Impact Charts: A Tool for Identifying Systematic Bias in Social Systems and Data

Darren Erik Vengroff
vengroff@gmail.com

ABSTRACT

We introduce impact charts and apply them to problems of systematic bias encoded in three different data sets. Impact charts are highly visual, making the effects they find easy to understand by both domain experts and non-experts. Impact charts are nonlinear and non-parametric, so they are able to identify structural biases whose functional forms are not *a priori* well understood.

Impact charts are based on SHAP, an interpretability method initially designed to interpret predictions made by Machine Learning (ML) models, which is in turn based on Shapley values, an approach to assigning responsibility for economic outcomes to different factors. Although impact charts use techniques from the ML community, they are intended for use in general settings, whether ML is present or not.

Impact charts provide valuable insights even when generated from aggregate data sets. Aggregate data sets typically provide the individuals whose data they are derived from an additional level of privacy as compared to the original unaggregated data. In this work, we relied predominantly on aggregate data from the U.S. Census Bureau, which is known to have a robust privacy program.

We introduce and evaluate impact charts using three examples of their use. Our first example uses impact charts to identify racial and ethnic bias in eviction rates. Our second example uses U.S. Census data to identify racial and ethnic bias in housing prices. Our third example assesses the impact of several factors on local access to supermarkets.

All three examples not only correct for the effects of income, but also clearly demonstrate the relative impact of income as compared to racial and ethnic features. For example, we demonstrate that in some areas like DeKalb County, GA, the fraction of the population that is Black impacts eviction rates three times more than income does.

In addition to the impact charts specifically discussed herein, we have produced thousands of geographically localized impact charts for the data sets mentioned above.

There is wide variation in the shape and structure of impact plots built using data from different local areas. We hypothesize that in the future work we will be able to categorize these and identify local policy decisions, whether *de jure* or *de facto* that cause the differences from one area to the next.

CCS CONCEPTS

• **Applied computing** → **Sociology**; *Economics*; • **Computing methodologies** → Classification and regression trees; *Machine learning*.

KEYWORDS

Impact Charts, Systematic Bias, Machine Learning

ACM Reference Format:
Darren Erik Vengroff. 2024. Impact Charts: A Tool for Identifying Systematic Bias in Social Systems and Data. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3630106.3658965

1 INTRODUCTION

In 1881, Frederick Douglass wrote an article entitled, "The Color Line" [21]. The color line was his name for the "moral disorder," of racial prejudice. He saw it as neither natural nor inevitable. Nineteen years after Douglass' work, at the 1900 Paris Exposition, W.E.B. Du Bois presented the color line in visual form. He exhibited charts and graphs that presented a wide variety of data on the lives and circumstances of Black Americans [4].

We take inspiration from Du Bois' work seeking to visually expose the color line, but using a new method inspired by and applied to the age of Machine Learning (ML). We need new tools because since the color line was first identified, it has been objectivized [40], systematized [46, 53] and digitized [6, 12, 23, 43]. Our contribution is the idea of *impact charts*, a new tool that builds on SHAP [37] to create a powerful means of interrogating data sets and exposing the biases they contain.

Impact charts as a bias detection mechanism are also inspired by prior work on helping users make less biased decisions [5, 59, 60], and interactive systems that help them to do so [8, 14, 33, 41, 62].

In this work, we will concentrate our attention on data sets that were initially collected to measure the status quo of social, political, and economic systems. Census data sets are good examples of this. When impact charts are built on top of this kind of data sets, they can expose the underlying systematic biases the data encode more directly and more flexibly than traditional tools such as scatter plots [15] and regression analysis [20]. Furthermore, as we shall see in examples below, impact charts can provide insights from aggregated data sets without the need for or risk of personally identifiable information. We deliberately used data sets whose raw inputs are predominantly from the U.S. Census Bureau, which has been a leader in deploying privacy preserving techniques in the generation of aggregate data [2, 55].

Impact charts can also serve a diagnostic role in ML [1] when built on top of data sets that are being used or might be used to train ML models that replicate their biases. Impact charts can be

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658965

generated as part of the testing phase of data set development [29] or as part of a data set nutrition label [28, 51].

The remainder of this work is organized as follows: In Section 2 we define the general problem of impact assesment, and define impact charts as a solution. Because impact charts are visualizations, we introduce them and work through how they are constructed and interpreted and what insights they can help us with in an example-driven way over the course of Sections 3-5. In Section 3, we introduce the visual vocabulary of impact charts and demonstrate how they show the persistence of the color line when applied to data on eviction rates, race, and ethnicity. In Section 4, we switch to data on home values, which we use not only to expose the color line again, but also to compare impact charts to alternatives. In Section 4.2 we compare impact charts to scatter plots [15]. In Section 4.3, we demonstrate the flexibility impact charts because of their lack of assumptions and flexibility of scoring metric vs. regression analysis. In Section 5, we show how a collection of impact charts, one for each of several features, can help us quickly refine hypotheses and identify the feature or features with the largest impact. Finally, Section 6 offers some conclusions and suggested next steps for the further research, development, and application of impact charts.

2 DEFINITIONS

2.1 General Problem Statement

Before we dig into impact charts as a solution, we begin with a general problem, stated as follows: given n observations of the values of $m+1$ variables, what is the relationship between m of them and the other one? Is there a relationship? How is it structured? What quantitative, qualitative, and visual characteristics does the relationship have? Within this relationship what is the contribution to or *impact* of each of the m variables on the final one.

In ML settings, whose conventions we will use herein, the m variables are called *features* and denoted $x_0, x_1, ...x_{m-1}$ while the final variable is called the *target* and is denoted y [9]. We will work with a set of n observations of the feature values $(x_0^{(j)}, x_1^{(j)}, \ldots, x_{m-1}^{(j)})$ and the target value $y^{(j)}$ for $0 \leq j < n$.

2.2 Impact Charts

Impact charts are a set of m plots, one for each feature x_i, that attempt to show the impact of different values of x_i that appear in the data on the corresponding values of y.

Impact charts are based on SHAP [36, 37], a method designed to interpret predictions made by ML models. SHAP is in turn based on Shapley values [48], an approach to assigning responsibility for economic outcomes to different factors. SHAP has also been used to identify racial bias in credit decisions [58]. Although impact charts use SHAP and other techniques from the ML community, they are intended for use in general settings, whether ML is present or not.

Impact charts and the Shapley values that underly them are structured such that the impacts of each of the features on the target for a given data point add up to exactly the difference between the target value of the point and the mean value of the target across the entire data set [36, 37, 48]. Even if feature values are correlated, we tease apart relative impact across the features such that each impact is independent of the impacts of other features. Mathematically,

$$\hat{y}(x_0, x_1, ...x_{m-1}) = \mathop{\mathbb{E}}_{0 \leq j < n}[y^{(j)}] + \sum_{0 \leq i < m} I(x_i)$$

where $I(x_i)$ is the impact of x_i. That is to say that the prediction $\hat{y}$ made by some ML model is simply the mean of the target y values it was trained on plus the sum of the impacts of each of the features x_i.

This property is important because it lets us look at a plot of $I(x_i^{(j)})$ vs. $x_i^{(j)}$ for all $0 \leq j < n$ showing the impact of a single feature for all n feature values seen in a data set. We don't have to worry that what we are seeing is actually the effect of some other feature, because that appears in an entirely separate impact chart for that other feature. In this respect it is fundamentally different and more powerful than an $(x_i^{(j)}, y^{(j)})$ scatter plot [15]. We will see this in Section 4.2.

Two of the examples we will discuss below come from residential real estate, which has been widely shown to harbor significant racial biases [27, 44, 46, 47, 53]. In Section 3, we consider a data set with x_i describing the median household income and racial and ethnic distribution of renters in different neighborhoods, and y indicating the eviction rate in each of those neighborhoods. We might expect an impact chart for the impact of renter income on eviction rate to be that low-income neighborhoods have higher rates of eviction and high-income neighborhoods have lower rates of eviction. And in an unbiased world, we would expect there to be no impact from the racial and ethnic features. Impact charts enable us to interrogate hypotheses like these and, if they are invalid, quickly come up with alternate hypotheses.

The additive nature of impacts can be easily seen in the other real estate example that we discuss in detail in Section 4. There, we use impact charts to say things like housing prices in neighborhoods that are 80% white are impacted by \$50,000. By this we mean that they are \$50,000 different than prices in otherwise comparable neighborhoods where the level of whiteness varies. Similarly, in a high income neighborhood, income may have a \$100,000 impact on housing prices. Even if this neighborhood is more white than average, and whiteness is correlated with higher incomes in the data, this impact is completely additive to the impact of the racial demographics of a neighborhood. So a neighborhood might have housing prices \$50,000 above the mean because of it's whiteness and an *additional* \$100,000 because of its high income for a total of \$150,000 above the mean. A similarly high-income neighborhood that is majority minority instead of majority white would see the income impact but not the racial impact, giving it housing prices \$100,000 above the mean.

Because of the way they break impact down into additive components, impact charts can be used as a tool for identifying and visualizing social biases embedded in a variety of data sets and the social, political, and economic systems that produced them. For example, in a hypothetical world where there is no systematic racism and the data were perfectly representative, the impact charts for the housing price data set would show that the whiteness of a neighborhood had zero impact on housing prices. The same would be true of the Blackness, Asianess, Hispacness, or any other racial or demographic *-ness* represented in the data.

The extent to which these impacts are non-zero in particular circumstances can help guide our understanding of how and where systematic racism [24] and other systematic biases exist and what policy changes might be necessary to mitigate them.

2.2.1 Machine Learning and Prediction vs. Explanation.

2.2.1 Machine Learning and Prediction vs. Explanation. Impact charts are built on top of ML models. The ones presented herein use a technique called gradient boosted trees [25] as implemented in the XGBoost package [17], but our software also supports using other ML techniques to generate them. We chose gradient boosted trees because they are generally regarded to be effective for modeling tabular data [39]. Tree-based Shapley values are also less computationally intensive than those computed on other types of ML models [36].

In general ML problems are posed as prediction problems, not explanation problems [49]. Given a set of n training data points $(x_0^{(0)}, x_1^{(0)}, ...x_{m-1}^{(0)}, y^{(0)})$ through $(x_0^{(n-1)}, x_{,1}^{(n-1)}, \ldots x_{m-1}^{(n-1)}, y^{(n-1)})$, the goal is to *learn* a function $\hat{y}(x_0, x_1, ...x_{m-1})$ that minimizes some function of the difference, or error, between $\hat{y}(x_0^{(j)}, x_1^{(j)}, ...x_{m-1}^{(j)})$ and $y^{(j)}$ across all $0 \leq j < n$. ML systems then use this function to make predictions $\hat{y}(x_0^{(j)}, x_1^{(j)}, ...x_{m-1}^{(j)})$ for $j \geq n$ of the value of the target for points in feature space that were not seen at training time.

While ML models can be good at prediction problems according to commonly-used metrics (with the caveat that errors are often concentrated in ways that disproportionately impact historically marginalized groups [13, 43]) they tend not to be particularly explainable, which might make them appear not to be applicable to our problem. The good news on this front is that interpretability techniques have been developed [36, 37, 45, 61] and among them SHAP [36, 37] in particular has the additive impact property we need for impact charts.

The SHAP approach has the additive characteristics we want impact charts to have, but it focuses on explaining the predictions made by a specific model trained on a specific data set, not on explaining the impact of features in a data set, much less the impact of the systematic social, political, or economic processes whose behavior is encoded in the data set.

To solve this problem, impact charts apply SHAP not to a single ML model, but to an ensemble of k ML models, each trained on a random subset of the initial data. This is a form of bagging [11] but we explain by aggregating Shapley values rather than predictions. A related approach is to use bagging with carefully designed weak but inherently explainable models [34, 35] rather than SHAP on a more general set of modeling techniques.

We use bagging of k models for two reasons. First, a single model can be prone to the problem of *over-fitting*, which means that it essentially memorizes the data it is trained on, but does not contain any structure that allows it to generalize to accurately predict $\hat{y}$ for feature values it has never seen [18]. By using an ensemble, we ensure that none of the individual models is overfit to the data set as a whole.

Second, and more critically important, when we have k models, we can look at the impact of a given feature on the target for a given data point in each of the models independently. It is often the case that models trained on different random samples of the same data set have a similar overall error rate across the data set, but concentrate their errors in different areas. The net result of this, which we will see visually in Section 3 is that we get not just a single estimate of the impact of a feature on the target, but an empirical distribution of it. The structure, and in particular the variance of this distribution helps us reason about how to interpret the estimate of the impact. Intuitively, if the models in the ensemble agree, to within a small amount of variance, on the impact of a feature on the target, we can more reasonably believe that the models have independently latched on to a relationship inherent in the data or the process that created it, rather than that each of member of the ensemble has overfit to the subset of the data upon which it was trained.

3 EVICTION, RACE, ETHNICITY, AND INCOME

3.1 A First Impact Chart: Eviction and Blackness

Our first set of impact charts were generated from a data set with a point for each census tract in DeKalb County, Georgia in each of ten years from 2009 to 2018 inclusive. The data set has ten features. Seven of the features indicate the fraction of renters in the tract who identify as belonging to each of the following racial groups: White Alone; Black Alone; American Indian or Alaskan Native Alone; Asian Alone; Native Hawaiian or Pacific Islander Alone; Some Other Race Alone; and Two or More Races. The eighth feature indicates the fraction of renters in the tract who identify as being White Alone not Hispanic or Latino. The ninth feature represents the fraction of renters who identify as being Hispanic or Latino of any Race. The tenth feature is median household income for renters in the tract normalized to constant 2018 dollars. These features were derived from U.S. Census American Community Survey (ACS) 5-year data [56].[1] The target of the impact charts is eviction rate, measured as the number of evictions in a year per 100 rental households in the tract. This data comes from the Eviction Lab at Princeton University [26]. Although we discuss only DeKalb County here, we have prepared similar eviction impact charts for 489 counties across the country and published them at http://evl.datapinions.com/.

As in all the examples herein, we used aggregate data rather than individual data, thus lessening both our risk of failing to preserve privacy [42] and our risk of introducing noise by imputing unknown racial features in data where they are missing [27].

We will start with Figure 1, an impact chart that looks at the impact of the percentage of renters who identify as Black on the rate of eviction filings in DeKalb County, Georgia between 2009 and 2018.

Let's look at this first impact chart and see what it tells us. We will start with the leftmost fifth of the chart, representing tracts that that are less than 20% Black. They are to the left of the light gray vertical grid line labeled 20% at the bottom. All have impacts

[1] For more on the variables that racial and ethnic features were based on, see the ACS variables B25003A_003E, B25003B_003E, ..., B25003I_003E as described at https://api.census.gov/data/2018/acs/acs5/variables/B25003A_003E.html and similar. For income data, see the variable B25119_003E as described at https://api.census.gov/data/2018/acs/acs5/variables/B25119_003E.html. Replace 2018 with other years in the URL as needed.

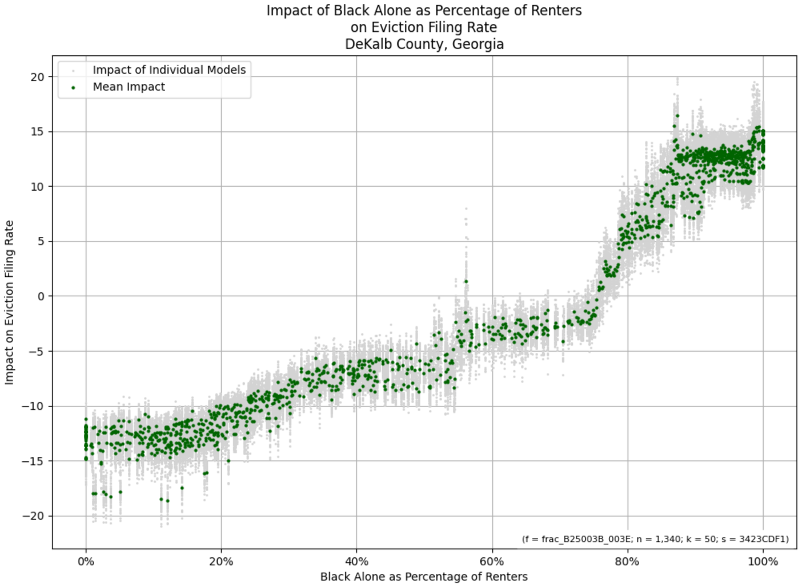


Figure 1: The impact of the percentage of renters who identify as Black on the rate of eviction filing in DeKalb County, Georgia between 2009 and 2018, measured at the census tract-year level. The horizontal axis is the percent of renters who identify as Black alone. It ranges from 0% to 100%. Some census tracts in the county are almost entirely non-Black and others are almost entirely Black. The vertical axis is the impact of the percent of renters who identify as Black on the eviction filing rate, as measured by number of eviction cases filed per 100 renters per year. Each green dot represents one of the census tracts in the county during a single year.

(green dots) below -10. Most are between -10 and -15. A handful are below -15. What this means is that the eviction rate for these tracts are 10 to 15 evictions per 100 renters lower than otherwise comparable neighborhoods (similar income and mix of other racial and ethnic groups). The impact chart makes this plainly visible.

An even more extreme impact can be seen at the right side of the chart, to the right of the vertical grid line labeled 80% at the bottom. These are tracts where more than 80% of renters identify as Black. If we look at the green dots in this region, they almost all have an impact greater than +5 on the vertical axis. Many, including almost all that are above 90% Black, have an impact greater than +10. This means that neighborhoods that are overwhelmingly Black have higher rates of eviction even though the underlying models can also assign impact to income and would do so if income had an impact across different racial demographics.

3.2 A Second Impact Chart: Eviction and Whiteness

The point of an impact chart is to look at the impact of one single variable, regardless of the statistical relationship it may have to other variables in the model. We can look at other variables as well; each gets its own impact chart. For example, Figure 2 is the impact chart for the percentage of renters who are white.

In this impact chart, the green dots for tracts where renters are less than 7% white are all in the positive impact range. This means that neighborhoods that are mostly non-white have higher rates of eviction than others. Again, this is corrected for income. What this impact chart is telling us is that regardless of what groups of non-white renters live in a tract, the simple fact that few of the renters are white is, on its own, sufficient to drive eviction filing rates up.

On the other hand, when neighborhoods are 50% or more white, there is an impact of −2 to −5 points on the eviction rate. Neighborhoods with a lot of white renters have fewer evictions than neighborhoods where renters are predominately not white.

When we looked at the impact chart for Black renters, the trend looked like it might reasonably be explained with a straight line. But the white impact chart has a more complicated behavior. This ability to identify effects with shapes that are not just straight lines or other fixed parametric shapes is one of the key things that distinguishes impact charts from regression analysis. Because we used ML instead of linear or other parametric regression, we are able to identify these nonlinear impacts, even when we don't know anything about the shapes we expect to see before we start the analysis.

3.3 What About the Impact of Income?

The impacts in Figures 1 and 2 demonstrate that the racial demographics of a neighborhood systematically impact eviction rates. But what about income? That is the one feature that we would hypothesize would have an impact, independent of the color line. Figure 3 looks at the impact of the income feature.

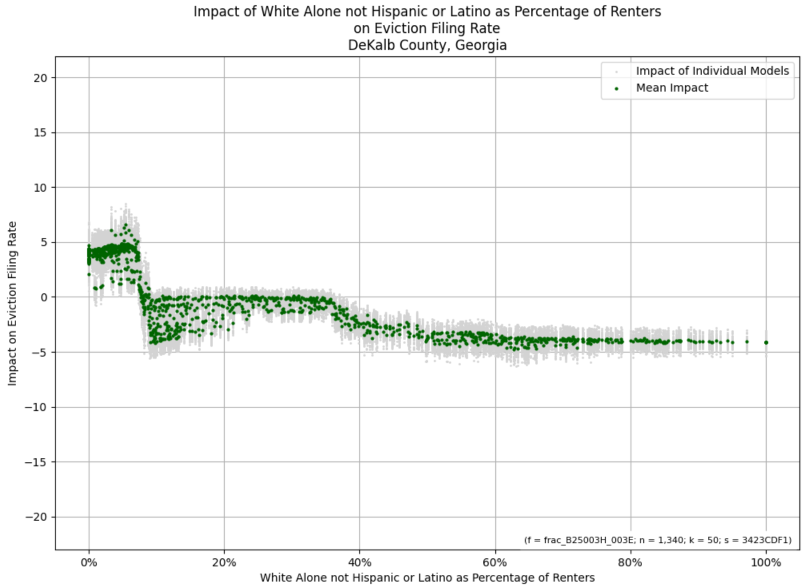


Figure 2: The impact of the percentage of renters who identify as non-Hispanic or Latino white on the rate of eviction filing in DeKalb County, Georgia between 2009 and 2018, measured at the census tract-year level. Compare to Figure 1, which is for the feature representing the percent of renters who identify as Black in those same tracts over the same time period.

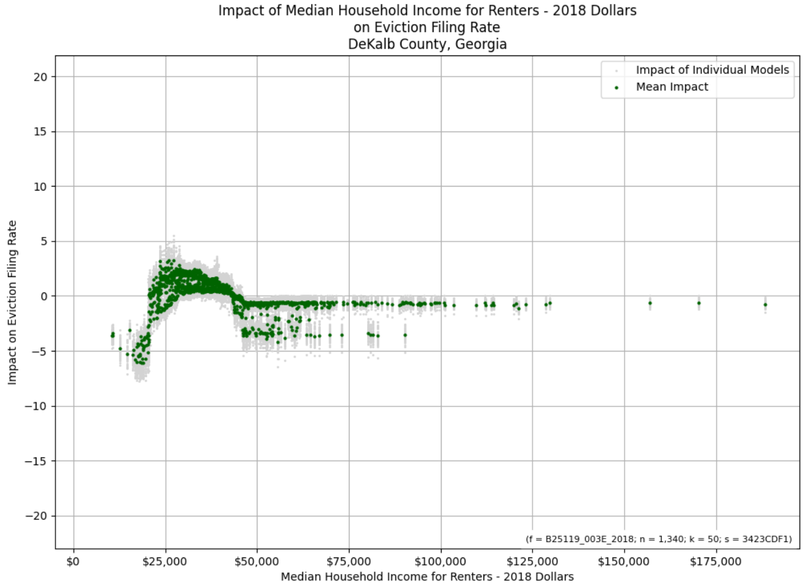


Figure 3: The impact of the median income for renters in a census tract of eviction filing in DeKalb County, Georgia between 2009 and 2018. Compare to Figures 1 and 2, which are for the features representing the percent of renters who identify as Black and non-Hispanic or Latino white in those same tracts over the same time period.

We hypothesized that low-income tracts would have higher eviction rates. But the impact chart tells us something more nuanced. There is a spike between $20,000 and $40,000, but it drops away on either side. For tracts on the right side of the spike, this fits the hypothesis that higher income areas have lower rates of eviction and above a certain income level ($75,000 here) income has essentially no effect on eviction rate. But on the left side, the opposite happens. Lower income tracts have stronger negative impact, including the lowest seen anywhere on the chart, below -5 for some tracts where median renter household income is around $20,000. It is possible that this is the case because low-income renters qualify for programs or housing with less stringent eviction practices than the open market. It is certainly something that points us in that direction for further inquiry.

Taken together, these three impact charts tell two compelling stories. The first is that the color line of systematic racism is alive and well in eviction filings in DeKalb County, Georgia. The impact of fraction of renters in a census tract who are Black (a range of ±15) is three times the size of the impact of the median income of renters in the tract (a range of ±5). The second, less complete but potentially more positive story of the existence of some kind of eviction safety net for very low income households.

3.4 What About the Gray Dots?

The green dots we have been looking at represent the average impact from 50 different machine learning models. Each of these models uses the same code, but is trained on a different random sample of 80% of the data. The grey dots are the impact of the variable in each of the 50 different models. So for each green dot, there are 50 gray dots behind it, in a vertical line.

The reason we plot the gray dots is that ML models can be fickle, giving very different results when trained on data that to humans looks relatively the same [50, 52]. By training many models, we can see if they agree on the impact of each input or if they are all over the place, indicating that our final estimate of the impact (the green dot) isn't that accurate.

The length of the vertical distribution of gray dots gives us a visual idea of how confident we can be on what the corresponding green dot tells us. They are like error bars. If all 50 models closely agree, the gray dots don't extend very far above or below the green dot. But in some cases they do. You can see a few examples in each of the impact charts above. Often, the green dots for these tracts look somewhat out of place relative to the trend of the green dots immediately around them. This means that the models are having trouble agreeing on what is going on in the tract, perhaps because it is heavily influenced by some other feature that we did not include in our model.

We also sometimes see cases, like for incomes between $40,000 and $80,000, where there are what look like two different impact lines of green dots, each surrounded by tight groupings of gray dots. In cases like these the difference between the two groups of green dots may lie in a missing feature that could explain the impact but was not given to the model.

4 HOUSING PRICES, INCOME, RACE, AND ETHNICITY

Our next data set moves us from rental markets to owner-occupied housing. The underlying data set in this case is derived entirely from U.S. Census ACS 5-year data [56] from 2021. Like our first data set, if has features for race and ethnicity, but for those of homeowners rather than renters, and a feature for the median income of homeowners [2]. The data is at the block group level, which is finer granularity than the census tract level and captures data about small neighborhood-sized areas. The target is median home value for owner-occupied homes in the block group. We will look at impact charts for data from the Los Angeles, California area (a CBSA in Census terminology [54]). As was the case with our eviction data, we generated hundreds of additional impact charts for other CBSAs. They can be found at http://rih.datapinions.com/impact.html.

Our goal in generating impact charts with this data is to get a more nuanced understanding of the impact of the racial and ethnic makeup of neighborhoods on home values than previous studies were able to do using regression analysis [44, 47]. We were also able to do this using only aggregate data, thus reducing the possibility of exposing private information about individuals. This does not mean that researchers using impact charts needn't be concerned with privacy and related ethical issues, but it does mean that they can often do their work with data that makes it easier to preserve privacy.

4.1 The Impact of Whiteness

Consider Figure 4, an impact chart which looks at the impact of whiteness on median home values at the block group level in the Los Angeles area CBSA.

The green dots are denser in Figure 4 than we saw in the impact charts in Figures 1, 2, and 3, mainly because there are a lot more of them (7,019 vs. 1,340). There are also a lot more grey dots, (7,019 × 50 = 350,950) but for most of the graph, the band of grey dots is between $50,000 and $75,000 wide from top to bottom. More importantly, for block groups that are below about 24% white, all of the grey dots are below zero.

We built $k = 50$ different models with random 80% subsets of our data, but not a single one of them ever indicated there was a single block group with white population under 24% where the impact of this low level of whiteness was positive. So homes in low-white population neighborhoods have lower value, even when income is included in the models and regardless of what races and ethnicities are present. Non-whiteness is a color line that systematically devalues a neighborhood.

At the other end of the scale, neighborhoods that are overwhelmingly white saw home values an average (green dots) of $75, 000 or more higher than comparable neighborhoods. With the exception of a few outliers, the error bars implied by the grey dots measured this impact as being at least $50, 000 in all cases.

[2] For more on the variables that racial and ethnic features were based on, see the ACS variables B25003A_002E, B25003B_002E, . . . , B25003I_002E as described at https://api.census.gov/data/2021/acs/acs5/variables/B25003A_002E.html and similar. For income data, see the variable B25119_002E as described at https://api.census.gov/data/2021/acs/acs5/variables/B25119_002E.html.

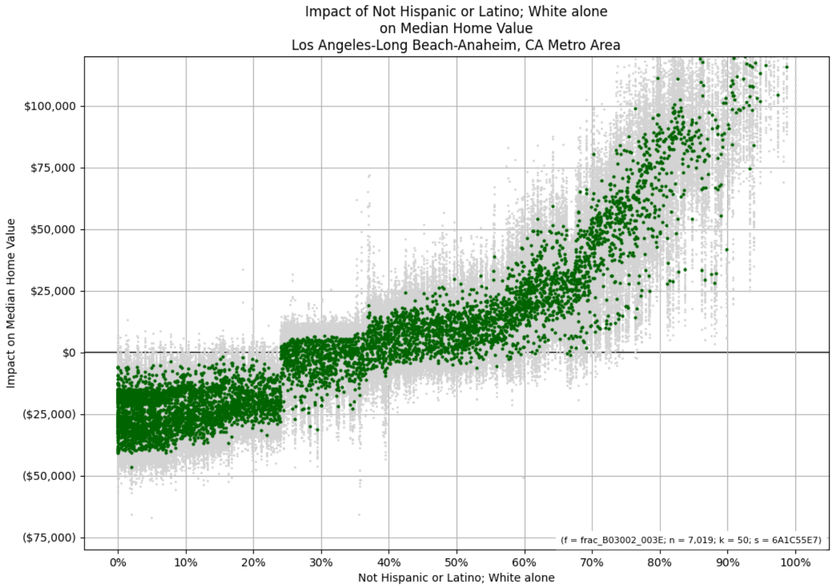


Figure 4: The impact of the fraction of homeowners in a block group who identify as white on median home value in that block group in the Los Angeles-Long Beach-Anaheim, CA CBSA.

4.2 Scatter Plots as an Alternative Visualization

The impact of whiteness on home values is made abundantly clear in the impact chart in Figure 4. It is so clear, in fact, that one might wonder why this kind of impact is more normally publicized through individual stories of specific individuals who were impacted. Rather than systematic studies, stories are regularly published that describe how Black families replaced family photos and other personal effects in their homes with artifacts chosen to imply that the home was owned by a white family. The result was that professional appraisers increased their assessment of the value of the home, sometimes by hundreds of thousands of dollars [22, 30, 31].

Impact charts let us go from individual anecdotes to compelling systematic conclusions using readily available privacy-preserving data. Recall that everything in Figure 4 was derived from publicly available aggregate U.S. Census data.

For the sake of comparison, the most widely used visualization that is designed to help us visualize the impact of a feature on a target is the scatter plot [15]. Figure 5 is a scatter plot of the data set behind the impact chart in Figure 4.

The fundamental problem with Figure 5 is that we are showing the home values $y^{(j)}$ on the vertical axis, not the impact $I(x_i^{(j)})$ of the single whiteness variable as we did in Figure 4. The scattering across $2,000,000 on the vertical axis obscures the effect of the single variable on the horizontal axis with the effects of all the other variables. We could add a regression line, or a regression curve if we fit a non-linear model, but even if we did so, the impact would not be explicitly shown. At best we could say something about impact based on the slope m of the line, like, "for every 10 points the percentage of homeowners who are white goes up, home values tend to go up by $\$0.1m$."

Alternatively, we could partition the block groups into those with high white (or Black, etc…) population, say 95% or more, and low white (or Black, etc…), say 50% or less. We could then use classical statistical techniques to evaluate the difference between the distribution of median home values in the different groups, correcting for median income [44]. We think that impact charts produce a richer and more compelling story than this approach.

Some existing approaches to understand bias using scatter plots [33] focus on helping users understand the relationship and impact of interacting features, which is complementary to the insights on the impact of individual features that impact charts offer.

4.3 Regression Analysis and Measures of Error

Another consideration, which is often useful in constructing impact charts, and which was critical in the construction of the impact charts for the housing value data set, is the choice of error function. All ML models try to minimize error on the data they are trained on, but how we decide to measure error can significantly influence their behavior. In regression models, Mean Squared Error (MSE) is widely used both because it often gives good results and because it is computationally efficient in many cases [20].

Among the assumptions that regression analysis makes is that of homoscedasticity [20], which means that the variance in the observations of the target, median housing price in our example, is the same regardless of the value of the variable’s value.

But it’s hard to imagine this condition holds in our case. Suppose, for example, that for homes that are actually worth $200,000, our input data captures them to within an accuracy of ±$20,000, or ±10%. What about homes worth $1,000,000? We don’t expect to be able to also measure their value to within ±$20,000, which would be ±2%.

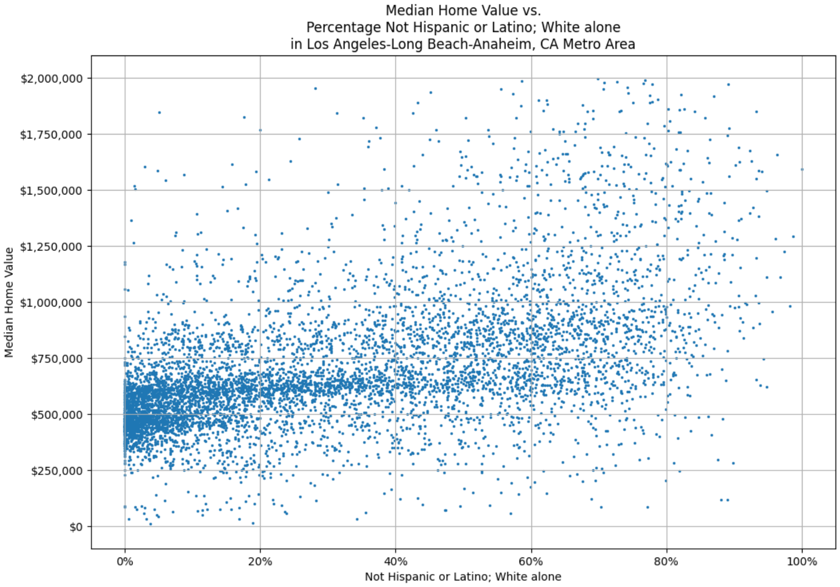


Figure 5: A scatter plot of median home value (vertical axis) vs. the fraction of homeowners in a block group who identify as white (horizontal axis) at the block group level in the Los Angeles-Long Beach-Anaheim, CA CBSA.

Instead, we might measure them to within the same ±10% as the lower priced homes, which would be ±$100,000. This is very much a classic case of heteroscedasticity, the opposite of homoscedasticity.

Similarly, if we built a model to predict median home prices, we would not expect its mean error to be the same in dollar terms on homes worth around $200,000 as homes worth around $1,000,000. At best, we might hope for the same mean relative error. For example, the model might have a mean absolute error of 20% of the actual home value regardless of whether it was $200,000 or $1,000,000. In the former the error would be ±$40,000 and in the latter it would be ±$200,000. This measure of error is called Mean Absolute Percent Error (MAPE).

Note that we use the mean of the absolute value of the error percentage rather than the mean of the percentage error. Otherwise a large positive error in half the cases (say an 80% overestimate) could be counteracted by a large negative error (an 80% underestimate) in the other half. The mean would be 0%, making this poor model look like an excellent one.

Optimizing model construction to minimize MAPE tends, especially in cases of heteroscedasticity, to produce very different models than optimizing for minimum MSE. If we optimize for MSE in cases like ours, the influence of the high-priced end of the market can overwhelm the influence of the lower priced end because absolute errors are larger there. Squaring compounds this effect. So in effect, we try really hard to be good at predicting prices at the high end of the market even if that means, in MAPE terms, we end of with a pretty bad model at the low end of the market. We'd like to have a good model at both ends of the market, so we'd prefer to optimize our model for MAPE rather than MSE.

XGBoost [17], the open-source implementation of gradient boosted trees that we used behind the impact charts herein, supports MAPE. We used it for all of the housing price impact charts we generated. For the eviction rate, we use Mean Average Error (MAE), since the target was already expressed as a rate of eviction per 100 renters.

5 ACCESS TO SUPERMARKETS, INCOME, RACE, AND ETHNICITY

Our final motivating example uses 2019 data from the U.S. Department of Agriculture (USDA) Food Access Research Atlas (FARA) [57] along with 2019 U.S. Census ACS data [56] to assess the impact of race, ethnicity, income and access to a vehicle on access to local supermarkets that carry fresh healthy food. Areas without such access are commonly referred to as *food deserts*. The target variable, as defined by the FARA data set is called `lapophalfshare`. It represents the fraction of the population of a census tract that lacks access to a supermarket within a distance of 1/2 mile of their home. This is expressed on a scale from 0, meaning everyone in the tract has a supermarket within 1/2 mile of their residence, to 100, meaning nobody does. 30 would means 30% of the tract's population does not have access to a supermarket within 1/2 mile but the other 70% does.

The data sets we constructed are at the census tract level for all census tracts within the most populated census places (cities and towns, typically) across the United States. It is sparser than the the previous two data sets, because the FARA data is itself more sparse. We will concentrate on the New York City data set herein.

As was the case for the two previous data sets, the impact charts we generated for food access in New York City indicated that racial and ethnic disparities exist, even when the model corrects for income. Some of these are shown in Figures 6(a)-(d). We could certainly dig into hypotheses about what these impact charts tell us, remembering that in a hypothetical post-racial society the green dots would form a horizontal line at an impact of zero. But having done that in Sections 3 and 4, we will instead look at another important and useful feature of impact charts: even at thumbnail scale, as in Figure 6, impact charts enable us to quickly, easily, and visually evaluate our *a priori* hypotheses.

Figure 6(e) shows the impact of median income. This is where our initial hypothesis that lower income people would be more affected by food deserts failed. All of the impacts are within ±5 percentage points, as most were for racial and ethnic features, but for household incomes below $75,000, the impacts are all negative, meaning that those census tracts have more access to grocery stores as a result of their lower income, not less.

This counterintuitive finding, however, makes more sense in light of Figure 6(f), which shows the impact of access to a vehicle, which we also included as a factor in our model. The range of impacts here went well beyond the ±5 we saw in Figures 6(a)-(e). For tracts where fewer than 40% of households have access to a vehicle, the impact hovers around -10. For tracts where more than 90% of households have access to vehicles the impact extends +7.

So what we see here is that although there are racial, ethnic, and income impacts, the feature with the biggest impact is vehicle access. This does not imply causality. More likely it shows that there are both high and low income tracts in New York City where there are high and low numbers of residents with access to vehicles. Where vehicle access is low, access to nearby supermarkets is high, and where vehicle access is high, access to nearby supermarkets is lower, regardless of income level, race, or ethnicity. Though validating it is beyond the scope of this work, we hypothesize that car-friendly development results in residents needing to obtain cars, which cycles back to encourage more car-dependent development over the course of many years.

In summary, in this application impact charts made it easy to invalidate our initial hypothesis that income would have a large impact on supermarket access. Instead, vehicle access is the feature that has the greatest impact. Impact charts also showed us the full nonlinear nature of the impacts across the spectrum of feature values without any need to constrain our exploration to specific parametric forms that the impacts might take on.

6 SUMMARY AND FUTURE WORK

We introduced impact charts and applied them to the problem of visualizing the impact of the color line in two housing-related applications. In the eviction analysis in Section 3, we showed using impact charts that the impact of Blackness was 3 times that of the impact of income on eviction rates. In the home valuation analysis in Section 4, we illustrated the impact of whiteness, introduced the use of MAPE as an error metric, and demonstrated the degree to which impact charts can clearly demonstrate impact when scatter plots and regression analysis do not. We also discussed impact charts relative to regression analysis using constructed subpopulations. Finally, in Section 5, we used USDA FARA data to illustrate how by visually scanning a collection impact charts, we can refine hypotheses.

Our current work is empirical, but we are also working on a theoretical analysis to characterize the impacts that impact charts expose vs. any they may miss by looking at their behavior on synthetic data produced by parametric and causal models.

Finally, we note that the case studies here each involved a data set representing a single geographic area (a county, a CBSA, and a city). Because we chose ML techniques that produce accurate models and also enable us to compute impact charts efficiently (as opposed to more general techniques that consume resources exponential in the number of features [37]) we have been able to generate thousands of impact charts for different geographies[3].

6.1 Code and Data Availability

Having introduced impact charts and shown their use on three data sets, we believe we have only scratched the surface. In order to enable further development and make it easy for researchers to bring their own data sets and generate impact charts from them, we have made a reference implementation of impact charts available at https://github.com/impactchart/impactchart. This repository is the recommended starting point and holds the most current version of the impact chart library. We hope that in using our code, researchers will both find additional insights like those discussed herein and identify limitations of either the approach or the software implementing it so that we can make appropriate improvements.

For completeness, the original code that constructed the data sets and produced the impact charts in Section 3 has been preserved at https://github.com/datapinions/evldata and https://github.com/datapinions/evlcharts respectively. The original code for Section 4 has been preserved at https://github.com/datapinions/rihdata and https://github.com/datapinions/rihcharts. The original code for Section 5 has been preserved at https://github.com/vengroff/faradata and https://github.com/vengroff/faracharts.

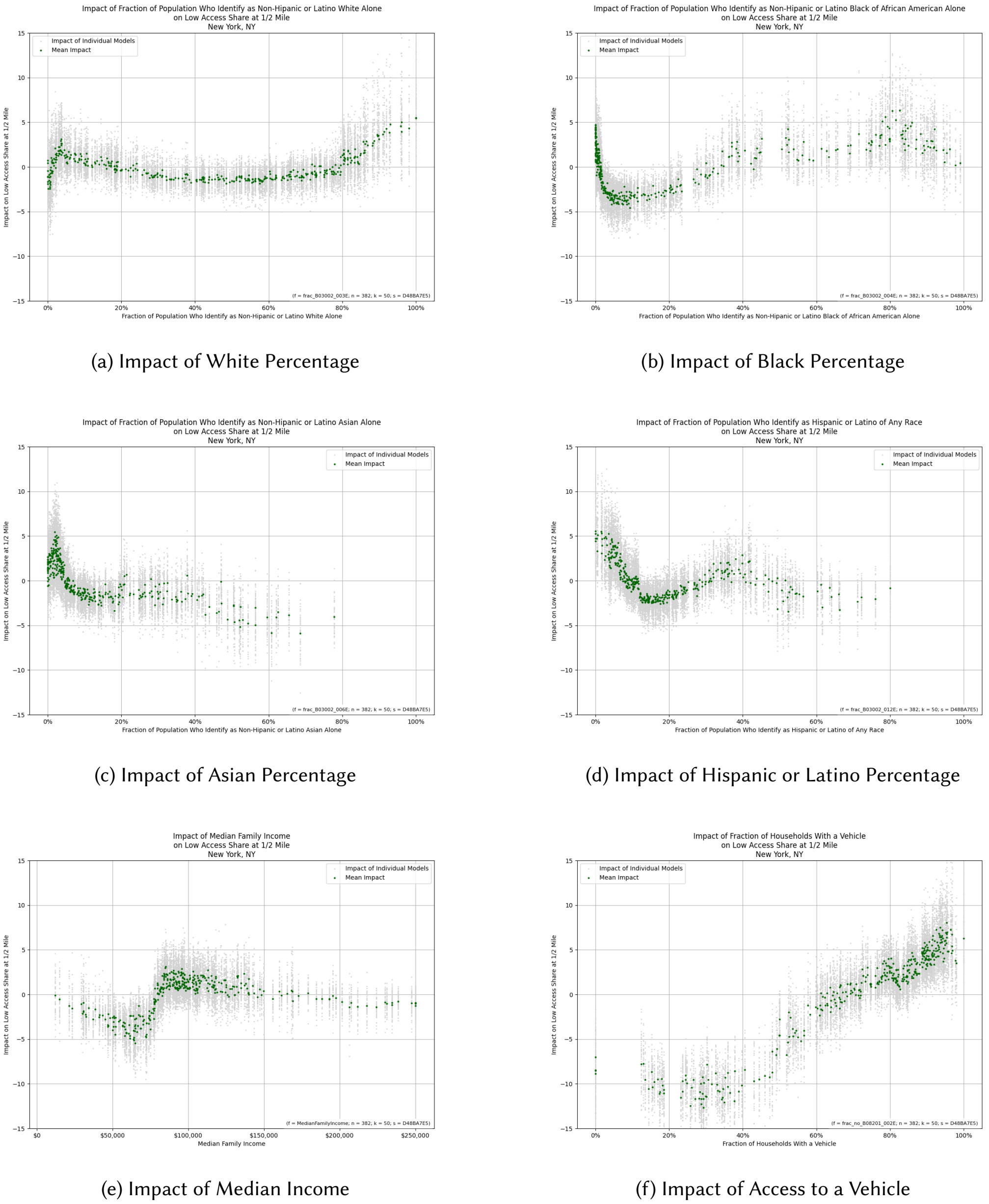
6.2 Future Work

In future work, we believe it will be possible to characterize the impacts we see across hundreds of geographies based on political features of the geographies themselves. For example, we will be able to say that the color line creates greater disparities in cities in one state or region than in another, or in cities that have implemented a particular program vs. those that have not. This, we believe, will provide policy makers and citizens alike with powerful information on how to address color line impacts in their communities.

We would also like to look into integrating the generation of impact charts into the lifecycle of data sets used to train ML models [29] and summary reports on such data sets, such as data set nutrition labels [28, 51]. Impact charts could also be integrated into existing systems for visually assessing bias in data sets and ML systems [5, 8, 14, 41, 59, 60, 62].

In engaging practitioners, whether ML practitioners, data engineers, data scientists, social scientists, or policymakers, it will be critical that we understand how they use and interpret impact charts. In studying this we intend to build on existing work on how

[3] For reference, a full run that generates almost 5,000 eviction impact charts runs in a matter of hours on a single laptop.



(a) Impact of White Percentage

(b) Impact of Black Percentage

(c) Impact of Asian Percentage

(d) Impact of Hispanic or Latino Percentage

(e) Impact of Median Income

(f) Impact of Access to a Vehicle

Figure 6: Impact charts showing the impact of the percentage of residents identifying as members of certain racial and ethnic groups ((a)-(d)), the impact of median income (e), and the fraction of residents with access to a vehicle (e) on the fraction of residents who lack of access to a supermarket within half a mile of home.

practitioners work with anti-bias tools and how tools can be constructed to be more useful to practitioners [3, 7, 10, 16, 19, 32, 38].

Coming full circle to Douglass' color line [21] as we introduced it at the outset, we hope that the use of impact charts can contribute in some small way to exposing the color line as it exists today and inspiring solutions that lead to its elimination.

7 ETHICS AND SOCIAL IMPACT

When working at the intersection of data and structural racism, we believe there are important ethical and social concerns that researchers must address.

7.1 Ethical Concerns and Mitigation

Whenever we deal with data including race, ethnicity, gender identity or any of many other sensitive variables or proxy variables, we have to be concerned about privacy. We were while developing impact charts. A common approach to managing privacy involves researchers gaining access to sensitive data in a highly controlled environment, for example via https://www.researchdatagov.org/, doing their analysis, and carefully structuring the results they publish so as to preserve the privacy or the (possibly unwilling or unknowing) participants whose data they studied. We deliberately did not take this approach, and instead challenged ourselves to develop an analysis technique that could effectively find structural impacts using aggregate data. In particular, most of our data comes from the U.S. Census, which has a robust privacy program [2, 55]. We cannot claim that aggregate data can never be manipulated in ways that expose individuals [42], but we can say that it can be analyzed with substantially less risk in most circumstances.

7.2 Author Positioning and Reflection

Having been trained as a Computer Scientist in an era when ethics were not a part of most undergraduate or graduate CS curricula, the author has historically been biased towards technology-centric solutions to problems. And without question impact charts are technology-centric.

However, like all technologies, we must come to view them as tools and evaluate the extent to which their uses are concentrated in applications that are socially useful or socially damaging. We believe that they are the former and that this has been demonstrated by the example in Sections 3, 4, and 5.

That having been said, impact charts are only one small tool. Pointing out adverse impacts does not eliminate them. At best, it is diagnostic [1]. But as such it can motivate those generating the charts to become a part of a larger movement, to engage with, to listen to, and to be led by those often less privileged individuals who bear the brunt of the impacts the charts highlight.

7.3 Adverse Impact

Impact charts, no less so than any other data-driven technological artifacts, can be useless or even dangerous when fed the wrong data. While at its core the technique is designed to identify and expose biased impacts in society as reflected in data, there is nothing stopping motivated actors from repurposing the vocabulary of impact charts to tell the stories they want told on top of data they have generated, manipulated, or commissioned to meet their own needs.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 252–260. https://doi.org/10.1145/3351095.3372871

[2] John M. Abowd. 2018. Protecting the Confidentiality of America's Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau. https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html

[3] Jacqui Ayling and Adriane Chapman. 2022. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics* 2 (2022), 405–429. https://doi.org/10.1007/s43681-021-00084-x

[4] Whitney Battle-Baptiste and Britt Rusert (Eds.). 2018. *W. E. B. Du Bois's Data Portraits: Visualizing Black America.* Princeton Architectural Press, Princeton, NJ. 144 pages.

[5] Emma Beauxis-Aussalet, Michael Behrisch, Rita Borgo, Duen Horng Chau, Christopher Collins, David Ebert, Mennatallah El-Assady, Alex Endert, Daniel A. Keim, Jörn Kohlhammer, Daniela Oelke, Jaakko Peltonen, Maria Riveiro, Tobias Schreck, Hendrik Strobelt, and Jarke J. van Wijk. 2021. The Role of Interactive Visualization in Fostering Trust in AI. *IEEE Computer Graphics and Applications* 41, 6 (2021), 7–12. https://doi.org/10.1109/MCG.2021.3107875

[6] Ruha Benjamin. 2019. *Race After Technology.* Polity, Cambridge, UK. 172 pages.

[7] Glen Berman, Nitesh Goyal, and Michael Madaio. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *CHI Conference on Human Factors in Computing Systems (CHI '24)* (Honolulu, HI). ACM, New York, NY, USA.

[8] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[9] Christopher Bishop. 2006. *Pattern Recognition and Machine Learning.* Springer, Berlin.

[10] Emily Black, Rakshit Naidu, Rayid Ghani, Kit Rodolfa, Daniel Ho, and Hoda Heidari. 2023. Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools. In *EAAMO '23: Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA). ACM, New York, NY, USA, 1–11. https://doi.org/10.1145/3617694.3623259

[11] Leo Breiman. 1996. Bagging Predictors. *Machine Learning* 24, 2 (1996), 123–140.

[12] Joy Buolamwini. 2023. *Unmasking AI: My Mission to Protect What is Human in a World of Machines.* Penguin Random House, New York, NY. 336 pages.

[13] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[14] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST).* IEEE, New York, NY, 46–56. https://doi.org/10.1109/VAST47406.2019.8986948

[15] John M. Chambers, William S. Cleveland, Beat Kleiner, and Paul A. Tukey. 1983. *Graphical Methods for Data Analysis.* Wadsworth, Belmont, CA. 395 pages.

[16] Jiyoo Chang and Christine Custis. 2022. Understanding Implementation Challenges in Machine Learning Documentation. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA) *(EAAMO '22).* ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3551624.3555301

[17] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16).* ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[18] Gerda Claeskens and Nils Lid Hjort. 2008. *Model Selection and Model Averaging.* Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511790485

[19] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22).* ACM, New York, NY, USA, 473–484. https://doi.org/10.1145/3531146.3533113

[20] Annette J. Dobson and Adrian G. Barnett. 2018. *An Introduction to Generalized Linear Models* (4th ed.). Chapman and Hall/CRC, Boca Raton, FL.

[21] Frederick Douglass. 1881. The Color Line. *The North American Review* 132, 295 (June 1881), 567–577.

[22] Brook Endale. 2021. Home appraisal increased by almost $100,000 after Black family hid their race. *USA Today* (13 September 2021). https://www.usatoday.com/story/money/nation-now/2021/09/13/home-appraisal-grew-almost-100-000-after-black-family-hid-their-race/8316884002/

[23] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press, New York, NY. 272 pages.

[24] Joe Feagin. 2006. *Systemic Racism: A Theory of Oppression.* Routledge, Milton Park, Abingdon-on-Thames, Oxfordshire, UK. 386 pages.

[25] Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 5 (2001), 1189 – 1232. https://doi.org/10.1214/aos/1013203451

[26] Ashley Gromis, Ian Fellows, James R. Hendrickson, Lavar Edmonds, Lillian Leung, Adam Porton, and Matthew Desmond. 2022. Estimating Eviction Prevalence across the United States. https://data-downloads.evictionlab.org/#estimating-eviction-prevalance-across-us/

[27] Peter Hepburn, Renee Louis, and Matthew Desmond. 2020. Racial and Gender Disparities among Evicted Americans. *Sociological Science* 7, 27 (2020), 649–662. https://doi.org/10.15195/v7.a27

[28] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677 [cs.DB]

[29] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

[30] Joe Johns, Laura Robinson, and Nicole Chavez. 2021. A Black couple had a White friend show their home and its appraisal rose by nearly half a million dollars. *CNN* (9 December 2021). https://www.cnn.com/2021/12/09/business/black-homeowners-appraisal-discrimination-lawsuit/index.html

[31] Debra Kamin. 2022. Home Appraised With a Black Owner: $472,000. With a White Owner: $750,000. *The New York Times* (18 August 2022). https://www.nytimes.com/2022/08/18/realestate/housing-discrimination-maryland.html

[32] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. ACM, New York, NY, USA, 702–714. https://doi.org/10.1145/3531146.3533135

[33] Hannah Kim, Jaegul Choo, Haesun Park, and Alex Endert. 2016. InterAxis: Steering Scatterplot Axes via Observation-Level Interaction. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 131–140. https://doi.org/10.1109/TVCG.2015.2467615

[34] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Beijing, China, 150–158. https://doi.org/10.1145/2339530.2339556

[35] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Chicago, IL, USA, 623–631. https://doi.org/10.1145/2487575.2487579

[36] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2 (2020), 56–67. https://www.nature.com/articles/s42256-019-0138-9

[37] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[38] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 52 (apr 2022), 26 pages. https://doi.org/10.1145/3512899

[39] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Benjamin Feuer, Chinmay Hegde, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. 2023. When Do Neural Nets Outperform Boosted Trees on Tabular Data? arXiv:2305.02997 [cs.LG]

[40] Khalil Gibran Muhammad. 2019. *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America*. Harvard University Press, Cambridge, MA. 416 pages.

[41] David Munechika, Zijie J. Wang, Jack Reidy, Josh Rubin, Krishna Gade, Krishnaram Kenthapadi, and Duen Horng Chau. 2022. Visual Auditor: Interactive Visualization for Detection and Summarization of Model Biases. In *2022 IEEE Visualization and Visual Analytics (VIS)*. IEEE, New York, NY, 45–49. https://doi.org/10.1109/VIS54862.2022.00018

[42] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* (Oakland, CA). IEEE, New York, NY, 111–125. https://doi.org/10.1109/SP.2008.33

[43] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, New York, NY. 248 pages.

[44] Andre M. Perry, Jonathan Rothwell, , and David Harshbarger. 2018. *The devaluation of assets in Black neighborhoods: The case of residential property*. Brookings. https://www.brookings.edu/articles/devaluation-of-assets-in-black-neighborhoods/

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG]

[46] Richard Rothstein. 2017. *The Color of Law: A Forgotten History of How Our Government Segregated America*. Norton, New York, NY. 368 pages.

[47] Jonathan Rothwell and Andre M. Perry. 2021. *Biased appraisals and the devaluation of housing in Black neighborhoods*. Brookings. https://www.brookings.edu/articles/biased-appraisals-and-the-devaluation-of-housing-in-black-neighborhoods/

[48] Lloyd S. Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

[49] Galit Shmueli. 2010. To Explain or to Predict? *Statist. Sci.* 25, 3 (2010), 289–310. https://doi.org/10.1214/10-STS330

[50] M. Stone. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* 36, 2 (1974), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

[51] Julia Stoyanovich and Bill Howe. 2019. Nutritional Labels for Data and Models. *Bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering* 42, 3 (2019), 13–23. http://sites.computer.org/debull/A19sept/p13.pdf

[52] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation* 23, 5 (2019), 828–841. https://doi.org/10.1109/TEVC.2019.2890858

[53] Keeanga-Yamahtta Taylor. 2021. *Race for Profit: How Banks and the Real Estate Industry Undermined Black Homeownership*. UNC Press, Chapel Hill, NC. 376 pages.

[54] U.S. Census Bureau. 2010. US Census Bureau Geographic Entities and Concepts. https://www.census.gov/content/dam/Census/data/developers/geoareaconcepts.pdf

[55] U.S. Census Bureau. 2019. A History of Census Privacy Protections. https://www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html

[56] U.S. Census Bureau. 2023. American Community Survey Data. https://www.census.gov/programs-surveys/acs/data.html

[57] U.S. Department of Agriculture. 2023. Food Access Research Atlas. https://www.ers.usda.gov/data-products/food-access-research-atlas/

[58] Ramon Vilarino and Renato Vicente. 2021. An experiment on the mechanisms of racial bias in ML-based credit scoring in Brazil. arXiv:2011.09865 [cs.CY]

[59] Emily Wall, Leslie M. Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Phoenix, AZ). IEEE, New York, NY, 104–115. https://doi.org/10.1109/VAST.2017.8585669

[60] Emily Wall, Arpit Narechania, Adam Coscia, Jamal Paden, and Alex Endert. 2022. Left, Right, and Gender: Exploring Interaction Traces to Mitigate Human Biases. *IEEE Trans. Vis. Comput. Graph.* 28, 1 (2022), 966–975. https://doi.org/10.1109/TVCG.2021.3114862

[61] Maksymilian Wojtas and Ke Chen. 2020. Feature Importance Ranking for Deep Learning. arXiv:2010.08973 [cs.LG]

[62] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M. Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376447