

Machine Learning for Design

Lecture 7

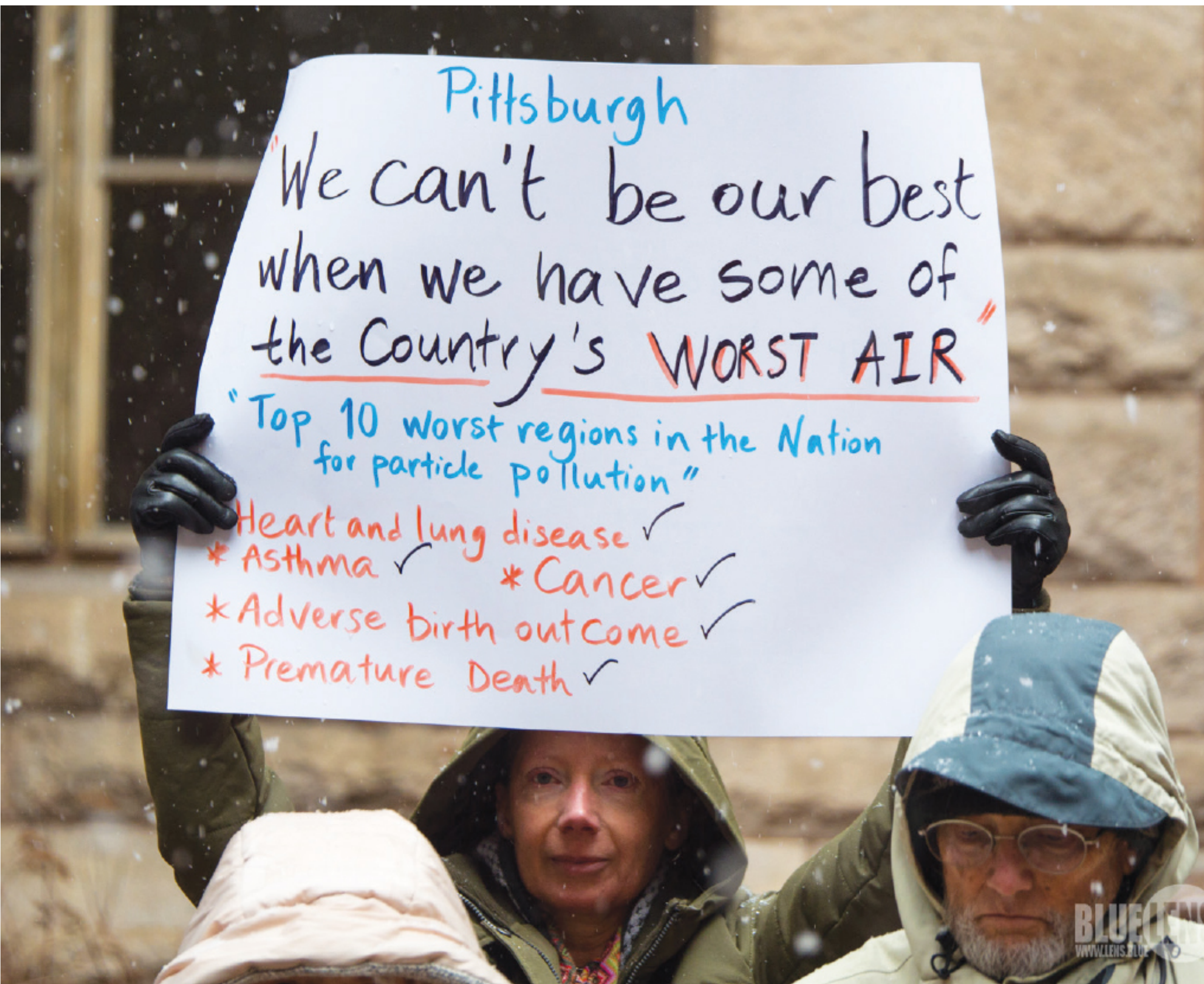
Design and Develop Machine Learning
Models - *Part 1*

**And now, let's
Smell Pittsburgh**

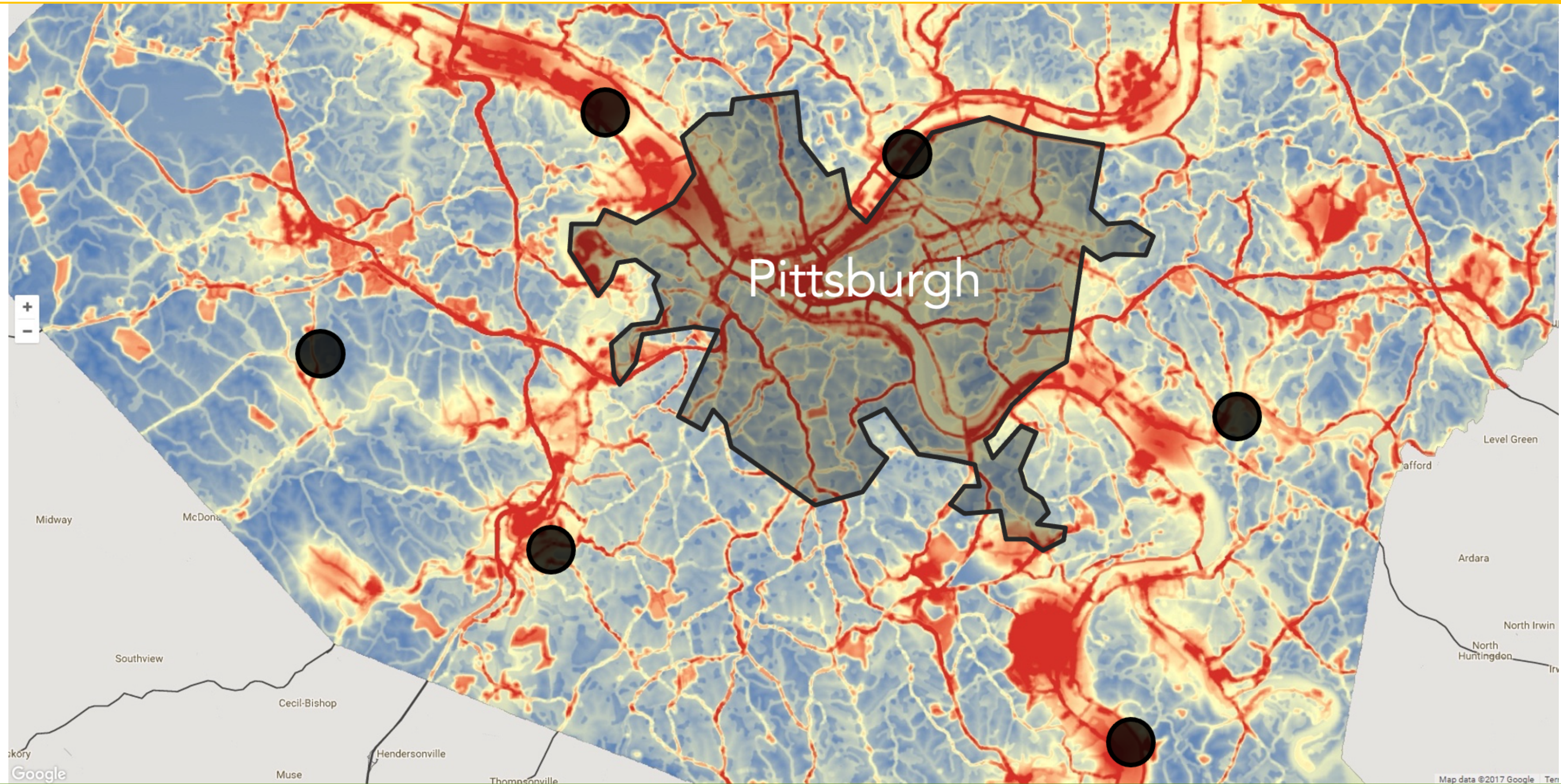
Credits: Yen-Chia Hsu

<https://yenchiah.me/>

According to the American Lung Association, **Pittsburgh is one of the ten most polluted cities (measured by particulate matter) in the United States.** Local residents have been fighting against air pollution for decades.

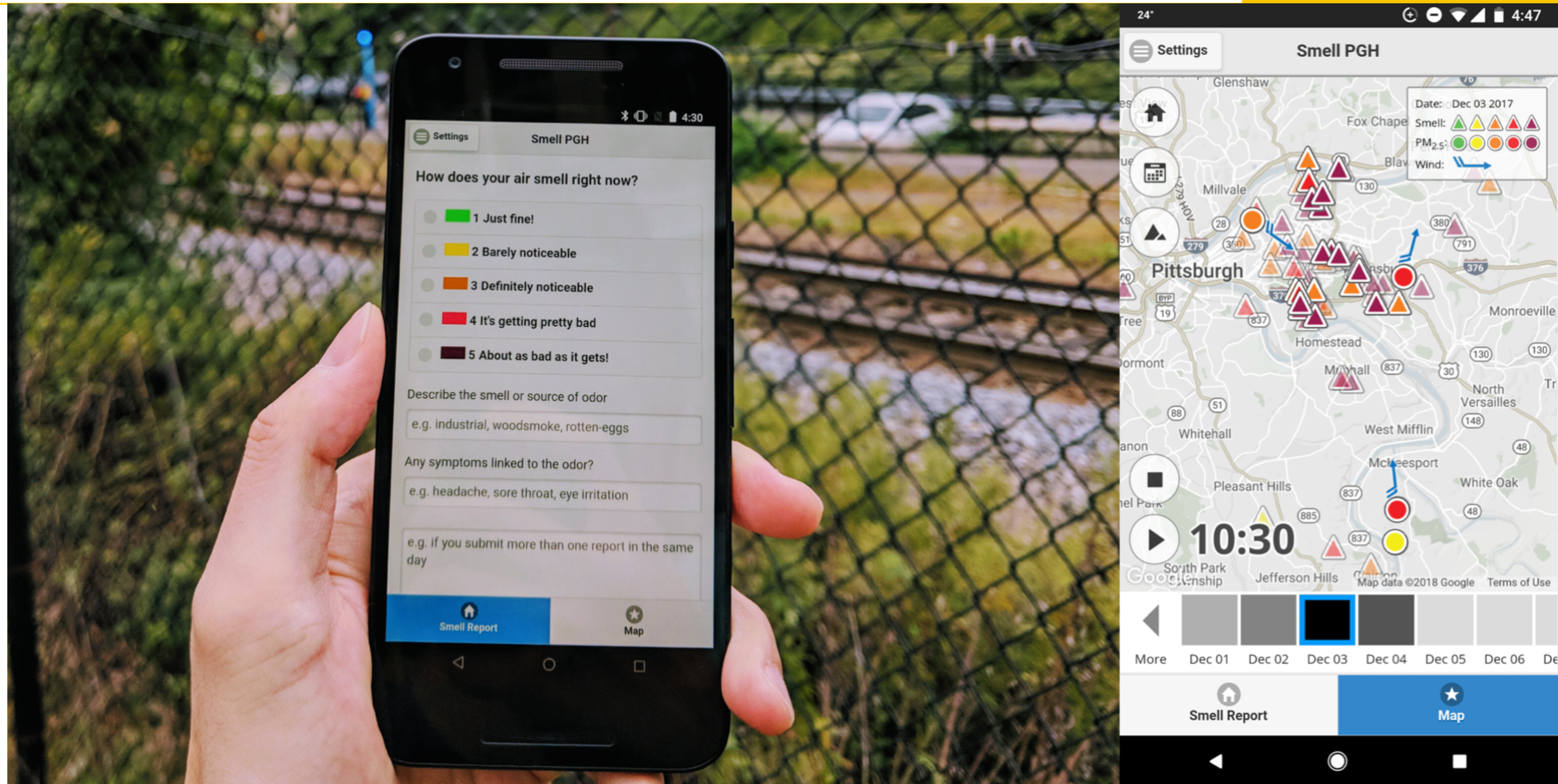


Local people have identified smell as an indicator of air pollution. But, how can we effectively **collect the smell experiences on a city-wide scale** with more than 300,000 residents over many years?








Link to the Pittsburgh pollution map — <https://breatheproject.org/pollution-map/>

Smell Pittsburgh is a mobile application that enables local communities to **contribute odor reports** in real-time (with accurate time and location information) and **visualize air pollution** collaboratively.



Link to the Smell Pittsburgh application — <https://smellpgh.org>

How does your air smell right now?

-  1 Just fine!
-  2 Barely noticeable
-  3 Definitely noticeable
-  4 It's getting pretty bad
-  5 About as bad as it gets!

Describe the smell or source of odor

Any symptoms linked to the odor?

Add a personal note to the Health Department

Smell Pittsburgh **predicts upcoming smell events** (based on the existing data at a certain time point) and **sends push notifications** to inform users while **encouraging engagement** in submitting odor data.

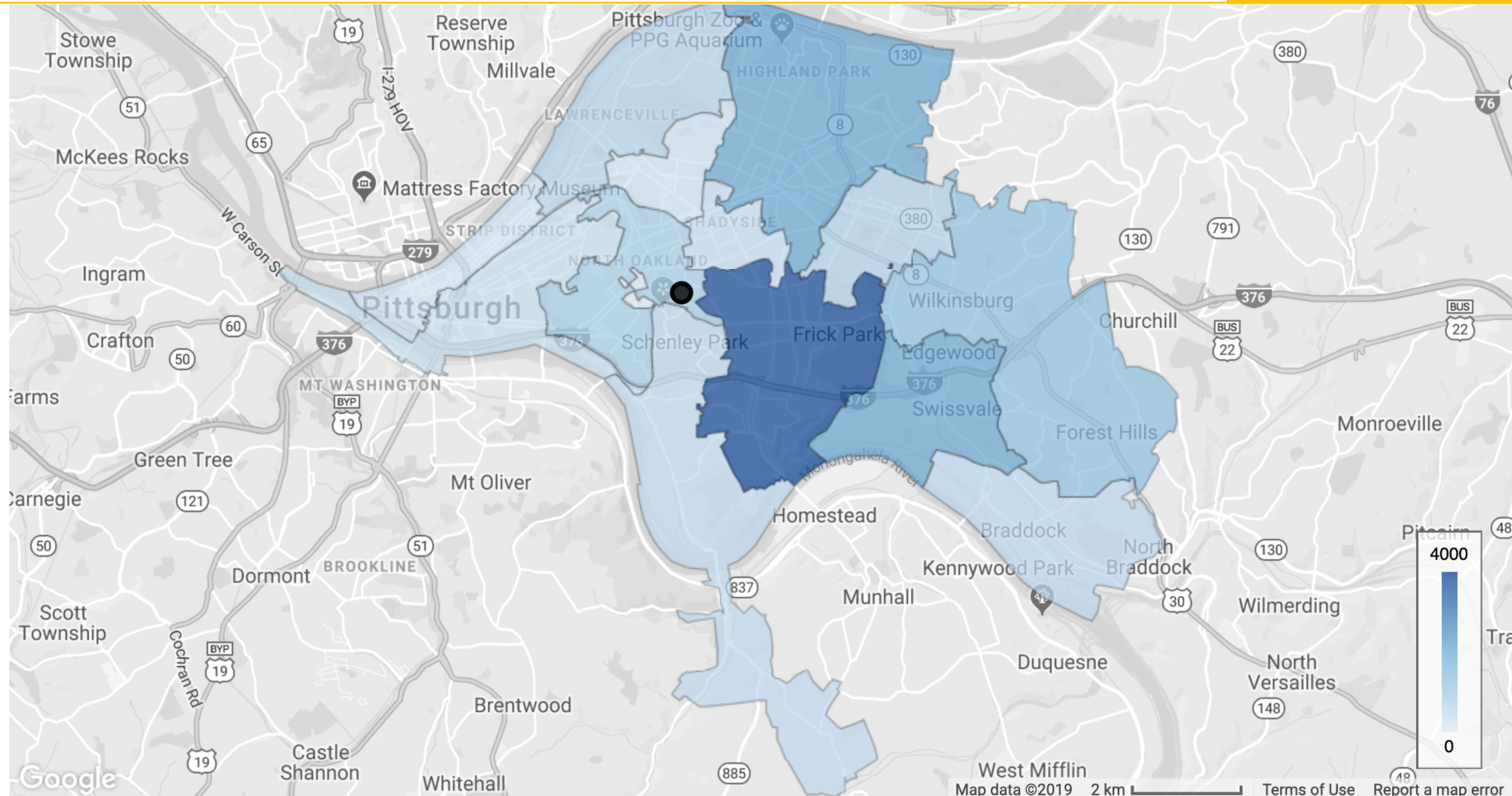


Smell Event Alert

Local weather and pollution data indicates there may be a Pittsburgh smell event in the next few hours.

Keep a nose out and report smells you notice!

A **geographic region in Pittsburgh** is manually selected when predicting the smell events. The black dot in the figure represents the location of Carnegie Mellon University.



Number of smell reports aggregated by zip codes in the dataset.

To predict the presence of bad odor within the next few hours, we need to **estimate a function that can map sensor measurements to smell events** as accurately as possible.

O ₃ : 26 ppb	CO: 127 ppb
H ₂ S: 0 ppb	PM _{2.5} : 9 µg/m ³
Wind: 17 deg	...

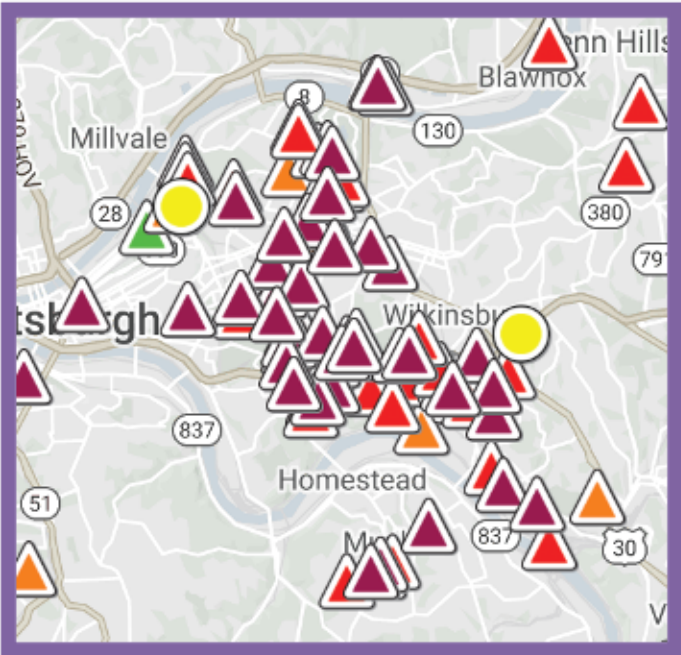
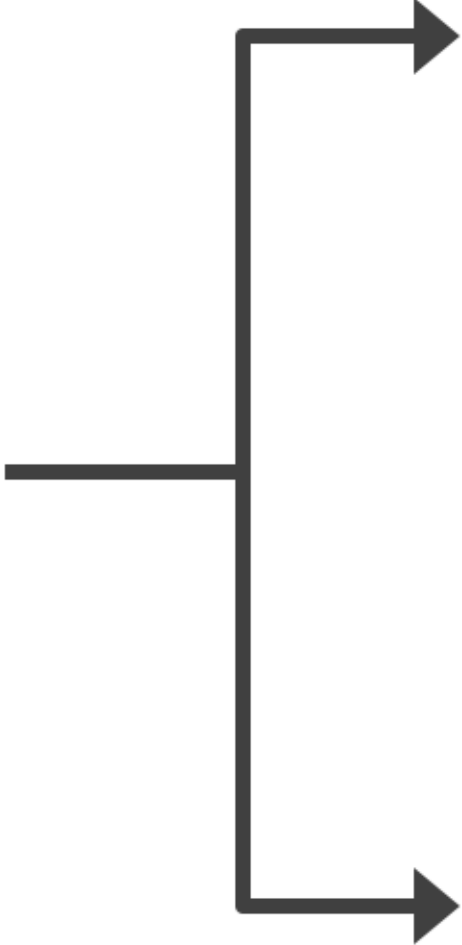
Observation 1

O ₃ : 1 ppb	CO: 1038 ppb
H ₂ S: 9 ppb	PM _{2.5} : 23 µg/m ³
Wind: 213 deg	...

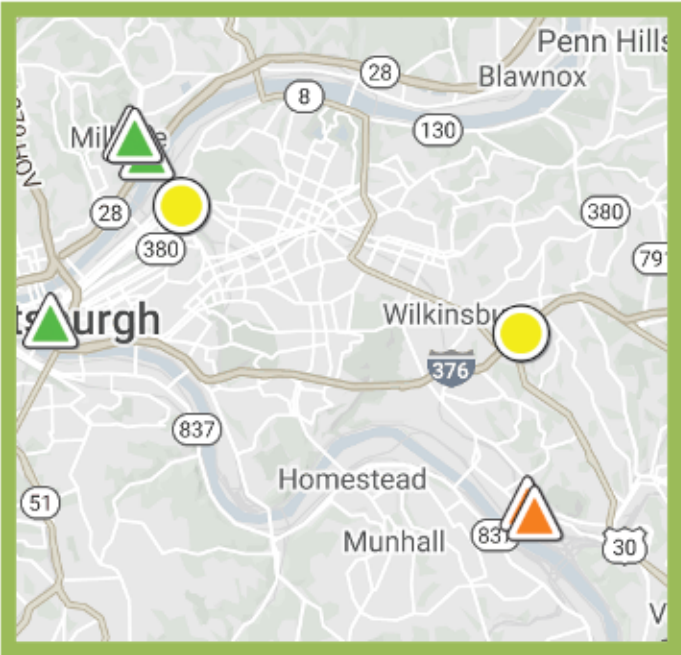
Observation 2



Sensors



☹️ Has Event



😊 No Event

One can technically use if-else rules to predict smell events. But such an approach can be laborious. **Can we do better than manually specifying these if-else rules** while minimizing human efforts?

O ₃ : 26 ppb	CO: 127 ppb
H ₂ S: 0 ppb	PM _{2.5} : 9 µg/m ³
Wind: 17 deg	...

Observation 1

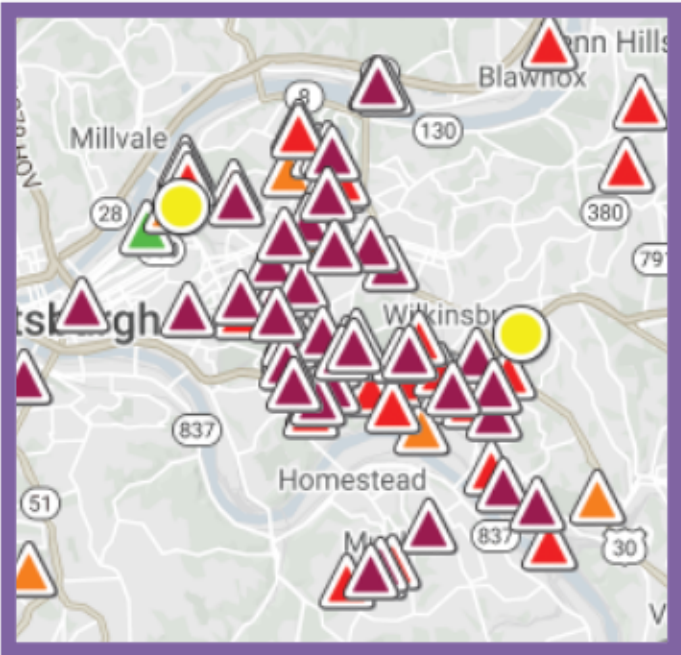
O ₃ : 1 ppb	CO: 1038 ppb
H ₂ S: 9 ppb	PM _{2.5} : 23 µg/m ³
Wind: 213 deg	...

Observation 2

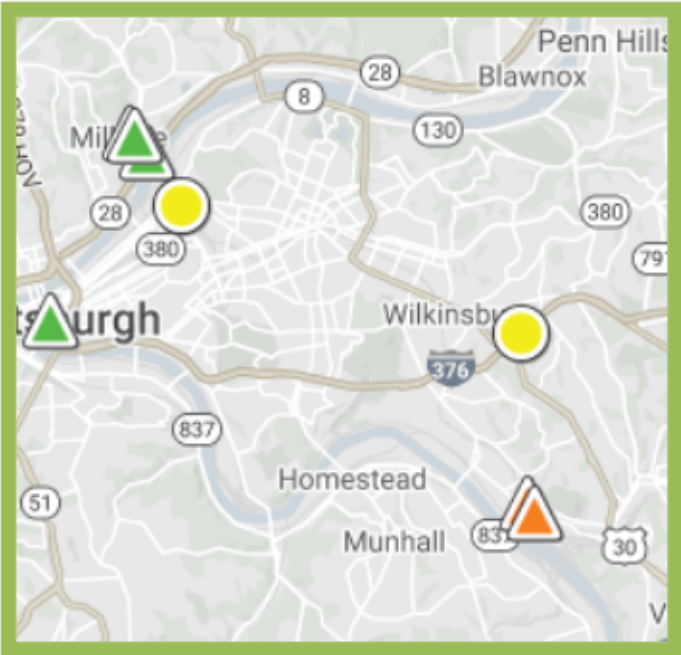


if H₂S > ?
and CO > ?
and PM_{2.5} > ?
and ...
then has event

else no event



☹️ Has Event



😊 No Event

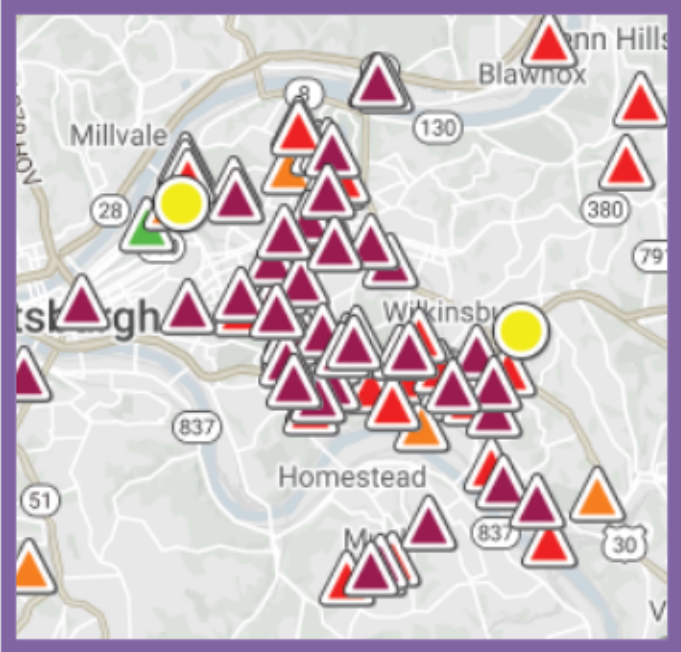
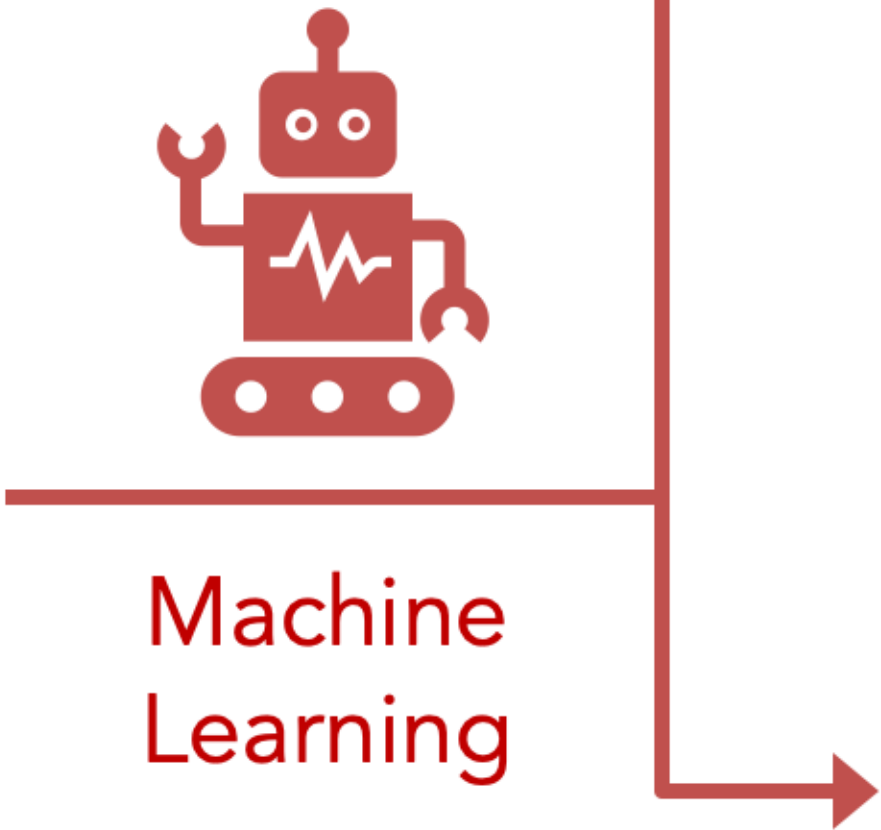
It turns out that we can use the Smell Pittsburgh dataset to estimate a function (i.e., train a machine learning model) that can predict smell events from sensor measurements.

O ₃ : 26 ppb	CO: 127 ppb
H ₂ S: 0 ppb	PM _{2.5} : 9 µg/m ³
Wind: 17 deg	...

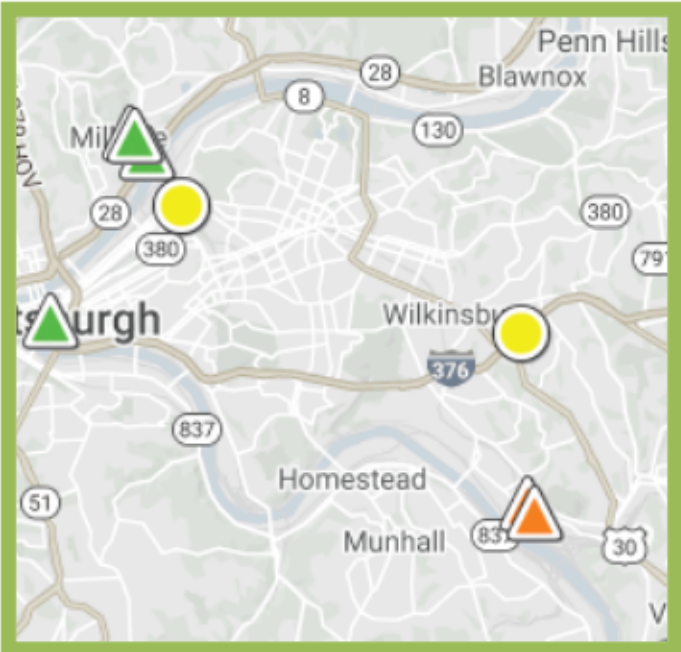
Observation 1

O ₃ : 1 ppb	CO: 1038 ppb
H ₂ S: 9 ppb	PM _{2.5} : 23 µg/m ³
Wind: 213 deg	...

Observation 2



☹️ Has Event



😊 No Event

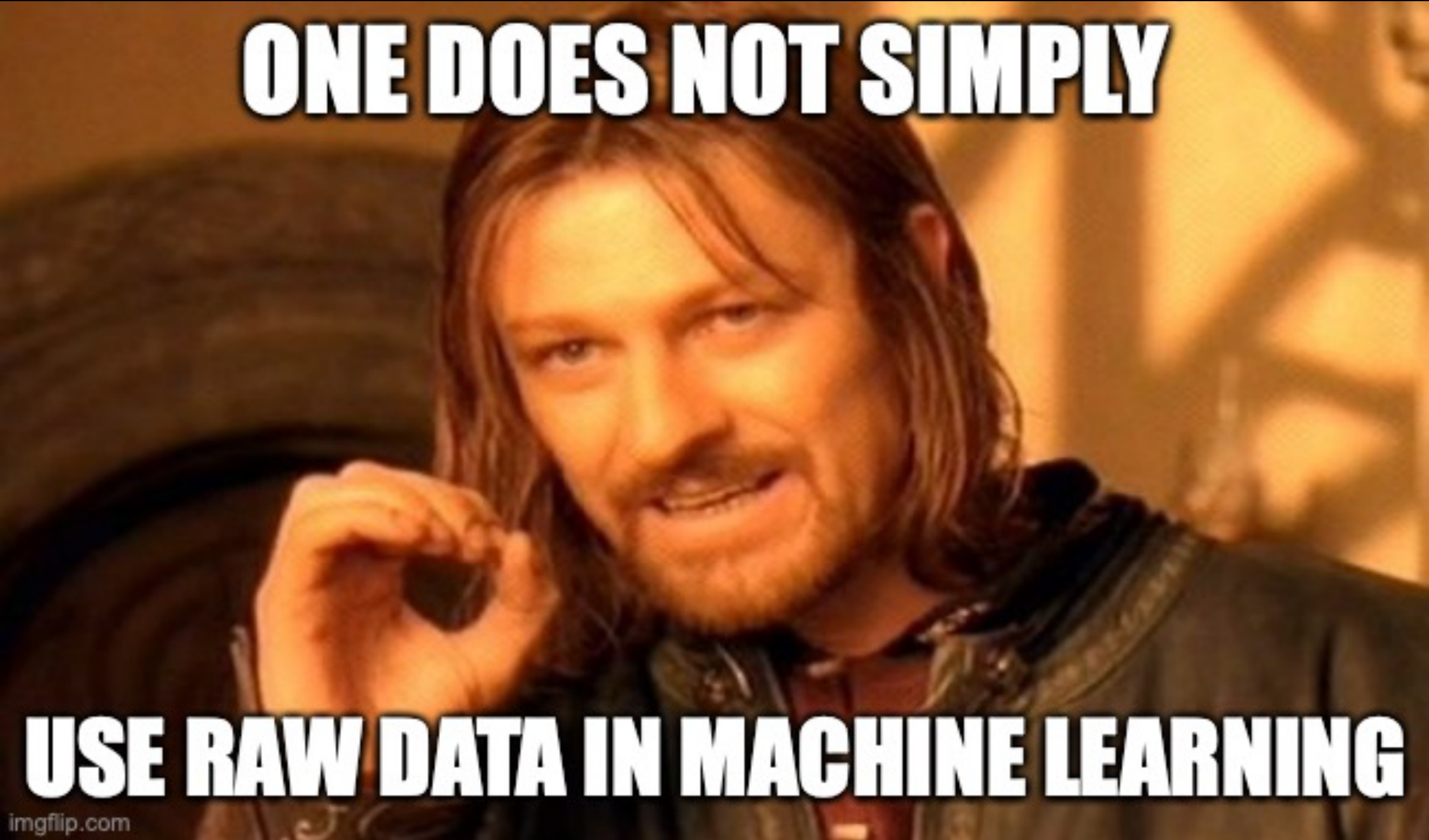
Researchers collected the **Smell Pittsburgh dataset**, including all the smell reports and sensor measurements (from air quality and weather monitoring stations) from October 31 in 2016 to September 30 in 2018.

■ Samples of Citizen-Contributed Smell Reports

EpochTime	feelings_symptoms	smell_description	smell_value	zipcode
...
1478353854	Headache, sinus, seeping into house even though it is as shut and sealed as possible. Air purifiers are unable to handle it thoroughly.	Industrial, acrid, strong	4	15206
1478354971		Industrial	4	15218
...

■ Samples of Air Quality Sensor Measurements

EpochTime	3.feed_28.H2S_PPM	3.feed_28.SO2_PPM	3.feed_28.SIGTHETA_DEG	3.feed_28.SONICWD_DEG	3.feed_28.SONICWS_MPH
...
1478046600	0,019	0,020	14,0	215,0	3,2
1478050200	0,130	0,033	13,4	199,0	3,4
...

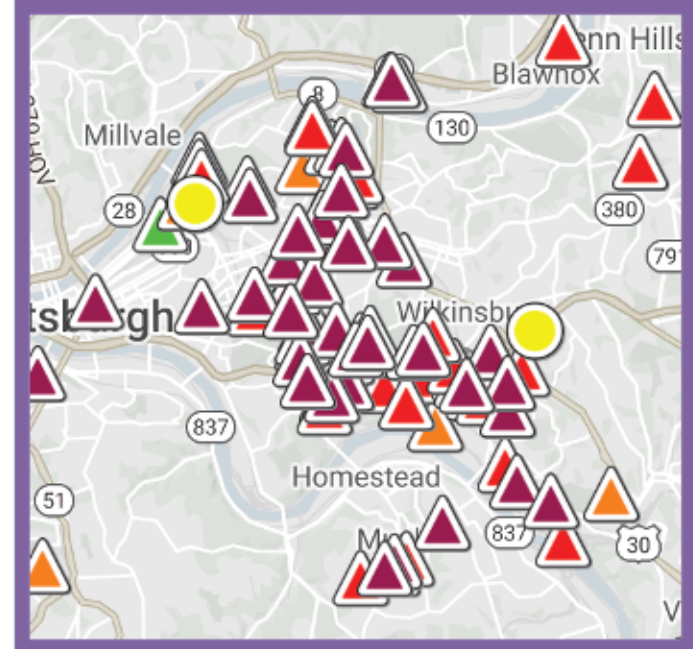
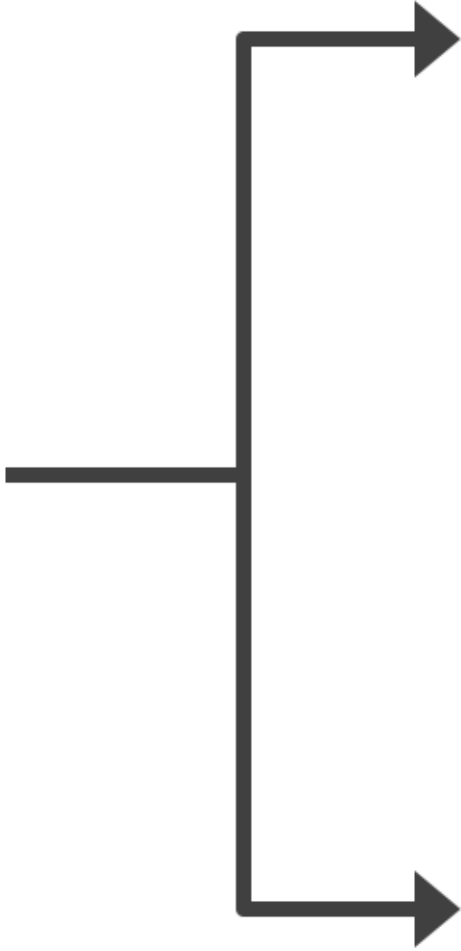


We need to **quantitatively define a smell event** (i.e., the presence of bad odor): whether the sum of smell values within a specific time range is larger than a particular threshold.

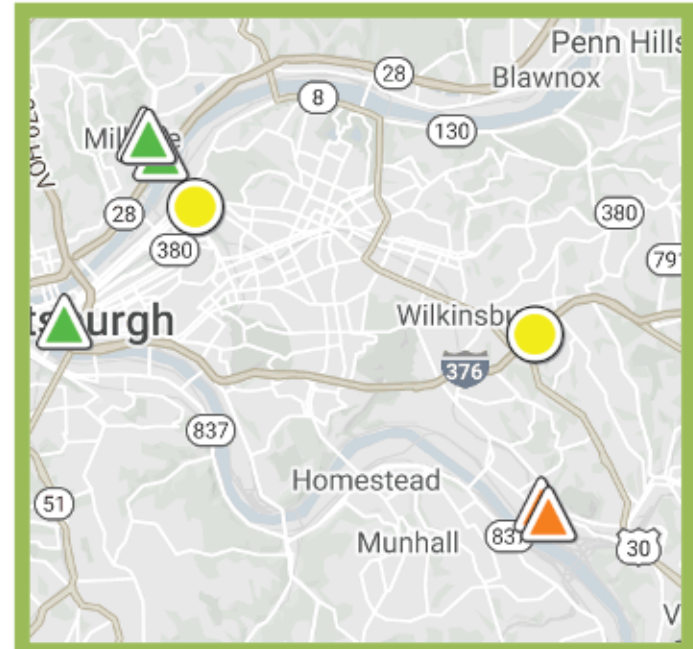
■ Samples of Citizen-Contributed Smell Reports

EpochTime	smell_value	zipcode
...
1478353854	4	15206
1478354971	4	15218
1478359473	4	15218
1478371179	3	15207
1478393585	3	15217
1478399011	4	15217
1478432399	4	15218
1478432502	2	15206
1478434105	4	15217
1478435133	4	15206
1478435313	4	15206
1478435748	3	15206
1478435801	5	15218
...

if the sum of smell values within H hours > V
 (need to define H and V)
 then has event
 else no event

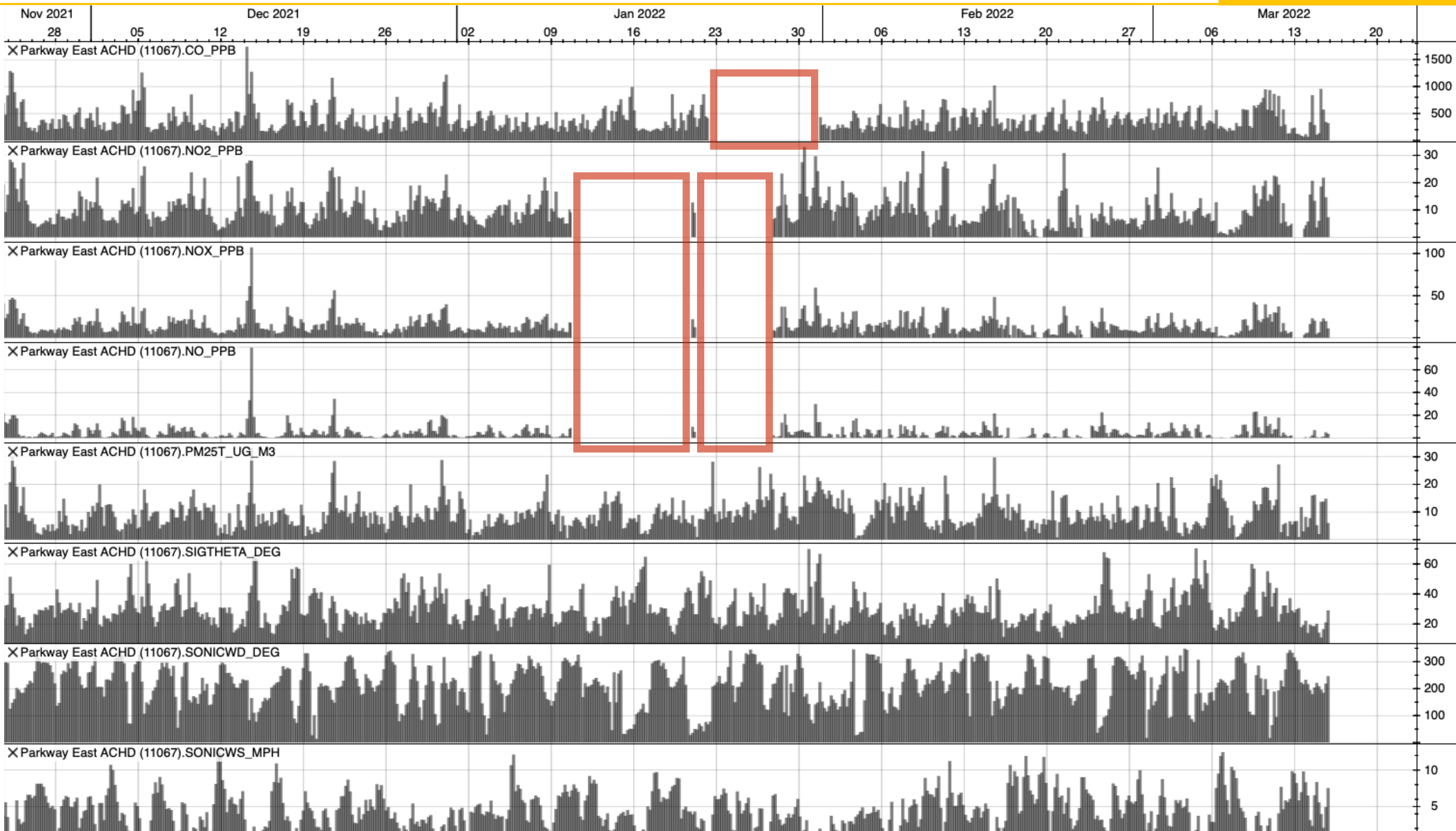


☹️ Has Event

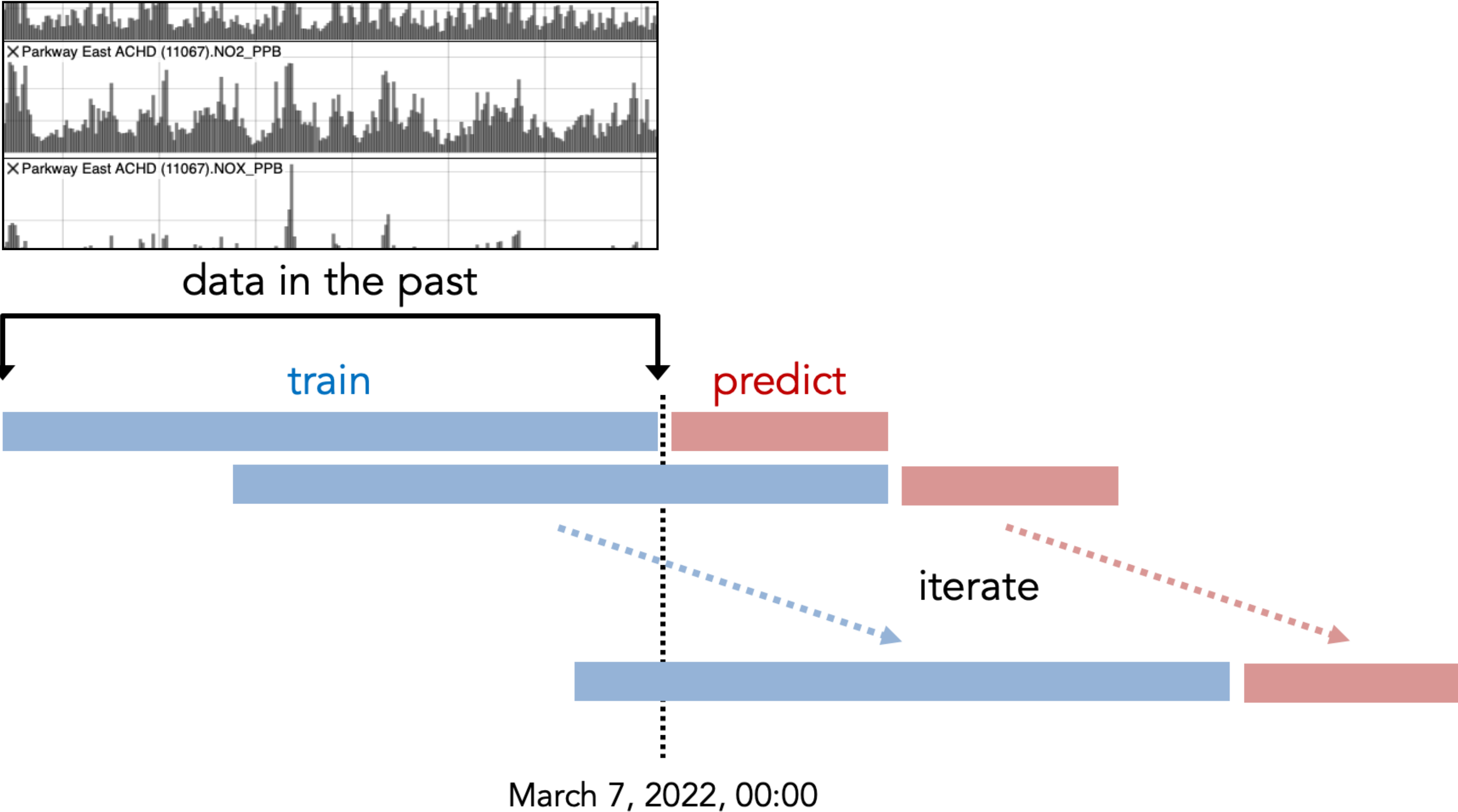


😊 No Event

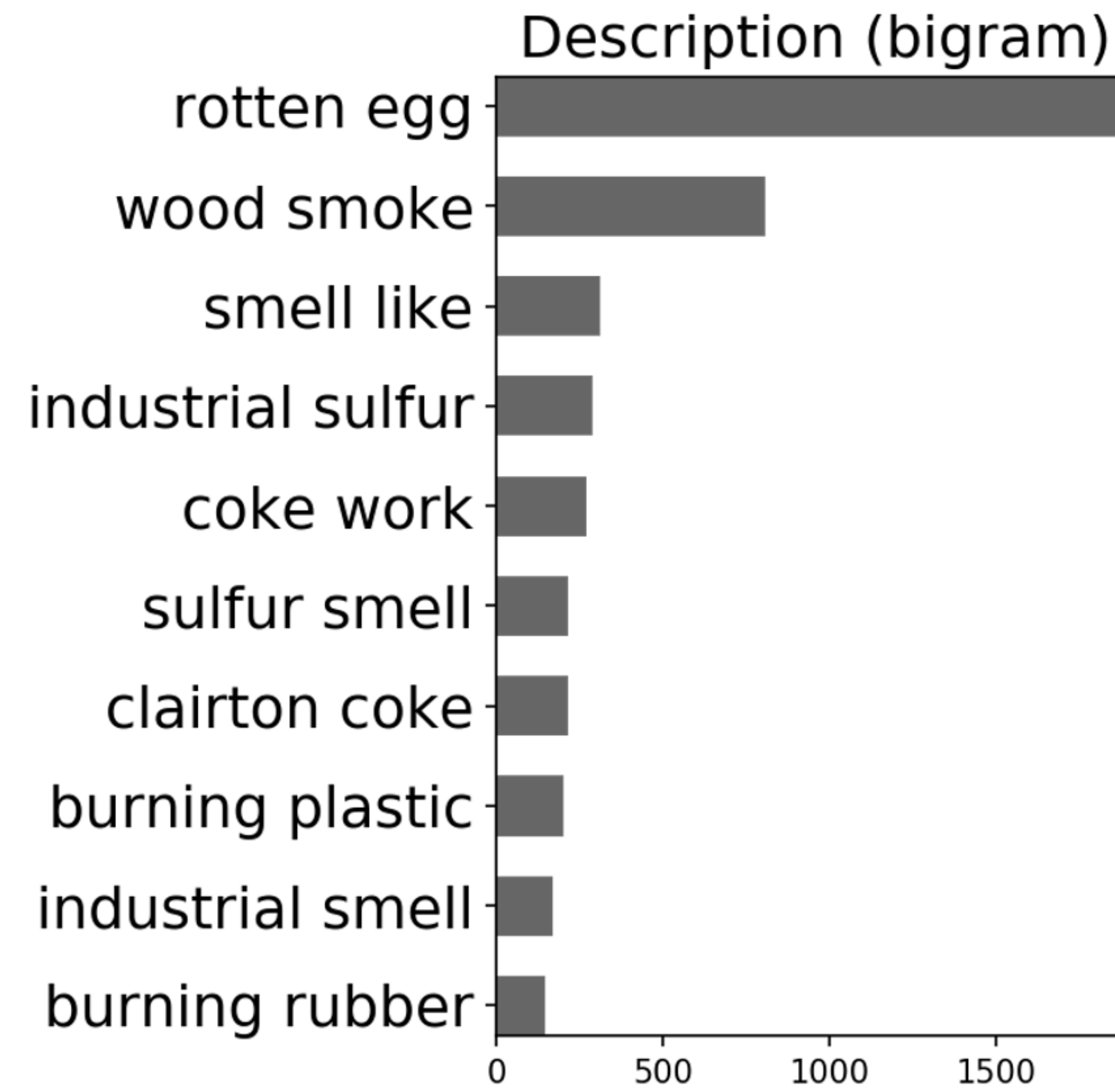
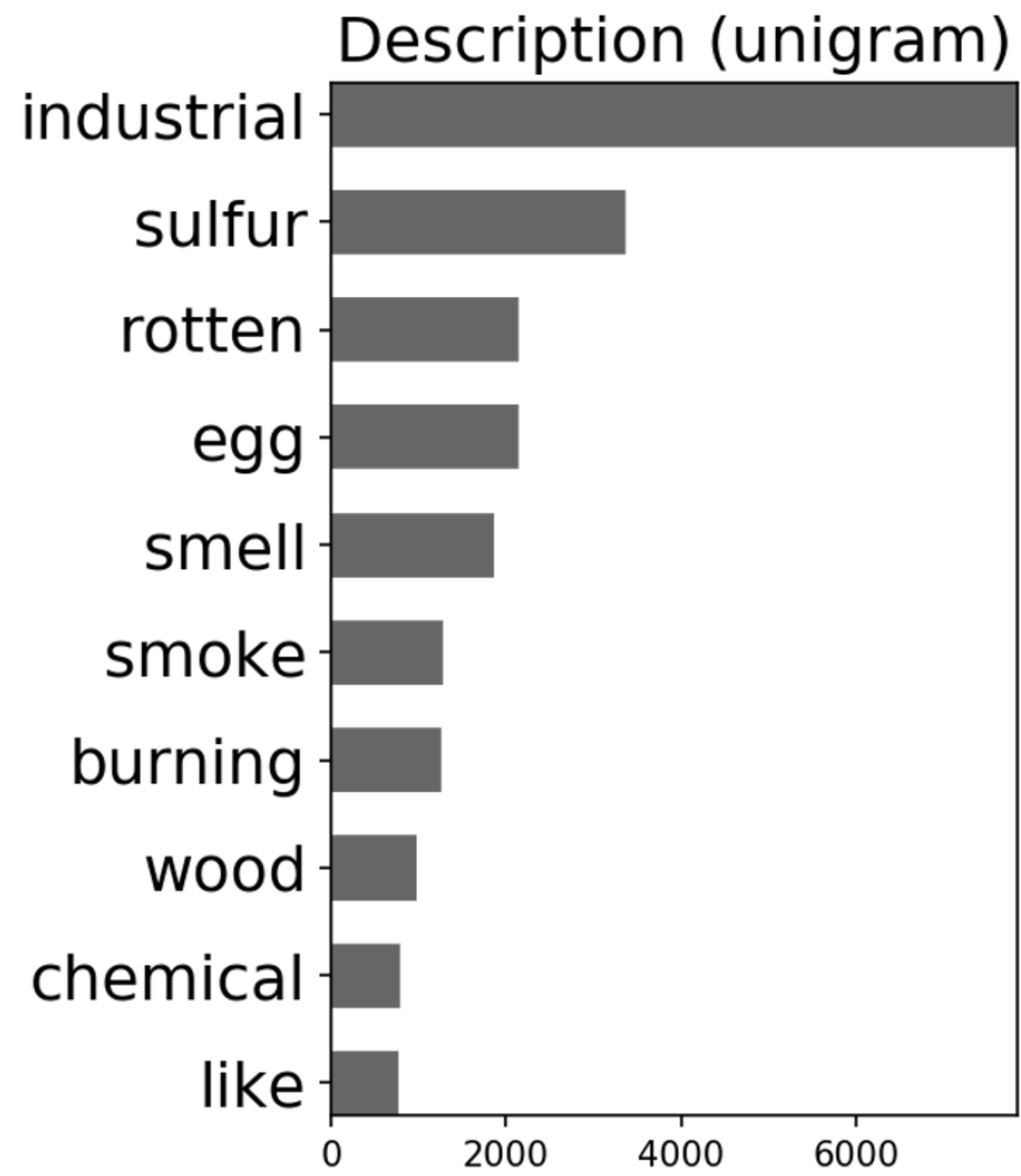
We need to **treat missing data**. The sensor measurements can be missing during some time periods since some air quality or weather monitoring stations may be down for maintenance.



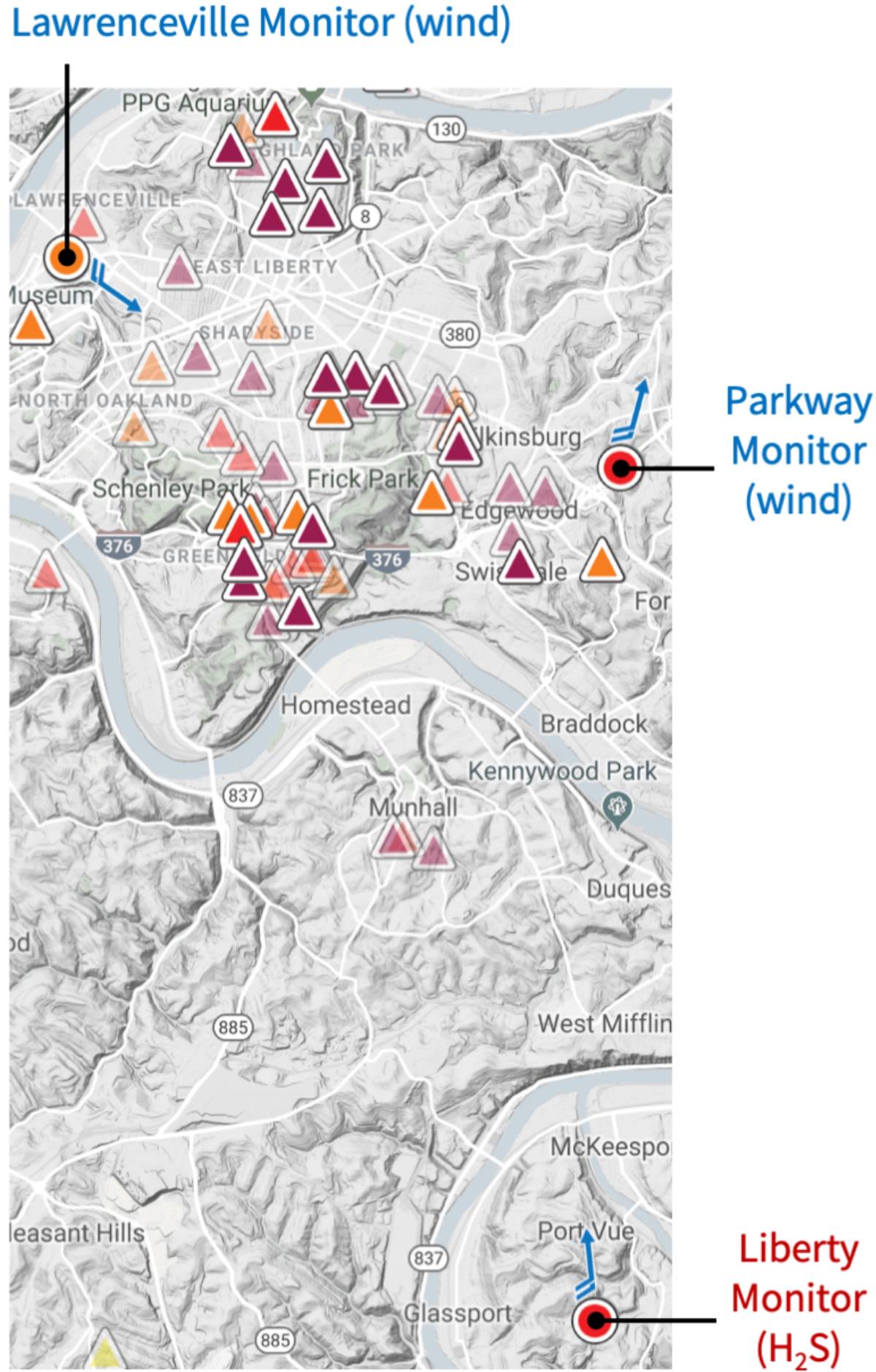
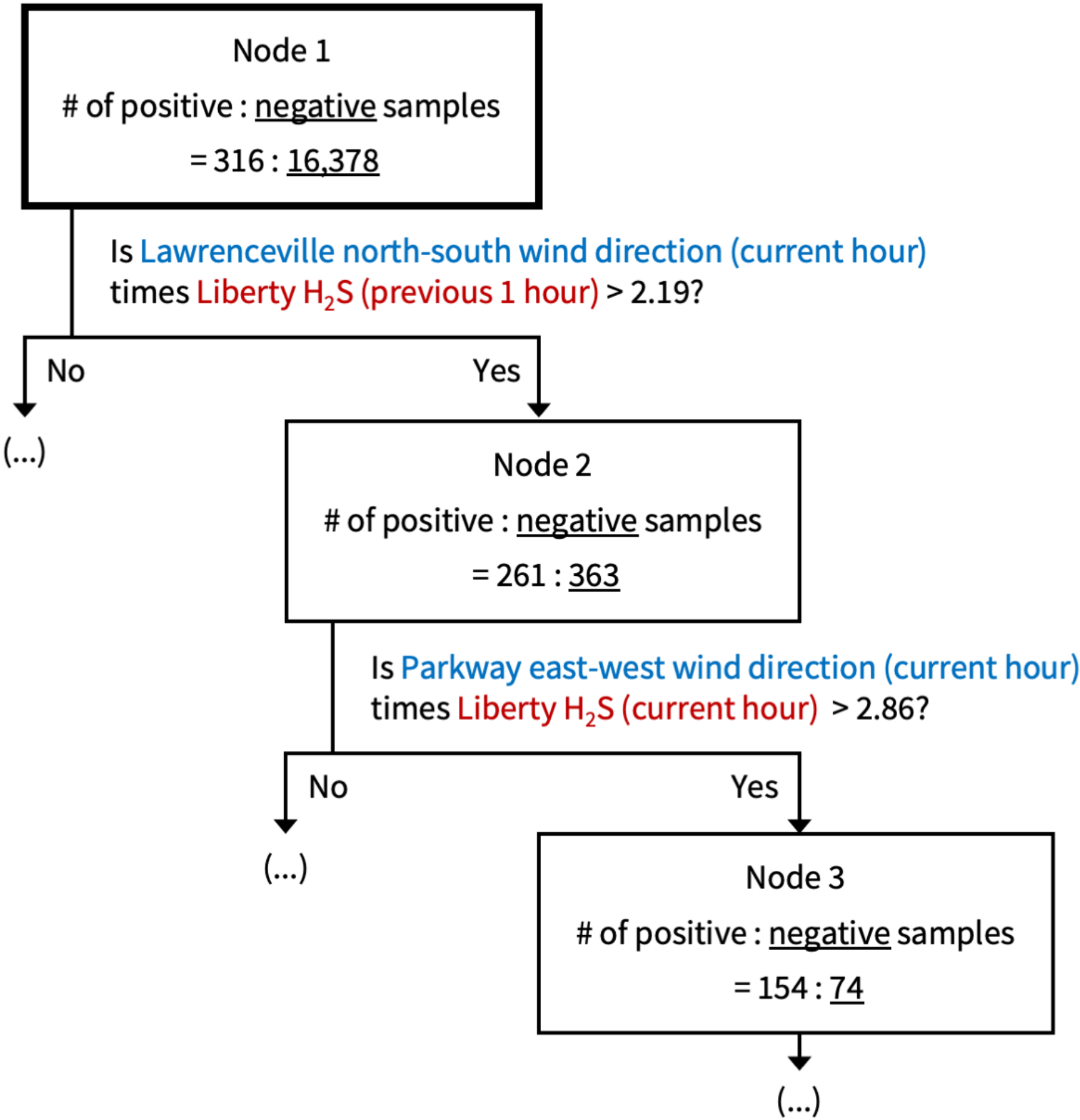
The dataset contains **time-series data**, which means each data point has a timestamp, and we can only use data in the past (i.e., data that exists for a specific time point) to train the model to predict the future.



How do we know **which variables from which monitoring stations** are effective in predicting the presence of bad odor? We can explore the data to get insights or rely on local knowledge of pollution sources.



We also need to **extract and decide the features** that we want to use when training the machine learning model. Such features can help us identify air pollution patterns in the Pittsburgh region.



Machine Learning for Design

Lecture 7

Design and Develop Machine Learning
Models - *Part 1*