Big Data Fundamentals and Applications

Data Preprocessing – Numerical Analysis II

Asst. Prof. Chan, Chun-Hsiang

Master program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Applied Artificial Intelligence, , Chung Yuan Christian University, Taoyuan, Taiwan

Outlines

1. Data Science Mindset
2. Visualization
3. Part IV Matrix computation (Numpy)
   Part V Table computation (Pandas)
   Part VI Visualization (Matplotlib & Seaborn & Bokeh)
   Part VII Statistics (Scipy)
4. Part X Seismic Risk Map

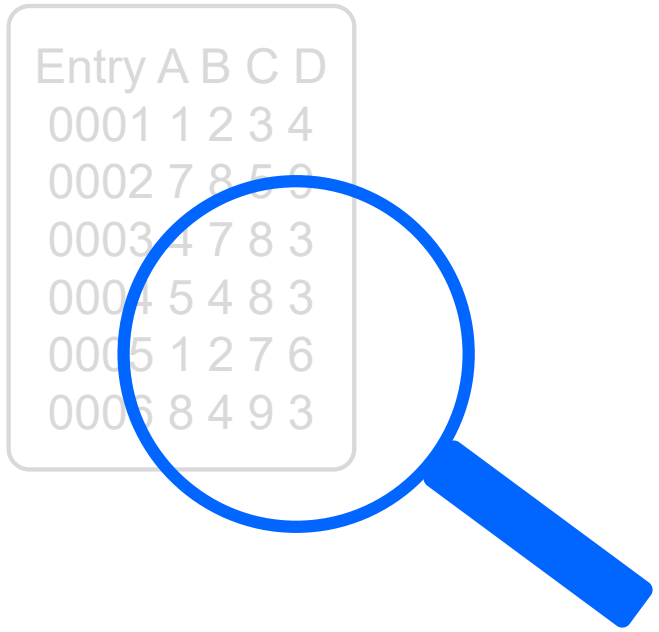
Data Science Mindset

- When we obtain a data science project, what will you do?
- Here is a simple situation, given a large numerical dataset… If you can ask all questions about the dataset, and then what kinds of questions you want to ask?
  - …
  - …
  - …
  - …
  - …

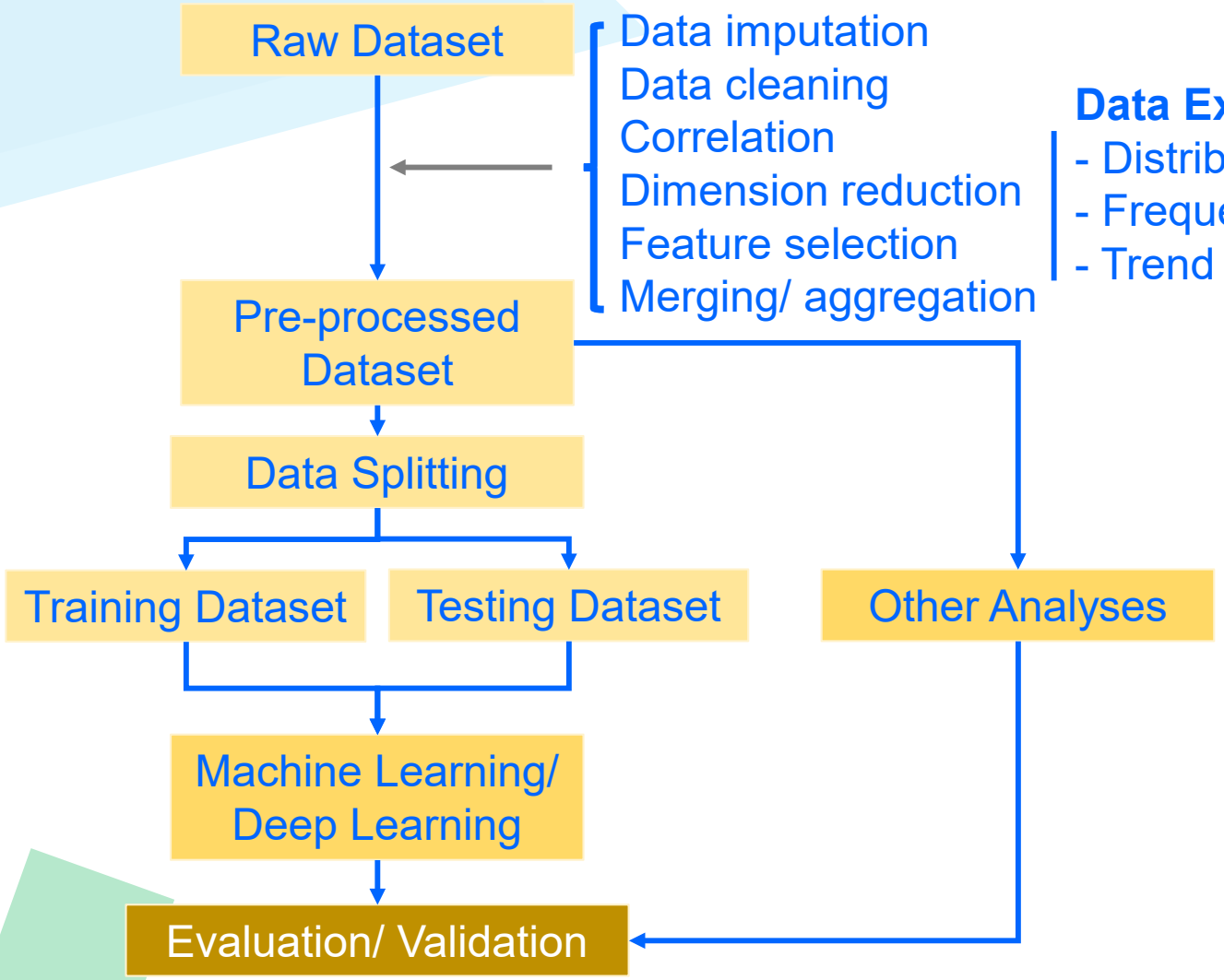
Question 1 List all questions and give a reason.

Data Science Mindset

- Basically, we need to overview and pre-check the dataset.
  - How many features?
  - Null values?
  - Data type of each column
  - Resolution?
  - Sampling rate?
  - …
  - …



Data Science Mindset



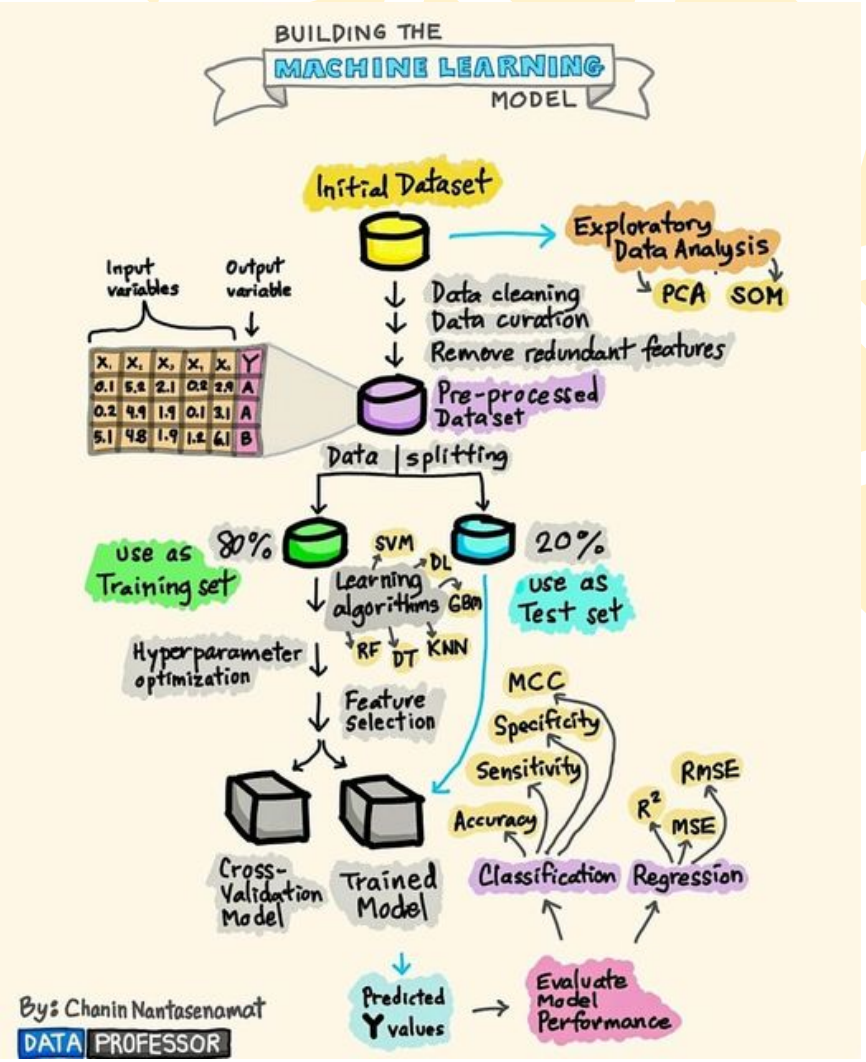


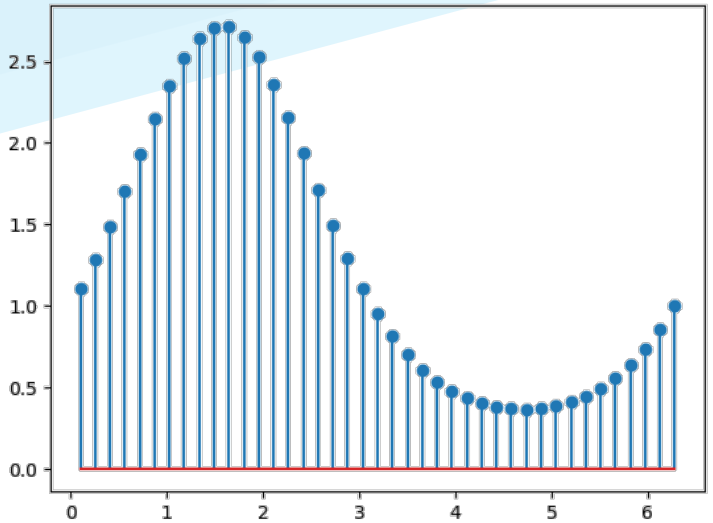
Figure source: https://bit.ly/3BJd2d4

Visualization – Methods

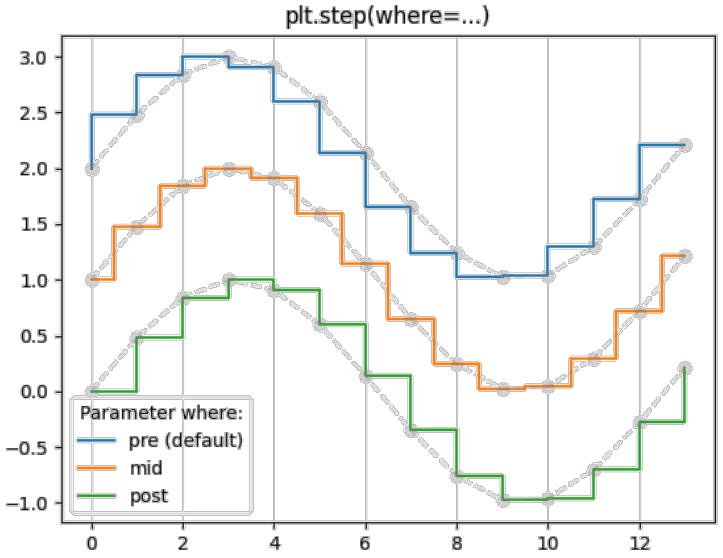
- Several visualization functions were developed in built-in package, such as line, scatter, boxplot, and histogram.

Distribution			
Continuous Data	**Discrete Data**	**Ordered/ Categorical Data**	**Proportional Data**
Scatter Bubble Histogram Violin Plot Box Plot Heatmap Density Map	Scatter Bubble Histogram Violin Plot Box Plot Heatmap Density Map	Group/ Stacked Bar Pie	Group/ Stacked Bar Pie

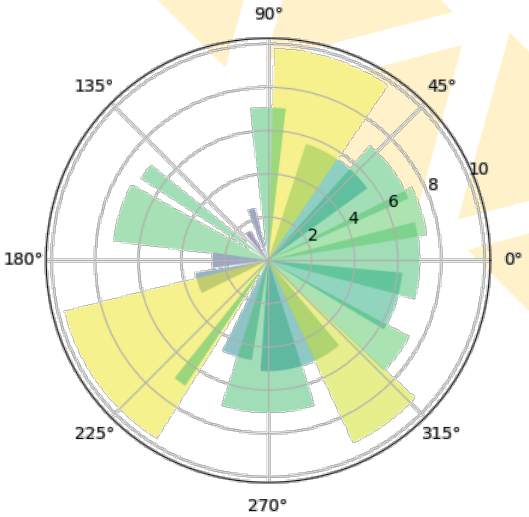
Visualization – Methods



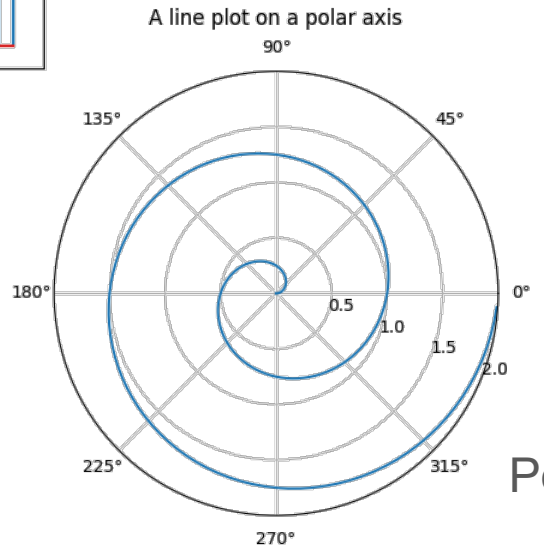
Stem plot



Step plot



Polar plot



Polar plot

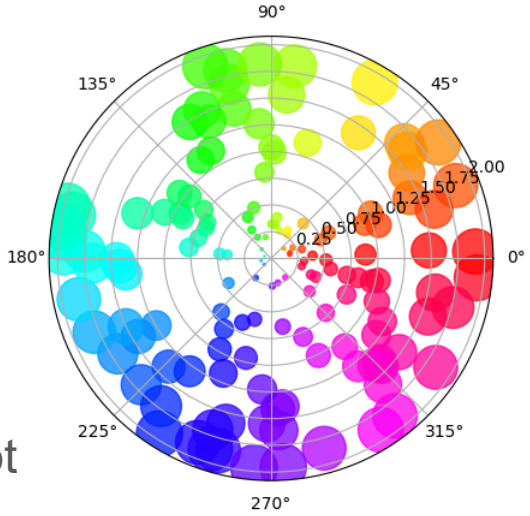
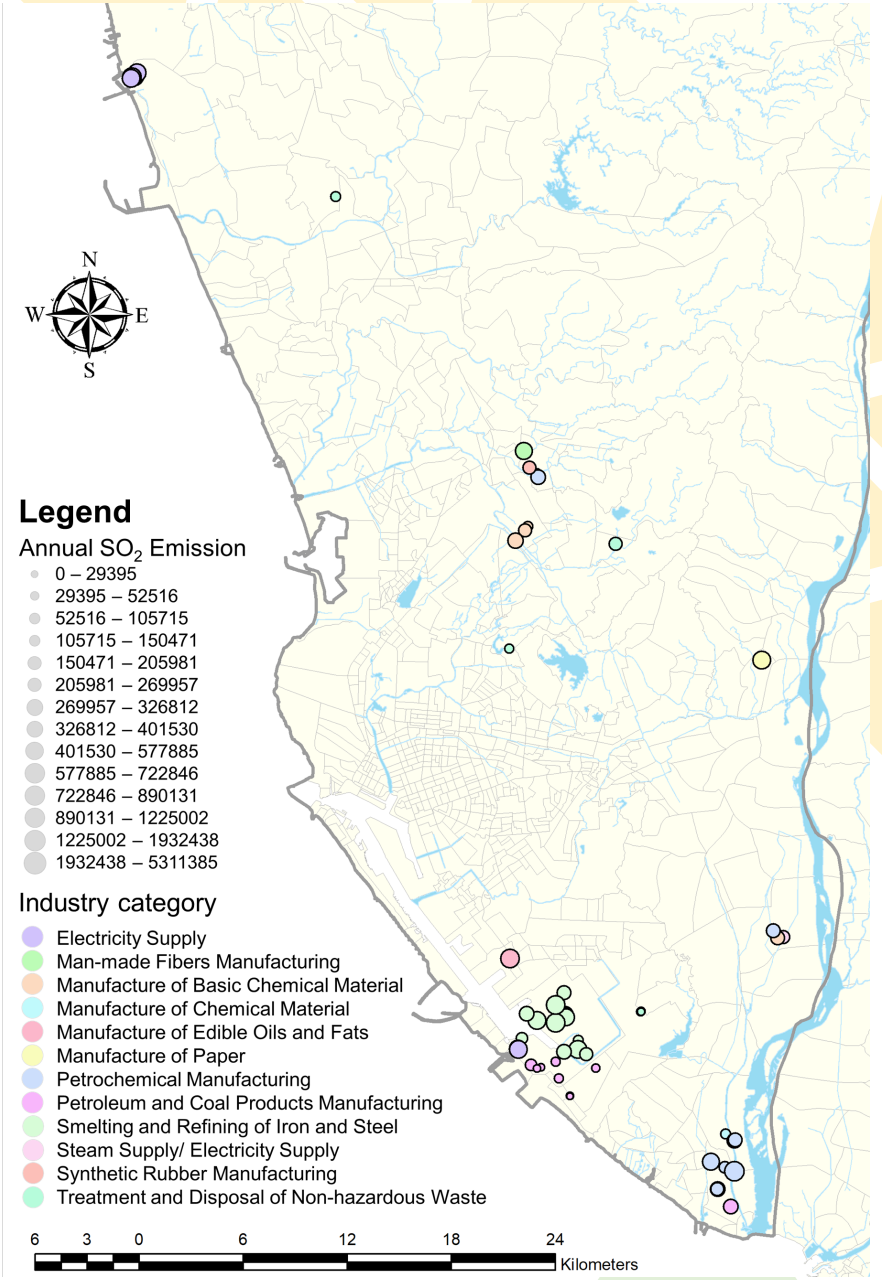


Figure source: https://matplotlib.org/stable/gallery/index

Visualization – Element

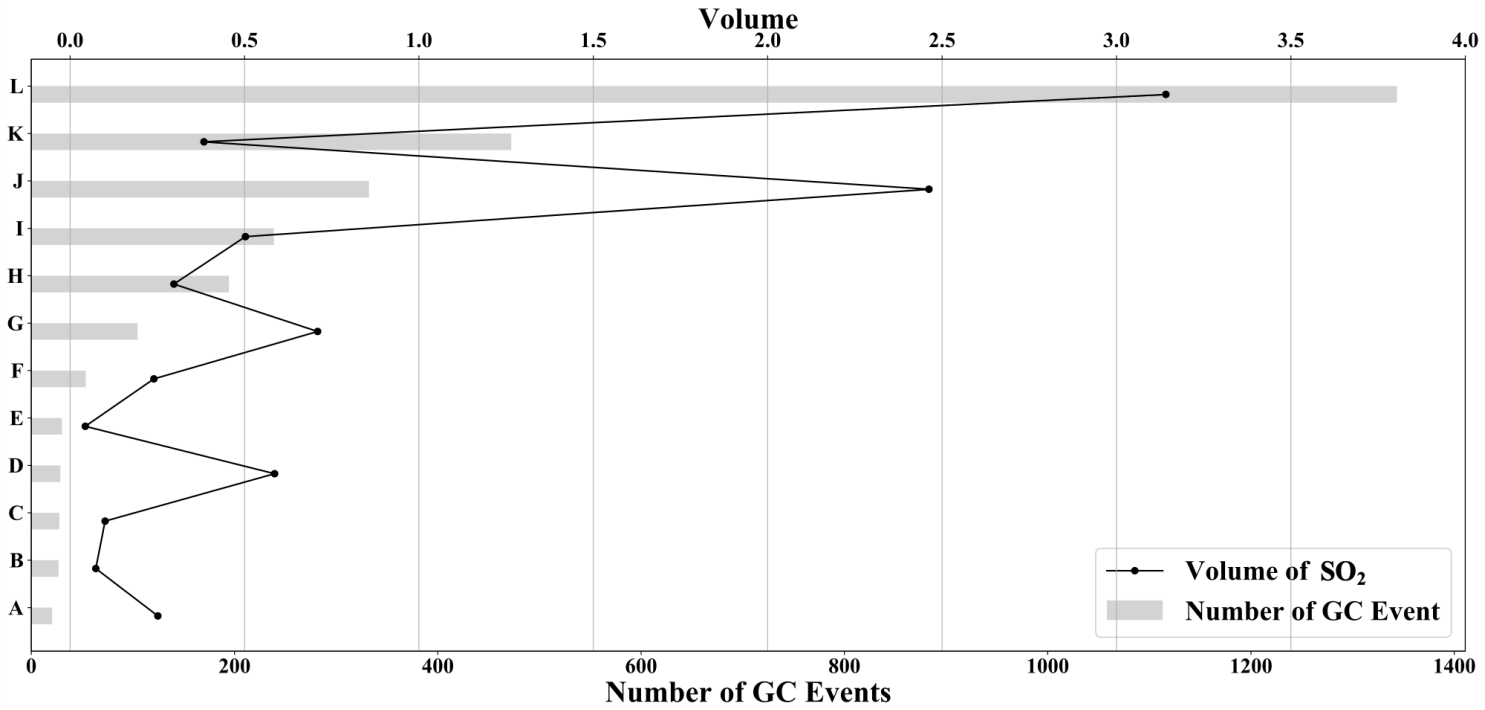
- Important elements in the figure.
  1. Title
  2. X/ Y tick label
  3. X/ Y label
  4. Legend (size, color, symbol)
  5. Grid (optional)
  6. Error bar (optional)
  7. Confidence interval (optional)
  8. Colormap (optional)
  9. Compass icon(optional)
  10. Scale bar(optional)

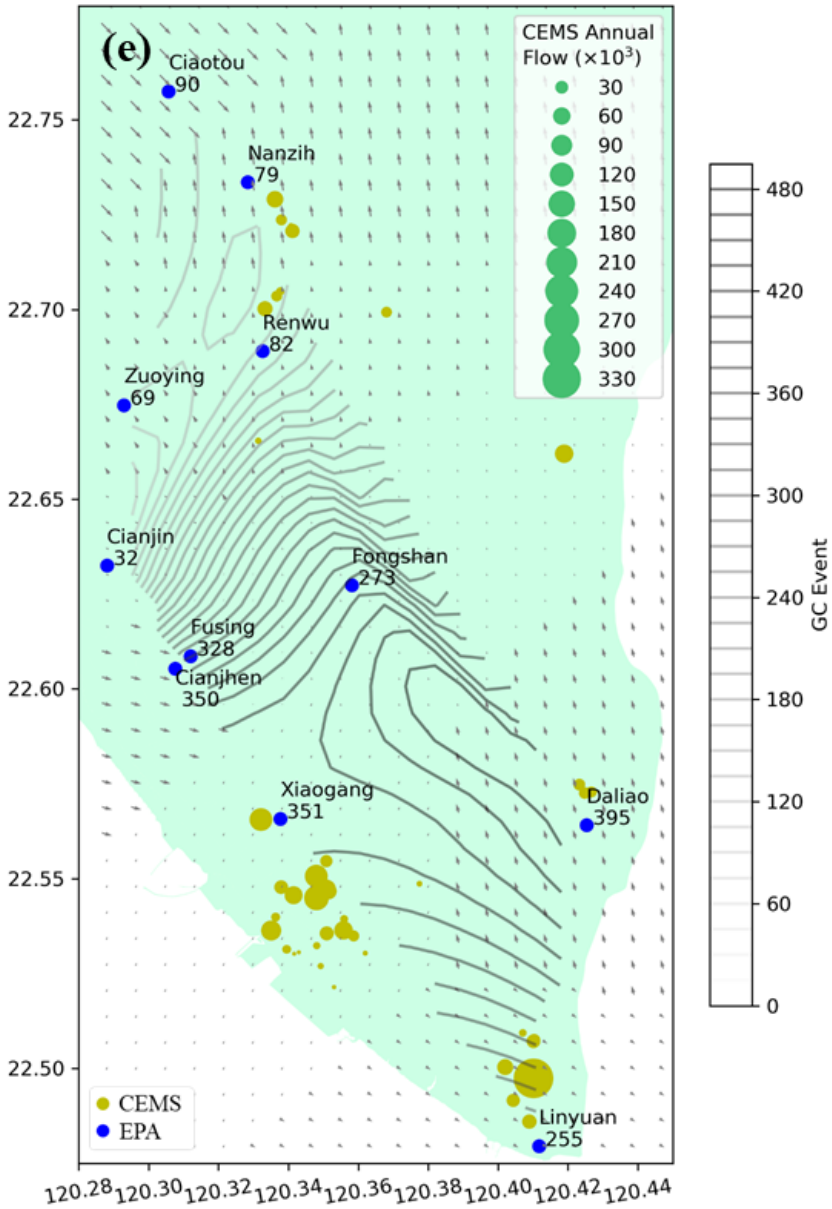


Figures are from Chan et al. (2022). A novel evaluation of air pollution impact from stationary emission sources to ambient air quality via time-series Granger causality. Earth Data Analytics for Planetary Health. Springer.

Visualization – Elements

More examples



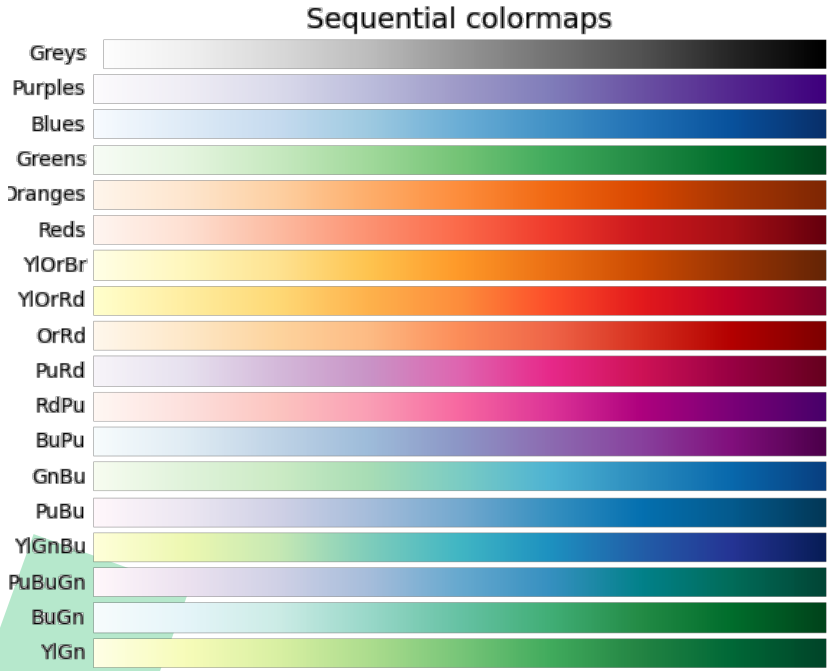


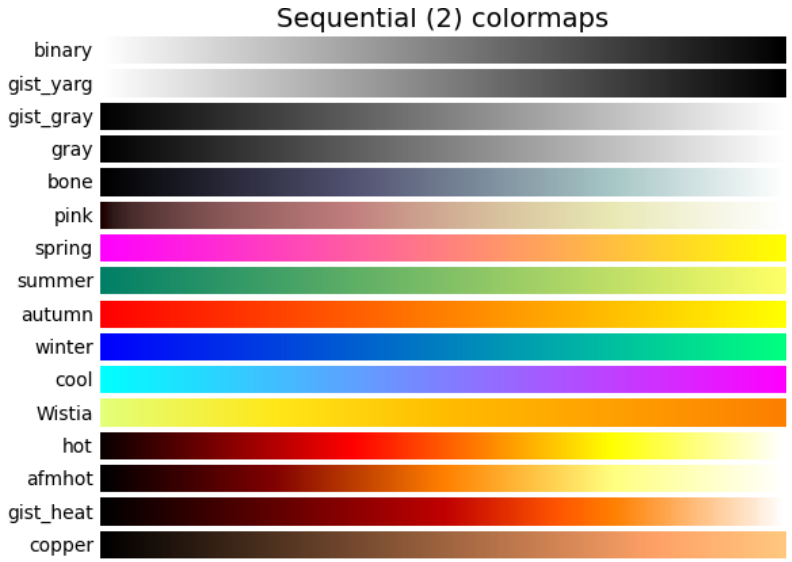
Figures are from Chan et al. (2022). A novel evaluation of air pollution impact from stationary emission sources to ambient air quality via time-series Granger causality. Earth Data Analytics for Planetary Health. Springer.

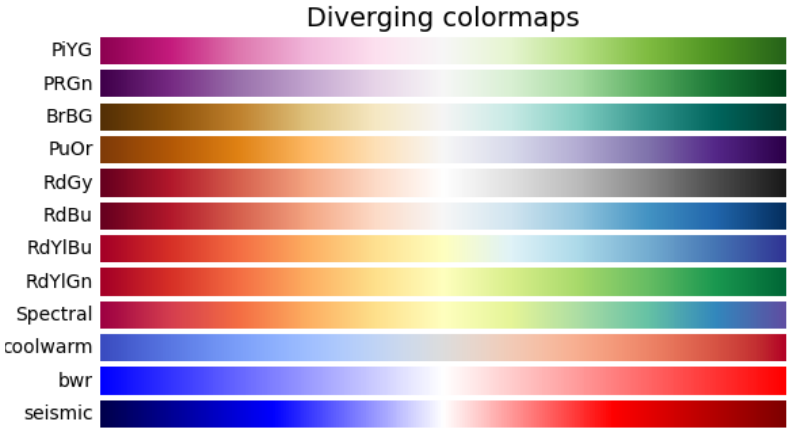
Visualization – Color

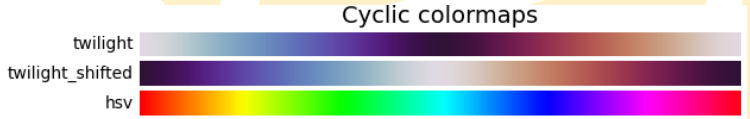
Figure source: https://matplotlib.org/stable/gallery/color/colormap_reference.html#sphx-glr-gallery-color-colormap-reference-py

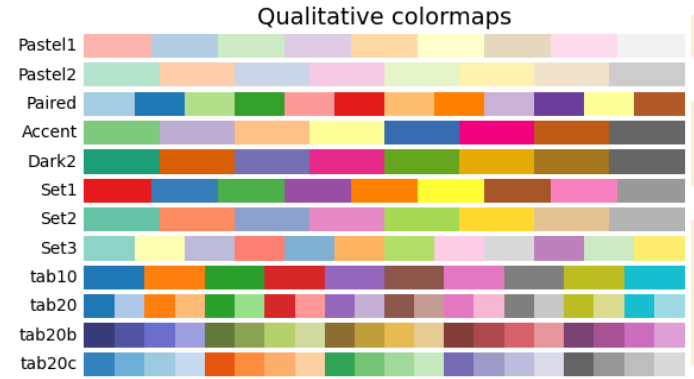
- Colormap selection

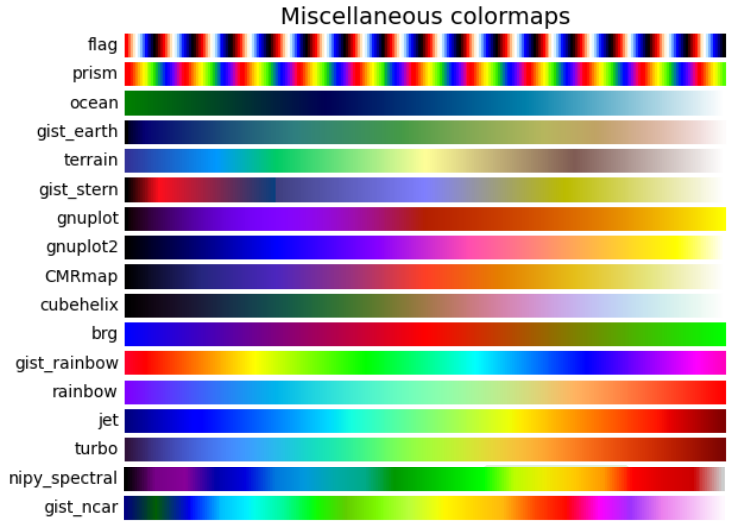












Visualization – Color

- Colormap selection
  - Do not pick more than 8 colors from graduated colormap
  - Graduated colormaps are for continuous values
  - Discrete colormaps are for categorical values
  - … (**think about it!**)

- To better the understanding of figures…
  - Use different size or symbol to represent different data
  - Plot different data into the same subplot/ figure
  - Use subplot with fixed x and y ranges

Question 2 Visualization

- We can visualize our dataset in different ways, such as 2D, 3D or animation. As a reader, please tell us your thoughts on 2D, 3D and animation in terms of practicality and applicability respectively.

Question Time

If you have any questions, please do not hesitate to ask me.

The End

Thank you for your attention))