Data Science

Lecture 1-2: Social Impact of Data Science

University of Amsterdam

Lecturer: Yen-Chia Hsu

Date: Sep 2025

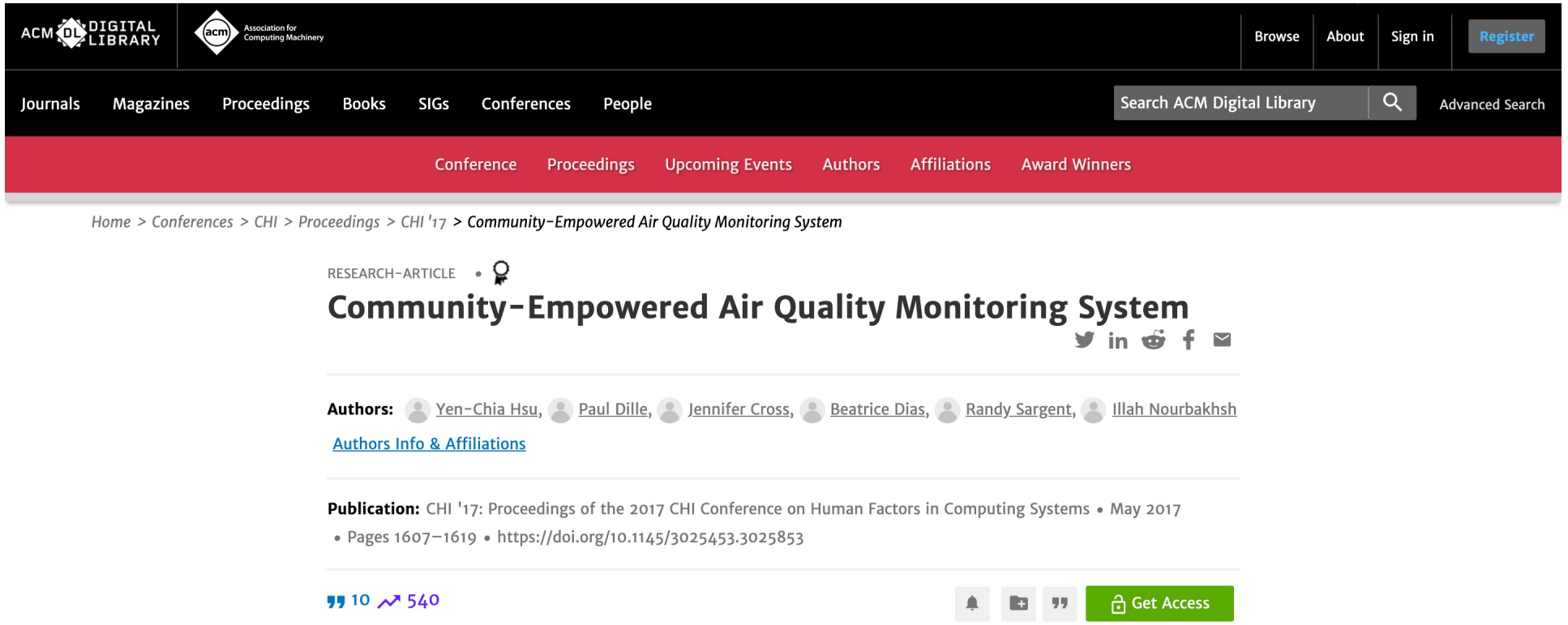
Data science can create a social impact to influence society positively!

This lecture uses my previous projects to show the social impact of data science.

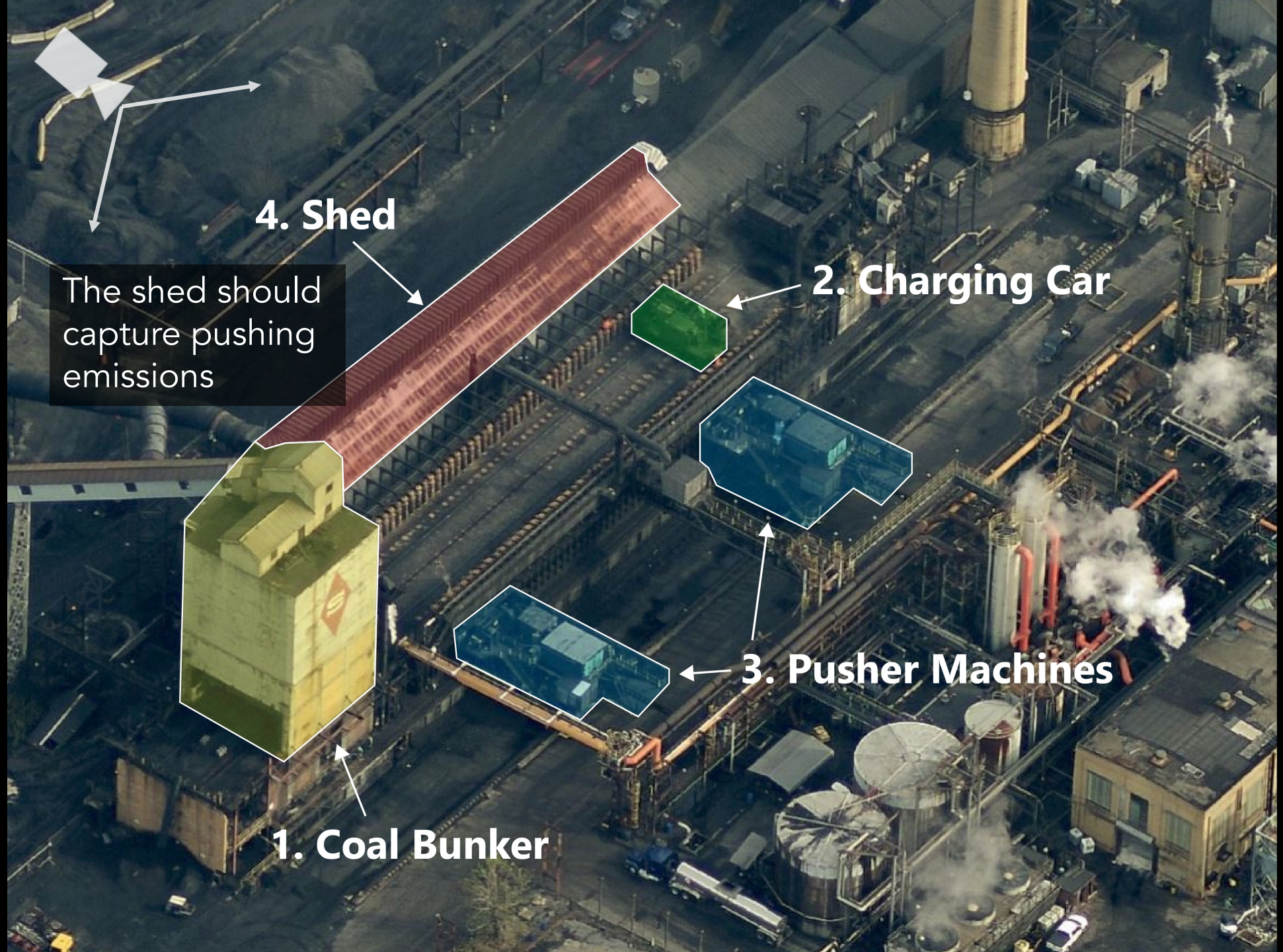
Air Quality Monitoring System

Integrating smoke videos, sensor readings, and smell reports

http://shenangochannel.org



Community
Coke Refinery
CMU
Ben Avon Heights
Ben Avon
Avalon
Bellevue
Stowe Township
Ohio River
Riverview Park
Reserve Township
Millvale
Etna
Allegheny River
LAWRENCEVILLE
GARFIELD
BLOOMFIELD
McKees Rocks
Brunot Island
Ingram
Thornburg
Crafton
EAST ALLEGHENY
STRIP DISTRICT
ALLEGHENY WEST
NORTH SHORE
PNC Park
CULTURAL DISTRICT
Pittsburgh
UPTOWN
NORTH OAKLAND
University of Pittsburgh
Carnegie Mellon University
Google


4. Shed
The shed should capture pushing emissions
2. Charging Car
3. Pusher Machines
1. Coal Bunker

- Co-design to collect evidences of air pollution
- Present them to the government or media
- Raise the public awareness of air quality issues

Residents can integrate the evidence of air pollution and their personal experiences into a story to influence regulators' attitudes and increase the community's confidence.

Hsu, Y. C., et al. (2017). Community-Empowered Air Quality Monitoring System. ACM CHI.

"But what I see in the video," the acting director of U.S. EPA Region III Air Protection Division said, referring to videos from the system that were projected on a screen at the front of the meeting room, "is totally unacceptable."

Community meeting on November 19, 2015 (https://youtu.be/-OxfkU-ql5M?t=4274)

2015-05-27 12:44:10

2015-10-11 17:40:42

2015-06-15 17:44:21

2015-06-11 17:54:00

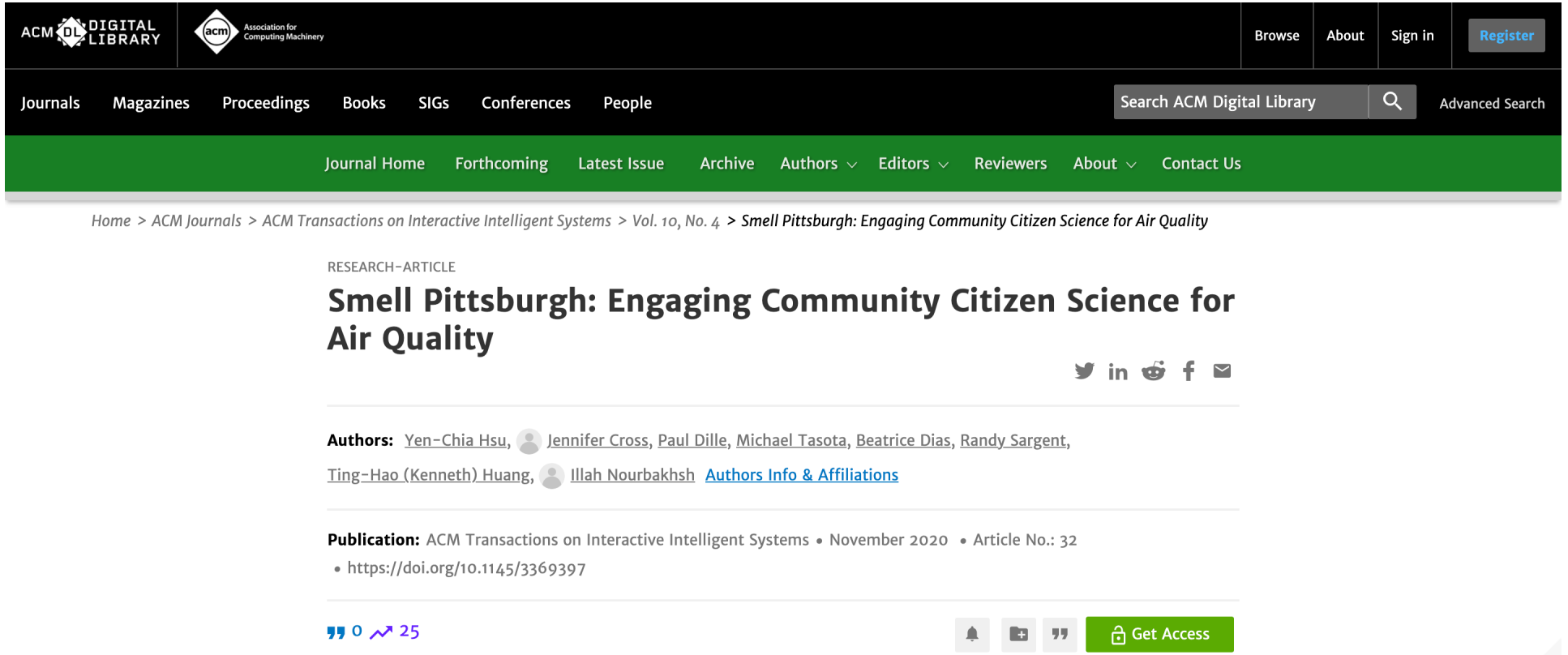
2015-05-02 10:43:46
2015-08-24 11:00:20
2015-05-10 11:30:19
2015-05-27 15:04:45
2015-05-01 15:41:13
2015-10-15 11:41:27
2015-10-23 11:21:06
2015-05-01 15:19:56
2015-05-27 13:32:20
2015-05-27 13:56:44
2015-05-04 18:54:07
2015-05-08 16:46:30
2015-05-13 08:45:00
2015-05-22 10:12:33
2015-06-15 17:44:06
2015-06-11 17:58:47
2015-05-02 18:24:48
2015-10-11 16:44:47
2015-10-19 17:25:44
2015-07-29 17:52:25
2015-07-21 17:33:22
2015-09-15 16:46:16
2015-10-20 17:15:33
2015-05-15 07:36:33
2015-05-02 08:26:32

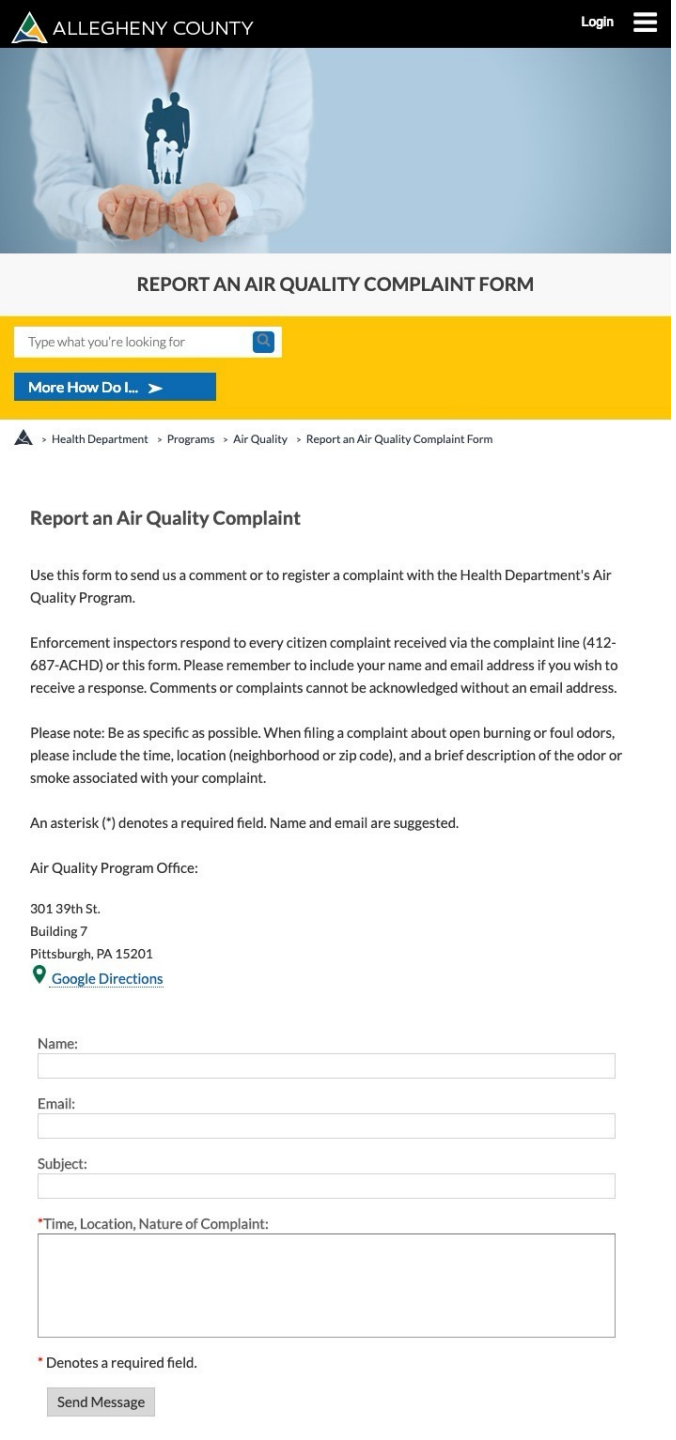
2015-07-02 13:13:06
2015-07-02 17:50:51
2015-07-02 19:45:26
2015-07-02 19:33:22
2015-07-03 10:47:37
2015-07-04 15:44:28
2015-07-04 18:12:09
2015-07-05 15:11:53
2015-07-05 13:32:32
2015-07-05 17:24:44
2015-07-05 19:44:25
2015-07-06 11:18:35
2015-07-06 14:38:39
2015-07-06 15:40:41
2015-07-08 14:04:50
2015-07-11 09:12:35
2015-07-11 13:07:50
2015-07-11 17:25:36
2015-07-12 15:26:00
2015-07-12 15:43:06
2015-07-14 12:22:11
2015-07-14 16:34:50
2015-07-15 19:01:43
2015-07-15 20:16:52
2015-07-16 14:20:45
2015-07-16 20:36:05
2015-07-17 09:00:43
2015-07-16 16:16:21
2015-07-16 16:38:36
2015-07-17 16:37:50
2015-07-18 15:17:53
2015-07-18 14:20:28
2015-07-18 18:42:41
2015-07-18 19:50:42
2015-07-19 17:32:45
2015-07-20 15:56:11
2015-07-20 18:04:46
2015-07-20 20:41:53
2015-07-22 17:23:09
2015-07-23 15:00:07
2015-07-23 15:14:30
2015-07-24 19:14:01
2015-07-25 16:01:44
2015-07-25 17:45:17
2015-07-25 19:17:43
2015-07-27 16:54:47
2015-07-28 15:07:40
2015-07-30 10:51:56
2015-07-30 20:18:02
2015-07-29 08:44:56
2015-07-31 17:53:43
2015-08-01 18:43:44
2015-08-02 11:56:06
2015-08-02 15:48:57
2015-08-02 17:43:54
2015-08-02 19:15:16
2015-08-03 09:49:17
2015-08-03 17:45:49
2015-08-05 19:34:29
2015-08-06 14:29:38
2015-08-06 19:47:21
2015-08-07 12:48:51
2015-08-29 18:12:03
2015-08-25 17:17:11
2015-08-26 14:08:31
2015-08-25 17:04:26
2015-08-22 17:13:44
2015-08-23 16:29:21
2015-08-28 17:43:28
2015-08-22 14:07:37
2015-08-21 19:30:00
2015-08-17 18:59:11
2015-08-21 16:40:54
2015-08-30 16:52:32
2015-08-30 16:39:47
2015-08-28 19:52:27
2015-08-28 14:44:01
2015-08-29 17:51:37
2015-08-22 13:09:04
2015-08-17 16:54:19
2015-08-09 06:32:27
2015-08-10 16:58:49
2015-08-10 19:17:32
2015-08-17 15:50:24
2015-08-15 17:03:00
2015-08-16 13:58:09
2015-08-12 18:26:28
2015-08-14 12:44:51
2015-08-13 17:36:04
2015-08-14 16:16:24
2015-08-09 17:16:26
2015-08-17 18:18:06
2015-08-14 19:07:19
2015-08-27 13:15:13
2015-08-15 17:23:27
2015-06-29 16:24:32
2015-06-29 15:39:47
2015-06-29 14:20:12
2015-06-29 14:04:32
2015-06-29 13:39:53

Smell Pittsburgh

Crowdsourcing and visualizing pollution odors

https://smellpgh.org



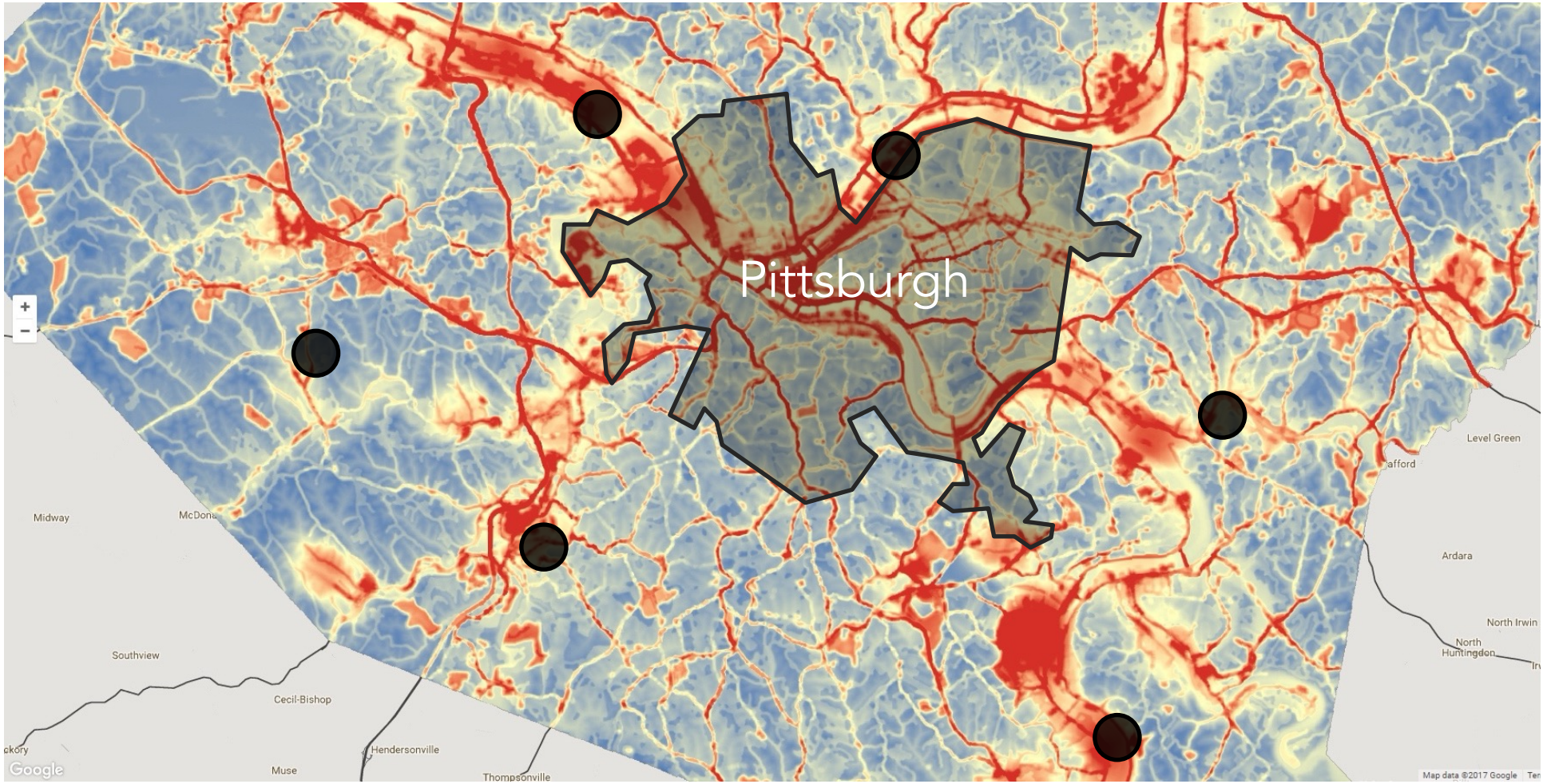


The prior approach that asks citizens to report odor complaints post hoc via forms or phone calls suffers from:

- poor data quality
- non-transparency

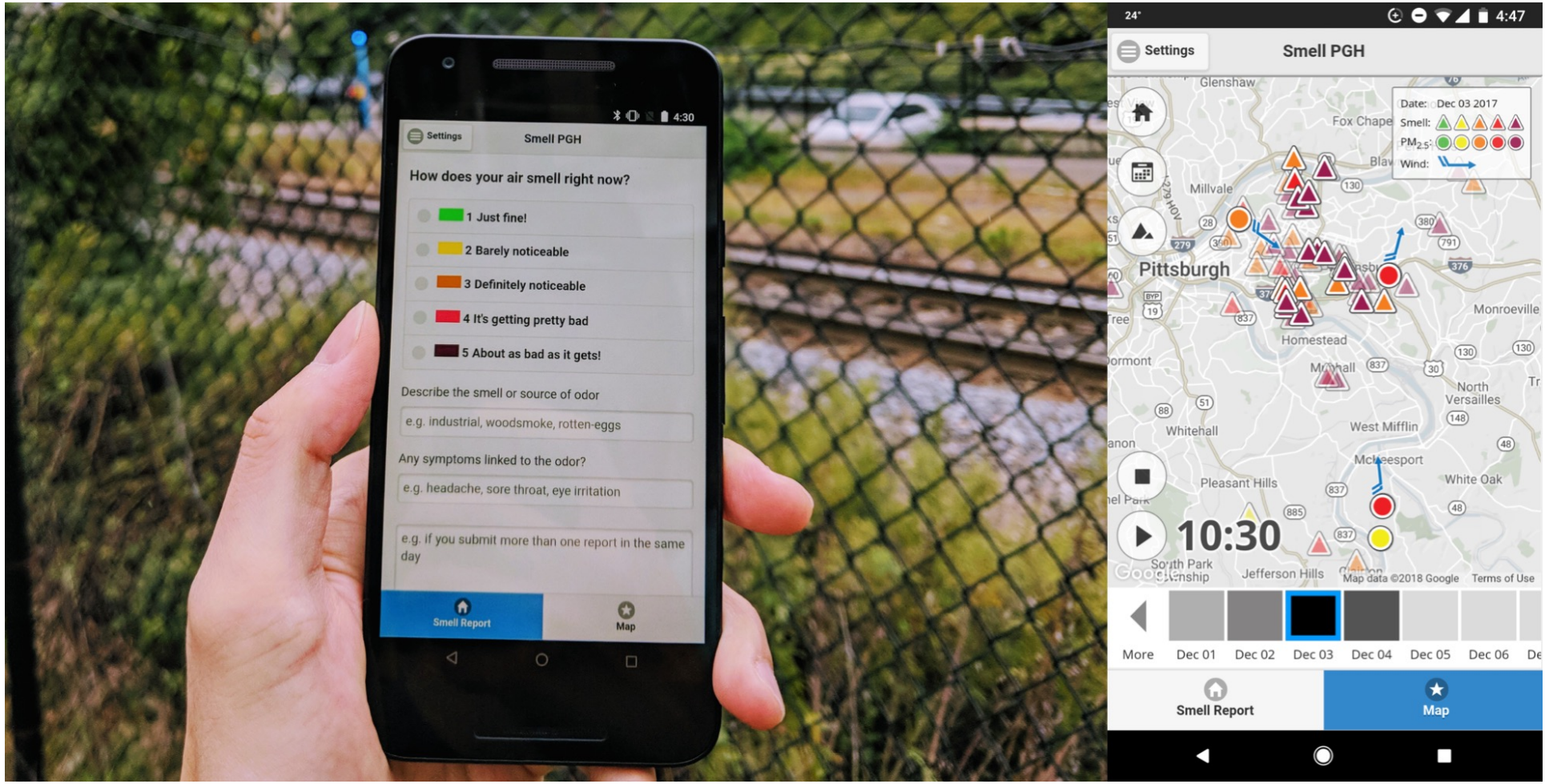
The odor reporting form -- https://www.alleghenycounty.us/Health-Department/Programs/Air-Quality/Report-an-Air-Quality-Complaint-Form.aspx

How can we effectively collect the smell experiences on a city-wide scale with more than 300,000 residents and many pollution sources over many years?



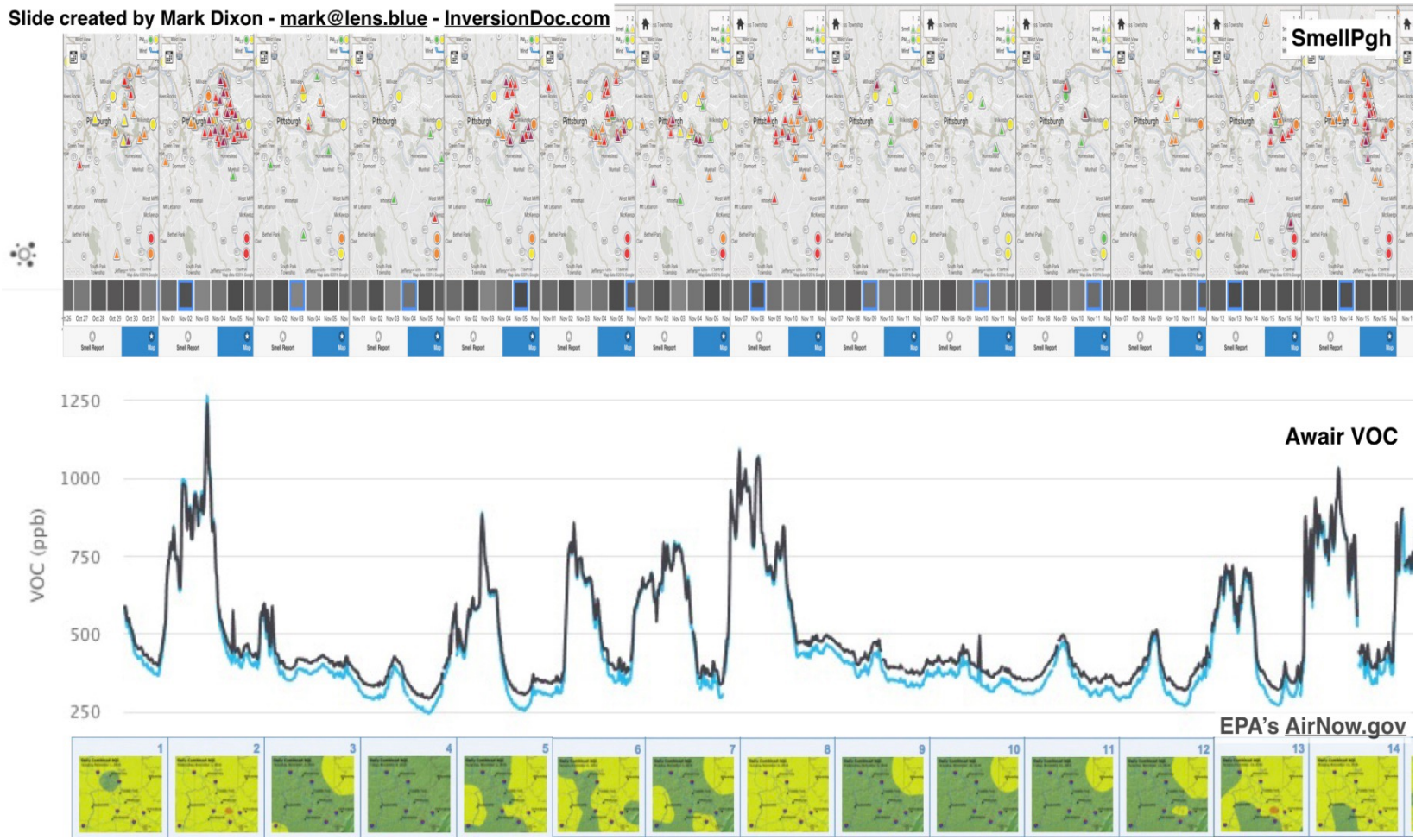
Pittsburgh pollution map -- https://breatheproject.org/pollution-map/

Smell Pittsburgh enables communities to collect data on a large scale. Also, visualizing multi-modal data in real-time can help communities understand local concerns.



Hsu, Y. C., et al. (2020). Smell Pittsburgh: Engaging Community Citizen Science for Air Quality. ACM TiiS.

Community members plot smell reports with self-operated VOC (volatile organic compounds) sensors to find correlations and inspect how pollution impacts them.



Decision-makers in the local health department mentioned that "Every aspect of the activity and operation of these coke plants will have a more stringent standard applied."

Pittsburgh Post-Gazette

Photo credit Don Hopey

Air advocates read 'scroll of smells' at health board meeting

Table 1: Distribution of Smell Reports

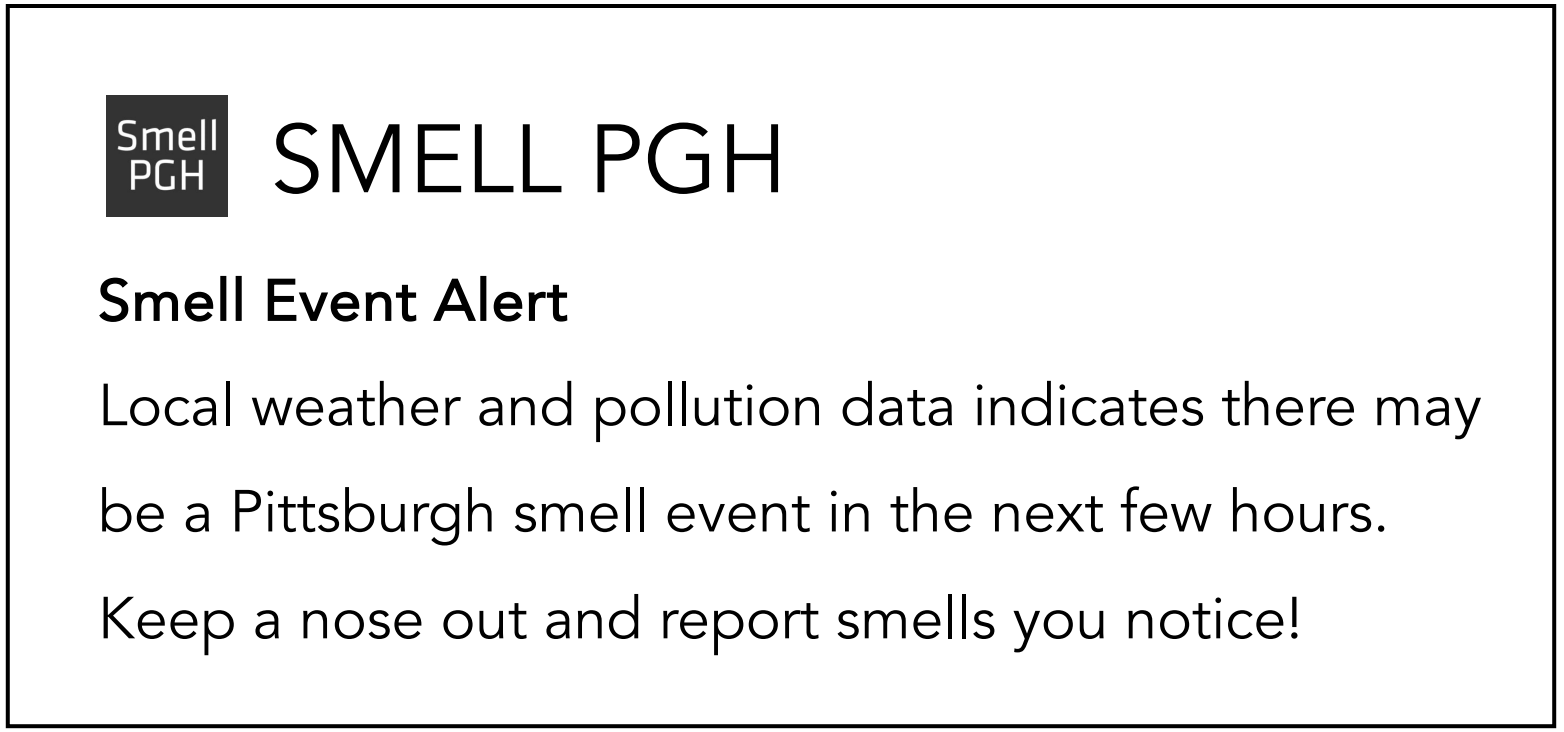
Smell Rating	Description	2022	2021	2020	2019	2018	2017
1	Just fine!	333 (3.6%)	707 (5.7%)	1,565 (8.2%)	1,711 (9.5%)	1,199 (13.0%)	1,658 (20.4%)
2	Barely noticeable	287 (3.1%)	447 (3.6%)	921 (4.8%)	798 (4.4%)	497 (5.4%)	665 (8.2%)
3	Definitely noticeable	1,922 (20.8%)	2,902 (23.3%)	4,436 (23.3%)	4,305 (23.9%)	2,649 (28.8%)	2,246 (27.7%)
4	It's getting pretty bad	3,185 (34.5%)	4,258 (34.2%)	6,014 (31.6%)	5,805 (32.3%)	2,932 (31.9%)	2,171 (26.8%)
5	About as bad as it gets!	3,506 (38.0%)	4,126 (33.2%)	6,082 (32.0%)	5,358 (29.8%)	1,918 (20.9%)	1,372 (16.9%)
Sum		9,233	12,440	19,019	17,977	9,195	8,112

Table 2: User Engagement with The Smell PGH App

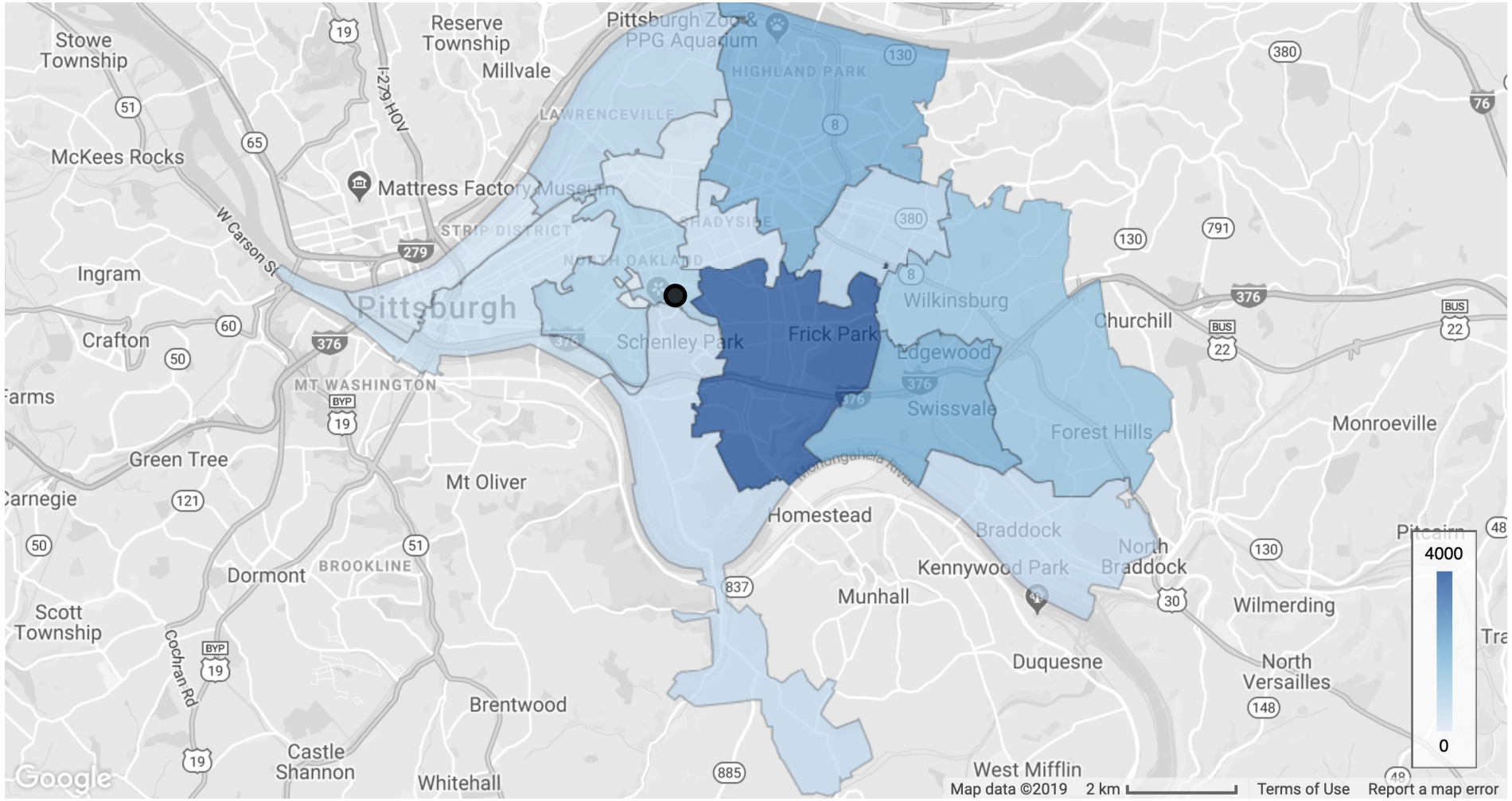
Number of Unique Users	2022	2021	2020	2019	2018	2017
Submitted Reports	1,242 (37.0%)	1,804 (41.5%)	2,688 (46.8%)	3,274 (50.9%)	1,769 (66.9%)	1,308 (58.4%)
Used the Map	3,054 (90.9%)	3,919 (90.2%)	5,227 (91.0%)	5,708 (88.8%)	2,248 (85.0%)	1,949 (87.0%)
Participated (N)	3,358	4,347	5,743	6,429	2,645	2,239

Is smell data useful in predicting local air pollution events and identifying patterns?

Smell Pittsburgh predicts upcoming smell events (based on the collected reports) and send push notifications to inform users and encourage engagement.



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



Number of smell reports aggregated by zip codes -- https://smellpgh.org/analysis#figure9

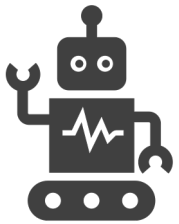
We use a Random Forest (a machine learning model) to predict smell events from air quality data (obtained from government-operated sensor stations).

O_3: 26 ppb	CO: 127 ppb
H_2S: 0 ppb	$PM_{2.5}$: 9 µg/m^3
Wind: 17 deg	…

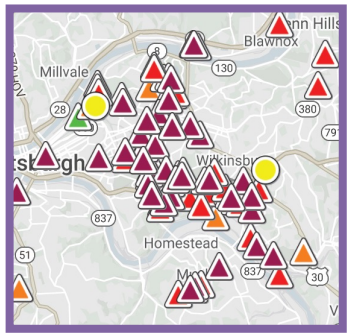
Observation 1

O_3: 1 ppb	CO: 1038 ppb
H_2S: 9 ppb	$PM_{2.5}$: 23 µg/m^3
Wind: 213 deg	…

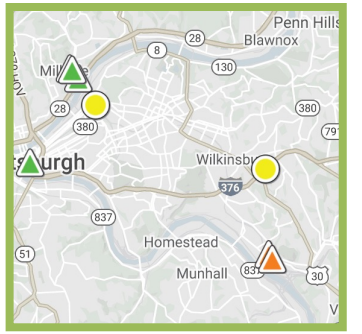
Observation 2



Machine Learning

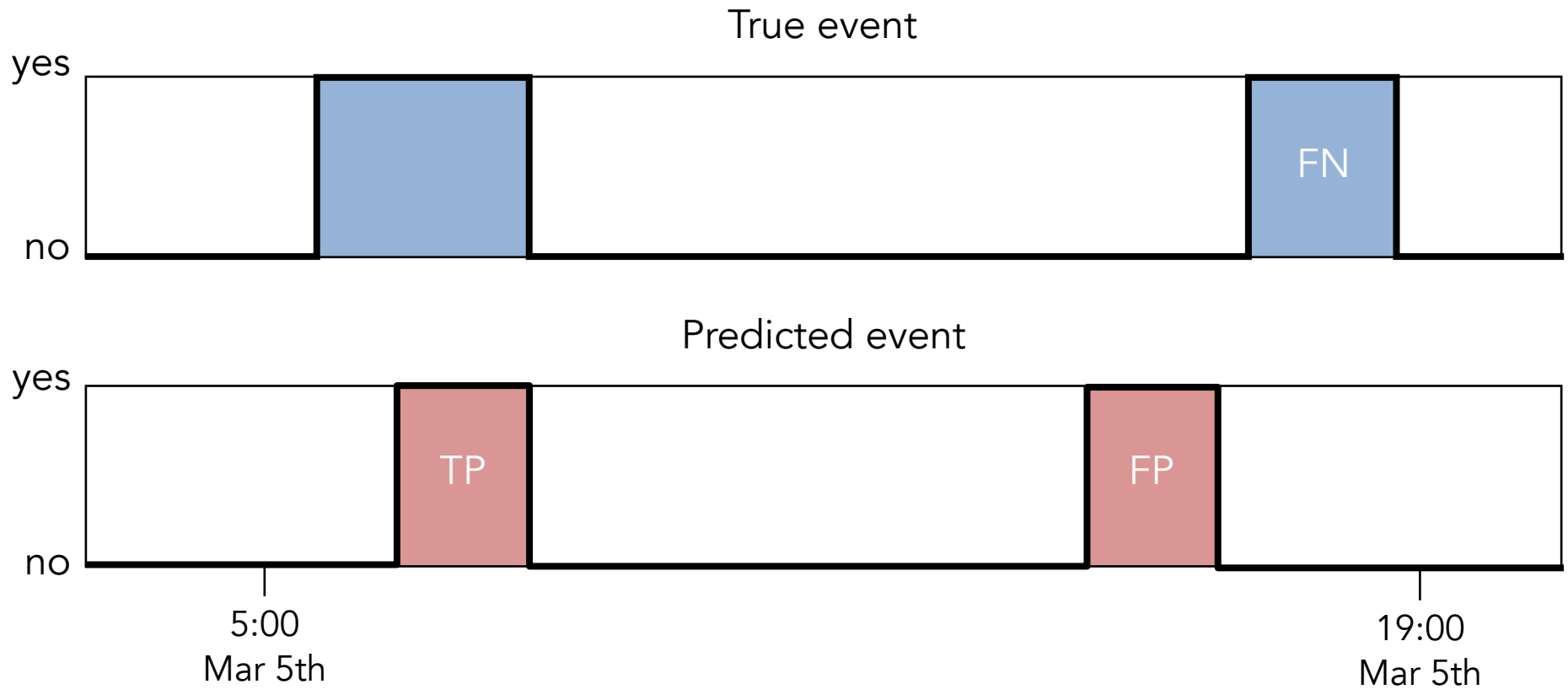


Has Event

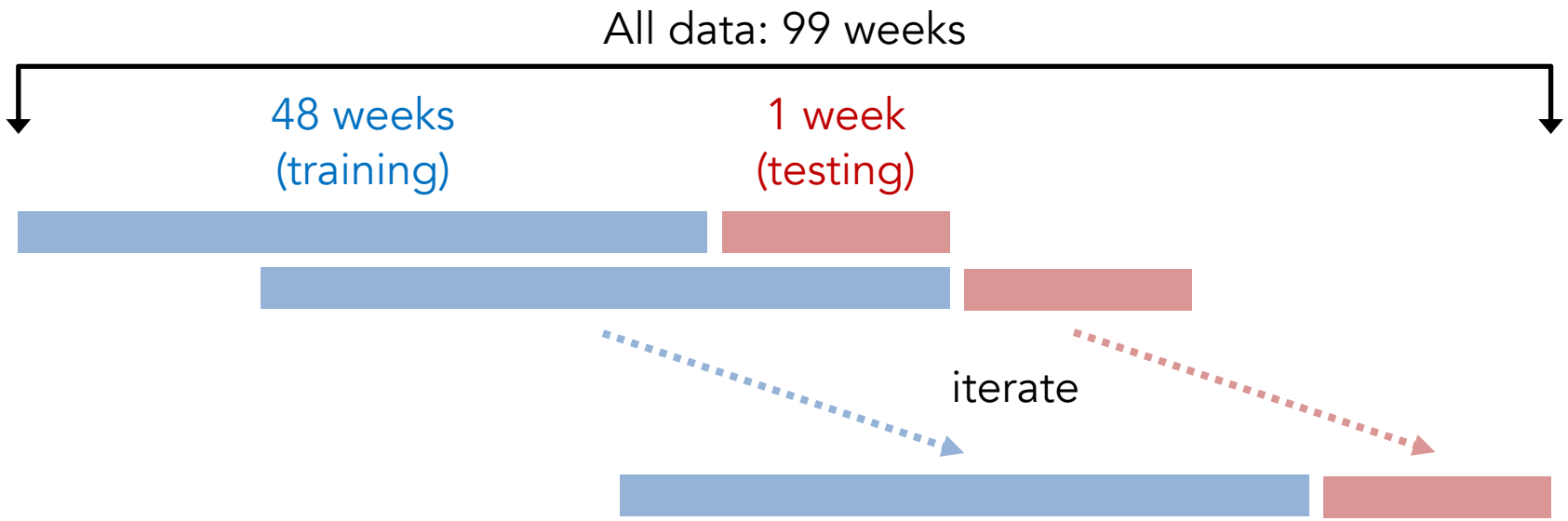


No Event

To evaluate models, we first compute true positives (TP), false positives (FP), and false negatives (FN) for smell events.

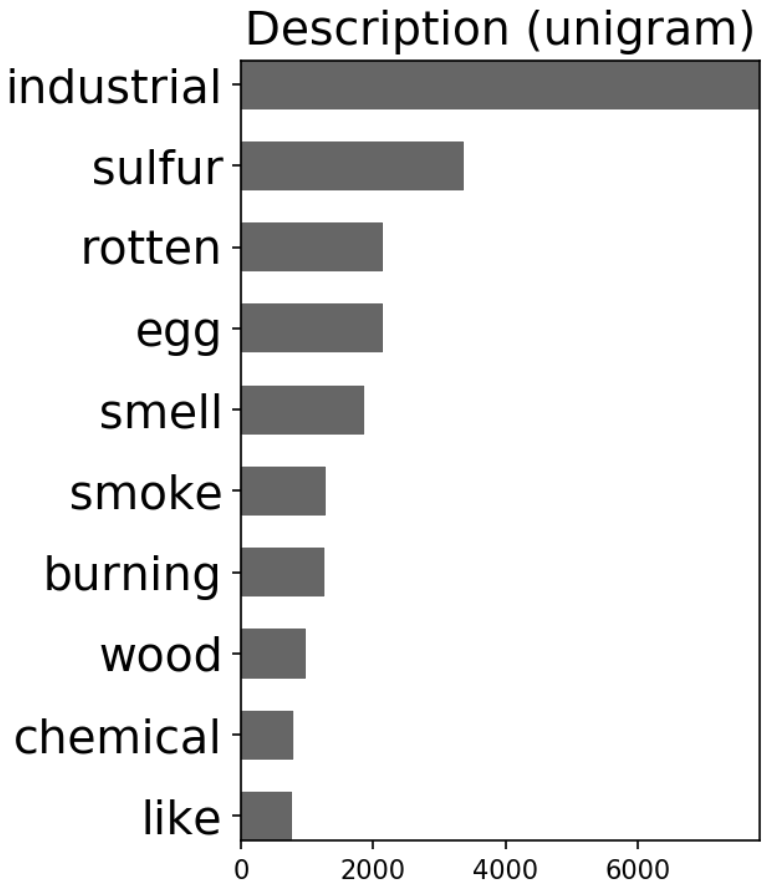


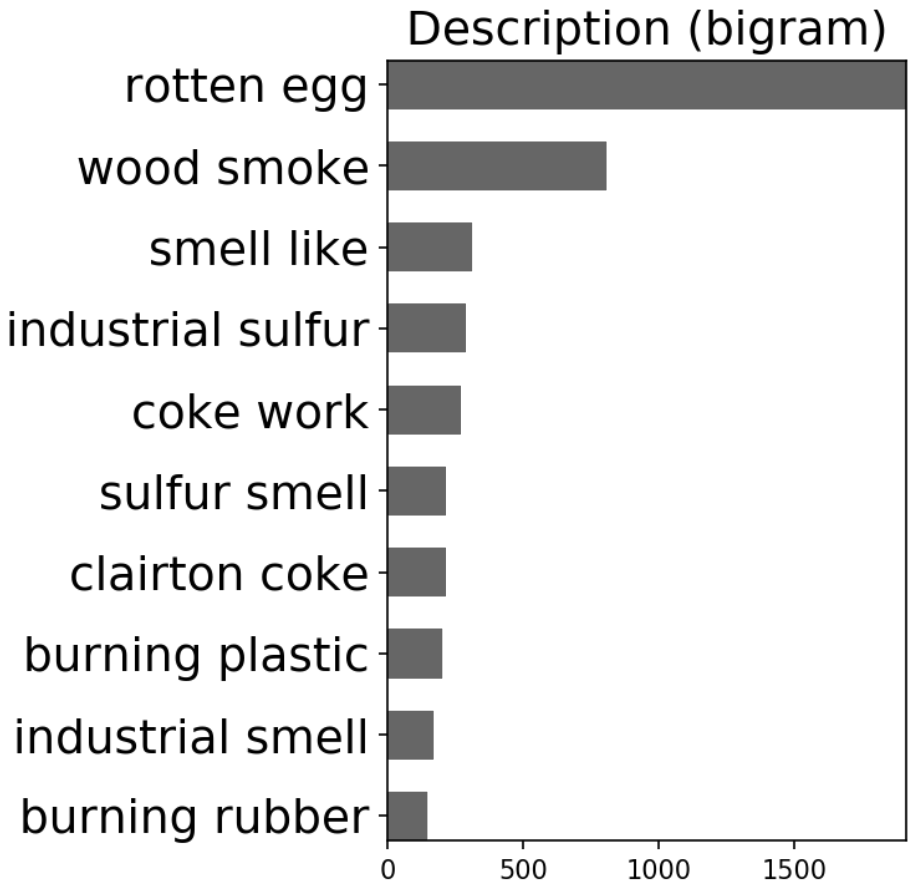
We apply time series cross-validation of several pairs of training and testing sets to evaluate the model performance.



	Precision	Recall	F-score
Our best model	0.87±0.01	0.59±0.01	0.70±0.01
Always yes	0.2	1	0.33

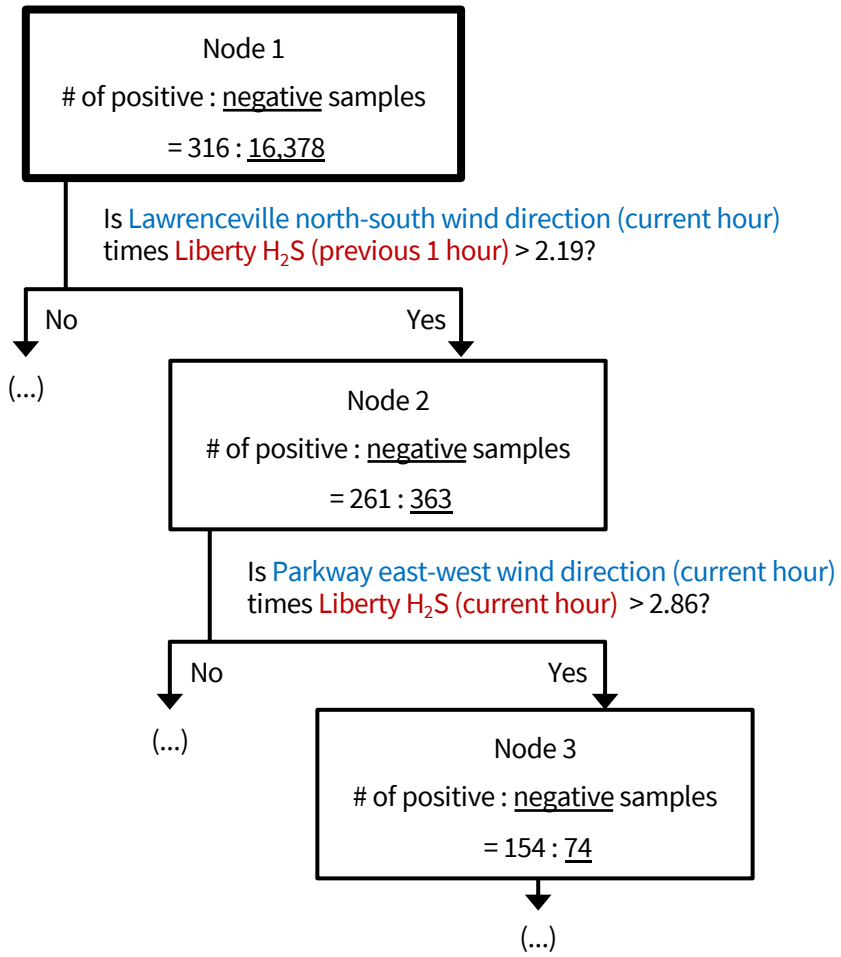
High-frequency words and phrases in smell reports mostly describe industrial pollution odors, like hydrogen sulfide.

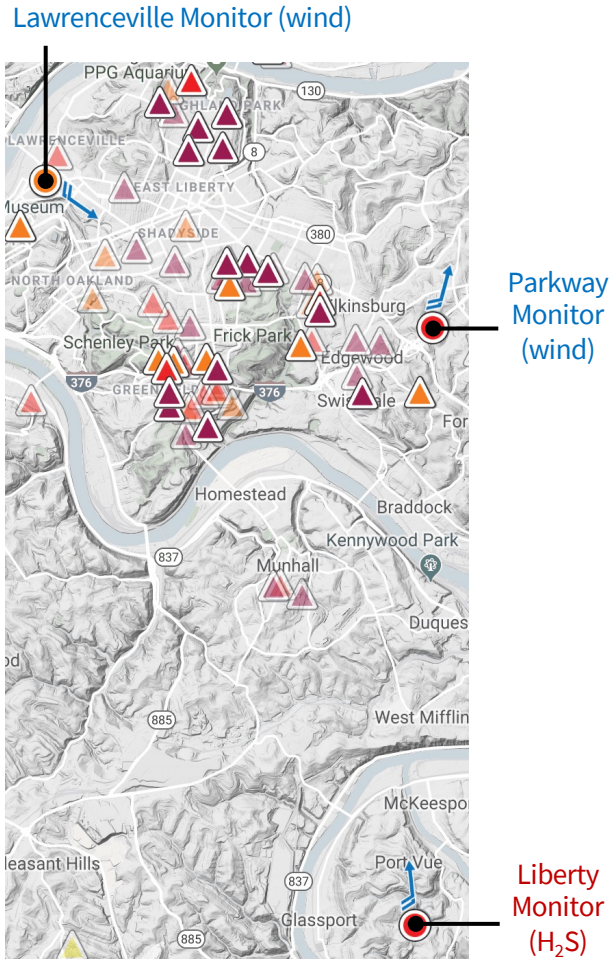




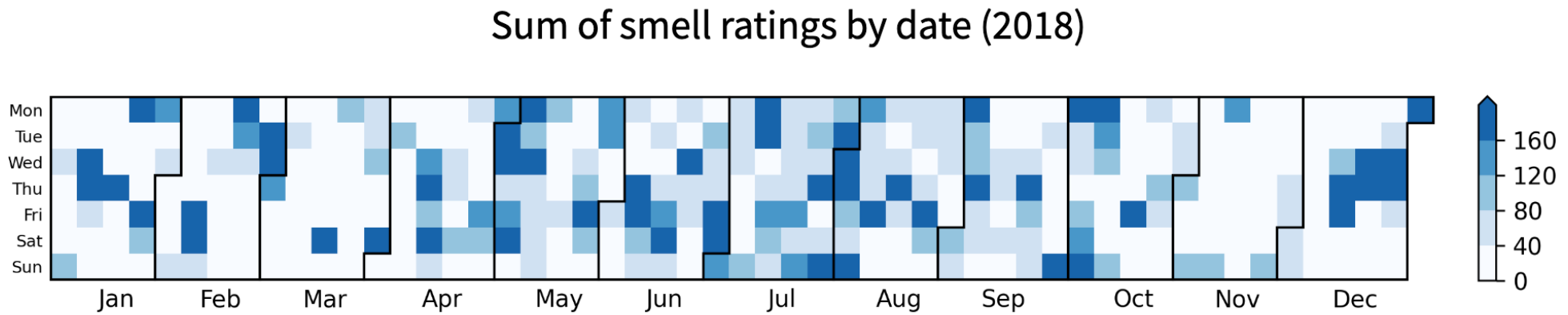
Content analysis of smell reports -- https://smellpgh.org/analysis#figure11

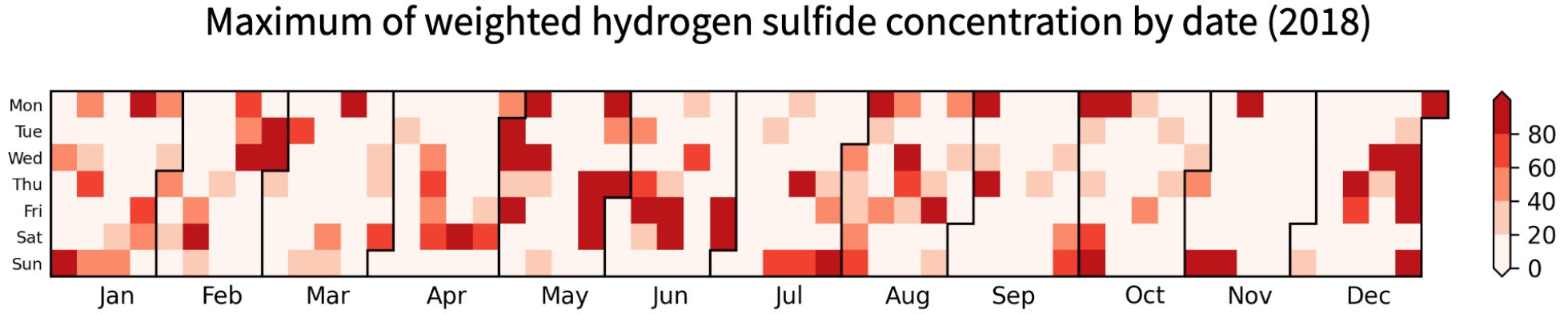
We also use Decision Tree (a machine learning model) to explain about 30% of the smell events, which is a joint effect of wind information and hydrogen sulfide.



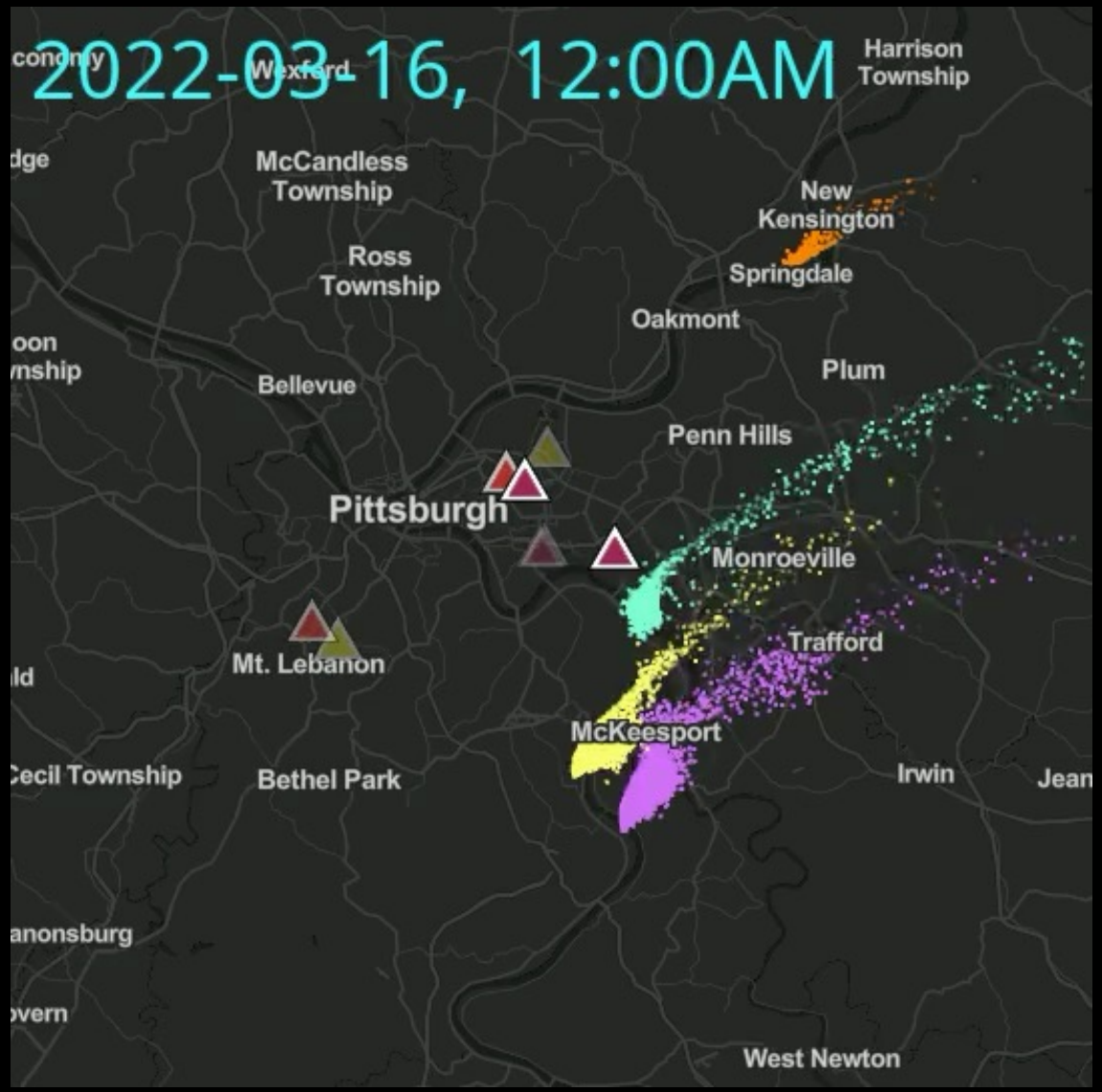


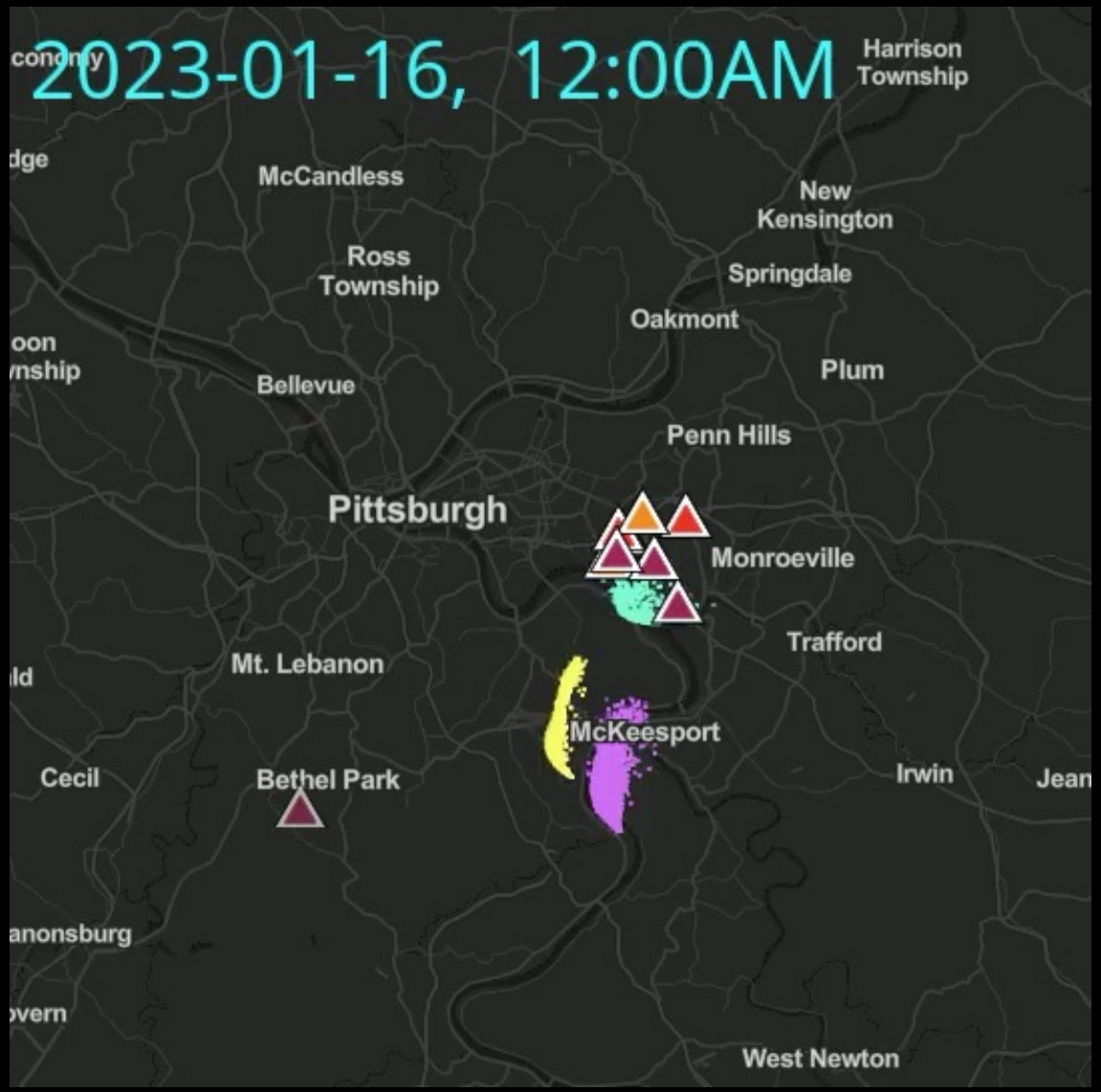
We further compute and visualize the relationship between the sum of smell ratings and the maximum concentration of weighted hydrogen sulfide for each day.



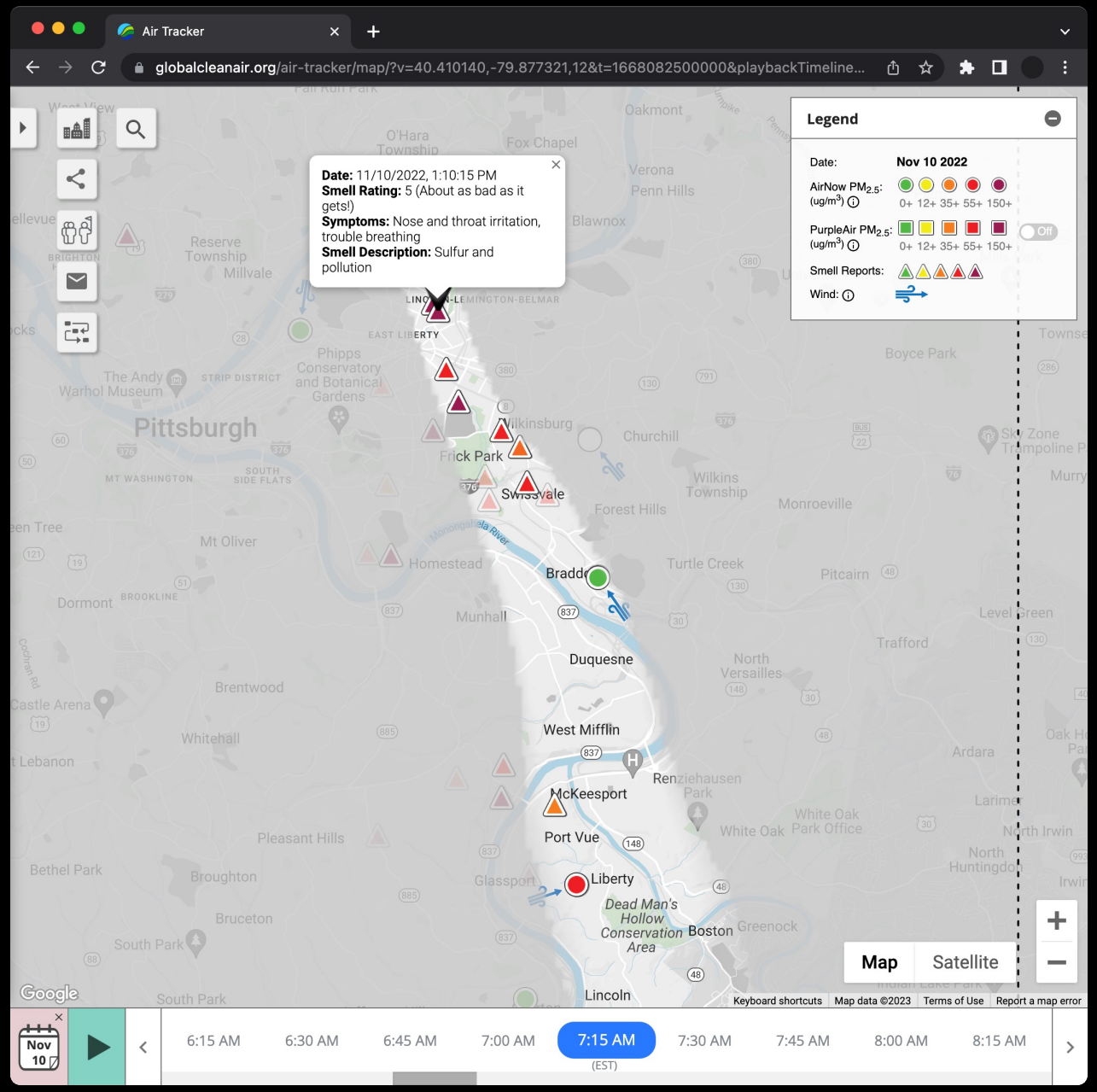


Relationship between smell reports and hydrogen sulfide -- https://smellpgh.org/analysis#figure14





Visualize how pollution emissions can align with smell reports -- https://plumepgh.org/?date=2022-03-16



Track potential air pollution sources -- https://globalcleanair.org/air-tracker/map/

Project RISE

Recognizing **I**ndustrial **S**moke **E**missions

https://smellpgh.org

Proceedings of the AAAI Conference on Artificial Intelligence

Current Archives About Search

Home / Archives / Vol. 35 No. 17: IAAI-21, EAAI-21, AAAI-21 Special Programs and Special Track / AAAI Special Track on AI for Social Impact

Project RISE: Recognizing Industrial Smoke Emissions

Yen-Chia Hsu
Carnegie Mellon University

Ting-Hao (Kenneth) Huang
Pennsylvania State University

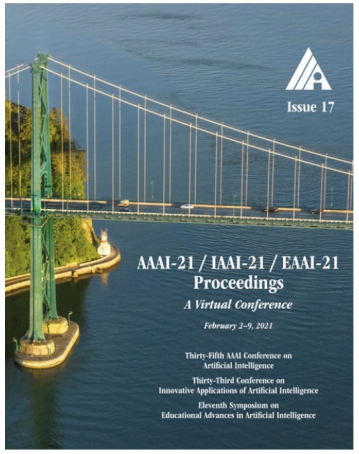
Ting-Yao Hu
Carnegie Mellon University

Paul Dille
Carnegie Mellon University

Sean Prendi
Carnegie Mellon University

Ryan Hoffman
Carnegie Mellon University

Anastasia Tsuhlares
Carnegie Mellon University



Information

For Readers

For Authors

For Librarians

Open Journal Systems

Subscription

Login to access subscriber-only resources.

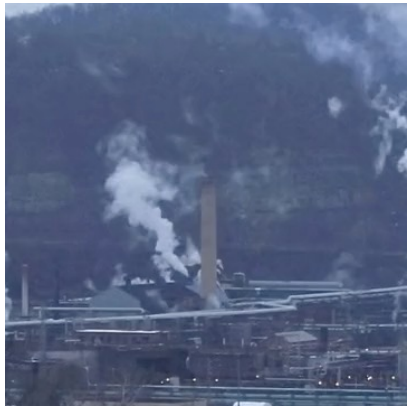
PKP|PS

This project aim to recognize industrial smoke emissions automatically on the videos obtained from a camera monitoring network.

http://smoke.createlab.org

Has smoke

Has smoke



No smoke

Has smoke

We invite communities to annotate if the videos have industrial smoke emissions using a web-based tool.

Project RISE

smoke.createlab.org/label.html

Project RISE

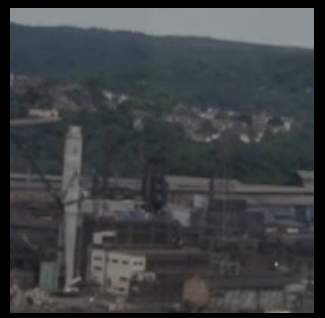
So far, 22929 (23.91%) out of 95879 videos are fully labeled, and 3403 (3.55%) videos are partially labeled (learn more). You have reviewed 431 pages, of which (researcher) have passed the quality check (learn more).

Sign Out | Interactive Tutorial | My Contribution

Each video is 3 seconds. Click or tap to select videos that have smoke. Click or tap again to deselect. Skip a video if you are not sure whether it has smoke.

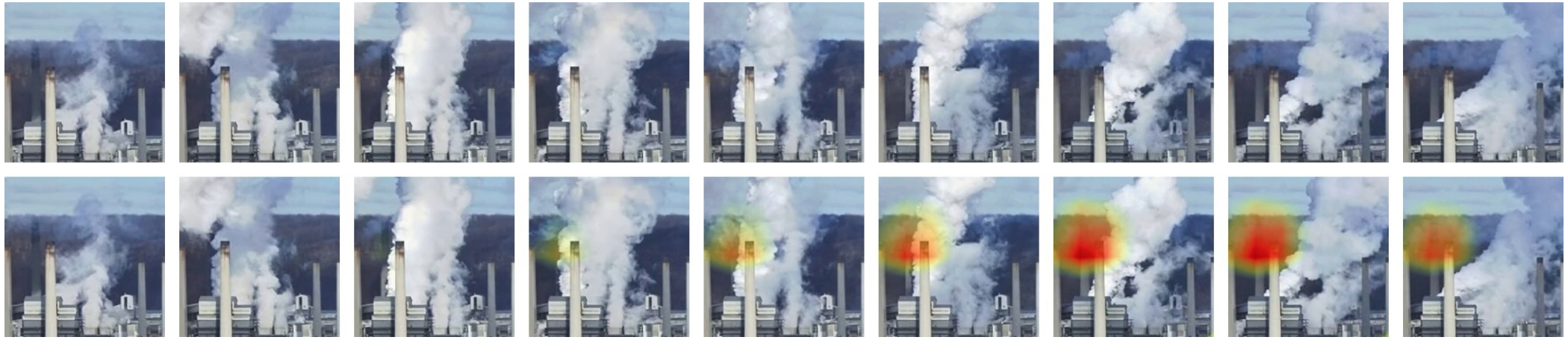
1

2



19
Views

The data enables Computer Vision applications using deep neural networks, and our baseline model (I3D+Timeception) can find emissions with reasonable performance.



	Precision*	Recall*	F-score*
Our best I3D model	0.86	0.79	0.82
Always yes	0.41	1	0.58

*Average of the metric for all data splits (training, validation, and test)

Hsu, Y. C., et al. (2021). Project RISE: Recognizing Industrial Smoke Emissions. AAAI.

Project RISE

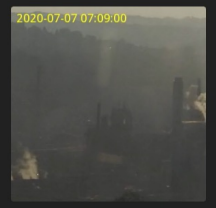
smoke.createlab.org/event.html?date=2020-07-07&...

Project RISE

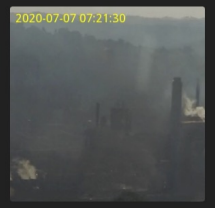
This page shows videos that the Artificial Intelligence model thinks have hazardous smoke emissions. The following image shows the camera view ID and location.

The timeline below shows smoke events (learn more), where the x-axis means hour of day. To show videos, select a date 2020-07-07 and compare the events with Smell Pittsburgh. Also, select a camera ID 2 and view ID all to filter videos.

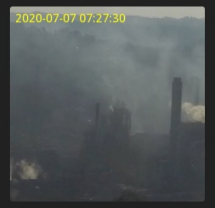
07/07/2020, 06:48
Duration: 4.7 min
View ID: 2-3
Link to Viewer



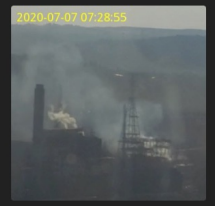
07/07/2020, 07:08
Duration: 7.8 min
View ID: 2-3
Link to Viewer



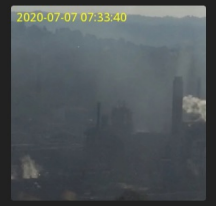
07/07/2020, 07:19
Duration: 3.3 min
View ID: 2-3
Link to Viewer



07/07/2020, 07:25
Duration: 6.2 min
View ID: 2-3
Link to Viewer

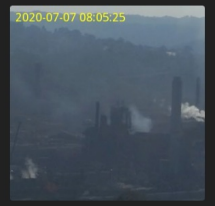


07/07/2020, 07:27



07/07/2020, 07:32

07/07/2020, 07:47

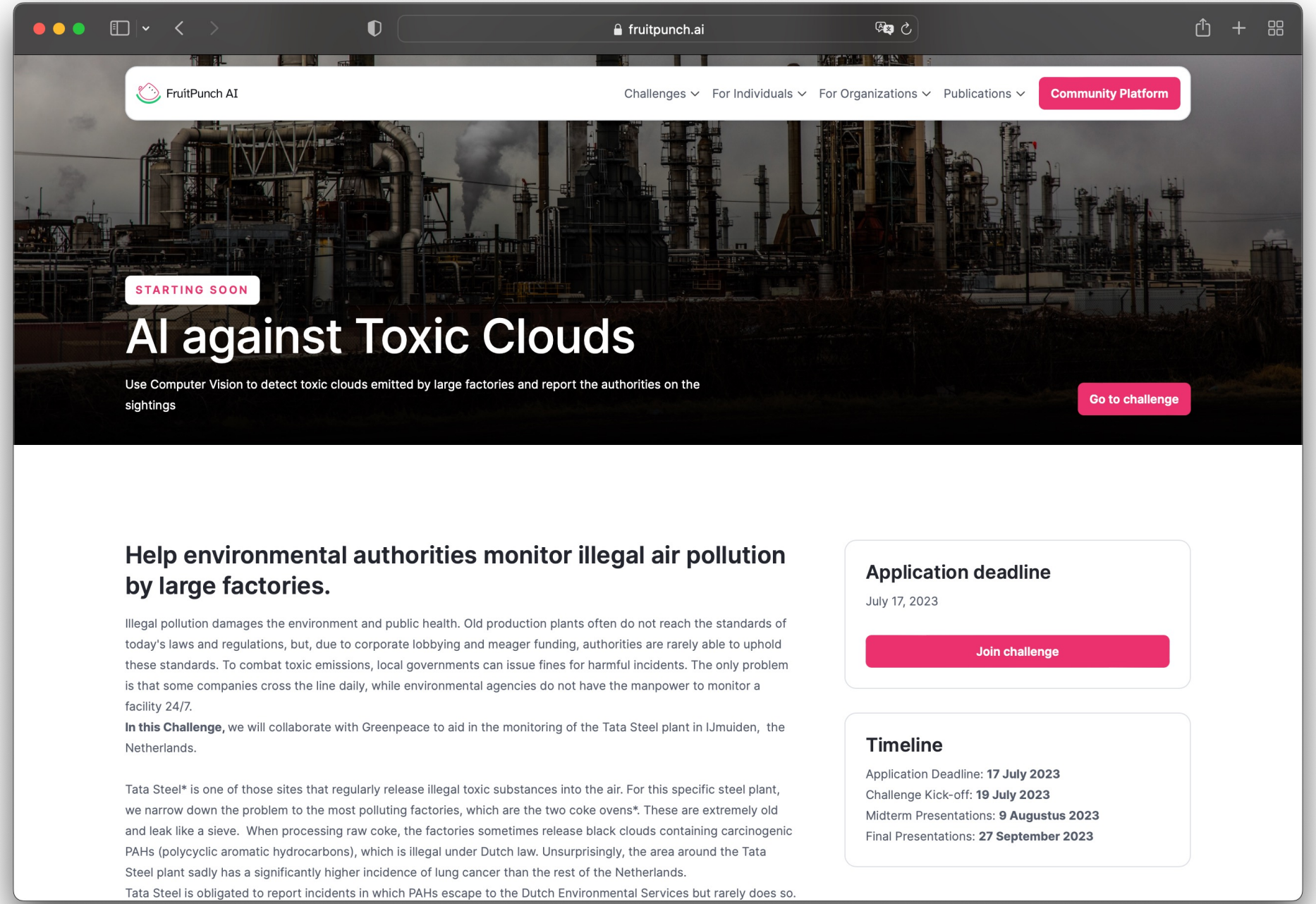


07/07/2020, 08:02

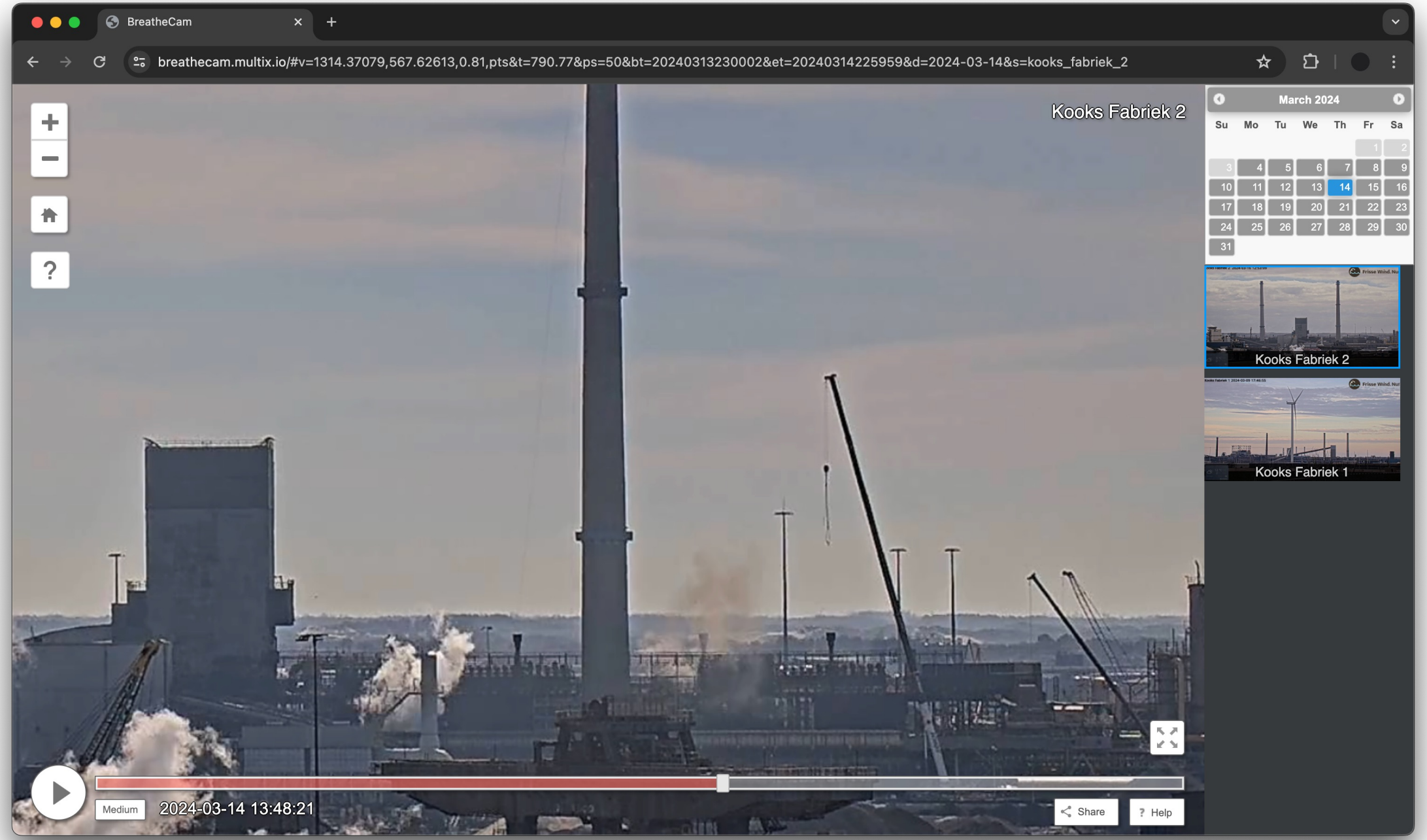
Dutch Context

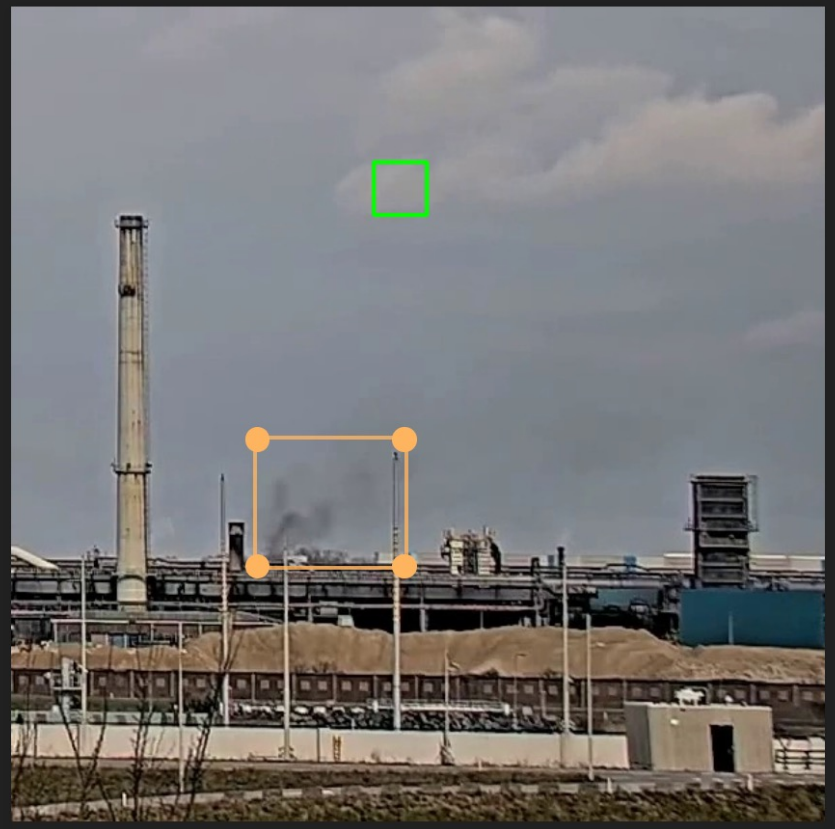
Frisse Wind. Nu!
Castricum
Middenbeemster
Heemskerk
Purmerend
Krommenie
WEIDEVENNE
Wijk aan Zee
Beverwijk
Assendelft
Velsen-Noord
ZAANSE SCHANS
Koog aan de Zaan
IJmuiden
Ilpervela
Zaandam
Oostzaan
Landsmeer
Bloemendaal
Spaarnwoude
WESTPOORT
AMSTERDAM-NOORD
Haarlem
SLOTERDIJK
Zwanenburg
Zandvoort
Amsterdam
OSDORP
Heemstede
Vijfhuizen
AMSTERDAM-ZUID
N203
A9
N244
N246
A22
A7
N235
N246
N203
N202
N208
r103
A8
A10
S102
s101
A10
A5
N200
A9
N200
N208
s116
S100
A205
N519
A10
N205
N206
N208
Google
Map data ©2025 Google
Netherlands
Terms
Privacy
Send Product Feedback
2 km

Spot the Toxic Cloud -- https://spotdegifwolk.nl/



AI against Toxic Clouds -- https://www.fruitpunch.ai/challenges/ai-against-toxic-clouds


BreatheCam
breathecam.multix.io/#v=1314.37079,567.62613,0.81,pts&t=790.77&ps=50&bt=20240313230002&et=20240314225959&d=2024-03-14&s=kooks_fabriek_2
Kooks Fabriek 2
March 2024
Su Mo Tu We Th Fr Sa
Kooks Fabriek 2
Kooks Fabriek 1
Medium
2024-03-14 13:48:21
Share
Help

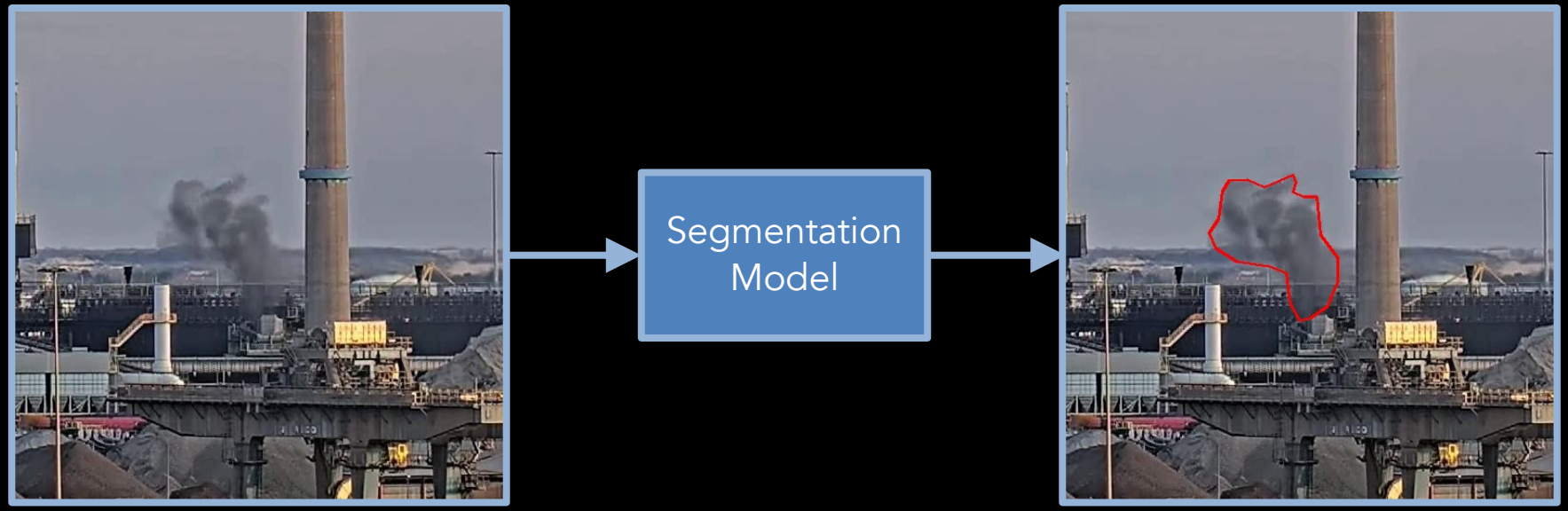


10/36

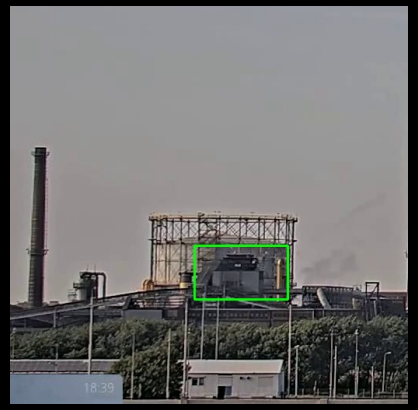
Let op: geef alleen feedback op de afbeelding met het groene kader, niet op de hele video.

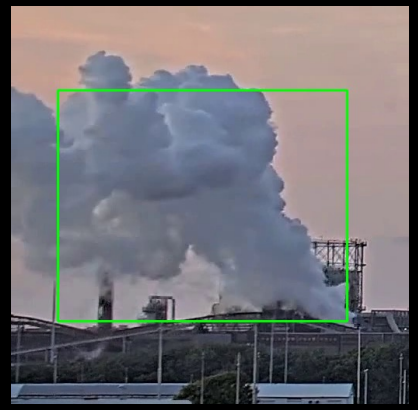
Beeld 5/8 Verwijder kader

Vorige Bevestigen en Volgende

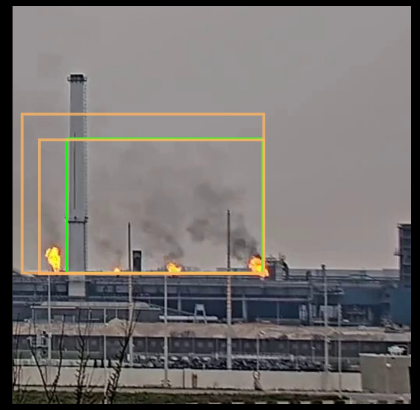


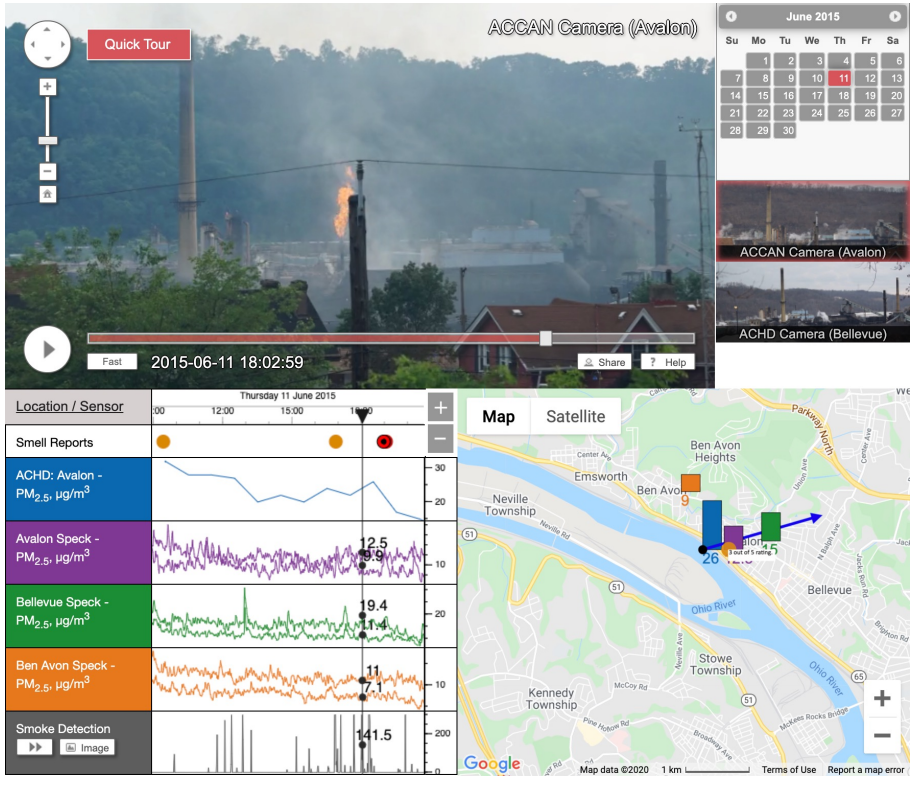
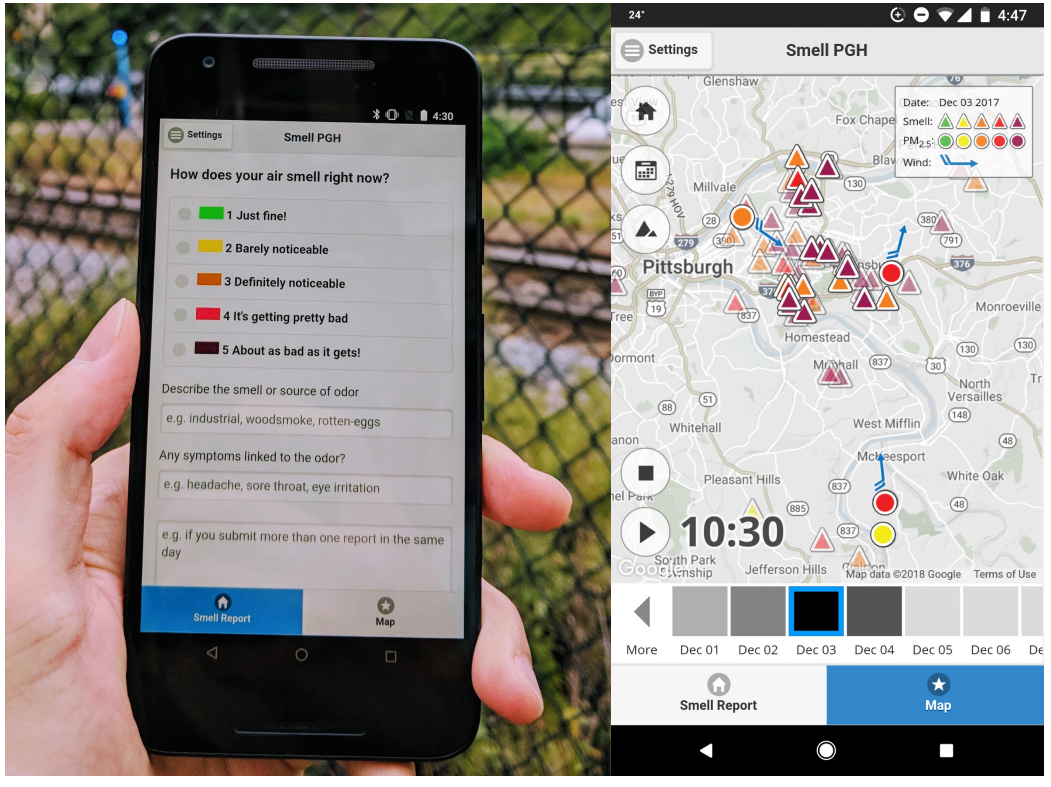
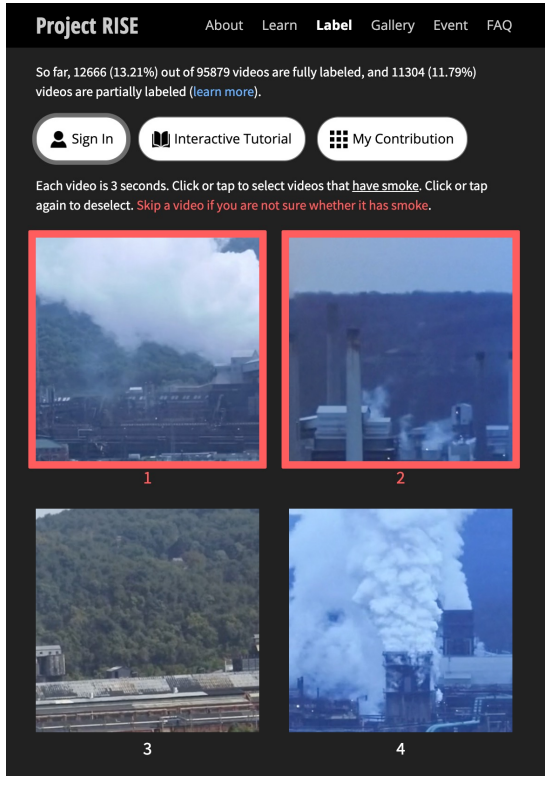
To what extent can community feedback help improve the model?


Wrong: No Smoke


Wrong: Steam

Too Small


Roughly OK

Questions?

- Yen-Chia Hsu – http://yenchiah.me
- Air Quality Monitoring – http://shenangochannel.org
- Smell Pittsburgh – https://smellpgh.org
- Project RISE – https://smoke.createlab.org
- IJmondCAM – https://ijmondcam.multix.io/