informs.
http://pubsonline.informs.org/journal/isre/

INFORMATION SYSTEMS RESEARCH
Vol. 29, No. 1, March 2018, pp. 4–24
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

Research Commentary

Mind the Gap: Accounting for Measurement Error and Misclassification in Variables Generated via Data Mining

Mochen Yang,[a] **Gediminas Adomavicius,**[a] **Gordon Burtch,**[a] **Yuqing Ren**[a]

[a] Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455

Contact: yang3653@umn.edu, http://orcid.org/0000-0001-5101-9041 (MY); gedas@umn.edu (GA); gburtch@umn.edu (GB); chingren@umn.edu (YR)

Received: October 6, 2015
Revised: July 20, 2016; January 28, 2017
Accepted: April 24, 2017
Published Online in Articles in Advance: January 29, 2018

https://doi.org/10.1287/isre.2017.0727

Copyright: © 2018 INFORMS

Abstract. The application of predictive data mining techniques in information systems research has grown in recent years, likely because of their effectiveness and scalability in extracting information from large amounts of data. A number of scholars have sought to combine data mining with traditional econometric analyses. Typically, data mining methods are first used to generate new variables (e.g., text sentiment), which are added into subsequent econometric models as independent regressors. However, because prediction is almost always imperfect, variables generated from the first-stage data mining models inevitably contain measurement error or misclassification. These errors, if ignored, can introduce systematic biases into the second-stage econometric estimations and threaten the validity of statistical inference. In this commentary, we examine the nature of this bias, both analytically and empirically, and show that it can be severe even when data mining models exhibit relatively high performance. We then show that this bias becomes increasingly difficult to anticipate as the functional form of the measurement error or the specification of the econometric model grows more complex. We review several methods for error correction and focus on two simulation-based methods, SIMEX and MC-SIMEX, which can be easily parameterized using standard performance metrics from data mining models, such as error variance or the confusion matrix, and can be applied under a wide range of econometric specifications. Finally, we demonstrate the effectiveness of SIMEX and MC-SIMEX by simulations and subsequent application of the methods to econometric estimations employing variables mined from three real-world data sets related to travel, social networking, and crowdfunding campaign websites.

History: Sabyasachi Mitra, Senior Editor; Gautam Pant, Associate Editor.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2017.0727.

Keywords: data mining • econometrics • measurement error • misclassification • statistical inference

1. Introduction

The application of data mining[1] methods creates appealing opportunities for research across multiple disciplines, such as information systems (IS), marketing, economics, and finance. The increasing availability of big data and unstructured data further contributes to the popularity of data mining methods (Agarwal and Dhar 2014, Chen et al. 2012, Varian 2014). Based on observed data, predictive data mining models can be used to automatically generate or estimate variables that researchers are interested in, making it an efficient and sophisticated approach to processing large amounts of structured and unstructured data. Recent examples include the use of text mining techniques to determine the sentiment of text (e.g., Pang et al. 2002, Das and Chen 2007), and the use of image classifiers to predict an individual's gender or race from a profile picture (e.g., Chan and Wang 2014, Rhue 2015), or to detect the presence (absence) of various objects in AirBNB property listings (Zhang et al. 2016).

Many IS studies have recently sought to combine data mining approaches with traditional statistical analyses or econometric modeling in a two-stage process. In the first stage, pretrained data mining models are deployed to generate new variables that are not readily available from existing data. In the second stage, these generated variables are added into regression models, usually as independent regressors. Several papers adopting this two-stage process have uncovered interesting insights and have been published in top IS journals (e.g., Gu et al. 2007, 2014; Aggarwal et al. 2012; Lu et al. 2013; Moreno and Terwiesch 2014; Wang et al. 2013; Archak et al. 2011). For instance, Aggarwal et al. (2012) adopted a text classification model to label sentiments of online blog posts as positive, negative, and neutral. They then estimated a regression to demonstrate the effect of message sentiment on venture financing outcomes.

However, an important issue with this two-stage process is that variables generated in the first stage almost certainly contain some amount of *predictive error*, because predictive data mining models are imperfect. Such error then carries over to the second-stage econometric models and manifests as *measurement error*, if the variable is continuous, or *misclassification*, if the variable is discrete. For example, suppose we have built a text classification model on a training data set, which predicts the sentiment of Facebook posts as either positive or negative, and the model has achieved a recall, or sensitivity, of 0.8 for the "positive" class on a holdout, testing data set. This means that 20% of posts that are actually positive are incorrectly classified as negative. These errors, if ignored, can introduce systematic biases into the second-stage estimations and may, therefore, threaten the validity of the subsequent statistical inferences.

The issue of measurement error and misclassification is not new and has received a great deal of attention from econometricians and statisticians (Greene 2003). However, it warrants special attention in the new context of big data and increasing interest in combining data mining with econometric modeling for the following reasons. First and foremost, measurement error is unobservable in many situations; however, the errors here, which originate from imperfect predictions by first-stage data mining models, can be observed and quantified using standard methods of model evaluation, stemming from confusion matrices or continuous measures of error. This provides a clear opportunity to diagnose the error and correct for the resulting bias. Second, many, if not most, studies in IS that have used the two-stage approach of combining econometric modeling with data mining have failed to acknowledge the potential estimation biases introduced by measurement error or misclassification. We believe that this may derive, at least in part, from a lack of understanding or awareness of the issue. Third, the variables obtained from the first-stage prediction typically enter the second-stage estimation as independent regressors. Unlike error in dependent variables, which typically leads to inflated variance of estimates and decreased model fit, error in independent variables generally introduces systematic biases into coefficient estimates (Greene 2003), and thus causes serious concerns.[2] Yet, most IS researchers seem to be unaware of either the potential biases from predictive errors or proper methods to mitigate the biases.

In this commentary, we hope to bridge this gap by addressing three key issues: (1) To what extent will measurement error or misclassification from data mining models bias estimations in econometric analyses that incorporate the output of those models? (2) How can we diagnose the structure of the measurement error or misclassification, and the resulting biases, in a particular research setting and data set? (3) How can we mitigate these biases?

Based on both theoretical reasoning and simulated data, we first demonstrate that measurement error and misclassification can indeed introduce considerable biases into several commonly used econometric models, such as linear regressions, generalized linear regressions (e.g., logit, probit, and Poisson models), and panel data regressions. Notably, our simulations are conducted based on commonly observed levels of predictive performance in data mining models, in terms of error variance for numeric predictions, or precision and recall for classifications. Hence, the errors we simulate and the biases we observe are likely to manifest in an actual study.

Having established the undesirable impact of error on econometric analyses, we then review several possible error-correction methods. We focus on two simulation-based methods that lend themselves well to mitigating the bias introduced by predictive measurement error and misclassification. The simulation-extrapolation (SIMEX thereafter) method applies to continuous variables with additive measurement error (Cook and Stefanski 1994). The misclassification-SIMEX (MC-SIMEX thereafter) method applies to discrete variables with misclassification (Küchenhoff et al. 2006). We focus on SIMEX and MC-SIMEX rather than other approaches, such as the instrumental variable approach or the method-of-moments approach, for two main reasons. First, SIMEX and MC-SIMEX can easily be applied to a variety of model specifications, whereas most other methods require model-specific assumptions. Second, SIMEX and MC-SIMEX can be configured based solely on the observable performance indicators of first-stage data mining models, whereas other methods typically require explicit modeling of errors in the second-stage estimations. We validate the effectiveness of SIMEX and MC-SIMEX using simulated data, and we then apply both methods to three real-world data sets. Our results demonstrate the effectiveness of these methods in mitigating estimation bias from measurement error and misclassification. Our results also reveal the limitation of these or any methods in addressing predictive measurement error issues, when first-stage data mining performance is problematically low. Finally, we provide a guiding procedure that researchers can follow to diagnose estimation biases and assess the efficacy of specific error-correction methods in consideration of their research settings, with specific data samples and data mining models.

This commentary contributes to the IS literature in three ways. First, we describe and raise awareness of the issue of measurement error and misclassification in the context of an increasingly prevalent methodological practice in IS research, i.e., the integration of data mining and econometric analyses. We show that, while predictive error can bias econometric estimations, the ability to quantify such error brings the opportunity to correct for estimation biases. Second, we review several existing remedial approaches that can address the identified issue, and demonstrate the effectiveness of two methods in particular, using both simulations and real-world empirical applications. Third, we propose a diagnostic procedure via which researchers can assess the characteristics of the measurement error and estimation bias in a particular scenario, with a given sample of data, and thereby choose the best approach to address the problem in that setting.

Measurement error and misclassification may arise in a variety of research settings and are very difficult to avoid completely. Therefore, we believe that awareness of the problem and the severity of its consequences can help researchers understand the potential risks of combining data mining with econometric analyses, and thus to improve the robustness of their conclusions. At the same time, we stress that the points raised in this commentary do not necessarily invalidate the results of any past work, because the predictive error in the first-stage data mining can have variable effects on the subsequent econometric estimation. The predictive error may cause attenuation of coefficients in some cases, amplification in others, and in some cases it may have little effect at all. Thus, going forward, our aim with this commentary is to highlight the unique opportunity of error correction in this setting and to provide IS scholars with guidance on the implementation of this integrated methodology in as robust a manner as possible.

2. The Common Practice of Combining Data Mining and Econometric Analyses

Studies that have adopted the two-stage methodology of combining data mining techniques with econometric estimations are becoming prevalent in the IS discipline. A cursory search of recently published issues of top IS journals and conference proceedings revealed at least 13 studies that have used this approach; we identified six recent studies in *Information Systems Research* (Gu et al. 2007, 2014; Aggarwal et al. 2012; Wang et al. 2013; Moreno and Terwiesch 2014; Singh et al. 2014), two in *Management Science* (Archak et al. 2011, Lu et al. 2013), two appearing in other journals (Ghose and Ipeirotis 2011, Ghose et al. 2012), and three in the *Proceedings of the International Conference on Information Systems* (Chan and Wang 2014, Rhue 2015, Zhang et al. 2016).[3] The two-stage methodology has also been adopted in several fields outside the IS community, such as marketing (e.g., Tirunillai and Tellis 2012), human-computer interaction (e.g., Liu et al. 2012; Zhu et al. 2011, 2012), economics (e.g., Jelveh et al. 2014), and finance (see Fisher et al. 2016 for a review). In this section, we report and discuss several patterns we have observed in these publications.

The most common application of data mining models in these studies was text classification that was used primarily for coding online user-generated content, such as consumer reviews. Another, less common use was image classification that was used to identify objects or persons from digital photographs (Ghose et al. 2012, Chan and Wang 2014, Rhue 2015, Zhang et al. 2016). Most of the papers followed the common approach to develop the classification models.[4] To build a classification model, researchers first draw a random subsample of observations from the data set and have them manually classified or labeled by human coders based on predefined rules. This manually classified subsample then becomes the ground truth for training and evaluating the classifier.[5] A classifier is trained using a portion of the labeled data and then its performance is evaluated using the remaining data, by comparing the classifier's predictions with the ground truth. Some studies (e.g., Ghose and Ipeirotis 2011, Ghose et al. 2012) have adopted a more advanced evaluation method, known as cross-validation, wherein the labeled set is partitioned into K folds, and classifiers are iteratively trained on different sets of $K - 1$ folds and evaluated using the remaining fold. The trained classifiers are then deployed on the unlabeled remainder of the data set to obtain predicted labels. This approach has the benefit of scalability, because hand coding an overwhelmingly large data set is often infeasible.

There exist many data mining techniques for building predictive models, including classification and regression trees, k-nearest neighbors, naïve Bayes, neural networks, support vector machines, Bayesian networks, and various linear and nonlinear regression techniques. Some of the techniques were developed to predict continuous outcomes (numeric prediction task), some to predict discrete outcomes (classification task), and others can be configured for either purpose. Several metrics are available to assess their predictive performance. For numeric prediction, evaluations are based on prediction errors, i.e., the differences between predicted and actual values. Commonly used metrics include MAE (mean absolute error) and RMSE (root mean squared error) (Aggarwal 2015). For classification models, commonly used metrics include overall accuracy (the percentage of correct predictions across all classes), precision (the percentage of predictions in a given class that are correct), and recall (the percentage of cases that truly belong to a given class that are correctly predicted by the model) (Aggarwal 2015). Figure 1 illustrates these performance metrics using a binary classification model as an example. All papers

Figure 1. Performance Metrics for a Two-Class Classification Model

	Actual	
	Positive	Negative
Predicted		
Positive	TP	FP
Negative	FN	TN

Accuracy = (TP + TN)/(TP + TN + FP + FN)
For positive class:
Precision = TP/(TP + FP), Recall = TP/(TP + FN)
For negative class:
Precision = TN/(TN + FN), Recall = TN/(TN + FP)

Notes. The left-hand panel is a confusion matrix obtained by evaluating a given predictive model. It summarizes the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The right-hand panel lists the performance metrics derived from the confusion matrix, including overall accuracy, precision, and recall rates.

we surveyed used classification models; six papers reported the predictive performance of their data mining models, with overall accuracy ranging from 60% to 87%, precision ranging from 70% to 100%, and recall ranging from 74% to 100%.

Another pattern we observed was that, in all papers, the variables generated via data mining were incorporated into second-stage regressions with many other covariates. Typical second-stage econometric models include linear regressions with fixed or random effects, logit or probit regressions, systems of equations, and vector autoregression (VAR). We also observed that the econometric models were typically estimated on a much larger sample than the one used to train the data mining model in the first stage. For example, Moreno and Terwiesch (2014) used a labeled sample to train their model that comprised 2% of the total data set. The trained model was then used to generate the variable of interest for the remaining 98% of the data set. As we will show in Section 3, the measurement error or misclassification that originates from data mining has the potential to introduce systematic biases into subsequent econometric estimations. These biases persist in large samples, and are generally harder to anticipate or predict as the specification of the econometric model becomes more complex.

3. Estimation Biases Because of Measurement Error or Misclassification

In this section, we present both analytical and simulation results regarding the biases in coefficient estimates caused by measurement error or misclassification, for several commonly used econometric models. First, we discuss a simple linear regression with one regressor, containing either measurement error or misclassification. In this scenario, the bias can be mathematically derived. Subsequently, for more complicated model specifications, we demonstrate the resultant biases using simulated data.

3.1. Bias in Linear Regression with One Regressor

Consider a simple linear regression with only one regressor: $Y = \beta_0 + \beta_1 X + \varepsilon$. If both the dependent and independent variables are precisely measured, ordinary least squares (OLS) would yield unbiased, consistent, and efficient estimates of β_0 and β_1 (Greene 2003). Now, suppose that instead of X we actually observe $\hat{X}$, which includes error. Regressing Y on $\hat{X}$ would yield biased estimates.

If X is a continuous variable, there are two broad types of measurement error: *classical* error and *nonclassical* error. If the measurement error, e, is *random* and *additive*—i.e., $\hat{X} = X + e$—and *independent* of both X and ε, such error is known as classical measurement error (Carroll et al. 2006). The error results in an *attenuation bias*, that is, the estimated $\hat{\beta}_1$ satisfies $E(\hat{\beta}_1 \mid \hat{X}) = \beta_1[\sigma_X^2/(\sigma_X^2 + \sigma_e^2)]$, which implies that the regression coefficient is underestimated (see Greene 2003 for the proof). Given X, the magnitude of the bias depends on the variance of the error, and larger error variance leads to greater bias. Measurement error that is not random, not additive, or not independent of X and ε is known as nonclassical measurement error, and we will discuss its impact in Sections 3.2 and 3.3.

If X is a discrete variable, the misclassification would also result in a systematic bias in the regression coefficient. For simplicity, we can assume $\hat{X}$ is a dummy variable and conditionally independent of Y given X (i.e., nondifferential misclassification),[6]then the estimated $\hat{\beta}_1$ will satisfy $E(\hat{\beta}_1 \mid \hat{X}) = \beta_1[\Pr(X = 1 \mid \hat{X} = 1) - \Pr(X = 1 \mid \hat{X} = 0)]$ (Gustafson 2003; see Online Appendix A1 for the proof). Using data mining performance measures, this relationship can be written as follows: $E(\hat{\beta}_1 \mid \hat{X}) = \beta_1[\Pr(X = 1 \mid \hat{X} = 1) + \Pr(X = 0 \mid \hat{X} = 0) - 1] = \beta_1[\mathit{Precision}(\mathit{class}\ 1) + \mathit{Precision}(\mathit{class}\ 0) - 1]$. That is, the magnitude of the bias is determined by the sum of the precision rates for the two classes. Online Appendix A2 provides an example and a graphical illustration of how misclassification can result in estimation bias. In extreme cases when the sum of the two precision scores is smaller than 1, the estimated coefficient may shift in the opposite direction from the true value, resulting in a coefficient of the opposite sign.

Finally, note that the above finite sample results also hold asymptotically. For continuous measurement error, $\operatorname{plim} \hat{\beta}_1 = \beta_1[\sigma_X^2/(\sigma_X^2 + \sigma_e^2)]$. For binary misclassification, $\operatorname{plim} \hat{\beta}_1 = \beta_1[\Pr(X = 1 \mid \hat{X} = 1) - \Pr(X = 1 \mid \hat{X} = 0)]$. Therefore, coefficient estimates with errors are inconsistent.

3.2. Bias in More Complicated Models: Theoretical Results

Considering the above discussion, one might be tempted to conclude that measurement error and misclassification will typically only produce an *attenuation bias* in coefficient estimates and, thus, will only lead to conservative results and Type II error. However, it is important to note that an *amplification bias* may also manifest. This can happen when either the error structure or the econometric specification in the second-stage regression model grows more complicated.

First, in the case of linear regression with one regressor, amplification bias can manifest under nonclassical measurement error whose error structure deviates from the random, additive, and independent error that we described in Section 3.1. For example, consider a continuous variable with an additive measurement error, where the error structure includes both a random component and a systematic component in the form of $\hat{X} = a + bX + e$, $E(e) = 0$, where a represents the additive systematic error, b represents the multiplicative systematic error, and e represents the random error. The resulting coefficient on $\hat{X}$ then satisfies $E(\hat{\beta}_1 \mid \hat{X}) = (\beta_1 b\sigma_X^2 + \rho_{e\varepsilon}\sigma_e\sigma_\varepsilon)/(b^2\sigma_X^2 + \sigma_e^2)$ (Carroll et al. 2006). Attenuation bias happens with a classical measurement error, as we illustrated in Section 3.1, only because we assumed (1) e is uncorrelated with ε, the regression error term (i.e., $\rho_{e\varepsilon} = 0$), and (2) there is no systematic error between X and $\hat{X}$, i.e., $b = 1$. In our scenario of interest, the form of the measurement error is determined by the data mining model. If the data mining model systematically underestimates the true value, i.e., $b < 1$, then the second assumption is no longer true, and the bias in β_1 may manifest as an amplification. An amplification bias can occur in misclassification as well. Gustafson (2003) notes that, if a categorical variable with more than two levels bears misclassification, no simple conclusion can be drawn about the direction of the bias in the estimated coefficient.

Second, when the second-stage regression specification becomes more complicated, the biases can be similarly difficult to anticipate. In multivariate regressions, even when the other variables (i.e., those not generated from data mining) are measured without error, the presence of a data mined variable with predictive error can cause the coefficient estimates of all variables to be biased in unknown directions (Greene 2003, Gustafson 2003, Buonaccorsi et al. 2005). In nonlinear regressions, the directions of biases associated with both the variable with error and the other precisely measured variables are also uncertain (Carroll et al. 2006).

3.3. Bias in More Complicated Models: Simulation Results

Because analytical, closed-form solutions are generally difficult to obtain for complicated regression models with measurement error or misclassification, we provide an illustrative numerical analysis based on simulation. The simulation was conducted as follows. First, we generated three variables having different underlying distributions: $X_1 \sim N(0, 1^2)$, $X_2 \sim \text{Bernoulli}(p)$, and $X_3 \sim \mathit{Uniform}(-10, 10)$. We modified X_1 and X_2 to introduce measurement error or misclassification (see details in Sections 3.3.1 and 3.3.3). Second, we generated another normally distributed variable as the error term as $\varepsilon \sim N(0, 0.5^2)$. Third, we generated a dependent variable as a function of the independent variables and the error term: $Y = 1 + 2 \times X_1 + 3 \times X_2 + 0.5 \times X_3 + \varepsilon$. The coefficients were fixed to quantify the magnitudes of estimation biases. In addition to linear regression, we also simulated logit, probit, and Poisson regressions, as well as a linear regression with fixed effects. For the three generalized linear models, we generated dependent variables based on the corresponding distributional assumptions.[7] For the linear panel data model with fixed effects, the regression estimated was $Y_{ij} = \alpha_i + 2 \times X_{1ij} + 3 \times X_{2ij} + 0.5 \times X_{3ij} + \varepsilon_{ij}$, where $\alpha_i = i$, $i \in \{1, 2, \ldots, 25\}$ and $j \in \{1, 2, \ldots, 200\}$. The α_i represented the panel-specific fixed effects. We also simulated a linear random-effects regression, where α_i were randomly drawn from a standard normal distribution. The results were qualitatively the same as the fixed-effect model, so we only reported the fixed-effect regression. The results that we report below are based on 5,000 observations. We repeated the analysis with 10,000 observations and got similar results with no qualitative differences. Below we present three simulation results, respectively, showing estimation biases caused by classical and nonclassical measurement error in X_1, and misclassification in X_2.

3.3.1. Classical Measurement Error in X_1.

We first simulated the impact of classical additive measurement error in X_1 with $\hat{X}_1 = X_1 + e$, where $e \sim N(0, \sigma_e^2)$ and is independent of X_1, ε, and the other covariates in the simulated regressions. Because X_1 follows a standard normal distribution, we considered three values of σ_e—0.1, 0.3, and 0.5—to capture different degrees of measurement error. For all simulations in this section, $X_2 \sim \text{Bernoulli}(0.3)$ and contained no misclassification, enabling us to isolate the impact of measurement error in X_1. Table 1 summarizes our simulation results. For each regression, the first column shows coefficient estimates without measurement error (denoted as b) and the second column shows coefficient estimates with measurement error in X_1 (denoted as b'). The third and fourth columns show the relative magnitude of estimation bias, calculated using estimated and true coefficient values, respectively. i.e., $\text{bias}_1 = (b' - b)/b$ and $\text{bias}_2 = (b' - b_{\text{true}})/b_{\text{true}}$. We mainly focus on bias_1 in our subsequent discussion, because this value captures the bias purely because of measurement error, whereas bias_2 reflects both bias from measurement error and error because of noise. For the linear fixed-effects regression, we omit the estimates of the fixed effects, because of space consideration. Because the data was simulated, all estimates were statistically significant. We therefore do not report standard errors or levels of statistical significance.

Several patterns emerged that are worth noting. While the coefficient on X_1 was consistently downward biased, coefficients of other variables were biased in different directions. As the magnitude of measurement error increased from 0.1 to 0.5, bias in the coefficient of X_1 also increased from −1.3% to −20% in the OLS case. Compared to OLS, biases in generalized linear models were greater. For example, with measurement error of $\sigma_e = 0.5$, the coefficient of X_1 in OLS was underestimated by 20%, compared to 27.2% in logit, 41.7% in probit, and 33.1% in Poisson regression. Bias in the linear fixed-effect model was comparable to bias in OLS, and estimates of the fixed effects were unbiased.

3.3.2. Nonclassical Measurement Error in X_1.

We simulated three types of nonclassical measurement error in X_1: (1) $\hat{X}_1 = X_1 + e$, $e \sim N(0, \sigma_e^2)$ and is independent of ε and other covariates, but is correlated with the true value X_1 with $\rho_{X_1 e} = 0.5$; (2) $\hat{X}_1 = X_1 + e$, $e \sim N(0, \sigma_e^2)$ and is independent of X_1 and ε, but is correlated with X_3 with $\rho_{X_3 e} = 0.5$; (3) $\hat{X}_1 = 1 + 0.5X_1 + e$, $e \sim N(0, \sigma_e^2)$ and is independent of X_1, ε, and other covariates. The first two scenarios represent random measurement error that is correlated with either the true value or another covariate in the second-stage regression, and the third scenario represents systematic independent measurement error. All three scenarios of error may occur in data mining model predictions. For simplicity, we only report simulation results for linear regressions in Table 2. We obtained similar results for other regressions.

Several patterns emerged that are worth noting. Under scenario (1), where measurement error was correlated with the true value of X_1, we observed greater downward bias in the coefficient on X_1 than the bias from classical measurement error. Under scenario (2), where the error was correlated with X_3, we observed biases in the coefficients of both X_1 and X_3. Under scenario (3), where the measurement error was systematic, we observed overestimation of the coefficient of X_1 for $\sigma_e = 0.1$ and $\sigma_e = 0.3$, but attenuation for $\sigma_e = 0.5$. In other words, as error variance became greater, the bias shifted from amplification to attenuation. As noted previously, this result demonstrates numerically that measurement error introduced during first-stage data mining tasks do not necessarily result in attenuation and conservative estimates; in some cases, it may result in *amplified* coefficient estimates.

3.3.3. Misclassification in X_2.

We simulated misclassification by modifying the value of X_2. We use a misclassification matrix to represent the magnitude

Table 1. Regression Results for X_1 with Classical Measurement Error

	OLS				Logit				Probit				Poisson				Fixed effect			
	b	b'	bias$_1$ (%)	bias$_2$ (%)	b	b'	bias$_1$ (%)	bias$_2$ (%)	b	b'	bias$_1$ (%)	bias$_2$ (%)	b	b'	bias$_1$ (%)	bias$_2$ (%)	b	b'	bias$_1$ (%)	bias$_2$ (%)
$\sigma_e = 0.1$																				
C	1.004	1.006	0.3	0.6	1.002	0.993	−0.9	−0.7	1.042	1.012	−2.8	1.2	0.999	1.232	23	23.2				
X_1	**1.995**	**1.970**	**−1.3**	**−1.5**	**1.996**	**1.954**	**−2.1**	**−2.3**	**1.979**	**1.903**	**−3.9**	**−4.9**	**2.000**	**1.911**	**−4.4**	**−4.5**	**1.994**	**1.977**	**−0.9**	**−1.2**
X_2	2.988	2.990	0.1	−0.3	2.890	2.868	−0.8	−4.4	2.897	2.822	−2.6	−5.9	3.000	2.956	−1.5	−1.5	2.986	2.987	0.03	−0.4
X_3	0.499	0.500	0.2	0.0	0.493	0.489	−0.8	−2.2	0.484	0.471	−2.7	−5.8	0.500	0.486	−2.8	−2.8	0.499	0.499	0	−0.2
$\sigma_e = 0.3$																				
C	1.004	1.005	0.2	0.5	1.002	0.962	−3.9	−3.8	1.042	0.911	−12.5	−8.9	0.999	1.482	48	48.2				
X_1	**1.995**	**1.833**	**−8.1**	**−8.4**	**1.996**	**1.748**	**−12.4**	**−12.6**	**1.979**	**1.567**	**−20.8**	**−21.7**	**2.000**	**1.760**	**−12.0**	**−12.0**	**1.994**	**1.824**	**−8.5**	**−8.8**
X_2	2.988	2.980	−0.3	−0.7	2.890	2.757	−4.6	−8.1	2.897	2.499	−13.7	−16.7	3.000	2.866	−4.5	−4.5	2.986	2.990	0.13	−0.3
X_3	0.499	0.500	0.2	0.0	0.493	0.472	−4.3	−5.6	0.484	0.421	−13.2	−15.8	0.500	0.474	−5.1	−5.2	0.499	0.496	−0.6	−0.8
$\sigma_e = 0.5$																				
C	1.004	0.996	−0.8	−0.4	1.002	0.899	−10.2	−10.1	1.042	0.765	−26.6	−23.5	0.999	2.445	144	144.5				
X_1	**1.995**	**1.595**	**−20.0**	**−20.3**	**1.996**	**1.453**	**−27.2**	**−27.4**	**1.979**	**1.155**	**−41.7**	**−42.3**	**2.000**	**1.337**	**−33.1**	**−33.2**	**1.994**	**1.589**	**−20.3**	**−20.6**
X_2	2.988	3.011	0.8	0.4	2.890	2.678	−7.3	−10.7	2.897	2.205	−23.9	−26.5	3.000	2.750	−8.3	−8.3	2.986	2.983	−0.1	−0.6
X_3	0.499	0.500	0.2	0.0	0.493	0.445	−9.7	−11.0	0.484	0.357	−26.2	−28.6	0.500	0.418	−16.4	−16.4	0.499	0.499	0	−0.2

Notes. For each regression, b stands for coefficient estimates when no error was introduced, b' stands for coefficient estimates when error was introduced in X_1. bias$_1$ and bias$_2$ stand for relative magnitude of estimation bias, calculated using estimated and true coefficient values, respectively.

Table 2. Regression Results for X_1 with Nonclassical Measurement Error

	Scenario (1)				Scenario (2)				Scenario (3)			
	b	b'	bias$_1$ (%)	bias$_2$ (%)	b	b'	bias$_1$ (%)	bias$_2$ (%)	b	b'	bias$_1$ (%)	bias$_2$ (%)
$\sigma_e = 0.1$												
C	1.004	1.003	−0.1	0.3	1.004	1.005	0.1	0.5	1.004	−2.817	−380	−381.7
X_1	**1.995**	**1.884**	**−5.6**	**−5.8**	**1.995**	**1.979**	**−0.8**	**−1.1**	**1.995**	**3.827**	**91.8**	**91.4**
X_2	2.988	2.983	−0.2	−0.6	2.988	2.983	−0.2	−0.6	2.988	2.977	−0.4	−0.8
X_3	0.499	0.498	−0.2	−0.4	0.499	0.481	−3.6	−3.8	0.499	0.498	−0.2	−0.4
$\sigma_e = 0.3$												
C	1.004	1.000	−0.4	0.0	1.004	1.006	0.2	0.6	1.004	−1.939	−293	−293.9
X_1	**1.995**	**1.641**	**−17.7**	**−18.0**	**1.995**	**1.863**	**−6.6**	**−6.9**	**1.995**	**2.903**	**45.5**	**45.2**
X_2	2.988	2.976	−0.4	−0.8	2.988	2.975	−0.4	−0.8	2.988	3.006	0.6	0.2
X_3	0.499	0.498	−0.2	−0.4	0.499	0.450	−9.8	−10.0	0.499	0.500	0.2	0.0
$\sigma_e = 0.5$												
C	1.004	0.997	−0.7	−0.3	1.004	1.005	0.1	0.5	1.004	−0.971	−197	−197.1
X_1	**1.995**	**1.412**	**−29.2**	**−29.4**	**1.995**	**1.667**	**−16.4**	**−16.7**	**1.995**	**1.941**	**−2.7**	**−3.0**
X_2	2.988	2.973	−0.5	−0.9	2.988	2.971	−0.6	−1.0	2.988	2.995	0.2	−0.2
X_3	0.499	0.497	−0.4	−0.6	0.499	0.425	−14.8	−15.0	0.499	0.498	−0.2	−0.4

Notes. For each regression, b stands for coefficient estimates when no error was introduced, b' stands for coefficient estimates when error was introduced in X_1. bias$_1$ and bias$_2$ stand for relative magnitude of estimation bias, calculated using estimated and true coefficient values, respectively.

of misclassification in X_2.[8] For a binary variable, the misclassification matrix can be denoted as $(M_{00}, M_{10}, M_{01}, M_{11})$, where $M_{ab} = \Pr(\hat{X}_2 = b \mid X_2 = a)$. It can also be written, equivalently as $(M_{00}, 1 - M_{11}, 1 - M_{00}, M_{11})$, where M_{00} is the recall rate for class 0 (true negative rate) and M_{11} is the recall rate for class 1 (true positive rate). We generate $\hat{X}_2$ by adjusting the value of X_2, changing it from 0 to 1 with a probability of M_{01} and from 1 to 0 with a probability of M_{10}. Using this method, $\hat{X}_2$ simulates predicted values from a binary classifier, with recall rate M_{00} for class 0 and recall rate M_{11} for class 1.

To examine the impact of different levels of misclassification, we simulated three scenarios: (1) $X_2 \sim$ Bernoulli(0.3), $M_{00} = 0.8$, and $M_{11} = 0.8$; (2) $X_2 \sim$ Bernoulli(0.3), $M_{00} = 0.6$, and $M_{11} = 0.5$; and (3) $X_2 \sim$ Bernoulli(0.5), $M_{00} = 0.6$, and $M_{11} = 0.5$. Scenarios (1) and (2) had a skewed Bernoulli distribution for the true value of X_2, with $\Pr(X_2 = 1) = 0.3$; and scenario (3) had a balanced distribution. Scenarios (2) and (3) also had greater misclassification than scenario (1). For all simulations in this section, $X_1 \sim N(0, 1^2)$ and contained no measurement error. Table 3 summarizes the results.

Several patterns emerged that are worth noting. First, even if a classifier achieved a reasonable level of performance in terms of precision and recall, the misclassification could still lead to severe bias in the coefficient estimates. For example, scenario (1) represented a binary classifier with an 80% recall rate for both classes, as well as 63% precision for the positive class and 90% precision for the negative class. Based on the published work we have surveyed, this level of performance would be considered good in many application domains. However, our simulation showed that the coefficient on X_2 was underestimated by 46.6% in the OLS regression. Second, we observed similar biases in scenarios (2) and (3) although the magnitude of the biases was greater than scenario (1). Because of the greater misclassification in scenarios (2) and (3), the coefficient on X_2 was reduced nearly to zero although it remained statistically significant. Third, in the linear fixed-effect model, the estimates of the fixed effects were also biased to various degrees, ranging from 4% to 33% overestimation. Overall, our simulation results demonstrate the biases associated with misclassification, and the risk of making inferences from the resultant estimates.

4. Bias Correction

Section 3 provides ample evidence that measurement error and misclassification, which can be introduced with the application of data mining techniques, may severely bias the estimates of econometric models. This poses serious challenges to the increasingly prevalent practice of combining data mining with econometric analysis. However, the good news is that, although data mining models produce predictions with error, the standard practice of model performance evaluation affords a readily accessible quantification of the error. Quantifying error allows one to employ corrective methods that can mitigate subsequent estimation biases. In this section, we first review several existing error-correction methods. Then, we focus on two simulation-based methods (SIMEX and MC-SIMEX), which were initially developed in the field of statistics

Table 3. Regression Results for X_2 with Misclassification

	OLS				Logit				Probit				Poisson				Fixed effect			
	b	b'	bias$_1$ (%)	bias$_2$ (%)	b	b'	bias$_1$ (%)	bias$_2$ (%)	b	b'	bias$_1$ (%)	bias$_2$ (%)	b	b'	bias$_1$ (%)	bias$_2$ (%)	b	b'	bias$_1$ (%)	bias$_2$ (%)
Scenario (1): $X_2 \sim$ Bernoulli(0.3), $M_{00} = 0.8$, and $M_{11} = 0.8$																				
C	1.004	1.292	28.7	29.2	1.002	1.163	16.1	16.3	1.042	0.943	−9.5	−5.7	0.999	1.699	70.1	69.9				
X_1	1.995	1.995	−0.0	−0.2	1.996	1.689	−15.4	−15.6	1.979	1.342	−32.2	−32.9	2.000	1.878	−6.1	−6.1	1.994	2.009	0.75	0.4
X_2	**2.988**	**1.596**	**−46.6**	**−46.8**	**2.890**	**1.106**	**−61.7**	**−63.1**	**2.897**	**0.979**	**−66.2**	**−67.4**	**3.000**	**1.722**	**−42.6**	**−42.6**	**2.986**	**1.583**	**−47.0**	**−47.2**
X_3	0.499	0.494	−1.0	−1.2	0.493	0.413	−16.3	−17.4	0.484	0.328	−32.3	−34.4	0.500	0.533	6.5	6.6	0.499	0.492	−1.4	−1.6
Scenario (2): $X_2 \sim$ Bernoulli(0.3), $M_{00} = 0.6$, and $M_{11} = 0.5$																				
C	1.004	1.768	76.1	76.8	1.002	1.370	36.7	37.0	1.042	1.131	8.6	13.1	0.999	2.671	168	167.1				
X_1	1.995	2.004	0.5	0.2	1.996	1.616	−19.0	−19.2	1.979	1.221	−38.3	−39.0	2.000	1.831	−8.5	−8.5	1.994	2.001	0.35	0.0
X_2	**2.988**	**0.282**	**−90.6**	**−90.6**	**2.890**	**0.304**	**−89.5**	**−89.9**	**2.897**	**0.148**	**−94.9**	**−95.1**	**3.000**	**0.205**	**−93.2**	**−93.2**	**2.986**	**0.271**	**−90.9**	**−91.0**
X_3	0.499	0.493	−1.1	−1.4	0.493	0.393	−20.4	−21.4	0.484	0.296	−38.9	−40.8	0.500	0.536	7.1	7.2	0.499	0.492	−1.4	−1.6
Scenario (3): $X_2 \sim$ Bernoulli(0.5), $M_{00} = 0.6$, and $M_{11} = 0.5$																				
C	1.002	2.352	135	135.2	0.980	1.744	78.0	74.4	0.917	1.260	37.4	26.0	1.000	3.662	266	266.2				
X_1	1.995	1.987	−0.4	−0.6	1.961	1.477	−24.7	−26.2	1.927	1.091	−43.4	−45.5	2.000	1.823	−8.8	−8.9	1.994	1.988	−0.3	−0.6
X_2	**2.997**	**0.332**	**−88.9**	**−88.9**	**2.923**	**0.266**	**−90.9**	**−91.1**	**2.910**	**0.243**	**−91.7**	**−91.9**	**3.000**	**0.032**	**−98.9**	**−98.9**	**2.999**	**0.272**	**−90.9**	**−90.9**
X_3	0.499	0.503	0.8	0.6	0.502	0.388	−22.8	−22.4	0.475	0.276	−41.7	−44.8	0.500	0.480	−4.0	−4.0	0.499	0.503	0.8	0.6

Notes. For each regression, b stands for coefficient estimates when no error was introduced, b' stands for coefficient estimates when error was introduced in X_2. bias$_1$ and bias$_2$ stand for relative magnitude of estimation bias, calculated using estimated and true coefficient values, respectively.

and can be used to mitigate bias in second-stage econometric estimations. We describe the general process, which researchers can follow to quantify and correct errors in their data sets. We then use simulations to show the effectiveness of SIMEX and MC-SIMEX methods.

4.1. Review of Bias Correction Methods

There have been at least five popular bias correction methods discussed in the research literature, including (1) instrumental variables, (2) method of moments, (3) likelihood-based methods, (4) regression calibration, and (5) simulation extrapolation (SIMEX).

The *instrumental variable approach* can be used to address all kinds of endogeneity issues in regression including measurement error. In a linear regression $Y = X\beta + \mathbf{Z}\boldsymbol{\gamma} + \varepsilon$, where X contains additive measurement error, i.e., $\hat{X} = X + e$, the regression model can be rewritten as $Y = \hat{X}\beta + \mathbf{Z}\boldsymbol{\gamma} + (\varepsilon - e\beta)$. Thus, the variable with error $\hat{X}$ is correlated with the error term, causing endogeneity. If the researcher can find an appropriate instrument, W, that is correlated with $\hat{X}$ but not with the error term, then a two-stage least squares (2SLS) estimator can be used to obtain the unbiased estimate of the coefficient of X.

Alternatively, if the researcher has accurate knowledge about the moments of measurement error and other variables in the econometric model, the unbiased coefficients may be recovered under some specifications, either analytically or numerically. This approach is known as the *method-of-moments approach*, or *functional approach* (Carroll et al. 2006). In linear regressions with only one regressor, this approach is very straightforward. If the values of σ_X^2 and σ_e^2 are known or can be estimated, one can easily calculate the corrected coefficient as $\hat{\beta}_1[(\sigma_X^2 + \sigma_e^2)/\sigma_X^2]$. In multivariate linear models or nonlinear models, one also needs knowledge of the covariance between the measurement error and other covariates.

Another option is the *likelihood-based method*, which involves explicit modeling of the error, that is, modeling the probability of observing the values of the dependent variable, given the values of the independent variables. Typically, to model this likelihood, researchers need to make distributional assumptions, such as the conditional distribution of a variable with error given its true values, and the distribution of the true values (Carroll et al. 2006). If such information is available, then the likelihood-based method can help recover unbiased estimates via maximum likelihood estimation.

There are also data-driven approaches such as *regression calibration*, which is a general-purpose bias correction method (Gleser 1990). Imagine that, for a subset of data, researchers can observe both the variable measured with error ($\hat{X}$) and its true value (X). Using this subset, it is then possible to fit a regression model of X on $\hat{X}$ and the other observed covariates ($\mathbf{Z}$), denoted as $f(\hat{X}, \mathbf{Z})$. For remaining data where X is not observable, it can be estimated via the model $f(\hat{X}, \mathbf{Z})$. Then, using the estimated values of X and other precisely measured covariates, the researcher can carry out the desired econometric analyses under an assumption that no measurement error remains. This method essentially views measurement error as a missing data problem. The true values for the variable with error are considered missing, and are imputed from a predictive model built on the subsample of data where true values are observed.

Finally, another general-purpose, data-driven approach to bias correction is *simulation extrapolation* or SIMEX (Cook and Stefanski 1994). As a simulation-based method, the SIMEX method has several advantages over the other methods, in dealing with measurement error and misclassification caused by data mining models. Compared to the first three methods outlined above, SIMEX requires relatively little information and fewer assumptions. For example, the instrumental variable approach requires the identification of an appropriate instrument. The method-of-moments approach requires knowledge of the moments of measurement error as well as the covariance between the error and other covariates. The likelihood-based method requires researchers to make distributional assumptions. By contrast, SIMEX requires only information on the variance of measurement error or a misclassification matrix, which is readily available from the performance evaluation measures of data mining models. Both the error variance and misclassification matrix can be calculated by comparing model predictions with true values using the test data set. Because the test set is typically a random subsample of the labeled data, the calculated error variance or misclassification matrix can be generalized to the broader, unlabeled data.

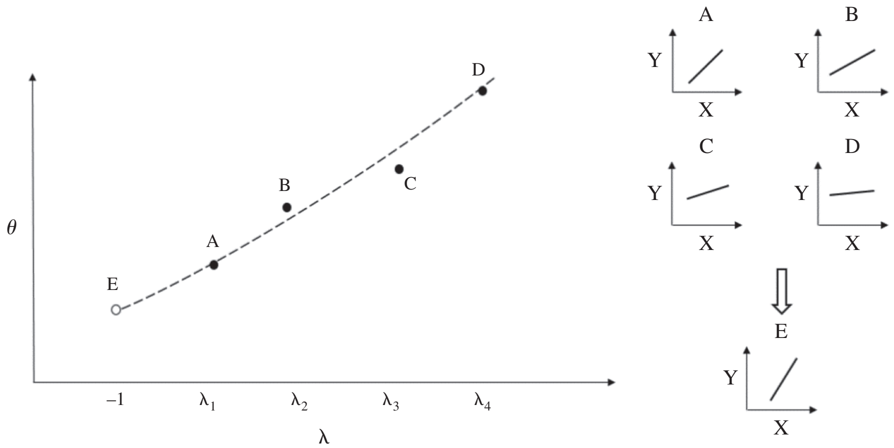
SIMEX demonstrated better performance than regression calibration under a number of scenarios, in particular under nonlinear econometric specifications. We experimented with both regression calibration and SIMEX for logit, probit, and Poisson regressions, where one independent variable was normally distributed ($\sigma_e = 0.3$) and contained classical measurement error (described in Section 3.3.1). The error biased the coefficient estimate from 2 down to 1.748, 1.567, and 1.760, respectively. SIMEX was able to correct the coefficient back to 1.999, 1.878, and 1.977, whereas regression calibration only corrected the coefficient back to 1.973, 1.815, and 1.927. Moreover, SIMEX requires less time and effort to execute because the procedure has been implemented in software packages that are commonly available for statistical analyses. For example, SIMEX is available in *R*, an open source statistical programming language, via the *simex* package, and also available in STATA, via the *simex* function (Hardin et al. 2003). Because of the above reasons, we focus on SIMEX as the primary correction method in this paper. Meanwhile, we encourage researchers to consider and evaluate multiple error-correction procedures including SIMEX to identify the best fit for their research setting, data, and data mining models, via the diagnostic procedure we outline in Table 4 (Section 4.3).

4.2. Introduction to SIMEX and MC-SIMEX

The SIMEX method was proposed by Cook and Stefanski (1994) to address additive measurement error in a *continuous* variable (i.e., $\hat{X} = X + e$) in models where the error variance σ_e^2 is known or can be accurately estimated. The SIMEX method consists of two steps: a simulation step and an extrapolation step. In the simulation step, a fixed set of nonnegative values $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ is selected (e.g., $\{1, 2, \ldots, m\}$). Then, multiple versions of $\hat{X}$ are generated as $\{\hat{X}(\lambda_1), \hat{X}(\lambda_2), \ldots, \hat{X}(\lambda_m)\}$, where $\hat{X}(\lambda_k) = X + e(\lambda_k)$, each with increasing error variance (specifically, $e(\lambda_k)$ has variance $(1+\lambda_k)\sigma_e^2$). In other words, the method simulates variables with increasingly larger measurement errors. Each $\hat{X}(\lambda_k)$ is associated with a set of coefficient estimates $\theta(\lambda_k)$. In the extrapolation step, a parametric model $\theta(\lambda)$ is estimated, which describes the relationship between the magnitude of the error and the coefficients. Then, extrapolating $\theta(\lambda)$ to $\theta(-1)$, one can approximate the coefficient estimates under zero measurement error (see Cook and Stefanski 1994 for more details). Figure 2 provides a graphical illustration of the SIMEX correction process. The parametric model $\theta(\lambda)$ may take several functional forms, including linear, quadratic, and nonlinear. Asymptotic methods have been proposed to estimate the standard errors for corrected coefficients following the application of the SIMEX method, including the delta (Carroll et al. 1996), jackknife (Stefanski and Cook 1995), and bootstrapping methods.

The MC-SIMEX method, an extension of the SIMEX method, was introduced by Küchenhoff et al. (2006) to accommodate misclassification in *discrete* variables when the misclassification matrix is known or can be estimated. It involves the same two basic steps as SIMEX. In the simulation step, $\hat{X}(\lambda_k)$ is generated by adjusting the values of $\hat{X}$ based on the λ_kth power of the misclassification matrix (see Section 3.3.3 for the procedure of adjusting values of $\hat{X}$ based on a given misclassification matrix). In the extrapolation step, a parametric function $\theta(\lambda)$ is estimated and extrapolated to $\theta(-1)$, to approximate coefficient estimates under conditions of zero misclassification. Küchenhoff et al. (2007) proposed an asymptotic standard error estimation method for MC-SIMEX. Online Appendix A3 provides the pseudocode for implementing both SIMEX and MC-SIMEX methods.

Figure 2. Graphical Illustration of the SIMEX Correction Process



Notes. In the simulation step, four versions of X with increasing error are generated. Each corresponds to a set of parameter estimates, marked by points A, B, C, and D. In the extrapolation step, a parametric model (as shown by the dotted curve) is fitted, and extrapolated to the case where no error is present, marked by point E. The subplots on the right show the changes in regression line (obtained during the second-stage econometric estimation) during the error-correction process.

For both classical measurement error and misclassification, SIMEX and MC-SIMEX can be directly applied, regardless of the second-stage model specifications. However, for nonclassical measure error that contains systematic error, SIMEX correction is unlikely to be effective. Take the error structure $\hat{X} = a + bX + e$ and $b \neq 1$ as an example. Although SIMEX can eliminate estimation bias caused by the random component e, it cannot fix the bias from the systematic component. To overcome this challenge, we propose a data *preprocessing* step in addition to the original SIMEX procedure. Because we can observe both X and $\hat{X}$ in the labeled training data used to build first-stage data mining models, we can fit a linear regression of $\hat{X}$ on X to obtain estimations $\hat{a}$ and $\hat{b}$. We can then generate a new variable: $\hat{X}' = (\hat{X} - \hat{a})/\hat{b}$, to reduce the nonclassical error structure to the classical form $\hat{X}' = X + e'$. From this relationship, we can calculate the modified error e' as the difference between $\hat{X}'$ and X and its standard deviation as σ'_e. Then, we can apply the standard SIMEX correction procedure, using $\hat{X}'$ as the (modified) variable with measurement error and σ'_e as the (modified) error standard deviation.

4.3. Diagnosing Error and Evaluating Correction Efficacy

Before error correction, researchers should first assess three things. First, it is important to understand the error's functional form. If the measurement error contains a systematic component, it may require special error-correction procedures such as the SIMEX procedure with data preprocessing we described in Section 4.2. Second, it is important to assess the severity of the bias in the second stage. While measurement error and misclassification may invalidate coefficient estimates and statistical inference, it is also possible to have trivial to minimal bias, which is of little concern. Third, it is important to evaluate the efficacy of the chosen error-correction methods. For example, both regression calibration and SIMEX are general-purpose methods applicable to many circumstances. Researchers should carefully compare the relative efficacy of each

Table 4. Procedure for Diagnosing Error and Evaluating Correction Efficacy

Error Diagnostics (Steps 1–4):

Step 1. Conduct planned second-stage econometric analysis on the labeled data set, using *true* labels.

Step 2. Conduct planned second-stage econometric analysis on the labeled data set, using *model-predicted* labels.

Step 3. If error is continuous, use true labels and model-predicted labels to estimate error functional form.

Step 4. Compare estimates from Steps 1 and 2 to understand the impact of measurement error, including but not limited to (1) the degree of bias, (2) the direction of bias, (3) changes in statistical significance, and (4) changes in model fit. Use the estimate from Step 3 to understand the characteristics of the continuous error.

Correction Diagnostics (Steps 5–6):

Step 5. Apply candidate error-correction methods (e.g., SIMEX) on the second-stage econometric model. Use the estimate from Step 3, if warranted, to modify the error-correction procedure(s) accordingly.

Step 6. Compare estimates from Steps 1, 2, and 4 to understand the efficacy of candidate error-correction methods, and choose the most effective error-correction method for actual analysis.

Table 5(a). SIMEX Correction for X_1 with Classical Measurement Error

	OLS			Logit			Probit			Poisson			Fixed effect		
	b	b'	b_{simex}	b	b'	b_{simex}	b	b'	b_{simex}	b	b'	b_{simex}	b	b'	b_{simex}
$\sigma_e = 0.1$															
X_1	1.995	1.970	1.992	1.996	1.954	1.988	1.979	1.903	1.960	2.000	1.911	1.933	1.994	1.977	1.996
$\sigma_e = 0.3$															
X_1	1.995	1.833	1.998	1.996	1.748	1.999	1.979	1.567	1.878	2.000	1.760	1.977	1.944	1.824	1.985
$\sigma_e = 0.5$															
X_1	1.995	1.595	1.946	1.996	1.453	1.910	1.979	1.155	1.591	2.000	1.337	1.646	1.944	1.589	1.944

correction procedure and choose the one that best fits their purposes.

In Table 4, we outline a basic procedure for diagnosing errors and choosing error-correction methods. Following the procedure, researchers can use the labeled data set from the first-stage data mining model to diagnose the functional form of the error, the severity level of bias, and the effectiveness of correction methods, because both the true values and model-predicted values of the variables are observed. Equipped with knowledge from the diagnostic procedure, researchers can proceed to actual analyses using the unlabeled data set and apply the chosen error-correction method. The increase in sample size using unlabeled data may help identify desired effects with greater power and more precision.

4.4. Using SIMEX and MC-SIMEX for Error Correction

To demonstrate the effectiveness of SIMEX and MC-SIMEX, we applied them to the simulated data from Section 3. For each model specification, we ran either SIMEX (for continuous measurement error in X_1) or MC-SIMEX (for discrete misclassification in X_2) and reported the corrected coefficient estimates associated with the variables containing measurement error or misclassification. The efficacy of both methods depends on an accurate estimation of the extrapolation function $\theta(\lambda)$. Through experiments with simulated and actual data, researchers have identified the quadratic and nonlinear extrapolation functions to be effective for a large number of model specifications (Cook and Stefanski 1994, Küchenhoff et al. 2006). We used the quadratic extrapolation function for all of our simulations. Researchers should experiment with alternative extrapolation functions to determine the one best suited to their situation.

Table 5(a) shows the correction results for classical measurement error models corresponding to our simulations in Section 3.3.1. Table 5(b) shows the results for nonclassical measurement error models, corresponding to the simulations in Section 3.3.2. For the systematic measurement error simulated in scenario (3), we applied the SIMEX procedure with preprocessing. Table 5(c) shows results for our discrete misclassification models, corresponding to simulations in Section 3.3.3. In all of the tables, the first two columns, respectively, contain coefficients without error and with error, denoted as b and b', and the third column contains the corrected estimation, denoted as b_{simex} in Table 5(a) and $b_{mcsimex}$ in Table 5(c). In Table 5(b), we report corrected estimates, obtained from standard SIMEX procedure both without and with data preprocessing, denoted as b_{simex} and b_{simex_pre}. All coefficients were statistically significant except those in parentheses.

Based on Tables 5(a) and 5(c), we can see that standard SIMEX and MC-SIMEX effectively reduced the bias in all regressions. In a number of cases, the correction procedure almost fully recovered the unbiased estimate. Even when misclassification was severe, such as in scenarios (2) and (3) in Table 5(c), MC-SIMEX enabled us to correct the coefficient of X_2 in the right direction, although the corrected coefficient was not statistically significant in the Poisson regressions. Our results from Table 5(c) also suggest that error-correction methods have limited effectiveness when

Table 5(b). SIMEX Correction for X_1 with Nonclassical Measurement Error

	Scenario (1)			Scenario (2)			Scenario (3)			
	b	b'	b_{simex}	b	b'	b_{simex}	b	b'	b_{simex}	b_{simex_pre}
$\sigma_e = 0.1$										
X_1	1.995	1.884	1.901	1.995	1.979	1.999	1.995	3.830	3.985	1.997
$\sigma_e = 0.3$										
X_1	1.995	1.641	1.757	1.995	1.863	2.036	1.995	2.903	3.739	1.866
$\sigma_e = 0.5$										
X_1	1.995	1.412	1.634	1.995	1.667	2.057	1.995	1.941	2.922	1.456

Table 5(c). MC-SIMEX Correction for X_2 with Misclassification

	OLS			Logit			Probit			Poisson			Fixed effect		
	b	b'	$b_{mcsimex}$	b	b'	$b_{mcsimex}$	b	b'	$b_{mcsimex}$	b	b'	$b_{mcsimex}$	b	b'	$b_{mcsimex}$
Scenario (1): $X_2 \sim$ Bernoulli(0.3), $M_{00} = 0.8$, and $M_{11} = 0.8$															
X_2	2.988	1.596	2.557	2.890	1.106	1.860	2.897	0.979	1.648	3.000	1.722	2.669	2.986	1.583	2.543
Scenario (2): $X_2 \sim$ Bernoulli(0.3), $M_{00} = 0.6$, and $M_{11} = 0.5$															
X_2	2.988	0.282	0.756	2.890	0.304	0.800	2.897	0.148	0.391	3.000	0.205	(0.565)	2.986	0.271	0.733
Scenario (3): $X_2 \sim$ Bernoulli(0.5), $M_{00} = 0.6$, and $M_{11} = 0.5$															
X_2	2.997	0.332	0.904	2.923	0.266	0.774	2.910	0.243	0.632	3.000	0.032	(0.065)	2.986	0.272	0.723

the performance of the data mining model is poor. In these cases, researchers should focus on improving predictions first, and only deploy the correction methods as a secondary, remedial action.

Results in Table 5(b) show that SIMEX was also effective for nonclassical measurement error. When measurement error was correlated with the true value of X_1, as in scenario (1), SIMEX corrected the coefficient of X_1, although the correction was not as good as in the case of independent error. When measurement error was correlated with X_3, as in scenario (2), SIMEX corrected the coefficients of both X_1 and X_3. When there was systematic error as in scenario (3), SIMEX correction without preprocessing failed and actually exacerbated the bias in the coefficient of X_1, moving it further away from its true value. However, applying SIMEX after our proposed preprocessing successfully corrected the coefficient on X_1, for $\sigma_e = 0.1$ and $\sigma_e = 0.3$. For $\sigma_e = 0.5$, SIMEX with preprocessing also performed better than SIMEX without preprocessing, though it is worth noting that the corrected coefficient (1.456) was still further from the true value (1.995) than the original "biased" estimate (1.941). This marks an important situation under which the SIMEX method may be not only ineffective, but detrimental. When continuous measurement error contains both a systematic component and a random component with large variance, the two combined can result in a smaller "net" bias than each component alone. Under this special scenario, error correction is incapable of resolving the bias. If the researcher first employs the diagnostic procedure outlined in Table 4, it would be possible to observe whether the chosen correction procedure is improving estimates, or in fact making matters worse.

Another important observation from Tables 5(a)–5(c) is that the effectiveness of SIMEX and MC-SIMEX corrections vary with (1) the amount of error and (2) the model specification. As the amount of measurement error or misclassification increases, the correction generally becomes less effective, i.e., the corrected coefficients shift further away from the true coefficients. Additionally, corrections for linear and logit models appear generally more effective than corrections for probit and Poisson models.

To understand the effectiveness of SIMEX and MC-SIMEX corrections under a wider array of circumstances, we conducted additional, more comprehensive simulation studies. We extended the simulation studies described above by systematically varying the distributions and variances of the precisely measured covariates (i.e., X_2 and X_3 for simulations of measurement error in X_1; X_1 and X_3 for simulations of misclassification in X_2). The results of these additional simulations can be found in Online Appendix A8. Based on these additional simulations, we were able to further validate our aforementioned observations. First, SIMEX and MC-SIMEX are able to mitigate the biases in almost all cases. Importantly, as the amount of error increases, the magnitude of bias generally becomes larger, and the correction tends to become less effective. Second, corrections for linear and logit models appear to be more effective than corrections for probit and Poisson models. Third, the effectiveness of corrections also depends on the distributions and variances of the error-free covariates. However, it is difficult to provide theoretical, a priori predictions about the correction's effectiveness for situations that we have not considered here. Accordingly, we would caution researchers to adopt the diagnostic procedure described in Table 4 to understand the nature of the error in their particular data set, for their particular data mining model and regression specifications, and thereby assess the efficacy of any correction procedures in their unique empirical contexts.

5. Application to Field Data: Three Real-World Data Sets

In this section, we apply SIMEX and MC-SIMEX methods to three real-world data sets. The three examples cover a variety of data types, model specifications, and research questions that are commonly seen in IS research. We use the first two examples to demonstrate the effectiveness of SIMEX and MC-SIMEX. We use the third example to illustrate a scenario under which the SIMEX correction is *not* effective, because of extremely poor performance of the predictive data mining model. In all three examples, we follow the diagnostic procedure outlined in Table 4, which helps to ascertain whether error correction is effective.

5.1. Review Helpfulness on TripAdvisor.com

In the first example, we apply the MC-SIMEX method to a real-world data set of online reviews from TripAdvisor.com. We examine the relationship between textual sentiment and perceived helpfulness, employing the two-stage approach of combining data mining and econometric modeling. We first built a textual classification model to predict the sentiment of written reviews as either positive or negative, and then estimated two econometric models controlling for several other factors. We drew on the star rating of a review as the ground truth for its sentiment.

5.1.1. Research Setting. There is an extensive body of literature on online reviews in the IS discipline. Researchers have investigated the effects of various review characteristics such as volume, valence, and reviewer identity on consumer behaviors (Mudambi and Schuff 2010, Forman et al. 2008, Dellarocas 2003). Some studies in this domain have also combined data mining with econometric analysis. Archak et al. (2011), for example, built a text classification model to identify product features from consumer product reviews, and then estimated the impact of specific product features on product sales.

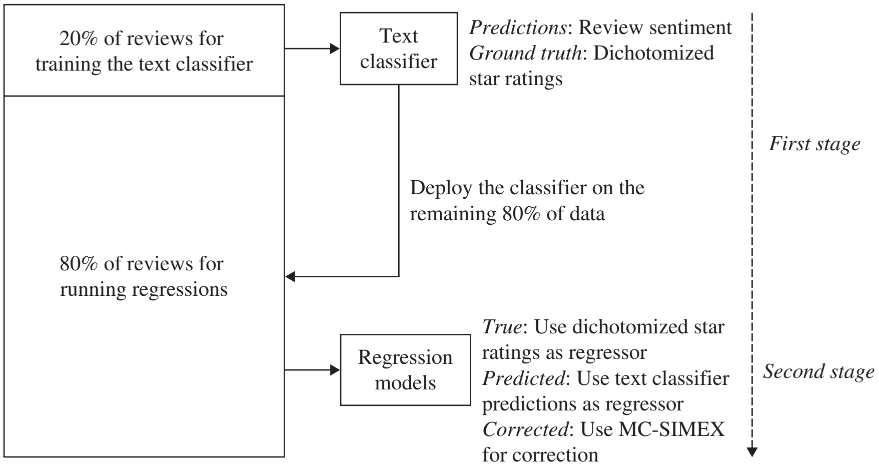
TripAdvisor.com is a travel-related website that hosts consumer reviews of service providers. Users can post reviews about hotels, restaurants, or resorts. Reviewers provide an overall rating on a five-star scale and, optionally, ratings on separate dimensions of the consumption experience. For example, reviewers can rate a hotel based on its price, service, or overall quality. Readers of a review can indicate its "helpfulness" by casting a vote. As a prominent site for consumer reviews, TripAdvisor.com has been examined in several studies (e.g., Huang et al. 2016, 2017; Mayzlin et al. 2014).

We collected 11,953 English-language reviews for 234 randomly selected U.S. restaurants. For each review, we gathered data on its textual content, star rating, the number of helpful votes it received, whether the review contained a photo, and the number of reviews posted prior to the focal review, which indicated the review's position in the sequence of all reviews for a restaurant. Using this data set, we examined the impact of review sentiment on perceived helpfulness. Figure 3 shows the two-stage process.

5.1.2. First Stage: Text Classification of Sentiment. To identify the sentiment of a review, we analyzed its textual content using natural language processing and textual classification techniques. In general, it is unnecessary to perform sentiment analysis when the star rating associated with a review is available, because a high rating often corresponds to positive sentiment and a low rating corresponds to negative sentiment. However, many online venues host consumer opinions and word of mouth as text, without the benefit of numerical ratings (e.g., Godes and Mayzlin 2004). In such a setting, researchers typically hire a team of human coders to manually label the sentiment of a small, random sample of text from a large data set. Using this labeled sample, one can then train a classifier and deploy it to classify the sentiment of the remaining unlabeled text. In this example, we treated the star rating of each review as the ground truth of its sentiment, for the purpose of training and evaluating a sentiment classifier that is based *only* on the textual content of reviews. Doing so allows us to quantify the misclassification and to illustrate the bias introduced in the second-stage econometric model because of error. If the reviewer gave a restaurant three or fewer stars, we coded the review as negative. If the reviewer gave four or five stars, we coded the review as positive. Using these criteria, 79% of the reviews in our sample were coded as positive and 21% were coded as negative, indicating a skewed distribution.

We followed standard practices in training the text classifier. First, we randomly selected 20% of the

Figure 3. Overview of the Two-Stage Process in Studying Review Helpfulness


20% of reviews for training the text classifier
Text classifier
Predictions: Review sentiment
Ground truth: Dichotomized star ratings
First stage
Deploy the classifier on the remaining 80% of data
80% of reviews for running regressions
Regression models
True: Use dichotomized star ratings as regressor
Predicted: Use text classifier predictions as regressor
Corrected: Use MC-SIMEX for correction
Second stage

original sample (i.e., 2,391 reviews) as the labeled data set for training and evaluating the performance of the model. Second, we followed standard natural language processing procedures (e.g., Jurafsky and Martin 2008) to convert each review into a word vector, in several steps. We transformed all text to lowercase, tokenized the text of each review into words, removed stop words, conducted stemming, and extracted bigrams and trigrams. We then applied the TF-IDF (term frequency-inverse document frequency) weighting scheme to rescale the word vector frequencies of occurrence (Jurafsky and Martin 2008). Third, we built a classifier using the linear support vector machine (SVM) technique (Vapnik 1995), and evaluated the classifier using fivefold cross validation. Our classifier achieved 93.03% precision and 92.93% recall for the positive class, and 73.97% precision and 74.28% recall for the negative class. This performance corresponds to the following misclassification matrix: $(M_{00}, M_{10}, M_{01}, M_{11}) = (0.74, 0.07, 0.26, 0.93)$. Finally, we deployed the trained classifier on the remaining, unlabeled sample, i.e., on 9,562 reviews. In the end, every review in the unlabeled data set had a predicted sentiment.

5.1.3. Second Stage: Econometric Analysis of the Impact of Sentiment on Perceived Helpfulness.

The dependent variable, *helpfulness*, was coded as a dummy variable indicating whether a review received any helpful votes. The independent variable, *sentiment*, was set to 1 if the review was positive and 0 if it was negative. We also included several control variables including (1) *photo*, a dummy variable indicating whether the review had a photo or not; (2) *words*, the number of words in the review; and (3) *sequence*, the number of reviews posted about a restaurant before the focal review. We estimated two models, as illustrated in the equations below: a linear probability model (LPM) and a logit model

$$\text{LPM:}\quad \mathit{helpfulness} = \beta_0 + \beta_1 \mathit{sentiment} + \beta_2 \mathit{photo} + \beta_3 \log(\mathit{words}) + \beta_4 \mathit{sequence} + \varepsilon,$$

$$\text{Logit:}\quad \text{Logit}(\mathit{helpfulness}) = \beta_0 + \beta_1 \mathit{sentiment} + \beta_2 \mathit{photo} + \beta_3 \log(\mathit{words}) + \beta_4 \mathit{sequence} + \varepsilon.$$

Before carrying out the actual regression analysis and the MC-SIMEX correction, we followed the diagnostic procedure outlined in Table 4 by running the two regressions on our 20% labeled data ($N = 2{,}391$). We used fivefold cross-validation to evaluate our first-stage SVM model. For each fold, we obtained the predicted sentiment label from the SVM model built off the other four folds. Our diagnostic analyses showed that misclassification in *sentiment* attenuated its effect on *helpfulness*, and that MC-SIMEX was effective in correcting the bias. We include these diagnostic results in Online Appendix A4. Table 6 shows our actual estimations, performed on the sample of 9,562 reviews.[9] For each model, we report three sets of results. The first column, labeled as "True," reports estimates obtained using the "true" values of the sentiment based on star ratings. The second column, labeled as "Predicted," reports estimates obtained using predicted sentiment from our text classifier. The third column, labeled as "Corrected," reports corrected estimates, by applying the MC-SIMEX method. We have provided the *R* code that was used to conduct the MC-SIMEX correction, in Online Appendix A5.

As shown in Table 6, sentiment was negatively associated with review helpfulness. Compared to positive reviews, negative reviews were more likely to receive helpful votes. In addition, reviews that contained photos were less likely to be perceived as helpful and

Table 6. Regression Results and Corrections of the TripAdvisor.com Data Set ($N = 9{,}562$)

	LP model			Logit model		
	True	Predicted	Corrected	True	Predicted	Corrected
Intercept	0.1707***	0.1538***	0.1763***	−1.5750***	−1.6703***	−1.5644***
	(0.0105)	(0.0111)	(0.0166)	(0.0665)	(0.0715)	(0.0965)
Sentiment	**−0.0693*****	**−0.0463*****	**−0.0684*****	**−0.4240*****	**−0.2843*****	**−0.3854*****
	(0.0097)	**(0.0099)**	**(0.0161)**	**(0.0611)**	**(0.0629)**	**(0.0934)**
Photo	−0.0167*	−0.0174*	−0.0158*	−0.1149*	−0.1203*	−0.1100
	(0.0077)	(0.0077)	(0.0070)	(0.0578)	(0.0580)	(0.0568)
Words	0.7986***	0.7893***	0.7282***	4.3924***	4.3185***	3.9686***
	(0.0494)	(0.0510)	(0.0664)	(0.3011)	(0.3114)	(0.3715)
Sequence	−0.0010	−0.0049	−0.0040	0.0095	−0.0159	−0.0132
	(0.0166)	(0.0166)	(0.0174)	(0.1114)	(0.1114)	(0.1161)
Log likelihood	−4,408.24	−4,422.64		−4,470.65	−4,483.95	
AIC	8,828.5	8,857.3		8,951.3	8,977.90	

Notes. The MC-SIMEX method does not provide log likelihood or AIC statistics. Standard errors are in parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

longer reviews were more likely to be perceived as helpful. *Sequence* did not have a significant relationship with helpfulness. These findings were generally consistent with those of prior work (e.g., Mudambi and Schuff 2010, Yin et al. 2014), which indicates that our second-stage model specification was appropriate and valid.

Comparing the "Predicted" regressions with the "True" regressions show that the misclassification in the predicted sentiment considerably biased the estimation, as expected. The coefficient associated with *sentiment* in the "Predicted" estimation was only two-thirds the magnitude of the coefficient in the "True" estimation (i.e., the estimation based on dichotomized star ratings). Had we relied directly on the *sentiment* variable generated by the data mining model and ignored the misclassification, we would have greatly underestimated the magnitude of the effect of review sentiment on perceived helpfulness. The presence of misclassification in the *sentiment* variable also biased the other coefficient estimates, to various degrees. We also observed that the "Predicted" regressions exhibited worse model fit than the "True" regressions, assessed based on the log likelihood and AIC. Overall, the analysis proved the effectiveness of MC-SIMEX in correcting estimation bias from misclassification. This is particularly true in the LP model, where MC-SIMEX almost perfectly recovered the true, unbiased coefficient estimate for *sentiment*.

To assess the impact of sample size on correction effectiveness, we repeated the above analyses for three random samples of 500, 2,000, and 5,000 observations. We observed three notable patterns. First, for each sample size, MC-SIMEX was able to mitigate the bias on *sentiment*. Second, as the sample size increased from 500 to 5,000, the relative magnitude of bias decreased and the effectiveness of correction increased. This indicates that a sufficiently large sample is necessary to obtain both precise estimations and good correction outcomes. Third, further increasing sample size from 5,000 to 9,562 (i.e., the full sample) did not reduce the relative magnitude of bias, but did produce better corrected coefficients for *sentiment*. This suggests that having an increasingly larger sample does not eliminate bias, but generally does benefit error correction. The results of these additional analyses are included in Online Appendix A6.

5.2. User Engagement on Facebook Business Pages

In the second example, we applied MC-SIMEX to another real-world data set on user-generated posts on Facebook business pages. We examined the relationship between post sentiment and user engagement with a post, measured as the number of comments the post had received. We first built a textual classification model to predict the sentiment of posts as either positive or negative, and then estimated two econometric specifications controlling for several other factors.

Facebook business pages is a feature that Facebook launched in 2007, which enables companies to interact with their customers on Facebook. Organizations use Facebook business pages primarily for marketing purposes by posting information about their products and services, offering coupons, as well as encouraging consumers to share positive word of mouth (Goh et al. 2013). Visitors of the business page can engage with both marketer-generated and user-generated posts through liking, commenting, or sharing (Goh et al. 2013). In this example, we examine how the sentiment of a user-generated post affects the number of comments it receives. We gathered 8,059 user-generated posts, all of which were created in 2012, from the Facebook business pages of 39 consumer-oriented Fortune-500 companies such as airlines, banks, and retailers. For each post, we collected its textual content, poster ID, and the number of comments the post attracted. We hired Amazon Mechanical Turk workers to label the sentiment of the posts. We had five independent workers code each post, and used the majority (modal) rule to determine the sentiment. In total, 2,751 posts were labeled as positive, and 5,308 posts were labeled as negative. These manually labeled sentiments served as the ground truth for building the sentiment classifier, and for validating the performance of the MC-SIMEX correction procedure in our second-stage estimation.

We built our sentiment classifier using a random sample of 10% of the labeled data (806 posts). We followed standard procedures in building the text classifier, as described in Section 5.1.2. The classifier was built using the linear SVM technique, and evaluated using fivefold cross-validation. Our classifier achieved 84.21% precision and 81.45% recall for the positive class (denoted as class 1), and 90.56% precision and 92.10% recall for the negative class (denoted as class 0). This performance corresponds to the following misclassification matrix: $(M_{00}, M_{10}, M_{01}, M_{11}) = (0.92, 0.19, 0.08, 0.81)$. We then deployed the trained classifier on the remaining 90% of our labeled sample (7,253 posts), and included the predicted sentiment in the second-stage econometric analysis.

In our econometric analysis, we examined the relationship between post sentiment and user engagement. The dependent variable, *comments*, was the number of comments each post received. The independent variable, *sentiment*, was coded as 1 if the post was positive and 0 if the post was negative. We controlled for several factors that may affect the level of engagement with a post including (1) log(*Words*), the log-transformed word count of each post; (2) *User Activeness*, the posting user's level of activeness, measured as the total number of posts that the user had created on the business page where the focal post appeared in 2012; (3) log(*Popularity*), the popularity level of the page on which the focal post was published, measured as the total number of user posts on the page in 2012; and (4) *Type*: the media type of the focal post assigned by Facebook such as link, photo, video, or status. Because our dependent variable is a count measure, we ran both linear regression and Poisson regression.

Next, we followed the procedure in Table 4 to conduct a diagnostic analysis before the second-stage estimation. We ran the proposed regressions with 10% of our labeled data ($N = 806$). The diagnostic analysis showed that misclassification in *sentiment* attenuated its effect on *comments*, and MC-SIMEX was able to correct the bias (detailed results in Online Appendix A7). We then conducted the regression analysis with the remaining 7,253 posts. Table 7 shows the regression results and corrected coefficients.

According to Table 7, positive posts received fewer comments than negative posts. Because of misclassification in *Sentiment*, all coefficient estimates were biased to various degrees and in different directions. The most important thing to note is that MC-SIMEX effectively mitigates the estimation biases in both OLS and Poisson models. For linear regression, MC-SIMEX almost fully recovered the unbiased coefficient of *Sentiment*. For Poisson regression, there was a slight overcorrection, i.e., the absolute value of the corrected coefficient of *Sentiment* was greater than its unbiased value, but the corrected estimate was still closer to the true value than the biased coefficient.

5.3. Campaign Organizer Age and Crowdfunding Outcomes

Our third and final example demonstrates the application of SIMEX correction to a real-world data set of crowdfunding campaign outcomes from a leading reward-based crowdfunding website (Agrawal et al. 2014). We examined the relationship between the age of a fundraising campaign organizer and the amount of money she was able to raise. We collected campaign organizers' profile pictures and used a third-party face recognition service to infer the age of the persons in those pictures. The predicted age was not all accurate and contained measurement error. We used user's self-reported age, which we obtained from the platform operator, as the ground truth. We estimated a linear regression model controlling for several other factors. While we chose the first two examples to demonstrate the effectiveness of our proposed error-correction methods, we chose this third example to show the limitations and boundary conditions of the methods, i.e., their effectiveness depends on a reasonable level of performance of the data mining models.

In recent years, crowdfunding has garnered a great deal of attention within the IS community (Burtch et al. 2013, 2015). On reward-based crowdfunding platforms like Kickstarter and Indiegogo, individuals can launch fundraising campaigns to raise money from the crowd to finance a project, a cause, or a venture. The money may be used to fund a new product or service or to support public goods and charitable endeavors. For each campaign, the organizer sets a fixed amount of money to be raised, and a fixed duration

Table 7. Regression Results ($N = 7{,}253$)

	OLS model			Poisson model		
	True	Predicted	Corrected	True	Predicted	Corrected
Intercept	−2.9475***	−3.2001***	−2.9373***	−2.2148***	−2.3401***	−2.1591***
	(0.4804)	(0.4810)	(0.4878)	(0.1053)	(0.1053)	(0.2750)
Log(Words)	0.5840***	0.6177***	0.5470***	0.3412***	0.3551***	0.3049***
	(0.0408)	(0.0420)	(0.0469)	(0.0087)	(0.0089)	(0.0319)
Activeness	0.0303***	0.0302***	0.0305***	0.0108***	0.0109***	0.0110***
	(0.0045)	(0.0045)	(0.0045)	(0.0005)	(0.0005)	(0.0016)
Log(Popularity)	0.3107***	0.3151***	0.3195***	0.1753***	0.1782***	0.1824***
	(0.0470)	(0.0471)	(0.0471)	(0.0102)	(0.0102)	(0.0258)
Type = Link	−0.7544*	−0.7428*	−0.7504*	−0.5955***	−0.6051***	−0.6081*
	(0.3335)	(0.3343)	(0.3341)	(0.0994)	(0.0994)	(0.2383)
Type = Photo	0.0265	−0.1490	−0.0451	−0.0065	−0.1307	−0.0462
	(0.2679)	(0.2669)	(0.2684)	(0.0699)	(0.0693)	(0.1342)
Type = Video	−1.0167	−1.0116	−0.9993	−0.7607*	−0.8191**	−0.8086
	(0.9282)	(0.9304)	(0.9302)	(0.3017)	(0.3017)	(0.5908)
Sentiment	**−0.7356***	**−0.4789***	**−0.7132***	**−0.5724***	**−0.4047***	**−0.6307***
	(0.0987)	**(0.1041)**	**(0.1333)**	**(0.0251)**	**(0.0262)**	**(0.1002)**
Log likelihood	−19,559	−19,576		−12,701	−12,859.5	
AIC	39,133	39,167		35,653	35,970	

Notes. The MC-SIMEX method does not provide log likelihood or AIC statistics. Standard errors are in parentheses.

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

for the fundraising. A campaign is deemed a success if the fundraising goal is reached or surpassed within the specified duration. In this example, we examine the following research question: How does the age of a campaign organizer affect a campaign's fundraising success? Although age information is not directly available on the website, it can be inferred from organizers' profile pictures. We gathered information on 1,368 crowdfunding campaigns, each with a unique organizer who had uploaded a high-quality profile picture. For each campaign, we collected data on its beginning and end dates, the fundraising goal, the amount of money it collected by the end of the campaign, and whether it had been featured on the homepage of the crowdfunding website. We also had access to self-reported demographic information for each campaign organizer, including gender and year of birth. We used the year of birth to calculate an organizer's actual age at the time of our data collection, which was used as the ground truth for the age variable. Next, we replicated the two-stage approach of combining data mining with econometric analysis.

In the first stage, we downloaded the profile pictures of the 1,368 campaign organizers. We used the Microsoft Face API,[10] a third-party face recognition service, to automatically infer the age of each organizer based on their profile picture. There were 63 profile pictures that contained more than one person. In those cases, we took the average of the predicted ages of all individuals appearing in the photo. Having both true and predicted ages for each organizer, we estimated the measurement error structure in the form of $\hat{X} = a + bX + e$, where $\hat{X}$ was the predicted age and X was the true age. We estimated the error structure on 30% of our sample, i.e., 410 randomly selected campaign organizers. This was done to mimic the reality that a researcher typically only has a small subsample of labeled data in practice. This analysis indicated that $\hat{a} = 18.78$, $\hat{b} = 0.36$, $\mathrm{SD}(e) = 9.96$, which signaled very high levels of both systematic error and random error. Using error measures in data mining, such error corresponded to a MAE value of 10.58 and a RMSE value of 14.14 (in years). In the context of age recognition, aside from the inherent difficulty of estimating one's age based on a photo, there were other sources of measurement error such as cosmetic or photo-retouching effects, the use of someone else's photos, or the use of photos from a younger age.

In the second stage, we fit a linear regression to examine the relationship between organizer age and campaign outcomes. The dependent variable, *percent*, is the percentage of the fundraising goal achieved by the end of a campaign. This variable can be greater than 1 if a campaign raised more than its fundraising goal. The independent variable, *age*, is either the true value or predicted value of the organizer's age. We included three control variables: (1) *gender*, representing the organizer's self-reported gender; (2) *featured*, a dummy variable indicating whether the campaign had been featured on the platform homepage at any point during the course of fundraising; and (3) *duration*, representing the number of days from the beginning to the end of the campaign. Given the existence of both systematic and random error components in measurement error, we used the SIMEX procedure with data preprocessing that we proposed in Section 4.

Before the second-stage regression analysis and the SIMEX correction, we followed the diagnostic procedure outlined in Table 4, by running the regression on the random subsample of 30% of our data, which was used to understand error functional form. The first four columns of Table 8 show the results of our diagnostic analysis, respectively, the true coefficients, the predicted coefficients, and two corrected estimations, from SIMEX procedures without and with preprocessing. Although we used SIMEX with preprocessing for

Table 8. Regression Results for Diagnostic ($N = 410$) and Actual Analysis ($N = 1{,}368$)

	Diagnostic analysis				Actual analysis			
	True	Predicted	Corrected (no preprocess)	Corrected (preprocess)	True	Predicted	Corrected (no preprocess)	Corrected (preprocess)
Intercept	0.4006^{***} (0.0805)	0.2826^{***} (0.0683)	0.2392^{*} (0.1076)	0.2972^{***} (0.0486)	0.4017^{***} (0.0415)	0.2542^{***} (0.0338)	0.2442^{***} (0.0488)	0.2573^{***} (0.0243)
age	**−0.0020** **(0.0018)**	**0.0013** **(0.0019)**	**0.0027** **(0.0034)**	**0.0007** **(0.0010)**	**−0.0035***** **(0.0009)**	**0.0003** **(0.0009)**	**0.0006** **(0.0015)**	**0.0002** **(0.0005)**
gender = male	-0.1190^{**} (0.0405)	-0.1190^{**} (0.0406)	-0.1219^{**} (0.0400)	-0.1207^{**} (0.0399)	-0.0598^{**} (0.0205)	-0.0557^{**} (0.0208)	-0.0567^{**} (0.0211)	-0.0565^{**} (0.0211)
featured = yes	1.0676^{***} (0.1155)	1.0634^{***} (0.1159)	1.0571^{**} (0.3532)	1.0608^{**} (0.3536)	0.8320^{***} (0.0535)	0.8291^{***} (0.0538)	0.8288^{***} (0.0538)	0.8288^{***} (0.0538)
duration	-0.0005^{*} (0.0002)	-0.0006^{**} (0.0002)	-0.0006^{***} (0.0001)	-0.0006^{***} (0.0001)	-0.0004^{***} (0.0001)	-0.0004^{***} (0.0001)	-0.0004^{***} (0.0001)	-0.0004^{***} (0.0001)
R-squared (%)	19.55	19.4			17.24	16.44		

Notes. The SIMEX method does not provide *R*-squared statistics. Standard errors are in parentheses.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

error correction, we also included the corrected coefficients from SIMEX without preprocessing for comparison purposes. We observed that, while the true relationship between age and fundraising outcomes was negative, measurement error caused the sign to flip to positive. Furthermore, our results suggest that, without data preprocessing, SIMEX correction failed to mitigate the bias. Interestingly, even with preprocessing, the SIMEX procedure was incapable of recovering the correct sign of the age variable. Diagnostic analysis suggests that the level of measurement error from the image classifier was too high to be corrected by SIMEX. We next carried out the second-stage analysis on the remainder of our data, and the results are shown in the last four columns of Table 8.

We observed similar results from the second-stage analysis. After controlling for other factors, age was negatively associated with the percentage of funding goal achieved. More specifically, an increase of 10 years in organizers' age can reduce fundraising by approximately 3.5%. Compared to younger people, older people were less likely to achieve their fundraising goals. Because of severe measurement error in age, the coefficient estimate of *age* became positive and insignificant when predicted age was included as the regressor. As such, measurement error introduced by the image classifier would, in this case, cause researchers to completely miss the effect of age. SIMEX correction without preprocessing further increased the bias, because it did not account for the systematic error component. Despite our best effort, even with preprocessing, SIMEX correction failed to recover the correct sign and significance of the coefficient on *age*, although it produced a coefficient that was closer to the unbiased value.

Our third example demonstrates an important lesson and insight about both the SIMEX method and error-correction procedures in general. When data mining models perform poorly and generate predictions with severe error, it is very challenging to fully recover the sign and statistical significance of the true coefficient using any error-correction method. In this example, our posthoc assessment suggests that the true age is only weakly correlated with machine-detected age ($\rho = 0.33$). In other words, we would like to note that error-correction methods do not have unlimited capability of uncovering the correct signal from any amount of error. Knowing this, researchers should prioritize reducing error and improving performance in their first-stage data mining models, rather than rely primarily on error corrections.

6. Discussion and Conclusions

An increasing use of data mining and econometrics as a two-stage analysis process provides many new opportunities for IS research. In the first stage, a wide variety of data mining techniques equip researchers with the tools to classify unstructured data, such as text or images, and to gather information that is not directly observable, such as sentiment. The output of these models can be subsequently incorporated into the second-stage econometric estimations to test hypotheses and make inferences. This combined approach, however, has potential pitfalls. In particular, this practice introduces challenges to statistical inference because of the well-known issues of measurement error or misclassification, which may compromise researchers' ability to draw robust conclusions. As we have demonstrated both analytically and empirically, ignoring measurement error or misclassification is likely to severely bias econometric estimations. We have also shown that (1) even a relatively low level of measurement error or misclassification from data mining models can result in substantial biases in subsequent econometric estimations, and (2) the biases are harder to anticipate when the error structures or econometric models grow more complex. This issue is particularly concerning, given the increasing focus on the magnitude of coefficients (i.e., the economic significance) in empirical studies, over and above mere statistical significance (Lin et al. 2013).

Fortunately, standard practices in data mining involve the evaluation of model performance using test data sets and provide established ways to quantify measurement error or misclassification. With this information, we can take actions to mitigate biases in the second-stage estimation. In this commentary, we reviewed several error-correction methods and focused on two simulation-based methods, SIMEX and MC-SIMEX, as promising remedies to correct for biases from measurement error and misclassification. We illustrated their effectiveness using both comprehensive simulations and three real-world data sets. In most cases, biases were reduced and the corrected coefficients were closer to the true values. In some cases, the corrected coefficients almost perfectly recovered their true values. We also identified two situations under which the effectiveness of error-correction methods may be either limited or unnecessary. First, when the level of measurement error or misclassification is very high, SIMEX or MC-SIMEX are not powerful enough to correct the bias. Solely relying on these methods could lead researchers to draw incorrect conclusions. Second, when continuous measurement error contains both a systematic component and a random component with relatively large variability, it can sometimes result in little bias in coefficient estimation. Using SIMEX in this situation could therefore be unnecessary or even detrimental. Finally, note that error-correction methods cannot account for biases caused by misspecification in the second-stage econometric model. For example, we simulated scenarios where the regression model had omitted variable bias, besides measurement error. The

SIMEX method was able to correct for the bias because of the measurement error, but it had no way of identifying the existence of omitted variable bias.

In addition to causing biases in regression coefficient estimates, measurement error and misclassification in independent variables can affect several other important aspects of econometric analysis, including confidence interval estimation, goodness-of-fit calculation, and hypothesis testing. For a linear regression where one of the independent variables contains *classical* measurement error, several asymptotic results are known in the literature. First, estimation of the error variance σ_ε^2 will be inconsistent (Wansbeek and Meijer 2000). More specifically, the sample estimate s_ε^2 will exceed the true value σ_ε^2 in the limit. As a result, the standard error for each regressor will also be overestimated in the limit. Consequently, the corresponding confidence interval will be wider than it should be, and the corresponding p-value will be larger than it should be. In general, classical measurement error in linear regressions makes OLS estimators more conservative. Second, as a direct ramification of the overestimation of error variance, R^2 is biased toward zero, indicating worse model fit. Third, the reliability of hypothesis testing may become questionable. For example, the commonly used F-statistic is biased toward zero, which means the null hypothesis that every coefficient is zero is not rejected often enough. In the case of misclassification, although limited theoretical results are available in the literature, our simulations showed that the presence of misclassification can inflate estimation of model error variance and can result in decreased model fit. Although we primarily focus on demonstrating and correcting biases in coefficient estimates in this commentary, we believe that readers should be aware of other consequences of error.

This commentary highlights both the opportunities and potential pitfalls of combining data mining and econometric modeling. Given the growing prevalence of the integrated approach, we hope to raise awareness of the fact that failing to account for measurement error or misclassification, which arises from the data mining process, could result in misleading findings. We chose SIMEX and MC-SIMEX as exemplary error-correction methods because they are easy to parameterize by using performance metrics from the data mining process and because they can be applied to a variety of econometric models. However, we do not claim that these two methods are superior to other error-correction methods in all situations. Instead, we acknowledge there are situations where other methods may be more appropriate. For example, the regression calibration method has been shown to produce consistent estimates for linear models (Carroll et al. 2006), and the instrumental variable approach can be used when valid instruments are available. We encourage researchers to evaluate and adopt error-correction methods on a case-by-case basis, depending on the nature of their data and research setting. We provide a diagnostic procedure to help researchers assess and deal with measurement error in their research practice. We propose that researchers use the labeled data set from first-stage data mining to diagnose the structure of the error, the severity of resulting bias, and the effectiveness of available correction methods. Conducting these diagnostic analyses before applying the error-correction procedure can help the researcher fully understand and address the issue.

Our commentary provides a first step toward addressing the challenges with measurement error from combining data mining techniques with econometric analysis. There are several promising avenues for future work. The first future direction is to continue improving existing error-correction methods. When applying the SIMEX and MC-SIMEX methods, we occasionally observed cases in which the coefficients of precisely measured (error-free) covariates were slightly overcorrected or shifted in the opposite direction. Although the mathematical underpinnings of the correction methods in no way would suggest that this result is a systematic or asymptotic property (but rather is a finite sample property), researchers should be aware of this potential issue. Future research should continue to improve the stability and robustness of the SIMEX and MC-SIMEX methods. Second, there are challenging scenarios when several variables, potentially of different types, are simultaneously measured with error, or when the measurement error takes complicated forms. Current error-correction methods may not be capable of mitigating biases in these challenging cases, which calls for more novel and powerful new methods. Third, researchers can seek to develop novel approaches to combine predictive data mining with econometric analysis that avoid the peril of measurement error. Through this commentary, we hope to raise awareness of these methodological challenges and opportunities and help IS scholars to better sharpen our collective toolkit and harness the power of data mining methods in empirical research.

Endnotes

[1] We use the general term "data mining" throughout the commentary, although the same methodologies are also referred to as "machine learning," "statistical learning," or "predictive analytics" in various contexts.

[2] In this commentary, we do not consider the issue of error in dependent variables, because it is rare for studies to employ predictive models to generate outcome variables for second-stage estimations. Indeed, during our review of the literature for this commentary, we did not come across any study in the IS literature that has taken this approach.

[3] We searched for papers that used predictive data mining methods (e.g., classification) and excluded studies that only employed dictionary-based natural language processing techniques (e.g., Johnson et al. 2015, Tetlock et al. 2008) and studies that used exploratory data mining methods (e.g., Wu 2013, Bao and Datta 2014). In this commentary, we do not discuss exploratory data mining models, such as topic modeling using latent Dirichlet allocation, because they generally do not have prediction-oriented evaluation metrics that can be used to make error corrections.

[4] For a comprehensive introduction to data mining or textual classification, readers may refer to Aggarwal (2015) or Provost and Fawcett (2013). Varian (2014) also provides an overview of data mining techniques for econometricians.

[5] In this commentary, we focus on predictive errors from data mining models. We do not consider intercoder disagreement or error introduced via the human-labeling process. We believe that disagreements among human coders are fundamentally different from predictive errors. Manual labeling is most often employed when there is no ground truth. Disagreements among coders typically reflect inherent ambiguity or subjectivity in the coding process, whereas predictive errors typically reflect the limited learning capacity of data mining models. For subjective or open-ended labeling tasks, the issue of coder-introduced error might be less concerning because the labels reflect researchers' subjective belief about "ground truth" and may not contain definitive error. The application of data-driven procedures to resolve intercoder disagreement falls outside the scope of this work. However, for an example that discusses the issue of intercoder disagreement and the use of SIMEX to mitigate its impact, see Hopkins and King (2010).

[6] This assumption is likely to hold if $\hat{X}$ is generated via a data mining model, because $\hat{X}$ is only determined by its true value, X, and the data mining model, which is usually a separate process from the data-generating process reflected by the second-stage regression equation.

[7] Let $Xb = 1 + 2X_1 + 3X_2 + X_3$. We draw Y_{Logit} from a Bernoulli distribution with $p = 1/(1 + e^{-Xb})$. We draw Y_{Probit} from a Bernoulli distribution with $p = \phi(Xb)$. We draw Y_{poisson} from a Poisson distribution with $\lambda = e^{Xb}$.

[8] The misclassification matrix, while different from a confusion matrix, is readily constructed from the confusion matrix by calculating the recall rates for each class.

[9] Incorporating the labeled (i.e., ground truth) sample that was used to build the classification model may bias the misclassification matrix. However, in most research, because the labeled sample is usually a very small portion of the entire data set, this bias in misclassification matrix generally will not invalidate the error-correction process.

[10] https://www.microsoft.com/cognitive-services/en-us/face-api.

References

Agarwal R, Dhar V (2014) Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research. *Inform. Systems Res.* 25(3):443–448.

Aggarwal CC (2015) *Data Mining: The Textbook* (Springer, Cham, Switzerland).

Aggarwal R, Gopal R, Gupta A, Singh H (2012) Putting money where the mouths are: The relation between venture financing and electronic word-of-mouth. *Inform. Systems Res.* 23(3-part-2): 976–992.

Agrawal A, Catalini C, Goldfarb A (2014) Some simple economics of crowdfunding. Lerner J, Stern S, eds. *Innovation Policy and the Economy*, 1st ed., Vol. 14 (University of Chicago Press, Chicago), 63–97.

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.

Bao Y, Datta A (2014) Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Sci.* 60(6):1371–1391.

Buonaccorsi JP, Laake P, Veierød MB (2005) On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics* 61(3):831–836.

Burtch G, Ghose A, Wattal S (2013) An empirical examination of the antecedents and consequences of contribution patterns in crowd-funded markets. *Inform. Systems Res.* 24(3):499–519.

Burtch G, Ghose A, Wattal S (2015) The hidden cost of accommodating crowdfunder privacy preferences: A randomized field experiment. *Management Sci.* 61(5):949–962.

Carroll RJ, Küchenhoff H, Lombard F, Stefanski LA (1996) Asymptotics for the SIMEX estimator in nonlinear measurement error models. *J. Amer. Statist. Assoc.* 91(433):242–250.

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) *Measurement Error in Nonlinear Models: A Modern Perspective* (CRC Press, Boca Raton, FL).

Chan J, Wang J (2014) Hiring biases in online labor markets: The case of gender stereotyping. *Proc. 35th Internat. Conf. Inform. Systems (ICIS), Auckland, NZ*, 1161–1178.

Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: From big data to big impact. *MIS Quart.* 36(4):1165–1188.

Cook JR, Stefanski LA (1994) Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* 89(428):1314–1328.

Das SR, Chen MY (2007) Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Sci.* 53(9):1375–1388.

Dellarocas C (2003) The digitization of word of mouth: Promise and challenges of online feedback. *Management Sci.* 49(10):1407–1424.

Fisher IE, Garnsey MR, Hughes ME (2016) Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems Accounting, Finance Management* 23(3):157–214.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* 19(3):291–313.

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge Data Engrg., IEEE Trans.* 23(10): 1498–1512.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.

Gleser LJ (1990) Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. *Contemporary Math.* 112:99–114.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4):545–560.

Goh KY, Heng CS, Lin Z (2013) Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content. *Inform. Systems Res.* 24(1):88–107.

Greene WH (2003) *Econometric Analysis* (Pearson Education, Delhi, India).

Gu B, Konana P, Rajagopalan B, Chen HM (2007) Competition among virtual communities and user valuation: The case of investing-related communities. *Inform. Systems Res.* 18(1):68–85.

Gu B, Konana P, Raghunathan R, Chen HM (2014) The allure of homophily in social media: Evidence from investor responses on virtual communities. *Inform. Systems Res.* 25(3):604–617.

Gustafson P (2003) *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments* (CRC Press, Boca Raton, FL).

Hardin JW, Schmiediche H, Carroll RJ (2003) The simulation extrapolation method for fitting generalized linear models with additive measurement error. *Stata J.* 3(4):373–385.

Hopkins DJ, King G (2010) A method of automated nonparametric content analysis for social science. *Amer. J. Political Sci.* 54(1): 229–247.

Huang N, Hong Y, Burtch G (2017) Social network integration and user content generation: Evidence from natural experiments. *MIS Quart.* 41(4):1035–1058.

Huang N, Burtch G, Hong Y, Polman E (2016) Effects of multiple psychological distances on construal level: A field study of online reviews. *J. Consumer Psych.* 26(4):474–482.

Jelveh Z, Kogut B, Naidu S (2014) Political language in economics. Working paper, New York University, New York.

Johnson SL, Safadi H, Faraj S (2015) The emergence of online community leadership. *Inform. Systems Res.* 26(1):165–187.

Jurafsky D, Martin JH (2008) *Speech and Language Processing* (Prentice Hall, Upper Saddle River, NJ).

Küchenhoff H, Lederer W, Lesaffre E (2007) Asymptotic variance estimation for the misclassification SIMEX. *Comput. Statist. Data Anal.* 51(12):6197–6211.

Küchenhoff H, Mwalili SM, Lesaffre E (2006) A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics* 62(1):85–96.

Lin M, Lucas HC Jr, Shmueli G (2013) Research commentary—Too big to fail: Large samples and the *p*-value problem. *Inform. Systems Res.* 24(4):906–917.

Liu Y, Chen R, Chen Y, Mei Q, Salib S (2012) I loan because...: Understanding motivations for pro-social lending. *Proc. 5th ACM Internat. Conf. Web Search Data Mining* (ACM, New York), 503–512.

Lu Y, Jerath K, Singh PV (2013) The emergence of opinion leaders in a networked online community: A dyadic model with time dynamics and a heuristic for fast estimation. *Management Sci.* 59(8):1783–1799.

Mayzlin D, Dover Y, Chevalier J (2014) Promotional reviews: An empirical investigation of online review manipulation. *Amer. Econom. Rev.* 104(8):2421–2455.

Moreno A, Terwiesch C (2014) Doing business with strangers: Reputation in online service marketplaces. *Inform. Systems Res.* 25(4):865–886.

Mudambi SM, Schuff D (2010) What makes a helpful review? A study of customer reviews on Amazon.com. *MIS Quart.* 34(1): 185–200.

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. *Proc. ACL-02 Conf. Empirical Methods Natural Language Processing*, Vol. 10 (Association for Computational Linguistics, Strousburg, PA), 79–86.

Provost F, Fawcett T (2013) *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking* (O'Reilly Media, Sebastopol, CA).

Rhue L (2015) Who gets started on Kickstarter? Demographic variations in fundraising success. *Proc. 36th Internat. Conf. Inform. Systems (ICIS), Fort Worth, TX*, 1303–1314.

Singh PV, Sahoo N, Mukhopadhyay T (2014) How to attract and retain readers in enterprise blogging? *Inform. Systems Res.* 25(1): 35–52.

Stefanski LA, Cook JR (1995) Simulation-extrapolation: The measurement error jackknife. *J. Amer. Statist. Assoc.* 90(432):1247–1256.

Tetlock PC, Saar-Tsechansky M, Macskassy S (2008) More than words: Quantifying language to measure firms' fundamentals. *J. Finance* 63(3):1437–1467.

Tirunillai S, Tellis GJ (2012) Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Sci.* 31(2):198–215.

Vapnik V (1995) *The Nature of Statistical Learning Theory* (Springer, New York).

Varian H (2014) Big data: New tricks for econometrics. *J. Econom. Perspect.* 28(2):3–28.

Wang T, Kannan KN, Ulmer JR (2013) The association between the disclosure and the realization of information security risk factors. *Inform. Systems Res.* 24(2):201–218.

Wansbeek T, Meijer E (2000) *Measurement Error and Latent Variables in Econometrics*, Vol. 37 (North-Holland, Amsterdam).

Wu L (2013) Social network effects on productivity and job security: Evidence from the adoption of a social networking tool. *Inform. Systems Res.* 24(1):30–51.

Yin D, Bond S, Zhang H (2014) Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quart.* 38(2):539–560.

Zhang S, Lee D, Singh P, Srinivasan K (2016) How much is an image worth? An empirical analysis of property's image aesthetic quality on demand at AirBNB. *Proc. 37th Internat. Conf. Inform. Systems (ICIS), Dublin, Ireland*, 168–188.

Zhu H, Kraut R, Kittur A (2012) Effectiveness of shared leadership in online communities. *Proc. ACM* 2012 *Conf. Comput. Supported Cooperative Work* (ACM, New York), 407–416.

Zhu H, Kraut RE, Wang YC, Kittur A (2011) Identifying shared leadership in Wikipedia. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (ACM, New York), 3431–3434.