AIRBOX ANALYSIS

劉昱劭 彭敬樺 周才錢 張翔中

OUTLINE

- Motivation
- Data Collection
- Data Preprocessing
- Data Visualization
- Training Results
- Conclusion
- Group Contribution

MOTIVATION

Air pollution is a serious environmental issue that bothers our homeland nowadays, thousands of sensors are deployed to keep an eye on air quality in real time. Through the observed data in the Airboxes, we try to figure out the evidence that affect air quality, using PM2.5 as the metric.

DATA COLLECTION

- EDIMAX AIRBOX

開放資料 API

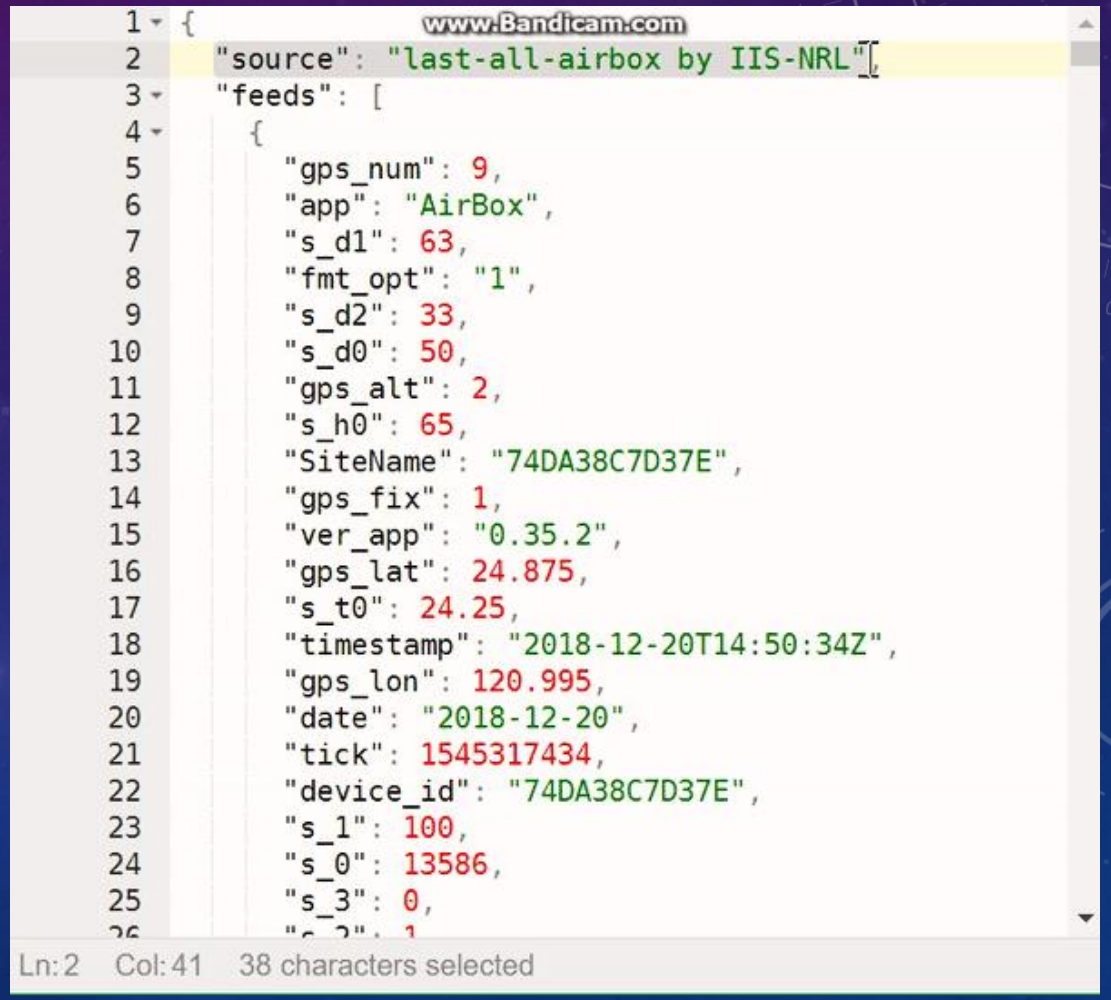
[請注意] 資料時區為：UTC+0

JSON 格式 API（更新頻率：每 5 分鐘）

EDIMAX AIRBOX 空氣盒子	中研院 MAPS 感測器	LASS 社群測站
LASS4U 感測器	G0V INDIE 感測器	G0V PROBECUBE 感測器

DATA COLLECTION

- JSON data
- Update every 5 minutes



DATA COLLECTION

- Use shell script to fetch

```
--2018-12-20 22:59:15--  https://pm25.lass-net.org/data/last-all-airbox.json
Resolving pm25.lass-net.org (pm25.lass-net.org)... 35.187.152.119
Connecting to pm25.lass-net.org (pm25.lass-net.org)|35.187.152.119|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 844654 (825K) [application/json]
Saving to: ‘2018-12-20-2259.json’

--2018-12-20 23:04:15--  https://pm25.lass-net.org/data/last-all-airbox.json
Resolving pm25.lass-net.org (pm25.lass-net.org)... 35.187.152.119
Connecting to pm25.lass-net.org (pm25.lass-net.org)|35.187.152.119|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 842583 (823K) [application/json]
Saving to: ‘2018-12-20-2304.json’

2018-12-20-2304.json                100%[=============================================================>] 822.83K

2018-12-20 23:04:16 (11.1 MB/s) - ‘2018-12-20-2304.json’ saved [842583/842583]

Fetch time: 2018-12-20-2304
```

DATA PREPROCESSING

- JSON to CSV

```
~/ML/m2/final/logfile/json(master*) » ipython toTable.py
not processed yet:

2018-12-20-2304.json
2018-12-20-2309.json
2018-12-20-2314.json
2018-12-20-2319.json
2018-12-20-2324.json
2018-12-20-2329.json
2018-12-20-2334.json
2018-12-20-2339.json
2018-12-20-2344.json
2018-12-20-2349.json
2018-12-20-2354.json
--------------------
(After) not processed yet:

--------------------
```

DATA PREPROCESSING (CONT.)

- JSON to CSV
- 1800 rows
- 10 days => over
  5 million rows

	SiteName	App	date	device_id	fmt_opt	gps_alt	gps_fix	gps_lat	gps_lon	gps_num	...	s_d0	s_d1	s_d2	s_h0	s_t0	tick
0	74DA38C7D37E	AirBox	2018-12-06	74DA38C7D37E	1	2.0	1.0	24.875	120.995	9.0	...	29.0	33.0	20.0	73.0	27.62	1.544083e+09
1	74DA38B0538C	AirBox	2018-12-06	74DA38B0538C	1	2.0	1.0	0.000	0.000	9.0	...	22.0	24.0	15.0	69.0	30.37	1.544085e+09
2	74DA38F20B88	AirBox	2018-12-06	74DA38F20B88	1	2.0	1.0	22.750	120.316	9.0	...	55.0	67.0	37.0	68.0	30.00	1.544083e+09
3	74DA38F20B86	AirBox	2018-12-06	74DA38F20B86	1	2.0	1.0	22.733	120.248	9.0	...	37.0	46.0	23.0	74.0	30.87	1.544084e+09
4	74DA38B0514E	AirBox	2018-12-06	74DA38B0514E	1	2.0	1.0	25.018	121.304	9.0	...	30.0	36.0	21.0	75.0	29.25	1.544084e+09
5	74DA38F20B82	AirBox	2018-12-06	74DA38F20B82	1	2.0	1.0	22.847	120.465	9.0	...	61.0	86.0	53.0	68.0	30.00	1.544084e+09
6	74DA38B051F0	AirBox	2018-12-06	74DA38B051F0	1	2.0	1.0	23.796	120.223	9.0	...	33.0	40.0	23.0	100.0	30.25	1.544084e+09
7	74DA38B051F4	AirBox	2018-12-06	74DA38B051F4	1	2.0	1.0	25.034	121.525	9.0	...	1.0	1.0	0.0	75.0	30.00	1.544085e+09
8	74DA38AF47B0	AirBox	2018-12-06	74DA38AF47B0	1	2.0	1.0	23.464	120.417	9.0	...	27.0	30.0	21.0	73.0	30.62	1.544083e+09
9	[illegible]	AirBox	2018-12-06	74DA38F20AA4	1	2.0	1.0	22.685	120.492	9.0	...	59.0	72.0	41.0	62.0	32.25	1.544085e+09
10	74DA38E2B586	AirBox	2018-12-06	74DA38E2B586	1	2.0	1.0	25.034	121.534	9.0	...	22.0	23.0	14.0	67.0	29.50	1.544085e+09
11	[illegible]	AirBox	2018-12-06	74DA38F20AA0	1	2.0	1.0	24.588	120.735	9.0	...	31.0	38.0	19.0	83.0	26.50	1.544082e+09

DATA PREPROCESSING (CONT.)

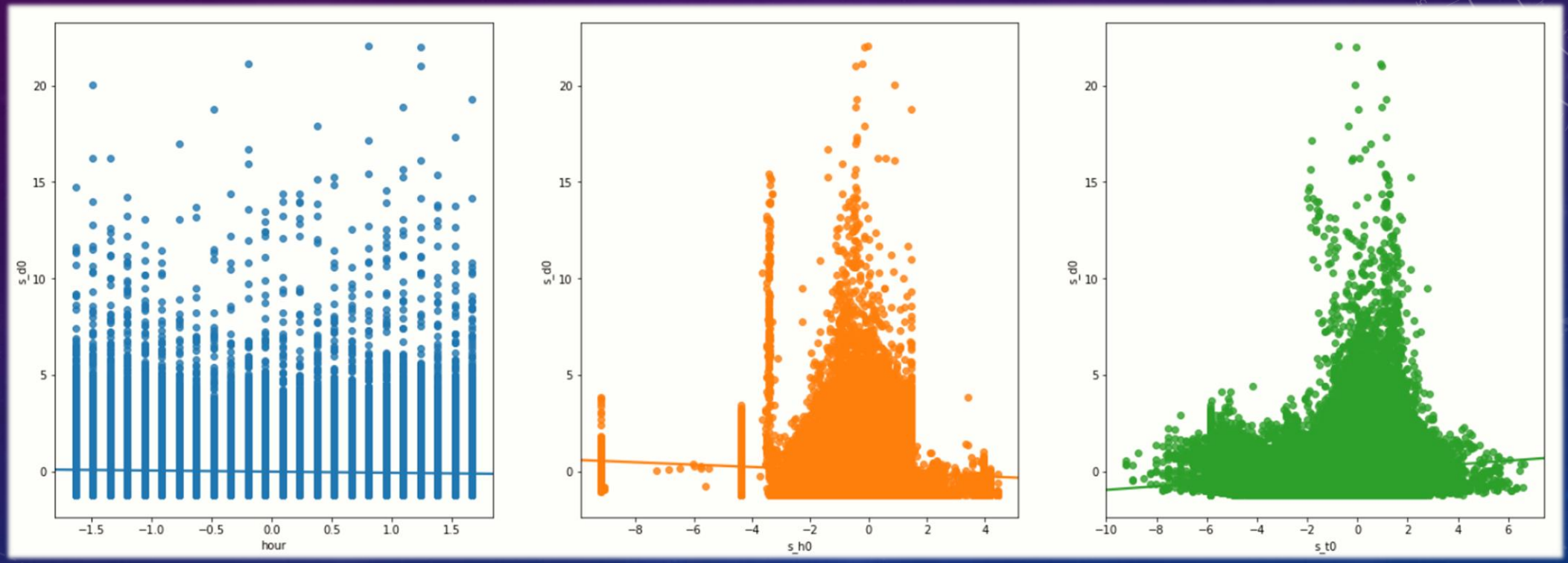
- Drop useless columns

	SiteName	date	humidity	temperature	time	hour	pm2.5	pm1.0	pm10.0
0	74DA38EBF78E	2018-12-07	70.0	24.62	20:29:58	20	0.0	0.0	0.0
1	苗栗縣縣立南埔國小	2018-12-07	75.0	21.50	20:46:40	20	0.0	0.0	0.0
2	苗栗縣縣立田美國小	2018-12-07	82.0	19.75	20:32:28	20	0.0	0.0	0.0
3	74DA38B0538C	2018-12-07	60.0	29.00	20:44:11	20	0.0	0.0	0.0
4	changhua18	2018-12-07	100.0	24.25	20:36:37	20	13.0	13.0	12.0

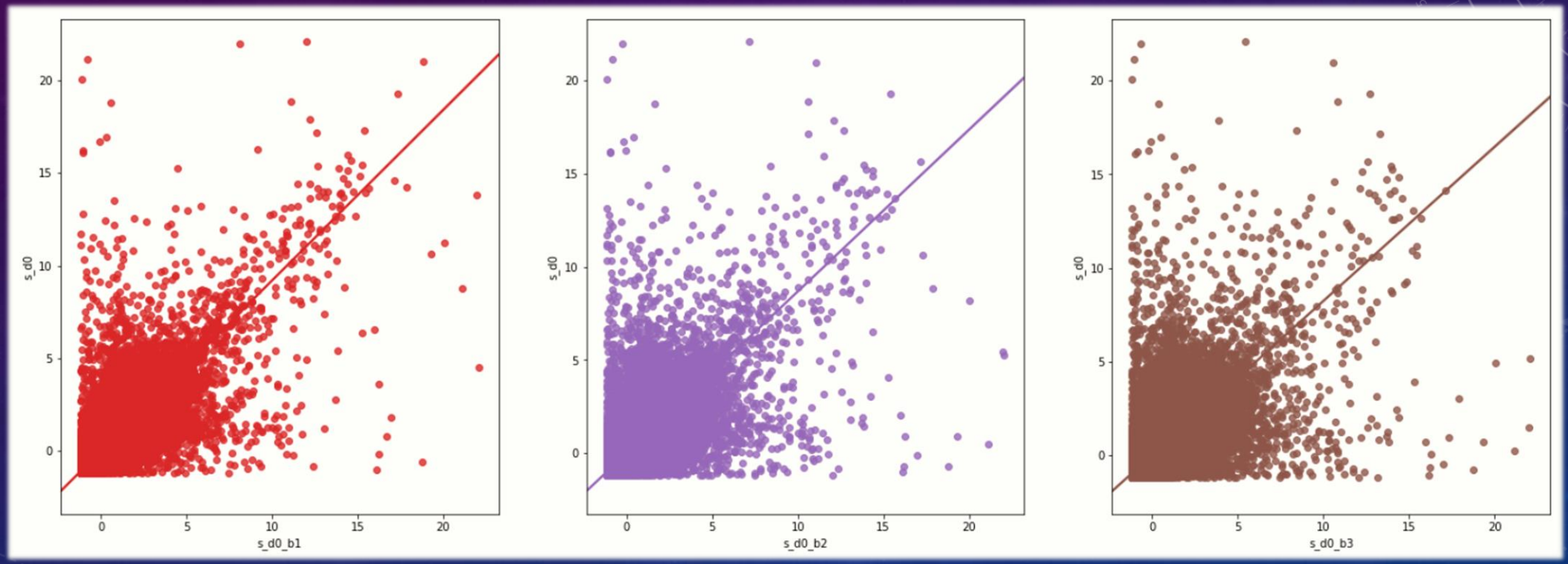
DATA PREPROCESSING (CONT.)

- **Group data**
  - ✓ SiteName > date > hour
  - ✓ multiple rows are averaged within groups
- **Create new columns**
  - ✓ s_d0_b1 / s_d0_b2 / s_d0_b3
  - ✓ averaged PM2.5 in previous 1/2/3 hour(s)

DATA VISUALIZATION



DATA VISUALIZATION (CONT.)



TRAINING RESULTS (BAD)

- Input features:
  - ✓ hour, humidity, temperature

model	MSE	R2
SGD	416.134126	0.011176
Nearest Neighbors	474.128305	-0.126630
Decision Tree	384.951370	0.085273
Random Forest	380.620162	0.095565
Neural Net	406.638213	0.033741

model	MSE	R2
Ridge1	415.779871	0.012018
Ridge2	409.395326	0.027189
Ridge3	404.176106	0.039591
Ridge4	400.773313	0.047677
Ridge5	394.678103	0.062160

TRAINING RESULTS (GOOD)

- Input features:
  - ✓ hour, humidity, temperature, s_d0_b1, s_d0_b2, s_d0_b3

model	MSE	R2
SGD	62.426707	0.851661
Nearest Neighbors	68.699790	0.836755
Decision Tree	63.552603	0.848985
Random Forest	68.824660	0.836458
Neural Net	58.989715	0.859828

model	MSE	R2
Ridge1	62.192023	0.852218
Ridge2	59.522047	0.858563
Ridge3	58.831554	0.860204
Ridge4	63.374947	0.849408
Ridge5	111.430257	0.735218

CONCLUSION

- Time and weather doesn’t have much apparent relationship with PM2.5
- Derivative attributes that we created make data more fit to the regression models

THANK YOU !