Prospectus Outline

Do modern political science theories effectively explain the phenomena that are being studied? How can political scientists assess the validity of their theories? I believe these two questions capture the driving force of today's research agendas. Each researcher wants to understand the phenomena they have dedicated themselves to, and as a result they also want to provide the best evidence that their theory and hypotheses are correct and predict best (Muchlinski et al. 2016).

Null-hypothesis statistical testing (NHST) is the standard quantitative analytic approach in modern political science. Since the advent of personal computers, and the subsequent availability of empirical modeling software such as Stata, R, Python, etc. it has become increasingly easy to estimate both simple and complex models to assess the validity of theories. This general process adheres to the following steps:

1. Theorize and hypothesize about some phenomena.

2. Establish some a priori heuristic to judge empirical analysis of the theory. Most commonly this is setting an alpha level of .05.

3. Identify relevant data and collect it.

4. Estimate models to assess the relevant null hypotheses (H_0) identified in step one.

5. Using the heuristic established in step two infer what the models say about the validity of the theory and hypothesis of step one. Most commonly, if the p-value is less than or equal to the alpha level then the corresponding coefficient is considered statistically significant.

This dissertation will focus on proposing a new methodological framework for quantitative analysis in political science that focuses on evaluating how well theories explain political events. The NHST approach over past decades offers only weak tools for analyzing how effectively variables of interest explain the outcome, and inference based on coefficients related to hypotheses is the central focus. Instead, I propose a paradigmatic shift in what political science should consider paramount. Specifically, the ultimate litmus test for whether a theory or hypothesis is good or not should be whether or not it improves out-of-sample prediction. Any variable that does not improve out-of-sample prediction is very likely not theoretically relevant to the event that occurs in its current operationalization. My framework, framework for out-of-sample

prediction (FOOSP), argues that it is imperative to move theoretical evaluation toward prediction, and away from NHST.

NHST depends on p-values and statistical significance to determine the validity of theories and hypotheses. The p-value is the proportion of the distribution of the test statistic (typically t-test or z statistic) that remains in the tails beyond the observed test statistc. For an explicit definition see Gailmard (2014, p.244): "The p-value is the probability, assuming H_0 is true, that the random test statistic is at least as extreme (relative to the expected value under H_0) as its observed value." Put into a mathematical expression a p-value is simply $P(Data|H_0)$ (Levine et al. 2008). Thus, a p-value represents strictly the probability of observing the test statistic of a coefficient produced from the model estimated on a sample given H_0 is true.

A p-value has very limited information to infer from when used correctly. At its best, it tells the researcher if the observed relationship is likely or unlikely given the null hypothesis. However, this interpretation is often stretched beyond this meaning and is interpreted in ways that are not accurate or scientific. First, it is commonplace for no strict alpha level to be set, and thus the criterion for statistical significance becomes ad hoc. Most authors use multiple levels of significance and stars to denote different p-values (e.g. Piazza 2012). This practice is so ingrained and common that it is the default of the 'esttab' package in Stata, which ouputs regression coefficient tables. Second, it is becoming common to include relatively large p-values as significant such as .1 (e.g. Daxecker and Prins 2013; Berrebi and Klor 2006) although it is questionable at this point whether one should feel safe rejecting the null hypothesis. Third, the interpretation of these different alpha levels as more or less evidence for a theory is also common. Take for example Bleek and Lorber (2014, p. 442, emphasis added)): "... security guarantees remained insignificant, although they came very close to the 10 percent threshold. For the reproduced logit model, security guarantees remained *extremely significantly* negatively correlated with all the three stages of proliferation behavior." In this example a small p-value is "extremely significant" in the authors' interpretation despite there being no such thing as an extremely significant p-value. In truth, they can only conclude with a high degree of certainty that it is unlikely to observe that relationship given that the null is true.

P-values offer very little in the way of scientific study of a phenomena. At best, they offer a probability that the null hypothesis is correct given data. This will be discussed in greater detail later, but I am not alone in my staunchness against p-values. Discussing his home field of psychology Meehl (1979, p.817) wrote: "I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychologgy." While his sentiments are stronger than mine, I agree in principal and will show that a fundamental analytic shift is needed. Statistical significance does not tell a researcher anything about how well a variable explains the

outcome, and this is fundamentally the concern of scientific research.

Notably, there are goodness of fit tests for each model type, and different way to analyze variance explained. In OLS there is R^2, the percent of variance explained by the model, and $\bar{R}^2$, the same as R^2 but penalized for the addition of variables to the model. Logistic regression has receiver operating characteristics (ROC) curves and the area under the curve (AUC). Each of these tools is useful in understanding how a model is fit to the sample used to train it, but none of them provide any information of how a model predicts or explains outside of the sample nor do they provide specific evidence to caution against overfitting.

Also notable is that political science research almost always uses the full sample to train the model, and does not take into consideration overfitting the data. Overfitting occurs when a model "follows the errors, or noise, too closely (James et al. 2017, p. 22)." When a model is fit on the entire sample, and nothing is withheld to test a model on, overfitting is near impossible to detect. Models that are overfit produce coefficients that are partially describing randomness, and thus are not generalizable. Inferences that are not generalizable are not useful outfside of the given sample.

One further broad critique of the most common approach of political science quantitative research is the general lack of discussion of prediction. Given the central goal of explaining political phenomena, it is not possible to evaluate the explanatory power of a theory effectively without considering out-of-sample prediction. Any models not tested using out-of-sample prediction are prone to unidentified overfitting, and offer no evidence of a theories predictive power. A simple solution to this problem is to consider a model's predictive ability out-of-sample both with and without the variables of interest to the theory at hand. Then, the base model that excludes the new theoretical contribution can be used as a control for whether

The implementation of even simple machine learning approaches can resolve these problems. My framework is proposed as a set of increasingly more difficult tools to adopt, proposing that all research adopts the minimally invasive strategy in all contexts, and suggesting that all research strive to adopt as many strategies as possible. This framework also poses a fundamental shift from variable-by-variable significance testing to entire model (and thus theory) predictive assessment. This framework is based on the belief that a theory that cannot be used to effectively predict outside of the training sample is a bad theory, and thus should be discarded.

Problems with NHST

The abuse of p-values as a metric for whether or not an idea or variable merits publication in a scientific setting is a sufficient problem in all quantiative fields that the American Statistical Assocaition recently released a statement on the use of p-values (Wasserstein and Lazar 2016). In this there are six principles,

which are important to denote at the outset of the discussion of why NHST is problematic (pp. 131-132):

1. "P-values can indicate how incompatible data are with a specified statistical model."

2. "P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone."

3. "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specified threshold."

4. "Proper inference requires full reporting and transparency."

5. "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result."

6. "By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis."

The first point suggests that p-values can indicate whether or not data is compatible with the null hypothesis. Small p-values indicate incompatability. Point two is elaborated to explain that the p-value contains information only about the relationship between the data and the specified null and alternative hypotheses, and that is it. Point three argues that binary decision-points are arbitrary, and adjudicating a coefficient's importance based on a p-value is misguided and incomplete. Point four encourages the sharing of all specifications, which guards against the practice of 'p-hacking', when a researcher estimates multiple models and only reports the most favorable ones. Points five and six are self-explanatory.

The ASA statement affirms my argument above that p-values are a measurement regarding the probability distribution of a test statistic relevant to a null hypothesis. The use of a p-value outside of that setting is irresponsible and misleading. Nevertheless, the publication of insignificant, or 'null,' results is rare. This practice suggests that a significant coefficient is indicative of something important, and also indicates a belief as a field that significant results illuminate explanatory or predictive mechanisms. P-values provide no such mechanism, as will be discussed in detail below.

Breiman (2001) further illuminates general problems with the NHST approach: what he calls data modeling (as opposed to algorithmic modeling) focuses on fitting models to full samples, and assessing their inferential value based on goodness of fit and p-values. As will be discussed further this approach is flawed because goodness of fit metrics are poor determinants of model fit in high-dimensional models (Breiman

2001; Landwehr et al. 1984; Bickel et al. 2006), and goodness of fits do not serve as good adjudicators or which model is 'correct' in a set of similarly well-fit models.

Together, the faults of p-values and the limited nature of goodness-of-fit (GoF) analysis serve as two of the stronger arguments against NHST. Below, I enumerate many arguments against NHST, and expand on these.

Null and alternative hypotheses

The null hypothesis in most political science research is that an explanatory variable has no relationship with a dependent variable: $H_0 : \beta = 0$ where β represents the coefficient on a variable of interest from some model type. The alternative hypothesis in this context is almost always $H_A : \beta \neq 0$. In almost every setting it is likely not true that any explanatory variable and a dependent variable have truly 0 relationship (Levine et al. 2008, p. 177). Even if a relationship is small or mostly random it is not 0. Thus, the null hypothesis is never truly false, so the NHST framework is predicated on a strawman concept (Levine et al. 2008; Meehl 1986).

Furthermore, it is important to further discuss the implications of a rejected null hypothesis. Given a sufficiently small p-value that one feels confident rejecting H_0 then the alternative hypothesis remains. In practice, the alternative hypothesis is purported to be of whatever formulation the author's theory leads to and is built on mechanisms sourced from a theory.

Consider the timeless piece of Fearon and Laitin (2003), it is foundational in the civil war literature and has been cited 2118 times (according to the Cambridge University Press website, 6/28/2019). Their first hypothesis (p. 78) argues that more ethnically and religiously diverse countries will experience greater risk for civil war. Formally, $H_0 : P(Civil\ War|Diversity) \leq P(Civil\ War|\neg Diversity)$ $H_A : P(Civil\ War|Diversity) > P(Civil\ War|\neg Diversity)$. They use Stata and their replication materials indicate they run a simple logistic regression and estimate two-tailed z-tests to get p-values because that is the defualt in the Stata logistic command. Thus, despite their hypotheses being unidirectional, and one-tailed, the real stated null is $H_0 : P(Civil\ War|Diversity) = P(Civil\ War|\neg Diversity)$ and the alternative is $H_A : P(Civil\ War|Diversity) \neq P(Civil\ War|\neg Diversity)$. The significance of the coefficients offers only sufficient information to reject the proposed relationship given the null of no difference between diverse and not diverse settings, and does not give sufficient information to confirm the proposed alternative hypothesis. Inasmuch as one has confidence in the model's specification and lack of biased coefficients, one can extend interpretation *independently of the p-values* to the coefficients, while still only concluding in terms of statistical significance that the relationship between civil war and diversity is not zero.

Moreover, note that none of the empirically validated discussion of these models' findings relate to the

theoretical mechanisms proposed in the paper. They argue that ethnic diversity creates tension between cultures because of governmental policies that create barriers between majority and minority groups. The more diversity one country experiences the more likely theses barriers spark conflict. In measurement, diversity is measured as the proportion of the population made up by the largest minority group. So, once more the null hypothesis is mispecified. In reality, the null is: $H_0 : P(Civil\ War|Large\ Minority) = P(Civil\ War|\neg Large\ Minority)$. And the alternative is: $H_A : P(Civil\ War|Large\ Minority) \neq P(Civil\ War|\neg Large\ Mino$

Thus, a further abstraction is drawn between what is theorized and hypothesized and what is actually tested. It is difficult to impossible to perfectly measure theoretical concepts and hypothetical mechanisms, but it is important nonetheless to assure as much as possible that analyses reflect the hypotheses, and vice versa. Despite the fact that their results do not support this hypothesis, their analysis still shows signs of their readiness to misinterpret coefficients based on p-values. In discussion of the per capita income coefficients they say "... is strongly significant in botha statistical and substantive sense ..."(p. 83). The notion of 'strong statistical signifiance' is completely misguided. A small p-value on the per capita income variable denotes that they are relatively confident in their ability to reject the null hypothesis that per capita income has no relationship with civil war onset, it does support the claim that per capita income definitely has a relationship with civil war onset.

In sum, the NHST setup severely limits the intrinsic connection between what the null hypothesis being tested is and what the authors propose. In the seminal work of Fearon and Laitin there are two manifestations. First, they hypothesize about ethnic diversity, but only measure size of the largest minority. Thus, whether the empirical null hypothesis given that operationalization and their written implied null hypothesis is up for debate. Second, the interpretation of coefficients as strongly significant and rejecting the null hypothesis as validating the alternative hypothesis is incorrect. In a following section I discuss how $P(Data|H_0) \neq P(H_0|Data)$.

The impossibility of meeting a model's assumptions

Also, this process is heavily dependent on the residuals of a model, which are dependent on meeting a model's assumptions. Consider simple linear regression (OLS) assumptions: (1) a model is linear and correctly specified, (2) there is not too much multicollinearity, (3) the disturbance term has zero expectation, (4) the disturbance term is homoscedastic, (5) the values of the disturbance term have independent distributions, (6) the disturbance term has a normal distribution (Dougherty 2011, pp. 159-160). If each of these assumptions is not met then either the coefficients or the standard errors will be biased or wrong, and inference based on that model is wrong.

As a simple proof of concept for why NHST with OLS (and all of its variants) are poorly suited for

the study of international relations consider a setting in which the unit of anlaysis is country-year and the dependent variable is trade outflow or some continuous economic indicator.

First, it is practically impossible on multiple fronts to satisfy the first assumption because it implies that all relevant variables are in the model. The only way the true model can be specified, and thus ommitted variable bias is eliminated, is if there is a way to know the true model. If there is no randomness in the dependent variable at all then it is possible to have a model with perfect fit, but that leads to the second near impossibility: identifying and gathering sufficient data and accurately measuring all of the relevant variables. Even given a list of what the perfect model is, obtaining complete data that is perfectly measured in political science is impossible at the country-year unit of analysis level. Especailly with economic data, many countries are unwilling to share.

Second, and more damning to NHST, it is not practical or possible to perfectly account for the inter-dependence of observations on at least three fronts: spatial, temporal, and relational. One can technically model temporal correlation, spatial correlation, and relational dynamics (via network models), but doing all three together is certainly not practical. Not only would the model be incredibly complex and difficult to interpret, but identifying that model correctly is very difficult. From a temporal standpoint, one must identify the lag structure of the dependent variable and the lag structure of the independent varaibles necessary to get white noise residuals. From a spatial standpoint, one must account for spatial correlation in the residuals according to a similar set of potential lag structures but in a spatial plane rather than temporal one.

Most importantly, without correctly accounting for problems with the residuals that include spatial and temporal correlation the standard errors of the ceofficients are incorrect and the t or z-statistics used to calculate p-values are then incorrect. In this setting, even if evaluating hypotheses based on p-values was a great idea it is very difficult to be sure that the p-values are correct in settings of complex data structures. This is just one of the many reasons that a common critique of p-values is that they are fickle to model specification (Basuchoudhary and Bang 2018; Colaresi and Mahmood 2017). Transition toward predictive models in machine learning and focusing on adding to predictive power of models avoids these problems because they are not reliant on strict assumptions because they do not have concern with NHST.

Limitations what p-values imply

- in large samples, it is likely that a model will show a variable as statistically significant even if the author's specified relationship is false (Levine et al. 2008; Meehl 1986).

Dangers of fitting models on full samples

Dangers of Goodness of Fit tests

Goodness of fit tests only consider how well a model fits the data, not how well it predicts the outcome (Breiman 2001).

- Breiman (2001) argues that there are two schools for analysis: data modeling and algorithmic modeling. Data modeling is political science on average: fitting models on full samples looking to draw inference from coefficients and assess the models based on goodness of fit without considering cross-validation or potential problems of overfitting. In this school, a model with a significant goodness of fit statistic is considered a good explainer. Alternatively, algorithmic modeling considers much more advanced and complicated to interpret models that predict outcomes, and thus represent nature's process, better.

In conclusion, the faults listed above of NHST cast significant doubt on its ability to meet the two fundamental goals of political science research: explanation of events, and assessment of theories. Strictly speaking, NHST does provide measures of explanation of variance *within the fit model*, but provides no such tool to test explanation out-of-sample. An overfit model can explain artificially high levels of variance because it tries to fit itself to randomnesss in the DV, which hampers generalizability and leads to incorrect inferences based on coefficients (Hill et al. 2013; Colaresi and Mahmood 2017). Aside from the problem of model fit and predictive performance, NHST provides no direct way to asses whether or not a model's variables actually add value to predicting the outcome of interest (Ward et al. 2010; Gelman and Loken 2014). Coefficients offer insight into the correlational relationship between independent and dependent variables, and give some indication of the effect size within sample. Within sample effect size does not translate to out-of-sample prediction, and a model that only fits its training data well while performing poorly out of sample is not effective at identifying good theoretical arguments. NHST provides no methodology for assessing predictive power of the model, and thus provide no mechanism for validating a theory's usefulness.

Given the conclusion that NHST fails to meet the needs of political scientists, I propose that whenever possible a new machine-learning based framework for out-of-sample prediction (FOOSP) supplants NHST. This framework meets the needs of political scientists for assessing theoretical validity and is actually simpler in understanding and implimentation than NHST because of it's weaker focus on standard errors and p-values and stronger focus on predictive accuracy. Ultimately, theories that predict poorly, or add no value to existing theoretically derived models, are bad theories. FOOSP provides a framework to more effectively identify predictive versus nonpredictive theories allowing the field in general to move forward in a more Lakatosian direction, and away from degenerative theory testing under NHST.

Next, I discuss examples of work already moving in this direction and utilizing machine learning of all kinds on political science and other social science problems. Then, I explain my framework for analysis moving forward.

Machine Learning in Political Science: examples of FOOSP in practice

Data collection/coding

- Several datasets that are commonly used in political science today use webscraping and machine learning a la natural language processing, specifically topic modeling and other classification algorithms (Pham and Hoang 2018; Nardulli et al. 2015; Boschee et al. 2019)

- Greene et al. (2018) use a variety of supervised machine learning techniques to assess the Fariss (2014) claim that as human rights standards have increased over time, the human-coded scales of political repression have become less reliable because the reports themselves have changed. They find that classifiers trained on older State Department Human Rights Reports are worse predictors of more modern classifications, indicating that there are indeed dynamic reporting processes of human rights violations that need to be accounted for.

- Minhas et al. (2015) use each stemmed and word (after removing stopwords) as features in an SVM to classify regime type according to the text the the state department Country Reports on Human Rights Practices and Freedom House's Freedom in the World reports. They achieve very high recall and accuracy according to the Geddes et al. (2014) and Hadenius and Teorell (2007).

- Saiya and Scime (2019) argue that decision-trees are dramatically under-utilized in predicting violence. They argue using the START GTD terrorism data that religious and secular terrorism are predicted using different factors, and are thus fundamentally different processes to predict. They advocate the use of gain-ratio analysis and proxy tree analysis.

- Al-Zewairi and Naymat (2017) uses four supervised machine learning approaches (distributed random forests, deep learning, naive bayes, and gradient boosting) to identify Islamic extremists from the Profiles of Individual Radicalization in the US (START)).

- Burscher et al. (2015) use the Passive Aggresive learning algorithm to create a classifier that correctly labels the policy type of a given news article. This is an example of using machine learning to assist coding, which is one of the most common uses in political science.

- Croicu and Weidmann (2015) use an ensemble method (built of SVM and NB, excluding random forests because of computational expense) to predict which articles are useful for identifying news articles that are relevant to protests to dramatically decrease the human-coding time needed to use news sources.

Evaluation of existing theory

- Muchlinski et al. (2016) shows that random forests out-predict other forms of common models used in political science to predict rare events (logit, rare-events logit, L1-logit). Given a primary goal of theorizing about an outcome is predicting an outcome, it would be best to use the most effectively predicting models.

- Basuchoudhary and Bang (2018) tests the predictive accuracy of what the NHST literature to date has found. Premised on the argument that parametric NHST suffers from four main problems: not interested in predictive accuracy, p-values are sensitive to model specification, cannot effectively compare variables' explanatory power, nonstandardized robustness checks, does not distinguish correlation from causation.

Out of Sample Prediction

- Gleditsch and Ward (2013) use logit models trained on data up to 1990 to develop a model to predict conflict outcome after 1990 using the Issue Correlates of War (ICOW) data.

- Blair et al. (2014) uses survey based local violence data across 240+ towns in Liberia to predict local violence in three different years. This is a prime example of high spatial variation but low temporal variation.

- **?** use poisson point process models to analyze the GED data (Sundberg and Melander 2013) at local levels. They use dirichilet tesselation to avoid information loss from data aggregation resulting from ad hoc gridding of data. They use leave-one-out cross-validation, leaving out one country to use for out-of-sample prediction. Their model fits reasonably well, but they find that more parsimonious (3 variable) models predict best.

- Witmer et al. (2017) have perhaps the most ambitious example of out-of-sample forecasting attempts. They used the ACLED data in 1-degree grid cells and simulated future climate data to predict subnational violence all the way into 2065. Obviously, this sort of work is impossible to validate, but it is quite a good example of the type of work that political science should strive for. While there is a strong argument to be made that this paper's forecasting model is relatively weak (most variables are

country-level besides weather), the principal behind advocating predictive modeling should be more central to political science. The theories and research of political science is useful only inasmuch as it can actually identify when and where events will happen.

- Hill et al. (2013) use random forests to predict culpability of of terrorism in the Philippines (where previously multinomial logit and naive bayes were used). They use cross-validation via out-of-bag prediction to guard against overfitting of the model and find that Random forests predict as well or better than the others as long as there is no class imbalance problem.

- Ding et al. (2017) use a variety of geo-coded data to predict terrorist events around the world in 2015. They achieve about a 96% accuracy rate using .1x.1 (latitude x longitude) grids around the world. They used neural networks, support vector machines, and random forests.

- Mueller and Rauh (2018) use Latent Dirichlet Allocation (LDA) models to asses topics in newspaper articles to predict the onset of civil war. They use LASSO machine learning techniques to predict the onset of civil conflict using the results of unsupervised machine learning. This is a very common sense finding from very complicated methods that made it into the APSR

- Brandt et al. (2014) suggest that instead of RMSE and MAE probabilistic predictive models should be evaluated not by point metrics such as RMSE and MAE, but but a variety of tools common in meteorology that are optimized to assess predictive accuracy or probabilistic models. Many of these methods (ranked probability score for ordered categories of a discrete random variable, continuous rank probability scores for continuous options, verification rank histograms, probability integral transform, etc.). Each of these evaluation strategies focuses on different ways of considering how the predicted probability of an event matches up with the true outcome, and generally ranking by which predictions are best.

Framework for out-of-sample prediction

1. Theorize and hypothesize about some phenomena.

2. Identify some a priori base model to which predictive accuracy will be compared.
3. Identify and collect relevant data. Split this data into a training set and a test set.
4. Fit a model(s) best suited for this data to a training set first without the new variables identifiedin step one, then with. Where practical and applicable, apply cross-validation strategies within the training set.

5. Use the resulting models from step four to predict the values of the test set. Evaluate the accuracy of the test set using appropriate predictive accuracy assessment tools (e.g. point metrics (root mean squared error, mean absolute error) or forecast density evaluation).
6. Compare the base model's predictive accuracy with the new model's predictive accuracy. Based on the change in predictive accuracy, infer about the validity of the proposed theory.

Step one in this process is the same as NHST: identify some explanation of a phenomena that is interesting to political science that can be evaluated empirically. Step two is where NHST and FOOSP diverge: under FOOSP it is critical that a base model is identified. Ideally, this base model is the best predicting model the literature has to offer to date. In practice, this model is very difficult to identify given the variance in methods, samples, and concept operationalization. Thus, evidence in this context should only be interpreted inasmuch as the sample is generalizable. Ideally, one can identify the broadest sample, and the most general model that the literature has some moderate level of agreement on. Over time, this framework will be able to parse down and remove variables that add no predictive value. Furthermore, methods such as penalized regression models (e.g. ridge regression, and LASSO) or random forests can be used to evaluate variable importance for prediction. Variables that are not important for prediction, should not be include in favor of parsimony and avoiding overfitting.

Step three is also similar in most ways to NHST, with the caveat of splitting the sample. A good rule of thumb is to keep anywhere from 33% (Colaresi and Mahmood 2017) to 10% of the data as a test set and use the rest to train the model. Within the training set, cross-validate where possible and practical. Cross-validation is the process in which a training set is split into k parts, and a model is fit on k-1 parts, while the remaining part of the training set is used to evaluate; this process iteratures within the training set until each remaining k has been used as an evaluation set (Colaresi and Mahmood 2017; Stone 1974; Breiman 2001; James et al. 2017, p 176). This process iteratively improves models by identifying flaws on each iteration, and generating a composite model after each iteraturion. This is generally referred to as k-fold cross-validation, and it is difficult to impliment in contexts where spatio-temporal dynamics are part of the modeling process because the way you split the data fundamentally affects the modeling process.

Other methods of cross-validating data are bootstrapping and bagging. Bootstrapping samples with replacement from the existing training set and estimates a model on this new bootstrapped training set. Bagging is best for high variance data, and its process involves bootstrapping several samples from the training data and averaging the resultant models' predictions to account for high variation (James et al. 2017, pp. 316-317). This is a common approach in decision-trees.

Step four is simple enough: just fit the two models: base and updated model. If there are multiple

hypotheses to be tested, perhaps multiple updated models should be estimated, carefully considering when to add more than one variable at a time.

Step five is where FOOSP shines as a staunch improvement over NHST. Simply take the cross-validated models and test their predictive accuracy on the left out test set. There are a variety of methods this can take, the simplest of which was described above: leaving aside about 1/3rd of the data for a test set. Other methods include leave-one-out cross-validation (LOOCV); during step three fit the model to the entire sample minus one, and test prediction on the left-out observation, repeat this process for all observations and average the mean-squared error across all iterations (James et al. 2017). This is obviously computationally intensive, but provides extensive cross-validation to learn and improve models from. In the cases of bootstrapping and bagging, it is normal to test on the training set, or 'out-of-bag.'

Step six simply involves evaluating the information retreived from the previous five steps. In an ideal case one finds that adding the variables measured for the theory and hypothesis have substantively significantly improved the model's predictive accuracy. In this setting statistical significance is *never* part of the discussion because standard errors and residuals in the GoF sense are completely irrelevant to the central concerns of FOOSP. Below, I elaborate on specific methods that can and should be used such as ridge regression, LASSO, k-nearest neighbors (KNN), Naive Bayes, Random Forests, and Support Vector Machines. Each of these models, as well as standard OLS and other generalized linear models (e.g. logistic regression) and much more also fit nicely into this general framework. Also below, I elaborate further on exactly why freedom from the tyranny of GoF, standard errors, and p-values makes sience easier.

The tyranny of residuals in NHST and FOOSP's liberation

Simply put, the foundation of NHST as discussed above is that models meet their strict assumptions. Under these met strict assumptions standard errors are correct and p-values can be calculated and null hypotheses can be rejected, but alternative hypotheses cannot be confirmed. NHST also does not allow the examination of predictive accuracy out-of-sample. NHST provides almost no value for scientific advancement of theories, despite science's best effort for decades to extract value.

FOOSP, alternatively, does not have any concern, or even discussion of standard errors, or p-values, and thus does not have to meet those strict assumptions. Predictive models either predict well out of sample, or they do not. Theories, hypotheses, and variables either add predicted value, or they do not. Whereas under NHST not properly modeling temporal or spatial dynamics can result in a virtually uninterpretable model from an inferential perspective, predictive-based modeling is largely unaffected. If space and time are important, and unmodelled, then the model predicts poorly, but it still predicts. Moreover, the predictions that it makes, while predicting poorly, are still interpretable even if they model performs poorly.

Thus, whereas under NHST researchers can fail to model important correlation in the residuals and fallaciously interpret the results when they contain little truth, FOOSP directly punishes for poorly modeling real dynamics by predicting poorly. In NHST, this factor is never even considered, and many journals and referees do not ask for spatio-temporal models on average.

Explanation of models

Penalized Regression

K-nearest neighbors

Naive Bayes

Random Forests

Support Vector Machines

Malek Al-Zewairi and Ghazi Naymat. Spotting the Islamist Radical within: Religious Extremists Profiling in the United State. *Procedia Computer Science*, 113:162–169, January 2017. ISSN 1877-0509. doi: 10.1016/j.procs.2017.08.336. URL https://www.sciencedirect.com/science/article/pii/S1877050917317465.

Atin Basuchoudhary and James T. Bang. Predicting Terrorism with Machine Learning: Lessons from "Predicting Terrorism: A Machine Learning Approach". *Peace Economics, Peace Science and Public Policy*, 24(4), 2018. ISSN 1554-8597. doi: 10.1515/peps-2018-0040. URL http://www.degruyter.com/view/j/peps.2018.24.issue-4/peps-2018-0040/peps-2018-0040.xml.

Claude Berrebi and Esteban F. Klor. On Terrorism and Electoral Outcomes: Theory and Evidence from the Israeli-Palestinian Conflict. *The Journal of Conflict Resolution*, 50(6):899–925, 2006. ISSN 0022-0027. URL http://www.jstor.org/stable/27638530.

Peter J. Bickel, Ya'acov Ritov, and Thomas M. Stoker. Tailor-Made Tests for Goodness of Fit to Semiparametric Hypotheses. *The Annals of Statistics*, 34(2):721–741, 2006. ISSN 0090-5364. URL http://www.jstor.org/stable/25463434.

Robert A. Blair, Christopher Blattman, and Alexandra Hartman. Predicting Local Violence. *SSRN Electronic Journal*, 2014. ISSN 1556-5068. doi: 10.2139/ssrn.2497153. URL http://www.ssrn.com/abstract=2497153.

Philipp C. Bleek and Eric B. Lorber. Security Guarantees and Allied Nuclear Proliferation. *Journal of Conflict Resolution*, 58(3):429–454, 2014. ISSN 0022-0027. doi: 10.1177/0022002713509050. URL https://doi.org/10.1177/0022002713509050.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, and James Starz. ICEWS Automated Daily Event Data. June 2019. doi: 10.7910/DVN/QI2T9A. URL https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QI2T9A. type: dataset.

Patrick T. Brandt, John R. Freeman, and Philip A. Schrodt. Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting*, 30(4):944–962, October 2014. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2014.03.014. URL http://www.sciencedirect.com/science/article/pii/S0169207014000612.

Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–215, 2001. ISSN 0883-4237. URL http://www.jstor.org/stable/2676681.

Bjorn Burscher, Rens Vliegenthart, and Claes H. De Vreese. Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? *The ANNALS of the American Academy of Political and Social Science*, 659(1):122–131, May 2015. ISSN 0002-7162. doi: 10.1177/0002716215569441. URL https://doi.org/10.1177/0002716215569441.

Michael Colaresi and Zuhaib Mahmood. Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, 54(2):193–214, March 2017. ISSN 0022-3433. doi: 10.1177/0022343316682065. URL https://doi.org/10.1177/0022343316682065.

Mihai Croicu and Nils B Weidmann. Improving the selection of news reports for event coding using ensemble classification. *Research & Politics*, 2(4):2053168015615596, October 2015. ISSN 2053-1680. doi: 10.1177/2053168015615596. URL https://doi.org/10.1177/2053168015615596.

Ursula Daxecker and Brandon Prins. Insurgents of the Sea: Institutional and Economic Opportunities for Maritime Piracy. *Journal of Conflict Resolution*, 57(6):940–965, December 2013. ISSN 0022-0027. doi: 10.1177/0022002712453709. URL https://doi.org/10.1177/0022002712453709.

Fangyu Ding, Quansheng Ge, Dong Jiang, Jingying Fu, and Mengmeng Hao. Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach. *PLOS ONE*, 12(6):e0179057, June 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0179057. URL http://dx.plos.org/10.1371/journal.pone.0179057.

Christopher Dougherty. *Introduction to Econometrics*. Oxford University Press, Oxford ; New York, 4 edition edition, April 2011. ISBN 978-0-19-956708-9.

Christopher J. Fariss. Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability. *American Political Science Review*, 108(2):297–318, May 2014. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055414000070. URL http://www.cambridge.org/core/journals/american-political-science-review/article/respect-for-human-rights-has-improved-over-time-modeling-the-changing-standard-of-accountability/0E7B51BE2CDA4A141779E594FF0F0EFF.

James D. Fearon and David D. Laitin. Ethnicity, Insurgency, and Civil War. *American Political Science Review*, 97(1):75–90, February 2003. ISSN 1537-5943, 0003-0554. doi: 10.1017/S0003055403000534. URL http://www.cambridge.org/core/journals/american-political-science-review/article/ethnicity-insurgency-and-civil-war/B1D5D0E7C782483C5D7E102A61AD6605.

Sean Gailmard. *Statistical Modeling and Inference for Social Science*. Cambridge University Press, New York, 2014.

Barbara Geddes, Joseph Wright, and Erica Frantz. Autocratic Breakdown and Regime Transitions: A New Data Set. *Perspectives on Politics*, 12(2):313–331, June 2014. ISSN 1537-5927, 1541-0986. doi: 10.1017/S1537592714000851. URL https://www.cambridge.org/core/journals/perspectives-on-politics/article/autocratic-breakdown-and-regime-transitions-a-new-data-set/EBDB9E5E64CF899AD50B9ACC630B593F.

Andrew Gelman and Eric Loken. The Statistical Crisis in Science. *American Scientist*, 102(6):460–465, 2014. ISSN 0003-0996. URL http://search.proquest.com/docview/1627946025?rfr_id=info%3Axri%2Fsid%3Aprimo.

Kristian Skrede Gleditsch and Michael D Ward. Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes. *Journal of Peace Research*, 50(1):17–31, January 2013. ISSN 0022-3433, 1460-3578. doi: 10.1177/0022343312449033. URL http://journals.sagepub.com/doi/10.1177/0022343312449033.

Kevin T. Greene, Baekkwan Park, and Michael Colaresi. Machine Learning Human Rights and Wrongs: How the Successes and Failures of Supervised Learning Algorithms Can Inform the Debate About Information Effects. *Political Analysis*, pages 1–8, 2018. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2018.11. URL http://www.cambridge.org/core/journals/political-analysis/article/machine-learning-human-rights-and-wrongs-how-the-successes-and-failures-of-supervised-learning-algorithms-can-inform-the-debate-about-information-effects/C4CB3FC4383A50129F6A0F0809BA2452.

Axel Hadenius and Jan Teorell. Pathways from Authoritarianism. *Journal of Democracy*, 18(1):143–157, January 2007. ISSN 1086-3214. doi: 10.1353/jod.2007.0009. URL http://muse.jhu.edu/article/209112.

Joshua B Hill, Daniel J Mabrey, and John M Miller. Modeling terrorism culpability: An event-based approach. *The Journal of Defense Modeling and Simulation*, 10(2):181–191, April 2013. ISSN 1548-5129. doi: 10.1177/1548512912455470. URL https://doi.org/10.1177/1548512912455470.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, 1st ed. 2013, corr. 7th printing 2017 edition edition, September 2017. ISBN 978-1-4614-7137-0.

James M. Landwehr, Daryl Pregibon, and Anne C. Shoemaker. Graphical Methods for Assessing Logistic Regression Models. *Journal of the American Statistical Association*, 79(385):61–71, March 1984. ISSN 0162-1459. doi: 10.1080/01621459.1984.10477062. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10477062.

Timothy R. Levine, René Weber, Craig Hullett, Hee Sun Park, and Lisa L. Massi Lindsey. A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. *Human Communication Research*, 34(2):171–187, April 2008. ISSN 0360-3989, 1468-2958. doi: 10.1111/j.1468-2958.2008.00317.x. URL https://academic.oup.com/hcr/article/34/2/171-187/4210725.

P E Meehl. What Social Scientists Don't Understand. In *Metatheory in social science: Pluralisms and subjectivities*, page 24. University of Chicago Press, Chicago, 1986.

Paul E. Meehl. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4):806, 1979. ISSN 1939-2117. doi: 10.1037/0022-006X.46.4.806. URL http://psycnet.apa.org/fulltext/1979-25042-001.pdf.

Shahryar Minhas, Jay Ulfelder, and Michael D Ward. Mining texts to efficiently generate global data on political regime types. *Research & Politics*, 2(3):2053168015589217, July 2015. ISSN 2053-1680. doi: 10.1177/2053168015589217. URL https://doi.org/10.1177/2053168015589217.

David Muchlinski, David Siroky, Jingrui He, and Matthew Kocher. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, 24(1):87–103, 2016. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpv024. URL http://www.cambridge.org/core/journals/political-analysis/article/comparing-random-forest-with-logistic-regression-for-predicting-classimbalanced-civil-war-onset-data/109E1511378A38BB4B41F721E6017FB1.

Hannes Mueller and Christopher Rauh. Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review*, 112(2):358–375, May 2018. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055417000570. URL http://www.cambridge.org/core/journals/american-political-science-review/article/reading-between-the-lines-prediction-of-political-violence-using-newspaper-text/4EABB473AFE18F157EEDE4339F34ABB0.

Peter F. Nardulli, Scott L. Althaus, and Matthew Hayes. A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data:. *Sociological Methodology*, 2015. doi: 10.1177/0081175015581378. URL https://journals.sagepub.com/doi/pdf/10.1177/0081175015581378.

T. S. Pham and T. Hoang. Modeling The Causes Of Terrorism From Media News: An Innovative Framework Connecting Impactful Events With Terror Incidents. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 364–369, November 2018. doi: 10.1109/KSE.2018.8573353.

James A. Piazza. Types of Minority Discrimination and Terrorism. *Conflict Management and Peace Science*, 29(5):521–546, 2012.

Nilay Saiya and Anthony Scime. Comparing Classification Trees to Discern Patterns of Terrorism*. *Social Science Quarterly*, 100(4):1420–1444, 2019. ISSN 1540-6237. doi: 10.1111/ssqu.12629. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/ssqu.12629.

National Consortium for the Study of Terrorism and Responses to Terrorism (START). Profiles of Individual Radicalization in the United States, 2018. URL http://www.start.umd.edu/pirus.

M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. ISSN 0035-9246. URL http://www.jstor.org/stable/2984809.

Ralph Sundberg and Erik Melander. Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research*, 50(4):523–532, July 2013. ISSN 0022-3433, 1460-3578. doi: 10.1177/0022343313484347. URL http://journals.sagepub.com/doi/10.1177/0022343313484347.

Michael D Ward, Brian D Greenhill, and Kristin M Bakke. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375, July 2010. ISSN 0022-3433. doi: 10.1177/0022343309356491. URL https://doi.org/10.1177/0022343309356491.

Ronald L. Wasserstein and Nicole A. Lazar. The ASA Statement on p-values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133, April 2016. ISSN 0003-1305, 1537-2731. doi: 10.1080/00031305.2016.1154108. URL https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108.

Frank DW Witmer, Andrew M Linke, John O'Loughlin, Andrew Gettelman, and Arlene Laing. Subnational violent conflict forecasts for sub-Saharan Africa, 2015–65, using climate-sensitive models. *Journal of Peace Research*, 54(2):175–192, March 2017. ISSN 0022-3433. doi: 10.1177/0022343316682064. URL https://doi.org/10.1177/0022343316682064.