Interpretable Machine Learning Methods

Zachary M. Jones

eScience Institute, University of Washington

Conventional Methods in (Observational) Political Science

- fixed model complexity (simple model classes)
- direct interpretation of model parameters
- goodness-of-fit based model selection

Problems with Conventional Methods

- assumed models are not based on strong belief or evidence
- what we learn from a fit model depends **heavily** on what we assumed is true
- goodness-of-fit suffers by restricting model adaptation

Interpretable Machine Learning Methods Solve this Problem

- large possible model classes (adaptive complexity)
- decomposition methods allow interpretation which depends much less on the (weaker) assumptions made
- out-of-sample goodness-of-fit is the focus

History

Beck, King, and Zeng (2000) already said this (but nobody listened)! Why?

- too method-specific
- no software
- didn't use social/statistical theory to make their point
- de Marchi, Gelpi, and Grynaviski... >:(
- they didn't keep doing this with their substantive work

Mine Contribution(s)

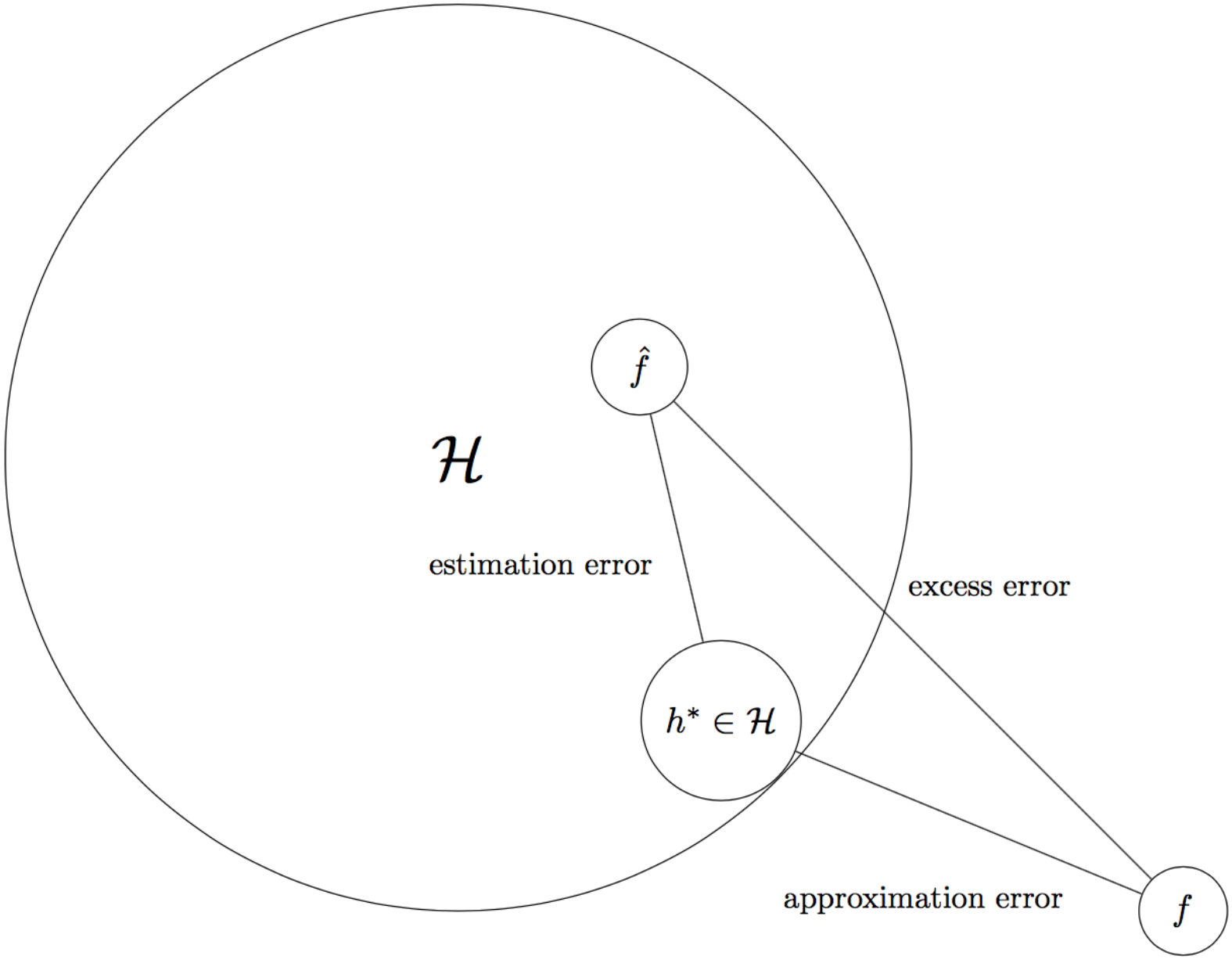
- general purpose interpretation software (`mlr`, `mmpf`, `edarf`, `fanova`) which works with *any* method
- more focus on theoretical background
- applications to cite so you don't have to make this argument (how convenient for me!)

The Machine Learning Demarcation Problem

an arbitrary but useful distinction! based on the size of the space of possible models (hypothesis space)

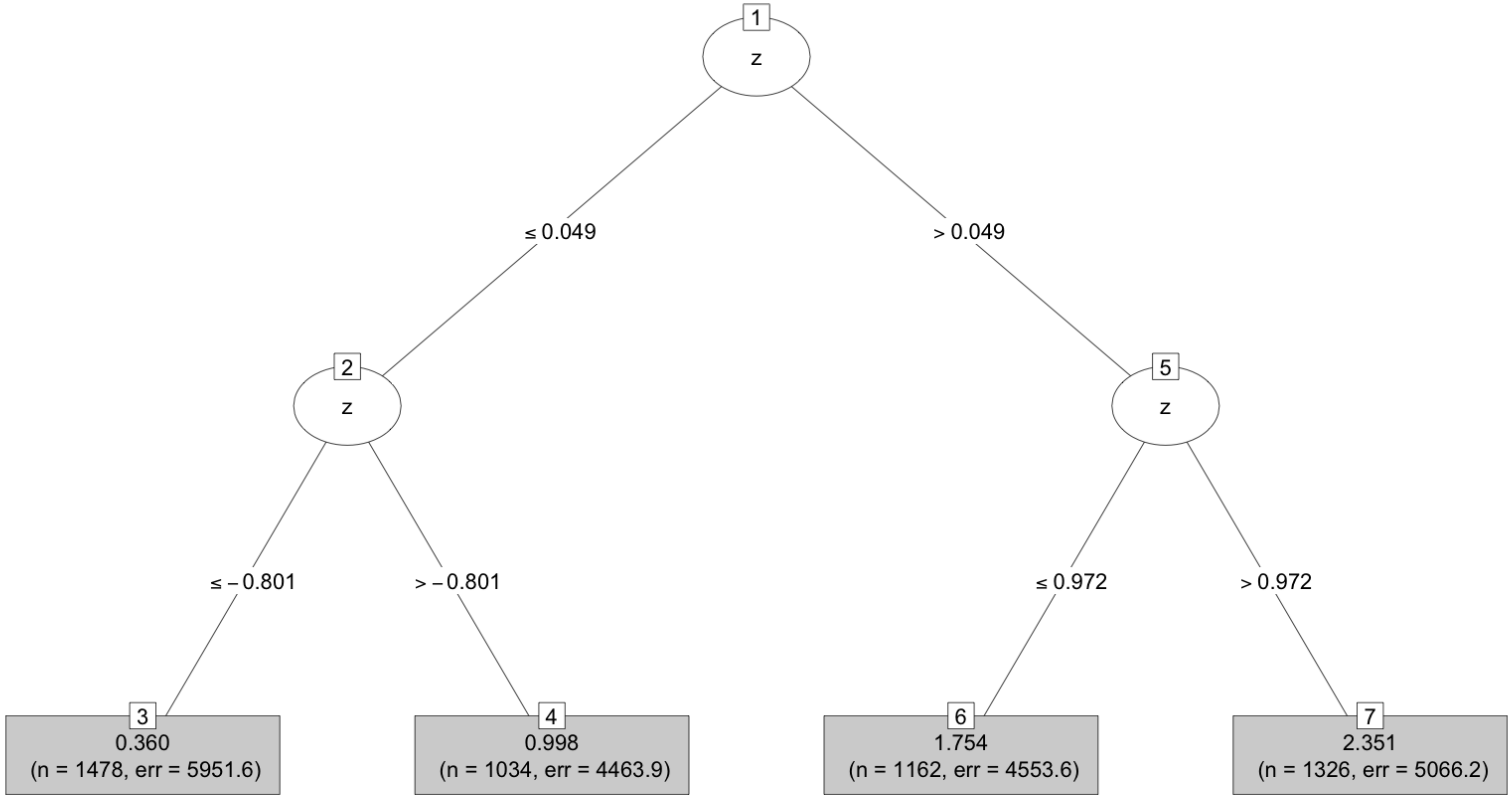
- a **conventional** method has a fixed-dimension hypothesis space
- a **machine learning** method has a hypothesis space which whose dimension is data-adaptive

(Some) Sources of Model Error



What Makes Models Uninterpretable?

assumptions $\longrightarrow$ estimation



What Do We Want to Learn From Our Models?

1. is this feature important
   - on its own? in combination with other features?
   - what does important mean?
2. what is shape of the relationship between this (these) feature(s) and the outcome(s)?
3. how reliable is my model's representation of these things?

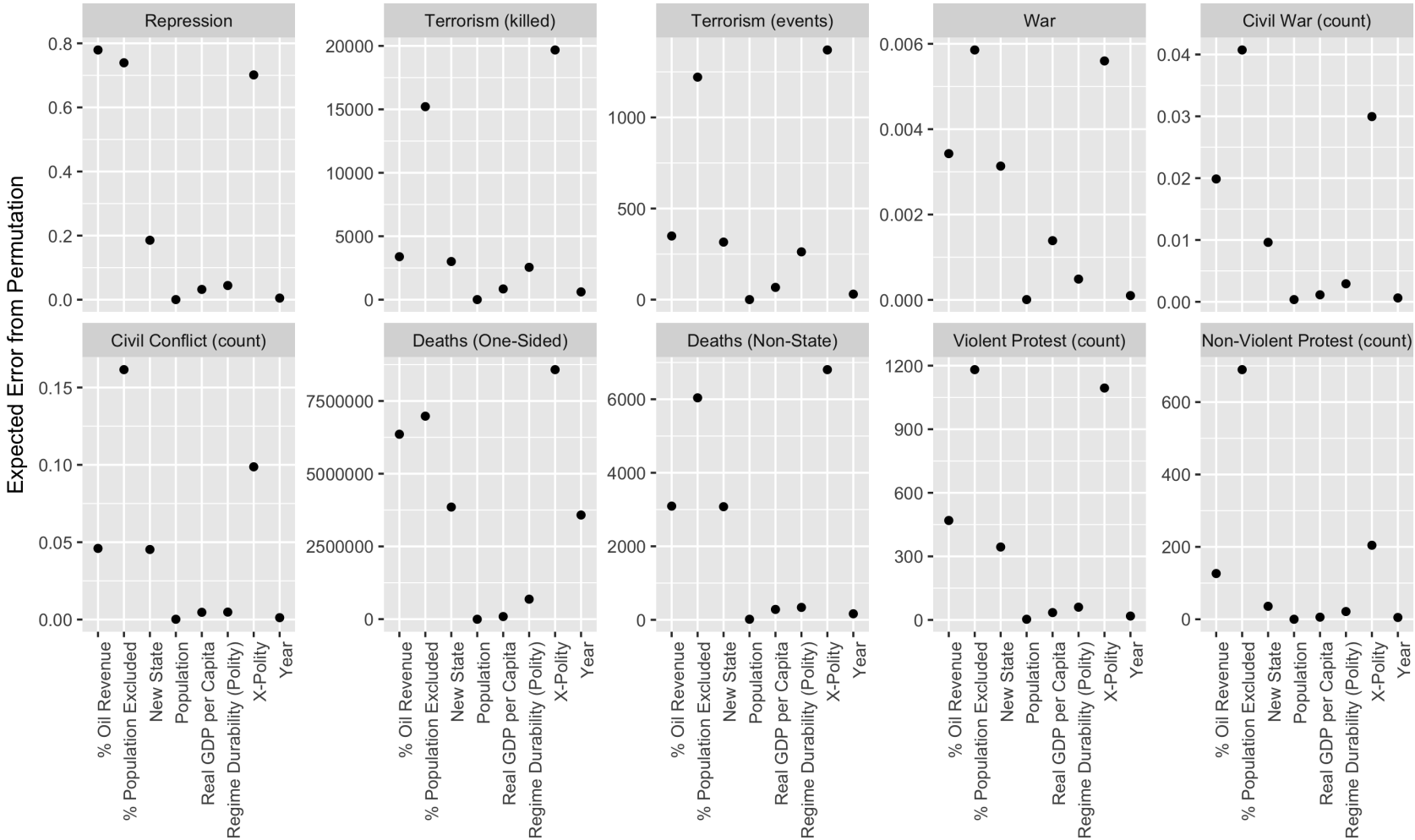
Permutation Importance

randomly shuffling covariates makes them useless to the model (if they are important)

what is the average change in prediction error from doing this?

$$I_{\mathbf{x}_u} = \frac{1}{M} \sum_{j=1}^{M} \mathcal{L}\left(\hat{f}(\mathbf{x}_{u\pi(j)}, \mathbf{x}_{-u}), \hat{f}(\mathbf{x})\right)$$

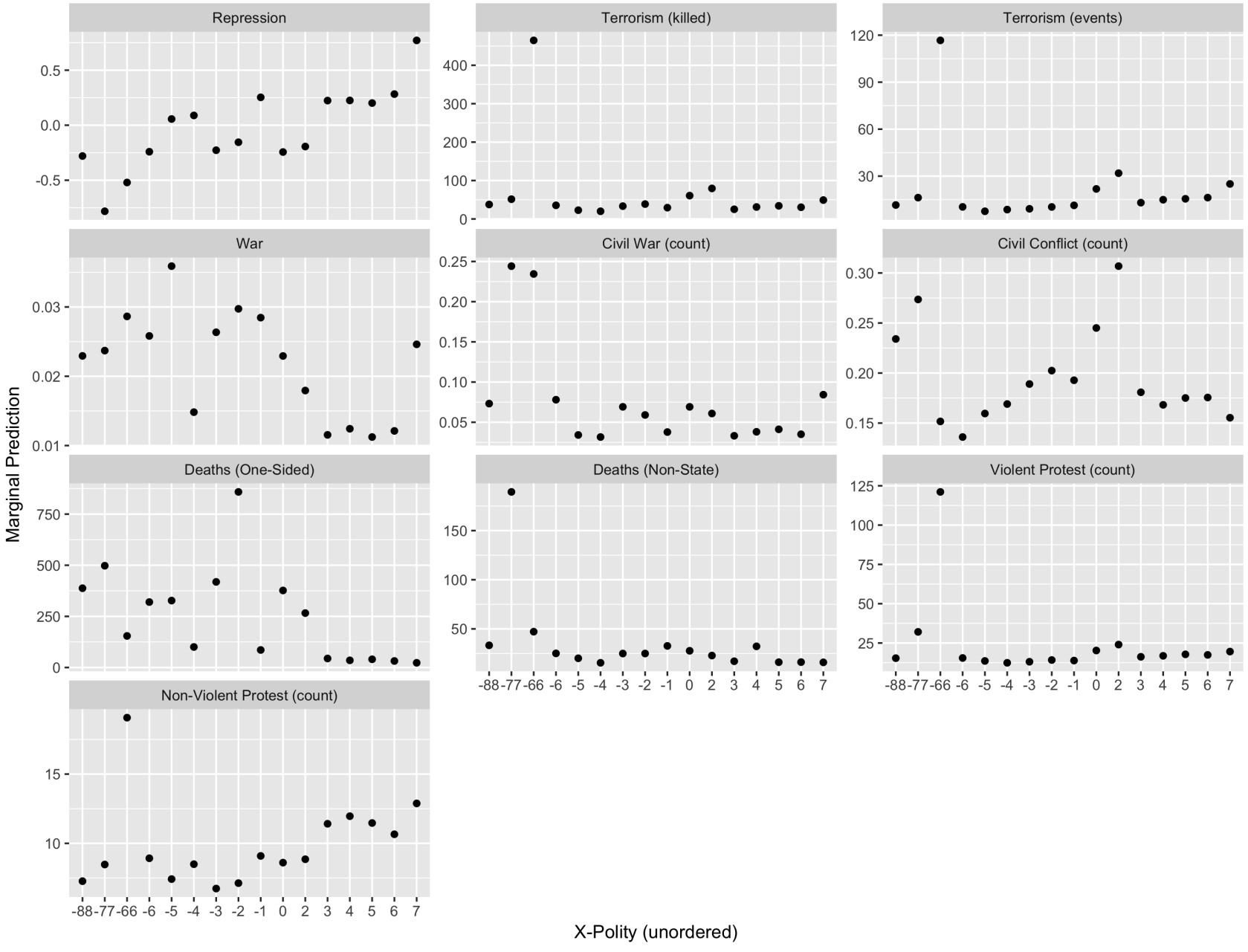
Feature Importance



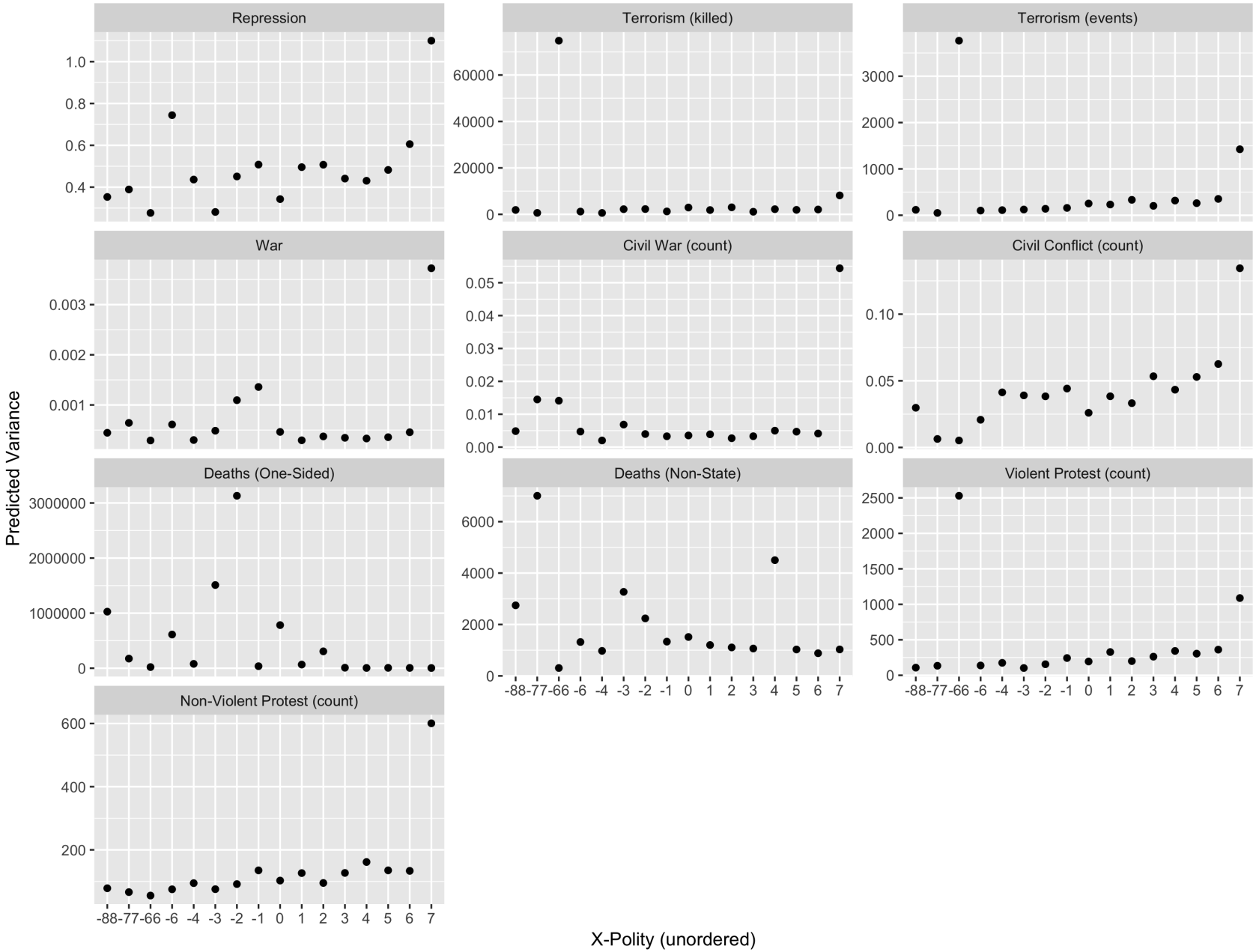
Partial Dependence

1. make a grid of values on the feature(s) you are interested in
2. take the cartesian product of that grid with the training data (or a subsample thereof)
3. evaluate the model on this design matrix
4. collapse the predictions over the training data (e.g., the sample mean or variance)

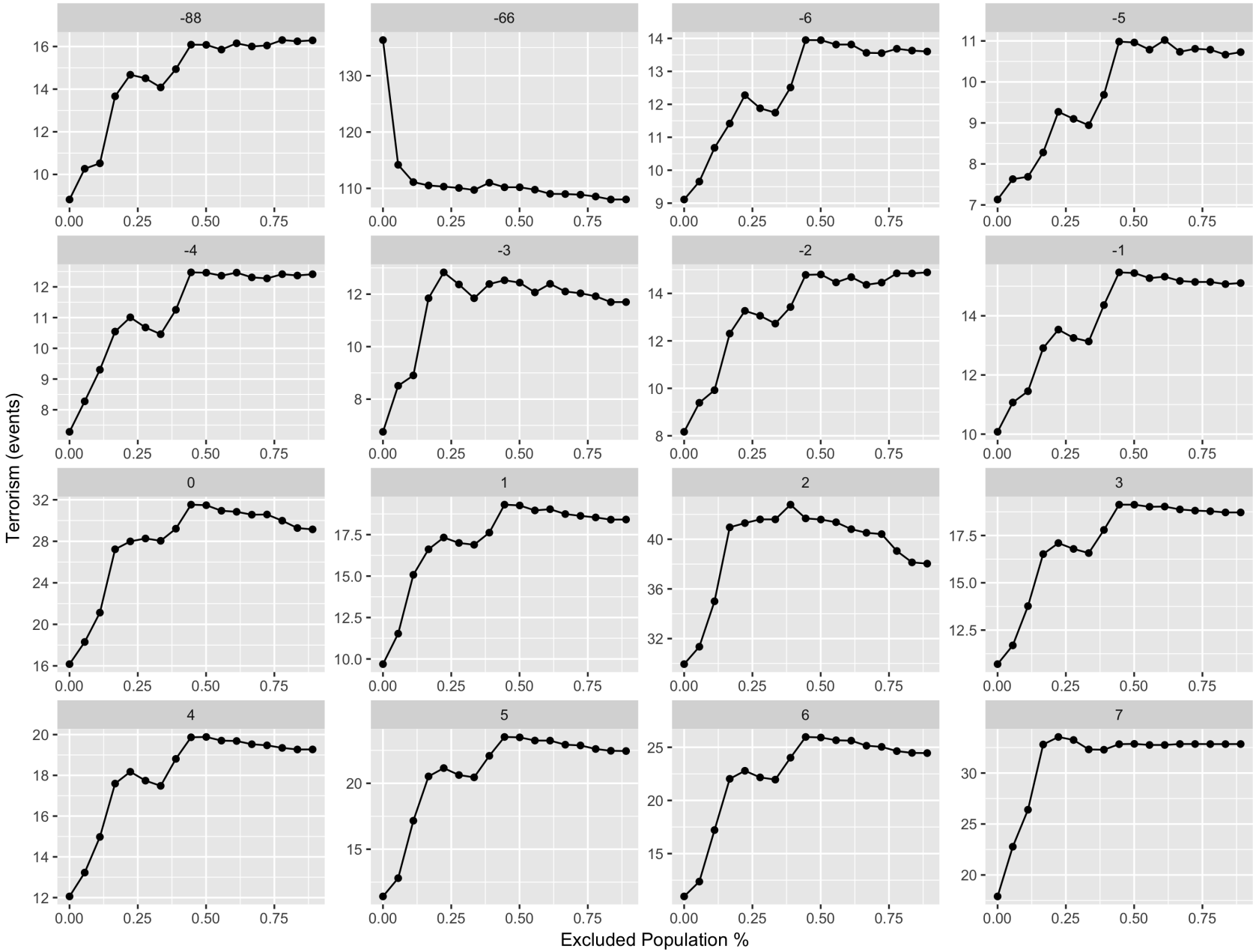
Functional Form



Interaction Detection



Interaction Detection II



Conclusions

- interpretable ml can help us find data/theory problems
  - generally more appropriate than conventional methods for our problems
  - it should be the default

Future Directions

- honest reliability assessment
- software improvement/extension

mmpf: Monte-Carlo Methods for Prediction Functions

- model-agnostic Monte-Carlo integration for prediction functions
- very configurable
- gihtub.com/zmjones/mmpf

edarf: Exploratory Data Analysis using Random Forests

- works with major random forest packages in R (randomForest, randomForestSRC, party, ranger, partykit)
- smart wrapper for mmpf with convenience functions for plotting

mlr: Machine Learning in R

- end-to-end solution for fitting, tuning, evaluating and interpreting machine learning methods
- github.com/mlr-org/mlr
- mlr-org.github.io/mlr/

fanova: Functional Analysis of Variance

- full decomposition of regression functions in terms of additive components
- currently doesn't scale very well

Why Not Directly Estimate the Interpretable Model?

- do predictions have to be made by an interpretable model?
- what sort of interpretations are required? does all of the model need to be interpretable?
- exploratory data anlysis and/or model evaluation (contrast prediction w/ explanatory model)

Interpretation Methods

to decompose $\hat{f}$ or $\mathcal{L}(\hat{f})$

methods fall into two classes

- point-wise (grid)
- surrogate models

usually additive decompositions

$$\hat{f}(\mathbf{x}) \approx g_u(x_u) + \sum_{v \subset u} g_v(x_v) + g_{-u}(x_{-u}) + \sum_{i \in u \subset v' \subseteq -i} g_{v'}(x_{v'})$$

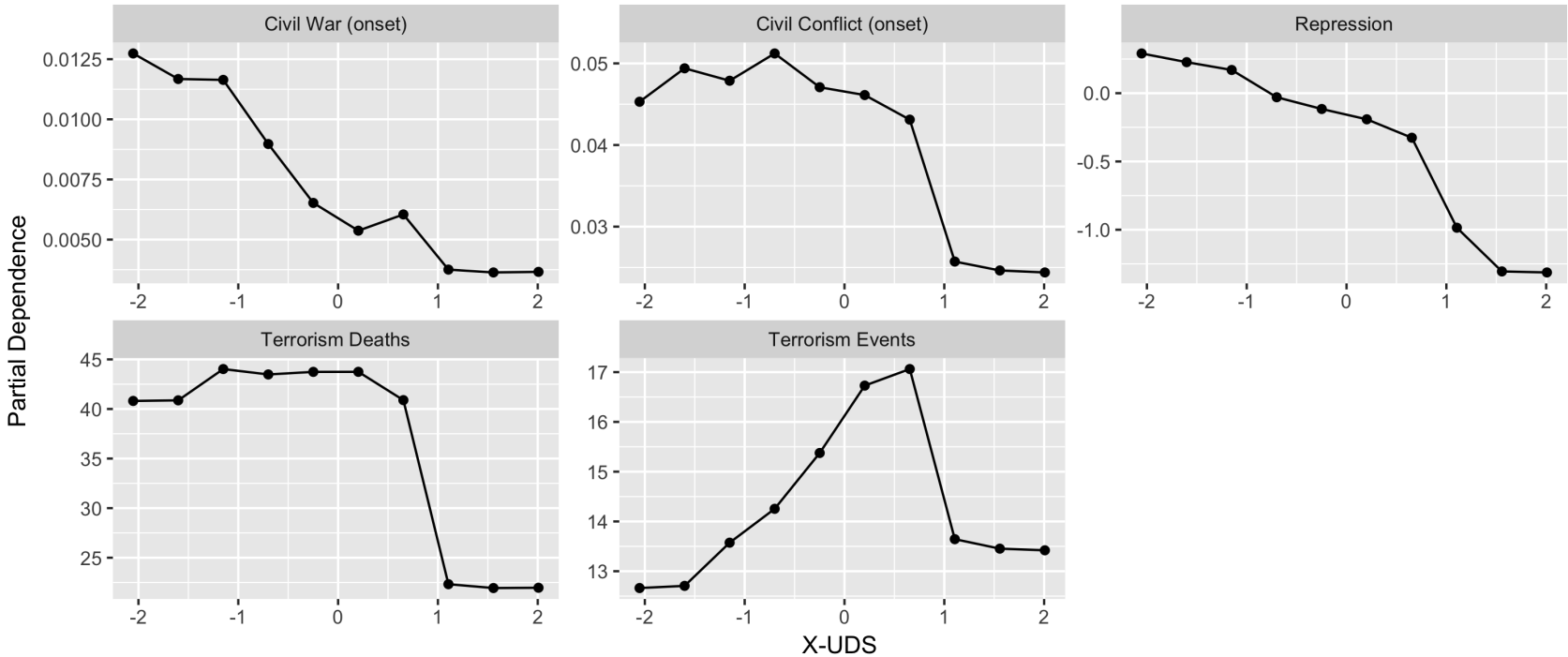
Grid Methods

1. create a design matrix
2. evaluate model on said matrix
3. aggregate

Surrogate Models

1. pick an interpretable model class
2. fit said model to x and $\hat{y}$
3. look at it

Lupu and Jones 2018



Hill and Jones 2014