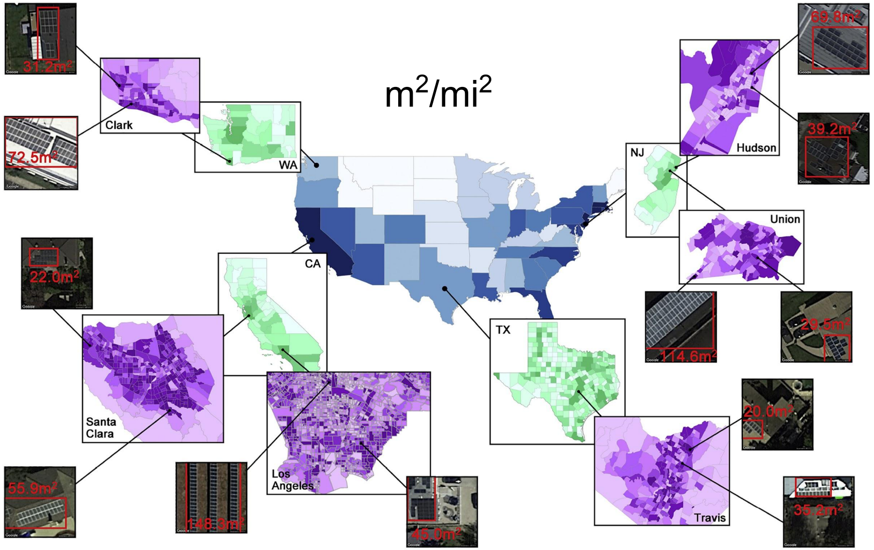
DeepSolar

Jordan Landers | Thinkful Unit 3 Capstone

Deep Solar Dataset



state level	county level				census tract level					
contiguous U.S.	CA	TX	NJ	WA	Santa Clara	Los Angeles	Travis	Union	Hudson	Clark
<1.34	<0.69	<0.03	<132	<0.11	< 255	< 618	< 70	< 323	< 11	< 0.84
1.34 - 3.45	0.69 - 6.53	0.03 - 0.15	132 -507	0.11 - 0.27	255 - 921	618 - 1181	70 - 205	323 - 598	11 - 327	0.84 - 22.53
3.45 - 4.34	6.53 - 21.18	0.15 - 0.27	507 - 729	0.27 - 0.51	921 - 1526	1181 - 1579	205 - 301	598 - 1092	327 - 743	22.53 - 49.59
4.34 - 7.92	21.18 - 90.44	0.27 - 0.46	729 - 1194	0.51 - 1.33	1526 - 2152	1579 - 2027	301 - 435	1092 - 1499	743 - 1460	49.59 - 79.60
7.92 - 14.24	90.44 - 165.9	0.46 - 1.03	1194 - 1325	1.33 - 1.90	2152 - 2864	2027 - 2531	435 - 647	1499 - 2022	1460 - 3331	79.60 - 95.80
14.24 - 24.12	165.9 - 281.9	1.03 - 1.71	1325 - 1814	1.90 - 5.95	2864 - 3753	2531 - 3176	647 - 842	2022 - 2942	3331 - 6598	95.80 - 131.59
24.12 - 76.74	281.9 - 446.2	1.71 - 3.98	1814 - 3027	5.95 - 11.68	3753 - 5071	3176 - 4079	842 - 1319	2942 - 4398	6598 - 9971	131.59 - 165.94
76.74 - 224.1	446.2 - 1088.1	3.98 - 11.95	3027 - 4148	11.68 - 24.20	5071 - 7490	4079 - 6366	1319 - 2197	4398 - 7960	9971 - 38058	165.94 - 281.21
> 224.1	> 1088.1	> 11.95	> 4148	> 24.20	> 7490	> 6366	> 2197	> 7960	> 38058	> 281.21

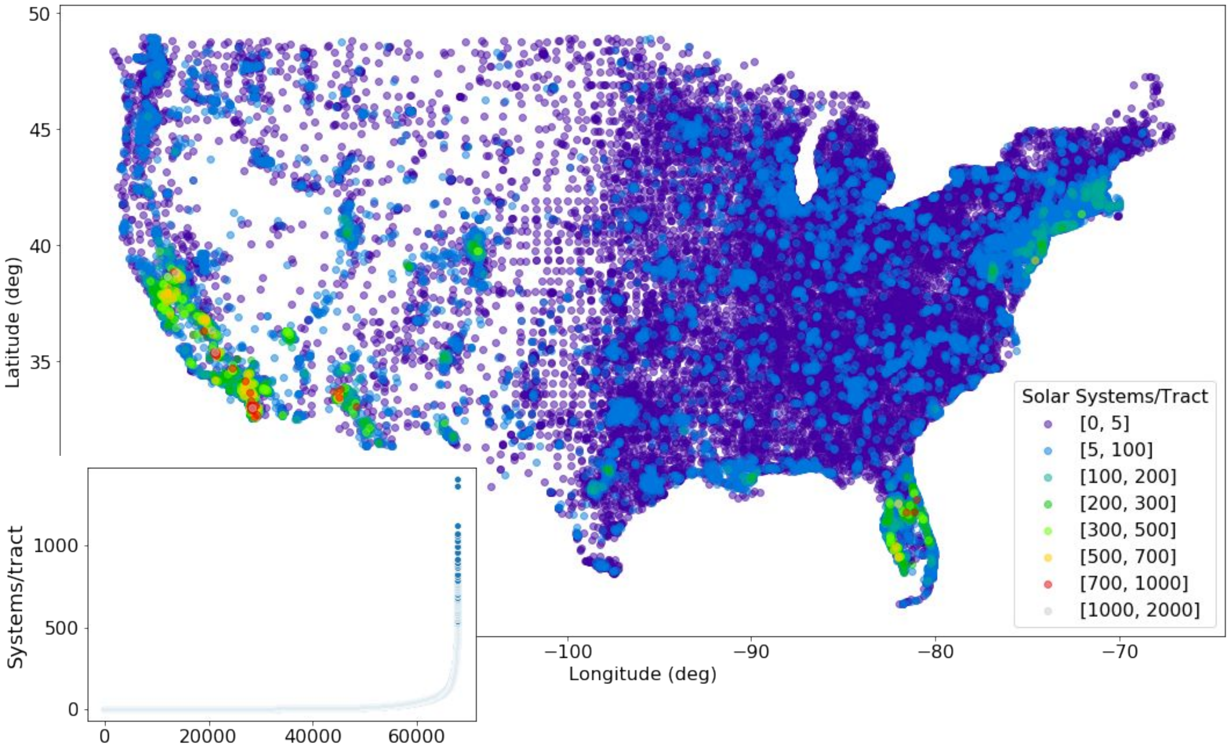
- Solar panel systems were counted from satellite images using a semi-supervised convolutional neural network
- Published dataset includes solar, geographic, and demographic data for 48 contiguous states at census tract resolution

Jiafan Yu, Zhecheng Wang, Arun Majumdar, Ram Rajagopal. **DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States**. *Joule*, 2018; 2 (12): 2605 DOI: 10.1016/j.joule.2018.11.021

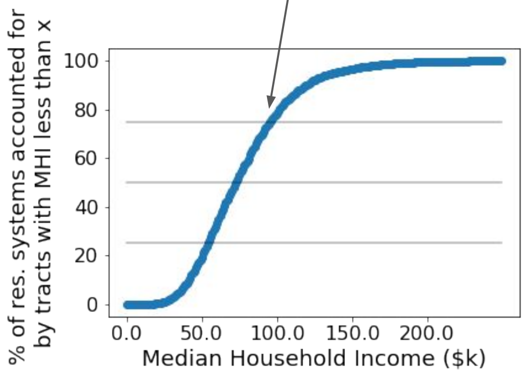
Data Prep

- `voting_2012_dem_percentage`, `voting_2012_gop_percentage` were dropped because they were not reported in CO, CT, FL, GA, SC, UT, WY.
- Geographic/climate related values missing from 5208 tracts, affecting as many as 25% of tracts in certain states
    - For sets of missing tracts accounting for less than a threshold percentage of the respective county land area, missing values were replaced with county averages. 0%, 25% and 30% were investigated as threshold percentages.
- Dropped `null`, `infinite`, and `NaN` values
- Dropped categorical variables

Target Variable: `solar_systems_count_residential`



~75% of solar installations accounted for by median household income < $100k

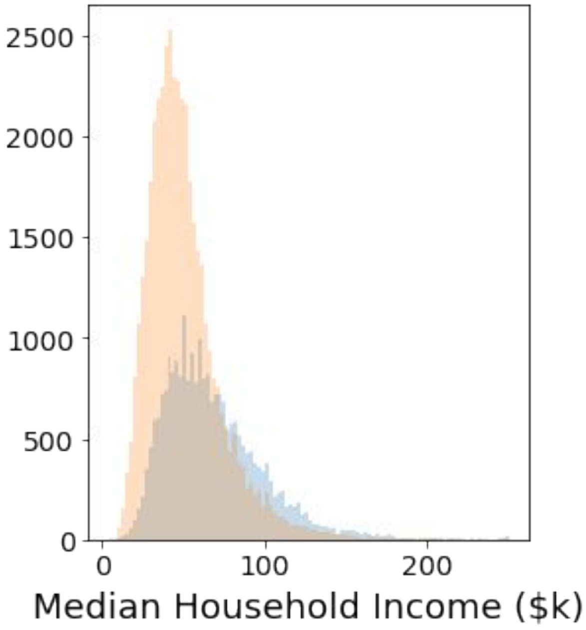


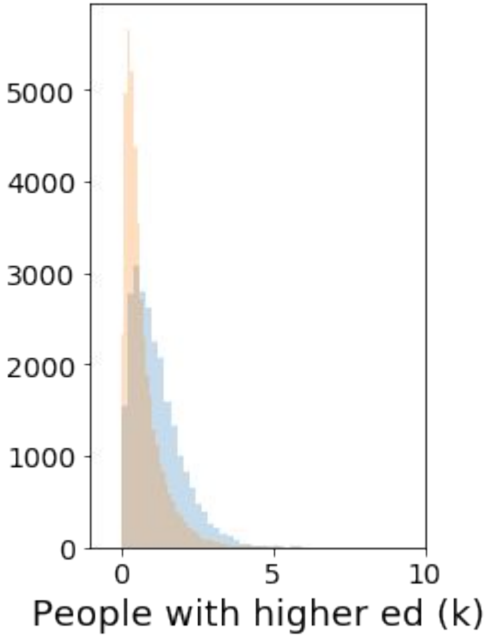
92% of households have income < $100k

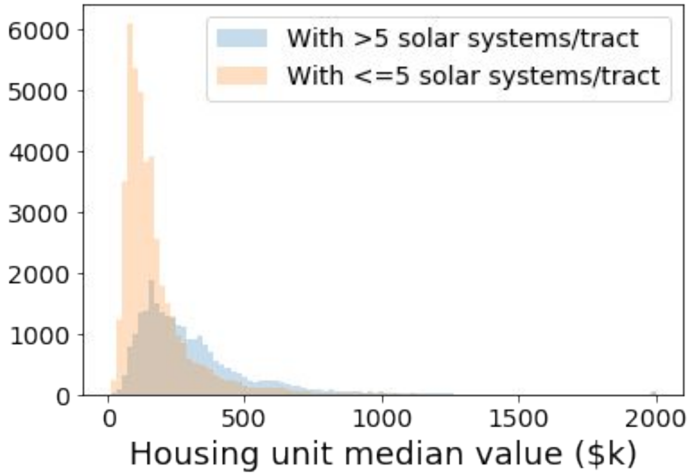
Challenges

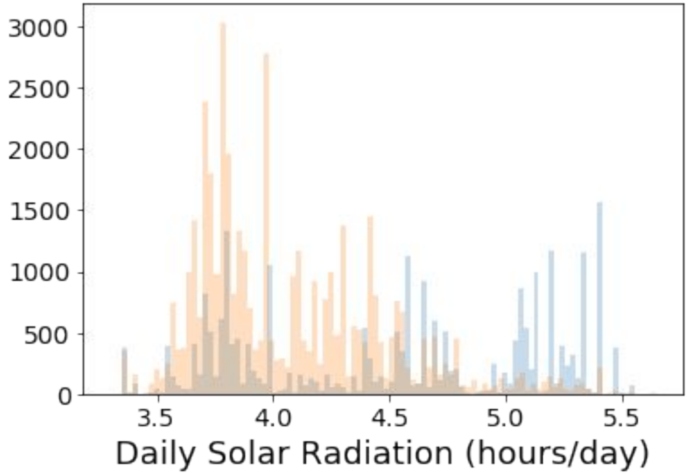
- Data reflect cumulative effect of various economic conditions and incentive programs, demographic data is a snapshot that may/may not reflect the characteristics of the population that installed solar systems
- Cost of living varies so money related variables are locally calibrated
    - $67k (average `median_household_income` in CA) represents less buying power in CA than it does in TN, but to the model $67k is $67k
- Class imbalance (1.13% of households in dataset have solar installations)
- Inconsistent distribution
    - 62% of residential solar systems are located in CA (skew of residential solar systems = 7.4)
    - CA represents 11% of households
- Data do not reflect a system at saturation
    - Solar is being adopted at an increasing rate, so a household without solar is not necessarily a “non-solar household” as much as “not *yet*-solar household”

Exploring the Data



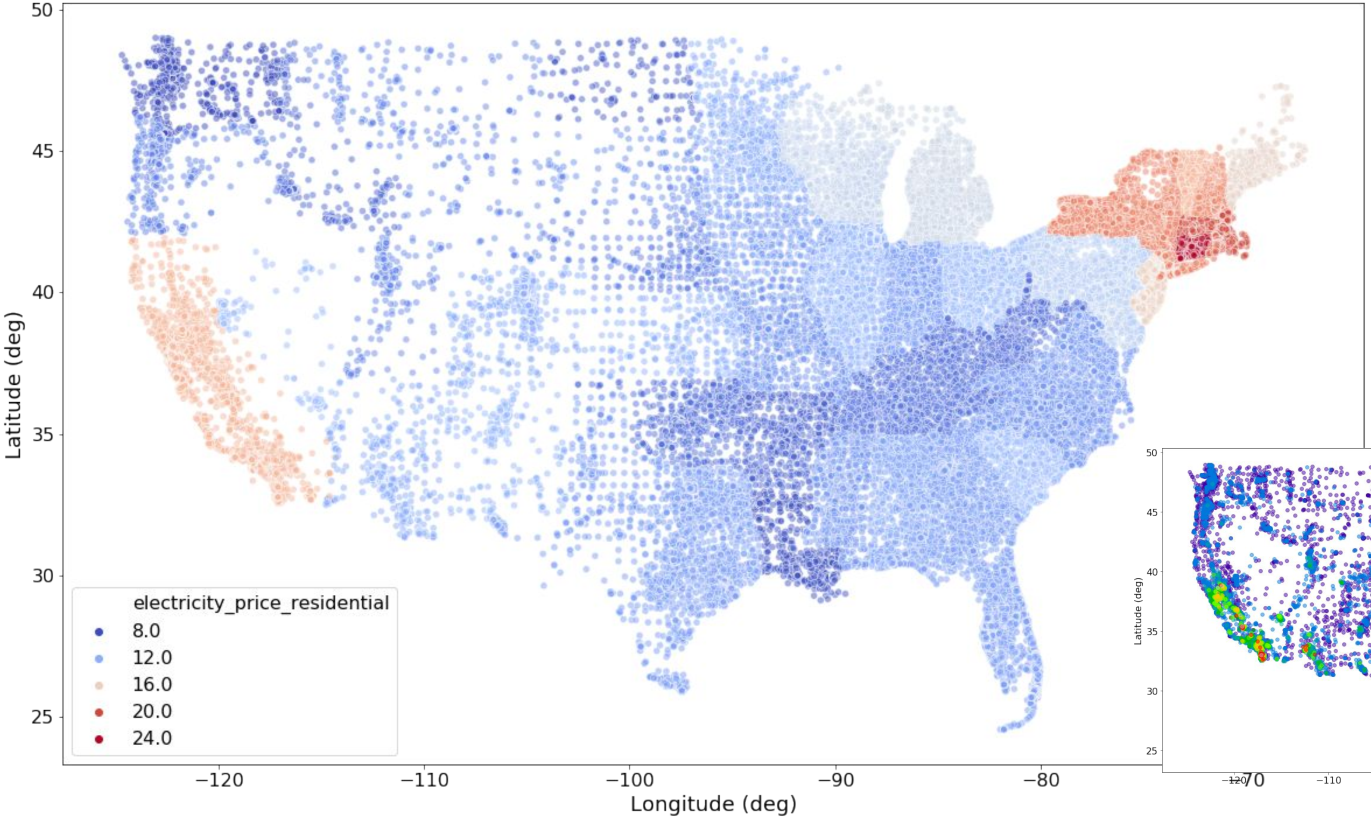


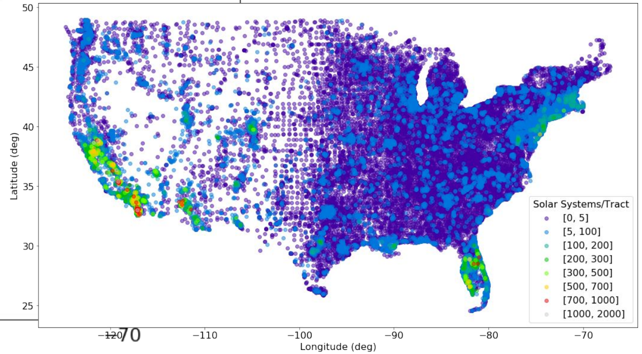




Populations are statistically different (p=0) but distributions clearly overlap.

	Median Household Income	Count Higher Ed	Housing Unit Median Value	Daily Solar Radiation
> 5 installations per tract				
count	24533	24533	24533	24533
mean	68653	1176	311485	4.54
Standard dev.	31159	884	233107	.63
min	9100	6	9999	3.35
max	250001	20482	2000000	5.65
<= 5 installations per tract				
count	43403	43403	43403	43403
mean	50779	702	173116	4.07
Standard dev.	23723	648	147227	.4
min	4147	0	9999	3.3
max	250001	10471	2000001	5.65

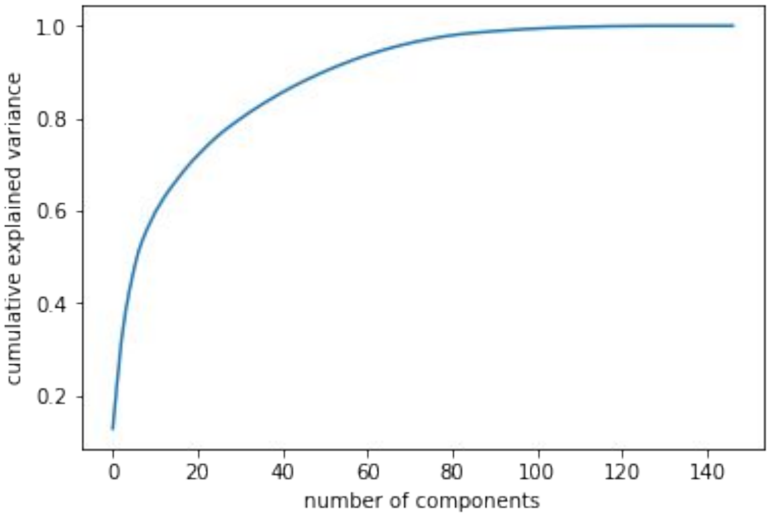




Indirect relationship between solar installations and electricity price

Feature Set

140 variables in published data set pertaining to residential Heating, Weather, Income, Education, Employment, Race, Age, Occupation, Commute, Political, Energy consumption/Pricing, Incentives



- PCA suggests that over 99% of variance can be accounted for using 92 principle components
- No strong feature loading in first 15 components

Modeling

Considerations

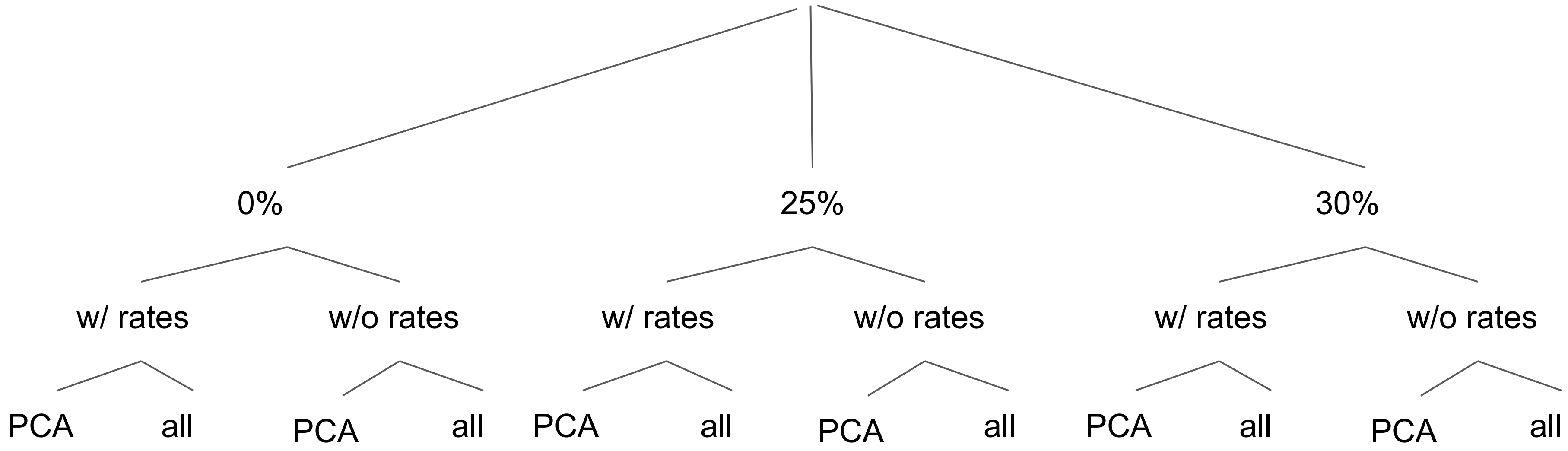
- Regression
- Complicated relationships between variables
- 100 - 140 variables
- 50,000 training examples

Model Candidates

- Ridge
- Random Forests
- Support Vector Machines
- Gradient Boosting

Model Sets

- 0%, 25%, 30% = threshold percent of county land requiring imputed geo values
- w/ rates, w/o rates = including/excluding rate variables from feature set
- PCA, all = reduce feature set using PCA, or retain all variables



Results: PCA (n=100)

* Same Training/Test sets used for all tests except where noted

	Ridge	Random Forests	SVM-R	Gradient Boosting-Regression
Training/Test 1	0.435 ∓ 0.028	.474 ∓ 0.045	0.608 ∓ 0.029	0.660 ∓ 0.035
Training/Test 2	0.434 ∓ 0.012	.471 ∓ 0.015	0.623 ∓ 0.042	0.640 ∓ 0.008
Training/Test 3	0.430 ∓ 0.031	.485 ∓ 0.018	0.627 ∓ 0.030	0.654 ∓ 0.034
CV score	0.433 ∓ 0.026	.477 ∓ 0.030	0.621 ∓ 0.034	0.652 ∓ 0.031
Total score	0.445 ∓ 0.004	.468 ∓ 0.019	0.594 ∓ 0.022	0.647 ∓ 0.012
Time to fit (s)	0.154 ∓ 0.006	448.2 ∓ 8.37	1961 ∓ 25.6	669.2 ∓ 6.6

Results: Full Feature Set

* Same Training/Test sets used for all tests except where noted

	Ridge	Random Forests	SVM-R	Gradient Boosting-Regression
Training/Test 1	.47 ∓ .031	.637 ∓ .029	.62 ∓ .025	.777 ∓ .02
Training/Test 2	.463 ∓ .015	.643 ∓ .02	.632 ∓ .039	.775 ∓ .028
Training/Test 3	.469 ∓ .011	.632 ∓ .02	.630 ∓ .01	.772 ∓ .015
CV score	.468 ∓ .021	.637 ∓ .02	.627 ∓ .028	.775 ∓ .022
Total score	.48 ∓ .011	.643 ∓ .023	.61 ∓.017	.771 ∓ .01
Time to fit (s)	.111 ∓ .005	390.8 ∓ 1.26	1452 ∓ 25.5	569.7 ∓ 2.04

Model Results: feature set comparison

* Same Training/Test sets used for all tests except where noted

	With rates	Without rates	30% threshold	No imputed data	No dropped columns
Training/Test 1	0.777 ∓ 0.02	0.775 ∓ 0.02	0.775 ∓ 0.027	0.77 ∓ 0.026	0.773 ∓ 0.019
Training/Test 2	0.775 ∓ 0.028	0.771 ∓ 0.026	0.776 ∓ 0.017	0.776 ∓ 0.014	0.781 ∓ 0.027
Training/Test 3	0.772 ∓ 0.015	0.767 ∓ 0.015	0.773 ∓ 0.036	0.774 ∓ 0.013	0.79 ∓ 0.011
CV score	0.775 ∓ 0.022	0.771 ∓ 0.021	0.775 ∓ 0.023	0.772 ∓ 0.019	0.781 ∓0.021
score	0.771 ∓ 0.01	0.77 ∓ 0.003	0.773 ∓ 0.005	0.781 ∓ 0.004	0.783 ∓ 0.006
Time to fit (s)	569.7 ∓ 2.04	584.2 ∓ 69.4	746 ∓ 45.4	1955.9 ∓ 1210.4	1573.5 ∓ 1215.0

Different training/test sets used

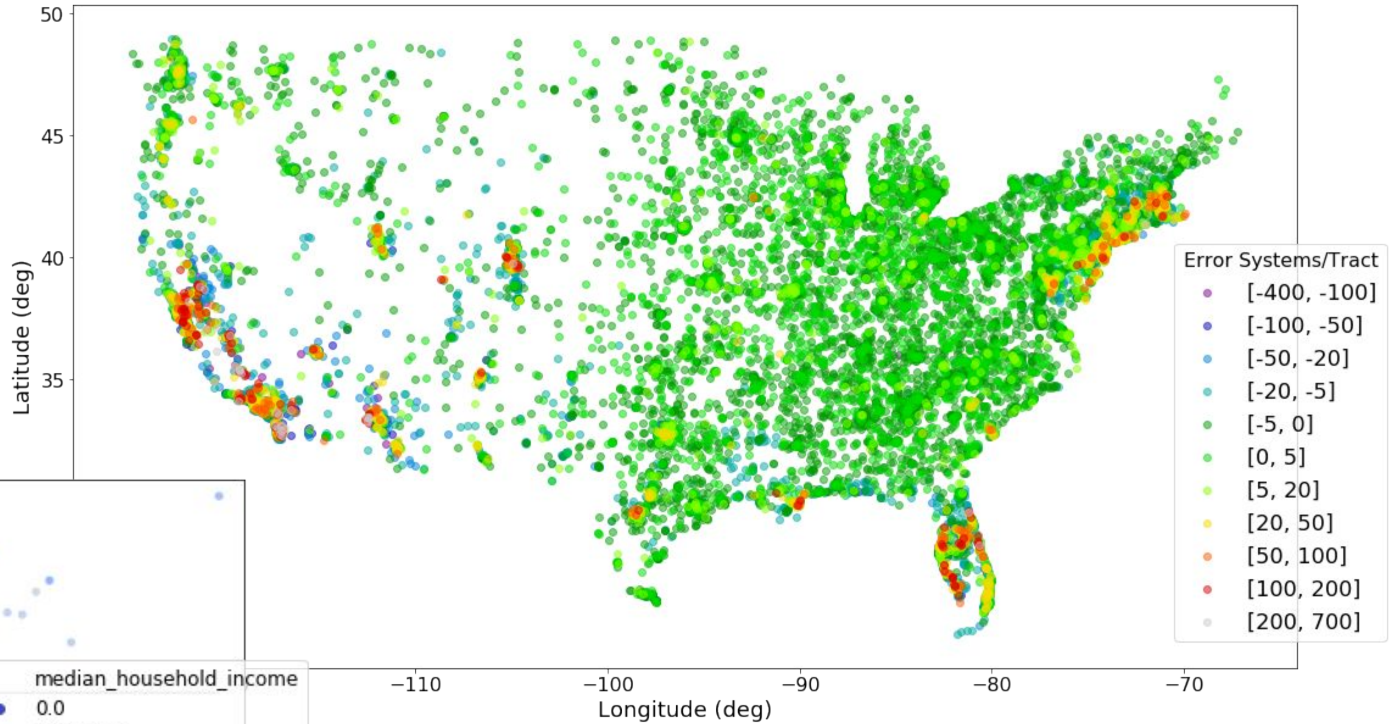
Model Optimization & Performance

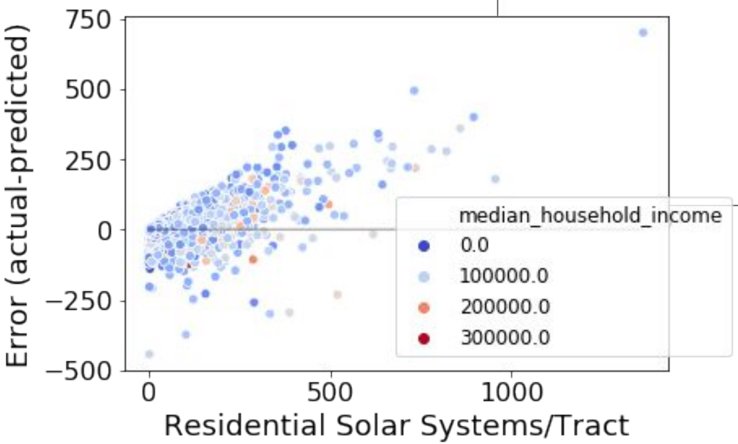
Used GridSearchCV to optimize performance of Gradient Boosting (~9 hrs)

	Option 1	Option 2
`criterion`	friedman_mse	mae
`loss`	ls	huber
`max_depth`	3	6
`n_estimators`	400	500
`learning_rate`	.01	.1
`min_samples_split`	2	4

Test score: 0.77 ∓ 0.007
CV score (n=6): 0.775 ∓ .023
Mean time to fit: 723.3 ∓ 113.3

Residuals & Conclusions





(-) → over predicted
Potential Markets?

(+) → under predicted
Unincorporated variables?

What's next?

Solar companies & public utilities have stakes in residential solar predictions

- Where to market solar?
  - Affinity Propagation to investigate groups of tracts that are similar to those that were overpredicted
- Where to prepare the grid for upcoming shifts in usage patterns?

Potential next steps:

- Clustering data and training separate models for each cluster
  - Separating data based on thresholds of single variables did not yield interesting results, but clusters based on unsupervised learning might do better