Valuation Technical Brief

Executive Summary

HouseCanary develops the most accurate and comprehensive valuations for residential real estate. This paper provides a transparent look at the methodology we use to combine the best data with the best models.

The objective of our valuation modeling is to accurately predict the current value of a given property within the United States: what that home would currently transact for at arm's length. HouseCanary strives to put a value on every single-family home, condo, townhouse, manufactured home, and multi-family property in the United States. The value range should be as narrow as possible while still providing an approximate 68% coverage probability (one standard deviation) on actual arm's-length sale prices.

The purposes of this paper are as follows:

1. Define HouseCanary's modeling principles and assumptions
2. Identify the data included in the HouseCanary valuation algorithm
3. Provide context for the logic behind the HouseCanary valuation algorithm
4. Define key model outputs and validation

At a Glance

106.5 million

US residential real estate properties

Models for five residential property types
Updated monthly

2.8%

median absolute prediction error (MdAPE) as of publication (July 2019)

Machine learning-based
Third-party validated
Performance reported online

Modeling Principles and Assumptions

The key to accurate valuations: comprehensive data + machine learning

At a high level, our valuation algorithm follows three steps: (1) query and clean relevant data from HouseCanary's database, (2) build localized price indices to adjust all past prices to current values, and (3) train valuation models from time-adjusted historical prices.

Traditional valuation models only consider recently transacted nearby comparable properties. While recently transacted comparable properties are an important piece of the valuation puzzle, we can achieve more accurate valuations by also considering all previous arm's-length sales for the subject property and neighborhood, as well as information from multiple other sources: macroeconomic data, capital markets data, mortgage records, search and social data, and house/parcel data.

The following statements are assumptions about the data inputs, model methodology, and resulting property valuations.

1. **Neighborhoods move together**: Property prices in a given neighborhood tend to move together through time.
2. **Deep histories are meaningful**: It is statistically more efficient to use all known historical price events instead of only those occurring within the past few years.
3. **Past value is a strong predictor**: After accounting for trends in price and changes in the condition and structure of the underlying asset, the best predictor of the value of a home is most often past price occurrences of that home.
4. **Machine learning optimizes modeling**: Machine learning-based algorithms are better than classical models at recognizing and exploiting higher-order complex relationships among input variables and their relationships to the response variable, and enable us to continually improve our models without directly programming them for improvement.
5. **Humans ensure data quality**: Human effort should focus on enhancing existing datasets to generate cleaner data that can be fed into the algorithm, further improving valuation accuracy.

Valuation Data

Nationwide county assessor, county recorder, MLS, and other property-level data

Property-level predictor variables entering the valuation algorithm are composed of public record data, multiple listing service (MLS) data, and other property-level information:

- **Public record data** include property characteristics sourced from 3,100+ county assessor offices and historical recorded sales prices sourced from 2,700+ county recorder offices over the previous 20 years, where available.
- **MLS data** include property characteristics, listed prices, and contract prices.
- **Other property information** includes data on mortgage balances, mortgage type, and measures of financial distress, along with other details that impact value, such as whether a property has a yard, a view, or is in proximity to a busy street versus a golf course.

For the purpose of non-disclosure states, contract prices from the MLS are used as a substitute for recorded sales prices if we can jointly verify that an arm's-length sale was actually recorded from the county recorder's office.

We refresh public records and MLS data daily. The only potential delay in our dataset would be due to delays from the original source. For example, if a particular county takes three months to add recorded transactions to its electronic data file, we will not have access to those transactions until three months after they were recorded. Market-level summary data are typically compiled and added into the database monthly or quarterly.

Ensuring Data Quality

Trusted data quality, with addresses certified by USPS CASS

HouseCanary ensures data quality with systems and processes developed and supported by a team of data engineers, data analysts, and domain experts. We design our system for complete end-to-end visibility of data flows, with layers of dynamic, intelligent controls.

Our process starts with full profiling of any data source that will feed into our platform. To profile data sources we use multiple full data instances over time to provide a complete picture of content and expected changes. We create field-by-field, value-by-value rules that determine how the data map into our content management system. Those rules, combined with intelligent monitoring, provide a first layer of control that flags suspicious data changes, anomalous and unexpected values, and inconsistent data use. Flagged content gets quarantined until a human review approves it as valid, the issue gets remedied by the source, or we implement new handling logic that solves the problem.

When validated data flow into the content management system, they are linked and normalized at the address, building, and census block levels. Once the content is linked, a second quality control pass uses multi-source comparison to arrive at a consensus view of the correct and usable data for a given object. Only these data are fed into products and models.

We use a United States Postal Service Coding Accuracy Support System (USPS CASS) certified service to validate, standardize, and match all addresses that feed into our system. The USPS CASS service handles all matching and standardization for all addresses, whether they come from data feeds or from user inputs. To protect against degradation related to updates, any change of any component of a full address triggers a new validation, standardization, and matching event for the subject address and all previously matched and related addresses.

Algorithm Methodology

Step 1: Query and clean relevant data from HouseCanary's database
Prior to modeling, we gather, combine, and filter all of our validated data. This step includes, but is not limited to, identifying valid historical arm's-length sale prices and market clearing list prices and using proprietary, localized models to determine valid property characteristics when data differ across sources. The filtered historical sales and list prices form the basis for our response variable and price indices.

We build and run each model at the census tract level. Census tracts partition the United States into non-overlapping sections, which are further subdivided into non-overlapping census blocks. When a single census tract does not have enough data to model, we incorporate data from neighboring tracts to supplement the target tract. If neighbors plus the target still do not yield enough information, we continue to pull in data from neighbors of neighbors until the required sample size threshold has been met. Neighboring census tract information is only used to train models, and not used for accuracy testing purposes. Separate models will be built for the neighboring tracts, with each individually treated as a subject tract.

Step 2: Build localized price indices to adjust all past prices to current values
Our proprietary machine learning models generate local historical price indices for each of five property types (Single Family Residential, Multi-family Residential, Condominium, Townhouse, Manufactured/Mobile). To illustrate, the time scatter plot in Figure 1 shows all transactions within a census tract over the last 20 years. The vertical axis represents price in dollars.

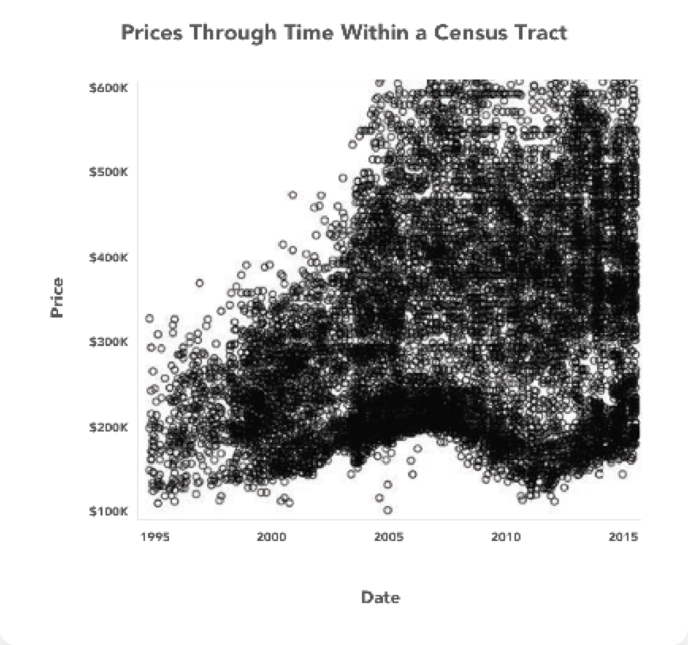

Figure 1: All transactions within a census tract over the last 20 years

As an example, Figure 2 plots all identified price instances by property type for four individual census blocks within the target census tract. Census blocks are the smallest quantifiable groups of properties in the US — there are approximately 10 million census blocks across the country. Single-family home prices are shown in Figure 2 as yellow dots, and condos are shown as blue dots. The yellow line represents the model-estimated single-family home price index for a particular block, and the blue line represents the model-estimated condo price index for that block.

We use proprietary machine learning methods to create smooth price indices at the census block level. These methods borrow information from surrounding areas to estimate the index for an individual block. By borrowing information we can generate price indices even for blocks with very small sample sizes. As an example, Block 4 only contains condos, and has a very limited number of historical price observations. However, the model is able to estimate an index over the entire 20-year time period by using the behavior of condo prices from other similarly behaved condos in the census tract.

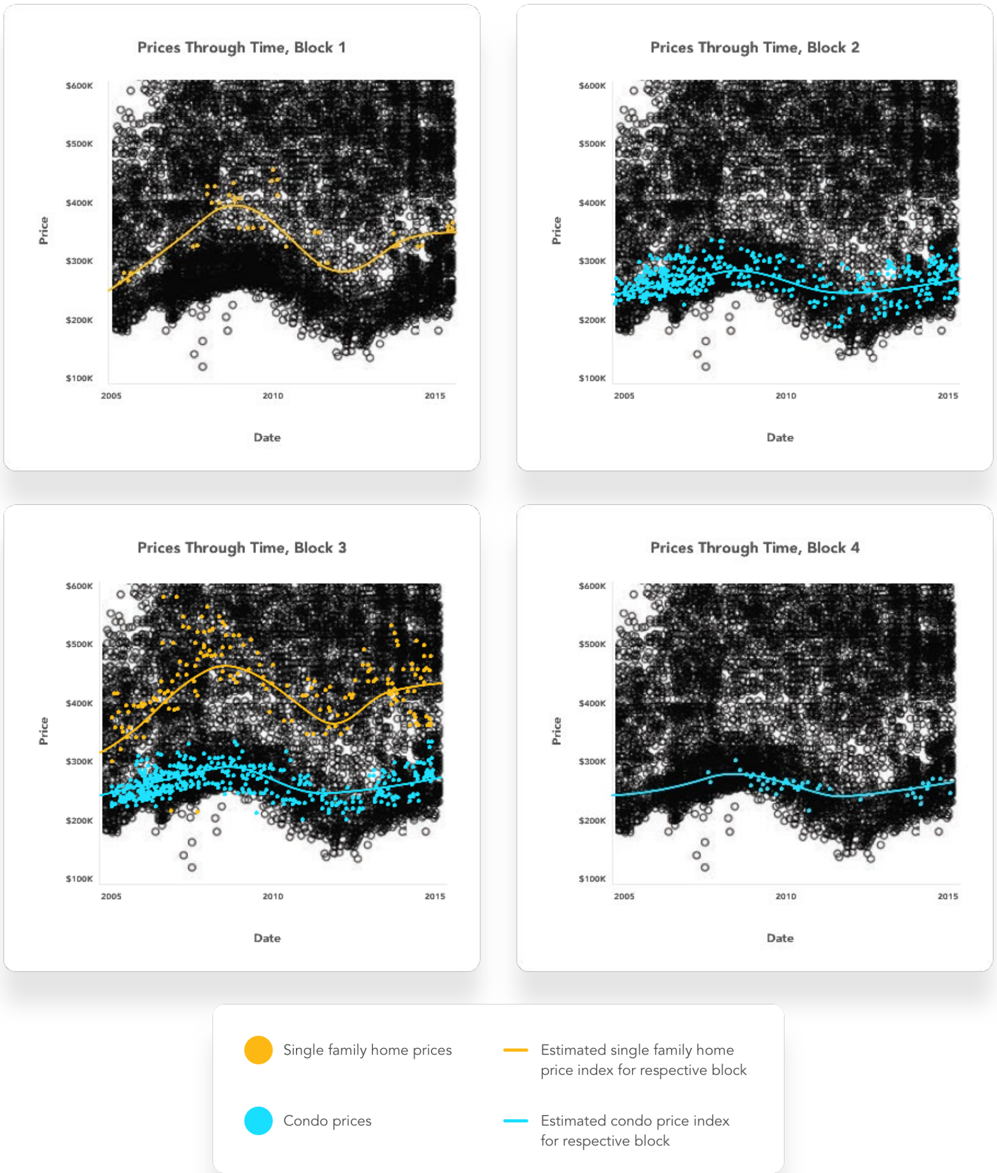

Figure 2: All price instances by property type for four individual census blocks

We generate two index models: one in terms of median price and another in terms of median price per square foot. Each model generates indices broken out by property type when more than one property type exists within the relevant block. We use these indices to bring the price of all valid historical sale and list prices to current values. By controlling for both time and location prior to fitting the valuation models, we obtain a much larger model dataset than if we limited ourselves to only using closed sales prices over the previous one or two years.

To show the effect of time adjustment at the block level, Figure 3 shows the empirical distribution of the resulting time-adjusted current prices in Block 3. Two distinct price populations emerge in this block: condos, with a median value of $210,000 (blue distribution adjusted from blue price index in Figure 2, Block 3), and single family homes, with a median value of $445,000 (yellow distribution adjusted from the yellow price index in Figure 2, Block 3).

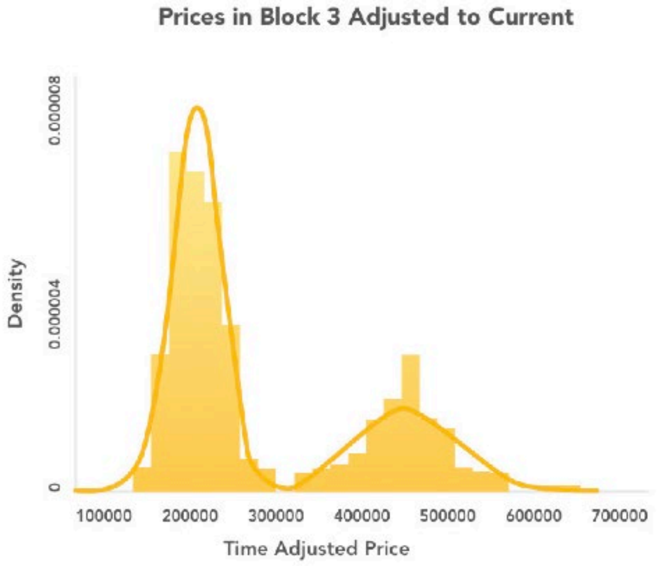

Figure 3: Effect of time adjustment at the block level

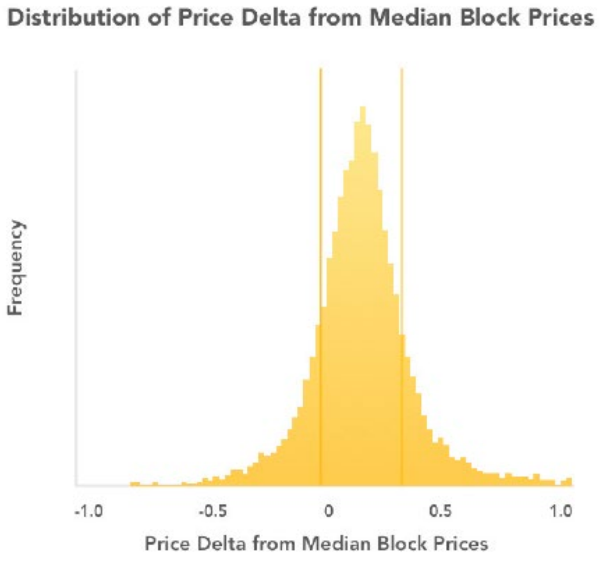

Figure 4: Time-adjusted price deltas from block median

Once the time-adjusted prices are calculated, the final step in generating the dependent variable is to apply a transformation to center the prices around their respective current block median price for each property type. We center as a percent deviation from the current block median estimated price for each property type.

Figure 4 shows the distribution of the price deltas across the entire census tract. The yellow vertical lines represent the empirical 16th and 84th percentiles, i.e., the bounds in which approximately 68% of the data resides (one standard deviation). In the example census tract, these values correspond to approximately +/-0.14. In other words, if we only used the estimated current median block price, the property type, and the date of occurrence to estimate all known historical prices in this census tract, we would be within 14% of actual historical price approximately 68% of the time. The quantity 14% is referred to as model forecast standard deviation (FSD), which is explained in-depth in the "Forecast Standard Deviation" section.

The deltas in Figure 4 represent the remaining unexplained variance among all historical prices in the census tract after we have accounted for the property type, time, and location in which those prices occurred. The final step seeks to further explain this unexplained variance in terms of many other observable data inputs.

Step 3: Train valuation models from time-adjusted historical prices
The third and final step involves fitting machine learning-based models to explain the price deltas shown in Figure 4. In this step we account for many more variables beyond the property type, time, and location than were accounted for through the index models. Depending on the data available, these models can include property characteristics, neighborhood characteristics, macro- and microeconomic data, spatial relationships, repeat price observations, and much more, as described in the "Valuation Data" section above.

Using the transformed response variables and varying subsets of the many predictor variables, we build and run multiple different machine learning models for each census tract. The data density of each predictor variable affects its significance level in value estimation for each census tract and property, which is accounted for in the different models. For example, one model may contain lot of listing data, but if a given subject property has not been listed or sold in recent history, that model may not be as strong as another model that contains more geo-location and property detail input variables. The multiple models are meant to leverage the factors that best relate to a given property, since not every property is the same or has the same amount of data available. Together, the models make up a single algorithm used to predict the value of a given property. Figure 5 shows how the final weighted historical price estimates compare to actual observed historical prices in this census tract.

The primary limitation for any of the individual component models revolves around the amount, density, and correctness of available data. Each model is statistically consistent in that it will get better with access to more localized data and more detailed property characteristic information. This consistency implicitly assumes that a large portion of the data is correct and representative of the actual underlying housing stock.

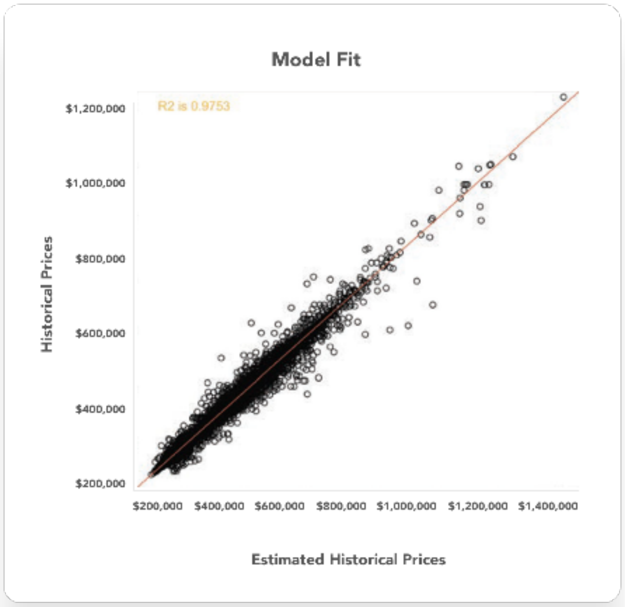

Figure 5: Model fit versus all observed historical prices

Model Validation and Outputs

Forecast Standard Deviation

Model-derived measure of uncertainty

We measure model uncertainty using the HouseCanary forecast standard deviation (FSD). This is a quantity that sets upper and lower bounds on the value estimate such that the range will actually capture the subsequent arm's-length sale price approximately 68% of the time (one standard deviation).

As an example, if the property FSD is 0.07, then the upper bound on value is given by P*(1+0.07) and the lower bound on value is given by P*(1-0.07), where P is the estimated price for that property. If an arm's-length sale were to occur shortly after the estimate was generated, there is a 68% probability that the estimated range would cover the actual realized price.

The confidence score is simply 1-FSD. If FSD equals 0.07, then the confidence score is 0.93. It is common to see both these quantities scaled up by a factor of 100, i.e., 7 and 93.

We train the FSD model on the census tract-level empirical error distribution after all individual component price estimates have been combined into a final estimate. At a high level, FSD depends both on the distributional spread in the tract-level empirical error distribution, and on how much agreement or disagreement exists among individual component price estimates for an individual property. Recall from step 2 of the "Algorithm Methodology" section that the empirical error distribution is the collection of percent deviations obtained by comparing estimated historical prices to actual historical prices. Figure 5 shows the actual historical prices on the vertical axis against the estimated historical prices on the horizontal axis. Figure 6 shows the empirical error distribution for these same observations.

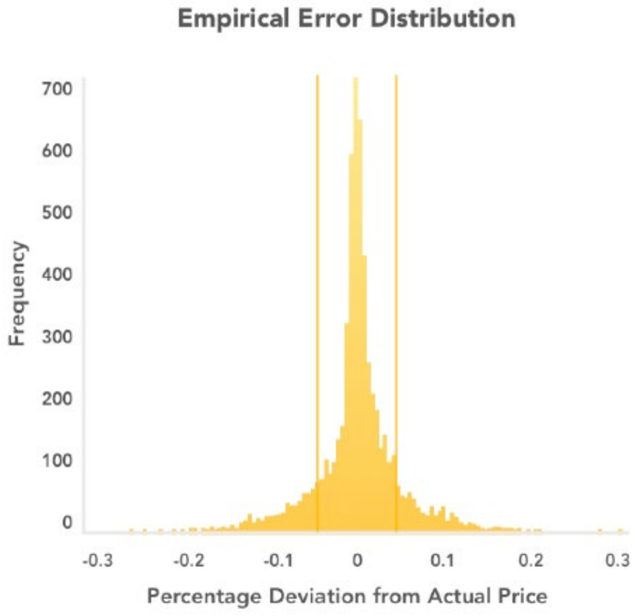

Figure 6: Empirical error distribution

The vertical yellow lines indicate the 16th and 84th error percentiles, so that 68% of all the deviations fall within these yellow lines. In this example, these yellow lines are approximately +/-0.045, meaning that the average FSD for properties within this census tract is approximately 0.045. An individual property's FSD could be much larger than 0.045 if the estimated value lies in a tail of the estimated price distribution and/or if there is a high degree of disagreement among the individual component models for that particular estimate.

We assume symmetry but not normality when looking at the empirical error distribution, and normality is often violated. Because normality is not assumed, the empirical rule cannot be applied to obtain other coverage probabilities; using 2*FSD to estimate a new interval is not guaranteed to yield an approximate 95% coverage probability.

Testing and Validation

Continuous internal testing + quarterly third-party testing
After we train a model, we test it using a new set of input variables, different from the data used to build the model. We again back-transform the predictions generated from the new data into value estimates from percent deviations. Using these results, we calculate accuracy measures including the median absolute percent error (MdAPE) rates. The complete list of accuracy measures is defined below:

1. **hit_rate**: The proportion of sold properties in which we had an estimate of value prior to the sale.
2. **Median_Abs_Pct_Err**: The 50th percentile of absolute error in percentage terms, or MdAPE. In other words, if this value equals 3.0%, then half our estimates were within +/-3.0% of actual sales price, and half were outside +/-3.0% of actual sales price.
3. **Median_PCt_Err**: The 50th percentile of actual error in percentage terms (not absolute error). Values close to zero imply that the estimator is unbiased.
4. **Within X%**: The percent of estimates that fell within +/-X% of actual sales price. HouseCanary produces this value for the 5%, 10%, and 20% bounds.
5. **Within_HC_Prediction_Interval**: The percent of actual sales prices that fell within HouseCanary's upper and lower estimates of value. The coverage probability of this interval is 68%. Therefore, this value should fall somewhere close to 68%, and values near 68% indicate that our model is accurately measuring the unexplained variance in price. As error rates continue to decrease, the width of the intervals will get smaller while still maintaining the target 68% coverage probability.

HouseCanary's valuation models are refreshed and tested internally monthly, and are also tested by a third party quarterly. A moving six-month testing window is used for the internal validation, meaning the test set property sale prices start six months back, and include the most recent month. The six-month moving test window allows for measurement of many geographic areas, even those with long delays in providing sales price data. It is not uncommon for some regions to have a three to four month delay in reporting sold prices via the county recorder and/or MLS.

As of July 23, 2019, HouseCanary's continuous internal testing over the previous six months yielded a national MdAPE of 2.8% on 1,994,203 transactions. In addition to internal testing, HouseCanary undergoes quarterly third-party testing. On a blind sample, the most recently completed third-party test measured over the first quarter of 2019 yielded a national MdAPE of 2.9%, compared to actual contract price.

Detailed test results, including MdAPE and the other metrics above, are available by request at the national, state, and MSA levels. At the national level, HouseCanary's internal results are further available by property type and by the month in which the closed sale price occurred.

Dynamic Capabilities

Real-time valuation and updates to existing valuations

As of this writing, the current algorithm gets trained and produces static nationwide estimates of value on a monthly cycle. In most cases, we store these values and use them as our best estimates of value until we run the algorithm again the following month and produce a new set of estimates.

There is one notable exception to the process above: when HouseCanary receives a new property listing from one of our direct MLS feeds, the new information from that listing is sent to our backend system. On the backend, a set of model objects stored from the most recent model run takes the new listing information as input and instantly updates the valuation based on the new information. This mid-cycle value estimate will remain in place until the next regularly scheduled model run.

About HouseCanary

HouseCanary's research team includes PhD statisticians, economists, and mathematicians. Learn more at **www.housecanary.com/about**.

Contact

Please contact us with any questions or comments at **sales@housecanary.com**.