Data Valuation: Use Cases, Desiderata, and Approaches

Mike Fleckenstein
Data Environments & Engineering,
The MITRE Corporation, Mclean, VA
USA
fleckenstein@mitre.org

Ali Obaidi
Data Environments & Engineering,
The MITRE Corporation, Mclean, VA
USA
ali@mitre.org

Nektaria Tryfona
Electrical and Computer Engineering,
Virginia Polytechnic Institute and
State University, USA
tryfona@vt.edu

ABSTRACT

Data valuation has been given increasing thought for the past 20 years. The importance of data as an asset in both the private and public sectors has systematically increased, and organizations are striving to treat it as such. However, this remains a challenge, as data is an intangible asset. Today, there is no standard to measure the value of data. Different approaches include market-based valuation, economic models, and applying dimensions to data. This paper summarizes 18 months of research on data valuation. First, we show how we developed a Data Valuation Framework by grouping past approaches to data valuation. Second, we describe how we built and scored our own Dimensional Data Valuation Model.

CCS CONCEPTS

• **Computing methodologies**; • **Modeling and simulation**; • **Model development and analysis**; • **Modeling methodologies**;

KEYWORDS

Data Valuation, Data Monetization, Data Valuation Framework, Economic Data Model, Market-based Data Model, Dimensional Data Models

ACM Reference Format:
Mike Fleckenstein, Ali Obaidi, and Nektaria Tryfona. 2023. Data Valuation: Use Cases, Desiderata, and Approaches. In *Second ACM Data Economy Workshop (DEC '23), June 18, 2023, Seattle, WA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3600046.3600054

1 INTRODUCTION

The phrase "data as an asset" is routinely used in business today and rightly so. Many organizations are increasingly using data to augment their products. John Deere markets data management, that includes "[o]n-the-go measuring of moisture, protein, starch and oil values in wheat, barley and rapeseed/canola, as part of its product line [1]. Airbus created the Skywise data platform that aims to include producer and consumer data in exchange for access to end-to-end supply chain data [2]. Government, non-profit organizations, academia, and many companies deal almost exclusively in data, so data is their primary asset.

DEC '23, June 18, 2023, Seattle, WA, USA

ACM ISBN 979-8-4007-0846-6/23/06.
https://doi.org/10.1145/3600046.3600054

Data is bought and sold legally and illegally, insured and breached, used to create alternative currencies, even tied to physical assets in the form of deeds, licenses, and NFTs. Thus, many marketplaces and regulations have evolved around data. Yet most organizations are unable to value their own data.

Data Valuation has been given increasing thought for the last 20 years. The importance of data as an asset in both the private and public sectors has systematically increased and organizations are striving to manage it as an asset. However, this remains a challenge, as data is an intangible asset. Today there is no standard to measure the value of data.

We examined the history of methods used for data valuation and found a finite set of approaches. Our framework groups these data valuation approaches into three models:

1. Market-based models, which calculate data's utility in terms of cost and revenue, or profit.
2. Economic models, which estimate data's utility in terms of economic and public benefit.
3. Dimensional models, which extend the above models to estimate utility based on categories – both data-specific and contextual.

We found that large organizations, such as our own and governments, have a need to value individual datasets. Therefore, we set out to design and test our own dimensional data valuation model.

2 RELATED WORK

We found a relatively long history of the dimensional approach to data valuation dating back more than 40 years. Each study proposes to value data through a variety of dimensions, both directly related to data (e.g., data quality) as well as contextual (e.g., usage). A few of the studies scored the application of dimensions against small data sets. Table 1 highlights the core of these research studies since 2017. For studies prior to 2017 please see "A Review of Data Valuation Approaches and Building and Scoring a Data Valuation Model" [3]

3 USE CASES

3.1 Context, Data Sets, Perspectives, and Scoring

Our research [3] extended prior work by designing and scoring an enhanced dimensional model against a variety of data sets. We focused on two use cases:

1) How to compare the value of similar data sets, and
2) How to assess the value of a data set against existing data.

For the first use case, we compared two similar data sets for flight scheduling and navigation, and two similar data sets for voter data. Based on our experience, we realized that this approach is beneficial for sizable organizations that aim to streamline their operations by reducing the number of similar data sources, or those

Table 1: Prior Dimensional Data Valuation Research

Study	Data Value Categories and Conclusion
Brennan, et al. [4]	Categories: Operational value, replacement cost, competitive advantage, regulatory risk, timeliness, and secondarily ease of measurement, and data quality Conclusion: Reinforces a hierarchy of data value dimensions, i.e., utility, context, usage and quality, cost, as well as the use of manual survey-based methods as useful for data valuation.
Laney [5]	Categories: Intrinsic value (validity, completeness, scarcity, lifecycle), business value (relevance, validity, completeness, timeliness), performance value (relative key performance indicator benefit when leveraging information assets), cost value, market value, and economic value Conclusion: Models are imperfect and have greater utility in combination than as stand-alone. Dimensions may be modified based on organization needs. Models provide an indicator of "information asset management" maturity, which results in increased value from data.
Fleckenstein and Fellows [6]	Categories: Cost, data type (quality), maturity of data stewardship, data architecture, and data lifecycle Conclusion: Principles from physical asset valuation may be used to extract relevant dimensions related to data valuation.
Harwich and Lasko-Skinner [7]	Categories: Quality, format, ability to link data, type of data, reason of data collection, quantity, actionability, use of data, market capitalization, and relative cost of getting data elsewhere Conclusion: Public authorities should develop a clear national strategy that seeks to optimize the value of data, help the public sector when accessed for commercial purposes, and ensure the value of data is optimized between data owners, public sector, and industry.
Viscusi and Batini [8]	Categories: Information quality (accuracy, accessibility, completeness, currency, reliability, timeliness, usability, credibility, believability, reputation, trustworthiness), information structure (abstraction, codification, derivation, integration), information diffusion (scarcity, sharing), information infrastructure abstraction, embeddedness, evolving (timeliness), flexibility, openness, sharing, standardization (codification), financial value, pertinence, transaction costs Conclusion: These metrics may be useful for measuring information value. Data valuation analysis produced was limited due to the complex and multidisciplinary nature of information value. Further studies recommended to clarify data valuation categories.
Nagle and Sammon [9]	Categories: Business value (cost reduction, revenue generation, risk mitigation), acquisition (cost and legitimate need of data), level of integration (existing vs. needed), analytics effectiveness, delivery (data quality and visual impact), and level of data governance Conclusion: Data value map can be used to gain a shared understanding.

looking to substitute an existing data set with a superior one (e.g., more cost-effective, more dependable, and with lower maintenance requirements).

Our second use case was developed by collaborating with various internal projects that wanted to assess the inclusion of new data into their current data pool. To accomplish this, we utilized COVID-19 baseline data sets and supplemented them with additional COVID data. Consequently, our second use case centered on contrasting the value of existing data with that of augmented data.

After conducting our research, we had a solid understanding of the significant dimensions. Ultimately, we achieved our optimal results by posing inquiries related to ownership, cost, utility, age, privacy, data quality, as well as volume and variety. The table below depicts our final dimensions.

We developed a survey containing about 30 questions. We used our dimensional model research as our baseline and built on prior work. Our proposed data valuation is based on questions in dimensions of ownership, cost, utility, age, privacy, data quality, and volume and variety.

We utilized three different types of data sets in our study. Regarding COVID-19, we employed data sets on cases/death rates, testing, and vaccination; for flight scheduling and navigation, we utilized similar vendor-sourced data sets. As for voter data, we drew on data from two states, Ohio and North Carolina. The data sets were either publicly accessible (COVID-19 from Johns Hopkins University in 2021, and voter data from the U.S. Election Assistance Commission in 2020) or made available to us (flight scheduling and navigation data).

Our decision to incorporate both paid and free data sets was motivated by the desire to account for costs in our data valuation assessments explicitly. This approach enabled us to verify whether more expensive data sets offered higher data quality or greater data volume. We chose to utilize COVID-19 and voter data sets because they are freely accessible, extensively used, abundant, of high quality, adaptable to different viewpoints, and easy to supplement with different sources of data. This strategy enabled us to conduct several types of comparisons.

Table 2: Dimensions Used In this Study to Value Data

Dimension	Description
Ownership	Outright data set ownership plus licensing restrictions and service agreements
Cost	Addresses the cost of data set acquisition, maintenance, and replacement
Usage	Data set mission criticality, ability to integrate, usage scope, usage frequency, metadata, additional resources, expected increase in demand, and diminishing value
Age	Refresh rate and available history
Privacy	Whether the data set contains sensitive data such as Personally Identifiable Information (PII) and Protected Health Information, and meets privacy standards
Data Quality	Completeness, accuracy, currency, consistency, duplication, trustworthiness, and timeliness
Volume & Variety	Number of records, scope of information for each record, and ability to answer needed questions

3.2 Scoring

First, we assigned a raw value to each survey question. We assigned a score of 1 to the answer with the least value and increased the score by 1 for each answer that contributed more to the value. We incorporated a conversion factor to standardize the scoring process, ensuring that questions with more answers did not receive an advantage over questions with fewer answers. Finally, we applied a weight factor ranging between 1 and 5, setting the significance of each question in relation to the other questions.

Figure 1 below illustrates the dimension of data quality when comparing two comparable data sets - specifically, flight scheduling data. It is evident that data set 2 possesses superior data quality compared to data set 1. Notably, in addition to the data quality score, data set 2 also displays higher scores for cost, usage, age, volume, and variety.

Figure 2 presents a snapshot of our volume and variety dimension in comparing baseline COVID-19 data with augmented COVID-19 data. We demonstrate how the inclusion of testing and vaccination data alongside case and death rate data enhances the overall value of our analysis. It's worth noting that, apart from the data-quality score, the combined data receives a significantly higher usage score. Furthermore, since both data sets are publicly available under the creative commons license, neither cost nor ownership are relevant factors.

3.3 Perspectives

In cases where there were different perspectives, we allowed for different weights by perspective. We arrived at this approach through trial and error, realizing that certain dimensions or survey questions within a dimension might be more relevant to some organizations than others.

In assessing the value of COVID-19 data, we analyzed it from the perspectives of government, a hospital, JHU, and a public service research organization. For flight scheduling and navigation data, we scrutinized a vendor, government, and a public service research organization.

Figure 3 illustrates how various organizations assign different levels of value to the COVID-19 data set. We provide a snapshot of the volume and variety dimension of our COVID-19 evaluation from the vantage points of four distinct perspectives: government, a research organization, a hospital, and JHU. These perspectives have been inferred based on our own informed estimations.

Dimensions	Survey Question	Survey Answer	Points	Weight	Conversion Factor	Data set 1		Data set 2	
						Raw Score	Weighted Score	Raw Score	Weighted Score
Data quality	How complete is the data set? (completeness)	The data set has all data available and accessible for every record.	4	5	1.25	3.00	18.75	4.00	25.00
		The data set is partially complete. Some data is missing or unusable.	3						
		Significant data is not completely represented.	2						
		The data set is barely complete. Lots of effort is needed to make it useful.	1						
	What is the level of accuracy of the data set?	High level	3	5	1.67	2.00	16.67	3.00	25.00
		Medium level	2						
		Low level	1						
	How does the accuracy of the additional data affect the overall level of accuracy?	It increases it	3	5	1.67	NA	NA	NA	NA
		No change	2						
		It decreases it	1						
	How current is this data set?	No gap. Data is up to date.	5	3	1.00	3.00	9.00	3.00	9.00
		Lags few days	4						
		Lags weeks	3						
		Lags months	2						
		Not at all current	1						
	What is the level of consistency of the data set?	Highly consistent	3	3	1.67	3.00	15.00	3.00	15.00
		Somewhat consistent	2						
		Not consistent	1						
	Are all entities represented in this data set unique?	Yes	3	3	1.67	3.00	15.00	3.00	15.00
		Some duplicate records exist but controlled	2						
		High percentage of duplicate records exist	1						
	How trustworthy is this data set?	High. All data parts can be attributed to the original data source and date.	3	3	1.67	3.00	15.00	3.00	15.00
		Medium. Some parts of the data can be attributed.	2						
		None. Cannot confirm the origination of this data set.	1						
					Sum	17.00	89.42	19.00	104.00

Figure 1: Example snapshot of comparing two similar data sets.

Dimensions	Survey Question	Survey Answer	Points	Weight	Conversion Factor	JHU COVID-19 Case/Deaths		JHU COVID-19 Cases/Deaths and Testing and Vaccination	
						Raw Score	Weighted Score	Raw Score	Weighted Score
Volume and variety	How much does the new data add to variety of data?	Significantly adds to variety	3	4	1.67	2.00	13.33	3.00	20.00
		Somewhat adds to variety	2						
		Does not add to variety	1						
	How much does the new data add to volume of data?	The number of additional records increases data significantly.	4	4	1.25	2.00	10.00	4.00	20.00
		The number of additional records increases data somewhat.	3						
		The number of additional records increases data slightly.	2						
		The data set adds no additional records.	1						
	To what extent does the new data or the additional records help answer more questions (significantly, somewhat, little, or none)?	Significantly	4	5	1.25	3.00	18.75	4.00	25.00
		Somewhat	3						
		Little	2						
		None	1						
					Sum	7.00	42.08	11.00	65.00

Figure 2: Example snapshot of adding data to existing data pool.

3.4 Findings and Discussion

Our scoring verified some typical assumptions. For example:

When comparing two similar data sets, higher cost is associated with greater data quality, increased usage, longer history, and a

Dimensions	Survey Question	Survey Answer	Points	Weight	Conv. Factor	JHU COVID-19		Weight	JHU COVID-19		Weight	JHU COVID-19		Weight	JHU COVID-19	
						Gov			Research Org.			Hospital			JHU	
						Raw Score	Weight Score		Raw Score	Weight Score		Raw Score	Weight Score		Raw Score	Weight Score
Volume and variety	How much does the new data add to variety of data?	Significantly adds to variety	3	4	1.67	3.00	20.00	4	3.00	20.00	4	3.00	20.00	4	3.00	20.00
		Somewhat adds to variety	2													
		Does not add to variety	1													
	How much does the new data add to volume of data?	The number of additional records increases data significantly.	4	4	1.25	4.00	20.00	4	4.00	20.00	4	4.00	20.00	4	4.00	20.00
		The number of additional records increases data somewhat.	3													
		The number of additional records increases data slightly.	2													
		The dataset adds no additional records.	1													
	To what extent does the new data or the additional records help answer more questions?	Significantly	4	5	1.25	4.00	25.00	5	4.00	25.00	5	3.00	18.75	5	3.00	18.75
		Somewhat	3													
		Little	2													
		None	1													
					Sum	11.00	65.00		11.00	65.00		10.00	58.75		10.00	58.75

Figure 3: Example of different perspectives.

higher volume and variety of data. This observation is demonstrated in the comparison of the two flight scheduling data sets.

When it comes to flight navigation, data set 1 was obtained under a free license, whereas data set 2 was purchased at a cost. Despite being more expensive to acquire, data set 2 exhibits notably higher usage rates, due to factors such as its inclusion of metadata, ease of integration with other data sets, availability of additional resources, and overall popularity. Furthermore, data set 2 was also rated higher in terms of data quality, volume, and variety.

When comparing data sets that augment existing data, the combined data sets generally receive higher scores. This trend is evident in the COVID-19 data, where the inclusion of testing and vaccination data alongside case and death rate data results in significantly higher usage rates. However, it should be noted that our example involved adding a relatively small data set to another small data set. We acknowledge that adding a small data set to a large data pool may not always yield the same outcome.

Context is important. For example:

1. For flight scheduling data, we scored three perspectives: vendor, government, and research organization. One of our usage questions revolved around frequency of use, which is daily for the government and the research organization but rare for the vendor. This implies a lower value for the vendor, which is counterintuitive since the vendor stands to profit from the data set. Thus, the vendor might give this question a low weight or no weight at all.
2. For privacy, we scored for Personally Identifiable Information (PII) and whether the data set met required privacy compliance. In the case of voter data, both data sets contained PII, which we valued higher. Such data is useful for a variety of analyses. However, meeting privacy compliance might require an organization to mask PII data, in which case it may value masked data higher.
3. The desire to answer new questions through analytics for COVID-19 data likely differs across stakeholders (e.g., government, research organization, hospital, and JHU). While we did not engage stakeholders from each of these organizations, we assumed that COVID-19 data sets were more likely to be used for analytics by the government and research organizations.

Our team experimented a lot with applying different weights. In the end, we applied weights that we thought were reasonable. We also concluded that weights are very context specific. For example, cost may matter much more to a particular stakeholder or in a particular context. We realized that weights may also differ by perspective. While our weights fell between 1 and 5, we encourage users to experiment with weights in ways that work in their context. The survey acts as a blueprint for stakeholders to register their professional opinion on the value of data sets.

There were instances we were not able to investigate. For example:

1. We were able to determine the relative value of two different data sets based on dimensions using a score-based approach. Translating that value into monetary terms likely requires the secondary application of a market-based or economic model to a given data set.
2. We anticipate that, given a sufficient database of survey responses, it will be possible to apply artificial intelligence and machine learning to these surveys so that they can be more automatically completed. This requires logging many additional use cases.

4 ADDITIONAL CONSIDERATIONS

4.1 Comparison to Real Estate Valuation

Early on, we saw a conceptual parallel to a real-estate appraisal. In real estate, assets – in this case homes – are valued and compared based on a simple set of questions, both inherent to the structure as well as its surroundings. This is similar to valuing data, where some value is based on the data itself and additional value is based on the context in which data is used.

To aid the mapping of questions from real estate to dataset, we classified real estate questions into categories like cost, structure, site, neighborhood, and so on. Some of these categories mapped more directly (e.g., cost) and others were more derived (e.g., neighborhood density and growth rate translated into availability of and demand for similar data). Below are some examples of this mapping[1].

While this type of comparison yields insight into possible data valuation dimensions, we understand that, in order for data to be valued like real-estate, a large database of comparables must exist and this is something we don't have for data.

4.2 Artificial Intelligence and Machine Learning

In our research, we worked with a finite collection of relatively small data sets and we manually scored answers to our questions. However, we envision a future in which questions can be answered more automatically based on artificial intelligence (AI) and machine learning (ML).

Successful application of AI/ML ultimately depends on a large data set to which answers to questions have already been scored. Such a score could be a point score or a dollar value score as, for example, in a database of real-estate comparables.

[1]For the full mapping please contact the authors

Table 3: Real-Estate Dimension to DV Dimension Mapping

Real Estate Valuation Dimension		Data Valuation Dimension
Structure	Square footage	Volume & variety
	Finished area	Data quality; completeness
	Builder / realtor / inspector rating	Data quality; trustworthiness
	Evidence of infestation / dampness / settlement; flood zone	Usage; diminishing value
	Age	Age
Ownership	Owner of public record	Ownership
Cost	Sale price; cost to build; replacement cost; price per sq. ft.	Cost
	Cost to build	Cost of data acquisition
	Cost to replace	Replacement cost
	Taxes, utilities, HOA, assessments	Maintenance cost
Neighborhood	Density; growth; property value trend; comparables	Demand; adoption; access frequency; usage
Site	Area	Volume & variety
	Zoning; easements; encroachments	Ownership; licensing restrictions
	Shape; view; improvements	Data Quality
	Off-Site Improvements (Street/Alley; Public/Private)	Usage

The introduction of new data sets can then leverage existing answers to questions using AI/ML. We discuss several possible approaches to creating and maintaining a large repository of valued data sets in the next section.

5 NEXT STEPS

We see continued increasing desire to value data and many opportunities for innovative approaches to data valuation going forward. A coordinated approach to real-estate valuation dates back to the late 1800s when brokers began compensating each other for help selling their properties [10]. Today the Multiple Listing Service (MLS) is a decentralized network of databases, governed by data standards [11], that has expanded into many enhanced listing services (e.g., Realtor.com, Yahoo, Google, Craigs List, Zillow, Trulia). In addition, local city and county databases exist for the purpose of tax-based real-estate valuation. These databases include the prices at which real-estate sells. This has resulted in the ability to value real-estate in monetary terms.

We might approach creating a similar integrated set of data valuation databases through the creation and application of standards – for example, by answering a standard set of questions – to a large collection of data sets. This might even allow data set suppliers to provide some of this information, in terms of metadata.

Some of the data needed is already available, albeit in proprietary pockets. Future research might attempt to gather and co-locate some of this data, including pricing and cost data.

For example:

1. Organizations that sell their data (e.g., credit score, market measurement) price their data on cost and profit.
2. Insurers that provide policies to cover data breaches have both policy pricing and claim (i.e., the cost of data breach) data.
3. Weather data is produced by public entities and enhanced by private ones.
4. Evaluations of the actual (vs. projected) impact on the economy due to the release of key government data can yield data set dollar valuations based on publicly available information.

Gathering this type of data merges the market-based, economic, and data valuation models. This provides the most comprehensive approach to data valuation, adding the benefit that data could be valued in monetary terms.

ACKNOWLEDGMENTS

The work behind this paper was funded by the MITRE Corporation Innovation Program (MIP). We thank Dr. Nitin Naik and Dr. Kris Rosjford for useful insights and discussions.

REFERENCES

[1] John Deere Data Management at https://www.deere.com/en/technology-products/precision-ag-technology/data-management/, accessed April 2023
[2] Airbus Skywise at https://aircraft.airbus.com/en/services/enhance/skywise, accessed April 2023
[3] Fleckenstein, M., Obaidi, A., & Tryfona, N. (2023). A Review of Data Valuation Approaches and Building and Scoring a Data Valuation Model. Harvard Data Science Review, 5(1). https://doi.org/10.1162/99608f92.c18db966
[4] Rob Brennan, *et al.*, "Exploring Data Value Assessment: A Survey Method and Investigation of the Perceived Relative Importance of Data Value Dimensions," Dublin City University, Trinity College Dublin, University College Cork, 2020
[5] D. Laney, "Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage," Gartner Research, 2018
[6] M. Fleckenstein and L. Fellows, "Modern Data Strategy," Springer, 2018
[7] Eleonora Harwich and Rose Lasko-Skinner, "Making NHS Data Work for Everyone," section 2.2.2., December 2018
[8] Gianluigi Viscusi and Carlo Batini, "Digital Information Asset Evaluation: Characteristics and Dimensions," Working Paper, EPFL and University of Milano-Bicocca, March 28, 2017
[9] T. Nagle and D. Sammon, "The Data Value Map: A Framework For Developing Shared Understanding On Data Initiatives," ECIS 2017 Proceedings
[10] "Multiple Listing Service (MLS): What Is It at https://www.nar.realtor/nar-doj-settlement/multiple-listing-service-mls-what-is-it, accessed Apr 2023
[11] Real Estate Standards Organization (RESO) Data Dictionary at https://www.reso.org/data-dictionary/, accessed Apr 2023.

DISCLOSURE STATEMENT

The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

Approved for public release. Distribution unlimited 23-00075-1