SEPTEMBER 2024

Defining Big Data

American Academy of Actuaries
Data Science and Analytics Committee

American Academy of Actuaries
actuary.org

American Academy of Actuaries
Data Science and Analytics Committee

Dorothy Andrews, MAAA, ASA—*Chairperson*

Mary Pat Campbell, MAAA, FSA

Reese Mularz, MAAA, FCAS

Belinda Nguyen, MAAA, FCAS

Dave Sandberg, MAAA, FSA

The American Academy of Actuaries is a 20,000-member professional association whose mission is to serve the public and the U.S. actuarial profession. For more than 50 years, the Academy has assisted public policymakers on all levels by providing leadership, objective expertise, and actuarial advice on risk and financial security issues. The Academy also sets qualification, practice, and professionalism standards for actuaries in the United States.

AMERICAN ACADEMY of ACTUARIES

AMERICAN ACADEMY OF ACTUARIES
1850 M STREET NW, SUITE 300, WASHINGTON, D.C.
20036202-223-8196 | WWW.ACTUARY.ORG

September 2024

Any references to current laws, regulations, or practice guidelines are correct as of the date of publication.

Defining Big Data

Contents

Introduction 2

What is big data? 3

The effects of changing technology 6

Lifecycle of data in analytics 9

Data transformation 15

Actuarial standards of practice and the responsibilities of the actuary surrounding big data 19

In summary: Know your data 23

Further work from DSAC 24

Additional resources 25

Introduction

"Big data" as a concept continues to grow in popularity. While aspects of this concept have existed in insurance and actuarial work for decades, especially in property and casualty personal lines of insurance, issues surrounding big data have extended to all major areas of the insurance industry.

Given the ubiquity of data-generation devices that people use and carry on their persons, such as smartphones or wearable biometric devices, actuaries and data scientists have seen opportunities for these new sources of information to be used in underwriting, pricing, risk modeling, valuation, fraud detection, claims adjustment, marketing, and much more. There is concern among regulators that these data sources may be improperly used or cause unforeseen problems for consumers and the financial stability of insurers and market. Consumers are concerned with how insurers may use their data and what decisions are made based on their data.

To address the complexities of these issues, there is a need for a common terminology and a common foundation of understanding of what big data is, what some of the key characteristics are, how the understanding of big data has changed, and what may change in the future. The ensuing discussion is intended to build a shared language and concepts for the actuarial community, regulators, and other interested parties. Additionally, the use of term "data" in the actuarial standards of practice (ASOPs) will be identified along with the challenges presented by its usage.

What is big data?

Defining data

The definition of "data" has changed over the years. For actuaries, the definition of "data" has been rooted in the ASOPs as practice has evolved in the industry. To provide an example of that evolution, listed below is a sampling of how the definition has evolved in ASOPs, starting in 1991. All of the following will note whether the ASOP has since been revised or is in effect as of March 2024.

ASOP No. 17, *Expert Testimony by Actuaries*, Doc No. 029, adopted by the Actuarial Standards Board (ASB) July 1991 [since revised]

> 2.6. Data—Statistical or other information that is generally numerical in nature or susceptible to quantification.

ASOP No. 23, *Data Quality*, Doc No. 044, adopted by the ASB July 1993 [since revised]

> 2.3 Data—For purposes of this standard, the term refers to numerical, census, or classification information and not to general or qualitative information. Assumptions are not data, but data are commonly used in the development of actuarial assumptions.

ASOP No. 23, *Data Quality*, Doc No. 185, adopted by the ASB December 2016 [in effect as of March 2024]

> 2.3 Data—Numerical, census, or classification information, or information derived mathematically from such items, but not general or qualitative information. Assumptions are not data, but data are commonly used in the development of actuarial assumptions.

ASOP No. 56, *Modeling*, Doc No. 195, adopted by the ASB December 2019 [in effect as of March 2024]

> 2.2 Data—Facts or information that are either direct input to a model or inform the selection of an input. Data may be collected from sources such as records, experience, experiments, surveys, observations, benefit plans or policy provisions, or output from other models.

As of March 2024, there are four ASOPs where the term “data” is defined. Interestingly, these four ASOPs are a few general ASOPs that address all actuarial areas of practice. This table shows how the term “data” is used differently over current ASOPs.

Currently Active ASOP (as of March 2024)	**Date Adopted by ASB**	**Definition of “data”**
ASOP No. 17, *Expert Testimony by Actuaries*	June 2018	Numerical, census, or classification information, or information derived mathematically from such items, but not general or qualitative information. Actuarial assumptions are not data, but data are commonly used in the development of actuarial assumptions.
ASOP No. 23, *Data Quality*	December 2016	Numerical, census, or classification information, or information derived mathematically from such items, but not general or qualitative information. Assumptions are not data, but data are commonly used in the development of actuarial assumptions
ASOP No. 38, *Catastrophe Modeling*	July 2021	Facts or information that are either direct input to a catastrophe model or inform the selection of input. Data may be collected from sources such as records, experience, experiments, surveys, observations, benefit plan or policy provisions, or output from other models.
ASOP No. 56, *Modeling*	December 2019	Facts or information that are either direct input to a model or inform the selection of input. Data may be collected from sources such as records, experience, experiments, surveys, observations, benefit plan or policy provisions, or output from other models.

In the most recent revision for ASOP No. 17, its definition for data had been updated to be in line with the definition of data for ASOP No. 23, and it is likely the next cycle of revisions will find these definitions once again updated.

When it comes to standards of practice and regulation, by their very nature they will lag the development of the technologies and the approaches being developed and used in practice. Some approaches and practices will fail to persist and never make it to the stage of being codified in an ASOP.

With the evolution of new data sources and technology, the conception of data has changed. At its heart, data is information, and from the actuarial and business perspective, the challenge is to extract usable information from it. “Usable” in this context is defined as influencing decision-making in some way, usually through the means of model input and/or assumption-setting. Typically, data is quantitative in nature, but it can be qualitative as well. Qualitative data is more often transformed into quantitative information for reference or further analysis. Within the context of data and analytics, some concepts can be very technical and esoteric, therefore only a sampling of knowledge around the topic of big data is discussed. As practice evolves, standards catch up in defining the term “data.” For example, earlier ASOPs exclude types of information from the definition of “data” that are typically considered data now, because they were qualitative and not quantitative.

What makes data big?

Given that the relatively objective concept of “data” has evolved with evolving technology, it is unsurprising that the subjective term “big” is even more difficult to define.

On the surface, the term “big data” suggests expansive datasets measured in petabytes of storage, but the “big” in “big data” is relative to the technology and resources to manage, process, and refine the data. The defining characteristics of big data were first coined by Doug Laney of META Group in 2001, where he described the three dimensions of big data: volume, velocity, and variety.[1] The definition has since expanded with the expanding big data capabilities and some sources cite as many as 42 “V’s” of big data.[2]

As noted in prior Academy works, the definition of big data generally includes the “5 V’s”:

Volume	Large amounts of data are collected and require processing.
Velocity	Data is available and must be processed at lightning speed, frequently instantaneously.
Variety	The data being used comes in different forms.
Veracity	The reliability of the data is not uniform.
Value	The data being extracted must be usable or be able to be monetized.[3]

Due to increased capacity for processing data, along with algorithms developed to make connections and derive patterns from data of disparate sources, there has been an unleashing of automated decision-making at a scale not possible before.

The importance of these five dimensions, in particular, is the unwieldiness in handling and extracting meaning from the data. With other industries, the focus is primarily just being able to collect and use the data for profitable means, but with insurance, the highly regulated aspects of the industry come into play as well. Having to consider multiple stakeholders is what makes big data more unwieldy for insurers compared to many other industries.

[1] “A Very Short History of Big Data”; *Forbes*; May 9, 2013. Accessed September 27, 2022.
[2] *Big Data and Algorithms in Actuarial Modeling and Consumer Impacts*; American Academy of Actuaries; Data Science and Analytics Committee; 2021.
[3] *Big Data and the Role of the Actuary*; American Academy of Actuaries; Big Data Task Force; 2018.

The effects of changing technology

When considering the exponential growth of data and analytics, it is helpful to consider the developments that made big data possible.

In the 1960s, the concept of a database—a system designed for storing and organizing data—emerged and underwent rapid transformations in the ensuing decades. By the 1970s and 1980s, the relational database model had emerged as the predominant data model, leveraging set theory and predicate logic to enable efficient data management. In the 1980s, the computer-buying boom advanced the commercialization of database management systems. The advance of the relational database model brought forth a new focus on data management, namely data structure and integrity. It was in the 1980s that the concept of data warehousing began to take form to help tame the first V of big data, volume, and emphasize the V of veracity. A data warehouse is a system that aggregates data from different sources into a single, central, consistent data store to support data analysis.[4,5]

The 1990s were marked by the appearance of the World Wide Web, making remote access to computer systems and legacy data possible. It was also in the '90s that personal productivity tools such as Microsoft Excel and Access were introduced. It was in the early 2000s that people began to understand the volume and velocity of data being generated by online services, and the complexity of managing it. In 2003, Jeff Dean and Sanjay Ghemawat, two developers at Google, published a paper on the Google File System, a scalable distributed file system that ran on inexpensive commodity hardware and was highly tolerant to component failures.[6,7]

For background, two features of the paper—discussion of the distributed file system and distributed computing across computing clusters—are key to understanding the expansion of data capabilities. In the years preceding the paper, Google had amassed more data than any supercomputer could handle. To overcome that challenge, they bought consumer computer hardware and built code to unite the components into a single system. This system consisted of a master computer or master node managing the replication and placement of data across the other nodes. By replicating and distributing data and workloads across nodes, the system allowed for higher computing efficiency and tolerance for component failures and data corruption.[8,9]

[4] "History of Databases"; *International Journal of Management & Information Systems*; December 31, 2012.
[5] "Data Warehouse"; IBM; March 2020. Accessed September 27, 2022.
[6] "History of Databases"; op. cit.
[7] "The Google File System"; *Proceedings of the nineteenth ACM symposium on Operating systems principles*; October 2003; pgs 29-43.
[8] Ibid.
[9] "The Friendship That Made Google Huge"; *The New Yorker*; December 3, 2018.

The 2003 Google File System paper was highly influential in the development of Apache Hadoop, the open-source software framework that accelerated the expansion of big data.[10] The popularity of open-source software, or software that is open to the public to use for any purpose without cost, made big data management and analytics cheaper and more accessible. This structure allowed for increased experimentation and contributions from the developer communities and ultimately broader application. Two other well-known open-source software projects that enhanced the value of data and big data are Python, a general-purpose programming language, and R, a statistical programming language.

As technology evolved, streaming data to and from personal devices became the norm and the coming online of consumer devices has become the internet of things (IoT). Storage capacity no longer was the primary cost driver for computing systems, especially as cloud storage became more commonplace. Raw, unstructured data was being generated far quicker than humans could curate or even review to extract additional value and meaning. Data was stored in unstructured repositories called data lakes rather than the very structured data warehouses previously mentioned. With the variety of data generated and applications being developed, developer time was now the primary cost factor.

NoSQL databases gained in popularity the late 2000s for their ability to manage unstructured data while allowing applications to quickly store and query documents, files, and other objects. The term "NoSQL" does not mean that SQL, the common relational database query language, is not used. More accurately, NoSQL is described as "not only SQL" relating to the fact that a NoSQL system can handle more than the structured data found in a relation database. Though not generally used for predictive analytics, an actuary may encounter this type of database in their data preprocessing, modernization, and process automation practices.[11,12,13]

Terms related to data technology

Database management system

A database management system (DBMS) is the software used to organize, support and maintain the information or data in a structure stored in a computer.[14]

[10] "What is Google BigQuery"; *Google BigQuery: A Definitive Guide*; 2019.
[11] "What is NoSQL?"; MongoDB. Accessed September 26, 2022.
[12] "Data Lake, Big Data, NoSQL—The Good, The Bad and The Ugly"; Gartner; July 29, 2017.
[13] "SQL vs NoSQL"; IBM; June 12, 2022.
[14] "Data Lake, Big Data, NoSQL—The Good, The Bad and The Ugly"; op. cit.

Relational database

A type of database that stores and provides access to data points that are related to one another, using a relational data model that represents data in tables.[15]

Data warehouse

A system that aggregates data from different sources into a single, central, consistent data store.[16]

Data lake

A collection of storage instances of various data assets including both curated and uncurated/raw data.[17]

SQL

Short for structured query language, SQL is a programming language used in managing data in relational database management systems.[18]

NoSQL database

A non-relational database. A database that allows different structures other than a typical row-and-column-based relational database, allowing more flexibility to use a format that best fits the data.[19]

Distributed file system

A distributed file system (DFS) is a file system where files (data) are divided, and in most cases replicated, across multiple locations.

Distributed computing

A process where computing workloads are grouped into smaller processes and spread across multiple locations.

Hadoop

An open-source software framework developed for reliable, scalable, distributed computing. One of the major frameworks in big data management.[20]

[15] Ibid.
[16] "Data Warehouse"; IBM; March 5, 2020.
[17] "Data Lake, Big Data, NoSQL—The Good, The Bad and The Ugly"; op. cit.
[18] "SQL vs NoSQL"; op. cit.
[19] Ibid.
[20] "Hadoop"; Apache. Accessed September 26, 2022.

Open-source software

Computer software that is released under a license in which the copyright holder grants users the rights to use, study, change, and distribute the software and its source code to anyone and for any purpose.[21]

Internet of things (IoT)

The network of physical objects that are embedded with technology for the purpose of connecting and exchanging data with other devices and systems over the internet.[22]

Lifecycle of data in analytics

The analysis of data is the primary interest of actuaries, and as such it is helpful to provide a framework for the analysis process which extracts valued business insights from the data. To generalize, let's consider the following stages: Data Generation & Collection, Exploratory Data Analysis, Assembly of the Modeling Dataset, Modeling, Model Implementation, and Post-Implementation Monitoring.

Data Generation & Collection

1. Today, data can be passively generated, in that a human need not make any additional actions for the data to be recorded. An example of this type of data would be a thermometer measuring air temperature in a given location.
2. Data can also be actively generated by humans, in that a human being needs to perform an action for the data to be recorded. An example of this would be the recording of someone's death in a death certificate, which involves a medical examiner recording details of the death, including cause of death. This involves human judgment. Other examples like this include data recorded in audio or visual media.
3. There is also data generated in an in-between state, such as a person wearing a smartwatch, and that smartwatch recording the number of steps taken by the person, and then that data shared with a health app. Other examples would be information on websites visited or aggregate purchasing information.
4. Finally, all the above generation methods can be collected from both private and public sources into a "big data" set to use as the basis for one's analysis.

[21] "The Free Software Alternative: Freeware, Open Source Software, and Libraries"; *Information Technology and Libraries*; September 25, 2014.
[22] "Internet of Things"; Oracle. Accessed Sept 26, 2022.

Exploratory Data Analysis (EDA)

The pre-model building step that investigates patterns and deficiencies in the modeling data that need to be addressed. Exploratory data analysis assesses aspects of modeling data such as missingness, relationships among variables, imbalance among attributes, and other anomalous patterns.

Assembly of the Modeling Dataset

The focus of this paper is on the modeling dataset assembly stage, where new data can be generated and existing data manipulated, while the other stages in the analytics process are left for future publications. The following sections will examine some of the data transformation techniques used to make data fit for use in modeling.

Terms based on sources of data

The following terms will be helpful to understand how the various methods of data generation/collection, exploration, and dataset assembly may impact data and model governance.

Internal data

Data that a company generates or collects from its own operations or activities. This can include data from various sources such as customer transactions, financial reports, and operational metrics.

External or third-party data

Data sourced from outside of the company. It can either be free or bought. Common examples are information about population demographics, the industry market conditions, and the general economy. It could also include census data and motor vehicle record data on drivers. It may also be the result of a third-party aggregation of internal data supplied by diverse organizations or individuals.

Public/private/proprietary data

Public data is readily available to the public, while private data is information that is intended to be kept restricted to certain parties. Proprietary data is owned by an individual or organization.

For example, census data collected by the government is public data within the economy. Within an organization, an example of private data may be protected health information which, if released, can be used to identify an individual. Sometimes, there is also an overlap between public and private data. For example, aggregated enterprise data such as experience data may be made public.

Personally identifiable information (PII)

Any representation of information that permits the identity of an individual to whom the information applies be reasonably inferred by either direct or indirect means.[23]

Government-generated data

Data collected and curated by the government. Government data is an example of publicly available data, though not all government data is made public. One source for this information is data.gov, where data is available on a wide array of topics such as geodemography, agriculture, climate, and energy.

Terms based on qualities of the data

Structured data

Data that conforms to a tabular format with relationships between the rows, columns, and tables. It is easy to access with programming code because of this organization. Relational database tables and tabular data in flat files are common examples of structured data.

Actuaries and data scientists most commonly use structured data in their work. This is typically found in data tables of policy and claims information such as losses, age, construction type, etc.

Semi-structured data

Data that has some organizing properties, making it easier to parse and analyze, but is not yet fully specified like data in a relational database. Specifically, semi-structured data contains internal tags and markings that allow for grouping and hierarchies. An example of this can be an email where sections are separately tagged but the content within the sections is still unstructured text.[24]

[23] "Guidance on the Protection of Personal Identifiable Information"; Department of Labor. Accessed September 26, 2022.
[24] "Unstructured Data"; MongoDB. Accessed September 26, 2022.

Unstructured data

Data that lacks the row, column, and table structure of structured data. It includes data stored as text, in images, recordings or video formats. Extracting meaning and reliable information from unstructured data requires specialized processing packages.[25]

Transactional data

Data that supports the daily operations or business events of an organization and are included in the application systems that automate a company's key business processes. For example, data recorded by the policy system when an endorsement is added to a policy is transactional data. Future sources of usable transactional data can include the written and pictorial documentation used to decide underwriting and claim adjustment transactions.

Analytical data

Data that supports decision-making, reporting, and analysis. This data is more often curated data whose raw form was taken from a transactional data source and structured for business purposes including business intelligence, financial reporting, research, and analytics.

Quantitative data

Data that is in the form of a numeric value. Examples include, but are not limited to, age of an insured, policy premium, and automobile value.

Categorical data

Data that describes a quality or characteristic that is not in the form of a numeric value. Examples include, but are not limited to, construction type, roof type, and the state an insured lives in.

Terms based on how data are used

Data governance

The process of managing the availability, usability, integrity, and security of the data in company operations and systems, based on internal data standards and policies that also control data usage.

[25] Ibid.

Data dictionaries

A data dictionary contains the names, definitions, and characteristics of data elements in a database. It aids in the interpretation of data fields and how to appropriately use them.[26]

Metadata

Data that describes aspects of the data and data artifacts. Examples include data describing a data field's data type, description of the data table, size of the dataset, etc.

Business intelligence

Business intelligence (BI) comprises the strategies and technologies used by enterprises for the data analysis and management of business information.

Data visualization

An interdisciplinary field that deals with the graphic representation of data and information. It is a particularly efficient way of communicating when the data or information is numerous as, for example, a time series.

Mathematical statistics

A branch of mathematics that applies probability theory to data statistics.

Machine learning

The use and development of computer systems that can learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data. Machine learning is generally considered a subset of artificial intelligence.

Artificial intelligence

The theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

Feature engineering

The process that takes raw data and transforms it into features that can be used to create a predictive model using machine learning or statistical modeling. This is discussed further in the section below.

[26] "What Is a Data Dictionary?"; University of California Merced Library. Accessed February 13, 2023.

Imputation

The process of assigning values to missing data. This is discussed further in the section below.

Cross-validation

A method to test the validity of a predictive model by which the modeling dataset is divided into two or more subsets, with some subsets used to train the model and the remaining subsets used to test or validate the model. There are generally three types of data subsets: train, validation, and test. Test subsets are often called hold-out data. These subsets are discussed below.

Training data

A dataset that is used to build a model and calculate the model form and parameter estimates. This dataset will commonly be 60%-75% of the total amount of modeling dataset.

Validation data set

A dataset that is used to tune the model. The model built on the training data used to predict values in the validation data. The model is iteratively tuned until prediction errors on the validation data are within an acceptable tolerance.

Test/hold-out dataset

A dataset that is used to evaluate a model. Hold-out data is not used directly to train the model. This means the model should not have seen the data, with the goal of evaluating how well the model generalizes to the unseen data.

Sampled data

In general, it is not possible to gather statistics and data on an entire population. Therefore, data used to train a model is collected on a subset or sample of the population. Data can also be sampled on a sample dataset, which is common when dealing with big data. Model performance is usually assessed on data that is randomly sampled (most common), stratified, or grouped by like-kind and sampled for improved homogeneity.

Another type of sampling is bootstrapping. If the data is too sparse to analyze, this process involves sampling with replacement from the original data until a large enough sample is obtained. In essence, the data is pulling up by its own "bootstraps" to improve model performance.

Data transformation

Data generally needs to be transformed or processed in some way to be made useful for modeling and analysis. While discussion of how to transform data is beyond the scope of this paper, some of the terms commonly used surrounding data transformation, especially in use with big data will be identified.

Feature engineering

Feature engineering involves creating new features out of existing features. For example, body mass index (BMI) is a feature engineered variable. It is the ratio of weight in kilograms divided by height in meters squared.

It is possible to transform data beyond its recognition, and this may pose problems when consumers get an insurance-related result (such as a denial of claim) that they want to dispute. The consumers' original data may have been transformed in such a way that the consumers will not understand the decision made or know how to dispute it. There are good reasons to use feature engineering, but it is important to make sure that more advantages than disadvantages are created with its use.

The advantages include imputing values when data is missing, preserving the variance structure on variables when the imputation method is random sampling imputation, identifying data missingness, and negating the effect of outliers.

Feature engineering can get very complicated, and it only works if there is good knowledge of the domain in which new variables are created. This can put the average consumer at a disadvantage in recognizing errors. Feature engineering has some other disadvantages as well. It can distort data metrics, such as means and correlations, and it can change the shape of the distribution of the data, leading to overrepresentation in some regions of the distribution. The predictive power of original variables may also be changed in some way and not necessarily for the better. A summary of several pros and cons is given below.

Features of Feature Engineering on Big Data	
Pros	***Cons***
Easy to implement to obtain complete datasets	Requires good domain knowledge to be effective and reliable
Preserves variance of original variable	Can be difficult to recognize data errors in transformed data
Captures the importance of missing data	Distortion of data metrics (e.g., ρ) and distribution
Nullifies the negative effects of outliers	May mask or create outliers
Improves model efficiency and signal detection	Loss of interpretation of engineered variables
Creates more meaningful metrics from raw data	May lead to overrepresentation skewing distribution
	May mask predictive power of original variable
	May introduce bias to data if features are engineered based on preconceptions about the data

It is important to test the effects of feature engineering on the model. Just as important, companies need to be prepared to disclose and explain feature-engineered variables, as explainability and transparency are becoming important issues to regulators.

Feature engineering approaches

Disclosure is important when using feature engineering. Provide clear explanations of how variables were feature engineered, numerical examples, and theoretical support that allows for replication of feature-engineered variables and rationales for why they are needed. Additional best practices include:

- Clearly identify all variables subject to feature engineering.
- Provide business and risk-related rationales to support.
- Provide the methodology and formulas to support all feature-engineered variables.
- Provide academic references to support their use.
- Include clear examples that show the calculation. The example should be detailed enough so that an independent reviewer can replicate the values for other observation in the dataset.

For insurance modeling purposes, there are some sound practices when using feature engineering. The first is to clearly identify all feature-engineered variables and provide the methodology and examples that show how the feature was derived. The calculation should be easy to replicate, although the calculation may be complicated. Second, it is important to provide business and risk-related rationales. A meaningful rationale relates the variable to the risk. Third, providing academic support is a good approach to supporting the use of the variable. It gives the reviewer research to assess in evaluating the feature.

Imputation methods

There are several approaches to handling missing data with imputation methods. The methods will always impose a distribution shape on your data that may not mirror the true shape if the data were known. Unfortunately, machine learning methods will not work with missing data. In many cases, simply removing data points where there are missing dimensions will not be acceptable, so imputing a value to the missing dimensions will be needed. It is important to disclose how missing data was handled and discuss that the imputation methods chosen may have biased model results.

The two categories of data that may be subject to imputation methods are continuous and categorical. With continuous data, numerical imputation methods include:

- **Arbitrary value imputation—**The method assigns an arbitrary number to the missing data determined at the discretion of the modeler.[27] While this method is simple to implement, it can seriously distort simple distribution metrics such as the mean, median, mode, and variance.
- **Start/end of distribution imputation—**The method uses the extreme of the distribution at plus/minus 3 standard deviations to impute missing data.[28] This method also has the unsatisfactory result of distorting distribution metrics, model predictive power, and the true effect of outliers.
- **Mean/median/mode imputation—**These approaches are more regularly used but each is preferrable in certain situations. The mean is preferred if the data is not skewed, the median is preferred if the data is skewed, and the data is categorical.[29] The disadvantages of these approaches are similar to the previously discussed approaches.
- **KNN Imputation—**The approach is more sophisticated the previous ones in that it uses the K-Nearest Neighbors (KNN) algorithm to impute missing values based on known nearest to data points.[30] The advantage of this method is that it can preserve the original structure of the data, minimizing distortions to variable distributions.[31] This method requires tuning of its k parameters and can be sensitive to outliers.

[27] ArbitraryNumberImputer, from Feature-engine: A Python Library for Feature Engineering for Machine Learning. Accessed September 26, 2022.
[28] "Feature Engineering: Handling Missing Data"; UDig. Accessed September 26, 2022.
[29] "Handling missing data: Mean, Median, Mode"; Naukri Learning; April 30, 2022.
[30] "Comparison of five imputation methods in handling missing data in a continuous frequency table"; *AIP Conference Proceedings*; 2021.
[31] "Nearest neighbor imputation algorithms: a critical evaluation"; *BMC medical informatics and decision making*; 2016.

- **MissForest imputation—**This is a nonparametric imputation method based on the Random Forest technique and can be applied to any type of data. The method is robust in the presence of noisy data and multicollinearity, but it does not work well with small datasets.[32]

For categorical data, imputation approaches include:

- **Mode imputation**—This is the most basic of imputation methods with the mode of the non-missing values imputed to the missing values.[33] As discussed above, this approach can distort the distribution on the variable and additional analysis should be engaged to determine the extent of the distortion.
- **Random sampling imputation—**As its name implies, random observations are drawn from non-missing data to impute missing data. This method has the advantage of preserving the distribution shape of variables and minimizing distortion of distribution metrics. It can also be used with continuous data.[34] A disadvantage is that randomly assigned values may be incoherent with other attributes in the data record.
- **Hot deck imputation**—This method is similar to KNN imputation. Missing values are imputed using data points that are similar in non-missing characteristics.[35] This approach is subject to modeler bias.

These are just a sampling of the approaches that can be used to impute missing data. There are other approaches.

Regulators are starting to focus their attention on why data is missing, and on what insurers are doing to secure the values for missing data rather than impute them. Imputation methods introduce noise into calculations, and that noise may have unintended consequences, including harm to insureds.

[32] "MissForest: The Best Missing Data Imputation Algorithm?"; *Towards Data Science*; August 31, 2020.
[33] "The ability of different imputation methods for missing values in mental measurement questionnaires"; *BMC Medical Research Methodology*; 2020.
[34] "Efficient random imputation for missing data in complex surveys"; *Statistica Sinica*; 2000.
[35] "A Review of Hot Deck Imputation for Survey Non-response"; *International statistical review*; 2010.

Actuarial standards of practice and the responsibilities of the actuary surrounding big data

ASOPs as promulgated by the ASB identify what the actuary should consider, document, and disclose when performing an actuarial assignment. Actuaries in professional practice may also handle new or nonroutine situations not anticipated by the ASOPs. As previously noted, the codification of ASOPs will necessarily lag the cutting edge as people experiment to prove the science—in this case data science and big data.

In all situations, according to the Code of Professional Conduct, Precept 3, Annotation 3-2, "where a question arises with regard to the applicability of a standard of practice, or where no applicable standard exists, an Actuary shall utilize professional judgment, taking into account generally accepted actuarial principles and practices." The Code of Professional Conduct sets forth what it means for an actuary to act as a professional. Technical and ethical guidelines such as the Code and the standards of practice uphold the integrity of the profession and professional responsibility to the public. When it comes to developing areas, professional judgment is brought to light and actuaries must use their professional judgment to consider not only the written word but the spirit of the codes and standards.

Below we will consider major ASOPs that apply to all actuarial practice areas and how these may be adapted to situations where big data come into play.

ASOP No. 41, *Actuarial Communications*

In the realm of actuarial data and analytics, teams with interdisciplinary backgrounds, incorporating a variety of skills and knowledge, are generally more effective than those without such diversity. Actuaries today are no longer simply technical experts in risk management but business professionals working in a variety of industries inside and outside the insurance industry. A hallmark of strong management and leadership is effective communication.

The definition of an actuarial communication is broad and encompasses any written, electronic, or oral communication issued by an actuary with respect to actuarial services, while an actuarial document includes any actuarial communication in any recorded form. An actuary working in an analytics function may write technical specifications or complex code. Do these constitute an actuarial communication and actuarial documents? Would they be considered actuarial communications per ASOP No. 41? Is this the intent of the ASOP with the changing practices to which actuaries engage? These questions are not answered here but instead are posed to the actuarial community.

Reproducible research has become a norm in academia and industry, and the methods to capture the journey have become more sophisticated with the growth of technology. One such tool is Jupyter Notebooks, which can execute code in line with explanatory text, output results, and render the entire document in a business artifact such as a PDF or HTML ready for posting to the web. To sustain the growth of human knowledge, personal productivity is advancing as data technology advances.

ASOP No. 41 applies to any actuary providing an actuarial communication in any practice area, and an actuary who makes an actuarial communication assumes the responsibility for it. The one exception is when the actuary making the communication disclaims that responsibility by stating reliance on other sources. Disclosures to be made in actuarial reports and communications as specified in ASOP No. 41 are with the intended users of these communications in mind. In subsection 3.7, ASOP No. 41 warns that actuarial documents may be misused:

> "The actuary should recognize the risks of misquotation, misinterpretation, or other misuse of such a document and should take reasonable steps to ensure that the actuarial document is clear and presented fairly."

Given the nature of some of the third-party data sources, the changing complexity of actuarial assignments, and the need for actuarial judgment with evolving practices, actuaries may need to be sensitive to potential misuses of their reports in communication, especially considering issues of data reliability. ASOP No. 41 provides guidance for disclosing the scope of the requested work; the methods, procedures, assumptions, data, and other information required to complete the work; and the development of the communication of the actuarial findings.

ASOP No. 41 refers to data a few times, cross-referencing ASOP No. 23, *Data Quality* (discussed below), in one part. But one portion in particular is section 3.2 on the Actuarial Report, which is the method by which the actuary is to communicate their actuarial findings to any intended users. As noted in ASOP No. 41:

> "An actuarial report may comprise one or several documents. The report may be in several different formats (such as formal documents produced on word processing, presentation or publishing software, e-mail, paper, or web sites). Where an actuarial report for a specific intended user comprises multiple documents, the actuary should communicate which documents comprise the report."

As mentioned above, that actuarial report may not be in what we consider a traditional document (such as what you are currently reading). It may include code, such as with a Jupyter notebook.

One key feature of the report is identifying the data by which the actuarial findings were made, "with sufficient clarity that another actuary qualified in the same practice area could make an objective appraisal of the reasonableness of the actuary's work as presented in the actuarial report."

With respect to big data, documenting the data well enough and communicating the findings so that other actuaries may review the results may be a challenge. Aspects of the 5 V's of big data can make the data on which models are built a moving target.

In ASOP No. 41, subsection 3.4.3, Reliance on Other Sources for Data and Other Information, further reference to data is made, with a cross-reference to ASOP No. 23. A discussion of ASOP No. 23 is contained below.

ASOP 23, *Data Quality*

The goal of the actuary in arguably all actuarial assignments is to objectively extract information from data. Although not addressing big data specifically, we can refer to ASOP No. 23, *Data Quality*, to provide guidance when working with data. In essence, big data is simply data and what is considered big by today's computing standards may not be in the future. One key element of ASOP No. 23 is the determination of appropriateness of the data for the intended use.

ASOP No. 23 provides guidance to actuaries when selecting data, performing a review of data, using data, relying on data supplied by others, preparing data, or making disclosures regarding data quality, in performing actuarial services. If the actuary is preparing the data for analysis, they must prepare it as if they were to use it for the intended purpose.

Actuaries are expected to select appropriate data, make certain reasonability checks on their data where practical, and disclose if and why no reasonability check was made. Any deviations from the any standards of practice must be disclosed—ASOP No. 23 or any other ASOP related to the assignment.

The actuary is expected to disclose the sources of their data and make a reasonable attempt to understand the fields used in the model. In the rapid experimentation environment of today, much of the review and the selection of data may be automated. Experimentation can be a messy process, but once findings are conveyed, the materiality of the conclusions dictates what must be disclosed. Peer review and socialization of the work product provides natural quality control if done appropriately. While the inputs into a training dataset may not be reviewed in detail at the onset of a modeling project, the final analysis will have a finite number of inputs or some framework to understand variables used and the relationships to the predicted output.

In the end, actuarial assignments require the actuary to understand the intended use of the actuarial work product and the relationship of the data to the results. In the absence of the technical knowledge for aspects of an assignment, disclosure is the best practice and consideration for existing recommended actuarial practices is needed.

ASOP No. 56, *Modeling*

Predictive modeling and advanced analytics go hand in hand with big data, giving relevance to ASOP No. 56, *Modeling*. This standard applies to actuaries in any practice area when performing actuarial services with respect to designing, developing, selecting, modifying, or using all types of models.

ASOP No. 56 (and ASOP No. 38, *Catastrophe Modeling*, which is similar) refers back to ASOP No. 23 with respect to data, and it is always in context of the use of data with a model, and specifically to make sure data are appropriate for the intended purpose of the model.

Data is mentioned in several places in this ASOP—most notably in subsection 3.1.5, Data, in which it is explicitly stated that "[t]he actuary should use, or confirm use of, data appropriate for the model's intended purpose" and continues in that vein. That is for data in general.

Items that could be of importance in reference to big data, however, are found in subsection 3.2, Understanding the Model, in subpart c—the actuary should understand the limitations of the data. Recall the "5 V's" of big data, and in particular the V's that may make big data suspect: the veracity and value in big data specifically. Just because there is a lot (volume) of data coming out of these systems quickly (velocity) with lots of distinct aspects (variety) does not mean it is giving us anything useful for our models (value) or anything we can actually rely on (veracity).

Two other areas where data are important in this ASOP are subsection 3.6.1, Model Testing, and subsection 3.6.2, Model Output Validation, the first of which involves testing models as data change, which is important with big data. Given the high velocity of big data, significant changes can occur in a short period of time, leading to the question of whether one can recalculate quickly enough to accommodate a testing cycle. For output validation, this involves "hold-out data," or data not used to calibrate or fit the model, and then one checks whether the model performs well on the "hold-out data." One of the benefits of big data is having sufficient volume to perform such tests. There are approaches, such as k-fold cross-validation, where there are overlapping training and testing sets to see how sensitive the model structure is to the data. This is one of many techniques that could be used for this sort of testing; the point is such testing by the actuary of the model against the data is expected.

In summary: Know your data

As noted in ASOP No. 23, the actuary is not necessarily expected to do a detail-level audit of the data. Specific techniques are not prescribed in ASOP No. 56 for testing models with data. However, the general expectation is this:

- Know your data
    - Where did it come from?
    - What are its limitations?
    - Is it appropriate to use for your purposes?
- Use your data appropriately
    - Document your data sources
    - Clean up the data so that it is fit for use
    - Document changes and data limitations
    - Test your model (output validation, when data changes)
- Disclose to intended users of actuarial services
    - Disclose to users your sources of information
    - State your reliance on third-party sources
    - Explain what level of reasonability checks have been performed on the data
    - Be aware of potential misuses of reports

Further work from DSAC

The American Academy of Actuaries Data Science and Analytics Committee is producing a series of papers for the actuarial community, regulators, and other interested parties on issues surrounding big data, algorithms based on these data, and how these are used in insurance. This paper was intended as a foundational paper for terminology as well as the actuarial role in data use.

Recent activity by DSAC:

The Data Science and Analytics Committee, in partnership with the Racial Equity Task Force, released a major issue brief, *An Actuarial View of Correlation and Causation—From Interpretation to Practice to Implications*. (July 6, 2022)

The Data Science and Analytics Committee released a major issue paper, *Big Data and Algorithms in Actuarial Modeling and Consumer Impacts*. (November 8, 2021)

The Data Science and Analytics Committee submitted comments in response to a joint request for information and comment from several federal agencies regarding financial institutions' use of artificial intelligence (AI) and machine learning. (June 30, 2021)

Data Science and Analytics Committee (DSAC) Chairperson Dorothy Andrews addressed the issue of race in insurance underwriting in a meeting convened by the National Council of Insurance Legislators (NCOIL) and shared an update on the DSAC's work on the topic. (December 9, 2020)

The Academy's Big Data Task Force released a new monograph, *Big Data and the Role of the Actuary*. (October 17, 2019)

Additional resources

Banerjee, Prashant (2020). "A Reference Guide to Feature Engineering Methods." Kaggle Notebook. Retrieved on September 15, 2022.

Dilmengani, Cem (April 18, 2022). "Feature Engineering in 2022: How to Get the most out of your Data". AI Multiple. Retrieved on September 15, 2022.

Patel, Harshil (August 30, 2021). "What is Feature Engineering—Importance, Tools and Techniques for Machine Learning." Towards Data Science. Retrieved on September 15, 2022.

Radečić, Dario (July 6, 2021). "Imputing Numerical Data: Top 5 Techniques Every Data Scientist Must Know." *Towards Data Science*. Retrieved on September 15, 2022.

Sheriff, Shanawaz (April 15, 2020). "The What, Why and How of Feature Engineering." Actuvate blog. Retrieved on September 15, 2022.

Turner, C.R., Fuggetta, A., Lavazza, L., & Wolf, A.L. (1999). "A conceptual basis for feature engineering." *Journal of Systems and Software*, 49(1), 3-15.

American Academy
of Actuaries

1850 M STREET NW, SUITE 300, WASHINGTON, D.C. 20036
202-223-8196 | **ACTUARY.ORG**