A THEORY OF COMPLEMENTARITY FOR EXTRACTING ACCURATE DATA FROM INACCURATE SOURCES THROUGH INTEGRATION

(Research-in-Progress)

Irit Askira Gelman
University of Arizona
Askira@eller.arizona.edu

Abstract: The purpose of this research is to develop a theory that helps produce accurate data integration output given multiple, overlapping, inaccurate sources. At present, the theoretical foundations of solutions that center on source selection and conflict resolution with a similar goal are limited. This paper introduces a new solution approach and theory that center on notions of complementarity, based on the assumption that errors are not random. The essence of the new approach is the following: instead of concentrating on sources that are individually highly accurate, center on sources that have a complementary nature and yield highly accurate output when integrated. The theory can offer guides for the characterization of accuracy, effective use of accuracy estimates, source selection, and conflict resolution strategies. Implementations will require information about error patterns.

Key Words: Data Integration, Data Quality, Accuracy, Source Selection, Conflict Resolution, Complementarity.

INTRODUCTION

Consider a situation in which needed data appear in each of *n* equally accessible relation instances, all sharing the same schema and same semantics. All the relation instances have the same number of tuples, designating the same set of real world objects. All object identification problems have been resolved. However, the data sources, in this case relation instances, are not entirely accurate. They suffer from errors, causing inconsistencies among related data values in different sources. The question of interest in this work is how to apply data integration to get the highest accuracy out of the available data sources.

Various solutions have been proposed to this question, aiming to maximize integration output accuracy through "good" source selection and conflict resolution (e.g., [8,11,13]). However, the theoretical basis of these solutions has often been neglected. The conditions that would guarantee a desired outcome have not been specified, and the properties of the desired outcome are not clear either. This paper introduces a new solution approach and theory. A major assumption of this work is that errors are not random. Causal factors may include, for example, deliberate supply of false data by humans, lack of understanding by those who contribute the data, equipment malfunctions, and so on. Regardless, factors that produce errors in one source may have different effect or no effect at all on alternative sources. Subsequently, errors in different sources may have a complementary nature that can be taken advantage of through integration, provided this nature is detected and understood. The essence of the new approach is the following: instead of concentrating on sources that are individually highly accurate, center on sources that have a complementary nature, one way or another, and yield highly accurate output when integrated.

This paper refers, in particular, to the complementarity that exists when error rates vary widely within each source, such that subsets of the data that have high error rates in one source match subsets with low error rates in other sources. More precisely, the paper addresses a simpler scenario, where, rather than having low error rate, some data subsets are completely error free, or, more generally, have what is called here "limited perfect accuracy." The method of study is analytical, employing the Information Structure (IS) model [2,6] to portray error distributions—this model has the advantage that it enables the representation of variations in errors rates. Such representations serve for defining notions of complementarity, which prove to be useful in the identification of conditions under which the output of integration is error free. The study also separates circumstances that necessitate utilization of data from multiple sources, from those where the use of data can be limited to one source. This distinction is achieved through the concept of *fusion*. Ultimately, this work demonstrates the potential value of a data integration approach that is guided by complementarity as an organizing design principle. Applications would require information about error patterns.

The structure of the paper is the following. The second section outlines related work, and the contribution of the current work in that context. Notation, definitions, and terms that are used throughout this study are presented next, as well as definitions of notions classified as "limited perfect accuracy." The latter serve as a basis for definitions of complementarity and derivation of conditions under which the output of integration is free of errors. A following section discusses implications to source selection, conflict resolution, and metadata requirements. In addition, since this paper's assumptions about error patterns are narrow compared to the varied possible patterns in practical application settings, future extensions of this theory are discussed as well. The paper ends with a brief conclusions section. Proofs are provided in an appendix.

RELATED WORK

In recent years there is a growing number of studies that address data integration under the assumption that data are not necessarily accurate, such that an important objective of the integration process is to minimize the number of errors. Various studies focus, in particular, on problems of source selection, source ranking, and conflict resolution, given the availability of multiple, overlapping data sources that are not error free. For the most part, such work has been conducted under the relational database framework.

Several of these studies apply models of information about quality (metadata) [1,5,7,9,10,11,13,14,15], typically viewing data quality as multi-dimensional such that accuracy as an important dimension. Current work emphasizes the value of highly accurate sources. Measures of accuracy include the standard deviation of a value (a lower standard deviation corresponds to higher accuracy) [9], the ratio of correct values to the total number of values [10], and percentage of correct values [5].

Proposed conflict resolution solutions are fully automatic, semi-automatic, or manual. A fully automatic conflict resolution method is described in [13]. This method is supported by quality estimates and their respective reliability estimates. A semi-automatic conflict resolution model is proposed in [9]. The model applies quality and performance metadata combined with user-supplied weights, to provide utility-based resolution. Tools that enable experts to express their preferred resolution strategy are suggested in [4,8]. For related work on source ranking and source selection see, for example, [1,5,11,16].

At present, the theoretical basis of solutions that aim to maximize integration output accuracy through source selection and conflict resolution is limited. In particular, in many cases there is no proof that a proposed approach would result in a desired outcome in any sense. This work takes a step towards addressing the existing shortage of theory, and introduces a theory that centers on notions of complementarity rather than on highly accurate individual sources. Such theory can offer guides for the characterization of accuracy, effective use of accuracy estimates, source selection, source ranking, and conflict resolution strategies.

The assumptions that underlie this paper, mainly that error rates can vary significantly in different data subsets, resemble the understanding that has been recommended by Motro and Rakov (e.g., [10,13]), and has also been accepted by other researchers. However, unlike other studies, this work highlights its positive potential as far as data integration solutions.

BASIC DEFINITIONS

Prior to any analysis, we begin by introducing fundamental concepts.

A data source is represented by a one-dimensional or multidimensional random variable, denoted by $\boldsymbol{Y}$ or $\boldsymbol{Z}$. Data values are taken to be instances of $\boldsymbol{Y}$ (or $\boldsymbol{Z}$). The actual variable is represented by a one-dimensional or multidimensional random variable denoted by $\boldsymbol{S}$. Correct values are instances of $\boldsymbol{S}$.

A distinction between a data source and the information that a data source provides about the correct values is achieved through the notion of an information structure (IS) [2,6]. An *information structure* is a function $f: S \times Y \rightarrow \Re^+$ where S denotes a set of *states of the world*, Y denotes a set of *signals*, and, for every element s of S, $f(y|s)$ is a probability density function over Y. The set of states, S, and the signal set, Y, are not restricted, e.g., they can be finite or infinite. The example that serves throughout this paper refers to finite sets, however, the results apply also, in particular, to real numbers. Similarly, the probability density functions are not restricted. In fact, they need not even be the same under different states of the world. This way the definition of an IS provides a means for expressing variations in error rates or error distributions.

Assuming $\boldsymbol{Y}$ and $\boldsymbol{S}$ that take values in the sets Y and S, respectively, if, for every element s of S and y of Y, $f(y|s)$ is the conditional density of $\boldsymbol{Y}=y$ given $\boldsymbol{S}=s$, then f models the information that $\boldsymbol{Y}$ provides about $\boldsymbol{S}$. If $\boldsymbol{Y}$ is multidimensional, i.e., $\boldsymbol{Y}'=(\boldsymbol{Y}_1,..,\boldsymbol{Y}_n)$, then, under equivalent conditions, f models the *integration*, or *aggregate*, information that $\boldsymbol{Y}_1,..,\boldsymbol{Y}_n$, provide about $\boldsymbol{S}$. Specifically, assume that, for $j=1,..,n$, $\boldsymbol{Y}_j$ takes values in the set Y_j. Then, an IS $f: S \times Y \rightarrow \Re^+$, where $Y=Y_1 \times .. \times Y_n$, models the integration information that $\boldsymbol{Y}_1,..,\boldsymbol{Y}_n$, provide about $\boldsymbol{S}$ if for every s in S and y_j in Y_j, $j=1,..,n$, $f(y_1..y_n|s)$ is the joint conditional density of $\boldsymbol{Y}_j=y_j$, $j=1,..,n$, given $\boldsymbol{S}=s$.

To clarify the notion of an IS, assume that $\boldsymbol{Y}$ is a one-dimensional variable that corresponds to a source about the occupations of customers of some organization, and $\boldsymbol{S}$ corresponds to their actual occupations. The information that $\boldsymbol{Y}$ provides about $\boldsymbol{S}$ is modeled by an IS as follows. Suppose, for the sake of simplicity, that there are only three occupation types: business, engineering, and education. The state set is, therefore, S={business, engineering, education}. The signal set is, for instance, Y={Buss, Eng, Edu}, and f, the IS, is described by the matrix:

(1)

Signal /State	Buss	Eng	Edu
Business	.97	.02	.01
Engineering	.03	.85	.12
Education	0	.10	.90

According to this IS (1), if a customer's occupation is in business, the probability that the reported value is "Buss" is 0.97, the probability that it is "Eng" is 0.02, and the probability that it is "Edu" is 0.01. When the customer is an engineer the probability that the recorded value is "Eng" is 0.85, the probability that it is "Buss" is 0.03, and the probability that it is "Edu" is 0.12, and so on.

NOTIONS OF (LIMITED) PERFECT ACCURACY

An important understanding that motivates this work is that data sources are generally not error free. Error free data are modeled here by a *perfect IS*. A perfect IS is an IS where every signal is a *perfect signal*, i.e., a signal that points unambigously to one state.

Definition 1: Perfect signal. Let *f* denote an IS defined over SxY. If *y* in Y is such that $f(y|s)>0$ implies $f(y|s')=0$ for every $s', s'\neq s$, then *y points to s with certainty*, or *y is a perfect signal*.

Definition 2: Perfect IS. Let *f* denote an IS defined over SxY. If, for every *y* in Y, *y* is a perfect signal, then *f* is a *perfect IS*.

The IS in the earlier example (1) is not a perfect IS. An IS representing an error free source under that scenario would be a 3x3 identity matrix, i.e., a square matrix whose diagonal elements are 1s and whose off-diagonal elements are all 0s. Such matrix associates every signal with exactly one state. For instance, the signal "Buss" would be exclusively associated with the state "Business," since the probability of the signal "Buss"given any other state would be zero.

It is easy to see that if the information provided by a source is a perfect IS, then an IS that models the integration information provided by that source and any other source(s) is a perfect IS as well.

We assume that data sources have errors, and errors are not randomly distributed. In particular, the analysis will focus on conditions in which errors demonstrate characteristics that may be classified, broadly, as "limited perfect accuracy." Definition 3 portrays one instance in this category: *perfect IS given state s*. Definition 3 designates a situation in which, even if the source as a whole has errors, the source is accurate when the correct value is *s*. Accordingly, an IS is perfect given state *s* if every signal that has positive probability given *s* is associated exclusively with that state, i.e., the signal has zero probability given any other state.

Definition 3: Perfect IS given state *s*. Let *f* denote an IS defined over SxY. If *s* in S is such that, for every *y* in Y, *y* points to *s* with certainty, then *f* is a *perfect IS given s*.

Evidently, if an IS is perfect given every possible state, then it is a perfect IS. In the following IS, the signal "Buss" is a perfect signal—it points to the state "Business" with certainty:

(2)

Signal /State	Buss	Eng	Edu
Business	.91	.05	.04
Engineering	0	.85	.15
Education	0	.10	.90

The signal "Buss" in IS (2) is not produced unless the state is "Business," since the probability of that signal given any other state is zero. However, IS (2) is not perfect given the state "Business" or any other state. In contrast, the IS below (3) is perfect when the state is "Business." Hence, the respective source is always accurate when the actual occupation is in business.

(3)

Signal /State	Buss	Eng	Edu
Business	1	0	0
Engineering	0	.85	.15
Education	0	.10	.90

It is easy to show that if the information that a source provides is a perfect IS given some state, then the IS that models the integration information provided by that source and any other source(s) is, too, a perfect IS given that state (see also Lemma 1 in the next section).

The conditions that Definition 3 stipulates are significantly weaker compared to the requirements on a perfect IS. Nonetheless, Definition 4 forms a second instance of limited perfect accuracy that involves weaker conditions than those of Definition 3. The concept of an *IS that has perfect distinction between states s and s'* targets situations in which a value that a source shows is not error free altogether, but it reduces the range of possibilities for the true value. Accordingly, an IS has perfect distinction between the states *s* and *s'* if every signal that has a positive conditional probability given *s* has zero conditional probability given *s'*.

Definition 4: An IS has perfect distinction between states *s* and *s'*. Let *f* denote an IS defined over SxY. A signal *y* in Y *enables perfect distinction* between states *s* and *s'* in S, if $f(y|s)>0$ implies $f(y|s')=0$. If every *y* in Y enables perfect distinction between *s* and *s'*, then *f has perfect distinction between s* and *s'*.

When the information that a source provides has perfect distinction between states, then, again, so does the aggregate information provided by that source and any other source(s) (see also Lemma 3). In addition, if an IS has perfect distiction between a state *s* and any other state, then it is a perfect IS given state *s*.

The following IS (4) is such that the signal "Buss" enables perfect distinction between the states "Business" and "Education" on one hand, and "Engineering" on the other. The validity of "Engineering" is ruled out given this signal—only "Business" and "Education" are possible.

(4)

Signal /State	Buss	Eng	Edu
Business	.91	.05	.04
Engineering	0	.85	.15
Education	.05	.05	.90

However, IS (4) does not have perfect distinction between "Business" and "Engineering," or any other state. In contrast, the IS below (5) has perfect distinction between "Business" and "Engineering," since every signal enables perfect distinction between them.

(5)

Signal /State	Buss	Eng	Edu
Business	.91	0	.09
Engineering	0	1	0
Education	.05	.05	.90

The concept of perfect distinction between states is somewhat related to the notion of "partial values" [3], which has been introduced in the database integration literature in the context of conflict resolution when inconsistencies arise due to semantic mismatch. Partial values are "a finite set of possible values such that the "true" or "real" value of the partial value is exactly one of the values in that set." The combination of two partial values is given by their intersection.

One way in which partial values differ from the concept of ISs that have perfect distinction between states, is that the latter refers to errors, while the former refers to semantic differences. Hence, their treatment is not always the same. For example, partial values are created part of a database integration process, through user-defined, one-many and many-many mappings between domains of actual attributes

and domains of virtual attributes. In contrast, implementation of the concept of perfect distinction between states will involve, in some cases, analysis of errors in individual databases that will lead to the creation of metadata. Second, Definition 4 is not limited to finite sets. When the state set and the signal set are subsets of the real numbers, perfect distinction between states may be expressed using, for instance, intervals that contain the actual state (e.g., $\boldsymbol{S}<n_1$, $n_1 \leq \boldsymbol{S} \leq n_2$).

Definition 3 and Definition 4 have outlined the error patterns that will be examined in this study. Subsequent analysis will center on the integration of overlapping data sources that are not error free, though errors display limited perfect accuracy in agreement with Definition 3 and/or Definition 4. The investigation will focus on conditions in which the integration of sources that exhibit such errors yields error free output, but it also illustrates how integration can generally improve accuracy in a known ways, given such sources.

INCREASING DATA INTEGRATION OUTPUT ACCURACY: COMPLEMENTARITY RELATIONS

Suppose that none of the available sources is error free, but some obey one or more of the error patterns defined above. Two alternative sets of conditions that enable error free output when sources of this kind are integrated are identified. These conditions are collectively recognized by the term "complementarity." The second set of conditions is a generalization of the first and is perhaps less congruent with existing data quality solutions. Both have an advantage as far as their implementation. For the most part, they refer to data quality properties of data in individual sources—such properties may be known, or studied, independent of any integration setting.

The analysis also addresses a property of the output of integration that is termed "fusion." The notion of fusion assists in distinguishing between settings in which the output of the integration of two (or more) values is determined based on both values, from settings in which it can be based on just one value. This distinction can be useful in guiding conflict resolution, and may have efficiency implications—when the identity of the preferred source is known, there may be no need to consult additional sources.

We begin with the definition of fusion, and proceed with a study under the assumption that sources adhere to Definition 3 (ISs are perfect given one or more states), followed by a more general analysis in agreement with Definition 4.

In essence, a signal in an integration IS is a fusion if none of the signals that it comprises suggests the same likelihood of states. An integration IS is a fusion if at least one of its signals is a fusion.

Definition 5: Fusion. Let f_j, j=1,..,n, denote ISs defined over SxY_j, respectively, such that, for every j, f_j models the information that $\boldsymbol{Y}_j$ provides about $\boldsymbol{S}$. Let h, defined over SxY, $Y=Y_1$x..xY_n, denote the integration information that $\boldsymbol{Y}_j$, j=1,..,n, provide about $\boldsymbol{S}$. A signal $(y_1..y_n)$ in Y such that $h(y_1..y_n|s)>0$ for some s in S is a *fusion* if, for every j, there exist s_{j1}, s_{j2}, in S such that $h(y_1..y_n|s_{j1})/h(y_1..y_n|s_{j2}) \neq f_j(y_j|s_{j1})/f_j(y_j|s_{j2})$. If there exists a signal in Y that is a fusion, then h is a *fusion*.

When ISs are perfect given a state

Assume several ISs where none is a perfect IS, but one or more is perfect given a state(s). We say that one IS complements another IS in a certain state, if, given that state, the former is perfect, while the latter is not perfect. Definition 6 corresponds to a situation in which a source that is consistently accurate when the actual value is a certain value is integrated with a second source that is not error free as far as that value.

Definition 6: Complementarity in state. Let *f*, *g*, denote ISs defined over SxY, SxZ, respectively. It is said that *f complements g in state s*, if *f* is perfect given *s*, and *g* is not perfect given *s*.

The notion of complementarity in a state creates an asymmetric, irreflexive, binary relation over the set of ISs over S.

When an IS complements another IS in a state, then an IS that models their integration information is perfect given that state. Lemma 1 asserts this understanding.

Lemma 1: Let *f*, *g*, denote ISs defined over SxY, SxZ, respectively. *f* models the information that $\boldsymbol{Y}$ provides about $\boldsymbol{S}$, *g* models the information that $\boldsymbol{Z}$ provides about $\boldsymbol{S}$, and *h* models the integration information that $\boldsymbol{Y}$ and $\boldsymbol{Z}$ provide about $\boldsymbol{S}$. If *f* complements *g* in state *s* in S, then *h* is a perfect IS given *s*.

Lemma 1 hints that by repeatedly adding sources of this kind the integration information can reach perfect accuracy. Proposition 1 refers to such conditions in detail. The integration information of any number of ISs is a perfect IS if for any IS that is not perfect given some state, there is another IS that complements it in that state.

Proposition 1: Let f_j, j=1,..,n, denote ISs defined over SxY_j, respectively, such that, for every j, f_j models the information that $\boldsymbol{Y}_j$ provides about $\boldsymbol{S}$, and f_j is not a perfect IS. Let *h* denote the integration information that $\boldsymbol{Y}_j$, j=1,..,n, provide about $\boldsymbol{S}$. Then, if, for every j and every s in S, f_j is not perfect given s implies that there exists k, $1 \le k \le n$, such that f_k complements f_j in s, then *h* is a perfect IS.

Turn to the fusion property, Lemma 2 suggests that any signal of an integration IS that comprises a signal which points to a state with certainty, is not a fusion. The intuition behind this lemma is that a signal that points to a state with certainty is just as accurate as a composite signal that comprises it. Subsequently, when every signal of the integration IS comprises a signal which is just as accurate, as is the case when the conditions of Proposition 1 are met, then that IS is not a fusion (Proposition 2).

Lemma 2: Let f_j, j=1,..,n, denote ISs defined over SxY_j, respectively, such that, for every j, f_j models the information that $\boldsymbol{Y}_j$ provides about $\boldsymbol{S}$. Let *h*, defined over SxY, $Y=Y_1$x..xY_n, denote the integration information that $\boldsymbol{Y}_j$, j=1,..,n, provide about $\boldsymbol{S}$. If y_j in Y_j points to s in S with certainty, then, every signal $(y_1..y_j..y_n)$ in Y (i.e., that comprises y_j) is not a fusion.

Proposition 2: Under the assumptions of Proposition 1 *h* is not a fusion.

Example:
Consider three overlapping sources about customer occupations, such that none is free of errors. The rates of errors in each of these sources vary. The variation is mainly due to special discounts and other bonuses that are given to customers in selected occupations, and motivate strict verification of those occupations. One of the sources is maintained in an environment in which customers from the education sector must show appropriate documents that verify their occupation. The other two sources are products of similar verification procedures, applied to business, and engineering, respectively. The ISs that describe these sources are given by:

(6)

Signal /State	Buss	Eng	Edu
Business	.1	0	0
Engineering	0	.83	.17
Education	0	.07	.93

(7)

Signal /State	Buss	Eng	Edu
Business	.85	0	.15
Engineering	0	1	0
Education	.08	0	.92

(8)

Signal /State	Buss	Eng	Edu
Business	.9	.1	0
Engineering	.06	.94	0
Education	0	0	1

Proposition 1 requires that for every state where an IS is not perfect, there is another IS that complements the former in that state. The ISs above satisfy this requirement. IS (6) complements the other two ISs in "Business," IS (7) complements the other ISs in "Engineering," and IS (8) complements the other ISs in "Education." The integration IS is:

(9)

Signal /State	Buss, Buss, Buss	Buss, Buss, Eng	Buss, Edu, Buss	Buss, Edu, Eng	Eng, Buss, Edu	Eng, Eng, Buss	Eng, Eng, Eng	Eng, Edu, Edu	Edu, Buss, Edu	Edu, Eng, Buss	Edu, Eng, Eng	Edu, Edu, Edu
Business	.765	.085	.135	.015	0	0	0	0	0	0	0	0
Engineering	0	0	0	0	0	.0498	.7802	0	0	.0102	.1598	0
Education	0	0	0	0	.0056	0	0	.0644	.0744	0	0	.8556

In accord with the conclusion of Proposition 1, this IS (9) conforms to the definition of a perfect IS, representing error free output. (The numbers in this matrix were derived based on an assumption of conditional independence.)

In accord with the conclusion of Proposition 2, IS (9) is not a fusion, since none of the signals of that IS is a fusion. Take, for example, the signal (Buss, Buss, Buss), which points to the state "Business" with certainty. According to Lemma 2, this signal is not a fusion. It consists of three signals, such that the first one is derived from the IS in (6). However, that signal, just like the composite signal, points to "Business" with certainty. Therefore, given that signal, the other two signals are redundant.

When ISs have perfect distinction between states

We now turn to conditions in which none of the ISs is perfect, yet one or more has perfect distinction between states. A second notion of complementarity is defined, that assists in portraying conditions in which a respective integration IS is perfect. More precisely, the ensuing analysis focuses on conditions under which the integration IS is perfect given a state. (This analysis can be easily combined with the earlier analysis to produce understanding regarding conditions that yield a perfect IS.) Unlike before, the findings show that when ISs are not perfect given a state but have perfect distinction between states, the outcome of integration is a fusion.

When ISs have perfect distinction between states we say that one IS complements another IS in distinguishing between two states if the former has perfect distinction between the states while the latter does not.

Definition 7: Complementarity in distinction. Let *f*, *g*, denote ISs defined over SxY, SxZ, respectively. It is said that *f complements g in distinguishing between states s and s'*, if *f* has perfect distinction between *s* and *s'*, and *g* does not have perfect distinction between *s* and *s'*.

Again, this definition creates asymmetric, irreflexive, binary relation over the set of ISs over S.

When an IS complements another IS in distinguishing between states their integration inherits the perfect distinction property. Lemma 3 asserts this understanding.

Lemma 3: Let *f*, *g*, denote ISs defined over SxY, SxZ, respectively. *f* models the information that $\boldsymbol{Y}$ provides about $\boldsymbol{S}$, *g* models the information that $\boldsymbol{Z}$ provides about $\boldsymbol{S}$, and *h* models the integration information that $\boldsymbol{Y}$ and $\boldsymbol{Z}$ provide about $\boldsymbol{S}$. If *f* complements *g* in distinguishing between *s* and *s'* , then *h* has perfect distinction between *s* and *s'* in S.

Lemma 3 suggests that by repeatedly adding sources of this kind the integration information can reach perfect accuracy. Proposition 3 centers on this issue directly, by pointing to conditions where the integration IS is perfect given some state. The integration IS is perfect given state *s* if, for any state *s'*, there is an IS that complements any IS that does not enable perfect distinction between *s* and *s'*.

Proposition 3: Let f_j, j=1,..,n, denote ISs defined over SxY_j, respectively, such that there exists *s* in S where for every j, f_j is not a perfect IS given *s*. Suppose that, for every j, f_j models the information that $\boldsymbol{Y}_j$ provides about $\boldsymbol{S}$, and let *h* denote the integration information that $\boldsymbol{Y}_j$, j=1,..,n, provide about $\boldsymbol{S}$. Then, if, for every j and every *s'* in S, f_j does not have perfect distinction between *s* and *s'* implies that there exists f_k, $1 \leq k \leq n$, such that f_k complements f_j in distinguishing between *s* and *s'*, then *h* is a perfect IS given *s*.

Turn, again, to the fusion property. Unlike the former scenario, the integration IS can be a fusion under the current assumptions. The (composite) signal of the integration may point to *s* with certainty although none of the signals that it comprises does. Such signal is a fusion, and, by definition, the IS is a fusion as well.

Proposition 4: Suppose, under the assumptions of Proposition 3, that y_j in Y_j, j=1,..,n, are such that none is a perfect signal. Then, if $h(y_1..y_n|s)>0$, *h* is a fusion.

Example:
Consider two, overlapping, sources about customer occupations, that are not free of errors. The scenario of the previous example still applies, with some variation. Despite the strict verification procedures as far as some occupations, the accuracy of occupation data is deliberately compromised in special cases. The motivation for such compromise could be to enable certain customers enjoy discounts, or bonuses, that they would not normally get due to their actual occupation. In particular, we assume that in one instance there are strict verification procedures as far as the business sector, but a few customers from the education sector are registered as business people (10 below). In a second case, there are strict procedures as far as educators, but a few engineering people are registered, nonetheless, as educators (11 below):

(10)

Signal /State	Buss	Eng	Edu
Business	1	0	0
Engineering	0	.83	.17
Education	.03	.04	.93

(11)

Signal /State	Buss	Eng	Edu
Business	.9	.1	0
Engineering	.04	.94	.02
Education	0	0	1

The ISs (10) (11) satisfy the requirements of Proposition 3. Especially, none of them is perfect given the state "Business," however, IS (10) has perfect distinction between "Business" and "Engineering," and IS (11) has perfect distinction between "Business" and "Education." Consequently, the two ISs complement each other. The integration IS is:

(12)

Signal /State	Buss, Buss	Buss, Eng	Buss, Edu	Eng, Buss	Eng, Eng	Eng, Edu	Edu, Buss	Edu, Eng	Edu, Edu
Business	.9	.1	0	0	0	0	0	0	0
Engineering	0	0	0	.0332	.7802	.0166	.0068	.1598	.0034
Education	0	0	.03	0	0	.04	0	0	.93

The IS above (12) is perfect given the state ”Business.” (Again, the numbers in this matrix were derived assuming conditional independence.)

IS (12) is a fusion. For example, the signal (Buss, Buss) points to “Business” with certainty, although, taken individually, the signal “Buss” is not a perfect signal in any of IS (10) or IS (11). When the signal “Buss” is received from a source that matches IS (10), the possibility that the state is “Engineering” is ruled out. When the signal “Buss” is received from a source that matches IS (11), the possibility that the state is “Education” is ruled out. Together, the two signals determine that the state is “Business.” Therefore, the signal (Buss, Buss) is a fusion.

DISCUSSION

If errors are not randomly distributed, such that, for example, error-rates vary significantly within each source, then errors in different sources may have a complementary nature that can be exploited through data integration. This is a major assumption of this research. This paper refers, in particular, to error patterns that are broadly described as “limited perfect accuracy.” One instance in this category is when a source as a whole has errors, but it is accurate within a subset of the values. A second instance, which is a generalization of the first, is when a source has errors, but a given value rules out a subset of the range of possible values. Although the examples in the paper refer to sources that are described by one-dimensional random variables, the analysis applies also to multidimensional variables.

The analysis demonstrates very intuitive points. Mainly, when sources that have errors as above are also complementary, namely, when error free subsets vary among different sources, or different sources rule out different possible values, their integration can increase the accuracy of the data in known ways. In fact, under best conditions the outcome will reach perfect accuracy. Nonetheless, such theory suggests a new approach to integration. This approach assumes the availability of information about errors, and instead of concentrating on data that are individually highly accurate, centers on data that have a complementary nature, one way or another, to produce accurate outcome through integration. These aspects, as well as future extensions of this theory, are discussed next.

Complementarity and source selection

With a fast-growing number of competing sources in numerous domains, effectiveness and efficiency considerations advise the importance of correct selection of sources for data integration. The traditional approach to source selection says that sources that score highest by some measure of accuracy are preferred (e.g., [11]). A source selection decision based on complementarity could lead to a substantially different choice. This argument is clarified by an example—the example does not apply any defined measure of accuracy, it merely aims to convey an underlying intuition.

Example:

Consider three overlapping sources that are modeled by the following ISs.

(13)

Signal /State	Buss	Eng	Edu
Business	1	0	0
Engineering	0	1	0
Education	0	.01	.99

(14)

Signal /State	Buss	Eng	Edu
Business	1	0	0
Engineering	0	1	0
Education	0	.02	.98

(15)

Signal /State	Buss	Eng	Edu
Business	.87	.02	.11
Engineering	0	1	0
Education	.12	0	.88

Suppose that only two sources are allowed to take part in data integration. The question then is which two of the three sources would offer the best outcome. IS (13) and IS (14) look more similar to a perfect (IS) than IS (15). Therefore, selection based on accuracy might point to the sources matching (13) and (14). In

contrast, IS (15) complements each of the ISs (13) and (14) in distinguishing between "Engineering" and Education." Therefore, the aggregate information of the sources represented by (13) and (15) is a perfect IS, and so is the aggregate information of the sources represented by (14) and (15). At the same time, both (13) and (14) reflect the same weakness—none enables perfect distinction between "Engineering" and "Education." Consequently, selection that obeys complementarity would prefer any of the pairs (13) and (15), and (14) and (15), over (13) and (14).

Assuming that a set of data sources satisfies the conditions of Proposition 1 or Proposition 3, the problem of finding a subset that satisfies these conditions such that data exchange cost is minimal can be formalized, when the state set S is finite, by an integer programming set covering model. An algorithm for optimal solution of such model would be exponential, however, the set covering problem has an efficient heuristic algorithm with a performance guarantee [12].

Complementarity and conflict resolution

Conflict resolution strategies can be guided by complementarity relations. Each complementarity relation introduced in this paper implies a different conflict resolution strategy. When complementarity in state is observed, conflict resolution can be limited to singling out the data of the suitable source. Handling complementarity in distinction would involve processing data from multiple sources, in order to eliminate all the impossible values.

Information requirements

Data integration that is guided by complementarity as a design principle requires information about error patterns. Furthermore, a detailed study of the errors may reveal pockets of high accuracy, which an aggregate measure of accuracy would not disclose. Therefore, for best results, application should be based on detailed understanding of error distributions (e.g., metadata). A good feature of the theory presented here is that it is largely based on properties of individual sources—such properties may be known, or studied, independent of any integration setting. In some cases, the source of the desired understanding may be domain experts, e.g., people that handle the creation of the data. However, in recent years there is also research about automatic approaches. For example, Motro and Rakov [10,13] present a data analysis approach for producing detailed estimates of accuracy and completeness. Surprisingly, existing source selection and conflict resolution work does not make direct use of detailed accuracy estimates even when such estimates are assumed available—aggregates are preferred over detailed estimates (e.g., [1,5,10,11, 13]).

Here is a simple example that illustrates the importance of detailed understanding of the data. The information source in this example is represented by a multidimensional variable rather than a one-dimensional variable.

Example:
Consider a situation in which data about occupations corresponds to an attribute in a relation part of a relational database, such that a subset of the ideal relation is as depicted by Table 1. ("S" in the age group column in Table 1 refers to seniors, "Y" refers to young customers, and "A" to adults.)

Rec #	First Name	Last Name	Age Group	Occupation	Rec #	First Name	Last Name	Age Group	Occupation
1	Jim	Davis	Y	Buss	11	Robert	Young	A	Edu
2	Jennifer	Duarte	A	Edu	12	David	Wood	A	Buss
3	Gerald	Gutierrez	A	Edu	13	Raina	Wiley	Y	Buss
4	Erin	Henderson	A	Eng	14	Joice	Spitz	Y	Buss
5	Tiffany	Knuth	A	Buss	15	Daniel	Sanders	S	Eng
6	Sam	Newell	Y	Eng	16	Andrew	Richards	S	Edu
7	Leslie Ann	Presnell	S	Edu	17	Michael	Campbell	A	Edu
8	Daniel	Reed	S	Eng	18	Elaine	Cook	S	Buss
9	Martin	Sawyer	S	Buss	19	Andrea	Billings	Y	Eng
10	Adele	McKinley	A	Edu	20	Bryan	Ross	S	Eng

Table 1: A subset of the correct relation instance.

Suppose that a customer's declared occupation is checked if the customer associates himself or herself with the business sector. In addition, although verification is strict with customers in the mid-age group, it is not as strict with people in the young age group and seniors. (Members of the latter age groups may receive age-specific bonuses, therefore they may be required to prove their claimed age group.)

Error rates could vary in this case both by occupation and by age group. Assume that the aggregate information that occupation and age data provide about customers' occupations is portrayed by the following IS:

(16)

Signal /State	Buss, Y	Buss, A	Buss, S	Eng, Y	Eng, A	Eng, S	Edu, Y	Edu, A	Edu, S
Business	.22	.39	.3	.02	.02	.01	.02	.01	.01
Engineering	.01	0	.01	.21	.45	.25	.01	.04	.02
Education	.03	0	.01	.01	.06	.01	.23	.39	.26

According to (16), the signal (Buss, A) points to the state "Business" with certainty. Therefore, whenever the data report that a customer's occupation is "Buss" and he/she is in the "A" age group, the data are free of errors. However, a less detailed portrayal of the data, which ignores the joint effect of age and occupation on errors and considers only the pattern of errors in the occupation data, would miss the above described, potentially useful, perfect signal:

(17)

Signal /State	Buss	Eng	Edu
Business	.91	.05	.04
Engineering	.02	.91	.07
Education	.04	.08	.88

Potential theory extensions

The assumptions that this work makes on error patterns do not cover the varied potential in practical application settings. However, the theory can be extended to comparable notions of complementarity based on weaker assumptions on error patterns. Such theory will apply to circumstances in which sources demonstrate less than perfect distinction between states, i.e., subsets of the data that have high error rates in one source match subsets with low error rates in other sources.

A more complete understanding of data integration accuracy should also involve the introduction of notions of complementarity associated with dependence between errors, or sources. An example scenario will clarify this direction:

Two sources show basic demographic data relating to a chosen population. The origins of the data are self-reports volunteered by users. Data in the first source are collected part of users' job seeking efforts, while in the second source data are collected in social circumstances. Age data form part of the information. Age data have errors in both sources, primarily because people often misreport their age. Errors are most common at the tails of the population, i.e., relatively young or relatively old people. Error patterns suggest that young people that tend to inflate their age in job seeking contexts take years off their age in social circumstances, such that, mainly, there is a negative correlation between respective errors in the two sources. On the other hand, older people take years off in both cases—correlation is positive.

Knowledge about dependence relationships of this kind may be useful for decision-making about source selection and conflict resolution. Take conflict resolution, for instance—when errors are negatively correlated, a conflict resolution strategy that combines the data (e.g., average) can produce highly accurate estimates of the true age. Subsequently, negative correlation between errors may be viewed as another form of complementarity that can contribute to higher integration output accuracy.

CONCLUSIONS

The purpose of this research is to develop a theory that can help produce the highest integration output accuracy, given multiple, overlapping, inaccurate sources. This paper introduces a theory that demonstrates the potential value of a data integration approach that is guided by complementarity as an organizing design principle. Such theory can offer guides as far as the characterization of accuracy, effective use of accuracy estimates, source selection, source ranking, and conflict resolution strategies.

Future work should be conducted in several directions. There is a need to extend this work, in particular, develop a corresponding theory under different assumptions on error patterns. Theory should be implemented and evaluated in practical scenarios—issues such as fitness to existing approaches and technological environment, costs, and gains in performance would be, in general, of interest. Importantly, regardless of the specific integration setting, implementation must be based on information about error distributions. Therefore, research in this direction is relevant too.

REFERENCES

[1] Avenali, A., Bertolazzi, P., Batini, C., and Missier, P., "A Formulation of the Data Quality Optimization Problem in Cooperative Information Systems." *International Workshop on Data and Information Quality in conjunction with CAISE'04*, Riga, Latvia, 2004.

[2] Blackwell, D. "Equivalent comparisons of experiments." *Annals of Mathematical Statistics* Vol. 24, No. 2, 1953, pp. 265-272.

[3] Demichiel, L., G., "Resolving Database Incompatibility: An Approach to Performing Operations over Mismatched Domains." *IEEE Transactions on Knowledge and Data Engineering*, Vol. 1, No. 4, 1989, pp. 485-493.

[4] Galhardas, H., Florescu, D., Shasha, D., and Simon, E., "An Extensible Framework for Data Cleaning."

[5] Holland, G. "Methods for Building Data Elements for Multi-sourced Data Products." *8th International Conference on Information Quality*, ICIQ 2003, Cambridge, Ma.

[6] McGuire, C. B., "Comperisons of Information Structures." In *Decision and Organization*, C.B. McGuire and R. Radner (Eds.), University of Minnesota Press, 2nd edition, 1986.

[7] Missier, P., and Batini, C. "A Multidimensional model for Information Quality in Cooperative Information Systems," *8th International Conference on Information Quality*, ICIQ 2003, Cambridge, Ma.

[8] Motro, A., Anokhin, P. "Resolving Inconsistencies in the Multiplex Multidatabase System." ISE-TR-99-07, 1999.

[9] Motro, A., Anokhin, P., and Acar, A. C. "Utility-based Resolution of Data Inconsistencies." In *Proceedings of IQIS 04, International Workshop on Information Quality in Information Systems* (at SIGMOD 2004, International Conference on Management of Data), Paris, France, June 2004, pp. 35--43.

[10] Motro, A., and Rakov, I. "Not All Answers Are Equally Good: Estimating the Quality of Database Answers." In *Flexible Query-Answering Systems* (T. Andreasen, H. Christiansen, and H.L. Larsen, Editors). Kluwer Academic Publishers, 1997, pp. 1-21.

[11] Naumann, F., Leser, U., and Freytag, J. "Quality-driven integration of heterogeneous information systems." In *Proceedings of the 25th International Conference on Very Large Data Bases*, Edinburgh, U.K., 1999.

[12] Nemhauser, G.L., and Wolsey, L.A. "Integer Programming." in Nemhauser, G.L., Rinnooy Kan, A.H.G., and Todd, M.J. (Eds.), *Handbooks in Operations Research and Management Science, Vol. 1*:*Optimization*, North-Holland: Elsevier Science B.V, 1989.

[13] Rakov, I., "Quality of Information in Relational Databases and its Use for Reconciling Inconsistent Answers in Multidatabases." *Fourth Doctoral Consortium on Advanced Information Systems Engineering*, 1997

[14] Wang, R.Y., Madnick, S.E. "A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective." In *Proceedings of the 16th International Conference on Very Large Data Bases*, 1990, pp. 519-538

[15] Wang, R.Y., Reddy, M.P., and H.B.Kon. "Toward quality data: An attribute-based approach." *Decision Support Systems*, 13, 1995, pp. 349-372.

[16] Yu, J., Li, Yong-Qing, and Wessels, D. "Fuzzy Algorithm for Selection of Reliable Information from a Collection of Different Data Sources." *5th International Conference on Information Quality*, ICIQ 2000, Cambridge, Ma.

APPENDIX

Proof of Lemma 1: $\forall y \in Y$, if $h(y,z|s)>0$, then, by the definition of marginal density, $f(y|s)>0$. However, by assumption, f is perfect given s. Therefore, by Definition 3 and Definition 1, $f(y|s')=0$, $\forall s' \in S$ s.t. $s' \neq s$. It follows, by the definition of joint density, that $h(y,z|s')=0$, $\forall s' \in S$ s.t. $s' \neq s$. We have proved that, $\forall y \in Y$, $h(y,z|s)>0 \Rightarrow h(y,z|s')=0$, $\forall s' \in S$ s.t. $s' \neq s$. Therefore, by Definition 3 and Definition 1, h is a perfect IS given s.

Proof of Proposition 1: $\forall s \in S$ and f_j $(1 \leq j \leq n)$, if f_j is not perfect given s, then, by assumption, $\exists k$, $1 \leq k \leq n$, s.t. f_k complements f_j in s. Therefore, by Definition 6, $\forall s \in S$, $\exists i$ $(1 \leq i \leq n)$ s.t. f_i is perfect given s. Now add a second source, $\boldsymbol{Y}_m$ $(1 \leq m \leq n)$. Consider f_i, f_m, and h_{im}, where h_{im} denotes an IS that models the integration information of $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_m$ about $\boldsymbol{S}$. The proof of Lemma 1 applies to f_i, f_m, and h_{im}, whether or not f_m is perfect given s. It follows that h_{im} is perfect given s. Adding one source at a time this logic can be repeatedly applied to show that h is perfect given s. We have proved that, $\forall s \in S$, h is perfect given s. Therefore, h is a perfect IS.

Proof of Lemma 2: Since $y_j \in Y_j$ points to s with certainty, then, by Definition 1, $f_j(y_j|s)>0 \Rightarrow f_j(y_j|s')=0$ $\forall s' \in S$ s.t. $s' \neq s$. If $f_j(y_j|s)=0$, then, by the definition of joint density, $h(y_1..y_j..y_n|s)=0$, so $(y_1..y_j..y_n)$ is not a fusion. Therefore, assume $f_j(y_j|s)>0$, $h(y_1..y_j..y_n|s)>0$. Since $f_j(y_j|s')=0$ $\forall s' \in S$ s.t. $s' \neq s$, it follows $h(y_1..y_j..y_n|s')=0$, $\forall s' \in S$ s.t. $s' \neq s$. But then, $\forall s' \in S$ s.t. $s' \neq s$, $h(y_1..y_j..y_n|s')/h(y_1..y_j..y_n|s)=f_j(y_j|s')/f_j(y_j|s)=0$. Moreover, since $h(y_1..y_j..y_n|s')=0$ and $f_j(y_j|s')=0$ $\forall s' \in S$ s.t. $s' \neq s$, an inequality as in Definition 5 is not possible as far as y_j. Therefore, $(y_1..y_j..y_n)$ is not a fusion.

Proof of Proposition 2: Consider $(y_1..y_n) \in Y_1 x..x Y_n$. If $(y_1..y_n)$ is such that $h(y_1..y_n\ |s)=0$ $\forall s \in S$, then $(y_1..y_n)$ is not a fusion. Therefore, assume that $h(y_1..y_n|s')>0$. According to Proposition 1, $\forall s \in S$ and f_j $(1 \leq j \leq n)$, if f_j is not perfect given s, then, $\exists k$, $1 \leq k \leq n$, s.t. f_k complements f_j in s. Therefore, $\exists i$ $(1 \leq i \leq n)$ s.t. f_i is perfect given s'. Consider $y_i \in Y_i$ s.t. y_i appears in the composite signal $(y_1..y_n)$. By Definition 3, y_i points to s' with certainty. Therefore, by Lemma 2, $(y_1. .y_n)$ is not a fusion. We have shown that, $\forall (y_1..y_n)$, $(y_1..y_n)$ is not a fusion. It follows, by definition, that h is not a fusion.

Proof of Lemma 3: Suppose that $h(y,z|s)>0$. Therefore, by the definition of marginal density, $f(y|s)>0$. However, by assumption, f has perfect distinction between s and s'. Therefore, by Definition 4, $f(y|s')=0$. It follows by the definition of joint density that $h(y,z|s')=0$. Therefore, (y,z) enables perfect distinction between states s and s'. We have shown that $\forall (y,z)$, (y,z) enables perfect distinction between states s and s'. Therefore, by Definition 4, h has a perfect distinction between s and s'.

Proof of Proposition 3: $\forall s' \in S$, if f_j does not have perfect distinction between s and s', then, by assumption, $\exists k$ $(1 \leq k \leq n)$ s.t. f_k complements f_j in distinguishing between s and s'. Therefore, by Definition 7, $\forall s'$, $\exists i$ $(1 \leq i \leq n)$ s.t. f_i has perfect distinction between s and s'. Now add a second source, $\boldsymbol{Y}_m$ $(1 \leq m \leq n)$. Consider f_i, f_m, and h_{im}, where h_{im} is an IS that models the integration information of $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_m$ about $\boldsymbol{S}$. The proof of Lemma 3 applies to f_i, f_m, and h_{im}, whether or not f_m has perfect distinction between s and s'. It follows that h_{im} has perfect distinction between s and s'. Adding one source at a time, the same logic can be repeatedly applied to show that h has perfect distinction between s and s'. We have proved that, $\forall s' \in S$, h has perfect distinction between s and s'. Therefore, h is a perfect IS given s.

Proof of Proposition 4: According to Proposition 3, h is a perfect IS given s. Since, by assumption, $h(y_1...y_n|s)>0$, then, by definition, $h(y_1...y_n|s')=0$, $\forall s' \in S$, s.t. $s' \neq s$,. Therefore, $h(y_1...y_n|s')/h(y_1...y_n|s)=0$, $\forall s' \in S$. In addition, by the definition of marginal density, $\forall j$ s.t. $1 \leq j \leq n$, $f_j(y_j|s)>0$. Since, by assumption, y_j, $j=1,..,n$, are such that none is a perfect signal, then, $\forall j$, $\exists s_j \in S$, $s_j \neq s$, s.t. $f_j(y_j|s_j)>0$. Therefore, $\forall j$, $f_j(y_j|s_j)/f_j(y_j|s)>0$, $h(y_1...y_n|s_j)/h(y_1...y_n|s)=0$. Therefore, by definition, $(y_1...y_n)$ is a fusion. Therefore, h is a fusion.