December 10, 2015

United States District Court for the Western District of Wisconsin

One Wisconsin Institute, Inc., et al. v. Nichol, et al.

Case No. 3:15-CV-324

Expert Report

Yair Ghitza

Catalist

1090 Vermont Avenue NW, Suite 300

Washington, DC 20005

Yair Ghitza

12/10/2015
Date

Biographical Information

My name is Yair Ghitza and I am the Chief Scientist at Catalist, LLC (hereinafter "Catalist"). I began working at Catalist in a part-time capacity as a Consultant in 2008, and have been serving in a full time capacity in my role as Chief Scientist since 2011. I have over eleven years of experience in statistics, political science, and computer science in both professional and academic settings. I received my PhD from Columbia University in Political Science. My dissertation—*Applying Large-Scale Data and Modern Statistical Methods to Classical Problems in American Politics*—was accepted with distinction and was nominated for the annual Savage Award in Applied Methodology in 2014. The Savage Award in Applied Methodology is awarded to a dissertation that makes outstanding contributions in the field of applied Bayesian statistics. My main areas of expertise are statistical methods and American politics, particularly focusing on the analysis of public opinion and voting. Prior to my time at Columbia, I was a visiting research assistant in the Media Lab at MIT, performing core research in artificial intelligence and computer vision.

Catalist is a leading data utility that provides services to civic engagement and advocacy organizations as well as political campaigns. Catalist compiles, enhances, stores, and updates person-level data for the entire U.S. adult population. The base data is obtained from official voting rolls in all 50 states and the District of Columbia, as well as from national commercial consumer databases. Combining these datasets with publicly available data, such as Census data, and private data from our clients and business partners, results in a rich, national database of civic and commercial behavior that is updated on an ongoing and regular basis. Catalist augments its database with a number of modeled predictions about each person's likely civic behavior, characteristics, or preferences on a number of subjects relevant to civic participation.

At Catalist, I am mainly responsible for constructing and overseeing the construction of thousands of predictive models. Usually, these models predict some characteristic, attitude, or likely behavior of the individuals in the database. I have built these types of models for a wide range of topics, from the likelihood that a particular voter will participate in a particular election, to the likelihood that a voter has children in the household, conditional on a wide range of other data points. I also lead research efforts at Catalist, often building new statistical methods to deal with different types of data, and developing new methods of leveraging the data to help our clients achieve their goals.

Nature and Scope of Retained Services

With respect to this matter, Catalist was retained by Plaintiffs' counsel to append probabilistic race and partisanship estimates to individual-level records from the Wisconsin voter registration file, which was provided by Plaintiffs' counsel.[1] For this work, Catalist is being compensated in the following manner: Catalist is receiving an hourly rate of $200.00 for Report and Preparation Fees (which is inclusive of deposition testimony, preparation and other work), and a File Processing Fee of $23,530 (which is inclusive of taking the file provided by Plaintiffs' counsel and performing the following tasks: (1) running it through address standardization/correction, (2)

[1] I signed a protective order before accessing these materials.

appending geocoding information, and (3) appending race and partisan estimates using Catalist's race and partisanship models).

With respect to the performance of services, Catalist received one text file from Plaintiffs' counsel, consisting of approximately 3.4 million records, with each record containing information for a single individual. Catalist took the following steps to append the required data to each record of the file: (a) standardized the file format using address standardization and CASS correction;[2] (b) appended geocodes, based on the standardized addresses; (c) appended 2010 Census geography; (d) appended probabilistic race estimates; (e) matched each record to the internal Catalist database, attempting to find a record of the associated person in our internal database; and (f) appended a probabilistic partisanship estimate for each person that was found in the match process. In preparation for our work, we discussed this process with Drs. Mayer and Barry Burden. Once all of these fields were appended, we were directed by Plaintiffs' counsel to provide the output file to Dr. Mayer.

With respect to geocodes, our geocode process appends a latitude and longitude to the address record. Geocode data was found for 99.98% of the addresses on the file, with the remaining records not geocoded because the address associated with the record was missing or not associated with a known geocode. 2010 Census geography was appended for 99.68% of the records, with remaining records missing because census geographies have a dependency on geocodes, with low fidelity geocode types sometimes not applicable.

The race estimates are calculated probabilities that each person falls into one of six racial categories—White, Black, Hispanic, Asian, Native American, or Other—which sum to 100% for each individual sent to Catalist for appending. We also include our best estimate of each record's race based on these six probabilities, and a confidence estimate. These racial probability estimates are based on a series of statistical models, built internally at Catalist. The models are developed using millions of records of self-reported racial self-identification from both voter registration records and high-quality survey data.

We use decision trees, regression, and other machine learning techniques to train the models, along with standard statistical processes such as a holdout validation set to ensure the validity of the modeled estimates. We have found that our race estimates out-perform other commercial and academic alternatives. This is due to three main methodological innovations. We improve upon existing methods for (1) extracting more reliable information from names, particularly in combination with age; (2) extracting more reliable information from census data as it relates to race estimates; and (3) leveraging information from both voter registration records and survey data in combination to provide the best possible estimates.

The partisanship estimate is a single number, indicating the probability that each person self-identifies as a Democrat, as a percentage of (Democrat + Republican). A score of 0 indicates the person is most likely to identify as Republican; a score of 100 indicates the person is most likely to self-identify as Democrat; and scores in between indicate a range of uncertainty, with a score of 50 indicating the least amount of certainty between the two. Although self-identified

[2] CASS is an address certification system offered by the U.S. Postal Service that improves the accuracy of carrier route, 5-digit ZIP, ZIP + 4, and delivery point codes that appear on mail pieces.

Independents are not explicitly indicated in the partisanship estimate, they are more likely to have scores closer to 50. Like the race estimates, this model was built internally at Catalist, using millions of records of self-identified party registration and party identification from both voter registration records and high-quality survey data. We use decision trees, regression, and other machine learning techniques to train the model, along with standard statistical processes such as a holdout validation set to ensure the validity of the modeled estimate. The partisanship estimate was appended to 3,333,574 records, 98.6% of the full file. The remaining records were not given a partisanship estimate because we did not find a record of them in the internal Catalist database.

List of Publications I Have Authored Within the Previous 10 Years

- Ghitza, Yair. *Applying Large-Scale Data and Modern Statistical Methods to Classical Problems in American Politics*. 2014. PhD Dissertation.

- Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57(3):762-776.

- Erikson, Robert and Yair Ghitza. 2012. "Setting the Agenda Setter." Prepared for *2012 Meeting of the American Political Science Association*, New Orleans, LA. August 30-September 2, 2012.

- Erikson, Robert, Yair Ghitza, and Christopher Wlezien. 2010. "Differential Campaign Effects in Battleground and Non-Battleground States? An Analysis of Recent Presidential Elections." Presented at the *Annual State Politics and Policy Conference*, Springfield, Illinois, June 3-5, 2010.

- Gelman, Andrew, Daniel Lee, and Yair Ghitza. 2010. "Public Opinion on Health Care Reform." *The Forum* 8(1).

- Gelman, Andrew, Daniel Lee, and Yair Ghitza. 2010. "A Snapshot of the 2008 Election." *Statistics, Politics, and Policy* 1(1).

- Gelman, Andrew, Jonathan P. Kastellec, and Yair Ghitza. 2009. "Beautiful Political Data." In *Beautiful Data: The Stories Behind Elegant Data Solutions*. O'Reilly Media.

- Ghitza, Yair, and Todd Rogers. 2009. "Data Driven Politics." In *The Change We Need: What Britain Can Learn from Obama's Victory*. Ed. Nick Anstead and Will Straw. Fabian Society.

List of Matters Where I Have Testified as an Expert Within the Previous 10 Years

- *State of Texas v. Holder,* C.A., No. 12-cv-00128 (RMC-DST-RLW) (D.D.C.), DJ 166-76-143.