TOWARDS A COMPLEMENTARITY THEORY FOR EXTRACTING ACCURATE DATA FROM INACCURATE DATABASES THROUGH INTEGRATION

Irit Askira Gelman
University of Arizona
askirai@email.arizona.edu

Abstract: A widely accepted assumption in the database community holds that databases are accurate. However, in reality, errors are widespread. The increasing connectivity of systems introduced new challenges and opportunities as far as data accuracy. The question of interest in this study is how to produce accurate data through integration, given multiple, overlapping, inaccurate databases. The theoretical foundations of current solutions that center on "good" source selection and inconsistency resolution with a similar purpose are limited. These solutions use high-level information about errors, and prefer sources whose total error rate is low. This paper introduces a new approach and theory that can guide to higher accuracy when detailed information about errors is available. The new theory centers on sources in which errors demonstrate complementarity with errors in other sources, such that they make accurate output when integrated. A major assumption of this work, motivated by common perceptions of data custodians and other stakeholders, is that errors are not random. The theory can offer guides for the design of source selection and inconsistency resolution strategies, and may contribute to effective characterization of errors.

1. INTRODUCTION

A widely accepted assumption in the database community holds that data stored in databases are accurate. Unfortunately, in reality, databases are often far from accurate—data errors are widespread [22]. The increasing connectivity of systems and growing popularity of the Internet in recent years introduced new challenges and opportunities as far as data accuracy [3]. These issues have emerged largely due to a new access to numerous, autonomous, data sources.

Consider, for example, a situation in which needed data appear in each of n equally accessible sources—in this case relation instances—which are not entirely accurate. To simplify, assume also that all the relation instances share the same schema and semantics, have equal number of tuples that designate the same real world objects, and all object identification problems have

been resolved. Evidently, autonomous sources do not necessarily exhibit the same errors, therefore errors induce conflicts among related data in different sources. The question of interest in this work is the following: Given a set of relation instances as described, how do we apply data integration to get the highest accuracy, subject, perhaps, to cost constraints?

Various solutions have been proposed to this question, aiming to maximize integration output accuracy through "good" source selection and conflict resolution strategies (e.g., [13,15,16,19]). Current solutions use high-level information about errors, and prefer sources whose total error rate is low [13,15,16,19]. However, the theoretical basis has often been neglected in these solutions. The conditions that guarantee a desired outcome have not been evaluated, and the properties of the desired outcome are not discussed either. This paper initiates a new solution approach and theory. The essence of the new approach is the following: *Instead of preferring sources whose total error rate is relatively low, seek sources in which errors exhibit complementarity with errors in other source(s) such that, when integrated, the sources yield highly accurate output*.

In fact, complementarities between sources have been taking an increasingly important role in strengthening the quality of data through integration. Current data-integration techniques take advantage of complementarities between sources to enhance the *completeness* of answers to queries [21]. Some data integration techniques that increase the *resolution* of answers by exploiting complementarities between sources have been explored as well (e.g., [5]). This paper is focused on the accuracy dimension of data quality. It refers, in particular, to such complementarity that is observed when error rates vary widely within each source so that a subset of the data that is characterized by a high error rate in one source corresponds to data with a low error rate in another source. Accordingly, a major assumption of this work is that errors are

not random. One example involves personal data that people are commonly asked to volunteer, and may often be reluctant to report truthfully (age, education, marital status, dates of significant events in life, email address, etc.). A subset of such data that is of lower significance to one organization can be of great importance to another, so that the latter would apply various measures to guarantee its correctness. This way, if the factors that produce errors are controlled differently in the creation of different sources, then data subsets with a significant amount of errors in one source may correspond to subsets with only few errors in other source(s). Such complementarity can be taken advantage of through integration, *provided it is detected and understood*.

For simplicity, in place of a low error rate, this initial inquiry assumes the existence of error-free subsets, or, more generally, sources that are characterized here by "limited perfect accuracy." The method of study is analytical, employing the Information Structure (IS) model [2,9] to portray error distributions—this model has the advantage that it enables the representation of variations in errors rates. The paper defines two notions of complementarity, and proves the potential value—in terms of output accuracy—of implementing such notions in data integration designs. While these notions are mainly relevant to source prioritization and selection, the issue of combining related data and conflict resolution is addressed as well, through the concept of *fusion*. The purpose of introducing this concept is to link the promised, good properties of the output of integration as far as accuracy, with a method that would guarantee the suggested benefits. Mainly, this concept distinguishes circumstances that call for combining related data from different sources, from those in which there is no need to combine data since an identified source stores equally accurate data. The theory is illustrated through examples, touching on the implications to source selection and conflict resolution strategies. The

paper also discusses implementation and future research issues. 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 in particular. Proofs will be provided upon request.

1.1 Example

The examples in this paper refer to a simple relation, whose schema consists of three attributes: Name, Age, and Occupation. A subset of such relation instance is depicted by Table 1.

Rec #	Name	Age	Occupation	Rec #	Name	Age	Occupation
1	Jim Davis	24	Buss	11	Robert Young	40	Edu
2	Jennifer Duarte	32	Edu	12	David Wood	29	Buss
3	Gerald Gutierrez	43	Edu	13	Raina Wiley	21	Buss
4	Erin Henderson	37	Eng	14	Joice Spitz	25	Buss
5	Tiffany Knuth	46	Buss	15	Daniel Sanders	60	Eng
6	Sam Newell	23	Eng	16	Andrew Richards	62	Edu
7	Leslie Ann Presnell	60	Edu	17	Michael Campbell	45	Edu
8	Daniel Reed	63	Eng	18	Elaine Cook	67	Buss
9	Martin Sawyer	61	Buss	19	Andrea Billings	22	Eng
10	Adele McKinley	51	Edu	20	Bryan Ross	62	Eng

Table 1: A subset of a relation instance.

2. 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. In particular, the rapid proliferation of computer networks has stimulated research interest in the problems of source ranking and selection, and conflict resolution strategies, given the availability of multiple, overlapping

sources that are not error-free (e.g., [1,7,10,11,13,14,15,16,19]). Often, such work has been conducted using the relational database framework.

A common premise of proposed solutions in this research stream is that information about the quality of data is obtainable. Solutions in this category exploit high-level quality estimates. Naumann et al. [15], for example, apply multiple data quality criteria for source selection. Their method associates each source with a set of aggregate quality data, and identifies a subset of "efficient" sources using Data Envelopment Analysis (DEA). Avenali et al. [1] describe an approach for optimizing source selection in terms of the cost of data exchange, given constraints on data quality requirements. The authors assume the availability of multiple, partly overlapping databases in a cooperative information system, and formulate an integer programming model that employs query-level data quality estimates. Naumann et al. [16] address a similar problem where data quality considerations guide source filtering and query plan preferences. Again, quality measures are given by aggregates. The work of Holland [7] is closer to the standpoint of the current paper. Holland proves the potential gain from accounting for variations in error rates, through a case study of source ranking and conflict resolution. He examines three different source-ranking procedures. The procedures vary in their provision for error rate variations within sources. Holland finds that a procedure that accounts better for error rate variations achieves better outcome in terms of integration output accuracy. Similarly, Motro and Rakov advocate the understanding that error rates can vary significantly among different data subsets [12,13,19]. They describe a data analysis method of identifying data subsets that are homogeneous in their soundness (/completeness). Their method produces respective soundness (/completeness) estimates. Motro and Rakov also demonstrate the use of such estimates in conflict resolution [13,19]. Interestingly, their conflict resolution specification does not make direct use of the

detailed estimates. Instead, it is based on aggregates of the estimates—they use query-level aggregates to construct weights on different answers.

Conflict resolution has also been studied extensively by researchers in the database community. Such research does not link inconsistencies with errors (e.g. [4,5,8,20]).

The assumptions that underlie this paper, mainly that error rates can vary significantly across different data subsets, resemble the understanding that has been recommended by Motro and Rakov, and has also been accepted by other researchers. However, unlike other studies, this work highlights the opportunity that such variations provide for increasing data accuracy through integration, assuming autonomous, overlapping sources. Furthermore, the theoretical basis of solutions that aim to increase integration output accuracy through source selection and conflict resolution is often limited. This work takes an initial step towards addressing the existing shortage of theory.

3. BASIC DEFINITIONS

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

A data source, e.g., the values of a single attribute of a relation, a relation instance, or a database view, is represented by a one-dimensional or multidimensional random variable, denoted by Y or Z. Data values are modeled as instances of Y (/ Z). Actual values are represented by a one-dimensional or multidimensional random variable denoted by S. Correct values are instances of 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) [2,9]. An

information structure is a function $f: S \times Y \to \Re^+$ 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, but 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 distributions and rates.

Assuming Y and S that take values in the sets Y and S, respectively, if, for every element *s* of S, $f(y|s)$ is the conditional density of Y given $S=s$, then *f* models the *information that Y provides about S*.

Given *n* random variables, $Y_1,..,Y_n$, if *f* models the information that $Y'=(Y_1,..,Y_n)$ provides about S, we say that *f* models the *integration* information that $Y_1,..,Y_n$, provide about S.

To clarify the notion of an IS, assume that Y is a one-dimensional random variable that corresponds to the values of the occupation attribute in the relation instance described above (Table 1), and S corresponds to the respective correct occupations. The information that Y provides about 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 $Y=\{Buss, Eng, Edu\}$, 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.

4. 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 by a *perfect IS*. A perfect IS is an IS where every signal is a *perfect signal*. A perfect signal points with *certainty* to one state, i.e., the probability of such signal is positive only under that state.

Definition 1: Perfect signal. Let f denote an IS defined over $S \times Y$. A signal $y \in Y$ is a perfect signal of f if, for every s and s' in S such that $s' \neq s$, $f(y|s) > 0$ if and only if $f(y|s') = 0$. If y is a perfect signal of f and $f(y|s) > 0$ then y points to s with certainty.

Definition 2: Perfect IS. Let f denote an IS defined over $S \times Y$. If, for every $y \in Y$, y is a perfect signal of f, 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.

We assume that data sources have errors, and errors are *not* random. In particular, the analysis will focus on conditions in which sources demonstrate characteristics classified here, broadly, as "limited perfect accuracy." Accordingly, any IS where one signal, or more, is perfect, but the IS is not a perfect IS, belongs in the category of "limited perfect accuracy." Definition 3 portrays a special subset of this category: *Perfect IS given a state*. Definition 3 designates a situation in which, even if the source as a whole has errors, the source is accurate in a selected value or range of values. Thus, an IS is perfect given a state if there exists a state such that every signal that has positive probability given that state is a perfect signal. The set of all such states is the *perfect set* of the IS.

Definition 3: Perfect IS given a state. Let f denote an IS defined over $S \times Y$. If there exists $s \in S$ such that, for every $y \in Y$, $f(y|s) > 0$ implies that y is a perfect signal of f, then f is perfect given a state. The set $S_f = \{s \in S |$ for every $y \in Y$, $f(y|s) > 0$ implies that y is a perfect signal of $f\}$ is the perfect set of f.

Evidently, if the perfect set of an IS contains every possible state, the IS is perfect.

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, the perfect set of IS (2) is empty—none of the states satisfies the conditions of Definition 3. In contrast, the perfect set of the IS below (3) is

not empty—it contains the state "Business." The interpretation is that 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

Given two ISs and their matching integration IS, it is easy to show that if a state is a member of the perfect set of any of these ISs, it is also a member of the perfect set of the integration IS. Intuitively, when a source has an error-free subset, then integration with an alternative, possibly inferior, source, should not have negative effect as far as accuracy—especially if the error-free subset is detected.

Lemma 1: Let f, g, denote ISs defined over $S \times Y$, $S \times Z$, respectively. IS f models the information that Y provides about S, g models the information that Z provides about S, and h models the integration information that Y and Z provide about S. Let S_f denote the perfect set of f, and S_h the perfect set of h. Then, $S_h \supseteq S_f$.

The conditions that the concept of a perfect IS given a state stipulates are significantly weaker compared to the requirement on a perfect IS. Nevertheless, the concept of an *IS* that has *perfect distinction between states* implies even weaker conditions. This concept applies the notion of a *weakly perfect signal*, which targets situations in which a value that a source shows is not necessarily error-free, but it reduces the range of possibilities for the true value. A signal is weakly perfect if there exist two states such that the conditional probability of the signal given one state is positive if and only if its conditional probability given the other state is zero. Thus,

the signal rules out exactly one of the two states. Subsequently, an IS has perfect distinction between two states if every signal that has positive conditional probability given one of them has zero probability given the second.

Definition 4: Weakly perfect signal. Let f denote an IS defined over $S \times Y$. A signal $y \in Y$ is a weakly perfect signal of f if there exist s and s' in S such that $f(y|s)>0$ if and only if $f(y|s')=0$. The set $T_{fy} = \{\{s,s'\}| s$ and $s' \in S, f(y|s)>0$ if and only if $f(y|s')=0\}$ is the distinction set of y.

Definition 5: IS has perfect distinction between states. Let f denote an IS defined over $S \times Y$. If there exist s and s' in S such that, for every signal $y \in Y$ such that $f(y|s)>0$, y is a weakly perfect signal of f and $\{s,s'\}$ is a member of the distinction set of y, then f has perfect distinction between states. The set $D_f = \{\{s,s'\}| s$ and $s' \in S$, and for every signal $y \in Y$ such that $f(y|s)>0$, y is a weakly perfect signal of f and $\{s,s'\}$ is a member of the distinction set of $y\}$ is the distinction set of f.

Any IS where one signal, or more, is weakly perfect, but the IS is not a perfect IS, belongs, again, in the category of "limited perfect accuracy." The notion of an IS that has perfect distinction between states delimits a special subset of this category.

If, for a given IS, a state s is such that the distinction set of the IS contains elements for s and any other possible state, then the IS is perfect given a state. Especially, s is a member of the perfect set of the IS. In other words, the notion of a perfect IS given a state is a special case of an IS that has perfect distinction between states.

The following IS (4) is such that the distinction set of the signal "Buss" contains {Business, Engineering} and {Education, Engineering}. 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, the distinction set of IS (4) is empty. In contrast, the distinction set of the IS below (5) is not empty—it contains {Business, Engineering}.

(5)

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

We conclude this section with the assertion that, given two ISs and their respective integration IS, a member of the distinction set of any of the ISs is also a member of the distinction set of the integration IS. In other words, the good property is "inherited" by the integration IS. (The interpretation is parallel to the interpretation of Lemma 1.)

Lemma 2: Let f, g, denote ISs defined over $S \times Y$, $S \times Z$, respectively. Suppose that f models the information that Y provides about S, g models the information that Z provides about S, and h models the integration information that Y and Z provide about S. Let D_f denote the distinction set of f. Let D_h denote the distinction set of h. Then, $D_h \supseteq D_f$.

The definitions in this section outline the error patterns that we examine in this study. Subsequent analysis will center on the integration of overlapping data sources that are not error-free, though sources display limited perfect accuracy consistent with the definitions in this section.

5. INCREASING ACCURACY USING COMPLEMENTARITY RELATIONS

Suppose that none of the available sources is error-free, but some obey one or more of the limited perfect accuracy conditions defined above. This section defines two notions of complementarity—the second of which is a generalization of the first—and proves some good properties, in terms of output accuracy, of the integration information when complementarity conditions hold.

The analysis also addresses a characteristic of the integration information named "fusion." The purpose of introducing this concept is to link the promised, good properties of the output of integration as far as accuracy, with an appropriate method of combining the data (i.e., a method that would guarantee the suggested benefits). The notion of a fusion distinguishes settings in which the output of the integration of two values can be based on just one of them, from settings where integration should involve some synthesis of the two values. This distinction can be useful in guiding conflict resolution, and may have efficiency implications—once the preferred source has been identified, there is 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 that are perfect given a state), followed by a more general analysis in agreement with Definition 5.

In essence, a signal of an integration IS is a fusion if none of the signals that it comprises implies the same likelihood of states, i.e., the signal offers new information.

Definition 6: Fusion. Let f, g, denote two ISs defined over $S \times Y$, $S \times Z$, respectively. IS f models the information that Y provides about S, g models the information that Z provides about S, and h models the integration information that Y and Z provide about S. A signal (y,z) is a fusion if

$h(y,z|s_0)>0$ for some $s_0 \in S$, and there exist s_1, s_2, s_3, and s_4 in S, such that $h(y,z|s_1)/h(y,z|s_2) \neq f(y|s_1)/f(y|s_2)$ and $h(y,z|s_3)/h(y,z|s_4) \neq g(y|s_3)/g(y|s_4)$.

5.1 Complementarity in state

Assume several ISs such that none is perfect, but some have non-empty perfect sets. We say that one IS *complements* another IS *in state* if there exists a state that belongs to the perfect set of the former but not to the perfect set of the latter. The set of all such states is the *perfect complementarity set*. When two ISs complements one another, the ISs are *complementary* in *state*. Definition 7 designates a situation in which a source that is accurate in a certain range of values is integrated with a source that is not error-free in that range.

Definition 7: Complementarity in state. Let f, g, denote ISs defined over $S \times Y$, $S \times Z$, respectively. Let S_f denote the perfect set of f. Let S_g denote the perfect set of g. It is said that f complements g in state if there exists $s \in S$ such that $s \in S_f$, $s \notin S_g$. The set $S_g^f = S_f - S_g$ is the perfect complementarity set of f and g. If f complements g in state and g complements f in state then f and g are complementary in state.

Proposition 1a says next that when an IS complements another IS in state, the perfect set of the latter is a proper subset of the perfect set of the matching integration IS. Proposition 1a implies that when two ISs are complementary in state, their perfect sets are both proper subsets of the perfect set of the integration information. In this sense, the integration information is strictly better, i.e., more accurate, than the information provided by any of the participating sources. In fact, proposition 1a provides a foundation for quantifying such superiority.

Proposition 1a: Let f, g, denote ISs defined over $S \times Y$, $S \times Z$, respectively. IS f models the information that Y provides about S, g models the information that Z provides about S, and h models the integration information that Y and Z provide about S. Let S_f denote the perfect set of f, S_g denotes the perfect set of g, and S_h denotes the perfect set of h. Suppose that f complements g in state, and let S_g^f denote the perfect complementarity set of f and g. Then, $S_h \supseteq S_g \cup S_g^f \supset S_g$.

   Proposition 1a hints that by repeatedly adding sources of this kind the integration information can reach perfect accuracy—or, more generally, can demonstrate any specified perfect set. Proposition 1b refers to such conditions in detail. The perfect set of the integration information includes a given subset if, for any IS and any state in this subset, if the state is not a member of the perfect set then there is another IS that complements it such that the same state is a member of the perfect complementarity set.

Proposition 1b: Let f_j, $j=1,..,n$, denote ISs defined over $S \times Y_j$, respectively, such that, for every j, f_j models the information that Y_j provides about S. Let h denote the integration information that Y_j, $j=1,..,n$, provide about S. Let S_h denote the perfect set of h, and S' is a subset of S. Then, $S_h \supseteq S'$ if, for every j and every $s \in S'$, s does not belong to the perfect set of f_j implies that there exists k, $1 \leq k \leq n$, such that f_k complements f_j in state and s belongs to the perfect complementarity set of f_k and f_j.

   Propositions 1a and 1b suggest the positive potential of complementarities in state, yet they do not make clear what strategy of combining the data would materialize such potential. This issue is addressed by proposition 2, which applies the concept of a fusion. Primarily, this proposition supports the simple understanding that, when a source shows data that are known to be perfect,

there is no need to combine such data with any other data. A signal that points to a state with certainty is just as accurate as a composite signal that comprises it—the composite signal does not offer any new information. Hence, Proposition 2 says that a signal of an integration IS that comprises a signal which points to a state with certainty, is not a fusion.

Proposition 2: Let f, g, denote ISs defined over $S \times Y$, $S \times Z$, respectively. IS f models the information that Y provides about S, g models the information that Z provides about S, and h models the integration information that Y and Z provide about S. If $y \in Y$ is a perfect signal of f, then, for any $z \in Z$, the signal (y,z) is not a fusion.

Next, an example will illustrate the theory in this section, and clarify some implications of these results to source selection and conflict resolution.

Example:

Consider three overlapping sources that show customer names and their occupations. The names are correct and consistent across the sources, but none of the sources is free of errors as far as occupation, and error rates vary within each source. 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 claimed occupation. This source is described by IS (6) below (the word "Name" should be interpreted as a wildcard). The other two sources portrayed by ISs (7) (8) below, are products of comparable verification procedures that are applied to business people.

(8)

Signal /State	Name; Buss	Name; Eng	Name; Edu
Name; Business	1	0	0
Name; Engineering	0	.95	.05
Name; Education	0	.02	.98

(7)

Signal /State	Name; Buss	Name; Eng	Name; Edu
Name; Business	1	0	0
Name; Engineering	0	.96	.04
Name; Education	0	.03	.97

(6)

Signal /State	Name; Buss	Name; Eng	Name; Edu
Name; Business	.83	.17	0
Name; Engineering	.15	.85	0
Name; Education	0	0	1

Since IS (6) complements IS (7) such that the state "Name; Education" is in the respective perfect complementarity set, then, according to Proposition 1a, the integration information that corresponds to IS (6) and IS (7) would have the state "Name; Education" in its perfect set. Furthermore, since IS (7) complements IS (6) such that the state "Name; Business" is in the perfect complementarity set, this state too would be contained in the perfect set of the integration IS. Hence, the perfect set of the integration IS would contains both "Name; Education" and "Name; Business." The integration IS is given by:

(9)

Signal /State	Name; Buss, Name; Buss	Name; Buss, Name; Eng	Name; Buss, Name; Edu	Name; Eng, Name; Buss	Name; Eng, Name; Eng	Name; Eng, Name; Edu	Name; Edu, Name; Buss	Name; Edu, Name; Eng	Name; Edu, Name; Edu
Name; Business	.83	0	0	.17	0	0	0	0	0
Name; Engineering	0	.144	.006	0	.816	.034	0	0	0
Name; Education	0	0	0	0	0	0	0	.03	.97

Although the conditions of Proposition 1b are not met, IS (9) is actually a perfect IS, representing error-free output. (The numbers in the matrix were derived based on an assumption of state-conditional independence. Nonetheless, the conclusion that the integration information is a perfect IS would have been the same, regardless of dependence relationships.) While this finding is not predicted by Proposition 1b, a generalization in a later section will provide the tools to predict it.

Integration of information as in IS (6) and IS (8) would, likewise, yield a perfect IS. But, as for ISs (7) and (8), the conditions of Proposition 1a only guarantee that the integration is perfect given the state "Name; Business." The integration information could be the following IS:

(10)

Signal /State	Name; Buss, Name; Buss	Name; Buss, Name; Eng	Name; Buss, Name; Edu	Name; Eng, Name; Buss	Name; Eng, Name; Eng	Name; Eng, Name; Edu	Name; Edu, Name; Buss	Name; Edu, Name; Eng	Name; Edu, Name; Edu
Name; Business	1	0	0	.0	0	0	0	0	0
Name; Engineering	0	0	0	0	.912	.038	0	.048	.002
Name; Education	0	0	0	0	.0006	.0194	0	.0294	.9506

The states "Name; Engineering," and "Name; Education" are not contained in the perfect set of (10). Although the values in (10) are specifically based on an assumption of state-conditional independence of the sources, no other assumption on the dependence relationship could have resulted in a perfect IS. Integration of a pair of sources that correspond to IS (7) and IS (8) is inferior, in this sense, to integration of sources that correspond to ISs (6) and (7), or (6) and (8).

This example offers an opportunity to illustrate the difference, in source selection preferences, between an approach that is guided by the proposed complementarity notions versus an approach that prefers sources with low aggregate error-rates. Suppose that the number of sources taking part in the integration is to be kept at a minimum. A "traditional" approach that prefers sources with low aggregate error-rates might register that the overall error rate of the source that matches IS (6) is significantly higher than the error rates of the other two sources. (This observation would be true if, for example, the prior probabilities of the states are not too far apart from each other.) Hence, such approach might favor the sources matching ISs (7) and (8) over (6), and limit integration to these two sources. In contrast, the complementarity theory in this paper indicates the clear superiority of any of the pairs (6) and (7), and (6) and (8), over the pair (7) and (8). A

18

choice along this theory would lead to integrating either one of those two pairs, but the pair (7) and (8) would be avoided.

Turn to the fusion property, application of the respective definition (Definition 6) supports the intuition that when data are known to be perfect, there is no need to synthesize the data with data from another source. For example, suppose that we inquire about the occupation of a certain person and we have three sources available, comparable to ISs (6)-(8). Having identified that the integration of IS (6) and IS (7) amounts to a perfect IS, we select the sources that match IS (6) and IS (7) for that purpose. Suppose that the first source (6) answers "Eng" while the second (7) answers "Buss." We can see that there is no need to synthesize the two answers—the second source determines the answer exclusively. IS (9) shows that the signal "Name; Eng, Name; Buss" points to the state "Name; Business" with certainty. Yet, this signal is not a fusion. According to IS (7), "Name; Buss," just like the signal that comprises it, points to "Name; Business" with certainty. Therefore, given the signal "Name; Buss," the signal "Name; Eng" is redundant.

5.2 Generalization: Complementarity in distinction

We now turn to conditions in which none of the ISs is perfect, yet one or more have perfect distinction between states. These conditions form a generalization of the conditions in the previous section. A second notion of complementarity is defined, which assists in portraying the good properties, in terms of output accuracy, of the integration information. Unlike before, integration of complementary sources typically involves a fusion, i.e., in order to materialize the promised gains there is typically a need to synthesize data from multiple sources.

We say that one IS *complements* another IS in *distinction* if there exist two states such that a subset that contains them belongs to distinction set of the former, but not to the distinction set of the latter. The set of all such subsets is called *distinction complementarity set*. When each of two ISs complements the other in distinction, they are *complementary* in *distinction*.

Definition 8: Complementarity in distinction. Let f, g, denote ISs defined over $S \times Y$, $S \times Z$, respectively. Let D_f denote the distinction set of f. Let D_g denote the distinction set of g. It is said that f complements g in distinction if there exist s and s' in S such that $\{s,s'\} \in D_f$, and $\{s,s'\} \notin D_g$. The set $D_g^f = D_f - D_g$ is the distinction complementarity set of f and g. If f complements g in distinction and g complements f in distinction then f and g are complementary in distinction.

Proposition 3a says that when an IS complements another IS in distinction, the distinction set of the latter is a proper subset of the distinction set of the matching integration IS. Proposition 3a implies that when two ISs are complementary in distinction, their distinction sets are both proper subsets of the distinction set of the integration information. In this sense, again, the integration information is strictly better than the information provided by any of the participating sources—Proposition 3a also provides a basis for quantifying such superiority.

Proposition 3a: Let f, g, denote ISs defined over $S \times Y$, $S \times Z$, respectively. IS f models the information that Y provides about S, g models the information that Z provides about S, and h models the integration information that Y and Z provide about S. Let D_f denote the distinction set of f, D_g the distinction set of g, and D_h the distinction set of h. Suppose that f complements g in distinction, and let D_g^f denote the distinction complementarity set of f and g. Then, $D_h \supseteq D_g \cup D_g^f \supset D_g$.

Proposition 3a suggests also that by repeatedly adding sources of this kind the integration information can reach perfect accuracy. Proposition 3b centers on this issue directly—it is a generalization of Proposition 1b. The distinction set of the integration information includes a given distinction set if, for any IS and any element of the latter set, if such element does not belong to the distinction set of the IS, there is another IS that complements it such that the element is a member of the distinction complementarity set.

Proposition 3b: Let f_j, $j=1,..,n$, denote ISs defined over $S \times Y_j$, respectively, such that, for every j, f_j models the information that Y_j provides about S. Let h denote the integration information that Y_j, $j=1,..,n$, provide about S. Let D_h denote the distinction set of h, and suppose that D' is such that $D' \subseteq \{\{s, s'\} \mid s$ and $s' \in S$, $s' \neq s\}$. Then, $D_h \supseteq D'$ if, for every j and every $\{s, s'\} \in D'$, $\{s, s'\}$ does not belong to the distinction set of f_j implies that there exists k, $1 \leq k \leq n$, such that f_k complements f_j in distinction and $\{s, s'\}$ belongs to the distinction complementarity set of f_k and f_j.

Turn, again, to the fusion property. Proposition 4 supports the intuitive understanding that, for better output accuracy, when data from one source rule out a subset of the possible values while data from another source rule out a different subset, the data should be combined. When two signals, each associated with a different source, distinguish between states such that the respective distinction sets are not included within each other, then a composite signal comprising the described signals is a fusion.

Proposition 4: Let f, g, denote ISs defined over $S \times Y$, $S \times Z$, respectively. IS f models the information that Y provides about S, g models the information that Z provides about S, and h

models the integration information that Y and Z provide about S. Suppose that y is a weakly perfect signal of f and z is a weakly perfect signal of g. Let T_{fy}, T_{gz}, denote the distinction sets of y, z, respectively. If $h(y,z|s_0)>0$ for some $s_0 \in S$, $T_{fy} \not\subset T_{gz}$, and $T_{gz} \not\subset T_{fy}$, then (y,z) is a fusion.

Example:

The previous example will now be extended and linked to the understanding about complementarity in distinction. ISs (6), (7), and (8) were formerly examined with regard to complementarity in state. However, a closer study of each of the pairs (6) and (7), and (6) and (8), shows that wherever the ISs are not complementary in state, they are complementary in distinction. First, take the pair IS (6) and IS (7). IS (6) complements IS (7) in distinction, such that {Name; Education, Name; Engineering} belongs to the associated complementarity set. IS (7), in return, complements IS (6) in distinction, such that {Name; Business, Name;Engineering} is in the relevant complementarity set. Proposition 3b can be applied to show that the integration information of IS (6) and IS (7) is a perfect IS, by defining D' to contain every possible pair of different states. There are three such pairs: {Name;Business, Name;Education}, {Name; Business, Name; Engineering}, and {Name; Engineering, Name; Education}. Proposition 3b indicates that the distinction set of the integration information includes D' in this example.

In the same way, IS (6) complements IS (8) in distinction, such that {Name;Education, Name; Engineering} is in the complementarity set, while IS (8) complements IS (6) in distinction, such that {Name; Business, Name; Engineering} is in the complementarity set. These ISs too satisfy the requirements of Proposition 3b such that the integration information is a perfect IS.

Signals of IS (9) that point to "Name; Engineering" with certainty are all fusions. For example, the signal "Name; Buss, Name; Eng" points to "Name; Engineering" with certainty, although the

signal "Name; Buss" is not a perfect signal of IS (6), and "Name; Eng" is not a perfect signal of IS (7). When the signal "Name; Buss" is received from a source that matches IS (6), the possibility that the state is "Name; Education" is ruled out. When the signal "Name; Eng" is received from a source like (7), the possibility that the state is "Name; Business" is ruled out as well. Together, the two signals determine, through repeated elimination, that the state is actually "Name; Engineering."

6. APPLICABILITY AND GENERALIZABILITY

Generalization of the theory

Error patterns: The definitions of complementarity that this paper introduces do not cover the varied potential in practical application settings. This way, for example, these definitions refer to states, instead of signals. Therefore, for instance, a source that matches IS (2), which includes a data subset that is perfectly accurate, is not addressed by the theory at this stage.

In addition, there is a need to develop comparable notions of complementarity based on weaker assumptions about error patterns, which will apply to settings in which sources demonstrate *less than perfect* distinction between states. This generalization will retain a similar focus on error rates that 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 another source.

A more complete understanding of data integration accuracy could also involve the introduction of notions of complementarity related to dependence between specific errors, e.g., dependence between error directions or error sizes. To clarify this argument, here is an example:

Two sources show basic demographic data of members of a selected group. The origins of the data are self-reports volunteered by the members. Data in the first source are collected in the context of job seeking efforts, while in the second source data are collected in social interaction circumstances. Age data form part of the demographic 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. The error patterns reveal that young people tend to inflate their age in job seeking contexts, and the same people often take years off their age in social circumstances. In other words, there is a negative correlation between respective errors in the two sources. On the other hand, older people often take years off in both cases—the correlation is positive. Knowledge about dependence relationships of this kind can be useful for decision-making about source selection and conflict resolution. Dependence between errors may form a basis for additional notions of complementarity that can contribute to higher integration output accuracy.

Other data quality dimensions: The idea of complementarity could be useful in strengthening additional quality dimensions through integration. The theory in this paper can be easily extended to apply to the dimensions of completeness and resolution. However, the idea of complementarity might benefit additional data quality dimensions, such as, perhaps, time-related dimensions.

Validity of complementarity assumptions

The perception regarding the possible commonness of complementarities has been implied by various data quality researchers. However, to the best knowledge of this author, direct evidence is limited. Ultimately, the prevalence of complementarities as described here should be assessed based on empirical evidence. The answer to this question can vary from one setting to another.

Information requirements

Obviously, the reliance of this theory on information about error distributions in different data subsets makes implementations sensitive to the state of understanding in this domain. This is a broad issue, with various open questions [3,17]. For the most part, current methods have used human beings as origins of relevant information. However, the potential for automated analyses has been recognized [17], and some actual developments in this direction exist, such as the data analysis method of Motro and Rakov ([12]) which is based on a sample of clean data.

The required depth of such analysis can vary. Below is an example that emphasizes this issue.

Suppose that a customer's declared occupation is checked if the customer associates himself or herself with the business sector. However, although verification is strict for customers in the mid-age group, it is not as strict for people in the young age group—up to 26 years old—and seniors—above 55 years old. (Members of the latter age groups may be offered age-specific bonuses such that 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:

(11)

Signal /State	Buss, age< 26	Buss, 26≤age≤55	Buss, age> 55	Eng, age< 26	Eng, 26≤age≤55	Eng, age> 55	Edu, age< 26	Edu, 26≤age≤55	Edu, age> 55
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

IS (11) indicates that, whenever the data report that a customer's occupation is "Buss" and he or she is in the middle age group, the data are free of errors (the corresponding signal points to the state "Business" with certainty.) 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:

(12)

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

In this IS (12), none of the signals is perfect.

Algorithms

Algorithms that implement this theory will vary depending on the characteristics of the specific application. As an example, suppose that the conditions of Proposition 3b are satisfied by a set of data sources, where the state set is finite (e.g., the domain of a relevant attribute is finite). Suppose also that there is overlap between different sources so that more than one source can have perfect distinction between two given states. In addition, each source is associated with a typical access cost. We want to find a subset of the candidate sources that, when integrated, will produce perfect information given a chosen state with minimal total access cost. This problem can be formalized by an integer programming set covering model. An algorithm for optimal solution of such model is exponential, however, the set covering problem has an efficient heuristic algorithm with a performance guarantee [18].

7. CONCLUSIONS

The question of interest in this study is how to produce accurate data through integration, given multiple, overlapping, inaccurate sources, subject, perhaps, to cost constraints. 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 integration. This is the major insight of this research. This paper refers, in particular, to sources that have what is broadly described as "limited perfect accuracy." The analysis demonstrates intuitive points. Mainly, when sources that have errors as above are also complementary, namely, when error-free subsets vary in different sources, or different sources rule out different possible values, their integration will have higher accuracy. Under best conditions, the outcome will even reach perfect accuracy. We have also developed a basis for quantitative evaluations of different integration alternatives. The new theory suggests the potential value of a data integration approach that is guided by notions of complementarity.

The examples touch on implications of the theory to source selection and conflict resolution. Especially, we have compared a strategy that obeys complementarity with a strategy that prefers sources with low aggregate error-rates. This comparison hints to the potential superiority of a strategy that complies with complementarity. We believe that a complementarity theory can offer a guide for the effective use of detailed information about errors, when such information is available. Strategies that prefer low total error-rates will apparently be at a disadvantage in these circumstances.

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 validity of assumptions about errors, 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] Dagstuhl Seminar No. 06332 report, "Data Quality on the Web." Schloss Dagstuhl International Conference And Research Center For Computer Science, 2003.

[4] Dayal, U., and Hwang, H.Y., "View Definition and Generalization for Database Integration in a Mulidatabase System, *IEEE Transactions on Software Engineering*, Vol. 10, No. 6, 1984, pp. 628-644.

[5] 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.

[6] Galhardas, H., Florescu, D., Shasha, D., and Simon, E., "An Extensible Framework for Data Cleaning." 16th *International Conference on Data Engineering*, ICDE 2000, 2000.

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

[8] Lim, E-P., Srivastava, J., Shekhar, S., An Evidential Reasoning Approach to Attribute Value Conflict Resolution in Database Integration." *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8 , No. 5, 1996, PP. 707 - 723.

[9] 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.

[10] Motro, A., and Anokhin, P., "Use of Meta-data for Value-level Inconsistency Detection and Resolution During Data Integration." 5th *World Multi-Conference on Systemics, Cybernetics and Informatics*, SCI 01, Orlando, Florida, 2001.

[11] 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.

[12] 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.

[13] Motro, A., and Rakov, I. "Estimating the Quality of Databases." In *Proceedings of FQAS 98, Third International Conference on Flexible Query Answering Systems* (T. Andreasen, H. Christiansen, and H.L. Larsen, Editors), Roskilde, Denmark, May 1998. Lecture Notes in Artificial Intelligence No. 1495, Springer-Verlag, pp. 298-307.

[14] Naumann, F., Haussler, M., "Declarative Data Merging with Conflict Resolusion." 7th *International Conference on Information Quality*, ICIQ 2002, Cambridge, Ma.

[15] Naumann, F., Freytag, J.C., and Spiliopoulou, M., "Quality Driven Source Selection Using Data Envelopment Analysis." In 3th *International Conference on Information Quality*, ICIQ 1998, Cambridge, Ma.

[16] 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*, VLDB 99, Edinburgh, U.K., 1999.

[17] Naumann, F., and Rolker, C. "Assessment Methods for Information Quality Criteria." Technical report 138, Humboldt-Universitat zu Berlin, Institut fur Informatik, 2000.

[18] 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.

[19] Rakov, I., "Quality of Information in Relational Databases and its Use for Reconciling Inconsistent Answers in Multidatabases." 4th *Doctoral Consortium on Advanced Information Systems Engineering*, 1997

[20]Tseng, F. S-C., Chen, A.L.P., Yang W-P. "A Probabilistic Approach to Query Processing in Heterogeneous Database Systems." *2nd International Workshop on Research Issues on Data Engineering: Transaction and Query Processing*, RIDE-TQP'92, 1992, pp. 176-183

[21] Ullman, J.D. "Information Integration Using Logical Views." *6th International Conference on Database Theory, ICDT97*, 1997, pp. 19-40.

[22] Wang, R.Y., Ziad, M., and Lee, Y. W. *Data Quality*, Kluwer Academic Publishers, 2001.