Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States

Timnit Gebru[a,1], Jonathan Krause[a], Yilun Wang[a], Duyun Chen[a], Jia Deng[b], Erez Lieberman Aiden[c,d,e], and Li Fei-Fei[a]

[a]Artificial Intelligence Laboratory, Computer Science Department, Stanford University, Stanford, CA 94305; [b]Vision and Learning Laboratory, Computer Science and Engineering Department, University of Michigan, Ann Arbor, MI 48109; [c]The Center for Genome Architecture, Department of Genetics, Baylor College of Medicine, Houston, TX 77030; [d]Department of Computer Science, Rice University, Houston, TX 77005; and [e]The Center for Genome Architecture, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved October 16, 2017 (received for review January 4, 2017)

The United States spends more than $250 million each year on the American Community Survey (ACS), a labor-intensive door-to-door study that measures statistics relating to race, gender, education, occupation, unemployment, and other demographic factors. Although a comprehensive source of data, the lag between demographic changes and their appearance in the ACS can exceed several years. As digital imagery becomes ubiquitous and machine vision techniques improve, automated data analysis may become an increasingly practical supplement to the ACS. Here, we present a method that estimates socioeconomic characteristics of regions spanning 200 US cities by using 50 million images of street scenes gathered with Google Street View cars. Using deep learning-based computer vision techniques, we determined the make, model, and year of all motor vehicles encountered in particular neighborhoods. Data from this census of motor vehicles, which enumerated 22 million automobiles in total (8% of all automobiles in the United States), were used to accurately estimate income, race, education, and voting patterns at the zip code and precinct level. (The average US precinct contains ∼1,000 people.) The resulting associations are surprisingly simple and powerful. For instance, if the number of sedans encountered during a drive through a city is higher than the number of pickup trucks, the city is likely to vote for a Democrat during the next presidential election (88% chance); otherwise, it is likely to vote Republican (82%). Our results suggest that automated systems for monitoring demographics may effectively complement labor-intensive approaches, with the potential to measure demographics with fine spatial resolution, in close to real time.

computer vision | deep learning | social analysis | demography

For thousands of years, rulers and policymakers have surveyed national populations to collect demographic statistics. In the United States, the most detailed such study is the American Community Survey (ACS), which is performed by the US Census Bureau at a cost of $250 million per year (1). Each year, ACS reports demographic results for all cities and counties with a population of 65,000 or more (2). However, due to the labor-intensive data-gathering process, smaller regions are interrogated less frequently, and data for geographical areas with less than 65,000 inhabitants are typically presented with a lag of ~ 2.5 y. Although the ACS represents a vast improvement over the earlier, decennial census (3), this lag can nonetheless impede effective policymaking. Thus, the development of complementary approaches would be desirable.

In recent years, computational methods have emerged as a promising tool for tackling difficult problems in social science. For instance, Antenucci et al. (4) have predicted unemployment rates from Twitter; Michel et al. (5) have analyzed culture using large quantities of text from books; and Blumenstock et al. (6) used mobile phone metadata to predict poverty rates in Rwanda. These results suggest that socioeconomic studies, too, might be facilitated by computational methods, with the ultimate potential of analyzing demographic trends in great detail, in real time, and at a fraction of the cost.

Recently, Naik et al. (7) used publicly available imagery to quantify people's subjective perceptions of a neighborhood's physical appearance. They then showed that changes in these perceptions correlate with changes in socioeconomic variables (8). Our work explores a related theme: whether socioeconomic statistics can be inferred from objective characteristics of images from a neighborhood.

Here, we show that it is possible to determine socioeconomic statistics and political preferences in the US population by combining publicly available data with machine-learning methods. Our procedure, designed to build upon and complement the ACS, uses labor-intensive survey data for a handful of cities to train a model that can create nationwide demographic estimates. This approach allows for estimation of demographic variables with high spatial resolution and reduced lag time.

Specifically, we analyze 50 million images taken by Google Street View cars as they drove through 200 cities, neighborhood-by-neighborhood and street-by-street. In Google Street View images, only the exteriors of houses, landscaping, and vehicles on the street can be observed. Of these objects, vehicles are among the most personalized expressions of American culture: Over 90% of American households own a motor vehicle (9), and their choice of automobile is influenced by disparate demographic factors including household needs, personal preferences, and economic wherewithal (10). (Note that, in principle, other factors such as spacing between houses, number of stories, and extent of shrubbery could also be integrated into such models.) Such street scenes are a natural data type to explore: They already cover

Significance

We show that socioeconomic attributes such as income, race, education, and voting patterns can be inferred from cars detected in Google Street View images using deep learning. Our model works by discovering associations between cars and people. For example, if the number of sedans in a city is higher than the number of pickup trucks, that city is likely to vote for a Democrat in the next presidential election (88% chance); if not, then the city is likely to vote for a Republican (82% chance).

Author contributions: T.G., J.K., J.D., E.L.A., and L.F.-F. designed research; T.G., J.K., Y.W., D.C., J.D., E.L.A., and L.F.-F. performed research; T.G. and J.K. contributed new reagents/analytic tools; T.G., J.K., Y.W., D.C., J.D., E.L.A., and L.F.-F. analyzed data; and T.G., J.K., E.L.A., and L.F.-F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: tgebru@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1700035114/-/DCSupplemental.

much of the United States, and the emergence of self-driving cars will bring about a large increase in the frequency with which different locations are sampled.

We demonstrate that, by deploying a machine vision framework based on deep learning—specifically, Convolutional Neural Networks (CNN)—it is possible to not only recognize vehicles in a complex street scene but also to reliably determine a wide range of vehicle characteristics, including make, model, and year. Whereas many challenging tasks in machine vision (such as photo tagging) are easy for humans, the fine-grained object recognition task we perform here is one that few people could accomplish for even a handful of images. Differences between cars can be imperceptible to an untrained person; for instance, some car models can have subtle changes in tail lights (e.g., 2007 Honda Accord vs. 2008 Honda Accord) or grilles (e.g., 2001 Ford F-150 Supercrew LL vs. 2011 Ford F-150 Supercrew SVT). Nevertheless, our system is able to classify automobiles into one of 2,657 categories, taking 0.2 s per vehicle image to do so. While it classified the automobiles in 50 million images in 2 wk, a human expert, assuming 10 s per image, would take more than 15 y to perform the same task. Using the classified motor vehicles in each neighborhood, we infer a wide range of demographic statistics, socioeconomic attributes, and political preferences of its residents.

In the first step of our analysis, we collected 50 million Google Street View images from 3,068 zip codes and 39,286 voting precincts spanning 200 US cities (Fig. 1). Using these images and annotated photos of cars, our object recognition algorithm [a "Deformable Part Model" (DPM) (11)] learned to automatically localize motor vehicles on the street (12) (see *Materials and Methods*). This model took advantage of a gold-standard dataset we generated by asking humans (both laypeople, recruited using Amazon Mechanical Turk, and car experts recruited through Craigslist) to identify cars in Google Street View scenes.

We successfully detected 22 million distinct vehicles, comprising 32% of all of the vehicles in the 200 cities we studied and 8% of all vehicles in the United States. After localizing each vehicle, we deployed CNN (13, 14), the most successful deep learning algorithm to date for object classification, to determine the make, model, body type, and year of each vehicle (Fig. 1). Using our human-annotated gold standard images, we trained the CNN to distinguish between different types of cars. Specifically, we were able to classify each vehicle into one of 2,657 fine-grained categories, which form a nearly exhaustive list of all visually distinct automobiles sold in the United States since 1990 (Fig. 1). For instance, our models accurately identified cars (identifying 95% of such vehicles in the test data), vans (83%), minivans (91%), SUVs (86%), and pickup trucks (82%). See *SI Appendix*, Fig. S1.

Using the resulting motor vehicle data, we estimate demographic statistics and voter preferences as follows. For each geographical region we examined (city, zip code, or precinct), we count the number of vehicles of each make and model that were identified in images from that region. We also include additional features such as aggregate counts for various vehicle types (trucks, vans, SUVs, etc.), the average price and fuel efficiency, and the overall density of vehicles in the region (see *Materials and Methods*).

We then partitioned our dataset, by county, into two subsets (Fig. 2). The first is a "training set," comprising all regions that lie mostly in a county whose name starts with "A," "B," or "C" (such as Ada County, Baldwin County, Cabarrus County, etc.). This training set encompasses 35 of the 200 cities, $\sim$ 15% of the zip codes, and $\sim$ 12% of the precincts in our data. The second is a "test set," comprising all regions in counties starting with the letters "D" through "Z" (such as Dakota County, Maricopa County, Yolo County). We used the test set to evaluate the model that resulted from the training process.

Using ACS and presidential election voting data for regions in our training set, we train a logistic regression model to estimate race and education levels and a ridge regression model to estimate income and voter preferences on the basis of the collection of vehicles seen in a region. This simple linear model is sufficient to identify positive and negative associations between

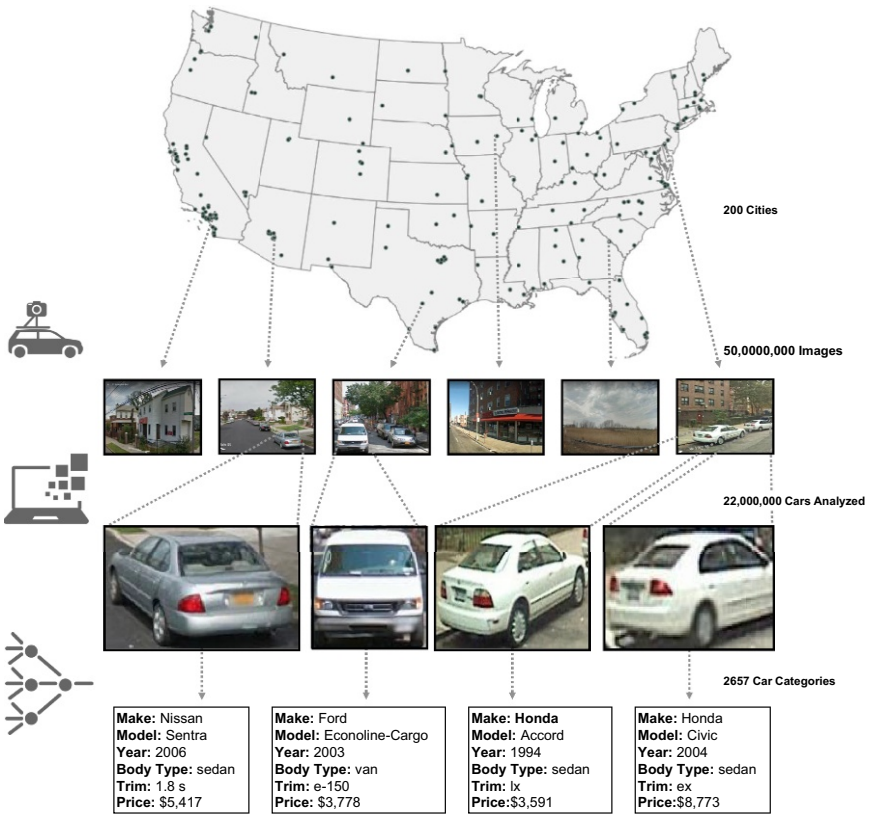

Fig. 1. We perform a vehicular census of 200 cities in the United States using 50 million Google Street View images. In each image, we detect cars with computer vision algorithms based on DPM and count an estimated 22 million cars. We then use CNN to categorize the detected vehicles into one of 2,657 classes of cars. For each type of car, we have metadata such as the make, model, year, body type, and price of the car in 2012. Images courtesy of Google Maps/Google Earth.

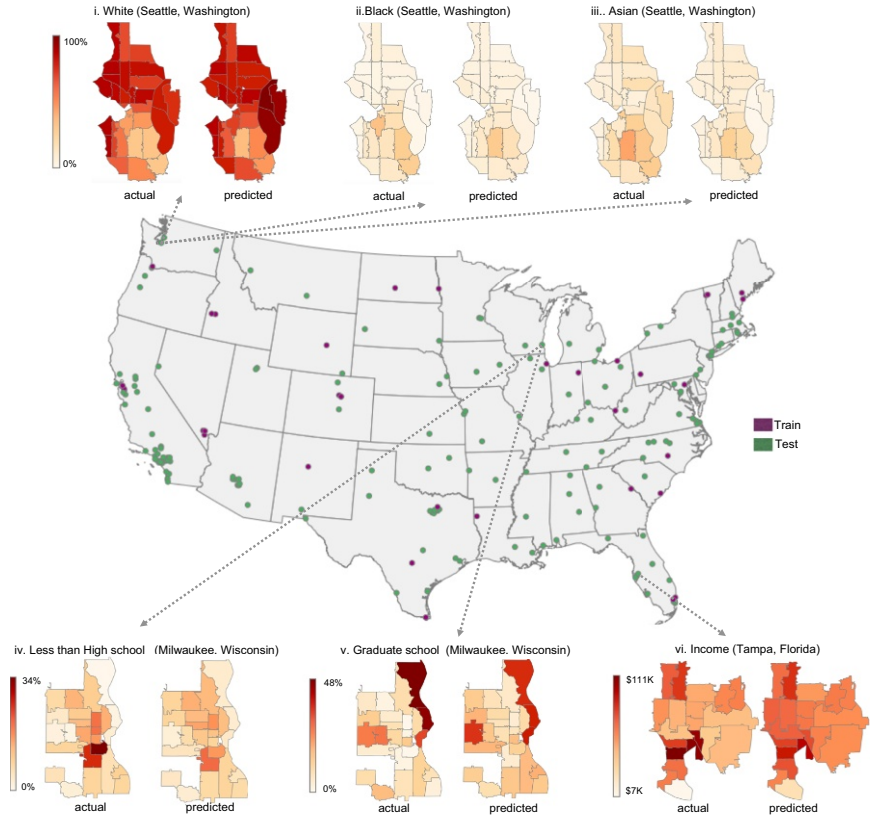

Fig. 2. We use all of the cities in counties starting with A, B, and C (shown in purple on the map) to train a model estimating socioeconomic data from car attributes. Using this model, we estimate demographic variables at the zip code level for all of the cities shown in green. We show actual vs. predicted maps for the percentage of Black, Asian, and White people in Seattle, WA (*i–iii*); the percentage of people with less than a high school degree in Milwaukee, WI (*iv*); and the percentage of people with graduate degrees in Milwaukee, WI (*v*). (*vi*) Maps the median household income in Tampa, FL. The ground truth values are mapped on *Left*, and our estimated results are on *Right*. We accurately localize zip codes with the highest and lowest concentrations of each demographic variable such as the three zip codes in Eastern Seattle with high concentrations of Caucasians, one Northern zip code in Milwaukee with highly educated inhabitants, and the least wealthy zip code in Southern Tampa.

the presence of specific vehicles (such as Hondas) and particular demographics (i.e., the percentage of Asians) or voter preferences (i.e., Democrat).

Our model detects strong associations between vehicle distribution and disparate socioeconomic factors. For instance, several studies have shown that people of Asian descent are more likely to drive Asian cars (15), a result we observe here as well: The two brands that most strongly indicate an Asian neighborhood are Hondas and Toyotas. Cars manufactured by Chrysler, Buick, and Oldsmobile are positively associated with African American neighborhoods, which is again consistent with existing research (16). And vehicles like pickup trucks, Volkswagens, and Aston Martins are indicative of mostly Caucasian neighborhoods. See *SI Appendix*, Fig. S2.

In some cases, the resulting associations can be easily applied in practice. For example, the vehicular feature that was most strongly associated with Democratic precincts was sedans, whereas Republican precincts were most strongly associated with extended-cab pickup trucks (a truck with rear-seat access). We found that by driving through a city while counting sedans and pickup trucks, it is possible to reliably determine whether the city voted Democratic or Republican: If there are more sedans, it probably voted Democrat (88% chance), and if there are more pickup trucks, it probably voted Republican (82% chance). See Fig. 3 *A*, *iii*.

As a result, it is possible to apply the associations extracted from our training set to vehicle data from our test set regions to generate estimates of demographic statistics and voter preferences, achieving high spatial resolution in over 160 cities. To be clear, no ACS or voting data for any region in the test set were used to create the estimates for the test set.

To confirm the accuracy of our demographic estimates, we began by comparing them with actual ACS data, city-by-city, across all 165 test set cities. We found a strong correlation between our results and ACS values for every demographic statistic we examined. (The r values for the correlations were as follows: median household income, $r = 0.82$; percentage of Asians, $r = 0.87$; percentage of Blacks, $r = 0.81$; percentage of Whites, $r = 0.77$; percentage of people with a graduate degree, $r = 0.70$; percentage of people with a bachelor's degree, $r = 0.58$; percentage of people with some college degree, $r = 0.62$; percentage of people with a high school degree, $r = 0.65$; percentage of people with less than a high school degree, $r = 0.54$). See *SI Appendix*, Figs. S3–S5. Taken together, these results show our ability to estimate demographic parameters, as assessed by the ACS, using the automated identification of vehicles in Google Street View data.

Although our city-level estimates serve as a proof-of-principle, zip code-level ACS data provide a much more fine-grained portrait of constituencies. To investigate the accuracy of our methods at zip code resolution, we compared our zip code-by-zip code estimates to those generated by the ACS, confirming a close correspondence between our findings and ACS values. For instance, when we looked closely at the data for Seattle, we found that our estimates of the percentage of people in each zip code who were Caucasian closely matched the values obtained by the ACS ($r = 0.84$, $p < 2e-7$). The results for Asians ($r = 0.77$, $p = 1e-6$) and African Americans ($r = 0.58$, $p = 7e-4$) were similar. Overall, our estimates accurately determined that Seattle, Washington is 69% Caucasian, with African Americans mostly residing in a few Southern zip codes (Fig. 2 *i* and *ii*). As another example, we estimated educational background in Milwaukee, Wisconsin zip codes, accurately determining the fraction of the population with less than a high school degree ($r = 0.70$, $p = 8e-5$), with a bachelor's degree ($r = 0.83$, $p < 1e-7$), and with postgraduate education ($r = 0.82$, $p < 1e-7$). We also accurately determined the overall concentration of highly educated inhabitants near the city's northeast border (Fig. 2 *iv* and *v*). Similarly, our income estimates closely match those of the ACS in Tampa, Florida ($r = 0.87$, $p < 1e-7$). The lowest income zip code, at the southern tip, is readily apparent.

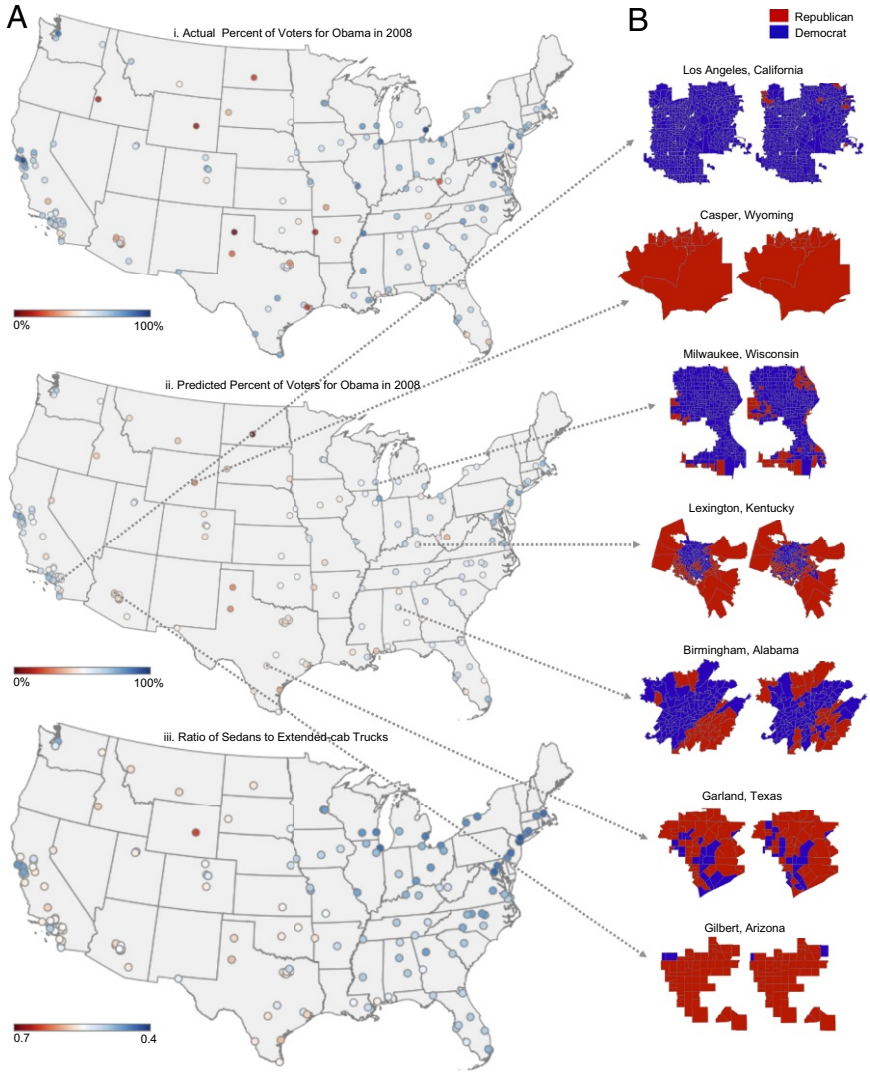

Fig. 3. Actual and inferred voting patterns. *A*, *i* and *ii* map the actual and predicted percentage of people who voted for Barack Obama in the 2008 presidential election (r = 0.74). *iii* maps the ratio of detected pickup trucks to sedans in the 165 cities in our test set. As can be seen from the map, the ratio is very low in Democratic cities such as those in the East Coast and high in Republican cities such as those in Texas and Wyoming. (*B*) Shows actual vs. predicted voter affiliations for various cities in our test set at the precinct level using our full model. Democratic precincts are shown in blue, and Republican precincts are shown in red. Our model correctly classifies Casper, WY as a Republican city and Los Angeles, CA as a Democratic city. We accurately predict that Milwaukee, WI is a Democratic city except for a few Republican precincts in the southern, western, and northeastern borders of the city.

While the ACS does not collect voter preference data, our automated machine-learning procedure can infer such preferences using associations between vehicles and the voters that surround them. To confirm the accuracy of our voter preference estimates, we began by comparing them with the voting results of the 2008 presidential election, city-by-city, across all 165 test set cities. We found a very strong correlation between our estimates and actual voter preferences ($r = 0.73$, $p << 1e-7$). See *SI Appendix*, Fig. S5. These results confirm the ability of our approach to accurately estimate voter behavior.

While city-level data provide a general picture, precinct-level voter preferences identify patterns within a particular city. By comparing our precinct-by-precinct estimates to the 2008 presidential election results, we found that our estimates continued to closely match the ground truth data. For instance, in Milwaukee, Wisconsin, a very Democratic city with 311 precincts, we correctly classify 264 precincts [85% accuracy (Fig. 3*B*)]. Most notably, we accurately determine that there are a few Republican precincts in the South, West, and Northeastern borders of the city. Similarly, in Gilbert, Arizona, a Republican city, we correctly classify 58 out of 60 precincts (97% accuracy), identifying one out of the two small Democratic precincts in the city (Fig. 3*B*). And in Birmingham, Alabama, a city that is 23% Republican, we correctly classify 87 out of the 105 precincts (83% accuracy). Overall, there was a strong correlation between our estimates and actual electoral outcomes at the single-precinct level ($r = 0.57$, $p < 1e-7$).

These results illustrate the ability of our machine-learning algorithm to accurately estimate both demographic statistics and voter preferences using a large database of Google Street View images. They also suggest that our demographic estimates are accurate at higher spatial resolutions than those available for yearly ACS data. Using our approach, zip code- or precinct-level survey data collected for a few cities can be used to automatically provide up-to-date demographic information for many American cities.

Thus, we find that the application of fully automated computer vision methods to publicly available street scenes can inexpensively determine social, economic, and political patterns in neighborhoods across America. By collecting surveys for a few cities—the type of data routinely obtained via ACS—and inferring data for other regions using our model, we can quickly determine demographic patterns.

As self-driving cars with onboard cameras become increasingly widespread, the type of data we use—footage of neighborhoods from vehicle-mounted cameras—are likely to become increasingly ubiquitous. For instance, Tesla vehicles currently take as many images as were studied here every single day. It is also important to note that similar data can be obtained, albeit at a slower pace, using low-tech methods: for instance, by walking around a target neighborhood with a camera and a notepad. Thus, street scenes stand in contrast to the massive textual

corpora presently used in many computational social science studies, which can be constrained by privacy and copyright concerns that prevent individual researchers from obtaining the raw data underlying published analyses.

The automated method we present could be substantially improved by expanding our object recognition beyond vehicles (17, 18) and incorporating global image features (7, 19–21). For instance, our experiments show that global image features extracted from CNN can also be used to infer demographics. But this approach requires more data—at least 50% of our dataset—rather than the 12% to 15% we use using our current method (see *SI Appendix*). The model could also be improved by integrating other types of data, such as satellite images (22), social networks (4), and economic data pertaining to local consumer behavior in particular geographic regions. Nevertheless, there are many characteristics that our methodology—which relies on publicly available data—may not be able to infer (see *SI Appendix*). For instance, the age of children in a neighborhood can be estimated with moderate accuracy ($r = 0.54$), while the percentage of farmers in a neighborhood was not successfully inferred using our method ($r = 0.0$).

Although automated methods could be powerful resources for both researchers and policymakers, their progress will raise important ethical concerns; it is clear that public data should not be used to compromise reasonable privacy expectations of individual citizens, and this will be a central concern moving forward. In the future, such automated methods could lead to estimates that are accurately updated in real time, dramatically improving upon the time resolution of a manual survey.

Materials and Methods

Here, we describe our methodology for data collection, car detection, car classification, and demographic inference. Some of these methods were partially developed in an earlier paper (12), which served as a proof of concept focusing on a limited set of predictions (e.g., per capita carbon emission, Massachusetts Department of Vehicles registration data, income segregation). Our work builds on these methods to show that income, race, education levels, and voting patterns can be predicted from cars in Google Street View images. In the sections below, we discuss our dataset and methodology in more detail.

Dataset. While learning to recognize automobiles, a model needs to be trained with many images of vehicles annotated with category labels. To this end, we used Amazon Mechanical Turk to gather a dataset of labeled car images obtained from edmunds.com, cars.com, and craigslist.org. Our dataset consists of 2,657 visually distinct car categories, covering all commonly used automobiles in the United States produced from 1990 onward. We refer to these images as product shot images. We also hired experts to annotate a subset of our Google Street View images. The annotations include a bounding box around each car in the image and the type of car contained in the box. We partition the images into training, validation, and test sets. In addition to our annotated images, we gathered 50 million Google Street View images from 200 cities, sampling GPS points every 25 miles. We captured 6 images per GPS point, corresponding to different camera rotations. Each Street View image has dimensions 860 by 573 pixels and a horizontal field of view of $\sim 90^\circ$. Since the horizontal field of view is larger than the change in viewpoint between the 6 images per GPS point, the images have some overlapping content. In total, we collected 50,881,098 Google Street View images for our 200 cities. They were primarily acquired between June and December of 2013 with a small fraction (3.1%) obtained in November and December of 2014. See *SI Appendix* for more detail on the data collection process.

Car Detection. In computer vision, detection is the task of localizing objects within an image and is most commonly framed as predicting the *x*, *y*, width, and height coordinates of an axis-aligned bounding box around an object of interest. The central challenge for our work is designing an object detector that is fast enough to run on 50 million images within a reasonable amount of time and accurate enough to be useful for demographic inference. Our computation resources consisted of 4 T K40 graphics processing units and 200 2.1 GHz central processing unit cores. As we were willing to trade a couple of percentages in accuracy for efficiency (12), we turned to the previous state-of-the-art in object detection, DPMs (11), instead of recent algorithms such as ref. 23.

For DPMs, there are two main parameters that influence the running time and performance, which are the number of components and the number of parts in the model. *SI Appendix*, Table S3 provides an analysis of the performance/time tradeoff on our data, measured on the validation set. Based on this analysis, using a DPM with a single component and eight parts strikes the right balance between performance and efficiency, allowing us to detect cars on all 50 million images in 2 wk. In contrast, the best performing parameters would have taken 2 months to run and only increased average precision (AP) by 4.5.

As discussed in ref. 12, we also introduce a prior on the location and size of predicted bounding boxes and use it to improve detection accuracy. Incorporating this prior into our detection pipeline improves AP on the validation set by 1.92 at a negligible cost. *SI Appendix*, Fig. S6*B* visualizes this prior. The output of our detection system is a set of bounding boxes and scores where each score indicates the likelihood of its associated box containing a car.

We converted these scores into estimated probabilities via isotonic regression (24) (see *SI Appendix* for details). We report numbers using a detection threshold of -1.5 (applied before the location prior). At test time, after applying the location prior (which lowers detection decision values), we use a detection threshold of -2.3. This reduces the average number of bounding boxes per image to be classified from 7.9 to 1.5 while only degrading AP by 0.6 (66.1 to 65.5) and decreasing the probability mass of all detections in an image from 0.477 to 0.462 (a 3% drop). *SI Appendix*, Fig. S8 shows examples of car detections using our model. Bounding boxes with cars have high estimated probabilities, whereas the opposite is true for those containing no cars. The AP of our final model is 65.7, and its precision recall curve is visualized in *SI Appendix*, Fig. S7*B*. We calculate chance performance using a uniform sample of bounding boxes greater than 50 pixels in width and height.

Car Classification. Our pipeline, described in ref. 12, classifies automobiles into one of 2,657 visually distinct categories with an accuracy of 33.27%. We use a CNN (25) following the architecture of ref. 14 to categorize cars. CNNs, like other supervised machine-learning methods, perform best when trained on data from a similar distribution as the test data (in our case, Street View images). However, the cost of annotating Street View photos makes it infeasible to collect enough images to train our CNN only using this source. Thus, we used a combination of Street View and the more plentiful product shot images as training data. We modified the traditional CNN training procedure in a number of ways.

First, taking inspiration from domain adaptation, we approximated the WEIGHTED method of Daumé (26) by duplicating each Street View image 10 times during training. This roughly equalizes the number of training Street View and product shot images, preventing the classifier from overfitting on product shot images.

Product shot and Street View images differ significantly in image resolution: Cars in product shot images occupy many more pixels in the image. To compensate for this difference, we first measured the distribution of bounding box resolutions in Street View images used for training. Then, during training, we dynamically downsized each input image according to this distribution before rescaling it to fit the input dimensions of the CNN. Resolutions are parameterized by the geometric mean of the bounding box width and height, and the probability distribution is given as a histogram over 35 different such resolutions. The largest resolution is 256, which is the input resolution of the CNN (see *SI Appendix* for additional details).

At test time, we input each detected bounding box into the CNN and obtain softmax probabilities for each car category through a single forward pass. We only keep the top 20 predictions, since storing a full 2, 657-dimensional floating point vector for each bounding box is prohibitively expensive in terms of storage. On average, these top 20 predictions account for 85.5% of the softmax layer activations' probability mass. After extensive code optimization to make this classification step as fast as possible, we are primarily limited by the time spent reading images from disk, especially when using multiple GPUs to perform classification. At the most fine-grained level (2, 657 classes), we achieve a surprisingly high accuracy of 33.27%. We classify the car make and model with 66.38% and 51.83% accuracy respectively. Whether it was manufactured in or outside of the United States can be determined with 87.71% accuracy.

We show confusion matrices for classifying the make, model, body type, and manufacturing country of the car (*SI Appendix*, Fig. S9 *A–D*). Body type misclassifications tend to occur among similar categories. For example, the most frequent misclassification for "coupe" is "sedan," and the most frequent misclassification for trucks with a regular cab is trucks with an

extended cab. On the other hand, there are no two makes (such as Honda and Mercedes-Benz) that are more visually similar than others. Thus, when a car's make is misclassified, it is mostly to a more popular make. Similarly, most errors at the manufacturing country level occur by misclassifying the manufacturing country as either "Japan" or "USA," the two most popular countries. Due to the large number of classes, the only clear pattern in the model-level confusion matrix is a strong diagonal, indicative of our correct predictions.

Demographic Estimation. In all of our demographic estimations, we use the following set of 88 car-related attributes: the average number of detected cars per image; average car price; miles per gallon (city and highway); percent of total cars that are hybrids; percent of total cars that are electric; percent of total cars that are from each of seven countries; percent of total cars that are foreign (not from the USA); percent of total cars from each of 11 body types; percent of total cars whose year (selected as the minimum of possible year values for the car) fall within each of 5 year ranges (1990–1994, 1995–1999, 2000–2004, 2005–2009, and 2010–2014); and percent of total cars whose make is each of 58 makes in our dataset.

Socioeconomic data were obtained from the ACS (2) and were collected between 2008–2012. See *SI Appendix* for more detail on ground truth data used in our analysis (e.g., census codes). Data for the 2008 US presidential election were provided to us by the authors of ref. 27 and consist of precinct-level vote counts for Barack Obama and John McCain. We ignore votes cast for any other person; that is, the count of total votes is determined solely by votes for Obama and McCain.

To perform our experiments, we partitioned the zip codes, precincts, and cities in our dataset into training and test sets as discussed in the main text, training a model on the training set and predicting on the test set. We used a ridge regression model for income and voter affiliation estimation. For race and education, we used logistic regression to use structure inherent in the data. Specifically, for each region, summing the percentage of people with each of the 5 possible educational backgrounds (or each race) should yield 100%. In all cases, we trained 5 models using fivefold cross-validation to select the regularization parameter and averaged the trained models. We normalize the features to have zero mean and unit SD (parameters determined on the training set). We also clip predictions to stay within the range of the training data, preventing our estimates from having extreme values. The geographical regions of interest are restricted to be ones with a population of at least 500 and at least 50 detected cars.

We compute the probability of voting Democrat/Republican conditioned on being in a city with more pickup trucks than sedans as follows. Let r be the ratio of pickup trucks to sedans. We would like to estimate $P(Democrat|r > 1)$ and $P(Republican|r < 1)$:

$$P(Democrat|r > 1) = \frac{P(Democrat, r > 1)}{P(r > 1)} \quad [1]$$

$$P(Republican|r < 1) = \frac{P(Republican, r < 1)}{P(r < 1)} \quad [2]$$

We estimate $P(Democrat, r > 1)$, $P(Republican, r < 1)$, $P(r > 1)$, and $P(r < 1)$ as follows. Let $S_d = \{c_i\}$ be the set of cities with more votes for Barack Obama than John McCain. Let $S_s = \{c_j\}$ be the set of cities with more sedans than pickup trucks. Let n_s be the number of elements in S_s and let $n_d s$ be the number of elements in $S_d \cap S_s$. Similarly, let S_p be the set of cities with more pickup trucks than sedans, S_r the set of cities with more votes for John McCain than Barack Obama, and $n_r p$ the number of elements in $S_r \cap S_p$. Finally, let C be the number of cities in our test set:

$$P(Democrat, r > 1) \approx \frac{n_d s}{C} \quad [3]$$

$$P(Republican, r < 1) \approx \frac{n_r p}{C} \quad [4]$$

$$P(r > 1) \approx \frac{n_s}{C} \quad [5]$$

$$P(r < 1) \approx \frac{n_p}{C} \quad [6]$$

Using these estimates, we calculate $P(Democrat|r > 1)$ and $P(Republican|r < 1)$ according to Eqs. **1** and **2**.

ACKNOWLEDGMENTS. We thank Neal Jean, Stefano Ermon, and Marshall Burke for helpful suggestions and edits; everyone who worked on annotating our car dataset for their dedication; and our friends and family and the entire Stanford Vision lab, especially Brendan Marten, Serena Yeung, and Selome Tewoderos for their support, input, and encouragement. This research is partially supported by NSF Grant IIS-1115493, the Stanford DARE fellowship (to T.G.), and NVIDIA (through donated GPUs).

1. Department of Commerce, US Census Bureau (2013) US census bureau's budget estimates. Available at www.osec.doc.gov/bmi/budget/fy13cbj/Census_FY2013_CongressionalJustification-FINAL.pdf. Accessed September 13, 2014.
2. Department of Commerce, US Census Bureau (2012) American community survey 5 year data (2008-2012). Available at https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk. Accessed September 13, 2014.
3. Department of Commerce, US Census Bureau (2010) Decennial census. Available at https://www.census.gov/data/developers/data-sets/decennial-census.html. Accessed September 13, 2014.
4. Antenucci D, Cafarella M, Levenstein M, Ré C, Shapiro MD (2014) *Using Social Media to Measure Labor Market Flows* (National Bureau of Economic Research, Cambridge, MA), Technical Report 20010.
5. Michel JB, et al. (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331:176–182.
6. Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. *Science* 350:1073–1076.
7. Naik N, Philipoom J, Raskar R, Hidalgo C (2014) Streetscore–Predicting the perceived safety of one million streetscapes. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE, New York), pp 793–799.
8. Naik N, Kominers SD, Raskar R, Glaeser EL, Hidalgo CA (2017) Computer vision uncovers predictors of physical urban change. *Proc Natl Acad Sci USA* 114:7571–7576.
9. American Association of State Highway and Transportation Officials (2013) *Vehicle and Transit Availability. Commuting in America 2013* (American Association of State Highway and Transportation Officials, Washington, DC), Report 7.
10. Choo S, Mokhtarian PL (2004) What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice. *Transport Res Pol Pract* 38:201–222.
11. Felzenszwalb P, Girshick R, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part based models. *IEEE Trans Pattern Anal Mach Intell* 32:1627–1645.
12. Gebru T, et al. (2017) Fine-grained car detection for visual census estimation in *AAAI*, in press.
13. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86:2278–2324.
14. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. Available at papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. Accessed November 9, 2017.
15. Bland M (2012) Asian consumers and the automotive market. Available at app.compendium.com/uploads/user/a33eed35-8a44-4da7-84c4-16f3751fe303/9855ee60-f764-43b4-84c4-40950ff36307/File/3e1e2e5d8d20fad49eaac919e38abc8e/polk_3af_05_17_2012_presentation.pdf. Accessed November 6, 2017.
16. Auto Remarketing Staff (2011) Which brands most attract African-American buyers? Available at www.autoremarketing.com/content/trends/which-brands-most-attract-african-american-buyers. Accessed October 24, 2016.
17. Simo-Serra E, Fidler S, Moreno-Noguer F, Urtasun R (2015) Neuroaesthetics in fashion: Modeling the perception of beauty. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York), pp 869–877.
18. Matzen K, Bala K, Snavely N (2017) Streetstyle: Exploring world-wide clothing styles from millions of photos. arXiv:1706.01869.
19. Ordonez V, Berg TL (2014) Learning high-level judgments of urban perception. *European Conference on Computer Vision* (Springer, Boston), pp 494–510.
20. Khosla A, An B, Lim JJ, Torralba A (2014) Looking beyond the visible scene. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York), pp 3710–3717.
21. Zhou B, Liu L, Oliva A, Torralba A (2014) Recognizing city identity via attribute analysis of geo-tagged images. *European Conference on Computer Vision* (Springer, Boston), pp 519–534.
22. Jean N, et al. (2016) Combining satellite imagery and machine learning to predict poverty. *Science* 353:790–794.
23. Ren S, He K, Girshick R, Sun J (2017) Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell* 39:1137–1149.
24. Barlow RE, Bartholomew DJ, Bremner J, Brunk HD (1972) *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression* (Wiley, New York).
25. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86:2278–2324.
26. Daumé H III (2007) Frustratingly easy domain adaptation. *Conference of the Association for Computational Linguistics* (ACL, Prague, Czech Republic).
27. Ansolabehere S, Palmer M, Lee A (2014) Precinct-Level Election Data. Available at hdl.handle.net/1903.1/21919. Accessed January 13, 2015.