# Visualizing Meta-Explanations in Early Intervention Systems for Police Departments

Damon Crockett* Joe Walsh† Klaus Ackermann‡ Andrea Navarrete§ Rayid Ghani¶

Center for Data Science and Public Policy
University of Chicago

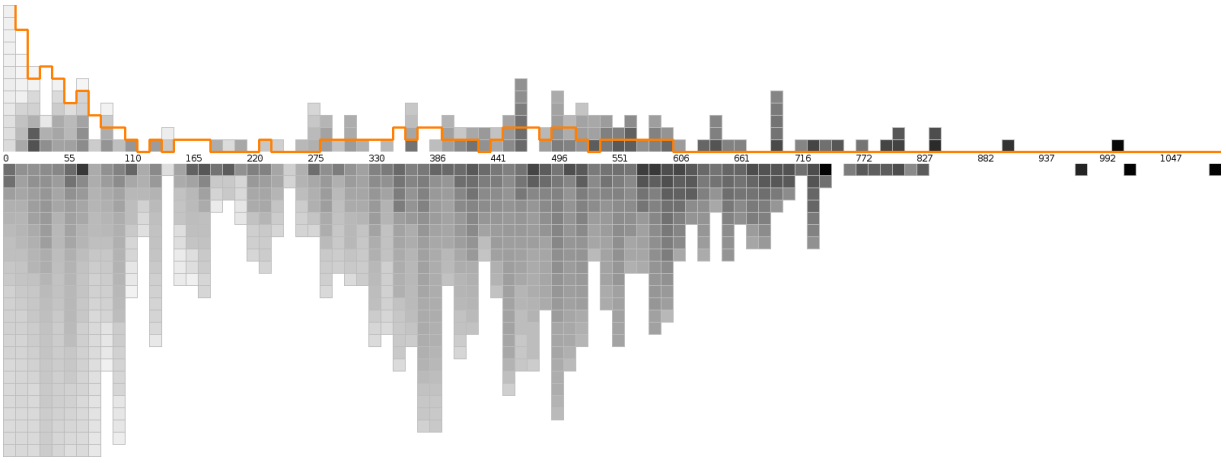


Figure 1: Positive (above) and negative (below, cropped) test set distributions on the variable **civilian-initiated dispatches**. Glyph brightness encodes modeled risk, ascending from white to black. Orange line traces negative distribution scaled by global positive-negative ratio. For middle and high values of the variable, officers are proportionately more likely to have had an adverse incident, which explains why this variable is correlated with high risk. However, the visualization also reveals that there are high risk officers with very low values for the variable. During interactive use, the user can select individual officers to discover which variables are most important in determining their level of risk and to see where they fall on the distributions of these and other variables.

## Abstract

The recent spread of machine learning methods into critical decision-making, especially in public policy domains, has necessitated a focus on their intelligibility and transparency. The literature on intelligibility in machine learning offers a range of methods for identifying model variables important for making predictions, but measures of predictor importance may be poorly understood by human users, leaving the crucial matter unexplained—viz., *why* the predictor in question is important. There is a critical need for tools that can interpret predictor importances in such a way as to help users understand, trust, and take action on model predictions. We describe a prototype system for achieving these goals and discuss a particular use case—early intervention systems for police departments, which model officers' risk of having "adverse incidents" with the public.

## 1 Introduction

There are (at least) two distinct aspects of the intelligibility of a statistical model. The first is predictor "importance", a general term comprising a highly diverse collection of measures. The second concerns understanding—an accounting of *why* a given predictor is important for the model or for one of its predictions. If we are domain experts, we hope, ultimately, to give such an account in terms of causal relationships, but we can make progress on the goal without appealing to causes and without simply pointing back to the definition of our chosen measure.

Importantly, understanding that falls short of substantive, causal explanation can nonetheless give us reason to *trust* the behavior of the model. On the assumption that measures of predictor importance provide explanation in some minimal sense, we promote understanding and trust by *explaining this explanation*—by giving, we might say, a "meta-explanation". Meta-explanations certify predictor importances by visually articulating the relationship between response and predictor using a minimum of algorithmic apparatus. Meta-explanations promise to illuminate model behavior both for model-builders and for domain experts who may lack statistical expertise and for whom importance measures are unintelligible on their own. In what follows, we will discuss a system that produces meta-explanations for a particular use case: early intervention systems (EIS) for police departments.

## 2 Police EIS

A police EIS uses internal departmental data to predict whether a given officer will have an "adverse incident" within the next year [2]. Our models are typically very large, using upwards of 4000 predictors, but we may report as few as 10 that are "important". Since these reports have the potential to influence departmental policy, it is essential that we are confident in our importance measure and can explain it to the department. These two constraints—model size and intelligibility—make our job difficult. Because our models are large, the marginal effect of any single predictor on the model is relatively

*e-mail: damoncrockett@gmail.com
†e-mail:jtwalsh@uchicago.edu
‡e-mail:ackermann@uchicago.edu
§e-mail:anavarreteriveras@gmail.com
¶e-mail:rayid@uchicago.edu

small, and such effects are therefore easiest to distinguish using precise, quantitative measures. On the other hand, because our reports must be understandable and trustworthy, we favor intuitive, visual presentations that depict only basic model elements. In view of these constraints, we are developing a system that is both understandable and capable of showing small effects.

## 3 System Design

### 3.1 Global Views

Global views show multiple variables at once, and there are two types: table and facet. In table views, each row is an officer and each column is a variable. Cells are colored according to variable group and their saturation levels encode normalized variable values. Rows and columns can be sorted to reveal patterns. Facet views contain small multiples of simplified variable views that show, for each predictor bin, the deviation in that bin—in officer units—from the global positive-negative ratio. The result is a single line that rises above the axis at bin locations where the ratio is greater and below the axis where it is lower than the global ratio. Both table and facet views create an interactive space for selecting variables for a closer look.

### 3.2 Variable Views

Variable views present binned distributions of variables as glyph histograms where each glyph represents an officer (cf. Amershi et al's *ModelTracker* [1]). The one-one mapping of officers to glyphs makes it possible visually to distinguish among officers in the same bins, creates for each officer an exclusive visual mark that can be interactively selected by the user, and encourages the user to think about the data in human terms native to the domain.

There are two types of variable views: response and predictor. Response view, for test sets and for the "active" set—the department's active roster of officers—shows the distribution of risk scores and, if available, depicts labels using glyph color. Predictor views are available for training, test, and active sets, and are described below.

Training Set Predictor views on the training set are used to articulate the specific ways that predictors help the model distinguish between positive and negative classes, and so are the primary sites for meta-explanation of global predictor importances. The data are split into positive and negative classes, and class distributions on the given predictor are plotted above (positive) and below (negative) the predictor axis. Because the negative class is always much larger than the positive class, it can be difficult to compare their shapes. We thus add a *third* distribution to the plot—viz., that hypothetical distribution having the relative bin heights of the negative class and the size of the positive class. This enables us both to see precisely the (possibly) complex way in which the positive and negative classes are split by the given predictor and to retain an officer-by-officer accounting of the data, all in the same plot. This re-scaled negative distribution is represented as a line, so it is not confused with the plotted glyphs representing actual officers (Fig. 1).

Variable Interactions It may turn out in some cases that although variable $V$ is deemed important, predictor view on distributions over $V$ does not show the expected split between positive and negative classes. In such cases, it may be that $V$'s importance depends on its interacting with another variable. We search for such interactions by first computing the Kullback Leibler (KL) divergence between the positive and negative class distributions over $V$, where $P$ is the positive and $Q$ is the negative class distribution:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{1}$$

We then consider all other variables in the data. For categorical variables, we use (1) to measure the distance between positive and negative classes in each category; for continuous variables, we do the same within each of some chosen number of bins. We then choose some number of candidate interactors and visualize the results as described in Section 3.2, one plot for each bin or category.

Test and Active Sets Predictor views on test and active sets allow us to see individual officers split into classes along a variable $V$, as before, but map an additional variable—risk, standardly—to glyph brightness (Fig. 1). For active sets, the 100 highest risk officers are usually treated as the "positive" class, but conversely we can treat all officers as belonging to the same class, in which case all are plotted above the axis.

Thus far, we've focused on the meta-explanation of global predictor importances, but because test and active sets include modeled risk, they introduce the additional problem of explaining (and meta-explaining) individual assignments of risk. If predictor importances are available for individual assignments of risk, glyphs can be colored to indicate whether or not the predictor is important for the assignment. Relatedly, users can select individual officers in variable views to produce facet views of either that officer's top predictors or of any subset of predictors (e.g., the globally important ones). In these special facet views, the officer's value for each predictor is indicated on the predictor axis.

Predictor views on the test and active sets are related to partial dependence plots (PDPs) [3], but PDPs have the disadvantages that they do not show the distribution over $V$, they do not represent individuals, and they average risk at each variable value over *every row in the data*, regardless of the actual value at that row. It's unclear what conclusions about the domain are justified by this counterfactual exercise, and the user may be tempted to view interventions on the relevant predictors as having known effects on risk.

## 4 Future Work

We are expanding on our work here in two ways. First, we are in the process of developing more detailed meta-explanations for individual model predictions (see Krause et al [4] for recent work on this problem). Presently, the system represents individuals as simple glyphs in a series of one-dimensional spaces. We are trying to determine whether greater representational complexity—both for individuals and for similarity spaces—can be made useful for us.

Second, our measures of usability and effectiveness are largely informal at present, and the system still needs input from the department before final design decisions are made. We are now undertaking a more formal user evaluation of the system and will deploy it alongside our existing system for model explanation, in order to compare the two.


## Acknowledgments

This work was supported by the Charlotte-Mecklenburg Police Department and the Eric Schmidt Foundation.


## References


[1] S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard, and J. Suh. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 337–346. ACM, 2015.

[2] S. Carton, J. Helsby, K. Joseph, A. Mahmud, Y. Park, J. Walsh, C. Cody, C. E. Patterson, L. Haynes, and R. Ghani. Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 67–76. ACM, 2016.

[3] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

[4] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5686–5697. ACM, 2016.