Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2007 Proceedings

Americas Conference on Information Systems (AMCIS)

12-31-2007

GIGO or not GIGO: Error Propagation in Basic Information Processing Operations

Irit Askira Gelman
University of Arizona

GIGO OR NOT GIGO: ERROR PROPAGATION IN BASIC INFORMATION PROCESSING OPERATIONS

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

Abstract

This work analyzes the sign of the relationship between input accuracy and output accuracy in two basic information processing operations – the Boolean binary logical OR and logical AND. These operations are often used in the course of decision-making and problem solving tasks. The analysis shows a surprising result: the sign of the relationship varies. Conditions that determine the sign are specified, and those under which the association is negative are explained and illustrated.

Keywords: *Data Management, Data Accuracy, GIGO.*

Introduction

The relationship between input accuracy and output accuracy is of great interest in numerous problem domains, and has been investigated—under different assumptions and titles—in various research areas. Some research areas are computer science, statistics, political science, econometric forecasting, physical sciences, and information systems. Nonetheless, our understanding of that relationship is only partial at present. In fact, even the sign of the relationship is not well understood. Many researchers have embraced the belief in GIGO (Garbage In, Garbage Out), and have largely treated GIGO as an axiom. Originally coined in the computer industry, this acronym, which indicates a strong positive link between input accuracy and output accuracy, is nowadays popular in general. However, there is a growing literature that hints to a more complex association.

One example emerges from a theory, established in several domains, that statistical dependence relationships among data sources, or data errors, can have a dramatic effect on the accuracy of the information that an integration process produces (e.g., Barabash 1965; Frantsuz 1967; Toussaint 1971; Cover 1974; Clemen and Winkler 1985; Berg 1993; Ladha 1995; Askira Gelman 2004; Kuncheva et al. 2003). In some cases, negative correlation between data sources or data errors has a remarkable positive effect on the accuracy of the output information (e.g., Clemen and Winkler 1985; Berg 1993; Ladha 1995; Kuncheva et al. 2003). Consequently, higher data accuracy can lead to higher, or lower, output accuracy, subject to variations in such dependencies. A second example is based on studies of prediction model-building paradigms, which indicate that adding noise to a data sample that serves in the construction of a model can improve the accuracy of the model (e.g., Bishop 1995; Raviv and Intrator 1996; Skurichina et al. 2000). Evidently, controlled levels of noise can compensate for limitations of the model-building algorithms. That is, information-processing optimality seems to be a factor that can affect the sign of the link between input accuracy and output accuracy.

This work addresses the question of the sign of the association between input accuracy and output accuracy. It is a theoretical investigation, part of a research project that aims to expand our understanding of the effect of errors in fundamental information processing operations. Two basic information processing operations are examined in this paper: the Boolean binary logical OR, and logical AND. These operations are often used in the course of decision-making (Einhorn 1970). The scenario assumed here is simple—an operation applies two inputs, both of which are not free of errors. The correct input values as well as error occurrences are random; there are no dependencies. The relationship between input accuracy and output accuracy is interpreted as a relationship between the probability of input error occurrence and the probability of output error occurrence, and analyzed using statistical properties of random variables. The analysis produces a surprising result: the sign of the relationship varies.

A description of the method and notation follows next. Later, the conditions that determine the sign are specified, and conditions in which the association is negative are explained. A subsequent section offers illustrations of conditions in which the association is negative.

Method and notation

Our analysis is based on statistical properties of random variables. The measure of accuracy that the analysis employs is probability of error occurrence (which is the same, in this instance, as the expected value of the corresponding variable).

The variables in use by this analysis are listed and defined below:

- U, V: The ideal, correct input; U and V are dichotomous random variables that accept the values 1 and 0, which correspond to *true* and *false*, respectively.
- W: The desired, correct output; W is a dichotomous random variable that accepts the values 1 (*true*) and 0 (*false*).
- U_a, V_a: The available, possibly incorrect input; U_a and V_a are dichotomous random variables that accept the values 1 (*true*) and 0 (*false*).
- D_U, D_V: The occurrence of an input error as reflected by the value of the available input. D_U and D_V are dichotomous random variables that accept the values 1 and 0, which correspond to *error* and *no error*, respectively.
- W_a: The output that is generated based on the available input; W_a is a dichotomous random variable that accepts the values 1 (*true*) and 0 (*false*).
- D_W: The occurrence of an output error as reflected by the value of the available output; D_W is a dichotomous random variable that accepts the values 1 (*error*) and 0 (*no error*).

Statistical parameters:

- p_U , p_V , p_{D_U} , p_{D_V} , p_{D_W} : Expected values; subscripts identify the relevant random variables. For example, the expected value of U is denoted by p_U , i.e., $p_U = E(U) = \Pr(U = 1)$. Note that the expected value of a random variable that represents the occurrence of an error is the same as the probability of occurrence of that error.
- σ_U , σ_V , σ_{D_U} , σ_{D_V} : Standard deviations; subscripts identify the relevant random variables. Additional notations, including $Stdev(UV)$ and similar notations, represent the standard deviations of the products of selected random variables (here, the standard deviation of the product of U and V).
- ρ_{UV} , $\rho_{D_U D_V}$, ρ_{UD_V} , ρ_{VD_U} , ρ_{VD_V} , ρ_{UD_U} : Correlation coefficients; subscripts identify the relevant random variables. Additional notations, including $Corr(UV, D_U D_V)$ and comparable notations, correspond to correlation coefficients involving products of random variables (in this case, the correlation coefficient between the product of U and V and the product of D_U and D_V).

The relationship among U_a, D_U, and U is given by:

$$U_a = (1 - D_U)U + D_U(1 - U) = U + D_U - 2UD_U \qquad (1)$$

If the value of D_U is zero, that is, if this variable indicates that no error has occurred, then (1) is reduced to U_a=U. However, if the value of D_U indicates an error, then (1) assigns a value of one to U_a if U is zero and a value of zero if U is one. An equivalent relationship exists among V_a, D_V, and V, and among W_a, D_W, and W:

$$V_a = (1 - D_V)V + D_V(1 - V) = V + D_V - 2VD_V \qquad (2)$$

$$W_a = (1 - D_W)W + D_W(1 - W) = W + D_W - 2WD_W \qquad (3)$$

Mean values are constrained:

$$0< p_{D_U} <0.5,\ 0< p_{D_V} <0.5,\ 0< p_U <1,\ 0< p_V <1 \tag{4}$$

These constraints are mostly natural. However, the assumption that both p_{D_U} and p_{D_V} are strictly positive, namely, that both inputs have errors, is crucial to the outcome of this model. The significance of this assumption will be clarified in the following sections.

Logical disjunction (OR)

An error-free disjunction operation is portrayed by:

$$W = U + V - UV \tag{5}$$

The consistency of (5) with the definition of logical disjunction can be easily verified through a systematic evaluation of W for each possible combination of the values of U and V.

Similar to real-world settings, the relationship among the actual output, W_a, and the actual inputs, U_a and V_a, is the same as the relationship among the correct output and inputs:

$$W_a = U_a + V_a - U_a V_a \tag{6}$$

Using (1)-(6), the connection between the probability of an output error, and statistical properties of the correct input and error terms, is described by Lemma 1. Lemma 1 asserts that the probability of an output error is equal to an aggregate of the expected values of D_U, D_V, VD_U (i.e., the product of V and D_U), UD_V, D_UD_V, and UVD_UD_V. In this aggregate, the sign of the expected value of D_U, D_V, and UVD_UD_V is positive, while the remaining elements are negative.

Lemma 1: Assuming (1)-(6):

$$p_{D_W} = E(D_U) + E(D_V) - E(UD_V) - E(VD_U) - E(D_U D_V) + 2E(UVD_U D_V) \tag{7}$$

A quick glance at (7) reveals that the outcome of an input error varies depending on both the correct values and error occurrence in the opposite input. We will use Lemma 1 to examine the direction of the effect of higher input accuracy on output accuracy. We first re-express (7) as a function of the input error probability p_{D_V} (although (8) focuses on p_{D_V}, an analog function applies to p_{D_U} due to a symmetry of the inputs):

$$\begin{aligned} p_{D_W} = {} & p_{D_V} + p_{D_U} - [\rho_{UD_V}\sigma_U (p_{D_V}(1-p_{D_V}))^{1/2} + p_U p_{D_V}] - E(VD_U) - [\rho_{D_UD_V}\sigma_{D_U}(p_{D_V}(1-p_{D_V}))^{1/2} + p_{D_U}p_{D_V}] \\ & + 2Corr(UV, D_U D_V) Stdev(UV)[(\rho_{D_UD_V}\sigma_{D_U}(p_{D_V}(1-p_{D_V}))^{1/2} + p_{D_U}p_{D_V})(1-(\rho_{D_UD_V}\sigma_{D_U}(p_{D_V}(1-p_{D_V}))^{1/2} + p_{D_U}p_{D_V}))]^{1/2} \\ & + 2E(UV)[\rho_{D_UD_V}\sigma_{D_U}(p_{D_V}(1-p_{D_V}))^{1/2} + p_{D_U}p_{D_V}] \end{aligned} \tag{8}$$

For the sake of simplicity, this work assumes that none of the random variables is involved in any dependence, such that all the correlation coefficients in (8) are zero. Notably, the main findings of this research can be produced under considerably weaker assumptions, such that the independence assumption is not critical.

Independence Assumption: None of the variables in $\{U, V, D_U, D_V\}$ or products of such variables is statistically dependent on any other variable in $\{U, V, D_U, D_V\}$ or any product of such variables.

The partial derivative of (8) with respect to p_{D_V} under this assumption is:

$$\partial p_{D_W} / \partial p_{D_V} = 1 - p_U - p_{D_U} + 2p_U p_V p_{D_U} = 1 - p_U(1 - 2p_V p_{D_U}) - p_{D_U} \tag{9}$$

The direction of the link between input error probability and output error probability is determined by the sign of (9). A positive sign of such derivative implies a positive link, while a negative sign indicates a negative link. Proposition 1 applies (9) for addressing the sign of the effect of input error probability on output error probability.

Proposition 1: A higher value of p_{D_V} implies higher value of p_{D_W} if and only if:

$$p_U + p_{D_U}(1-2p_U p_V) < 1 \tag{10}$$

Surprisingly, (10) does not necessarily hold. Conditions in which (10) does not hold are illustrated in a later section. Roughly, when U has a low probability of a zero value and V has a high probability of a zero value, then an increase in input error probability p_{D_V} can produce lower output error probability. This is specifically true if the actual input U_a is not highly accurate. Intuitively, if both inputs have errors when the means of the correct values of the inputs are unequal enough, then errors in the data source with the low mean have the role of "good errors." That is, they offset the "bad errors" in the other data source. Therefore, a higher error rate in the data source with the low mean can actually enhance output accuracy.

Logical Conjunction (AND)

In the case of logical conjunction, the ideal logical conjunction operation—where inputs are error-free—is captured by:

$$W = UV \tag{11}$$

The consistency of (11) with the definition of logical conjunction can be verified through a systematic evaluation of W for each possible combination of the values of U and V. The relationship between the actual input U_a and the ideal input U is described by (1). Similarly, the relationship between the actual input V_a and the ideal input V is described by (2). The relationship among the actual output W_a, and the actual inputs, is the same as the relationship among the correct output and inputs:

$$W_a = U_a V_a \tag{12}$$

The relationship between the actual output W_a and the ideal output W is specified by (3). We assume, again, certain constraints on mean values (4).

Using (1)-(4), (11), and (12), the link between the probability of output error and statistical properties of the correct input and respective error terms is described by Lemma 2. Lemma 2 asserts that the probability of an output error is equal to an aggregate of the expected values of the products VD_U, UD_V, D_UD_V, UD_UD_V, VD_UD_V, and UVD_UD_V. In this aggregate, the sign of the expected value of VD_U, UD_V, D_UD_V , and UVD_UD_V, is positive, and the remaining terms are negative.

Lemma 2: Assuming (1)-(4), (11), and (12):

$$p_{D_W} = E(VD_U) + E(UD_V) + E(D_U D_V) - 2E(UD_U D_V) - 2E(VD_U D_V) + 2E(UVD_U D_V) \tag{13}$$

We see again that the outcome of input errors varies depending on the correct values and error occurrence in the opposite input. For studying the relationship between input accuracy and output accuracy, we re-express (13) as a function of the input error probability p_{D_V} :

$$\begin{aligned} p_{D_W} = {} & [\rho_{UD_V}\sigma_U (p_{D_V}(1-p_{D_V}))^{1/2} + p_U p_{D_V}] + E(VD_U) + [\rho_{D_U D_V}\sigma_{D_U}(p_{D_V}(1-p_{D_V}))^{1/2} + p_{D_U} p_{D_V}] \\ & - 2E(UD_U)p_{D_V} - 2Corr(UD_U, D_V)Stdev(UD_U)(p_{D_V}(1-p_{D_V}))^{1/2} - 2E(VD_U)p_{D_V} \\ & - 2Corr(VD_U, D_V)Stdev(VD_U)(p_{D_V}(1-p_{D_V}))^{1/2} + 2Corr(UV, D_U D_V)Stdev(UV) \\ & \times [(\rho_{D_U D_V}\sigma_{D_U}(p_{D_V}(1-p_{D_V}))^{1/2} + p_{D_U} p_{D_V})(1-(\rho_{D_U D_V}\sigma_{D_U}(p_{D_V}(1-p_{D_V}))^{1/2} + p_{D_U} p_{D_V}))]^{1/2} \\ & + 2E(UV)[\rho_{D_U D_V}\sigma_{D_U}(p_{D_V}(1-p_{D_V}))^{1/2} + p_{D_U} p_{D_V}] \end{aligned} \tag{14}$$

The sign of the link between the input error probability and the output error probability is determined by the sign of the partial derivative of (14) with respect to p_{D_V} . Such derivative is calculated assuming statistical independence, as before:

$$\partial p_{D_W} / \partial p_{D_V} = p_U + p_{D_U} - 2p_{D_U}(p_U + p_V - p_U p_V) \tag{15}$$

Next, Proposition 2 applies (15) for addressing the sign of the effect of input error probability on output error probability.

Proposition 2: A higher value of p_{D_V} implies higher value of p_{D_W} if and only if:

$$p_U + p_{D_U}(1 - 2p_U - 2p_V + 2p_U p_V) > 0 \tag{16}$$

Analog to Proposition 1, Proposition 2 hints that higher input accuracy may or may not produce higher output accuracy. Conditions in which (16) does not hold are illustrated next. Similar to logical disjunction, we will see that a higher input error probability can produce lower output error probability when the means of the correct values of the data sources are unequal enough. However, contrary to logical disjunction, errors are constructive when the mean of the correct data is high. Technically, the derivative (15) can be negative if p_U is low enough and p_V is high enough. In this case, errors in the data source with the high mean may play the role of "good errors" by offsetting the "bad errors" in the other data source.

Illustration

Decision scenarios that entail dichotomous decision criteria in which the means of the correct values demonstrate great inequality are, in fact, common. We first consider an organizational decision in that class that uses logical disjunction. A decision of that kind generally aims to affect a large section of some target population. For that purpose, it applies a criterion for inclusion that is widely satisfied in the target population. Mainly, such criterion is accompanied by another criterion that ensures that a chosen subset of the members that are left out in this way is included as well. A specific example is an operational decision regarding a periodic machine checkup that applies the following criteria to determine if a given machine should undergo this preventative checkup: (a) the machine is at least two years old, and (b) the machine has been exceptionally heavily utilized since any preceding checkup (e.g., more than *n* hours). A machine is serviced if any of these conditions is met. Consider an organization where most of the machines have been bought a few years ago such that (a) is true for a large majority of the machines in use. In addition, by definition (b) is only true for a small number of the machines. Now, suppose that U is derived from (a) such that U=1 corresponds to a machine that is at least two years old and U=0 otherwise. Likewise, V is derived from (b) such that V=1 corresponds to a machine that has been heavily utilized and V=0 otherwise. Clearly, the means of these variables, p_U and p_V, can differ greatly. We will assume that p_U =0.95 and p_V =0.05.

To illustrate conditions in which higher input error probability produces lower output probability under this scenario, we assume that the available data are inflicted with errors. In addition, faithful to the independence assumption of this paper, none of the variables or variable products is statistically dependent on any of the other variables or their products. Suppose, for example, that the probability of error in the machine age criterion, p_{D_U}, is 0.12. The effect of increasing the probability of error in the utilization criterion, p_{D_V}, from 0.01 to 0.49 is demonstrated in Figure 1. Note that the increase in error rate occurs in the input with the lower mean of correct values.

Since inequality (10) does not hold under the outlined assumptions ($\partial p_{D_W} / \partial p_{D_V}$ = -0.0586), our analysis directs that higher values of p_{D_V} do not produce higher values of p_{D_W}. Figure 1 portrays p_{D_W} as a function of p_{D_V}. The values were computed based on equation (7). Essentially, as input error probability grows from 0.01 to 0.49, output error probability decreases from 0.1134 to 0.0853.

Conditions in which higher input error probability produces lower error output probability under logical conjunction are not hard to depict either. The suitable decision targets a relatively small section of some population and applies a criterion for inclusion that is not commonly satisfied in the target population. However, in addition to such unique criterion, it applies a broad criterion in order to ensure that some basic requirement is met. A specific example that remains in the organizational setting described earlier is a decision regarding a more rare, and possibly more costly service in which the criteria to determine if a given machine should undergo the service or not are (a) and (b) as above. A machine is serviced only if both conditions are met.

Again, p_U =0.95 and p_V =0.05. However, this time the probability of error in the data about machine utilization will be fixed at p_{D_V} =0.12, and we will examine the effect of increasing the probability of error in the data about the machine age, p_{D_U}, from 0.01 to 0.49. Notably, the increase in the error rate occurs in the source with the higher mean of correct values.

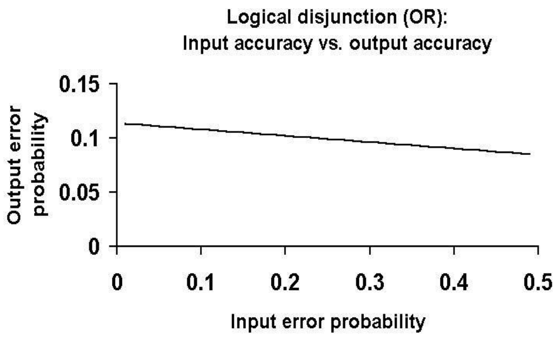

Figure 1. **p_{D_W} as a function of p_{D_V} under the logical disjunction operation**

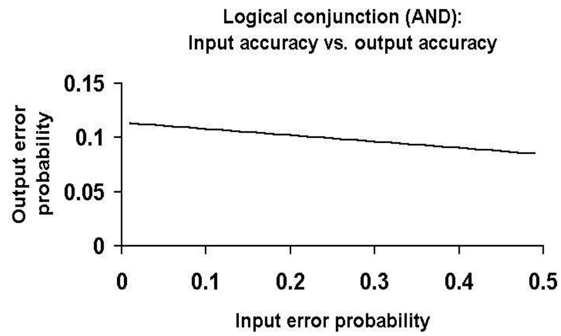

Figure 2. **p_{D_W} as a function of p_{D_U} under the logical conjunction operation**

Table 1. **The decrease in p_{D_W} as p_{D_V} increases under logical disjunction (a sample)**

p_U	p_V	p_{D_U}	$\frac{\partial p_{D_W}}{\partial p_{D_V}}$	p_U	p_V	p_{D_U}	$\frac{\partial p_{D_W}}{\partial p_{D_V}}$	p_U	p_V	p_{D_U}	$\frac{\partial p_{D_W}}{\partial p_{D_V}}$	p_U	p_V	p_{D_U}	$\frac{\partial p_{D_W}}{\partial p_{D_V}}$
0.99	0.01	0.02	-0.01	0.99	0.33	0.21	-0.06	0.91	0.17	0.33	-0.14	0.83	0.33	0.49	-0.05
		0.11	-0.10			0.29	-0.09			0.41	-0.19		0.38	0.49	-0.01
		0.17	-0.16			0.37	-0.12			0.49	-0.25	0.75	0.01	0.26	-0.01
		0.25	-0.24			0.49	-0.16		0.25	0.18	-0.01			0.33	-0.08
		0.33	-0.31		0.41	0.09	-0.01			0.25	-0.05			0.41	-0.15
		0.41	-0.39			0.15	-0.02			0.33	-0.09			0.49	-0.23
		0.49	-0.47			0.21	-0.03			0.41	-0.13		0.09	0.3	-0.01
	0.09	0.03	-0.01			0.29	-0.04			0.49	-0.18			0.37	-0.07
		0.09	-0.06			0.37	-0.06		0.33	0.25	-0.01			0.49	-0.17
		0.15	-0.11			0.49	-0.08			0.37	-0.06		0.17	0.35	-0.01
		0.21	-0.16		0.48	0.09	-0.01			0.49	-0.11			0.41	-0.06
		0.29	-0.23			0.15	-0.01		0.41	0.41	-0.01			0.49	-0.12
		0.37	-0.29			0.21	-0.01			0.49	-0.03		0.25	0.41	-0.01
		0.49	-0.39			0.29	-0.01	0.83	0.01	0.18	-0.01			0.49	-0.06
	0.17	0.03	-0.01			0.37	-0.01			0.21	-0.04		0.33	0.49	-0.01
		0.09	-0.05			0.49	-0.01			0.29	-0.12	0.67	0.01	0.34	-0.01
		0.15	-0.09	0.91	0.01	0.10	-0.01			0.37	-0.19			0.41	-0.07
		0.21	-0.13			0.17	-0.08			0.49	-0.31			0.49	-0.15
		0.29	-0.18			0.25	-0.16		0.09	0.21	-0.01		0.09	0.39	-0.01
		0.37	-0.24			0.33	-0.23			0.29	-0.08			0.49	-0.10
		0.49	-0.32			0.41	-0.31			0.37	-0.14		0.17	0.44	-0.01
	0.25	0.03	-0.01			0.49	-0.39			0.49	-0.25			0.49	-0.05
		0.09	-0.04		0.09	0.12	-0.01		0.17	0.25	-0.01		0.23	0.49	-0.01
		0.15	-0.07			0.17	-0.05			0.31	-0.05	0.55	0.01	0.47	-0.01
		0.21	-0.10			0.25	-0.12			0.37	-0.10			0.49	-0.03
		0.29	-0.14			0.33	-0.19			0.49	-0.18		0.03	0.48	-0.01
		0.37	-0.18			0.41	-0.25		0.25	0.31	-0.01			0.49	-0.02
		0.49	-0.24			0.49	-0.32			0.37	-0.05		0.05	0.49	-0.01
	0.33	0.07	-0.01		0.17	0.15	-0.01			0.49	-0.12		0.06	0.49	-0.01
		0.15	-0.04			0.25	-0.08		0.33	0.39	-0.01	0.53	0.01	0.49	-0.01

Inequality (16) does not hold under these circumstances ($\partial p_{D_W} / \partial p_{D_U}$ = -0.0586), such that, according to the analysis, higher values of p_{D_U} do not produce higher values of p_{D_W}. Figure 2 describes p_{D_W} as a function of p_{D_U}. The values of p_{D_W} were computed based on (13). Evidently, as input error probability grows from 0.01 to 0.49, output error probability decreases, again, from 0.1134 to 0.0853.

Table 1 shows the value of $\partial p_{D_W} / \partial p_{D_U}$ for a sample of the values of p_U, p_V, and p_{D_U} where the partial derivative under logical disjunction is negative. Table 1 echos the fact that the link between p_{D_V} and p_{D_W} can be negative for any p_V <0.5 and any p_{D_U} <0.5, if the value of p_U is high enough. Table 1 also suggests that higher values of p_U and p_{D_U}, and lower values of p_V, drive the partial derivative down. These facts are easily derived from (9) and (10).

Likewise, Table 2 shows the value of $\partial p_{D_W} / \partial p_{D_U}$ under logical conjunction for a sample of the values of p_U, p_V, and p_{D_V} where the partial derivative is negative. Table 2 reflects the fact that the link between p_{D_U} and p_{D_W} can be negative for any p_U >0.5 and any p_{D_V} <0.5, if the value of p_V is low enough. Higher values of p_U and p_{D_U}, and lower values of p_V, drive the partial derivative $\partial p_{D_W} / \partial p_{D_U}$ down.

Table 2. The decrease in p_{D_W} as p_{D_V} increases under logical conjunction (a sample)

p_U	p_V	p_{D_U}	$\frac{\partial p_{D_W}}{\partial p_{D_V}}$	p_U	p_V	p_{D_U}	$\frac{\partial p_{D_W}}{\partial p_{D_V}}$	p_U	p_V	p_{D_U}	$\frac{\partial p_{D_W}}{\partial p_{D_V}}$	p_U	p_V	p_{D_U}	$\frac{\partial p_{D_W}}{\partial p_{D_V}}$
0.99	0.01	0.02	-0.01	0.99	0.41	0.42	-0.01	0.91	0.33	0.49	-0.10	0.75	0.01	0.49	-0.24
		0.11	-0.10			0.49	-0.07		0.41	0.47	-0.01		0.09	0.3	-0.07
		0.17	-0.16		0.47	0.49	-0.01			0.49	-0.03			0.37	-0.11
		0.25	-0.24	0.91	0.01	0.02	-0.01	0.83	0.01	0.18	-0.11			0.49	-0.18
		0.33	-0.31			0.12	-0.09			0.21	-0.13		0.17	0.35	-0.03
		0.41	-0.39			0.25	-0.20			0.29	-0.18			0.41	-0.07
		0.49	-0.47			0.33	-0.26			0.37	-0.24			0.49	-0.12
	0.09	0.1	-0.01			0.41	-0.33			0.49	-0.32		0.25	0.41	-0.01
		0.15	-0.06			0.49	-0.39		0.09	0.21	-0.06			0.49	-0.06
		0.21	-0.12		0.09	0.12	-0.01			0.29	-0.11		0.31	0.49	-0.01
		0.29	-0.19			0.17	-0.05			0.37	-0.17	0.67	0.01	0.34	-0.11
		0.37	-0.27			0.25	-0.12			0.49	-0.25			0.41	-0.13
		0.49	-0.39			0.33	-0.19		0.17	0.25	-0.01			0.49	-0.16
	0.17	0.18	-0.01			0.41	-0.25			0.31	-0.05		0.09	0.39	-0.07
		0.21	-0.04			0.49	-0.32			0.37	-0.10			0.49	-0.11
		0.29	-0.12		0.17	0.21	-0.01			0.49	-0.18		0.17	0.44	-0.03
		0.37	-0.19			0.25	-0.04		0.25	0.35	-0.01			0.49	-0.05
		0.49	-0.31			0.33	-0.11			0.37	-0.03		0.23	0.49	-0.01
	0.25	0.26	-0.01			0.41	-0.18			0.49	-0.12	0.55	0.01	0.47	-0.04
		0.29	-0.04			0.49	-0.25		0.33	0.44	-0.01			0.49	-0.04
		0.37	-0.11		0.25	0.3	-0.01			0.49	-0.05		0.03	0.48	-0.03
		0.49	-0.23			0.33	-0.04		0.38	0.49	-0.01			0.49	-0.03
	0.33	0.34	-0.01			0.41	-0.10	0.75	0.01	0.26	-0.12		0.05	0.49	-0.02
		0.37	-0.04			0.49	-0.17			0.33	-0.16		0.06	0.49	-0.02
		0.49	-0.15		0.33	0.39	-0.01			0.41	-0.20	0.53	0.01	0.49	-0.02

Concluding remarks

Understanding the relationship between input accuracy and output accuracy is important for effective and efficient data and information system design and management. However, our understanding of that relationship is limited. This study addresses the question of the sign of the relationship in two basic information processing operations, the Boolean operations of logical OR and logical AND. The results suggest that when the correct values of the input variables vary widely in their means, the sign of the relationship between input accuracy and output accuracy can be negative. These results are surprising and troubling. Their practical implications may be of substantial interest, especially since the assumptions of the theoretical model are not very restrictive for the most part. In particular, although the independence assumption is a strong assumption, it can be shown that the results are not highly sensitive to the independence assumption; the pre-requisite that errors are random can be significantly relaxed. A critical assumption of this model is that both inputs have errors. If only one of the inputs has errors and their rate is reduced, then it can be easily proved that the output error rate will decrease too, consistent with GIGO. However, unfortunately, the assumption that both inputs have errors may hold in numerous real-world settings.

The findings of this research imply that, in essence, errors should not all be treated equally. Of course, if accuracy could be improved to the extent that data are error-free, such distinction would be immaterial. However, when resources are limited, the ability to set priorities while taking into account the intended use of the data can be valuable. For instance, a potentially useful strategy in the example of the machine checkup—if the economic consequence of the application that uses logical disjunction is dominant—is to set high priority for improving the accuracy of the machine age data (where the mean of the correct value of the respective indicator is high). Alternatively, if the economic consequence of the application that uses logical conjunction takes control, then an opposite strategy can have better outcome. The accuracy of the machine utilization data—where the respective indicator has a low correct input mean—should be a priority. An attempt to decrease the error rate of the machine age data before the utilization data have been corrected would be inefficient in this case, and can cause actual losses.

References

Askira Gelman, I. "Simulations of the relationship between an information system's input accuracy and its output accuracy." *Proc. 9th Int'l Conf. Information Quality* , 2004, pp. 99-110.

Barabash, T. L. "On properties of symbol recognition." *Engineering Cybernetics*, 1965, pp. 71-77.

Berg, S. "Condorcet's jury theorem revisited." *European Journal of Political Economy*, 9(3), 1993, pp. 437-446.

Bishop, C.M. "Training with noise is equivalent to Tikhonov regularization." *Neural Computation*, 7(1), 1995, pp. 108-116.

Clemen, R.T., and Winkler, R.L. "Limits for the precision and value of information from dependent sources." *Operations Res.*, 33(2), 1985, pp. 427-442.

Cover, T. "The best two independent measurements are not the two best." *IEEE Transactions on Systems, Man and Cybernetics*, Vol. SMC-4, No. 1, 1974, pp. 116-117.

Einhorn, H.J. "The use of nonlinear, noncompensatory models in decision making." *Psychological Bulletin*, 73(3), 1970. pp. 221-230.

Frantsuz, A.G. "Influence of correlations between attributes on their informativeness for pattern recognition." *Engineering Cybernetics*, No 4, 1967.

Ladha, K. "Information pooling through majority-rule voting: Condorcet's Jury Theorem with correlated votes." *J. Econ. Behavior and Organization*, Vol. 26, 1995, pp. 353-372.

Kuncheva, L.I., Whitaker, C.J., Shipp, C.A., and Duin, R.P.W. "Limits on the majority vote accuracy in classifier fusion." *Pattern Analysis and Applications*, 6(1), 2003, pp. 2-31.

Raviv, Y., and Intrator, N. "Bootstrapping with noise: An effective regularization technique." *Connection Science*, Special issue on Combining Estimators, 8, 1996, pp. 356-372.

Skurichina, M., Raudys, S., and Duin, R.P.W. "K-Nearest neighbours directed noise injection in multilayer perceptron training. " *IEEE Transactions on Neural Networks*, 11(2), 2000, pp. 504-511.

Toussaint, G.T. "Note on optimal selection of independent binary-valued features for pattern recognition." *IEEE Transactions on Information Theory*, Vol. IT-17, 1971, p. 618.

Appendix

Proof of Lemma 1: Using (1), (2), (4), and (5), we derive from (3) that:

$$D_W = (D_U + D_V - 2UD_U - 2VD_V - UD_V - VD_U - D_U D_V + 2UVD_V + 2UVD_U + 2UD_U D_V + 2VD_U D_V - 4UVD_U D_V)/(1-2(U+V-UV)) \tag{A.1}$$

We show that: .

$$\begin{aligned}&(D_U + D_V - 2UD_U - 2VD_V - UD_V - VD_U - D_U D_V + 2UVD_V + 2UVD_U + 2UD_U D_V + 2VD_U D_V - 4UVD_U D_V)/(1-2(U+V-UV))\\&= D_U + D_V - UD_V - VD_U - D_U D_V + 2UVD_U D_V\end{aligned} \tag{A.2}$$

by calculating the value of the left-hand-side expression of (A.2) and the value of the right-hand-side expression of (A.2) given each of the possible variable-value combinations, and demonstrating that the expressions have the same value.

D_U	D_V	U	V	LHS of (A.2)	RHS of (A.2)
0	0	0	0	**0**	**0**
0	0	0	1	**0**	**0**
0	0	1	0	**0**	**0**
0	0	1	1	**0**	**0**
0	1	0	0	**1**	**1**
0	1	0	1	**1**	**1**
0	1	1	0	**0**	**0**
0	1	1	1	**0**	**0**
1	0	0	0	**1**	**1**
1	0	0	1	**0**	**0**
1	0	1	0	**1**	**1**
1	0	1	1	**0**	**0**
1	1	0	0	**1**	**1**
1	1	0	1	**0**	**0**
1	1	1	0	**0**	**0**
1	1	1	1	**1**	**1**

It follows that:

$$\begin{aligned}p_{D_W} &= E(D_W) = E((D_U + D_V - 2UD_U - 2VD_V - UD_V - VD_U - D_U D_V + 2UVD_V + 2UVD_U + 2UD_U D_V + 2VD_U D_V - 4UVD_U D_V)/(1-2(U+V-UV)))\\&= E(D_U + D_V - UD_V - VD_U - D_U D_V + 2UVD_U D_V) = E(D_U) + E(D_V) - E(UD_V) - E(VD_U) - E(D_U D_V) + 2E(UVD_U D_V)\end{aligned} \tag{A.3}$$

End of proof.

Proof of Lemma 2: Using (1), (2), (11), and (12), we derive from (3), that:

$$D_W = (UD_V - 2UVD_V + VD_U - 2UVD_U + D_U D_V - 2VD_U D_V - 2UD_U D_V + 4UVD_U D_V)/(1-2UV) \tag{A.4}$$

We show that:

$$
\begin{aligned}
&(UD_V - 2UVD_V + VD_U - 2UVD_U + D_U D_V - 2VD_U D_V - 2UD_U D_V + 4UVD_U D_V)/(1-2UV) \\
&= VD_U + UD_V + D_U D_V - 2UD_U D_V - 2VD_U D_V + 2UVD_U D_V
\end{aligned} \tag{A.5}
$$

by calculating the value of the left-hand-side expression of (A.5) and the value of the right-hand-side expression of (A.5) given each of the possible variable-value combinations, and demonstrating that the expressions always have the same value.

D_U	D_V	U	V	LHS of (A.5)	RHS of (A.5)
0	0	0	0	**0**	**0**
0	0	0	1	**0**	**0**
0	0	1	0	**0**	**0**
0	0	1	1	**0**	**0**
0	1	0	0	**0**	**0**
0	1	0	1	**0**	**0**
0	1	1	0	**1**	**1**
0	1	1	1	**1**	**1**
1	0	0	0	**0**	**0**
1	0	0	1	**1**	**1**
1	0	1	0	**0**	**0**
1	0	1	1	**1**	**1**
1	1	0	0	**1**	**1**
1	1	0	1	**0**	**0**
1	1	1	0	**0**	**0**
1	1	1	1	**1**	**1**

It follows that:

$$
\begin{aligned}
p_{D_W} &= E(D_W) = E((UD_V - 2UVD_V + VD_U - 2UVD_U + D_U D_V - 2VD_U D_V - 2UD_U D_V + 4UVD_U D_V)/(1-2UV)) \\
&= E(VD_U + UD_V + D_U D_V - 2UD_U D_V - 2VD_U D_V + 2UVD_U D_V) \\
&= E(VD_U) + E(UD_V) + E(D_U D_V) - 2E(UD_U D_V) - 2E(VD_U D_V) + 2E(UVD_U D_V)
\end{aligned} \tag{A.6}
$$

End of proof.