Machine Learning for the ARM Climate Research Facility

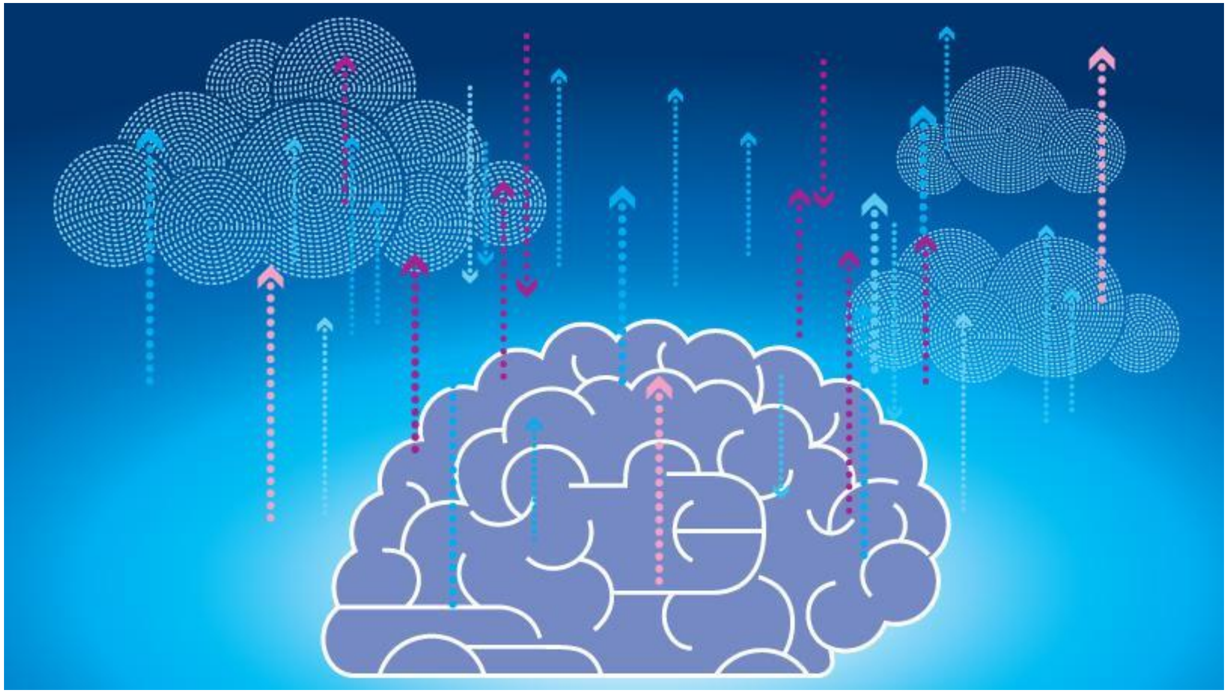
JEFFERY T. MITCHELL

Brookhaven National Laboratory

2017 ARM/ASR PI Meeting, Vienna, VA

Outline

- CIMEL Sun Photometer Instrument Anomaly Detection
  - Random Forest
- MFRSR Instrument Anomaly Detection
  - Multivariate Regression
- AOS Local Source Emission Identification
  - Neural Network
  - Support Vector Machine



CIMEL Sun Photometer Instrument Anomaly Detection

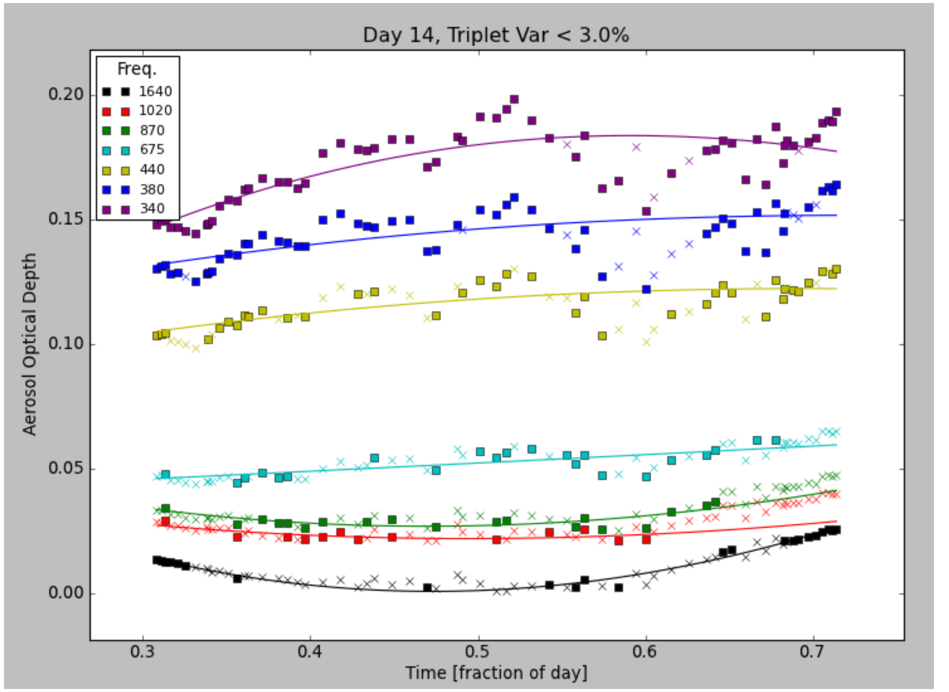
- The CIMEL Sun Photometer is a multi-channel automatic sun and sky scanning radiometer that measures the direct solar irradiance and sky radiance at the Earth's surface.
- The sampling rate is typically 10 minutes.
- Measurements are taken only in daylight hours without precipitation. The measurements are sensitive to cloud conditions.
- THE PROJECT: Apply machine learning algorithms to detect anomalies due to instrument failure modes with a fast, automated application. Failure modes include obstructions and filter degradation.

CIMEL Sun Photometer Feature Extraction

ARM
CLIMATE RESEARCH FACILITY

- A machine learning model considers inputs from multiple data features simultaneously.
- There are large variations due to weather conditions in a given day, so features are extracted on a daily basis.
- Example features include the coefficients of daily fits (A_0, A_1, A_2) of the aerosol optical depth (AOD) measurements for each filter to the curve:

$$AOD(t) = A_0 + A_1 t + A_2 \cos(\Theta(t))$$

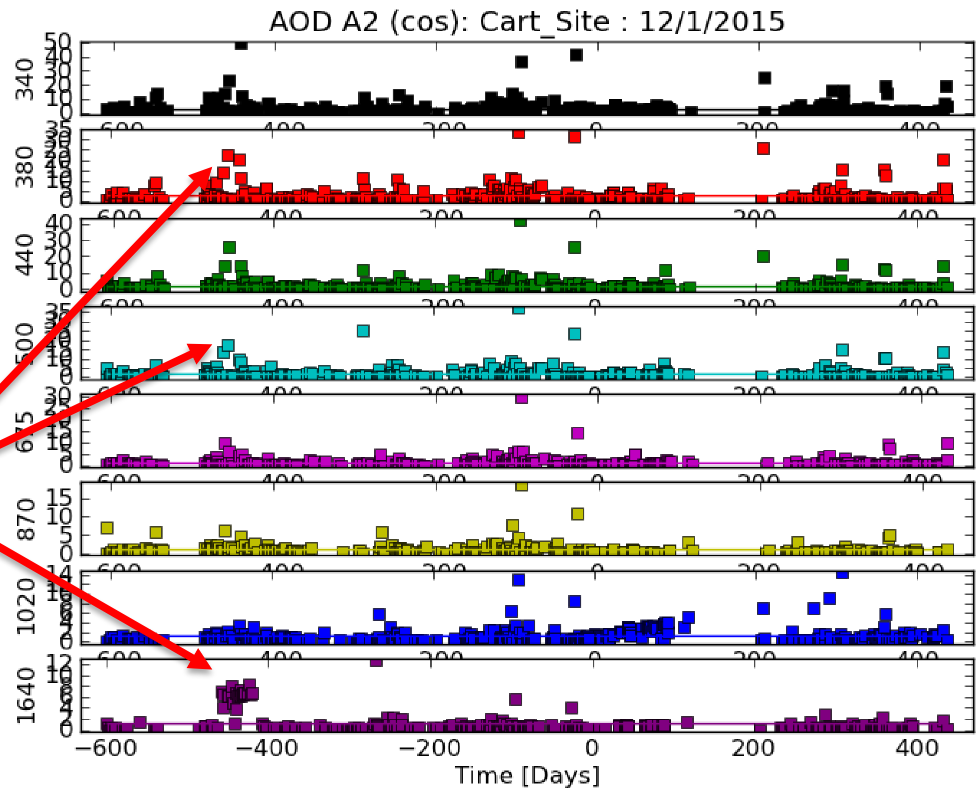


An example of fits to the data for one day. Points marked by an "x" are influenced by clouds and not included in the fit.

The Cosine Coefficient Features for the CIMEL Sun Photometer

ARM
CLIMATE RESEARCH FACILITY

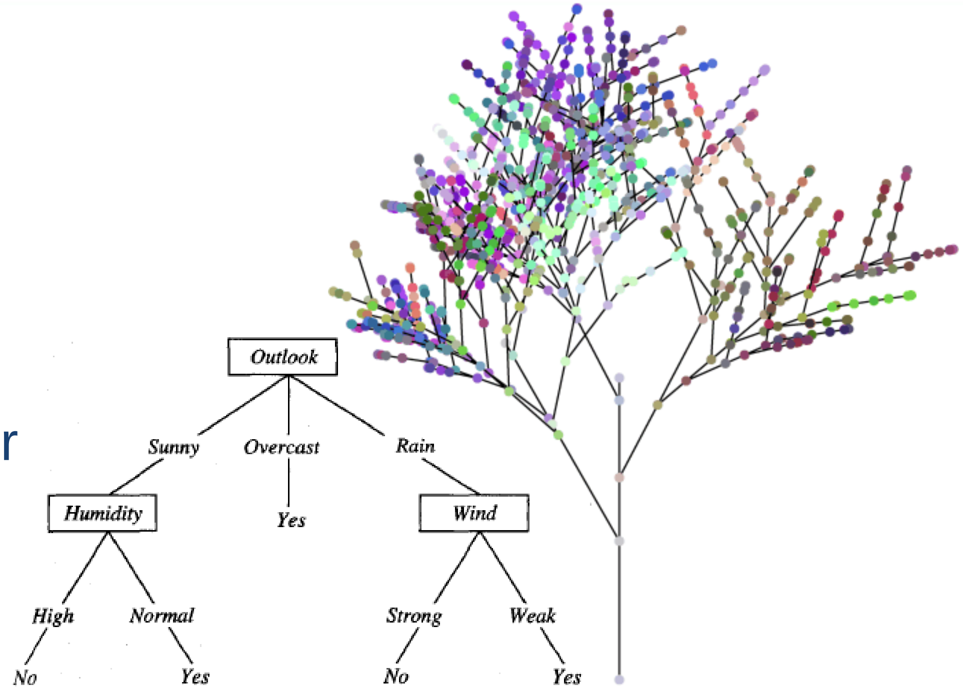
- Almost 3 years of data are processed from the SGP site in Oklahoma (4/1/14 – 2/12/17).
- Shown here is the cosine coefficient for each day from the instrument's 8 filters.
- The arrows point to days where a spider web was obstructing the measurements.
- Correlating multiple features can be more sensitive to problems then considering single features.



Anomaly Detection Using a Random Forest Regressor

ARM CLIMATE RESEARCH FACILITY

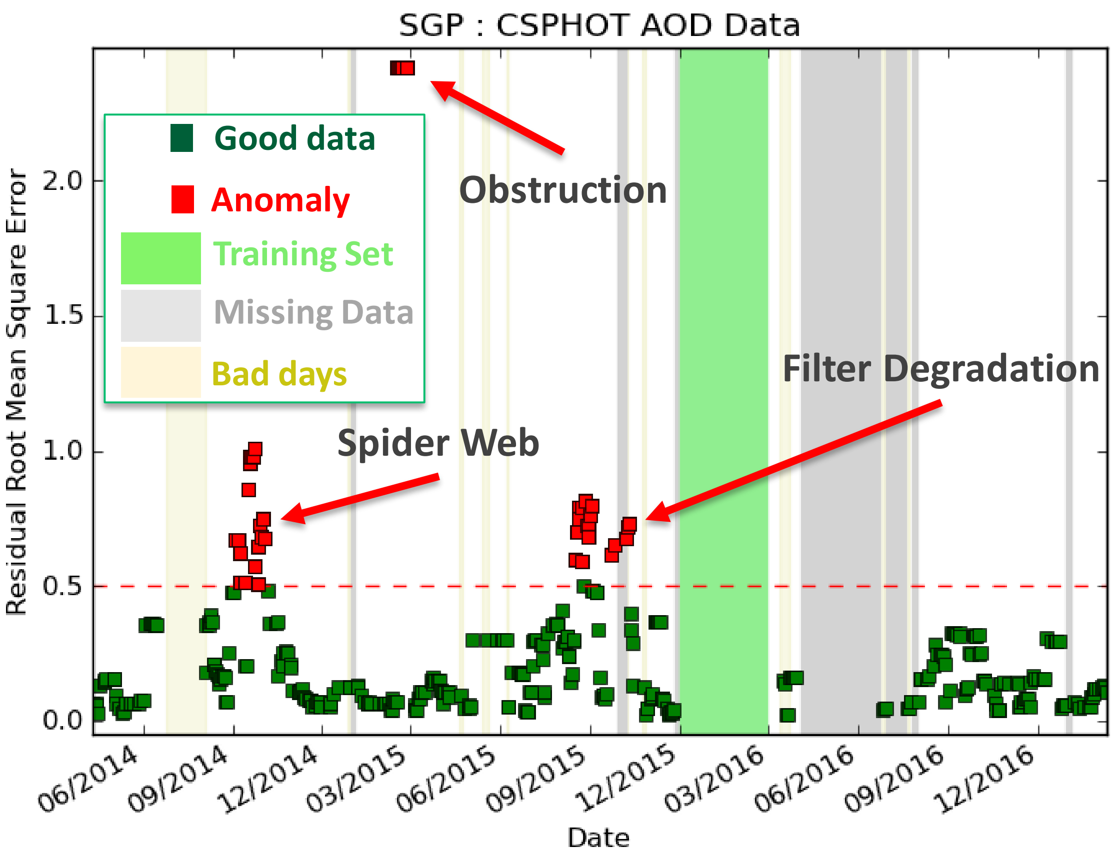
- A random forest model is an ensemble method that builds a set of decision trees from subsets of data and subsets of features. The final result is the average of the results from all of the trees.
- A random forest model is chosen for the following reasons:
  - It generalizes well.
  - The input data does not need to be scaled or processed.
  - Results are easy to interpret and provide information on the nature of the problem.



For the CIMEL sun photometer, the random forest is asked to predict the value of a reference measurement, the AOD value at noon for the 500 nm channel.

CIMEL Sun Photometer Anomaly Detection Results

- A training set covering a period of “good” instrument operation is defined.
- The model is trained using that dataset. It learns the features of the data.
- The trained model is asked to predict the rest of the data. The residual root mean square (RMSE) is reported.
- High RMSE values indicate anomalous days.
- The model translates well to other sites with little tuning.



The application detected all known problems with the instrument over this period.

Running time = 15 seconds per year of data.

Multifilter Rotating Shadow-band Radiometer (MFRSR) Anomaly Detection

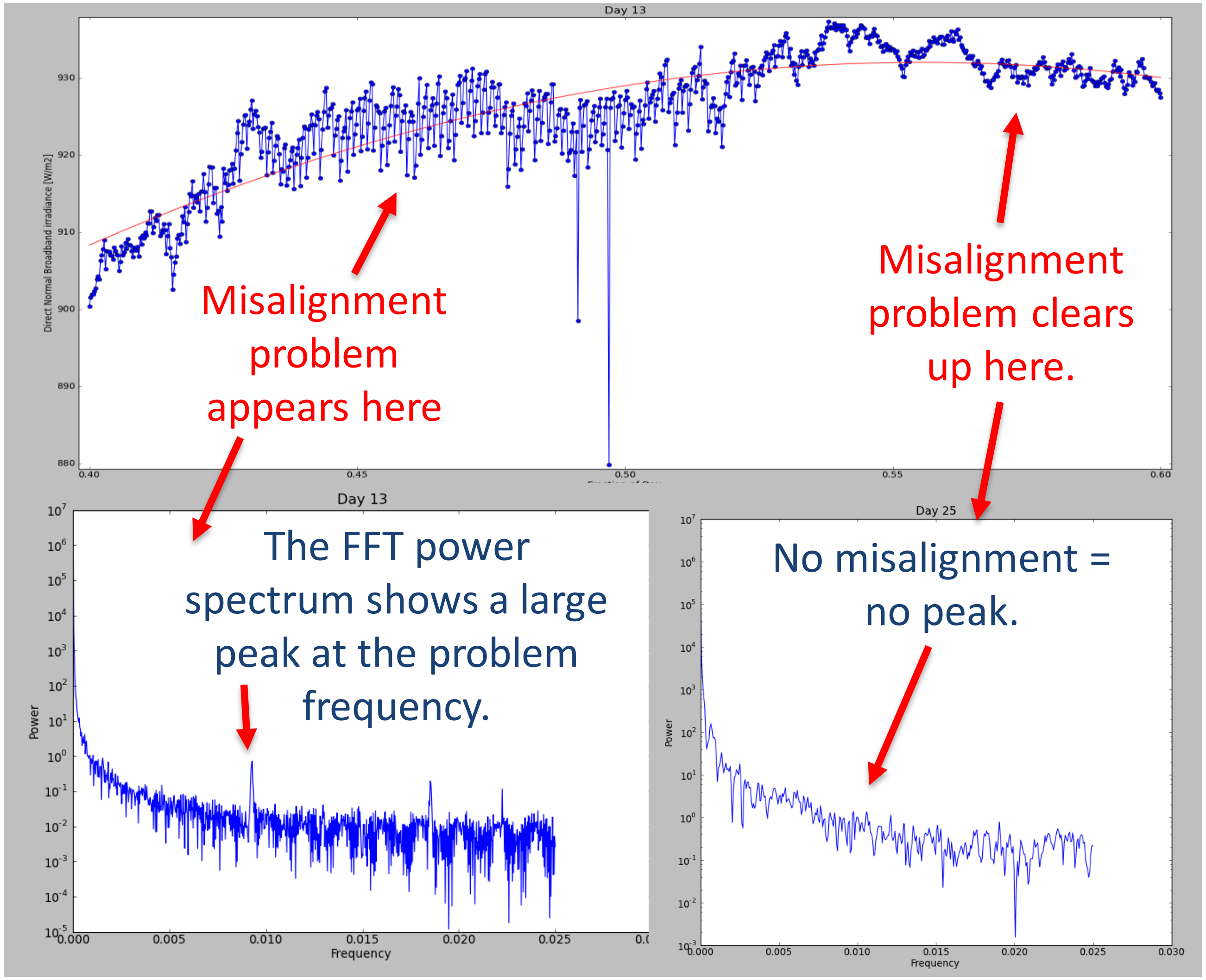
ARM
CLIMATE RESEARCH FACILITY

- The MFRSR takes spectral measurements of direct normal, diffuse horizontal, and total horizontal solar irradiances.
- The sampling rate is 20 seconds.
- Measurements are taken in daylight hours and are affected by clouds.
- A machine learning application similar to that for the CIMEL sun photometer has been developed.
- The application also contains a filter algorithm to detect a common problem due to misalignment of the shadow band.

Detecting Shadow Band Misalignment in the MFRSR

ARM
CLIMATE RESEARCH FACILITY

- This problem mode creates an oscillating pattern in the data.
- A Fast Fourier Transform (FFT) has been shown to be effective to detect this (*M.D. Alexandrov et al., Applied Optics 46, 8027 (2007)*).
- The FFT algorithm is automated here. **One year of data can be analyzed in 2 minutes.**



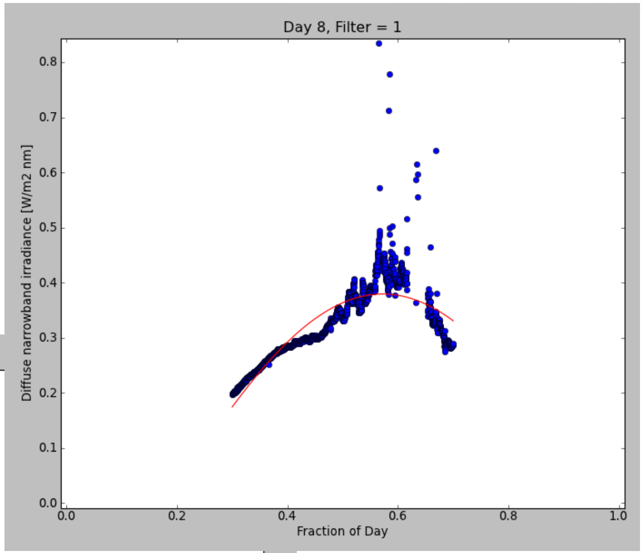
MFRSR Shadow Band Misalignment Detection Results

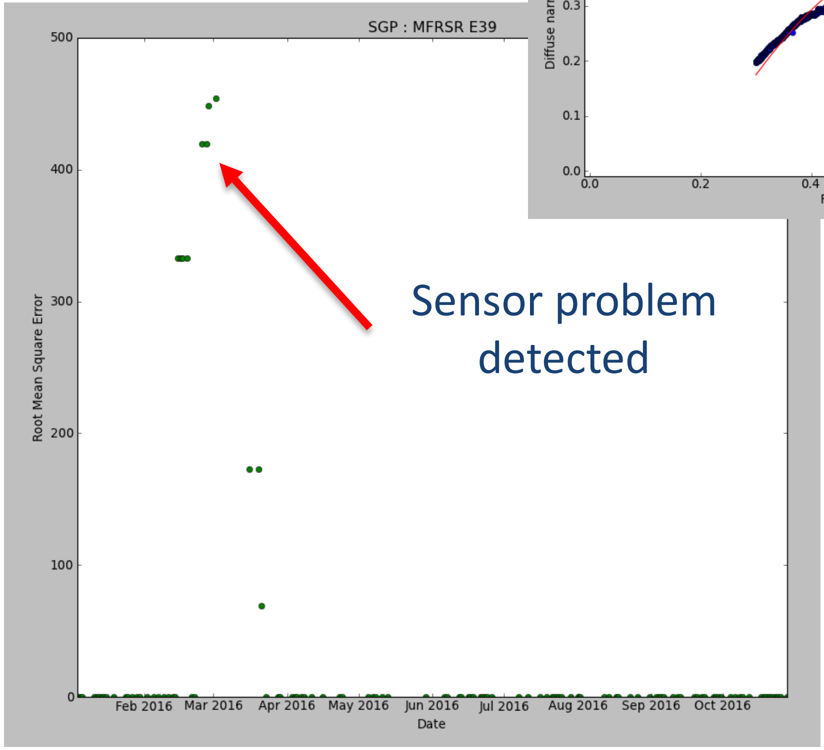
- This shows the output of the FFT algorithm per day from the E33 instrument at the SGP site over one year.
- There is a DQR documenting a misalignment problem covering the red shaded area.
- The FFT algorithm successfully identifies the problem. There are no false positives reported for this year.

MFRSR Anomaly Detection Results

- A multivariate regression model is trained to predict the diffuse narrowband irradiance for filter 1.
- The model is trained on 3 months of nominal operation.
- The model is then compared to the data in the test set and the RMSE is reported.
- A sensor problem in one channel that was reported in a DQR is successfully detected.

A typical day of MFRSR measurements.





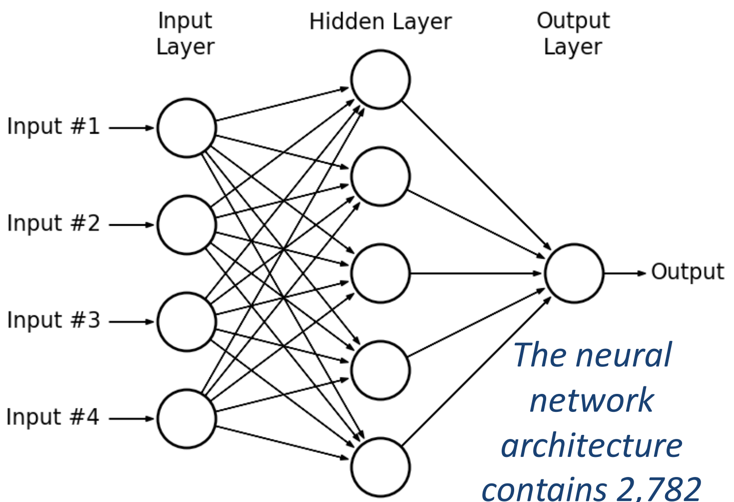
Identifying Local Emission Sources in the Aerosol Observing System (AOS)

ARM
CLIMATE RESEARCH FACILITY

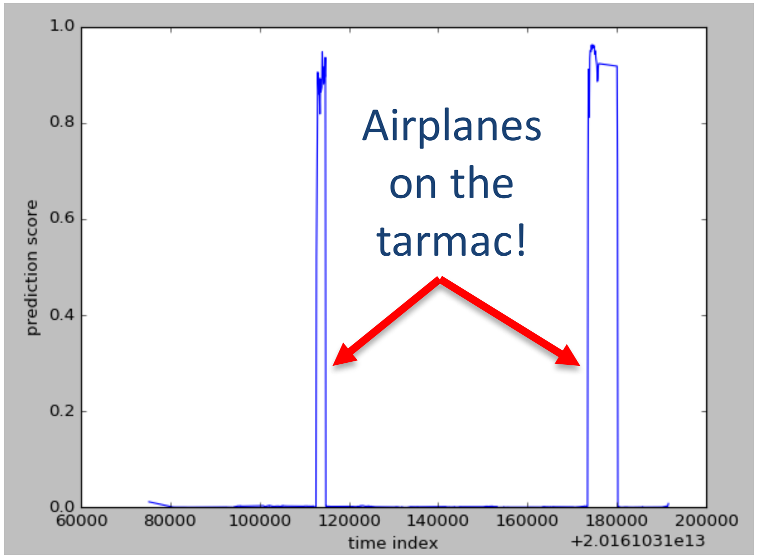
- THE PROJECT: AOS instruments at the ENA site are located next to an airport. Develop an automatic machine learning application to identify emissions from the airport using multiple AOS instruments.

A Neural Network to Identify Airplanes in Tower Camera Images

- For supervised machine learning, the data must be tagged before training.
- Tower camera images of the tarmac are used to tag local emission sources.
- A neural network was developed to automatically identify airplanes in the images.
- Once trained, the neural network can process a day of images (700 of them) in 10 seconds on a laptop.
- The accuracy per image for airplane identification is 96%.



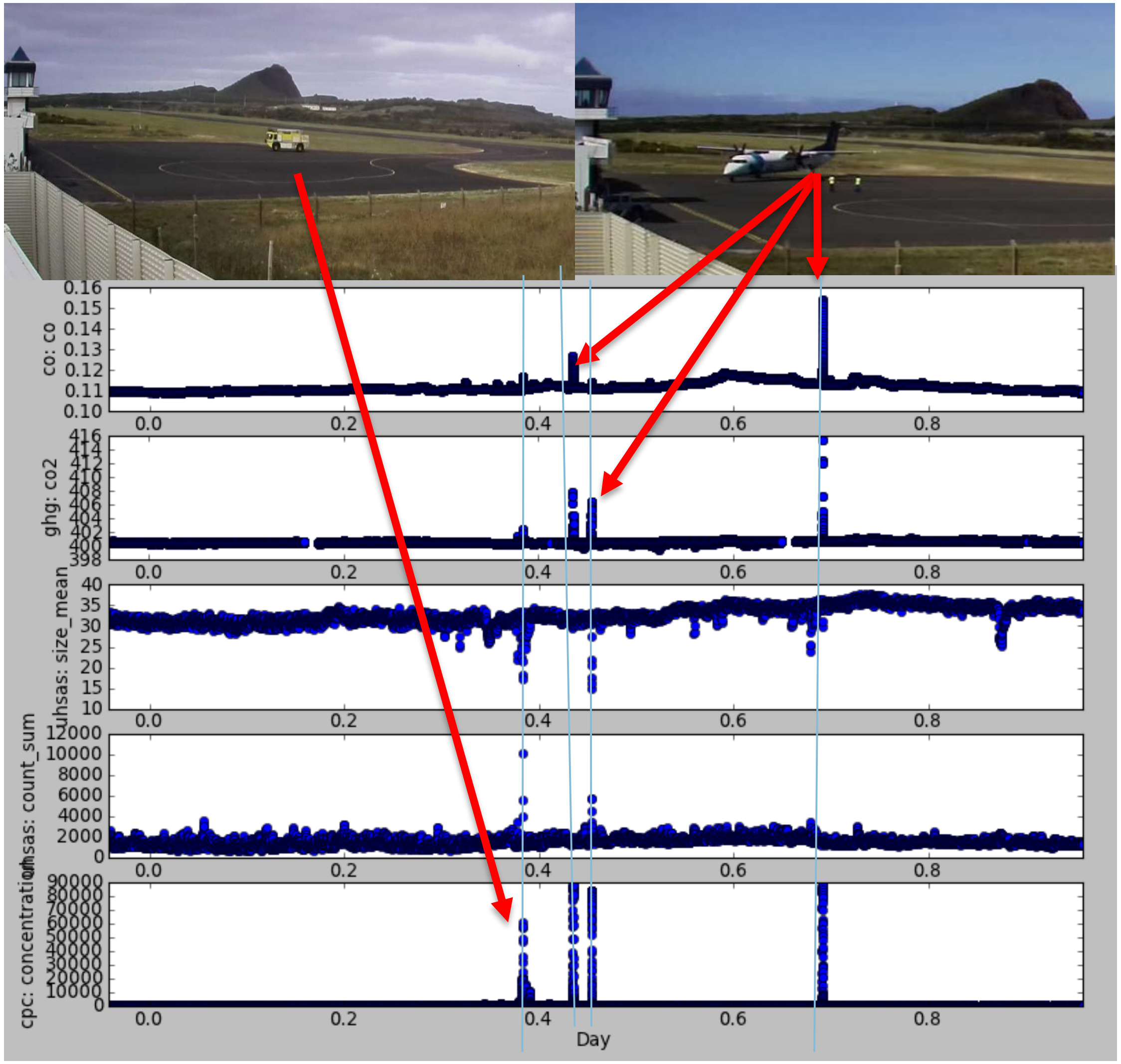
The neural network architecture contains 2,782 input units and 360 hidden units.



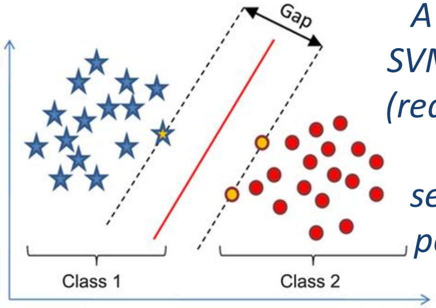
The neural network output of the probability that an airplane is present in each image.

Local Emission Sources in the AOS Data

- This shows 5 simultaneous data streams from 4 different instruments over one day:
  - CO Monitor Carbon monoxide concentration
  - Greenhouse Gas Monitor Carbon dioxide concentration
  - Ultra-high Sensitivity Aerosol Spectrometer (UHSAS) mean particle size
  - UHSAS particle count integral
  - Condensation Particle Counter aerosol concentration
- Most instruments have a sampling rate of 1 second. The UHSAS has a sampling rate of 10 seconds This is more than 8000 measurements per day per data stream.
- Notice that not all events are seen in all data streams. Multiple instruments are necessary.

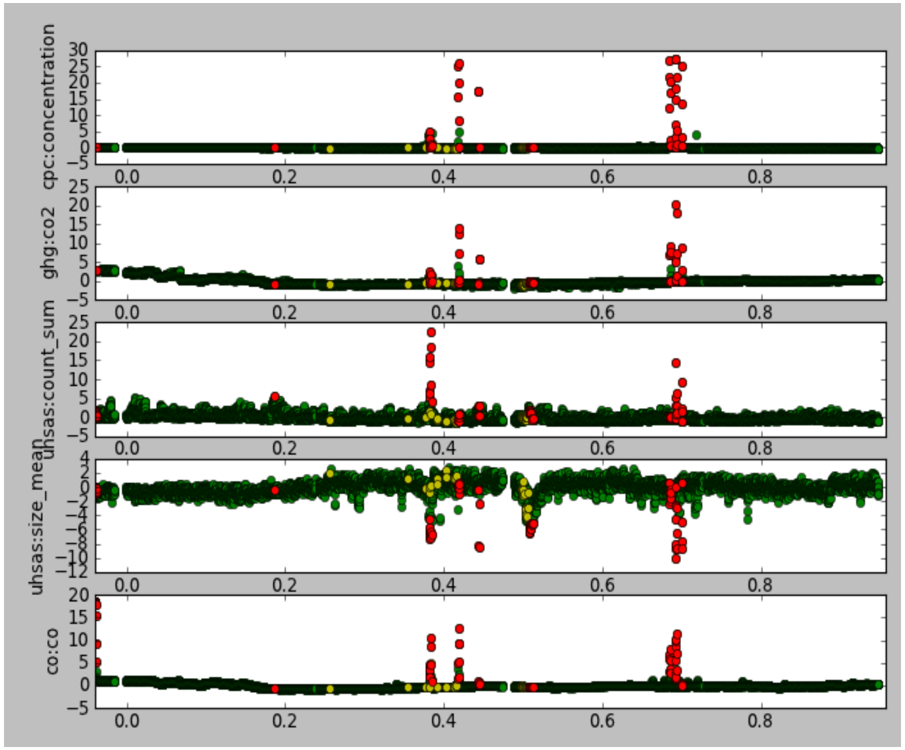


AOS Local Emission Source Detection Results



A 2-D illustration of an SVM model. A hyperplane (red line) is constructed to maximize the class separation. The yellow points are the support vectors.

Results from the SVM model over one day of AOS data (11/21/16). Events from fire trucks and two airplanes are correctly identified (red dots). The yellow dots represent falsely reported anomalies.



- A training data set is defined over 5 days of AOS operation by removing data corresponding to times when local emission sources are present as identified by the neural network.
- A one-class Support Vector Machine (SVM) model is trained. The SVM defines a single class describing good data.
- The model is then compared to data in the test set. The model will report data that lies outside of it's class as anomalies.
- The model is better than 99% accurate for identifying local emission sources.
- **Running time for 1 day of data: 15 seconds**

Summary

- Machine learning algorithms have been applied to anomaly and local source emission detection in several ARM instruments.
- The algorithms are powerful because they can make inferences based upon multiple measurements from multiple instruments simultaneously.
- A fast and accurate assessment of data quality has been demonstrated and can be interpreted at a glance.
- Further evaluation and implementation of the applications is underway.
- Many exciting ARM analyses are possible with the power of machine learning!

Acknowledgements

- I would like to acknowledge everyone who has contributed to these projects including Laurie Gregory, Lynn Ma, Richard Wagener, Alice Cialella, Scott Smith, Stephen Springston, Thomas Madigan, Art Sedlacek, Andrew McMahon, Laura Riihimaki, and Connor Flynn.

- More details can be found in the following posters:
  - A2 Poster #101 (AOS) and B2 Poster # 136 (CIMEL Sun Photometer)