PREPRINT

Big Data, Analysis of

Damian Trilling
University of Amsterdam
d.c.trilling@uva.nl

Trilling, D. (2018). Big Data, Analysis of. In: Matthes, J. (ed.), *International Encyclopedia of Communication Research Methods*. Hoboken, NJ: Wiley.
doi:10.1002/9781118901731.iecrm0014

Abstract
This entry describes what so-called Big Data are and how they can be analyzed in communication science and the computational social sciences more broadly. It briefly addresses the epistemological questions around this type of analysis and its historical development in communication science and related disciplines. The entry then moves on to a more practical description of the technical requirements, the structure of such data sets, and their storage in different types of databases. It gives an overview of different analytical approaches, including, but not limited to, visualizations, supervised and unsupervised machine learning, natural language processing, and network analysis. Ethical implications and future directions are discussed.

Keywords
computer science, communication research methods, content analysis, digital media, new media, quantitative methods, social networks

Main text

The analysis of Big Data is the analysis of datasets that are by orders of magnitude larger than it is traditionally the case in the social sciences. In communication science this refers mostly, but not necessarily, to digital trace data like those data that users produce on social network sites, writing reviews, or simply by browsing the Internet; or to large collections of documents.

It is important to note, though, that "Big Data" is a catchall term that is used differently in different fields. While a computer scientist might have in mind supercomputers and server farms, scholars in the humanities or social sciences sometimes use the term for datasets that can be processed on a consumer laptop. Top-notch supercomputing is not used within communication science yet and lies beyond the scope of this entry. At the other end of the spectrum, datasets whose size is only one or two orders of magnitude larger than most traditional communication science datasets are not of core interest here. For example, a dataset that consists of 50,000 tweets does not require a fundamentally new approach in terms of tools or methods. Big Data as it is understood in this entry refers to datasets that are too large to be processed using common office or statistics programs. These large datasets allow answering new questions, but

require new methods, as classical significance-based hypothesis tests lose their meaning.

Not only among disciplines, but also within communication science, there is no consensus on how to define Big Data. While some embrace a definition that characterizes Big Data as a "phenomenon that rests on the interplay of […] [t]echnology […] [, a]nalysis [, … and m]ythology" (boyd & Crawford, 2012, p. 663), others use a more technical definition that is widely used in the commercial sector and that describes Big Data as the combination of increasing volume, velocity, and variety of data. This "3-Vs-definition", coined in 2001 by Doug Laney, an employee of a consultancy firm, is still widely used in its original or in a slightly modified form. While the first v, volume, speaks for itself, the second v, velocity, refers to the dynamic nature and the fast pace of changes of such data, and the third v, variety, to the huge variety of data sources and data types that are used and combined in many Big Data analyses. It is mainly the combination of these three features, not necessarily each one separately, that calls for new methodological approaches. Often, a fourth v, veracity, is added, which alludes to the uncertainty of data: in Big Data research, it can be difficult or even impossible to assess the accurateness of all data points.

Although often associated with social media data, Big Data approaches in communication science encompass other types of data as well. Examples include data collected by sensors, GPS, clickstream data, the analysis of large-scale text corpora, or any kind of data retrieved from databases or scraped from websites. Often, such data are noisy and lack the rigid structure of quantitative data gathered with traditional means, which are typically stored in tables and where clearly defined variables have clearly defined values. In contrast, data acquired via a Big Data approach may be incomplete and may contain errors and inconsistencies. At the same time, they can be very rich and contain information that was previously unavailable to researchers.

Historical development

Since the 1970s or even earlier, social scientists have been using computers for analyzing their data. The type of data analyzed was limited and generated by a fixed set of methods, in particular surveys, experiments, and content analyses. With few exceptions, these data were explicitly generated for social-scientific research. In such a paradigm, the researchers have full control over the data creation process: After all, it is their study design that generates the data, and without their study, the data would not exist. With the proliferation of information and communication technologies in all domains of society from the mid-1990s onwards, a new type of data emerged: Digital traces left behind when people engage in any kind of online activity, use loyalty cards or mobile phones, or are captured by CCTV. The digitalization and interconnectedness of these technologies have made it possible to create as a byproduct datasets that are of interest to social scientists, but which were not collected for this purpose in the first place. Nowadays, researchers still influence how the dataset is constructed, but the primary process of data collection is at least partly out of their hands: Instead of asking people questions that the researcher has developed, people produce data out of themselves by sharing or re-distributing content, by writing comments, posts, or tweets. The unobtrusiveness of the data collection is an obvious advantage, as well as the fact that data on social phenomena that were

simply unobservable before have suddenly become observable. As online and offline life have become more and more intertwined in the new millennium, such digital traces have become omnipresent, with the rise of social network sites as a very prominent example.

The general trend towards digitalization of data in combination with widely available broadband internet has made it also easier to get access to large collections of datasets of all kinds that were much harder to access, if accessible at all: news databases reaching back for decades, but also documents and datasets released by authorities or organizations – a development also referred to as a move towards *open data*.

This availability of data has sparked the interest of different disciplines in Big Data analysis. Naturally, computer science has played a leading role. But above all, new fields have emerged in the new millennium, most notably the field of data science, that operates on the boundaries of statistics and computer science (Cleveland, 2001). Similarly, the field of computational social science aims at bridging the gap between the traditional social sciences on the one hand and computer science and data science on the other hand. Employing a data-driven approach, computational social sciences try to answer questions about individual and group behavior (Lazer et al., 2009). In an introduction to a special issue of the *Annals of the American Academy of Political and Social Science* on Big Data in the social sciences, Shah, Cappella, and Neumann (2015) suggest to speak analogically of computational communication science when referring to such approaches within the discipline of communication science. Similarly, the field of Digital Humanities incorporates computing tools and answers humanistic research questions employing the analysis of digital data. Nevertheless, computer science, digital humanities, and computational social science come from different epistemological traditions (Kitchin, 2014), and thus often address different questions and make different methodological choices.

Consequently, the advent of Big Data analysis in the social sciences and humanities has sparked a debate on the epistemological implications of Big Data research. Central to this debate is the role of theory. Some take a firm position on this and argue that in an era of an abundance of available data, theory has become superfluous. Mayer-Schönberger and Cukier (2013, p. 7) state that "society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what". This notion has been challenged by many who underline the importance of theory in the social sciences and argue that data without theoretical interpretation are just data, but do not offer real insights (e.g., boyd & Crawford, 2012). Still, the inductive approach implied by Big Data research can lead to new and potentially surprising insights, and, ultimately, also to theory development.

The case of computer translation is one practical example of how, given sufficiently large data sets and enough computing power, an inductive statistical approach can drastically outperform theory-driven deductive approaches. For decades, it had been promised that a translation software for daily use would be available in near future. The real breakthrough, however, did not come until Google Translate was introduced, a system that radically broke with the tradition of machine translation based on linguistic rules and introduced an algorithm that was based on statistical principles. In this case, it is indeed not important for the system to understand why a certain word with multiple potential meanings is

translated one way or the other, as long as the 'what' is correctly estimated. Similarly, recent in advances in image recognition have led to algorithms that can distinguish types of pictures, like pictures of cows, sailboats, cats, birthday parties, and so on. These algorithms do not work based on a definition of cows or sailboats, but by inferences drawn from statistical similarities within collections of such pictures.

Technical requirements

Due to technical advancements, many techniques that only few years ago required expensive specialist hardware, can nowadays be run on consumer computers. On the other hand, also the amount of data available has grown and, with it, the demands in terms of storage and computing power. While the price of both storage and computing power is declining steadily and hardware requirements can often be met within reasonable financial limits, this does not mean that consumer software can be used. For example, an Excel sheet cannot contain more than 1,048,576 rows by 16,384 columns, which means that even a simple operation like calculating a mean and a standard deviation cannot be accomplished once a dataset contains more than a million cases – at least not without some tricks like splitting the dataset with another program first. Similarly, if one wants to count the occurrence of specific words in a set of news articles, one cannot do so once one wishes to take into account more than around 16,000 different words. Common statistical software like SPSS or STATA may not always enforce such a strict limit, but keeps a copy of the data in the computer's memory (RAM), effectively limiting the size of the dataset in a similar way. This is also the case for the widespread statistical programming language R, although there are R packages that to a certain extent circumvent this restriction. For a Big Data project, both hardware and software thus may be limiting factors. Often, writing tailor-made software is necessary, and computing facilities have to fit the scale of the analysis. The necessary computing power could be provided by some high-end servers that a university department might purchase to this end. Only a few thousand Euros are needed to purchase a server, that, for instance, is capable of collecting 190.000 tweets per hour over the course of a whole year (Murthy & Bowman, 2014). However, next to the purchasing costs, having a set of own servers also involves administrative costs for maintenance.

Another considerable drawback is the need to purchase an infrastructure that is capable of handing the most demanding analysis that is to be run, even if most of the time, a less expensive infrastructure would be sufficient. In many cases, though, an investment in own servers has become obsolete by the advent of cloud computing, which offers much more flexible solutions. Researchers can rent resources on a larger infrastructure, which makes it possible to have access to exactly those resources needed to accomplish a certain task without having to maintain an own server farm. Commercial parties, most notably Amazon Web Services, provide such an infrastructure, but in some countries, researchers can apply for making use of similar publicly funded services for free. Many BigData tools are designed to integrate with these services, and can spin up new instances of virtual machines as needed. Next to monetary advantages, also legal considerations might play a role, because storing data with a private company in another jurisdiction can be problematic in some cases.

The possibilities range from setting up a single virtual machine with a performance comparable to a desktop computer to running hundreds of these instances at the same time, involving for example the so-called MapReduce technique. This can be done using Apache Hadoop, a framework for distributed storage and processing of data that exceeds a size of several terabytes. As outlined in the next section, data are often retrieved continuously – a fact that is also alluded to by the *velocity* aspect in the aforementioned definition of Big Data as being high volume, high velocity, and high variety data. This means that the infrastructure used must not only have the room to store large datasets (a requirement that is rather easy to satisfy in times of ever-decreasing storage prices), but also the capability of retrieving and processing large amounts of data continuously and without interruptions – a requirement that is much harder to meet.

In contrast to many quantitative methods that are common in the social sciences, where researchers are often able to use the same one statistics program for all types of research they conduct, Big Data analyses are usually not conducted using an existing one-size-fits-all software package. This is a direct consequence of the great variation of requirements, both in terms of hardware and software. While tools and software exist for specific Big Data-related tasks, by and large, a certain level of programming skills is inevitable for doing Big Data analysis. It is conceivable, though, to hire a programmer who develops both a back end for data handling (where the actual calculations are performed) and a front end, which provides an easily accessible interface. In this way, even a researcher lacking the necessary programming skills can make use of the calculations provided by the back end.

Data acquisition, management, and storage

Application Programming Interfaces (APIs) and other data sources

The analysis of Big Data is necessarily intertwined with the way in which it can be stored and retrieved. A convenient way to retrieve data is the use of a so-called API (Application Programming Interface). If a service offers an API, researchers can easily write a program that retrieves data from the other party. As social network sites, but also Wikipedia or a lot of Google services, provide an API, Big Data researchers can directly access the data provided by these services, without any need to either manually download data or to write a program that scrapes and parses the data. When using an API, the data are often created at the moment of delivery: For example, using the Twitter Streaming API, one can retrieve tweets sent at that very moment, a principle that one could compare to a video recorder. This illustrates that the concept of a "file" that is to be analyzed is inappropriate in many Big Data analyses. While an API provides easy access to a large amount of data, the dependence on the delivering party is one downside. In particular, most parties impose strict limits on the amount of data a researcher is allowed to retrieve, and some offer full access only for money.

Other ways of retrieving data for Big Data analysis include scraping and crawling. Scraping means extracting information from a web page and storing it in a suitable structured format. Crawling means retrieving a web page, searching it for links, following these links, retrieving these pages, and so on. It is especially the combination of both that makes it possible to construct extensive datasets. For example, one can write a program that identifies all ratings, reviews, and

prices on a page on a product comparison website, identifies links to similar products, downloads the ratings, reviews, and prices on that page as well, and so on.

Another data source is the re-usage of existing datasets that were collected for other purposes, but that can offer additional insights when linked to other data sources. Developments like the open data movement and the trend towards open government data are important data sources in this context.

One important characteristic of Big Data analysis is that it derives its value largely from linking several datasets. For example, the revision history of Wikipedia entries of companies (which are openly accessible) might be of limited interest, but once linked to a dataset that allows linking both the contributors and the companies to geographical locations, and possibly integrating other data like stock ratings, a number of possibilities for new types of analyses emerge. An example of such a linkage would be using Wikipedia data to identify topics in Twitter data (Yıldırım, Üsküdarlı, & Özgür). Here, the advantage of APIs becomes evident: It is not necessary to have a copy of each of multiple datasets at the same place. Instead, it is possible to retrieve exactly those pieces of data that are needed by linking to the appropriate API.

SQL and NoSQL databases

Whatever the source, ultimately, the data have to be stored in one way or another to be further analyzed. One might conceive of possible analyses that do not require storing a full dataset but rather acquire the data on-the-fly on an as-needed-basis. However, this approach will not be considered further, as it is unlikely to play a significant role within social-scientific research and as it is in conflict with the aim of getting reproducible and checkable results. Therefore, it can be assumed that the first step in any form of analysis is to find a way of storing the data. In traditional quantitative methods used in the social sciences, data are saved in tables, with each row representing a case and each column representing a variable. These tables, which are often stored in proprietary file formats (e.g., SPSS, STATA, SAS), are then loaded into the computer's memory and analyzed. As datasets become larger, or as they exhibit a more complex structure, this approach is not feasible any more. The conceptualization of "the dataset" as one file that represents all necessary data in the form of one table, thus, is often not appropriate any more. More than that, the very notion of the dataset having to be stored on the computer of the researcher has become obsolete. Rather, data collection and storage are increasingly done on one or several servers. Some analyses might also be performed there; for others, a filtered or aggregated subset of data might be downloaded.

A technically simple approach would be the storage of the data in a large number of separate files, like plain text files, CSV tables, JSON files, or XML files. If all of them are structured in the same or at least a similar way, the researcher can easily write a script that processes these data to extract the needed information or to conduct an analysis on the data. In fact, when raw text files are stored on the Apache Hadoop Distributed File System (HDFS), they are readily available for analyzing using the MapReduce framework, for instance. An advantage of storing the data in such files is the transparency and simplicity of the method, but especially when analyses involve looking for a specific piece of information in this large collection of files, the approach can be extremely

inefficient. Even if in a first stage, data are collected this a way, and even if they are stored the same way for archiving purposes, the ultimate goal often is to feed them into a database for more efficient processing and analysis.

One approach is to rely on a relational database management system (RDBMS) like MySQL. In such a system, several tables are linked by keys. For example, one could have a table of hundreds of thousands of posts to some online platform, with one column indicating the sender's user name. Rather than replicating the specific information for each user (email address, homepage, member since…) in each row, this information is stored in another table in which it can be looked up.

None of the tables in such a database have to be fully loaded into a computer's memory at any time. This makes it possible to store, process, and analyze large amounts of data, while being able to access it in multiple ways at the same time. A user can add new data to the database while others are conducting some analyses or exporting data. This is why researchers who analyze Twitter data often use such a database. For example, the Twitter content analysis toolkit DMI-TCAT (Borra & Rieder, 2014) collects tweets continuously by querying the Twitter API and stores them in a MySQL database. This does not interfere with the toolkit's analysis functions, which allow for calculating a number of metrics and plotting graphs based on data from the very same database. Such a database therefore should not only be seen as a means of storing data. It also provides the infrastructure for first analyses. For example, using the SQL query language, one can calculate descriptive statistics, cross tabulations, aggregate data, and the like in a very efficient way.

Data collected via the APIs of social networks often have a rather rigid structure that can be organized reasonably well in tables, and hence in a relational database. However, as outlined above, Big Data analysis often involves data from a variety of sources that might not be as tidy. Tables are a less suitable form of storage for messier types of data, like large chunks of text (or images or videos) scraped from the Internet, or any form of data with a structure that is not uniform across cases. Here, NoSQL databases come into play. Databases like MongoDB or CouchDB do not use tables to store data, but store data in documents or key-value pairs. An entry could consist of a key named "text", with a whole blog entry as associated value, another key named "title" with its title as value, and so on. In such NoSQL databases, data are considered to be semi-structured. This implies that not all entries have to consist of the same keys, and the data structure of the values is neither centrally defined nor enforced. This makes NoSQL databases very flexible and less sensitive to messy or incomplete input data.

The organization of entries in key-value pairs rather than conceptualizing each entry as a row in a table is the main difference with relational databases, where a strict format is defined and enforced, and where it is difficult to change the structure of the database once it has been created. It is also easy to store nested data in a NoSQL database: The value of a key-value pair can consist of a key-value pair (or even a list of several key-value pairs) itself. This data representation closely resembles (or even is identical to) the JSON format, the format in which most APIs deliver their data, which makes exchange in both directions easy. Once a database grows really big, NoSQL databases scale well and can be run distributed on multiple servers.

Like relational databases, also NoSQL databases allow for analyses by efficiently calculating basic statistics or by aggregating data. Next to that, most systems allow for efficient processing of textual data, which seems particular interesting for communication science. This includes sophisticated indexing and searching facilities, which can be used to create subsets of the data for further analysis. In particular, it is possible to use so-called regular expressions to look for text strings that follow a specified pattern, or to take stopwords and stemming (see below) already into account when searching the database and before the actual analysis is conducted. Next to this, for some databases like ElasticSearch, interfaces have been developed that allow exploring the data visually in a web browser. Tools like xtas (De Rooij, Vishneuski, & De Rijke, 2012) provide extensive tools for text mining and natural language processing and rely on ElasticSearch as backend, which makes it possible to make use of all other interfaces to ElasticSearch. ElasticSearch can handle very large amounts of data, as it can be installed on a cluster of multiple computers.

Due to the amount of data concerned, a distinction has to be made between the full dataset as stored in the database and a subset used for the analysis. Especially operations like searching, filtering, and creating aggregated datasets can be done very efficiently on the database level. This means that while statistical analyses in the social sciences often take place within one dataset and using one software package, a Big Data approach to data analysis can involve several layers: for example, a set of commands in the database's query language might be used to compute averages and frequency tables, which can be done much more efficiently on the database layer. Also, many databases have advanced indexing abilities, which makes it possible to search for specific search strings in huge amounts of text within a minimum of time. For the more advanced statistical analysis, then, filtering mechanisms are used to export the relevant data and to pass them to the program responsible for this analysis. Programming languages like Python make it possible to connect to all major databases directly, and also to conduct a wide range of statistical tests, so that researchers can make optimal use of both. But also some statistical packages can directly connect to at least some databases.

It has to be noted that by no means it is necessary that database and analysis layer are physically on the same machine. It is very common to run the database on an optimized server, while the programs used for the actual analysis run on a different computer and connect to the database server for getting access to the data. For really large datasets of tens of terabytes or more, distributed file systems are used, in which both data and the computational tasks are spread among a number of physical or virtual machines. The technique used for this is called MapReduce, in which a task (calculating a mean, counting word frequencies, and the like) is split into smaller sub-tasks that are then distributed to the different machines.

Analytical approaches

Due to the amount of data available, the emphasis on statistical significance, which is prevalent in traditional quantitative social sciences, is replaced by other approaches. In many analyses, there is not even a sampling stage (often referred to as “N = all”), which makes it rather pointless to generalize findings from a “sample” to a “population”. But even in research designs where the argument can

be made that there is a population from which a sample was drawn, the high N leads to p values close to zero even for tiny and de-facto meaningless effects. Therefore, Big Data analysis often focuses on inductive approaches that are meant to discover relationships and patterns in a dataset, and in doing so is rather situated in an explorative than a hypothesis-testing framework. This does not mean, however, that statistical techniques used in traditional social-scientific contexts are not used any more. The tasks for which they are employed differ, though. Furthermore, fields such as Information Retrieval and Artificial Intelligence have influenced the set of techniques commonly employed in the analysis of Big Data.

Before employing any of the analytical methods described below or a combination of them, attention has to be paid to the potential messiness of the data. While in small-scale datasets, input errors and the like can be spotted manually, other measures have to be taken to ensure the integrity of datasets in Big Data analysis. It is very likely that in a dataset of millions of cases some entries will contain errors, which means that the researcher has to pay special attention to defining how to deal with implausible values, missing values, or errors like spelling mistakes in string variables.

This also implies that researchers have to take into account that Big Data analysis is a much more iterative process than it is the case with many traditional research methods: While the exact analysis strategy for a social-scientific experiment can be determined during the design phase, and while a content analysis codebook might undergo one or two rounds of improvement during pre-testing, preprocessing steps needed for the analysis of natural language will only deliver useful results after an iterative process of continuous refinement. The same holds true for the model specification stage.

Visualization, dashboards, and descriptive statistics

When data are stored in a database on a server, and thus in a form that is less intuitively accessible than a simple table that is opened in a statistics or office program, they have to be made accessible to the researcher. To make optimal use of them, some programming knowledge will be necessary for accessing the database directly. To simplify routine tasks, however, it has become common to use so-called dashboards, web pages that connect to the server and display descriptive statistics and charts, and allow exporting subsets of the data. For instance, a first step of Big Data analysis is often an explorative look at the data to for example trace how trends evolve over time. Plotting the use of words in a dataset of texts or the number of messages sent via a social network site over time can be done in such a way very efficiently.

More broadly, data visualization for offering quick insights in complex datasets has become increasingly popular. This reaches from creating simple scatterplots, line graphs or pie charts to word clouds, social network graphs, and geographical maps. Approaches using a dashboard that provides instant descriptive statistics and visualizations are also popular in commercial applications, often referred to as “Business Intelligence” or “Analytics”. However, it should be noted that without proper preprocessing of the data, such visualizations can be misleading. Also, social scientists usually do not want to stop with such an often mainly descriptive analysis. While the methods employed vary widely depending on the data to be analyzed and depending on

the question to be answered, some techniques can be distinguished that are particularly relevant to Big Data analysis.

Natural Language Processing

Although Big Data can refer to any kind of data, in communication science, most – but not all – datasets will contain some sort of more-or-less formally written text, so-called natural language. In traditional quantitative content analysis, human coders are used to transform natural language into numbers by coding the variables of interest. In a Big Data context, such tasks involve Natural Language Processing (NLP). Some basic text operations, like word counts, are even used as canonical example to explain and compare the performance of large-scale data processing infrastructures like Hadoop MapReduce or Apache Spark, which distribute this task among a large number of computers. However, many analyses in communication science use packages that do not use distributed computing, like the quasi-standard NLTK (Natural Language ToolKit) package for Python (Bird & Loper, 2009).

Generally, two ways of handling text can be distinguished: one that takes into account the grammatical structure of a sentence and one that does not. In recent years, remarkable accomplishments have been made in NLP. Nowadays, it is not only possible to parse sentences, but also to recognize named entities. When doing so, part-of-speech tags can be assigned to each word (token) which identify not only whether a token is a noun, a verb, an adjective, and so on, but also to which other part of the sentence it refers. With such techniques, one can conduct a semantic network analysis (Van Atteveldt, 2008) that identifies actors and codes the relationships between them. Part-of-speech tagging can also be used at the feature selection phase, to for example only include nouns and adjectives in further analyses. A widely used toolkit for this is Stanford CoreNLP (Manning et al., 2014), which offers part-of speech tagging and named entity recognition. It remains problematic, though, that such packages have only been developed for relatively few languages yet.

Ultimately, however, many Big Data approaches typically use a so-called bag-of-words (BOW) representation of text. This means that each text is only described in terms of the frequency of each word, disregarding word order. Often, weighting schemes are applied. For example, the tf-idf scheme weights each term frequency in a given document by the number of documents in which it appears. By penalizing a term for occurring in multiple documents, the tf-idf-weighted BOW-representation of a document reflects the most characteristic words for this document. When constructing a matrix of such term frequencies per document, the dataset shrinks by orders of magnitude compared to the full text of all documents. In addition, as many words will occur zero times in many of the documents, the document can be stored and processed as a sparse matrix. In such a way, even an immense amount of textual data can be reduced to a manageable representation that can be subjected to statistical analysis.

When creating a BOW representation, and before the actual analysis, a number of preprocessing steps are usually taken. These include the removal of so-called stopwords (words without an interpretable meaning, like “the”, “a”, but – depending on the context – also “says”, “go”, etc.) or stemming, in which words are reduced to their stem. To overcome the problem that some words have a completely different meaning in certain combinations, texts are sometimes not

split into single words, but into so-called n-grams, mostly bi-grams (two consecutive words) and occasionally tri-grams (three). The most popular application of this among a lay audience is probably the Google Ngram Viewer, that allows plotting the frequency of n-grams in books over time.

It is advisable, though, to perform some preprocessing steps before transforming texts to a BOW representation. This includes using so-called regular expressions to replace multi-word expressions or synonyms, but can also involve part-of-speech tagging or named entity recognition. For example, it is common to keep only specific parts of speech (like nouns and adjectives) and discard all other words when one is interested in identifying topics or frames in large corpora of texts.

Despite their simplicity, BOW representations have proven to be highly useful in many areas of Big Data analysis. For instance, by calculating the similarity of the BOW representations of two documents by measures like the cosine distance, one can identify near-duplicates, which is extremely useful to identify texts that are published in multiple outlets and which enables communication scientists to trace how information spreads. In essence, and in line with the Big Data analysis paradigm, such applications show that the mere correlation of features like word frequencies is enough to answer many research questions, and a deeper understanding of the meaning of a text is not necessary. BOW representations are also used as input for supervised and unsupervised machine learning algorithms, word co-occurrence analyses, and many more. As a rule of thumb, one might say that for tasks like identifying topics or calculating document similarity, a BOW representation is the most appropriate one, while its assumptions are problematic in cases were exact meaning of specific words in their context plays a role, like in sentiment analysis, were the meaning of words changes drastically if they are preceded by a negation.

Unsupervised machine learning

One of the key tasks in Big Data analysis is reducing the amount of information. For textual data, a first step can be applying the Natural Language Processing mechanisms described above.

Even after such preprocessing steps are taken, and also for all other types of Big Data-datasets, it remains necessary to reduce the amount of information to be able to conduct meaningful analyses. For potentially millions of cases, there might be hundreds or thousands of variables, and one aims at reducing these numbers by finding hidden structures in the data. These approaches are referred to as “unsupervised machine learning”.

The techniques for unsupervised machine learning approaches are often also applied in other – small scale – contexts, and hence, many communication scientists are familiar with them. Two examples for this are principal component analysis (PCA) and cluster analysis. Often, PCA or the related singular value decomposition (SVD) are used to reduce the dimensionality of the data, i.e. to reduce the number of variables to be taken into account in future analysis. Shifting the focus from similar variables to similar cases, cluster analysis can be used to identify groups of cases. Often referred to as pattern recognition, such clustering in which essentially each case is assigned a categorical label is the goal of some Big Data research. The composition of the categories then can be presented with descriptive statistics; or category membership can be used as

either dependent or independent variable in some statistical model. Whatever approach one chooses, both principal component analysis and cluster analysis can help reducing the sheer amount of information to interpretable entities.

Reducing the amount of data can be necessary not only to ease interpretation, but also for technical reasons. Unsupervised machine learning approaches often serve as an intermediate step to make the data manageable for further analyses. It is important to note that a huge variety of different kinds of data could be subjected to such an analysis, no matter whether it is about representations of text, numbers or something else. To give just one example: PCA is not only popular in the social-scientific context, but also in areas like image compression.

Even if the underlying statistical rationale is the same, the algorithms and implementations that are used in Big Data analysis can differ from what researchers might be familiar with in traditional statistics programs. While the latter represent and analyze data in a two-dimensional matrix (a so-called dense matrix), many Big Data-datasets have an enormous amount of cells that have a value of zero. Therefore, these data can be much better represented using a sparse matrix. This makes computation more efficient and, in addition, addresses the limitation that a gigantic amount of memory would be required to load the whole matrix. This means that one can conduct a principal component analysis on a matrix that – in the standard, dense form – would not even fit in the computer's memory.

Related to topics as cluster analysis and dimensionality reduction is the question of how to deal with enormous amounts of variables when constructing a statistical model. This step is referred to as "feature selection". A researcher who ultimately wants to build a regression model can hardly use thousands of independent variables in this model. One could, for example, use a PCA as a step to determine which variables (or "features") are most important; or one could use the component scores of the PCA themselves as features (which would be referred to as "feature extraction"). Other techniques for reaching this include, in the case of regression models, the lasso method, which can be used to determine which variables to include and which to drop.

Another way of reducing the amount of data is the field of topic modeling, which provides a model of the text that is not too difficult to interpret. Topic modeling is an inductive technique that aims at identifying shared topics in a collection of texts. Latent dirichlet allocation (LDA) is the most prominent technique here, even though it was only introduced in 2003 (Blei, Ng, & Jordan, 2003). It uses a BOW representation of texts and models this representation as being generated by a mixture of topics. Each topic, then, generates words with a specific probability that define the topic. When estimating such a topic model, a researcher can thus identify a number of latent constructs (topics) that can be described by the words that are typical for this topic. The procedure estimates a topic score for each topic for each document in the dataset, which allows for further analysis. Next to that, visualization packages are available that allow interactive exploration of the topics.

The researcher has to specify the number of topics beforehand, and it is advisable to compare the performance of differently fine-grained models. As a rule of thumb, when analyzing text, between 30 and 150 topics are common choices. The topics can be sorted by their coherence, using for example their so-

called Umass value or their perplexity, and evaluated based on their interpretability. When giving a substantive interpretation to these topics, it is important to let humans with domain knowledge assess whether the results are consistent with what one would expect, and conducting a formalized test to compare the results with results of human coding can be a useful step, too. Topic models are sensitive to proper preprocessing: if not done carefully, the topics will be characterized by mainly meaningless stopwords. Jacobi, Van Atteveldt, and Welbers (2016) give detailed recommendations on how to use LDA to analyze large corpora of journalistic texts.

Supervised machine learning

Supervised machine learning comprises a set of techniques that can be used to build a model when both dependent and independent variables are known for a subset of the data. One supervised machine learning technique that communication scientists are usually familiar with is regression analysis: Once a regression equation has been estimated, it can be used to predict an outcome based on a vector consisting of values for the independent variables. Even if no case with this specific combination of values exists in the dataset, the expected value for the dependent variable can be estimated by simply filling in the values of the independent variables into the equation.

Supervised machine learning is used when a so-called training dataset is available, in which both dependent and independent variables are available, but when the dataset of interest, for which the dependent variable is unknown, is larger by orders of magnitude. For example, movie reviews typically consist of a longer text (and possibly some additional information, like the release year, the director, or a list of actors) and a rating (for example, one to five stars). Given a dataset of 2,000 of such reviews, one could use 1,000 of the cases to estimate a model (to "train a classifier") that predicts the star rating based on some lexical features of the text (or the list of actors). The remaining 1,000 cases can be used to assess how well the classifier performs. Once trained, the classifier can be used to predict the rating of an arbitrary number of movies, for which the rating is unknown.

Such an approach can be used in sentiment analysis, the attempt to find out how positive or negative the tone of a text is (see below). Others have trained classifiers to assess automatically the topic of news articles or the frames that occur in them. For example, Burscher, Odijk, Vliegenthart, De Rijke, and De Vreese (2014) have successfully employed supervised machine learning to code four generic frames: conflict, economic consequences, human interest, and morality. Comparing different approaches, they find that a holistic approach, in which the final judgment (frame is present or not) is predicted directly, performs better than an indicator approach, in which each of the several specific indicators present in the codebook are predicted separately and then combined. They use tf-idf weights as features (independent variables) in their models, but note that that the performance of their classifiers could probably be improved further by using more sophisticated representations that take semantic or syntactic features of the text into account.

For many applications, it is sufficient to have training datasets in the order of magnitude of 1,000 cases, which means that even if a training dataset does not exist yet, a training dataset can be constructed with limited resources.

This means that machine learning can be an extremely powerful tool for communication scientists, as it allows to code automatically crucial information (like topics, genres, or frames) in large datasets for further analysis.

There are a number of different supervised machine learning techniques, all of them having different advantages and disadvantages, which cannot be discussed within the scope of this entry. The most common techniques are naïve Bayes classifiers, support vector machines, decision trees and random forests, and the already mentioned regression approaches. While they differ in terms of assumptions, efficiency, and accuracy, there is not one approach that is clearly superior to the others. It is common and can be considered to be best practice to train different classifiers and compare their performance, for example using so-called *k*-fold cross validation. Also, a so-called ensemble of classifiers can be used, that combines their predictions.

Sentiment analysis

One popular type of analysis within Big Data analysis is sentiment analysis. Sentiment analysis aims at determining the tone of a text and it is widely used in commercial applications (for example, to determine how positive or negative the sentiment towards a company is on social media) as well as in academic research. While the descriptive statistics of sentiment scores and their development over time can offer interesting insights in themselves, sentiment scores are frequently used as dependent or independent variables in statistical models. For example, sentiment scores have been used to predict reactions to messages on social media.

In its most simple form, sentiment analysis simply counts the number of positive and negative words from pre-defined lists, possibly attaching a weight to account for different levels of intensity for each of the positive and negative words. Such a bag-of-words approach has some serious shortcomings for sentiment analysis. In particular, it cannot deal with negation: “not good” is counted as positive, because the information that “not” belongs to “good” is lost in a BOW representation. This is why this approach is only used for pedagogical purposes nowadays. Purely dictionary-based BOW sentiment analysis has been replaced by more powerful algorithms that, while still working with pre-defined lists of lists of positive and negative words, have abandoned the BOW approach and take sentence structure into account. Approaches like the popular Sentistrenght algorithm (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) can deal with negation, but also with intensifiers: words like “very” imply multiplying the score of the attached word with a specific factor. A third approach is to abandon the use of pre-defined lists altogether and use supervised machine learning instead. Next to empirical arguments that demonstrate how successful this approach can be compared to the former ones (González-Bailón & Paltoglou, 2015), it also solves the methodological problem that the composition of predefined lists of words inherently has some validity issues. On the downside, machine learning can be sensitive to the specific training dataset and therefore, it seems more problematic to reuse a trained classifier for a completely different dataset than it is with the list-based approaches, which can be considered to be more universally applicable. With all methods, though, recognizing irony and sarcasm remains a largely unsolved problem.

Next to the general technique (BOW, parsing sentences, supervised machine learning), sentiment analysis approaches also differ in terms of the dimensions on which the texts are evaluated. The most straightforward approach is to use a one-dimensional scale reaching from negative to positive. Others use two axes for negativity and positivity, arguing that a text can potentially contain a high amount of positivity and negativity at the same time. Along similar lines of reasoning, other algorithms provide a scale ranging from objective to subjective along with a negative–positive scale. This indicates that while the negative/positive distinction lies at the core of sentiment analysis, it is by no means restricted to this. For example, the frequently used LIWC framework (Pennebaker, Booth, & Francis, 2007) even measures a set of different emotions and cognitive processes.

It should be noted that both unsupervised and supervised machine learning can be conducted on different scales. While many packages like the popular scikit-learn (Pedregosa, Varoquaux, Gramfort, Michel, & Thirion, 2011) are mainly designed for use on single machines, other packages, like Apache Spark's MLib scalable machine learning library, are explicitly designed to run on clusters of multiple computers.

Network analysis

With the availability of large-scale datasets of user interactions, mainly on social media, the method of social network analysis has received increased attention. Common are analyses of networks of retweets, @-mentions, or similar forms of interaction where actors are conceptualized as nodes and the interactions as edges. The latter can be directed (A is mentioning B) or undirected (A and B are occurring together).

Such network analyses are conducted on datasets of various sizes, not necessarily big ones, and some of the tools popular in that field are only capable of analyzing small datasets (e.g., --> NodeXL, in this encyclopedia). Tools have been developed that are suitable for larger networks (e.g., SNAP by Leskovec & Sosič, 2014); and in fact, they offer an excellent way of analyzing Big Data. Both the graphical visualization as the calculation of key metrics like centralities or clustering coefficients are good ways to analyze immense datasets and to reduce them to an interpretable representation. While the analysis of social media data is an obvious application, network analysis has been applied to a much wider range of data, including online shopping data, communication structures on online forums, and networks of words co-occurring in large text corpora (e.g., Sudhahar, Veltri, & Cristianini, 2015). Given the fact that an increase in the number of nodes exponentially increases the number of potential edges, researchers that employ network analysis can quickly get into a situation where the hardware and software restrictions discussed above apply. This is why using one off-the-shelf program is not an option for conducting these analyses on large datasets. Instead, by using one of the many network analysis modules for programming languages like Python, researchers can split the analysis into smaller tasks and thus create more efficient processes. It might even become necessary to make use of one of the Hadoop-based graph databases mentioned above.

Recently, researchers have started combining such network-analytical approaches with other techniques. One could think of weighing edges in a

network by a score obtained by sentiment analysis, or using regression techniques for predicting the structure of a network or its change over time. If such datasets grow in size, it is possible to make use of so-called graph databases, which allows not only storing data in key-value pairs like in the aforementioned NoSQL-databases, but additionally allows freely defining relationships between elements.

Challenges

Ethical concerns

The use of Big Data in social-scientific analysis raises a number of ethical concerns, which arise from the difficulties of applying standard procedures to safeguard ethical standards to Big Data research. For example, while the standard of getting *informed consent* from research participants is widely embraced, some Big Data projects are confronted with the practical impossibility of having tens of thousands of people signing such forms. Here, a gap exists between conventional research ethics and daily practices – especially when compared to the commercial sector, where A/B testing is common and perceived as unproblematic. In such tests, a service is modified for some people and their reactions are compared to the reactions of those using the unmodified version, which effectively makes people research participants without their knowledge. Another potential ethical issue is the question of privacy and anonymization. Even if information like usernames is stripped from a dataset, people might be identifiable because of the amount of data known about them. Researchers are not always aware of which inferences can be made, which may lead to de-facto de-anonymization of parts of a dataset. Related to this is the question of what is considered to be *public*. While traditionally, a conversation has been considered private and a newspaper article public, opinions differ on how to treat tweets or blog posts. Although they are de facto accessible to everyone, researchers tend to err on the side of caution and sometimes consider them private – which, in combination with the practical impossibility of getting an informed consent from all social media user – raises the question of how they can be analyzed in an ethically responsible way. This is even more problematic in environments like Facebook, where there are a lot of shades between public and private: a post might be visible to everyone, to members of specific groups, to friends of friends, to friends only, and so on.

Social-scientific research standards

Social-scientific research standards demand research to be reproducible, and in fact, journals and funding agencies increasingly demand sharing reproduction datasets and code. On the one hand, the fact that Big Data analyses are conducted automatically and custom-written scripts and programs are used, makes it easy to share the exact procedures employed. On the other hand, if data are too large to store with normal means or some other type of specialist hardware is required, this can be tricky – although it can be solved by using the above-mentioned cloud services. Specific attention to this is necessary.

Reproducibility is endangered by legal restrictions as well. For instance, Twitter prohibits sharing datasets of tweets. Also in other cases, measures that would be necessary to enable reproducibility may violate terms-of-services of the providers of the data sources used.

A third threat to reproducibility lies in the fact that Big Data is often dynamic in nature. There might be not *one* version of record, especially when data are retrieved on-the-fly using APIs of online services to collect or process the data. The dependence on third parties is problematic as well. When a company like Twitter, Facebook, or LinkedIn changes its API – something that happens rather frequently – it may happen that a specific analysis cannot be conducted any more. In particular, replication may become impossible. As the exact working of the underlying mechanisms remains unknown, essentially introducing a black-box in the research process, researchers cannot be completely sure that they measure what they want to measure. The dependency on third parties creates a divide between those who do have access and those who do not – because of technical skills, because of financial means (access to some data is sold by commercial companies), but also because of cooperation with insiders: Companies like Facebook publish research on their own data, working together with academic partners – which means access to data that are inaccessible to the rest of the academic community.

Future directions

Big Data analysis has been introduced to communication science only recently, which is one important reason why best practices still have to emerge. In addition, is likely that the way Big Data are analyzed will change, and it is hard to predict how the maturing process will develop. Nevertheless, several trends can be identified. First of all, Big Data analysis becomes increasingly common in the discipline, which is illustrated by the increasing cooperation with computer scientists, but also by the increasing number of communication science departments that offer courses on programming and Big Data-related methods courses. This also means that the boundaries of the discipline have to be revisited, and that it has to be addressed how a communication science approach to Big Data analysis differs from the approaches taken by other disciplines. Big Data methods still have to find their place in the methodological toolkit. It has to be addressed in which cases these new methods are appropriate to use, how they can be combined with and how do they relate to traditional methods.

Second, while Big Data analysis in communication science is often understood as analysis of social network data or other forms of user-generated content on the internet, the scope will broaden as more data become available and accessible for social scientists. One interesting development in this regard is the raise of the so-called *Internet of Things*, the connection of devices like fridges, heating, or wearable technology to the internet. Also this development addresses the self-definition of the discipline, as it has to be re-evaluated in how far such types of communication fall within the domain of communication science. This also means that old theories have to be revisited and new ones might have to be developed to account for both the changing communication environment and the changing methodological possibilities.

Third, Big Data analysis also becomes more relevant for fields that communication science students come to work in. As follows from the points above, this is true for those staying in academia, but it is not limited to that. In the commercial sector, the field of social media monitoring has become an important aspect of organizational communication. In the journalistic field, an increasing demand for data journalists can be observed. This means that it has to

be discussed in how far and in which form Big Data analysis skills have to be integrated into the curriculum.

Fourth and finally, both underlying algorithms and available toolkits are being improved. For example, in the field of topic modeling, algorithms are being developed that take into account the specific problem of identifying topics in shorter texts; and others work on algorithms that relax unrealistic assumptions of such models (for example, the assumption that word order does not matter). The steady increase of R packages and Python modules that implement algorithms used in Big Data research makes these advances accessible to communication scientists. Yet, in spite of the general accessibility of cloud computing, solutions that make the analysis of datasets of non-trivial size accessible without technical skills are still not available.

See also: Big Data, Collection of (IECRM0015); Cluster Analysis (IECRM0025); Content Analysis, Automatic (IECRM0043); R (software) (IECRM0201); Social Network Analysis, General (IECRM0235); Social Network Analysis (Social Media) (IECRM0236);

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borra, E., & Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, *66*(3), 262–278. doi:10.1108/AJIM-09-2013-0094

boyd, D., & Crawford, K. (2012). Critical questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Burscher, B., Odijk, D., Vliegenthart, R., de Rijke, M., & de Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures, 8*(3), 190–206. doi:10.1080/19312458.2014.937527

Cleveland, W. S. (2001). Data Science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review / Revue Internationale de Statistique*, *69*(1), 21–26. doi:10.1111/j.1751-5823.2001.tb00477.x

De Rooij, O., Vishneuski, A., & De Rijke, M. (2012). xTAS: Text analysis in a timely manner. *Proceedings of the 12th Dutch-Belgian Information Retrieval Workshop*, 89–90.

González-Bailón, S., & Paltoglou, G. (2015). Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. *The ANNALS of the*

American Academy of Political and Social Science, *659*(1), 95–107. doi:10.1177/0002716215569192

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism, 4*(1), 89–106. doi:10.1080/21670811.2015.1093271

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. doi:10.1177/2053951714528481

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., … van Alstyne, M. (2009). Computational social science. *Science*, *323*(February), 721–723. doi:10.1126/science.1167742

Leskovec, J., & Sosič, R. (2014). SNAP: A general purpose network analysis and graph mining library in C++. Retrieved from http://snap.stanford.edu/snap

Manning, C. D., Bauer, J., Finkel, J., Bethard, S. J., Surdeanu, M., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. Retrieved from http://aclweb.org/anthology/P14-5010

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

Murthy, D., & Bowman, S. a. (2014). Big Data solutions on a small scale: Evaluating accessible high-performance computing for social research. *Big Data & Society*, *1*(2), 1–12. doi:10.1177/2053951714559105

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., &Thirion, B. (2011). Scikit-Learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC. Austin; TX: LIWC.net.

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big Data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13. doi:10.1177/0002716215572084

Sudhahar, S., Veltri, G. a., & Cristianini, N. (2015). Automated analysis of the US presidential elections using Big Data and network analysis. *Big Data & Society*, *2*(1), 1–28. doi:10.1177/2053951715572916

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558. doi:10.1002/asi.21416

Van Atteveldt, W. (2008). *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content*. Charleston, SC: BookSurge.

Yıldırım, A., Üsküdarlı, S., & Özgür, A. (2016). Identifying topics in microblogs using Wikipedia. *Plos One, 11*(3). doi:10.1371/journal.pone.0151885

Further reading

Freelon, D. (2014). On the cutting edge of Big Data: Digital politics research in the social computing literature. In S. Coleman & D. Freelon (Eds.), *Handbook of Digital Politics*. Northampton, MA: Edward Elgar.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. doi:10.1093/pan/mps028

Mahrt, M., & Scharkow, M. (2013). The value of Big Data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20–33. doi:10.1080/08838151.2012.761700

Russel, M. A. (2013). *Mining the social web. Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more* (2nd ed.). Sebastopol, CA: O'Reilly.

Russell, S. J., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd international ed.). Harlow, UK: Pearson.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47. doi:10.1145/505282.505283

Bio

Damian Trilling is an Assistant Professor of Political Communication and Journalism at the Amsterdam School of Communication Research, Department of Communication Science, University of Amsterdam, Netherlands. His research interests include the use of new media in political communication and journalism, selective exposure, and personalized communication. Both in research and teaching, he focuses on the use of innovative computational social science methods.