Universal Data Structures for Air Quality Data
Target Audience

- Sensor manufacturers
- Data aggregators
- Local AQ agencies
- National AQ agencies
- NGOs
- Software companies
- Data analysts
- Air quality advisors
- Post docs
- Grad students
- IT teams
- Software developers

Basically, anyone working with AQ data.
Why trust me?

- Grad school experience building instruments
- 30 years doing data visualization
- 12 years running a business writing operational software
- 10 years working with air quality data
- 4 years working with sensor data
- I maintain the **PWFSLSmoke** and **AirSensor** R packages
Data Producers & Data Consumers

Producers

Hardware & Software Engineers

Concerns

- Electronics (amps, ADCs, wifi chips)
- Firmware
- Data transfer protocols
- Real-time data storage and retrieval
- Cost / size / reliability
- Single device type

Consumers

Scientists, Analysts & Statisticians

Concerns

- Data access
- Data usability
- Quality Control
- Statistics
- Data visualization
- Multiple device types
Scientific Data Management

Goal

- Meet needs of engineers and analysts

Concerns

- Engineering variables, units and formats
- Instrument specific concerns
- Analyst general questions
- Raw data ingest
- Data harmonization
- QC algorithms
- Data aggregation
- Data access
Scientific Data Management

1. Standardize/harmonize/correct low level data
   a. Download
   b. Parse
   c. Harmonize
   d. Add metadata
   e. Quality Control

2. Combine low level data into useful summaries
   a. Aggregate to hourly
   b. Combine multiple time series
   c. Use a common data format

3. Make data easily accessible
Google Maps -- low level data

Lowest level has lots of details
Each pixel represents ~15 cm
Zoom level 21 has ~25,000 Terabytes of data

Great for diving into the details.
Google Maps -- useful summary 1

Higher level summary

Each pixel represents ~15 m

Zoom level 13 has ~4.4 Terabytes of data

Enhanced with spatial metadata

Great for driving.
Google Maps -- useful summary 2

Even higher level summary

Each pixel represents ~1.0 klm

Zoom level 7 has ~1.1 Gigabytes of data

Enhanced with elevation data

Great for regional planning.
For Air Quality data, people want maps and time series
| Date               | CO       | O3       | NOx      | Temp  | Humidity| Pressure| Wind Speed | Wind Direction | AQI   |
|--------------------|----------|----------|----------|-------|---------|---------|------------|----------------|-------|
| 2021-10-07T07:01:00Z | 27.09    | 27.82   | 71       | 53    | 973.3   | 18.6    | 27.01      | 27.82          | 18.6  |
| 2021-10-07T07:01:03Z | 26.95    | 28.04   | 71       | 53    | 973.28  | 18.22   | 26.95      | 30.09          | 18.22 |
| 2021-10-07T07:01:05Z | 26.03    | 29.18   | 71       | 53    | 973.28  | 17.78   | 26.03      | 27.82          | 17.78 |
| 2021-10-07T07:01:07Z | 26.37    | 29.52   | 70       | 54    | 973.28  | 17.87   | 26.37      | 30.05          | 17.87 |
| 2021-10-07T07:01:09Z | 27.09    | 29.58   | 71       | 53    | 973.3   | 18.25   | 27.09      | 31.07          | 18.25 |
| 2021-10-07T07:01:11Z | 28.11    | 31.84   | 70       | 54    | 973.34  | 19.3    | 28.11      | 31.58          | 19.3  |
| 2021-10-07T07:01:13Z | 27.66    | 29.53   | 70       | 54    | 973.24  | 18.53   | 27.66      | 30.16          | 18.53 |
| 2021-10-07T07:01:15Z | 28.32    | 30.21   | 70       | 54    | 973.21  | 18.85   | 28.32      | 31.75          | 18.85 |
| 2021-10-07T07:01:17Z | 28.52    | 29.78   | 70       | 54    | 973.28  | 18.97   | 28.52      | 31.35          | 18.97 |
| 2021-10-07T07:01:19Z | 29.03    | 29.89   | 70       | 54    | 973.26  | 18.79   | 29.03      | 31.57          | 18.79 |
| 2021-10-07T07:01:21Z | 29.07    | 30.32   | 70       | 55    | 973.26  | 18.78   | 29.07      | 31.73          | 18.78 |
| 2021-10-07T07:01:23Z | 29.56    | 30.74   | 70       | 54    | 973.26  | 18.78   | 29.56      | 31.73          | 18.78 |
| 2021-10-07T07:01:25Z | 29.28    | 31.37   | 69       | 55    | 973.34  | 19.46   | 29.28      | 31.91          | 19.46 |
| 2021-10-07T07:01:27Z | 29.02    | 30.43   | 69       | 56    | 973.35  | 19.66   | 29.02      | 31.95          | 19.66 |
| 2021-10-07T07:01:29Z | 29.03    | 32.74   | 69       | 56    | 973.38  | 19.64   | 29.03      | 31.95          | 19.64 |
| 2021-10-07T07:01:31Z | 29.65    | 32.33   | 69       | 56    | 973.41  | 19.13   | 29.65      | 32.77          | 19.13 |
| 2021-10-07T07:01:33Z | 28.84    | 31.93   | 70       | 56    | 973.4   | 19.1   | 28.84      | 32.08          | 19.76 |
| 2021-10-07T07:01:35Z | 28.51    | 32.33   | 70       | 56    | 973.43  | 19.25   | 28.51      | 32.81          | 19.93 |
| 2021-10-07T07:01:37Z | 28.07    | 30.07   | 68       | 56    | 973.49  | 18.73   | 28.07      | 31.81          | 19.19 |
| 2021-10-07T07:01:39Z | 27.09    | 27.82   | 71       | 53    | 973.3   | 18.6    | 27.09      | 30.13          | 18.6  |

**Plus 822 more lines**

All Parameters, 1 Day, 1 Sensor (112 Kilobytes)

Great for diving into details.
1 Parameter, 1 Day, 1 Sensor

Raw = 112 Kilobytes
Summarized = 606 bytes

Great for plotting time series.
| Date         | Value1 | Value2 | Value3 | Value4 | Value5 | Value6 | Value7 | Value8 | Value9 | Value10 | Value11 | Value12 |
|-------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|---------|---------|
| 2021-10-07T07:00:00Z | 28     | 28     | 26     | 29     | 28     | 26     | 27     | 24     | 20     | 17      | 19      |         |
| 2021-10-07T08:00:00Z | 31     | 31     | 28     | 34     | 33     | 28     | 27     | 24     | 20     | 23      |         |         |
| 2021-10-07T09:00:00Z | 32     | 31     | 31     | 38     | 36     | 30     | 29     | 29     | 31     | 24      | 24      |         |
| 2021-10-07T10:00:00Z | 36     | 31     | 36     | 39     | 37     | 35     | 31     | 33     | 35     | 36      | 28      | 24      |
| 2021-10-07T11:00:00Z | 37     | 33     | 35     | 38     | 37     | 34     | 33     | 34     | 34     | 35      | 35      | 25      |
| 2021-10-07T12:00:00Z | 36     | 28     | 36     | 40     | 38     | 36     | 32     | 33     | 36     | 34      | 27      | 23      |
| 2021-10-07T13:00:00Z | 38     | 32     | 37     | 39     | 39     | 36     | 35     | 35     | 34     | 35      | 34      | 25      |
| 2021-10-07T14:00:00Z | 38     | 36     | 39     | 40     | 38     | 38     | 34     | 36     | 39     | 39      | 39      | 29      |
| 2021-10-07T15:00:00Z | 37     | 36     | 39     | 42     | 38     | 38     | 32     | 34     | 39     | 40      | 30      | 31      |
| 2021-10-07T16:00:00Z | 35     | 34     | 35     | 40     | 38     | 35     | 31     | 33     | 36     | 37      | 28      | 32      |
| 2021-10-07T17:00:00Z | 15     | 32     | 31     | 32     | 31     | 31     | 16     | 16     | 31     | 31      | 21      | 30      |
| 2021-10-07T18:00:00Z | 8      | 27     | 24     | 22     | 15     | 23     | 7      | 7      | 24     | 27      | 17      | 25      |
| 2021-10-07T19:00:00Z | 7      | 20     | 22     | 25     | 21     | 22     | 8      | 8      | 22     | 21      | NA      | 21      |
| 2021-10-07T20:00:00Z | 23     | 12     | 15     | 22     | 21     | 16     | 21     | 22     | 15     | 13      | NA      | 11      |
| 2021-10-07T21:00:00Z | 17     | 11     | 13     | 18     | 16     | 13     | 16     | 17     | 13     | 12      | 10      | 9       |
| 2021-10-07T22:00:00Z | 15     | 12     | 12     | 14     | 13     | 12     | 14     | 15     | 11     | 11      | NA      | 10      |
| 2021-10-07T23:00:00Z | 14     | 12     | 11     | 14     | 12     | 11     | 13     | 13     | 11     | 11      | 8       | 9       |
| 2021-10-08T00:00:00Z | 12     | 9      | 11     | 13     | 11     | 11     | 12     | 10     | 9      | 7       | 9       |         |
| 2021-10-08T01:00:00Z | 12     | 9      | 11     | 9      | 7      | 8      | 8      | 9      | 8      | 7       | 6       | 5       |
| 2021-10-08T02:00:00Z | 9      | 8      | 9      | 10     | 9      | 8      | 8      | 8      | 8      | 7       | 7       |         |
| 2021-10-08T03:00:00Z | 8      | 9      | 9      | 11     | 9      | 9      | 8      | 8      | 9      | 11      | 6       | 7       |
| 2021-10-08T04:00:00Z | 6      | 12     | 11     | 12     | 11     | 10     | 7      | 7      | 10     | 11      | 7       | 8       |
| 2021-10-08T05:00:00Z | 7      | 13     | 12     | 13     | 12     | 12     | 7      | 7      | 12     | 12      | 8       | 8       |
| 2021-10-08T06:00:00Z | 11     | 14     | 15     | 16     | 15     | 11     | 11     | 15     | 15     | 9       | 11      |         |

Air Quality Data -- summary 2

1 Parameter, 1 Day, 12 Sensors
Raw = 1.34 Megabytes
Summarized = 1.58 Kilobytes
Great for maps AND time series.
Air Quality Data – high level summary (compact!!)

2021-10-07T07:00:00Z, 28, 28, 26, 29, 28, 26, 26, 27, 24, 20, 17, 19
2021-10-07T08:00:00Z, 31, 31, 28, 34, 33, 28, 27, 28, 27, 24, 20, 23
2021-10-07T09:00:00Z, 32, 31, 31, 38, 36, 30, 29, 29, 31, 31, 24, 24
2021-10-07T10:00:00Z, 36, 31, 36, 39, 37, 35, 31, 33, 35, 36, 28, 24
2021-10-07T11:00:00Z, 37, 33, 35, 38, 37, 34, 33, 34, 34, 35, 25, 28
2021-10-07T12:00:00Z, 36, 28, 36, 40, 38, 36, 32, 33, 36, 34, 27, 23
2021-10-07T13:00:00Z, 38, 32, 37, 39, 39, 36, 34, 35, 35, 34, 25, 28
2021-10-07T14:00:00Z, 38, 36, 39, 40, 38, 38, 34, 36, 39, 39, 29, 32
2021-10-07T15:00:00Z, 37, 36, 39, 42, 38, 38, 32, 34, 39, 40, 30, 31
2021-10-07T16:00:00Z, 35, 34, 35, 40, 38, 35, 31, 33, 36, 37, 28, 32
2021-10-07T17:00:00Z, 15, 32, 31, 32, 31, 31, 16, 16, 31, 31, 21, 30
2021-10-07T18:00:00Z, 8, 27, 24, 22, 15, 23, 7, 7, 24, 27, 17, 25
2021-10-07T19:00:00Z, 7, 20, 22, 25, 21, 22, 8, 8, 22, 21, NA, 21
2021-10-07T20:00:00Z, 23, 12, 15, 22, 21, 16, 21, 22, 15, 13, NA, 11
2021-10-07T21:00:00Z, 17, 11, 13, 18, 16, 13, 16, 17, 13, 12, 10, 9
2021-10-07T22:00:00Z, 15, 12, 12, 14, 13, 12, 14, 15, 11, 11, NA, 10
2021-10-07T23:00:00Z, 14, 12, 11, 14, 12, 11, 13, 13, 11, 11, 8, 9
2021-10-08T00:00:00Z, 12, 9, 11, 13, 11, 11, 12, 10, 9, 7, 9
2021-10-08T01:00:00Z, 9, 7, 9, 9, 7, 8, 8, 9, 8, 7, 6, 5
2021-10-08T02:00:00Z, 9, 8, 9, 10, 9, 8, 8, 8, 8, 7, 7
2021-10-08T03:00:00Z, 8, 9, 9, 11, 9, 9, 8, 8, 9, 11, 6, 7
2021-10-08T04:00:00Z, 6, 12, 11, 12, 11, 10, 7, 7, 10, 11, 7, 8
2021-10-08T05:00:00Z, 7, 13, 12, 13, 12, 12, 7, 7, 12, 12, 8, 8
2021-10-08T06:00:00Z, 11, 14, 15, 16, 16, 15, 11, 11, 15, 15, 9, 11

Time Series

Map
A Maximally Compact “Universal” Data Model

For “stationary” time series only

All time dependent measurements go into a ‘data’ table

All static, spatial/instrument metadata goes into a ‘meta’ table

A unique ‘deviceDeploymentID’ connects the tables
# Air Quality Metadata – high level summary

| Field                        | Description                                                                 |
|------------------------------|------------------------------------------------------------------------------|
| deviceDeploymentID           | device ID for each deployment                                                |
| deviceID                     | device ID for each device                                                    |
| deviceTypeID                 | device ID for each device                                                    |
| device_type                  | device type                                                                  |
| pollutant                    | pollutant                                                                    |
| units                        | units                                                                        |
| locationID                   | location ID for each location                                                |
| units                        | units                                                                        |
| latitude                     | latitude                                                                     |
| longitude                    | longitude                                                                    |
| elevation                    | elevation                                                                    |
| countryCode                  | country code                                                                 |
| stateCode                    | state code                                                                   |
| countyName                   | county name                                                                  |
| street                       | street                                                                       |
| houseNumber                  | house number                                                                 |
| zip                           | zip                                                                           |
| AQSID                        | AQS ID                                                                       |
| airnow_siteCode              | airnow site code                                                             |
| airnow_agencyName            | airnow agency name                                                           |
| airnow_FIPSMSACode           | airnow MSAC code                                                             |
| airnow_EPARegion             | airnow EPA region                                                            |
| airnow_MSAName               | airnow MSAName                                                               |
| airnow_agencyID              | airnow agency ID                                                             |
| airnow_GMTOffsetHours        | airnow GMT offset hours                                                     |
| address                      | address                                                                       |
| wrcc_type                    | WRCC type                                                                    |
| wrcc_monitorType             | WRCC monitor type                                                            |
| wrcc_serialNumber            | WRCC serial number                                                           |

Only 1 entry per “device-deployment”.
## Compact ‘meta’ table – ‘ID’ is the primary key

| ID  | locationName                | longitude | latitude | elevation | countryCode | stateCode | county      | timezone          |
|-----|-----------------------------|-----------|----------|-----------|-------------|-----------|-------------|-------------------|
| 1   | Fairhope, Alabama           | -87.9     | 30.5     | 37.2      | US          | AL        | Baldwin     | America/Chicago   |
| 2   | Ashland                     | -85.8     | 33.3     | 344.      | US          | AL        | Clay        | America/Chicago   |
| 3   | Muscle Shoals               | -87.6     | 34.8     | 122       | US          | AL        | Colbert     | America/Chicago   |
| 4   | Muscle Shoals               | -87.6     | 34.8     | 122       | US          | AL        | Colbert     | America/Chicago   |
| 5   | Crossville                  | -86.0     | 34.3     | 500       | US          | AL        | DeKalb      | America/Chicago   |
| 6   | Brewton (Closed 12/30/07)   | -87.1     | 31.1     | 50        | US          | AL        | Escambia    | America/Chicago   |
| 7   | Gadsden C. College          | -86.0     | 34.0     | 50        | US          | AL        | Etowah      | America/Chicago   |
| 8   | Dothan                      | -85.4     | 31.2     | 102       | US          | AL        | Houston     | America/Chicago   |
| 9   | Dothan (Civic Center)       | -85.4     | 31.2     | 264       | US          | AL        | Houston     | America/Chicago   |
## Compact ‘data’ table

| datetime     | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|--------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 2020-01-01 05:00:00 | NA | 5.1| 1.5| 4.4| NA | NA | NA | 4.5| 8.8| NA | NA | NA | NA | NA | 2.9| 4.6| NA | NA |
| 2020-01-01 06:00:00 | NA | 4.2| 0.5| 5.7| NA | NA | NA | 4.3| 7.6| NA | NA | 11.0| NA | NA | 7 | 2.7| 6.6| NA | 3.3 |
| 2020-01-01 07:00:00 | NA | 3.0| 0.3| 5.5|-2 | NA | NA | 4.3| 5.2| NA | NA | 4.3 | 349.0| NA | 5 | 2.2| 4.8| NA | 4.8 |
| 2020-01-01 08:00:00 | 2  | 3.3| 0.7| 5.8|-3 | 12 | 4.1| 4.5| 6.5| 11 | NA | 4.8 | 462.9| 105| 4 | 1.9| 3.0| 16 | 4.2 |
| 2020-01-01 09:00:00 | 3  | 3.0| 1.0| 5.8| 1 | 2.7| 4.2| 5.4| 7.2| 7  | NA | 6.4 | 549.8| 118| 4 | 1.9| 2.4| 14 | 4.5 |
| 2020-01-01 10:00:00 | 4  | 3.8| 0.8| 5.8| 1 | 2.7| 22 | 5.6| 8.4| 9  | NA | 7.4 | 550.0| 70 | 1 | 1.8| 3.3| 9  | 6.5 |
| 2020-01-01 11:00:00 | 3  | 3.8| 1.6| 6.1|-1 | 7  | 24 | 5.7| 9.2| 6  | NA | 5.3 | 498.6| 66 | 7 | 1.7| 3.5| 8  | 7.5 |
| 2020-01-01 12:00:00 | 3  | 3.5| 2.7| 6.1| 0 | 16 | 19 | 5.9| 5.7| 2  | NA | 7.3 | 342.1| 76 | 3 | 2.0| 4.0| 5  | 7.2 |
| 2020-01-01 13:00:00 | 4  | 3.2| 2.6| 6.4| 1 | 11 | 15 | 4.1| 6.7| 5  | NA | 5.8 | 195.1| 70 | 3 | 2.5| 3.8| 5  | 7.9 |
| 2020-01-01 14:00:00 | 2  | 2.6| 1.5| 5.5| 0 | 13 | 23 | 2.6| 8.1| 5  | NA | 5.2 | 142.9| 55 | 8 | 2.3| 3.3| 6  | 8.0 |
| 2020-01-01 15:00:00 | 1  | 2.0| 0.5| 5.6| 0 | 9  | 13 | 2.6| 5.5| 1  | NA | 2.8 | 134.9| 54 | 4 | 2.5| 3.3| 7  | 3.9 |
Advantages of Meta/Data “Universal” Structure

Simple & Understandable

Maximally Compact

Multiple monitors in a single file

Sufficient for both Maps and Time Series

CSV file format is well understood

Simple web server can serve static files
What about low-level, engineering data?

**Quality Control Parameters**

- Flow Rate (lpm)
- Air Temp. (°C)
- Internal R.H. (%)
- PM2.5 (mg/m³)
Data model for low-level, engineering data

Assume interest in a single monitor

‘Meta’ table is the same (but only has one row)

‘Data’ table has one column per engineering parameter

Similar advantages:

- Simple, understandable data structure
- Maximally Compact
- CSV file format is well understood
- Simple web server can serve static files
Data Access

Jon’s favorite data access – download static files

- Easy
- Fast
- All the data at once
- No programming required
- No authentication required

Jon’s favorite time series format – CSV

- XML 😞😞
- JSON 😐😐
- CSV 😃😃 🎉🎉 🎊🎊
http://data-monitoring_v2-c1.airfire.org/monitoring-v2/

- airnow
  - airnow-latency
  - airsis
  - daily
  - epa-aqs
  - known-locations
  - latest
  - s3-logs
  - wrcc

- latest/data
  - airnow_PM2.5_latest_data.csv
  - airnow_PM2.5_latest_data.csv.gz
  - airnow_PM2.5_latest_meta.csv
  - airnow_PM2.5_latest_meta.csv.gz
  - airnow_PM2.5_nowcast_latest_data.csv
  - airnow_PM2.5_nowcast_latest_data.csv.gz
  - airnow_PM2.5_nowcast_latest_meta.csv
  - airnow_PM2.5_nowcast_latest_meta.csv.gz
  - airsis_PM2.5_latest_data.csv
  - airsis_PM2.5_latest_data.csv.gz
  - airsis_PM2.5_latest_meta.csv
  - airsis_PM2.5_latest_meta.csv.gz
  - wrcc_PM2.5_latest_data.csv
  - wrcc_PM2.5_latest_data.csv.gz
  - wrcc_PM2.5_latest_meta.csv
  - wrcc_PM2.5_latest_meta.csv.gz
Reading in ‘csv.gz’ data

R

```r
meta <- readr::read_csv("meta.csv.gz")
data <- readr::read_csv("data.csv.gz")
```

Python

```python
meta = pandas.read_csv("meta.csv.gz")
data = pandas.read_csv("data.csv.gz")
```
Thanks for listening!