FairAir

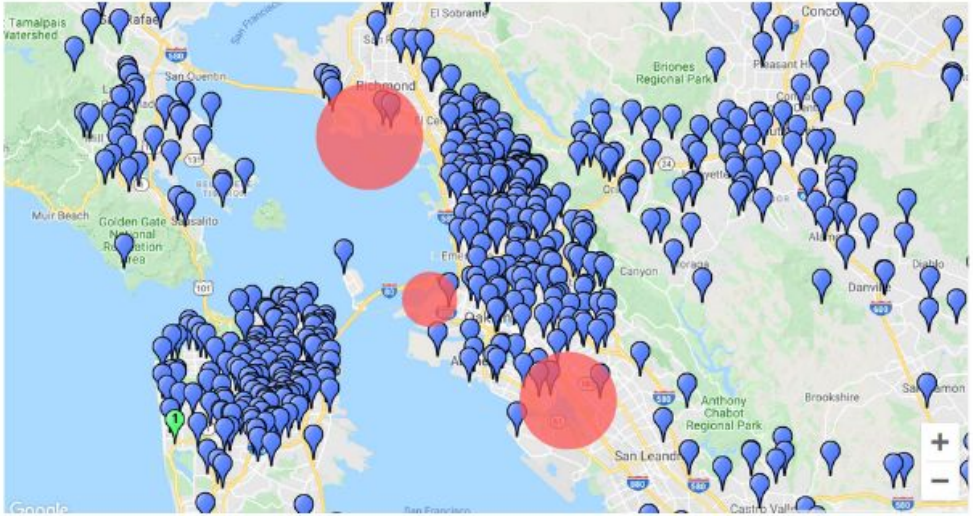
Filling the gaps in air pollution monitoring

Ben Arnoldy, Jake Miller, Angshuman Paul, Sameed Musvee, Mark Paluta

Air pollution isn't monitored where it is worst

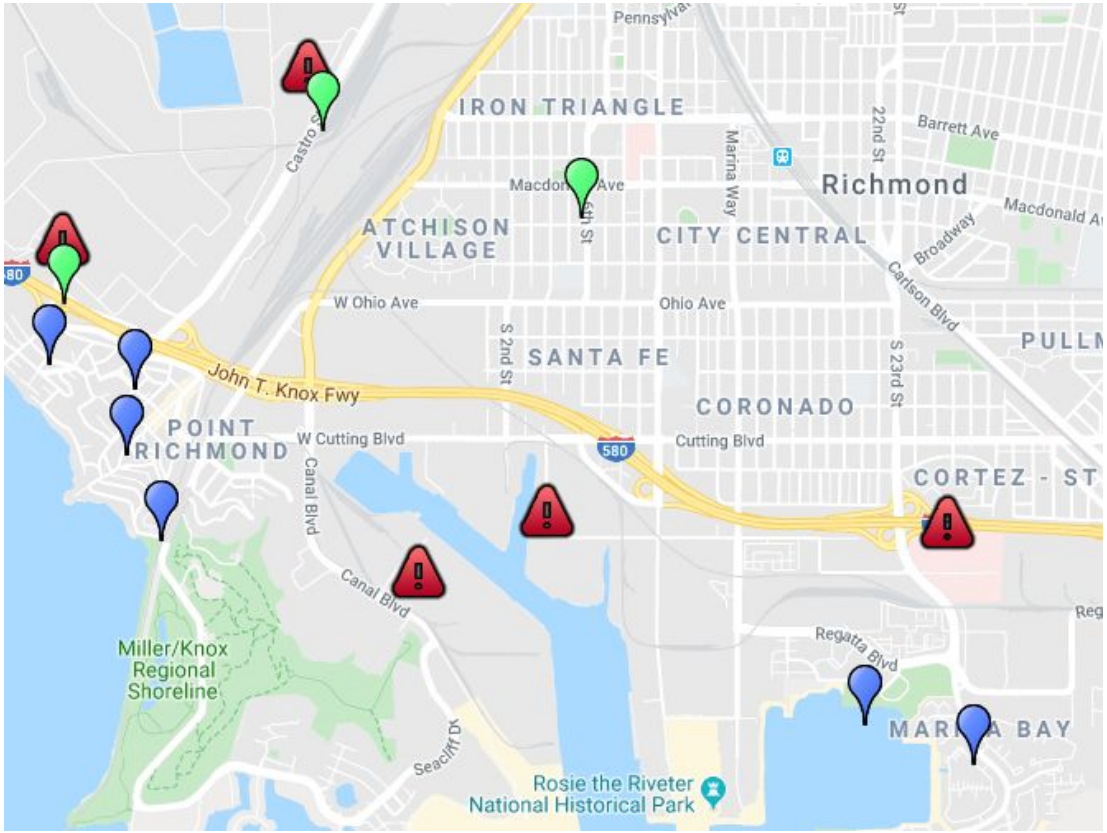
Sensor

Polluted area



Impact: 4.2M deaths worldwide every year

Demo of MVP

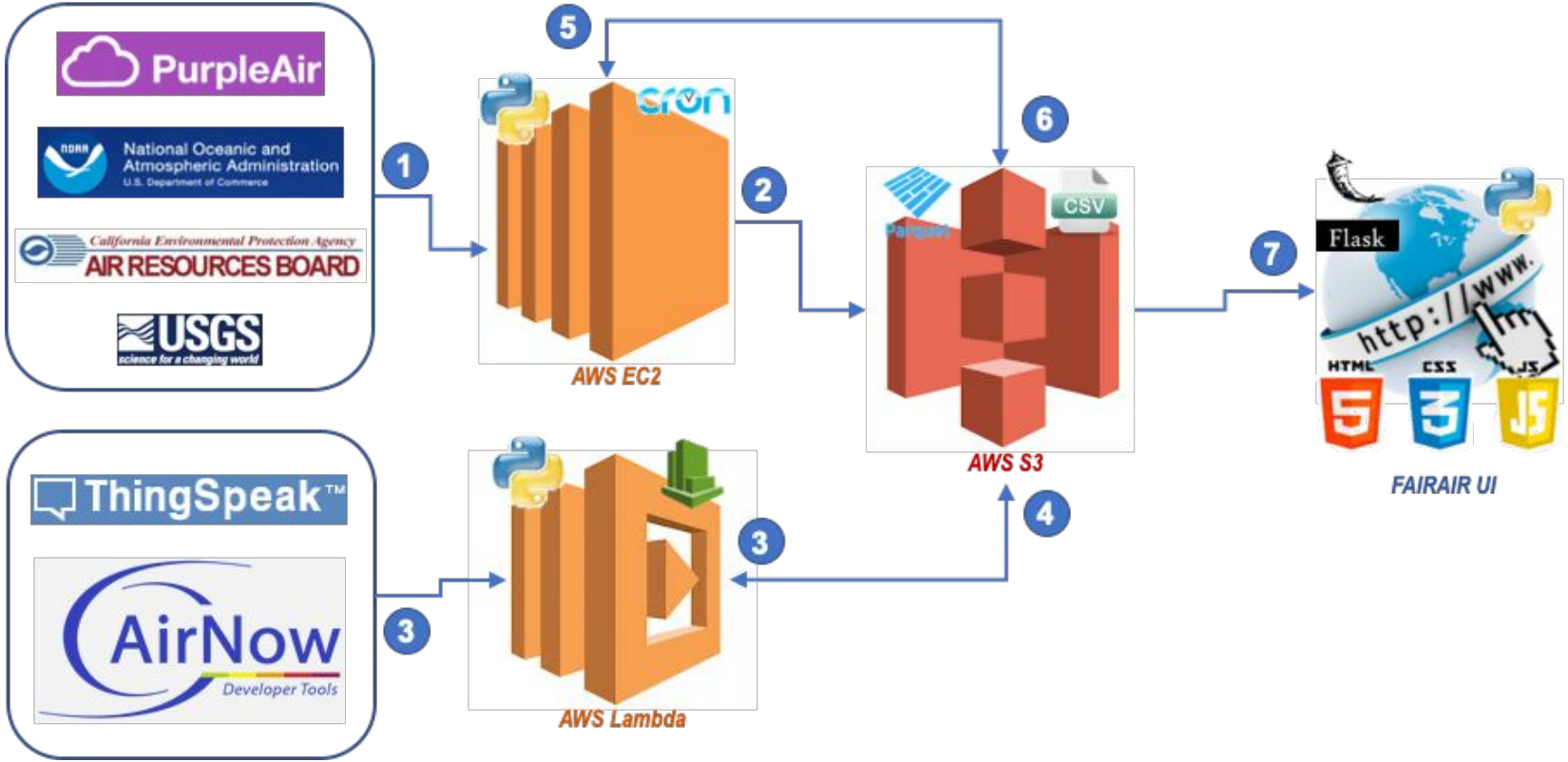


http://ec2-34-217-122-20.us-west-2.compute.amazonaws.com:8083/

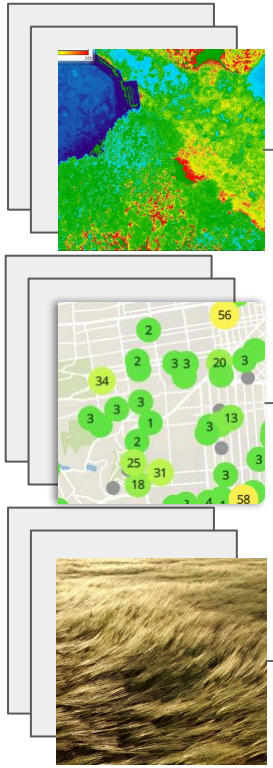
Data Sources

Source	Data Elements	Refresh Frequency	Time Grain in Data
Purple Air	Sensor data	Every 5 minutes	5 minutes
Thingspeak	Particulate , temperature and humidity data	Daily	10 minutes
National Oceanic and Atmospheric Administration	Wind and weather station data	Daily	5 minutes
Air Now	Ambient air data	Hourly	10 minutes
US Geological Survey	Topography, Elevation and Land Use data	One Time	N/A
California Air Resources Board	Known Polluters data	One Time	N/A

Data pipeline



Model Overview



Static Features (e.g., topography, vegetative cover)

Existing PA sensors

Hourly dynamic features (weather, ambient air quality)

Plotted on grid of 500m squares. Of 28,000 non-water squares, 508 have PA sensors

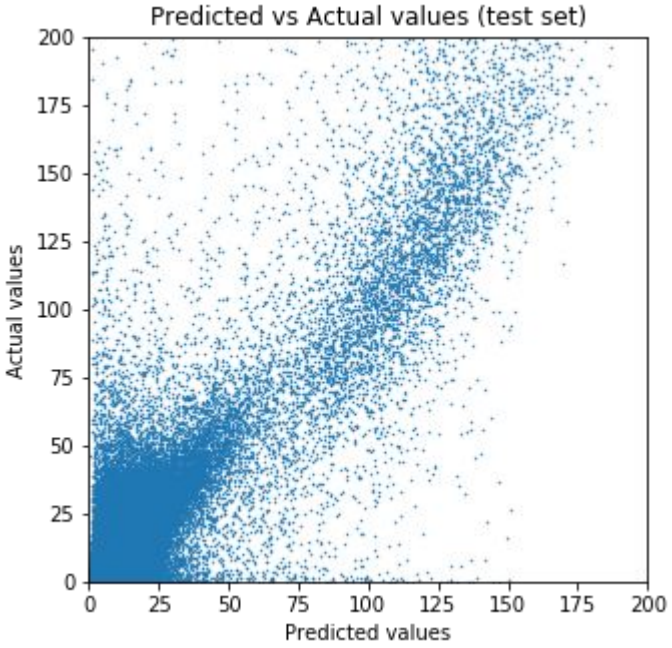
Random Forest Regressor

Proximity to each grid square to nearby sensors ("lonely factor")

Predicted PM2.5 level

UI

Model Selection and Evaluation

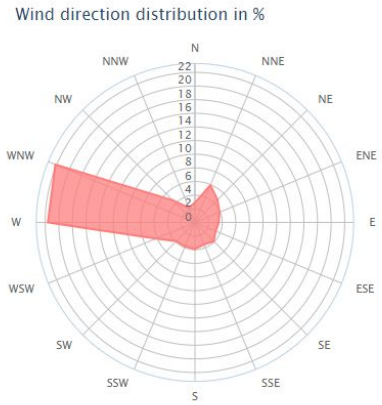


Most important features (using permutation importance): Ambient air quality, neighboring sensors, temperature. (Additional information in appendix).

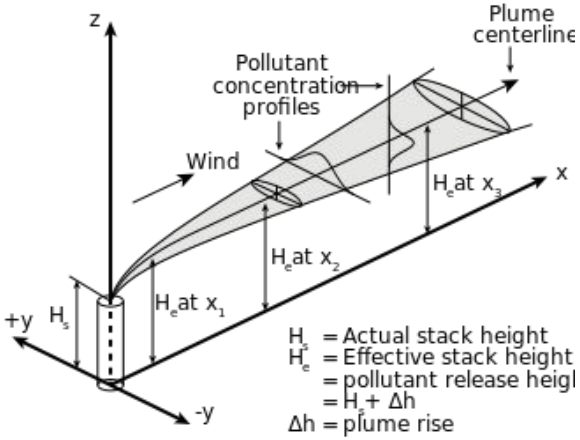
	FairAir	**Song and Han, 2019**
Location	Bay Area	Beijing
Sensor Type	Home	Mobile
Resolution	500m	1 km
Regressor	Random Forest	XGBoost-LR Hybrid*
IDW baseline RMSE**	14.8	17.8
Types of Features	6	62
RMSE	5.75	13.28
R^2	0.795	0.903

** XGBoost for encoding high-dimensional feature space; leaf path features then used as features in linear regression*
***RMSE for the test set when using inverse distance weighting spatial interpolation, no features; used only as baseline metric.*

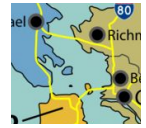
Key challenges / pivot



In certain areas, there was little variability among sensors with respect to wind.



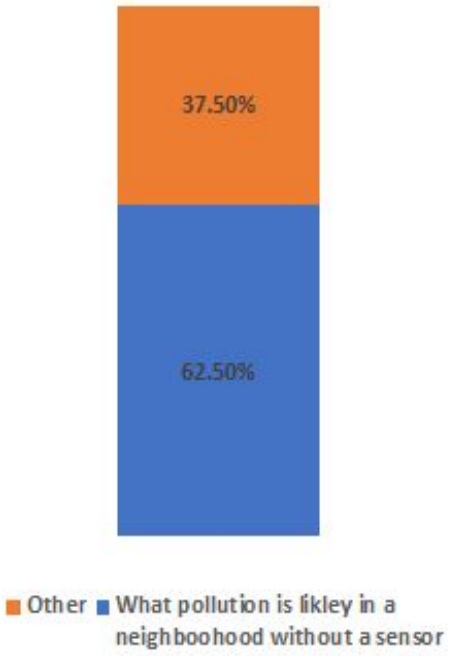
Gaussian Plume Dispersion: One of many air dispersion models that were not useful.



Increasing our area meant refactoring certain data processing scripts, but our solution is now more scalable and more generalizable.

User feedback

Very strong feedback to our updated idea, particularly predicting pollution in neighborhoods without sensors



User Quotes

“I like that the tool is dynamic, that it is adaptable to different regions which lets you decide what neighborhoods to focus on" - Dan Sakaguchi, Communities for a Better Environment.

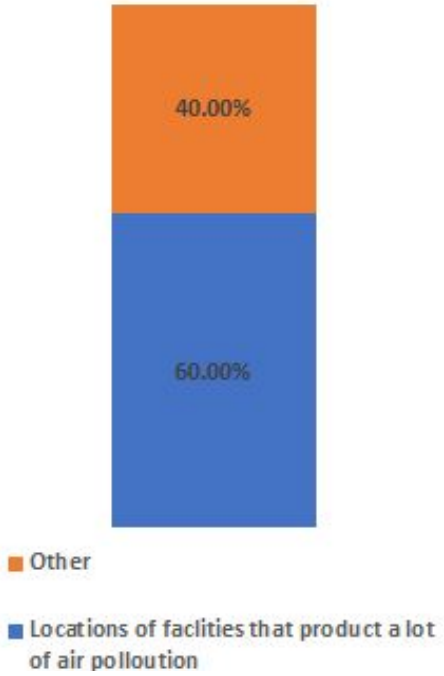
“How badass it is. This is such a cool idea and I love the interactive component you have included.”

“It is really neat to see that you are still recommending the same locations for sensors rather than just adjusting with the slightest shift in map zoom or pan. That really builds confidence that these aren't just willy nilly calculations.”

“It seems to be pretty responsive. Latency between updating the no. of sensors desired and those appearing on the map seems low.”

Future improvements

Polluters location was the #1 requested feature addition followed by real time air quality prediction



User Quotes

“It would be helpful in understand where pollution is coming from or a neighborhood I want to live.”

“I am more likely to care about this sort of tool from the perspective of social justice, advocacy, etc. Absent an a priori commitment to air quality issues, I am unlikely to think it's so chronically important that it should influence major life choices like where to live”

“I could imagine using this as a means to track how things like pollution from fires are affecting my area. I realize there are already tools for this, but it would be neat to include details like this. Or if you could automate sending warnings to subscribers when the air quality is particularly bad in their area then that would be neat.”

Wrap up

- Outdoor air pollution is a major killer.
- **FairAir** uses machine learning to predict pollution levels where there are no sensors — and to recommend where to deploy new sensors.
- Go to fairair.netlify.com to use the San Francisco Bay Area map tool and find local organizations fighting for clean air.

A low-cost PurpleAir sensor

Appendix: Feature Importance

Feature	feature_importance	permutation_importance	feature_importance_rank	permutation_importance_rank
imputed_epa_pm25_value	0.557002	1.02331	1	1
neighbors_total	0.160226	0.143997	2	2
imputed_temperature	0.0490909	0.102274	7	3
ndvi	0.0524636	0.096655	6	4
imputed_hum	0.0539146	0.096074	5	5
elevation	0.053955	0.0860544	4	6
wind_total	0.0733475	0.0644375	3	7

Notes:

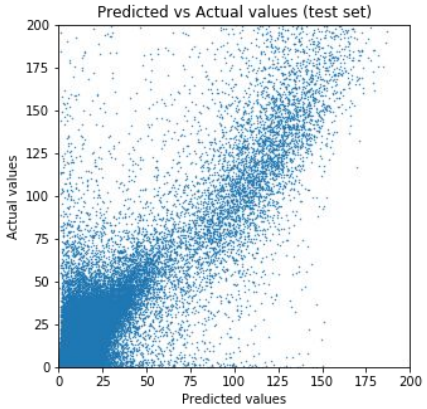
- Permutation importance is considered a less biased than scikit-learn's RF feature importance
- EPA PM2.5 comes from the nearest ambient EPA-sanctioned sensor. There were 8 in our bounding box
- The real purple air values from a 5x5 grid around the target grid were each used as features. Individually, each grid square was less important, but together they are second most important. Empty grid squares were filled with 0.
- Temperature was not judged important by scikit-learn's feature importance, but was ranked third by permutation importance.
- NDVI stands for normalized difference vegetation index is a satellite image-based measurement of vegetation for a given location.
- In the model, wind was broken out into vector components of x and y.
- Missing EPA, temperature, and humidity values were imputed using means from that timestep

Appendix: Notes on other models

- Several NN architectures scored around 15 RMSE (not good -- mean actual value was usually between 6-7, although a long tail past 300). Most simply guessed around the mean every time. They did not outscore inverse distance weighting (IDW) spatial interpolation.
- K Nearest Neighbors showed promise (RMSE around 8), but again mostly tended toward the mean and was computationally expensive. Other regressions (e.g., OLS) did not capture variance either.
- XGBoost was the closest competitor to the final random forest winner, but visualizing its predicted values shows how much it falls short:

RF winner:

- RMSE: 5.75
- R^2: 0.795



XGBoost, its nearest competitor:

- RMSE: 7.73
- R^2: 0.63