Journal of International Technology and Information Management

Volume 24 | Issue 3 Article 1

2015

Optimizing Data Management with Disparities in Data Value

G. Shankaranarayanan
Babson College

Adir Even
Ben Gurion University of the Negev

Paul D. Berger
Bentley University

Recommended Citation

Shankaranarayanan, G.; Even, Adir; and Berger, Paul D. (2015) "Optimizing Data Management with Disparities in Data Value," *Journal of International Technology and Information Management*: Vol. 24: Iss. 3, Article 1.
Available at: http://scholarworks.lib.csusb.edu/jitim/vol24/iss3/1

Optimizing Data Management with Disparities in Data Value

G. Shankaranarayanan
Babson College
USA

Adir Even
Ben Gurion University of the Negev
ISRAEL

Paul D. Berger
Bentley University
USA

ABSTRACT

When there is a disparity in the value of different data records and fields, there is a need for an optimization of data resources. Not all data necessarily contribute the same value. It depends on the usage of the data, as well as a variety of other factors. This paper presents models for optimizing data management in the presence of a disparity between the values contributed by different data. We expound on what disparity of data value represents and illustrate models to derive a numerical measure of such disparity. We then use real-world data from a large data resource used to manage alumni relations, and demonstrate our optimization methods and results. We then discuss the tradeoffs involved between value and cost, and the implications for data management, both in this real-world context and in general.

KEYWORDS: Disparity, Data Value, Data Management, Optimization

INTRODUCTION

Organizations make large investments in technology to manage data, a critical organizational resource. In making these investments organizations have to carefully evaluate the information systems and technology that are used to manage the data. Typically, such evaluations are primarily based on technical requirements such as storage capacities and processing speeds and on functional requirements such as presentation formats (e.g., dashboards and visualization) and business needs such as speed of delivery and search capabilities. In this paper, we suggest that evaluating data management systems must consider yet another perspective, economic aspects. In no way minimizing the importance of technical and functional aspects, we suggest that the design and management of data resources ought to also consider the cost-benefit tradeoffs associated with managing data resources. To emphasize this perspective we argue that all data should not be treated as contributing equally to the benefit derived from using a data resource and that some records in a dataset may contribute more to benefit than others. We refer to this as disparity in the value derived from the data (or value disparity). We believe that understanding value disparity and the associated value/cost tradeoffs, has important implications for data management. Not only can it impact how we use data, it can also impact the design and management of data resources and associated information systems.

We examine value disparity in a large data resource used to manage alumni relations. Using this as a context, we first evaluate value disparity and show that it is significantly large in this data resource. We then describe the current data acquisition and management policies, discuss the relationship between the policies and the value disparity identified, and in this context, highlight related implications for data management including how the data resource may be optimized considering an economic perspective. The rest of the paper is organized as follows. First we describe the research relevant to value/cost tradeoffs in data management to define the scope of our research. We then develop our models for assessing the magnitude of value disparity. We further illustrate the application of our models using sizable samples from the data resource for managing alumni relations. We finally discuss the implications of understanding value disparity for data management and conclude with a discussion on the limitations of our research.

BACKGROUND

Managing data efficiently and effectively helps organizations realize the business value of the data. Data contributes to business value in multiple different ways. Data supports managing operational activities such as tracking supply chain activities (Gattiker & Goodhue, 2004) and customer relationships (Roberts and Berger, 1999). Data also helps organizations gain competitive advantage through analytics (Davenport, 2006) and decision support (March & Hevner, 2007, Ramakrishnan et al., 2012). Analytics is supported today by business intelligence tools and data warehouses that collect and manage large data resources to provide a foundation for analytics and visualization (Mannino et al., 2008; March & Hevner, 2007). Finally, organizations acquire and sell data (e.g., AC Neilsen) and data serves as a revenue generator. To understand the business value of data, in this paper, we examine the value gained by the use of the data, similar to what is described above.

Managing data also involves costs. Organizations invest in sophisticated storage systems to store, manage, and efficiently retrieve data, complex applications for processing and delivering data to analysts and end-users, business intelligence tools to gain competitive intelligence through effective analysis of data, as well as procedures and policies to manage the quality and security of data. Complex data environments, such as a data warehouse supplemented by business intelligence tools involve a variety of cost factors that significantly affect operational efficiency (Mannino, 2008; West, 1994). Of course, there is also the cost associated with data acquisition. We take these and other similar cost factors to examine the costs associated with managing data. We classify the costs into fixed and variable costs. In this paper, we assume fixed costs as costs that do not change with the volume (measured by number of records) of data managed (acknowledging that when capacity is exceeded by volume, fixed costs are impacted to some degree). Examples of fixed cost include infrastructure costs, purchase costs for commercial software, design costs, development costs, and overheads associated with management of these. We assume variable costs as costs that monotonically increase with the data volume such as data acquisition costs, costs for cleansing, and even costs associated with adding storage.

From a technical perspective, the design of data management environments looks at design for storage capacity requirements, processing speed, type and quality of visualization and the choice of analytical tools. From a functional perspective the design looks at what data to capture and how it might be used. Literature in data management covers a number of methods to identify design and implementation requirements for data use (such as right data, right format, right delivery etc.)

(e.g., Garcia-Molina et al., 2002; Date, 2004) These methods help manage data from an end-user's perspective ensuring effective support for the end user to use the data. Functional perspective may further include the need for operation efficiency (Mannino et al., 2008), improved decision capabilities (March & Hevner, 2007), and competitive pressure from the industry (Ramakrishnan et al. 2012). Today, this includes the burgeoning area of "big data" (Weinberg et al., 2013).

Decisions regarding the design and administration of data management environments not only affect the cost of such environments but also the value gained from the data. It is necessary to examine the interplay between the cost (involved with implementing the technical requirements) and benefit (derived from the functional use of the data) perspectives and hence we posit the need for the economic perspective and corresponding optimization. We further state that optimizing the economic outcome should be an important objective in data management. Economic outcome can be improved by increasing business benefits derived from data, conceptualized as *value*, and by reducing *costs*. Studies have shown that economic tradeoffs can direct data management decisions – e.g., the optimization of data processes (Ballou et al., 1998), data retention policies (Kalfus et al., 2004), the configuration of data environments (Even et al., 2007) and the acquisition of data (Saar-Tsechansky & Provost, 2007; Zheng & Padmanabhan, 2006).

Information resources contribute to value through usage and experience. This value reflects benefits such as improved decisions or willingness to pay (Ahituv, 1980). The value is often viewed as the difference between the benefits derived from having full information versus partial or no information (Boland, 1985). It depends on the context of use and requires successful integration with complementary resources. The value of complete information in the context of data collected from users' activity for e-CRM models has been assessed empirically (Padmanabhan et al., 2006), where the benefits of using data from multiple sites (more complete information) was shown to be as high as 50%, depending on the problem context and performance metrics adopted. We, in this paper, attempt to determine the value of data based on the data contribution in some specific context. Our determination of the value is context dependent. Further, our approach is similar to that of Padmanabhan et al. (2006) in that we attempt to determine attributes that may have a bigger impact on user behavior than others – thus, classifying attributes to differentially manage them, as part of the optimization process.

We attempt to identify the contribution of data to value (value contribution) and associate this contribution to data management decisions. We specifically examine the magnitude of *value disparity* – whether value contribution is similar for all records in a dataset, or concentrated in a relatively small subset, and emphasize how that affects the optimization process. The value function (called "utility" in (Ahituv, 1980)) can serve as a tool for mapping the configuration of information technology/systems attributes to tangible value within specific usage. For modeling the distribution of value and analyzing the magnitude of value disparity, we adapt Lorentz's curve, and Gini's index - statistical tools used to analyze social and economic inequality in large populations. The Gini index is frequently used for assessing data irregularities in data mining applications (Schechtman, 2008), and for predicting customer contributions (Even et al., 2010).

A fundamental notion in statistics is that for inference purposes, a sample (a subset of the population) is superior to a "census" (entire population) from a cost/benefit perspective (Boland, 1985). The expected value of sample information (EVSI) increases with the number of samples, but at a decreasing rate. It eventually reaches a point at which the marginal cost exceeds the marginal value. EVSI, and the associated ENGS (expected net gain from sampling) curves, are established tools for evaluating these tradeoffs (Jagannathan, 1985)). Although, in general, our value/cost evaluation follows a similar approach, ours has two different features. First, optimal sample-size assessment using EVSI and ENGS does not embrace value differences, but rather, weights all samples equally. It further suggests that decision performance may not be significantly improved by increasing *the number* of samples beyond a certain point. We argue that there is an inherent disparity in the value of data items as some offer a higher contribution in certain decision contexts than others (e.g., more recent data is most often more valuable than less recent data). We suggest that, for data management, usage is not affected solely by *the number of items*, but by the identification of *the right items* - those with the highest value contribution - and possibly managing them differently. Second, in EVSI/ENGS studies, the common context for assessing optimal sample size is data *yet to be collected.* Although our value/cost assessment does apply to the acquisition of *new data resources*, it also has significant implications for decisions regarding data that *has already been collected*.

The notion of *value* has been used in active learning research to improve data acquisition (Provost & Fawcett, 2013; Saar-Tsechansky & Provost, 2007; Zheng & Padmanabhan, 2006). Applying a trained-model that considers costs and benefits, Provost (2005) suggests acquiring data that maximizes the expected-value for data mining applications. As noted in Zheng and Padmanabhan (2006), techniques to determine the value of data fall under two classes: heuristic and optimization-based. An exemplar of the heuristic approach is Query-By-Committee (QBC) (Freund et al., 1997). It obtains predictions on the unseen data from several models (members of the committee). The set of data viewed as offering the highest value is one that creates the most disagreement amongst the committee members. In optimization-based determination of value, an objective function is employed and the data that offers the highest value is one that optimizes the objective function.

Arguing that firms need to augment their data with additional data to build a superior model, Saar-Tsechansky and Provost (2007) present an active learning solution to selectively acquire data that optimizes model performance (maximize value while keeping the costs down). The solution they propose applies Goal-oriented Active Learning (GOAL), for a predictive model that targets decision-making in a direct-marketing context. Here, value is measured using estimates of benefit (from a new customer) and the costs in acquiring new customers.

Our model requires an estimate for value and (as explained later) in our study of donor-records, we estimate value based on past donations. We attempt to assign value to records in a dataset and use value to distinguish between records in that dataset. As our objective is to highlight the disparity in value of records and its implications for managing records, we do not consider varying the accuracy of estimates. Different from the research described above, the objective of our framework is to classify the data records for differentially managing them and not for inducing predictive models. Further, while we believe it can be incorporated, our approach does not include active learning. We have hence not compared active learning methods to our approach described in this paper.

DISPARITY IN THE VALUE OF DATA RESOURCES

The value of data is a measure of the contribution of that data to business value considering both current and potential contributions. The question is whether the overall value derived from the use of a dataset depend primarily on the entire dataset or only on a smaller subset, given that all records in that dataset are not equally important? Or, does a variation of the 80/20 rule essentially apply? We interpret this question as reflecting the magnitude of disparity in the value contributed by the records within the dataset. We first describe value disparity and illustrate it with a simple example. We then develop analytical tools for modeling and measuring value disparity in large datasets.

Value Attribution and Disparity

Consider a tabular dataset with all records having the same set of attributes but having different values for those attributes – identical structure with varying content. The variation in content may affect the relative importance of records and hence their contribution to value. Our assessment methodology is based on attributing value to data records. We attribute as value a numeric measure that reflects the relative contribution of each record for business use. The context in which the dataset is used determines the estimate of value and the attribution of value to the records in the dataset. This is not new as examples of attribution methods, reflecting relative value, have been discussed in the literature and may be adapted for the purpose of attributing value - e.g., Customer Lifetime Value assessment (Berger & Nasr, 1998), Recency/Frequency/Monetary (R.F.M. reflecting recent purchases, frequency of purchases, and the monetary value of purchases) analysis in database marketing (Berger and Magliozzi,1993; Roberts and Berger, 1999), and new customer acquisition (Saar-Tsechansky & Provost, 2007). The estimated and attributed value can reflect actual monetary value in some contexts (e.g., potential to generate revenue). As explained later, assessing economic tradeoffs requires measuring both value and cost along the same monetary scale. And, indeed, the tools for modeling and assessing disparity described in this section do not depend on the value units. For brevity, this work describes the attribution of a single value variable, which represents one usage context, or an aggregation of multiple usages (e.g., by considering a sum of independent random variables, each reflecting a different usage.)

We consider a dataset with N records (indexed by *[n]*), and assign each record a non-negative value measure ($v_n \geq 0$), reflecting the relative value of record *[n]* in the evaluated usage. We assume no interaction effects between usages and, hence, sum the record values, to arrive at the overall dataset value: $v^D = \Sigma_n v_n$. The dataset value v^D is at its maximum when the entire dataset is available/used and may be reduced if some records are chosen not to be used (perhaps by their cost not warranting them to be used, if the dataset is not "wholly owned" by the user.) A simple value allocation may assign an identical value per record (i.e., a constant $v_n = v^D/N$). This "naive" allocation rarely reflects real-world use, as records differ in importance and value contribution. Another, and perhaps less naïve allocation, might, in a given application, assign value in proportion to "recency". The issue with this is more subtle, and reflects an underlying assumption of temporal linearity which may not be entirely accurate.

For a large dataset (large N), the distribution of value is modeled as a random variable v with a probability density function (PDF) $f(v)$. From the PDF we can calculate the mean $\mu = E[v]$, the cumulative distribution function (CDF) $F(v)$, and the percent point function (PPF, the inverse of CDF), $G(p)$. Here we demonstrate the computations first for the continuous *Pareto distribution*, and then for a *discrete distribution* - both used later to analyze disparity of value in a real-world database. Similar computations can be applied to virtually any other statistical distribution (e.g., Uniform, Exponential, Weibull).

The Pareto distribution is commonly used in economic, demographic, and ecological studies. It is characterized by two parameters: the highest probability is assigned to the lowest possible value of $Z>0$ (see equation (1); Z can be arbitrarily close to 0). The probability declines as v grows and the parameter $w \geq 1$ defines the rate of decline:

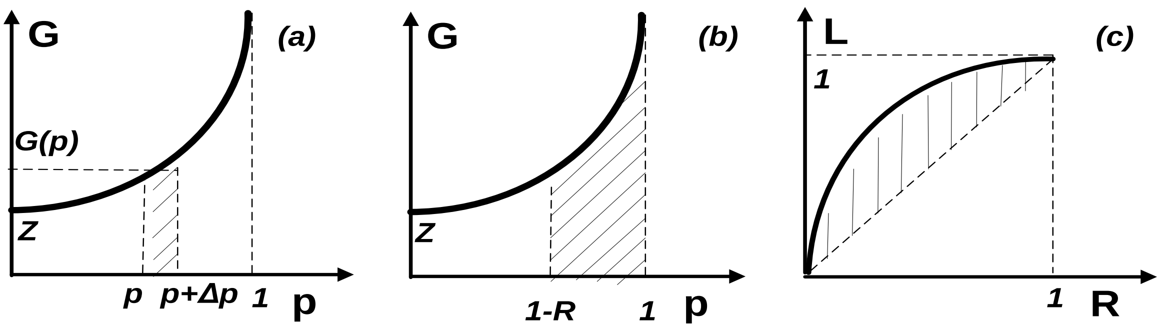
$$f(v) = wv^{-(w+1)} \big/ Z^{-w} \;\; v \geq Z; \;\; F(v) = 1 - (v/Z)^{-w} \;\; v \geq Z; \;\; G(p) = Z \big/ (1-p)^{1/w}; \;\; \mu = wZ/(w-1) \qquad \textbf{(1)}$$

A variable with a discrete distribution takes a value from a finite set of J possible values $v_1 \ldots v_J$ (the index reflects sorting in increasing order), with probabilities of $p_1 \ldots p_J$, respectively $(\Sigma_j\, p_j=1)$. It is characterized by:

$$f(v) = \begin{cases} p_j & v = v_{j=1..J} \\ 0 & otherwise \end{cases} \qquad\qquad F(v) = \begin{cases} 0 & v < v_1 \\ \sum_{k=1}^{j} p_k & v \in \left[v_j, v_{j+1}\right)_{j=1..J-1} \\ 1 & v \geq v_j \end{cases}$$

$$G(p) = v_j \quad p \in \left(\sum_{k=1}^{j-1} p_k, \sum_{k=1}^{j} p_k\right]_{j=1..J} \qquad \mu \;=\; \sum_{j=1}^{J} p_j v_j \qquad \textbf{(2)}$$

Figures: 1a, 1b, and 1c: Obtaining the Cumulative Value Curve.



To assess the extent to which records vary in their value, we define R, the proportion of highest-value records, as a $[0,1]$ ratio between the N^* records of highest value (i.e., the top N^* when rank-ordered in descending order) and N, the total number of records (e.g., $R=0.2$ for a dataset with $N=1{,}000{,}000$ records and $N^*=200{,}000$ records that offer the highest value of the $1{,}000{,}000$). The *cumulative value curve L(R)* is a $[0,1]$ proportion of the overall value as a function of R. $L(R)$ can be calculated from the percent point function $G(p)$. For a large N, the added value for a small probability interval $[p, p+\Delta p]$ can be approximated by $N \bullet G(p) \bullet \Delta p$ (Figure 1a).

Letting *Δp*→0, and integrating the PPF over *[1-R, 1]* (Figure 1b), and dividing the result by the total value (approximated by μN), we get the cumulative value curve *L(R)* (Figure 1c):

$$L(R) = N\int_{1-R}^{1} G(p)dp \Big/ N\mu = \int_{1-R}^{1} G(p)dp \Big/ \mu \text{, where,} \qquad \textbf{(3)}$$

R – The *[0,1]* proportion of highest-value records
L(R)- The cumulative value curve of the value variable *v* as a function of *R*, within *[0,1]*
N - The number of dataset records
v, μ - The value variable and its mean
G(p) - The proportion point function of the value variable *v*

The curve *L(R)* is defined for *[0,1]*, where *L(0)=0* and *L(1)=1,* and does not depend on *N* or on the value unit. The curve is calculated by "backwards integration" over *G(p),* which is monotonically increasing; hence, it is monotonically increasing and concave within *[0,1]*. The first derivative of *L(R)* is therefore positive and monotonically decreasing, and the second derivative is negative. The maximum point of the curve (i.e., *L(1)=1*) corresponds to the maximum possible dataset value v^D, and the curve reflects the maximum portion of overall value that can be obtained by the partial dataset – i.e., when only a portion *R* of the dataset is available, the value of $v^D L(R)$ can be achieved at best.

The cumulative value, *L(R),* is equivalent to the Lorentz's curve, a statistical tool for modeling disparity in value distributions. The Gini index (or coefficient), which is derived from Lorentz's curve, is a commonly used measure of disparity. This index *(φ)* measures the relative area between the curve and the 45^o line (i.e., *f(R)=R*). This area is highlighted in Figure 1c, and can be calculated by:

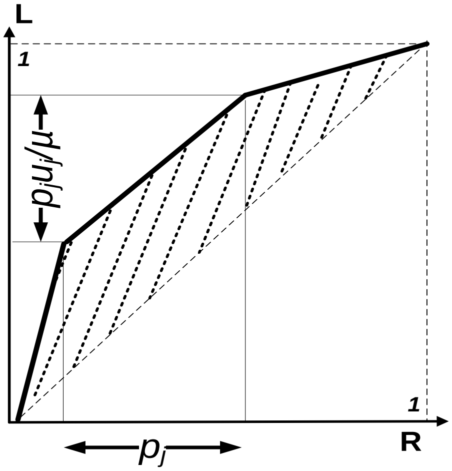
$$\varphi = \left(\int_0^1 L(R)dR - \int_0^1 RdR\right)\Big/\int_0^1 RdR = \left(\int_0^1\left((1/\mu)\int_{1-R}^1 G(p)dp\right)dR - 0.5\right)\Big/0.5 =$$
$$2\int_0^1\left((1/\mu)\int_{1-R}^1 G(p)dp\right)dR - 1 = (2/\mu)\int_0^1 pG(p)dp - 1 \qquad \textbf{(4)}$$

The value of *φ* is within *[0,1]*, where a higher value indicates a greater disparity. The lower bound, *φ→0,* indicates perfect equality – dataset records with identical and deterministic values and a curve that approaches *L(R)=R*. The upper bound, *φ→1,* indicates a high degree of disparity - a small portion of records with a relatively high value, while the value of most other records is substantially lower. The corresponding curve in this case approaches *L(R)=1* (with, technically, a vertical line at *R=0*, rising to *L(R)=1*). The curve and the index can be further evaluated for specific distributions and can be often expressed using a closed analytical form. For the Pareto distribution, the evaluations are:

$$L(R) = \frac{1}{\mu}\int_{1-R}^{1} G(p)dp = \frac{w-1}{wZ}\int_{1-R}^{1}\frac{Z}{(1-p)^{1/w}}dp = R^{1-\frac{1}{w}};\ \varphi = \frac{2}{\mu}\int_0^1 pG(p)dp - 1 = \frac{1}{2w-1} \qquad \textbf{(5)}$$

The Pareto curve and the index do not depend on the minimum value *Z*, but only on the decline rate *w*. Disparity decreases with *w*, where *w=1* indicates the highest possible disparity *(L(R)=1, φ=1*). Conversely, with *w*→∞, *L(R)*→*R* and *φ*→*0*. The value now is approximately identical for all instances: *f(v)≈1*, for *v≈Z* and *~0* otherwise (i.e., *v≈Z* with probability *~1).*

Figure 2: The Cumulative Value Curve for a Discrete Distribution.



The cumulative value for a discrete distribution is a piecewise-linear curve (Figure 2), in which each segment is associated with a single value in the set of *J* possible values. The curve is obtained by backwards integration of the PPF; hence, the segments are sorted in decreasing order of value (i.e., in a reverse order of the index *[j]*). The length of the horizontal axis per segment is the relative proportion of the dataset, or the probability p_j associated with the value v_j. The length of the vertical axis of each segment is the relative value *[j]*: $(p_j*v_j)/ \Sigma_j(p_j*v_j)=(p_j*v_j)/\mu$. It can be shown that the corresponding disparity index (the relative size of the shaded area in Figure 2) can be calculated as:

$$\varphi = 1 + \frac{1}{\mu}\sum\nolimits_{j=1..J}\left(p_j^2 v_j\right) - \frac{2}{\mu}\sum\nolimits_{j=1..J}\left(p_{J-j+1} v_{J-j+1}\sum\nolimits_{w=1..J-j} p_w\right) \qquad \textbf{(6)}$$

When using a binary classification, ("High" vs. "Low"), this expression can be simplified to $\varphi=p_2(v_2/\mu-1)$, where v_2 is the higher value among the two, and p_2 is the associated probability. When both values are equal (i.e., $v_1=v_2=\mu$), $\varphi=0$; and when p_2→*0* and v_1→*0*, *φ*→*1*.

EVALUATING VALUE DISPARITY IN ALUMNI DATA

To illustrate assessment of value disparity, we use data samples from a real-world data resource used for managing alumni relations. This resource and associated system are critical for the organization, as gifts by alumni, parents and friends account for a majority of its revenue. It is used for managing donors, tracking gift history and managing pledge campaigns. We evaluate large samples from two key datasets:

(a) Profiles (358,372 records) is a dataset that captures donor profile. It has a unique identifier (Profile ID), and includes a number of descriptive attributes.

Table 1: Evaluated Attributes in the Profile dataset.

Category	Attribute	Distinct Values	Description
Graduation	Graduation Year	1864 to 2007	The year in which a person has graduated
	Graduation School	33	The primary school of graduation
Demographics	Gender	2	Male or Female
	Marital Status	7	Marital status
	Ethnicity	7	Ethnic group
	Religion	31	Religion
	Occupation	117	The person's occupation
	Income	3	Income-level (High, Medium, or Low)
Location	Home Country	212	The country of the home address
	Home State	75	The state of the home address
	Business Country	212	The country of the business address
	Business State	75	The state of the business address

We selected the profile attributes used in this evaluation and shown in Table 1 based on inputs from key users who specified that these attributes were the most relevant and important. These attributes were extensively used in classifying profile-data and in managing alumni relations. These attributes can be placed into three categories: *(a) Graduation attributes*: *Year* and *School* are typically included when a record is added to the dataset. These two rarely change. *(b) Demographic attributes*: Some of these attributes exist when a record is added (e.g., *Gender, Marital Status, Religion*, and *Ethnicity*). Some others such as *Income* and *Occupation* are added later on. Some demographics (e.g., *Marital Status* and *Income*) may change over time, and *(c) Location attributes*: *Home* address (including *Country* and *State*) is typically included when a record is added. *Business* address is added subsequently; and both addresses may change. Each of these attributes is associated with a value domain, which includes a discrete set of possible values. For most attributes (except for the *Graduation Year*), the associated value domains are defined in lookup tables that list all the possible values along with some descriptive information.

Another attribute, *Prospect*, which is binary (0 or 1) and used in classification, reflects two fundamentally different usages of the data. "Prospects" (*11,445* records, *~3%* of the dataset) are donors who have made large contributions or are assessed to have a potential for substantial gift in the future. Prospects are not approached via regular campaigns. Each prospect is assigned a staff responsible for maintaining an ongoing contact (such as, for example, invitations to special fund-raiser dinners and tickets to shows/games). Donors classified as "non-prospects" (*~97%* of the dataset) are typically approached via pledge campaigns, each targeting a large donor base (e.g., via phone, mail, or Email).

(b) Gifts (1,415,432 records): this dataset captures the gift transactions. It has a unique identifier (*Gift ID*), a *Profile ID* (foreign key that links each gift transaction to a specific profile), *Gift Date, Gift Amount* and a few administrative attributes that describe payment procedures. Importantly, in this study we evaluate disparity in the *profiles* dataset. The *gifts* dataset is used for assessing the *value* of each profile.

The sample data is from a 24 year period, and represents approximately *40%* of the data volume in the actual system. A large number of records (data from an older system) was added (*203,359* profiles, *405,969* gifts) right after the implementation of the new system. Subsequently, both datasets have experienced a gradual growth. While *Profiles* grows by *7,044* records (STDEV: *475*) annually, G*ifts* grows by *45,884* records annually (STDEV: *6,147*). For confidentiality reasons, some attribute values are masked in these samples (e.g., actual addresses and phone numbers, graduation school, gender and ethnicity codes) and all gift amounts are multiplied by a positive constant.

ATTRIBUTING VALUE AND ASSESSING DISPARITY IN PROFILE RECORDS

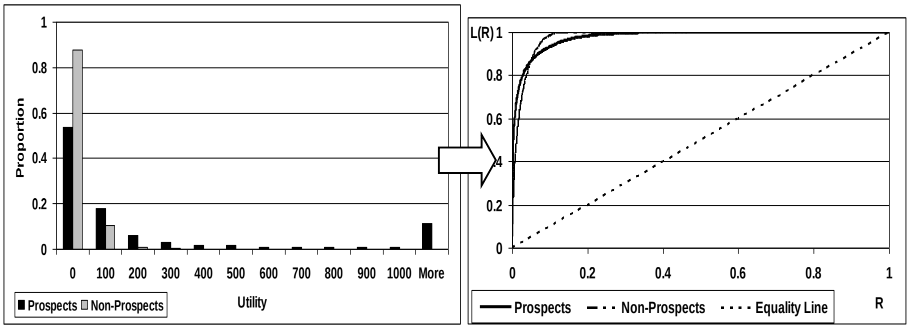
The value of using alumni data is reflected by the transactions in the gifts dataset. Profile data, along with past gift transactions, is used to identify and approach alumni with high donation potential. Gift transactions reflect the outcome of these efforts and can be linked to individual profile records. A common assumption in using CRM systems is that future purchases (gifts, in this case), to a large extent, can be predicted by past activities. This assumption is supported by the correlations between annual donation amounts and "inclinations" (Table 2). Inclination was coded as 1 for a profile if there was at least one donation (in *Gifts)* in the most recent 5 years of data and 0 if not. The correlations between annual inclinations are positive and significant. For amounts, the correlations are much lower, while still positive and significant. These amounts values are much lower for prospects compared to non-prospects. The most recent 5 years of data are illustrated in Table 2, the most recent denoted "Y."

Table 2: Correlations between Annual Inclinations and Amounts.*

		PROSPECTS (11,445 RECORDS)				**NON-PROSPECTS (346,927 RECORDS)**			
	Year	**Y-4**	**Y-3**	**Y-2**	**Y-1**	**Y-4**	**Y-3**	**Y-2**	**Y-1**
Inclination	**Y-3**	0.566				0.529			
	Y-2	0.545	0.563			0.510	0.521		
	Y-1	0.519	0.555	0.550		0.473	0.504	0.503	
	Y	0.493	0.508	0.533	0.516	0.442	0.466	0.498	0.503
Amount	**Y-3**	0.240				0.399			
	Y-2	0.199	0.139			0.359	0.389		
	Y-1	0.157	0.061	0.146		0.301	0.351	0.412	
	Y	0.016	0.010	0.062	0.020	0.271	0.315	0.341	0.386

All correlations reject H0: ρ =0 at p-Value ~0

Figure 3: Alumni Profiles Value (a) Histogram and (b) Cumulative Value Curve.



From these correlation results, we use the *average annual dollar amount donated in the 5 years (Y-4 to Y)*, as a proxy for the value of profiles. Value is *0* if a person has made no donations and positive otherwise. The value distribution (Figure 3a) shows high inequality. For non-prospects, the mean value is *\$6.7*, the standard deviation is *\$38.1*, and the proportion of profiles with *0* value (i.e., no gifts within the *5* year period) is very high *(~88%)*. For prospects, the mean and the standard deviation are substantially higher (*\$1,303.5* and *\$15,506* respectively) and the proportion of profiles with *0* value *(~54%)* is substantially lower. The corresponding cumulative value curves are shown in Figure 3b.

Assuming a Pareto distribution (Equations 1 and 5), the curves and the disparity coefficients are estimated with Log-Log regression. For non-prospects, the approximated curve is $L(R) = R^{0.111}$ (p-value: *~0*, Adjusted R-Sq: *0.535*). The equivalent Pareto parameter $w=1/(1-0.111)=1.124$, and the disparity (Gini) coefficient is $\varphi=1/(2w-1)=0.8$. The approximate curve for prospects is even steeper: $L(R) = R^{0.053}$ (p-value: *~0*, Adjusted R-Sq: *0.546*). The equivalent Pareto parameter $w=1.056$, and the disparity (Gini) coefficient is $\varphi=0.9$. In both cases, the Pareto distribution appears to be a reasonable fit for curve-approximation, though other asymmetric distributions (e.g., Weibull or Exponential) may work as well.

The disparity scores suggest a high magnitude of disparity in gift-giving, both for prospects $(\varphi=0.9)$ and for non-prospects $(\varphi=0.8)$. This has important business implications as it may suggest that a large portion of the data resource is underused - an opportunity for increasing gifts (indeed, *54%* of prospect records and *88%* of the non-prospect records are associated with *0* value). It may also highlight the need to differentially manage records in this data resource. As further discussed later, a better understanding of the business implications of value disparity requires recognition of data management costs and the associated value/cost tradeoffs.

VALUE DISPARITY ALONG ATTRIBUTE VALUES

The disparity in the value of records can be further linked to specific attributes. An attribute in a tabular dataset would have the same structure and data type for all records. However, the value of the same attribute in different records may not be the same. This variability may differentiate the relative importance and associated value of records. To illustrate this argument, we first consider *Prospect* attribute in the *Profiles* dataset. This has a value of *1* for donors classified as prospects and *0* for non-prospects. The *11,445* prospect records, *3.2%* of the dataset, are associated with a value of *\$14.92* million, *86.5%* of the overall value. On the other hand, the *346,927* non-prospect records (*96.8%*), offer a value of *\$2.32* million, *13.5%* of the overall value. A big difference!! A vast majority of the value can be attributed to a very small proportion of alumni records. The disparity coefficient for this value distribution (Equation 6) is $\varphi = 0.833$ (relatively close to *1*), indicating a very high magnitude of disparity.

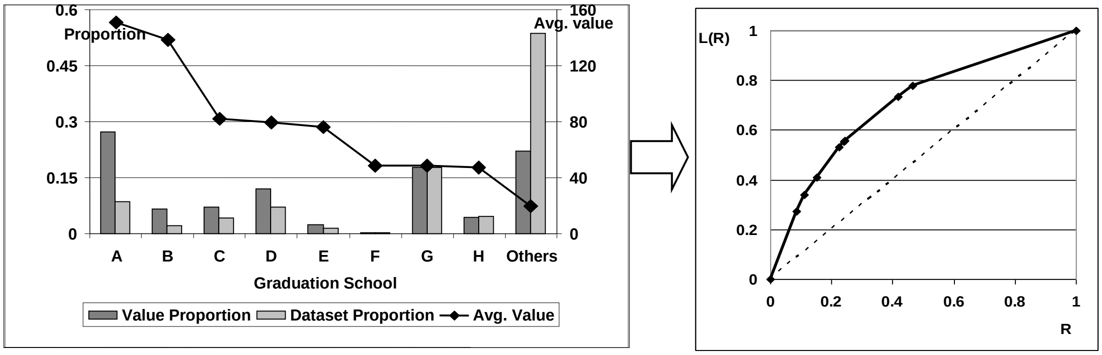
Other attributes may also offer the capability to differentiate records in terms of value, some to a greater extent than others. This can by illustrated, for example, by comparing the *School of Graduation* and the *Home State* attributes. Table 3 summarizes the number of records and the value for the different schools within the university from which the donors graduated. The table summarizes only records with a valid school code and, for brevity, groups the schools with the lowermost value averages into one category (Others.) It is sorted in descending order of average value per record. For each value of the *School of Graduation* attribute we calculate the dataset proportion (number of records in category, divided by the total), and the value proportion (value associated with records in category, divided by the total).

Table 3: Dataset Records and Value Distribution along School of Graduation.

School of Graduation	Records	Dataset Prop.	Cum. Dataset Prop.	Value ($)	Value Prop.	Cum. Value Prop.	Average Value per Record
A	31,182	0.087	0.087	4,711,490	0.273	0.273	**151.10**
B	8,148	0.023	0.110	1,129,014	0.065	0.339	**138.56**
C	15,144	0.042	0.152	1,237,464	0.072	0.411	**81.71**
D	25,899	0.072	0.224	2,059,409	0.119	0.530	**79.52**
E	5,320	0.015	0.239	403,287	0.023	0.553	**75.81**
F	1,189	0.003	0.242	57,773	0.003	0.557	**48.59**
G	63,091	0.176	0.419	3,046,884	0.177	0.734	**48.29**
H	16,471	0.046	0.464	778,476	0.045	0.779	**47.26**
Others	191,904	0.536	1	3,813,305	0.221	1	**19.87**
Total	**358,348**	**1**		**17,237,102**	**1**		**(Avg.) 48.10**

Figure 4a shows the proportion of value, the proportion of the dataset and the average value per school of graduation. The variability in the value contribution associated with each school is high, as is the variability in the average value per record. For schools *A* to *E* the value proportion is much higher than the dataset proportion. For *F* to *H*, the proportions are nearly equal, and for the *25* combined schools under *"Others,"* the value proportion is significantly *lower* than the dataset proportion *(0.221* versus *0.536)*. The associated cumulative value curve (Figure 4b) reflects this relatively high disparity (the relatively large area between the curve and the *45^0* equality line).

Figure 4: Graduation School: (a) Dataset and Value Distribution, (b) Cum. Value Curve.



A similar analysis of *Home State* values paints a different picture (see Table 4 and Figure 5 - only records with valid state codes were analyzed (i.e., from USA and Canada). Home State A has a significantly higher average value, but the overall differences in value between other categories of *Home State* are smaller. Further, the differences between the value proportion and the dataset proportion for the top most states are not as high as in the case of *School of Graduation*. Accordingly, the associated cumulative value curve (Figure 5b) shows a lower magnitude of disparity (relatively smaller area between the curve and the 45^0 equality line).

We now quantify disparity for the evaluated attributes (listed in Table 1), treating each value distribution as a discrete variable (this time, not lumping data together). We calculate the associated disparity (Gini) coefficient (Equation 6), separately for prospects and non-prospects, considering only records with valid quantities (i.e., not missing, and listed in the associated lookup table).

Table 4: Dataset Records and Value Distribution along Home State.

Home State	Records	Dataset Prop.	Cum. Dataset Prop.	Value ($)	Value Prop.	Cum. Value Prop.	Average Value per Record
A	5,030	0.016	0.016	1,073,838	0.064	0.064	**213.49**
B	21,839	0.070	0.086	2,180,912	0.130	0.194	**99.86**
C	4,899	0.016	0.102	476,576	0.028	0.222	**97.28**
D	12,443	0.040	0.142	1,192,655	0.071	0.294	**95.85**
E	450	0.001	0.144	42,083	0.003	0.296	**93.52**
F	1,454	0.005	0.148	125,535	0.007	0.304	**86.34**
G	123	0.000	0.149	8,991	0.001	0.304	**73.10**
H	5,931	0.019	0.168	409,385	0.024	0.328	**69.02**
Others	258,691	0.832	1	11,264,068	0.672	1	**43.54**
Total	**310,860**	**1**		**16,774,044**	**1**		**(Avg.) 53.96**

Figure 5: Home State (a) Dataset and Value Distribution, (b) Cum. Value Curve.

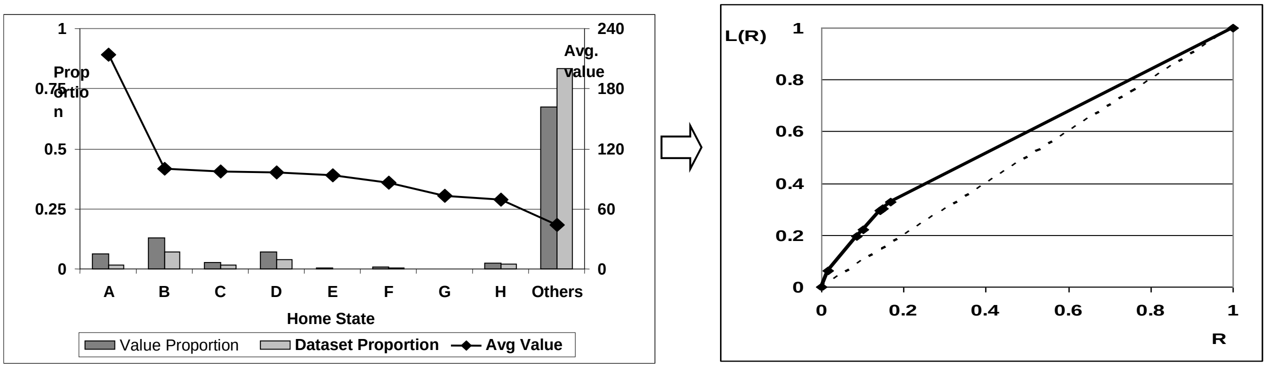


Table 5: Disparity Coefficients for the Evaluated Profile Attributes.

Attribute	**Prospects (11,445 Records)**			**Non-Prospects (346,927 Records)**		
	Prop. of Valid Quantities	**Actual Quant-ities**	**Disparity Coefficient**	**Prop. of Valid Quantities**	**Distinct Quantities**	**Disparity Coefficient**
Graduation Year	1.000	79	0.555	0.999	121	0.282
Grad. School	1.000	23	0.291	0.999	27	0.240
Gender	0.997	2	0.084	0.991	2	0.072
Marital Status	0.972	7	0.172	0.891	7	0.214
Ethnicity	0.665	6	0.079	0.593	7	0.062
Religion	0.757	24	0.328	0.600	38	0.113
Occupation	0.344	94	0.613	0.144	103	0.226
Income	0.891	3	0.029	0.623	3	0.088
Home Country	0.992	37	0.034	0.922	181	0.056
Home State	0.992	57	0.241	0.922	73	0.149
Business Country	0.822	39	0.053	0.565	146	0.042
Business State	0.822	57	0.267	0.565	69	0.108

The results (summarized in Table 5) highlight key differences between attributes and their association with value contribution and disparity. Some attributes (e.g., *Graduation Year*, *Graduation School*, and *Occupation*) are associated with a relatively high disparity, both for prospects and for non-prospects. For some attributes (e.g., *Religion*, *Home* and *Business State*) the disparity for prospects is relatively high, but somewhat lower for non-prospects. It is high for non-prospects and lower for prospects for other attributes (e.g., *Marital Status*). For yet others (e.g., *Gender*, *Ethnicity*, *Income*, *Home* and *Business Country*) the magnitude of disparity is very low, both for prospects and non-prospects.

This variability in disparity scores highlights the differing importance of attributes in different usage contexts. Some attributes have a stronger capability for differentiating records along (and can possibly help predict) value for all usages. In other words - all profile records with certain specific worth of this attribute are more likely to be associated with high value, while profile records with different specific worth are likely to be associated with low value. Certain attributes may differentiate records along value for some usages, but not for others. Yet others may have low capability to differentiate records along value for any usage. The variability in scores also highlights the need to manage certain attributes differently. Importantly, developing differentiating policies for managing records and attributes requires recognizing possible value/cost tradeoffs, which are discussed next.

IMPLICATIONS FOR DATA MANAGEMENT

There are important implications for data management due to the disparity in the value of dataset records. Using an economic perspective, we can evaluate it by assessing the effect of disparity on value/cost tradeoffs and the overall net-benefit (Even et al., 2010). We consider *v*, the aggregated value variable with corresponding maximum value v^D and cumulative value curve *L(R)*. We define *V(R)*, the maximum possible value as a function of *R*:

$V(R) = v^D L(R)$, where, **(7)**

V(R) - The maximal possible value as a function of R (the proportion of highest-value records)

v^D - The maximal possible value for the entire dataset (i.e., for *R=1*)

L(R) - The cumulative value of the aggregated value variable *v*, as a function of *R*

A characteristic that is provides critical support for our argument on value/cost tradeoffs is that both value curves *L(R)*, and consequently *V(R)*, are monotonically non-decreasing with a declining marginal return. Our definition of *R as the sorting of records in descending order of value* explains this property.

There are costs associated with managing data records. We assume identical variable cost per record, uncorrelated to the record's value. Accordingly, we initially model the variable cost as a linear curve. This curve has a v̲ariable component c^v that is linearly proportional to the dataset size (and, hence, to *R*), and a f̲ixed component c^f, that is independent of the dataset size:

$C(R) = c^f + c^v R$, where, **(8)**

C(R) - The dataset cost for R (the proportion of highest-value records)

c^f, c^v - Fixed cost and unit variable cost, respectively

Scaling both value and cost to the same monetary unit, the net-benefit contribution *B(R)* of the dataset is defined as the difference between value and cost[1]:

$$B(R) = V(R) - C(R) = v^D L(R) - \left(c^f + c^v R\right) \qquad \textbf{(9)}$$

Due to c^f, *B(R)* is negative at *R=0* (the entire curve may be negative if *C>U* for all *R*). It is concave and has a single maximum within *[0, 1]*. An optimum, R^{OPT}, can be obtained by comparing the first derivative of *(9)* to *0*:

$$dB(R)/dR = v^D\left(dL(R)/dR\right) - c^v = 0\text{, or } dL(R)/dR = c^v/u^D \qquad \textbf{(10)}$$

Below R^{OPT} the net-benefit can be improved by increasing *R,* since the added value is higher than the added cost. Beyond R^{OPT}, the marginal cost exceeds the marginal value and increasing *R* reduces the net-benefit. For a steep curve (i.e., $L(R) \rightarrow 1$, $\varphi \rightarrow 1$), or when the variable cost is significantly higher than the maximal value *(i.e.,* $c^v >> v^D$*)*, the optimum approaches a low record proportion (i.e., $R^{OPT} \rightarrow 0$*)*. If no positive R^{OPT} exists, the dataset cannot provide a positive net-benefit due to the fixed cost c^f. Conversely, if the variable cost is relatively low *(i.e.,* $c^v << v^D$*)*, R^{OPT} is obtained at a high record proportion (i.e., $R^{OPT} \rightarrow 1$*)*. With high equality (i.e., $L(R) \rightarrow R$, $\varphi \rightarrow 0$), the solution will be at one of the edges – either near $R^{OPT}=0$, or near $R^{OPT}=1$. Notably, regardless of whether the R^{OPT} solution is within the *[0,1]* range or at the edges, a positive net benefit is not guaranteed and has to be verified.

The optimality equation *(10)* can be extended for the Pareto distribution (Equation 4):

$$dL(R)/dR = (1 - 1/w)R^{-1/w} - c^v/v^D\text{, and } R^{OPT} = \left[(1 - 1/w)v^D/c^v\right]^w \qquad \textbf{(11)}$$

For *w>1*, the optimum R^{OPT} for the Pareto distribution is always positive. It is within *(0,1]* when the value/cost ratio is $c^v/v^D \geq 1-(1/w)$, otherwise, the optimal net-benefit is obtained at $R^{OPT}=1$. The optimum approaches *0* for a high degree of disparity*(*$w \rightarrow 1$*)*, i.e., when the great majority of the value is obtained from a relatively small number of records. The dependence of R^{OPT} on the value/cost ratio grows with less disparity (i.e., greater *w*). When the variable cost is relatively small (i.e., $c^v << v^D$*)*, the optimal net-benefit is more likely to be obtained when the entire dataset is included *(*$R^{OPT}=1$*)*. When the variable cost is substantially large, the optimal net-benefit is more likely to be at $R^{OPT}<1$. For a high a degree of disparity *(i.e.,* $w \rightarrow \infty$*)*, the expression $(1\text{-}1/w)^w$ converges to a constant *1/e*. If the value is higher than the variable cost $(v^D>c^v)$, $(u^D/c^v)^w \rightarrow \infty$, and the optimum is obtained for the entire dataset (i.e., $R^{OPT}=1$). If $v^D<c^v$, $(v^D/c^v)^w \rightarrow 0$, $R^{OPT} \rightarrow 0$, and the dataset is unlikely to yield any positive net-benefit.

[1] Here we explicitly use the notions of EVSI and ENGS alluded to earlier in the paper.

Value/cost tradeoffs for a Pareto distribution can also be assessed at the record level. Record values in this distribution are non-negative (Equation 1), where Z represents the lowest possible value (which can be arbitrarily close to *0*). If the *variable* cost per record is lower than Z, the value will always supersede it; hence, R will be maximized at *1* (although, it is still possible that the entire dataset will not be implemented if the *fixed* cost is too high). On the other hand, if the variable cost per record is greater than Z, the optimal R^{OPT} is likely to be lower than *1*.

For a discrete distribution, the net-benefit curve is piece-wise linear - being the difference between a piece-wise linear value curve (Figure 2) and a linear cost curve. The curve is bounded and its optimal point can be obtained using linear-optimization. Depending on value and cost parameters, the optimum may be an end point ($R=0$ or $R=1$), or an interior solution.

There are capacity limits imposed on real-world systems for the volume of data they can effectively process and store. If this capacity limit is exceeded upgrades to the system may be needed involving a higher fixed cost and possibly at a different variable cost. This may necessitate our adjusting the cost model by possibly representing it as a piece-wise linear curve. Despite such adjustments, the cost curve will always be monotonically increasing (or non-decreasing) with volume. Therefore, the argument about the existence of a maximum net-benefit point, which can possibly be internal to the evaluated range, still holds.

We have stated that there are key implications for managing data due to the magnitude of disparity and the associated value/cost tradeoffs. We first demonstrate the contribution of value/cost assessment toward cost-effective data quality improvement (an important data management activity) of the alumni data, and then discuss the potential contribution with respect to other data management activities.

VALUE/COST ANALYSIS FOR ASSESSING DATA QUALITY IMPROVEMENT POLICIES

Value assessment can define superior measurements that reflect quality assessment in context for common data quality dimensions (e.g., completeness, currency, and accuracy.) Differentiating records based on value contribution can help prioritize quality management efforts and make them more efficient. Value-cost analysis can be used for evaluating data-quality improvement policies and help identify cost-effective solutions from an economic perspective. Here, we demonstrate this concept with a post-hoc analysis of a policy for improving the quality of the alumni data. CRM data is particularly vulnerable to quality defects (Heinrich et al., 2009). Datasets that capture customer profiles and transactions grow rapidly and it is challenging to maintain high quality in datasets. Data-quality defects (e.g., missing, inaccurate, and/or outdated values) might prevent managers and analysts from having the right picture of customers and their purchase preferences and, hence, negatively impact marketing efforts.

There are serious data quality issues in the alumni profiles dataset as indicated by our preliminary evaluation. There are missing values for key attributes in approximately 84% of the prospect profiles and 94% of the non-prospect profiles, including some that are crucial for alumni-relation management and solicitation efforts (e.g., Income, Profession, Home and Business Address). Further, data had not been audited or updated in five years in ~22% of the prospect profiles and ~50% of the non-prospect profiles. Hence there are obsolete and inaccurate data values in a large proportion of the profiles dataset (for reasons such as changes in address, marital status, income level, etc.). A large majority (~97%) of alumni profiles is classified as non-prospects, and these are associated with relatively low contribution (~88% of the non-prospect alumni have made no contribution within the last 5 years).

We learned that the alumni managers link the large proportion of zero-value profiles to the high rate of data-quality defects. We will demonstrate a value-cost analysis with a subset of profile records (~50% of the profiles dataset) of alumni who have graduated within the last 30 years. All the records in this subset have some data quality defects (i.e., not updated within the last 5 years and/or with missing values in key attributes) and have no associated value (no donations) within the last 5 years. We define *Record Age* as a variable that defines the number of years from when the record was added to the database (the year of graduation) to the point of evaluation (e.g., the age of a record that was added last year would be 1). The number of records in this subset declines with *Record Age*. For evaluation purpose, we assume that the curve that represents the number of defective records *(N)* versus the record age *(A)* is approximately linear: $N = n_1A+n_0$, where the slope n_1 reflects the annual change in the number of records (here, $n_1=-100$, a negative number, as the number of records decreases with age), and n_0 is a positive number that reflects the intersection at $A=0$ (here, $n_0=8000$). The total record number for 30 years, based on this estimated curve, is 193,500.

To assess the value contribution potential of the targeted profiles we evaluated the value (average annual contribution within the last 5 years) associated with non-prospect alumni who have made some contribution, per record age. While the number of records decreases with record age, the value increases with record age – alumni who have graduated many years ago would typically have higher income and financial resources and, hence, are usually more willing to make higher contributions than recent graduates. For evaluation purpose, we assume that the curve that represents the annual value *(V)* versus the record age *(A)* is approximately linear: $V = v_1A+v_0$, where the slope v_1 is a positive number (here, we take $v_1 = 2.5$), that reflects the annual increase in value, and v_0 is a positive number that reflects the intersection at $A=0$ (here, we take $n_0 = 10$). Approximately 20% of those non-prospect alumni with complete and accurate data have made some donations within the last 2-year period. Accordingly, the evaluation assumes that 20% of the alumni with corrected data will make some annual donation within the next 2 years, with a donation rate similar to the average annual contribution of non-prospect alumni with the same record age, who made some donations.

The dataset proportion variable R corresponds to the number of years (out of 30) and the number of profiles associated with each year. As the less-recent profiles have higher contribution potential, it would be reasonable to improve the data quality of profiles with high record age first, and go "backwards" to the more recent profiles. For example – the *"R"* corresponding to record age 30 is 5000/193,500 = *0.026* (where 5000 is the number of records with record-age of 30 and 193,500 is the total number of records), the "R" corresponding to profiles of age 29 and 30 is (5100+5000)/193,500 = 0.052, and so on. To estimate the overall value-contribution potential per age, we first multiplied the estimated value by the number of targeted records, and then multiply it by the expected donation rate (20% of the records, for two years). Following these calculations, the potential value of age 30 is $51,000, and the overall annual value potential was estimated at $1,064,550. Using this estimation, $L(R)$ was calculated as the cumulative value proportion per R – e.g., $L(0.026)=0.048$, $L(0.052)=0.095$. Using a log-regression (F-Value=1713.16, P-Value=~0, Adjusted R-SQR=0.948), we estimate the cumulative value curve based on the 30 points (one per record age) - $L(R)=R^{0.733}$, with disparity index of $\varphi=0.154$.

We now use this curve to evaluate two potential schemes to improve data quality, the first being an alumni survey. The survey will be mailed to the targeted alumni and the surveyed person will be asked to update his/her personal details. Let's assume that the response rate of such a survey is estimated at 30%, and the average cost per record is estimated at $5 (a maximum variable cost of C^v= *$965,500*), including printing and mailing cost, and the time needed to handle the delivery and update the database. The survey also involves some fixed costs (e.g., campaign planning and initiation, managerial overhead), but they are relatively small and considered negligible for the matter of our analysis. Based on the 30% response-rate assumption, the maximum contribution potential for the alumni survey is estimated at v^D= *$1,064,550* and the optimum *R* is at $R^{OPT}=0.447$ (Equation 11) – equivalent to the subset of profiles with record age between 15 and 30, with a corresponding maximum net-benefit (Equation 9) of *B= $157,576*.

The second scheme is a comprehensive investigation of alumni details. Updating data on an individual can be done by searching the web, hiring external agencies, or assigning a contact person. Such updates are commonly done to prospect profiles, but not to non-prospects, due to the high cost. Let's assume that the cost of such an investigation would be $20 per record (a maximum variable cost of C^v= *$3,870,000,* plus some negligible fixed costs), with a success rate of 90%. For the comprehensive investigation, the maximum potential value-contribution is estimated at v^D=*$2,193,550*. The optimum is at $R^{OPT}=0.152$ (Equation 11) – equivalent to the subset of profiles with record age between 25 and 30, with a corresponding maximum net-benefit (Equation 9) of *B=$214,708*. As the maximum net-benefit of the second treatment is higher, the recommendation would be to run a comprehensive investigation for all profiles with records age between 25 and 30. However, some addition net-benefit can be gained by surveying alumni with profiles of record age between 15 and 24 (corresponding, approximately, to $0.152 \leq R \leq 0.447$), as within this range the marginal value per record is higher than the variable cost. The estimated added value within this age is $1,064,550*((0.447)^{0.733} - (0.152)^{0.733}) = \$322,232$, the estimated added cost is $967,500*(0.447-0.152) = \$285,244$, and the added net-benefit is *$36,987*.

The above analysis emphasizes the need for differential policies for improving the quality of customer data. Expensive schemes to improve quality should be used only for a small subset of the customer profiles that have a higher contribution potential that can justify the cost. For example, according to this analysis, it would be recommended not to apply any of the analyzed schemes for profiles of record age between 1 and 14. Obviously a real-world application will require a more thorough evaluation and more precise estimations of value and costs (e.g., by utilizing decision calculus with knowledgeable managers, or surveying vendors who specialize in customer list enhancements). The alumni managers indicated other possible data-quality improvement schemes (e.g., email surveys, automated search in public databases) that can be analyzed in the same manner and may be applicable to the entire record-age range.

To improve cost-effectiveness, instead of correcting all values, we may choose to correct only those attributes that are better predictors of contribution potential. Pending further analysis, we may also choose to correct errors in such attributes only for those records that offer high value. The completeness of profile records appears low, as some key attributes have high missing-value rates (attributes with low proportions of "valid values" in Table 5). Should the organization invest in fixing these (e.g., by contacting the donor directly or by paying a list-enhancement bureau)? Fixing all missing values (and/or correcting errors in existing values) is expensive. We may consider fixing only attributes that can better predict contribution. A good example for this argument is the Occupation. This attribute is associated with a relatively high magnitude of disparity, (0.613 for prospects and 0.226 for non-prospects). On the other hand, it appears to have a very high rate of missing or invalid values (0.656 and 0.856, for the two groups). This implies that the potential for improving value by addressing quality defects in Occupation is relatively high.

VALUE/COST ANALYSIS IN OTHER DATA MANAGEMENT CONTEXTS

Beyond the potential contribution to better data-quality management, value/cost analysis can have important implications for other data management contexts as well.

Data Usage: Understanding disparity in the value of data records can improve the utilization of a data resource for decision-making. As shown in our empirical evaluation, analyzing disparity can identify subsets of records that offer higher value and attributes that can differentiate records based on value. Users can be directed to examine these records and attributes closely or use them more often in their decision process. In our alumni data, for example, certain attributes (e.g., *Graduation Year*, *Graduation School*, *Occupation*) are associated with higher disparity in donations, while disparity scores for others (e.g., *Gender*, *Ethnicity*) are lower. This may suggest that, when categorizing potential donors and designing pledge campaigns, users should make more use of those attributes with high disparity scores and significant capability to differentiate records and less use of attributes that do not differentiate records. Understanding the attributes with high (or low) disparity may also affect the selection of software tools and/or the implementation of applications that aid data usage. For example, proprietary software packages can determine gender by examining names. In the alumni data, *Gender* is associated with low value disparity. Benefits gained by investing in such software solutions might fail to justify the cost in that usage context.

The Design of Datasets and Data Environments: Disparity in the value of records may impact the design of datasets. Low disparity (i.e., $L(R)$ converging to the 45^o line and $\varphi \rightarrow 0$) implies records with similar business value contribution. Accordingly, for optimality, one has to choose between implementing the entire dataset and not implementing it. For datasets with high value disparity (i.e., $L(R) \rightarrow 1$, and $\varphi \rightarrow 1$), depending on value/cost tradeoffs, for optimality, the designer may exclude low-value records or manage them separately. Disparity assessment may also affect attribute structure. When data is imported from an external data source (e.g., data warehouse), attributes that differentiate value strongly should be included. The designer may consider excluding other attributes that offer weak differentiation (Even et al., 2007).

Differential value may also address the design of data environments and affect higher-level design choices (Even et al., 2007), especially, for large datasets (e.g., in a data warehouse). Managing large datasets requires high investments (Mannino et al., 2008) in IT infrastructure (e.g., more powerful database and network servers) and data delivery platforms (e.g., sophisticated business intelligence tools). Investing in a powerful infrastructure will be harder to justify if a majority of the value comes from a small fraction of the data, which can be managed effectively by a less powerful (expensive) system. Better understanding of values and disparities can inform design decisions in data environments such as investments in storage and processing capacity, the configuration of data repositories, and data marts for departmental use.

Our assessment of disparity in alumni data was triggered by a data warehousing initiative. The alumni data is currently managed in a legacy system that does not permit the analytical use of this data. To support this analytical capability for managing alumni relationships, the organization is considering a data warehouse. The intent is to support a variety of data presentation and delivery capabilities that can permit sophisticated and multi-dimensional analysis of the alumni data. The analysis described here sheds light on key design decisions in terms of choice of attributes to extract into the warehouse, the volume of data (number of records) to use, and the choice of attributes to purchase from external sources.

Data Acquisition, Retention and Pricing: If records significantly differ in value, it makes sense to invest more in acquiring and maintaining records that offer higher value. A typical example of such differentiation is the archiving older data, as maintenance costs can be avoided or reduced by archiving or deleting records with lower value. Further, data vendors typically apply bulk pricing, based on characteristics such as data volume and/or the number of data retrieval activities involved (West, 2000). If the value distribution is better understood, vendors can price data, based on its potential overall value to the buyer. Also, data is purchased to enrich a customer dataset. Agencies offer list-enhancement and the pricing is typically a step function (e.g., *$X* for up to *5,000* records, *$Y* for *5,001-15,000* records, and so on). Should the entire list be enriched or should we focus only on specific records? Understanding disparity in contribution potential can identify economically superior solutions for such decisions.

For example, our analysis indicates high disparity in the donations associated with each profile. Should the organization consider enhancing data only for donors with high gift potential? Should it avoid enhancing data for profiles that have demonstrated low value contribution so far (notably, *226,508 ≈ 63.2%* of the profiles are not associated with any gifts)? Further evaluations of disparity in contribution potential and value/cost tradeoffs are necessary before a final recommendation is made. However, based on our current analysis, the organization can benefit by focusing on enhancing the data associated with the value-contributing subset of the profiles (such as those profiles that are associated with "Prospects").

COST STRUCTURE AND THE SCOPE OF EVALUATION

When evaluating value/cost tradeoffs and their impact on data management decisions, which data management activity (e.g., usage, design, acquisition, and/or quality improvement) should be examined? Should all activities receive the same attention? We suggest that identifying the scope of evaluation is linked to the cost structure. Notably, the optimal configuration (the solution to Equation 7) is affected by the variable costs (monotonically increasing with the number of records), but not by the fixed costs. However, the fixed cost is important for assessing whether or not the optimal net-benefit (as determined by Equation 7) is positive. It is therefore important to identify which cost factors should be treated as fixed and which ones as variable, in addition to the *magnitude* of each factor, before the evaluation.

Data storage, processing and delivery are IT-intensive and associated with high fixed costs due to requirements gathering, design, initial investment in hardware and networking, software licensing (e.g., DBMS, ETL tools, Business Intelligence platforms), software development and customization. Once the infrastructure is established, the added variable costs associated with storage (e.g., increasing disk space), processing (e.g., upgrading processors), and delivery (e.g., renting higher bandwidth) are relatively low, given the “cloud” and the declining costs of IT for data storage and processing. Hence, in IS environments that manage relatively small data volumes, economics-driven optimization of storage, processing, and delivery capacities may have negligible impact and the effort may not be justified.

On the other hand, the variable costs associated with data acquisition and maintenance, being labor-intensive activities, are often high. Data acquisition involves manual data entry or fees to vendors. These costs typically grow monotonically with the number of data records. Variable data quality maintenance costs may also be high. In CRM environments, for example, certain attributes (of a donor or customer) may be missing when a new record enters the system (e.g., occupation and income), and others might become outdated and unfit to use (e.g., address, marital status, and credit score) if not audited and corrected frequently. The cost of auditing and enhancing records are not be negligible. Such efforts involve contacting the person, or hiring an agency that specializes in collecting demographic data.

The alumni-data samples evaluated here are relatively small and the variable costs associated with data storage, processing and delivery are relatively low. However, in environments that manage significantly higher volumes, the variable costs of processing, storage and delivery may be non-trivial. Federal regulations that mandate capture and maintenance of data increase data volumes further. Being uncertain about the value of the data, organizations are unwilling to jettison data as data that has no value now may have significant value in the future as new usages emerge. Hence, data volumes increase because a large part of the data acquired is stored even if unused. Social media and the Internet of Things have made it easy for organizations to access data from external sources that were inaccessible and even unknown before. Hence, while storage costs have dropped, the demand for capacity has gone up. While the cost *per unit* capacity (e.g., storage space or processing speed) may decrease, the *overall* cost may actually increase. Typically, managing data volumes beyond a certain threshold may be possible only by switching to more advanced and pricier technologies. In data environments that manage very large volumes, evaluating and optimizing these capacities may significantly impact on the overall net-benefit.

CONCLUSIONS

Data management must be examined from an economic perspective. By linking data management decisions and the associated economic tradeoffs to the distribution of value in large datasets, and developing analytical tools for assessing value and its disparity, the study justifies the need for the economic perspective. Managing alumni data, a form of a CRM environment, is used as a context to demonstrate the analytical tools developed here. The study shows that the results of such assessment have important implications for cost-effective management of alumni data.

Modeling and assessing value and its disparity can reflect the current state of data resources, identify improvement targets, and help track progress toward these targets through periodic evaluations. Further, the assessment highlights subsets of records and attributes associated with higher (or lower) value. It can, hence, guide exploration and experimentation of alternative usages and/or administration of specific subsets. Disparity assessments may be interpreted in different ways. They can serve as a tool for identifying opportunities for improvement, establishing differentiating data management policies, indicating incorrect usage of data subsets, or detecting over-investments in data with low value. Disparity assessment alone does not provide a full picture of the current state of data resources. It can be complemented, for example, with measurements of data quality which reflect the presence of defects. However, assessing the disparity in data value may offer insights that can direct prioritization of data management efforts.

This study has its limitations. We do not present measurement of the actual costs involved in managing the alumni data resource - this data is yet to be collected and analyzed. To get a true picture of economic tradeoffs, value and cost must be assessed for the entire data resource, not just for a subset. Generally, the costs considered and discussed in this study are associated directly with data management. However, managing alumni relations in real-world settings involves other costs, such as those associated with solicitation (e.g., mailing, phone calls) and customer retention (e.g., fund-raisers, and show/game tickets). Though not associated directly with managing the data, these costs may significantly affect alumni data management decisions. Our cost model assumes an equal variable cost per dataset record. Variable costs may not always be linear with number of records (e.g., purchasing bulk data at a discounted price). The current value model considers a static "snapshot" of value. Value contribution may dynamically change over time and modeling the effect of these variations on data management decisions will require different analytical tools (e.g., time-series analysis). We develop disparity measurements assuming a Pareto or a discrete distribution. While these appear reasonable for modeling donor behavior, other scenarios may be better represented by other distributions.

In real-world environments, attributing value is context-specific, subjective at times, and can be challenging (e.g., value attribution in banking is likely to be different from that in healthcare). A data resource can be used by multiple consumers, each using the data for a different context resulting in different (and possibly conflicting) value assignments. Even within the same context, two different users may assign value differently to the same data. Further, the benefit from some usages may be unknown when the dataset is established. Assessing the value in complex business settings and attributing it to records requires further study. Here we have used gift amounts as a proxy for value. This is not unreasonable in the context of alumni data as gift transactions do provide a good measure of value based on the revenue generated by each profile. We have used a specific and relatively simple assessment of value, the average of the most recent gifts made per year. Alternative assessments do exist such evaluating Recency, Frequency and Monetary (R.F.M.) differentials in donation behavior (Roberts & Berger, 1999), or estimating Customer Lifetime Value (Berger & Nasr, 1998). Finally, our research draws attention to the association between data management decisions and associated economic outcomes. Given increasing data volumes and data management costs, we believe that the economic perspective is important and beneficial.

REFERENCES

Ahituv N., (1980). A systematic approach towards assessing the value of information system, *MIS Quarterly,* 4(4), 61-75.

Ballou D. P., Wang R., Pazer H., & Tayi G. K. (1998). Modeling information manufacturing systems to determine information product quality, *Management Science,* 44(4), 462-484.

Berger, P. D., & Magliozzi, T. (1993). List segmentation strategies in direct marketing. *OMEGA - International Journal of Management Science*, 21(1), 1-61.

Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: marketing models and applications. *Journal of Interactive Marketing*, 12(1), 17-30.

Boland, T. W. (1985). The use of the expected-net-gain chart to illustrate various aspects of sampling and sample design. *The American Statistician*, 39(1), 47-52.

Date, C. J. (2004). An Introduction to Database Systems (Eighth Edition), Pearson Education Inc., Boston, USA.

Even, A., Shankaranarayanan, G., & Berger, P. D. (2007). Economics-driven design for data management: an application to the design of tabular datasets, *IEEE Transactions. on Knowledge and Data Engineering,* 19(6), 818-831.

Even, A., Shankaranarayanan, G. & Berger, P. D. (2010). Inequality in the value of customer data: implications for data management and usage. *Journal of Data Base Marketing and Customer Strategy Management*, 17(1), 17-35.

Freund, Y., Seung, H., Shamir, E. & Tishby, N. (1997). Selective sampling using query by committee algorithm, *Machine Learning*, 28, 133-168.

Garcia-Molina, H., Ullman, J. D., & Widom, J. (2002). Database Systems: The Complete Book, Prentice Hall, Upper Saddle River, New Jersey, USA.

Gattiker, T. F., & Goodhue, D. L. (2004). Understanding the local-level costs and benefits of ERP through organizational information processing theory. *Information and Management,* 41(4), 431-443.

Heinrich, B., Kaiser, M. & Klier, M. (2009). A procedure to develop metrics for currency and its application in CRM. *ACM Journal of Data and Information Quality*, 1(1), 1-28.

Jagannatahan, R. (1985). Use of sample information in stochastic resources and chance-constrained programming models, *Management Science*, 31(1), 96-108.

Jain, S., & Kannan, P. K. (2002). Pricing of information products on online servers: issues, models, and analysis, *Management Science*, 48(9), 1123-1142.

Kalfus O., Ronen B., & Spiegler, I. (2004). A selective data retention approach in massive databases. *Omega,* 2004(32), 87-95.

Mannino, M., Hong, S. N., & Choi, Jun. (2008). Efficiency Evaluation of Data Warehouse Operations, Decision Support Systems 44 (2008) 883-898.

March, T. S., & Hevner, A. R. (2007). Integrated decision support systems: a data warehousing perspective. *Decision Support Systems*, 43(3), 1031-1043.

Padmanabhan, B., Zheng, Z., & Kimbrough, S. O. (2006). An empirical analysis of the value of complete information for eCRM models. *MIS Quarterly*, 30(2), 247-267.

Provost, F. (2005). Towards economic machine learning and value-based data mining, First International Workshop on Value-Based Data Mining, 2005, Chicago, IL, USA.

Provost, F., & Fawsett, T. (2013). Data Science for Business. O'Reilly Media, Sebastopal, CA.

Ramakrishnan, T., Jones, M. C., & Sidorova, A., (2012). Factors influencing business intelligence (BI) data collection strategies: an empirical investigation. *Decision Support Systems,* 52, 486-496.

Roberts, M. L., & Berger, P. D. (1999). Direct Marketing Management (2nd edition), Prentice-Hall, Englewood Cliffs, NJ, 1999.

Saar-Tsechansky, M. & Provost, F. (2007). Decision-centric active learning of binary outcome models. *Information Systems Research,* 18(1), 4-22.

Schechtman, E., Yitzhaki, S. & Artsev, Y. (2008). Who does not respond in the household expenditure survey: an exercise in extended gini regressions? *Journal of Business and Economics Statistics*, 26(3), 329–344.

West, L. A., Jr., (1994). Researching the cost of information systems. *Journal of Management Information Systems*, 11(2), 75-107.

West, L. A., Jr. (2000). Private markets for public goods: pricing strategies of online database vendors, *Journal of Management Information Systems,* 17(1), 59-84.

Weinberg, B., Davis, L., & Berger, P. D. (2013). Perspectives on big data. *Journal of Marketing Analytics*, 2(1), 187-201.

Zheng, Z. & Padmanabhan, B. (2006). Selectively acquiring customer information: a new data acquisition problem and an active learning-based solution. *Management Science,* 52(5), 697-712.