Perspectives on Behavior Science (2024) 47:203–223
https://doi.org/10.1007/s40614-023-00376-z

SI: BIG DATA AND BEHAVIOR SCIENCE

Tutorial: Lessons Learned for Behavior Analysts from Data Scientists

Leslie Neely[1] · **Sakiko Oyama**[1] · **Qian Chen**[1] · **Amina Qutub**[1] · **Chen Chen**[2]

Accepted: 30 April 2023 / Published online: 25 May 2023

Abstract

Big data is a computing term used to refer to large and complex data sets, typically consisting of terabytes or more of diverse data that is produced rapidly. The analysis of such complex data sets requires advanced analysis techniques with the capacity to identify patterns and abstract meanings from the vast data. The field of data science combines computer science with mathematics/statistics and leverages artificial intelligence, in particular machine learning, to analyze big data. This field holds great promise for behavior analysis, where both clinical and research studies produce large volumes of diverse data at a rapid pace (i.e., big data). This article presents basic lessons for the behavior analytic researchers and clinicians regarding integration of data science into the field of behavior analysis. We provide guidance on how to collect, protect, and process the data, while highlighting the importance of collaborating with data scientists to select a proper machine learning model that aligns with the project goals and develop models with input from human experts. We hope this serves as a guide to support the behavior analysts interested in the field of data science to advance their practice or research, and helps them avoid some common pitfalls.

Keywords Data science · Behavior analysis · Big data

Introduction

Behavior analysis is a natural science focused on explaining the behavior of organisms, including humans (Cooper et al., 2019). Within the science are the four domains: theory and philosophy, experimental analysis of behavior, applied

✉ Leslie Neely
leslie.neely@utsa.edu

[1] Department of Educational Psychology, University of Texas at San Antonio, 501 West Cesar Chavez, San Antonio, TX 78207, USA

[2] University of Central Florida, Orlando, FL, USA

behavior analysis (ABA), and practice/therapy based on the science of ABA. One of the hallmarks of behavior analysis, as applied to humans in particular, is the collection of observational data on individuals, often in single-subject research designs (Kazdin, 2011). For example, in the ABA-based behavioral therapy, a clinician will operationalize a behavior so that it is observable and measurable. The clinician then collects data on the individual patient, graphs the data, and uses visual analysis to make data-based decisions regarding the effects of the therapy on the target behavior (Kipfmiller et al., 2019). The clinician may also consider data from other sources, such as caregiver reports, or potentially, other sources such as sensors (e.g., wearable technologies). This process repeated over multiple sessions generates a large volume of rich data for just the single patient. With a growing patient population, the ABA-based behavioral therapy generates what can be termed as "big data."

Big data is a computing term used to refer to large and complex data sets (Cox & Ellsworth, 1997). Big data does not refer to just a sheer volume of data. Rather, in the context of health data, it often refers to multimodal (variety of different sources), heterogeneous, high-dimensional (many data points for a single subject), and large volumes of data that is rapidly produced (velocity; Zikopoulos et al., 2012). For example, daily behavior data collected by behavior analysts, combined with other sources of data such as medical records and lifestyle data (e.g., sleep pattern data), would generate big data meeting the requirements of variety (i.e., multiple data sources), volume, and velocity. Although behavior analysts typically use visual analysis to analyze data, these data sets are too complex to be analyzed by humans; even too complex for traditional computational models.

The field of data science combines computer science with mathematics/statistics and leverages artificial intelligence aimed to extract meaning from complex datasets (Dhar, 2013). Data science addresses the capacity issues of traditional mathematics by leveraging artificial intelligence, particularly machine learning. Machine learning, a branch of artificial intelligence, uses machines to analyze data at a capacity much larger than human abilities (Alloghani et al., 2020). The basic premise is that systems can learn from data, identify patterns, and make decisions with minimum human input, although humans are necessary to determine the model, set and tune the parameters of the model, and interpret what the results mean and contextualize the results for their field. The field of data science has already led to impressive breakthroughs in other fields. For example, data scientists were able to design a neural network that can predict whether an individual has Parkinson's disease from their nighttime breathing patterns (Yang et al., 2022). A second recent advancement used classification techniques to identify the risk of dementia can be predicted using gait analysis (i.e., how the person was walking) combined with cognition tests (Collyer et al., 2022). As a third example, recent bioinformatics analyses of high-dimensional molecular sequencing data enabled identification of new de novo genetic risks for cognitive symptoms of autism and other neurodevelopment disorders from a cohort of > 40 K individuals (Fu et al., 2022; Zhou et al., 2022). With breakthroughs like these enabled by computational approaches, it stands that the field of data science holds great promise for the field of behavior analysis in abstracting meaningful information from the complex data humans and traditional computational methods have not been able to analyze. For example, large-scale analysis of child demographic and molecular data combined with initial standardized and curriculum-based assessments might allow for prediction of responsiveness to therapy based on the science of ABA. Analyses, such as these, may also allow for more precise dosing of therapy. The goal of this article is to present some basic lessons for the behavior analytic researchers regarding integration of data science into the field of behavior analysis. We hope this serves as a guide to support the behavior analysis interested in the field of data science to advance their practice or research, and helps them avoid some common pitfalls.

Lesson 1: Collect Clean Data: Garbage in, Garbage out

Garbage in, garbage out (GIGO) was first attributed to George Fuechsel, an IBM educator and programmer, and is now a concept in computer science meaning the quality of the result is determined by the quality of the input (Butler et al., 2010). In essence, "unclean," inaccurate, incomplete, or inconsistent data can lead to inaccurate, incomplete, and inconsistent results. The best way to ensure a smooth research or clinical consumption of big data is to spend time thinking upfront about how to collect data to ensure accurate and reliable data.

The field of ABA typically relies on human observers recording data manually or recording data with the support of electronic data collection systems. One of the most important lessons in data collection is to ensure the data is accurately/reliability capturing the behavior of interest (Kazdin, 2011). When using human raters, that means recording measures of reliability, such as interobserver agreement measures. Unfortunately, within the practice of ABA, it is not necessarily standard to collect reliability data during clinical practice. Therefore, a large volume of ABA clinical data may not meet quality requirements to be included in larger analyses. In an ideal situation, the practice of ABA would make steps to improve collection of reliable data in our clinical practice. This comes at a cost though, with human capital resources necessary to commit to the collection of reliability data. Therefore, an alternative first step might be to leverage the extensive high-quality ABA research because it is standard practice to collect reliability data in ABA research (Neely et al., 2015; Vollmer et al., 2008). High-quality ABA research also collects treatment fidelity data, ensuring the data on the critical features of intervention and environment are implemented (Falakfarsa et al., 2022).

Although a majority of clinical data may not meet the quality standards to be included in larger scale analysis, some data might be easier to validate retrospectively. For example, practitioners may be interested in patterns of attendance/cancellations (e.g., utilization) and the potential role social determinants of health play (i.e., economic stability, education access and quality, healthcare access and quality, neighborhood and built environment, and social and community context; U.S. Department of Health & Human Services, 2023). This data could be used in multiple ways, such as identifying new service offerings (e.g., telehealth) or to identify what other service providers might be needed to support a particular family (e.g., social worker or health-care navigator). Attendance data can be easily validated and then analyzed with other data, such as age, distance family lives from service area, attendance at school, and socioeconomic status.

In addition to leveraging the ABA literature base, a second solution might be to identify approaches to automatize data collection in research and clinical practice. A review by Bak et al. (2021) of the ABA literature base identified less than 10% of behavior analytic research used automatic data collection, with less than half of those articles being applied research. Automation of data collection in the practice of ABA may be an important step in collecting valid, reliable, and accurate data in research and clinical populations. The field of data science is well-equipped to work with large data volumes from automatic data collection sources. Equipment used to capture such data can include video cameras, inertial measurement units (IMUs), accelerometers, activity monitors, sleep mats, smartwatches, etc. For example, IMU devices are used to characterize movement/activity patterns by integrating data captured by multiple sensors (i.e., accelerometer, gyroscope, and magnetometer). Integrating IMU with other smart devices and biosensors into the Internet of Things (IoT; Gubbi et al., 2013), a network of connected devices and cloud computing technology, can allow instant storage and processing of multimodal data informative for health decisions. Given that human behavior can be affected by the environmental factors, such as vocal-verbal behavior from a communication partner, collecting these data using validated instruments (e.g., language captured with microphones and processed using speech recognition algorithms) may be helpful in interpretation of the data. With these instruments collecting data at relatively high sampling frequency, the recorded data can amass quickly. Therefore, streamlining the data collection and storage process is important to manage the data.

Similar to reliability data necessary for human raters, investigators must select devices with appropriate technical specifications to collect valid data. For example, when collecting data via IMU sensors, researchers must use devices with appropriate sampling frequency and maximum data range. This is a particularly important consideration when using devices for a novel purpose. For example, IMU sensors, similar to those found in popular wearable technologies such as Apple watches, are typically used to track posture and slow movements. Therefore, the sampling frequency and range of data captured by the sensors on the device (i.e., accelerometer, gyroscope, and magnetometer) are not optimized to capture large magnitudes of impact force (e.g., how hard a boxer is punching). Collecting data at a lower than ideal sampling rate can lead to inaccurate data representation (i.e., aliasing bias), whereas data capturing at unnecessarily high sampling frequency will increase the data storage demand. Ensuring that the device has a range to capture all relevant data is important, because the data can plateau once it exceeds the maximum limit of the device. See Fig. 1 below for a screen shot of wrist acceleration during the hitting behavior measured using IMU and the motion capture system. The two systems (i.e., IMU system and the motion capture system) produced similar acceleration values for hits (i.e., impacts) below 6 g, as indicated by the generally overlapping lines on the graph. However, the acceleration values from the two systems started to deviate from one another for the hits exceeding 6 g, which was the max limit of acceleration that the IMU device could capture.

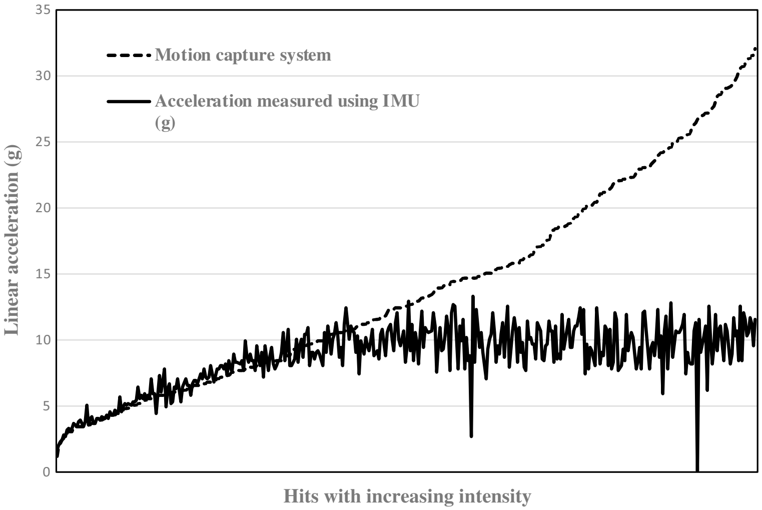

Fig. 1 Wrist linear acceleration during hitting captured using IMU and motion capture system

When using more than one input source simultaneously to collect the data, it is extremely helpful if the sources are automatically time synchronized so that there is a correspondence between the data points on all devices. For example, in our study aimed at quantifying the intensity of the problem behaviors during functional analysis sessions (Neely et al., 2022), we captured the patients' motion with IMU sensors, human raters, and videos. The recording of the IMU, human rater data collection, and video were started about the same time but had to be manually synchronized based on the timing of the first hit to ensure that the rest of the data were synchronized. This procedure was fine for this project, because we did not plan for a large sample ($n=3$). However, when considering larger samples, we need to synchronize the data.

To properly time synchronize the two data sources, the sampling rate used for one device is ideally the same or a multiple of the rate used by the other. The most accurate way to synchronize the data is through a universal time stamp or electronic synchronization, where an electrical signal is sent to all devices/software simultaneously to mark the synchronization timepoint. As an alternative, event synchronization using artificially introduced data signals that all devices can capture may be used to synchronize the data. In addition to choosing the right device and the device settings, collection of valid data requires the use of consistent data collection methods across time. Similar to retraining of human raters to ensure reliable data collection, it is recommended that researchers and clinicians conduct calibration of data collection sources at the onset of data collection and at specified intervals during data collection.

Lesson 2: You Need to Store the Data

With the collection of big data comes the need to store the data. When considering options for storing the data, the researcher or clinician might consider a couple of things including their intended use for the data, how long they need to store the data due to legal or regulatory requirements, who needs to access the data, size of the data, and their budget to pay for the storage. For example, if a behavior analytic researcher is working with data scientists, they may want to use the data to inform algorithm development, store the data for 3 years as required by their Institutional Review Board, only want the research team to access the data but want to share across institutions, collect large data sets, and have little budget. In this situation, they may consider cloud storage solutions supported by their university (such as Microsoft OneDrive, DropBox, or Box) or local solutions (such as external hard drives, local servers, and database management systems such as Microsoft SQL Server, Oracle Database, and Microsoft Access). Cloud-based storage and search systems (e.g., Google Cloud Storage, Google BigQuery, Amazon Web Services) that are also quickly integrated with machine learning and artificial intelligence tool platforms, and enable secure sharing provide advantages for collaborative team projects. Potential storage solutions are available in Table 1.

Lesson 3: Don't Skip the Pre-Processing Stage

The quality of the data directly determines the prediction capacity and generalizability of the machine learning model (Jain et al., 2020). In real world applications, such as behavior analysis, the collected data may contain a large number of missing values, a lot of "noise," or there may be abnormal points due to manual input errors, which is unfavorable for the training of reliable machine learning models. Therefore, data preprocessing becomes a standard and necessary step to transform or encode data so that it may be easily and accurately parsed by the machine (García et al., 2015). In general, there are four main data pre-processing steps including data cleaning, data integration, data transformation, and data dimensionality reduction (García et al., 2016). These data pre-processing steps are important for downstream machine learning applications such as classification, clustering, and regression. In the following, we overview the commonly used methods in each pre-processing step. Because these topics are extensive, we also provide some recommendations for further reading at the end of each section.

Step 1: Data Cleaning

The main idea of data cleaning is to "clean" data by filling in missing values, smoothing noisy data, resolving the inconsistency, and removing outliers.

Table 1 Potential Storage Solutions

Storage Solutions	Example	Positives	Considerations
Commercial Cloud Storage	Amazon Web Services; Microsoft OneDrive; Dropbox; Box	Readily available; cybersecurity protections included; can handle large data sets; easy to share data between collaborators	Subscription; subject to cyberattacks; data saving can be long dependent on internet connectivity; synchronization of files between collaborators; data owner has to share control of data with cloud storage business
Local Storage Solution	External Hard Drive	Less vulnerable to cyber-attacks with encryption; can store large data; one-time cost; data saving is quick; data owner has sole control of the data	Difficult to share between collaborators; physical security is a concern (e.g., stolen device); the vulnerable computing systems where the external hard drives are connected to can also undermine data security (e.g., malware); insider threats/attacks; data owners/researchers might be malicious or negligent insiders who modify or delete data illegitimately
	Local Hard Drive on Computer or a Local Server	Less vulnerable to cyberattacks if the PC and local area network are well protected (e.g., if the organization has a strong onsite cybersecurity team/IT office); can store large data; one-time cost; data saving is quick; data owner has sole control of the data	Easy for internal data sharing. External users can use VPN to access the internal network to access the data; physical security and cybersecurity are still big concerns such as stolen devices, compromised computing network, not patched frequently patched PCs, and malicious or negligent insider attackers

Table 1 (continued)

Storage Solutions	Example	Positives	Considerations
Customized block-chain database	Using blockchain platforms such as Hyperledger Fabric	Enhanced data security and more resilient to cyber-attacks because of blockchain's default transparency, immutability properties; can store large data; data owner has sole control of the data if the blockchain systems are private	Need technical expertise to design, deploy, manage and maintain blockchain-based databases and their security.@ongoing cost to maintain database;@data saving requires additional steps (process to correct format).@Blockchain databases are still suffering from cyber threats to information systems (PCs)
Database management system (DBMS)	MySQL, PostgreSQL, Microsoft SQL Server, Oracle Database, and Microsoft Access	Significantly strengthened data exchange; data integrity is maintained; data is backed up	Hardware and software expenditure (requiring a high-speed CPU and a large memory to perform the DBMS software); multiuser DBMS can be more expensive

Table 2 Skill Acquisition of Handwashing with Task Analysis

Step	Trial 1	Trial 2	Trial 3
Approaches sink	Independent	Independent	Independent
Steps up on stool	Independent	Independent	Independent
Turns on cold water tap	Independent	Independent	Independent
Pumps soap onto hand	Gestural Prompt	Independent	Independent
Rubs hands together for 20 s	Physical Prompt	Gestural Prompt	Independent
Turns off cold water tap	Physical Prompt	Physical Prompt	Gestural Prompt
Steps off of stool	Physical Prompt	Physical Prompt	Physical Prompt
Takes a paper towel from pile	Physical Prompt	Physical Prompt	Physical Prompt
Wipes hands until dry	Physical Prompt	Not recorded	Physical Prompt
Throws paper towel into trashcan	Physical Prompt	Physical Prompt	Physical Prompt

Table 3 Skill Acquisition of Handwashing Unclean Data

Step	Independent	Gestural	Physical Prompt
Approaches sink	1	0	0
Steps up on stool	1	Not Recorded[1]	0
Turns on cold water tap	1	0	0
Pumps soap onto hand	0	1	0
Rubs hands together for 20 s	0	0	1
Turns off cold water tap	Not Recorded[1]	0	0
Steps off of stool	0	0	1
Takes a paper towel from pile	0	0	1
Wipes hands until dry	0	0	Not Recorded[1]
Throws paper towel into trashcan	0	0	1

[1] Deleted data from Table 2 example for illustration purposes

Dealing with Missing Values Data collected using human rater data is often plagued with missing values due to human error or technology failures. Inappropriate handling of the missing values in the data analysis may introduce bias and lead to misleading conclusions. There are several commonly used techniques to handle missing values. (1) Using central tendency (mean, median) to replace the missing values. This is often referred to as data imputation. There are more sophisticated imputation methods such as k nearest neighbor (k-nn) imputation (Zhang, 2008) and imputation using multivariate imputation by chained equation (Azur et al., 2011). Using interpolation methods (e.g., linear interpolation, spine interpolation, simple moving average, and weighted moving average) to impute the missing values. For example, Table 2 contains raw data from a handwashing program where the behavior analyst conducted three trials and recorded the level of prompting for each step. In Table 3, we converted to binary data, removed some of the recordings, and deleted some of the data. We then labeled the deleted data as "not recorded" for illustration. One way to deal with missing data is to calculate the mean/median of the data point

Table 4 Clean Data using Mean/Median of Before and After Missing Data

Step	Independent	Gestural	Physical Prompt
Approaches sink	1	0	0
Steps up on stool	1	0	0
Turns on cold water tap	1	0	0
Pumps soap onto hand	0	1	0
Rubs hands together for 20 s	0	0	1
Turns off cold water tap	0	0	0
Steps off of stool	0	0	1
Takes a paper towel from pile	0	0	1
Wipes hands until dry	0	0	1
Throws paper towel into trashcan	0	0	1

immediately before and after the missing data point. Table 4 contains the “cleaned” data using this approach. Readers are referred to García et al. (2015; Chapter 4) and Han et al. (2011; Chapter 3, Sect. 3.2) for more details on various techniques for handling missing values in the data.

Dealing with Noisy Data The data collected from IoT devices can be noisy, which is caused by various factors including electronic equipment disturbance, IoT device failure, and other causes of error. The noisy data points can potentially corrupt the knowledge that can be extracted by machine learning algorithms. In order to treat noise in data analysis, one can exploit existing outlier detection techniques for handling data with extremely high levels of noise (Xiong et al., 2006). Noise filtering is another widely used method which identifies and removes the noisy instances in the data (García et al., 2015). For example, in one of our behavior analysis studies, we used an RGB-depth camera and an accelerometer for hand gesture recognition (Liu et al., 2014). However, due to the presence of various noise sources in an actual operating environment, general movements (such as high fives) often appear in the collected sensor data (e.g., inertial signals). To smooth out the noisy input data, a moving average filtering is applied in a local sliding-window fashion (e.g., a window of 10 consecutive data points). Figure 2 of Liu et al. (2014) shows an example of the raw and filtered signals from the RGB-depth camera (capturing the position information of a moving hand) and the inertial sensor (capturing the acceleration information of a moving hand). Our experiments showed that this data filtering step can improve the hand gesture recognition accuracy considerable by adequately reducing jitters in the signals.

Step 2: Data Integration

For behavior analysis, data can be collected from multiple sources (e.g., human rater data, sleep data, pharmaceutical data, wearable IMU sensors, fixed cameras), data integration has become a vital part of the process. We provided some guidance on the need to synchronize data collection in Lesson 1. However, typical operations for data integration also include the identification and unification of variables and domains, and the analysis of attribute correlation (García et al., 2015). The readers are referred to (García et al., 2015) and (Han et al., 2011) for a comprehensive set of methods for data integration. Data integration is an essential step for behavior analysis, especially when dealing with multiple sources. The key to effective data integration is understanding the nature of each data source and ensuring that they are combined in a meaningful way based on the following two main steps: understand the data sources and determine the integration strategy.

Understand the Data Sources Before integrating data, understand the nature, format, and structure of each data source. For example, human rater data might be in the form of questionnaires or interview transcripts, sleep data could be time series data, pharmaceutical data might be in spreadsheets, wearable IMU sensors could provide continuous data, and cameras may generate video data.

Determine the Integration Strategy Decide how the data sources will be combined. This might involve merging, concatenating, or joining the datasets. Consider if the integration will be done horizontally (combining features or variables) or vertically (combining instances or observations).

Here we provide a walk-through example: Suppose we want to integrate sleep data and wearable IMU sensor data for behavior analysis. Sleep data is collected at the end of each day, whereas IMU sensor data is collected every minute in a typical setting. To integrate these two datasets, we need to have a common feature (also called a key) on which we can join them. In this case, the common feature is the date. However, before we can join the datasets, we need to aggregate the IMU sensor data to match the daily granularity of the sleep data. Aggregating the IMU sensor data means summarizing the minute-by-minute data into daily values. This can be done using different aggregation methods, such as calculating the average, sum, or maximum value for each day. A step-by-step explanation of the process is presented as follows:

1) Aggregate the IMU sensor data: For each day, calculate the desired summary statistic for the IMU sensor data. For example, you could calculate the daily average, total, or maximum value for each IMU sensor variable. This will create a new dataset with daily values for each IMU sensor variable.
2) Add a date feature to both datasets: Ensure that both the sleep data and the aggregated IMU sensor data have a date feature in a consistent format (e.g., YYYY-MM-DD). The date feature will serve as the key for joining the datasets. Join the datasets on the date feature: Combine the sleep data and the aggregated IMU sensor data by matching the date feature in both datasets. This can be done using a join operation, where each row in the sleep data is matched with the corresponding row in the IMU sensor data based on the date. After completing these steps, we will have an integrated dataset with daily sleep data and daily aggregated IMU sensor data. This dataset can then be used for further analysis,

Table 5 Categorical Data from a Functional Analysis

Participant	Topography of Behavior	Identified Function
A1BL	Aggression	Social Positive: Tangible
A2BL2	Aggression	Social Negative: Escape
B3BL3	Self-injury	Social Positive: Tangible
A4BL4	Self-injury	Social Negative: Escape

Table 6 Converted Data

	Aggression	Self-Injury	Tangible	Escape
A1BL1	1	0	1	0
A2BL2	1	0	0	1
A3BL3	0	1	1	0
A4BL4	0	1	0	1

such as identifying relationships between sleep patterns and physical activity levels, or building predictive models for behavior analysis.

Step 3: Data Transformation

Data transformation includes many techniques include binarization and normalization of data for the machine learning models to properly process the data (García et al., 2015). To note, there are other techniques (e.g., data discretization) not discussed. Readers can reference Ramírez-Gallego et al. (2016) and Yang et al. (2009) for guidance on additional techniques.

Data Binarization Data binarization refers to the transformation of categorical data or continuous data into binary (0 s and 1 s) for ease in analysis. For example, Table 5 presents data for four participants with two "categories" (i.e., topographies) of behavior (i.e., aggression and self-injurious behavior) and two "categories" for function (i.e., access to tangible and escape from demand). A behavior analyst might clean the data by converting categorical data (Table 5) into binary data as presented in Table 6. Also, Table 3 provides a binarization of the previous handwashing task analysis example presented in Table 2.

Data Normalization The dimensions of different features in the data may be inconsistent, and the difference between the values may be very large. If not processed, the results of the data analysis may be affected. Therefore, the data needs to be scaled according to a certain ratio to make it fall in a specific range, e.g., [-1, 1]. In particular, the data/features have to be normalized before feeding them to distance-based mining methods, clustering algorithms, and supported vector machine (SVM). In the following, we introduce two commonly used data normalization techniques. More details on the data normalization methods can be found in (García et al., 2015) and (Han et al., 2011).

(1) Min–Max normalization applies a linear transformation on the original data and maps the data to new range. Suppose that min$X$ and max$X$ are the minimum and maximum values of an attribute X. To map a value x of X to $\bar{x}$ in a new range $[new\ \min\ X, new\ \max\ X]$, the Min–Max normalization is computed as follows:

$$\bar{x} = \frac{x - \min\ X}{\max\ X - \min\ X}(new\ \max\ X - new\ \min\ X) + new\ \min\ X$$

Min–Max normalization is able to preserve the relationship among the original data values.

(2) Z-score normalization processes the data to have a mean of 0 and a variance of 1 based on the following equation (we adopt the same notations from Min–Max normalization):

$$\bar{x} = \frac{x - \overline{X}}{\sigma_X}$$

where $\overline{X}$ and σ_X are the mean and standard deviation, respectively.

In the context of our previous example (mentioned in Step 2: Data Integration) with sleep data and IMU sensor data, let's assume we have the following features after integrating the datasets: 1. "total_sleep_hours"—total hours of sleep per day (range: 0–12 h); 2. "avg_activity_level"—average activity level per day from the IMU sensor (range: 0–100, arbitrary activity units).

To perform data normalization, we can apply either Min–Max scaling or Z-score normalization on the two features (i.e., "total_sleep_hours" and "avg_activity_level"), the values of the features will be on a common scale, making it easier to compare them and use them in machine learning algorithms.

Step 4: Data Dimensionality Reduction

When datasets become large in the number of predictor variables or the number of samples/instances, machine learning algorithms face the curse of dimensionality problem (Verleysen & François, 2005). It will impede the operation of most machine learning algorithms as the computational cost rises. The purpose of data dimensionality reduction is to achieve a condensed representation of the dataset that is smaller in volume, while maintaining the integrity of the original data set. A few dimension reduction algorithms are described as follows.

Principal Component Analysis (PCA) Principal component analysis, or PCA, is a statistical method that reduces the numbers of attributes by combining highly correlated attributes together (Abdi & Williams, 2010). It can effectively transform the data from the original high dimensional space to a low-dimensional subspace spanned by a set of linearly uncorrelated features (i.e., principal components). PCA is a widely used technique for dimension reduction when many of the variables are highly correlated with each other. It is an unsupervised approach. The reader is referred to Abdi and Williams (2010) for more details and procedures of applying PCA on the data.

Linear Discriminant Analysis (LDA) Linear discriminant analysis, or LDA, is a supervised dimensionality reduction method that can be performed before applying a machine learning model for supervised learning tasks (e.g., behavior classification; Tharwat et al., 2017). In supervised learning tasks, the training data is typically accompanied with label information. Take behavior classification using IMU sensor data for instance, each training data sample has a label, i.e., the category of the behavior like "aggression" or "self-injury." LDA transforms the data into a lower dimensional space, where samples belonging to the same class are pulled close to each other and samples from different classes are pushed away, thereby maximizing class separability. The reader is referred to Tharwat et al. (2017) for more details on LDA. For implementation of the methods discussed in the data pre-processing steps, the readers are recommended to use Scikit-learn(https://scikit-learn.org/stable/), which is a free software machine learning library for the Python programming language.

Feature Selection Unlike PCA and LDA methods that transform the original data or feature to a lower dimensional space, feature selection technique aims to reduce the data dimension by selecting a subset from the original feature set according to certain feature selection criterion, while still providing good prediction results. The readers are referred to Cai et al. (2018) for a comprehensive review on the feature selection methods in machine learning.

For illustrate, Table 7 presents the results of reducing the detailed prompting data from Table 4 into two categories: independent and prompted.

Table 7 Feature Selection Prompted Versus Independent

Step	Independent	Prompted
Approaches sink	1	0
Steps up on stool	1	0
Turns on cold water tap	1	0
Pumps soap onto hand	0	1
Rubs hands together for 20 s	0	1
Turns off cold water tap	0	1
Steps off of stool	0	1
Takes a paper towel from pile	0	1
Wipes hands until dry	0	1
Throws paper towel into trashcan	0	1

Table 8 Functional Analysis Results with Binary Data and Gender

	Gender Raw (Binary)	Aggression	Self-Injury	Tangible	Escape
A1BL1	M (0)	1	0	1	0
A2BL2	M (0)	1	0	0	1
A3BL3	F (1)	0	1	1	0
A4BL4	F (1)	0	1	0	1

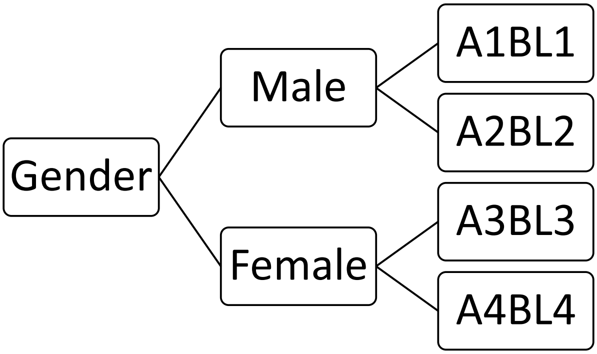

Fig. 2 Results of clustering

Lesson 4: Determine Your Use Case

Although behavior analysts may be responsible for the collection and organization of the data, they might not necessarily be developing the model. However, behavior analysts do need to have a basic understanding of the different approaches to analysis so that they can identify the right data scientist and/or know what kind of analysis they are needing for their data. Therefore, this section is meant to provide a basic introduction so that the behavior analyst can know what to ask for and who to seek out.

Algorithms (the product of machine learning) can be developed via unsupervised learning or supervised learning training methods (Alloghani et al., 2020; Sandhu, 2018). Starting with unsupervised learning, a behavior analyst would submit "unlabeled" data, or data that is not already clustered. The behavior analyst may then ask a question to which they do not already have an explanation or hypothesis for how to group the data. The machine learning model is then applied to discover patterns. For example, using Table 6, we might add gender to the table (Table 8). The algorithm might then cluster A1BL1 and A2BL2 together and A3BL3 and A4BL4 together (see Fig. 2). If we do the analysis, we might identify the machine learning algorithm has clustered according to the gender attribute. To note, we would not want to make strong statements about gender predicting topography of problem behavior using such a small dataset.

Supervised learning aims to predict OR classify values based on inputs (Alloghani et al., 2020). Supervised learning requires labeled data (e.g., coded data with classifications) and a specific question. The behavior analyst would create a "training data set" with the labeled data and train the algorithm based on the labeled data set. For example, the behavior analyst may have data on child sleep patterns and daytime problem behavior demonstrating increased problem behavior following nights of restless sleep. They would classify this data (e.g., poor night of sleep versus good night of sleep) and the resulting daytime problem behavior (e.g., low level of behaviors vs. high level of behaviors). The more data they provide (or "feed") to the algorithm, the hope is the algorithm will become more accurate (depending on the model). They can then use the nighttime data to train the model to predict daytime behavior. They could allow the behavior analyst to modify their treatment for days following poor sleep behavior. Using existing data sets to develop predictive algorithms can have many applications in behavior analysis, for example, identifying patterns of behavior and responsiveness to treatment to help inform dosing of therapy sessions. However, this approach does require the behavior analyst to prepare a training data set and classify the data prior to sending to the data scientist.

Lesson 5: Model Development is Not Static

Although collecting and analyzing big data using artificial intelligence can allow for identification of patterns and development of prediction models beyond the capacity of a human scientist, the model development should include the context experts to ensure applicability and generalizability to the field. An example of model development "gone wrong" is the case of the IBM Watson for Oncology. In 2017, *STAT News* published a report highlighting the counter therapeutic and potentially harmful recommendations originating from Watson for Oncology (Ross & Swetlitz, 2017). Upon investigation, it was revealed that in early iterations of Watson's predictions, hypothetical data, rather than real patient data, were utilized to train the model. This undermined the training of the model, resulting in an inaccurate model. To avoid this very situation, model development is best structured as an iterative process involving all stakeholders: behavioral scientists, clinicians, biologists and the modelers/artificial intelligence experts, etc. An effort by all parties to educate themselves in the terminology and paradigms of each other's' fields accelerates progress. As such, a simple primer on the arch of model development is provided in the Fig. 3.

In addition to the steps identified in Fig. 3, a vital piece of model development involves sharing back the data to inform clinical/educational decisions and dissemination (Hosny et al., 2019). Forms of dissemination can be broadly classified into three categories: (1) dissemination within the field of behavioral and clinical experts; (2) dissemination and sharing with other modelers; and (3) dissemination to the public. Dissemination with the field can take typical forms, such as publications and conference presentations. Novel advancements in data visualization and modeling are also enabling ways to present modeling results

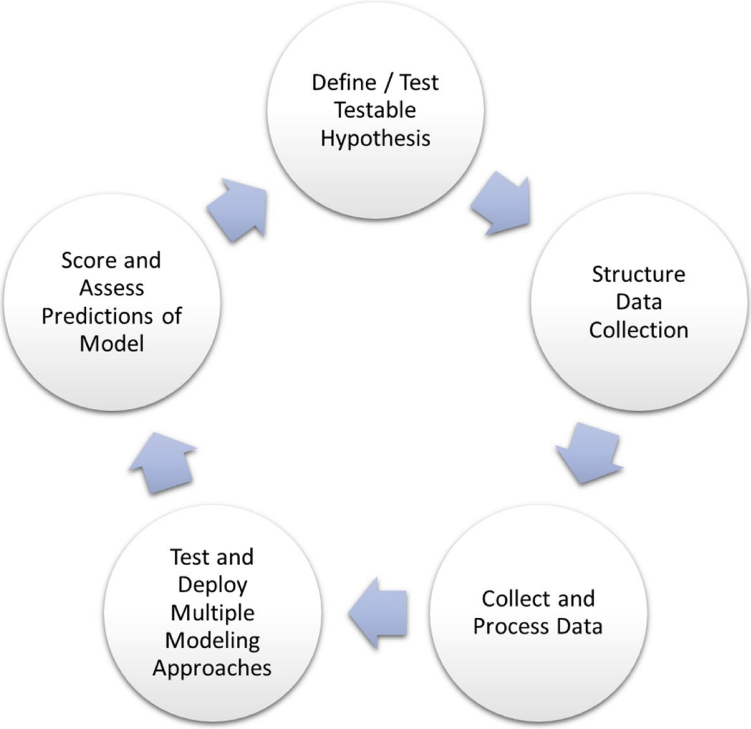
Define / Test Testable Hypothesis
Structure Data Collection
Collect and Process Data
Test and Deploy Multiple Modeling Approaches
Score and Assess Predictions of Model

Fig. 3 Model Development Process

in real-time and obtain expert curation by a mouse-click or voice command that is incorporated back into the model and updates data (Zhang & Gao, 2021). For disseminating and sharing with modelers, dissemination of data and code within the fields of systems biology and artificial intelligence benefits from a long history and tools to enabling sharing and development of models. These include coding standards and version control, code repositories (e.g., Github) and data repositories (e.g., the National Sleep Research Resource, sleepdata.org). Behavior analysts should consider, as a field, contributing to data repositories and code repositories in a similar fashion. Dissemination to the public, or translation of knowledge into lay terms and clear communication to the public who can benefit from the study's findings, enable actionable outcomes from the work. Unfortunately, this communication is often delayed (Munro & Savel, 2016). Unique steps to engage the public and computational audiences already in use in the data science field include: (1) hosting crowd-sourced competitions on the data (e.g., through platforms like Kaggle or DREAM; Boudreau & Lakhani, 2013); and (2) sharing the data and models at public forums (Parrott, 2022). Behavior analysts might consider adopting these unique approaches to dissemination.

Lesson 6: You Need to Protect the Data

One of the major responsibilities of the researcher and clinician when working with human patient data is to protect that data. Unfortunately, protection of the data is not always implemented with fidelity. Take, for example, the massive breach of Verkada, a company specializing in cloud-based security systems (Randolph & Hunt, 2021). In March 2021, hackers were able to penetrate Verkada data, gaining access to live security feeds from over 150,000 hospital, school, and prison cameras. Through this breech, the hackers were able to view private videos from women's health appointments and inside of a mental health clinic where a man was experiencing a manic episode requiring restraint. In July 2019, one Alabama hospital's network was shut down for 8 days due to cyberattacks, which resulted in a baby's death. According to the *Wall Street Journal*'s report (Poulsen et al., 2021), this attack is now allegedly linked to the first hospital death caused by a ransomware attack. These attacks continue to happen with approximately 40 million patients across different health care networks falling victim to cyberattacks in 2021 (Jerich, 2021).

These cybersecurity incidents highlight one of the most important lessons for behavior analytic researchers and clinicians: you must ensure you have systems/resources in place to protect the data you collect as well as ensure the privacy of your patients/participants. Although recently the cybersecurity and health-care communities has been collaborating to enhance cybersecurity awareness (Joseph, 2018) the commercial-off-the-shelf cybersecurity solutions are still in their early stage. Like the healthcare industry, the behavior analytic researchers, clinicians and their patients along with the underlying human behavior data, technologies, and AI-based data analytics approaches are vulnerable to cyber threats. It is very likely behavior science may be one the most targeted discipline of the healthcare industry for cybercrime in the future when many more researchers and clinicians adopt state-of-the-art technology. Therefore, it is highly recommended that behavior analysts partner not only with data scientists, but cybersecurity experts to ensure to identify potential solutions. The cybersecurity expert can help them identify potential off the shelf solutions and/or to design and monitor the platform/mechanism. In addition, the cost associated with protection of the data (e.g., liability insurance and cybersecurity expert support) is negligible compared to potential violations of the Health Insurance Portability and Accountability Act of 1996, which can be penalized at $50,000 per incident if considered "willful neglect" (Alder, 2023).

Conclusion

The purpose of this article was to provide some basic lessons (or "tips") for researchers or clinicians considering collecting data from multiple sources and integrating data science techniques into their data analysis. We discussed five different lessons spanning from considerations when collecting the data to iterative machine learning model development. Although these lessons are not a comprehensive list of everything the behavior analyst should know we hope it provides some guidance and a foundation for more researchers and clinicians to start partnering with data scientists to advance our knowledge of organism behavior.

Funding This work was supported in part by the U.S. Department of Energy/National Nuclear Security Administration (DOE/NNSA) #DE-NA0003985 and UTSA Strategic Initiative Fund Grant #SIF010.

Declarations

Conflict of Interest We have no known conflict of interest to disclose.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433–459. https://doi.org/10.1002/wics.101

Alder, S. (2023). What are the penalties for HIPAA violations? *HIPAA Journal*. https://www.hipaajournal.com/what-are-the-penalties-for-hipaa-violations-7096/. Accessed 19 May 2023.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. In M. Berry, A. Mohamed, & B. Yap (Eds.), *Supervised and unsupervised learning for data science: Unsupervised and semi-supervised learning*. Springer. https://doi.org/10.1007/978-3-030-22475-2_1.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*(1), 40–49. https://doi.org/10.1002/mpr.329

Bak, M. Y. S., Plavnick, J. B., Dueñas, A. D., Brodhead, M. T., Avendaño, S. M., Wawrzonek, A. J., Weber, E., Dodson, S. N., & Oteto, N. (2021). The use of automated data collection in applied behavior analytic research: A systematic review. *Behavior Analysis: Research & Practice, 21*(4), 376–405. https://doi.org/10.1037/bar0000228

Boudreau, K. J., & Lakhani, K. R. (2013, April). Using the crowd as an innovative partner. *Harvard Business Review*. https://hbr.org/2013/04/using-the-crowd-as-an-innovation-partner. Accessed 19 May 2023.

Butler, J., Lidwell, W. & Holden, K. (2010). *Universal principles of design* (2nd ed.). Rockport Publishers. http://books.google.com/books?id=l0QPECGQySYC&pg=PA112#v=onepage&q&f=false. Accessed 26 Apr 2011.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing, 300*, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Collyer, T. A., Murray, A. M., Woods, R. L., Storey, E., Chong, T., Ryan, J., Orchard, S. G., Brodtmann, A., Srikanth, V. K., Shah, R. C., & Callisaya, M. (2022). Association of dual decline in cognition and gait speed with risk of dementia in older adults. *JAMA Network Open, 5*(5), e2214647. https://doi.org/10.1001/jamanetworkopen.2022.14647.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2019). *Applied behavior analysis* (3rd ed.). Pearson Education.

Cox, M., & Ellsworth, D (1997). Managing big data for scientific visualization. In *ACM Siggraph*, *97*(1), 21–38). https://www.researchgate.net/profile/David-Ellsworth-2/publication/238704525_Managing_big_data_for_scientific_visualization/links/54ad79d20cf2213c5fe4081a/Managing-big-data-for-scientific-visualization.pdf.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73. https://cacm.acm.org/

Falakfarsa, G., Brand, D., Jones, L., Godinez, E. S., Richardson, D. C., Hanson, R. J., Velazquez, S. D., & Willis, C. (2022). Treatment integrity reporting in *Behavior Analysis in Practice*, 2008–2019. *Behavior Analysis Practice, 15*, 443–453. https://doi.org/10.1007/s40617-021-00573-9

Fu, J. M., Satterstrom, F. K., Peng, M. Brand, H., Collins, R. L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S. P., Stevens, C. R., Cusick, C., Babadi, M., Banks, E., Collins, B., Dodge, S., Gabriel, S. B., Gauthier, L., Lee, S. K. . . . Talkowski, M. E. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism**.** *Nature Genetics*, *54*, 1320–1331. https://doi.org/10.1038/s41588-022-01104-0.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: Methods and prospects. *Big Data Analytics, 1*(1), 1–22. https://doi.org/10.1186/s41044-016-0014-0

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72). Springer International.

Gubbi, J., Buyya, R., Marusic, S., & Palaiswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *FutureGeneration Computer Systems, 29*(7), 1645–1660. https://doi.org/10.1016/j.future.2013.01.010

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.

Healthy People 2030, U.S. Department of Health and Human Services. (2023). *Office of Disease Prevention and Health Promotion*. Retrieved from https://health.gov/healthypeople/objectives-and-data/social-determinants-health

Hosny, A., Schwier, M., Berger, C., Örnek, E. P., Turan, M., Tran, P. V., Weniger, L., Isensee, F., Maier-Hein, K. H., McKinley, R., Lu, M. T., Hoffmann, U., Menze, B., Bakas, S., Fedorov, A., & Aerts, H. J. (2019). Modelhub. ai: Dissemination platform for deep learning models. *arXiv preprint* arXiv:1911.13218. *https://arxiv.org/ftp/arxiv/papers/1911/1911.13218.pdf.*

Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumbar, S., Mittal, R. S., & Munigala, V. (2020, August). Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on knowledge discovery & data mining* (pp. 3561–3562). ACM. https://doi.org/10.1145/3394486.3406477.

Jerich, K. (2021). The biggest healthcare data breaches of 2021. *Healthcare IT News*. https://www.healthcareitnews.com/news/biggest-healthcare-data-breaches-2021. Accessed 19 May 2023.

Joseph, T. (2018). CyberMed summit highlights vulnerabilities of medical technology. *Arizona Board of Regents*. https://phoenixmed.arizona.edu/newsroom/news/cybermed-summit-highlights-vulnerabilities-medical-technology. Accessed 19 May 2023.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings*. Oxford University Press.

Kipfmiller, K. J., Brodhead, M. T., Wolfe, K., LaLonde, K., Sipila, E. S., Bak, M. Y., & Fisher, M. H. (2019). Training front-line employees to conduct visual analysis using a clinical decision-making model. *Journal of Behavioral Education, 28*(3), 301–322. https://doi.org/10.1007/s10864-018-09318-1

Liu, K., Chen, C., Jafari, R., & Kehtarnavaz, N. (2014). Fusion of inertial and depth sensor data for robust hand gesture recognition. *IEEE Sensors Journal, 14*(6), 1898–1903. https://doi.org/10.1109/JSEN.2014.2306094

Munro, C. L., & Savel, R. H. (2016). Narrowing the 17-Year research to practice gap. *American Journal of Critical Care, 25*(3), 194–196. https://doi.org/10.4037/ajcc2016449

Neely, L., Cantrell, K., Svoboda, M., Graber, J., Wimberley, J., & Oyama, S. (2022). Feasibility of wearable technology to supplement measurement of behavioral intensity. [Manuscript submitted for publication]

Neely, L., Davis, H., Davis, J., & Rispoli, M. (2015). Review of reliability and integrity trends in autism-focused research. *Research in Autism Spectrum Disorder, 9*(2), 1–12. https://doi.org/10.1016/j.rasd.2014.09.011

Parrott, M. (2022). *The AI model share project*. Columbia University: Institute for Social & Economic Research & Policy in the Faculty of Arts & Sciences. https://iserp.columbia.edu/center/ai-model-share-project. Accessed 19 May 2023.

Poulsen, K., McMillan, R., & Evans, M. (2021). A hospital hit by hackers, a baby in distress: The case of the first alleged ransomware death. *Wall Street Journal*. https://www.wsj.com/articles/ransomware-hackers-hospital-first-alleged-death-11633008116. Accessed 19 May 2023.

Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2016). Data discretization: Taxonomy and big data challenge. *Wiley Interdisciplinary Reviews: Data Mining & Knowledge Discovery, 6*(1), 5–21. https://doi.org/10.1002/widm.1173

Randolph, K., & Hunt, M. (2021). Security incident report. *Verkada*. https://docs.verkada.com/docs/Security_Incident_Report_Version1.2.pdf. Accessed 19 May 2023.

Ross, C., & Swetlitz, I. (2017). IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. *STAT*. https://www.statnews.com/2017/09/05/watson-ibm-cancer/. Accessed 19 May 2023.

Sandhu, T. H. (2018). Machine learning and natural language processing: A review. *International Journal of Advanced Research in Computer Science, 9*(2), 582–584. https://doi.org/10.26483/IJARCS.V9I2.5799

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications, 30*(2), 169–190. https://doi.org/10.3233/AIC-170729

Yang, Y., Webb, G. I., & Wu, X. (2009). Discretization methods. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 101–116). Springer.

Yang, Y., Yuan, Y., Zhang, G., Wang, H., Chen, Y., Liu, Y., Tarolli, C. G., Crepeau, D., Bukartyk, J., Junna, M. R., Videnovic, A., Ellis, T. D., Lipford, M. C., Dorsey, R., & Katabi, D. (2022). Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Natural Medicine (online First)*. https://doi.org/10.1038/s41591-022-01932-x

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. In J. Cabestany, A. Prieto, & F. Sandoval (eds) *Computational Intelligence and Bioinspired Systems. IWANN 2005. Lecture Notes in Computer Science* (Vol. 3512, pp. 758–770). Berlin, Heidelberg: Springer. https://doi.org/10.1007/11494669_93.

Vollmer. T. R., Sloman, K. N., & St. Peter Pipkin, C. (2008). Practical implications of data reliability and treatment integrity monitoring. *Behavior Analysis in Practice, 1*(2), 4–11.https://doi.org/10.1007/BF03391722.

Xiong, H., Pandey, G., Steinbach, M., & Kumar, V. (2006). Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge & Data Engineering, 18*(3), 304–319. https://doi.org/10.1109/TKDE.2006.46

Zhang, J., & Gao, R. X. (2021). Deep learning-driven data curation and model interpretation for smart manufacturing. *Chinese Journal of Mechanical Engineering, 34*, 71–92. https://doi.org/10.1186/s10033-021-00587-y

Zhang, S. (2008). Parimputation: From imputation and null-imputation to partially imputation. *IEEE Intelligent Informatics Bulletin*, *9*(1), 32–38. http://www.comp.hkbu.edu.hk/~iib/2008/IIB08Nov/feature_article_4/TRANS-JOUR-parimputation_finish_.pdf.

Zhou, X., Feliciano, P., Shu, C., Wang, T., Astrovskaya, I., Hall, J. B., Obiajulu, J. U., Wright, J. R., Murali, S. C., Xu, S. X, Brueggeman, L., Thomas, T. R., Marchenko, O., Fleisch, C., Barns, S. D., Snyder, L., G., Han, B., Chang, T. S., Turner, T. T., . . . & Chung, W. K. (2022). Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nature Genetics*, *54*, 1305–1319. https://www.nature.com/articles/s41588-022-01148-2#citeas.

Zikopoulos, P., Deroos, D., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D. (2012). *Harness the power of big data The IBM big data platform*. McGraw-Hill Professional.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.