Model Research/Development

# A) Adjustable parameters

* Before selecting an algorithm it is important to understand exactly what different variations in data it will need to accommodate, aka what user can select

1. Location- can they select only 1 sensor or many?
   1. *Team discussion*- eval each sensor separately
2. Time/date range- is there a limit to how long/short? Different data (daily vs hourly averages) used for different time scales? E.g. if using daily average weeks/years months? If using hourly, day(s)/week(s)?
   1. Also are we looking for local outliers or global outliers? Aka if a timespan is selected, are we looking for outliers just within that period (+ some time before/after to level set?) or in terms of all 43 years?
      1. If global or if we provide just a few options to toggle between rather than something more open-ended, we might be able to pre-compute everything and just update viz based on user selection
   2. *Team discussion*- time range = 1st selection
      1. Then see on map see which sensors had anomalies in that period
      2. Compute anomaly based on range of dates selected (not all 43 years)
      3. If limit number of options user can select in terms of date range, may be able to precompute- this may or may not be feasible (because causes tool to be less useful)- goal => realistic date range
      4. Maybe only do most recent 20 years instead of all 43 years
         1. Limit maximal date range if run into compute issues
3. Types of air quality/pollutant/temperature data: univariate or multivariate? Can the user select which variables they care about? If more than 1 variable is selected are the variables evaluated separately (univariate) or combined together (multivariate)
   1. *Team discussion*- focus on 1 variable, rest are not used in modeling but shown in visualization as descriptive
      1. Default AQI- but could also could select 1 on others (e.g. PM10 or PM2.5 or 1 of toxins)
      2. NO multivariate anomaly detection

# B) Dimension Reduction (HDDA techniques) - Marco

* ***Goal*** = to reduce the size of data we are storing. We will run our anomaly detection algorithms on subsets of this data in real-time as users interact with the visualization.
* Is this understanding correct?
* Is there any concern that dimension reduction might reduce or eliminate outliers?
* *Team discussion*- need for hourly data!
  + Hourly data- B-splines to reduce 24 daily observations to maybe 6

## Spines

* 1. B-splines- computationally efficient but have issues at boundaries
  2. Cubic splines- require fitting # of knots
  3. Smoothing splines- require fitting lambda parameter (degree of smoothing)
  4. *Resources*:
     1. HHDA book= Elements of Statistical Learning, chapter 5
     2. From proposal- Anomaly detection PM10 in Malaysia: <https://www.sciencedirect.com/science/article/pii/S1309104215302440?via%3Dihub>

## FPCA

* 1. Often used for sparse/missing signals, can also be used for dimension reduction
  2. *HDDA paper*: <https://doi.org/10.1198/016214504000001745>
  3. *Team discussion*- Maybe for PM2.5 (because has highest number of measurements)

## Tensor Decomposition

* 1. CP or Tucker- may perform better than splines or FPCA (especially Tucker)
  2. *HDDA Resources*:
     1. Survey article: <https://www.jstor.org/stable/25662308>
     2. Tensors with image (has section specifically on dimension reduction with Cp and with Tucker): <https://doi.org/10.1080/00401706.2019.1527727>
     3. Point cloud: <https://doi.org/10.1080/00401706.2018.1529628>

## Compressive Sensing

* 1. Option if signal are sparse (potentially in frequency if not in time)
  2. *HDDA:* Intro article <https://ieeexplore.ieee.org/document/4472240>

## Robust PCA

* 1. Can potentially find both low rank approximation and (sparse) outliers in same step
  2. *HDDA article*: <https://proceedings.neurips.cc/paper/2009/file/c45147dee729311ef5b5c3003946c48f-Paper.pdf>

# C) Outlier detection

## CUSUM

* 1. Standard visualization = control chart
  2. Univariate
  3. *Team discussion*: worth trying as first approach
  4. **Issues**: Initial attempt not great at detecting anomalies, see [plots](https://docs.google.com/presentation/d/14znmNesc48c85mMtkOfSEPdpoxEAAz0DCC3XCDF2DI0/edit?usp=sharing)
     1. Seasonality
     2. CUSUM requires previously calculated value to calculate the next value
        1. Use a for loop, probably a more elegant way to code this?
        2. Still only O(n) so probably not prohibitive time-wise especially if we use a faster library (possibly polars)?
     3. *Slack discussion:* Possible fix might be time-varying threshold, something based off the rolling mean and standard deviations

## ARIMA

* 1. Again is univariate
  2. German bank more complicated version: <https://www.bundesbank.de/resource/blob/763892/f5cd282cc57e55aca1eb0d521d3aa0da/mL/2018-10-17-dkp-41-data.pdf#:~:text=In%20their%20implementation%20in%20various,a%20higher%20than%20monthly%20frequency>
  3. *Slack discussion*: because of seasonality issue in data, ARIMA possible solution
     1. Relevant Python packages:
        1. <https://pypi.org/project/pmdarima/>
        2. <https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.ARIMA.html#pmdarima.arima.ARIMA>
        3. <https://www.statsmodels.org/dev/generated/statsmodels.tsa.seasonal.seasonal_decompose.html>
     2. Frequency and period parameter:
        1. <https://robjhyndman.com/hyndsight/seasonal-periods/>
           1. This blog post discusses R functions, but the tables could possibly be helpful to figure out what integer or number to pass into the frequency and/or period parameter(s)

## Mahalanobis distance

* 1. Used in Malaysian study that also uses splines
  2. Pro and con = is just simple stat calc
  3. *Team discussion*: 2ndary possibility after CUSUM and DBSCAN

## Isolation Forest

* 1. Builds trees and measures number of partitions needed to isolate instance, anomalies expected to require fewer partitions
  2. *Article from Proposal*: <https://www.e3s-conferences.org/articles/e3sconf/pdf/2023/73/e3sconf_iced2023_10005.pdf>
     1. Also used
        1. **kernel density estimation (KDE)** to understand distribution of different variables- could be useful for us too
        2. **Box plots**- to id outliers - another method we might consider or use to check our results
           1. Assuming data follows a normal distribution we could also do another statistical test like Generalized ESD (like the Grubbs Test, but Grubbs is for only 1 outlier): <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm>

## Sliding window anomaly detection algorithm (SWAD)

* 1. Univerariate detection using combo of IQR and differences
  2. *Article from proposal*: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9867200/>

## Long short-term memory (LSTM)

* 1. Requires anomaly free training data
  2. *Same article as SWAD*: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9867200/>
  3. Combo LSTM with autoencoder (AE): paper from proposal: <https://ieeexplore.ieee.org/abstract/document/10011213>
     1. This paper has a good round-up of a lot of variations on ML anomaly detection
  4. Stacked LSTM (multiple LSTMs): <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7582627/>
  5. LSTM in Industry:
     1. <https://publications.aston.ac.uk/id/eprint/40857/1/08896029.pdf>,
     2. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9195000>
  6. A different AE approach (no LSTM): <https://ieeexplore.ieee.org/abstract/document/9877800>
     1. If we get Purple Air data this article has a network spacial correlation approach
     2. Also has good round up of papers for stat vs clustering vs ML air quality anomaly detection
     3. More papers with similar roundups (but that use custom methodolgies I think it would be hard for us to implement/replicate):
        1. <https://ieeexplore.ieee.org/abstract/document/8081731>
        2. <https://arxiv.org/pdf/2103.12910.pdf>

## SVM

* 1. <https://ieeexplore.ieee.org/abstract/document/7857445>

## Clustering

* 1. AE + k-means (allow multivariate model): <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8400863/>
  2. Modified k-means (this article was univariate only): <https://ieeexplore.ieee.org/abstract/document/9288694>
  3. **DBSCAN** - mentioned in lectures 😀
     1. Allows multi-variate + is non-parametric, some parameter tuning
     2. Papers:
        1. <https://dl.acm.org/doi/pdf/10.1145/2733381>
        2. <https://cdn.aaai.org/KDD/1996/KDD96-037.pdf>
        3. <https://dl.acm.org/doi/pdf/10.1145/3068335>
     3. *Team discussion*: Possibly pick a clustering method if CUSUM not sufficient (maybe DBSCAN)
     4. *Exploratory attempt*: Better out of box than CUSUM, see [plots](https://docs.google.com/presentation/d/14znmNesc48c85mMtkOfSEPdpoxEAAz0DCC3XCDF2DI0/edit?usp=sharing), still not account for seasonality but may adjust for that (add factor variables, etc.)

## Other assorted:

* 1. Bayesian: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6210001/>
  2. Survey article on challenges of outlier detection on high-dimensional data: <https://onlinelibrary.wiley.com/doi/epdf/10.1002/sam.11161?saml_referrer>

# D) Validation

* In addition to validating with case studies as we previously discussed, other options include:

1. Create a synthetic dataset and use that to evaluate model performance, LSTM paper
2. Compare 2 methods, as in Isolation forest paper (compare with Box plots)