


# Modeling and Decision Making with Social Systems: Lessons Learned from the Fragile Families Challenge

Journal Title  
XX(X):1–11  
©The Author(s) 2017  
Reprints and permission:  
sagepub.co.uk/journalsPermissions.nav  
DOI: 10.1177/ToBeAssigned  
www.sagepub.com/  


Brian J. Goode<sup>1</sup>, Debanjan Datta<sup>1</sup> and Naren Ramakrishnan<sup>1</sup>

## Abstract

We created predictive models for six outcomes of 4,292 participants during the Fragile Families Challenge. Our submission was ranked 5th and 9th in the Material Hardship and Layoff categories. Motivated by reproducible and replicable research, predictive modeling is attracting attention for enabling transparent and verifiable models that are more robust to overfitting. However, model error is not guaranteed to be an adequate indicator when considering individual samples. Our Material Hardship model with low MSE and 21% baseline improvement will misclassify 69% of individual training samples using the measure definition. Assuming minimizing individual risk to be a basic requirement of utility in social systems, a baseline is derived to fix model performance to use case tolerance. We present the submission methodology and the lessons learned from this exercise. Model prediction error alone may be insufficient performance criteria for high-risk systems, and we advocate for broader analysis of model utility: reductive leverage.

## Keywords

social systems, reductive leverage, computational social science, model evaluation

The proliferation of available data and increase of computational power has made quantitative modeling a focus of social science research. This capability has helped usher in a trend of understanding by predictive modeling due to the lack of reproducibility and replicability in more traditional explanatory reasoning, cf. (Collaboration 2015). A predictive model trains by updating its parameters using data with known outcomes, and predicts the outcomes of unseen data. Research in predictive modeling has made inroads into a number of social and behavioral science topics. By studying a number of surrogates for factors leading to massive protests, researchers Ramakrishnan et al. (2014) have shown the ability to predict protests with a lead time of over one week. In psychology, the case is being made to consider more predictive approaches in the course of scientific reasoning (Yarkoni and Westfall 2017). The epistemological argument is that knowledge of a phenomena can be validated through predictive value, but it may not have the same causal mechanisms of explanation. However, predictions are still the result of a subjective process that, like any research method, is open for scrutiny. In what follows, we analyze the impact of error choice and model correspondence on the results of our own submission in the Fragile Families Challenge (FFC). The lessons we learn from this exercise suggest that more consideration be given to the system within which models are anticipated to operate.

The goal of the Fragile Families Challenge was to predict six outcomes of 4,292 individual participants at the age of 15. One predictive model was submitted for each outcome: GPA, Grit, Material Hardship, Layoff, Eviction, and Job Training. Over 12,000 longitudinal features were collected by the Challenge organizers from parents and the participants themselves over a timeline spanning from birth to age 9. There are two broad classes of models (Brieman 2001)

to formulate predictions from these features: data-driven (statistical) and algorithmic. The first effectively treats the system as a black box and improves as parameters are trained to capture the relevant signals in the data, and not the noise (e.g., overfitting). The second identifies the internal workings of the system and derives from specific theories in social science (e.g., compartmental models or structural equation models). Given the vast amount of data and time constraints to produce a submission, the models are largely data-driven with imputed data values that derive from the survey procedures. It is not necessary in this Challenge that the models themselves correspond to actual social science artifacts as long as the output sufficiently matches the outcome for which the model is derived. We show that this correspondence criteria can be relaxed if more is known about the model application, and that design decisions such as choice of error necessarily effect correspondence.

Models can be evaluated along several dimensions: falsifiability, explanatory adequacy, interpretation, faithfulness, goodness of fit, generalizability and complexity. The Challenge criteria evaluate goodness of fit on test data using the mean square error (MSE). In this paper, we will extend the analysis of our FFC results to include performance in an application scenario in addition to conventional error metrics. The results of the submissions are presented in two reference frames: reduction and leverage. Reduction is a traditional performance measure like mean squared error that

<sup>1</sup>Discovery Analytics Center, Virginia Tech

## Corresponding author:

Brian J. Goode, Discovery Analytics Center Virginia Tech, Arlington, VA 22203, USA

Email: bjgoode@vt.edu

determines how well model parameters fit the Challenge data (e.g., proximity measures of outputs). Leverage measures the performance of a model when applied in a system and is meant to exhibit utility. From the utility perspective, there are many different reasons for having a fitted model, cf. [Epstein \(2008\)](#). Forming predictions similar to the training set is one possible use case. A model trained to accurately predict can also be useful for sensitivity analysis and understanding parameter effects. To demonstrate leverage, we form a scenario where model outputs are required to form groups of samples using a threshold criterion. When the threshold corresponds to the definition of the outcome measure we find that MSE is not always a guarantee of performance as the choice of error forces a change in model correspondence. A driving motivation is that our Material Hardship model with MSE 20% below the baseline and ranked 5th in the Challenge will actually misclassify 69% of individual samples when the output has a threshold corresponding to the definition of Material Hardship. These subjective decisions in constructing the model-system can become obfuscated in the modeling process (e.g., [Wilson and Hardgrave \(1995\)](#)). Such systemic inconsistencies do not typically show up in model analysis, and will potentially brush up against ethical boundaries defining individual risk if placed in practice. We detail a different baseline measure based on output tolerance to use to ensure that use-case and model training choices are consistent.

The manuscript first outlines the novel aspects of the approach taken to form a FFC submission. Data imputation strategies are discussed in detail as well as the model training and validation steps. The paper concludes with the model analysis organized by four major lessons learned throughout the FFC exercise.

## Approach to the Fragile Families Challenge

The method used to create a submission followed these main steps:

1. **Imputing missing data.** When possible, missing values from parental and primary child-care providers were substituted with values from similar questions in other surveys. This assumes that the responses among the missing data are similar between parents and from year to year. Replacing these values accounted for roughly 31% of the missing data. The remaining missing values were replaced with the most frequent value.
2. **Question reconstruction.** For some of the outputs, custom features were created based on the construction of the survey question. To illustrate, Material Hardship is defined as an average of 11 survey questions (see Appendix A1 for details). This feature was reproduced in the input data. Additionally, this also acted as an OR feature by summing along all of the features that correspond to a particular outcome.
3. **Identify model classes.** Model classes for this submission were chosen purely on the output type. First, model classes for categorical outputs were chosen but had little success. Then, linear regressions and logistic regressions were chosen to directly

tune for the MSE metric. These became our final submission.

4. **Test models and choose the best given the MSE metric.** Using k-fold (10 fold) method that is customary to machine learning model development, the finalized models were formed into prediction submissions.

The novel aspects of this procedure are discussed below. In particular, human intervention was used to add features to match particular outcomes. A semi-automated process for identifying similar questions was used to impute data across survey respondents and waves.

### *Imputing by comparing across surveys.*

After removing features from the data that were either all null or exhibiting no entropy, roughly 44 million data points remained. Nearly 55% of these data points had null or missing values. There are a number of ways this can be handled using machine-learning only techniques. One of the approaches is to discard rows with missing attributes. This simple approach can have negative consequences resulting in skewed data and a significant loss of potentially useful information. Another simple approach is to use the mean which results in little statistical deviation in aggregate. But the assumption that the attributes are independent of each other can be wrong and negatively affect modeling. A popular approach is forming a regression on available data points to predict the missing variables (cf., [Takahashi \(2017\)](#); [Honaker et al. \(2011\)](#)). This may be not be a valuable approach in certain cases when data and model coupling lead to incorrect analysis.

Another way of thinking about data is that it is the result of a process. A close inspection of the surveys themselves show a number of repeated questions that exist across multiple pathways through the survey. For example, in the mother baseline survey, questions B5, B11, and B22, are all the same and read:

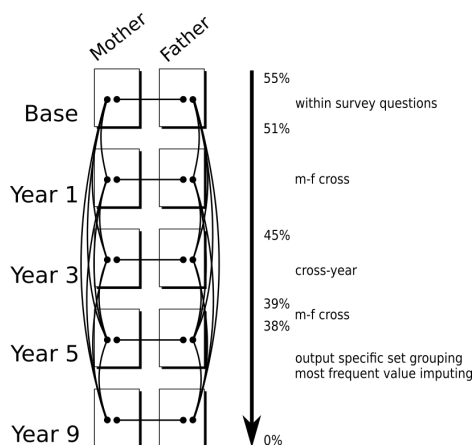
I'm going to read you some things that couples often do together. Tell me which ones you and [BABY'S FATHER] did during the last month you were together.

In any given survey only one of these questions would be answered depending on the path that the interview took, namely:

1. "Questions for mothers who are not romantically involved with baby's father."
2. "Questions for mothers who are in a romantic or 'on again, off again' relationship."
3. "Questions for married mothers only."

Therefore, it made sense to impute any missing values using the question that was actually answered, given that the other two would be missing by survey procedure. Matching questions on all surveys and using related answers within the same survey reduced the missing data percentage to 51%.

The surveys each year for the mother, father, and primary caregiver also had many similar questions. To account for situations where only one parent or primary caregiver could



**Figure 1.** Data reduction scheme for the Fragile Families Challenge submission. 1. Inner-survey question matching was used to fill in missing data from related questions within a given survey. 2. Cross-MF surveys matched across the same questions from the mother/father survey. 3. Cross-year surveys matched same MF surveys across years. 4. Question specific features were analyzed for having all of the necessary answers to support building the question. 5. The most frequent value was chosen for the remaining missing data.

be reached to take the survey, same questions were identified across surveys. If a question was missing a value, the value from the same question on the other survey was then substituted. This strategy effectively creates a “representative caregiver” feature wherein all rows can be compared with the same questions, but not placing emphasis on the gender or relation of the parent. For questions that involved household conditions and financial matters, it was assumed that this constructed feature would suffice. To correctly identify the same questions, ranges of related questions were made for each of the five survey blocks (Baseline, Year 1, Year 3, Year 5, Year 9). The structure of the survey made this a simple task. We label this substitution “Cross-MF” imputing, and it reduced the missing data percentage to 45%.

The last major operation performed on the data was a cross-year comparison of the same survey. Although the mother, father, and primary caregiver surveys varied from year to year, there are many related questions. To fill in missing data responses to similar questions from previous years were averaged and used to impute the missing values. This was primarily a factor in more recently conducted surveys. The assumption is that the average survey value would not change dramatically in the missing years. This is termed, “Cross-Year” imputing. It was more complicated to perform, because it involved comparing questions across five different surveys. The surveys also had similar questions with sometimes very different wording, and the task was too large to complete only by hand. However, the survey data was too small to adequately train an algorithm to detect similar features with a useful degree of accuracy.

A simple recommender was designed to suggest clusters of related questions across each of the survey years. The output used the NLTK toolbox (Bird and Klein 2009) text difference measure, and two thresholds were used to identify candidate questions. The first threshold was a strict threshold that determined the preliminary admittance into the set of potential candidates. Each set was delineated by a marker

f2j4	Are you/children currently covered by private health plan?
f3j4	Are you/your child(ren) currently covered by private health insurance?
f4j4	Are you or your child currently covered by a private health insurance plan?
f5g2e	G2E. You are currently covered by any type of health insurance?

f2h3	-	Approx how much could you sell this home for today?
f3i3	23.0	Approximately how much could you/they sell home for?(\$)
x f4c20a	22.0	Approx how much child support did you pay?
* f5i13	30.0	i13. how much you earn in that job, before taxes

**Figure 2.** Two samples of candidate sets shown from the recommender output are given. Questions without markers will be clustered and their values will be averaged to fill in missing data. Questions with an \* were labeled as not part of the set automatically and questions with an x were manually labeled. This procedure was a simple way to find related questions without resorting much manual effort writing code or comparing surveys.

such as that shown in the Appendix. The user would not see any related question clusters outside of the set. To reduce the chances of false-negatives, the primary threshold was set high so that less similar words would be admitted. However, if the user had to manually remove these entries, that would take more time. Therefore, a second threshold was set lower so that any entry included in the set, but also reasonably not similar in text was marked with an \*. As a result, the user need only remove the \* for questions that are to be included, and put an “x” next to questions that should not be included. This made it easy for a user to scan the sets fairly quickly and have only minimal input into the process by providing corrections. Once completed, a mapping was made with the output results to replace missing values. The result of this effort was to reduce the missing values to 39%. Another pass at the Cross-MF imputation reduced the missing values to 38%.

### Output-Specific Features.

Most of the outputs required of the models in the Fragile Families Challenge were based directly from questions or compositions of questions in the surveys. Specifically, these output are: Grit, Material Hardship, Job Training, Eviction, and Layoff. The mapping of questions to outputs are given in the Appendix. Assuming that an autoregressive model would be a useful predictor, each of the constituent questions were manually mapped between all surveys to ensure that missing data in these questions is substituted. For compositions of questions such as Grit and Material Hardship, each of the constituent survey questions was summed. This created an additional feature that both mimicked the result of the output and provided an OR logic in the feature data.

### Model Selection.

The scikit-learn package (Pedregosa et al. 2011) written in Python was the platform used to create and test candidate models. The significant details of the approach are given as follows. Initially, since all of the outputs appeared to be either binary or categorical, support vector machines (SVM), naïve Bayes, and logistic regressions were used to produce predicted outputs. Using k-fold validation on the training data, none of these models performed with any improvement over simple baseline measures such as choosing a stratified output from training samples on the non-binary outputs. Recognizing that the error measures required in the competition, mean squared error (MSE) and

Brier score, were both continuous measures, support vector regression and linear regression were used instead. We make particular note that this decision was not based on output matching, but rather on improving the chosen error score. This is a topic that will be revisited in the below sections. Ultimately, the submission used a linear regression for the categorical (ordinal) outputs, and a logistic regression for the binary outputs.

## Results

To evaluate how well the resulting models and data adjustments performed, the results focus on two components: aggregate error (reduction of the dataset) and individual error (model leverage in application). The Fragile Families Challenge used MSE to compare the model submissions, and is a common practice in machine learning. This represents the overall conformation of the model parameters to the dataset. We extend the error analysis to include the error of individual samples being located within a partition. Threshold partitions are used to evaluate model leverage, because it is the basic use-case requirement of any of these models to be able to discern groups at various output resolutions.

### Model Performance - Aggregate Reduction

The mean squared error is calculated as follows:

$$e = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (1)$$

where  $N$  is the number of samples,  $\hat{y}_i$  an individual sample from model output, and  $y_i$  the actual value. The Brier score is calculated similarly, but  $y_i$  is either 0 or 1. This is an aggregate error estimator, because it provides a point value that can be used to compare model performance over multiple datasets, sample sizes, and model types. In this way, it is a “black box” measure that is agnostic to how the output is generated as long as it was generated with the same input data used for comparison.

The modeling strategy had varying degrees of success in both the intermediate (i.e., leaderboard) and final test rounds shown in Table 1. Only one submission was given to the leaderboard, so in our results, both the intermediate and final test rounds offer points of comparing unseen data. This only applies to error and baseline comparisons as rank improvement results can include other submissions that could potentially have tuned to the leaderboard. Because we can compare both the final and intermediate results, we were able to make two judgments based on the data: a.) Was the model error consistently below the baseline? and b.) How elastic is the model error to changes in the baseline? The first question evaluates the traditional metric of model performance. The second, elasticity, evaluates the degree to which model performance can be attributed to model outputs from data compared to the baseline. Elasticity is calculated as follows

$$\epsilon = \frac{\Delta e_{model}}{\Delta e_{baseline}} \quad (2)$$

where  $\Delta$  is the difference between final and intermediate values. As elasticity approaches 0, the model is either using

the data to efficiently predict if MSE is low compared to the baseline, or it can be an indicator of weak weighting of input data. As elasticity approaches 1, the model is behaving more like the baseline measure which indicates a possible dependence on more frequent outputs or highly skewed output distributions. In this type of analysis, a “good” model will have consistently high baseline improvements, consistently low mses, and elasticity not in the neighborhood of 1.

In general, the aggregate error did not deviate very much in this model from the intermediate to final test rounds. Most metric volatility is seen in the baseline improvement and elasticity measures. In looking at baseline improvement, only the Material Hardship model showed a consistent prediction error well below the baseline measure with adequate elasticity relative to the sample. With respect to the range of models, it still showed consistent performance in terms of ranking. At  $\epsilon = 1.3$ , it does not show an elasticity near 1, when compared to the other output models. Although ranked high in the final test with the Layoff model, the baseline improvement is not consistent. The Grit model showed very little baseline improvement. Eviction and Job Training models showed more elasticity near 1, which indicates that their performance is closely tied to the training dataset output samples, because the error varies almost proportionally to the baseline with near equal MSE. These models will not be responsive to major changes in output sample frequency, and will show more error if there is model creep due to changes in the data. In practice, this results in more retraining and less confidence in the model itself. Finally, GPA is showing not to have performed very well, despite the increase in baseline improvement in the final round. In this case, the input variables with high weights in the GPA model are unlikely to correspond to changes in GPA.

Although we can only consider two data points, what we learn from this exercise is that the Material Hardship model is the best performing model from our submission in terms of MSE. It also appears to be able to use the input data with consistent performance relative to the baseline model without necessarily showing unit elasticity. We do note that the output data for Material Hardship did show heavy skewing with distribution mass close to 0. This impact on results will be discussed in more detail.

### Model Performance - Thresholding Leverage

Aggregate measures are agnostic to model type and make error comparisons subject to valid outputs being a subset of the measure support. For example, if the output is a rank, then it has values that are integers. Mean squared error has real numbers as support, and since the integers are a subset, MSE is a viable measure. In the design of the experiment, however, the choice of error can potentially have a dramatic impact on how the model is tuned and the conclusions that are drawn. To illustrate, the output type and chosen error measure for each outcome is shown in Table 1. The ordinal output type indicates that the output set can be placed into order and has the property of transitivity. The binary output type indicates that the output can either be one of two values: 0 or 1. The ordinal output types in this study are further constrained in that they are finite sets. Both Grit and GPA can take one value



**Table 1.** Results of the final and (intermediate) submissions using aggregate measures. These results do not include the “MDRC” submission updated on September 29th, 2017.

Output Name	Output Type	Prediction Error	Error Measure	Model Rank	Baseline Error	Model-Baseline Improvement (%)	Elasticity $\epsilon$
GPA (intermediate)	ordinal	0.365 (0.397)	MSE	36 of 124 (89 of 143)	0.425 (0.393)	14.04% (-0.99%)	-0.96
Grit (intermediate)	ordinal	0.255 (0.217)	MSE	77 of 111 (30 of 113)	0.253 (0.220)	-0.96% (1.20%)	1.15
Material Hardship (intermediate)	ordinal	0.020 (0.025)	MSE	5 of 113 (15 of 111)	0.025 (0.029)	21.51% (14.48%)	1.30
Eviction (intermediate)	binary	0.055 (0.053)	Brier	32 of 89 (33 of 86)	0.055 (0.053)	1.07% (0.94%)	0.95
Layoff (intermediate)	binary	0.164 (0.177)	Brier	9 of 91 (50 of 96)	0.167 (0.174)	1.66% (-1.34%)	1.72
Job Training (intermediate)	binary	0.185 (0.202)	Brier	46 of 94 (35 of 98)	0.185 (0.202)	-0.01% (0.02%)	0.99

in the set  $\{1.0, 1.25, 1.5, \dots, 4.0\}$ . Material Hardship can take one value in the set  $\{0, 1/11, 2/11, \dots, 1\}$ . In all cases, order is important from one group to the next. Both the MSE and Brier score error measures allow for models to output real-valued predictions and still produce valid comparisons. However, there is no guarantee that order is preserved with these continuous measures or that they will be within a given neighborhood of the true-to-definition value.

To demonstrate this point, results are given in Table 2 for how many individual samples will be incorrectly classified based on a chosen threshold,  $\tau$ . In each output, the threshold is chosen to be  $1/2$  of the distance between one value in the set and its adjacent value. A misclassification occurs when the residual of a sample,  $r = |\hat{y}_i - y_i|$ , is less than  $\tau$ . Misclassifications are therefore two-sided in this case, except for the output boundaries. For non-binary results, the residual mean,  $\sigma$  and standard deviation  $\sigma$  are given as approximate indicators of the shape and location of the residual distribution. These estimators are not given for the binary outputs as they are not applicable. The data used to calculate the residuals are the training datasets, which is a more conservative estimate than the intermediate and final test sets of which we do not have access. Only non-null samples were used in the calculations. Detailed figures of these errors are given in the Appendix.

The results indicate that all ordinal output types, regardless of scale, have more incorrect sample classifications than the binary outcomes. In other words, if these models were to collapse their output values to the defined output sets, they are incorrect for 68-82% of the samples, depending on the output considered. This is further illustrated by the 0 mean residual distributions that all show  $\sigma \gg \tau$ . The binary outputs appear to have less of an effect with respect to misclassifications, and this is likely to improve by using a posteriori threshold correction (Berger 1985). However, this is a correction for data and is not derived from output specification.

This result underscores the challenge that arises when designing models for high consequence decision making. In the case of this submission, our best performing model

for Material Hardship had the best aggregate characteristics relative to the other models in the submission. It was ranked in the top 5 models in the final testing round. However, following the definition of Material Hardship (see Appendix), the model is incorrect for 69% of the training samples. In traditional machine learning applications, like ranking search engine results (ordinal) or predicting whether an e-mail is SPAM (binary) these results do not carry as much consequence. However, in social systems as we present below, the ethical boundaries (Metcalf and Crawford 2016) force us to consider the potential high consequence and high risk in predicting outcomes related to human lives.

## Lessons Learned from the FFC

The results above show the importance of individual sample error when the output use has potentially high consequences, such as when making predictions regarding human lives. This places machine learning in quite an ethical and practical predicament. On one hand, in theory, machine learning can help reason about large amounts of data in social systems to lead to better decision making. On the other, it has the potential to make systemic errors of real consequence that may be unknown or obfuscated in regular practice. These results can be attributed to three issues: correspondence, aggregate v. individual error, and model-system design. In the case of correspondence we have seen where model outputs corresponding to certain value types can change based on model design and use. We have shown where aggregate error and its performance over the baseline does not necessarily translate into a proportional response for individual samples. The choice of error measure in modeling system design can affect both correspondence and error performance. We will ask four questions to help us draw lessons about how these models were developed and then used in the context of social systems like Fragile Families. The result of this analysis will be a more rigid roadmap for social modeling system design.

1. When is correspondence necessary in a social modeling system?

**Table 2.** Number of incorrectly partitioned outputs given the threshold,  $\tau$ .

Output Name	$\tau$	$\mu$ - residual	$\sigma$ - residual	# Misclassified Samples	# Samples	% Incorrect
GPA	0.125	4.463e-12	0.560	982	1165	82.58%
Grit	0.125	4.584e-12	0.430	1114	1418	78.56%
Material Hardship	0.045	1.894e-12	0.138	1001	1459	68.61%
Eviction	0.5	n/a	n/a	87	1459	5.96%
Layoff	0.5	n/a	n/a	343	1461	23.48%
Job Training	0.5	n/a	n/a	265	1277	20.75%

2. Does aggregate error reasonably reflect the individual sample error within a social modeling system?
3. Do model comparison techniques change in a social modeling system?
4. What can we learn about the design of future social modeling systems?

These questions will be answered using a methodology we term *reductive leverage* (RL): analyzing model fit and usage. We will expand analysis of a model to include aspects of the model-system. The model-system is defined to be the entire system that supports model development and usage including: data generation processes, human interactions with the experiment (i.e., assumptions, design), and the connection to a use-case (action taken based on model output). A social modeling system is defined to be a model-system that is used for studying social systems (i.e., longitudinal outcomes, crowd dynamics, etc.). In the following discussion, we assume the following:

1. Social systems are high consequence: individual sample error is important.
2. In machine learning models, the aggregate error remains persistent throughout model uses. Regardless of whether or not a given data set is useful, model parameters for a given model will reliably minimize an aggregate error metric. This results from validation in traditional machine learning.

We will present the lessons we learned from our participation in the Fragile Families Challenge. To illustrate our points, we will focus on our best performing model in the competition, Material Hardship, and illustrate how it functions within the context of a social modeling system.

### *When is correspondence necessary in a social modeling system?*

Correspondence, in this context, is the capability of a symbol to equate to a process or state. In the case of the Fragile Families Challenge, strict correspondence in model output for Material Hardship would mean enforcing that the Material Hardship model output only the set of defined Material Hardship values. We enforced this correspondence when the threshold,  $\tau$ , was set to map all of the model output values to one of the pre-defined outputs. This action assumed that the model outputs were to be as close as possible to the true output variables, as evaluated by the MSE. However, by using the MSE, this enabled our Material Hardship linear regression to actually perform better because the output correspondence was reduced. The values output

by the linear regression are close enough as measured by MSE. The correspondence requirement cannot be known until the output is used. This appears to be a very innocuous point, however, we will discuss how correspondence is not strictly required as long as the use case in the model-system is specified.

To illustrate the point, let us consider two different model-systems through conceptual metaphor (Lakoff and Johnson 2008). The first metaphor we consider in machine learning is *modeling is teaching*. Models are trained, they are tested, and then put to work. For statisticians, the same model-system can be cast as *modeling is shaping*. They are looking for a good model *fit*. In both of these metaphors, models are standalone entities whose output is subject to correspondence. In this view, the the output is evaluated by a process (cf. Kitcher (2002)) that compares to an entity that is meant to be the output itself. The expectation is that the resultant model will have parameters trained to allow the model to provide the correct output data. We have already seen how easy it is to have unmatched correspondence; the choice of error in the model-system design is enough to alter the correspondence. One only needs to admit predictions that are not defined by the output (e.g., continuous values for GPA, Grit, Material Hardship).

The second model-system we consider is a metaphor more common to control theory: *models are artifacts*. In this view, models are tools, where the functioning of a model takes place within the context of a system. The performance of a model need not be based on the model alone, but rather on the stability of the entire system. And, it need not locally have a correspondence criterion, and the correspondence that it does exhibit can be useful by virtue of the design and practice, cf. Winsberg (2006). For example, an adaptive flight controller may use a neural network, but the outputs are not required to be commands that a pilot would give (e.g., set heading 320 degrees). Instead, electrical signals suffice, so long as the downstream actuator or model in the system can use the signal without failing. The performance of this model is certainly analyzed, but within the context of the broader system reaching a goal. Even with humans-in-the-loop, correspondence has been elevated to the performance of the broader system (cf., Hutchins (1995)). The actual model outputs just have to be good enough (e.g., *satisfice*) to act on.

These are two different views on modeling and vary on the level at which correspondence occurs, and the amount of information known about the system. In the first, models are evaluated locally at the model output without considering how the model output will be used. The idea is that once

trained, a model can be applied to provide a similar output in a variety of systems (subject to retraining/refitting). By enforcing correspondence, less thought has to go into downstream interpretation (cf., Zipf (1949)). The model is useful by virtue of being trained accurately. However, for complex systems or inadequate data, this may not always be possible as specified by the output. In the second case, models are evaluated within the context of the system. The model does not need to have direct correspondence to a particular meaning, but it does need to produce outputs that allow the system to function. This weakens the restrictions on model training and accuracy, so long as the system-level correspondence holds. However, the success of these systems is in the ability to explain the system-level processes. For physical systems that are reproducible and repeatable, this is possible with some effort. For social systems, the prediction horizon can be much shorter due to more complexity in the dynamics. In the first case, the focus of evaluation is on how well the data is reduced to model parameters. In the second, the focus is on how well the model leverages output to drive the model-system components.

In the Fragile Families Challenge, the focus of the competition was to improve the model MSE. Models like our Material Hardship linear regression model output continuous variables, and correspondence was lost with respect to the original Material Hardship definition. As we have shown, when correspondence is again reinforced, the process has to be specified and the error of the model is not guaranteed to have the same performance with respect to MSE. If we were to adopt the second system metaphor, we do not have enough information on the social system to accurately specify the system dynamics with which the model would be interfacing. As discussed in the methods section, models with categorical output did not perform well, but this does not mean that the models are not useful. Rather, in the case of non-correspondence, we must specify how such a model can be leveraged when its output is used. Given what is known about the FFC model outputs (real), order preservation and the correct partitioning of categories is the basic requirement for further activity in the system. Therefore, we want to evaluate and specify the correspondence error due to partitioning at a particular threshold,  $\tau$ . As  $\tau$  is varied, we will see an inverse proportional relation between the output resolution and model accuracy. It follows that if model correspondence does not have to be exact in order to make improvements to accuracy, the model use case must be specified. In the case of the Fragile Families Challenge, we can relax correspondence with MSE, but we need to identify a threshold  $\tau$  and the likelihood of correctly partitioning the samples given the error measure.

*Lesson 1: Model outputs do not need a strict correspondence constraint. However, when lacking correspondence, the criteria on the output within the model-system must be specified.*

### ***Does aggregate error reflect the individual sample error within a social modeling system?***

In a social modeling system, individual samples are high-consequence and error in individual samples is generally more critical. However, in the Fragile Families

Challenge models are evaluated with aggregate error measures. Regardless of correspondence implications, we now investigate whether or not the aggregate error is an appropriate measure for model utility with respect to individual samples. We have already shown that our submission in the FFC does not have commensurate results between aggregate and individual error measures. In particular, we can simply look at the residual mean and residual standard deviation (i.e., Table 2) and reason that there is a high likelihood of misclassification. However, that is just one data point in terms of a model-system and it relies on the data in order to make the judgment. In what follows, we present a method for considering the relationship between the MSE aggregate error and individual threshold error. The procedure is agnostic toward the specific types of models and data that constitute the model-system. Specific details are provided in the Appendix.

The first step is to identify the different error topologies. The MSE is a pointwise measure that enables transitive comparisons of model performance along a line. Residual errors of  $N$  samples can be cast as vectors,  $\mathbf{r}$ , in an  $N$ -dimensional space. The key insight here is that for a given MSE error,  $e$ , there are an infinite number of  $\mathbf{r}$  that satisfy the MSE equation

$$R^2 = eN = \langle \mathbf{r}, \mathbf{r} \rangle \quad (3)$$

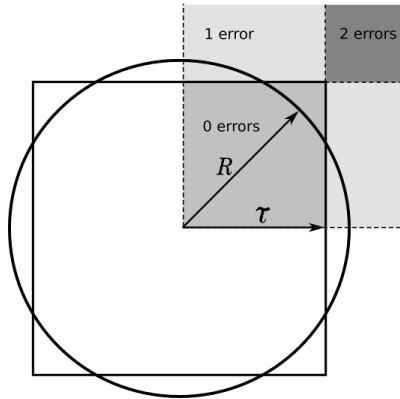
which is derived in the Appendix. With parallels to statistical mechanics,  $e$ , acts as a macrostate that is used to describe the many possible microstates  $\mathbf{r}$  much like temperature and particle velocity (cf., Krauth (2006)). Moreover, the  $\mathbf{r}$  vector components take values from an  $N$ -dimensional sphere,  $S_N(R)$  with radius  $R = \sqrt{e \cdot N}$ . In this case  $S_N(R)$  is the  $e$ -level set manifold of all possible residual values for each individual sample. This is shown for  $N = 2$  in Figure 3. By switching from aggregate error,  $e$ , to individual error  $\mathbf{r}$ , the topology of the error has shifted from being a point to an  $N$ -sphere.

To evaluate whether or not the aggregate error reflects individual sample errors we start by considering that the radius of  $S_N(R)$  is a function of both  $N$  and  $e$ . It also is directly related to the magnitude of each of the residual vectors,

$$R = \|\mathbf{r}\| \quad (4)$$

With uniform spatial deviations on the  $N$ -sphere, then as  $R$  contracts for a set  $N$ , then the residual magnitudes will reduce uniformly. However, there is no guarantee that all residuals will reduce uniformly. This is important when the location of a residual on the  $N$ -sphere leads to different errors. To illustrate, consider thresholding the residuals to determine when a sample will be incorrectly labeled based on threshold,  $\tau$ . For  $N$  samples, this threshold can be cast as an  $N$ -cube, where residuals outside of the cube will have some non-zero number of individual errors. This is shown in Figure 3, where for the case of  $N = 2$ , the two dimensional space is partitioned by crossing on the 2-cube (square). For  $R \leq \tau$ , then we will see correct partitioning for all samples. However, for  $R \geq \tau\sqrt{N}$ , there must be at least 1 sample that is incorrectly partitioned. With the condition,  $\tau < R < \tau\sqrt{N}$ , then incorrect partitioning will be linked to the residual location on the  $N$ -sphere. The probability,  $P(n_e >$

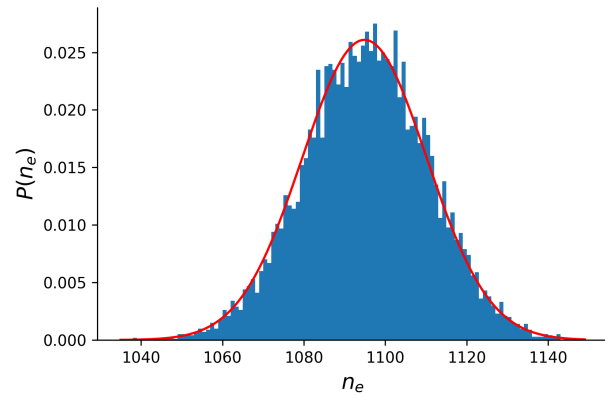
0), of any incorrect partitioning occurring corresponds to the ratio of the N-sphere manifold that is located outside of the N-cube. Likewise, the probability of a certain number of incorrect partitioning,  $P(n_e)$ , corresponds to the surface area of the sphere located in partitions corresponding to  $1, 2, \dots, N$ , dimensional threshold crossings. These partitions are shown in Figure 3 with the number of errors labeled in each region.



**Figure 3.** Illustrative diagram of the integral with  $N = 2$ . The circle (2-sphere) shows the level-set values for a given MSE error,  $e$ , where  $R = \sqrt{e \cdot n}$ . Without knowledge of the data or model (i.e., during system design), any point on the surface of the sphere is just as likely to be a valid microstate for a given macrostate,  $e$ . In other words, uniform priors are assumed. The partitioning action, in this case, is represented by the square (2-cube). Misclassifications occur when the microstate has a radius that exceeds the surface of the cube. The number of misclassifications depends on where the microstate is located with respect to the lower dimensional cubes on the overall threshold surface (shown in grey). These error regions have an area with respect to the unit cube that is equal to the binomial function. The probability density for the number of incorrect partitions is found by integrating the surface of the n-sphere within each of these regions and dividing over the total area of the hypersphere (e.g., microstate partition function). This integral is easier for lower dimensions, but becomes more difficult once  $N > 4$ . Therefore, direct sampling is used as an estimator. Results indicate that in the limit as  $N \rightarrow \infty$ , the density function exhibits Gaussian behavior. The distribution has characteristics of a general binomial distribution.

In general, calculating  $P(n_e)$  is not trivial for higher  $N$  due to the discontinuities between the N-sphere and inscribed N-cube. In the Appendix, we present the problem in a form that is more similar to a general binomial distribution (e.g., Drezner and Farnum (1993); Fu and Sproule (1995)). However, we use direct sampling to create an estimator,  $\hat{P}(n_e)$ . An example output is given in Figure 4 with  $e = .020$ ,  $\tau = .045$ , and  $N = 1459$ . Under these conditions, if a model has an MSE of .020 and binning threshold of .045, the probability density (blue) indicates that it is extremely unlikely that fewer than half of the individuals will be correctly partitioned. The upper bound for a finite number of samples is located at around 1150 incorrect classifications. Similar to the binomial distribution, in the case of propensity as  $N \rightarrow \infty$  or an infinite number of samples, the distribution appears to become Gaussian (shown in red).

**A new baseline for social systems.** What these results show is that, without using data in the formulation of the basis

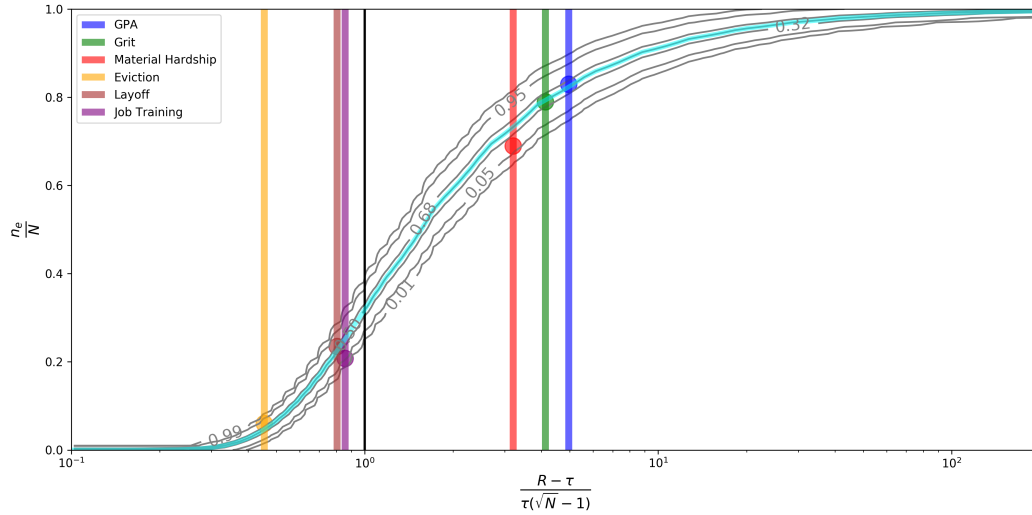


**Figure 4.** Sampling Result [MSE = .020,  $\tau = .045$ , and  $n = 1459$ ]. Results of direct sampling (Monte Carlo) estimation of the integral described above. Each cross-section of the lower dimensional n-cubes was used to find the probability of incorrectly thresholding  $k$  samples at MSE,  $e$ . The resulting distribution is approximately Gaussian (fit shown in red). In the design of a model design procedure, one would want to use this distribution to determine either a.) is the anticipated number of incorrect sample thresholds correct assuming that the MSE error remains persistent?, or b.) is the data having a significant effect on model training such as seeing evidence for improbable microstates over repeated applications.

for comparison, we can predict the probability distribution of individual sample residual outcomes given an aggregate error,  $e$ . In the above, the  $N$ -dimensional space is partitioned to threshold the residuals to indicate incorrect classification based on a uniformly applied threshold,  $\tau$ . As a result, we are able to obtain a probability distribution  $P(n_e)$ , that given any possible model satisfying the constraint of having aggregate error  $e$ , that  $n_e$  misclassifications will occur. This probability exists due to the effect of the spatial distribution of residuals in the partitioned space. However, any such criteria, such as likely individual residual value can be calculated and need not be restricted to the partitioned space mentioned here.

To answer the question as posed, aggregate error will correctly reflect individual error in the case when the spatial variation in the residual space is nonexistent or at least homogeneous. Such a case would occur in the FFC challenge for our Material Hardship model, if  $e < .045$  and we were interested in misclassifications of the Material Hardship ranking as defined. However, any greater MSE would result in spatial variation in the output space with respect to thresholding - the MSE would no longer represent the set of individual residual errors. Therefore, we present a new baseline for aggregate error,  $e_b = \arg \max_e \{P(n_e > \alpha) \leq \beta\}$ , where  $\alpha$ ,  $\beta$  and  $n_e$  must be established *a priori* based on use requirements. The probability distribution is shown in Figure 5 in normalized coordinates. A model-system designer would look to this distribution to find the probability shown in gray curves that corresponds to the individual error,  $n_e$  that is acceptable. This baseline would ensure that the probability of having more than  $\alpha$  misclassifications is less than  $\beta$ . For social systems of high-consequence to individuals, both  $\alpha$  and  $\beta$  would be rather low.





**Figure 5.** The approximate probability distribution for  $P(\frac{n_e}{N} < \alpha)$  [ $N=128$ ] is shown. Direct sampling was used to sample this surface. The curves in gray show the level set probabilities spanning the normalized error-threshold on the x-axis. The vertical black line shows the inflection point in the surface where error is guaranteed to occur. The lines with color represent the baseline chosen for each of the output measures. The circles show the output from Table 2. More accurate partitioning occurs in the region at the bottom-left of the axes.

*Lesson 2: When the spatial variation of residual vectors is non-existent or homogeneous with respect to the model-system output, then the aggregate measure will reflect changes in the individual sample. Otherwise, there will be some difference and variation in individual samples with respect to the aggregate measure. The degree to which this is acceptable depends on the model-system, and can be evaluated with the baseline measure proposed.*

### *Do model comparison techniques require a change in social modeling systems?*

In social modeling systems, we assume that individual samples or manifold locations of residuals are important. Are traditional aggregate measures adequate for social modeling systems? The advantages of the approach are numerous. The machine learning approach to aggregate error measure is agnostic toward model type and data input. This allows for the comparison of multiple models and data studies (e.g., ablation studies), but reduces the enforcement of adherence to model assumptions and correspondence to actual processes. The focus becomes predictor proximity to actual test measurements. The approach considered in the FFC reduces the entire modeling process to a single data point. This allows for a transitive comparison among models, making for easy ranking. The assumption is that the model has been tuned so that the aggregate error remains persistent subject to minor variations across test data sets.

A different type of analysis would be simply to look at the distribution of residuals as shown in the Appendix and in Table 2. By enforcing such a comparison, the assumption for model use is that the residual distribution becomes persistent across data sets. Such an approach would be a rigorous exercise, because the comparison would require comparisons across distributions. For normal distributions this is a parametric exercise, and for more complex distributions tests

for similarity like Komolgorov-Smirnov (KS) tests can be used. However, given variations in data and that one residual distribution is one vector,  $\mathbf{r}$ , the set of residuals covered by this error procedure is significantly reduced making it a much more difficult criterion to satisfy. At the other extreme, error forecasting can be avoided altogether, raising limited resource and ethical issues. Aggregate measures like MSE that reduce a model to one data point are useful in that they are an efficient way to make comparisons across models, while not enforcing criteria that is too strict and costly to produce.

So far, we have considered the comparison technique, but another major component is the baseline used for comparison. This method compares candidate models to a baseline model that does not use input data to generate predictions (e.g., the “dart throwing chimp” (Tetlock 2005)). The assumption in the baseline comparison is that improvements from the baseline are due to improvements in the model from use of data. We have shown in the results in Table 1 where the Material Hardship model performs well with respect to the baseline aggregate measure. However, the baseline measure does not translate into equivalent performance in the use case. Specifically, the model will misclassify roughly 69% of the individual samples included in the training set if restricted to the definition of Material Hardship. A non-data driven baseline derived from model output requirements would produce different conclusions. The baseline  $e_b$  is developed based on the probability  $P(n_e)$  of misclassifying individual samples. Such an approach maintains the merits of traditional model comparisons including transitivity and efficient use of data, but also takes into account the individual residual probabilities without being too restrictive. It has a direct correspondence to the number of individual samples being misclassified. In other words, instead of relying on assumptions of data quality,  $e_b$

only requires that the model-system design consider how the output will be used a priori.

*Lesson 3: There are numerous ways to compare models and they need not reduce to a single dimension. However, due to limited resources or ethical concerns it is unlikely that multiple models can be tested over different error measures. Therefore, if we are to choose one best model based on error, the error measure needs to satisfy the transitive property. The comparisons do not necessarily have to change for social systems, but greater consideration can be given to a model's leverage to establish new baselines of acceptable model tolerance.*

### **What can we learn about the design of future social modeling systems?**

The previous questions have centered around two main themes: how much is correspondence a requirement of a model and how do we account for model performance when individual samples are high-consequence? Our answers thus far have focused one particular model-system, our Fragile Families Challenge submission. We now try to generalize these concepts into a methodology that can be followed for the design of future related model-systems. We term this methodology reductive leverage.

Much of what has been discussed so far is the traditional viewpoint that a model's utility can be measured by its performance in the train, validate, test paradigm. Error measures of this kind are essentially measures of model reduction. Reduction in this case is how well the choice of model parameters can capture the relation of the input/output data pairs into model principles. For example, a linear regression encodes in its weights the input-output relation relative to linear proportionality. Likewise, if we were to deploy a neural network model, training the weights reduces the input-output pairs to connectionist reasoning. Traditional error measures such as MSE are estimators of this reductive process.

In contrast, many of the conclusions we have drawn for social systems subject to the assumptions given is that the utility of a model-system may not be reflected in a given use case. We argue that for high-consequence model-systems, we need to consider not just model reduction, but also how we expect it to be used within the more broadly defined model-system. The less that is known about the use case, the more correspondence becomes important. The more that is known about the use case, the more individual errors can be estimated given an aggregate error measure. We term the utility of the model within the model-system, its leverage. Together, a model's capability in terms of reductive leverage shifts from evaluating on a given objective truth at the model level to considering how well a model *resonates* with its environment.

Based on work presented so far, we have learned the following with respect to designing for reductive leverage:

#### **1. Decide on a model-system before analyzing data.**

- (a) Identify the use case.
- (b) Are the model outputs high-consequence?
- (c) What are the tolerances and requirements for model outputs?

#### **2. Choose an aggregate baseline relative to the use case.**

- (a) If the baseline depends on data, does the data reflect the use case?
- (b) If the system is high-consequence, does the baseline reflect a reasonable tolerance for individual sample error?
- (c) Does the baseline error measure have the proper correspondence needed for the model-system?

#### **3. Train, Validate, Test**

This procedure requires a priori output specification, which then specifies a model-system. The model-system satisfying these criteria is one that maintains a consistent aggregate error in a basin less than the output tolerance. In this way, models can be efficiently compared and still maintain guarantees of individual risk tolerance within decided bounds.

*Lesson 4: Both output correspondence and the significance of individual samples becomes of greater concern in social systems where the outcome of modeling can have potentially higher consequences. To better understand the implications of a model-system, we posit looking not only at the reductive capabilities of a model, but also its leverage in use cases. Further, we posit that this should be done in the design phase while making no assumptions on the data itself.*

## **Conclusions**

This paper presents a submission to the Fragile Families Challenge and the subsequent lessons that were learned during development, testing, and post hoc analysis. The novelty in the development approach lies in the use of the survey questionnaires over the various waves in order to match similar questions for more process-oriented data imputing. Using this approach and selecting features based on the output measure definitions, this imputation strategy reduced the number of unavailable data from 55% to 38% before filling in with most frequent values. As a result, the models in this submission performed mostly better than the baseline with Material Hardship performing very well in both error and ranking. Subsequently, we analyzed the submission along several dimensions and discovered that there is a discrepancy between the MSE measure and individual partitioning according to the output measure definitions.

To account for our results, we analyzed the submission in terms of a model-system. Using reductive leverage methodology, we explored model correspondence requirements as well as accounting for different types of error discrepancies across use cases. Based on our findings, we developed a model-system strategy for developing a baseline that does not depend necessarily on data, but on required model tolerances based on defined use cases. As a result, when considering models in the context of social systems where individual sample outcomes can be of high-consequence, we advocate for the development and specification of how the models will be used to determine whether the results are to be considered acceptable. Specifically, we advocate for using a baseline derived from output tolerances rather than training data.

## References

- Berger J (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer. ISBN 9780387960982.
- Bird EL Steven and Klein E (2009) *Natural Language Processing with Python*. O'Reilly Media Inc.
- Brieman L (2001) Statistical modelling : Two cultures. *Statistical Science* 16.
- Collaboration OS (2015) Estimating the reproducibility of psychological science. *Science* 349(6251). DOI:10.1126/science.aac4716. URL <http://science.sciencemag.org/content/349/6251/aac4716>.
- Drezner Z and Farnum N (1993) A generalized binomial distribution. *Communications in Statistics - Theory and Methods* 22(11): 3051–3063. DOI:10.1080/03610929308831202. URL <http://dx.doi.org/10.1080/03610929308831202>.
- Epstein JM (2008) Why model? *Journal of Artificial Societies and Social Simulation* 11(4): 12. URL <http://jasss.soc.surrey.ac.uk/11/4/12.html>.
- Fu J and Sproule R (1995) A generalization of the binomial distribution. *Communications in Statistics - Theory and Methods* 24(10): 2645–2658. DOI:10.1080/03610929508831639. URL <http://dx.doi.org/10.1080/03610929508831639>.
- Honaker J, King G, Blackwell M et al. (2011) Amelia ii: A program for missing data. *Journal of statistical software* 45(7): 1–47.
- Hutchins E (1995) *Cognition in the Wild*. A Bradford book. MIT Press. ISBN 9780262082310.
- Kitcher P (2002) On the explanatory role of correspondence truth. *Philosophy and Phenomenological Research* 64(2).
- Krauth W (2006) *Statistical Mechanics: Algorithms and Computations*. EBSCO ebook academic collection. Oxford University Press. ISBN 9780198515364.
- Lakoff G and Johnson M (2008) *Metaphors We Live By*. University of Chicago Press. ISBN 9780226470993.
- Metcalf J and Crawford K (2016) Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society* 3(1): 2053951716650211. DOI:10.1177/2053951716650211. URL <https://doi.org/10.1177/2053951716650211>.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.
- Ramakrishnan N, Butler P, Muthiah S, Self N, Khandpur R, Saraf P, Wang W, Cadena J, Vullikanti A, Korkmaz G, Kuhlman C, Marathe A, Zhao L, Hua T, Chen F, Lu CT, Huang B, Srinivasan A, Trinh K, Getoor L, Katz G, Doyle A, Ackermann C, Zavorin I, Ford J, Summers K, Fayed Y, Arredondo J, Gupta D and Mares D (2014) 'beating the news' with embers: Forecasting civil unrest using open source indicators. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*. New York, NY, USA: ACM. ISBN 978-1-4503-2956-9, pp. 1799–1808. DOI:10.1145/2623330.2623373. URL <http://doi.acm.org/10.1145/2623330.2623373>.
- Takahashi M (2017) Multiple ratio imputation by the emb algorithm: Theory and simulation. *Journal of Modern Applied Statistical Methods* 16(1): 34.
- Tetlock P (2005) *Expert Political Judgment: How Good is It? how Can We Know?* Princeton paperbacks. Princeton University Press. ISBN 9780691123028.
- Wilson RL and Hardgrave BC (1995) Predicting graduate student success in an mba program: Regression versus classification. *Educational and Psychological Measurement* 55(2): 186–195. DOI:10.1177/0013164495055002003. URL <https://doi.org/10.1177/0013164495055002003>.
- Winsberg E (2006) Models of success versus the success of models: Reliability without truth. *Synthese* 152(1). URL <https://doi.org/10.1007/s11229-004-5404-6>.
- Yarkoni T and Westfall J (2017) Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science* 0(0): 1745691617693393. DOI: 10.1177/1745691617693393. URL <https://doi.org/10.1177/1745691617693393>. PMID: 28841086.
- Zipf G (1949) *Human behavior and the principle of least effort: an introduction to human ecology*. Addison-Wesley Press.

## Acknowledgements

The analysis portion of this research was partially supported by DARPA Cooperative Agreement D17AC00003 (NGS2). Thank you to both Dichelle Dyson and Samantha Dorn.