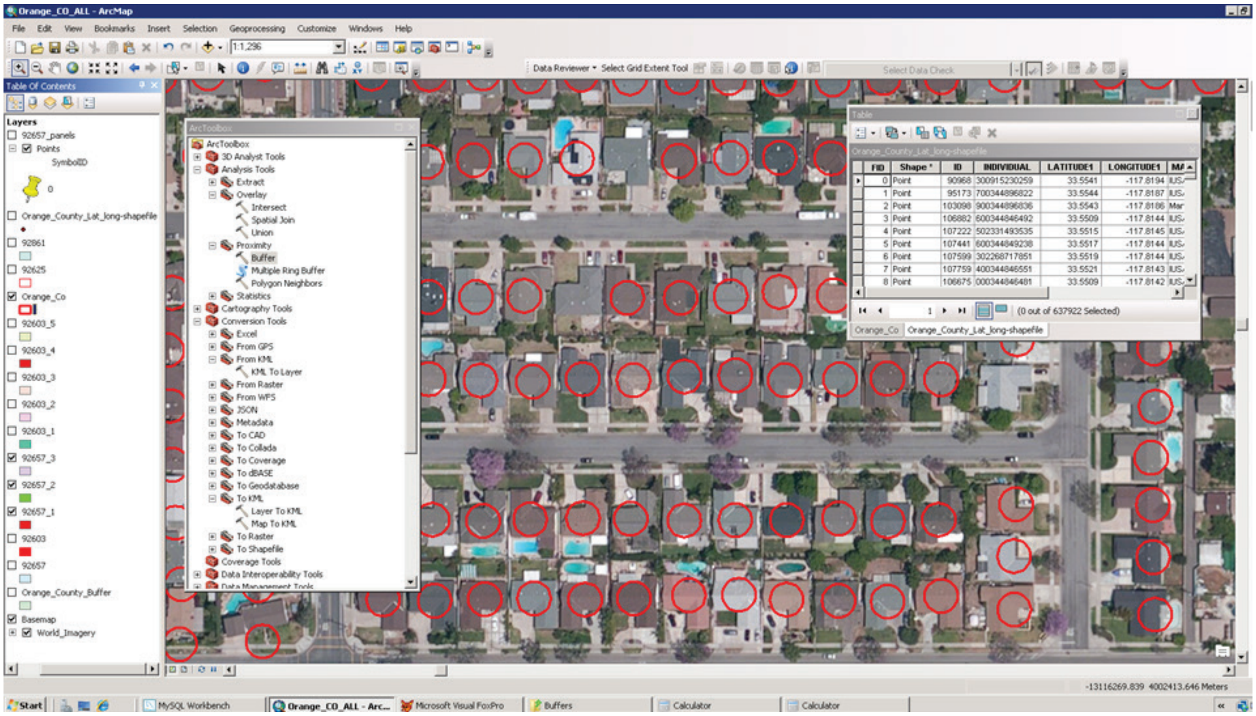
Using Amazon’s Mechanical Turk & Machine Learning to Identify & Model Owners of Solar Panels

Less than one percent of the 125M plus residential buildings in the United States currently have solar panels installed. This low penetration exists despite a 30% federal tax credit and other state and local incentives. At the same time, according to the framework of the recently-signed Paris Accords, the United States will need to reduce its greenhouse gas emissions by 26% below 2005 levels by 2025. 30% of US greenhouse gas emissions currently come from electrical generation. While many reductions in this category will come from new solar and wind power plants, residential solar will also play a large role in meeting the Paris obligations. The financial benefit a consumer could realize for installing residential solar differs according to a geography’s level of solar suitability (sunlight) and a state’s financial incentives, but there are a number of states where it is already economically worthwhile to install solar (AZ, CA, CO, DE, FL, HI, MA, MD, NJ, NV, NY, etc.). This raises the question of how do we find the individual home owners most likely to buy or lease solar panels, particularly in these target states.

HaystaqDNA’s interest in this area is not entirely academic. While the company’s origins are as a left-leaning political modeling firm, there is an immediate value in finding residential solar buyers for other industries as well. Individuals who buy in one category of green energy will usually buy in others. Haystaq’s existing clients in the in the automotive industry face the increasing need to find buyers of electric and electric-hybrid vehicles. In the coming years, manufacturers of LED lights, tankless water heaters, high efficiency appliances, etc. will be able to market to these same individuals.

Haystaq’s founder and CEO, Ken Strasma perfected his microtargeting skills and techniques in John Kerry’s 2004 democratic primary run and in Barack Obama’s 2008 presidential campaign. In 2013, HaystaqDNA was formed to take these techniques and technologies used in the political arena and use them to help companies find and understand their customers in the corporate world. We knew that if we could find a sample of existing solar panel owners, we could use our advanced statistical algorithms to find other consumers who would behave in a similar way, just as we have in politics. Where there is sufficient data available, Haystaq has successfully applied microtargeting techniques to verticals including: automotive, healthcare, television programming, professional sports, consumer package goods and retail. Unfortunately, outside of the agencies that regulate state incentive programs (which don’t share data), there is no centralized source of solar panel owners. Haystaq needed to create its own method of identifying solar owners.

While solar panel ownership data is not freely available, there are a number of sources for satellite and aerial photos. Through early in-house experiments, we found that if satellite images were overlaid with GPS coordinates (either provided by the vendor or geocoded in-house) we could match images of structures to the owners of those structures. We did this by using a commercial database and either using the provided latitude/longitude values for each household or by geocoding the addresses when this information was not provided. We were then able to manually review images rooftop by rooftop and determine which had solar panels. This initial effort was successful, but time-intensive and wearying for our analysts.

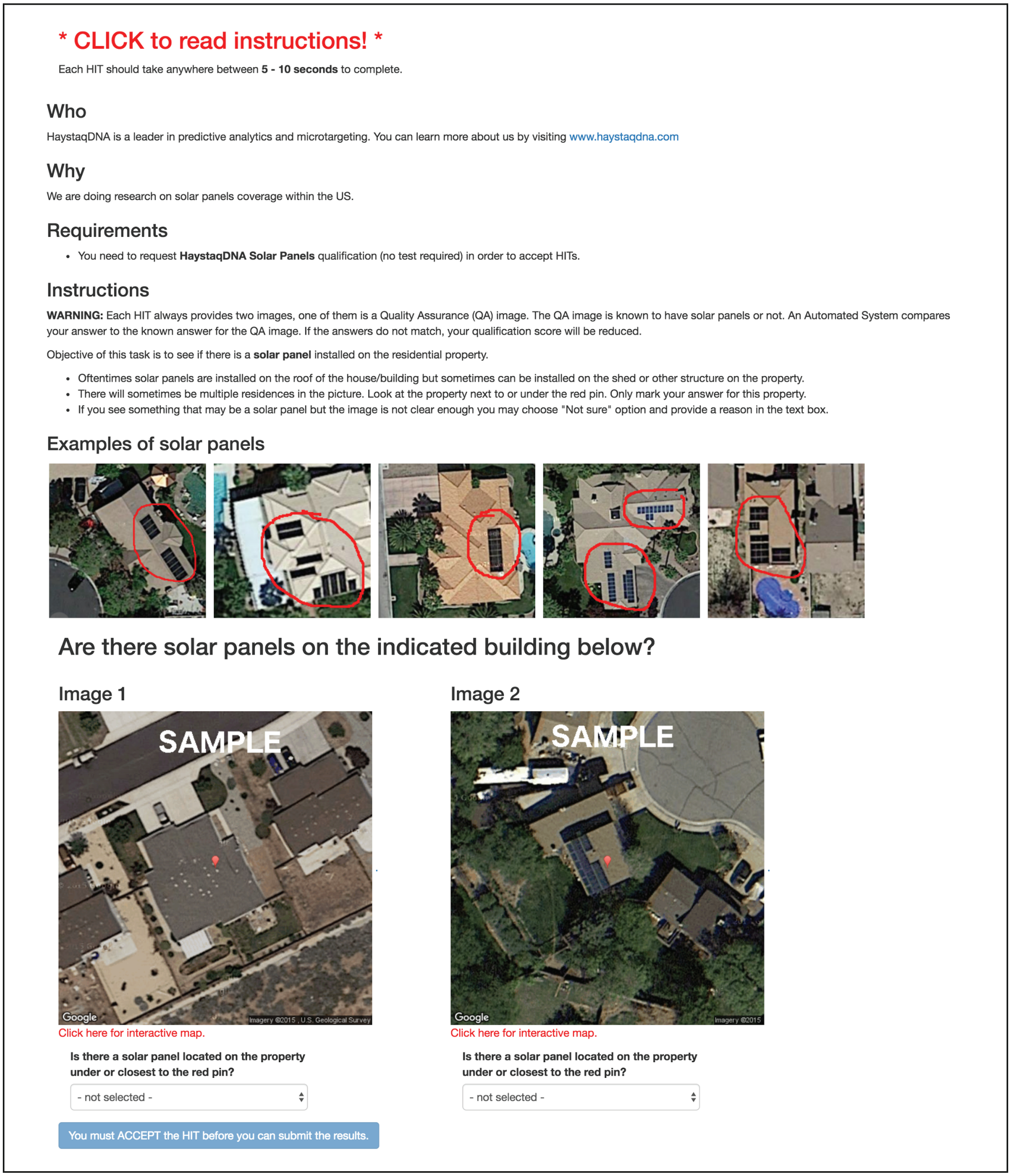


FID	Shape *	ID	INDIVIDUAL	LATITUDE1	LONGITUDE1	MA
0	Point	90968	300915230259	33.5541	-117.8194	USA
1	Point	95173	700344896822	33.5544	-117.8187	USA
2	Point	103098	900344896836	33.5543	-117.8186	Mar
3	Point	106882	600344846492	33.5509	-117.8144	USA
4	Point	107222	502331493535	33.5515	-117.8145	USA
5	Point	107441	600344849238	33.5517	-117.8144	USA
6	Point	107599	302268717851	33.5519	-117.8144	USA
7	Point	107759	400344846551	33.5521	-117.8143	USA
8	Point	106675	000344846481	33.5509	-117.8142	USA

(0 out of 637922 Selected)

An image from Haystaq's early internal interface for finding rooftop solar panels

The next step was to turn to Amazon's Mechanical Turk (MTurk) crowdsourcing marketplace. MTurk allows Haystaq to put out a request for work to be completed by remote workers willing to work for the incentive offered. In this case we wanted users to look at images of roofs and mark whether each one appears to contain a solar panel or not. We automated a backend to carve images into individual residential buildings, match those buildings to households on our consumer file and feed those images into MTurk's API. We slip in images, that we know contain solar panels and we use these images as our Quality Assurance. If a worker cannot correctly identify the QA images, we disregard their work and prevent them from accepting future work from us. For the non-QA images, we feed each image to two different workers, if they agree that an image does or does not contain solar panels, we mark it as such. If the workers disagree, we send the image to a third worker for arbitration. Due to a high prevalence of rooftop solar panels, we have collected test samples from areas of Orange County, CA; Los Angeles County, CA; and Nevada.



* CLICK to read instructions! *

Each HIT should take anywhere between **5 - 10 seconds** to complete.

Who

HaystaqDNA is a leader in predictive analytics and microtargeting. You can learn more about us by visiting www.haystaqdna.com

Why

We are doing research on solar panels coverage within the US.

Requirements

- You need to request **HaystaqDNA Solar Panels** qualification (no test required) in order to accept HITs.

Instructions

WARNING: Each HIT always provides two images, one of them is a Quality Assurance (QA) image. The QA image is known to have solar panels or not. An Automated System compares your answer to the known answer for the QA image. If the answers do not match, your qualification score will be reduced.

Objective of this task is to see if there is a **solar panel** installed on the residential property.

- Oftentimes solar panels are installed on the roof of the house/building but sometimes can be installed on the shed or other structure on the property.
- There will sometimes be multiple residences in the picture. Look at the property next to or under the red pin. Only mark your answer for this property.
- If you see something that may be a solar panel but the image is not clear enough you may choose "Not sure" option and provide a reason in the text box.

Examples of solar panels

Are there solar panels on the indicated building below?

Image 1

Click here for interactive map.

Is there a solar panel located on the property under or closest to the red pin?

- not selected -

Image 2

Click here for interactive map.

Is there a solar panel located on the property under or closest to the red pin?

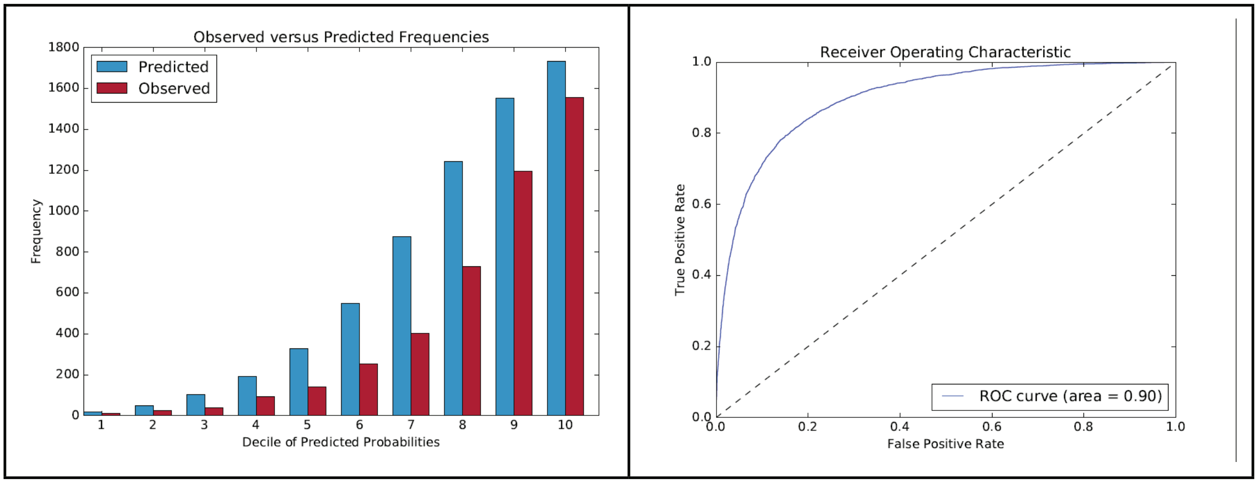
- not selected -

Above is the initial screen Haystaq's MTurk workers see.

Our MTurk interface has provided us with huge efficiency gains over our initial method. There are still several drawbacks. 1. With novice MTurk workers, we get some false positives for things easily mistaken for solar panels, like skylights or solar water heaters (this second one is not as problematic as it is still a green product). 2. While this gives us a great way to efficiently sample a geography, it is still too expensive and slow to comprehensively examine the entire country. We can, however, create a model with the geographic samples we have collected.

At this point in the process we turn back to Haystaq's tried and tested data analytics techniques and code. The consumer file mentioned above, consists of roughly 260M US adult consumers. This file contains all of the Personally Identifying Information (PII) as well as over 1200 fields of additional data — Census Data, Property Data, Survey Data, Modeled Data and aggregated data bought from sources like magazines, retailers, airlines, hotels, insurance companies, financial institutions, etc. All of this data is converted to 'indicator' or 'independent variable' form where text fields are converted to binary flags and false numeric data is discarded. The solar panel sample data from MTurk is already matched to this dataset at a household level. We create the 'dependent variables' or 'DVs' for each house using an assigned head of household. If a roof had a solar panel, the owner of that house is designated as a 1, the owner of a house without a solar panel is assigned a 0. We then use Python (and its SciKit-Learn and Pandas libraries) to model these dependent variables. We use a variety of algorithms (Logistic Regression, Decision Trees, Nearest Neighbor, Neural Networks, etc) to create the initial models. We often blend the results of multiple models as we find that doing this tends to amplify the underlying signal of each model and cancel out the random errors (noise). No one indicator field will determine a person's score, our typical models will use in excess of 100 coefficients. The final model provides an algorithm that scores each individual in the consumer database for relative likelihood to become a solar panel buyer.

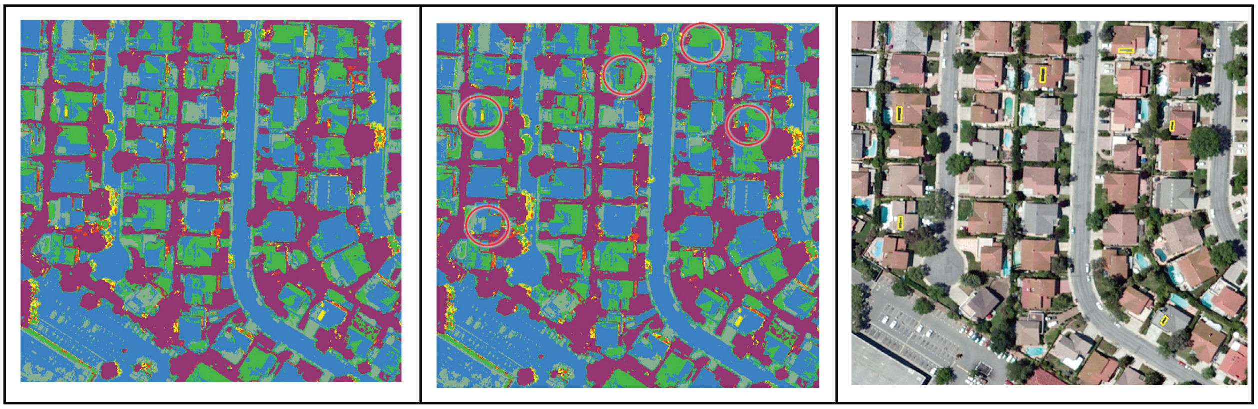
Having a model is meaningless, unless we can validate that it works. Towards this end we withhold one third of the sample of known solar panel owners from the modeling process which becomes the 'test set'. We then verify that our scores accurately classified the individuals in the test set who are known to own solar panels. Haystaq's solar panel models have proved to be highly predictive of the test set.



An example of two of our QA checks against the test sample — a Hosmer-Lemeshow step chart and a Receiver Operating Characteristic chart are featured above.

Once we have a model that validates well, we score everyone on the consumer file within a similar geography. At this point we have a rank ordering of consumers ranging from most likely to buy solar (or other green products) to those least likely to buy. We believe this model has direct value to people marketing these product, but at this point we use this model as a filter on or feeder model for our electric and electric-hybrid car models.

NEXT STEPS: Ideally, we would not rely on a model to find solar panel owners, but instead we would be able to identify all users. With our existing process using MTurk, classifying all 125M US rootfops as solar or non-solar would be time- and cost-prohibitive. This is exacerbated by the need to resurvey periodically to monitor solar growth. To solve this problem, we are writing code to have our AWS cluster environment attempt to categorize the rooftop images before we send them to MTurk workers for verification. We can feed a server a set of images known to have solar panels and a set of images that do not. Using image sensing and machine learning algorithms, it will then attempt to categorize new images. Those images are scored, the server learns from its mistakes and this process continues iteratively until the server can correctly categorize the rooftops with solar installs. Using this process, only the images identified by the server as containing solar panels would be sent to MTurk for human validation.



Some images from trial runs of using machine learning techniques to identify solar panels on rooftops.

There are a number of challenges with having a computer correctly identify a specific set black rectangles surrounded by other dark rectangles, but the early results seem promising.

There are two potential challenges specific to solar that might limit the window in which we can use this technology. One is that our models are making the assumption that a home's current owner is the owner that installed the solar (or secondarily that the owner valued the solar panels equivalently when buying the home). While rooftop solar penetration is low, that is a fairly safe assumption, but as the efficiency of solar panels increase and their penetration increases that assumption will eventually break. We can get around this challenge by having snapshots in time of solar coverage and comparing this to changes in homeownership that are visible in the consumer file. The other challenge is that our method assumes that solar panels are the easy to identify black rectangles we currently see. Recently Tesla announced solar panels that look and act like traditional roofing tiles — aside from Tesla it is likely that future panels will derivate into different shapes and arrangements. That said, solar is only one potential application for this technology as it could also be easily adapted to find things like swimming pools, boats or RVs.