Issues in Information Systems
Volume 23, Issue 4, pp. 135-141, 2022

DOI: https://doi.org/10.48009/4_iis_2022_112

Making sense of data: toward a general taxonomy

Andrew Banasiewicz, *Merrimack College, banasiewicza@merrimack.edu*

Abstract

One of the aspects of data analytics that continues to receive scant research attention is pre-analysis data processing and preparation. With the goal of contributing to that aspect of general data analytic literacy, the research summarized here outlines a general framework geared toward developing computational data familiarity, framed here as understanding of individual variables' encoding characteristics as a necessary precondition to engaging in analytically robust analyses of data. Set in the context of two general classificatory dimensions of 'data type' and 'data origin', the three-tier general data classification taxonomy outlined here supports quick classification of all data into distinct, MECE-compliant groupings, by expressly considering data element-, data file-, and data reservoir-level characteristics. The framework discussed here is intended to serve as a foundation of a more explicit data learning approach to be developed as a part of a larger data analytic literacy initiative.

Keywords: data analytics, literacy initiative, data classification

Introduction

Nowadays, data are not just and voluminous – they are also quite varied, and that miscellany poses a challenge to fostering the increasingly essential data analytic literacy (Banasiewicz, 2021); in fact, the idea of general data familiarity feels almost unattainable in view of the seemingly endless arrays of sources and types of data. And indeed, when thinking of data in terms of narrowly defined source-type conjoints, just the task of identifying all available data seems overwhelming. That line of reasoning, however, is premised upon the often implicit assumption that data familiarity manifests itself as a product of understanding of data's computational characteristics and their informational content (Banasiewicz, 2019), but is that indeed the case? For instance, what specific data knowledge would an analyst tasked with summarizing product sales details need to be able to correctly complete the assigned task? Most certainly s/he would need to be able to discern individual variables' computational characteristics to apply correct – in data processing sense – mathematical operations, but that analyst would more than likely not need an equally robust knowledge of those variables' informational content, because that facet of data knowledge is not essential to correctly completing the aforementioned task of computing product sales summaries. Stated differently, the general idea data familiarity can manifest itself either as the ability to correctly manipulate data, or as the ability to validly interpret the informational content of data (or, of course, as both). It thus follows that when the idea of data familiarity is considered within the confines of data analytic literacy it should be framed within a narrower context of *computational data familiarity,* an idea that encapsulates understanding of individual variables' encoding properties, which is what ultimately determines which mathematical operations can be performed (Banasiewicz, 2019; Mikalef et al., 2018).

It is important to note the difference between how individual data elements can be analytically processed and what type of data analytic outcomes might offer the most informative insights. To continue with the earlier example of computing product sales summaries, the outcome of such analysis could express sales details as a sum or a mean or a median – from a computational perspective all could be correct, though each might support somewhat different informational takeaway, which underscores the importance of expressly differentiating between the earlier mentioned computational characteristics and informational content of data. Knowing what is *permissible*, analysis-wise, is distinct and different from knowing what might be *desirable,* information usage-wise – computational data familiarity always demands the former but not necessarily the latter.

One of the key benefits of focusing only on what operations are permissible has the benefit of greatly reducing the otherwise overwhelming variety of data to just a small handful of types that exhibit shared computational characteristics. For example, automotive accident data typically encompass structured numeric values in the form of accident codes, as well as unstructured text adjuster notes – while informationally similar (both describe different aspects of auto accidents), structured numeric and unstructured text data require fundamentally different processing and analytic approaches; conversely, point-of-sales-captured product codes and automotive accident codes are informationally quite different but computationally indistinct because of shared computational characteristics (i.e., both use structured numeric values). A more in-depth exploration of those simple interdependencies is at the core of a comprehensive computational data familiarity focused taxonomy of data types discussed here.

The core data classification considerations

The considerations outlined above are of key importance to building broad based data analytic competencies, an essential skillset in the evermore data-driven modern society (Banasiewicz, 2021; Hains et al., 2019). Building a robust foundation of data analytic competencies requires a robust approach to developing foundational data familiarity, which is an area that thus far received comparatively little attention. The readily available (i.e., online) data descriptions paint an inconsistent, frankly a confusing picture of data types – even those sources that look at data from the perspective of encoding related considerations tend to be inconsistent: One such classification schema divides data types into integer, floating-point number, character, string, and Boolean, another one into integer, real, character, string, and Boolean values, and yet another into integer, character, date, floating point, long, short, string, and Boolean values. Those inconsistencies aside, at least some of the differences implied by those and other typologies are immaterial if they do not call for different data processing related computational steps.

The preceding discussion considers data familiarity from the perspective of individual data elements which is a core, but not the only consideration. For reasons of convenience and practicality, virtually all data elements are stored in collections which can be thought of falling into one of two general forms: data files, which combine related data elements, and data repositories, which link together multiple data files. Consequently, the development of computational data familiarity also calls for addressing considerations that emanate from file- and repository-level understanding of computational data characteristics. When considered within that broader context, analytic understanding of data can be framed in the context of data *type-* and *origin*-derived characteristics.

Given that the notions of 'type' and 'origin' can be interpreted in a variety of different ways in the context of data analytics, some definitional level-setting might be warranted. As used here, *data type* delimits computationally distinct category, where computational distinctiveness manifests itself in the need for specific mathematical and logical data manipulation operations. Starting from the premise that data can be conceptualized as encodings of actual or presumed facts, typically states or events such as demographic details or sales transactions, computational distinctiveness arises out of the manifest nature of data encodings (e.g., some data are encoded using numeric values, others using text), the range of values individual data elements are allowed to assume, and mathematical operations such as addition or subtraction that can be performed on those values; moreover, since individual data elements are typically grouped into sets (data files) and larger collections (data reservoirs), designs and layouts of those sets and collections (e.g., unstructured vs. structured data files, databases vs. data warehouses) impose additional data processing requirements.

Complementing the computational distinctiveness-minded type of data dimension of data analytic knowledge is the analyses-informing *origin of data* dimension, which accounts for familiarity with generalizable informational origins of data, a critical facet of sound analytic insight creation. For example, the widely used geodemographic data aggregates are derived from the US Census Bureau's demographic details – understanding that the former are not actual captured measurements but rather values that were computed using the detailed Census Bureau data (by means of averaging census block, which is the smallest geographic unit used by the Census Bureau) values is essential to producing valid and reliable data analytic outcomes.

Data type

While in the everyday sense of the word the term 'data' conjures up images of numeric values, that particular encoding represents just one type of data, which is one where facts are encoded using digits; data, however, can also be encoded using non-numeric values (Quantzing, 2021; Rumbold & Pierscionek, 2018). (While it could be argued that the terms 'numeric' and 'digital' are, in principle, interchangeable since data encoding-wise numeric values are ultimately strings of digits, the common usage of the latter of the two terms now refers specifically to the binary 0-1 system used in modern electronics, whereas the former encompasses much more than just the binary system. Still, that intersection of implied and common usages of those two terms poses a definitional problem as the term 'number' – the root of 'numeric' – is commonly understood to represent a quantity, which in turn implies that all numeric data are expressions of some measurable amounts, which is not the case since data encoded using digits can also represent discrete categories. In view of those considerations, the terms 'digit' and 'digital' are used in this research to denote values that are encoded using only digits, as opposed to letters or other forms of expression, and may or may not represent quantities.) In fact, according to industry sources, some 80% to 90% of data is something other than numeric – text, images, web server logs, etc. (Dialani, 2020; Quanzing, 2020). In a more general sense, data can be encoded using numeric values only, letters, including mixed characters (e.g., combinations of letters and digits often used as unique identifiers), visual images, or some combination of all of those types. With that in mind, all data types can nonetheless be simplified to the three core groupings of numeric, text, and image. It is important to note that the 3-part data taxonomy outlined below is based on the overt appearance of data, not on how data elements may be stored in computer memory, a consideration that is particularly cogent for visual image data, which are represented and stored in starkly different formats.

In keeping with the earlier general definition of data, *numeric* data can be defined as actual or presumed facts encoded using digits; it is important to note, however, that digital encoding does not necessarily imply a quantity. For instance, the value of '5' could represent either a quantity, e.g., 5 units of product X, or it could represent a category or a label, e.g., Region 5, but in order for that analytically consequential difference to be recognized, such digit-encoded value needs to be further contextualized by considering the remaining two data type characteristics of allowable range of values and permissible operations. In other words, the development of robust knowledge of data types calls for seeing data as a conjoint of the three distinct characteristics of 1. the nature of encoding, 2. permissible operations, and 3. the range of allowable values. All considered, it is important to keep in mind that numeric data may or may not represent measurable quantities – in general, while all numeric data are encoded using digits, not all digit-encoded values represent quantities.

In a manner similar to the earlier made numeracy-literacy distinction, *text* data can be characterized as actual or presumed facts encoded using letters. Paralleling the quantitative vs. non-quantitative distinction of numeric data, while the everyday conception of the notion of 'text' conjures up images of words arranged into sentences in accordance with applicable syntax rules, when considered from the perspective of data types, text also includes alpha-numeric and special character values, or combinations of letters, digits and symbols that may or may not be arranged into informative (from the human perspective) syntactical entities. One of the key defining characteristics of text data is that they are explosively high-dimensional in the sense of potential elements of meaning (Banasiewicz, 2019; Quantzing, 2021), where an element of meaning could be an explicit word, a multi-word expression, or an implied idea. For instance, a document which is *k* words long and where each word comes from a vocabulary of *m* possible words will have the dimensionality of k^m; simply put, a relatively short text file (e.g., a couple of so standard typed pages) can contain a surprisingly large number of potential elements of meaning. It follows that text data tend to be highly informationally-nuanced, primarily because of their syntactical structure but also because like terms can take on different meanings in different contexts, and the use of punctuation, abbreviations and acronyms, and the occurrence of misspellings can further change or confuse computerized text mining efforts.

The third and final broad form of data, *image*, encompasses visual representations of anything that could range from an abstract form, such as a corporate logo, to a direct visual depiction of an object of interest, such as a picture of a product. Moreover, to the extent to which video is – at its core – a sequence of single images, the scope of what constitutes image data includes static images as well as video. As noted earlier, the focus of the 3-part data typology outlined here is on overt representation, which implicitly sidesteps questions such as how any of the three data types discussed here might be represented in internal computer storage (electronic images are composed of pixels which are stored in computer memory as arrays of integers – in other words, within the confines of computer storage, image data can be considered numeric; that said, analysis-wise, and thus from the perspective of data analytic literacy, image data are not only visually different, but working with images calls for a distinct set of competencies, all of which warrants treating those data as a separate category).

The tripart *numeric—text—image* classificatory typology highlights individual data element-level distinctions; additional differences arise when data are aggregated – for storage and usage – into data files or tables. The most familiar data file layout is a two-dimension grid, in which rows and columns are used to delimit individual data records and data elements (i.e., features or variables); typically, rows delimit data records (e.g., transactions, customer records, etc.) while columns delimit individual variables, producing matrix-like data layout. Known as *structured* data format, that layout's persistent order makes it ideal for tracking of recurring events or outcomes, such as retail transactions as exemplified by point-of-sales (e.g., the ubiquitous UPC scanners) data capture. Overall, structured data are easy to describe, query and analyze, thus even though, according to numerous industry sources such as IDC, a consultancy, those data only account for about 10% of the volume of captured data nowadays, structured data are still the main source of organizational insights.

Many other data sources, however, yield naturally *unstructured* data, primarily because what they record simply does not lend itself to persistent, fixed record-variable format, as exemplified by Twitter records or Facebook posts. In principle, any layout that does not adhere to the two-dimensional persistently repetitive format can be considered unstructured; consequently, the largely non-uniform text and image data are predominantly unstructured. That lack of persistent structure means that unstructured data are considerably more challenging to describe, query and analyze, and thus in spite of being almost overwhelmingly abundant (i.e., account for the 'remaining' 90% or so of data captured today), their analytic utilization has been comparatively low. That said, advances in text and image data mining technologies are slowly making unstructured text and image data more analytically accessible.

Data origin

Framed here as cognizance of generalizable informational origins of data, origin is the second of the two dimensions of data knowledge. To be generalizable, the informational data origins need to be context-nonspecific in the sense of being equally applicable to diverse organizational settings, such as commercial vs. nonprofit, and have to be equally applicable to different industry segments, such as retail, healthcare, or information technology. When considered from that perspective and using the general logic of the MECE (mutually exclusive, collectively exhaustive) principle (Banasiewicz, 2019; Blokdyk, 2020), data can be grouped into five broad informational origin categories of passively observational, actively observational, derived, synthetic, and reference.

Passively observational data, as exemplified by point-of-sales product scanner recordings or RFID sensor readings, are typically a product of automated transaction processing and communication systems; those systems capture rich arrays of transactional and communication details as an integral part of their operational characteristics. Hence those data are captured 'passively' because recording of what-when-where type details is usually a part of those systems' design, and they are 'observational' because they are an ad hoc product of whatever transaction of communication happens to be taking place. From the informational perspective, the scope and the informational content of passively observational data are both determined by the combination of a particular system operating characteristics and the type of electronic interchange; and lastly, given the ubiqutuous nature of electronic transaction processing and communication systems coupled with the sheer frequency of commercial transactions and interpersonal communications, the ongoing torrents of passively observational data flows are staggering.

Actively observational data share some general similarities with passively observational data, but are noticeably different in terms of the key aspects of what they represent and how they are captured. Perhaps best exemplified by consumer surveys, actively observational data can be characterized as purposeful and periodic: Those data are purposeful because they capture pre-planned or pre-determined measurements of attitudinal, descriptive or behavioral states, characteristics and outcomes of interest; they are periodic because they are captured either on as-needed or recurring but nonetheless periodic basis.

Derived data, as suggested by their name, were created from other data, perhaps best exemplified by the earlier mentioned US Census Bureau-sourced geodemographics. A detailed counting of all US residents, undertaken by the Census Bureau every 10 years (as mandated by the US Constitution), produces more that 18,000 variables spanning social, economic, housing, and demographic dimensions, but those detailed data can only be used for official governmental purposes; the US law, however, allows census-derived block level averages to be used for non-governmental purposes. Thus geodemographics, so named because they represent geography-based demographic averages, are data that were sourced from the detailed Census Bureau's data, and though those averages are based entirely on the census details, informationally and analytically they represent a distinct class of data. The same reasoning applies to other derived data categories, as such brand-level aggregates sourced from SKU-level (stock keeping unit, a distinct variant of a given brand, such as differently sized and packaged soft drink brand's varieties) details.

Synthetic is a yet another variant of data as seen from the perspective of analytic origin. Broadly defined, synthetic data are artificially, typically algorithmically generated, rather than representing real-world outcomes or events; their generation parameters, however, are usually derived from real-life data. One of the key benefits associated with synthetically generated data is that those data can be structured to replicate the key informative characteristics of sensitive or regulated data enabling organizations to leverage the informational value of those restricted data without the risk of running afoul of the numerous (and expanding) data access, sharing, and privacy considerations. Consequently, synthetic data are commonly used to train machine learning algorithms and build and validate predictive models. For example, healthcare data professionals are able to use and share patient record-level data without violating patient confidentiality; similarly, risk professionals can use synthetic debit and credit card data in lieu of actual financial transactions to build fraud-detecting models.

The last data origin-framed data variety is *reference* data. In keeping with the everyday definition of the term ‘reference’, which is a source of information used to ascertain something, reference data is a listing of permissible values and other details that can be used to ascertain validity of other data. For example, reference data could include a listing of wholesale and retail product prices set by the manufacturer which can then be used in determining discounted or sales prices; reference data definitions also commonly spell out a number of data layout considerations, such as the fact that the last two digits might represent decimal values. That particular data facet tends to be confused with similar-but-distinct master data, typically framed as agreed upon definitions shared across an organization. While both reference and master data serve as general benchmarks in the data sensemaking process, the former can be seen as a source of objective and/or technical specifications (e.g., set product prices, units of measurement, etc.), while the latter usually represent the currently agreed upon – within the confines of individual organizations – meaning and usage of different data elements.

A general three-tier data classification taxonomy

The preceding overview provided a general outline of the core data classificatory considerations, focusing primarily on individual data elements, while also addressing basic data aggregates, commonly referred to as data tables or files. Be it safekeeping, ongoing maintenance or simple convenience, individual data tables are typically linked together – in accordance with the logic of a chosen data model (e.g., relational or network) – into data reservoirs such as databases, data marts, or data warehouses. While familiarity with those differently organized data collections might seem tangential to developing robust analytic understanding of available data, and the broader goal of data analytic literacy, it is in fact important to correctly interpreting the contents of individual data tables, primarily because data elements in table X might be analytically linked with data elements in table Y, etc. And while an in-depth discussion of competing data models and types of data reservoirs falls outside of the scope of this overview, understanding of more general data aggregates – known as data lakes and data pools – plays an important role in developing a more complete understanding of the potential informational value of available data as it helps to shed light on how those often large and diverse data collections impact data sensemaking.

Broadly characterized as centralized repositories of data, data *lakes* and data *pools* are similar insofar as both typically encompass multiple data files, often representing different facets of organizational functioning, such as product, sales, distribution, and promotional details, customer records, etc. The key difference separating data lakes from data pools is standardization, which is the extent to which data captured from different sources were transformed into a single, consistent format – the former typically are not while the latter usually are standardized.

When considered jointly with the earlier discussed data elements and data tables, data repositories round off a general, tripart data framing taxonomy of data elements – data tables – data repositories, graphically summarized in Figure 1. The framework depicted below is meant to serve as a general data sensemaking blueprint enabling analytically-skilled individuals to rapidly acquire computational data familiarity with new to them data sets.

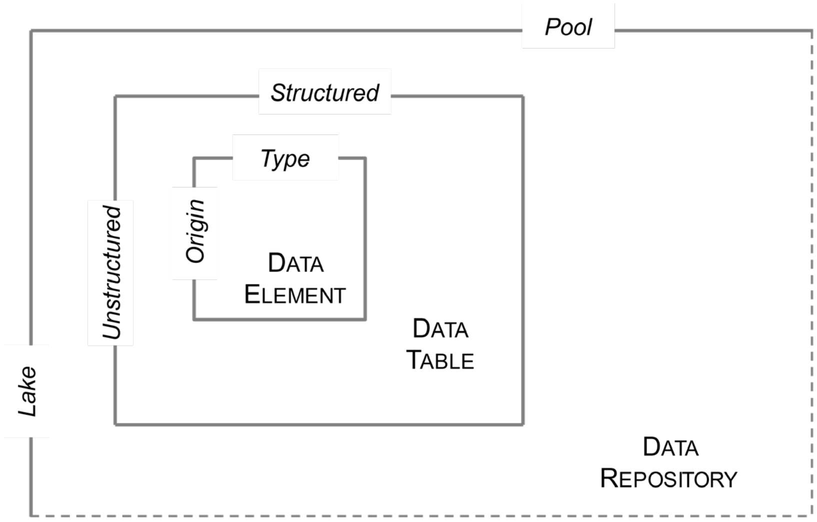

Figure 1
Making sense of data: elements, tables, repositories

As discussed earlier, the general data classification taxonomy is rooted in the premise that engaging in analyses of available data does not require deep informational content knowledge – it requires the ability to correctly identify the key computational characteristics of data, seen here as necessary prerequisites to valid and reliable analyses. That means that a particular set of data can be used to produce multiple analytically sound informational outcomes, and it is for the ultimate users of data analytic outcomes to decide which of those outcomes are most relevant to their informational needs.

References

Banasiewicz, A. (2019). *Evidence-based decision-making: how to leverage available data and avoid cognitive biases*. Routledge.

Banasiewicz, A. (2021). *Organizational learning in the age of data*. Springer Nature.

Blokdyk, G. (2020). *MECE principle 2021: a complete guide*. Lightning Source.

Hains, D., Intindola, M., Lepisto, D., & Wagner, B. (2019). Scrimmage! Teaching quantitative literacy through a multidimensional simulation. *The International Journal of Management Education*, 17(1), 119–129.

Mikalef, P., Giannakos, M. N., Pappas, I. O., & Krogstie, J. (2018). The human side of big data: understanding the skills of the data scientist in education and industry. *2018 IEEE Global Engineering Education Conference,* IEEE, 503–512.

Rumbold, J. M. M., & Pierscionek, B. K. (2018). What are data? a categorization of the data sensitivity spectrum. *Big Data Research*, 12, 49–59.