

## Big Data and Analytics: Issues and Challenges for the Past and Next Ten Years

Stephen Kaisler  
SHK & Associates  
[skaisler1@comcast.net](mailto:skaisler1@comcast.net)

J. Alberto Espinosa  
American University  
[alberto@american.edu](mailto:alberto@american.edu)

Frank Armour  
American University  
[fjarmour@gmail.com](mailto:fjarmour@gmail.com)

William H. Money  
The Citadel  
[wmoney@citadel.edu](mailto:wmoney@citadel.edu)

### Abstract

*In this paper we continue the minitrack series of papers recognizing issues and challenges identified in the field of Big Data and Analytics, from the past and going forward. As this field has evolved, it has begun to encompass other analytical regimes, notably AI/ML systems. In this paper we focus on two areas: continuing main issues for which some progress has been made and new and emerging issues which we believe form the basis for near-term and future research in Big Data and Analytics. The Bottom Line: Big Data and Analytics is healthy, is growing in scope and evolving in capability, and is finding applicability in more problem domains than ever before.*

### 1. Introduction

Our 2013 HICSS article, *Big Data: Issues and Challenges Moving Forward* (Kaisler et al., 2013) examined aspects of Big Data and Analytics (BD&A) and identified key issues and challenges affecting its employment. Our inaugural minitrack article sought to introduce and focus discussions and work in key areas of BD&A. Because this was a fast-moving field, we followed up with various articles and a book (Kaisler et al., 2016; Kaisler et al., 2014; Kaisler et al., 2017). Later, in 2019 we wrote a HICSS assessment article (Espinosa et al., 2019) that analyzed the 500+ papers that cited our original 2013 article. As of August 2022, this article's discussion of the issues and challenges enjoys over 1,280 Google Scholar citations, and research in this field continues to grow. This is most definitely a vibrant field full of challenges and innovations. As minitrack co-chairs, we seek to recognize and document the changes in the field and to provide guidance to the broad BD&A community regarding research findings, trends and needs. This article reports on the "state" of research in the field.

We start by briefly summarizing our 2019 HICSS paper findings and observing which issues and trends have matured or resolved, which are still unresolved

(or ignored by researchers) and highlight what appear to be a new set of emerging issues and trends deserving future research attention. Because of the overwhelming number of articles that have cited our series of papers on this topic, we reviewed a selected set of articles citing our 2013 and 2019 papers, published in premier journals and conference proceedings since 2019.

### 2. State of BD&A Research

The first observation below addresses Doug Laney's infamous 3Vs (Laney, UNK), which define big data – Volume, Velocity and Variety. The published research shows that many scholars, including ourselves, have proposed several other defining Vs, such as veracity, value, validity, etc. We observe that this gold rush to identify and name the next V or Vs muddles rather than resolves what Vs are important from a theoretical perspective.

#### 2.1 Stability of the 3Vs

Over the past twenty years, the number of Vs used to characterize Big Data has been augmented from the original three to more than ten. The latter number represents a significant dilution from the basic idea of simplifying our understanding of what Big Data is. The appeal of the original 3Vs was their simplicity and how it captured the essence of what big data is and isn't. For example, small datasets that don't change much and are of the same data type cannot be considered big data. As the magnitude of each of the Vs increases, we enter the big data world. As such, perhaps big data is best conceptualized as a matter of degree, more than an absolute.

#### 2.2 Proliferation of V's

The multiple other Vs that have been proposed by various authors (e.g., value, variability, veracity, validity, etc.) are perhaps best characterized, not as fundamental determinants of big data, but as important attributes that affect how the data can be managed, accessed and analyzed. We observed a consensus in the literature that the original 3Vs provided a stable

defining framework for big data, and provided common ground for research in BD&A.

### 2.3 Additional V's as Big Data Attributes

We welcome proposals for additional Vs, not as defining properties of big data, but as additional attributes with relevance for particular datasets and projects, or as a focus of a BD&A research project. For example, Veracity by itself does not define Big Data, but it is an important attribute when researching issues of predictive accuracy. Similarly, Value does not really define big data, but it is a useful attribute for research related to clarifying results, hypothesis acceptance (or rejection) and “explainability.”

### 2.4 Evolution of BD&A

The staggering tsunami-like growth rate of BD&A research, application interest, and projects continues. We note that there appears to be a movement in the field focusing on BD&A functional tools, technologies algorithms and methods. This focus overlooks many unattended (and important) research questions, including talent development, governance, data ownership, compliance, management, security, privacy, BD&A processes and lifecycle, teaming and collaboration in BD&A, and functional domain limitations of BD&A applications. Limitations of space prevent us from discussing these topics in this paper.

Issues and challenges come in both components in this field: data and analytics. With respect to data challenges, there are many unresolved Big Data issues, including data quality, validation, provenance, data storage and transport, and the variability and complex representation of some types of data. Table 1 presents some of our observations.

**Table 1. Selected Data Challenges**

Data are being generated at increasingly faster rates than can be utilized and analyzed for relevance to the problem domain. Intake rates, value variance attributed to speed, and missing data are not well researched.
The effect of dataset size on explainability and interpretation of results (e.g., is more data better?).
The effect of data order (occurrence) in a data stream on analytical robustness.
The effect of data perishability given limited computational and storage resources. Some data is relevant over the long-term (e.g., product sales) and some data is fleeting (e.g., RFID tag location).
Integration of data from multiple sources in multiple formats, volumes, and periodicity.

Analytics challenges arise from application proliferation across different techniques and modalities. The eclectic variety of analytical methods is encouraging, but it also muddles attempts to match

appropriate methods and tools to specific problems and domains. Research attention has been focused on demonstrating algorithms, methods, and tools, but little on the reasoning and relevance of these individual elements to the specific business and organizational problems discussed. Table 2 presents some of our observations.

**Table 2 Selected Analytics Challenges**

Scaling effects on application performance and computational and storage resources.
Deriving explainable and interpretable results from complex analytical processes (almost absent in black box modeling).
Generalization of results to dynamic domain specific datasets.
Data descriptions showing the relevance of analytical tools, methods, and processes to specific domains and datasets.
Assessing the value of BD&A in domain decision-making
Developing success criteria for evaluating BD&A analytics for specific problems or domains.

The BD&A field continues to evolve with methods and tools that extend into new domains. Essentially, every business domain is being affected by BD&A, with new issues and challenges emerging as the field matures. This paper – 10 years on – re-examines BD&A to assess the recognition and extent of this evolution and assist researchers by focusing research and theory on the new emerging issues and challenges.

### 3. Research Questions

Table 3 lists the main BD&A issues and challenges we identified in our 2013 and 2019 HICSS articles. Some overlapping categories have been combined as a result of our analysis.

**Table 3. BD&A Issues and Challenges**

Big Data	Analytics
Volume	Cleansing & Curating
Velocity	Enrichment
Variety	Population Imbalance
Value	Analytic Strategy
Veracity	Domain Drift
Storage	Tipping Point
Transport	Provenance
Management	Adaptive Analytics
Input/Output	Analytics Tools and Methods
Quality vs. Quantity	Descriptive, Predictive and Prescriptive Analytics
Growth vs. Expansion	Analytic Science
Security, Access, Privacy and Governance	Ethical Challenges
Processing Scale	Analytics Architectures

Validation	Machine Learning and Cross-Validation
------------	---------------------------------------

This paper categorizes the issues and trends we have identified into the following: (1) technical – e.g., tools, algorithms; (2) data – collection/generation, pre-processing and use; (3) analytics – e.g., methods, approaches; (4) maturity – i.e., accepted vs. diverging practices; and (5) management – e.g., explainability, governance, talent development, etc. Consequently, this article seeks to answer these research questions:

*RQ1: Which BD&A issues and trends have achieved maturity, attention, and resolution acceptance?*

*RQ2: Which BD&A issues remain unaddressed or unresolved, thus requiring further research attention?*

*RQ3: What new and emerging issues and trends deserve close attention by scholars?*

#### 4. Technical Approach

To address the research questions, we identified over 700 papers published since our 2019 paper. Our criteria for selecting papers to review was based on relevance to issues and challenges included in our 2019 paper as well as recognition of emerging issues that raised new issues not previously covered. We reviewed over 150 papers – too numerous to include in our references herein, but included in our extended bibliography. See <https://tinyurl.com/yxchy2ya> for a complete list of papers published since 2019.

Many of these papers focused on analytic exercises or technique demonstrations but did not address some of the key issues and challenges that we noted previously. We found new emerging issues as well as some progress in previously mentioned issues and challenges.

Table 4 summarizes the key emerging BD&A issues we identified. They are organized by categories in our literature review for this paper, and further discussed below.

**Table 4. BD&A Key Emerging Issues**

Issue Category	Issues
Structural	AI/ML Components
	Quantitative/Symbolic AI/ML
	Ontology-Based Supervised Learning
	Analytic Architectures
Maturity	Proliferation of V's
	Transfer of ML Models
	Continuous Learning
	Big Knowledge: Big Data
Evaluation	Domain Expansion
	Reproducibility
Developmental	Data Set Quality
	Data Cleansing

Management	BD&A Governance
	Big Data Growth

## 5. Structural Issues

*Structural issues* are related to how analytics are implemented and how they fit into an architectural framework that provides the end-to-end systematic processing of data and reporting of results. Processing incorporates elements of data storage and data curation. The published literature focuses on individual elements. Little attention is directed toward the analytics life cycle or the analytic framework.

As noted in (Kaisler et al., 2019; Kaisler et al., 2014), complex problems require a deeper understanding of structural issues in developing, choosing, and employing advanced analytics. Underlying this area is the assumption that single methods cannot solve complex problems. Solutions to these problems will require the integration of various systems, protocols, and technologies to collect, store, and analyze data. The term “Quantitative” rather than “Statistical” is used in this context because many ML techniques have a numerical structure but are neither statistical nor probabilistic.

### 5.1 AI/ML Components

Many artificial intelligence (AI) and machine learning (ML) models are custom designed, based on the belief that each problem is relatively unique. In software development, this is not supported since many design patterns have been identified, documented, and catalogued (Gamma et al. 1995). This process, which began with the earliest software component libraries, has evolved for basic ML methods, and has emerged in domain-specific problem-based pattern libraries (7wData, 2017). One approach that has received considerable interest is the ML Commons Foundations ML Cube. It focuses on developing a set of “...simple and interchangeable building blocks that can easily be shared...” for developing new ML systems (ML Commons, 2022).

Automated ML (AutoML) tools are proposed to allow users to develop deployable ML models given a dataset and some configuration parameters. A few systems have been developed with some success, but a comprehensive evaluation of these tools and the claims made for them has not yet been completed. Tenenbaum (Tenenbaum, 2019) argues that this “black box approach” to model-building hides the mathematical models, data cleaning, and feature, model and parameter selection. As a result, bias can be introduced, and lack of knowledge of the algorithm and results can lead to incorrect inferences. This approach makes it easy to apply a wide range of modeling methods to a single problem and generate

predictive accuracy metrics, thus simplifying the model method and specification selection process. However, the lack of understanding of how the ML model is constructed and operated can affect the ability to explain the results, which is a well-known problem with many black box methods.

## 5.2 Integrating Quantitative/Symbolic AI/ML

There has been a significant disconnect between the quantitative and symbolic AI/ML reasoning communities due to perceptions of the efficacy of the respective sets of techniques and methods. Yoshua Bengio, the 2019 AM Turing Award winner, stated that *“Deep learning, as it is now, has made huge progress in perception, but it hasn’t delivered yet on systems that can discover high-level representations — the kind of concepts we use in language. Humans are able to use those high-level concepts to generalize in powerful ways. That’s something that even babies can do, but machine learning is very bad at.”* (Saba, 2021). However, he apparently insists that deep learning will be able to perform high-level reasoning *without resorting to symbolic and logical reasoning*. We remain unconvinced that this is possible, given the absence of demonstrations of high-level reasoning using domain knowledge in quantitative approaches.

This dissonance has affected progress in complex problem domains because neither side recognizes the benefits of applying the techniques and methods of the other side. An AI/ML system that can generalize with high-level representation of knowledge must depend on symbolic reasoning. To date, no theory has demonstrated that human reasoning can escape symbolic reasoning. We argue that this is a critical research area if significant progress is to be made in solving complex domain problems automatically. The integration of AI and ML methods is necessary to achieve evolutionary learning and improvement in self-learning mechanisms.

## 5.3 Ontology-based Unsupervised Learning

Many of the methods used in ML are based on supervised learning models requiring the dataset to be appropriately labeled for quantitative, categorical or binomial (positive vs. negative) outcomes. Labeling requires data pre-processing before it can be used to build a data model appropriate for predictive analytics and/or prescriptive analytics. Data labeling is a cognitively and labor-intensive process.

Integrating a domain ontology with an unsupervised learning system may avoid the manual label process by using domain knowledge to determine good and bad examples of domain phenomena. This is an example of an emerging field called *augmented analytics* (Tableau, 2022). ML techniques may be applied to automate data labeling via an ontology.

## 5.4 Analytic Architectures

Kaisler (2005) described two paradigms that must be considered in developing analytics systems for complex problems: software architectures and software frameworks. As Big Data problems become more complex and datasets more diverse, organizations need to consider the architecture and infrastructure supporting their Big Data processing environments. They also need to consider all aspects of the life cycle from collecting the raw data through to presenting the end results.

The need for a data fabric or unified foundation for ingesting and storing composable data increases with complexity and the scaling of data. An analytics fabric, which can facilitate the application of multiple analytics applied to integrated datasets is also needed. Developing end-to-end AI/ML systems requires a deep understanding of all aspects of the life cycle and consideration of factors such as scaling, performance, accuracy, and explainability.

DevOps for AI/ML (Meenu et al., 2021) is emerging as a critical skill in AI/ML development because data pipelines are becoming more complex, emphasizing tool integration, and requiring governance to ensure reliability. This movement has strong implications for some of the other issues mentioned in this paper. Table 5 presents a few structural issues identified in our reviews.

**Table 5. Structural Issues**

Where latency is an issue, an architecture must reflect constraints due to locational processing whether at edge servers, centralized locations, or distributed across multiple, possibly problem specific, servers.
At an Exabyte scale, ingesting data into repositories becomes an acute problem and can overwhelm physical and logical systems with inadequate performance.
At an Exabyte scale, getting data to analytical systems in a timely and efficient manner to support decision-making can restrict Big Data and AI/ML processing.
As connectivity to the Internet increases, the ability to access and use real-time data becomes a major tool for assessing analytic results and decision-making.
Developing a standard ML Framework will enable comparison of different approaches to ML (Akkiraju, 2018)

Note: “real-time” has many definitions, but we use it in the context of the beholder.

## 6. Maturity Issues

Maturity relates to the evolution of understanding which tools, methods and techniques to apply to problem types/domains and how effective they are at solving exemplary and complex problems. High levels of maturity can lead to reduced cycle times, reduction of costs, lower error rates, and increased agility in

problem solving. Recently, maturity models for AI and ML have begun to emerge (Alsheibani et al., 2019).

### 6.1 Transfer Learning of Models

Machine learning (ML) aims to provide automated extraction of insights from data by means of a predictive model (Tramer et al., 2016) with optimal accuracy. Machine learning models are output by algorithms and are comprised of model data and a prediction algorithm. A machine learning model is **trained to recognize certain types of patterns** in the set of data, providing it with an algorithm that it can use to reason over and learn from the data. Models can be selected, specified, and tuned to maximize their predictive accuracy through cross-validation, which involves testing the models for accuracy with different data than was used to train the model.

Most machine learning models are developed on specific data for specific tasks. Thus, the resulting models are task-specific. This problem was observed in early AI systems and resulted in loss of confidence in AI as a significant tool for automating many human tasks. Because the expense of developing AI/ML is increasing, reusability of such systems would allow such systems to be exploited for a wider range of tasks.

Developing “generic models” may yet be a “bridge too far” in the near term. However, Transfer Learning (TL) (Ruder, 2017) has emerged as a research area to develop systems and models in scenarios for which the models were not originally developed and trained. As humans do not discard knowledge previously learned each time they start a new task, e.g., “persistent knowledge”, TL has the potential to port knowledge learned by one model to new or other models delivering reusability of knowledge, and model technology with a significant impact on the creation, deployment and use of AI/ML.

**Table 6. Transfer Learning Issues**

Developing an architecture and framework for “generic models” is an open research problem.
What protocol, e.g., knowledge format, could be used to transfer knowledge from one model to the next?
If knowledge transfer is implemented as a dynamic feature by linking models, what protocol(s) should be developed to affect this exchange of information?
When linking two or (more?) models, is bidirectional communication an effective means for improving each model’s performance and results?
As Transfer Learning becomes a major tool for applying ML methods, how do we assess successful application of “generic” ML models to similar or different domains?

### 6.2 Continuous Learning

Building a Big Data set presents significant challenges in acquiring, organizing, integrating, and managing the dataset. The integrity of an

organization’s data repository and analysis process can be dependent on its dataset(s) being constantly updated. As data is continuously collected, users must consider replacement strategies for the collected data. Replacement may also require development of new or revised data structures to store the new data. It may also force the development of new analytical systems or enhancement or retraining of existing models to improve diagnostic, predictive or prescriptive capability and to enhance performance.

In complex domains, data may be collected from multiple types of sensors using different sensory modes. By modes, we mean vision, smell, touch, kinesthetic, and auditory. Integration of data across a set of modal sensors may be as difficult as integration across sensors from different modalities.

Associated with continuous data collection is the issue of continuous machine learning (CML) (Minku, 2022). Reasoning, prediction, and learning processes must update the results from AI/ML systems to keep up with the evolving domain situation as new data are ingested. One approach is to periodically retrain a model using the new dataset, but as more data becomes available this may not be feasible due to scale or training time.

**Table 7. Continuous Learning Issues**

For multi-model analytics, how do we synchronize inputs from different models arriving at different rates?
What is the best method: incremental, periodic versus dynamic (perhaps evolutionary) learning? Or both?
How do we provide feedback in a continuous, dynamic learning environment?
How do we provide hints to direct a continuous learning system? (Ramprasaanth et al., 2019).
As data is collected from disparate data sources, synchronization of data collection must be addressed.
How can one handle concept drift over time as new data is dynamically ingested?

### 6.3 Big Knowledge: Big Data

Big Data has been described and discussed for over thirty years as a means of improving an organization’s understanding of its business environment and operations. Big Data was often used to describe and diagnose the current state, and predict future state possibilities. Big Data Analytics have been utilized in many domains to assist in strategic planning and decision making. The next stage of Big Data usage we believe is to develop Big Knowledge sets (Ruqian et al., 2019) which capture the domain knowledge in a representational form capable of being manipulated by predictive and prescriptive decision-making systems.

Within the past decade, integration of multiple diverse Big Data sets has become a standard practice in order to provide an enriched dataset that is more descriptive of a problem domain’s phenomena. Integration often occurs at a syntactic level focusing

on data characteristics rather than data meaning. Data lakes emphasize the collection of data rather than its integration into a coherent descriptive and diagnostic structure. Different collection methods produce data that may be used for joint analysis but are often limited by different underlying semantics that must be logically associated, or alternatively, disambiguated and deconflicted. Automated integration algorithms can combine data with limited domain knowledge leading to a disparate representation of a problem space.

**Table 8. Big Knowledge Issues**

An organization’s primary focus should be to determine what to do with the data it has – whether raw or derived
Fusion algorithms are required to merge data into problem-focused data structures, possibly after cleansing and transformation.
Big Knowledge repositories need to focus on how data will be used rather than how it is stored, even if this means some inefficiency in the use of storage systems.
Developing automated integration algorithms to combine datasets with little domain knowledge while avoiding disparate representation of a problem space.
As knowledge volume increases, explainability of results will become more difficult, but, perhaps, more comprehensive. But a key issue is how much/detailed does a user want an explanation to be?
Beyond Data As-a-Service (DAAS), Knowledge As-a-Service (KAAS) will become a capability that can replace or augment web browsers and similar systems.

## 6.4 Big Data Domain Expansion

A critical issue is the emergence of deep fake news and images using AI/ML techniques to create and/or modify existing data to present disinformation (Chorasa et al., 2021). The results of this activity, called *generative AI* (Longoni et al., 2022), vary from disinformation to libelous, potential harmful, information. But generative AI is also useful for creating synthetic data when there is a sparsity of actual data. Big Data may help detect and defend against such misappropriation of technology, but the effort has only begun.

## 7. Evaluation Issues

An *evaluation issue* is related to the assessment of the efficacy of a set of analytic techniques, measurement of the quality of the result, and delivery of the value contribution to decision making. It also encompasses the comparison of models and/or approaches to problem solving to determine the best one based on specified criteria, such as accuracy, performance, and fidelity.

As Liao and colleagues note (2021), “*evaluating algorithmic progress is a double-edged sword*”. Benchmarks are designed with specific performance

metric(s) that allow objective assessment(s) of different algorithms. But, characterizing a new algorithm or system with a single performance metric “creates an illusion of simplicity that ignores (or hides) underlying assumptions in the learning problem”.

The selection of algorithms for a given problem, whether Big Data, AI, or ML, is more complex than just selecting a particular technology. Many factors must be considered, including the type, quality, and quantity of the data; the type of problem being addressed; the accuracy and performance required, and supporting infrastructure—hardware and software.

Simple techniques have been proposed for evaluating quantitative ML methods based on well-known statistical metrics that assess the accuracy of the models by computing error or deviance metrics. This approach is primarily a statistical exercise, lacking domain knowledge to assess what the model’s relevance is to the business problem and how well the problem has been understood, specified in the model and solved. These techniques deal largely with static datasets, tend to saturate quickly, and are susceptible to overfitting. Large non-stationary datasets need dynamic algorithms to handle the temporal (and spatial) aspects of the domain.

Rigorous evaluation techniques and metrics exist for descriptive and diagnostic analytics, some for predictive analytics (Calixto et al., 2020; Kananda, 2019), but almost none for prescriptive analytics. As noted in (Espinosa et al., 2019), many analytics are idiosyncratic, which leads to restricted evaluation techniques and methods. As Aroyo et al (2022) noted, the goodness of the fit of the model to the dataset is about how accurately the model performs with a test dataset. However, how well the dataset or model represents the real world or the actual business problem is never assessed. Lacking this “validity” metric, we must hold suspect the quality and validity of the results.

One important aspect of the evaluation process is the tension between explainability (Xu et al, 2019)—how correctly a model describes the true effects on the focal outcome, and predictive accuracy—how much the model predictions depart from the actual outcomes. Often, explainable models are not the most accurate (e.g., ordinary least squares regression) and, conversely, the most accurate models are not interpretable (e.g., neural networks). This tension is one basis for assessing how much people will trust and select AI/ML systems. The authors recall that in one past HICSS presentation in this mini-track, one author stated “*the only thing that matters in ML is the accuracy of the models*”, followed by this statement by the next presenter “*we evaluate loans for applicants and we are accountable to regulators, so our models must be explainable*”.

Quantitative models cannot explain the true relationship between predictors and outcomes by themselves. Furthermore, they cannot explain causality, but only the quantitative association of a given set of predictors with an outcome. These predictions can be affected by multiple things like nuisance or noise variables, spurious correlation, multicollinearity, and measurement errors. Thus, a suggested solution is to integrate symbolic formalisms and semantic inferencing with them.

### 7.1 Reproducibility in Big Data & AI/ML

In the physical sciences, reproducibility of results is often a gold standard for an experiment, especially if it improves techniques, precision, and other metrics. Unfortunately, in AI/ML, we have few reproducible experiments or projects and comparisons between similar projects is often extremely difficult (Ghanta et al, 2018). Lack of reproducibility is a severe constraint on acceptance of the quality of such projects combined with a lack of comprehensive documentation of processes, procedures, and results. This affects our ability to assess the quality of algorithms in these areas.

**Table 9. Evaluation Issues**

How do/will we measure the improvements in prediction with continuous learning?
In advanced analytics, how do/will we determine if one prediction is better than others?
How do/will we assess the value contribution of different modal data to decision making?
How do we measure data quality in a consistent and unambiguous way? [24]
Benchmarking: science, application, or system? Separately or combined? [26]
How do we ensure that benchmark data captures the natural ambiguity of the real world?
New frameworks are required to handle repeated analyses, gather and process results, and assess and explain the characteristics of results (semi-) automatically.

### 8. Development Issues

A *development issue* is one that relates to the creation of a system for BD&A (Thiygalingam et al, 2021). It is concerned with three development aspects, the: science or business; application; and the system. As we have noted, there are many factors to be considered in developing a solution approach. Hard constraints, such as data volume, time to decision, processing speed, and communication bandwidth, have led to the dichotomy of “small data analytics” and “Tiny ML”. These are downsized versions of algorithms and systems that can fit within those hard constraints. A critical problem to be resolved is how to downsize existing algorithms and systems or must we

create new ones while retaining a specified degree of accuracy and precision in the results.

### 8.1 Data Set Quality

Assessing dataset quality is an open-ended research problem. When problems arise in assessing the performance of AI/ML models, most efforts focus on improving the use and implementation of algorithms. Some methods have been proposed for assessing dataset quality on a domain basis, but there seems to be no universally agreed-upon method for determining the reliability of dataset quality.

As we noted in (Kaisler et al, 2013), issues like noise, data imbalance, skewed distribution of records, missing or inconsistent data, among many others can lead to imperfect, inaccurate or erroneous results. Methods for identifying, assessing, and mitigating these problems tend to be domain-specific with limited cross-domain applicability.

One aspect of this is a custom-designed data collection technique for particular domains. Standard software engineering, methodologies for dataset collection are needed to improve data quality. One approach that has worked very well is NIST’s Text Retrieval Conference (TREC, see <https://trec.nist.gov/>). It has pioneered advances in the assembly of a large store of practical experience in collecting, assembling, labeling, and measuring the quality of datasets.

One aspect that has not been very well-researched is the occurrence of errors in crowdsourcing. Measuring errors due to the many causes of data inaccuracy in crowdsourcing environments represents a significant scaling problem. It is attributed to the number and types of sources, the amount of data from different sources, and the perceived reliability, veracity, and accuracy of each of the sources.

Erroneously labeled datasets used for supervised learning systems introduce imperfect or poor performance into such systems. Northcutt et. al. surveyed errors in 10 most commonly used CV, NLP, and audio datasets to assess the occurrence of errors (Bendre et al, 2016). They have implemented a tool called Cleanlab as an open-source solution to begin identifying such errors (<https://github.com/cleanlab/cleanlab>).

Data cleansing is critical to successful analysis of datasets by improving dataset quality. Methods for dataset cleansing also tend to be domain-based with limited applicability to other domains. Data cleansing can be labor-intensive, both cognitively and manually, but recent research has begun to address the automation of data cleansing.

Aroyo et. al. (2021) identified critical properties of data that affect the quality of the results in Big Data and AI/ML systems. They noted that “*bad data begets*

*bad results*". As we noted (Kaisler et al, 2013), good algorithms cannot overcome bad quality data. Table 10 addresses issues with data quality.

**Table 10. Data Quality and Cleansing Issues**

How do we assess the amount of preprocessing work to prepare data for consumption by Big Data and AI/ML systems?"
Standardized data labeling conventions are needed to eliminate erroneous labeling.
Crowdsourcing standards are required for comparable set development and results.
Applying AutoML shows significant promise in data cleansing and the imputation of missing data because the skills and methods in this area are often applied repeatedly.
Using generative AI to improve the breadth and depth of datasets to ensure adequate domain coverage.

## 9. Management Issues

In a recent study, Abraham et al (2019) identified data governance as a critical issue. Numerous articles have focused on the importance of data governance as a factor for success in data science and AI/ML systems. Several articles in the extended bibliography address different aspects of data governance.

An emerging area is MLOps, which is a set of processes designed to transform experimental Machine Learning models into productionized services ready to make decisions in the real world (Ippolito, 2022). Similar to DevOps, it also includes data validation and continuous training/evaluation of models. A number of design patterns have been proposed for architecting ML systems that offer many of the benefits described by Kaisler in [paradigms]. The MLOps approach emphasizes management of ML systems design from an end-to-end systems perspective.

### 9.1. Data Governance

As data grows and becomes more distributed, use and management, physical storage limitations, and effective data governance policies and procedures will be required to ensure the quality and security of big data (Judah et al, 2021). The implication for ethical use of Big Data results have not been adequately explored. But, as (Chorasa et al, 2021; Longoni et al, 2022) pointed out, this has become of significant concern in today's political arena. We previously addressed ethical and criminal issues in (Kaisler et al, 2014), but have observed little resolution of them.

**Table 11. Data Governance Challenges**

Ensure equitable access based on specific organizational usage and ownership criteria.
--

Establish and enforce security and privacy mechanisms to protect against corruption, unintentional release, and destruction.
Ensuring accountability for data and analytics governance decisions.
Designing systems to scale enterprise big data.
Tracking governance investments against business value.
Tracking governance enforcement and auditing across BD&A.

## 9.2 Big Data Growth

Curating and managing Big Data repositories means that critical issues need to be addressed with consistent policies that maintain the usefulness of the data. These previously identified issues (Kaisler et al, 2014; Espinosa et al, 2019) continue to affect the use of Big Data in many existing and emerging domains.

As datasets grow in size, both individually and as collections, the question of scaling becomes critically important. With terabyte disk volumes, this wasn't as critical, albeit performance issues, but with petabyte and exabyte data sizes, we observe several emerging challenges:

1. If we can't store the data in one system, what criteria do we use to distribute across multiple systems?
2. How can we construct ML models that can be distributed across multiple systems, given (1)?
3. How do we scale ML tools and techniques to handle increased dataset sizes, e.g., will ML techniques scale with increased data sizes?

**Table 12. Big Data Growth Challenges**

Cycling out old information at the same time new information is being inserted into a dataset.
Determining and enforcing limits on the lifetime of temporal longitudinal information.
Ensuring effective cleansing algorithms are uniformly applied across similar data.
Introduction of new or revised data models to handle changes in the diversity of data.
Will current AI/ML tools scale with increasing dataset size? If not, what modifications will need to be made or will new algorithms need to be developed.
How much data is enough to solve a problem to a given degree of accuracy, reliability, and usefulness?

## 10. Conclusion: What has Changed?

Our previous literature review (Espinosa et al, 2019) convinced us that, the discipline is well along the path to maturity in many areas, but lagging in others, which is common with complex, fast-growing disciplines. With new applications, techniques, tools, and issues emerging, we believe that we need to continually assess how the discipline is doing with respect to our three RQs:

*RQ1: BD&A issues and trends that have achieved maturity, attention, and resolution acceptance are presented in Table 13.*

**Table 13. Resolved or Understood Issues**

Defining and characterizing Big Data by the three V's.
The use of some ML models for continuous learning (Rugian et al, 2019).
Attention to security and privacy of Big Data and its utilization is a maturing area.
Attention to critical issues in data cleansing, data governance, and data management are being resolved.
Recognition of how to use the results of multidisciplinary BD&A to solve interdisciplinary social and technical problems is a maturing area.

*RQ2: BD&A issues, tools and techniques that remain unaddressed or unresolved, thus requiring further research attention are presented in Table 14.*

**Table 14. New and Continuing Critical Issues**

How to determine if incorrect conclusions have been reached by AI/ML tools?
How to assess the risks and likelihood of using incorrect conclusions in decision making?
Applying AutoML to enhance the quality of large (>10 <sup>9</sup> records) datasets with traceability and reasonable time performance.
Detecting and handling concept drift in ML systems.
Developing tools and technologies to support dynamic AI/ML model evolution from streaming data.
As ML models are commercially deployed, from a black box perspective, the emergence of attacks to subvert or steal these models are emerging (Kananda, 2019).

*RQ3: New and emerging issues and trends that deserve closer attention by scholars are summarized in Table 15.*

**Table 15. New and Emerging Issues**

How to address, associate, and integrate the quantitative and symbolic methods of AI/ML analysis?
How to characterize quantitative AI/ML optimization?
How to characterize & assess qualitative AI/ML results?
Understanding how the use of DevOps (e.g., AIOps and MLOps) in AI/ML can lead to mature applications.
Although Quantum Computing has made some significant progress, the role of Quantum Computing in BD&A is an open research issue—uncertainty continues to reign.

## 11. Future Work

The issues and challenges raised in this paper have suggested research areas that should be pursued in the near-term as described in Table 16. BD&A has matured substantially in the last 10 years, but as in most complex disciplines, some old issues become well understood, other old issues are still unresolved, and new challenges and issues emerge much faster than we can find solutions.

This review has helped us re-assess our understanding of where the BD&A field is today, but we hope that it can also help members of this minitrack community identify fertile areas for much needed research, although some of the topics in Table 15 (RQ3) represent long-term research issues. We seek collaboration with individuals and teams interested in these issues and challenges over the next few years.

**Table 16. Near-term Tasks**

Develop a standard metric and systemic approach in a framework to measure data quality and its fit to the real world.
Develop a method(s) to apply properties from software engineering excellence to assess BD&A model excellence.
Conduct replication studies to validate the applicability of BD&A tools to particular problems to eliminate “tool bias”.
Develop mechanisms for explaining how results are derived from initial raw data, including a “traceback” capability.
Develop mechanisms for detecting and correcting bias in AI/ML systems.
Develop mechanisms for introducing personality into interactive AI/ML systems acting as symbiotic agents with human users.
Develop protocols for information exchange among interacting AI/ML systems.
Develop a road map for long term research in BD&A and AI/ML systems.

As we noted, space limitations prevented us from addressing issues in several areas, such as data governance, legal compliance, and data management among others. We intend to continue this series of papers to address some of the issues and challenges in these other areas.

**Extended Bibliography:** We reviewed many papers beyond those selected for the primary review. An extended bibliography is available upon request from the authors.

## References

- Abraham, Rene, Johannes Schneider and Janvom Brocke, 2019. “Data governance: A conceptual framework, structured review, and research agenda”, *International Journal of Information Management*, 49:424-438.
- Akkiraju, R., V. Sinha, A. Xu et al. 2018. “Characterizing Machine Learning Process: A Maturity Framework”, arXiv:1811.04871
- Alsheiabni, S., Y. Cheun, and C. Messon. 2019. “Towards an Artificial Intelligence Maturity Model; From Science Fiction to Business Facts”, PACIS 2019.
- Aroyo, L., M. Lease, P. Paritosh, and M. Schaekermann. 2021. “Data Excellence for AI: Why Should You Care”, Retrieved from arXiv: <https://arxiv.org/abs/2111.10391>, 30 March 2022.

- Bendre, M.R. & V.R. Thool. 2016. "Analytics, challenges and applications in big data environment: a survey", *Journal of Management Analytics*, 3(3):206-239.
- Calixto, Nelito and João Ferreira. 2020. "Salespeople Performance Evaluation with Predictive Analytics in B2B", *Applied Sciences*.
- Choraša, M., K. Demestichas, A. Gielczyk, et al. 2021. "Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study", *Applied Soft Computing*, Volume 101.
- Espinosa, J.A., S. H. Kaisler, F. Armour, and W. Money. 2019. "Big Data Redux: Issues and Challenges Moving Forward," in *52nd Hawaii International Conference on System Sciences*, Maui, HI.
- Gamma, E., R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley. 1995.
- Ghanta, S., S. Subramanian, and S. Sundaraman. 2018. "Interpretability and Reproducibility in Production Machine Learning Applications", *17th IEEE International Conference on Machine Learning and Applications*.
- Ippolito, P. 2022. Design Patterns in Machine Learning for MLOps, <https://www.kdnuggets.com/2022/02/design-patterns-machine-learning-mlops.html>
- Judah, S., A. White, Strome, A. Toncheva, The State of Data and Analytics Governance: IT Leaders Report Mission Accomplished; Business Leaders Disagree, Gartner Research, 6 December 2021 - ID G00758465.
- Kaisler, S. 2005. *Software Paradigms*, John Wiley & Sons, New York, NY.
- Kaisler, S. H., F. Armour, J. A. Espinosa, and W. Money, "Big Data: Issues and Challenges Moving Forward," in *46th. Hawaii International Conference on System Sciences*, Maui, Hawaii, 2013.
- Kaisler, S.H., F. Armour, W. Money, and J. A. Espinosa, "Big Data: Issues and Challenge," in *Encyclopedia of Information Science and Technology, Third Edition*, M. Khosrow-Pour, Ed., Third ed: Information Resources Management Association, 2014, pp. 363-670.
- Kaisler, S.H. , F. Armour, W. Money, and J. A. Espinosa., "Advanced Analytics: Issues and Challenges", *47th Hawaii International Conference on System Sciences*, Hilton Waikoloa, Big Island, HI, 2014.
- Kaisler, S. H., F. Armour, A. Espinosa, and W. Money. 2014. "Advanced Analytics: Issues and Challenges", *Encyclopedia of Science and Technology*, M. Khosrow-Pour, Ed., Third ed: Information Resources Management Association, 2014.
- Kaisler, S.H., F. Armour, J. A. Espinosa, and W. Money, *Obtaining Value From Big Data for Service Delivery*, Third ed. New York, NY: Business Expert Press, 2016.
- Kaisler, S.H. , W. H. Money, F. Armour, and J. A. Espinosa, "Big Data: The Path to Maturity," *International Journal of Systems and Service-Oriented Engineering*, 7:23, 2017.
- Kananda, G. 2019. "An Evaluation of Big Data Analytics Projects and The Project Predictive Analytics Approach", *Oriental Journal of Computer Science and Technology*, 12(4): 132-146.
- Laney, D. Unk. "3D Data Management: Controlling Data Volume, Velocity, and Variety," Stamford, CT, February 6.
- Liao, T.I., R. Taoi, I.D Raji, and L. Schmidt. 2021. "Are we learning yet? A Meta\_Review of Evaluation Failures across Machine Learning", *35th Conference on Neural Information Processing Systems*.
- Longoni, C., A. Fradkin, L. Cian, and G. Pennycook. 2022. "News from Generative Artificial Intelligence Is Believed Less", *Proceedings of the 2022 Conference on Fairness, Accountability and Transparency*.
- Meenu, M., H. Olsson, and J. Bosch. 2021. "Towards MLOps: A Framework and Maturity Model", 47th EuroMicro Conference on Software Engineering and Advanced Applications, pp. 1-8.
- Minku, L.L. 2022. "Principles of Continuous Learning", <http://www.cs.bham.ac.uk/~minkull/slidesCISE/25-principles-continuous-learning.pdf> on May 19, 2022.
- ML Commons. 2022. ML Cube, retrieved from <https://mlcommons.org/en/mlcube/> on March 15, 2022.
- Northcutt, C., A. Athalye, and J. Mueller. 2021. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks", *Proc.of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*.
- Ramprasaath R.S., S. Lee, Y. Shen, et al.. 2019. "Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded". *Proceedings of the International Conference on Computer Vision (ICCV)*.
- Ruder, S. 2017. Transfer Learning – Machine Learning’s Next Frontier, retrieved from: <https://ruder.io/transfer-learning/>
- Ruqian, L, X. Jin, S. Zhang et al. 2019. "A Study on Big Knowledge and Its Engineering Issues," in *IEEE Transactions on Knowledge and Data Engineering*, 31(9): 1630-1644.
- Saba, W. 2021. AI Cannot Ignore Symbolic Logic, and Here’s Why., <https://medium.com/ontologik/ai-cannot-ignore-symbolic-logic-and-heres-why-1f896713525b>
- Tableau. 2022. Augmented analytics explained: definition, use cases, benefits, features, and more, <https://www.tableau.com/learn/articles/augmented-analytics>.
- Tenenbaum, J. Automated Machine Learning, <https://towardsdatascience.com/automated-machine-learning-d8568857bda1>, 10/21/2019.
- Thiyagalingam, J., K. Leng, S. Jackson, et al. 2021. "SciMLBench: A Benchmarking Suite for AI for Science, <https://github.com/stfcsciml/sciml-bench>.
- Tramèr, F., F. Zhang, A. Juels, et al. 2016. "Stealing Machine Learning Models via Prediction APIs" 25<sup>th</sup> USENIX Security Symposium, Austin, TX.
- Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J. (2019). "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges". In: Tang, J., Kan, MY., Zhao, D., Li, S., Zan, H. (eds) *Natural Language Processing and Chinese Computing*. Lecture Notes in Computer Science, Vol. 11839. Springer.
- 7wData. 2017. <https://7wdata.be/artificial-intelligence/design-patterns-for-deep-learning-architectures/>.