# Crime Hotspot Prediction and Visualization in Atlanta

Aravind Rajeev Nair,* Karan Nahar,† Vamsi V Kalidindi,‡ Krishna Raj,§ and Naveen Sethuraman¶
*Team 40*
(Dated: December 3, 2023)

## I. INTRODUCTION

The city of Atlanta grapples with various crime-related challenges, demanding a comprehensive strategy for crime prevention and citizen safety. The motivation behind this project is to employ advanced data analytics to understand crime patterns, predict high-risk areas, and assign risk scores to different neighborhoods in Atlanta. By leveraging technological advancements and data-driven approaches, the project aims to offer actionable insights to law enforcement for effective resource allocation and proactive policing. The dataset we are using for our analysis is a **crime dataset** from the official Atlanta PD website. The dataset we are using has information on the day/month/year when the crime occurred, the type of the crime as well as latitude and longitude coordinates specifying the exact location of the crime, the data is from 2009 to 2020.

## II. PROBLEM DEFINITION

### A. Formal Problem Definition

The formal problem statement entails employing data and visual analytics techniques to predict crime hotspots and assign risk scores to different neighborhoods in Atlanta. The project aims to utilize machine learning algorithms and time series modeling to analyze crime data from 2009 to 2020, cluster crimes geographically, assign risk scores based on crime severity, and forecast future risk scores. The end goal is to visualize these risk scores on a choropleth map, facilitating smarter policing strategies.

We have broken down our project into the following steps -

- Basic exploration of the data (EDA) and sanity checks to ensure data consistency. This would also include identifying missing values/outliers and understanding the distribution of crime types.
- The crimes were clustered using the GMM algorithm based on the latitude and longitude of the crime. Based on Atlanta geography, We then decided on the number of clusters.
- Post obtaining geographical clusters, we assigned a risk score for each cluster based on the crime type. This was done based on the severity of the crime (eg: 10 for Homicide and 1 for Petty Theft etc). Thus we obtained a risk score for each cluster for each month/year.
- We used time series modeling techniques like Exponential Smoothing to predict the risk score for future periods based on the existing data for each geographical cluster.
- Then we visualized this result on a choropleth map of Atlanta showing each geographical cluster, risk score associated with it, and the top 5 types of crime in that area for different time periods.

This risk score can then be utilized for citizen safety and for smart policing where we can send more police officers to areas that have a higher risk score.

Technologies/Algorithms Used: Python, k-means, GMM, Exponential Smoothing, Plotly, Dash

### B. Objective

This project seeks to understand and predict crime hotspots in Atlanta by analyzing historical crime data. By using machine learning and statistical models, we aim to group crimes based on location, assign risk scores considering the severity of different crime types, and forecast future risk scores. The visualization of these scores on a map will aid law enforcement in making informed decisions about resource allocation and enhancing citizen safety in high-risk areas.

## III. LITERATURE SURVEY

Several existing research has been done in the field of crime detection and prediction using machine learning techniques, we are summarizing some of these methods.

[1] discusses the use of LSTM and environmental data to predict crime hotspots. It is valuable for predicting crime locations, but it falls short in providing a safety score for neighborhoods.

[2] explores ensemble methods for crime prediction,

* anair@gatech.edu
† knahar3@gatech.edu
‡ vamsivk@gatech.edu
§ kraj34@gatech.edu
¶ nsethuraman3@gatech.edu

with a stacked SVM Classifier as the top performer. It offers insights into improving predictive accuracy but may have issues with high time complexity in real-time implementation.

[3] uses deep learning to predict crime types from descriptions, introducing the idea of risk scores. It is helpful for understanding crime types but lacks a visual crime map and focuses on individual crimes.

[4] presents CrimeVis, a system for visualizing crime and socioeconomic data. It offers tools for pattern exploration and policy evaluation, which can inform machine learning models, but might need more advanced modeling techniques.

[5] employs big data analytics and machine learning to predict crime patterns. It focuses on extensive data preprocessing and predictive modeling but might need broader testing across different locations.

[6] evaluates data mining methods for understanding criminal behavior. It aligns with our project goals but could benefit from a larger dataset and more diverse algorithms.

[7] predicts crime severity and visualizes trends using various machine-learning models. It's helpful for predicting future crime hotspots but lacks a discussion on model interpretability.

[8] uses XGBoost and SHAP to predict and interpret crime predictions. It offers insights into understanding the causes of crime but needs testing in different cities.

[9] proposes modeling crime prediction as a recommendation problem for finer granularity. It can aid in understanding when and where crimes occur in detail.

[10] utilizes KMeans clustering to identify crime hotspots, primarily in San Francisco. It provides a solid foundation for hotspot identification but could incorporate temporal data.

[11] employs KNN for predicting crime types with temporal features. It offers guidance on feature investigation but doesn't predict future crime hotspots.

[12] surveys various crime forecasting methods, offering a broad perspective. It provides a list of methods to explore but may not delve deeply into technical details.

[13] discusses predictive policing and risk assessment, emphasizing the importance of accuracy, fairness, and transparency. It highlights the need to address biases in historical data but lacks detailed technical solutions.

[14] introduces a multivariate prediction model for hotspots based on criminal preferences. It offers insights into predicting hotspots and comparing model performance with existing practices.

[15] emphasizes the complexity of ensuring fairness in ML systems and the need to consider socio-technical contexts. It identifies potential pitfalls but doesn't provide explicit technical solutions.

Based on literature, We have tried using time-series modeling for more accurate hotspot prediction ([1], [2], [10], [11]), predicting risk scores ([3]), and leverage data visualization tools ([4], [5]) to get a visual representation of hotspots.

## IV. METHODOLOGY AND EXPERIMENTS

### A. Intuition

The project's approach revolves around several innovative methodologies and practices aimed at enhancing crime analysis, prediction, and community engagement. Below we elaborate on each aspect and explain why this approach is expected to surpass the current state of the art in crime analysis:

- **Clustering Based Approach**: Leveraging Gaussian Mixture Models (GMM) for clustering enables the identification and adaptation of crime clusters, providing a more nuanced and up-to-date understanding of geographical crime distribution.
- **Predictive Policing:** By utilizing predictive models, law enforcement can proactively deploy resources in high-risk areas. Forecasting crime in each geographical cluster using exponential smoothing helps anticipate future crime rates, allowing for preemptive measures and resource allocation.
- **Interactive Visualization**: We developed an intuitive and interactive interface that facilitates data exploration and risk score analysis for a broader audience. Accessibility to a user-friendly platform enables policymakers, law enforcement, and the community to comprehend and utilize crime data effectively.

Overall, the project's emphasis on clustering, predictive policing, community engagement, and user-friendly visualization establishes its superiority over existing methods by offering more adaptable, accurate, and inclusive approaches to crime analysis and prevention.

### B. Methodology

- **Exploratory Data Analysis**:

  We started out by understanding the structure of our **crime dataset**.
  When we analyzed the count of crimes by month and year, we saw a decreasing trend of the crimes with an added seasonality factor in the occurrence of the crimes as shown in figure 1.

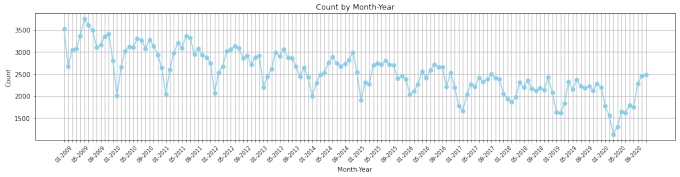


FIG. 1. Count of Crimes by Month/Year

We looked at the distribution of crimes for different days of the week in figure 2 to check for any patterns, however, the distribution looks even.

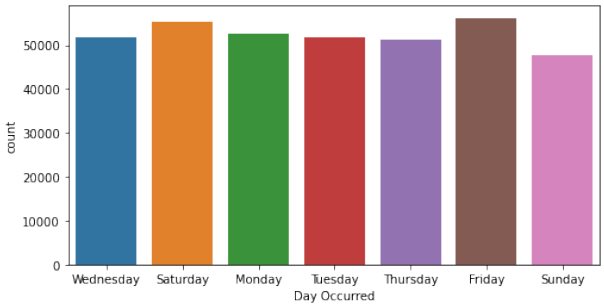


FIG. 2. Count of Crimes by Day of Week

Larceny (vehicular and non-vehicular), burglary, and auto theft are the three highest types of crime as shown in figure 3.

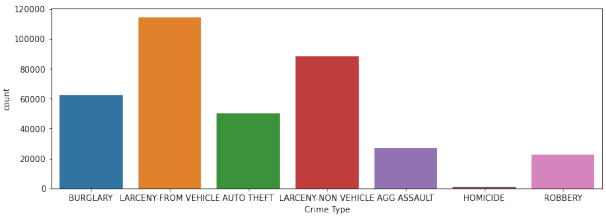


FIG. 3. Count of Crimes by Type

We also looked at the trend across time for these types of crime and saw that crimes like larceny-non vehicle have seen a dip in recent years, particularly after COVID-19 as seen in figure 4.

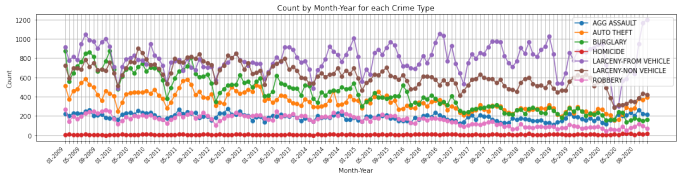


FIG. 4. Crime Types across Months/Years

We also experimented with the folium library in Python to render an interactive map that has the count of crimes in each geographic area of Atlanta as shown in figure 5.

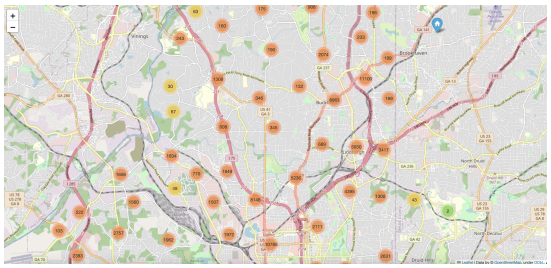

FIG. 5. Crime Heatmap by Location

- **Clustering crimes by Location**

  We have tried using both K-means and Gaussian Mixture Models to cluster the crimes by geography and have used the latitude and longitude of the crimes to cluster them. K-means is a non-probabilistic, hard-clustering technique that assigns data points to the nearest cluster, while GMM is a probabilistic, soft-clustering method that models data as a mixture of Gaussian distributions, allowing for more complex cluster shapes and providing uncertainty estimates.

  We first plotted the data points to see the geographic spread of the data. There were a few outliers (a couple of crimes occurring far south of the city in Jonesboro and Stockbridge). As these few points would be insufficient for analysis of crimes in those areas, we opted to exclude them from the dataset.

  To cluster the data into crime hotspots, we initially considered KMeans clustering, selecting an optimal Kvalue using the elbow method on the clustering inertias obtained from different values of K. This yielded an optimal K value of 5. On further consideration, we observed the 5-Means clustering approach to be deficient in that it failed

to capture crimes occurring around the Hartsfield-Jackson Atlanta International Airport as its own cluster, though the points were visually identifiable as being geographically distinct. Instead, the model tended to merge the airport data with other clusters.

As we would expect the airport to have a different distribution of crimes compared to urban and suburban areas, we thought it imperative that our clustering model identified it as its own cluster. After trying different combinations, we settled on a Gaussian Mixture model with a K value of 30. The large k-value allowed for smaller clusters, and hence proper clustering of the airport data. Using a Gaussian Mixture instead of K-Means allowed the clustering to capture long stretches of higher crime areas (eg. crimes along Hwy 20 to the west of Atlanta).

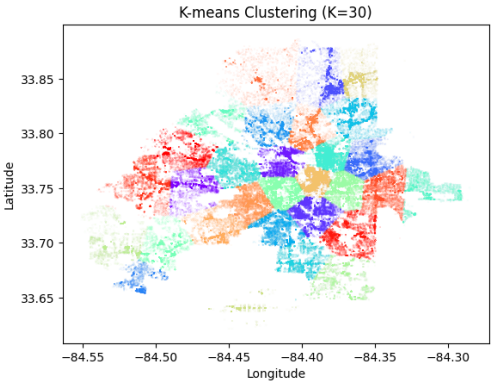


FIG. 6. Clusters using k-means

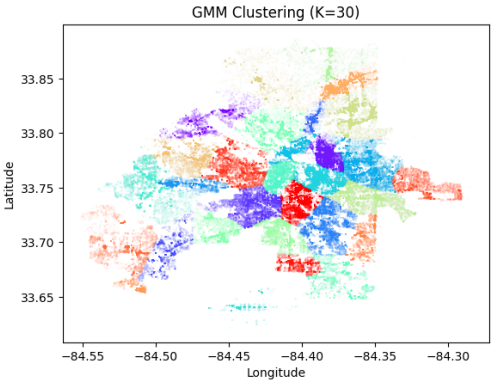


FIG. 7. Clusters using Gaussian Mixture Models

- **Crime Rating Assignment** : We have used the following scales for the 7 crime types:
  - Homicide- 5
  - Agg Assault - 4
  - Auto Theft- 3
  - Burglary- 3
  - Robbery- 3
  - Larceny from vehicle- 2
  - Larceny non-vehicle- 2

  The severity of these are from 5 to 1; 5 being the most severe crime. These scores are only illustrative, however, these scores can be changed based on the general sentiment or an official score can be introduced.

  Based on score assignment, each cluster and month is assigned a weighted score based on the count of crimes that occurred in the particular month indicating the severity of crimes in those areas during specific periods.

  This rating is normalized per day, allowing for a more accurate assessment of crime severity across different timeframes.

- **Time series modeling**: Basis the score assignment to each cluster across each month, the score can be visualized for one specific cluster (cluster 10 encompassing regions of English Ave and Vine City) as shown below in fig 8.

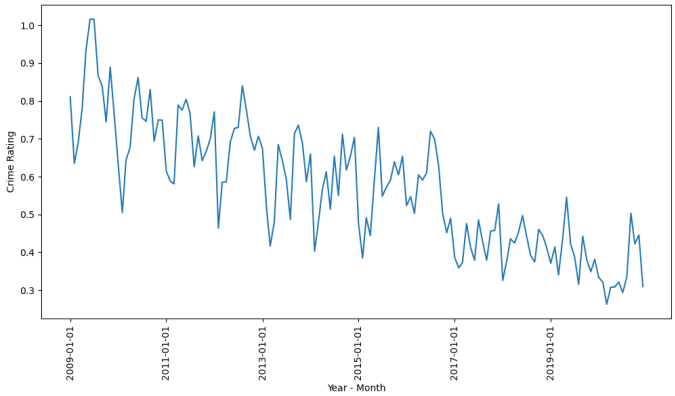


FIG. 8. Risk Score across months for cluster 10

From the plot, we can see a seasonality of 1 year with a trend, based on which we build a time series modeling technique using exponential smoothing for predicting the risk scores/crime ratings for each cluster.

In our crime risk score prediction model, we divided the dataset into training (2009-2018) and testing (2019-2020) periods. Employing an exponential smoothing model with a seasonality term of 1 year and a trend term, we capture recurring patterns in criminal activity and trends over time. This model choice allows adaptability to short-term fluctuations while discerning the long-term trajectory of crime rates.

The crime rating predictions for cluster 10 can be visualized in the below figure 9.

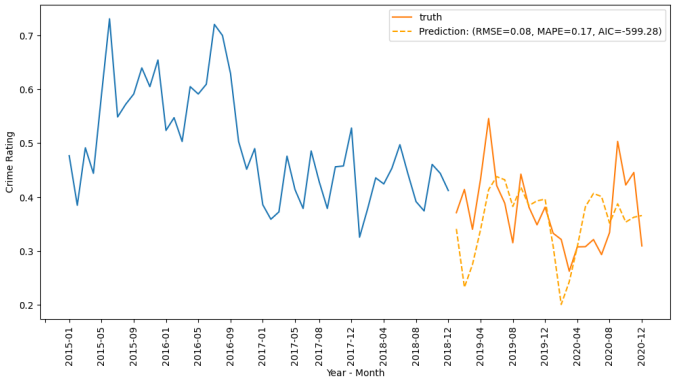


FIG. 9. Risk Score Predictions across months for cluster 10

- **User Interface and Visualization**: The interactive map, designed using Plotly and Dash, incorporates a suite of features aimed at providing users with a comprehensive and insightful exploration of crime data. Firstly, a dropdown menu allows for the selection of a specific Month-Year, enabling targeted analysis. The map dynamically generates visual representations of crime clusters, complete with a heatmap showcasing the intensity of criminal activity across different locations. Enhancing user engagement, hovering over a particular cluster triggers a tooltip displaying both the actual and predicted crime ratings for that specific cluster during the chosen Month-Year.

  Moreover, this interactive map also integrates a dynamic bar chart functionality. When users hover over a crime cluster, a corresponding horizontal bar chart shows below the map. This chart illustrates the most prevalent crimes within that cluster for the selected Month-Year, offering a granular breakdown of criminal activities. The integration of these features not only transforms the map into a powerful analytical tool but also enriches the user experience by providing meaningful insights into crime patterns and trends. A snapshot of the interactive visualization map is shown in figure 10.

**Conclusion**: Conclusively, the undertaken steps have successfully yielded risk scores across different clusters and months, providing a comprehensive overview of crime risks in our analysis.

### C. Experiments and Observations

The major questions that our experiments were designed to answer are as follows.

- What is the best way to cluster the regions in the city?
- How do we assign and predict the risk scores?
- What are the insights we obtain from visualizing the predictions?
- Which regions have the highest and lowest crimes?
- What is the overall trend in crimes over the years?

**Clustering**: We utilized a Gaussian Mixture Model (GMM) with a K value of 30 to delineate distinctive crime clusters within the dataset. The GMM algorithm effectively identified the airport area as a unique cluster owing to its specific crime distribution characteristics.

GMM's selection over K-means stemmed from its adeptness in handling clusters characterized by diverse shapes and densities. This model's adaptability to complex and irregular data distributions was pivotal, especially in accurately mapping the intricate crime patterns observed across different areas. It notably outperformed K-means due to its capacity to accommodate the nuanced features inherent in the crime dataset.

**Risk Score Prediction and Accuracy**: The approach (mentioned in the Methodology section) provides a reliable method for predicting crime trends in different clusters, which is essential for proactive crime management and resource allocation in community safety initiatives.

Across the clusters, our model showcased consistent performance metrics on the test dataset. The averaged low Sum of Squared Errors (SSE) of 0.06 and Mean Absolute Percentage Error (MAPE) of 31% collectively illustrate the model's effectiveness. These metrics serve as a robust measure of the model's practical utility in predicting crime trends accurately.

**Insights from Visualization**: The tool harnesses the capabilities of Dash and Plotly, offering an interactive dashboard for in-depth analysis of crime data. It includes several features designed to enhance usability and insight generation:

- **Dynamic Filtering**: Users can select specific months and years using a dropdown menu, dynamically updating the displayed data for precise analysis.


FIG. 10. Interactive Dashboard

- **Geographical Visualization**: An interactive map visually represents crime clusters, utilizing color-coded points based on crime ratings. Hover functionality provides detailed cluster information, enhancing geographical insights.

- **Cluster-Specific Crime Breakdown**: Hovering over a cluster on the map triggers a horizontal bar chart, displaying crime types and their counts for the selected cluster and time period.

- **Map Functionality**: The map employs a carto-positron Mapbox style, ensuring clarity and readability. It also offers zoom functionality for detailed exploration, enabling a closer examination of specific areas.

Despite minimal variation observed month by month in the clusters, notable patterns emerge from the data analysis. Areas close to Five Points and the Hotel District exhibit higher risk scores and crime ratings, indicating higher crime prevalence. Conversely, regions like Carey Park, Kirkwood, and East Lake Highlands, situated farther from Atlanta's city center, portray lower crime ratings.

The concentration of crimes appears centered around the Atlanta city center, particularly in locations such as Home Park, Atlantic Station, and Midtown. Vehicular larceny and auto theft emerge as the predominant crimes in these regions, suggesting a need for heightened policing and preventative measures tailored to combat these specific crime types.

Additionally, an overall decline in crime ratings over the years is observable, signifying potential positive trends in the city's safety landscape.

## V. CONCLUSION

This project aimed to better understand historical crime data from the city of Atlanta and explore ways to improve crime prevention and citizen safety. We planned to look at the crime trends in different regions of Atlanta and make predictions of future crime rates and risk levels, empowering law enforcement agencies to be better prepared.

Our project provided substantial insights into city crime patterns, addressing key research questions. The Gaussian Mixture Model (GMM) with a K value of 30 effectively delineated crime clusters, displaying adaptability to complex data distributions and accurately identifying unique clusters like the airport area.

Our proposed approach for predicting crime trends demonstrated reliability, exhibiting strong performance metrics on the test dataset, with low Sum of Squared Errors (SSE) and a Mean Absolute Percentage Error (MAPE) of 31%. The interactive dashboard, powered by Dash and Plotly, offered a valuable tool for crime data analysis, featuring dynamic filtering, geographical visualization, and cluster-specific crime breakdowns.

Identification of higher-risk areas in the city center, such as Five Points, Hotel District, Home Park, etc

emphasizes the need for targeted interventions. The observed decline in crime ratings over time suggests potential positive shifts in the city's safety landscape, underscoring the impact of our predictive analysis.

While our project presented valuable insights, it is essential to acknowledge certain limitations. The visualization tool, while robust, may benefit from enhancements for deeper interactive exploration. Additionally, the model's effectiveness in predicting specific crime types could be further refined for targeted interventions. Looking ahead, the implications of our work advocating for tailored crime prevention strategies, especially in combatting prevalent crimes like vehicular larceny and auto theft. Future extensions of this project could involve incorporating additional data sources for a more comprehensive analysis, as well as using more complex machine learning models for risk score prediction ultimately fostering safer communities through data-driven interventions.

All team members have contributed equally to this project.

## VI. REFERENCES


[1] X. Zhang, L. Liu, L. Xiao, and J. Ji, Comparison of machine learning algorithms for predicting crime hotspots, IEEE Access **8**, 181302 (2020).

[2] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim, and G. R. Sinha, An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach, IEEE Access **9**, 67488 (2021).

[3] M.-S. Baek, W. Park, J. Park, K.-H. Jang, and Y.-T. Lee, Smart policing technique with crime type and risk score prediction based on machine learning for early awareness of risk situation, IEEE Access **9**, 131906 (2021).

[4] L. J. S. Silva, S. Fiol-González, C. F. P. Almeida, S. D. J. Barbosa, and H. C. V. Lopes, Crimevis: An interactive visualization system for analyzing crime data in the state of rio de janeiro, in *International Conference on Enterprise Information Systems* (2017).

[5] M. Feng, J. Zheng, J. Ren, A. Hussain, X. Li, Y. Xi, and Q. Liu, Big data analytics and mining for effective visualization and trends forecasting of crime data, IEEE Access **7**, 106111 (2019).

[6] O. Llaha, Crime analysis and prediction using machine learning, in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)* (2020) pp. 496–501.

[7] A. Tamir, E. Watson, B. Willett, Q. Hasan, and J.-S. Yuan, Crime prediction and forecasting using machine learning algorithms, International Journal of Computer Science and Information Technology **12**, 26 (2021).

[8] X. Zhang, L. Liu, M. Lan, G. Song, L. Xiao, and J. Chen, Interpretable machine learning models for crime prediction, Computers, Environment and Urban Systems **94**, 101789 (2022).

[9] Y. Zhang, P. Siriaraya, Y. Kawai, and A. Jatowt, Predicting time and location of future crimes with recommendation methods, Knowledge-Based Systems **210**, 106503 (2020).

[10] G. Hajela, M. Chawla, and A. Rasool, A clustering based hotspot identification approach for crime prediction, Procedia Computer Science **167**, 1462 (2020), international Conference on Computational Intelligence and Data Science.

[11] A. Kumar, A. Verma, G. Shinde, Y. Sukhdeve, and N. Lal, Crime prediction using k-nearest neighboring algorithm (2020) pp. 1–4.

[12] N. Shah, N. Bhagat, and M. Shah, Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention, Visual Computing for Industry, Biomedicine, and Art **4** (2021).

[13] R. A. Berk, Artificial intelligence, predictive policing, and risk assessment for law enforcement, Annual Review of Criminology **4**, 209 (2021), https://doi.org/10.1146/annurev-criminol-051520-012342.

[14] H. Liu and D. E. Brown, Criminal incident prediction using a point-pattern-based density model, International Journal of Forecasting **19**, 603 (2003).

[15] A. D. Selbst, d. boyd, S. Friedler, S. Venkatasubramanian, and J. Vertesi, Fairness and abstraction in sociotechnical systems, in *FAT* 2019* (2020).