# Predicting Financial Crime: Augmenting the Predictive Policing Arsenal

Clifton, Brian
Threat Network Integrated

Lavigne, Sam
Threat Network Integrated

Tseng, Francis
Threat Network Integrated

## Abstract

Financial crime is rampant but hidden threat. The predictive policing community disproportionately focuses on blue collar crime while white collar crime remains untargeted by their algorithms. We propose and devleop a white collar crime predictive model, the White Collar Crime Early Warning System (WCCEWS), to fill this gap.

## I. Introduction

White collar crimes are those "committed by a person of respectability and high social status in the course of his occupation". Unlike blue collar, or "normal" crime, which is already the focus of many traditional and predictive policing initiatives, white collar criminals remain largely under-policed and underserved. The development of machine learning techniques provides police departments with new and exciting enforcement opportunities.

We propose and develop a predictive policing algorithm, the White Collar Crime Early Warning System (WCCEWS) for identifying and assessing the risk of large-scale financial crime at the city block level. With our model the police can more efficiently direct their limited resources for the greatest return on investment. The model is further integrated into tools for citizen policing and awareness, alerting users when they are in a high-risk area for financial crime.

Our model achieves 90.1% accuracy at predicting the activity of white collar crime in an area. The model is augmented to predict the severity of the crime (in terms of expected fines) and the nature of the crime.

Our model does not include predictive capacity for when financial crimes occur. As such we assume with a high degree of confidence that, for the predicted locations, financial crime occurs continually.

## II. Related Work

Our work is inspired by other predictive policing efforts, such as HunchLab [1] and PredPol [2]. These services overwhelmingly target "blue collar" or "normal" crime, overlooking the rich opportunity in targeting financial crimes with their technologies.

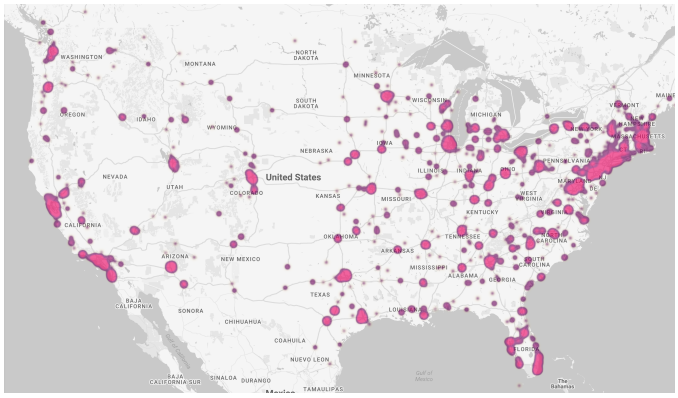


Fig. 1. Financial Crime Risk Surface

These services take the general approach of training some classification model with geospatial and/or historical crime data. They typically focus on predicting crime for a particular geographical region rather than at the level of an individual [1]. We adopt a similar approach, known as "Risk Terrain Modeling", or RTM, an approach to spatial risk analysis first proposed by Les Kennedy and Joel Caplan at Rutgers University. RTM is "used to identify risks that come from features of a landscape and model how they co-locate to create unique behavior settings for crime."

## III. Data

We collected data provided by the Financial Regulatory Authority (FINRA) [2] to compile incidents of financial malfeasance going back to 1964. Using these data we were able to match financial crimes to the location of the perpetrating individual or organization. Financial crimes were geographically clustered

[1] https://www.hunchlab.com/
[2] http://www.predpol.com/

Fig. 2. Features of a landscape that create unique behavior settings for crime

according to geohashes or zipcodes (depending on the specific model) computed from these locations.

To make our predictions we compiled auxiliary geographical data from a variety of sources. In particular, we looked at: 1) the locations of broker dealers [3]; 2) business employment statistics [4]; 3) active liquor licenses [5]; 4) lobbyists [6]; and 5) US government direct spending payments [7].

### A. Geohashes

In the geohash model, our geographical data started in the form of street addresses, which we converted to latitude/longitude coordinates with a geocoding service.

Coordinates are too fine-grained for useful predictions so we converted our coordinates to *geohashes*. A geohash is a set of characters that all coordinates in some region map to. For example, the coordinates `(40.15, 74)`, `(40.1, 74.1)`, `(40.1, 73.9)` all map to the geohash `txhs` (with a precision of 4).

The *precision* of a geohash is the size of a region coordinates are mapped to. A more precise geohash represents a smaller region and maps to a longer set of characters (increased precision requires increased specificity, and thus more characters).

For example, those same coordinates, when using a precision of 6, instead map to `txhs7v`, `txhsn5`, `txhs1e`, respectively. Note that the first four characters of each are still `txhs`.

For our model we use geohashes with a precision of 7, which map to regions of a 0.076km radius.

### B. Zipcodes

In the zipcode model, the geographical data starts in zipcode form, so no conversion step is required.

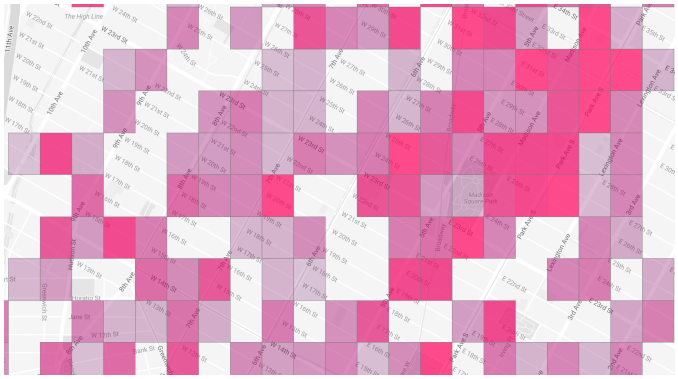

Fig. 3. Geohashes are shown as rectangles

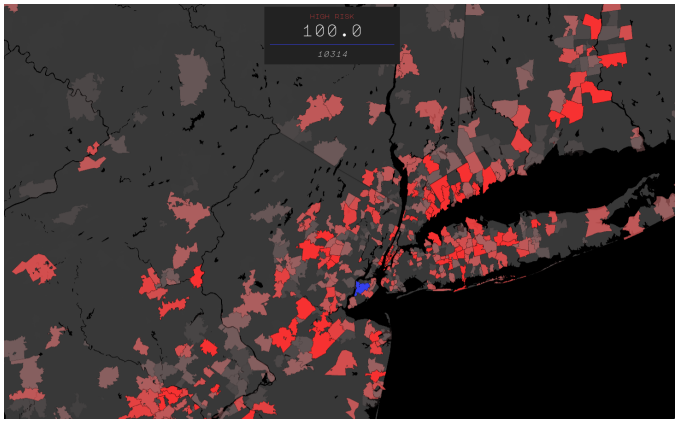


Fig. 4. Zipcode version of the model

## IV. The Model

Our model is composed of three sub-models, each trained to generate a separate prediction:

$M_{\text{crime}}$ predicts the probability of any financial crime occurring for a geohash. We use a forest of decision trees each generated by a bootstrap sample (i.e. a random forests model). The final prediction probability is the average of each tree's predicted probability. This is similar to the approach HunchLab uses [8]. Our model is trained on the aforementioned data.

$M_{\text{fine}}$ predicts the expected fine were a financial crime to take place in a geohash. It is a linear regression model trained on the same auxiliary data, with some additional polynomial features generated from the same data.

$M_{\text{type}}$ predicts a distribution over the types of financial crimes likely to occur in a geohash. It is a multilabel (one-vs-rest) random forest model, again trained on the same data.

## V. Conclusion & Future Work

In this paper we have presented our state-of-the-art model for predicting financial crime. By incorporating public data sources with a random forest classifier, we are able to achieve 90.1% prediction accuracy. It is difficult to know how this

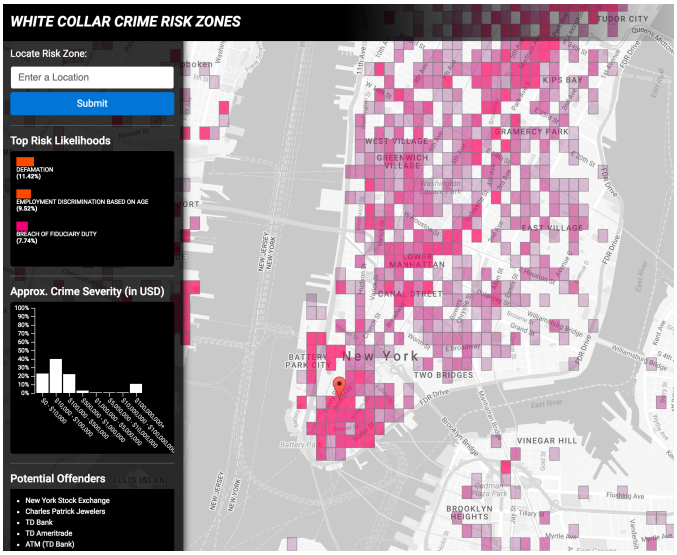


Fig. 5. The WCCEWS UI

compares with existing predictive policing algorithms as predictive policing accuracy is inconsistently defined and seldom independently measured [1]. We, however, are confident that our model's accuracy outperforms similar efforts and is ready for live deployment.

Our current model relies solely on geographical information. It does not consider other factors which may provide additional information about the likelihood of financial criminal activity.

Crucially, our model only provides an estimate for a particular region. It does not go so far as to identify which individuals within a particular region are likely to commit the financial crime. That is, all entities within that region are treated as uniformly suspicious.

Therefore we plan to augment our model with facial analysis and psychometrics to identify potential financial crime at the individual level. Recently standard machine learning techniques were shown to be an effective physiognomical approach, quantifying the *criminality* of an individual based on facial features [9]. Other predictive policing services incorporate astronomical features such as moon phases [10]. In light of these findings we are also considering incorporating phrenological data, and environmental factors such as contrails present in a region as additional predictors.

As a proof of concept we have downloaded the profile pictures of 300 random individuals whose LinkedIn profiles suggest they work for financial organizations in a high-risk area, and then averaged their faces to produce a generalized white collar criminal subject. Future efforts will allow us to predict white collar criminality through real-time facial analysis.

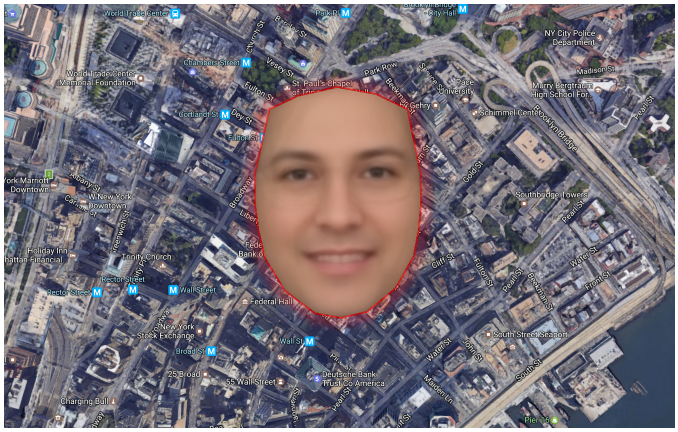

Fig. 6. Predicted White Collar Criminal for 40.7087811, -74.0064149

## References

[1] Upturn, "Predictive policing systems," " Jul 2016.
[2] FINRA. (2019) Financial industry regulatory authority.
[3] Enigma. (2012) Currently active - august 2012.
[4] ——. (2014) 2014 zipcode business practices - totals.
[5] ——. (2015) Active liquor licenses.
[6] ——. (2012) Primary report 2012.
[7] ——. (2009) U.s. government spending - direct payments - 2009.
[8] Azavea, "Hunchlab: Under the hood," " 2015.
[9] X. Wu and X. Zhang, "Automated inference on criminality using face images," *CoRR*, vol. abs/1611.04135, 2016.
[10] M. Chammah, "Policing the future," *The Verge*, Feb 2016.