Data Quality – The Role of Empiricism

Shazia Sadiq
The University of Queensland, Australia
shazia@itee.uq.edu.au

Tamraparni Dasu
AT&T Labs-Research, USA
tamr@research.att.com

Xin Luna Dong
Amazon, USA
lunadong@amazon.com

Juliana Freire
New York University, USA
juliana.freire@nyu.edu

Ihab F. Ilyas
University of Waterloo, Canada
ilyas@uwaterloo.ca

Sebastian Link
The University of Auckland, New Zealand
s.link@auckland.ac.nz

Renée J. Miller
University of Toronto, Canada
miller@cs.toronto.edu

Felix Naumann
Hasso Plattner Institute, University of Potsdam, Germany
felix.naumann@hpi.de

Xiaofang Zhou
The University of Queensland, Australia
zfx@itee.uq.edu.au

Divesh Srivastava
AT&T Labs-Research, USA
divesh@research.att.com

ABSTRACT

We outline a call to action for promoting empiricism in data quality research. The action points result from an analysis of the landscape of data quality research. The landscape exhibits two dimensions of empiricism in data quality research relating to type of metrics and scope of method. Our study indicates the presence of a data continuum ranging from real to synthetic data, which has implications for how data quality methods are evaluated. The dimensions of empiricism and their inter-relationships provide a means of positioning data quality research, and help expose limitations, gaps and opportunities.

1. INTRODUCTION

Effectiveness and efficiency have been critical to the success of data management, data integration and data analytics technologies over the years. Effectiveness ensures that the result serves the purpose for which it was obtained, while efficiency ensures that the process of obtaining the result does not waste critical resources. Obviously, one without the other, while possible, is not desirable, especially in this age of Big Data, where critical decisions need to be made correctly and quickly.

Empiricism postulates the fundamental role of experiments and measurements in the advancement of science [33]. Historically, it has been very important to improving the efficiency of data management technology. The TPC family of benchmarks (tpc.org) has contributed to measuring and continually improving the efficiency of data management technology over several decades. For example, the TPC-C and TPC-E benchmarks measure the performance of on-line transaction processing applications, and the TPC-H and TPC-DS benchmarks measure the performance of decision support systems. Although the focus has been mostly on relational or structured data, efficiency-oriented benchmarks exist for non-relational data models too. Recently, the TPC benchmarks have been expanded to consider the efficiency of data integration (TPC-DI) [39] and big data processing [9].

As long as one assumes that the input data are trustworthy and of high quality, and the transformations performed on the input to produce the result are well understood and match expectations, one can happily regard the result obtained as being of high quality as well. In the big data world, however, data sources are not always trustworthy, and complex, ill-understood pipelines are used to transform the data. Consequently, the quality of the results should be

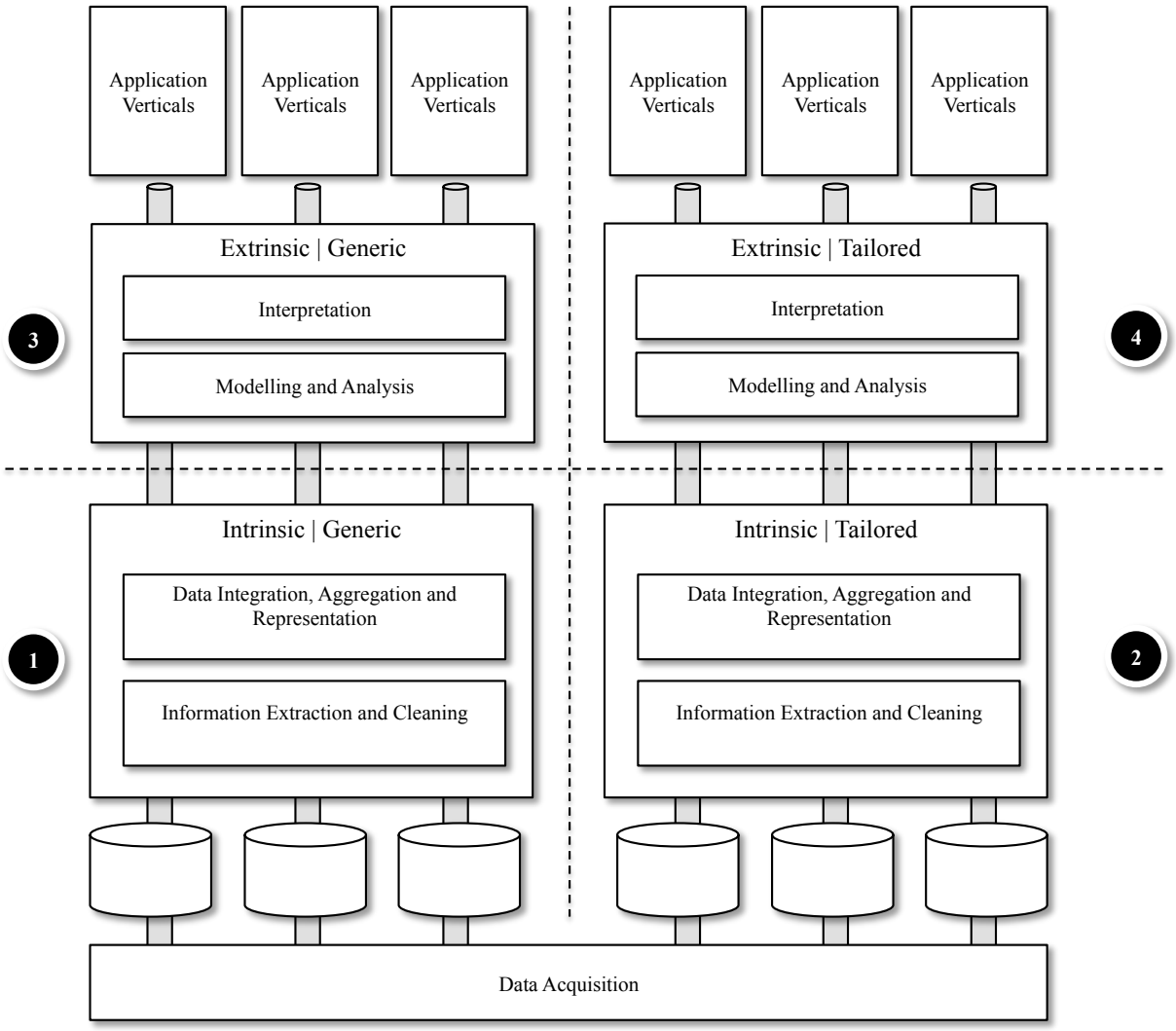

Figure 1. Typical Data Processing Pipeline

viewed with the appropriate level of skepticism. Being able to empirically evaluate the trustworthiness of the data sources and effectiveness of the data processing pipelines, and hence the quality of the obtained results, would go a long way towards ameliorating this undesirable situation. However, it is not immediately evident which aspects of the pipeline contribute more significantly to an authentic empirical evaluation.

In 2015, a group of global thought leaders from the database research community outlined several grand challenges in getting value from big data [3]. A key message was the need to develop the capacity to "understand how the quality of data affects the quality of the insight we derive from it". The role of data quality is recognized as pivotal to the effectiveness of data pipelines.

The notion of quality is highly contextual and tied deeply to fitness for use [25]. In determining the effectiveness of these pipelines, it therefore becomes critical to evaluate the fitness of the data for its intended use. Similar to how TPC benchmarks help measure efficiency in data management; empirical evaluations of data quality will help measure the effectiveness of data pipelines. However, balancing the purposefulness (depth) of data quality detection and cleaning methods with their capacity for wider applicability (scope) [17] remains a challenge.

In this paper, we identify two inter-related dimensions of empiricism that help locate the sweet-spot for empiricism in advancing data quality research and practice. These are the *type of metric*, and the *scope of method*. We explain these dimensions of empiricism in the next section.

While type of metric and scope of method have direct implications for the technology stack that implements a data processing pipeline (see Figure 1), a third aspect,

namely, the *nature of the data*, exposes a data continuum that defines the setting in which the data quality metrics and methods can be evaluated. In Section 3 we outline the data continuum and discuss the properties of real data, synthetic data and everything in between.

In Section 4, we present the various ways in which the dimensions of empiricism can be positioned, thus providing a lens through which the role of empiricism in data quality research can be studied. In order to gain a deeper insight into each of these positions, we reached out to thought leaders in data quality research [44, 45] to help elaborate on the motivation and rationale, key approaches, and possible challenges against each position. The viewpoints presented are extracted from a series of interviews conducted with the experts and are supplemented with a review of relevant literature.

Finally, in Section 5, we present a set of recommendations on promoting empiricism in data quality research and practice. These recommendations have been synthesized from the findings reported in this paper.

2. DIMENSIONS OF EMPIRICISM

Figure 1 presents a typical data processing pipeline from acquisition to analytics. The components of the pipeline have been extracted from [23], and include five steps of the data processing pipeline, namely, (i) Data Acquisition, (ii) Information Extraction and Cleaning, (iii) Data Integration, Aggregation and Representation, (iv) Modelling and Analysis, and (v) Interpretation.

Data quality considerations are rooted throughout the pipeline, from the time data are acquired, through various transformations within the pipeline, to their eventual interpretation for a given application.

Figure 1 presents four quadrants that represent four distinct positions where type of metric and scope of method influences the way in which the data processing tasks are handled. The background to these positions is introduced below and the positions are further detailed in Section 4.

2.1 Type of metric

Prior works have identified many metrics to measure specific data quality characteristics [54], such as completeness, timeliness, consistency, etc. Essentially, metrics can be intrinsic or extrinsic to the characteristics of the data [24].

- **Intrinsic metrics** are application-independent, and can be declaratively defined and measured, such as the format-consistency of a date/time attribute. Intrinsic metrics are expected to be handled in Quadrants 1 and 2 in Figure 1.
- **Extrinsic metrics**, on the other hand, are application-dependent, such as the fidelity of a specific analytical report. Thus, extrinsic metrics are exposed and managed in Quadrants 3 and 4 of Figure 1.

Even though intrinsic metrics can typically be implemented without reliance on external reality, there are some caveats to the assumption. For example, timeliness of event data can be measured from the update logs, however the notion of timeliness (comparing to the time at which the event occurred in reality) relies on external reality. Similarly, completeness can have multiple interpretations. For example, null values for mandatory attributes can be counted without reference to an external source, but missing records require reference to a trusted external source, such as master data. Missing records can also be application-dependent, for example a public transport dataset may be complete for city planning but incomplete for scheduling [41, 42].

Regardless of whether a metric is intrinsic or extrinsic, based on rules or statistical notions, it has to be tied to the underlying data to be relevant to measure data quality.

Thus, whereas intrinsic metrics are focused on the properties of the data only and not on how applications use it, the aim of intrinsic metrics is indeed to eventually contribute to achieving an extrinsic metric. An extrinsic metric also has to be tied to the underlying intrinsic metrics of data quality, even if its measurement is application-dependent. The **part-of** (or **aggregation**) relationship between an intrinsic and extrinsic metric is thus rather nuanced. The value of the data (an extrinsic metric) may be composed of not only multiple intrinsic metrics, such as completeness, lack of duplicates, format consistency etc., but also the relative importance of each metric for the task at hand.

For example, consider postal delivery services, wherein the completeness of customer address can be considered an intrinsic metric of customer data quality. At the same time, there can be an extrinsic metric on how well the data supports the business objective of priority deliveries to be completed within 72 hours. The two are clearly inter-related (i.e., incomplete customer addresses can result in delivery delays), but the relationship needs to be defined and measured in a precise way to demonstrate how the investment on making customer address data complete, impacts on the reduction in delayed or failed delivery attempts.

A data quality (detection or cleaning) method may be designed to optimize metrics of either intrinsic or extrinsic type. However, as shown in the examples above, in empirical evaluations of data quality systems, both intrinsic and extrinsic metrics need to be considered together, along with their possible dependencies.

2.2 Scope of method

There have been significant contributions from research and practice towards developing methods that assist in various phases of data quality management, including methods for detection, assessment, and repair of data quality problems [46]. Further, a variety of approaches have been proposed towards the design of these methods, for example interactive [29], exploratory [14], and autonomous [1, 4, 10, 22, 43] approaches. The scope of these methods can range from being tailored for specific use-cases, to being generically applicable.

There are two aspects that significantly influence the scope of the methods: (1) the type of data, for example structured, text, graph, etc., and (2) the application domain for which the data quality methods are being designed and developed.

- **Generic methods** can be reused in a variety of application contexts or applied to a number of data types, for example detecting similarity through tokenization and set similarity measures can be applied to strings [51], records [16], and videos [32]. As such, generic methods target problems that emerge in Quadrants 1 and 3 of Figure 1.
- **Tailored methods**, on the other hand, are specific for a particular data type or application domain, for example improving the usability of RDF data [2]. Tailored methods are developed to handle problems relating to Quadrants 2 and 4 of Figure 1.

The scope of a method will influence the design of evaluations and consequently the way in which the results of the method can be utilized. The scope of the method is independent of the type of metric. For example, a duplication detection method for relational data can be measured from both intrinsic and extrinsic perspectives.

We observe further that tailored methods can be considered as a **specialization** of a respective generic method, that is, specialized for a certain application domain or data type. Whereas traditional methods have relied on well-defined constraints and design principles, e.g., functional dependencies and normalization process for relational data, in the big data scenario these constraints are largely unknown, and data of different types can be integrated and re-purposed for different applications. This makes it difficult to navigate the spectrum of methods from applicable (generic) to purposeful (tailored), indicating a need to better understand how large collections of tailored methods (e.g., duplicate detection for specific data types) can be generalized for wider applicability.

3. The Data Continuum

Data quality research and practice have been empirically evaluated with both real and synthetic data. Synthetic data can be created through a perturbation of real data that represents ground truth [36]. Alternatively, synthetic data can be entirely created through a data generator that mimics real data properties and/or through the design of a generative model by learning parameters from real data [11, 47].

Both real and synthetic data can be of a variety of data types, such as structured or unstructured, streaming or historical etc. However, there are some key features that distinguish the two:

- **Real data** are data created in the 'wild', where there is little or no influence on its generative process from the data quality method being studied. Real data provide both meaning and impact to data quality research. However, the overheads, technical and legal, in the acquisition of real data can sometimes be prohibitive. Further, in using real data to test the efficacy of a system, ground truth is not always available or may require considerable time investment to create.
- **Synthetic data**, on the other hand, are created with specific schema and data characteristics in mind. More importantly, ground truth can be easily manufactured for synthetic data, whereas it is not readily available for real data.

In general, the absence of ground truth for real data is considered an impediment in measuring the effectiveness of data quality methods. Thus, differentiating between the discovery of actual and spurious data characteristics [27] becomes difficult. Recent work on using crowd sourcing to establish ground truth for real data has helped alleviate this problem to some extent [53].

Synthetic data can provide fertile ground to study specific problems of data quality relating to accuracy (availability of ground truth) as well as performance (large volume). However, synthetic data may not adequately capture the characteristics of the problem domain the data are supposed to represent.

Table 1. Contrasting Positions

	Generic	Tailored
Intrinsic	IG	IT
Extrinsic	EG	ET

Hence, the authenticity of the results obtained on synthetic data may be questioned. Note that one may argue that real data can also suffer from similar shortcomings attributed to design decisions made at the time of real data collection.

We note that synthetic data are an **abstraction** of real data with certain well-defined properties that help to remove unnecessary complexities and provide a controlled environment for the study of specific data quality problems. Whereas real or wild data start off as being rather opaque, through various transformations, annotations, and/or creation of ground truth, they start to become more transparent.

The process of acquiring metadata whether by profiling, or talking to experts, or augmenting with other data sources, also enhances the understanding of the data and moves the data toward the transparent end of the scale.

The process of tackling the problem of opaqueness of data, especially in the absence of ground truth, is challenging and currently under-studied. There is a need to understand the implications of these abstractions of real or wild data towards the creation of curated real data or generated synthetic data, and how these abstractions impact the overall authenticity of the data pipeline.

4. CONTRASTING POSITIONS

A number of contrasting positions emerge from the dimensions discussed above. Table 1 presents a summary of these positions relating to type of metric (I: Intrinsic and E: Extrinsic) and scope of method (G: Generic, T: Tailored), corresponding to the respective quadrant of the data processing pipeline shown in Figure 1.

Note that the four positions are designed as an aid for discussing properties of data quality methods and their evaluations. It is possible, and indeed desirable, to conduct an empirical evaluation that considers both intrinsic and extrinsic metrics, considering the use of the method in both a tailored and in a more generic setting or scope, and using the data continuum from real to synthetic data.

Table 2. Relevant Papers

Position	Papers
IG	Discovering Meaningful Certain Keys from Incomplete and Inconsistent Relations [28] Data Anamnesis: Admitting Raw Data into an Organization [30] Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources [15]
IT	Quality-Aware Entity-Level Semantic Representations for Short Texts [20] Data Quality for Temporal Streams [13]
EG	Effective Data Cleaning with Continuous Evaluation [21] Benchmarking Data Curation Systems [6]
ET	Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips [18]

Nonetheless, these positions and their inter-relationships present a means of interrogating the body of knowledge on data quality and allow us to expose the role of empiricism in data quality research and practice.

In Table 2, we present a list of papers published in a recent special issue of the Data Engineering Bulletin [45] that focused on empirical research in data quality management. The list is not meant to be exhaustive, but an exemplification of the positions discussed below.

4.1 IG: Intrinsic and Generic

Generic methods for intrinsic metrics are positioned within Quadrant 1 of Figure 1. Several intrinsic characteristics of data, such as duplicates [37] and anomalies [7, 12], are generalizable across many uses of the data. In fact, any numerically representable profile of the data [8] can be generically understood and reasoned with. Thus, generic methods can handle data quality management for various data processing tasks such as extraction, integration, and aggregation (see Figure 1).

Often more sophisticated or tailored metrics use these generic methods as building blocks [48]. Re-use of such generically applicable methods prevents re-invention and more importantly provides a uniform way to compare extensions in tailored methods. Promoting the re-use of generic methods based on intrinsic metrics leads to knowledge sharing, limits unnecessary re-invention, and should be encouraged.

The exposition of intrinsic generic metrics when applied to real data can get buried in the contextual details of real data, making the results difficult to share and re-use. Whereas there are no obvious distinguishing aspects of using synthetic data for evaluating intrinsic generic metrics, the separation of the scope of the method from the nature of data may indeed present improved mechanisms for comparative studies, knowledge sharing and real progress rather than re-invention. It can be argued that evaluations on both real and synthetic data should be equally valued.

Example: Generic methods for intrinsic metrics include the discovery of certain keys, functional dependencies and other meta-data from real data, e.g., Gene Ontology [27], and/or from synthetic data, e.g., fd-reduced-30 [38].

4.2 IT: Intrinsic and Tailored

Even when intrinsic characteristics of data are generalizable across many data types, they often need to be refined in a type-specific way to get the most out of the data [47]. Tailored methods for intrinsic metrics are positioned within Quadrant 2 of Figure 1. For example, although the notion of near-duplicates is meaningful across many data types (strings, images, videos), the specifics of the data type would dictate what is considered a near-duplicate; while an edit-distance based distance measure is meaningful for strings and short text [20], allowing for differences in resolution is important for videos. Care should be taken to avoid unnecessary proliferation of metrics that are data type specific, although sometimes they may be unavoidable or even desirable.

While many generic and tailored methods can be applied on synthetic data, such data sets are often generated for specific purposes, and may not share all the characteristics of real data, so only a subset of the methods may be meaningful for a specific synthetic data set. Indeed, use of synthetic data for any purposes other than evaluation needs to be carefully considered. The specificity of the synthetic data also raises the need for a clear separation between the generative models, data quality methods, and persons responsible for the injection and the subsequent detection of errors.

Example: Tailored methods for intrinsic metrics include the method of [13] for anomaly detection that was tailored for temporal streams and applied to real NYSE data. It is worth noting that conducting robust scalability evaluations of such methods will need carefully crafted synthetic data with tunable parameters.

4.3 EG: Extrinsic and Generic

Data acquires value only when it is successfully used, whether by applications or by humans. Hence it is essential to provide extrinsic metrics that quantify the impact of poor data quality on the tasks that make use of the data, such as modeling, analysis and interpretation as depicted in Quadrant 3 of Figure 1. Quantification of how well a data quality method succeeds in delivering data of value is an important externally focused metric, but is, in general, difficult to study for real data due to its contextual distinctions and complexity. Externally focused metrics, such as putting a monetary cost on the process of data cleaning that makes the data "fit for use" for the given tasks, can play a unifying role towards making it easier to understand and measure the impact of poor data quality.

Synthetic data, in combination with generic methods, presents a controlled and simplified setting through which both internally and externally focused metrics can be studied. Further synthetic data can facilitate comparative studies in terms of the quality of the output from the method as well as its performance on larger scale [48]. It can also assist in the development of community agreed benchmarks for extrinsic metrics.

For example, data curation tools [35] can assist in automating the curation process and reducing the human effort cost (an extrinsic metric). However, quantifying the cost of automating curation against the reduction of human effort for real data can carry an un-manageable complexity [34], and thus may need robust evaluations based on synthetic data.

Example: A method for evaluating the quality of a data exchange system against required user-effort using synthetic data with community agreed parameters such as schema and relation size [6].

4.4 ET: Extrinsic and Tailored

Since the value of data is in its successful use by external tasks, one might argue that it is the task that should drive the metrics [18]. Thus, task-specific, externally focused metrics demand evaluation experiments to be based on real data. This would naturally lead to tailored methods that depend on the type of data and its use. Tailored methods for extrinsic metrics are positioned within Quadrant 4 of Figure 1.

For example, the needs of a network engineer may be significantly different from that of a market analyst; the same data set may meet the needs of one but not the other, necessitating tailored methods and user involvement. A new wave of methods and tools are emerging that endorse human-in-the-loop thinking. The ability to maintain provenance in such iterative

and interactive data curation methods thus becomes particularly important [19], as it improves both the explainability as well as refinement of the method. The resulting Assess-Clean-Evaluate cycle [21] has been adopted by current commercial data cleaning and curation platforms such as [49, 50].

Although real data are necessary to study task specific, externally focused metrics, there is little evidence of large scale sharing of real scenarios (applications, metrics and data) due to the proprietary nature and privacy concerns. The specificity of the real scenarios can also be a prohibiting factor in engaging researchers due to the infeasibility of investing significant effort to evaluate just one use-case [40]. Community approved synthetic but realistic data, meta-data and quality metrics can help overcome this problem and facilitate the development of publicly available benchmarks (like the open source iBench [5, 31]). However, the transparency of the data and meta-data generators is imperative to ensure the integrity and repeatability of the evaluation processes.

Example: Exploratory methods to guide data cleaning in spatio-temporal urban data, e.g., NYC taxi data, towards meeting analytical needs of end users [18].

5. WHAT NEXT?

We have presented two dimensions of empiricism that provide a lens to study data quality research, namely type of metrics, and scope of method, along with the data continuum based on the nature of data.

The dimensions and their inter-relationships provide a means of positioning data quality research, and help to expose limitations, gaps and opportunities. We assert that the classification serves both academics and practitioners in evaluating their contributions.

Academics and researchers can use this classification to reflect on the role of empiricism in their research, and identify gaps in their work towards a more comprehensive and impactful research agenda. Especially researchers who have focused on intrinsic metrics might consider expanding their evaluation to extrinsic metrics in order to study the impact on business or a respective application area. Similarly, practitioners can use the classification to identify opportunities to steer data quality practices into new directions and/or to achieve robust outcomes. In particular, practitioners, who primarily work with real data towards development of tailored solutions, may consider the role and benefits of carefully designed, community accepted synthetic data to convey the broader applicability of their data quality methods.

Based on our investigations, we propose a call to action to promote empiricism in data quality research. Below we outline three immediate and specific actions that can and should be taken:

Share. Enable empirical research by sharing data, metadata, code, application scenarios, and benchmarks. Such sharing is not absent, but can be further promoted by recognizing contributions of data products and benchmarks as high value. We note the recent addition in the PVLDB experiments and analysis paper track [52], and encourage other publication venues to also consider ways to recognize similar contributions from the research community.

Guide. Synthetic data can facilitate rigorous and reproducible evaluations. However, it is necessary to ensure the transparency of results drawn from synthetic or heavily curated real data. The data quality research community needs to develop guidelines for experimental design that stipulate clear separation between error creation and measurement. Such guidelines are well accepted in other disciplines where the stipulations on experimental design are widely accepted, e.g., [26].

Expand. In spite of several decades of data quality research and a large number of outstanding contributions, data quality remains one of the biggest challenges in data management, and has been further exaggerated in the age of big data. A shift towards embracing the spectrum of positions outlined above is needed, which presents a step change from current approaches that tend to be focused on specific extreme positions. Thus, the continuum from real to synthetic data is necessary for robust evaluations; both intrinsic and extrinsic metrics are needed to fully capture a data quality problem; and both generic and tailored methods are needed to balance purpose and applicability.

We invite the community to use, challenge, and refine the classification presented in this paper, and work with us to further promote empiricism in data quality research.

6. ACKNOWLEDGMENTS

We would like to acknowledge all the co-authors of the papers published in [45]. This work is partially supported by the Australian Government through the ARC-DP project DP140103171 (Sadiq, Zhou and Srivastava), and the NSF award OAC-1640864 (Freire).

7. REFERENCES

[1] Abedjan, Z., Chu, X., Deng, D., Fernandez, R.C., Ilyas, I.F., Ouzzani, M., Papotti, P., Stonebraker, M., and Tang, N. 2016. Detecting Data Errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*. 9, 12, 993-1004.

[2] Abedjan, Z. and Naumann, F. 2013. Improving RDF data through association rule mining. *Datenbank-Spektrum*. 13, 2, 111-120.

[3] Abiteboul, S., Dong, L., Etzioni, O., Srivastava, D., Weikum, G., Stoyanovich, J., and Suchanek, F. 2015. The elephant in the room: getting value from Big Data. In *Proceedings of the 18th International Workshop on Web and Databases* (2015), ACM, 1-5.

[4] Ananthakrishna, R., Chaudhuri, S., and Ganti, V. 2002. Eliminating fuzzy duplicates in data warehouses. In *Proceedings of the 28th International Conference on Very Large Data Bases* (2002), VLDB Endowment, 586-597.

[5] Arocena, P., Glavic, B., Ciucanu, R., and Miller, R.J. 2015. The iBench integration metadata generator. *Proceedings of the VLDB Endowment*. 9, 3, 108-119.

[6] Arocena, P., Glavic, B., Mecca, G., Miller, R.J., Papotti, P., and Santoro, D. 2016. Benchmarking Data Curation Systems. *IEEE Data Eng. Bull.* 39, 2, 47-62.

[7] Berti-Equille, L., Dasu, T., and Srivastava, D. 2011. Discovery of complex glitch patterns: A novel approach to quantitative data cleaning. In *27th International Conference on Data Engineering (ICDE)* (2011), IEEE, 733-744.

[8] Berti-Equille, L., Loh, J.M., and Dasu, T. 2015. A Masking Index for Quantifying Hidden Glitches. *Knowledge and Information Systems*. 44, 2, 253-277.

[9] Cao, P., Gowda, B., Lakshmi, S., Narasimhadevara, C., Nguyen, P., Poelman, J., Poess, M., and Rabl, T. 2016. From BigBench to TPCx-BB: Standardization of a Big Data Benchmark. In *Technology Conference on Performance Evaluation and Benchmarking* (2016), Springer, 24-44.

[10] Chalamalla, A., Ilyas, I.F., Ouzzani, M., and Papotti, P. 2014. Descriptive and prescriptive data cleaning. In *Proceedings of the International Conference on Management of Data (ACM SIGMOD 2014)* (2014), ACM, 445-456.

[11] Christen, P. 2005. Probabilistic data generation for deduplication and data linkage. In *International Conference on Intelligent Data Engineering and Automated Learning* (2005), Springer, 109-116.

[12] Chu, X., Ilyas, I.F., and Papotti, P. 2013. Holistic data cleaning: Putting violations into context. In *Proceedings of the IEEE International Conference on Data Engineering, ICDE 2013* (Australia2013), IEEE, 458-469.

[13] Dasu, T., Duan, R., and Srivastava, D. 2016. Data Quality for Temporal Streams. *IEEE Data Eng. Bull.* 39, 2, 78-92.

[14] Dasu, T. and Johnson, T. 2003. *Exploratory data mining and data cleaning*. John Wiley & Sons.

[15] Dong, X.L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., Sun, S., and Zhang, W. 2016. Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. *IEEE Data Eng. Bull.* 39, 2, 106-117.

[16] Elmagarmid, A.K., Ipeirotis, P.G., and Verykios, V.S. 2007. Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*. 19, 1, 1-16.

[17] Floridi, L. and Illari, P. 2014. *The philosophy of information quality*. Springer.

[18] Freire, J., Bessa, A., Chirigati, F., Vo, H.T., and Zhao, K. 2016. Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips. *IEEE Data Eng. Bull.* 39, 2, 63-77.

[19] Freire, J., Glavic, B., Kennedy, O., and Mueller, H. 2016. The exception that improves the rule. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics* (2016).

[20] Hua, W., Zheng, K., and Zhou, X. 2016. Quality-Aware Entity-Level Semantic Representations for Short Texts. *IEEE Data Eng. Bull.* 39, 2, 93-105.

[21] Ilyas, I.F. 2016. Effective Data Cleaning with Continuous Evaluation. *IEEE Data Eng. Bull.* 39, 2, 38-46.

[22] Ilyas, I.F. and Chu, X. 2015. Trends in cleaning relational data: Consistency and deduplication. *Foundations and Trends® in Databases*. 5, 4, 281-393.

[23] Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., and Shahabi, C. 2014. Big data and its technical challenges. *Communications of the ACM*. 57, 7, 86-94.

[24] Jayawardene, V., Sadiq, S., and Indulska, M. 2013. The curse of dimensionality in data quality. In *24th Australasian Conference on Information Systems (ACIS)* (2013), RMIT University, 1-11.

[25] Juran, J.M. 1989. *Juran on leadership for quality*. New York: The Free Press.

[26] Kirk, R.E. 2003. *Experimental Design*.

[27] Köhler, H., Leck, U., Link, S., and Zhou, X. 2016. Possible and certain keys for SQL. *The VLDB Journal*. 25, 4, 571-596.

[28] Köhler, H., Link, S., and Zhou, X. 2016. Discovering Meaningful Certain Keys from Incomplete and Inconsistent Relations. *IEEE Data Eng. Bull.* 39, 2, 21-37.

[29] Krishnan, S., Wang, J., Wu, E., Franklin, M.J., and Goldberg, K. 2016. Activeclean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment*. 9, 12, 948-959.

[30] Kruse, S., Papenbrock, T., Harmouch, H., and Naumann, F. 2016. Data Anamnesis: Admitting Raw Data into an Organization. *IEEE Data Eng. Bull.* 39, 2, 8-20.

[31] The iBench Project: http://dblab.cs.toronto.edu/project/iBench/

[32] Liu, J., Huang, Z., Cai, H., Shen, H.T., Ngo, C.W., and Wang, W. 2013. Near-duplicate video retrieval: Current research and future trends. *ACM Computing Surveys (CSUR)*. 45, 4, 44.

[33] Markie, P. 2004. *Rationalism vs. empiricism*.

[34] Mecca, G., Papotti, P., and Santoro, D. 2014. IQ-METER-An evaluation tool for data-transformation systems. In *30th International Conference on Data Engineering (ICDE)* (2014), IEEE, 1218-1221.

[35] Miller, R.J. 2014. Big Data Curation. In *COMAD* (2014), 4.

[36] Naumann, F. CORA Dataset: https://hpi.de/naumann/projects/repeatability/datasets/cora-dataset.html

[37] Naumann, F. and Herschel, M. 2010. An introduction to duplicate detection. *Synthesis Lectures on Data Management*. 2, 1, 1-87.

[38] Papenbrock, T., Ehrlich, J., Marten, J., Neubert, T., Rudolph, J.-P., Schönberg, M., Zwiener, J., and

Naumann, F. 2015. Functional dependency discovery: An experimental evaluation of seven algorithms. *Proceedings of the VLDB Endowment*. 8, 10, 1082-1093.

[39] Poess, M., Rabl, T., Jacobsen, H.-A., and Caufield, B. 2014. TPC-DI: the first industry benchmark for data integration. *Proceedings of the VLDB Endowment*. 7, 13, 1367-1378.

[40] Popivanov, I. and Miller, R.J. 2002. Similarity search over time-series data using wavelets. In *18th International Conference on Data Engineering (ICDE)* (2002), IEEE, 212-221.

[41] Razniewski, S., Korn, F., Nutt, W., and Srivastava, D. 2015. Identifying the extent of completeness of query answers over partially complete databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (2015), ACM, 561-576.

[42] Razniewski, S., Sadiq, S., and Zhou, X. 2016. Exploiting Hierarchies for Efficient Detection of Completeness in Stream Data. In *Australasian Database Conference* (2016), Springer, 419-431.

[43] Rekatsinas, T., Chu, X., Ilyas, I.F., and Ré, C. 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *Proceedings of the VLDB Endowment*. 10, 11, 1190-1201.

[44] Sadiq, S. and Papotti, P. 2016. Big data quality-whose problem is it? In *32nd International Conference on Data Engineering (ICDE)* (2016), IEEE, 1446-1447.

[45] Sadiq, S. and Srivastava, D. 2016. Special Issue on Data Quality *Bulletin of the Technical Committee on Data Engineering*. 39 2.

[46] Sadiq, S., Yeganeh, N.K., and Indulska, M. 2011. 20 years of data quality research: themes, trends and synergies. In *Proceedings of the Twenty-Second Australasian Database Conference-Volume 115* (2011), Australian Computer Society, Inc., 153-162.

[47] Santoro, D., Arocena, P., Glavic, B., Mecca, G., Miller, R.J., and Papotti, P. 2016. BART in Action: Error Generation and Empirical Evaluations of Data-Cleaning Systems. In *Proceedings of the 2016 International Conference on Management of Data* (2016), ACM, 2161-2164.

[48] Soliman, M.A., Ilyas, I.F., and Chang, K. 2007. Top-k query processing in uncertain databases. In *23rd International Conference on Data Engineering (ICDE)* (2007), IEEE, 896-905.

[49] Tamr: https://www.tamr.com/

[50] Trifacta: https://www.trifacta.com/

[51] Ukkonen, E. 1992. Approximate string-matching with q-grams and maximal matches. *Theoretical computer science*. 92, 1, 191-211.

[52] VLDB. 44th International Conference on Very Large Data Bases 2018: http://vldb2018.lncc.br/call-for-research-track.html

[53] Wang, J., Kraska, T., Franklin, M.J., and Feng, J. 2012. Crowder: Crowdsourcing entity resolution. *Proceedings of the VLDB Endowment*. 5, 11, 1483-1494.

[54] Wang, R.Y. and Strong, D.M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*. 12, 4, 5-33.