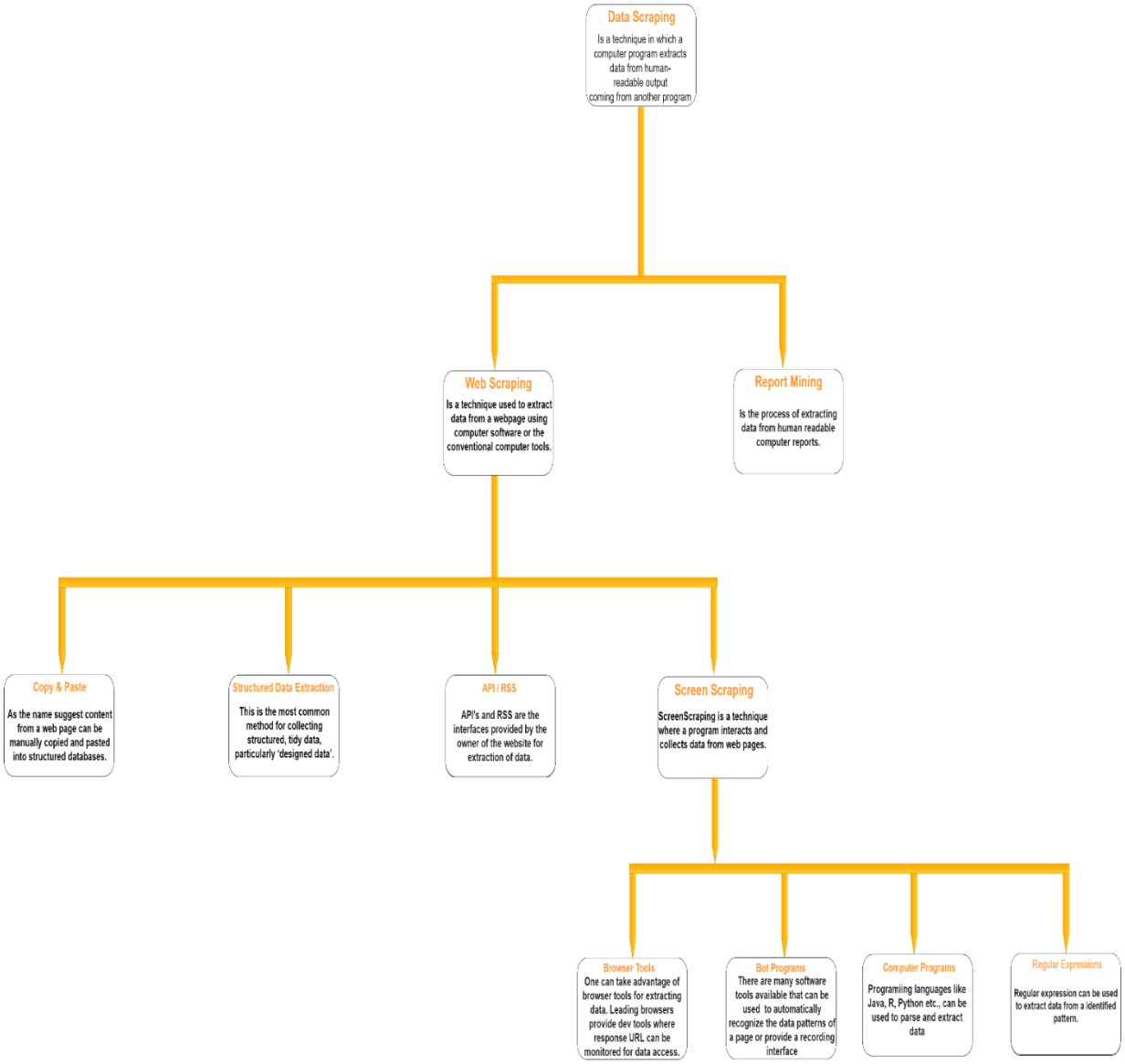

Figure 8: Hierarchical Tree for Data Collection Methods

Report Mining: Report mining is the process of extracting data from human readable digital reports. Digital reports can be PDFs, word processor formatted documents, images, or any other human readable documents. Each report reading will require distinctive data collection methods. A simple regular expression can do the job when it comes to pdf or word processing documents but data extraction from an image requires more advanced technical expertise and data are not always consistent. When data are extracted from images, the images are analyzed by software that specializes in optical character recognition (OCR). OCR datafies imaged reports so that the data

can be parsed, stored in databases, and analyzed by conventional data analysis software.

Web Scraping: Web scraping is a technique used to extract data from a web page using computer software or conventional computer tools. With the growing availability of public data online, researchers increasingly turn to web scraping techniques to extract data from web pages for data analysis. Such methods are most commonly used for ad hoc projects involving custom program/script development.

Copy and paste: As the name suggests, content from a web page is sometimes manually copied and pasted into structured databases. An advantage of this method is that it requires essentially no technical knowledge or expertise. A shortcoming of manually copying and pasting data from web pages is that the method is not entirely reproducible. When other methods are unavailable, copying and pasting, as a method of last resort, enables data extraction where it might otherwise not be possible.

Structured Data Extraction: This is the most common method for collecting tidy data, particularly 'designed data'. In other words, this is the primary method by which social scientists acquire data collected (produced) by members of the scientific research community (e.g. General Social Survey, World Values Survey, Panel Study of Income Dynamics, etc.). This method typically comes in two variants. In the first, users browse through a website authentication and download the entire dataset/database. In addition to gathering a structured and tidy dataset, designed data typically also included project data, including codebooks, survey instruments, and a project manifest. In the second, researchers use filtering criteria to select a subset of the data for extraction. These methods typically grant researchers access to the highest quality, analysis-ready scientific data due to the fact that data made public in this way are typically already structured, tidied, weighted, and in many cases, validated for various

scientific uses. Two limitations to this method are that a) it is more difficult to replicate data extraction (point and click navigation through a custom dataset build) and b) it is more likely that researchers who use this method will conduct analysis on out-of-date data. This is so because many large-scale datasets are versioned throughout the data life-course as errors are found, metadata and contextual data added, and as new data collections expand the size and scope of the original dataset. Any changes, enhancements, or alterations made to the data since the last user download will not be contained in previously downloaded data sets. With this method, a researcher's data is only as current as the last time it was downloaded.

RSS / API: Rich Site Summary (RSS) along with Application Programming Interface (API) are techniques for structured data extraction. RSS is primarily a format for delivering frequently changing web content. RSS feeds allow a user to subscribe to their favorite news sources, blogs, websites, and other digital properties, and then receive the latest content from all those different areas or sites in one place, without having to repeatedly visit each individual site [8]. The data can use Atom / RDF / RSS standard formats which are typically XML structured data.

- RSS Feeds facilitates access to data by using these standard formats. However, the user can only get the fixed content provided by the content administrator. RSS Feeds generally only provide the current data and it is difficult to find RSS Feeds which provide historic data. Nonetheless, RSS Feeds are very useful when researchers can capture the data in a structured database at regular intervals. In this way, researchers can build contemporary data pipelines for a wide range of public data. Many scientific journals like Oxford, Sage, APA and Springer etc. provide RSS Feeds.

- API: This method is in many ways similar to structured data extraction, except that rather than relying on point and click user effort, custom programs are written to automate the functions. The API method also allows researchers to access a much large pool of structured data sources, especially organic data. API's are an increasingly popular method for extracting designed data (e.g. World Development Indicators Database) and many of the standard data management and statistical analysis programs used by social scientists (Stata, R, SAS) have user-written programs that leverage API calls to directly fetch data from repositories in real-time. This method ensures that the data being analyzed are the most recently updated version. APIs also provisionally provide historic data. API's are the unrestrained version of RSS Feeds. API's allow people to search for a query string to get any combination of data including articles/items that have long rolled off the end of the typical RSS feed.

Screen Scraping: *Screen scraping* is a technique for gathering screen display data from one application and translating it so that another application can make use of it. Generally, screen scraping is the term associated with the program that translates between legacy application programs and new user interfaces so that the logic and data associated with the legacy programs can continue to be used. This is especially useful when there is no RSS feed or programmatic access to the API for gathering the data. Quite often, screen scraping refers to a web client that parses the HTML pages of targeted website to extract formatted data. Since the need for screen scraping programs is growing, there are abundant programs readily available today. Some of these programs are designed to collect data from websites by mimicing human behaviors (e.g. user clicks), these are typically called as bots. Mozenda, import.io, and Scraper Wiki are among the more notable off-the-shelf screen scraping computer

programs.In other domains, screen scraping is sometimes referred to as advanced terminal emulation [16]. A screen scraping program can reform data IBM Mainframe 3270 to data which is accessible in Windows-based systems.

A simple XPATH query written in Google Docs also can extract data from the web. The fact that Screen Scrapers are easy to write and low on maintenance cost makes it stand out. With knowledge of HTML, XPATH, Python or Java etc. user can design their own screen scraper. Unlike structured data extraction the data collected using screen scrapers is not out-of-date.

A limitation of screen scraping is that web pages change from time to time and, in turn, the data model and underlying architecture also changes, which enforces the changes to the current program. When such changes occur, screen scraping programs are at risk of losing some or all of their functionality. Because of the possibility of accessing dataset that is not always synthetic in nature, it can cause legal issues. One instance of such case involved American Airlines (AA) and a firm called FareChase. One other major issue with screen scraping is it often requires substantial tidying of data and data wrangling. Such activities require additional resources, technical knowledge, and greater risk of processing errors.

Tools: Tools come in two flavors, or categories. One is *browser tools* and the second is *internet bots*. Browser tools and internet bots collect data by running scripts/tasks and behaving similarly to a human user. Browser tools can be very valuable for extracting data. Leading web browsers provide dev tools where request URL can be monitored for data access. This requested URL can be further coupled with any of programming language and used for accessing data. Data extracted from request URL are most commonly delivered in JSON or XML format. Although this method works for many websites, data collected with these methods are typically not tidy meaning that

researchers can typically anticipate substantial preprocessing data prior to conducting meaningful analysis. Second, the bots will behave like a human and run scripts. The data collected from this method might not be always correct. It is subjected to errors and hard to reproduce. There are many such tools already available.

Other: This section includes any computer programs or regular expressions. Computer programs can be HTML Parsers, DOM parsers, socket programs etc. Semantic annotation recognizing or regular expression can also be used to extract data from a web page. Machine learning techniques can also be used to extract the data. Sometimes writing the standalone computer program is not sufficient, it might require the request URL to fetch data, which can be retrieved using browser tools. Though data from most web sites can be accessed using this method, the extracted data can be inconsistent.

Classification of Web Scraping Techniques

As can be seen from the preceding discussion, an immense amount of work has already been done to facilitate web data collection, the area is nonetheless still in its infancy and we anticipate that data scraping techniques will continue to evolve and standardize as a greater share of the scientific community becomes involved. A wide variety of these techniques have been established to obtain quality data from various sources. Though data scraping can be done in many ways we observed that there is a pattern common in many of these. To establish a classification for these tools we did a thorough review of the scientific literature and of relevant online forums[28][29]. Complete understanding of working of various online tools also provides excellent perspective on the patterns of working of these tools. Existing methods or techniques like Iguana, ReportMiner, dexi.io and Content Grabber have been analyzed for better

perspective. With the help of this analysis, we have established four classifications for web data extraction. These four classifications are comprehensive but there are certain techniques that are a hybrid of two or more classifications.

As explained in the previous sections, multiple types of data extraction techniques for various purposes are in vague. Using the appropriate web scraping technique for serving purpose is vital to derive optimal results. In the previous section we provided a conceptual overview of the several data collection methods currently deployed. In the next section, formally evaluate each of these methods along several dimensions, which we report in tabular form below. The evaluated dimensions include a) availability of data collection techniques, b) quality of collected data, c) ease of use, d) reproducibility, and e) technical skills required to perform data collection. We rate each method on each dimension and then provide summary scores of each data scraping technique.

Even though average score of a data collection technique will aid in choosing a particular data scraping technique for a purpose, there are multiple factors which might influence in choosing a data scraping technique. Consolidated results on these factors add value to this research. For instance, social networking sites like Twitter® offer API to encourage and facilitate the collection of analysis of its data. Extracting data via API requires moderate technical capabilities that might be beyond the skill set of the average social scientist. Data scraping techniques like *structure data extraction* is well-suited to people without technical skills although, but because this mode of data collection is not available on all the websites, additional skills are required. Similarly, HTML parsing require high technical capabilities but the results and reproducibility can be inconsistent.

An evaluation of the existing web scraping techniques has been proposed based on the metrics like required technical capabilities, availability, etc. to aid the users in choosing one which is appropriate for their necessity. Further research in this area could further refine the metrics we have outlined here and will likely include different classifications of data scraping techniques as the discipline evolves and matures. In this way, I see my work here as being as a much a ‘work in progress’ as are the tools currently being deployed to collect organic data via the internet.

As defined in previous sections, few web sources are yet to be explored, which adds lot of value to understand human behaviors and find patterns. Developing scraping techniques to such web source can help research community. Developing data collection techniques based on language specific packages might cause inhibitions to researchers who are not familiar to a language. Instead, developing Interfaces or generic web tools, will make development of these techniques independent of languages or environment constraints, will largely benefit the community of user who might not be familiar to the language been used.

Web Science Data Collection Methods							
Methods	**Average Score**	**Availability**	**Data Quality**	**Ease of use**	**Reproducibility**	**Technical Skills**	**Type of Data**
Structured Data Extraction	3.9	2	5	6	4	1	Structured
Cut & Paste	3.8	4	4	6	4	1	Structured/Semi Structured/Un Structured
API/Tools	3.6	2	5	2	6	3	Structured/Semi Structured
Web Scraping	3.2	5	3	1	3	4	Structured/Semi Structured/Un Structured
HTML Parsing	3	5	3	1	1	5	Structured/Semi Structured/Un Structured
Screen Scraping	3.6	5	1	3	3	6	Un Structured
Report mining	3.2	2	2	5	3	4	Un Structured
NOTES: 1=Low quality/Difficult; 6= High Quality / Easy;							

Conclusions

In this research, we have given attention to the emerging field of Data Science, with specific focus to the terms and methods, as presently constituted, involved in the collection and curation and data for scientific purposes. As with any new field, common methods are called by different names, and newly developed method are given names that sometimes overlap with existing terms and established meanings. We have attempted to inventory, organize and categorize terms in such as way that those who are new to the field of DS will have a working vocabulary with which to interact with fellow data scientists working in disparate disciplines. It is our hope that our work here facilitates more intensive and efficient collaborations between

computational social scientists, computer scientists, and software engineers, for example.

Our results show that much of the ‘language’ of DS is quite new, with many terms only entering the scientific literature after 2000. We also found that general terms such as data collection and data cleaning, continue to dominate the literature, though alternate terms such as data fusion, munging, and wrangling, are making inroads into popular scientific usage. As such processes are a natural part of the evolution and integration of various branches of the sciences, we suspect there will continue to be emergence of new methods, new meanings, and new terms, and that such emergence will be concurrent with convergence in the core actitivies of DS among disparate disciplinary practitioners of DS.

We also provide a simple rating schema for web scraping methods. We propose that such a schema can be of use to DS practitioners, especially those who are new to DS and those coming from disciplines that have traditionally been more distal to computational methods. We hope that future work will improve upon our findings and schema.

REFERENCES

[1] Wikipedia. (2016). Main Page. [online] Available at: http://en.wikipedia.org [Accessed 14 Nov. 2016].

[2] Fincom.at. (2016). FinCom Consulting | Quality in Service. [online] Available at: http://fincom.at [Accessed 14 Nov. 2016].

[3] Singh, J. (2013). Extract Class Refactoring by analyzing class variables. Proquest.

[4] Shmoop.com. (2016). Shmoop: Homework Help, Teacher Resources, Test Prep. [online] Available at: http://www.shmoop.com [Accessed 14 Nov. 2016].

[5] Arxiv.org. (2016). arXiv.org e-Print archive. [online] Available at: http://arxiv.org [Accessed 14 Nov. 2016].

[6] Bisio, R. and Seshadri, S. (2016). Data Informed | Big Data and Analytics in the Enterprise. [online] Data Informed. Available at: http://data-informed.com [Accessed 14 Nov. 2016].

[7] Informatica.com. (2016). Informatica: Data integration leader for Big Data & Cloud Analytics | Informatica US. [online] Available at: http://www.informatica.com [Accessed 14 Nov. 2016].

[8] GovDelivery. (2016). GovDelivery - Home. [online] Available at: http://www.govdelivery.com [Accessed 14 Nov. 2016].

[9] Cs.iastate.edu. (2016). Department of Computer Science. [online] Available at: http://www.cs.iastate.edu [Accessed 14 Nov. 2016].

[10] Dqglossary.com. (2016). GRC Data Intelligence - Data quality glossary. [online] Available at: http://www.dqglossary.com [Accessed 14 Nov. 2016].

[11] Mode Community. (2016). The SQL Tutorial for Data Analysis. [online] Available at: http://sqlschool.modeanalytics.com [Accessed 14 Nov. 2016].

[12] Emilio.ferrara.name. (2016). Emilio Ferrara, Ph.D. | Research Asst. Prof. @ Univ. of Southern California. [online] Available at: http://www.emilio.ferrara.name [Accessed 14 Nov. 2016].

[13] Whangarei.co.nz, W. (2016). | Websites to suit mobile searches. [online] Googleexperts.co.nz. Available at: http://googleexperts.co.nz [Accessed 14 Nov. 2016].

[14] Ramnathv.github.io. (2016). Addicted to R · academic/hacker. [online] Available at: http://ramnathv.github.io [Accessed 14 Nov. 2016].

[15] Cnpmjs.org. (2016). cnpmjs.org: Private npm registry and web for Company. [online] Available at: http://cnpmjs.org [Accessed 14 Nov. 2016].

[16] News, N. (2016). Technology. [online] National Mortgage News. Available at: http://www.mortgage-technology.com [Accessed 14 Nov. 2016].

[17] Datasolutions.searchdatamanagement.com. (2016). Data Management/Data Warehousing information, news and tips - SearchDataManagement. [online] Available at: http://datasolutions.searchdatamanagement.com [Accessed 14 Nov. 2016].

[18] Ijsrp.org. (2016). International Research Journal, Journal for Scientific Research. [online] Available at: http://www.ijsrp.org [Accessed 14 Nov. 2016].

[19] Qaworks.com. (2016). Home Page - QAWorks. [online] Available at: http://www.qaworks.com [Accessed 14 Nov. 2016].

[20] Searchdatacenter.techtarget.com. (2016). Data Center information, news and tips - SearchDataCenter. [online] Available at: http://searchdatacenter.techtarget.com [Accessed 14 Nov. 2016].

[21] Dl.acm.org. (2016). ACM Digital Library. [online] Available at: http://dl.acm.org [Accessed 14 Nov. 2016].

[22] Umb.edu. (2016). University of Massachusetts Boston - a student-centered urban public research university - University of Massachusetts Boston. [online] Available at: http://www.umb.edu [Accessed 14 Nov. 2016].

[23] PD-Rx Pharmaceuticals, I., Endo Pharmaceuticals, I., Inc., M., Endo Pharmaceuticals, I. and Corporation, M. (2016). MedLibrary.org: FDA Approved Prescription Medication Information. [online] MedLibrary.org. Available at: http://medlibrary.org [Accessed 14 Nov. 2016].

[24] Lee, K. (2014). Visual-based web page analysis. Proquest.

[25] Blog.professorcoruja.com. (2016). Professor Coruja - Business and Open Source Technology. [online] Available at: http://blog.professorcoruja.com [Accessed 14 Nov. 2016].

[26] Anon, (2016). [online] Available at: http://www.petisud.com [Accessed 14 Nov. 2016].

[27] Predictive Analytics Today. (2016). Predictive Analytics Today - Predictive Analytics, Data Mining, Big data, Text Analytics, Business Intelligence, Social Media Analytics, Emerging Technology. [online] Available at: http://www.predictiveanalyticstoday.com [Accessed 14 Nov. 2016].

[28] Ferrara, E., De Meo, P., Fiumara, G. and Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. Knowledge-Based Systems, 70, pp.301-323.

[29] Laender, A., Ribeiro-Neto, B., da Silva, A. and Teixeira, J. (2002). A brief survey of web data extraction tools. ACM SIGMOD Record, 31(2), p.84.

[30] Richardson, Matthew, 2012. “Citography: the visualization of nineteen thousand journals through their recent citations” Research Trends:26(January)

[31] E. H. Chi, "A taxonomy of visualization techniques using the data state reference model," IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings, Salt Lake City, UT, 2000, pp. 69-75.

[32] Robert M. Groves, Lars Lyberg; Total Survey Error: Past, Present, and Future. Public Opin Q 2010; 74 (5): 849-879.

Paul P. Biemer; Total Survey Error: Design, Implementation, and Evaluation. Public Opin Q 2010; 74 (5): 817-848.

[33] Hadley Wickham. 2014. “Tidy data”. The Journal of Statistical Software, vol. 59, 2014.

[34] Harzing, A.W. (2007) Publish or Perish, available from http://www.harzing.com/pop.htm

[35] Quantitative Analysis of Culture Using Millions of Digitized Books.2011.

[36] Robert M. Groves, Lars Lyberg; Total Survey Error: Past, Present, and Future. Public Opin Q 2010; 74 (5): 849-879.

[37] Robert J. Fisher; Social Desirability Bias and the Validity of Indirect Questioning. J Consum Res 1993; 20 (2): 303-315.