Air Sensor Data Analysis

Webinar 1:

Data Wrangling: Accessing data, data formats, and quality control

This guide is intended to be a resource and does not necessarily reflect U.S. EPA policies

Webinar Logistics

Automated closed captioning

"Switch to Phone Audio" if your computer's microphone/speakers are not working

Submit questions for Presenters, Zoom Technical Assistance

- **Closed captioning** is available by clicking the "CC" button in your control panel
- This webinar is being recorded.
- **To ask a question:** Type your question in the Q&A Box.
- **Technical difficulties:** If you are having technical difficulties, please send a message through the Q&A Box

Overview of Webinar Series

Webinar #1:

Data Wrangling: Accessing Data, Data Formats, and Quality Control

Webinar #2:

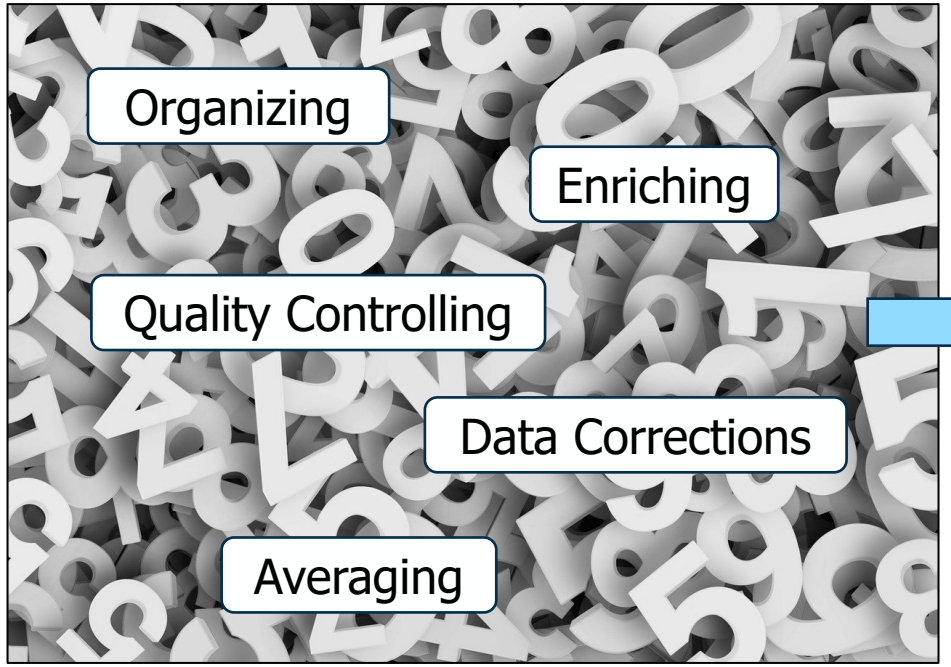
Gaining Insights from your Data: Visualization and Interpretation

Webinar 1 Agenda

1. What is Data Analysis?
2. Accessing & Organizing Your Data
3. Enriching Your Data
4. Quality Controlling & Validating Data
5. Publishing Data
6. Recap
7. Next Webinar

What is data analysis?

1. Data Wrangling



2. Data Insights

Distilling large amounts of information into meaningful visuals and summaries

Image Sources: Pixabay

Why do we analyze data?

- Understand your own data - insights and meaning
- Create helpful visuals and summaries
- Communicate findings to different audiences
- Spread awareness and gather interest on important issues
- Help make informed decisions and support action

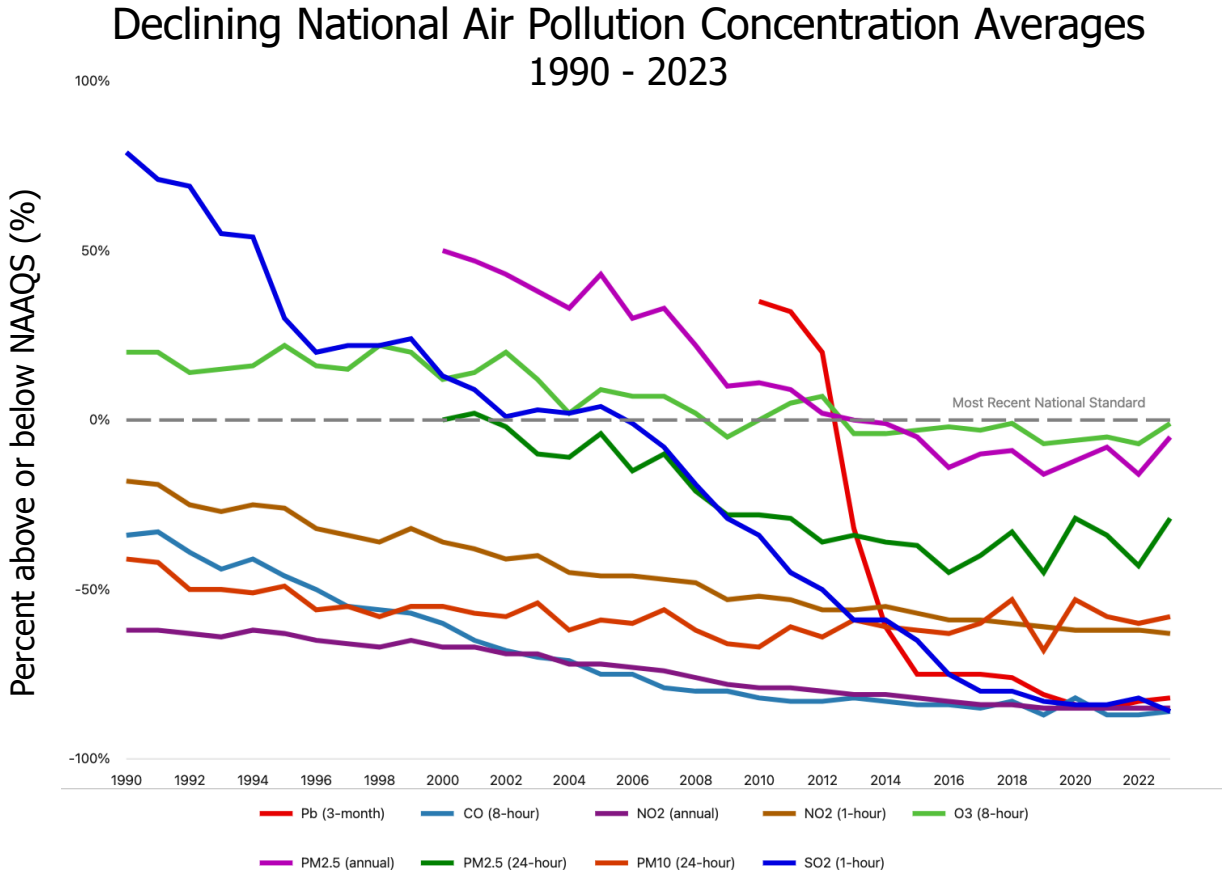
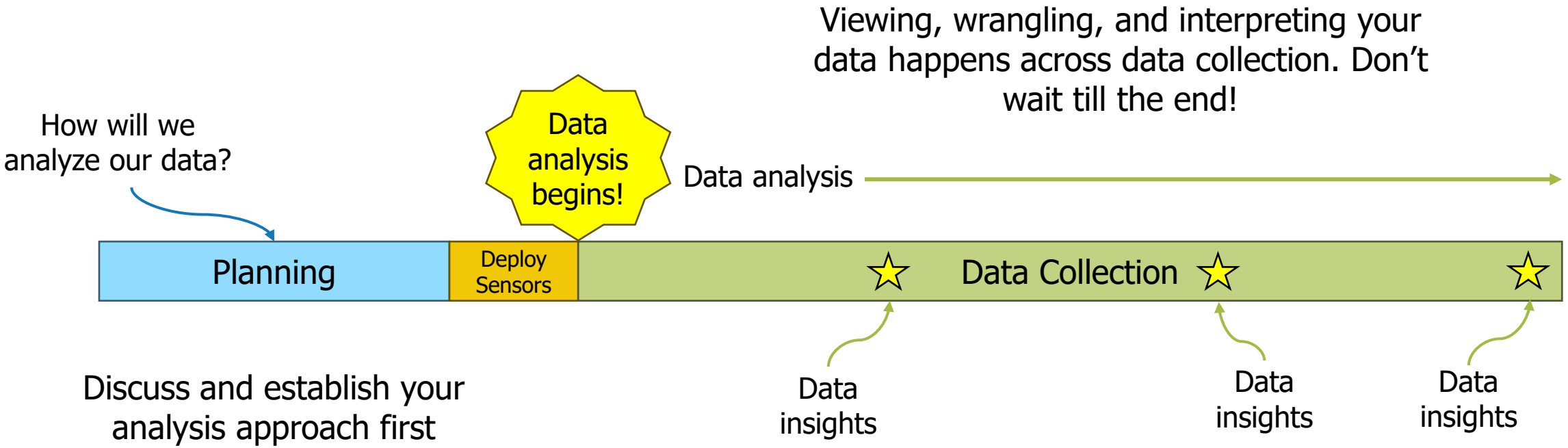


Image Source: EPA

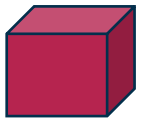
When do we do data analysis?



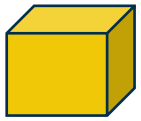
Critical Elements of Data Analysis

Building blocks of data analysis

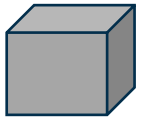
Display & Understand Your Data



Interpreting (What does it show? Not show?)

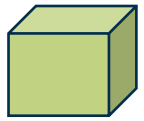


Visualizing (How to show data & information?)

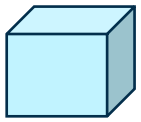


Creating Metrics (How to average data? How to aggregate data?)

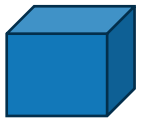
Organizing Your Data



Supporting Data (What else is needed? Who else needs to be involved?)



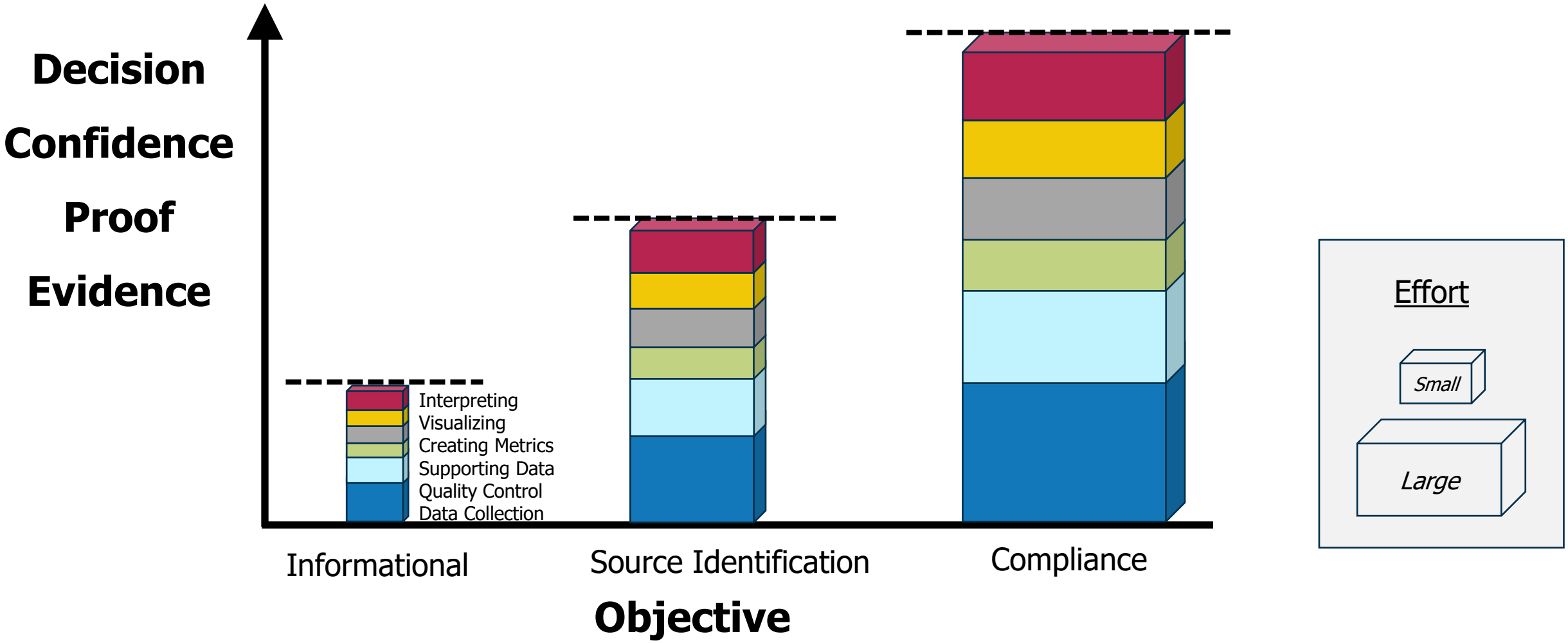
Quality Control (How robust? How to operate & maintain? How are data validated?)



Data Collection (How long? How many sites?)

Critical Elements of Data Analysis

Building blocks of data analysis



Accessing & Organizing Your Data

Accessing and Organizing Your Data

1. Access your data
   a. Variety of sources (e.g., API, download, data dashboard, physical SD card)
   b. May be from more than one source
2. Organize your data
   a. Many different formats and structures (e.g., file type, labeling, contents)
   b. The format you get your data in may not be the easiest to work with!
   c. Explore your data and reformat to meet your needs

	A	B	C	D	E	F
1		Heritage Elementary School	Heritage Elementary School	Lomax Junior High School	Lomax Junior High School	Deer Park South High School
2	Date/Time	O3/NO2 CairClip Sensor Data (ppb)	NO2 CairClip Sensor Data (ppb)	O3/NO2 CairClip Sensor Data (ppb)	NO2 CairClip Sensor Data (ppb)	O3/NO2 CairClip Sensor Data (ppb)
3	9/4/13 0:00	4.02	No data	2.67	0.27	1.72
4	9/4/13 1:00	4.27	No data	2.82	0.07	1.82
5	9/4/13 2:00	3.97	No data	2.63	0.07	1.65
6	9/4/13 3:00	4.20	No data	2.77	0.00	1.97
7	9/4/13 4:00	3.93	No data	2.77	0.07	1.87
8	9/4/13 5:00	4.07	No data	2.93	0.00	2.13
9	9/4/13 6:00	4.17	No data	2.53	0.27	2.10
10	9/4/13 7:00	4.57	No data	3.08	0.33	3.28
11	9/4/13 8:00	12.90	No data	11.60	1.53	13.57
12	9/4/13 9:00	8.37	No data	8.97	14.53	14.00
13	9/4/13 10:00	7.63	No data	7.98	5.43	20.57
14	9/4/13 11:00	5.43	No data	6.37	4.08	21.87
15	9/4/13 12:00	2.93	No data	2.33	8.03	36.80
16	9/4/13 13:00	3.13	No data	1.42	10.28	37.50
17	9/4/13 14:00	3.20	No data	1.20	6.65	36.20
18	9/4/13 15:00	3.17	No data	2.05	9.88	43.83
19	9/4/13 16:00	3.10	No data	2.22	5.80	42.80
20	9/4/13 17:00	3.22	No data			
21	9/4/13 18:00	3.40	No data			
22	9/4/13 19:00	3.30	No data			
23	9/4/13 20:00	3.52	No data			
24	9/4/13 21:00	3.48	No data			
25	9/4/13 22:00	3.40	No data			

Data Source: Duvall, R. et al. 2016

```
PID,Mo,Dy,Yr,Hr,Mi,VOC,lat,lon,spd,ME,jday
11,11,27,2018,12,3,1242,-999,-999,0.57,2,331.5021
11,11,27,2018,12,4,1169,-999,-999,0.57,2,331.5028
11,11,27,2018,12,5,1126,-999,-999,0.61,2,331.5035
11,11,27,2018,12,6,1082,-999,-999,0.39,2,331.5042
11,11,27,2018,12,7,1038,-999,-999,0.79,2,331.5049
11,11,27,2018,12,8,1006,-999,-999,0.83,2,331.5056
11,11,27,2018,12,9,972,-999,-999,0.47,2,331.5063
11,11,27,2018,12,10,925,-999,-999,0.88,2,331.5069
11,11,27,2018,12,11,904,-999,-999,0.61,2,331.5076
11,11,27,2018,12,12,891,-999,-999,0.82,2,331.5083
11,11,27,2018,12,13,943,-999,-999,0.51,2,331.509
11,11,27,2018,12,14,992,-999,-999,0.9,2,331.5097
11,11,27,2018,12,15,971,-999,-999,0.78,2,331.5104
11,11,27,2018,12,16,952,-999,-999,0.82,2,331.5111
11,11,27,2018,12,17,943,-999,-999,0.96,2,331.5118
11,11,27,2018,12,18,836,-999,-999,0.86,2,331.5125
11,11,27,2018,12,19,823,-999,-999,0.99,2,331.5132
11,11,27,2018,12,20,860,-999,-999,0.8,2,331.5139
11,11,27,2018,12,21,846,-999,-999,0.75,2,331.5146
11,11,27,2018,12,22,846,-999,-999,0.94,2,331.5153
11,11,27,2018,12,23,817,-999,-999,0.79,2,331.516
11,11,27,2018,12,24,708,-999,-999,0.96,2,331.5167
11,11,27,2018,12,25,679,-999,-999,0.74,2,331.5174
11,11,27,2018,12,26,688,-999,-999,0.7,2,331.5181
```

Data Source: Breen, M. et al. 2021

Organizing your Data - Tools

Familiarize yourself with how your data is structured and reorganize it to support your work!

Tools that help you view, edit, and reorganize your data

- Your data management or air sensor vendor!
- Excel or Google Sheets
- Program (e.g., Python, R) – complex, requires expertise

Organizing Your Data – Tips (1/3)

1. Data file type (.xlsx, .csv, .txt, etc.)

a. What file types will be easier for you to work with for analysis?
   i. .xlsx and .csv are most common and can be used with Excel and Google Sheets
b. Try to keep data all in the same file type, makes merging easier later on

2. Contents

a. Sensors output a lot of information – exclude data that isn't relevant to your analysis (e.g. Particle Bin Count, Fan Speed) **Never fully delete data**
b. Remove duplicate columns (e.g., Datetime Start and Datetime End often the same)
c. Remove blank columns (datasets may have placeholders for certain sensor configurations)

This helps reduce the file size, making your data easier to work with!

Organizing Your Data – Tips (2/3)

3. Labeling

a. What is easiest to understand and reference during analysis?
b. Use unique names (e.g., if you have two different PM2.5 data feeds, distinguish between them)
c. Simpler is better – "PM2.5" instead of "PM_2_5_(ug/m3)"
d. You can keep unit information separate in metadata (more in next slide!)
e. Include time zone in your date column label! (datetime PST, datetime LST)

4. Units

a. Keep units standard so you can easily compare data to health thresholds
b. Keep units uniform across datasets
c. Store unit information related to pollutants and other parameters in metadata (more in next slide!)

5. Significant Figures

a. Round data depending on the parameter (e.g., 2 for $PM_{2.5}$, at least 4 for coordinates) to help with interpretation

Organizing Your Data – Tips (3/3)

6. Date and time records

a. Include all time information available from your sensor/device (i.e., don't remove minutes or seconds)
b. Use a 0-23 clock - working with AM/PM can be tricky (e.g., 6 = 6AM, 18 = 6PM)

7. Time zones

a. Convert timestamps to your local project time zone (if needed)
b. Keep your data in standard time – daylight savings will create duplicate timestamp issues

8. File merging

a. Merge datasets to support more streamlined analysis (where appropriate) (e.g., one dataset for network of 20 sensors instead of 20 separate files)
b. Make sure all data features (e.g., time zone, units) are uniform before merging

Adding Metadata

Metadata is any data that helps describe the air quality data you are collecting, such as…

- Site Information
  - Site name
  - Address
  - Coordinates
  - Surroundings
  - Observations
- Device Information
  - Monitor manufacturer and model
  - Monitor ID (e.g., serial number)
  - Parameter data unit (e.g., ug/m^3, ppm)

Site_ID	Date_Created	Site_Name	Model	Hardware	Firmware_Ve	Latitude	Longitude	Elevation	County
1742	5/29/17 15:16	Beach at 48th Avenue	PA-II	2.0+1M+PMSX003-B+PMSX003-A	6.06b	37.75827	-122.5082	29	San Francisco
1872	7/10/17 14:23	Somerset Hills	PA-II	2.0+1M+BME280+PMSX003-B+PMSX003-A	6.06b	38.10359	-122.1887	455	Solano
1874	7/10/17 14:23	Glen Cove Ridge	PA-II	2.0+1M+BME280+PMSX003-B+PMSX003-A	6.06b	38.066784	-122.22	207	Solano
1882	7/10/17 14:26	Navone St.	PA-II	2.0+1M+BME280+PMSX003-B+PMSX003-A	6.06b	38.077988	-122.2305	178	Solano
2031	7/11/17 13:29	St Mary's Park	PA-II	2.0+1M+BME280+PMSX003-B+PMSX003-A	6.06b	37.733208	-122.4236	149	San Francisco
2063	7/12/17 10:44	Kings Mountain SW	PA-II	2.0+1M+BME280+PMSX003-B+PMSX003-A	6.06b	37.39356	-122.3472	1311	San Mateo

Data Source: Bay Area Air Sensor Dataset Metadata

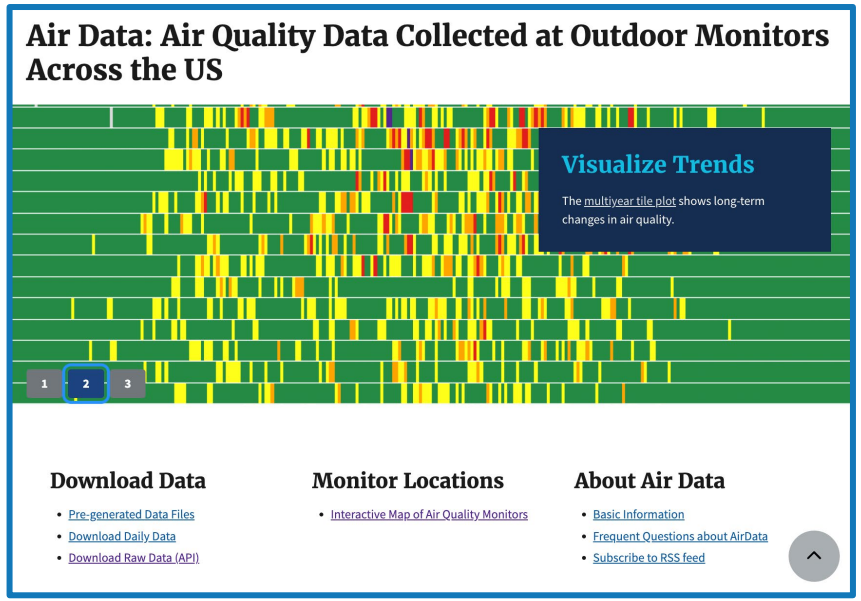
Field	Description	Units
Date and Time	Eastern Standard Date and Time	
Sensor1_PM_1HR_Env	1-hour Averaged $PM_{2.5}$ Concentration from Sensor 1. Data when RH>90% has been excluded.	µg/m^3
Sensor2_PM_1HR_Env	1-hour Averaged $PM_{2.5}$ Concentration from Sensor 2. Data when RH>90% has been excluded.	µg/m^3
Day	Days since the start of the field deployment	

Data source: Clements, A. et al. 2019

Metadata provides supporting information that helps organize and analyze your data

Enriching Your Data

Enriching your dataset



Why enrich your data? Allows you to…

- Validate the quality of your data (e.g., collocation)
- Confirm trends you see during data review
- Support analysis and data insights

What additional data would be helpful for your project?

- Regulatory data – compare data to other monitoring in the area
  - Resources: EPA's Air Quality System (AQS), AirData, state/local air agency website
  - Fields: Datetime, pollutant concentration, metadata
- Meteorology – investigate how weather conditions impact air quality
  - Resources: Local weather stations, local/state air agency, MesoWest, OpenWeather API, NCEI Past Weather
  - Fields: Wind speed, wind direction, temperature, relative humidity, pressure
- Emissions source information – provide context to your findings
  - Resources: (e.g., traffic, emissions inventory, permitted facilities)
  - Fields: Traffic count, location of relevant facilities

Image source: EPA AirData

Break

Quality Control and Data Validation

QA/QC Activities*

Quality Assurance (QA) - the management tasks that ensure you are collecting quality data.

- Planning
- Documenting
- Reporting

Quality Control (QC) - technical activities that measure and check the quality of your data, ensuring you can reliably meet project data quality goals.

- Collocating sensors
- Routine maintenance
- Frequently reviewing data
- Validating data – determining the quality of your data

- Ensure the data you collect (and ultimately analyze) is accurate!
- Build familiarity with your data
- Allow yourself (and others) to build trust in your data

*You may often hear QA/QC used together.

Data Review

Goal: Identify data issues and determine whether your data is valid, suspect, or invalid data. Ultimately, you will remove invalid data from analysis through validation.

Valid
Data is real and accurate

- Trends are expected or can be explained

Suspect
Data is atypical, needs further investigation

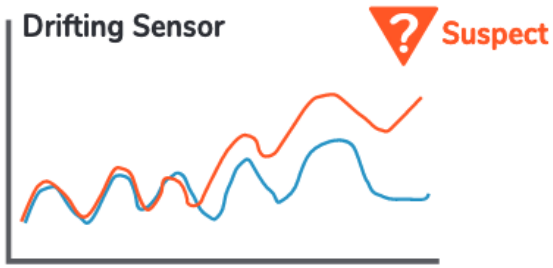
- Source biases unrelated to project objective (e.g., smoking near monitor raised levels, but wildfire smoke is interest)
- Patterns in the data that are difficult to explain

Invalid
Data is not real nor accurate

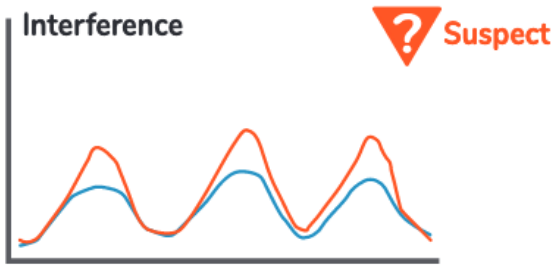
- Outside of the operational range
- Sensor malfunction
- Maintenance activities

There are many data review methods and tools – we'll focus on basic time series review.

Basic Data Review
Common data issues



Bias
Gradual positive or negative change in a sensor's response over time (e.g., aging of the sensor component)

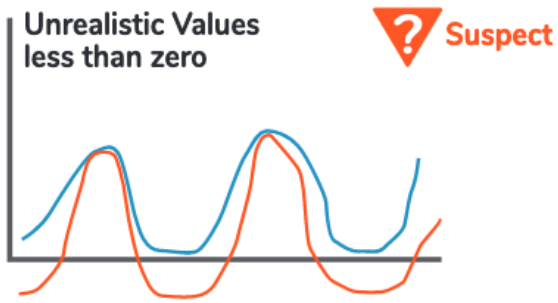


Interferences
Other pollutants that interfere with the measurement of the target pollutant

Note: Time series shown here are just one way to review data. Other forms (data summaries, AQI, and maps) can also be helpful for QC.

Unexpected downtime
Affects completeness of data

Missing parameter data
Some sensors may report more than one parameter important to check



Unexpected problems
e.g., large negative values, high values caused by sensor failure and outliers

Reference Instrument
Air Sensor

Image source: Bay Air Center

Data Review Steps

What next?

1. Review your data on a daily/weekly basis.
2. Investigate suspect data with team members – is it invalid?
    a. Can the levels be explained by an actual source, activity, or weather conditions?
    b. If the data is real (i.e., caused by a source), is the data relevant to your project?
    c. Compare data to nearby regulatory sites or other sensors
3. Remove invalid data from analysis. Tips:
    a. Keep a version of your data that is un-QC'ed!
    b. Don't remove or convert small negative data to zeros – this biases data.
    c. Check whether vendor has already tagged invalid data (e.g., -999)
4. Keep track of your QC procedure – sharing this with others builds confidence in your data

Data Corrections

Data corrections are adjustments made to sensor data that account for sensor error, uncertainty, or resolve bias.

- Developed by the sensor manufacturer, local collocations, or extensive studies
- Allows you to analyze sensor data across different data types
- Supports data acceptance and confidence in data quality

Make sure you're working with corrected data!

- Keep a version of your data before correction

More on sensor collocations: Air Sensor Toolbox - Air Sensor Collocation

Example: PurpleAir Wildfire Smoke Correction

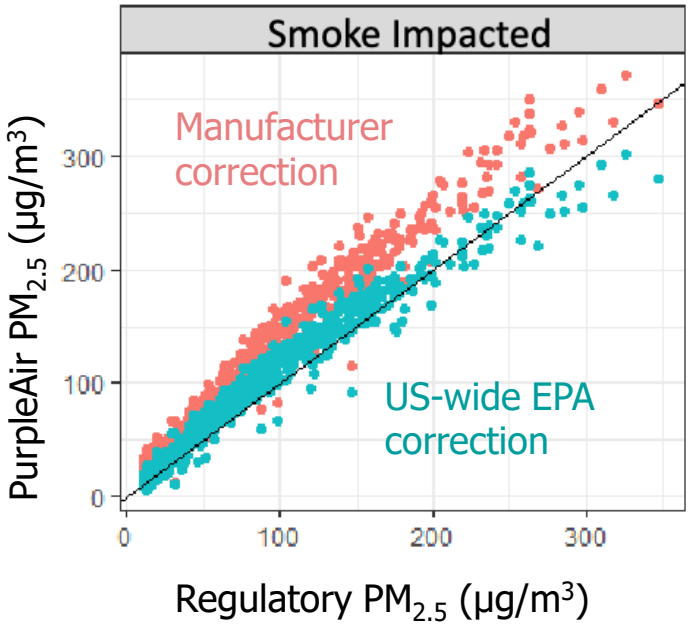


Image source: Barkjohn et al. 2021

Averaging

Allows you to evaluate the "big picture" signals in your data, supporting data review!

Typical averaging methods:

- Hourly
- 8-hour
- Daily (24 hours)
- Annual

Averaging best practices:

- Quality control data before averaging!
- Use begin-time averaging for air quality data
  e.g., 10:00-10:59 → 10:00 hourly average
- Make sure averages are representative and complete
  ≥75% of the data preferred for average

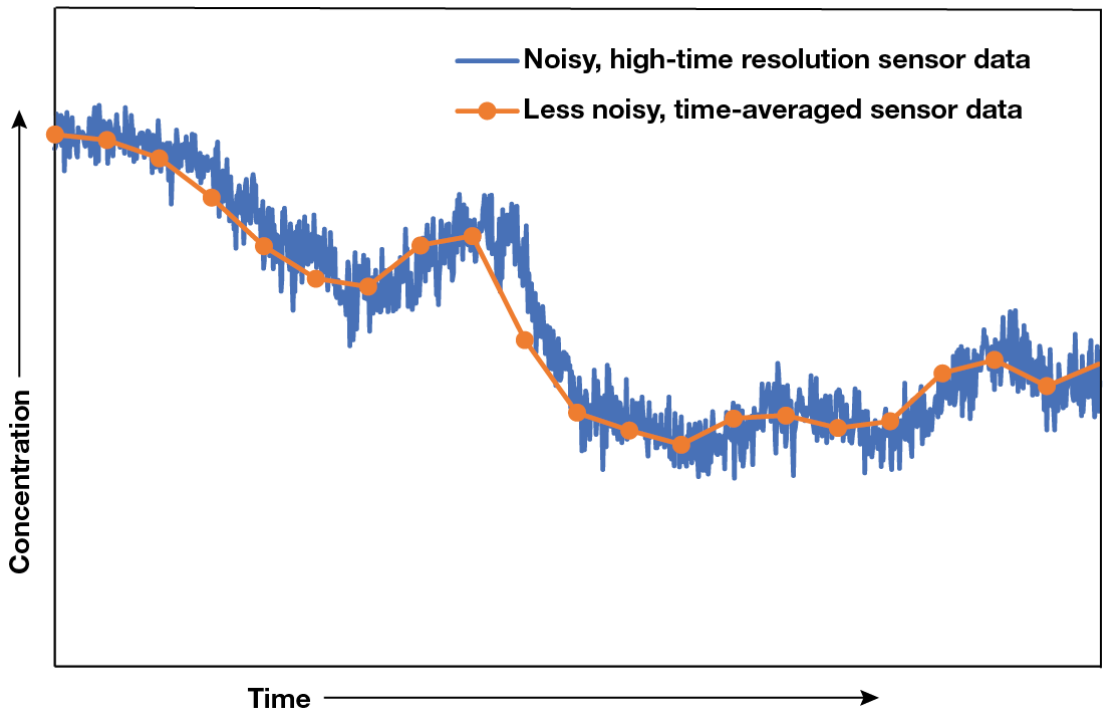


Image source: EPA

Averaging also prepares your data for interpretation

Allows you to compare your data to other standards:

- Air Quality Index (AQI)
- National Ambient Air Quality Standards (NAAQS)*

*Sensor data cannot be used to determine NAAQS compliance

QC Tools on your Sensor Dashboard

- View time series to diagnose issue and understand trends
- Compare datasets (e.g., across different pollutant channels or sites)
- Set up automated alerts (e.g., offline sensor, data out of range)
- Log issues to keep track
- Average and review your data

Choose your data dashboard based on available review tools!

Publishing Data

Why Should I Publish My Data?

- Share what you've found!
- Greater transparency in your findings
- Allow others to benefit from your data
- Create opportunities to collaborate with other governments, groups, and researchers
- It might be required by your grant

*Note: This section is discussing publishing **data**, not publishing analysis or results (i.e., a peer reviewed article).*

What Data do I Publish?

Real-time	Unvalidated Historical	Validated Historical
• Data that is delivered to a system immediately after it was collected. Often not quality controlled or validated.	• Data from past periods that has not been quality controlled or validated.	• Data from past periods that has been quality controlled or validated.

Who Do I Share My Data With?

Completely Private

- Real-time only accessible by select project team members
- Unvalidated historical only accessible by select project team members
- Validated historical only accessible by select project team members
- Data platform is password protected
- Data is not shared outside of this approved group

At Project Team Discretion

- Real-time only accessible by select project team members
- Unvalidated historical data can be requested by residents and external organizations
- Validated historical data can be requested by residents and external organizations
- Project representatives approve/deny data requests

Real-time Public, Historical Requested

- Real-time data is publicly available on a website with appropriate caveats about its data quality
- Unvalidated historical data is available upon request
- Validated historical data is available upon request
- Project representatives approve/deny data requests

All Data Publicly Available

- Real-time data is publicly available on a website
- Unvalidated historical data is downloadable from a public website or API
- Validated historical data is downloadable from a public website or API

Other Data Sharing Considerations

- If you choose to make data available upon request, a protocol will need to be established for requesting the data. This would require considering who would receive and approve the request and how (email address, website form, etc.).
- If you have a specific audience in mind, like a regulatory agency, make sure you discuss what their data sharing requirements are. Some agencies may only be able to use data that is shared publicly.
- You may also want to make some **data use guidelines**, that make all potential users of the data aware of the conditions, type, and intent of the data. These could be things like noting that all data is unvalidated and should be considered preliminary or who noting credit should be given to if the data is used by outside parties.

Where Can I Share My Data?

Public-facing website

- Instrument manufacturer
- Develop your own
- Use another organization's

Cloud drive

Google drive, Microsoft SharePoint, Box folder

Email

- Users can request via email
- Share compressed files

Application Program Interface (API)

Data exchanges between two applications over the internet; external application requests data from API; the API retrieves data from DMS and makes it available to that application

What
Who
Where

of data publishing

Conclusion

There's a lot of work *before* interpretation!

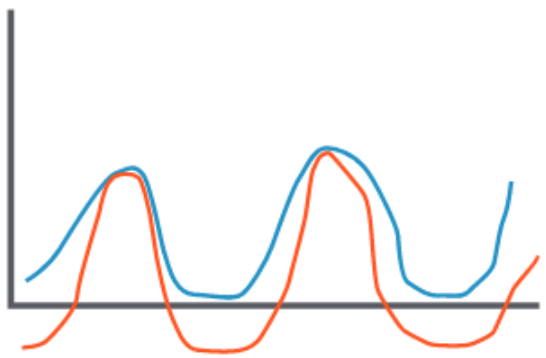
Don't skip these steps!

If you don't Data Wrangle:

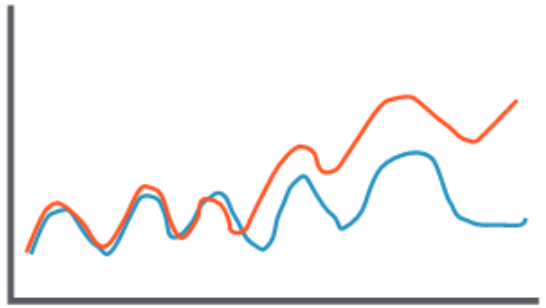
- May be drawing conclusions from your data that are not valid
- Can't back up your insights
- Data has lower credibility

Interactive End-of-Session QUIZ

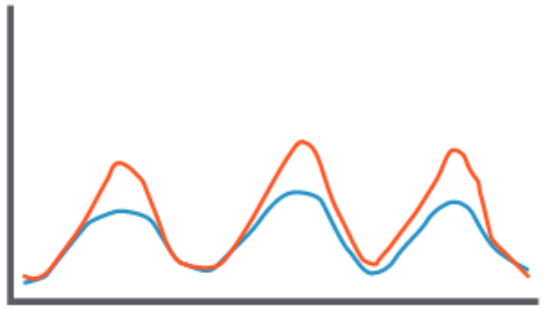
Which graph is showing sensor bias increasing over time?



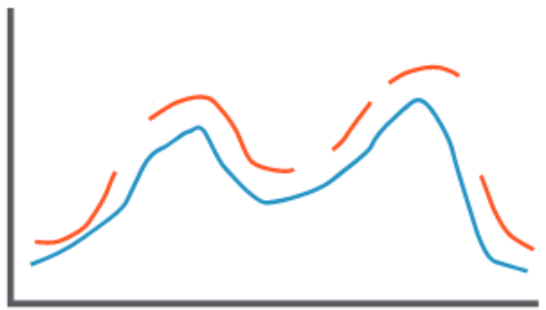
A



B



C



D

Reference Instrument
Air Sensor

Image source: BAC

You're analyzing 1-minute data from your air sensor. Which data timestamp is most helpful to use?

A

2025-06-16 19

B

2025-06-16 19:01:33

C

6/16/25 7PM

D

Thursday 2025-06-16 19:01:33.22 PDT

Image source: BAC

You've been collecting 1-minute $PM_{2.5}$ data for one month. You're interested in analyzing your data. What should your first step be?

A

Average your data

B

Quality control your data

C

Share your data with local residents

D

Draw insights from your data

Image source: BAC

Webinar 1 Recap

What is Data Analysis? Why do we do data analysis? When do we do data analysis?

Accessing your data

Tips on organizing your data

Enriching your data

Quality controlling data

Validating data

Publishing data

Interactive End-of-Session Feedback

What's the most useful thing you learned today?

What’s up next

Webinar 2

Gaining Insights from your Data: Visualization and Interpretation

Matching monitoring objectives and analysis, visualizing data, interpreting visualizations, drawing conclusions, analysis tools

Q & A