An Introductory Guide for Conducting Psychological Research with Big Data

Michela Vezzoli and Cristina Zogmaister

Department of Psychology, University of Milano-Bicocca

Author Note

Michela Vezzoli https://orcid.org/0000-0003-2122-2764

Cristina Zogmaister https://orcid.org/0000-0002-1540-7503

We have no known conflict of interest to disclose. The funding for this research was provided by eGlue SRL (Segrate, MI, Italy). We are grateful to Prof. Douglas Steinley, Editor in Chief of Psychological Methods, and the two anonymous reviewers for their valuable comments on earlier drafts of this manuscript.

Correspondence concerning this article should be sent to Cristina Zogmaister, Psychology Department, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milan. Email: cristina.zogmaister@unimib.it

This manuscript is a part of the first author's doctoral dissertation (discussed in February 2020).

The final article is available via its DOI: 10.1037/met0000513

Abstract

Big Data can bring enormous benefits to psychology. However, many psychological researchers show skepticism in undertaking Big Data research. Psychologists often do not take Big Data into consideration while developing their research projects because they have difficulties imagining how Big Data could help in their specific field of research, in imagining themselves as "Big Data scientists", or for lack of specific knowledge. This paper provides an introductory guide for conducting Big Data research for psychologists who are considering using this approach and want to have a general idea of its processes. By taking the Knowledge Discovery from Database steps as the *fil rouge*, we provide useful indications for finding data suitable for psychological investigations, describe how these data can be pre-processed, list some techniques to analyze them, and programming languages (R and Python) through which all these steps can be realized. In doing so, we explain the concepts with the terminology and take examples from psychology. For psychologists, familiarizing with the language of data science is important because it may appear difficult and esoteric at first approach. As Big Data research is often multidisciplinary, this overview helps build a general insight into the research steps and a common language, facilitating collaboration across different fields.

Translational abstract

Technological advances have led to an abundance of widely available data on every aspect of life, called Big Data. Big Data provide psychologists with new means to research psychological constructs in new ways. Psychologists often show excitement when talking about these new data opportunities. However, this enthusiasm has not yet led to extensive use of Big Data in the psychological community. One reason for this is that psychologists usually do not acquire the various skills and competencies that successful Big Data research requires during their training. This work provides an introductory guide for conducting Big Data research for psychologists who are considering using this approach and want to have a general idea of its processes and techniques. It gives useful indications for finding data suitable for psychological investigations, describes how these data can be pre-processed, and enlists some techniques and tools to analyze them.

Keywords: Big Data; Tutorial; Data science; Data mining; Psychology.

An Introductory Guide for Conducting Psychological Research with Big Data

Data are everywhere. Right now, millions of data of different nature are being recorded and stored somewhere. A term often used for this phenomenon is Big Data. The concept of Big Data was initially associated with three Vs: *volume, variety*, and *velocity* (Laney, 2001). *Volume* refers to the large amount of data. There is no specific threshold above which data become *big*. Rather, data become big when their scale exceeds the computing capacities of generally available hardware and software. *Variety* refers to the many forms that Big Data can take (e.g., text documents, geolocalization, video, audio). Each of these potential forms requires its own method of acquisition, management, and analysis. *Velocity* refers to the high speed with which data are often acquired and need to be managed. However, a univocal and systematic definition of Big Data is missing. To examine how researchers conceptualize this concept, Favaretto and colleagues (2020), interviewed 39 psychology, sociology, and data science researchers involved in Big Data. They found that most participants, besides focusing on the attributes ascribed to Big Data (i.e., the three Vs), identified other aspects, such as the nature of the data (e.g., digital data sources; data generated through people's daily activities) and processing techniques (e.g., need of specific algorithmic or computational processes), as determinant components of Big Data. A relevant step toward a conceptualization of Big Data within psychological and cognitive science was made by Paxton and Griffiths (2017). They recognized a considerable overlap between the concept of Big Data and that of naturally occurring datasets, which are typically gathered as observations of people, behaviors, or events by nonscientists for nonscientific, nonexperimental purposes (Goldstone & Lupyan, 2016). Within this paper, we frame Big Data as data characterized by high volume, high variety, and high velocity, which necessitate specific algorithmic and computational processes to be analyzed. Also, we frame Big Data as data that "were not collected for experimental purposes but could - with a little creativity and the right tools - provide insight into cognition and behavior" (Paxton & Griffiths, 2017, p.

1631). Considering Big Data as naturally occurring implies that they are not collected for scientific purposes.

Big Data provide psychologists with new means to research psychological constructs in new ways. A further understanding of people's psychological processes and behaviors may lie at the tips of people's fingers, the keyboards of their computers, or the touch screens of their smartphones. Every time people interact with an informatic device, data are registered somewhere: from a text message on their mobile phone to an Amazon order to the use of loyalty cards in supermarkets. Data like these can say something about individuals, their behavior, and, perhaps, why they behave the way they do and their personality. Thus, digital footprints, traces of behaviors people leave while going through cyberspace, often unintentionally, may be useful for answering psychological questions.

Many scientific fields have leveraged Big Data (e.g., computer science, business, physics), but psychology research has only recently started to tap into Big Data (e.g., Beaton et al., 2016). Most published materials discuss the use of Big Data in general, as a new and valuable methodological framework in psychological research (e.g., Adjerid & Kelley, 2018). Other works deal with the use of Big Data in subfields such as health psychology (e.g., Yetton et al., 2019), industrial & organizational psychology (e.g., Kobayashi et al., 2018), consumer psychology (Vezzoli et al., 2020), clinical psychology (e.g., Hollon et al., 2019), and cognitive psychology (e.g., Stevens & Soh, 2018). Three leading journals have dedicated special issues to the topic (*Psychological Methods*, *Current Opinion in Behavioral Sciences*, *Behavior Research Methods*, and *Zeitschrift Für Psychologie*). Psychological research organizations have organized conferences (Big Data in Psychology 2018, 2019, and 2021 by Leibniz Institute for Psychology Information - ZPID[1]) and forums (Big Data in personality and social psychology, by the Society of Personality and Social Psychology[2]).

[1]https://conferences.leibniz-psychology.org/index.php/index/index/index/index

[2]http://www.spsp.org/events/SPF/2019

Although psychologists often show excitement when talking about these new data opportunities, this enthusiasm has not yet led to extensive use of Big Data in the psychological community for several reasons. Paxton and Griffiths (2017), summarized these reasons in three categories: the "imagination gap," the "cultural gap," and the "skills gap."

The "imagination gap" refers to the observation that psychologists often do not consider their research questions as Big Data problems. Despite the availability of massive amounts of data, few psychologists can envision a dataset that would address an important theoretical question in their field. Bridging the imagination gap will require some work to adjust the psychologists' idea of the possible scope of data beyond data generated through research. For example, researchers interested in understanding categorization might investigate tagging behavior in the Yahoo Flickr Creative Commons 100M dataset (Thomee et al., 2015). Play-by-play sports records might be useful for studying team dynamics (Horowitz, 2015). Decades of online chess game records could shed light on expertise and decision-making (e.g., Free Internet Chess Server Database, https://www.ficsgames.org/).

The "cultural gap" refers to psychologists' difficulty envisioning themselves as Big Data scientists and Psychology as a science that can benefit from Big Data. Thus, the discontinuity between interest in Big Data research and the use of Big Data in research can be partially attributable to a lack of role models and acceptance of these new data resources.

Finally, the "skills gap" refers to the psychologists' lack of competencies and skills required to deal with Big Data. Indeed, Big Data are too large to be opened or handled in standard spreadsheet software. Similarly, they are too complex to be adequately analyzed by simple inferential statistics. Skills like data acquisition (e.g., data scraping), data cleaning, machine learning, and scientific programming are essential to Big Data research but are not often taught in traditional undergraduate and graduate courses in psychology. Even though Big Data scientists have designed a wealth of training materials (e.g., massive open online courses, tutorials) that can provide excellent jumping-off points for researchers from any

domain, effective training for cognitive scientists and psychologists requires the creation of tailored training opportunities.

This paper aims to help bridge the skills gap (Paxton & Griffiths, 2017) by taking the reader through the journey of Big Data research, from the initial phase of data acquisition to the result of the analysis, using the Knowledge Discovery from Database process as the *fil rouge*. We will describe a broad set of concepts related to Big Data acquisition (e.g., data sources), preprocess (e.g., data integration, outlier identification, feature selection, data reduction), and analysis (e.g., supervised learning techniques, model validation, and interpretability). In this way, the reader can get a general vision of the process of Big Data research, its steps, and some methodological issues of this research. In describing the steps, we will use examples of psychological Big Data research. These will help the reader better comprehend the concepts presented and get an idea of the types of questions that were asked, the data used, and the results obtained. By showing some of the possibilities of Big Data research for psychology, this article may contribute to bridging the imagination gap (Paxton & Griffiths, 2017). Given its purposes, this paper is aimed at those psychologists who have little or no prior experience with the fundamental concepts and processes of Big Data research and want to understand the steps to conduct it. Thus, to grasp the concepts described in this article, no technical knowledge or programming experience is required, but some statistical knowledge would be useful.

In addition to this guide, the interested readers may also find Chen and Wojcik's (2016) Big Data guide for psychologists helpful, especially if they are interested in working on text data. Compared to it, our guide provides an *end-to-end* roadmap of Big Data research and a much broader set of concepts related to data preprocessing and transformation (e.g., data integration, outlier identification, feature selection, data reduction) and analysis (e.g., supervised learning techniques, model validation and interpretability).

The Role of Theory in Big Data Research

Before starting our journey in the process of conducting Big Data research, we spend a few words clarifying the role of theory. Indeed, psychologists might be hesitant to approach Big Data research due to the false belief that Big Data research is a-theoretical (Anderson, 2008), or that Big Data investigation is at odds with psychologists' typical way of conducting research, which mainly consists in expressing hypotheses and subsequently testing them empirically. Based on a vision of Big Data research as collecting large datasets, feeding them to a statistical technique that finds out the relations and generates prediction models, one might think that, at best, the role of theory is to help make sense of the results. However, this would be an incorrect depiction of Big Data research, especially when it comes to its applications in psychology.

Theory has a central role in Big Data research, whether this is data-driven or theory-driven. When we investigate the traces of human behaviors in natural settings, theories guide the choice of data sources and variables, the ways to treat them, and the conclusions that are reached. A given dataset may become interesting for investigation exactly because there is a theory, or a set of hypotheses, regarding the variables it contains.

As discussed by Goldstone and Lupyan (2016), psychologists can use naturally occurring datasets to perform theory-driven research, which can provide external validation and ecological validity to the results discovered through experimental studies. For example, Berger (2016) investigated primacy effects in the citations of scientific journal articles and found that the articles positioned first in journal issues were cited more frequently than the others. Besides validation, theoretically driven Big Data research has the added advantage of helping researchers understand whether the effects observed in the laboratories have practical significance in real-world behaviors. Moreover, Big Data experiments can be created by manipulating key variables and observing how this affects natural behavior in real-life settings (albeit, as we will discuss later, this can pose informed-consent problems that need

careful consideration). Kramer and colleagues (2014), for example, manipulated the valence of Facebook posts and examined how this affected the mood of the users.

Furthermore, the analysis of naturally occurring datasets can often be conceived as a form of exploratory research, through which psychologists can conduct an initial investigation of the variables that might be involved in a phenomenon and get insight for further experimental research. With this strategy, as noted by Goldstone and Lupyan (2016), researchers can be more confident that what they are studying in the laboratory has relevance to everyday life. In all of these ways, Big Data research becomes a companion, not a substitute, to the traditional ways of doing psychological science, helping theoretical development in exploratory and confirmatory research.

The remainder of this paper is structured as follows. We will first describe the Knowledge Discovery from Database process (KDD; Fayyad et al., 1996), which provides a structured step-by-step path for analyzing large and complex datasets. Next, we will address, in separate sections, data acquisition sources and methods and the fundamental techniques for data preprocessing and transformation. Later, we will introduce data mining techniques. Finally, we will discuss two programming languages to implement them. Expanding these competencies and skills will be key for psychologists to make an impact in the Big Data era.

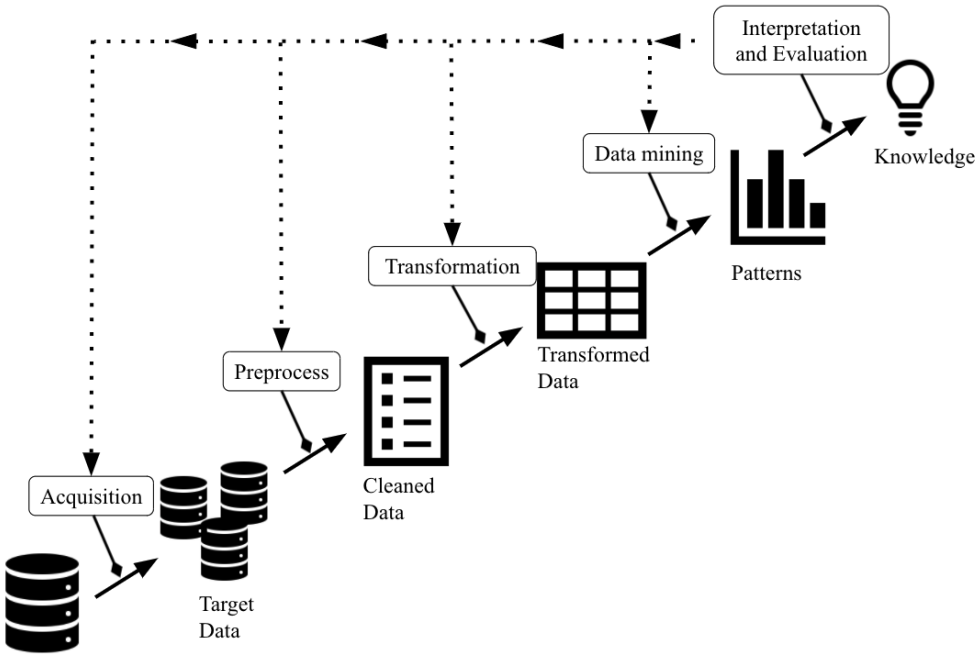
The Knowledge Discovery in Databases Process

To find valid, novel, useful, and understandable patterns from data, it is helpful to follow the KDD process (Fayyad et al., 1996). The KDD process is developed in five steps, as depicted in Figure 1. The first, *data acquisition*, concerns the identification and extraction of data from relevant sources (see “Data Acquisition and Sources” section). *Data preprocessing* aims to increase data quality by supplementing missing attributes, removing duplicates, and resolving inconsistencies (see “Data Preprocessing” section). *Data transformation* concerns the selection of the relevant variables and adapting them to the needs of the data mining algorithm (see “Data Transformation” section). *Data mining* aims to model the data and

derive patterns, learn functions, or make predictions from the data. In the last step, the results are checked for validity (see "Data Mining" section). Finally, in the *interpretation and evaluation* step, the user examines the validity of the results and interprets them (see "Validation and Interpretation" section).

Figure 1.

The KDD process.



As the dotted arrows in Figure 1 show, KDD is iterative because it often involves repetition and user interaction. Thus, one can move back to adjust previous steps whenever required. Consequently, KDD has many "degrees of freedom," meaning that one cannot present one formula or make a complete taxonomy for the right choices for each step and application type. Thus, it becomes essential to understand the process as a whole and know the aims of each step.

Usually, data mining is considered the key step of the entire KDD. This emphasis makes it difficult to distinguish KDD from data mining, especially in practical terms. However, KDD is the overall process (i.e., the five steps in Figure 1), and data mining is a particular step within it. Data mining is the application of specific algorithms for extracting

patterns from data. Therefore, the terms data mining and KDD will be used as separate concepts here.

Data Sources and Acquisition

Once the research aim is established, the next steps consist in determining which kinds of data might be useful for our project, identifying possible data sources, and collecting the data. Data acquisition is critical because data mining techniques learn and discover from the available data: If crucial variables or sources of cases are overlooked, the entire knowledge discovery process will fail. While it is sensible considering as many variables and cases as possible, collecting, organizing, and operating large datasets can be expensive and error-prone. Thus, the selection of cases and variables is a trade-off between exhaustivity and manageability. Here, psychologists' theoretical training can be of great value because it can drive the selection of useful data.

Social media data may be of great interest to psychologists because in social media people express preferences, opinions, attitudes, and interact with others. Social media platforms record a lot of information on these behaviors. However, they often lack user-friendly interfaces to ease data extraction from their systems and require using a set of protocols and technologies known as Application Programming Interfaces (APIs). For a broad idea of how to use APIs, we suggest Chen and Wojcik (2016).

Another valuable source of data is smartphones. As they accompany individuals everywhere and contain numerous sensors, such as accelerometers, GPS, Wi-Fi, light sensors, and microphones (Lane et al., 2010), smartphones collect a great deal of information about human behaviors. They routinely record data related to the social and mobility sphere: for instance, calls, texts, locations, and movements. Research with smartphone data is flourishing in computer science, and, recently, smartphones have entered the methodological toolkit of psychologists (Gosling & Mason, 2015; Wrzus & Mehl, 2015). For example, smartphone data have been used to study the relationship between partisanship and social distancing during the

COVID-19 pandemic (Gollwitzer et al., 2020), sleeping patterns and postures (Wrzus et al., 2012), interpersonal behaviors in group settings (Schmid Mast et al., 2015), emotional variation in daily life (Rachuri et al., 2010; Sandstrom et al.,2017), the links between mobility patterns and depression (Chow et al., 2017), academic performance (Wang et al., 2015), and happiness (Lathia et al., 2017). Researchers can detect more complex behaviors when they combine data from different sensors and sources. For example, combining data from microphones and accelerometers enables detecting common behaviors such as clapping, vacuuming, or taking out the trash (Lu et al., 2009). Merging GPS data and signals from Wi-Fi transmitters or accelerometers can be used to track individuals within a building (Chon & Cha, 2011), detect modes of transportation (Hemminki et al., 2013), and monitor pedestrian behavior (Wang et al., 2016). Information about smartphone usage, combined with data from other sensors, such as ambient light sensors, microphones, or accelerometers, can reveal users' sleep patterns (Hao et al., 2013) or alertness (Murnane et al., 2016).

Turning to other Big Data storing systems, companies' databases record information about large population segments that may be of psychological interest. However, accessing these data could be complicated for people external to the company. Further, the laws governing the use of companies' data (e.g., in the European Union, the General Data Protection Regulation) and the terms of service of the companies themselves may impose limitations on the types of data (e.g., personal information) that can be analyzed. Thus, companies and researchers must come to a mutually beneficial agreement for the use of these data.

Several companies sell data that might be relevant for psychological research, such as Acxiom (demographic data for direct marketing), Nielsen (audience measurement for TV, radio, music, and newspapers), and Equifax (demographic data on people around the world). However, if budget limits prevent data purchase, an alternative is to use archival repositories that allow acquiring data for free. Archival repositories are large database infrastructures set

up to manage, share, access, and archive datasets. Among these, there are the archives of political institutions (e.g., the European Union) and non-profit international organizations (e.g., World Health Organization, OpenPsychometrics, Word Association Lexicons). In these repositories, data owners often provide interfaces that facilitate data retrieval. In Appendix A, we have listed some data repositories along with a description of the downloadable data. A useful tool for searching data repositories is Google Data Search (https://datasetsearch.research.google.com/), a Google search engine that helps researchers find online and freely available data.

Specific Challenges for Big Data Quality

Traditional data acquisition methods most commonly used in psychology (i.e., experiments, questionnaires, and interviews) allow researchers to design their studies carefully, determine how to measure the variables of interest, and hence have a certain control on data acquisition. With Big Data, researchers must often rely on data collected by others. The well-known sources of errors and noises (e.g., missing values, outliers) that, if not identified and corrected, negatively affect the results' quality in more traditional psychological research, are present also in Big Data research. This means that, in general, psychologists are equipped with adequate knowledge to address these kinds of issues. However, they need to be aware that the likelihood of data quality issues with Big Data is substantially higher, due to the much lower or absent control of the experimenter on the measurement process.

However, in Big Data research there are also specific challenges. For example, working with Big Data usually entails integrating datasets from different sources. This can produce redundancies and inconsistencies, which can negatively affect the quality of the results. Another common issue with Big Data is the presence of a high number of variables (i.e., high-dimensionality), which can introduce bias in the analysis.

A further source of bias in Big Data research derives from sampling. Imagine we want to detect potholes in the streets. We could create a smartphone app that uses accelerometer data to acquire such knowledge, ask our citizens to download it, and have real-time knowledge on which roads need fixing. The city of Boston did exactly this and surprisingly enough, potholes were being disproportionately reported in affluent and young neighborhoods. As further analysis revealed, residents in these neighborhoods were more likely to own smartphones with network access (and cars) than residents living in lower-income neighborhoods (Harford, 2014). Psychologists are aware of the consequences of sampling bias, as it is also a problem of small data research (e.g., heavy reliance on White, Educated, Industrialized, Rich, and Democratic samples; Henrich et al., 2010). Even though Big Data may help reach larger and more diverse samples, there is the risk that some researchers confuse Big Data with the representativity of the population. Indeed, big sample size does not necessarily mean that the data are representative of the target population. Thus, whether with small or with big samples, there is always a problem when we infer the characteristics of the population from the results of non-representative samples.

Ethical Considerations on Big Data

The increasing popularity and richness of Big Data has prompted new questions about ethics, and efforts to answer them are still developing. Controversy over Big Data research ethics has been exacerbated by previous Big Data research. In one example (reported in Paxton, 2020), academic researchers scraped over 70,000 profiles from a dating website, conducted their study, and subsequently made this corpus of data (containing potentially identifying information) openly available, all without participants' consent or oversight by an Ethical Committee, and claimed that ethical objections were unwarranted as the collected data were already public. This study alarmed the public mainly because users were not given the opportunity to provide informed consent.

In general, Big Data research challenges the traditional approach to obtaining informed consent. For example, it is impractical to obtain consent to participation in the research from every person who posts a tweet of interest, or transits through a certain subway turnstile if we are investigating mobility behaviors. If consent is offered by agreeing with the service terms and conditions, only a few people read them. Also, Big Data studies often disregard other crucial principles in research involving people, such as that participants should be allowed to withdraw from a study at any time and be informed of all the uses of their data.

Under certain circumstances, for example, when the data we use are already anonymized and aggregated, informed consent might not be necessary ("Time to Discuss Consent in Digital-Data Studies," 2019). However, the way to ethically handle data available in the public domain is debated. The lack of regulations has led some practitioners to scrape online personal data in questionable ways that, in some cases, have resulted in legal litigations (e.g., HiQ and LinkedIn case; Stringam et al., 2021). Since informed consent in Big Data research can be problematic and scraping data from websites may be tricky in some instances, researchers may opt to follow some good practices to make informed decisions on data collection. For example, if it is planned to collect data from Twitter, reading its Terms of Use (TOU) statement and the robot.txt beforehand would be recommended (both can be found on the website of Twitter). The TOU statement indicates how the website should be accessed and used, and what are the acceptable methods for gathering and using the collected information (Chen & Wojcik, 2016; Davies, 2020).

Researchers should evaluate ethical issues related to their Big Data projects from the very beginning. Drawing from the theory of "privacy as contextual integrity" (Nissenbaum, 2010), Zimmer (2018) provided useful suggestions to guide ethical decision-making in Big Data projects. Big Data poses ethical concerns that researchers may not be aware of, as they leverage new sources of information and methodologies (e.g., scraping) that are not typical of

traditional research. Researchers should ask for support from institutionalized agencies trained and tasked to handle ethical issues. Thus, research ethics committees can play a central role in extending and adapting ethical principles to Big Data research and inducing researchers to think critically about the ethics of their projects.

Data Preprocessing

After receiving the raw data, the first thing researchers must do is validate them. Similar to small datasets, noise hinders the quality of the results. To allow valid conclusions, the data have to be complete, consistent, and accurate (Han et al., 2012). Indeed, it is a myth that we can apply algorithms to raw data and expect to receive useful insights as output (Lohr, 2014). Data preprocessing and transformation are the steps that require the most time and effort. For more information on data preprocessing, we recommend García, Luengo & Herrera (2015).

Data Integration

In Big Data research, data from many sources must often be integrated into one dataset. For example, to get a view of individuals' health status, the information to be combined can include data from wearable watches (e.g., heart rates, sleep rhythms) and smartphones (e.g., mobility data, food and water intake from apps), but also data from external systems, such as digital health records (e.g., laboratory test results, allergies) from National Healthcare Systems. Data integration consists in merging data from multiple sources into a unified dataset. Imagine being interested in investigating whether there is a relationship between specific types of crime and the weather conditions: we could relate data from an archive containing crime data in different cities and another archive containing weather data in those cities. To get a unified dataset of these data, we could merge them based on postcodes. However, data collected from different data sources may not be ready for integration and need prior processing. For example, imagine that the crime data present a more granular measurement unit (e.g., district instead of city-level) than the weather data. In

this situation, we may need to change the unit of crime data by aggregating districts into cities before using the postcode to merge the datasets.

Outlier Identification

Outliers are values that are markedly different from other members of the sample in which they occur (Barnett & Lewis, 1994). These can be atypical values in a variable (univariate outliers), or atypical combinations of values from different variables (multivariate outliers). Outliers are a problem that is extensively discussed in the psychological literature (e.g. Aguinis et al., 2013; Leys et al. 2019). The most straightforward methods for identifying outliers are well-known to psychologists and are typically used also for smaller datasets: visualization tools (e.g., histograms, box plots, and scatter plots), capping methods (e.g., points that fall outside of the 5th-95th percentile range), and statistical indicators (e.g., points above three standard deviations from the mean value). Thudumu (2020) discusses more advanced techniques that are particularly suited for outlier detection in Big Data, such as distance, clustering, density, and classification based techniques.

Some data mining techniques can tolerate outliers (e.g., decision trees), but with many other techniques, noisy data may deliver unreliable results. In those cases, action should be taken before performing any data analysis through either noise removal (Teng et al., 1990) or noise accommodation (Rousseeuw & Leroy, 1987).

For determining the actions to take, it is important to identify the reason for outlier existence. Some outliers are due to human or instrument errors. Others are valid values, which represent genuine anomalous characteristics or combinations of characteristics. To determine if a measurement is authentic, we can check whether it makes sense in the investigated context (e.g., if we know that the maximum population age is 115 years old, a data point of 130 is plausibly the result of incorrect data entry). If outliers are due to errors, they should be removed. If outliers are valid data points, we might consider conducting the same analysis with and without the outliers and then decide whether to remove or trim them.

For example, if the outliers create an artifact association between two variables, it is advisable to eliminate them.

One advantage of dealing with big datasets is that outliers can be more than odd values causing noise in the analysis. Sometimes their presence may indicate new phenomena of great value, making their identification even more critical. Novelty detection (Markou & Singh, 2003) plays a central role in many applications, including fraud identification (van Capelleveen et al., 2016), illegal parking spotting (Xie et al., 2017), and patient monitoring and alerting (Hauskrecht et al., 2013). Novelty detection represents a great opportunity for psychologists, as it allows them to detect new phenomena. For example, Pan and colleagues (2019) have recently developed a model based on novelty detection to capture anomalous behaviors in crowd video recordings (e.g., prison fights).

Missing Values

Missing data are a common problem in traditional and Big Data research and can negatively impact the conclusions drawn from the data. The probability of missing data increases with the number of variables in the dataset and the sample size. This makes their treatment extremely important in Big Data research (Petrozziello et al., 2018), to preserve the quality of the data and to avoid losing too much information in the preprocessing stage.

Various methods for handling missing data exist. The choice of which method to use should be guided by an understanding of what mechanism has caused a value to be missing. Little and Rubin (2002) have distinguished three mechanisms:

- Missing Completely at Random (MCAR): The probability of missing values depends neither on the variable value nor on the value of other variables for that data point. For example, when data are not recorded or have been lost because of bugs in the recording system.
- Missing at Random (MAR): The probability of missingness depends on the values of other variables. For example, the presence of missing values in income is not linked to the value

of income itself, but to education, if people with higher education are more reluctant to disclose their income.

- Missing Not at Random (MNAR): The probability of missingness depends on the value of the variable with missings. This mechanism is common in situations where people do not want to reveal personal or embarrassing information. For example, people with very high or very low income might be less likely to disclose it.

Albeit this distinction holds true both in small and in large datasets, in the smaller ones it is more difficult to inspect the presence of relations between the missingness and other variables.

Also, the planned analytical strategy is relevant to the choice of how to handle missing data. For example, to use time-series analysis, missing data should be handled with methods that preserve the temporal relations among the observations, such as Recursive Neural Network (Fang & Wang, 2020).

There are various ways to deal with missing data: they can be deleted (listwise or pairwise) or replaced through simple imputation (i.e., a single estimated value for the missing observation is obtained). However, these methods are mostly biased (Emmanuel et al., 2021). To avoid their shortcomings, multiple imputation methods can be used (Rubin, 1987), which create several different plausible imputed datasets and combine their results (Little & Rubin, 2002). There exists a wide variety of multiple imputation methods, such as Multivariate Imputation by Chained Equation, Predictive Mean Matching, and deep learning model (Jäger et al., 2021; Khan & Hoque, 2020). The third way to handle missing values is using analytical techniques robust to their presence, such as decision trees. These techniques require neither elimination nor imputation of missing values.

Both imputation and elimination assume, to some extent, that missing values are errors to get rid of. However, the fact that values are missing may in itself carry valuable information. Therefore, it may be useful to have indicators signaling the missing values.

These missingness indicators might assume the value 0 (i.e., value present) or 1 (i.e., value missing) and could be used as variables in the analyses.

Data Transformation

Big Data must often be adapted to the data mining algorithm. For example, some algorithms work better with normalized variables values, such as artificial neural networks or clustering algorithms. It is also common to transform variables to improve their properties (e.g., normalization) or to reduce the complexity of the information.

Normalization

Often, numerical variables are measured on scales that are entirely different from each other. For some data mining algorithms, differences in the ranges can lead to an undue influence on the results of the variable with the greatest range. For instance, this happens to neural networks and algorithms that use distance measures, such as clustering techniques. To avoid these detrimental effects, data have to be normalized (Han et al., 2012).

Feature Selection

Large datasets often entail a wide range of variables (i.e., features, in data science jargon). As mining on a reduced set of attributes increases the reliability of the results, boosts computational efficiency, and improves the understanding of the results, it becomes crucial to select the most appropriate ones for a given research project. Feature selection aims to find a minimum set of important variables in the dataset that will help answer the research question at hand and discard the redundant and irrelevant ones.

There are two types of strategies to identify relevant variables: domain knowledge-driven and data-driven feature selection. The knowledge-driven approach utilizes researchers' background knowledge (i.e., the relevant theories regarding the behaviors of interest) to establish features relevance, while the data-driven approach relies solely upon the data being analyzed to determine what features to retain. Purely data-driven strategies can result in

model overfitting (Groves, 2013), which is the excessive adaptation of the model to the data and incorporation of noise.

Feature selection methods based on data-driven techniques can be divided into filter, wrapper, and embedded methods (Guyon et al., 2005).

Filter methods select variables based on their statistical link with a certain variable. In the case of predicting whether a given individual will participate in political protests, the psychologist may want to include only those variables in the prediction model that do have a statistical relationship with the behavioral outcome. Thus, the analyst first tests which variables are related to the outcome (e.g., using correlation, chi-square statistics, latent discriminant analysis, information gain, and ANOVA) and keeps only these in the subsequent analyses.

While filter methods explore dyadic relations between variables, wrapper methods (Kohavi & John, 1997) investigate how subsets of variables predict an outcome. The best subset (i.e., the one providing the best predictive performance) is identified through an iterative process of trial-and-comparison between sets of predictors akin to the stepwise regression methods many psychologists are familiar with.

Embedded methods select the variables within the construction of the model in the data mining phase. While the model is constructed using an algorithm, the selection method embedded in that algorithm computes the importance of each variable in predicting the outcome. A well-known embedded method is the Least Absolute Shrinkage and Selection Operator (LASSO) method for regression, which penalizes the regression coefficients, shrinking many of them to zero, and selects all the variables with non-zero coefficients after the penalization.

The wrapper methods are computationally expensive due to their iterative nature, which makes them impractical when the number of features is large. In such situations, filter methods may be a better solution, possibly used for the initial sorting of the variables and

coupled with the embedded methods in the data mining stage for a more fine-grained feature selection.

Regarding their effect on the results, wrapper and embedded methods generally show higher prediction accuracy than filter methods (Li et al., 2017), because with these methods the variables are evaluated based on model accuracy. In addition, as filter methods consider each variable separately, they ignore variable covariances, which may also negatively impact the quality of the results. Finally, since with all of these methods the choice of variables is driven by the relations within the data at hand, the risk of overfitting is always looming in such circumstances, particularly for the wrapping methods (Loughrey & Cunningham, 2004). For a review of feature selection methods, we suggest García et al. (2015).

Data Reduction

Data reduction techniques are applied to reduce the volume of the dataset while keeping most of the integrity of the original data (Han et al., 2012). The aim is to make the data mining process more efficient. There are three types of data reduction methods: dimensionality reduction, cardinality reduction, and sample numerosity reduction methods.

Dimensionality Reduction

Dimensionality reduction aims to overcome the "curse of dimensionality" (Bellman, 1961, p. 94): various problems that arise when analyzing and organizing data with a high number of variables. High dimensionality hinders the efficiency and effectiveness of most data mining algorithms because of their computational complexity. Besides the feature selection techniques discussed above, another way to reduce the number of variables consists in summarizing groups of variables into a more restricted number of indicators. Besides the well-known Principal Component Analysis (Dunteman, 2001) and Factor Analysis techniques (Kim & Mueller, 1978), the analyst can also use Multidimensional Scaling (Kruskal, 1964) and some more recently developed nonlinear alternatives such as Locally Linear Embedding (Roweis & Saul, 2000) and Isometric Feature Mapping (Tenenbaum et al., 2000).

Cardinality Reduction

Data cardinality is the number of distinct values in a variable. With Big Data, we frequently have categorical variables with a large number of categories that often carry morphological or semantic links (Cerda & Varoquaux, 2019). Variables with many distinct values are called high-cardinality variables (e.g., ZIP codes). Reducing the number of categories in a variable decreases the complexity of that information, which is especially helpful in predictive modeling, where many algorithms hardly cope with high cardinality. For example, to reduce the cardinality of the ZIP code variable, we may group values according to the region where the ZIP codes are located, downsizing the number of values from thousands to a few. The easiest way to handle high cardinality is semantic grouping, which aggregates values into higher-ordered categories (e.g., ZIP codes and regions). For further details on cardinality reduction, we recommend the book by Casas (2019).

Sample numerosity reduction

These are a heterogeneous group of methods aimed at easing the analysis of large datasets by replacing the original dataset with a smaller data representation. They include data sampling, data grouping (e.g., condensation, squashing, clustering), and instance selection. For a description of such methods, we refer the reader to Garcia et al. (2015).

Data Resampling

One of the advantages of leveraging Big Data is the chance to predict rare occurrences. This is something that we can do only with Big Data (or at least medium ones) because with small data rare occurrences are too few to be analyzed properly. However, datasets with rare occurrences need special treatment. Imagine being interested in studying the factors that predict the occurrence of a rare disease. As we are dealing with a rarity, in our data the frequency of the minority class (e.g., individuals with the disease) is far lower than the frequency of the majority class (e.g., individuals without the disease). Such an issue, defined as *class imbalance*, can influence many data mining algorithms as they may ignore

the minority class entirely. This is an issue because the minority class is often the class we are most interested in (e.g., mental disorders, extreme IQ values).

Many methods have been developed to solve class imbalance (Weiss, 2009). Most of them eliminate or minimize rarity by altering the original data distribution through resampling, which decreases the overall level of imbalance by decreasing the rarity of the infrequent class. This can be done through oversampling, which increases the number of minority-class instances, and undersampling, which eliminates some of the majority-class instances. Both methods come with risks: Undersampling might remove instances important in model construction. As oversampling methods replicate minority instances, they may increase the likelihood of overfitting by learning data idiosyncrasies that might fail to generalize. The Synthetic Minority Oversampling Technique (Chawla et al., 2002) overcomes this risk by generating fabricated instances from existing ones.

Data Mining

Data mining is an ensemble of approaches, tasks, and techniques that aim to turn large amounts of data into useful information. Data mining can be used for top-down analysis (i.e., verification-driven data mining), but most data mining applications follow a bottom-up approach, where researchers dig into the data in search of new information (Han et al., 2012) using machine learning techniques (Bishop, 2006). Machine learning is a branch of artificial intelligence and computer science that focuses on the use of data and algorithms to imitate the way humans learn, gradually improving its accuracy.

This section describes the options available, their characteristics, and some methodological issues that psychologists approaching data mining should consider.

Data Mining Tasks

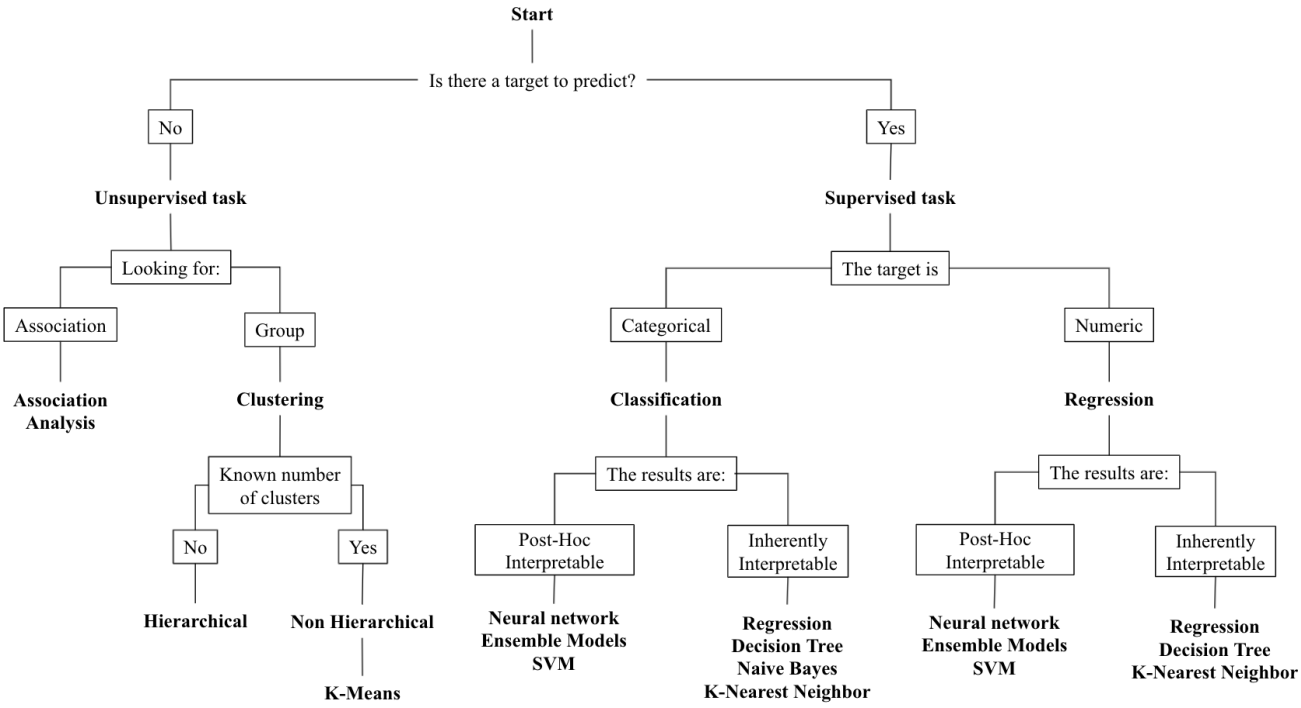
When deciding which type of algorithm to use, it is important to consider the research question, the nature of the data, and the resources available: Is there an outcome that we want to predict? What are the characteristics of the variables? What are the computational

resources and technical skills available? To help make a choice, we propose the decision-making algorithm in Figure 2.

An important distinction is between supervised and unsupervised learning. This is based on the presence in the dataset of a variable indicating an aspect of the data points that the model must predict (e.g., predicting individuals' happiness; Maimon & Rokach, 2010). Supervised learning algorithms need such a variable and are trained to predict it. Unsupervised algorithms, instead, group variables or observations based on their similarity or covariation (e.g., grouping tweets based on their contents).

Figure 2.

A possible decision-making process for choosing the algorithm.



Various unsupervised learning techniques are commonly used in psychological research, such as Principal Components Analysis, and cluster analysis. Another, less known, but equally interesting method is Association Rules. Unsupervised learning algorithms have been leveraged in psychological research, for example, to examine COVID-19–related discussions, concerns, and sentiments expressed in tweets (Jia et al., 2020) and to identify

discussion topics of fathers on fatherhood from social media data (Teague & Shatte, 2018). As many of the unsupervised algorithms (e.g., k-means, hierarchical clustering) are well-known among psychologists, we believe that discussing supervised algorithms may be of greater interest to the reader. Thus, we will not further discuss unsupervised learning techniques and refer to Han et al. (2012) for a detailed treatment.

Supervised learning techniques are called regression and classification, depending on the type of outcome variable. If the outcome is categorical, the supervised task is called classification (binary or multinomial depending on the number of distinct values of the outcome). Classification models have been used in psychological research for predicting gender, ethnicity, and sexual orientation from Facebook likes (Kosinski, et al., 2013). Dabek and Caban (2015) used a classification model to determine the probability of developing psychological conditions following a concussion. If the outcome is numerical, the task is called regression. This name should not be confused with the term used to indicate the homonymous statistical technique. Regression as a machine learning task indicates the prediction of continuous variables, independent of the specific statistical technique used for this endeavor. In the realm of Big Data, regression models have been used, for instance,, to predict psychological traits, such as materialism and self-control, from spending records (Gladstone et al., 2019), to evaluate the detrimental effect of exposure to a turbulent stock market on people's mental health (Qin et al., 2019), to predict self-assessments of the Big Five personality traits from smartphone sensing data (Stachl et al., 2019).

For further details, we recommend Rosenbusch and colleagues (2021) and Bramer (2020).

Supervised Learning Algorithms

Many algorithms perform supervised tasks. This section describes the most well-known (see also Table 1 for an overview).

Table 1.

List of the characteristics of the most used supervised learning algorithms

	KNN	Naive Bayes	Regression	Decision Tree	Ensemble	Neural Networks	SVM
Parametric or Non-Parametric	Non-Parametric	Parametric	Parametric	Non-Parametric	Non-Parametric	Non-Parametric	Non-Parametric
Assumptions on data	None	Attribute values independence	Many assumptions	None	None	None	None
Task Type	Classification and Regression	Classification	Classification and Regression	Classification and Regression	Classification and Regression	Classification and Regression	Classification and Regression
Ease of implementation	Easy	Easy	Easy	Average	Difficult	Difficult	Difficult
Interpretability	Inherent	Inherent	Inherent	Inherent	Post-hoc	Post-hoc	Post-hoc
Variable type	It performs better on numerical than categorical variables	It performs better on categorical than numerical variables	It works well on both categorical and numerical variables	It works well on both categorical and numerical variables	It works well on both categorical and numerical variables	It works well on both categorical and numerical variables	It works well on both categorical and numerical variables
Normalization	Required	Not needed	Helps to interpret and compare the coefficients	Not needed	Not needed	Helps the convergence of training	Helps in solving linear equations
Sensitivity to outliers	High	High	High	Low	High	Low	Low
Missing data handling capability	It needs complete data	It needs complete data	It needs complete data	Yes (e.g., surrogate splits)	It needs complete data	It needs complete data	It needs complete data

Regression. This consists of a set of methods that allow predicting the value of a numerical or categorical outcome based on one or several predictors. Thus, they can perform both regression and classification tasks. For example, Schwartz and colleagues (2014) built a regression model to predict Facebook users' depression. Combining 28,749 users' Facebook status with survey data, they found that users' mood worsened in the transition from summer to winter. Regression techniques are widely used because they are easy to implement, efficient to train, do not require many computational resources, and offer good baselines for judging the quality and debugging more sophisticated approaches (Kosinski et al., 2016). Their results are inherently interpretable because the estimated coefficients directly reflect variable contributions; hence, these models can be explained through these coefficients. Two

major shortcomings of regression are its reliance on various assumptions about the variables and their relationship (which, in some cases, can be circumvented by using different forms of regression like nonlinear regression techniques, such as polynomial or spline regression) and its difficulty in dealing with many predictors (which can be overcome through regularization techniques, like the LASSO method mentioned in the Feature Selection paragraph, that help avoid overfitting).

K-Nearest Neighbor (KNN). KNN is a nonparametric technique used for classification and regression. For a given data point, KNN first finds the *K* elements of the dataset (where *K* is a parameter chosen by the analyst) that are most similar to the data point. Similarity is conceptualized as proximity and operationalized with one of the various existing distance measures (e.g., Manhattan, Euclidean, or another Minkowski distance). Next, if classification is the goal, the algorithm assigns to a data point the value of the categorical outcome that is most common among its *K* most proximal neighbors (i.e., the mode). If regression is the goal, the algorithm assigns to a data point the average value of the *K* nearest data points on this variable. KNN may suffer from several disadvantages when tackling big datasets, such as high computational cost and sensitivity to noise. However, recent developments permit using the algorithm with Big Data (e.g., Flink Machine Learning KNN; Chatzigeorgakidis et al., 2018).

Naïve Bayes. This is a probabilistic algorithm based on the Bayes Theorem used for classification. Naïve Bayes models are easy and fast to build, with no complicated iterative parameter estimation (Bramer, 2020). They are inherently interpretable because each predictor is associated with an estimated posterior probability, which provides information on its contribution to the classification. In psychological research, Naïve Bayes has been used for classifying infants into high or low-risk groups for Autism Spectrum Disorder (Bosl et al., 2011) and to classify social media posts into engagement levels (Hwong et al., 2017). Naïve Bayes algorithms rely on the assumption of independence between the predictors. When this

assumption holds, a Naive Bayes classifier can perform better than other models like logistic regression. However, this assumption is nearly always violated, and, as a consequence, it hinders the predictive performance of the classifier. A further problem with Naive Bayes is that when a predictor in the test dataset has a value that is not present in the training dataset, the model is unable to make predictions. To overcome this issue, analysts can use sample smoothing techniques, such as Laplacian Correction (Manning et al., 2008).

Decision Tree methods. These are a set of nonparametric methods that can perform both classification and regression tasks. A decision tree predicts the outcome of interest, based on a set of if-then rules that are associated with the values of the predictors. They have been used, for instance, by Lin and colleagues (2020) for identifying patients at risk for suicide, and by Song and Song (2021) for detecting risk factors of cyberbullying. Decision trees come with a series of positive characteristics: they closely resemble human reasoning, their mechanism is easy to understand, and they are inherently interpretable (Kotsiantis, 2013). They do not rely on specific assumptions regarding the distribution of the predictors or their relations (Rokach & Maimon, 2005), and are robust to data transformation and outliers (Breiman et al., 1984; Steinberg, 2009). They handle missing values (without requiring imputation) and heavily skewed data (Song & Lu, 2015). They have also some disadvantages. Firstly, decision tree models tend to overfit, and therefore they may not generalize to new data. Secondly, they may show high variance (i.e., they can get unstable due to small variations in data). Thirdly, building decision trees, especially large ones with many branches, may be complex and time-consuming. Finally, too large trees may be difficult to interpret and pose visualization difficulties.

Ensemble methods. These methods initially create multiple models, and next combine them to produce an enhanced solution. For instance, in Random Forests (Breiman, 2001), several decision trees are constructed, based on randomly selected observations from the training datasets and random subsets of predictors. To make a new prediction, each of

these trees is used separately and makes its predictions. Then, the predictions of the different trees are aggregated into the final prediction of the Random Forest. Other ensemble methods exist for different machine learning algorithms (Bonaccorso, 2018; Bramer, 2020).

Ensemble methods were used, for example, by Gladstone et al. (2019) for predicting Big Five personality traits, materialism, and self-control from online spending transactions. Usually, ensemble models (e.g., Random Forest) provide better predictions and higher generalization than single models (e.g., decision tree; Dietterich, 2000). However, the results of an ensemble model are not directly interpretable. Thus, to understand what relationship links a predictor to an outcome, post-hoc methods for interpreting the results are necessary (see the Models Interpretability section). Random Forest is one of the best performing ensemble techniques available (Fernández-Delgado et al., 2014) as it is very robust and can limit overfitting (Breiman, 2001).

Artificial Neural Networks. Also called Neural Networks, these are a set of algorithms used both for classification and regression. A Neural Network is based on a collection of interconnected computational units, called nodes. Some of these nodes receive input from the dataset, some receive it from other nodes. Based on their input, the nodes make computations, and feed their results to other nodes or provide them as outcomes of the elaboration. Neural networks can model non-linear relationships and provide reliable results also with large amounts of variables. Neural networks have been used in various psychological studies, such as forecasting individuals' mood (Mikelsons et al., 2017; Taylor et al., 2017), classifying social interactions (Vieira et al., 2017), and predicting sexual orientation from facial images (Wang & Kosinski 2018).

In their simplest form, Neural Networks have three layers of nodes (also called Shallow Neural Networks). Neural Networks with more than three layers of nodes are defined as Deep Learning Networks. The biggest advantage of Deep Learning Networks over other learning algorithms is that they can preprocess and transform raw variables, thus, data do not

need prior preprocessing. However, differently from other techniques that can be applied also to smaller datasets, Deep Learning Networks need very large datasets (e.g., millions of data points). Also, even though using more layers is associated with higher model flexibility and accuracy, it also increases the risk of overfitting and challenges model interpretability. One of the most popular Deep Learning algorithms is the Convolutional Neural Network, commonly applied to analyze visual data (e.g., facial recognition, image search, natural language processing). For more information on Deep Learning Networks, we recommend Urban and Gates (2021).

Like ensemble methods, neural networks can be interpreted through post-hoc techniques (see Models Interpretability section). In addition, much effort has been recently put into the development of inherently interpretable neural networks, such as the Explainable Deep Neural Network (Angelov & Soares, 2020).

Support Vector Machine (SVM). This machine learning algorithm can perform both regression and classification tasks, but it is more often used for classification. SVMs have been applied, to provide a few examples, in the classification of emotional states from real-life spoken human-human interactions (Devillers et al., 2005) and the classification of personality traits based on eyes' iris position (Ramli & Nordin, 2018). An SVM model is a representation of the instances of a dataset as points in an n-dimensional space, mapped so that the instances of separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall. In their basic form, SVMs are linear classifiers, meaning that the classification is based on the value of a linear combination of the predictors. However, they can also construct nonlinear models. Like ensemble methods and neural networks, SVMs results can be interpreted through post-hoc methods (see Models Interpretability section).

Performance Measures

Evaluating the performance of a predictive model is fundamental for judging models quality, refining choices in the KDD iterative process (e.g., feature selection), and selecting the most accurate model from a given set of models built with different algorithms (Maimon & Rokach, 2010). Model ability to correctly predict the criterion of interest with new instances can be evaluated through criteria called performance measures. Choosing the performance measure to use is an important task because different measures tell different parts of the story and, thus, can affect the conclusions we can draw from the results (Kuhn & Johnson, 2019). This choice depends on the type of supervised learning task to be performed (i.e., classification vs. regression).

Classification algorithms can produce two types of outcomes: classes (e.g., binary classification output will be either "0" or "1") and probabilities (i.e., the output is the probability that a specific case belongs to each of the possible categories). Different algorithms produce either the former or the latter. For example, SVM and KNN return the predicted class. Logistic Regression, Random Forest, and Naïve Bayes provide probability outputs, which can be converted into class output by setting a threshold probability (e.g., instances with a probability below 0.5 can be assigned the "0" class, and those above 0.5 to the "1" class). The most straightforward instrument to assess the performance of a classification model is the confusion matrix (Kohavi & Provost, 1998), known as classification table by psychologists familiar with logistic regression. This is a table with as many rows as the categories observed and as many columns as the categories predicted. The cells of the matrix summarize the frequencies of correct and incorrect predictions. These frequencies are useful for computing various performance measures, such as accuracy, specificity, precision, recall, and F1.

A visual way to display the performance of a binary classifier is the Receiver Operating Characteristic (ROC; Fawcett, 2006) curve. The ROC plot is a popular and intuitive tool. It displays on the x-axis the 1 – specificity metric and on the y-axis the recall

metric at all threshold levels. The combination of such metrics produces a line that represents the performance. A classification model with random performance shows a straight line from the origin to the top right corner of the ROC space. A classification model with perfect performance shows a combination of two straight lines: From the origin to the top left corner and further to the top right corner. From the ROC plot, we can obtain the index known as Area Under ROC Curve (AUC), which is the portion of the area under the ROC performance line. AUC values range from 0.5 (random predictive model) and 1 (perfect predictive model). Compared to performance measures, such as accuracy, where the analyst must set a probability threshold for determining the predicted class, AUC provides an aggregate measure of performance across all possible probability thresholds.

As concerns regression algorithms (i.e., those that predict a continuous outcome), widely known performance measures are the Root Mean Squared Error (RMSE), Root Mean Squared Logarithmic Error (RMSLE) and Mean Absolute Error (MAE). The value of these metrics is usually interpreted as either the average distance of the residuals from zero or the average distance between the observed and predicted values (Kuhn & Johnson, 2013). Another commonly used metric is R Squared (R^2 or coefficient of determination), which indicates the proportion of information that is explained by the model. Even though it is easily interpretable, R^2 is dependent on the variation in the outcome variable: if we validate a model on new data where the outcome variable shows less (or more) variance than the outcome used to build that model, the R^2 performance will be affected by this variance change.

Validation and Interpretation

In the following section we discuss two key aspects of model construction, which are model validation (i.e., the process of confirming that the model reliably achieves its intended purpose, e.g., predicting a behavior) and model interpretability (i.e., how easy it is for researchers to understand how the output of the algorithm is related to the values of the input variables).

Model Validation

To check if the predictive model will be able to make accurate predictions with similar but new instances, it must undergo a validation process. The validation process prevents the risk of using a model that has overfitted or underfitted the data. Overfitting happens when the model learns idiosyncratic relationships that do not hold when applied to data that are similar, but not identical to that used to construct it (Hastie et al., 2009). Underfitting occurs when a model does not account for the true complexity (e.g., nonlinear effects and interactions) in the data and, therefore, cannot grasp the systematic variance. Overfitting and underfitting hamper predictive performance.

Validation consists of a set of techniques for subdividing the dataset into two complementary subsets (Han et al., 2012; Kohavi, 1995): The first is used to build the model and the second to test its performance and cannot be used in any way to build the model. There are various common ways to perform validation that we will now describe.

Holdout Validation

This strategy concerns the subsetting of the dataset in two parts: the training set, used to build the model, and the test set, used to prove its predictive ability. There is no definitive rule on the best proportion for this division, but it is advisable to build the model on 70%-80% of the cases and test it on the remaining portion since the more data the algorithm uses to build the model, the better the model can learn. To split the data into training and testing sets, it is recommended to use random sampling (e.g., completely or stratified random sampling). Some algorithms involve two training stages (e.g., neural networks), one to come up with the basic structure of the model and the second to optimize the parameters. In such cases, three datasets are involved: The training set to build the basic structure, the validation set to optimize the parameters, and the test set to evaluate the performance of the optimized model.

K-Fold Cross-Validation

Cross-validation extends the idea of holdout validation by repeating the train-test splitting process several times. With k-fold validation, the original dataset is divided randomly into k parts, named folds. In general, k-fold cross-validation is performed by taking one fold as the test set and the remaining k-1 folds as the training set. The validation process is repeated with each fold as the test data, and all predictions are averaged to obtain the overall model predictive performance. Another type of cross-validation is called leave-one-out cross-validation. This is k-fold cross-validation taken to its logical extreme where k is equal to N-1, where N is the number of cases in the dataset. Thus, the model is trained on all the instances except one, and the left instance is used as a test set. The results of all N predictions, one for each instance of the dataset, are averaged, and that average represents the predictive performance. This strategy has the advantage that the highest amount of data is used for training the model, and no random sampling is involved. However, using this method is time-consuming, especially for large datasets.

Holdout vs. Cross-Validation: a Comparison

Compared to cross-validation, the holdout method can be more effective and computationally inexpensive, therefore it is particularly useful with very large datasets and in case of limited computational resources. This is because model validation is performed only once, whereas, with cross-validation, the splitting process is repeated several times.

On the other hand, the holdout method produces less accurate models than cross-validation (Tantithamthavorn et al., 2017). As the holdout validation splits data into train and test sets, part of the data is not used during model construction. However, the test set could contain important observations that, if disregarded during model construction, may have detrimental effects on model performance. By using all observations for model construction, cross-validation produces more accurate models.

Models Interpretability

Machine learning algorithms can be distinguished based on their interpretability, which is the degree to which a human can understand the cause of a decision (Miller, 2017). We can distinguish two classes of interpretable machine learning models depending on the time when the interpretability is obtained (Molnar, 2019): inherently interpretable (e.g., regression and decision tree) and post hoc interpretable models (e.g., ensemble models and neural networks). Intrinsic interpretability characterizes those models that incorporate interpretability directly into their structures (e.g., the beta coefficients of a regression model). Post hoc interpretability is achieved by using methods that indicate what predictors have the biggest impact on predictions. There are different methods for interpreting the results of a model. Permutation importance tells the importance of a specific predictor to the overall predictive performance of a model by calculating how the model performance deviates after altering (i.e., permuting) the values of that predictor (e.g., Partial Dependence Plot and Shapley Additive Explanation). For a comprehensive review of such methods, we suggest Molnar (2019).

Several reasons drive the demand for interpretability (Doshi-Velez & Kim, 2017). In scientific fields, like psychology, which aim to gain knowledge of why something happened, the model itself becomes the source of knowledge, as its results can suggest new research questions in exploratory studies, and provide a test for the hypotheses when the research has a confirmatory quality. If the goal is prediction per se (e.g., the probability that an intervention is effective, or the probability that a client accepts a commercial offer), the model's interpretability may not be an issue. However, to know why predictions have been made (e.g., why there is a relationship between a predictor and intervention effectiveness, why clients with certain characteristics are more prone to accept the commercial offer), an interpretable model is needed. Interpretability becomes important in relation to the consequences of its predictions. Some models may not require explanations because they are used in low-risk environments, where mistakes will not have serious consequences (e.g., movie recommender

systems). In high-risk environments (e.g., disease diagnosis), instead, understanding why a prediction is made is often as important as achieving high predictive accuracy. The more a model's prediction affects people's lives, the more critical it becomes to explain the model's results. For example, imagine a doctor of an intensive care unit who wants to predict the likelihood of a patient's readmission or mortality for deciding about discharge. A predictive model helps the doctor pick the right moment to move the patient from the unit. If the doctor understands what the model is doing, the doctor is more likely to trust the model's predictions in making the final decision.

Finally, an important reason why interpretability is critical is based on the premise that environments change. If we do not understand how a model works, we will not be able to predict if it will stop making accurate predictions. A relevant example was provided by the failure of the Google Flu Trends algorithm, which was based on the terms used by users in the Google search engine and provided near real-time estimates of flu spread in the USA. When it was released in 2008, everyone was amazed by the accuracy of Google Flu Trends predictions and by the fact that those predictions were made without any scrutiny on how the model made them, for instance, which queries were predictive of influenza. However, for the 2012-2013 flu seasons, Google Flu Trends failed and predicted more than double the proportion of doctor visits for flu-like symptoms compared to those registered by the Centers for Disease Control and Prevention (Butler, 2013). A reason for the failure may be traced back to the fact that the environment in which Google Flu was first applied changed (Lazer et al., 2014). Thus, it is possible that not knowing how the algorithm worked did not allow practitioners to anticipate the difficulties of the model in a new environment, and hence its debacle.

R and Python

To conduct Big Data projects, it is necessary to know a programming language, because data are often too large and complex to be processed and analyzed with the usual

statistical tools, such as SPSS and Excel. Many programming languages exist (for an overview, see Ronin, 2019), but the most used and recommended are R and Python (Kaggle Survey, 2018).

Nowadays, R is the most widely used within psychology research. Thus, psychologists approaching Big Data research may find it useful to use R because they are already familiar with it, or they intend to easily discuss their analyses within the psychologists' community. However, besides being the most widely used alternative to R in Big Data research, some Python characteristics make it more suitable when dealing with some Big Data tasks, as discussed later in this section.

R (R Development Core Team, 2021) is an open-source programming language and statistical environment. R is rooted in statistics, data analysis, data exploration, and data visualization. There are more than 15000 packages on the Comprehensive R Archive Network (CRAN), which is a network of web servers around the world where users can find the source code, manuals, and documentation on those packages.

Python (von Rossum & Drake, 1995) is an open-source programming language rooted in computer science and mathematics. It was designed to be easy to read and cover multiple programming paradigms and applications. The Python library is called PyPI and has more than 113000 packages, which makes it the largest ecosystem among programming languages. Both R and Python can be implemented in integrated development environments (IDE) that facilitate application development and offer a central interface featuring all the tools a developer needs, such as code editor, compiler, and debugger. There are many IDEs that can support R (e.g., RStudio or StatET) and Python (e.g., Ipython Notebook IDE or Spyder). As they are open-source projects, updates are released frequently. While the core developers control the primary structure, many users around the world contribute with new features (e.g., packages) and bug fixes.

In the next sections, we describe and compare the two programming languages and introduce some resources that may be helpful when learning to use them.

Similarities and Dissimilarities

Analytical Functionalities

Both R and Python run on almost any standard computing platform and operating system. There are many R packages suitable for acquiring data (e.g., scraping tweets through *twitteR* package; Gentry, 2015), preprocessing, and building machine learning models. It can communicate with other languages (e.g., Python and C++), and can be used in common Big Data analytics engines like Apache Spark (Zaharia et al., 2016) or Hadoop (Apache Software Foundation, 2010). Similar to R, Python's versatility includes database connectivity, cross-language communication, data scraping, machine learning modeling, text, and image processing. The methods and techniques that have been described in the previous sections can be implemented in both languages. Compared to R, Python has a much more extensive suite of specialized advanced machine learning, Deep Learning, and Artificial Intelligence libraries, like scikit-learn and TensorFlow, which enable researchers to develop sophisticated data models.

Computational Speed

As for computational speed, even though both programming languages are capable of handling Big Data operations, Python is on average much faster when performing the same task than R (Korstanje, 2020). However, to overcome R speed limits, it may be advisable to use more efficient languages both in terms of implementation and execution time, such as C, C++, or Fortran. Researchers who may have access to some really fast and efficient Fortran code, can execute it within R. As these programming languages are hugely faster than both R and Python, they might decrease the speed advantage of Python over R.

Data Visualization and Reporting Capabilities

Compared to Python, R has more sophisticated graphics capabilities, as its graphics system allows for good control over every aspect of a plot or graph. R graphics packages allow for complex and sophisticated visualizations of high-dimensional data.

When it comes to communicating the findings, R truly stands out compared to Python. R has a fantastic range of tools that allows sharing the results in the form of a presentation or a document, such as *shiny* (Chang et al., 2021; a tool for building prototype web applications) and *rmarkdown* (Allaire, et al., 2021; a method for integrating code, graphical output, and text into a journal-quality report).

Advice for Starting Your First Project

Knowing how to program can be an important skill that can affect career advancement, both in academia and in the business world (Sijbrandij, 2017). The best kind of learning is learning by doing. One of the first things a beginner should do is to install the language environment and its IDE on the computer (e.g., R Studio or Jupyter) from the respective websites. Next, get acquainted with the many online and offline resources available.

Learning Tools

Various platforms provide learning paths (e.g., Codecademy) and online courses (e.g., Coursera). Most platforms supply basic content for free, requiring a fee for accessing advanced content (e.g., DataQuest). Table 2 offers an overview of some of the available online courses.

Table 2.

Online courses on programming languages.

Name	Pay or Free	Link to course
Codecademy	Free	www.codecademy.com/catalog/language/python www.codecademy.com/learn/learn-r
DataCamp	Pay (basic contents for free)	www.datacamp.com/courses/intro-to-python-for-data-science www.datacamp.com/courses/free-introduction-to-r
Coursera	Free auditory only	www.coursera.org/specializations/python www.coursera.org/learn/r-programming
DataQuest	Pay (basic contents for free)	www.dataquest.io/course/python-for-data-science-fundamentals/ www.dataquest.io/path/data-analyst-r/

EdX	Free auditory only	www.edx.org/course/introduction-to-python-fundamentals-5 www.edx.org/course/introduction-to-r-for-data-science-5
Udemy	Pay (basic contents for free)	www.udemy.com/course/complete-python-bootcamp/ www.udemy.com/course/r-programming-for-statistics-and-data-science/
Real Python (Python only)	Pay (basic contents for free)	realpython.com/courses/
CodeAvengers (Python only)	Pay	
Udacity	Pay (basic contents for free)	www.udacity.com/course/introduction-to-python--ud1110 www.udacity.com/course/programming-for-data-science-nanodegree-with-R--nd118
Lynda	Pay	www.lynda.com/search?q=python www.lynda.com/search?q=r
Analytics Vidhya (Python only)	Pay (some courses are free)	courses.analyticsvidhya.com/courses/introduction-to-data-science

Books are also essential. There are many great books and, unfortunately, not enough time to read them all. In Table 3, we have enlisted good reads, some of which are available online for free. Some provide an introduction on how to use languages from the beginning (e.g., Grolemund, 2014; Shaw, 2017), whereas others deal with programming languages applied to the Big Data methods and techniques discussed in this paper (e.g., McKinney, 2017; Wickham & Grolemund, 2017).

Table 3.

Reads on programming languages and main themes.

Authors	**Basics of language**	**Data acquisition**	**Data preprocessing**	**Data transformation**	**Data mining**	**Model validation**
Python						
Shaw, 2017	x					
Swaroop, 2015	x					
Mckinney, 2017	x		x	x		
Kazil & Jarmul, 2016	x	x	x	x		
Müller & Guido, 2016				x	x	x
Raschka & Mirjalili, 2019					x	x
Chollet, 2021					x	
R						
Grolermund, 2014	x					
Matloff, 2011	x					
Boehmke, 2016	x	x	x	x		
Zumel et al., 2019	x		x	x	x	x
Shu, 2020			x		x	x

James, et al., 2021	x	x
Ghatak, 2019	x	

Finally, if a particular concept does not make sense or a code continuously ends with an error, it would be useful and, hopefully, decisive to look for alternative online resources to disentangle that content. The online resources to learn computer programming are endless, and they are provided by many online tutorials (e.g., Data Flair or Learn Python) as well as online communities (e.g., Stack Overflow, KDNuggets, RBloggers,).

Conclusion

Big Data are promising for psychology because they allow to conduct a quantitative form of naturalistic field research. However, only a small proportion of psychologists are currently engaging in Big Data projects. Paxton and Griffiths (2017) identified three main reasons why: the "skills gap," the "imagination gap," and the "cultural gap." To contribute directly to bridge the "skills gap," we described the KDD process and outlined a broad set of concepts, techniques, and tools that are relevant to psychologists who want to conduct Big Data projects. Throughout the paper, we have also discussed methodological issues and highlighted related pitfalls that need to be considered when applying Big Data techniques. In addition, since the use of Big Data in psychological research is expected to increase and as collaborations with data scientists may become necessary, psychologists interested in Big Data research must familiarize themselves with the concepts and methodology of the KDD steps. As this overview shows, many concepts and techniques that we can use for the treatment of Big Data are not completely new to psychologists and psychometrists. Sometimes it is simply a matter of learning a new jargon. To mention just a few examples, think about supervised and unsupervised learning, which might be described as techniques for forecasting and techniques to summarize the data; or about the term "features" which refers to what we usually call variables or predictors. Therefore, one initial step for bringing psychologists closer to the study of Big Data is to make them aware that they already know

some of the analytical tools that they could use. Of course, there are statistical techniques that are known, but not much used in the psychological field, like prediction trees, which may be better suited for big datasets than for the ones typically used by psychologists. And, then, there are methods and techniques that are new to the psychological field, such as the Deep Learning Networks, which are outside traditional psychological research.

Besides contributing to the bridging of the "skill gap," by providing examples of applications of psychological Big Data research throughout the paper, this work contributes also to bridging the "imagination gap." As concerns the "cultural gap", or in other words the difficulty in getting the Big Data perspective adopted by individuals and institutions within psychology, by giving editorial space to contributions like ours (see also the special issues dedicated to the topic published by Psychological Methods and other outlets), psychological journals are providing important role models that might bring psychologists closer to Big Data research.

We believe that Big Data concepts and techniques, such as data reduction, out-of-sample validation, machine learning techniques, and methods of interpretable machine learning (e.g., permutation) will contribute to the generalizability and robustness of psychological studies. Indeed, several elements of robust science are common features of Big Data methods (Grand et al., 2018). Most psychological studies suffer from small underpowered samples, and publication pressures incentivize Questionable Research Practices (QRP's) like p-hacking (Albritton & Tonidandel, 2020). Big data methods, in contrast, tend to have large samples making statistical significance testing superfluous, thereby eliminating the need for p-hacking. Indeed, conclusions based on statistical inference can be misleading, because, with very large datasets, even minuscule effects can become statistically significant. With Big Data the question is no longer whether results are "significant" (in large samples, they nearly always are), but whether they are relevant and interesting (i.e, practically significant).

Despite the many advantages, it is also important to remark that Big Data research is not inherently more robust than traditional research. As in traditional research, we cannot expect to obtain robust results if they are based on flawed premises (i.e., garbage in, garbage out). Moreover, QRPs may also occur in Big Data research. For example, a researcher might apply one or more unsupervised learning techniques to identify groupings in a dataset, leverage various supervised learning techniques to identify potential predictors of cluster membership, and then produce an interpretation for notable relationships while ignoring those that appear less promising. Finally, they subsequently develop and reports a conceptual narrative that fits the results. This process closely resembles the practice of HARKing (i.e., hypothesizing after the results are known). Thus, in many respects, recommendations for limiting the impact and prevalence of QRPs in Big Data research are similar to those proposed for improving the robustness of more traditional psychological research and its replicability (e.g., improving the transparency, preregistration and alternative publication mechanisms).

This last observation brings us to spend some words on the relationship between Big Data research, open science practices, and the reproducibility crisis. As illustrated in the previous section, the practices promoted by the Open Science Movement can have a strong beneficial effect on the robustness of Big Data research. It should also be noted that Big Data techniques may facilitate adhering to open science practices, and hence contribute to resolving the reproducibility crisis that is affecting the field (Open Science Collaboration, 2015; Ioannidis, 2005). Thus, we believe that Big Data and open science practices can provide mutual benefit in contrasting the reproducibility crisis that affects both fields.

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*(7), 899–917. https://doi.org/10.1037/amp0000190

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers : *Https://Doi.Org/10.1177/1094428112470848*, *16*(2), 270–301. https://doi.org/10.1177/1094428112470848

Albritton, B. H., & Tonidandel, S. (2020). How Can Big Data Science Transform the Psychological Sciences? *The Spanish Journal of Psychology*, *23*, 1–5. https://doi.org/10.1017/SJP.2020.45

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2021). *rmarkdown: Dynamic Documents for R*. https://rmarkdown.rstudio.com

Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Networks*, *130*, 185–194. https://doi.org/10.1016/J.NEUNET.2020.07.010

Apache Software Foundation. (2010). *Hadoop*.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data* (3rd ed.). Wiley & Sons.

Beaton, D., Dunlop, J., & Abdi, H. (2016). Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. *Psychological Methods*, *21*(4), 621–651. https://doi.org/10.1037/met0000053

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. In *Journal of Electronic Imaging*. Springer. https://doi.org/10.1017/CBO9781107415324.004

Boehmke, B. C. (2016). *Data Wrangling with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-45599-0

Bonaccorso, G. (2018). *Machine Learning Algorithms : Popular Algorithms for Data Science and Machine Learning, 2nd Edition.* Packt Publishing Ltd.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature 2012 489:7415*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

Bosl, W., Tierney, A., Tager-Flusberg, H., & Nelson, C. (2011). EEG complexity as a biomarker for autism spectrum disorder risk. *BMC Medicine*, *9*(1), 18. https://doi.org/10.1186/1741-7015-9-18

Bramer, M. (2020). *Principles of Data Mining*. Springer London. https://doi.org/10.1007/978-1-4471-7493-6

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random forests. *Machine Learning*. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Stone, C. J., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. In *Chapman & Hall/CRC Texts in Statistical Science Series*. Chapman & Hall/Crc. https://doi.org/10.1002/widm.8

Butler, D. (2013). When Google got flu wrong. *Nature*, *494*(7436), 155. https://doi.org/10.1038/494155a

Casas, P. (2019). *Data science live book*. https://livebook.datascienceheroes.com

Cerda, P., & Varoquaux, G. (2019). Encoding high-cardinality string categorical variables. In *arXiv*. https://doi.org/10.1109/tkde.2020.2992529

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *shiny: Web Application Framework for R*.

Chatzigeorgakidis, G., Karagiorgou, S., Athanasiou, S., & Skiadopoulos, S. (2018). FML-kNN: scalable machine learning on Big Data using k-nearest neighbor joins. *Journal of Big Data 2018 5:1*, *5*(1), 1–27. https://doi.org/10.1186/S40537-018-0115-X

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, E. E., & Wojcik, S. P. (2016). A Practical Guide to Big Data Research in Psychology. *Psychological Methods*, *21*(4), 458–474. https://doi.org/10.1037/met0000111

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Chollet, F. (2021). *Deep Learning With Python*. Manning Publications. https://www.amazon.it/Deep-Learning-Python-Franc¸ois-Chollet/dp/1617296864

Chon, J., & Cha, H. (2011). LifeMap: A smartphone-based context provider for location-based services. *IEEE Pervasive Computing*, *10*(2), 58–67. https://doi.org/10.1109/MPRV.2011.13

Chow, P. I., Fua, K., Huang, Y., Bonelli, W., Xiong, H., Barnes, L. E., & Teachman, B. A. (2017). Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. *Journal of Medical Internet Research*, *19*(3), e62. https://doi.org/10.2196/jmir.6820

Dabek, F., & Caban, J. J. (2015). Leveraging Big Data to Model the Likelihood of Developing Psychological Conditions After a Concussion. *Procedia Computer Science*, *53*(1), 265–273. https://doi.org/10.1016/J.PROCS.2015.07.303

Davies, B. (2021). *Is Web Scraping Legal in 2020?* ScrapeDiary. https://scrapediary.com/is-web-scraping-legal/

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, *18*(4), 407–422. https://doi.org/10.1016/j.neunet.2005.03.007

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1857 LNCS*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *American Statistician*. https://doi.org/10.1080/00031305.1982.10483055

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. http://arxiv.org/abs/1702.08608

Dunteman, G. H. (2001). *Principal Components Analysis*. Sage. https://books.google.it/books?id=447RngEACAAJ

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(140), 1–37. https://doi.org/10.1186/s40537-021-00516-9

Fang, C., & Wang, C. (2020). *Time Series Data Imputation: A Survey on Deep Learning Approaches*. https://arxiv.org/abs/2011.11347v1

Favaretto, M., Clercq, E. De, Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLOS ONE*, *15*(2), e0228987. https://doi.org/10.1371/JOURNAL.PONE.0228987

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/J.PATREC.2005.10.010

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *Knowledge Discovery and Data Mining: Towards a Unifying Framework*. www.aaai.org

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., & Fernández-Delgado, A. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, *15*, 3133–3181. http://www.mathworks.es/products/neural-network.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/JCSS.1997.1504

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. In *The Annals of Statistics* (Vol. 29, pp. 1189–1232). Institute of Mathematical Statistics. https://doi.org/10.2307/2699986

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining*.

Gentry, J. (2015). *twitteR: R Based Twitter Client*.

Ghatak, A. (2019). Deep learning with R. *Deep Learning with R*, 1–245. https://doi.org/10.1007/978-981-13-5850-0

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can Psychological Traits Be Inferred From Spending? Evidence From Transaction Data. *Psychological Science*, *30*(7), 1087–1096. https://doi.org/10.1177/0956797619849435

Goldstone, R. L., & Lupyan, G. (2016). Discovering Psychological Principles by Mining Naturally Occurring Data Sets. In *Topics in Cognitive Science* (Vol. 8, Issue 3, pp. 548–568). https://doi.org/10.1111/tops.12212

Gollwitzer, A., Martel, C., Brady, W., Pärnamets, P., Freedman, I., Knowles, E., & Van Bavel, J. J. (2020). *Partisan Differences in Physical Distancing Predict Infections and Mortality During the Coronavirus Pandemic*. https://doi.org/10.31234/osf.io/t3yxa

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*(1), 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Grand, J. A., Rogelberg, S. G., Allen, T. D., Landis, R. S., Reynolds, D. H., Scott, J. C., Tonidandel, S., & Truxillo, D. M. (2018). A Systems-Based Approach to Fostering Robust Science in Industrial-Organizational Psychology. *Industrial and Organizational Psychology*, *11*(1), 4–42. https://doi.org/10.1017/IOP.2017.55

Grolemund, G. (2014). *Hands-on programming with R*. O'Reilly Media. https://www.oreilly.com/library/view/hands-on-programming-with/9781449359089/

Groves, W. (2013). Using domain knowledge to systematically guide feature selection. *IJCAI International Joint Conference on Artificial Intelligence*, 3215–3216. https://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/viewPaper/6999

Guyon, I., Bitter, H.-M., Ahmed, Z., Brown, M., & Heller, J. (2005). Multivariate Non-Linear Feature Selection with Kernel Methods. In *Soft Computing for Information Processing and Analysis* (pp. 313–326). Springer-Verlag. https://doi.org/10.1007/3-540-32365-1_12

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc. https://doi.org/10.1016/C2009-0-61819-5

Hao, T., Xing, G., & Zhou, G. (2013). ISleep: Unobtrusive sleep quality monitoring using smartphones. *SenSys 2013 - Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. https://doi.org/10.1145/2517351.2517359

Harford, T. (2014). Big data: A big mistake? *Significance*, *11*(5), 14–19. https://doi.org/10.1111/J.1740-9713.2014.00778.X

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. In *Springer Series in Statistics* (2nd ed.). Springer-Verlag. https://doi.org/10.1007/978-0-387-84858-7

Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G. F., & Clermont, G. (2013). Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics*, *46*(1), 47–55. https://doi.org/10.1016/j.jbi.2012.08.004

Hemminki, S., Nurmi, P., & Tarkoma, S. (2013). Accelerometer-based transportation mode detection on smartphones. *SenSys 2013 - Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. https://doi.org/10.1145/2517351.2517367

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? In *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X0999152X

Hollon, S. D., Cohen, Z. D., Singla, D. R., & Andrews, P. W. (2019). Recent Developments in the Treatment of Depression. *Behavior Therapy*, *50*(2), 257–269. https://doi.org/10.1016/j.beth.2019.01.002

Horowitz, M. (2015). *Detailed NFL Play-by-Play Data 2015*. https://www.kaggle.com/maxhorowitz/nflplaybyplay2015

Hwong, Y.-L., Oliver, C., Van Kranendonk, M., Sammut, C., & Seroussi, Y. (2017). What makes you tick? The psychology of social media engagement in space science

communication. *Computers in Human Behavior*, *68*, 480–492. https://doi.org/10.1016/j.chb.2016.11.068

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/JOURNAL.PMED.0020124

Jäger, S., Allhorn, A., & Bießmann, F. (2021). A Benchmark for Data Imputation Methods. *Frontiers in Big Data*, *0*, 48. https://doi.org/10.3389/FDATA.2021.693674

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning : with applications in R*. Springer.

Jia, Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *J Med Internet Res 2020;22(11):E20550 Https://Www.Jmir.Org/2020/11/E20550*, *22*(11), e20550. https://doi.org/10.2196/20550

Kazil, J., & Jarmul, K. (2016). *Data Wrangling with Python and Pandas*. O'Reilly Media. https://www.cs.tufts.edu/comp/150VAN/demos/DataWrangling.pdf

Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: an improved missing data imputation technique. *Journal of Big Data*, *7*(1), 1–21. https://doi.org/10.1186/S40537-020-00313-W

Kim, J., & Mueller, C. W. (1978). *Factor analysis: statistical methods and practical issues*. SAGE Publications.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text Mining in Organizational Research. *Organizational Research Methods*, *21*(3), 733–765. https://doi.org/10.1177/1094428117722619

Kohavi, R., & Provost, F. (1998). Glossary of Terms. *Machine Learning*, *30*, 271–274. https://doi.org/10.1023/A:1017181826899

Kohavi, Ron. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference of Artificial Intelligence*. https://www.researchgate.net/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection

Kohavi, Ron, & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324. https://doi.org/10.1016/s0004-3702(97)00043-x

Korstanje, J. (2020). *Is Python faster than R? R vs Python Speed Benchmark on a simple Machine Learning Pipeline*. Towards Data Science. https://towardsdatascience.com/is-python-faster-than-r-db06c5be5ce8

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21*(4), 493–506. https://doi.org/10.1037/met0000105

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. https://doi.org/10.1007/BF02289565

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer Nature.

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection : a Practical Approach for Predictive Models.* CRC Press LLC.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., & Campbell, A. T. (2010). A survey of mobile phone sensing. *IEEE Communications Magazine*, *48*(9), 140–150. https://doi.org/10.1109/MCOM.2010.5560598

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies*. https://doi.org/10.1016/j.infsof.2008.09.005

Lathia, N., Sandstrom, G. M., Mascolo, C., & Rentfrow, P. J. (2017). Happier People Live More Active Lives: Using Smartphones to Link Happiness and Physical Activity. *PLOS ONE*, *12*(1), e0160589. https://doi.org/10.1371/journal.pone.0160589

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration. *International Review of Social Psychology*, *32*(1). https://doi.org/10.5334/IRSP.289

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys*, *50*(6). https://doi.org/10.1145/3136625

Lin, G. M., Nagamine, M., Yang, S. N., Tai, Y. M., Lin, C., & Sato, H. (2020). Machine Learning Based Suicide Ideation Prediction for Military Personnel. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 1907–1916. https://doi.org/10.1109/JBHI.2020.2988393

Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. In *Wiley, New York.* John Wiley & Sons, Inc. https://doi.org/10.1002/9781119013563

Lohr, S. (2014). *Google Flu Trends: The Limits of Big Data*. The New York Times.

Loughrey, J., & Cunningham, P. (2004). Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets. *Research and Development in Intelligent Systems XXI*, 33–43. https://doi.org/10.1007/1-84628-102-4_3

Lu, H., Pan, W., Lane, N. D., Choudhury, T., & Campbell, A. T. (2009). SoundSense: Scalable sound sensing for people-centric applications on mobile phones. *MobiSys'09 - Proceedings of the 7th ACM International Conference on Mobile Systems, Applications, and Services*, 165–178. https://doi.org/10.1145/1555816.1555834

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-09823-4

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Markou, M., & Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, *83*(12), 2481–2497. https://doi.org/10.1016/j.sigpro.2003.07.018

Matloff, N. S. (2011). *The art of R programming : tour of statistical software design*. No Starch Press.

McKinney, W. (2017). *Python for data analysis : data wrangling with pandas, NumPy, and IPython*. Oreilly & Associates Inc.

Mikelsons, G., Smith, M., Mehrotra, A., & Musolesi, M. (2017). Towards Deep Learning Models for Psychological State Prediction using Smartphone Data: Challenges and Opportunities. *ArXiv*. http://arxiv.org/abs/1711.06350

Miller, T. (2017). *Explanation in Artificial Intelligence: Insights from the Social Sciences*. http://arxiv.org/abs/1706.07269

Molnar, C. (2019). *Interpretable Machine. A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/

Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python and Scikit-Learn. In *O'Reilly Media, Inc.* O'Reilly Media.

http://kukuruku.co/hub/python/introduction-to-machine-learning-with-python-andscikit-learn

Murnane, E. L., Abdullah, S., Matthews, M., Kay, M., Kientz, J. A., Choudhury, T., Gay, G., & Cosley, D. (2016). Mobile manifestations of alertness: Connecting biological rhythms with patterns of smartphone app use. *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2016*, 465–477. https://doi.org/10.1145/2935334.2935383

Nissenbaum, H. (2011). *A Contextual Approach to Privacy Online*.

Open Science Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Pan, L., Zhou, H., Liu, Y., & Wang, M. (2019). Global event influence model: integrating crowd motion and social psychology for global anomaly detection in dense crowds. *Journal of Electronic Imaging*, *28*(2), 1–18. https://doi.org/10.1117/1.JEI.28.2.023033

Paxton, A., & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods*, *49*(5), 1630–1638. https://doi.org/10.3758/s13428-017-0874-x

Petrozziello, A., Jordanov, I., & Sommeregger, C. (2018). Distributed Neural Networks for Missing Big Data Imputation. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2018.8489488

Qin, X., Liao, H., Zheng, X., & Liu, X. (2019). Stock Market Exposure and Anxiety in a Turbulent Market: Evidence From China. *Frontiers in Psychology*, *0*(FEB), 328. https://doi.org/10.3389/FPSYG.2019.00328

R Development Core Team. (2021). R: A Language and Environment for Statistical Computing. In *R Foundation for Statistical Computing* (4.0.5). https://www.r-project.org/

Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., & Aucinas, A. (2010). EmotionSense: A mobile phones based adaptive platform for experimental social psychology research. *UbiComp'10 - Proceedings of the 2010 ACM Conference on Ubiquitous Computing*, 281–290. https://doi.org/10.1145/1864349.1864393

Ramli, S., & Nordin, S. (2018). Personality Prediction Based on Iris Position Classification Using Support Vector Machines. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(3), 667. https://doi.org/10.11591/ijeecs.v9.i3.pp667-672

Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine Learning & Deep Learning with Python, Scikit-Learn and TensorFlow 2, Third Edition. In *Packt Publishing Ltd* (Issue January 2010). Packt Publishing.

Rokach, L., & Maimon, O. (2005). Decision Trees. *Data Mining and Knowledge Discovery Handbook*, 165–192. https://doi.org/10.1007/0-387-25465-X_9

Rosenbusch, H., Soldner, F., Evans, A. M., & Zeelenberg, M. (2021). Supervised machine learning methods in psychology: A practical introduction with annotated R code. *Social and Personality Psychology Compass*, *15*(2), e12579. https://doi.org/10.1111/SPC3.12579

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471725382

Roweis, S. T., & Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, *290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. In D. B. Rubin (Ed.), *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Schmid Mast, M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., & Choudhury, T. (2015). Social Sensing for Psychology. *Current Directions in Psychological Science*, *24*(2), 154–160. https://doi.org/10.1177/0963721414560811

Schwartz, H., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., Kosinski, M., & Ungar, L. (2014, January 1). Towards Assessing Changes in Degree of Depression through Facebook. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. https://doi.org/10.3115/v1/W14-3214

Shaw, Z. (2017). *Learn Python 3 the hard way : a very simple introduction to the terrifyingly beautiful world of computers and code*. Addison-Wesley Professional.

Shu, X. (2020). *Knowledge discovery in the social sciences a data mining approach*. University of California Press.

Sijbrandij, S. (2017). *Coding Careers: Developers As The Next Mass Profession*. Forbes. https://www.forbes.com/sites/forbestechcouncil/2017/12/12/coding-careers-developers-as-the-next-mass-profession/#3e7ca02febd9

Song, T. M., & Song, J. (2021). Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data. *Telematics and Informatics*, *58*, 101524. https://doi.org/10.1016/J.TELE.2020.101524

Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S., & Bühner, M. (2019). *Personality Research and Assessment in the Era of Machine Learning*. PsyArXiv. https://doi.org/10.31234/OSF.IO/EFNJ8

Steinberg, D. (2009). CART: Classification and Regression Trees. In X. Wu & V. Kumar (Eds.), *The Top Ten Algorithms in Data Mining* (pp. 193–216). Chapman and Hall/CRC. https://doi.org/10.1201/9781420089653-17

Stevens, J. R., & Soh, L.-K. (2018). Predicting similarity judgments in intertemporal choice with machine learning. *Psychonomic Bulletin & Review*, *25*(2), 627–635. https://doi.org/10.3758/s13423-017-1398-1

Stringam, B., Gerdes, J. H., & Anderson, C. K. (2021). Legal and Ethical Issues of Collecting and Using Online Hospitality Data. *Https://Doi.Org/10.1177/19389655211040434*. https://doi.org/10.1177/19389655211040434

Swaroop, C. H. (2013). *A Byte of Python*. https://open.umn.edu/opentextbooks/formats/945

Tantithamthavorn, C., McIntosh, S., Hassan, A. E., & Matsumoto, K. (2017). An Empirical Comparison of Model Validation Techniques for Defect Prediction Models. *IEEE Transactions on Software Engineering*, *43*(1), 1–18. https://doi.org/10.1109/TSE.2016.2584050

Taylor, S. A., Jaques, N., Nosakhare, E., Sano, A., & Picard, R. (2017). Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing*, 1–1. https://doi.org/10.1109/TAFFC.2017.2784832

Teague, S. J., & Shatte, A. B. R. (2018). Exploring the Transition to Fatherhood: Feasibility Study Using Social Media and Machine Learning. *Journal of Medical Internet Research - Pediatrics and Parenting*, *1*(2), e12371. https://doi.org/10.2196/12371

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, *290*(5500), 2319–2323. https://doi.org/10.1126/science.290.5500.2319

Teng, H. S. H. S., Chen, K., & Lu, S. C. Y. S. C. (1990). Adaptive real-time anomaly detection using inductively generated sequential patterns. *Proceedings of the Symposium on Security and Privacy*, 278–284. https://doi.org/10.1109/risp.1990.63857

Thomée, S. (2018). Mobile Phone Use and Mental Health. A Review of the Research That Takes a Psychological Perspective on Exposure. *International Journal of Environmental Research and Public Health*, *15*(12). https://doi.org/10.3390/ijerph15122692

Thudumu, S., Branch, P., Jin, J., & Singh, J. (Jack). (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, *7*(1), 1–30. https://doi.org/10.1186/S40537-020-00320-X

Time to discuss consent in digital-data studies. (2019). *Nature*, *572*(7767), 5. https://doi.org/10.1038/D41586-019-02322-Z

Urban, C. J., & Gates, K. M. (2021). Deep learning: A primer for psychologists. *Psychological Methods*. https://doi.org/10.1037/MET0000374

van Capelleveen, G., Poel, M., Mueller, R. M., Thornton, D., & van Hillegersberg, J. (2016). Outlier detection in healthcare fraud: A case study in the Medicaid dental domain. *International Journal of Accounting Information Systems*, *21*, 18–31. https://doi.org/10.1016/j.accinf.2016.04.001

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica.

Vezzoli, M., Zogmaister, C., & Van den Poel, D. (2020). Will they stay or will they go? Predicting customer churn in the energy sector. *Applied Marketing Analytics*, *6*(2), 136–150. https://www.ingentaconnect.com/content/hsp/ama/2020/00000006/00000002/art00006

Vieira, S., Pinaya, W. H. L., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, *74*, 58–75. https://doi.org/10.1016/J.NEUBIOREV.2017.01.002

Wang, Q., Guo, B., Peng, G., Zhou, G., & Yu, Z. (2016). CrowdWatch: Pedestrian safety assistance with mobile crowd sensing. *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 217–220. https://doi.org/10.1145/2968219.2971433

Wang, R., Harari, G., Hao, P., Zhou, X., & Campbell, A. T. (2015). SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 295–306. https://doi.org/10.1145/2750858.2804251

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257. https://doi.org/10.1037/pspa0000098

Wickham, H., & Grolemund, G. (2017). *R for data science : import, tidy, transform, visualize, and model data*. O'Reilly Media.

Wrzus, C., Brandmaier, A. M., von Oertzen, T., Müller, V., Wagner, G. G., & Riediger, M. (2012). A New Approach for Assessing Sleep Duration and Postures from Ambulatory Accelerometry. *PLoS ONE*, *7*(10), e48089. https://doi.org/10.1371/journal.pone.0048089

Wrzus, C., & Mehl, M. R. (2015). Lab and/or field? measuring personality processes and their social consequences. *European Journal of Personality*, *29*(2), 250–271. https://doi.org/10.1002/per.1986

Xie, X., Wang, C., Chen, S., Shi, G., & Zhao, Z. (2017). Real-Time Illegal Parking Detection System Based on Deep Learning. *International Conference on Deep Learning Technologies*, 23–27. https://doi.org/10.1145/3094243.3094261

Yetton, B. D., Revord, J., Margolis, S., Lyubomirsky, S., & Seitz, A. R. (2019). Cognitive and physiological measures in well-being science: Limitations and lessons. *Frontiers in Psychology*, *10*(JULY), 1630. https://doi.org/10.3389/fpsyg.2019.01630

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I. (2016). Apache spark: A unified engine for big data processing. *Communications of the ACM*, *59*(11), 56–65. https://doi.org/10.1145/2934664

Zimmer, M. (2018). Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity: *Https://Doi.Org/10.1177/2056305118768300*, *4*(2). https://doi.org/10.1177/2056305118768300

Zumel, N., Mount, J., Howard, J., & Thomas, R. (2019). *Practical Data Science With R*. Manning Publications

Appendix

List of some online data repositories.

Name	Link	Description
APA Responsible Conduct of Research	http://www.apa.org/research/responsible/data-links.aspx	It provides a collection of public datasets and repositories related to psychological fields, such as income dynamics, aging, and child abuse.
Awesome Public Datasets	https://github.com/caesar0301/awesome-public-datasets	It contains data related to various fields including psychology, cognition, and social sciences.
Data on the mind	http://www.dataonthemind.org/applicable-fields	It provides a collection of public datasets and repositories related to psychological fields, such as attention, decision making, and spatial cognition.
European Union Open Data Portal	https://open-data.europa.eu/	It contains datasets from institutions and other entities within the European Union.
Gateway to Global Aging	https://g2aging.org/	It contains datasets aging-related social and behavioral research data.
Kaggle	https://www.kaggle.com/datasets	It contains datasets from a variety of academic fields, such as consumer behavior and personality.
LearnSphere	http://learnsphere.org/index.html	It provides a collection of public datasets and repositories related to learning-related research.
National Institute on Aging - U.S. Department of Health and Human Service	https://www.nia.nih.gov/research/dbsr/publicly-available-databases-aging-related-secondary-analyses-behavioral-and-social	It provides a collection of public datasets on aging-related research.
Open-Source Psychometric Project	https://openpsychometrics.org/_rawdata/	It contains data related to personality.
openMorph	https://github.com/cMadan/openMorph	It contains data on brain morphology.
Our World in Data	https://ourworldindata.org/	It contains data related to various worldwide issues, such as poverty, human rights, and violence
PsychData	https://www.psychdata.de/index.php?main=none&sub=none&lang=eng	It contains data related to psychological fields, such as clinical, social and personality psychology
Roper Center	http://ropercenter.cornell.edu/polls/dataset-collections/	It contains public opinion data.
Wordbank	http://wordbank.stanford.edu/	It contains data on child language acquisition and vocabulary growth.