HCEO WORKING PAPER SERIES

# Potential Value of Data and Free Access to Data

Amos Golan
Spiro Stefanou

Working Paper

**2024-008**

04/2024

HUMAN CAPITAL AND
ECONOMIC OPPORTUNITY
GLOBAL WORKING GROUP

The University of Chicago
1126 E. 59th Street Box 107
Chicago IL 60637

www.hceconomics.org

# Potential Value of Data and Free Access to Data

Amos Golan[*1] and Spiro Stefanou[2]

This Version: December 28, 2023

**Abstract**

In this study we develop measures of the potential value of information with an emphasis on observed information – data. Though value is a relative concept, developing approximate and applicable measures is essential. Such a measure (or set of measures) allows us to evaluate the potential value of public and privately available datasets, and the value of accessing each. There are several benefits to having such measures. First, providers of data can perform a cost-benefit analysis. Second, policy makers can better determine the benefits of different data when deciding whether to invest in its collection, production and release. The proposed measures are derived from information-theoretic principles as well as other statistics, in conjunction with relative measures based on semantic arguments. These measures are functions of attributes that can be aggregated into three basic blocks: (*i*) data reliability, integrity and accuracy, (*ii*) data quality, and (*iii*) potential value. We provide detailed empirical examples applying these measures to three data sets, each of which is different in context, size and complexity.

*Key Words*: Benford's Law, Compressibility, Condition Number, Data, Information, Mutual Information, Potential Value, Relative Entropy, Shannon Entropy, Simple Statistics, Value

[1] American University and Santa Fe Institute; agolan@american.edu
[2] Administrator, ERS, USDA; Spiro.Stefanou@usda.gov
[*]Corresponding author.

We thank Danielle Wilson for all her work and help with this project, including data work, development of the codes and reviewing most of the related literature, as well as many helpful discussions. Golan also thanks Alan Isaac, Radu Balan, participants at the ERS meeting, and colleagues at the Santa Fe Institute for insightful discussions and suggestions. Golan thanks ERS-USDA for partially supporting that project.

## 1 Introduction

Individuals, researchers, and policy makers need information to make informed and educated decisions. These decisions are improved if the information used is of high quality, and if the inferential approach used to transform the information into knowledge and decisions is efficient and logical. In this work we concentrate on the first issue: the quality of information and its value. Our emphasis is on observed information: data. Unlike much of the literature on the quality and access to data, we are interested in building a set of complimentary measures to evaluate the *potential value* of data. It is somewhat like the 'option value' of the data. We define the potential value as the overall value that society may obtain from a certain dataset, assuming all the information and knowledge embedded in that data are extracted. It is not a value based on past use of the data, but rather the complete potential of that data, if indeed it materializes. The potential value of data is based on its quality, and on the meaning of the informational content in the data under different contexts. These measures are independent of the inferential approaches used (or to be used) when converting the information in the data into knowledge.

The motivation for this work is both philosophical and practical. The philosophical one is a special case of the more abstract approach for dealing with information and its value (see Dunn and Golan, 2021 and the references provided there) that concentrates on the value of observable information used for inference and decision making. The practical motivation stems from the need for a simple and applicable way of evaluating the potential of new and existing datasets. This need has been underlined recently by policy makers who require federal agencies to collect, develop and produce data for public use. The tools, however, for such evaluations are yet to be defined and developed. In this paper, we propose such measures and tools.

Federal agencies collect and produce data. These datasets provide the needed information used by public and private researchers, practitioners, and decision makers. These data are often variations of restricted use data and are only made available when conditions for protecting identifying information are met. The dollar cost of such data (for purchase and collection) is known or can be easily assessed and calculated. The cost of protecting these data – including storing and maintaining their integrity and confidentiality – are also known or can be approximated. But the real value (or potential value) of these data to society, researchers, and policy makers has yet to be defined or determined. This is not an easy task as data and information are special goods. In addition, their values – as any value – are relative. Assessing their full potential value is quite complicated, but necessary for fully understanding their importance to society.

In this study we develop an approach for evaluating the approximate potential value of datasets (including those that are publicly available) and the value of maintaining public access to these datasets. The measures we propose are applicable, interpretable, and relatively easy to compute and evaluate. We test our measures using simple 'toy' data and two distinct datasets provided by the Economic Research Service of USDA. Throughout this paper we treat the

concept of *value* as *potential value*. It is the complete (long-term) potential of a data set; not just what it was used for in the past. Figuring out the potential value of data will also allow for comparison of datasets which, in turn, will facilitate more accurate cost-benefit analyses of these products. From now on (unless otherwise specified) the word 'value' means potential value as defined above.

In the next section we briefly define and discuss the concepts of information, data, and value. We then touch on the notion of relative vs. absolute value and relate it to the value of data. In Section 3 we provide a brief summary of existing literature on the value (not 'potential' value) of data and information. In Section 4 we summarize our basic building blocks for measuring the potential value of data. We also define and discuss the notion of, and inter-relationships between, meaning and context, and the way they impact value. In Section 5 we develop and discuss the measurable quantities we use, including information-theoretic measures. In Section 6 we define the complete set of attributes we use to evaluate datasets, which include measures based on the semantics and meaning of the information. In Section 7 we define our proposed measure of the potential value of data (including public data), which is an aggregated measure of the attributes defined in Section 6. In Section 8 we provide a detailed example using a small dataset dealing with social media and new movie release buzz. The following two sections provide two additional case studies using medium and large sized datasets. In Section 11 we provide a graphical comparison of value, quality and attributes, across these three datasets. In Section 12 we discuss the value of access to these data. In Section 13, we comment on the monetary value of the proposed value measure. Finally, in Section 14 we provide concluding remarks and a short list of open questions. The Appendix provides detailed tables from the case studies, a breakdown of how the respective datasets were evaluated using the attributes discussed in Section 6, and an outline of the software used, and codes developed, to produce the different information theoretic and other measures.

## 2 Information, Data and Value

In their work on information and its value, Dunn and Golan (2021) write that *information* may be defined (though somewhat circular) as anything that informs us. From a decision making and inference point of view, being informed means a certain input – objective or subjective – enters our decision process and affects our inference and decision. We take the same view here. We also take the view that, in general, information is true and is not intended to be false, though it is frequently noisy and imperfect, and its meaning may be subject to interpretational and processing errors. On a more practical level, information may be thought of as this 'thing' that provides us with the means to reduce the bounds of uncertainty about possible outcomes. It is this 'thing' that informs us; it puts us in the condition of "having information." But "having information" is a weaker notion than having knowledge, or even beliefs. We can have information because we observed something, were told something, or even because it is in a document we were given. This does not mean that we really know what the information is or that we understand it.

However, even though having information is not the same thing as having knowledge, having information may end up contributing to one's stock of knowledge; however, measured and of whatever quality. For the applied researcher who is interested in modeling, inference and learning this means that information is anything that may affect one's estimates, the uncertainties about these estimates, or decisions. It is "meaningful content."

In this work we are interested in observed (hard) information, known as data. Our evaluation of that data is based on the premise that these data are used, by all possible users, for decision making and inference to convert it into knowledge. In fact, it is the potential inferred knowledge, conditional on the data, that we care about. But it is independent of the inferential approach.

### 2.1 Absolute vs. Relative Value

In Section 6 we discuss the main applicable attributes that we believe determine the value of data. Before we do so, it is necessary to highlight the fundamental issue of 'relativity' that must be discussed when assessing the value of data. This issue deals with the concepts of *objective* vs. *subjective*, *intrinsic* vs. *extrinsic*, and *absolute* vs. *relative*. Broadly speaking the first two are special cases of the latter. *Subjective* value is relative to the person making the value judgment. *Extrinsic* value is relative to something outside of the thing being valued. These distinctions may be applied to both the user of the data and the dataset itself. Since the last pair includes the others as special cases we briefly discuss the notions of *relativity* and *subjectivity*.

The different attributes that determine the value, whatever they are, are all relative to the decision maker, or the user. To mathematically show that the value of information is relative, or subjective, it is sufficient to provide the following argument. If at least one factor that determines the value of information is relative (subjective), the value of information must be relative (subjective) as well. Stating this differently, there is no unique way of defining a relatively (subjectively) based absolute value of information.

We can also think of this in a different way which is similar to deciding on the value of life. See for example, the critical review of Viscusi and Aldy (2003). The value, for example, that we put on our lives as individuals is far different from that of the insurer, the actuarial or the government. If there exists a dataset that will allow a doctor to cure someone's disease, they (and their family) will be willing to, theoretically, pay any amount for this. Others that do not have that disease will be willing to pay much less, or even nothing.

We argued above that the value of data is subjective and relative. It is not objective; It is not unique. It is relative to other information and the user, and it is subjective to the decision maker – the user. For a detailed discussion, see for example, Dunn and Golan (2021) . We complete this discussion by logically showing that an absolute value of data cannot be subjectively based.

Define users' preferences over the datasets. As with all preferences, they are subjective. We require the following properties from an objective, or absolute, value-of-data function:

A1. *Individual preferences are complete (any pair of datasets can be compared), transitive (trivial) and reflexive (any information/data set is at least as good as itself). The value-of-data measure should satisfy the same properties*.

A2. *If everybody prefers dataset X to Y, then the value-of-data measure should rank X ahead of Y.*

A3. *The preferences between datasets X and Y should depend only on the way individuals rank X vs. Y, regardless of how they rank other datasets.*

Following directly on the classical Arrow's Impossibility Theorem, (Arrow, 1963; Fishburn, 1970) it is trivial to show that if a value-of-data measure satisfies properties A1 – A3 and the set of users is finite, then it must be a 'dictatorship' or an expert opinion (all value-of-data rankings are rankings of a single user – the expert or the dictator). Under the above three requirements on users' preferences, there is no unique way of defining the "absolute" value of data (that is subjectively based). Even if this unique way is an expert opinion, it does not resolve the basic issue here: a single intrinsic (or objective, or absolute) value of information, that is free of the users' preferences, cannot be based on subjective evaluations. Though this argument shows that an objective value of information cannot be subjective based, there are a few trivial (uninteresting) exceptions. First, if all users have the same preferences toward the value of all datasets. Second, all experts value all datasets similarly (and consistent with all other potential users).

## 3 A Brief Summary of the Current Literature

The study of the value of information has a long history in both the philosophical and the social science literatures. The former is much more theoretical while the latter is typically empirical. We discuss both with an emphasis on the way value has been dissected into types – especially within the economics literature – and how these types relate to datasets.

Often the definitions of value (of information) have been contextualized with concepts of communication and understanding (Dretske, 2008), or are grounded in decision making (Gould, 1974). With this in mind, economists have generally defined three main types of value: use, existence, and option. Use value can be defined as the current value directly gained from receiving a good or a piece of information. For users of government data, for example, the value of this public good comes from the direct gain of any subsequent understanding, knowledge, or analyses. Existence value, on the other hand, captures additional value associated with knowing that, for example, a rare ecosystem, species (Davidson, 2013) or even dataset merely exists (Krutilla, 1967). This type of value can also be intuitively described as "passive" use value (Carson, 2012).

The third general type of value is option value, which is defined relative to the concept of potential and uncertainty. It can be categorized into different subtypes: real option value (Dixit & Pindyck,1994), quasi-option value (Arrow & Fisher, 1974), or, simply, option value (M. W.

Hanemann, 1989). Real option value is the value associated with the timing of an investment or decision given learned information about its risk and returns. Quasi-option value, on the other hand, is the value of learning from the postponement of a decision (Mensink & Requate, 2005; Traeger, 2014). More relevant to the context of value and potential value of data is option value. A common example of option value is the existence of a public park. Even if a local resident has no immediate intention of visiting it, they may still feel as though they benefit from having the option to potentially do so in the future.

The types of values defined above can often be difficult to conceptualize and interpret without a numeric or monetary representation. The contingent valuation method is arguably the most practical tool for indirectly estimating the monetary equivalent of existence and option value (Brookshire, 1982; Carson, 2012; W. M. Hanemann, 1994). This method uses carefully crafted surveys to illicit either an individual's maximum willingness to pay for, or minimum compensation needed to give up, a good (Carson & Hanemann, 2005). Respondent's willingness to pay can then be collectively used to estimate society's value (Ciriacy-Wantrup, 1947) of, for example, a publicly available dataset. Contingent valuation surveys are not, however, limited to valuing public goods. These surveys are often designed to gauge the value of individual health and risk, and collective results are used to calculate value of life statistics (Kip Viscusi, 2014; Lanoie et al., 1995; Magat et al., 1988). The contingent valuation approach is not without its critiques (Carson, 2012; Diamond & Hausman, 1994; Hausman, 2012). Aside from assumptions regarding the quality of surveys, their ability to illicit accurate responses and the quality of respondent data, contingent valuations are arguably dependent on perceptions and preferences rather than, for example, the informational content of a resource or good.

These definitions provide different perspectives on value and guide subsequent approaches to measurement. However, they cannot be directly, and consistently, applied to assessing the full (potential) value of datasets or publicly accessible datasets. Like any publicly available information, these datasets are nonrivalrous and nonexcludable public goods. When made accessible, digital data inherently meet these criteria as they can be simultaneously used by various people, firms, and organizations without affecting their quality or value. In Repo's review (1989) of the valuing of information literature, discussion on information in the form of a public good is limited. We believe it is because of the unique challenges of that ask. Dunn and Golan (2020) highlight the most prominent of challenges: differences in how data users value information. In particular, individuals, organizations and societies value information, and subsequently data, differently. These differences make measuring the total value of public data particularly difficult, especially in comparison to private information or traditional public goods. It would be quite challenging to identify, yet alone consistently quantify and disentangle, the denumerable ways in which data are valued by the universe of beneficiaries.

Other disciplines outside of information and communication sciences have proposed alternative theoretic approaches to valuing information. The economics literature, for example,

has developed models to explicitly consider information markets, and, in turn, the monetary value of data (Bergemann & Bonatti, 2019). As summarized by Veldkamp & Chung (2023), the macroeconomic literature, in particular, has incorporated the use of public data into its growth models. Jones and Tonettti (2020), Farboodi & Veldkamp (2022) and Freeman et al. (2023) all propose theoretical frameworks that depend on, or acknowledge, particular characteristics of data and data usage. These characteristics include the nonrivalry of digital data, the potential for data to depreciate, incentives for protecting data, and the heterogeneity of data usage in the private sector. Although interesting and novel, these theoretical models are context specific. Resulting valuations are also dependent on the economic influence of the data or the value of data-informed choices, not the actual content of the data.

On the empirical front, valuations of public data do exist, but these are either relative to alternative choices or counterfactual scenarios, or are reliant on the ability of the data to reduce uncertainty about certain preidentified outcomes. Hughes-Cromwick and Cornado (2019) provide a nice review of the innovative ways in which, for example, the private sector use and value public data. Outside the private sector, resource and environmental economists have built a robust literature on valuing public resources and the value of data used for their preservation. For example, Macauley (2006) and the Council on Food, Agriculture and Resource Economics (C-FARE) (2013) have each outlined a framework for empirically measuring the value of public goods and datasets, but these valuations are context dependent and do not capture *potential* value outside the scope of a particular application. The National Aeronautics and Space Administration's VALUABLEs initiative has also recently provided a variety of empirical analyses valuing different types of data. As part of this initiative Stroming et al. (2020), for example, show how public satellite data can be used for environmental impact assessments. These assessments are then used to quantify the socioeconomic value of publicly available satellite data but are only relative to potential policy outcomes.

The computer science and engineering literature has, however, focused on measuring the quality (and part of the value) of data based directly on its content. The most common approach to doing so begins with appropriately defining a dataset and its characteristics (Gebru et al., 2021; Holland et al., 2018; Tufis¸et al., 2020). The quality of these characteristics, or "facets", is measured using a predefined criteria and then mapped to a relative value (Kannan et al., 2018). The choice of "facets", the relative importance or weighting of these "facets", and the mapping method used to produce a valuation from these "facets" are, however, arguably subjective. Alternative approaches using standardized predicates for measuring data quality have been proposed (Bronselaer et al., 2018; Kaiser et. al., 2007). Though that approach may be the closest to our objective, these innovative alternatives are also based on inference of the data (an undesirable property for our valuation) and have yet to reconcile the subjective transformation of data quality to potential value.

## 4 Potential Value of Data – The Basic Building Blocks

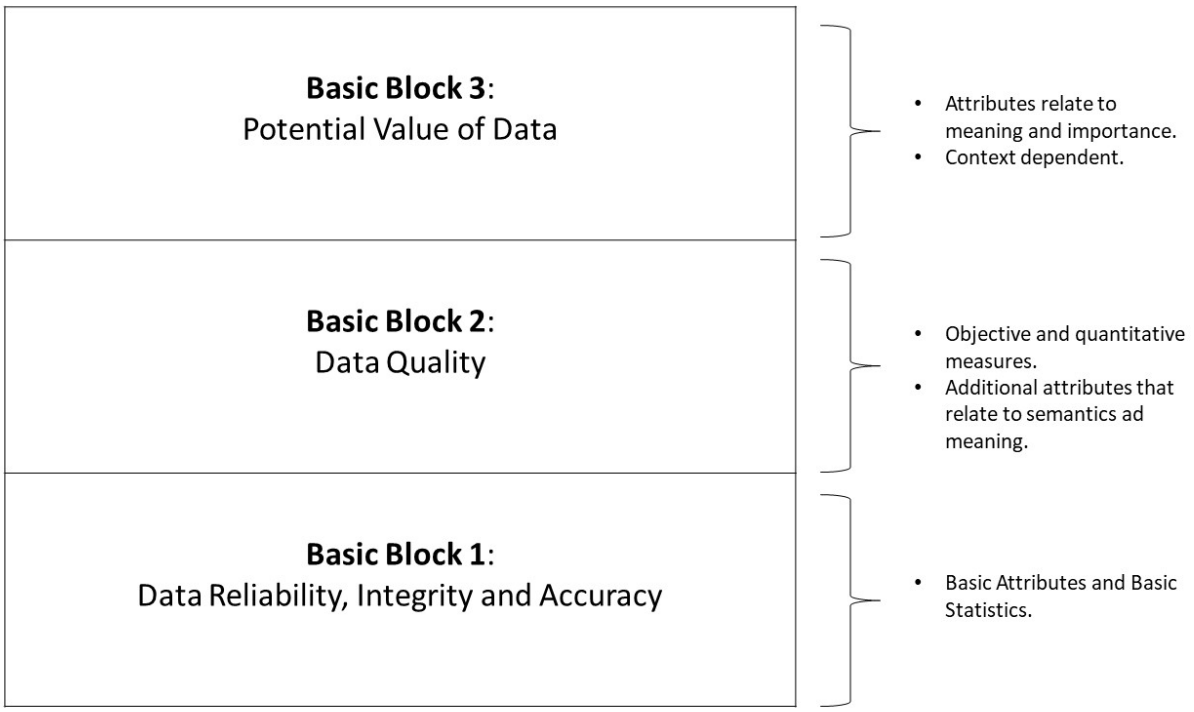
We now summarize the basic ideas and building blocks of our proposed value measure. We want our measure to satisfy a minimal set of requirements (attributes). We organize these requirements as three hierarchical building blocks, each containing several attributes. Some of these attributes are measurable and can be objectively quantified. Others are qualitative or ordinal and at times subjective, and some are fuzzier.

At the bottom of the hierarchy is the first building block: *Data Reliability, Integrity, and Accuracy*. It comprises measures identifying the basic attributes of the data, including basic statistics. The second building block is *Data Quality*. It comprises objective and quantitative measures as well as more complex attributes related to the meaning and semantics of the data. The third building block, at the highest level, is the *Value of Data* – or more specifically, *The Potential Value of Data*. It comprises the first two building blocks as well as other relative (subjective) attributes related to meaning and importance. That last part is the most complicated, as we must use meaning to evaluate value, but for 'meaning' we need a context.

We use the word meaning to capture the notion of what one intends to convey especially by language: the thing, action, feeling, idea, etc. that a word or words represent. This meaning is also affected by the context – the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood and assessed. Pragmatically, meaning and context are interrelated (Nouraldeen, 2015). The more information that is embodied in the context, the less words (or linguistic utterances, as linguists describe this) are needed. Consequently, information from both context and the words themselves simultaneously produces meaning (Johnson, 1974). To relate this to the more linguistic, philosophical, and information-theoretic literature, rather than 'meaning,' often the word semantics is used, which implies relating to meaning in language or logic.

Within our objective of assessing the value of data, the context may change with the user, time, and the state of nature (say the state of the economy, politics, etc.). Consider for example, a dataset summarizing farms' production in the Midwest. The context in this case describes the (*i*) dataset (crops, inputs, region, period, etc.), (ii) any additional related information if such exists (say, demand for the crop, technology, or water supply), and (iii) the user (say, researcher, policy maker, private firm, or the USDA research specialist). Within that context, we have the potential set of questions the data may shed light on. Think of the subset of questions related to the production process. For the researcher, a question of interest may be the exact structure of production and its relative efficiency. For the policy maker a main question may be whether there are some economies to scale that created a monopolistic power that demands regulation. For the USDA specialist, the interesting question may be water efficiency, quality of the soil and minimum pollution. In each one of these cases, the potential value is attained in the context-question pair.

Figure 1. Basic Building Blocks




But how should we put value, and potential value, on this pair? The value is relative to the context (possible questions and users) and the environment of that context. The overall potential value depends on the particular contexts (the state of nature together with the existing and potential user questions), or the universe of these particular contexts. Similarly, the value can be thought of as an 'option value,' or the perceived potential impact of that data.

With the above in mind, the overall structure is as follows. We require the *Reliability* block to exceed a minimal level. Otherwise, the data may be unusable. The exact minimal level depends on the potential use of the data. Conditional on that, the *Quality* block is calculated. If it satisfies our desires, we calculate the potential value according to the *Value* block. Keeping in mind that data are scarce resources, however, most often it is impractical and illogical to provide these minimal thresholds. This is because we want to use the data we have, regardless of its quality. But we should always calculate and evaluate the *Reliability* and *Quality* blocks, as part of the measures that approximate the value.

Although values are relative measures, at times one can normalize those in such a way that comparisons across datasets (in certain contexts) is possible. Assuming this is done, the last major question left is how to convert that value to a monetary one. This is discussed in Section

13. In the next section we define the less familiar quantitative measures, most of which emerged from information theory. We then discuss the attributes and measures used for each one of the three building blocks.

# 5 Measurable Quantities – Basic Definitions

We start with measurable quantities that emerged from information theory. We then discuss two more measures.

## 5.1 Information Measures

### *Information and Entropy of a Single Random Variable*

The *entropy* (Shannon, 1948) is the expected information content of an outcome of a discrete random variable $X$ whose probability distribution is $P$:

$$H(P) \equiv \sum_{k=1}^{K} p_k \log_2 \frac{1}{p_k} = -\sum_{k=1}^{K} p_k \log_2 p_k \tag{1}$$

where $k$ is the number of events, and with $x \log_2(x)$ tending to zero as $x$ tends to zero, or simply, take $p_k \log_2(p_k) \equiv 0$ if $p_k = 0$. Entropy is a function of the probability distribution $P$ and not a function of the actual values taken by the random variable. Therefore, entropy is free of semantics (meaning) of the information. In other words, the entropy of a random variable would remain the same if the variable's outcomes changed, but the number of outcomes and distribution did not.

Looking at the above definition it is clear that the entropy is a relative number (between zero – perfect certainty – and log($K$) – maximal uncertainty). However, it can be normalized such that systems of different dimensions and variables can be compared (Golan, 1988). The normalized entropy is $S(P) = H(P)/\log_2(K)$, where $S(P) \in [0,1]$. This measure provides an overall idea of the information contained in the data (but not about the meaning of that information). Note that from now on, rather than using $\log_2$ we will use the notation log (and in practice we will use the natural log).

### *Information and Entropy of Multiple Random Variables*

Let $X$ and $Y$ be two discrete random variables with possible outcomes $x_1, x_2, ..., x_K$ and $y_1, y_2, ..., y_J$ respectively and $P(X,Y)$ be their joint probability distribution. Define $P(X = x_k) \equiv p_k$, $P(Y = y_j) \equiv q_j$, $P(X = x_k, Y = y_j) \equiv w_{kj}$, $P(X \mid Y) = P(X = x_k \mid Y = y_j) \equiv p_{k|j}$, and $P(Y \mid X) = P(Y = y_j \mid X = x_k) \equiv q_{j|k}$ (with "|" standing for "conditional on"), where $p_k = \sum_j w_{kj}$, $q_j = \sum_k w_{kj}$ and the conditional probabilities satisfy $w_{kj} = q_j p_{k|j} = p_k q_{j|k}$.. The *joint entropy* of $X$ and $Y$ is

$$H(X,Y) \equiv \sum_{k,j} w_{kj} \log_2 \frac{1}{w_{kj}} = -\sum_{k,j} w_{kj} \log_2 w_{kj}. \tag{2}$$

As shown previously, this measure can also be normalized to the zero – one range.

The *conditional entropy* $H(X \mid Y)$ is

$$H(X \mid Y) = \sum_j q_j \left[ -\sum_k p_{k|j} \log_2 p_{k|j} \right] = \sum_j q_j \left[ -\sum_k \left( \frac{w_{kj}}{q_j} \right) \log_2 \left( \frac{w_{kj}}{q_j} \right) \right] = \sum_{k,j} w_{kj} \log_2 \left( \frac{q_j}{w_{kj}} \right), \tag{3}$$

which is the total information in $X$ conditional on $Y$ having a certain value $y_j \left( Y = y_j \right)$.

The interrelationship between entropy and joint entropy is embodied by the *chain rule for entropies* (the entropy of a composite event equals the sum of the marginal and conditional entropies):

$$H(X,Y) = H(Y) + H(X \mid Y) = H(X) + H(Y \mid X). \tag{4}$$

If $X$ and $Y$ are independent $\left( w_{kj} = p_k q_j \right)$, then $H(X,Y) = H(X) + H(Y)$.

So far, we have discussed the concepts of joint, marginal, and conditional entropies. Using these quantities, we now define the reduction in uncertainty of $X$ due to our knowledge of $Y$. The amount of information contained in a random variable $X$ about another random variable $Y$ is called the *mutual information* between these two random variables:

$$I(X;Y) \equiv \sum_{k,j} w_{kj} \ln \frac{w_{kj}}{p_k q_j} = D\left( w_{kj} \parallel p_k q_j \right) = H(X) - H(X \mid Y) \tag{5}$$

where $D\left( w_{kj} \parallel p_k q_j \right)$ is the relative entropy (known as the Kullback-Leibler divergence) between two probability distributions, and $I(X;Y) \geq 0$ with equality if and only if $X$ and $Y$ are independent. It is the *marginal additional information* the observer analyzing $X$ gains from knowing $Y$. The mutual information is the relative entropy between the joint distribution, $w_{kj}$, and the product of the marginal distributions, $p_k q_j$: $D\left( w_{kj} \parallel p_k q_j \right)$. In general

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{6}$$

where $I(X;Y) = I(Y;X)$ and $I(X;X) = I(X)$.

A basic interpretation of the above quantities is that having additional information from another correlated random variable reduces the uncertainty we have about our original random variable. Conditioning reduces entropy (and increases information) for dependent random variables. We use these measures, in conjunction with correlation measures, to identify dependence properties and quality of the data.

***Data Integrity: Entropy Convergence or Cumulative Entropy***

Cumulative entropy is another diagnostic used to test data integrity. It is used in a variety of disciplines to test for the convergent behavior of populations (Edwards & Tuljapurkar, 2005; Talaat et al., 2020). We adjust it for our needs. Following equation (1), the entropy of a random variable is cumulatively and sequentially calculated starting with the first two observations, then the first three observations, and so on. The data, however, need not be ordered. (For continuous random variables we use the approximate discrete distribution.) The cumulative entropy of well-behaved data is expected to exhibit a converging behavior as the number of data points increases. If, instead, this entropy keeps fluctuating as additional observations are considered, there may be issues with the integrity of the data. For example, think of the binary variable Female/Male. Say, the first 10 values are: F, F, M, F, F, M, M, M, M, F. Using log base 2, the 'cumulative entropies are: 0.00, 0.00, 0.92, 0.81, 0.72, 0.92, 0.99, 1.00, 0.99, 1.00. That is, the entropy converges to one which captures the fact that the number of females and males is practically equal for the complete sample.

***Data and Information Compression: The Shannon Limit***

The Shannon limit is the maximum possible compressibility of bits (1's and 0's) of information without loss of information – the exact meaning of the data remains unchanged. Compression of information eliminates redundancies by reducing the number of bits needed to describe the informational content of the data. Technically, it reduces the number of bits describing the informational content in a data set to a minimum. One can view the Shannon limit as a good proxy for the amount of nonredundant information contained in a dataset. Or even better, the level of predictability the data may provide.

Consider the following example. We want to transfer a very large data set between users. Since the dataset is very large, we first want to compress it to its 'limit' without any loss of information. But what is that limit? Let the original dataset be of $N$ bits (0's and 1's). Assume that out of the $N$ bits, there are $K$ 1's. Thus, the number of ways to express the information is the multiplicity:

$$W = N!/K!(N-K)! \tag{7}$$

We want to compress that amount ($W$) as much as possible without loss of information (lossless compression). We are looking for $Z \leq N$ bits, defined as the minimal number of bits

that still contain the same information as in the original dataset with $N$ bits. To find $Z$, let $W=2^Z$ and then calculate $Z=\log_2(W)$ from the above equation. Using the Stirling's approximation, we specify it in terms of entropy as

$$Z = -\left(N/\ln 2\right)\left[\pi \ln \pi + \left(1-\pi\right)\ln\left(1-\pi\right)\right], \tag{8}$$

where $\pi=K/N$. The value $Z$ is called the Shannon limit. For all non-uniform distributions $\left(\pi \neq 0.5\right)$ Z<W. For further discussion and examples see Golan (2018).

This measure is often expressed as a compression ratio, which is the uncompressed size ($N$) divided by the compressed size ($Z$). It can also be expressed in terms of space savings which is just one minus the ratio of $Z$ to $N$. The Shannon limit is a 'lossless' compression measure: it maintains all of the information in the data. As such it is not expected to exceed a ratio of approximately 2. A lossy compression – a compression with loss of information – may exceed that ratio, but it is outside the scope of our interest.

### 5.2 Other Measures

The condition number measures the degree of multicollinearity among exogenous variables in regressions such as $\mathbf{y} = f\left(X_1,\ldots, X_K\right)$, where $\mathbf{y}$ is an $N$-dimensional vector ($N$ observations) and $X$ is an $N \times K$ design matrix with $K$ variables.

A simple measure of singular values for measuring the degree of multicollinearity in data (Belsley, 1991) is:

$$\kappa\left(X'X\right)=\frac{\pi_1}{\pi_K}\ , \tag{9}$$

which is the ratio of the largest ($\pi_1$) and smallest ($\pi_K$) singular values of $X$ (with column scaled to unit length). If the design matrix $X$ is orthogonal (the $K$ variables are linearly independent of each other), then $\pi_i = 1$ for all $i=1,\ldots, K$ and $\kappa = 1$. As the degree of collinearity increases the condition number goes to infinity. If it exceeds approximately 900, the collinearity level is considered harmful for the inference. (Note that some statistical software report the square of the condition number.)

We use it here to investigate the level of multicollinearity in the data. It complements some of the multi-variable information theoretic measures discussed previously.

***Data Integrity: Benford's Law***

We use this law for confirming the data integrity. Benford's Law is a natural law about the distribution of digits of almost any numerical data. It describes the frequency distribution of the first (or other) digit of numerical data that are not dimensionless. These are numbers that carry dimensions (such as measurements), yet a universal law (independent of the units of measurement) for the distribution is scale free. Abidance to this law is often used as a diagnostic to test the integrity of data (Judge & Schechter, 2009). We apply it here for the first non-zero digit of the data. Higher order (say 2nd, 3rd, etc.) digits can also be calculated and tested.

Let $D$ be the leading (first) digit. To make it scale free, we let the logarithm of the numbers be uniformly distributed. Then the probability distribution of $D$ for integers between 1 and 9, known as Benford's law, is:

$$p(D) = \int_{D}^{D+1} \frac{dx}{x} \Big/ \int_{1}^{10} \frac{dx}{x} = \log_{10}(D+1) - \log_{10}(D) = \log_{10}\left(\frac{D+1}{D}\right) \tag{10}$$

and by "scale free" we mean that scaling $D$ by a factor $C$ results in a proportional scaling of $p(D)$. Note that the probability distribution $p(D)$ is proportional to the space between $D$ and $D+1$ on a logarithmic scale (power law).

Benford's law states that we observe the number 1 as the first digit approximately 30% of the time, while larger numbers are observed as the leading digit with lower and lower frequencies. This phenomenon happens irrespective of the unit of measurement and is scale invariant. We can test it in a simple way by applying the information-theoretic maximum entropy procedure (Jaynes, 1957) with the geometric mean: $\mu_g = \sum_{D=1}^{9} \log(D) p(D)$ where $\mu_g$ is the geometric expected value. The optimization problem (where the Shannon entropy is defined over the $D$'s) is then:

$$\begin{aligned} &\underset{\{P\}}{Maximize}\ H(P) = -\sum_D p(D) \log p(D) \\ &\quad subject\ to \\ &\mu_g = \sum_{D=1}^{9} \log(D) p(D) \\ &\sum_D p(D) = 1,\ p(D) \geq 0 \end{aligned} \tag{11}$$

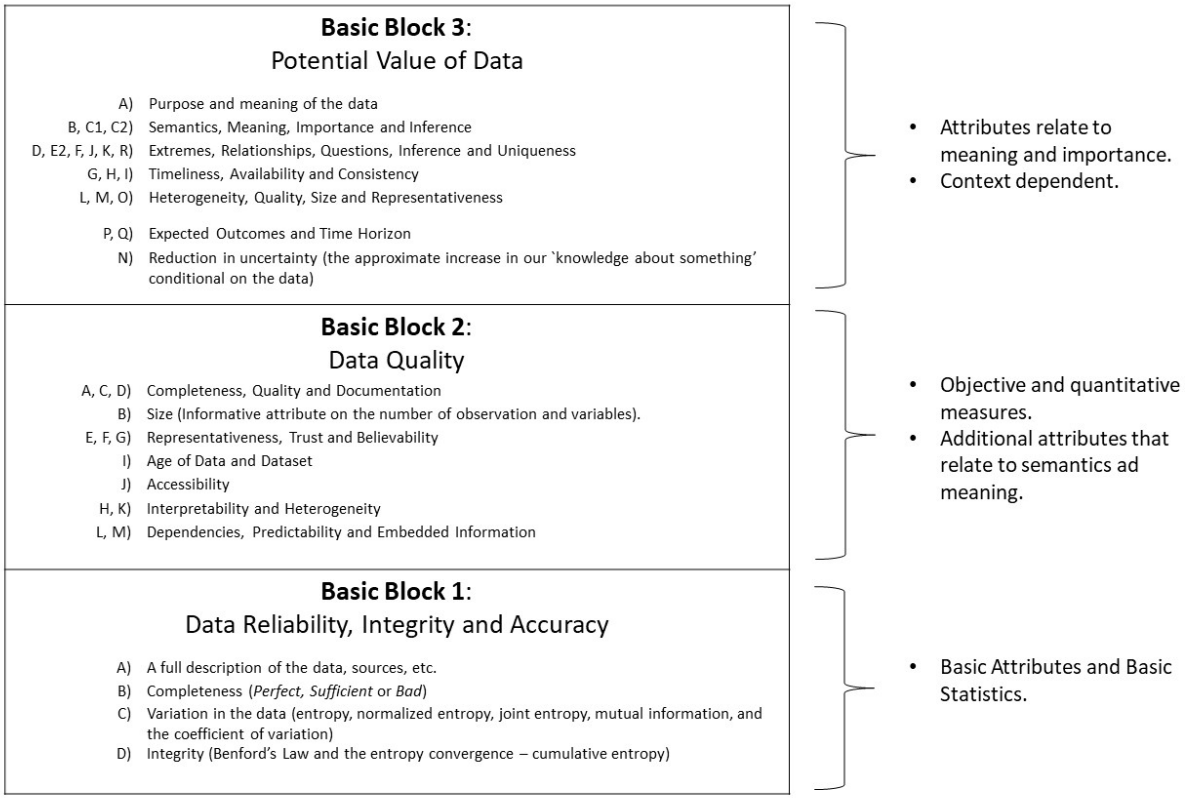
and the solution is $p(D) = D^{-\lambda} \Big/ \sum_D D^{-\lambda}$ where the exponent $\lambda$ is the Lagrange multiplier associated with the geometric mean constraint. Differences between the 'true' Benford's Law (Eq. 10) distribution and that derived via the maximum entropy procedure are indicative of potential abnormalities in the data. See Golan (2018) for more derivations and examples using Benford's Law. We note that Benford Law should not be used for studying data that do not span

at least a small number of magnitudes, say individuals' heights. We use this measure to investigate the integrity of datasets.

## 6 Attributes

The attributes used to evaluate the value of data are summarized for each one of the basic blocks described in Section 4. Figure 2 illustrates each basic block and the attributes within that block.

Figure 2. The Basic Building Blocks and their Attributes




### *Basic Block 1: Data Reliability, Integrity and Accuracy*

A) A full description of the data, sources, etc.
B) Completeness (of observations and variables). That measure takes three possible labels: *Perfect* (no missing information), *Sufficient* (there is some missing information, but it does not impact our knowledge, say not observing the zip code of an address), and *Bad* (essential information is missing). In most cases, detailed summary statistics (including missing values, and other problems) are suitable enough for evaluating this attribute.
C) Variation in the data (A basic summary of the variations, including entropy, normalized entropy, joint entropy, mutual information, and the coefficient of variation: standard-deviation/mean,).

D) Integrity. (Benford's Law and the entropy convergence – cumulative entropy.)

*The Score (Total):* 'Yes' (the data passed the minimal requirement) and 'No' (these data are worthless). This measure does not enter in the final tally of the overall value.

***Basic Block 2: Data Quality***

A) Completeness (of data as a whole: This is a theoretical concept with respect to the questions the dataset is supposed to be able to answer, whenever the questions preceded data collection, or with respect to how the data intend to be used). (1 – 10, with 10 complete).
B) Size (Number of observations and number of variables; This is an *informative* attribute that, together with other attributes, has an impact on the quality).
C) Documentation: do all proper documents about the data exist and are understandable? (0 – 2; no documents (0), some/all and partially understandable (1), perfect (2)).
D) All variables are defined correctly, including their units and meanings. (Each variable has a scale of 1 – 3, (not good) to perfect). Overall, the total normalized value: 3K/3K = 1. (This ensures that the total is in 0 – 1 and comparable across datasets.)
E) Are the data representative (with/without weights) of the underlying population (Yes (1) – No (0), sampling issues: Description).
F) Do we trust the data collected, the way it were collected, and the agency/individuals that publish the collected dataset? (Trust: 0 – 1, Collect: 0 – 1, Agency: 0 – 1).
G) Believability (this complements E and F above). Yes (1) – No (0).
H) Interpretability: Is it possible to provide coherent, logical, and consistent interpretation of each one of the variables (and for the resulting potential inference)? (Yes – No for each variable). Overall, the total normalized value: K/K = 1. (This ensures that the total is in the range 0 – 1 and comparable across datasets.)
I) Age:
    1. Age of the dataset availability (first date these data are available); using a discount factor from that date. The current year has the highest value of 1, then a discount factor for each previous year.
    2. Age of the data in the dataset (the most recent date of the main variable/s of interest); using a discount factor from that date; same scale as I1).

    The discount factor can be any positive number between 0 – 1 that is deemed reasonable or appropriate so long as it is consistent between datasets.
J) Accessibility: how easy it is to access the data, manage and manipulate it (1 – 5 (easiest/best)).
K) Heterogeneity – Does the dataset capture (approximately) the full heterogeneity of the underlying population. Yes (1) – No (0).
L) Data Dependencies (Condition number to capture the level of collinearity, 1 – 3: low (3) – very high (1), Correlation low (3) – very high (1)). Low values of collinearity and correlation are indicative of more information thus given higher scores.

M) Data Predictability and embedded information in the data. We use Shannon limit to provide an approximate notion of the non-redundant information in the data. This is a good proxy for the potential quality of prediction, but not how to do it. (See Golan, 2018, Chapter 3 for discussion.): scale 2 – 0, with 2 (compression ratio of about 2), 1 (compression ratio of about 1.5), 0 (compression ratio of about 1).

*The Score (Total):* Maximal value: 29 + Discounted Age of Data and Dataset (Criteria I)

***Basic Block 3: Potential Value***

A) Purpose and meaning of the data (Summary – in words).
B) Semantics and meaning: All the possible mutually exclusive types/families of questions that we can answer with these data given our current information and knowledge. See discussion of context and meaning in Section 4. (Scalar). Note that there are possibly different meanings under different contexts. Scale: 1 – 10 with 10 the highest. This is a relative measure based on the data user's subjective understanding. If there is more than a single user, the *median* of all users should be used.
C)
    1. Importance of questions to society. (Scale: 1 – 10 with 10 extremely important). Note, this is subjective and context dependent. (See discussion in B above.)
    2. Are some of the questions we can answer important for private entities? (Yes – No; If Yes scale 1 – 10 with 10 extremely important).
D) How many observations of extreme events are in the data? Scale: 0 – **Φ**. Calculation of **Φ**: the number of observations more than 3 standard deviations (sd) from the mean *plus* the number of observations between 2sd and 3sd from the mean multiplied by 0.5.
E)
    1. Is there a clear deterministic observed structure (story) in the data? (A description – no scale/weight).
    2. Is there information in the data that will allow potential studies of cause and effect? (Yes (1) – No (0))
F) Can we answer questions that we could not answer previously? (Yes (1) – No (0); If 'Yes,' How many?) Scale: 0 – # Questions.
G) Are the data timely (up to date)? (Yes (1) – No (0)).
H) For a repeated data, how fast are they published and made available for the users? (Scale: 0 – 2, too slow (0), somewhat regular (1), fast (2)).
I) Is the information (say, the variables) in repeated data consistent across time? (Scale: 0 – 1).
J) Do the data have the potential to shed new light on older questions-answers? (No – Maybe – Yes: 0, ½, 1).
K) Is it a completely new dataset – a data set that no one has ever used? (Yes (1) – No (0))
L) Size and representativeness of population affected. (Scale: 1-5 with 5 a large population is represented and effected).

M) Do the data present heterogenous information to properly answer questions? (No – Somewhat - Yes: 0 – ½ – 1; See also Quality block, J)

N) Approximate measure of reduction in uncertainty from the new data. It is the approximate increase in our 'knowledge about something' conditional on the data (Yes – No; If yes, how much; normalized entropy and relative information – Kullback-Leibler divergence). Scale: normalized entropy 0 – 1, normalized relative information (0 – 1). Note: The Yes – No are absolute, but the 'how much' is relative.

O) Data collection: experiment, administrative survey, or administrative/privately collected data? (Scale: 0 – 1 with all but private (1) and private (0)).

P) What are the potential expected outcomes/inferences in the short-run (data specific) and those in the long-run. (Description, expected importance: 1 – 10). This is a relative measure.

Q) Time horizon of resulting inferred outcomes, answers, and decisions. Scale: short (3), medium range (2), long range (1).)

R) Does another 'quite similar' dataset exist? (Yes (1) – No (0)); If yes, can combining both will increase the joint value by more than adding the values of both: economy to scale. That is, new questions and dimensions open up a higher level of precision. (Note, combining datasets is super-additive in terms of the number of questions that may be asked. It increases quadratically, just like the number of possible correlations. See also U below.) Scale (if 'yes') 1 – 10 with 10 the highest value.

***Related Descriptive Attributes***

S) Data Additivity (linear, super additivity, etc.; Define 'additivity' here and its dimension (say – more observations, more variables, etc.). No scale – discussion in the paper.

T) Size (number of observations vs. number of variables). Note that more observations may not increase the value as more variables. No scale/weight: this is a descriptive discussion, and the information is captured by other measures.

*The Score (Total):* Maximal value: 68 + **Φ** (from Criteria D).

## 7 The Potential Value of Data

Building on Section 6, the overall (relative) potential value of the data is the total sum of Blocks 2 and 3. That value is accompanied by some descriptive discussions as is explicitly shown in the attributes' specifications of the three blocks. It is important to consider the following when evaluating the potential value of a dataset:

- It is not recommended to assign a certain value to Block 1 (Data Reliability, Integrity and Accuracy). However, it is important to look at the proposed attributes of that block and decide whether the dataset should be used. It is expected that the answer will be 'yes' to most available datasets.

- As emphasized earlier, all values are relative (see discussion of semantic and context as well as relativity and subjectivity). Though some of the proposed attributes can be perfectly calculated, some are relative to the evaluator. If more than a single evaluator assigns the attributes' values, constructing the median (from values provided by the different evaluators) is recommended.
- The total sum of our proposed measure is composed of the sum of the different attributes. Each one of these attributes has a finite and bounded value. If the number of possible values is finite, say $Q$, then, the value of each attribute is bounded by $2^{|Q|}$. Aggregation in that case is trivial.
- Using the same evaluation scale provided here, and the same time discount factor, for different datasets, provides a ranking of these datasets even though these potential values are relative.

## 8 Empirical Example 1: Movie Releases

### 8.1 Description and Basic Statistics

We now show the way some of the measures and attributes proposed can be implemented and interpreted using a (toy) simple dataset from Craig, Greene, & Versaci (2015). This dataset contains characteristic information on 62 movie releases, along with box office performance statistics and metrics characterizing online "buzz" about movies prior to their release. The data contain a combination of discrete and continuous variables, which are discretized in such a way that the resulting distribution approximates the original distribution. Table 1 details the variables included in this data set and Table 2 presents their summary statistics.

The set of potential questions is described in Table 4 that presents the different variables used for different models aims to answer a variety of potential questions (Part B in the potential value block above).

Table 1. Description of Variables from Craig, et. al. (2015)

| *Variable* | *Description* |
|---|---|
| Opening Week Box Office Revenue ($USD) | First run U.S. box office. |
| Motion Picture Association Rating | MPAA Rating code, 1=G, 2=PG, 3=PG13 and 4=R. |
| Movie Budget ($USD in Millions) | Movie's production budget. |
| Star Power (Index) | Index created by *Forbes* magazine and published as part of its "Star Currency" list. The index ranges from 0 to 10 for each actor. The more famous the movie's stars are, the higher the index. The Star Power variable is the sum of indices among actors and actress in each movie. |
| Sequel | 1 if movie is a sequel, 0 if not. |
| Action | 1 if action film, 0 if not. |
| Comedy | 1 if comedy film, 0 if not. |
| Animated | 1 if animated film, 0 if not. |

| | |
|---|---|
| Horror | 1 if horror film, 0 if not. |
| Addict (Trailer Views) | Trailer views at traileraddict.com. |
| ComingSoon Website Comments | The number of message board comments at comingsoon.net about upcoming movies. Note that comments can be both positive and negative. |
| Fandango Votes | Public attention at fandango.com in the form of voting participation (i.e., the number of people that partook in voting). Participants can vote that they either “Don’t care” for the movie or that they “Can’t Wait” to see the movie. |
| Can’t Wait to See (Percent) | Percentage of Fandango voters that can't wait to see a movie. |

Table 2. Summary Statistics of Variables from Craig, et. al. (2015)

Panel A.
Discrete Variables

| Variable | Mean | Median | Max | Min | SD | CV |
|---|---|---|---|---|---|---|
| Sequel | 0.1 | 0 | 1 | 0 | 0.4 | 4.0 |
| Action | 0.2 | 0 | 1 | 0 | 0.4 | 2.0 |
| Comedy | 0.3 | 0 | 1 | 0 | 0.5 | 1.7 |
| Animated | 0.1 | 0 | 1 | 0 | 0.3 | 3.0 |
| Horror | 0.1 | 0 | 1 | 0 | 0.3 | 3.0 |
| Motion Picture Association Rating | 3 | 3 | 4 | 1 | 0.8 | 0.3 |

Continuous Variables

| Variable | Mean | Median | Max | Min | SD | CV |
|---|---|---|---|---|---|---|
| Opening Week Box Office Revenue ($USD) | 20,720,651.40 | 16,930,926.00 | 70,950,500.00 | 511,920.00 | 17,492,442.70 | 0.84 |
| Movie Budget ($USD in Millions) | 53.3 | 37.4 | 200 | 5 | 42.9 | 0.80 |
| Star Power (Index) | 18 | 18.1 | 36.8 | 0 | 8.9 | 0.49 |
| Addict (Trailer Views) | 5,933.80 | 3,480.00 | 45,865.70 | 568 | 7,674.60 | 1.29 |
| ComingSoon Website Comments | 78.2 | 36.5 | 594 | 2 | 124.6 | 1.59 |
| Fandango Votes | 522.3 | 430.5 | 1,778.00 | 35 | 390.7 | 0.75 |
| Can't Wait to See (Percent) | 0.5 | 0.5 | 0.8 | 0.1 | 0.2 | 0.40 |

Panel B.
Discretized Continuous Variables (Into 10 Bins)

| Discretized Variable | Mean | Median | SD | CV | Bin Size | |
|---|---|---|---|---|---|---|
| Opening Week Box Office Revenue | 3.3 | 3 | 2.4 | 0.7 | 7,043,858.00 | $USD |
| Movie Budget | 2.9 | 2 | 2.2 | 0.8 | 19.50 | $USD |
| Star Power | 5.4 | 5 | 2.4 | 0.4 | 3.68 | Index |
| Addict | 1.7 | 1 | 1.6 | 0.9 | 4,529.77 | Views |
| ComingSoon Website Comments | 1.8 | 1 | 2 | 1.1 | 59.20 | Comments |
| Fandango | 3.4 | 3 | 2.2 | 0.6 | 174.30 | Votes |
| Can't Wait to See | 5.7 | 6 | 2.4 | 0.4 | 0.07 | Percent |

## 8.2 The Attributes and Measures

### *Entropy of a Single Random Variable*

The entropy and the normalized entropy of each variable in the Craig, Greene, & Versaci (2015) dataset is calculated. The closer the distribution of a certain variable to uniform, the higher its entropy (and normalized entropy). This translates to a high level of uncertainty for that variable.

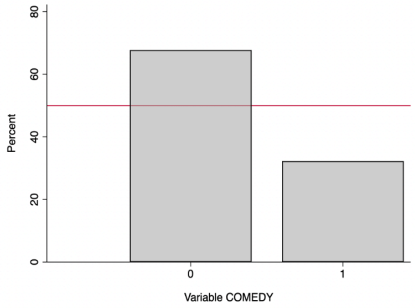
Figure 3, Panel A lists the entropy and the normalized entropy of each variable in the box office dataset. Panels B and C present the distributions of two binary variables identifying movies as of the comedy or animated genre, respectively. In each histogram, a red horizontal line corresponds to 50 percent or the uniform distribution for a binary variable. As is seen in the entropy values, the distribution of the comedy indicator is closer to uniformity than the distribution of the animated indicator. See Appendix Table 1 for sensitivity analyses using alternative entropy and normalized entropy calculations of the discretized continuous variables using 8 and 12 intervals for the discretization.

Figure 3. Entropy *and Normalized Entropy* of Variables from Craig, et. al. (2015)

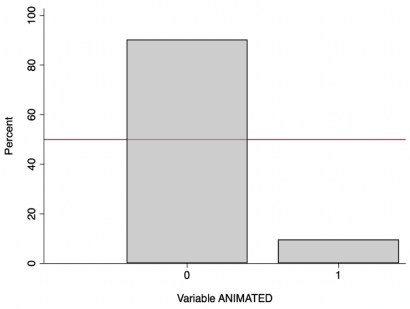
Panel A.

| *Variable* | *Entropy* | *Normalized Entropy* |
|---|---|---|
| Sequel | 0.598 | 0.598 |
| Action | 0.771 | 0.771 |
| **Comedy** | 0.907 | 0.907 |
| **Animated** | 0.459 | 0.459 |
| Horror | 0.459 | 0.459 |
| Motion Picture Association Rating | 1.685 | 0.842 |
| Discretized "ComingSoon Website Comments" | 1.404 | 0.423 |
| Discretized "Addict" | 1.526 | 0.460 |
| Discretized "Movie Budget" | 2.409 | 0.725 |
| Discretized "Opening Week Box Office Revenue" | 2.747 | 0.827 |
| Discretized "Fandango" | 2.791 | 0.840 |
| Discretized "Star Power" | 3.187 | 0.959 |
| Discretized "Can't Wait to See" | 3.199 | 0.963 |

Panel B. Distribution of Comedy




Panel C. Distribution of Animated




### *Information in a Single Random Variable*

Information is measured in bits. We get one unit, called a bit, of information when a choice is made between two alternatives. For example, we receive a bit of information when we are given a precise reply to a (binary) question answered by "yes" or "no." Shannon defines the information content of a single outcome $x_k$ as

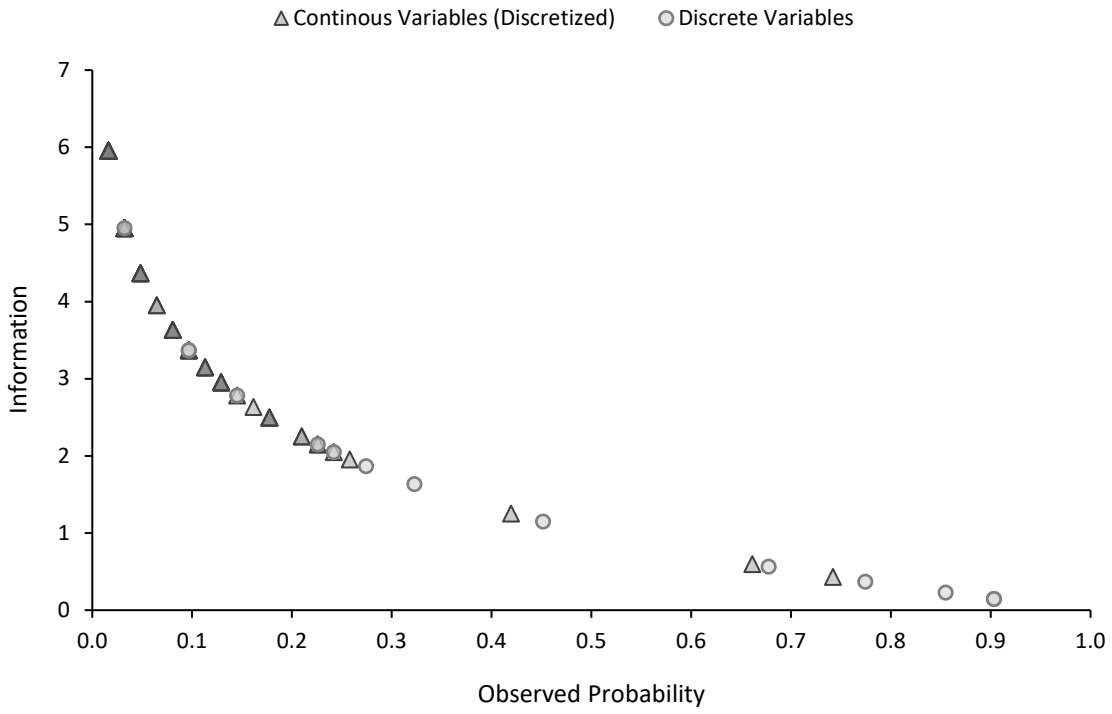
$$h(x_k) = h(p_k) = \log_2(1/p_k) = -\log_2(p_k). \quad (12)$$

Observing a rare outcome/event – at the tails of the distribution – provides much more information than observing an outcome around the mean of the data. Table 3 presents this inverse relationship among discrete variables in the box office dataset. Very few movies in the dataset are of either the animated or horror genre. The a-priori probability that a movie is of either genre is less than 10 percent. Consequently, the information associated with a new movie release being of the horror genre is over 20 times higher than the information associated with a new movie *not* being of this genre. Figure 4 shows the inverse relationship between probability and information for all 42 possible events across all discrete and discretized variables in the dataset. For a detailed discussion and graphical analyses of information, probabilities and entropy see Golan (2018).

Table 3. Information Associated with Discrete Variable Outcomes from Craig, et. al. (2015)

| *Variable* | *Outcome/Event* | *Information* | *Observed Probability* |
|---|---|---|---|
| Action | 0 / No | 0.369 | 0.774 |
| | 1 / Yes | 2.147 | 0.226 |
| Animated | 0 / No | 0.147 | 0.903 |
| | 1 / Yes | 3.369 | 0.097 |
| Comedy | 0 / No | 0.562 | 0.677 |
| | 1 / Yes | 1.632 | 0.323 |
| Horror | 0 / No | 0.147 | 0.903 |
| | 1 / Yes | 3.369 | 0.097 |
| Motion Picture Association Rating | 1 / G Rated | 4.954 | 0.032 |
| | 2 / PG Rated | 2.047 | 0.242 |
| | 3 / PG-13 Rated | 1.147 | 0.452 |
| | 4 / R Rated | 1.867 | 0.274 |
| Sequel | 0 / No | 0.226 | 0.855 |
| | 1 / Yes | 2.784 | 0.145 |

Figure 4. Relationship Between Information and Probability from Craig, et. al. (2015)



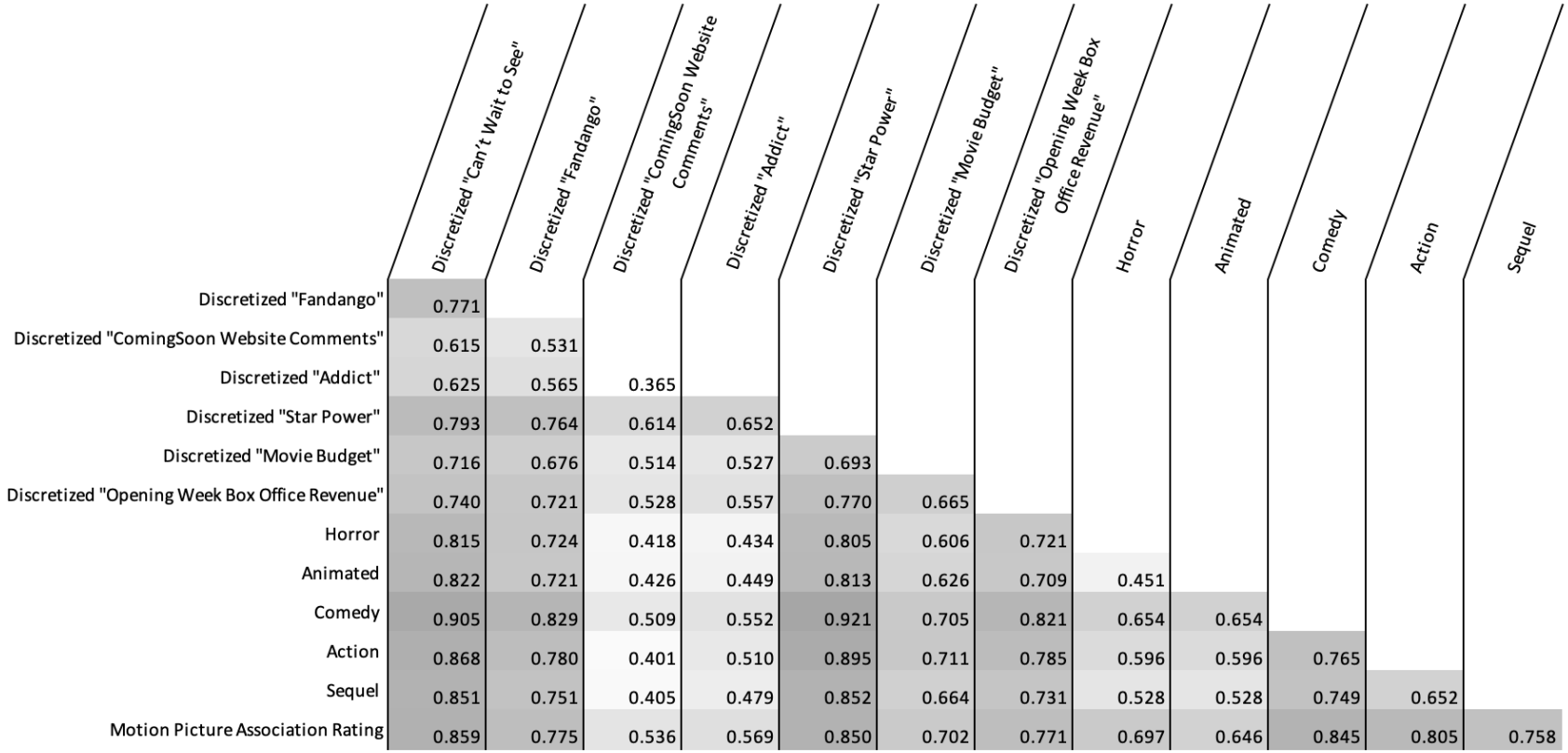

***Entropy of Multiple Random Variables***

Following equation (2), we calculate the joint entropy of each pair of variables in the dataset. These joint entropies are presented in Figure 5 with different gray scale highlighting combinations of variables with high entropy values and little joint information (and vice versa).

As an example, we show in Figure 6 the joint distribution of the 'comedy' genre with two other variables: the discretized star power (Panel A), and the trailer views on Addict (Panel B). For each variables-pair, there are 20 potential events. With uniform distribution the probability of each event is 0.05. A simple comparison of Panels A and B shows that the joint distribution of the comedy - star power pair is more uniform than that of the comedy - Addict trailer view pair. Consequently, the joint entropy of the former pair is higher than that of the later pair.

From an information, and surprise, point of view, consider Panel B. The comedy receives between approximately 9,628 to 14,157 trailer views (category 3). The probability of that event is approximately 0.016. Recalling that information is inversely related to the probability of an event, if a new comedy release were, for example, to receive more than 14,000 trailer views, such an event would be very surprising and informative. On the other hand, a new comedy release that receives less than 5,000 trailer views is much less informative as its probability is about 0.242.

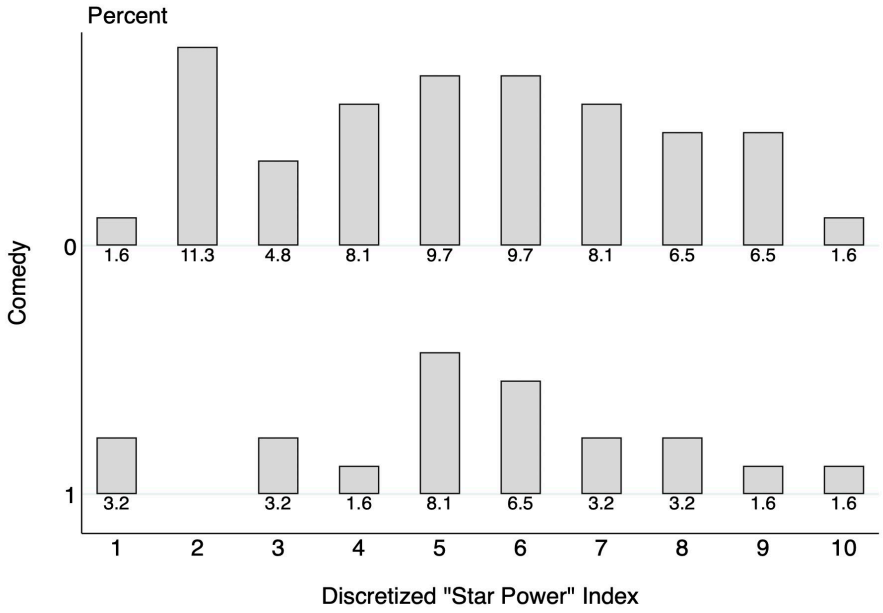
Figure 5. Normalized Joint Entropy of Variables from Craig, et. al. (2015)



| | Discretized "Can't Wait to See" | Discretized "Fandango" | Discretized "ComingSoon Website Comments" | Discretized "Addict" | Discretized "Star Power" | Discretized "Movie Budget" | Discretized "Opening Week Box Office Revenue" | Horror | Animated | Comedy | Action | Sequel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Discretized "Fandango" | 0.771 | | | | | | | | | | | |
| Discretized "ComingSoon Website Comments" | 0.615 | 0.531 | | | | | | | | | | |
| Discretized "Addict" | 0.625 | 0.565 | 0.365 | | | | | | | | | |
| Discretized "Star Power" | 0.793 | 0.764 | 0.614 | 0.652 | | | | | | | | |
| Discretized "Movie Budget" | 0.716 | 0.676 | 0.514 | 0.527 | 0.693 | | | | | | | |
| Discretized "Opening Week Box Office Revenue" | 0.740 | 0.721 | 0.528 | 0.557 | 0.770 | 0.665 | | | | | | |
| Horror | 0.815 | 0.724 | 0.418 | 0.434 | 0.805 | 0.606 | 0.721 | | | | | |
| Animated | 0.822 | 0.721 | 0.426 | 0.449 | 0.813 | 0.626 | 0.709 | 0.451 | | | | |
| Comedy | 0.905 | 0.829 | 0.509 | 0.552 | 0.921 | 0.705 | 0.821 | 0.654 | 0.654 | | | |
| Action | 0.868 | 0.780 | 0.401 | 0.510 | 0.895 | 0.711 | 0.785 | 0.596 | 0.596 | 0.765 | | |
| Sequel | 0.851 | 0.751 | 0.405 | 0.479 | 0.852 | 0.664 | 0.731 | 0.528 | 0.528 | 0.749 | 0.652 | |
| Motion Picture Association Rating | 0.859 | 0.775 | 0.536 | 0.569 | 0.850 | 0.702 | 0.771 | 0.697 | 0.646 | 0.845 | 0.805 | 0.758 |

Figure 6. Comparing the Joint Distribution of Select Variables from Craig, et. al. (2015)

Panel A. Joint Distribution of The Comedy Indicator
and the Discretized "Star Power" Variable




Panel B. Joint Distribution of The Comedy Indicator
and the Discretized "Addict" Variable

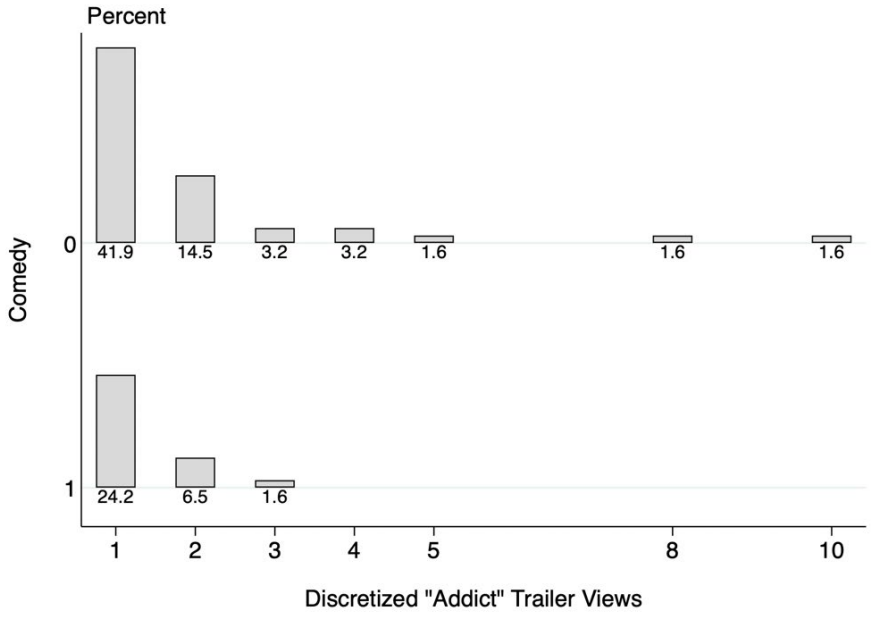
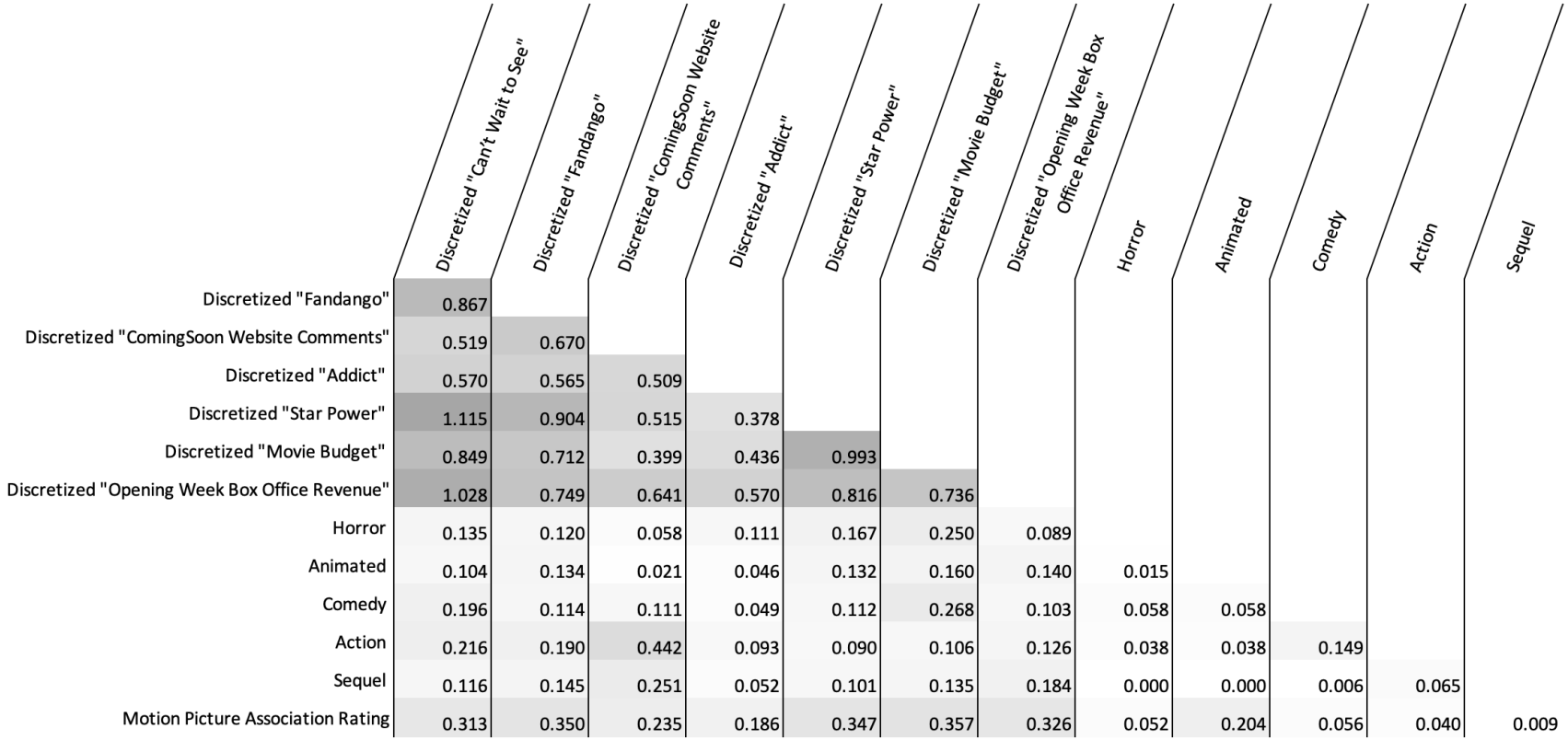


Figure 7. Mutual Information of all Pairs of Variables from Craig, et. al. (2015)

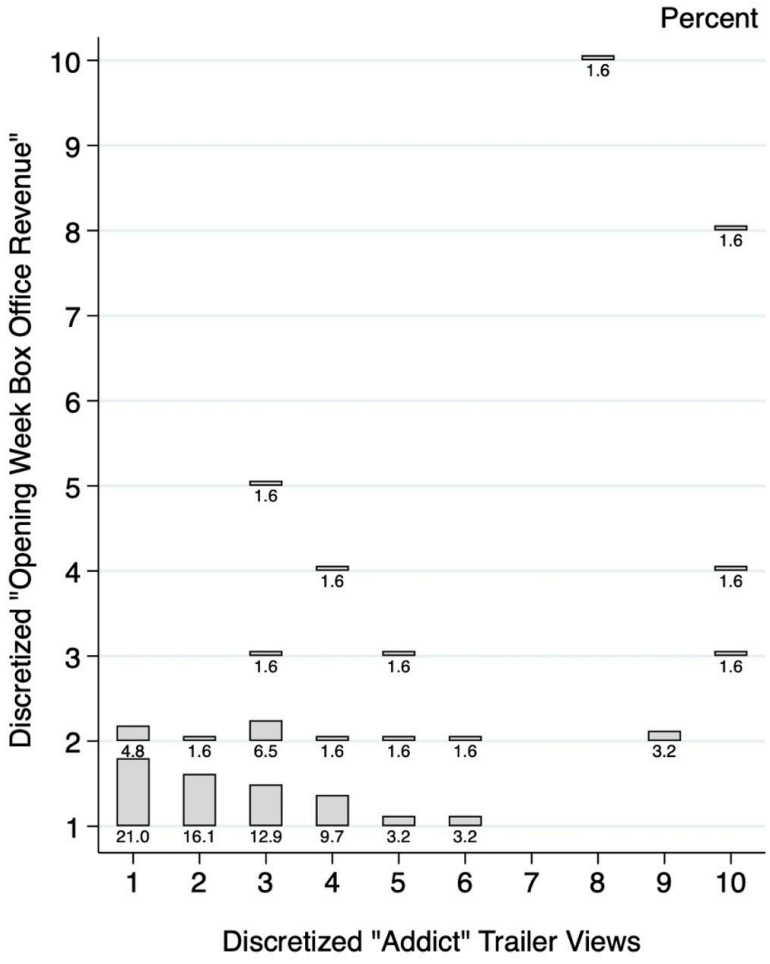


| | Discretized "Can't Wait to See" | Discretized "Fandango" | Discretized "ComingSoon Website Comments" | Discretized "Addict" | Discretized "Star Power" | Discretized "Movie Budget" | Discretized "Opening Week Box Office Revenue" | Horror | Animated | Comedy | Action | Sequel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Discretized "Fandango" | 0.867 | | | | | | | | | | | |
| Discretized "ComingSoon Website Comments" | 0.519 | 0.670 | | | | | | | | | | |
| Discretized "Addict" | 0.570 | 0.565 | 0.509 | | | | | | | | | |
| Discretized "Star Power" | 1.115 | 0.904 | 0.515 | 0.378 | | | | | | | | |
| Discretized "Movie Budget" | 0.849 | 0.712 | 0.399 | 0.436 | 0.993 | | | | | | | |
| Discretized "Opening Week Box Office Revenue" | 1.028 | 0.749 | 0.641 | 0.570 | 0.816 | 0.736 | | | | | | |
| Horror | 0.135 | 0.120 | 0.058 | 0.111 | 0.167 | 0.250 | 0.089 | | | | | |
| Animated | 0.104 | 0.134 | 0.021 | 0.046 | 0.132 | 0.160 | 0.140 | 0.015 | | | | |
| Comedy | 0.196 | 0.114 | 0.111 | 0.049 | 0.112 | 0.268 | 0.103 | 0.058 | 0.058 | | | |
| Action | 0.216 | 0.190 | 0.442 | 0.093 | 0.090 | 0.106 | 0.126 | 0.038 | 0.038 | 0.149 | | |
| Sequel | 0.116 | 0.145 | 0.251 | 0.052 | 0.101 | 0.135 | 0.184 | 0.000 | 0.000 | 0.006 | 0.065 | |
| Motion Picture Association Rating | 0.313 | 0.350 | 0.235 | 0.186 | 0.347 | 0.357 | 0.326 | 0.052 | 0.204 | 0.056 | 0.040 | 0.009 |

### *Mutual Information*

Figure 7 shows the mutual information (Eq. 5) for each pair of variables. Consider, for example, the Addict Trailer Views and Box Office pair. Their mutual information (Figure 7) is 0.57. To shed more light on the meaning of that number, consider their joint distribution (Figure 8). Most movies (approximately 66%) garnered less than 21.6 million dollars during their first run (categories 1-3). About 92% of all movies had between 500 to 14,157 trailer views on Addict (categories 1-3). As shown in Figure 8 and consistent with the findings of Craig, et. al. (2015), a movie rarely collects more than 21.6 million dollars at the box office (category 3 or higher), and when it does, most of the time it has over 32,276 trailer views on Addict (category 8 or higher). If, for example, we are interested in trying to predict the opening week revenues, given the number of trailer views, the marginal additional information gained from knowing the trailer views is relatively large. This is driven by the rare and extreme occurrence of box-office "hits" with, for example, over approximately 22 million dollars in revenue and over 32,300 trailer views.

Figure 8. Joint Distribution of the Discretized "Addict" and "Opening Week Box Office Revenue" Variables from Craig, et. al. (2015)



### *Set of Possible Questions and Related Condition Numbers*

We now summarize (Table 4) the set of possible questions one can answer using this dataset. We also show the potential variables that can be used to answer these questions, and we report their respected condition numbers.

Recalling that condition number captures the level of multicollinearity between combinations of variables, it is important to calculate those when thinking of the independent (including control) variables to be used in a model. The larger this number, the more ill-conditioned, or ill-behaved, the group of variables are, meaning they are more sensitive (less stable) for small changes in the model (Nath Datta, 2004). Within the context of the basic set of questions this dataset can tackle, we identified five possible characterizations and models (types), each is described by a subset of the variables. Variables may describe the costs, benefits, serve as inputs, attributes, or a signal. The results are shown in Table 4. Naturally, these five types are not mutually exclusive. Overall, the condition numbers are quite low and much below the threshold value discussed in Section 6.

Table 4. The Potential Set of Questions, and Condition Number Among Types of Variables from Craig, et. al. (2015)

| Type of Variables | Variables | Condition Number of Variables | Possible Question | Possible Dependent Variable |
|---|---|---|---|---|
| Measures or Determinants of Costs | Movie Budget ($USD in Millions), Star Power (Index), Motion Picture Association Rating, & Type of Movie (Sequel, Action, Comedy Animated and/or Horror) | 16.75 | What components of making a movie most influence its profitability? | Box Office Revenues ($USD) |
| Measures or Determinants of Short Run Benefits/Success | Fandango Votes, Can't Wait to See (Percent), ComingSoon Website Comments, Addict (Trailer Views), & Type of Movie (Sequel, Action, Comedy Animated and/or Horror) | 20.63 | Are movies that successfully garner view attention profitable? | Box Office Revenues ($USD) |
| Inputs | Movie Budget ($USD in Millions) & Star Power (Index) | 5.44 | What components of a movie capture viewer's attention? | Addict (Trailer Views) |
| Attributes | Motion Picture Association Rating, & Type of Movie (Sequel, Action, Comedy Animated and/or Horror) | 12.62 | What type of movies are viewers most interested in? | Can’t Wait to See (Percent) |
| Signals | Fandango Votes, Can't Wait to See (Percent), ComingSoon Website Comments, & Addict (Trailer Views) | 9.87 | Does media buzz signal a movies success at the box office? | Box Office Revenues ($USD) |

***Data Integrity: Benford's Law***

We now check whether the distributions of the first digit, of each variable, satisfy Benford Law. We use the maximum entropy approach (Jaynes, 1979), with the observed geometric mean as a constraint, to infer that distribution (see Section 6 and Golan, 2018). We repeat this process for the first non-zero digit of all continuous variables in the entire dataset. The resulting distribution is presented in Figure 9. The empirical distribution of all non-zero digits in the entire data set is close to the true Benford distribution suggesting that overall, the data satisfy our integrity requirement.

We repeated that analysis for each continuous variable in the data set. This is shown in Appendix Figure 1. All variables except for one, the percentage of Fandango participants that voted that they "Can't Wait to See" a movie, are close to the "true" Benford distribution. If the empirical distribution unexpectedly diverges from the Benford distribution or is abnormal in some way, then the data may be nonrandom. But, as discussed earlier, there are exceptions, such as if the data do not span at least several orders of magnitudes, which is the case with the 'Can't Wait' variable. This is also expected as the mean of that variable is equal to its median, meaning it cannot be a power law (where the mean is larger than the median). For completeness, Figure 10 presents the "Can't Wait to See" digit distribution. Figure 11 presents examples of variables that are consistent or inconsistent with Benford law.

Figure 9. Comparing Theoretical Benford Law Distribution to that of All Continuous Variables from Craig, et. al. (2015)

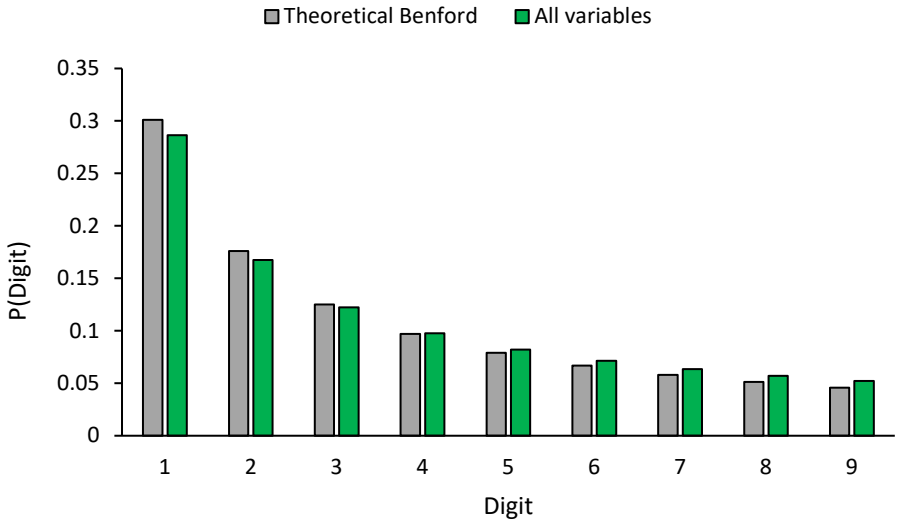


Figure 10. Comparing Theoretical Benford Law Distribution to the Can't Wait to See (in Percent) Variable from Craig, et. al. (2015)

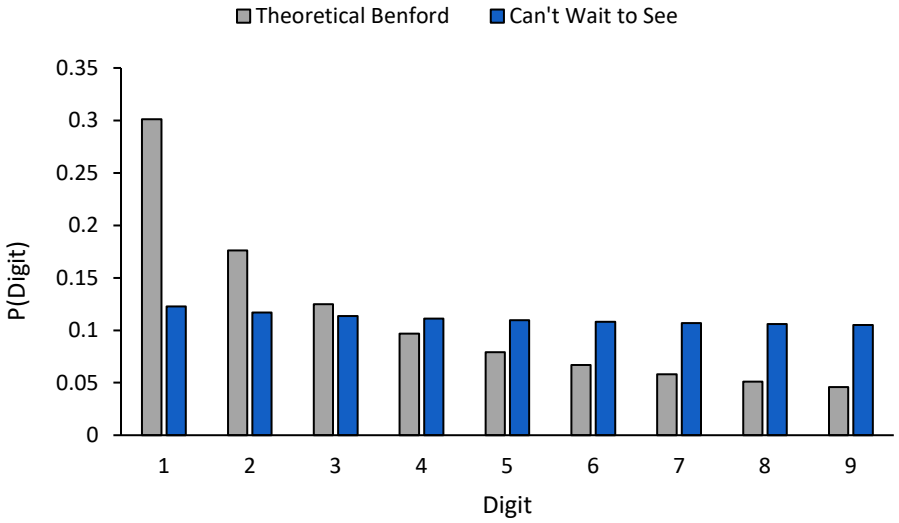
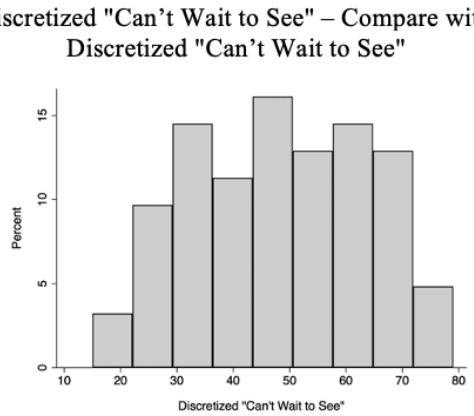
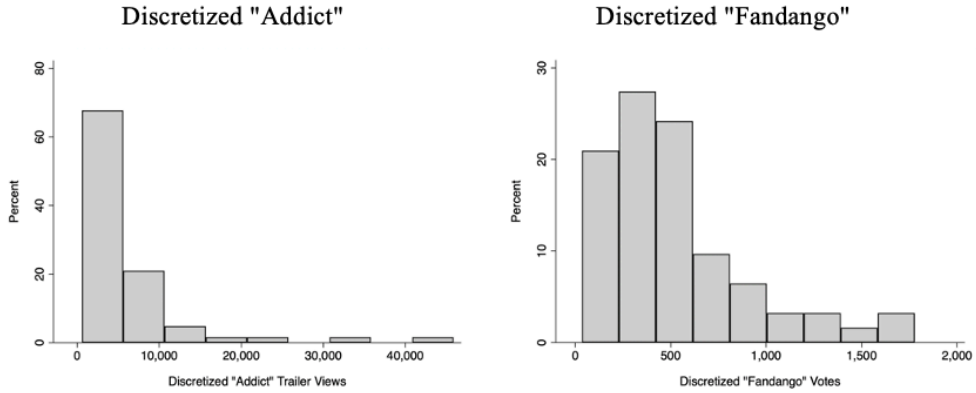



Figure 11. Comparing Empirical Digit Distributions of Variables that Do and Do Not Have Benford Distribution

Panel A. Discretized "Can't Wait to See" – Compare with Figure 8.



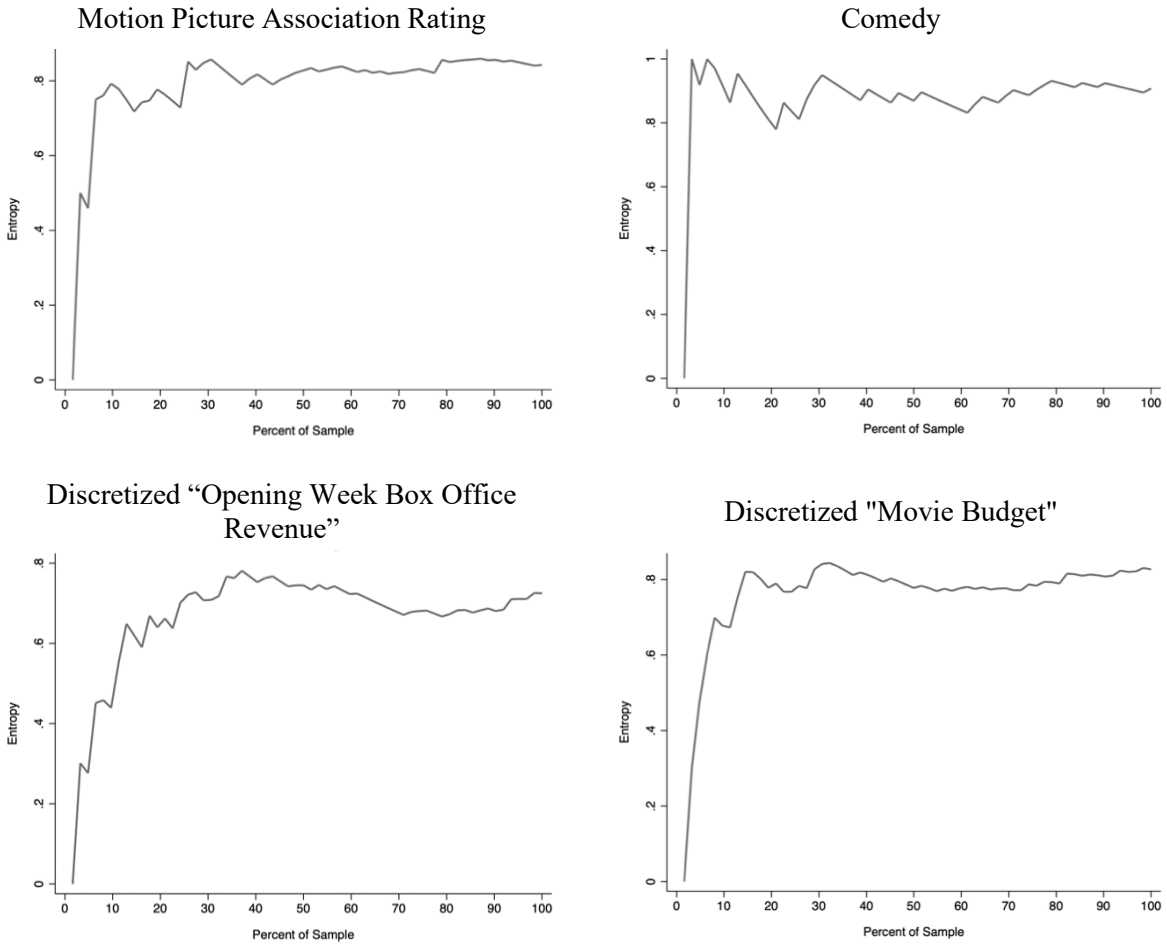

Panel B. Variables that Have a Benford Distribution of Their Digits - Compare with Figure 7



### *Cumulative Entropy or Entropy Convergence*

The cumulative entropy analyses shown below are consistent with the Benford Law analyses, and support our conclusion of the integrity of the data. Appendix 3, Figures 2 and 3 show the cumulative entropies for the rest of the variables in the box office dataset.

Figure 12. Cumulative Entropies for Select Variables from Craig, et. al. (2015)




### *Information Compression: The Shannon Limit*

The results are shown in Table 5. Overall, the data were compressed by approximately 31 percent = (1 – (Z/N)). This implies that the dataset can provide a decent level of inference and prediction. A simple calculation from Table 5 shows that the compression ratio is approximately 1.45.

Table 5. Shannon Limit Estimates Using Data from Craig, et. al. (2015)

| Shannon Limit | N | K | $\pi$ |
|---|---|---|---|
| 29,593.510 | 42,908.000 | 7,914.000 | 0.184 |

# 9 Empirical Example 2: Rural Urban Continuum Codes

## 9.1 Description and Basic Statistics

Here we evaluate a set of datasets called the Rural Urban Continuum Codes (RUCCs) published by the USDA. Over the last five decades (1974, 1983, 1993, 2003 and 2013) RUCC datasets have been produced to classify U.S. counties along the rural to urban spectrum based on population and physical adjacency to the nearest metropolitan area. The definition of each RUCC is provided in Table 6. County codes are assigned based on decennial population data, and metro and nonmetro delineations from the Office of Management and Budget (OMB). Each RUCC dataset contains each county's name, Federal Information Processing Standard (FIPS) code and RUCC. No other nonredundant characteristic information is consistently provided in these datasets.

Generally speaking, these datasets can provide a summary of how urbanized counties across the US are distributed, so one can argue that they have value on their own. However, the RUCC datasets are rarely used on their own; rather they are used together (time series of the RUCC's) or to supplement other datasets. Most of their value materializes when supporting (or complementing) other detests. For example, studying the geographically disadvantaged areas, where in conjunction with other data, say hospitals or welfare, can help the policy maker determine appropriate policies. The set of potential questions that this dataset can answer is described at the end of this section.

Table 6. Rural Urban Continuum Codes

| Code | Rural-Urban Classification |
|---|---|
| 0 | Central counties of metro areas of 1 million population or more (Prior to 2003) |
| 1 | Fringe counties of metro areas of 1 million population or more (Prior to 2003) |
| 1 | Counties in metro areas of 1 million population or more (from 2003 onward) |
| 2 | County in metro area of 250,000 to 1 million population |
| 3 | County in metro area of fewer than 250,000 population |
| 4 | Nonmetro county with urban population of 20,000 or more, adjacent to a metro area |
| 5 | Nonmetro county with urban population of 20,000 or more, not adjacent to a metro area |
| 6 | Nonmetro county with urban population of 2,500-19,999, adjacent to a metro area |
| 7 | Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area |
| 8 | Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area |
| 9 | Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area |

In this empirical example, we present different measures for the latest 2013 RUCC dataset and a merged dataset containing RUCC codes from the last five decades. Unlike the first (movie) example, these datasets have missing information. Table 7 provides summary statistics for the 2013 RUCC dataset. In that case, two of the 3,234 counties in this 2013 dataset are missing classification codes. Removing these two counties from the dataset has no meaningful impact on the informational content of the data. The distribution of each variable is practically unchanged.

Table 7. Summary Statistics of 2013 RUCC Dataset

| Variable | N | Mean | Median | Max. | Min. | SD | CV |
|---|---|---|---|---|---|---|---|
| Rural Urban Continuum Codes (RUCC), 2013 | 3,232 | 4.9 | 6.0 | 9.0 | 1.0 | 2.7 | 0.6 |
| County FIPS codes | 3,234 | 31,544.7 | 30,038.0 | 78,030.0 | 1,001.0 | 16,425.5 | 0.5 |
| 2010 Census population | 3,234 | 96,736.7 | 26,074.0 | 9,818,605.0 | 0.0 | 308,718.3 | 3.2 |
| Encoded state variable | 3,234 | 29.9 | 30.0 | 56.0 | 1.0 | 15.4 | 0.5 |
| Encoded description of Rural Urban Continuum Codes | 3,234 | 4.5 | 5.0 | 10.0 | 1.0 | 2.4 | 0.5 |

On the other hand, when all five RUCC datasets are merged, 106 of 3,241 counties have at least one piece of missing information or one variable with missing values. Table 8 presents summary statistics for the fully merged dataset. The first column of Table 8 shows the total number of counties that do not have missing data for each variable. Approximately, a quarter of the 106 counties with at least one piece of missing information are from Puerto Rico, whose municipios were first classified into RUCCs in 2013. Due to the large number of counties in the U.S., summary statistics are very similar between the full dataset, shown in Table 8, and the same dataset excluding the 106 counties with missing information.

In cases where missing data are not at random, dropping the observations with missing information is unadvisable. We can still employ the measures developed here if we treat these observations with care, such as creating a new category (dummy variable) for observations with missing values. Note that measures computed with and without missing data should be consistent when information is missing at random, or the dataset is similarly and meaningfully representative after the exclusion of missing information. Unless otherwise specified, the results presented here are computed without dropping any observation from the data; we treat missing information as a separate category. Consequently, the sum of probabilities corresponding to each observed event and the probability of observing no data is one.

Thus far, we have shown that the RUCC data are relatively complete and almost complete for mainland counties (Basic Block 1). In the next section we report measures of entropy and mutual information, which capture the information and variation in the data. We also

show the cumulative entropy, which confirms the integrity of RUCCs. These datasets are inherently complete, representative, and as heterogenous as possible given that all US counties are included (Basic Block 2, Criteria E, and K). The size of each dataset is limited to the number of counties in existence (over three thousand). In that way, the dataset captures the complete population. But only a few variables are consistently reported in each decade's dataset, thereby limiting its 'stand-alone' value (Basic Block 2, Criteria A). However, as mentioned earlier, that dataset is often used as a reference, supplement, or crosswalk to other datasets. We show in the next section that the amount of non-redundant information in these RUCCs is quite small (Basic Block 2, Criteria L and M).

These data are at the aggregate county level and no personal identifiable information is included. The USDA makes this complete dataset available and accessible to the public (Basic Block 2, Criteria J). They also provide detailed documentation on the definition of each code, how codes are constructed, discussion of the data used to designate codes, and changes over time[3] (Basic Block 2, C, F and I).

Table 8. Summary Statistics of Merged RUCC Datasets (1974, 1983, 1993, 2003, and 2013)

| Variable | N | Mean | Median | Maximum | Minimum | SD | CV |
|---|---|---|---|---|---|---|---|
| Rural Urban Continuum Codes (RUCC), 2013 | 3,232 | 4.9 | 6.00 | 9.0 | 1.0 | 2.7 | 0.6 |
| Rural Urban Continuum Codes (RUCC), 2003 | 3,142 | 5.1 | 6.00 | 9.0 | 1.0 | 2.7 | 0.5 |
| Rural Urban Continuum Codes (RUCC), 1993 | 3,142 | 5.6 | 6.00 | 9.0 | 0.0 | 2.7 | 0.5 |
| Rural Urban Continuum Codes (RUCC), 1983 | 3,141 | 5.8 | 6.00 | 9.0 | 0.0 | 2.6 | 0.4 |
| Rural Urban Continuum Codes (RUCC), 1974 | 3,141 | 6.0 | 7.00 | 9.0 | 0.0 | 2.5 | 0.4 |
| County FIPS codes | 3,241 | 31,515.0 | 30,035.00 | 78,030.0 | 1,001.0 | 16,451.2 | 0.5 |
| 2010 Census population | 3,234 | 96,736.7 | 26,074.00 | 9,818,605.0 | 0.0 | 308,718.3 | 3.2 |
| 2000 Census population | 3,141 | 89,596.3 | 24,595.00 | 9,519,338.0 | 67.0 | 292,462.2 | 3.3 |
| Percent of workers in nonmetro counties commuting to central counties of adjacent metro areas in 2013 | 3,142 | 5.0 | 0.00 | 64.5 | 0.0 | 9.2 | 1.8 |
| Encoded state variable | 3,241 | 29.9 | 30.00 | 56.0 | 1.0 | 15.4 | 0.5 |
| Encoded description of Rural Urban Continuum Codes (RUCC), 2003 | 3,142 | 4.7 | 5.00 | 9.0 | 1.0 | 2.3 | 0.5 |
| Encoded description of Rural Urban Continuum Codes (RUCC), 2013 | 3,234 | 4.5 | 5.00 | 10.0 | 1.0 | 2.4 | 0.5 |

[3] Documentation is available at <https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/documentation/>.

## 9.2 Attributes and Measures

We now report some of the proposed measures and attributes discussed earlier. Measures were computed under two scenarios of missing information: dropped or treated as a separate category or event. As previously discussed, excluding missing information did not meaningfully change any of the results emerging from that dataset. For example, Table 9 presents the distribution of RUCC codes for 2013 with and without missing information. The probability of observing each type of county changes only marginally. Consequently, the information content from each RUCC remains consistent.

Table 9. Information Content of RUCC with and without Missing Information, 2013

| With Missing Information | | RUCC | Without Missing Information | |
|---|---|---|---|---|
| Info. Content | Prob. | | Info. Content | Prob. |
| 2.780 | 0.146 | Counties in metro areas of 1 million population + | 2.866 | 0.137 |
| 3.037 | 0.122 | County in metro area of 250,000 to 1 million population | 3.048 | 0.121 |
| 3.135 | 0.114 | County in metro area of fewer than 250,000 population | 3.139 | 0.114 |
| 3.901 | 0.067 | Nonmetro county w/ urban pop. of 20,000 +, adjacent to metro area | 3.873 | 0.068 |
| 5.048 | 0.030 | Nonmetro county w/ urban pop. of 20,000 +, not adjacent to a metro area | 5.091 | 0.029 |
| 2.441 | 0.184 | Nonmetro county w/ urban pop. of 2,500-19,999, adjacent to a metro area | 2.402 | 0.189 |
| 2.894 | 0.135 | Nonmetro - Urban pop. of 2,500 - 19,999, not adjacent to metro area | 2.856 | 0.138 |
| 3.881 | 0.068 | Nonmetro - Completely rural or < 2,500 urban pop., adjacent to metro area | 3.833 | 0.070 |
| 2.921 | 0.132 | Nonmetro - Completely rural or < 2,500 urban pop., not adjacent to metro area | 2.907 | 0.133 |
| 8.492 | 0.003 | Missings | | |

Table 10 presents the normalized entropy of RUCC from 1974 to 2013 with and without missing information. Differences are no more than six percent and the relative entropies between the codes by decade remain quite similar.

Table 10. Normalized Entropy of RUCC By Decade with and without Missing Information

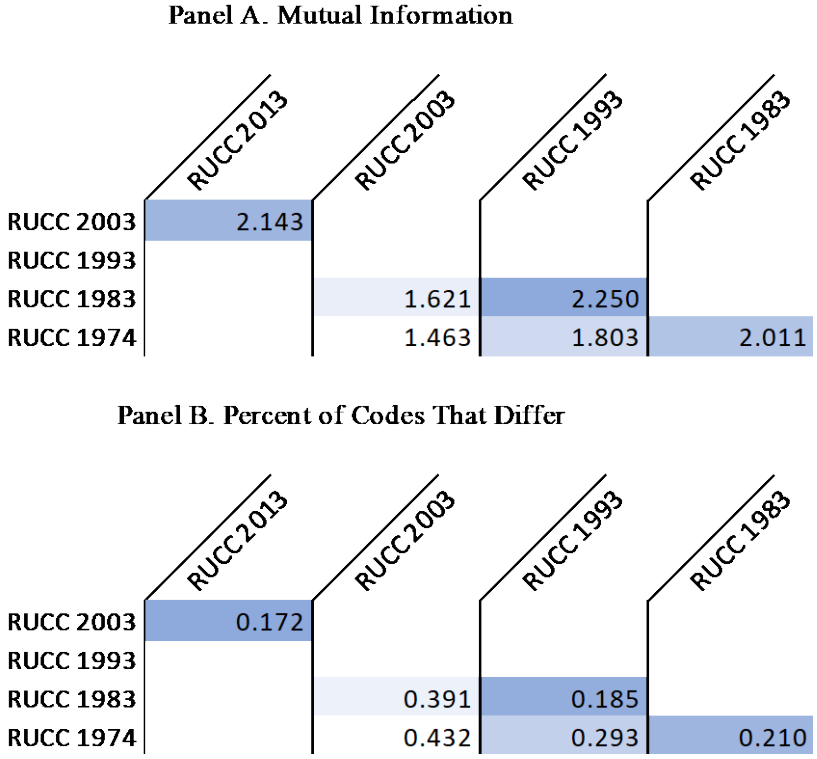
| Variable | With Missing Information | Without Missing Information |
|---|---|---|
| RUCC 1974 | 0.844 | 0.894 |
| RUCC 1983 | 0.850 | 0.900 |
| RUCC 1993 | 0.867 | 0.918 |
| RUCC 2003 | 0.901 | 0.960 |
| RUCC 2013 | 0.913 | 0.958 |

When using the merged dataset of RUCC, users are able to observe the distribution of counties during each decade and how this distribution has changed over time. However,

comparisons over time should be made with caution as the meaning of some codes, particularly the codes characterizing counties in metro areas with more than one million residents, changed in 2003.

Direct distributional comparisons can be made between codes in 2013 and 2003, and 1974 to 1993. Figure 13, Panel A presents the mutual information from such comparisons. For example, the mutual information between RUCC in 2003 and 2013 represents the marginal additional information gained from using RUCC in 2003 to analyze RUCC in 2013. As shown in Figure 13, Panel B, mutual information is inversely related to the percent of counties whose code changes between any two years. When changes are less frequent, the marginal information gained is greater, especially when changes are concentrated among certain types of counties.

Figure 13. Mutual Information Between Comparable RUCC Throughout the Decades




Note that RUCC are comparable between 2013 and 2003, and 1974 to 1993. See Table 7 for additional details.

Figure 14 presents the cumulative entropy of RUCC for 2003 and 2013. The entropy of each variable very quickly converges. For both variables, the measure converges after only considering ten percent of the sample suggesting no issues with the data's integrity.

Figure 14. Cumulative Entropy of RUCC in 2003 and 2013

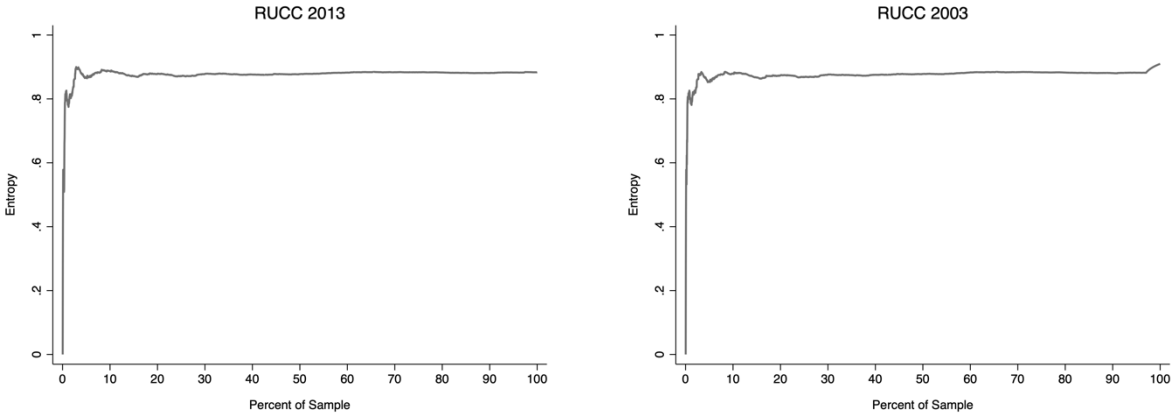



Table 11 summarizes the Shannon limit of the merged RUCC dataset with codes for five decades, and the 2013 RUCC dataset. The information in the former dataset can only be compressed by approximately 6 percent. Meanwhile the information in the later dataset, whose information is a subset of the former dataset, can be compressed by approximately 19 percent. The compression ratios are 1.066 and 1.229 respectively for the first and the second.

Table 11. The Shannon Limit for Two Datasets

| | Compressed By | Shannon Limit | N | K | $\pi$ |
|---|---|---|---|---|---|
| Merged Dataset for RUCC 1974 to 1993 | 6.2 % | 5,530,401 | 5,897,901 | 2,088,536 | 0.354 |
| RUCC 2013 Dataset | 18.6% | 5,046,081 | 6,199,276 | 1,560,427 | 0.252 |

As previously discussed, these RUCC datasets can be characterized as support datasets with metadata characterizing political geographies (Basic Block 3, Criteria A). Their meaning lie in the exact way the RUCCs are defined, and only questions on urbanization and the distribution of urban locations can be answered with this dataset alone. The merged RUCC dataset also provides information and changes over time, but without additional context cannot be used to summarize, predict, or explain causal relationships (Basic Block 3, Criteria B, C1 and E2). Consequently, measures like the condition number have no meaning (and value) in this case. In the next section we present the set of questions this dataset can be used to answer.

Combined, however, with other micro or macro level data will increase the value of both datasets and will provide the basis for answering a whole family of new questions. For example, RUCCs can serve as geographic controls for micro economic analysis using individual or house level data and help answer questions ranging from urban-rural migration to socioeconomic wellbeing. Or consider another simple example, where the RUCCs are combined with additional aggregate county level indicators for studying the patterns between local gross domestic product and urbanization. Thus, as emphasized earlier, the potential value of this (support) dataset

increases when combined, presumably as designed, with other datasets and causal relationships can be disentangled by conditioning on urban-rural status.

The set of possible questions these data can be used to answer is discussed below. However, in this case, to answer some of these questions may require supplementary data.

1. *What is the distribution of counties (by population density) within states?*
   (Note that this could be extended for distribution of urban and rural counties. However, the dataset has multiple classifications, and there are different designations of "rural" and not all non "rural" designations are necessarily urban.)
2. *How has the distribution of counties changed over time?*
3. *Which States have large rural and less populated areas?*

   (Note that this question may not be answered using the RUCC dataset alone. One may need to merge the RUCC with another dataset.)
4. *What states are primarily comprised of rural counties?*
5. *Are there rural areas (in a State) that are far (say, more than 50 or 100 miles) from big metropolitan areas?*
   (Note: Again, it is unclear if such a question can be answered using the RUCC data alone, as we would likely need geospatial data in the form of distance and size to determine (a) which counties neighbor each other and (b) how far counties are from the closest or largest metropolitan area in their state.)
6. *Could certain counties be defined (or designated) as less advantaged?*

   (Note: Assuming, for example, that population is used a as a proxy for economic activity we could determine which counties have less economic activity and are likely less advantaged.)

## 10 Empirical Example 3: Agricultural Resource Management Survey Data

### 10.1 Description and Basic Statistics

In this example we apply our approach to a very large and informative dataset. It is a survey data on the U.S. farming sector from 2003 to 2021. Industry income and expenditure information are collected by the National Agricultural Statistics Service (NASS) of the USDA via the Agricultural Resource Management Survey (ARMS). These data are used in a variety of USDA reports evaluating the state of the farming sector and its changes over time. The primary report summarizing survey results is the Farm Production Expenditures Report, which includes state and national estimates on average and total annual income earned, and expenses incurred, by farmers. Many other researchers and government organizations also use these data to produce their own estimates and to study the sector's strength and contribution to the overall economy. Farm level micro data from the survey are not published by NASS, rather only the averages at the state and national level are publicly available. This sample has state level data for the

following States: Arkansas, California, Florida, Georgia, Illinois, Indiana, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Carolina, Texas, Washington, and Wisconsin.

The set of questions, accompanied by the condition number for each potential question, is discussed in Section 10.3.

Table 12 presents summary statistics of the panel of ARMS data provided by NASS. Each variable (except for number of farms) is a farm-level average produced using survey data from each of the 15 states. The dollar values are not adjusted for inflation. Appendix B, Table 2 contains the ARMS definitions for all income and expenses variables. All variables are in $USD and complete except for the "value of inventory change" variable, which captures changes in the market value of crops produced over the calendar year. Eight of the 285 observations in the panel are missing data for this variable. Missing data are not consistently observed from any one year or state, and summary statistics remain meaningfully equivalent when corresponding observations are dropped. When comparing summary statistics between the full panel with and without missing information, no single statistic differs by more than five percent and on average all summary statistics differ by less than one percent.

Table 12. Summary Statistics of the ARMS Data, 2003 - 2021

| | Mean | Median | Maximum | Minimum | SD | CV |
|---|---|---|---|---|---|---|
| Number of Farms | 74,778.1 | 62,001.0 | 249,002.0 | 34,001.0 | 48,541.5 | 0.65 |
| Gross cash farm income | 190,754.3 | 157,773.0 | 739,452.0 | 40,646.0 | 127,539.5 | 0.67 |
| Livestock income | 54,193.3 | 36,554.0 | 222,741.0 | 7,247.0 | 42,033.7 | 0.78 |
| Crop sales | 104,676.0 | 83,945.0 | 492,628.0 | 11,403.0 | 81,196.6 | 0.78 |
| Government payments | 6,835.4 | 5,575.0 | 39,717.0 | 463.0 | 5,377.0 | 0.79 |
| Other farm-related income | 25,049.6 | 20,909.0 | 121,253.0 | 4,363.0 | 17,794.8 | 0.71 |
| Total cash expenses | 142,369.0 | 116,318.0 | 544,858.0 | 33,765.0 | 92,776.7 | 0.65 |
| Variable expenses | 113,536.8 | 94,016.0 | 473,447.0 | 26,128.0 | 78,390.3 | 0.69 |
| Livestock purchases | 9,668.4 | 2,928.0 | 111,548.0 | 146.0 | 15,887.9 | 1.64 |
| Feed | 16,214.4 | 11,559.0 | 77,174.0 | 1,968.0 | 14,263.4 | 0.88 |
| Other livestock-related expenses | 2,906.5 | 1,929.0 | 31,462.0 | 302.0 | 3,366.6 | 1.16 |
| Seed and plants | 11,386.1 | 9,464.0 | 30,279.0 | 1,521.0 | 6,811.9 | 0.60 |
| Fertilizer and chemicals | 22,312.7 | 18,796.0 | 66,351.0 | 4,187.0 | 12,450.1 | 0.56 |
| Utility expenses | 4,550.9 | 2,764.0 | 39,979.0 | 864.0 | 5,696.2 | 1.25 |
| Labor expenses | 20,335.6 | 9,060.0 | 179,227.0 | 1,639.0 | 29,662.2 | 1.46 |
| Fuels and oils | 7,843.9 | 7,493.0 | 19,788.0 | 1,802.0 | 3,447.6 | 0.44 |
| Repairs and maintenance | 8,646.6 | 7,884.0 | 22,140.0 | 3,029.0 | 3,903.3 | 0.45 |
| Machine-hire and custom work | 3,725.0 | 2,395.0 | 27,234.0 | 640.0 | 4,193.1 | 1.13 |
| Other variable expenses | 5,974.0 | 4,189.0 | 32,848.0 | 1,532.0 | 5,353.4 | 0.90 |
| Fixed expenses | 28,832.2 | 22,571.0 | 94,595.0 | 7,023.0 | 18,592.7 | 0.64 |
| Real estate and property taxes | 4,550.0 | 3,647.0 | 18,559.0 | 1,271.0 | 3,064.6 | 0.67 |
| Interest | 5,918.1 | 5,511.0 | 15,509.0 | 1,447.0 | 2,869.0 | 0.48 |
| Insurance premiums | 4,863.4 | 4,045.0 | 15,766.0 | 1,389.0 | 2,821.2 | 0.58 |
| Rent and lease payments | 13,500.7 | 8,770.0 | 52,734.0 | 1,410.0 | 11,175.4 | 0.83 |
| Net cash farm income | 48,385.4 | 37,345.0 | 206,700.0 | -915.0 | 37,210.8 | 0.77 |
| Nonmoney income | 7,632.5 | 6,937.0 | 24,517.0 | 3,603.0 | 2,744.9 | 0.36 |
| Value of inventory change | 4,943.0 | 3,393.0 | 32,616.0 | -13,386.0 | 7,471.4 | 1.51 |
| Depreciation | 13,794.2 | 12,152.0 | 41,483.0 | 4,057.0 | 7,157.2 | 0.52 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Labor, non-cash benefits | 297.4 | 213.0 | 2,465.0 | 13.0 | 268.1 | 0.90 |
| Adjusted breeding livestock income | 607.0 | 162.0 | 8,465.0 | 0.0 | 1,149.1 | 1.89 |
| Net farm income | 46,128.5 | 36,743.0 | 194,770.0 | -132.0 | 34,659.6 | 0.75 |

All variables except "Value of inventory change" contain N=285 observations. The "Value of inventory change" variable contains 277 estimates. See Appendix Table 2 for definitions of each variable provided by NASS.

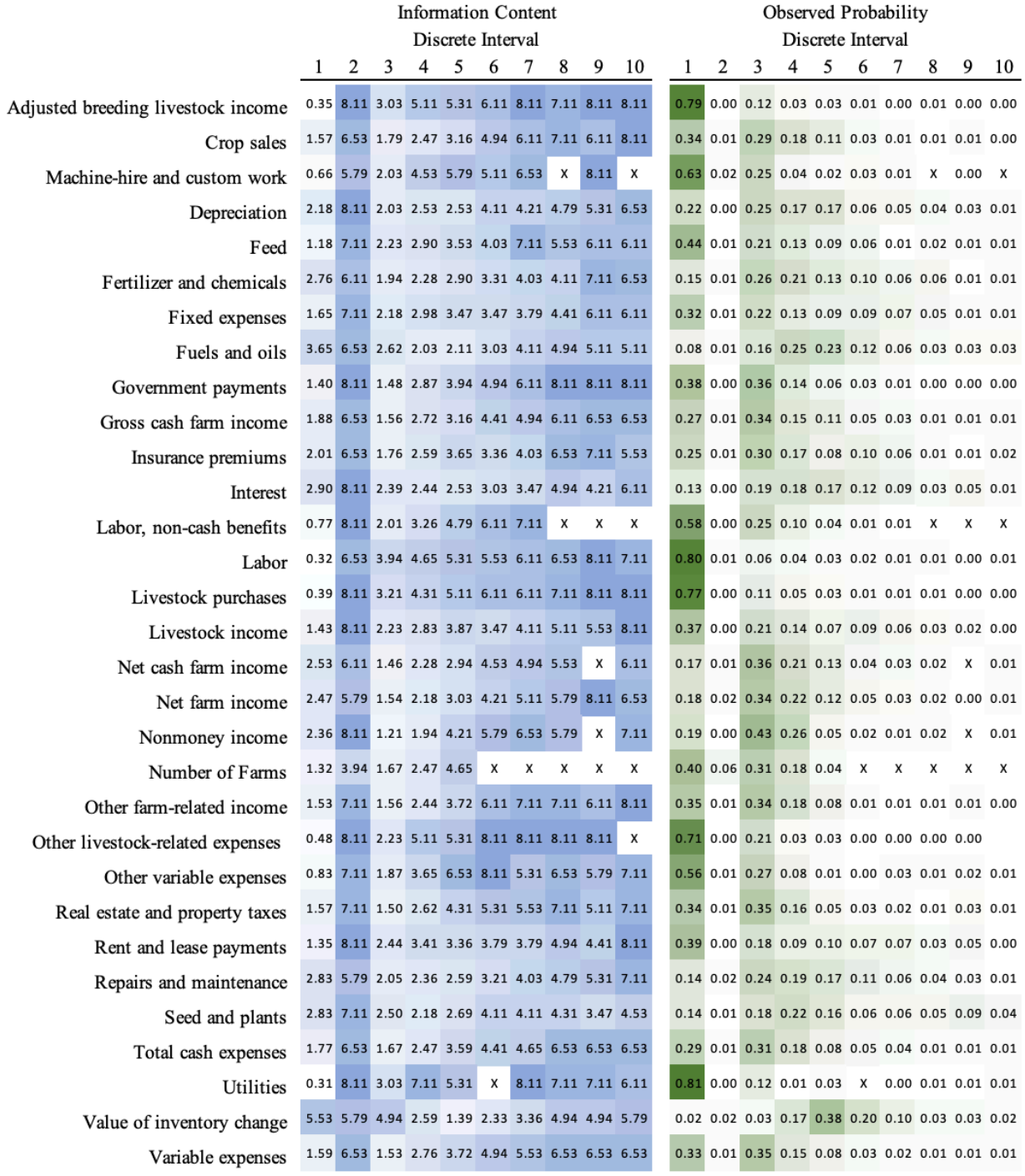
To compute the previously defined and discussed measures, all variables are discretized into ten intervals as done in prior empirical examples. Table 13 presents the mean, standard deviation and median of all discretized variables for the subset of 277 observations with complete information. The discretized average of all income and expenditure variables is less than five, and the discretized median of most variables is less than or equal to three. These statistics suggest the existence of observations that are very far from the means. Figure 15 presents the informational content and probability of each variable's discretized state. The probability that an average farm in any one state in the panel is garnering income or accruing operating costs in the top interval is less than one percent.

Table 13. Summary Statistics of Variables Discretized into Ten Intervals

| | Median | Mean | SD |
|---|---|---|---|
| Other livestock-related expenses | 1 | 1.5 | 1 |
| Utility expenses | 1 | 1.5 | 1.4 |
| Adjusted breeding livestock income | 1 | 1.5 | 1.2 |
| Livestock purchases | 1 | 1.6 | 1.3 |
| Labor | 1 | 1.6 | 1.6 |
| Machine-hire and custom work | 1 | 1.7 | 1.6 |
| Labor, non-cash benefits | 1 | 1.7 | 1.1 |
| Other variable expenses | 1 | 1.9 | 1.7 |
| Government payments | 2 | 2.1 | 1.4 |
| Other farm-related income | 2 | 2.3 | 1.5 |
| Number of Farms | 2 | 2.4 | 2.2 |
| Variable expenses | 2 | 2.4 | 1.8 |
| Feed | 2 | 2.4 | 1.9 |
| Real estate and property taxes | 2 | 2.4 | 1.8 |
| Nonmoney income | 2 | 2.4 | 1.4 |
| Crop sales | 2 | 2.5 | 1.7 |
| Total cash expenses | 2 | 2.6 | 1.8 |
| Gross cash farm income | 2 | 2.7 | 1.8 |
| Livestock income | 2 | 2.7 | 2 |
| Insurance premiums | 2 | 2.9 | 2 |
| Rent and lease payments | 2 | 2.9 | 2.2 |
| Net cash farm income | 2 | 2.9 | 1.8 |
| Net farm income | 2 | 2.9 | 1.8 |
| Fixed expenses | 2 | 3 | 2.1 |
| Depreciation | 3 | 3.1 | 1.9 |
| Fertilizer and chemicals | 3 | 3.4 | 2 |
| Repairs and maintenance | 3 | 3.5 | 2 |
| Interest | 3 | 3.7 | 2.1 |
| Seed and plants | 3 | 3.9 | 2.4 |
| Fuels and oils | 4 | 3.9 | 2 |
| Value of inventory change | 4 | 4.5 | 1.7 |

Note: Sample of non-missing data includes 277 observations at the state level.

Figure 15. Informational Content and Probability of Each Variable's Discretized State




Note: Cells with an "X" imply no observed data in that variable's interval.

## 10.2 Attributes and Measures

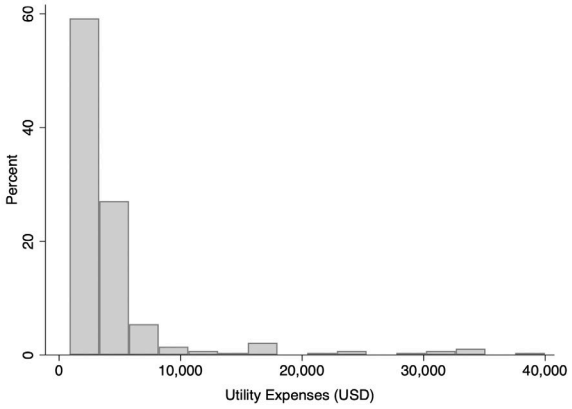
Table 14 presents the normalized entropy of all discretized variables in the dataset. Many of the average expense variables, such as labor, livestock and utility expenses, have the lowest normalized entropy values in the dataset. Low normalized entropy values for expense variables point toward the possible existence of 'outliers' (values far away from the means – at the tails of the distribution) among the sample of states within the panel. These 'outliers' can be seen in Figure 16, Panel A.

Table 14. Normalized Entropy of All Variables without Missing Information

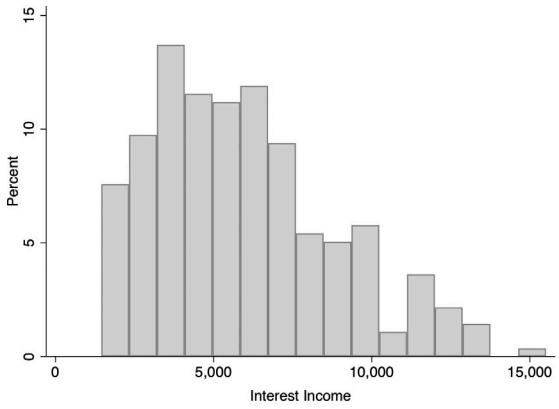
| Variable | Normalized Entropy |
|---|---|
| Seed and plants | 0.90 |
| Interest | 0.88 |
| Repairs and maintenance | 0.86 |
| Fuels and oils | 0.85 |
| Fertilizer and chemicals | 0.84 |
| Depreciation | 0.82 |
| Fixed expenses | 0.82 |
| Insurance premiums | 0.79 |
| Rent and lease payments | 0.78 |
| Livestock income | 0.77 |
| Value of inventory change | 0.77 |
| Net farm income | 0.76 |
| Gross cash farm income | 0.74 |
| Net cash farm income | 0.74 |
| Total cash expenses | 0.74 |
| Crop sales | 0.70 |
| Feed | 0.70 |
| Variable expenses | 0.70 |
| Real estate and property taxes | 0.68 |
| Other farm-related income | 0.65 |
| Government payments | 0.63 |
| Nonmoney income | 0.62 |
| Number of Farms | 0.58 |
| Other variable expenses | 0.54 |
| Labor, non-cash benefits | 0.49 |
| Machine-hire and custom work | 0.47 |
| Livestock purchases | 0.40 |
| Labor expenses | 0.38 |
| Other livestock-related expenses | 0.38 |
| Adjusted breeding livestock income | 0.36 |
| Utility expenses | 0.32 |

Figure 16. Comparing Variables with the Highest and Lowest Normalized Entropy Values

*Panel A.*
Distribution of a Variable with the Lowest Normalized Entropy, Utility Expenses

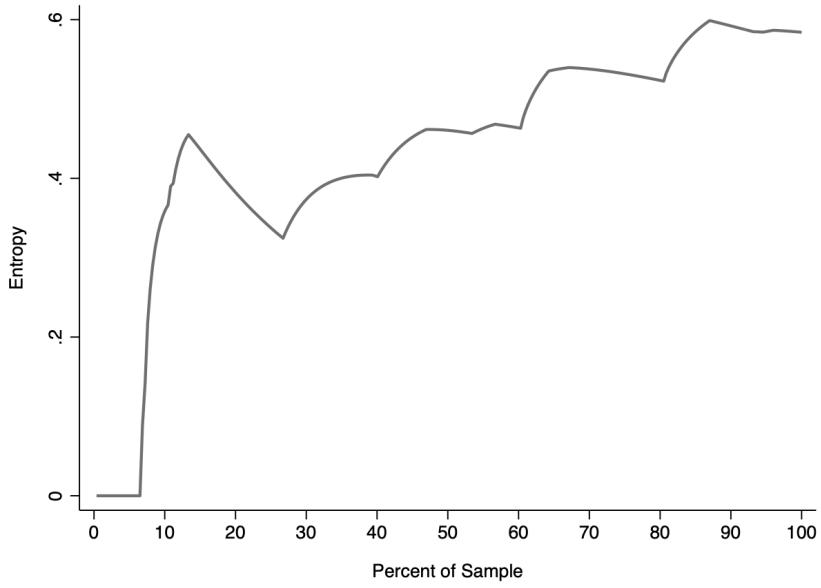



*Panel B.*
Distribution of a Variable with the Highest Normalized Entropy, Interest Income




The cumulative entropy of every single discretized random variable was calculated to study its convergence behavior. All variables exhibited distinguishable convergence except for the variable containing the number of total farms in each state (as shown in Figure 17). Slow convergence or the lack of convergence may signal poor data quality or some sharp fluctuations due to some deterministic or stochastic effects. Further examination of this variable reveals that values at the tails of the distribution were observed for a single State, Texas, which also exhibited a one-time 8 percent increase in the number of farms in 2007. No other abnormalities in this variable were discovered.

Figure 17. Cumulative Entropy of Number of Farms Per State, 2003 – 2021



Next, we apply Benford's Law. As shown in Figure 18, the data as a whole are well behaved and follow Benford's Law. However, when examining variable specific digit distributions, few deviations from the natural law exist. As shown in Figure 19, Panel A, and consistent with the cumulative entropy results in Figure 17, the digit distribution corresponding to the number of farms in each state deviates from Benford's Law. This is expected as Benford law holds for variables that there values vary by an order of more than a single magnitude.

Figure 18. Benford's Law Distribution of All Variables

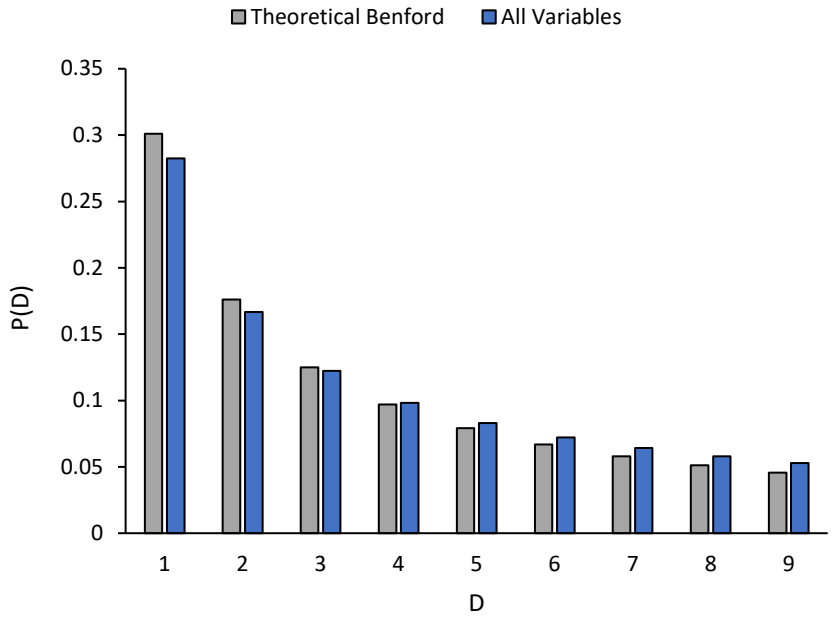
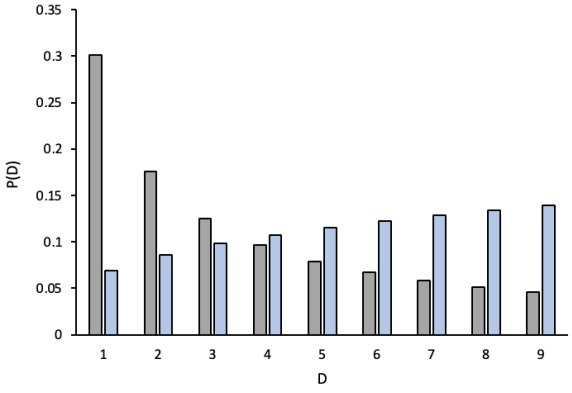



Figure 19. Variables that Deviate from Benford's Law

Panel A. Number of Farms




Panel B. Fuel Expenses

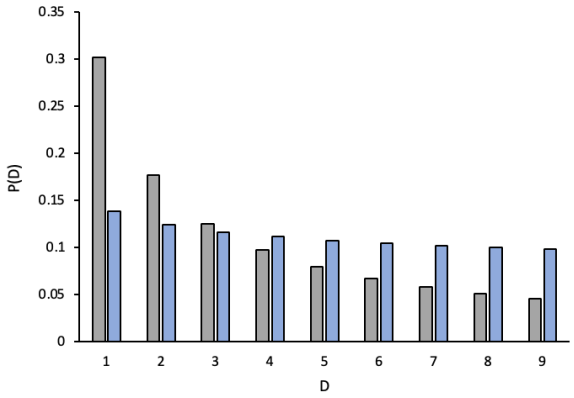


Figure 20. Mutual Information of Pairs of Variables

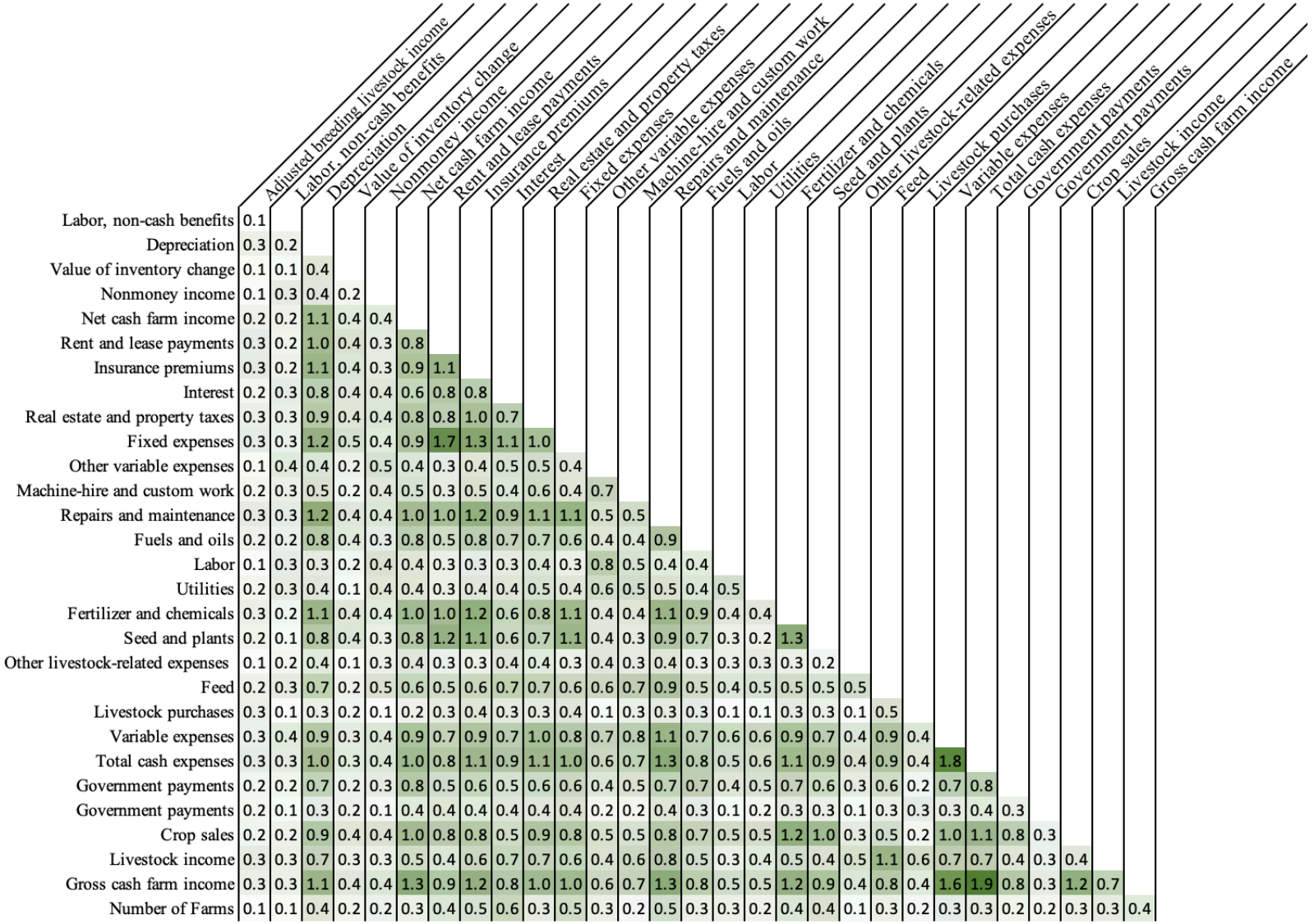


Figure 21. Joint Discretized Distributions of Discretized Gross Cash Income and Total Cash Expenses

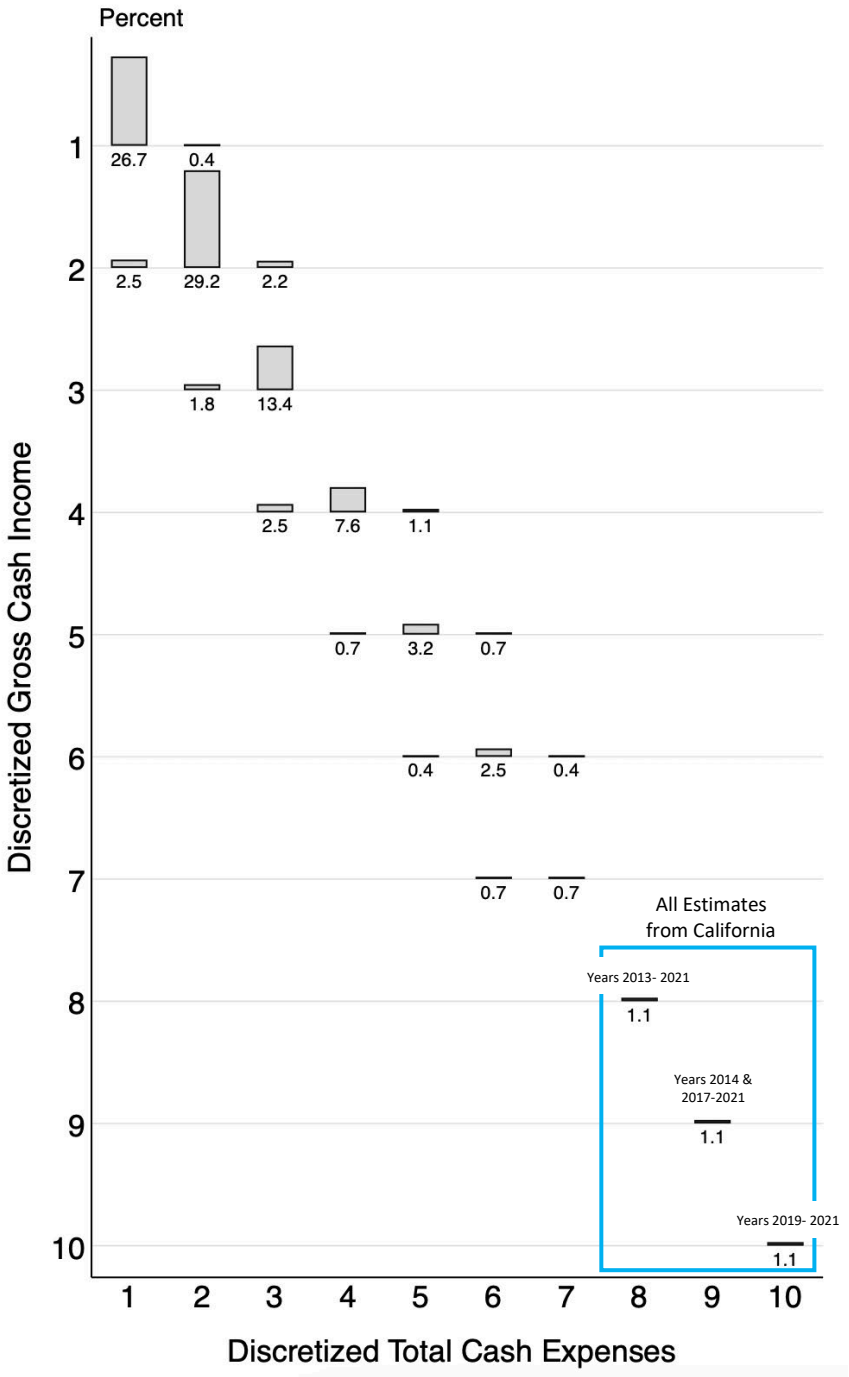



Figure 20 presents the mutual information between all pairs of income and expense variables. The largest values of mutual information come from the following pairs of variables: variable expenses and gross cash farm income, variable expenses and total cash expenses, and

total cash expenses and gross cash farm income. Figure 21 illustrates the joint distribution for the last pair of variables; those with the greatest mutual information. Consistent with prior results, some states across the panel have on average farms that incur relatively extreme expenses, but also earn relatively large incomes when compared to the average farm in other states. From 2013 to 2021, the top three intervals of both total cash expenses and gross cash farm income (in the box outlined in blue in Figure 21) were fully comprised of average estimates from California.

Table 15 summarizes the Shannon limit of the state-level panel. The information in this dataset can be compressed by approximately 35.2 percent. Of all other empirical examples presented, this dataset has the most redundancies. The compression ratio is 1.54.

Table 15. The Shannon Limit of ARMS Income and Expenditure Data, 2003 - 2021

| Compressed | Shannon Limit | N | K | $\pi$ |
|---|---|---|---|---|
| 35.2% | 370879.365 | 572,133 | 94,917 | 0.166 |

### 10.3 Questions and Condition Numbers: A Disaggregated Panel

So far we have used an aggregated panel of the ARMS dataset to investigate some of its basic quantitative properties. Here, we use a more detailed and disaggregated subset of the ARMS data to highlight some of the basic families (types) of questions that these data have the potential to study and answer. We use an unbalanced panel (2003 to 2021) by type of farm. Though many farms produce multiple products, they are classified in the dataset according to their primary product. Thirteen types of product classifications are considered over 15 States and 19 years. Each observation in this dataset is a farm of a certain type, within a specific State, for a specific year. Overall, there are 2,691 observations in this panel.

The degree of balance within each State and type of farm varies. Table 16 shows that variation. Each cell in the table presents the percent (State-farm-type) of the panel that is observed. Cash crops, cattle, and other field crops and livestock are relatively and consistently observed within each state over the sample period. However, wheat, tobacco, cotton and peanuts are inconsistently observed over the sample period. As expected with an unbalanced panel, there are observations with missing values (less than 0.5 percent of data). We note, however, that missing data do not imply zeros, as zeros are observed.

Table 17 characterizes key variables of the ARMS data (by state and type of farm) described above. Farms are categorized by type where the categories include cash grains, wheat, corn, soybeans, specialty crops (fruits and vegetables), cattle, hogs, poultry, dairy, other field crops, other livestock, and tobacco, cotton and peanuts. By design, farm income and expenses are broken down by type, are additive (Basic Block 3, Criteria E1), and can be used to calculate proportions of totals. Multicollinearity is thus high among variables, especially those related to

farm expenses and should not be jointly used for inference. This is consistent with the redundancies in the data as shown by its Shannon limit.

In addition to income, expense, production, relative efficiency, productivity, and potential environmental impact (pollution) information, the ARMS data capture how federal and local policies impact farms' behavior, such as operating costs via taxes or subsidies. Besides type of product, little characteristic information on farms is provided. No distributional information (say by type and/or State) is given. The observations are representative of aggregated survey data (Basic Block 3, Criteria N). Users are able observe the operating costs and conditions of an average farm, in a specific State, at a certain point in time, but are unable to observe individual farms (Basic Block 3, Criteria D and C2) or control for variation in farm specific characteristics.

As shown in Table 17, the ARMS data have the potential to answer a rich set of questions on the average state and annual progression of the farming sector from 2003 onward (Basic Black 3, Criteria O and Q). Since the ARMS data only provide information on averages, it is not useful for detailed distributional analyses, including distributional dynamics, by State and type.

Table 16. Percent of Panel Observed by State and Farm Product Type

| Product | State | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Arkansas | California | Florida | Georgia | Illinois | Indiana | Iowa | Kansas | Minnesota | Missouri | Nebraska | North Carolina | Texas | Washington | Wisconsin | **Percent of Product** |
| *Cash Crops* | 100 | 100 | 0 | 58 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 58 | 100 | **88** |
| *Cattle* | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | **100** |
| *Corn* | 68 | 16 | 0 | 74 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 21 | 100 | **79** |
| *Dairy* | 53 | 100 | 84 | 100 | 89 | 100 | 100 | 84 | 100 | 100 | 42 | 100 | 100 | 100 | 100 | **90** |
| *Hogs* | 58 | 11 | 16 | 37 | 100 | 100 | 100 | 68 | 100 | 95 | 74 | 100 | 21 | 11 | 42 | **62** |
| *Other Field Crops* | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | **100** |
| *Other Livestock* | 95 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 95 | 100 | 100 | 100 | 100 | **99** |
| *Poultry* | 100 | 95 | 100 | 100 | 16 | 100 | 95 | 11 | 95 | 100 | 21 | 100 | 100 | 100 | 95 | **82** |
| *Soybeans* | 100 | 0 | 5 | 58 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 5 | 0 | 100 | **71** |
| *Specialty Crops* | 95 | 100 | 100 | 100 | 100 | 100 | 89 | 79 | 100 | 95 | 74 | 100 | 100 | 100 | 100 | **95** |
| *Tobacco Cotton and Peanuts* | 95 | 63 | 100 | 100 | 0 | 11 | 0 | 0 | 0 | 74 | 0 | 100 | 100 | 0 | 0 | **43** |
| *Wheat* | 16 | 11 | 0 | 5 | 5 | 5 | 0 | 100 | 79 | 0 | 89 | 11 | 100 | 100 | 5 | **35** |
| **Percent of State** | **82** | **66** | **59** | **78** | **76** | **85** | **82** | **79** | **89** | **89** | **75** | **93** | **86** | **66** | **79** | |

Table 17. ARMS Data: Potential Questions and Condition Numbers

| Type of Variables | Variables | Condition Number of Variables | Possible Question | Possible Dependent Variable |
|---|---|---|---|---|
| Determinants of Costs, production, productivity, pollution and efficiency | Expenses relating to livestock, breeding livestock, utilizes, seeds and plants, repairs and maintenance, fuels and oils, fertilizer and chemicals, labor (and non-cash benefits), and feed; Fixed expenses; Insurance premiums; Rent and lease payments; Total cash expenses; Real estimate and property taxes; Other variable expenses; Machine-hire and custom work Interest (paid); | N/A[1] | Are certain types of farms more productive than others?<br>Stochastic frontiers of different industries;<br>Productivity – Pollution and Sustainability of farms and firms;<br>Has farm productivity changed over the last decade? | Net farm income |
| | | | Are there barriers to entry? If so, what?<br>Do changes in property taxes (i.e., property tax policy) affect the number of farms? | Number of farms |
| Short Run Benefits/Success | Livestock income; Net farm income; Income from crop sales; Other farm-related income; Nonmoney income | 113.45 | Has farm labor become more productive/profitable over the last decade?<br>Is pollution per unit of production increases? If so, what are the major causes for this? | Labor expenses |
| Federal and Local Policies | Government payments; Real estimate and property taxes | 2.48 | What types of farms receive the most government aid? Does it impact the structure of the industry? | Type of farm |
| | | | Do government payments keep farms in business? | Number of farms |
| | | | Do farms experiencing higher depreciation (in capital and/or the value of their output) receive government payments? | Depreciations and Change in the Value of Inventory |
| Attributes | Number of farms; Type of farm (encoded) | 4.78 | Which farming businesses are most lucrative (in terms of income and profit)? Is that connected to productivity or size? Has this changed over the last decade? | Net farm income |
| Relationship between the market/economy and farms/firms | Capital (Machine) Depreciation; Value of inventory change | 1.77 | Demand and supply;<br>Is the number of operating farms responsive to changes in the market value of inventory? | Number of farms |

[1] *No condition number was calculated. The expense variables categorized total incurred expenses by type and are by design collinear.*

## 11 Value and Quality: Graphical Comparison of the Datasets

Having summarized the different data sets in much detail, we now compare them. Figure 22 shows the aggregated values (by Quality and Value Blocks) for each one of the datasets. Figure 23 provides a detailed view of the attributes' distributions within each block. Table 18 shows the exact value of the attributes, by sub-categories defined in the table. The exact scores of each attribute and their definitions are shown in Appendix Table 3 and 4.

Figure 22. Datasets Comparison by Blocks: Quality and Value

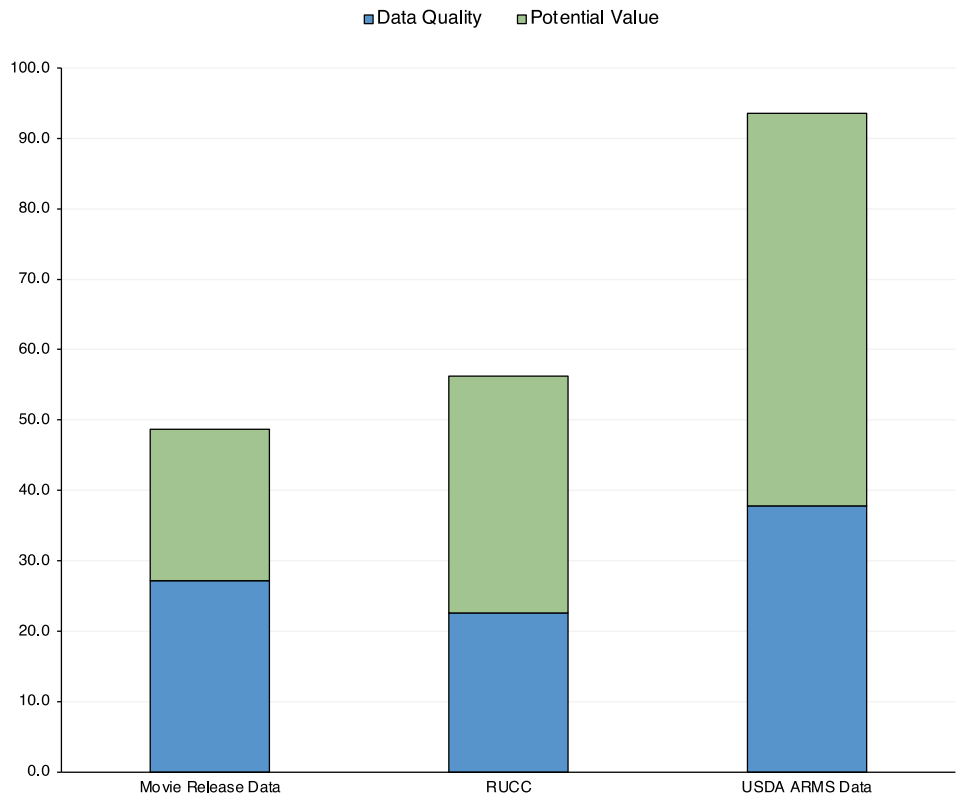
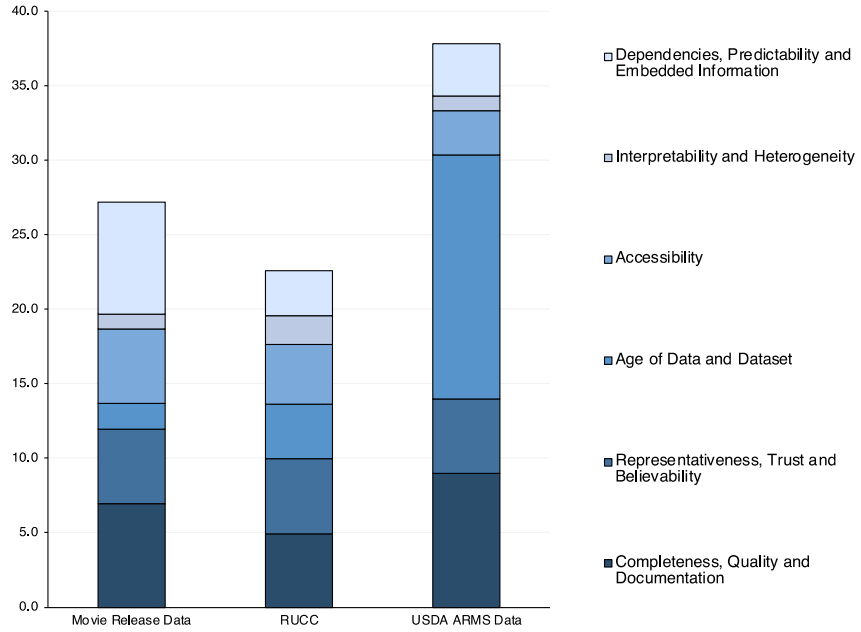


Figure 23. Datasets Comparison by Blocks and Attributes

Panel A. Attributes of Data Quality: Basic Block 2




Panel B. Attributes of Potential Value: Basic Block 3

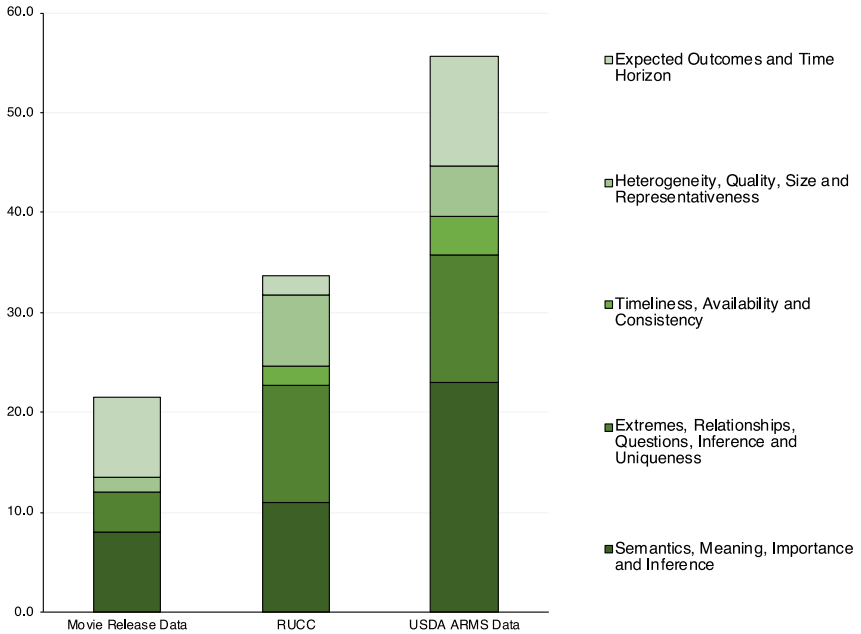


Table 18. Case Study Valuations Disaggregated by Attributes

Panel A. Data Quality: Basic Block 2

| | Criteria | Datasets | | |
|---|---|---|---|---|
| | | Movie Release Data | RUCC | USDA ARMS Data |
| **A, C, D** | Completeness, Quality and Documentation | 6.97 | 4.95 | 9.00 |
| **E, F, G** | Representativeness, Trust and Believability | 5.00 | 5.00 | 5.00 |
| **I** | Age of Data and Dataset | 1.71 | 3.70 | 16.33 |
| **J** | Accessibility | 5.00 | 4.00 | 3.00 |
| **H, K** | Interpretability and Heterogeneity | 1.00 | 1.92 | 1.00 |
| **L, M** | Dependencies, Predictability and Embedded Information | 7.50 | 3.00 | 3.50 |
| | **Score** | 27.2 | 22.6 | 37.8 |

Panel B. Potential Value: Basic Block 3

| | Criteria | Datasets | | |
|---|---|---|---|---|
| | | Movie Release Data | RUCC | USDA ARMS Data |
| **B, C1**, **C2** | Semantics, Meaning, Importance and Inference | 8.00 | 11.00 | 23.00 |
| **D, E2**, **F**, **J**, **K**, **R** | Extremes, Relationships, Questions, Inference and Uniqueness | 4.02 | 11.69 | 12.68 |
| **G, H, I** | Timeliness, Availability and Consistency | 0.00 | 2.00 | 4.00 |
| **L, M, O** | Heterogeneity, Quality, Size and Representativeness | 1.50 | 7.00 | 5.00 |
| **P, Q** | Expected Outcomes and Time Horizon | 8.00 | 2.00 | 11.00 |
| | **Score** | 21.5 | 33.7 | 55.7 |

## 12 Value of Access to Data

Like the value of data, the value of access to the data is also relative. However, the main difference is that the value to access the data is positive only if the potential value of the dataset itself is positive. (Given the arguments discussed in Section 2, we cannot think of a case where the data itself will have a negative value.) In a way, the potential value of the data can only be realized if we have access to that data. If that access is free (publicly available) for all, and straightforward, then the full potential of the data may be realized. This can be summarized as follows: The probability of extracting the full potential value embedded in the data is conditional on the accessibility to that data. The more accessible and open (for public use) are the data, the higher the probability of reaching the full potential of that data; the probability is a positive function of the accessibility.

## 13 Notes on Monetary Value

The policy maker may not be satisfied by just looking into the potential value of the data. Rather, they may want the monetary value of the data, capturing the potential benefit from the data. Since the cost is relatively simple to calculate or approximate, it will allow the policy maker to do a cost-benefit analysis. But converting the relative value to a monetary value is not trivial. It can be done in some cases.

Given the approach proposed in this paper, we can provide a certain (relative) value to a dataset. One way of converting that value to monetary units is as follows. Assume, for now, that we were able to answer all of the potential questions that a dataset can answer (which is a part of Block 3 attributes). That is, the complete potential of the data materialized. With that new knowledge, we can calculate the potential benefit to society. We can think about it as a variation of the compensation principle or Pareto improvement. Did the new knowledge that has transformed from the data (and access to that data) increase society's welfare? For example, if the data allowed us to construct a new policy that reduces inequality, and thereby increases society's welfare, then we can translate this, approximately, into monetary units. To summarize, assuming the potential of the data materialized, can it be transformed into a Pareto improvement for society? If the interest lies in a certain group within society, the question then is if the data allowed us to improve the welfare of that group, even if it is on account of others. But this is, of course, problem and policy specific.

## 14 Concluding Thoughts and Open Questions

We proposed a way to approximate the relative potential value of data and access to that data. Though value is a relative concept, we construct our proposed measure as a sum of different attributes. Each one of these attributes has a finite and bounded value. If the number of possible values is finite, say $Q$, then, the value of each attribute is bounded by $2^{|Q|}$. Aggregation in that case is trivial. The (relative) weights of these quality and value attributes provide insights into the different constituents of the overall potential.

However, as emphasized throughout this work, not all the attributes can be perfectly quantified. In fact, some of the most important attributes are defined on meaning and semantics. Assigning values to such attributes should be handled with much care. We proposed some ways of doing that, but as we discussed previously, these assignments are relative. However, given that (*i*) there are many attributes, and the (*ii*) assignments is done by the same expert, or a set of experts, one can argue that, even though the calculated values of the different datasets are relative, it is still possible to rank the datasets in terms of their approximate values. Furthermore, if several experts decide on these values independently, then an improved (relative) potential value measure is the median, or an interval range, based on the experts' assignments.

The three empirical examples studied here demonstrate the potential use of our proposed measures. But this is just the beginning. In future research, we plan to refine and update our measure, with special effort on the non-quantifiable, or fuzzier, attributes.

We conclude this paper with a short list of questions that we believe are still open.

- We did not discuss AI in this paper. The question of how to evaluate the additional contribution (if such exists) of AI to the potential value of data, is still an open question. But regardless of the answer to that question, one must keep in mind our basic assumption here: the value of the data must be independent of the inference. Will AI impact the value while satisfying our requirement?
- Is it possible to improve the way attributes related to meaning and semantics are defined, evaluated and (relatively) quantified?
- What other attributes contribute to the value?
- Should the attributes be mutually exclusive and independent of one another?
- What is the best way to transform the value to a monetary amount (From data to knowledge to monetary value)?
- We assigned some scales for each value, but is there a better way to assign values to attributes?
- Is it possible to develop a set of axioms for evaluating the potential value? (Certain axioms may yield similar attributes to what is proposed here; other axioms may yield different, or partially different, attributes.)
- What is the impact of data aggregation on the value? Is it problem specific, or is there a general 'law' to follow?

These are just some of the open questions. Some of these questions are new, others are as old as science itself. The search for answers for the older, and tougher, questions can be found in the philosophy and philosophy of science literature, as well as the more recent literature across many disciplines, including economics, environmental and climate science, medicine, computer science, and more.

**References**


Arrow, K. J. (1963). *Social choice and individual values.* (2nd ed.). Yale University Press.

Arrow, K. J., & Fisher, A. C. (1974). Environmental Preservation, Uncertainty, and Irreversibility. *The Quarterly Journal of Economics*, *88*(2), 312–319.

Belsley, D. A. (1991). *Conditioning diagnostics : collinearity and weak data in regression*. Wiley-Interscience.

Bergemann, D., & Bonatti, A. (2019). Markets for Information: An Introduction. *Annual Review of Economics*, *11*(1), 85–107. https://doi.org/10.1146/annurev-economics-080315-015439

Bronselaer, A., De Mol, R., & De Tre, G. (2018). A Measure-Theoretic Foundation for Data Quality. *IEEE Transactions on Fuzzy Systems*, *26*(2), 627–639. https://doi.org/10.1109/TFUZZ.2017.2686807

Brookshire, D. S. (1982). Valuing Public Goods: A Comparison of Survey and Hedonic Approaches. *The American Economic Review*, *77*, 165–177.

Carson, R. T. (2012). Contingent Valuation: A Practical Alternative when Prices Aren't Available. *Journal of Economic Perspectives*, *26*(4), 27–42. https://doi.org/10.1257/jep.26.4.27

Carson, R. T., & Hanemann, W. (2005). Contingent Valuation. In K.-G. Maler & J. Vincent (Eds.), *Handbook of Environmental Economics* (Vol. 2, pp. 821–936). North Holland.

Ciriacy-Wantrup, S. V. (1947). Capital Returns from Soil-Conservation Practices. *Journal of Farm Economics*, *29*(4), 1181. https://doi.org/10.2307/1232747

Council on Food, A. and R. E. (C-F. (2013). *From Farm Income to Food Consumption: Valuing USDA Data Products*.

Davidson, M. D. (2013). On the relation between ecosystem services, intrinsic value, existence value and economic valuation. *Ecological Economics*, *95*, 171–177. https://doi.org/10.1016/j.ecolecon.2013.09.002

Diamond, P. A., & Hausman, J. A. (1994). Contingent Valuation: Is Some Number Better than No Number? *Journal of Economic Perspectives*, *8*(4), 45–64. https://doi.org/10.1257/jep.8.4.45

Dixit, A., & Pindyck. Robert. (1994). *Investment Under Uncertainty*. Princeton University Pres.

Dretske, F. (2008). Epistemology and Information. In *Philosophy of Information* (pp. 29–47). Elsevier. https://doi.org/10.1016/B978-0-444-51726-5.50007-8

Dunn, J. M., & Golan, A. (2020). Information and Its Value. In M. Chen, M. J. Dunn, A. Golan, & A. Ullah (Eds.), *Advances in Info-Metrics* (pp. 3–31). Oxford University PressNew York. https://doi.org/10.1093/oso/9780190636685.003.0001

Edwards, R. D., & Tuljapurkar, S. (2005). Inequality in Life Spans and a New Perspective on Mortality Convergence Across Industrialized Countries. *Population and Development Review*, *31*(4), 645–674. https://doi.org/10.1111/j.1728-4457.2005.00092.x

Farboodi, M., & Veldkamp, L. (2022). *A Model of the Data Economy*.

Fishburn, P. C. (1970). *Utility Theory for Decision Making*. Wiley.

Freeman, R. B., Yang, B., & Zhang, B. (2023). Data deepening and nonbalanced economic growth. *Journal of Macroeconomics*, *75*, 103503. https://doi.org/10.1016/j.jmacro.2023.103503

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets. *Communications of the Association for Computing Machinery (ACM)*, *64*(12), 86–92.

A. Golan. *A. (1988). Discrete Stochastic Model of Economic Production and A Model of Fluctuations in Production - Theory and Empirical Evidence*. Thesis, UC Berkeley.
B. Golan, A. (2018). *Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information*. Oxford University Press.
Gould, J. P. (1974). Risk, stochastic preference, and the value of information. *Journal of Economic Theory*, *8*(1), 64–84. https://doi.org/10.1016/0022-0531(74)90006-4
Hanemann, M. W. (1989). Information and the Concept of Option Value. *Journal of Environmental Economics and Management*, *16*, 23–37.
Hanemann, W. M. (1994). Valuing the Environment Through Contingent Valuation. *Journal of Economic Perspectives*, *8*(4), 19–43. https://doi.org/10.1257/jep.8.4.19
Hausman, J. (2012). Contingent Valuation: From Dubious to Hopeless. *Journal of Economic Perspectives*, *26*(4), 43–56. https://doi.org/10.1257/jep.26.4.43
Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards*. https://doi.org/https://doi.org/10.48550/arXiv.1805.03677
Hughes-Cromwick, E., & Coronado, J. (2019). The Value of US Government Data to US Business Decisions. *Journal of Economic Perspectives*, *33*(1), 131–146. https://doi.org/10.1257/jep.33.1.131
Jaynes, E. T. (1979). Concentration of distributions at entropy maxima. *ET Jaynes: Papers on Probability, Statistics and Statistical Physics*, 315. http://bayes.wustl.edu/etj/articles/entropy.concentration.pdf
Johnson, M. G. (1974). Context, Flexibility and Meaning: Some Cognitive Aspects of Communication. *Journal of Advertising*, *3*(1), 16–20. https://doi.org/10.1080/00913367.1974.10672507
Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, *110*(9), 2819–2858. https://doi.org/10.1257/aer.20191330
Judge, G., & Schechter, L. (2009). Detecting Problems in Survey Data Using Benford's Law. *Journal of Human Resources*, *44*(1), 1–24. https://doi.org/10.3368/jhr.44.1.1
Kannan, K., Ananthanarayanan, R., & Mehta, S. (2018). *What is my data worth? From data properties to data value*. https://doi.org/https://doi.org/10.48550/arXiv.1811.04665
Kip Viscusi, W. (2014). *The Value of Individual and Societal Risks to Life and Health* (pp. 385–452). https://doi.org/10.1016/B978-0-444-53685-3.00007-6
Krutilla, J. V. (1967). Conservation Reconsidered. *The American Economic Review*, *57*(4), 777–786.
Lanoie, P., Pedro, C., & Latour, R. (1995). The value of a statistical life: A comparison of two approaches. *Journal of Risk and Uncertainty*, *10*(3), 235–257. https://doi.org/10.1007/BF01207553
Macauley, M. K. (2006). The value of information: Measuring the contribution of space-derived earth science data to resource management. *Space Policy*, *22*(4), 274–282. https://doi.org/10.1016/j.spacepol.2006.08.003
Magat, W. A., Kip Viscusi, W., & Huber, J. (1988). Paired comparison and contingent valuation approaches to morbidity risk valuation. *Journal of Environmental Economics and Management*, *15*(4), 395–411. https://doi.org/10.1016/0095-0696(88)90034-4
Mensink, P., & Requate, T. (2005). The Dixit–Pindyck and the Arrow–Fisher–Hanemann–Henry option values are not equivalent: a note on Fisher (2000). *Resource and Energy Economics*, *27*(1), 83–88. https://doi.org/10.1016/j.reseneeco.2004.04.001

Nath Datta, B. (2004). Some Fundamental Tools and Concepts From Numerical Linear Algebra. In *Numerical Methods for Linear Control Systems: Design and Analysis* (pp. 33–78). Academic Press.

Nouraldeen, A. S. (2015). Meaning and Context-Three Different Perspectives. *British Journal of English Linguistics*, *3*(2), 13–17.

Repo, A. J. (1989). The Value of Information: Approaches in Economics, Accounting, and Management Science. *Journal of the American Society for Information Science*, *40*(2), 66–85.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Stroming, S., Robertson, M., Mabee, B., Kuwayama, Y., & Schaeffer, B. (2020). Quantifying the Human Health Benefits of Using Satellite Information to Detect Cyanobacterial Harmful Algal Blooms and Manage Recreational Advisories in U.S. Lakes. *GeoHealth*, *4*(9). https://doi.org/10.1029/2020GH000254

Talaat, K., Cowen, B., & Anderoglu, O. (2020). Method of information entropy for convergence assessment of molecular dynamics simulations. *Journal of Applied Physics*, *128*(13), 135102. https://doi.org/10.1063/5.0019078

Traeger, C. P. (2014). On option values in environmental and resource economics. *Resource and Energy Economics*, *37*, 242–252. https://doi.org/10.1016/j.reseneeco.2014.03.001

Tufiș, M., Borrato, L., Kassa, Y., & Paraschiv, M. (2020). *Safe-Deed Private and Public Data Value*. Horizon 2020 European Union Funding for Research and Innovation.

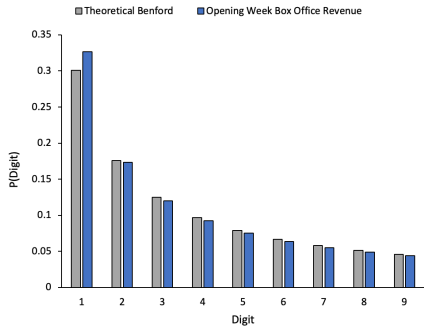
Veldkamp, L., & Chung, C. (2023). Data and the Aggregate Economy. *Forthcoming in the Journal of Economic Literature*.

Viscusi, W. K., & Aldy, J. E. (2003). The Value of a Statistical Life: A Critical Review of Market Estimates Throughout the World. *Journal of Risk and Uncertainty*, *27*(1), 5–76. https://doi.org/10.1023/A:1025598106257

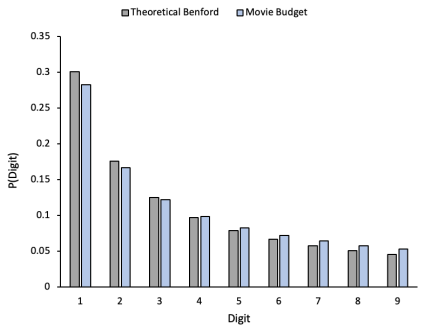
## Appendix A: Additional Figures From the Empirical Studies

Appendix Figure 1. Benford's Law Distributions for All Continuous Variables from Craig, et. al. (2015)

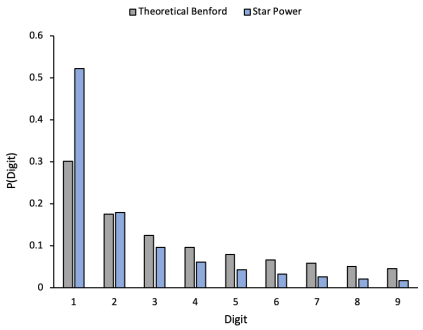
Opening Week Box Office Revenue ($USD)



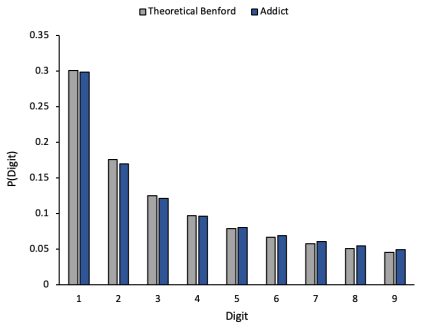

Movie Budget ($USD in Millions)



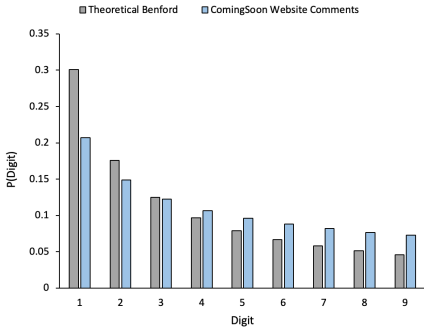

Star Power (Index)



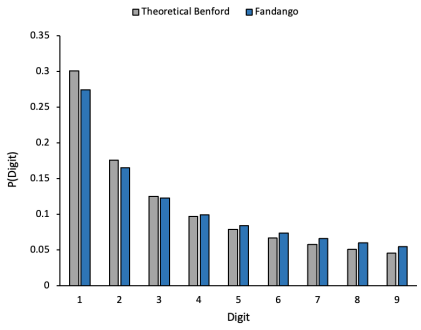

Addict (Trailer Views)



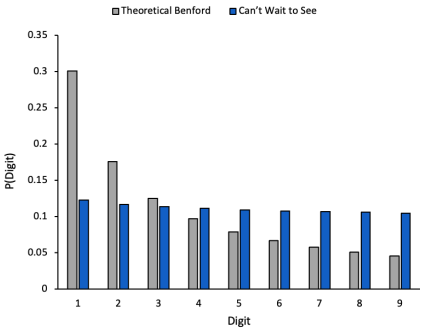

ComingSoon Website Comments




Fandango Votes

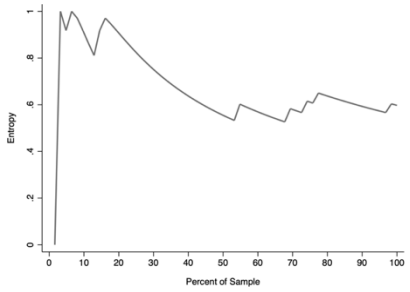



Can't Wait to See (Percent)

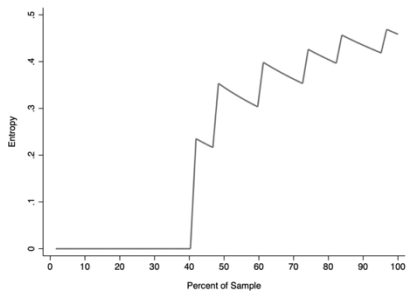


Appendix Figure 2. Cumulative Entropies for Discrete Variables from Craig, et. al. (2015)

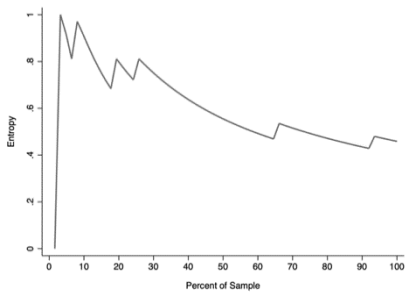
Sequel



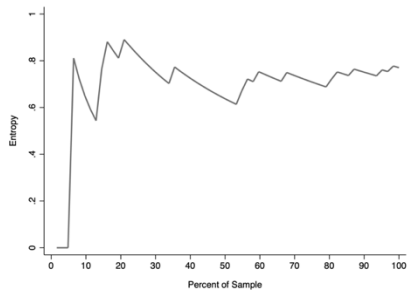

Horror



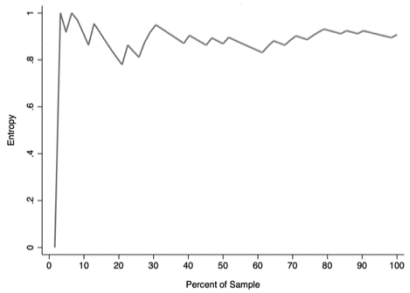

Animated



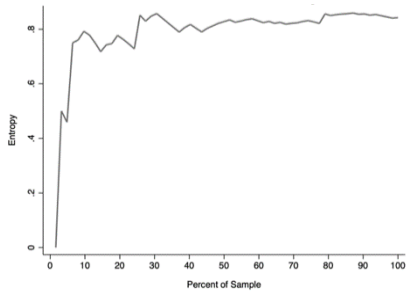

Action




Comedy

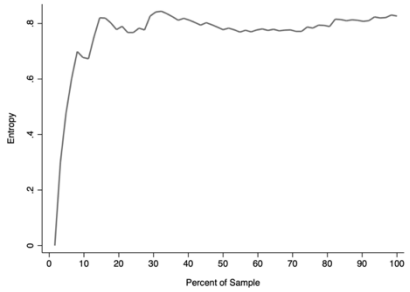



Motion Picture Association Rating

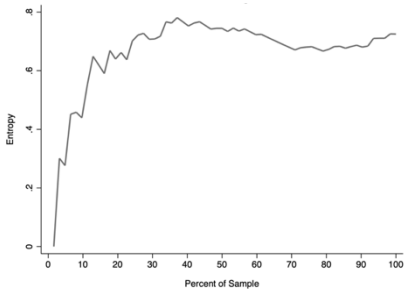


Appendix Figure 3. Cumulative Entropies for Continuous Variables from Craig, et. al. (2015)

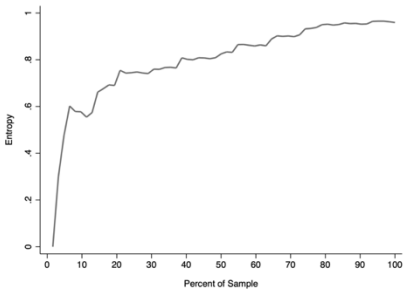
Opening Week Box Office Revenue ($USD)



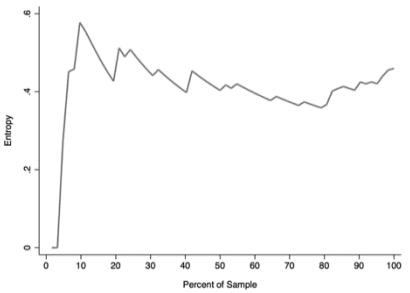

Movie Budget ($USD in Millions)



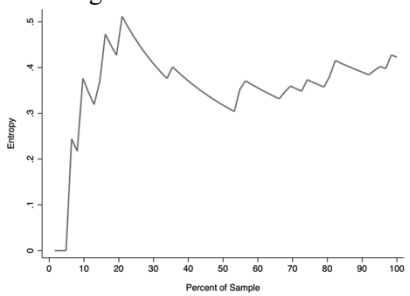

Star Power (Index)



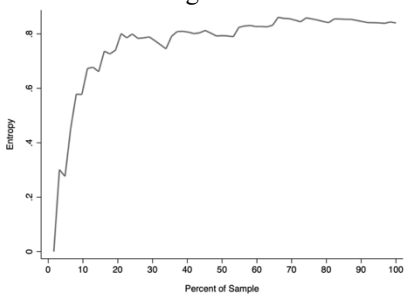

Addict (Trailer Views)



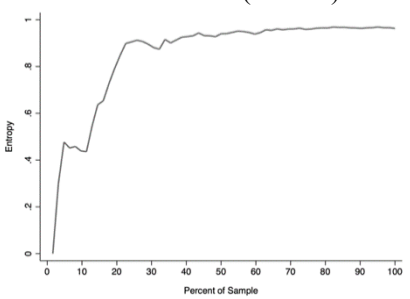

ComingSoon Website Comments




Fandango Votes




Can’t Wait to See (Percent)



## Appendix B: Additional Tables From the Empirical Studies

Appendix Table 1. Sensitivity of Entropy Calculations to Discretization of Continuous Variables from Craig, et. al. (2015)

| | Number of Intervals Used to Discretize | | | | | |
|---|---|---|---|---|---|---|
| | 8 | | 10 | | 12 | |
| Variable | Entropy | Normalized Entropy | Entropy | Normalized Entropy | Entropy | Normalized Entropy |
| Addict | 1.357 | 0.452 | 1.526 | 0.460 | 1.776 | 0.495 |
| Opening Week Box Office Revenue ($USD) | 2.450 | 0.817 | 2.747 | 0.827 | 2.989 | 0.834 |
| Movie Budget ($USD) | 2.146 | 0.715 | 2.409 | 0.725 | 2.629 | 0.733 |
| ComingSoon Website Comments | 1.070 | 0.357 | 1.404 | 0.423 | 1.571 | 0.438 |
| Can't Wait to See | 2.836 | 0.945 | 3.199 | 0.963 | 3.478 | 0.970 |
| Fandango | 2.427 | 0.809 | 2.791 | 0.840 | 2.969 | 0.828 |
| Star Power | 2.897 | 0.966 | 3.187 | 0.959 | 3.423 | 0.955 |

Appendix Table 2. Definition of Variables from the Agricultural Resource Management Survey

| Variable | Definition |
|---|---|
| Gross cash farm income | For farms participating in government programs, gross cash farm income is the total amount of cash received by the farm operation from the sale of agricultural products, services rendered, or government payments received during a given a calendar year. |
| Livestock income | The value of all livestock and poultry sold from the farm operation net of any marketing charges. All sales of livestock through marketing contracts or payments received from prior years' contracts are also included. For integrated operations, which do not sell the livestock but pass them on to another phase of the operation, an estimate of the value of the livestock moving through the operation is captured as sales. |
| Crop sales | The amount received from cash sales and marketing or forward contracts for all crops plus the difference between Commodity Credit Corporation crop placements and redemptions. Payments received for crops produced in previous years or delivered under prior year's marketing contracts are included. |
| Government payments | Gross value of direct government payments received by farm operations during the calendar year. Programs for which payments are received include: direct payments, counter-cyclical payments, loan deficiency payments (LDPs), marketing loan gains, peanut quota buyout program, milk income loss contract payments, disaster payments, conservation reserve program (CRP), wetlands reserve program (WRP), environmental quality incentive program (EQIP), and all other Federal and State programs. |
| Other farm-related income | Includes income from machine-hire, custom work, livestock grazing, land rental, contract production fees, outdoor recreation, timber sales, hedging profit or losses, insurance indemnities, cooperative patronage dividends and refunds, leasing of livestock and machinery or equipment, and any other farm-related source. |
| Total cash expenses | Cash expenses represent the total amount of funds paid out by the farm operation during a calendar year. Expenses paid by landlords or parties for which a contractual agreement existed during the year are excluded from cash expenses reported for farm operations. Marketing charges such as commissions, storage, inspection, insurance, drying, check-offs, yardage, and auction are excluded from cash expenses. |
| Variable expenses | Expenses incurred in the production process that vary with the quantity and prices of inputs used. |
| Livestock purchases | The total amount paid for livestock and poultry including commission, yardage, insurance and other associated fees. |

| | |
|---|---|
| | Breeding livestock treated as a depreciable capital is not included. |
| Feed | Amount paid by the farm operation for all feed grains, hay, forages, mixed or formula feeds, concentrates, supplements, premixes, salt, minerals, animal byproducts, and all other feed additives and ingredients. |
| Other livestock-related expenses | Amount paid by the farm operation during the calendar year for livestock leasing; custom feed processing, grinding or mixing of feed; bedding, litter, or straw; pasturing, grazing or custom feeding, veterinarian services or supplies, amount spent for sprays, dips, dusts or any other chemicals, sheep sheering, horse-shoeing, removal of dead animals, branding, artificial insemination and breeding fees, and performance testing. |
| Seed and plants | Expenses for purchases of seeds, plants, and related expenses such as seed cleaning, inoculation, rooting hormones, bagging, germinating, and delinting. |
| Fertilizer and chemicals | Amount the farm operation paid for commercial fertilizers, lime and soil conditioners, pesticides, insecticides, herbicides, fungicides, defoliants, nematicides, fumigants, growth regulators, and rodenticides used on crops, pasture, acreage idled under government programs, seeds, crop storage buildings or seed beds. |
| Utility expenses | Farm operations' share of expenses for electricity, telephone, and water including charges for irrigation water and electricity. |
| Labor expenses | Expenses for contract and hired labor engaged in "agricultural work" during the calendar year. Total cash wages include bonuses for all hired workers (including paid family members) and employees' share of social security taxes. Costs associated with fringe benefits such as insurance, pensions, workers compensation, and unemployment compensation are also included in labor expense. Wages paid to operators that are not hired managers are excluded. |
| Fuels and oils | Farm share of the operations' purchases of diesel, gasoline, liquid petroleum, natural gas, other fuels (kerosene, coal, wood), and motor oils and fluids. Includes fuels used for heating farm buildings and offices, drying or curing crops, and all machinery and equipment (including irrigation pumps). |
| Repairs and maintenance | Amount paid by the operation for repairs and maintenance of motor vehicles, machinery and equipment, irrigation and frost protection equipment, farm buildings, the operator's dwelling, and any other labor dwellings, fencing, soil conservation structures, drainage structures, and any other farm or ranch structures (corrals, feedlots, feeding floors). |
| Machine-hire and custom work | Amount spent by the operation for custom hauling and other custom work such as land tillage, planting or seeding, harvesting, and soil testing. Custom work is defined as work preformed by machines and labor hired as a unit. |
| Other variable expenses | Includes supplies, motor vehicle registration fees, transportation and storage, and all other general business expenses (for example, fees paid to accountants or attorneys, |

| | |
|---|---|
| | registration of purebred animals, travel expenses, postage, and magazine subscriptions). |
| Fixed expenses | Represents costs incurred by the farm operation during the calendar year, even when there is no production. |
| Real estate and property taxes | Expenses for taxes on farm land, buildings, capital improvements, machines, livestock, and other property. Includes all taxes paid during the calendar year even though they may have been levied in another year. |
| Interest | Amount paid by the farm operation for interest on farm business loans and mortgages, land contracts, and other farm loans secured by real estate, finance charges for operating loans, machinery and equipment loans, or any other interest on non-real estate loans. |
| Insurance premiums | Amount spent by the farm operation for Federal Crop Insurance, and the farm share of motor vehicle, liability, and blanket policies which provide more than one year's coverage. |
| Rent and lease payments | Expenses for land rental (including Public Industrial Grazing Association land), all vehicles, tractors, farm machinery, equipment, and structures leased. |
| Net cash farm income | This measure indicates the amount of net cash earnings from all business sources that a farm generates during the year. These funds can be used to repay principal on indebtedness, purchase new machinery or equipment, expand the farm business, or pay for family consumption or other obligations. |
| Nonmoney income | An estimated value of items produced and consumed on the farm, and the imputed rental value of farm dwellings. |
| Value of inventory change | The change in the market value of all crops, livestock, or purchased inputs from January 1 of the calendar to December 31. |
| Depreciation | An allocation of the portion of the original cost of a capital asset to each of the estimated years in which the asset will be used. The type of depreciation reported on the survey typically resembles the figure reported on tax returns. |
| Labor, non-cash benefits | Includes an estimate of the value of housing or lodging provided to workers, meals, fuel, vehicles, utilities, or payment in kind. An estimate of the value of products produced and consumed on the farm is provided by the survey respondent. The imputed rental value of farm dwellings is calculated for dwellings located on the farm and is based on rent-to-value ratios for different value ranges of the dwelling. |
| Adjusted breeding livestock income | The amount of unrecovered investment in breeding livestock upon sale, reflecting the difference between gross proceeds from the sale and the recognized gain after accounting for investment costs. |
| Net farm income | Net farm income indicates the profit or loss associated with current production. It represents the return (both monetary and nonmonetary) to farm operators for their labor, management and capital after all production expenses have been paid (that is, gross farm income minus production expenses). It includes net income from farm production as well as net income |

attributed to the rental value of farm dwellings, the value of commodities consumed on the farm, depreciation, and inventory changes.

All definitions are provided by the ARMS Data Analysis page, available at <https://my.data.ers.usda.gov/arms/data-analysis>.

Appendix Table 3. Detailed Basic Block 2 Evaluation of Datasets Used in Case Studies

| **Basic Block 2.** Data Quality | | Datasets | | |
|---|---|---|---|---|
| | | Movie Release Data | RUCC | USDA ARMS Data |
| A. | **Completeness of the Data**<br>Completeness (of data as a whole: This is a theoretical concept with respect to the questions the dataset is supposed to be able to answer, whenever the questions preceded data collection, or with respect to how the data intend to be used). (1 – 10, with 10 complete). | 5 | 2 | 7 |
| B. | **Size of the Dataset**<br>Size (Number of observations and number of variables; This is an informative attribute that, together with other attributes, has an impact on the quality). | | | |
| C. | **Documentation**<br>Do all proper documents about the data exist and are understandable? (0 – 2; no documents (0), some/all and partially understandable (1), perfect (2)). | 1 | 2 | 1 |
| D. | **Variable Quality**<br>All variables are defined correctly, including their units and meanings. (Each variable has a scale of 1 – 3, (not good) to perfect). Overall, the total normalized value: 3K/3K = 1. (This ensures that the total is in 0 – 1 and comparable across datasets.) | 0.97 | 0.95 | 1 |
| E. | **Representativeness**<br>Are the data representative (with/without weights) of the underlying population (Yes (1) – No (0), sampling issues: Description). | 1 | 1 | 1 |
| F. | **Trust**<br>Do we trust the data collected, the way it were collected, and the agency/individuals that publish the collected | | | |
| | Trust the data (0-1). | 1 | 1 | 1 |
| | Trust the way the data were collected (0-1). | 1 | 1 | 1 |
| | Trust the agency/individuals publish the data (0-1). | 1 | 1 | 1 |
| G | **Believability**<br>Complements representativeness and trust (Yes – No (1 – 0)). | 1 | 1 | 1 |
| H | **Interpretability**<br>Is it possible to provide coherent, logical, and consistent interpretation of each one of the variables (and for the resulting potential inference)? (Yes – No for each variable). Overall, the total normalized value: K/K = 1. (This ensures that the total is in the range 0 – 1 and comparable across datasets.) | 1 | 0.92 | 1 |
| I | **1. Age of the Dataset**<br>Age of the dataset availability (first date these data are available; using a discount factor from that date. The current year has the highest value of 1, then a discount factor for each previous year. | 0.85 | 2.88 | 15.37 |
| | **2. Age of Data in Dataset**<br>Age of the data in the dataset (the most recent date of the main variable/s of interest; using a discount factor from that date; same scale as D1).<br>The discount factor can be any positive number between 0 – 1 that is deemed reasonable or appropriate so long as it is consistent between datasets. | 0.85 | 0.82 | 0.96 |
| J | **Accessibility.**<br>How easy it is to access the data, manage and manipulate it (1-5 (easiest/best)). | 5 | 4 | 3 |
| K | **Heterogeneity**<br>Does the dataset capture (approximately) the full heterogeneity of the underlying population. Yes (1) – No (0). | 0 | 1 | 0 |
| L | **Data Dependencies** | | | |
| | Condition number to capture the level of collinearity, 1 – 3: low (3) – very high (1) | 3 | 1 | 1 |
| | Correlation low (3) – very high (1) | 3 | 1 | 1 |
| | Low values of collinearity and correlation are indicative of more information thus given higher scores. | | | |
| M. | **Data Predictability and Embedded Information**<br>We use Shannon limit to provide an approximate notion of the non-redundant information in the data. This is a good proxy for the potential quality of prediction, but not how to do it. (See Golan, 2018, Chapter 3 for discussion.): scale 2 – 0, with 2 (compression ratio of about 2), 1 (compression ratio of about 1.5), 0 (compression ratio of about 1). | 1.5 | 1 | 1.5 |
| | **Score** | 27.2 | 22.6 | 37.8 |

Appendix Table 4. A Detailed Basic Block 3 Evaluation of the Three Datasets

| **Basic Block 3.** Potential Value | | | | |
|---|---|---|---|---|
| | | Datasets | | |
| | | Movie Release Data | RUCC | USDA ARMS Data |
| A. | **Purpose and Meaning**<br>Provided in as a summary - in words - of the dataset. | | | |
| B. | **Semantics and Meaning**<br>All the possible mutually exclusive types/families of questions that we can answer with these data given our current information and knowledge. See discussion of context and meaning in Section 4. (Scalar). Note that there are possibly different meanings under different contexts. Scale: 1 – 10 with 10 the highest. This is a relative measure based on the data user's subjective understanding. If there is more than a single user, the median of all users should be used. | 5 | 2 | 8 |
| C. | **1. Importance of Questions to Society**<br>(Scale: 1 – 10 with 10 extremely important). Note, this is subjective and context dependent. (See discussion in B above.) | 2 | 4 | 7 |
| | **2. Audience**<br>Are some of the questions we can answer important for private entities? (Yes – No; If Yes scale 1 – 10 with 10 extremely important). | 1 | 5 | 8 |
| D. | **Extreme Events**<br>How many observations of extreme events are in the data? Scale: 0 – $\Phi$. Calculation of $\Phi$: the number of observations more than 3 standard deviations (sd) from the mean plus the number of observations between 2sd and 3sd from the mean multiplied by 0.5. | 0.52 | 0.19 | 1.18 |
| E. | **1. Structure**<br>Is there a clear deterministic observed structure (story) in the data? (A description – no scale/weight). | | | |
| | **2. Cause and Effect**<br>Is there information in the data that will allow potential studies of cause and effect? (Yes (1) – No (0)) | 1 | 0 | 1 |
| F. | **Questions**<br>Can we answer questions that we could not answer previously? (Yes (1) – No (0); If 'Yes,' How many?) Scale: 0 – # Questions. | 1 | 1 | 1 |
| G. | **Timely**<br>Are the data timely (up to date)? (Yes (1) – No (0)). | 0 | 0 | 1 |
| H. | **Release and Availability**<br>For a repeated data, how fast are they published and made available for the users? (Scale: 0 – 2, too slow (0), somewhat regular (1), fast (2)). | 0 | 1 | 2 |
| I. | **Consistency**<br>Is the information (say, the variables) in repeated data consistent across time? (Scale: 0 – 1). | 0 | 1 | 1 |
| J. | **Updated Inference**<br>Do the data have the potential to shed new light on older questions-answers? (No – Maybe – Yes: 0, ½, 1). | 0.5 | 0.5 | 0.5 |
| K. | **New**<br>Is it a completely new dataset – a data set that no one has ever used (say, at the tails of the data distributions)? (Yes – No) | 1 | 0 | 1 |
| L. | **Size and Representativeness**<br>Size and representativeness of population effected. (Scale: 1-5 with 5 a large population is represented and effected). | 1 | 5 | 4 |
| M. | **Heterogeneous**<br>Do the data present heterogenous information to properly answer questions? (No – Somewhat - Yes: 0 – ½ – 1; See also Quality block, J) | 0.5 | 1 | 0 |
| N. | **Reduction in Uncertainty**<br>Approximate measure of reduction in uncertainty from the new data. It is the approximate increase in our 'knowledge about something' conditional on the data (Yes – No; If yes, how much; normalized entropy and relative information – Kullback-Leibler divergence). Scale: normalized entropy 0 – 1, normalized relative information (0 – 1). Note: The Yes – No are absolute, but the 'how much' is relative. | | | |
| O. | **Data Collection Type**<br>Data collection: experiment, administrative survey, or administrative/privately collected data? (Scale: 0 – 1 with all but private (1) and private (0)). | 0 | 1 | 1 |
| P. | **Expected Outcomes**<br>What are the potential expected outcomes/inferences in the short-run (data specific) and those in the long-run. (Description, expected importance: 1 – 10). This is a relative measure. | 5 | 1 | 8 |
| Q. | **Time Horizon**<br>Time horizon of resulting inferred outcomes, answers, and decisions. Scale: short (3), medium range (2), long range (1).) | 3 | 1 | 3 |
| R. | **Uniqueness**<br>Does another 'quite similar' dataset exist? (Yes (1) – No (0)); If yes, can combining both will increase the joint value by more than adding the values of both: economy to scale. That is, new questions and dimensions open up a higher level of precision. (Note, combining datasets is super-additive in terms of the number of questions that may be asked. It increases quadratically, just like the number of possible correlations. See also U below.) Scale (if 'yes') 1 – 10 with 10 the highest value. | 0 | 10 | 8 |
| S. | **Data Additivity**<br>Data Additivity (linear, super additivity, etc.; Define 'additivity' here and its dimension (say – more observations, more variables, etc.). No scale – discussion in the paper. | | | |
| T. | **Number of Variables and Observations**<br>Size (number of observations vs. number of variables). Note that more observations may not increase the value as more variables. No scale/weight: this is a descriptive discussion, and the information is captured by other measures. | | | |
| | **Score** | 21.5 | 33.7 | 55.7 |

## Appendix C: Code Description (Codes Developed by Danielle Wilson)

This section outlines the way the codes were constructed for each of the discussed attributes. All attributes were coded in either Stata or Jupyter Notebook using Python. (*Exact link will be provided*.)

### *Entropy of a Single Random*

The entropy of each variable was computed using Stata. The following preliminary steps were taken:

1. Data were imported and each variable was specified as being in a discrete or continuous variable list.
2. A global was defined to set the number of intervals all continuous variables should be discretized to.

The following steps were applied in a loop over each discrete variable:

3. The frequency of each variable's outcome was computed. Each frequency was then used to compute corresponding probabilities ($p_k$'s, where each $p_k$ was the observed probability of event or outcome $k$).
4. The probabilities computed in Step 3 were confirmed to be proper (sum up to one)..
5. The following was computed for each variable's outcome and stored:$-p_k \log_2 p_k$.
6. Following equation (1), the entropy of each variable was computed by summing the results from Step 5, $-p_k \, log_2 \, p_k$, across all $k$ events.
7. The normalization, $\log_2 k$, for each variable was computed. The entropy of each variable (from Step 6) was normalized using the resulting value and stored accordingly.

The following steps were then applied in a loop over each continuous variable:

8. Each continuous variable was discretized into the specified number of intervals (from Step 2).
9. Steps 3 through 7 were repeated for each discretized continuous variable.

Stored results from the discrete and discretized continuous variables were reshaped into a table for export. The final table contained three columns. The first column contained the names of every variable in the data set, and the second and third columns contained the entropy and normalized entropy of each corresponding variable, respectively.

### *Entropy of Multiple Random Variables (Joint Entropy)*

The joint entropy of every combination of two variables in the dataset was computed using Stata. The following preliminary steps were taken:

1. Data were imported and each variable was specified as being in a discrete or continuous variable list.

2. A global was defined to set the number of intervals all continuous variables should be discretized to.
3. All continuous variables were discretized according to the specified global in Step 2. The original continuous variables were subsequently dropped.
4. Each of the remaining variables was indexed by appending a number at the end of each variable name.

The following steps were taken to compute the normalized joint entropy of every combination of two variables:

5. The joint probability of every two combinations of events was computed. A double loop was created to call upon combinations of variables using the indexing created in Step 4. The outer loop went through all variables, $k = 1, \dots, r$ while the inner loop went through all but the first variable, $j = 2, \dots, r$. The frequency of unique joint events was collected and used to compute the joint probability of every combination of two variables. These probabilities were subsequently exported as a separate Stata dataset.
6. Using the same double looping technique described in Step 5, the number of possible joint events, between discrete and discretized continuous variables ($k_{Varibale1} \times k_{Variable2}$) was computed and subsequently exported as a separate Stata dataset.
7. Exported results from Steps 5 and 6 were merged together. Every row in the dataset of joint probabilities from Step 5 was a combination of outcomes or events between every possible pair of variables. Every row in the dataset of possible joint events from Step 6 corresponded to a pair of variables. Consequently, a many to one merge was used to combine the later dataset (from Step 6) to the former (Step 5).
8. Using the merged dataset from Step 7, every joint probability was used to compute $-w_{kj} \log_2 w_{kj}$ for every combination of events.
9. Joint entropies were computed by following equation 2 and summing the components computed in Step 8 over every pair of variables.
10. The number of possible joint events was then used to compute a normalization, $\log_2(k_{Varibale1} \times k_{Variable2})$.
11. The normalization computed in Step 10 was applied to the joint entropies computed in Step 9.
12. Results from Step 11 were formatted into a lower triangular matrix where each cell contained the normalized joint entropy of two variables. This matrix was subsequently exported.

### *Mutual Information of Multiple Random Variables*

The mutual information of every two combinations of variables in the dataset was computed using Stata. The same four preliminary steps outlined in calculating the entropy of multiple variables were taken in preparation. Once completed, the following steps were taken:

1. The joint probability of every two combinations of events was computed. A double loop was similarly created to call upon combinations of variables using the indexing created during the last preliminary step. The outer loop went through all variables, $k = 1, \dots, r$ while the inner loop went through all but the first variable, $j = 2, \dots, r$. The frequency of unique joint events was collected

and used to compute the joint probability of every combination of two variables. These joint probabilities were exported as a sperate Stata dataset.

2. The frequency of each individual variable's outcomes was computed. These frequencies were then used to compute corresponding probabilities ($p_k$'s). These individual probabilities were similarly exported as a separate Stata dataset.
3. The exported dataset from Step 1 was opened and combined with the exported dataset from Step 2 using a many to one merge. The resulting dataset contained both joint and individual probabilities.
4. The joint and individual probabilities were used to compute $w_{kj}\ln\left(\frac{w_{kj}}{p_k q_j}\right)$, where $w_{kj}$ is the joint probability of outcome $k$ from the first variable and outcome $j$ from the second variable, $p_k$ is the individual probability of outcome $k$ from the first variable, and $q_j$ is the individual probability of outcome $j$ from the second variable.
5. Following equation 5, the mutual information between every two combination of variables was then computed by summing the components calculated in Step 4 over the corresponding pair of variables.
6. Results from Step 5 were formatted into a lower triangular matrix where each cell contained the mutual information between two variables. This matrix was subsequently exported.

### *Condition Number*

The condition number of combination of variables was computed using Stata. Data were imported and a list of variables by type (costs, benefits, inputs, attributes, and signals) were created using globals. Note that these lists were not mutual exclusive.

The "collin" command authored by Philip B. Ender from the University of California's Statistical Consulting Group was used to extract the condition number of each combination of variables by type[4]. The "collin" command computes a variety of diagnostic measures, including the condition number. Once run, the results for all measures are stores as scalars. The scalar for the condition number can be called and saved accordingly.

### *Data Integrity (Benford's Law)*

The empirical distribution of each variable with respect to Benford's law was computed in Jupyter Notebook using Python. The following packages were imported:

- sympy
- math
- pandas
- numpy
- fsolve (spicy.optimize)

[4] Additional details can be found in the command's help file available at <https://stats.oarc.ucla.edu/stat/stata/ado/analysis/collin.hlp>.

The following steps detail how the metric for each continuous variable was constructed. These steps can be easily modified to compute the metric for the entire dataset.

First, the following preliminary steps were taken:

1. Data were imported and each variable was specified as being in a discrete or continuous variable list. A data frame containing only continuous variables was created as this analysis cannot be applied to discrete variables.
2. Every element in the data frame of continuous variables (from Step 1) was converted to string format.
3. Using the resulting data frame (from Step 2), for each continuous variable, the first digit of every observation was extracted and stored as a new variable with the suffix "_first".
4. Using the "series.value_counts" function, a loop was run over all variables with the "_first" suffix. This loop collected the frequency of observed unique values. For each variable, the observed unique value was a non-zero digit, and the corresponding frequency (of observance) was in percent form. The frequency data for each variable was stored as a separate data frame.
5. The data frames with stored frequency data (corresponding to each variable from Step 4) were merged into one large data frame. The first column of this merged data frame contained the possible unique values of non-zero digits (1 to 9). The remaining columns contained the corresponding observed frequencies for each variable.

The subsequent steps were taken to compute the empirical distribution of digits for each continuous variable:

6. Using the merged data frame of frequencies from Step 5, the geometric mean of each variable was computed and stored.

The functional form of the empirical distribution using the variable's geometric mean is the solution to the optimization problem detailed in equation (9). The Lagrange multiplier within the functional form of this normalized solution, $p^*(D) = D^{-\lambda}/\sum_D D^{-\lambda}$, needs to first be solved for. This can be done by substituting $p^*(D)$ back into the observed moment constraint, i.e., the geometric mean, and solving for $\lambda$.

7. The "spicy.optimize.fsolve" function was used to solve for $\lambda$. This step was repeated for each variable using its corresponding geometric mean from Step 5. The resulting value for each variable's $\lambda$ was stored in a new data frame. The first column of this data frame contained each variable's name and the second contained the variable's solved value of $\lambda$.
8. The Lagrange multipliers were then used to calculate the denominator of $p^*(D)$, $\sum_D D^{-\lambda}$. This denominator was added as an additional column to the data frame created in Step 7.
9. The numerator for every digit, $D^{-\lambda}$, was then calculated for each variable and divided by the previously solved for denominator to compute $p^*(D)$. The results for $p^*(D)$ were stored and appended as an additional columns to the data frame from Step 8. The resulting data frame appears as follows:

| Variable | Lagrange Value | Denominator | Probability D = 1 | … | Probability D = 9 |
|---|---|---|---|---|---|
| variable1 | Created in Step 7. | Added in Step 8. | Added in Step 9. | … | Added in Step 9. |
| … | | | | | |
| variable N | | | | | |

10. For each variable, the solved values of $p^*(D)$ for all digits were summed to ensure that they together equal one.

The resulting data frame from Step 9 was exported as a CSV file.

### *'Cumulative Entropy' or 'Entropy Convergence' of Data*

The cumulative entropy of every variable in the dataset was computed using Stata. The following preliminary steps were taken:

1. Data were imported and each variable was specified as being in a discrete or continuous variable list.
2. A global was defined to set the number of intervals all continuous variables should be discretized to.
3. An "id" variable was used to numerate the observations in the dataset.

The following steps were applied in a loop over each discrete variable:

4. The normalization, $\log_2 k$, for each variable was computed.
5. An empty variable for the variable's cumulative entropy was created. This empty variable will be filled over the course of the subsequent steps.
6. The total number of observations in the dataset was stored as a local macro and used to define the range of an inner loop. This inner loop, which started from the first observation and ran until the last observation (defined by the local macro), was indexed by $n$ and did the following:
    a. Computed the frequency of each variable's outcome among observations with an "id" value (see Step 3) less than or equal to $n$. Using these frequencies, probabilities ($p_k$'s, where each $p_k$ is the observed probability of event or outcome $k$) were computed and stored. These probabilities were summed to ensure that they together equal one.
    b. Using the probabilities from Step 6a, the following was computed for each variable's outcome[5]: $-p_k \log_2 p_k$.
    c. Following equation (1), the cumulative entropy of each variable was computed by summing the results from Step 6b, $-p_k \, log_2 \, p_k$, across all $k$ events.
    d. The entropy calculation from Step 6c was normalized using $\log_2 k$ calculated in Step 4. The normalized entropy of each variable was saved in row $n$ of the (originally empty) cumulative entropy variable created in Step 5.
7. The cumulative entropy variable created in Step 5 (now filled) was then plotted as a connected line graph. The vertical axis of this plot was cumulative entropy value corresponding to the first $n$

[5] Note that steps 6b through 6d will use the subsample of observations with an "id" value less than or equal to $n$.

observations of the sample. The horizontal axis details the percent of the sample used to compute the corresponding cumulative entropy value.

The following steps were applied in a loop over each continuous variable:

8. Each variable was discretized into the specified number of intervals from Step 2.
9. Steps 4 through 7 are repeated for each discretized continuous variable.

### *Information Compression – The Shannon Limit*

Computation of the dataset's Shannon Limit was done in Jupyter Notebook using Python. The following packages were imported:

- sympy
- math
- pandas
- numpy
- struct
- collections
- groupby

The following preliminary steps were taken:

1. Data were imported and each variable was specified as being in either a discrete, continuous or string variable list.
2. A function to convert text to bits was defined[6]. This function was not used when computing the Shannon limit for the sample dataset from Craig, Greene, & Versaci (2015) as all variables were numeric (either discrete or continuous). However, this function was added for future use for any dataset the may contain string variables.

Once data were prepared and organized, and all needed functions defined, the following steps were taken to compute the Shannon Limit:

For continuous variables:

3. All data values across all continuous variables were concatenated into a single list. There were 62 observations and seven continuous variables in the sample dataset. Thus, the resulting list contained 434 (=62 x 7) elements.
4. Every element from the resulting list (from Step 3) was converted to double-precision floating-point format (IEEE 754 binary64) and stored as an element of a new list.

[6] This function was recommended by a user from Stack Overflow. Discussion of text to bit conversion and the recommended definition is available at <https://stackoverflow.com/questions/7396849/convert-binary-to-ascii-and-vice-versa>.

Note that alternative binary formats can be used with alternative degrees of precision. The resulting Shannon Limit should be robust to whatever chosen format.

For discrete variables:

5. All data values for all discrete variables were concatenated into a single list. The sample dataset contained 6 discrete variables, resulting in a list containing 372 (=62 x 6) elements.
6. Every element from the resulting list (from Step 5) was similarly converted to double-precision floating point format and stored as an element of a new list.

Note that if the data contained string variables, steps 3 and 4 would be repeated for this subset of variables and the text to bit function (defined in Step 2) applied accordingly.

7. Lists with data converted into binary form (from Steps 4 and 6) were concatenated into one large list.
8. The number of zeros and ones from the resulting list (Step 7) were counted. The number of zeros and ones were then summed together to determine the total number of bits.
9. The formula for the Shannon limit (equation 11) was then followed.
10. The elements needed to compute the formula, including $\pi$, K and N, and the Shannon Limit, Z, were organized into a data frame (with one observation) and exported as a CSV file.