# NGS2/Montage Monthly Report (09/15/2017) Supplement

## *Reductive Leverage Analysis*

**Motivational Narrative:**

An example of navigation is expertly written by Thomas Gladwin [1] in his discussion of Polynesian navigation in the Puluwat atoll in the Caroline Islands of Micronesia. Their form of navigation uses stars, waves, and “invisible” islands for dead reckoning. It works. So does a GPS and satnav. Each method, in its own way is a model of location and set out to accomplish the same end and satisfy existence through utility. However, each form of navigation has a different set of requirements for infrastructure, teamwork, and navigator training. They are completely different systems (c.f., [2]). How do we reason about them? Modern techniques would apply operationalism. If measuring by instantaneous location, GPS will have the better metric. If measuring by total effort and energy required in man-hours to date (power generation, satellite launching, development, maintenance, etc.), the invisible islands will be more efficient assuming). If measuring by success in reaching their destination, they will be the same. This line of reasoning can quickly degrade into a chaos of counterexamples and counterfactuals (e.g., GPS is more accessible, Puluwat navigation is more robust to technological failure, what if you are stuck on an island, etc.). The point, however, remains that whatever perspective of line of thought when evaluating these two models must take into consideration the environment within which they operate. In conclusion, operational measures are not guaranteed to be unbiased or agnostic to the ends of the agent or system implementing such measures. This is perhaps not very controversial (hopefully), but as far as I know, there is no methodology that systematically explores the interaction between a model and its environment. Reductive leverage analysis (RLA) seeks to make these interactions explicit and not reliant on vague correspondence (e.g., “this model has a better fit… to or for what?”). Furthermore, RLA seeks to avoid making false dichotomies resulting from general and possibly overreaching comparisons.

**Fragile Families Challenge:**

The Fragile Families Challenge is a data challenge design that utilises a common machine learning paradigm: provide training data, withhold test data, and compare model predictive performance based on common understood measure. Virginia Tech was provided with 4,292 rows corresponding to individual cases with over 12,000 features of survey data, objective testing, and observational data. For binary outcomes (True/False), the program utilized a Brier Score to evaluate relative model performance. For outcomes corresponding to output on a number line, the mean squared error (mse) was used. Both measures, by definition and challenge design, have a topology wherein model comparisons become transitive pointwise comparisons on the number line (see Fig. 1).

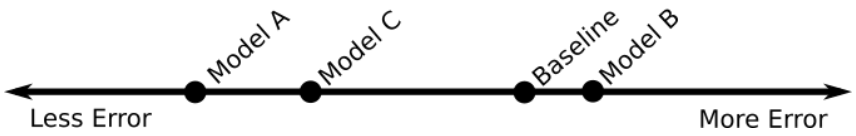


**Figure 1.** Error comparisons for models aggregate sample errors to provide an estimate of the model error that can be compared pointwise. If this value is the true measure of model adequacy when functioning in an applied environment (e.g., “making decisions”), then the measure holds. However, we show below that only slightly changing the *way the model is used* can alter the topology of the error surface considerably.

**Reductive Leverage Analysis:**
Reductive leverage analysis asks the question: “How well do the reductions specified by a model represent the data through the ***process*** of choosing parameters, and then how well does the model ***process*** perform in the context of an application environment?” In a way, the challenge measures (mse and Brier score) as conducted in the study satisfy the conditions measuring the reduction process (e.g., mse measures the degree that linear proportionality describes experimental data in a linear regression, and how well it can reproduce output on test data). We will focus on the leverage aspect of the problem.

*Model Leverage.* Models are used for a variety of reasons (c.f. [3]). Whether it is for explanation, prediction, or sensitivity studies, the assumption is that the output of a model will be used or acted upon in some way - existence through utility. When we look at a model’s leverage, we are looking at the degree that the model will be useful given some metric within the context of an application environment. The part of the environment that is considered must be specified in advance as shown in Figure 2.

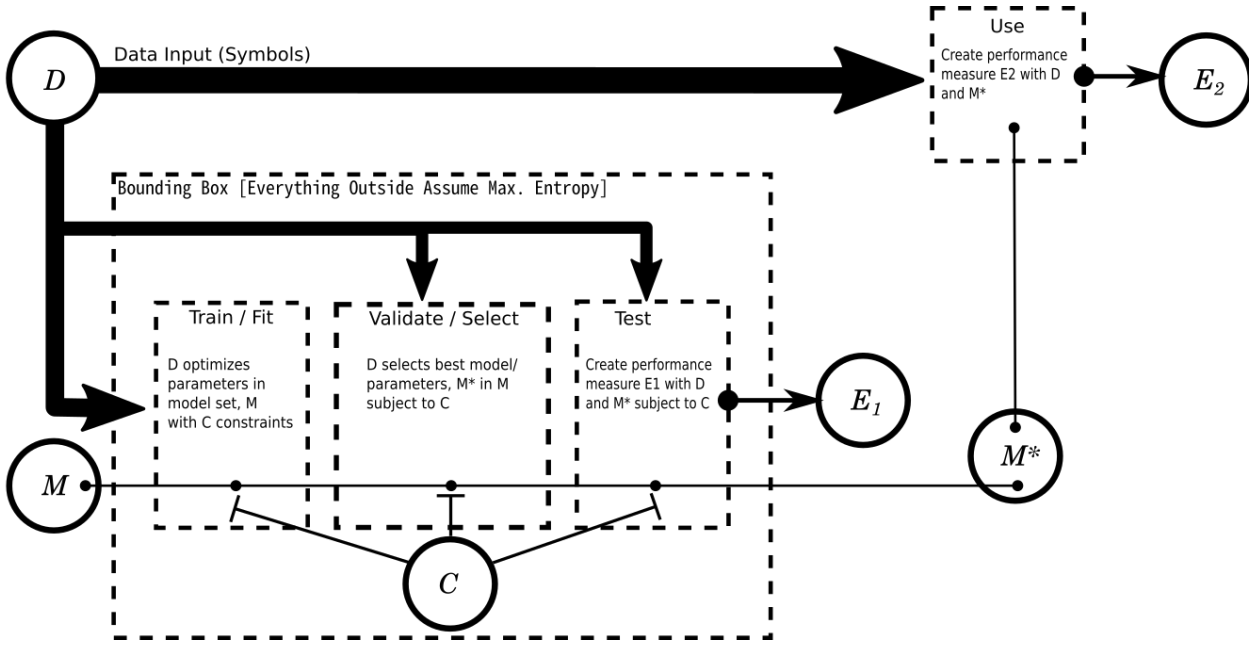


**Figure 2.** The modeling system. Typically, a modeling system will consists of the inner bounding box that focuses on the train/validate/test paradigm. In the traditionals setting, a set of models, M, is provided along with a set of symbolic data, D. Each of the models undergoes training - parameters are selected that best reproduce the data D split by input/output pairs. The choice of parameters is governed by a set of constraints, C. An example of a constraint is the type of output based on an error measure along with the pragmatic operation of the system subject to memory, time, and processing requirements. One of the models is then selected (incl. hyperparameters) in the validation phase, and the final performance is given by E1 in the test phase with data in D not yet seen. Reductive leverage analysis specifies this system, and by doing so, we can specify use cases that based on D and M*, the trained model, produce another performance metric E2. RLA is an analysis tool to specify the assumptions of the modeling system to better understand the relationship of M, D and E2. We show that E1 performance is not a guarantee of E2 performance.

In context of the Fragile Families Challenge, most of the focus was in tuning the material hardship model. Material hardship evaluates the answers to 11 parental survey questions and ranks the amount of material hardship in the set {0, 1/11, 2/11,...,1}. There are 12 possible values that the material hardship measure can have, by definition. With respect to the challenge criterion, we chose as our model the one that minimized the mse, which happened to be a linear regression model with L1-penalized weights (LASSO). Once submitted, our study appears as having beat the baseline measure by 14%, and is the top 15 of the models submitted in the category (most were at the baseline as shown in the histogram).

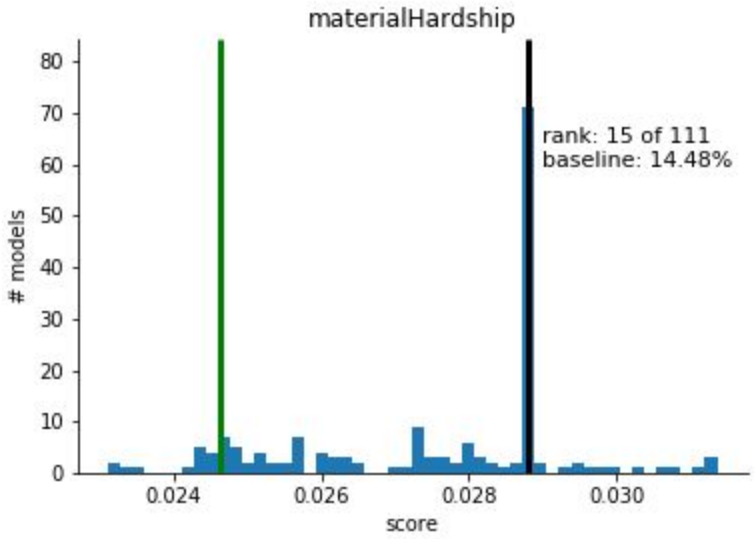


**Figure 3.** Performance of the Material Hardship model (green) with respect to the baseline (black). The error score (mse) frequency histogram is shown in blue. The majority of the models were centered around the baseline. Our model showed a 14% improvement over the baseline and was the top 15 model. In what follows, we will show that this information alone is not always sufficient to reason about the model performance in the context of actual use. However, within the bounds of the competition, this is exactly what is done. Therefore, we make the case that model design, testing, and evaluation should consider the entire process of model development and usage.

At first pass, one might conclude that given the mse scores of the baseline (black vertical line), the odds of making a correct prediction are already fairly good. Furthermore, by adding in data, our approach was able to improve upon the baseline by 14%, a broad jump given the overall distribution of scores among all the participants as shown in the histogram. And, if the process by which the model is used is ***exactly*** as is performed in the training/testing process, this conclusion has a higher probability of being true. More rigorously, if the model's use corresponded to the continued reliance of the mse L2 measurement error and the data is assumed to add value a priori, then the pointwise transitive comparison continues to hold. However, when people's lives are at stake, or populations are being selected for funding or appropriations, individuals are usually the unit of focus - meaning L_\infty measure over aggregated samples.

Let us, therefore, consider an alternate scenario to show the point. Consider an application where funding will be appropriated to different bins of people facing different levels of material hardship. In this case, the specifics of the funding are not important, as long as the rank order of individuals remains correct in the partitions. Therefore, to leverage this model, individuals must

be correctly partitioned, or else face a difference in their allocated funding. Although contrived in this document, there are many real scenarios that are similar that have potentially long-lasting and persistent effects on people's lives (e.g., https://www.fns.usda.gov/snap/eligibility#Income).

In doing our leverage analysis, we want to find the number of people that will be out of rank order in a given partition, where a partition is defined by a threshold \tau. For the purposes of illustration, consider Figure 4, where the partition threshold corresponds to the definition of material hardship itself (±1/22). All of the samples from the model above falling outside of the threshold are shown outside of the bounds of the vertical bars. In total, this amounts to 1001 individuals or 68.61% of the population in the training dataset being outside of their requisite partition.

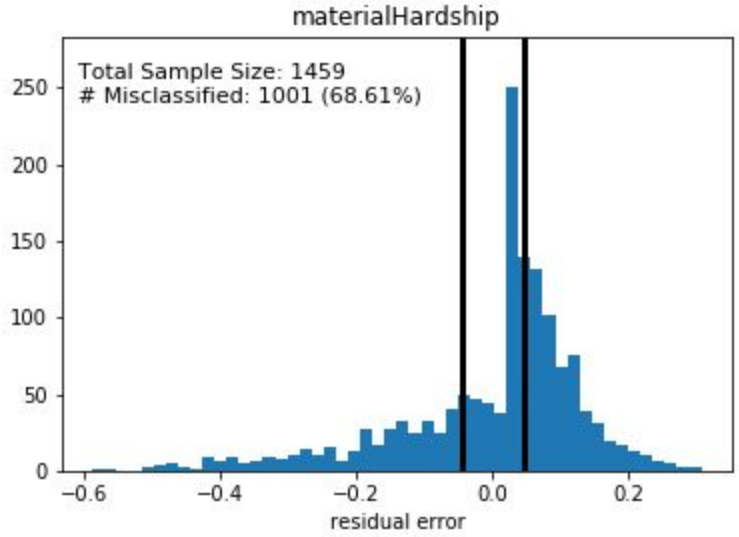


**Figure 4.** Errors that exceed the specified threshold are shown as outside the threshold bars (black). These samples will be partitioned incorrectly.

It is important to consider the entire system in which the model is involved. The above example illustrates the simple case where models are required to be evaluated by one metric - mse (L2), but when faced with another application, partitioning, a similar proximity error shows drastically different results. If we were to use the latter error to reason about the models, a large amount of data would be required as the error is calculated off of one data instance and does not have much power (it is "anecdotal"). In both calculations, the training data (input/output) is assumed to have some value in aiding in model convergence, at least in aggregate.

*A Priori Leverage Calculation.* Assume that we do not place any assumptions on the data used to train a model. The model will have some error, e. For the model comparison scenario, e, satisfies the pointwise comparison along a line, and in the Fragile Families Challenge, e is an L2 measure. However, if we consider applications that rely on individual outcomes as in the second scenario, the pointwise measure becomes a level-set of magnitude R=\sqrt(n*e) in a space of n-dimensions where n corresponds to the number of samples. The partition function, which defines all of the possible values for that particular level set is determined by type of measure. In the case of mse, it is a n-sphere of radius, R. With no assumptions placed on the data, any single value on the surface of the n-sphere is equally as likely to occur. To calculate the

probability of k miscalculations of the individual thresholding, we have to evaluate the integral of the surface of the n-sphere that extends beyond the surface of the n-cube with binomial partitioning. In practice, this is a tough integral to evaluate in high dimensions.

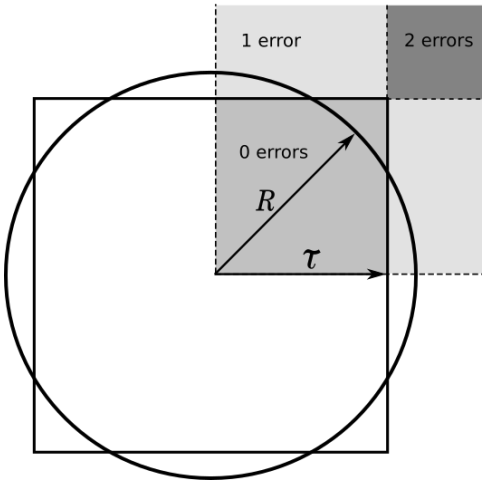


**Figure 5.** Illustrative diagram of the integral with n=2. The circle (2-sphere) shows the level-set values for a given mse error, e, where R = \sqrt(n * e). Without knowledge of the data or model (i.e., during system design), any point on the surface of the sphere is just as likely to be a valid microstate 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 -> \infty, the density function becomes Gaussian indicating that the overall distribution is most likely a generic binomial function - subject to proof.

Direct sampling can be used as an estimator (for algorithmic details, see: [4]). An example output is given below with mse = .025, \tau = .09, and n = 400. Under these conditions, if a model has mse = .025 and binning threshold of .09, 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 260 incorrect partitions. In the case of propensity, or an infinite number of samples, the distribution, which is a general binomial distribution (non fixed, p) appears to become Gaussian (shown in red). However, this has not been proved, yet.

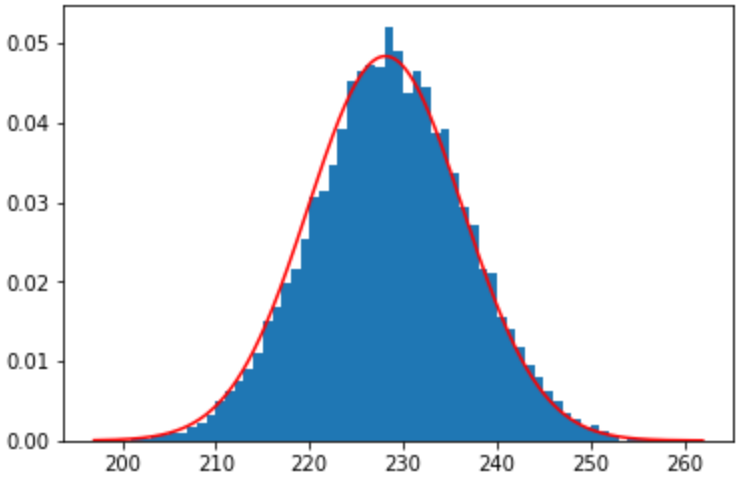


**Figure 6.** Sampling Result [mse = .025, \tau = .09, and n = 400]. Results of direct sampling (Monte Carlo) estimation of the integral described above. Each cross-section of the lower dimensional n-cubes was used to find the probability of incorrectly thresholding k samples at mse, e. The resulting 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.

***Preliminary Conclusions.*** Once again, this calculation is both model independent and data independent. The identification of this distribution only takes into account the change in topology of the error comparison from a linear to n-dimensional space. Specifically, this identifies the number of individual threshold errors given a persistent L2 error. If one were to use the models output from this competition for the purposes of identifying “at-risk” (or, alternatively “not-at-risk”) populations based on the material hardship measure, they would be hard pressed not to require an error low enough to minimize the chance of misallocating the partitions.

***What does this mean?*** There are three major components to designing a model. The model, the procedure, and the data. Much emphasis is given to the model and getting the right “fit”. This is a very specific metaphor that only applies to the model as a standalone entity. In the case of the above example, the design of the procedure - using mse error - played a significant role in how models were evaluated and reasoned. Furthermore, we must question the utility of data. Using reductive leverage analysis allows us to investigate these areas, because the overall system and interactions must be specified.

- **Model Evaluation.** Models are still evaluated in the traditional manner. The difference is that by including the use of the model, the topology of the error space can change quite significantly as shown (point to n-sphere). Reductive leverage analysis takes into account what the error can mean in different scenarios.

- **Procedure Design.** Reductive leverage analysis better informs modeling procedure design, because performance is no longer based on a single model fit with specific data, but analysis of all the possible models that could result and the probability of an individual instance subject to application.

- **Data Evaluation.** Does a dataset have value for an application? Improvement over a baseline cannot necessarily answer this question except if the training and application processes are identical, because the baseline includes the data in question. The calculation outlined above does not involve the use of data. Instead, it assumes maximum entropy over the individual model instances that could result with a given error level-set. Data can really only be evaluated over several instantiations. Therefore, if there are consistently lower probability error configurations (microstates), then this is additional evidence that the data is making an impact. Traditional procedures assume the best in data; this anticipates the worst, but tests for the best.

In summary, depending on the application environment for a model, the "fit" of the model may alter the error topology, and therefore our ability to properly anticipate the effect of the model's intended or unintended consequences. Although certainly not applicable to every problem, the procedures discussed above provide an outline for how to consider the greater system within which the model acts. This is goal of reductive leverage analysis.

**Objections:**
There are three potential objections of varying strengths that we consider.

1. One possible objection to this work would be that it does not take into account individual model behavior - the assumption of uniform sampling over the n-sphere. However, this is precisely what a pointwise measure does. There simply has not been enough work in this area to properly reason across models as described in the motivating narrative at the beginning. Indeed, our method does provide a mechanism for including individual model characteristics going forward. These individual model characteristics and tuning processes will affect the a priori probability that a particular microstate (set of sampling errors) is chosen off of a steady state surface.

2. A second objection is the assumption that the error level-set will be persistent from the training phase. At this point we can only say that this is merely an extension of the current methodology. We do not attempt to test or correct this at this time, but it is wholly consistent with existing modeling procedures.

3. A third possible objection is that this is just one system, and one that can be easily corrected by utilising a different error metric and using different (ordinal) models. It is true that a change in the design, such as using a classification error metric would solve the particular problem mentioned here. It is also true that using an ordinal regression would be better suited to the data. However, reductive leverage forces one to consider the design and application ***before*** work begins on modeling. So, in order to properly evaluate this object, we have to consider the timeline of information and the agency over different parts of the system. In this case, mse was chosen before involvement, and without an analysis like the above (before work began), it would be difficult to state the claims for a different metric. The second point to consider and more difficult to see is that although an output may not match *as defined*, it does not preclude it from having utility in the overall system. In this case, choosing a categorical error measure, which would enforce the definition of material hardship exactly on the output, may eliminate several models that could still have utility; e.g., rank order preserving continuous models. **As such, we are careful not to state the "correctness" or "incorrectness" of features**

**of the system - we only claim to show effects of those decisions on other parts of the system that may ordinarily not be considered.** To properly evaluate the output usability, we have considered a statistical mechanics approach. However, there are many approaches that could be used such as system feedback as a control system (if known and observable) to more qualitative measures such as conceptual metaphorical compatibility.

**What is next:**

1. Finish paper with FFC results (Oct. 1, 2017) [this has the mathematical details that were only sketched here]
2. Would be nice to come up with analytic solution to the integral discussed above - the properties can be very useful and easy to implement in real model-system development.
3. Generalize theoretic framework and assumptions more.

**References**

[1] Gladwin, Thomas. East is a Big Bird: Navigation and Logic on Puluwat Atoll. Harvard University Press, 2009.

[2] Hutchins, Edwin. Cognition in the Wild. MIT Press, 1995.

[3] Epstein, Joshua M., “Why Model?” Journal of Artificial Societies and Social Simulation. vol. 11, no. 4 12.

[4] Krauth, Werner. Statistical Mechanics: Algorithms and Computations. Oxford Master Series in Statistical, Computational, and Theoretical Physics. Oxford University Press, 2006.