Does Decomposing Losses Improve Our Understanding of the Financial Impact of Data Breaches?

Justin Theriot * and Benjamin Gowan †

Abstract

Existing research modeling the financial impact of data breaches focuses on the total amount lost. Our study asks if decomposing losses can provide accurate overall loss estimates with greater compositional detail of how and if different types of losses will materialize. We analyzed 17,796 unique cyber events from the Advisen loss data set, splitting total financial losses into three FAIR (Factor Analysis of Information Risk) subcategories; primary response costs (PRC), fines & judgments (F&J), and secondary response costs (SRC). The last two forms of loss do not always occur; thus, we independently model the probability of them occurring. Each model uses the following seven variables: record count, firm revenue, region (North America or European Union), threat access (external or internal), threat type (error or malicious), data type (PII, PCI, or PHI), and industry (2-digit NAICS). Our approach uses standard log-log linear regression and explores three complementary penalized models using ridge, lasso, and elastic net. We find that PRC increases by 5% per 10% increase in the number of records breached, while F&J increases by 2%. Firm revenue is passively associated with all forms of loss but does not impact the likelihood of incurring F&J or SRC. Threat access (external vs. internal) increases PRC but reduces SRC and the probability of F&J occurring. In comparison, threat type reduces PRC but increases SRC and F&J with no impact on the likelihood of either secondary form of loss occurring. PCI data raise PRC but substantially increases the probability of SRC while not affecting the loss amount for SRC or F&J. Events involving PHI data are twice as likely to incur F&J. The industry does not impact PRC, with only finance and retail having higher SRC. The victim industry impacts the probability of a secondary loss occurring to varying degrees. Our findings indicate variables affect the financial impact or the likelihood of losses differently, such that decomposing losses provides greater detail and better total loss estimates. Firms can use our results to better model how losses will materialize and efficiently enact cyber security controls to reduce risk exposure.

*RiskLens Data Science Manager - jtheriot@risklens.com
†RiskLens Senior Data Scientist - bgowan@risklens.com

1 Introduction

Companies collecting, storing, and using data still carry unknown financial risks from a potential data breach. Past research has focused on the total cost of a data breach, providing an overall estimate but not enabling users to understand how that loss will materialize. Our paper seeks to understand if decomposing losses into specific subgroups yield insights that allow companies to mitigate their risk exposure better. Our approach uses the FAIR (Factor Analysis of Information Risk) [1] ontology to break total losses into three subcomponents: primary response costs (PRC) and conditional secondary losses of fines & judgments (F&J), and secondary response costs (SRC). The number of records breached, firm revenue, region (North America or European Union), threat access (external or internal), threat type (error or malicious), data type (Personal Identifiable Information (PII), Personal Credit Information (PCI), and Protected Health Information (PHI), and industry (2-digit NAICS) are variables used to analyze their impacts on each of the forms of loss as well as the probability of secondary forms of loss occurring.

We begin our analysis by establishing a log-log linear regression model. We complement that model with three penalized regressions: ridge, lasso, and elastic net. The consequence of imposing this penalty is reducing the coefficient values to zero, emphasizing the most salient variables for more parsimonious models. We use logistic regression for the probability of secondary forms of loss occurring. We conduct a standard cross-validation analysis to evaluate and compare the models by dividing the data into two segments, 70% to train and 30% to validate the models.

We find differing effects on losses as the number of records breached increases; PRC sees a 5% increase for each 10% rise in the number of records breached, while F&J and SRC see a 2% and 2.5% increase. The number of records breached minimally impacts the probability of either secondary form of loss occurring. Firm revenue has a somewhat more even effect, with PRC increasing by 2.5%, F&J 1.5%, and SRC 1.25% for each 10% increase in revenue without impacting the probability of the secondary forms of loss occurring. The region does not significantly impact PRC or SRC. More importantly, F&J are 95% less likely to occur in North America than in the European Union but are 2.75 times more costly when they occur.

An external threat actor could cause PRC to be up to four times more expensive and increase F&J by 52% but is associated with reducing SRC by 42%. Additionally, the probability of F&J occurring is reduced by 65%, and external actors have no impact on SRC. Malicious events reduce PRC by 72% though the lasso and elastic net reduce the coefficient to 0. Furthermore, malicious events reduce the probability of F&J by 58% but are associated with an 18% increase in F&J costs. Malicious threat events do not impact the likelihood of SRC but could increase by 130%.

PCI data could increase PRC by 9% or more depending on the model. However, SRC is 195% more likely to occur with minimal impact on the amount. Whereas breaches involving PHI data are twice as likely to incur F&J, though only see a 2.3% increase in F&J.

The industry does not impact PRC. However, F&J within the finance and healthcare industry are 72%, and 52% more likely to occur. Although, the loss amounts were similar across industries. In contrast, the industry impacts the probability of SRC occurring to a varying degree, with accommodation, finance, information, public, and retail seeing a 60% to 190% increase. Aside from finance and retail, the magnitude coefficients are small, with 23% to 40%.

The approach outlined in this study demonstrates differing effects of loss event variables on decomposed forms of loss. Total loss modeling for data breaches lacks explanatory details. It ignores substantial differences in the magnitude and directional effect of variables across the forms of losses and the probability of secondary losses.

2 Literature Review

Ten years ago, Ponemon developed an initial model that estimated the cost per record of data breach to be at $150 [2]. However, this implies a data breach with 10 million records will have $1.5 billion in losses! Jay Jacobs subsequently used a log-log model on cost and the number of records breached [3], and to better effect, in the Information Risk Insights Report 2020 [4]. In his model, a breach of 10 million records would accrue $448 million in costs. While he improved on Ponemon's method, more factors contribute to loss than just a count of breached records. Romanosky's work [5] then presented a log-log model with the following features: revenue, repeat incident, malicious, probability of lawsuit, firm type, and industry to provide a more robust framework to understand the cost of a data breach—a significant improvement to understanding the factors that impact losses. The outcome of a cyber incident is multidimensional, as stated by Kesan and Zhang in [6]. Thus, nuances are not addressed by modeling only the final aggregated financial losses.

To aid with our loss decomposition, we introduce the FAIR ontology fig. 1. The top node provides the overall Risk based on the Loss Event Frequency and Loss Magnitude. Each of these nodes can be further broken down as seen and estimated. Our focus lies in the Loss Magnitude portion. Six forms of loss can occur in the primary and secondary loss nodes: productivity, response, replacement, competitive advantage, fines and judgments, and reputation. Productivity and replacement costs mainly occur as primary losses, and the latter materialize as secondary losses. Response costs commonly occur as both primary and secondary losses as described by Jones [1]. Our present analysis focuses on primary and secondary response costs, fines & judgments. Productivity and replacement costs are assumed not to occur in confidentiality data breaches (they appear in availability or integrity events). Reputation damage and competitive advantage are additional areas of potential loss, but they are much more complex and potentially impossible to estimate because of the theft of confidential strategic information as noted by Wolff and Lehr in [7]. Thus, these were considered out of scope for this analysis, and we do not attempt to analyze the factors contributing to them.

Figure 1: FAIR Model

FAIR defines *primary response costs* as those associated with gathering the incident and computer security incident response teams along with various departments to manage an incident. For example, Anthem incurred $281 million in PRC [12], and [13] for a 2015 data breach of 82 million PII records. Their costs included implementing security improvement, credit protection, notification to the public, and affected individuals engaging consultants. Equifax accrued $353 million in PRC, including infrastructure improvements to their application, network, and data security related to their 2017 data breach of 145 million PII records [8] and [8]. The Buckle Inc. was the victim of a data security incident in 2017 of two-thousand PCI records, and [10] and [11] resulting in $2.3 million PRC to improve their security.

Secondary response costs are those activities and expenses incurred while dealing with secondary stakeholders. For example, data breaches will include notifications, credit monitoring expenses, and customers choosing to participate in class-action lawsuits that require the firm and attorneys to defend against these litigations. For example, Anthem accrued $115 million SRC [13] involving counsel feed, settlement funds, reimbursement of expenses, and other minor expenses. Target had $172 million in SRC related to their 2013

data breach of 40 million PCI records [14], [15], and [16]. These costs were for card replacement, coming to an average of $10 per card, comprised of [15] reissuing the card, informing consumers of the reissuing, shipping the card, activating the card, and supplemental communication via call centers. In another case, Quest Diagnostics had a data breach in 2016 of 34 thousand PHI records resulting in $195,000 in SRC [17] from a class-action lawsuit. Similarly, Aetna Inc. had a data breach in 2017 of 12,000 PHI records due to a software error resulting in $19 million in SRC [18]. In both cases, costs were settlement funds for those impacted.

Fines & judgments are losses such as a regulatory body fine, a civil case judgment, or a fee based on contractual stipulations. For example, Anthem incurred $16 million in F&J from the US Department of Health and Human Services (HHS) for their breach. Aetna Inc. incurred $1.9 million in F&J from the US District Court for the Eastern District of Pennsylvania [18]. In 2013 the University of Rochester Medical Center had a data breach of 537 PHI records due to physical theft resulting in $3 million F&J from the HHS [19]. These cases align with Romanosky et. al [20] findings that the odds of a firm being sued are 3.5 times greater when individuals suffer financial harm. However, Target nor The Home Depot incurred any F&J from their respective breaches in 2013 and 2016 [14], [15], [16] [21], [22], and [23] but did incur class-action lawsuits as those are SRC.

The FAIR framework and our analysis show that losses do not accrue uniformly. Nearly all events will have PRC, but not all lead to SRC or F&J. Furthermore, if an event does incur F&J that is independent of SRC. Due to this, we model the three forms of losses separately with a distinct two-part model for F&J, and SRC to determine the probability of them occurring and the subsequent losses if they do occur. Our focus is on data breaches pertaining to confidentiality events concerning PII, PCI, and PHI.

We expand Romanosky's variable selection in [5], following Malavasi, Peters, Shevchenko et al. selection in [24]. However, we do not include detailed cyber-threat information such as studies [25] and [26] have included. We have grouped those into internal and external threat access and error and malicious threat types to reduce the variables in our modeling. Specifically, our modeling uses record counts, firm revenue, region, threat access, threat type, data type, and the industry's 2-digit North American Industry Classification System (NAICS) codes. Due to the sparsity of data for specific sectors, we have combined them into one industry resulting in the following variables for our model.

1. Record Count
2. Revenue
3. Region
   (a) European Union
   (b) North America
4. Threat Access
   (a) Internal
   (b) External
5. Threat Type
   (a) Error
   (b) Malicious
6. Data Type
   (a) PII
   (b) PCI
   (c) PHI
7. Industries
   (a) Information - 51
   (b) Finance & Insurance - 52
   (c) Professional - 54
   (d) Healthcare - 62
   (e) Accommodation - 72
   (f) Public - 92
   (g) Retail - 44, 45
   (h) Other Industries
      i. Agriculture - 11
      ii. Mining - 21
      iii. Utilities - 22
      iv. Construction - 23
      v. Manufacturing - 31, 32, 33
      vi. Trade - 42
      vii. Transportation - 48, 49
      viii. Real Estate - 53
      ix. Management - 55
      x. Administrative - 56
      xi. Educational - 61
      xii. Entertainment - 71
      xiii. Other Services - 81

Additional developments have been made in better modeling losses. For example, Martin and Wirfs [27] argue that the cost of cyber risk should not be measured based on the standard actuarial tools but using the

more complex extreme value theory, widely used in operational risk analysis. Wheatley et al. [28] describes a similar method: malicious data breach events are heavy-tailed, characterized by a highly heavy-tailed truncated Pareto or log-normal distribution. Both approaches provide insight but are geared towards only significant breaches or events useful for insurance companies following the Realistic Disaster Scenario. However, both models fail to help us understand the factors that contribute to typical losses and the composition of those losses.

3 Data

We used data from the Advisen Cyber Loss Dataset, which provides a historical view of 17,796 unique events as of 31 December 2021. Shevchenko, Jang, Malavasi et. al., in [26] found no distinct pattern or clear-cut relationship between the frequency of events, the loss severity, and the number of affected records in the Advisen loss dataset. Palsson, Gudmundsson, and Shetty in [29] also stated the Advisen cyber loss data feed is not sufficiently detailed to build a model that can predict the financial cost of a cyber incident with high accuracy. However, both parties utilized the entire dataset without qualification of events. Our analysis focuses strictly on **data breaches**, the unintentional disclosure of personal information stemming from the loss or theft of digital or printed information.

Data processing and filtering for data breaches results in 2,983 cyber events with aggregated case information. We remove security incidents related to disrupting corporate IT systems or losing intellectual property. We also remove privacy violations events that are due to the unauthorized collection, use, or sharing of personal information in violations of information protection statutes such as the Drivers Privacy Protection Act (DPPA), Video Privacy Protection Act (VPPA), Telephone Consumer Protection Act (TCPA), Children's Online Privacy Protection Act (COPPA), Do-Not-Call, Song-Beverly Act, and the Privacy Act. It also includes unsolicited communication from spam emails, other mass marketing communication (robocalling, texts, emails), or debt collection. However, privacy incidents play a more prominent role in cyber-security with regulations like the California Invasion of Privacy Act (CIPA). These are not data protection regulations but data collection and tracking regulations. While these incidents pose exciting research questions, they do not fit our definition of a data breach and, thus, are excluded from our analysis.

We map to the FAIR Forms of Loss as described in § 2 from our data set based on the following definitions from Advisen:

1. Primary Response Costs
   (a) FAIR
      i. These costs are the dollar amount paid by the company due to the event.
   (b) Advisen
      i. Other with Response Cost Present
         A. Response costs are the initial costs to the firm to remedy the cause of the incident.
2. Secondary Response Costs
   (a) FAIR
      i. These costs are presented as the amount equal to the financial loss suffered by the plaintiff or the company due to the event. These amounts can include money spent by the plaintiff in prosecuting the case for lawyers, law firms, legal representation, and other related expenses, along with any amounts outside of legal representation spent by the company in defense.
   (b) Advisen Data
      i. Financial Damages Amount
         A. The financial loss suffered by the plaintiff, or by the company, as a result of the incident.
      ii. Plaintiff Legal Fees

A. The amount spent by the plaintiff in prosecuting the case for lawyers, law firms, legal representation, and other related expenses when there is an indication that the plaintiff's legal fees are included in the settlement amount.

iii. Any other amount paid by the defendant due to the case or incident.

3. Fines and Judgments
   (a) Fair
      i. These costs are amounts charged to the company as court-ordered penalties or settlements.
   (b) Advisen
      i. Punitive Exemplary Damages
         A. Charges to the company as punishment, in addition to the actual value of the case.
      ii. Pain and Suffering
         A. Amount charged to the company to offset the pain and suffering of the plaintiff, which resulted from the incident.
      iii. Other Fines and Penalties
         A. Fines or penalties paid by the company due to the case.

Based on our review of SEC filings of post-breach incidents, as seen in § 2, the most common miss-mapping occurs in credit monitoring. Companies have provided that service upfront to impacted individuals, thus being a primary response cost, whereas other times, they have only offered it after litigation, thus being a secondary response cost. Even with those minor discrepancies, Wolff and Lehr in [7] note the Advisen dataset helps estimate incident-level costs as they aggregate information across the range of different data sources available, ideal for our analysis.

We use events between 2005 and 2021 as losses take time to accrue, reach public reports, and then make their way into the data set. Anything before 2005 is less relevant to our current understanding of data breaches. FAIR defines a loss event as a threat event where loss materializes, non-events are addressed by the primary loss event frequency portion of the FAIR model, not explored in this paper. Thus, for each independent loss model, we keep only those events with losses greater than \$0 for each dependent variable: PRC, F&J, and SRC.

Lastly, we adjust losses for inflation from when they accrued using the latest monthly CPI data from the Federal Reserve Economic Data (FRED) at the Federal Reserve Bank of St. Louis. The CPI is one of the most frequently used statistics for identifying periods of inflation. Our formula to convert losses from nominal to real is,

$$\ell_{i,j,t} = \frac{CPI_t}{CPI_{t-n}} \, \ell_{i,j,t-n} \tag{1}$$

Where ℓ_i is an individual company's unique event with a specific *real* loss j at time t. CPI is the current consumer price index at time t, and time $t-n$ is the CPI at the time of the loss. $\ell_{i,j,t-n}$ is the left-hand loss but in *nominal* terms at the time of the loss, $t-n$. Since each event is comprised of multiple losses at various times, we aggregate those losses.

$$\mathcal{L}_i = \sum_{j=1}^{n} \ell_j \tag{2}$$

Where $\mathcal{L}_i$ is the aggregated loss value for all losses over the event's time frame.

4 Methodology

Our analysis seeks to understand how the independent variables impact our decomposed forms of loss and the probability of secondary forms of loss occurring. We begin with a discussion of the loss magnitude models.

4.1 Loss Magnitude

The cost of the breach, against the number of records breached, and firm revenue have a non-linear form amongst themselves. Using a log/log-linear regression, the model becomes linear while allowing the variable coefficients to be elastic such that the percentage change in y is associated with a 1 percent change in x_k. Our initial model is:

$$\begin{aligned} ln(cost_{i,t}) = \beta_0 + \beta_1 * ln(records_{i,t}) + \beta_2 * ln(revenue_{i,t}) + \beta_3 * region_{i,t} + \beta_4 * access_{i,t} + \\ \beta_5 * type_{i,t} + \alpha * data_{i,t} + \lambda * industry_{i,t} + \epsilon_i \end{aligned} \tag{3}$$

where *cost* is the cost of the incident by firm i for loss type t. *records* is the log number of compromised records from the incident. *revenue* is the log of the firm's revenue. *access* and *type* are binary variables coded as 1 if the firm suffered an external or malicious event and 0 otherwise. *data* is a vector describing whether the impacted data type was PII, PCI, or PHI; by convention, we omit a reference category, PII. *industry* is a vector describing the impacted firm's industry as either information, finance & insurance, professional, healthcare, accommodation, public, retail, or other industries; again, by convention, we omit a reference category; other industries.

We introduce three complementary penalized regressions to guide our analysis. The consequence of imposing this penalty is reducing the coefficient values towards zero but not always zero. We test the first penalized regression technique, the **ridge regression**. Ridge regression shrinks the regression coefficients so that variables with minor contributions to the outcome have their coefficients close to zero. Compared to OLS, an advantage of ridge regression is that it avoids over-fitting and is more robust to multicollinearity. However, the regression will include all predictors in the final model. The ridge model does not differ structurally from a linear regression; however, the penalty term changes. As proposed by Hoerl and Kennard [30], the potential instability in the LS estimator could be improved by adding a small constant value λ to the diagonal entries of the matrix $X'X$ before taking its inverse.

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{4}$$

The result is the ridge regression estimator.

$$\hat{\beta}_{ridge} = (X'X + \lambda I_p)^{-1}X'Y \tag{5}$$

Ridge regression shrinks the coefficients towards zero, but it will not set any of them exactly to zero. Thus, we introduce our third model, **lasso regression**, as an alternative that overcomes this drawback. It shrinks the regression coefficients by penalizing the regression model with a penalty term called L1-norm, the sum of the absolute coefficients. It allows individual coefficients to go to zero and be removed from the model. The penalty reduces the coefficient estimates, with a minor contribution to the model, some to zero. An advantage of the lasso regression is that it produces a simpler and more interpretable models that incorporate only a reduced set of predictors. Thus, the Lasso estimates of the coefficients are

$$\hat{\beta}_{lasso} = min_{\beta}(Y - X\beta)'(Y - X\beta) + \lambda \sum_{j=1}^{p} \beta_j \tag{6}$$

so that the L2 penalty of the ridge regression $\sum_{j=1}^{p} \beta_j^2$ is replaced by an L1 penalty, $\sum_{j=1}^{p} \|\beta_j\|$ as described by Tibshirani [31]. Generally, lasso might perform better when some of the predictors have large coefficients and the remaining ones have tiny ones.

Our last loss magnitude model is the **elastic net** that produces a regression model penalized with both the Ridge L1-norm and Lasso L2-norm penalties. The consequence is to shrink coefficients effectively and set some coefficients to zero. Zou and Hastie [32] describe the estimates from the elastic net method are defined by

$$\hat{\beta}_{elastic-net} = min_{\beta}\|(Y - X\beta\|^2)'(Y - X\beta) + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1 \tag{7}$$

4.2 Secondary Loss Event Frequency

We tested a **logistic regression** model as a standard econometric model to understand how variables impact the probability of secondary losses. The model uses a binary classification variable for the secondary form of loss occurring or not and includes all independent variables previously described. The number of records breached and firm revenue variables are log-normal transformed with the independent variables as binary operators. Thus, a transformation to the response variable is applied to yield a continuous probability distribution over the output classes bounded between 0 and 1. The parameter estimates inform whether there is an increase or decrease in the predicted log odds of the response variable that would be predicted by one unit increase or decrease in one of the explanatory variables while holding all other explanatory variables constant.

$$logit(Y_i) = ln\frac{\pi}{1-\pi} = \beta_0 + \beta_1 ln(x_{it}) + ... + \beta_n * x_{it} \tag{8}$$

Therefore,

$$\pi = \frac{e^{\beta_0+\beta_1 ln(x_{i,1})+\beta_2 x_{i,2}+...+\beta_4 x_{i,n}}}{e^{\beta_0+\beta_1 ln(x_{i,n})+\beta_2 x_{i,n}+...+\beta_n x_{i,n}}} \tag{9}$$

4.3 Cross-Validation

To ensure the models analyzed here perform well on new out-of-sample data, we conduct simple cross-validation to evaluate and compare the models by dividing the data into two segments, 70% to train and 30% to validate the model. We reviewed the mean absolute error, median absolute error, R^2, and explained variance for the loss magnitude models on the holdout data. We review the mean accuracy score and the area under the curve for the secondary loss event frequency models.

5 Summary Statistics

We examine the decomposition of losses, compromised records, and firm revenues. Specifically, table 1 shows that the number of observations with losses differs significantly, with PRC having the least and SRC having the most at 698. Losses between PRC and SRC are distributed evenly, with F&J having an outlier in the Facebook Cambridge Analytica case. Records comprised are evenly distributed again between PRC and SRC, the maximum of 3 billion in the August 2013 Yahoo! data breach. Finally, revenue differs between the forms of losses.

In table 2, we see that PRC and SRC are concentrated in North America. We removed the region from our PRC analysis since only five of the 120 events occurred in the European Union. At the same time, F&J are evenly split between the two regions. Threat access shows that PRC has a more significant portion of external events, F&J internal events, and SRC, an even split between the two. Threat types between the three forms of loss have more malicious than error events. Firms might view errors as non-cyber events, thus failing to report them accurately, if they report them at all. Data type indicates a more significant portion of events involving PCI records. The finance and healthcare industry are heavily represented in the data, but we expect these firms to report cyber events at a higher rate since they are heavily regulated.

In table 3, the records compromised, and revenues differ significantly from those with reported secondary

Table 1: Loss Magnitude Data Summary Statistics

Log-Normal Variables					
Variable (millions)	N	Mean	Median	Min	Max
Primary Response Costs	120				
Loss		0.084	1.00	0.003	1,000
Records Compromised		0.094	0.064	<0.001	3,000
Revenues		304	332	0.314	132,000
Fines & Judgments	371				
Loss		0.224	0.200	<0.001	5,600
Records Compromised		0.003	0.003	<0.001	382
Revenues		95	110	0.002	182,000
Secondary Response Costs	698				
Loss		0.380	0.460	<0.001	1,170
Records Compromised		0.002	0.001	<0.001	3,000
Revenues		149	170	0.040	5,600

Table 2: Loss Magnitude Data Summary Statistics

Binary Variables						
Variable	**PRC**		**F&J**		**SRC**	
Region	*No.*	*Percent*	*No.*	*Percent*	*No.*	*Percent*
North America	115	95.8	203	57.7	656	93.9
European Union	5	4.2	168	42.3	42	6.1
Threat Access						
External	99	82.5	115	31.0	386	55.3
Internal	21	17.5	256	69.0	312	44.7
Threat Type						
Error	13	10.3	85	22.9	50	7.2
Malicious	107	89.7	286	77.1	648	92.8
Data Type						
PCI	79	65.8	93	25.1	386	55.3
PHI	22	18.3	109	29.2	106	15.2
PII	18	15.9	169	45.7	206	29.5
Industry						
Accommodation	7	5.8	6	1.6	46	6.6
Finance	31	25.8	65	17.5	149	21.3
Healthcare	8	6.7	80	21.6	88	12.6
Information	6	5.0	43	11.6	64	9.2
Professional	4	3.3	27	7.3	50	7.2
Public	15	12.5	38	10.2	59	8.5
Retail	11	9.2	25	6.7	74	10.6
Other	38	31.7	87	23.5	168	24.0

losses, though the data has increased four-fold. More details are seen in table 4 that closely match our loss data, though they have a better distribution within the industry.

Overall, the strength in the data lies in F&J, which is Advisen's focus. We have 371 events with losses, a relatively even split in the region, threat access, threat type, data type, and industry. SRC is another vital area from a data perspective, but given how these losses accrue, these are also the most complex in how they occur and are reported. Finally, the PRC is the weakest, with a small number of observations concentrated

Table 3: Secondary Loss Event Frequency Data Summary Statistics

Log-Normal Variables					
Variable (millions)	N	Mean	Median	Min	Max
SLEF	2,983				
Records Compromised		<0.001	0.001	<0.001	3,000
Revenues		80	68.5	0.002	559,000

in North America. The small number of events with PRC is not an indicator that they do not occur for every event but that firms are not required to report these numbers. SEC filings provide information related to the initial costs accrued after a breach, though often becoming muddled as more time passes after the incident and additional information is added. However, non-public companies are not required to report those costs. Whereas F&J and SRC stem from either government fines, litigation, or class-action lawsuits, which are public records, thus we are more assured of these being a complete picture. If companies were willing or able to reveal their information, everybody could be better off, as stated by Rothschild and Stiglitz [33] in the 1970s regarding insurance data.

Table 4: SLEF Data Summary Statistics

Binary Variables		
Variable, N=2,983	**No.**	Percent
F&J	371	12.4
SRC	698	23.4
Region		
North America	2677	89.7
European Union	306	10.3
Threat Access		
External	1,286	43.1
Internal	1,697	56.9
Threat Type		
Error	229	7.7
Malicious	2,754	92.3
Data Type		
PCI	992	33.3
PHI	611	20.5
PII	1,450	46.2
Industry		
Accommodation	90	3.0
Finance	465	15.6
Healthcare	508	17.0
Information	301	10.0
Professional	318	10.6
Public	172	5.8
Retail	236	7.9
Other	855	30.1

6 Analysis

Before we begin our analysis, let us review the model performance. All model results and coefficients can be found in table 5, table 6, table 7, and table 8. First, our model exploration suggests different losses should not be modeled uniformly. The elastic net performed best on PRC with an explained variance of 0.560 compared to OLS with a score of 0.467 even though the median absolute error increased from 1.46 to 1.66. However, OLS and ridge regression performed best for F&J with an explained variance of 0.391 compared to 0.366 for the elastic net with a median absolute error of 1.23 compared to 1.29. SRC showed no transparent model as explained, variances were between 0.274 and 0.280 for all four models. The SLEF model indicates the logistic regression is suitable for both forms of loss as accuracy scores were above 0.8 and 0.7, respectively.

Table 5: Primary Response Costs Coefficients

Primary Response Costs Model Results				
Variable, N=115	**OLS**	**Ridge**	**Lasso**	**Elastic Net**
Coefficients				
const	1.557	1.687	3.506	3.191
Record Count ***	0.518	0.516	0.500	0.497
Revenue ***	0.244	0.244	0.225	0.232
External ***	1.614	1.568	0.000	0.188
Malicious *	-1.272	-1.182	0.000	0.000
Payment Card Information ***	1.330	1.284	0.000	0.086
Protected Health Information	0.951	0.902	0.000	0.000
Accommodation	0.348	0.323	0.000	0.000
Finance	-0.325	-0.341	0.000	0.000
Healthcare	0.086	0.073	0.000	0.000
Information	0.929	0.862	0.000	0.000
Professional ***	3.939	3.634	0.000	0.000
Public	-0.596	-0.600	0.000	0.000
Retail	0.072	0.048	0.000	0.000
Cross-Validation				
Mean Absolute Error	1.91	1.86	1.88	1.86
Median Absolute Error	1.46	1.45	1.67	1.66
R^2	0.445	0.469	0.533	0.536
Explained Variance	0.495	0.495	0.560	0.560

Second, the number of records breached impacts each form of loss differently, not uniformly as implied by previous total loss models from Ponemon, Jacobs, and Romanosky. For example, PRC increases 5% for each 10% increase in the number of records for all models. Whereas F&J and SRC only increase by 2% and 2.5% for all models [1]. This implies that a larger number of records leads to a more complex breach, requiring a more significant effort to remedy and that as losses accrue, later they are dampened from the effect. Furthermore, a 10% increase in the number of records breached minimally impacts the probability of F&J or SRC occurring with a 1% or minor increase. [2]

[1] For the tables in § 6 *** indicates significance at the 0.01, ** at the 0.05, and * at the 0.1 level respectively.

[2] The mean and median absolute error results from our base e log-log model. For example, $100,000 has a natural log value of 11.51; adding an error of 2 (our results range from 1.59 to 2.03) for 13.51 equates to $736,747. A much smaller difference compared to a log base 10 model which would indicate two orders of magnitude.

Table 6: SLEF Coefficients & Cross-Validation

Secondary Loss Event Frequency Model Results				
Variable, N=2,983	**F&J**		**SRC**	
Logistic Regression Coefficients				
const	0.378		3.281	
Record Count	0.103 ***		0.028 **	
Revenue	-0.008		0.032 ***	
North America	-2.884 ***		0.684	
External	-1.042 ***		0.070	
Malicious	-0.867 ***		0.030	
Payment Card Information	0.221		1.083 ***	
Protected Health Information	0.983 ***		0.092	
Accommodation	0.284		1.050 ***	
Finance	0.545 ***		0.550 ***	
Healthcare	0.422 *		0.116	
Information	0.043		0.320 *	
Professional	0.005		-0.037	
Public	0.182		1.104 ***	
Retail	-0.073		0.481 ***	
Cross-Validation	μ Accuracy	AUC	μ Accuracy	AUC
Logistic Regression	0.89	0.80	0.77	0.70

Third, firm revenue appears to have a more uniform impact on PRC, F&J, and SRC with 2.5%, 1.5%, and 1.25% increases for each 10% increase in revenue across models. Large revenue firms have more complex organizational structures, leading to increased costs, and likely make them a more prominent target for reprisal litigation. Firm revenue increases F&J by 1.5% could be attributed to GDPR in the European Union and California Consumer Privacy Act, but neither impact the probability of SRC.

Fourth, the firm's region in North America or the European Union has been removed from PRC, given that only five of the 120 events were in the EU. The F&J model diverges from the other loss models here with a persistent, significant effect for the region. North American firms are 95% less likely to experience F&J than their EU counterparts, but they will be 275% more costly when they incur. The region was deemed insignificant for SRC. However, the OLS and ridge model indicates a 67% increase for North America. Given the nature of SRC, we suspect labor costs play a factor, though we cannot provide a theory without a breakdown of costs to see how they are accruing.

Fifth, external threat actors cause PRC to increase by up to four times more in the OLS and ridge regression models through the lasso and elastic net models, reducing the coefficient to 0. The increased cost can be attributed to external events requiring more effort to find and patch the vulnerability while also requiring third-party forensics all driving up costs. In comparison, external events reduce the probability of F&J occurring by 65%. This is coherent with firms being held more accountable for internal errors than external attacks. However, when F&J costs occur, external events are associated with a 52% increase in the OLS and ridge methods, with the lasso and elastic net reducing the value to 0. In comparison, external threat actor decreases SRC by 42%.

Sixth, the OLS and ridge models show that the malicious threat type reduces PRC by 72% though the lasso and elastic net reduce the coefficient to 0. In comparison, malicious events reduce the probability by 58% of incurring F&J and could attribute an 18% increase in the loss. This is consistent with Kesan and Zhang in [6] where they find that if the actor is the company itself, it is likely to be fined for violating regulations. The malicious threat type does not impact the probability of SRC. However, a malicious attack increases SRC by 130%. In practical terms, a malicious attack scenario would not presume higher than

Table 7: Fines and Judgments Coefficients

Fines and Judgments Model Results				
Variable, N=371	**OLS**	**Ridge**	**Lasso**	**Elastic Net**
Coefficients				
const	7.183	7.183	8.024	7.730
Record Count ***	0.199	0.199	0.216	0.214
Revenue	0.150	0.149	0.136	0.145
North America ***	1.325	1.317	0.131	0.424
External ***	0.418	0.417	0.000	0.000
Malicious	0.163	0.161	0.000	0.000
Payment Card Information	-0.160	-0.159	0.000	0.000
Protected Health Information	0.022	0.024	0.000	0.000
Accommodation	-0.750	-0.709	0.000	0.000
Finance	0.246	0.252	0.000	0.000
Healthcare	-0.456	-0.447	0.000	0.000
Information	-0.120	-0.113	0.000	0.000
Professional	-0.661	-0.646	0.000	0.000
Public	0.538	0.537	0.000	0.000
Retail **	-1.022	-1.006	0.000	0.000
Cross-Validation				
Mean Absolute Error	1.59	1.59	1.67	1.64
Median Absolute Error	1.27	1.23	1.45	1.29
R^2	0.383	0.383	0.334	0.358
Explained Variance	0.391	0.391	0.343	0.366

typical PRC nor assume an increased likelihood of SRC but should consider the potential for them to be higher than typical when they do occur.

Seventh, OLS, ridge, and elastic net weight PCI data as more costly for PRC, but to what extent? Is it nearly three times more expensive, as deemed by the OLS and ridge model, or closer to the 9% as deemed by the elastic net? Given additional upfront costs such as notification, replacement, and shipment when dealing with PCI data, we must assume an increase in PRC. PCI was statistically significant in impacting the probability of F&J, attributing a 25% rise in the likelihood. In contrast, breaches involving PHI are nearly twice as likely to experience F&J, though they only see a 2.3% increase in the loss. In comparison, PCI data is a strong indicator in incurring SRC with a 195% increase in probability, and PHI data has a more negligible increased impact of 10%. However, data type plays a minor role in the magnitude of SRC, with the OLS and ridge indicating a 2% decrease and PHI a 6% increase though both reduced to 0 within the lasso and elastic net.

Industry coefficients were deemed insignificant to PRC, aside from professional. This is reasonable considering that patching the vulnerability, hiring forensics, and notification costs associated with PRC are unlikely to differ by industry. However, a wage by occupation by industry study would be needed to confirm this hypothesis. Furthermore, the industry does not impact the actual F&J as no industries were statistically significant, except retail, with a 65% reduction. The coefficients indicate potentially different amounts, with finance and the public seeing higher F&J than others. The second most significant factor in the probability of accruing SRC is industry. Accommodation, finance, information, public, and retail see a 60% to 190% increase in SRC probability. However, the industry does not factor into the magnitude of SRC. The coefficients are small aside from finance and retail, with 23% and 40% increases in the OLS and ridge model. The lasso and elastic net model push them all to 0. Recouping credit card replacements is associated with costs considering both industries utilize PCI data. However, this does not explain the 52% reduction in SRC for

Table 8: Secondary Response Costs Coefficients

Secondary Response Costs Model Results				
Variable, N=698	**OLS**	**Ridge**	**Lasso**	**Elastic Net**
Coefficients				
const	7.291	7.297	8.453	8.317
Record Count ***	0.272	0.272	0.252	0.255
Revenue ***.	0.130	0.130	0.129	0.136
North America	0.515	0.512	0.000	0.000
External *	-0.542	-0.540	0.000	0.000
Malicious **	0.833	0.877	0.000	0.000
Payment Card Information	-0.024	-0.025	0.000	0.000
Protected Health Information	0.060	0.059	0.000	0.000
Accommodation *.	-0.736	-0.731	0.000	0.000
Finance	0.208	0.209	0.000	0.000
Healthcare	-0.049	-0.0481	0.000	0.000
Information	-0.087	-0.086	0.000	0.000
Professional	-0.114	-0.112	0.000	0.000
Public	-0.120	-0.118	0.000	0.000
Retail	0.332	0.322	0.000	0.000
Cross-Validation				
Mean Absolute Error	2.03	2.03	2.05	2.05
Median Absolute Error	1.79	1.79	1.77	1.76
R^2	0.278	0.278	0.273	0.274
Explained Variance	0.280	0.280	0.274	0.276

accommodation, another sector where PCI data is stored.

7 Discussion

While we are aware price-per-record estimates are faulty, the number of records does remain the strongest predictor of the cost of a breach, positively associated with primary costs, the probabilities of secondary losses, and the amount of those costs as summarized in table 9. Revenue is another strong predictor of the cost of a breach, though it does not impact the probability of incurring F&J. However, we have to ask if revenue should be an interactive term with record count, as a larger company most likely has a more significant number of records; thus, the probability of a substantial breach goes up. Through our decomposition, we have demonstrated key variables do not uniformly impact the forms of loss. The region is essential for the probability and amount of F&J but does not meaningfully influence the other forms of loss. External events only impact PRC, and the likelihood of incurring F&J. Malicious events affect the probability of F&J. PCI data impacts PRC and the probability of incurring SRC. Whereas PHI data positively impacts the likelihood of incurring F&J.

Table 9: Elastic Net & Logistic Regression Coefficients

Complete Scenario Table - Elastic Net					
Variable	**PRC**	**SLEF F&J**	**F&J**	**SLEF SRC**	**SRC**
const	3.191	0.378	7.730	3.281	8.317
Record Count	0.497	0.103	0.214	0.028	0.255
Revenue	0.232	—	0.145	0.032	0.136
North America	—	-2.884	0.424	—	—
External	0.188	-1.042	—	—	—
Malicious	—	-0.867	—	—	—
Payment Card Information	0.086	—	—	1.083	—
Protected Health Information	—	0.983	—	—	—
Accommodation	—	—	—	1.050	—
Finance	—	0.545	—	0.550	—
Healthcare	—	0.424	—	—	—
Information	—	—	—	0.320	—
Professional	—	—	—	—	—
Public	—	—	—	1.104	—
Retail	—	—	—	0.481	—

While this modeling approach of decomposing costs is informative and robust, there are limitations and areas for improvement. For limitations, firstly, as researched by Sangari, Dallal, and Whiteman [34], cyber research has struggled with the growing problem of under-reporting cyber incidents. Advisen data neither includes undetected incidents nor incidents detected but not disclosed. A secondary effect of under-reporting is the absence of a further breakdown of PRC and SRC; information that would be valuable. The information gathered through SEC filings and court cases lets us understand that these costs differ, but firms are reluctant to provide a clear breakdown. Determining where costs will occur would be beneficial to predicting how losses accrue. Thus, a tertiary effect forced us not to consider certain forms of loss, such as reputation damage, competitive advantage, or more complex extortion payment cases.

Second, we need a better understanding of how and when legal cases will impact a company beyond our initial estimates of F&J SLEF. Future research could distinguish both claims brought forth by government agencies, class-action lawsuits, and other legal filings. As Kesan and Zhang noted in [6], the scenario or intent matters in understanding their outcomes with losses having different dependence structures. Events can fall under various regulations, such as the General Data Protection Regulation or the California Invasion of Privacy Act which provides data rights to individuals. Other events fall under Electronic Fund Transfer Act, requiring financial institutions to adopt certain practices to transfer funds; the Electronic Communications Privacy Act, which prohibits the interception of electronic communications; or the Stored Communications

Act, which prohibits electronic communication service providers and remote computing service providers from knowingly disclosing the contents of customers' electronic communications or subscriber records.

Furthermore, as technology improves, data types are being expanded to include facial recognition, fingerprint, and voiceprint, leading to the Biometric Information Privacy Act, which regulates the collection, use, and handling of biometric identifiers and information by private entities. Biometric data is categorized as PII. The regulation does not treat biometric information data like email and home addresses, yet the data is classified as such by the courts; we need to account for this more sensitive information in our modeling.

Lastly, we need to understand the relationship between SLEF and associated costs better. While we can capture a portion of the implications of SLEF, there are still many unknowns of what event characteristics will trigger them and how those will influence losses when they occur. Several variables were associated with a lower probability of secondary losses but higher costs when those losses did occur, suggesting a nuanced relationship that bears further investigation. In [7], Wolff and Lehr stated that we need to understand how the costs are incurred and by whom to empower firms and policymakers to make good decisions around cyber investment. However, firms' lawyers frequently refuse to share written documentation regarding a breach with third parties like insurers, regulators, and law enforcement. They expand, stating that law firms overseeing breach investigations increasingly instruct forensic firms not to craft any final report regarding a breach, as Schwarcz, Wolff, and Woods noted in [35]. Bringing us full circle to our first limitation in data limitations due to under-reporting.

Understanding the financial impact after an event is the first step to understanding how a firm can adequately invest in reducing its risk exposure. Decomposed losses allow us to measure the effects of controls or anything that can be used to reduce the frequency or magnitude of loss [36]. Firms can then understand when investments in their cyber security program stop reducing risk exposure.

Our research provides avenues for practical applications in cyber risk quantification today. First, we base our estimates on firm revenue, region, data type, and industry. For example, we can develop a loss estimate specific to a company based in North America, in the Healthcare industry, with $4 billion in revenue, with a particular number of PHI records. Second, scope variables can improve scenario comparisons to help drive prioritization exercises. For example, should budget increases be allocated to improving internal or external access controls across assets? Third, since this research separately modeled various costs, this should improve final estimates and facilitate communication with stakeholders, enabling us to express the real risk and proper decompositions of operational costs vs. legal fees to enterprise risk management. Data limitations are unbound and a known problem in cyber security research, but as the community continues to push for transparency and SEC regulations begin to require reporting about material cybersecurity incidents, we can enhance our research to equip firms and analysts to make data-driven estimates of their cyber data breach exposure using readily available industry firmographics.

References

[1] Jones, Jack and Freund, Jack (2015) *Measuring and Managing Information Risk: A FAIR Approach.* Oxford: Elsevier Inc.

[2] Ponemon Institute (2020), Cost of a Data Breach Report 2020 https://www.ibm.com/blogs/ibm-anz/the-rising-cost-of-a-data-breach-in-2020/, accessed 20 January 2020.

[3] Jacobs, Jay. Analyzing ponemon cost of data breach. http://datadrivensecurityinfo/blog/posts/2014/Dec/ponemon/ (January 2020, date last accessed).

[4] Cyentia Institute (2020), Information Risk Insights Study 2020, https://www.cyentia.com/wp-content/uploads/IRIS2020_cyentia.pdf.

[5] Romanosky, Sasha, Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, Vol. 2 No. 2 : 121 - 135, 2016.

[6] Kesan, J., and Zhang, L. Analysis of Cyber Incident Categories Based on Losses. *University of Illinois College of Law Legal Studies Research Paper* Paper No. 20-08. 2020. (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489436).

[7] Wolff J., and Lehr, W. Degrees of Ignorance about the Costs of Data Breaches: What Policymakers Can and Can't Do about the Lack of Good Empircal Data. *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2943867* 18 Aug 2017.

[8] Equifax (2018). Form 10-K 2018. Retrieved from `https://otp.tools.investis.com/clients/us/equifax/SEC/sec-show.aspx?FilingId=12595195&Cik=0000033185&Type=PDF&hasPdf=1`

[9] Equifax (2019). Form 10-K 2019. Retrieved from `https://otp.tools.investis.com/clients/us/equifax/SEC/sec-show.aspx?FilingId=13243973&Cik=0000033185&Type=PDF&hasPdf=1`

[10] The Buckle Inc (2017). Form 10-Q 2017. Retrieved from `https://otp.investis.com/clients/us/buckle/SEC/sec-show.aspx?FilingId=12271176&Cik=0000885245&Type=PDF&hasPdf=1`

[11] The Buckle Inc (2018). Form 10-K 2018. Retrieved from `https://otp.investis.com/clients/us/buckle/SEC/sec-show.aspx?FilingId=12670024&Cik=0000885245&Type=PDF&hasPdf=1`

[12] Anthem (2015). Form 10-K 2015. Retrieved from `https://ir.antheminc.com/static-files/fc6f1db3-3faf-41f7-af08-af6fdbe7dc82`

[13] Anthem (2018). Form 10-K 2018. Retrieved from `https://ir.antheminc.com/static-files/d5cb23b3-d741-499e-9efa-f623a14c525a`

[14] Target (2014). Form 8-K 2014. Retrieved from `https://investors.target.com/static-files/309321b8-a04b-4455-a882-6720091add8e`

[15] Target (2014). Form 10-K 2014. Retrieved from `https://investors.target.com/static-files/63996355-8bc5-443c-927c-61617d9ded77`

[16] Target (2016). Form 10-K 2016. Retrieved from `https://investors.target.com/static-files/5d8f1a56-b6ff-420b-a111-6dc87ad08942`

[17] HIPPA CraticRx (05 November 2019). Retrieved from `https://www.hipaacompliance-ny.com/single-post/2019/11/05/Judge-Approves-Quest-Diagnostics-195K-Settlement-for-2016-Breach`

[18] Andrew Beckett, Arizona Doe, California Doe, S.A., Colorado Doe, Connecticut Doe, DC Doe et al vs Aetna, Inc., Case No. 2:17-CV-3864-JS (United States District Court for the Eastern District of Pennsylvania, 2018)

[19] The United States Department of Health and Human Services v. University of Rochester Medical Center, Resolution Agreement (2019)

[20] Sasha Romanosky Journal of Cybersecurity, Volume 2, Issue 2, December 2016, Pages 121–135, https://doi.org/10.1093/cybsec/tyw001

[21] The Home Depot (2014). Form 8-K 2014. Retrieved from `https://otp.tools.investis.com/clients/us/home\_depot/SEC/sec-show.aspx?FilingId=1021059&Cik=0000354950&Type=PDF&hasPdf=1`

[22] The Home Depot (2016). Form 10-Q 2016. Retrieved from `\https://otp.tools.investis.com/clients/us/home\_depot/SEC/sec-show.aspx?FilingId=11701717&Cik=0000354950&Type=&hasPdf=1`

[23] The Home Depot (2017). Form 10-Q 2017. Retrieved from `https://otp.tools.investis.com/clients/us/home\_depot/SEC/sec-show.aspx?FilingId=12085326&Cik=0000354950&Type=PDF&hasPdf=1`

[24] Malavasi, M., Peters, G., Shevchenko, P., Truck, S., Jang, J., Sofronov, G. Cyber Risk Frequency, Severity, and Insurance Viability. *Insurance: Mathematics and Economics* Vol. 106: 90-114. September 2022.

[25] Paters, G., Shevchenko, P., Truck, S., Malavasi, M., Sofronov, G., and Jang, J. Cyber Loss Model Risk Translates to Premium Mispricing and Risk Sensitivity. January 16, 2022. Available at SSRN: https://ssrn.com/abstract=4009941 or http://dx.doi.org/10.2139/ssrn.4009941

[26] Shevchenko, P., Jang, J., Malavasi, M., Peters, G., Georgy, S., Truck, S. The Nature of Losses from Cyber-Related Events: Risk Categories and Business Sector. *Journal of Cybersecurity* Vol. 9, Issue 1. 2023. (https://doi.org/10.1093/cybsec/tyac016)

[27] Eling, Martin and Wirfs, Jan H. What Are the Actual Costs of Cyber Risk Events? *European Journal of Operational Research*, Vol. 272 No. 3 : 1109 - 1119, February 2019.

[28] Wheatley S., Hofmann A., and Sornette D. Data breaches in the catastrophe framework & beyond. *eprint arXiv:1901.00699*, 2019.

[29] Palsson, K., Gudmundsson, S., and Shetty, S. Analysis of the impact of cyber events for insurance. *The Geneva Papers on Risk and Insurance - Issues and Practice* Computer Science. 4 June 2020.

[30] Hoerl, Arthur E., and Kennard, Robert W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *American Statistical Associate and the American Society for Quality Technometrics*, Vol. 42 No. 1, February 2000.

[31] Tibshirani, Robert Regression Selection and Selection via the Lasso. *Journal of the Royal Statistical Society Statistical Methodology Series B* Vol. 58 No. 1 : 267 - 288. 1996.

[32] Zou, Hui and Hastie, Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Statistical Methodology Series B* Vol. 67, Part 2: 301 - 320. 2005.

[33] Rothschild, Michael and Stiglitz, Joesph. Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information, *The Quarterly Journal of Economics*, Vol. 90 No. 4 : 629 - 649, November 1976.

[34] Sangari, S., Dallal, E., and Whitman, M. Modeling Under-Reporting in Cyber Incidents. *Risks* Vol. 10, No. 200. 2022. (https://doi.org/10.3390/risks10110200)

[35] Schwarcz, D., Wolff J., and Woods, D. How Privilege Undermines Cybersecurity. *Harvard Journal of Law & Technology* Vol. 36, No. 2. 2023.

[36] Jones, Jack (2021) *An Introduction to the FAIR Controls Analytics Model; FAIR-CAM.* FAIR Institute.