# Anomaly Detection for ARM Radiometers using Machine Learning Algorithms

LAURIE GREGORY[A], JEFFERY MITCHELL[A], RICHARD WAGENER[A], LYNN MA[A] AND LAURA RIIHIMAKI[B]

[a]*Brookhaven National Laboratory;* [b]*Pacific Northwest National Laboratory*

ARM/ASR PI Meeting, Tysons, Virginia

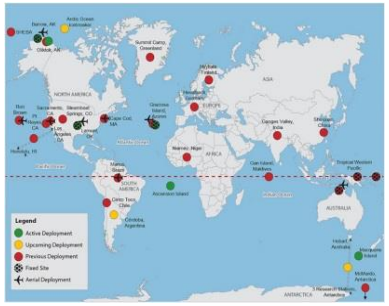
# Outline

- ARM Cimel Sun Photometer
- Mentorship – Daily data review- daunting task
- Why Machine Learning offers a solution
- Machine learning workflow and methodology
- Machine Learning results
- ARM wide applications- a coordinated effort
- Future Directions

# ARM Radiometer- Cimel Sun Photometer

- The CIMEL Sun Photometer (CSPHOT) is a multi-spectral automatic sun and sky scanning radiometer that measures the direct solar irradiance and sky radiance at the Earth's surface.
- Key Primary Measurement: Aerosol Optical Depth (AOD)
- Deployed at all ARM past and current sites - 16
- Part of NASA AERONET program – 100's of Cimel's deployed operationally
- Sensitive to the environment

(rain, snow, clouds, extreme temperature, etc..)

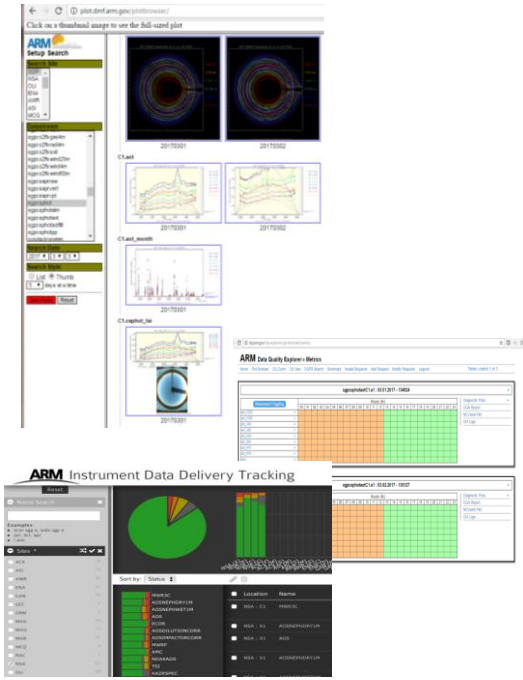



CSPHOTs at ARM sites

CSPHOT in Antarctica

# CSPHOT Data Quality Review- A Daunting Task

- **Many CSPHOTs** at a variety of ARM sites and conditions around the world
- Essential that ARM Instrument Mentors provide accurate and consistent data quality reports for all ARM data
- Daily data plot reviews are currently performed by humans- **time consuming, inefficient, and higher risk of inaccuracies**
- Many tools to review: The ARM Instrument Data Delivery Tracking tool, DQR's, DQA's, DQPR's, comparisons to other instruments, DQ Plot Browser, OSS, ARM DQ Explorer, ENGs, site reports




**Review process would be improved with automation techniques**

# Machine Learning in the Real World

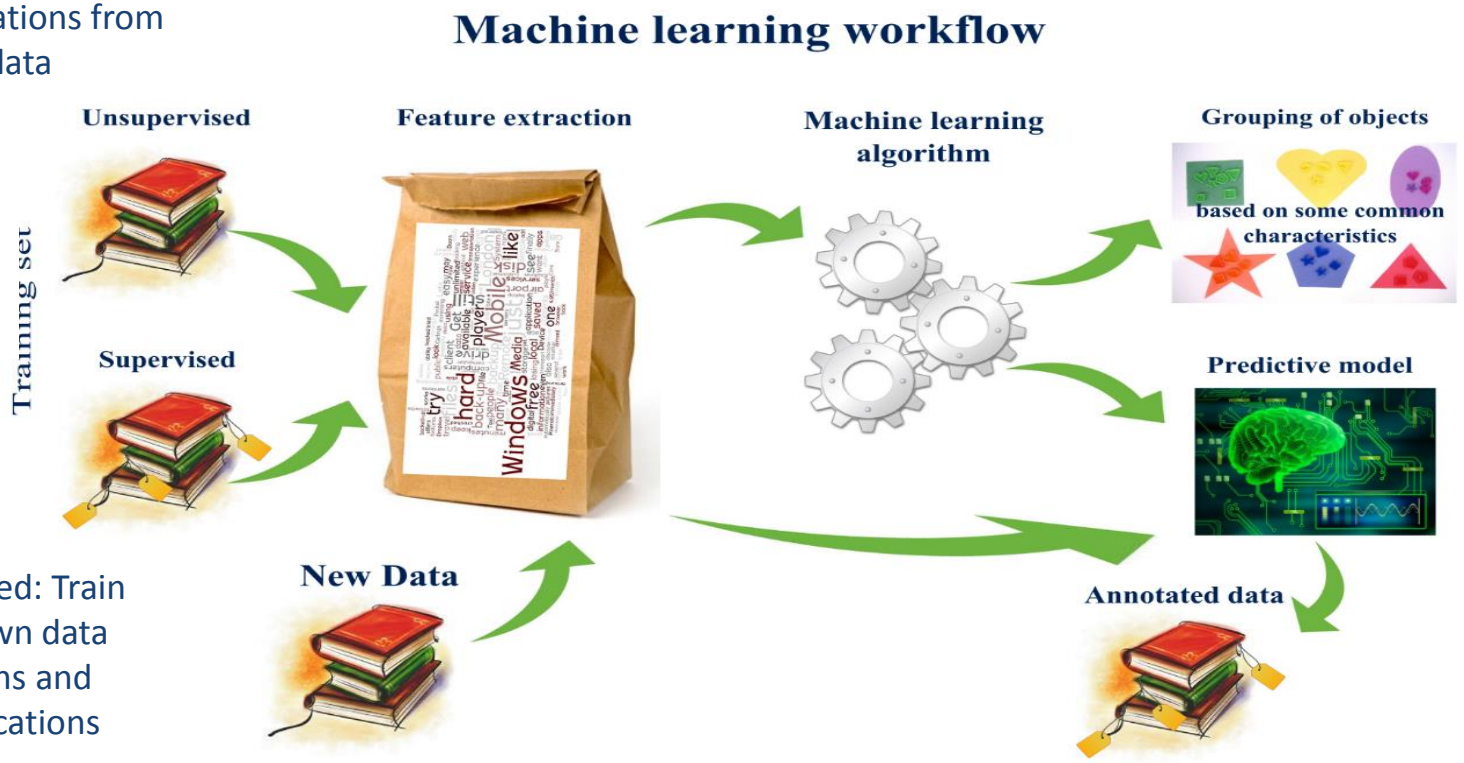
The use of machine learning applications has exploded over the past several years

- Movie / Purchase Recommendation Engines (Netflix, Amazon)
- Computer Vision and Image Recognition (Facebook)
- Self-driving cars (Google, Uber)
- Malware Detection
- Fraud Detection
- Financial Market Prediction
- Natural Language Processing (Siri)

**Why not try machine learning to detect ARM instrument anomalies as well**

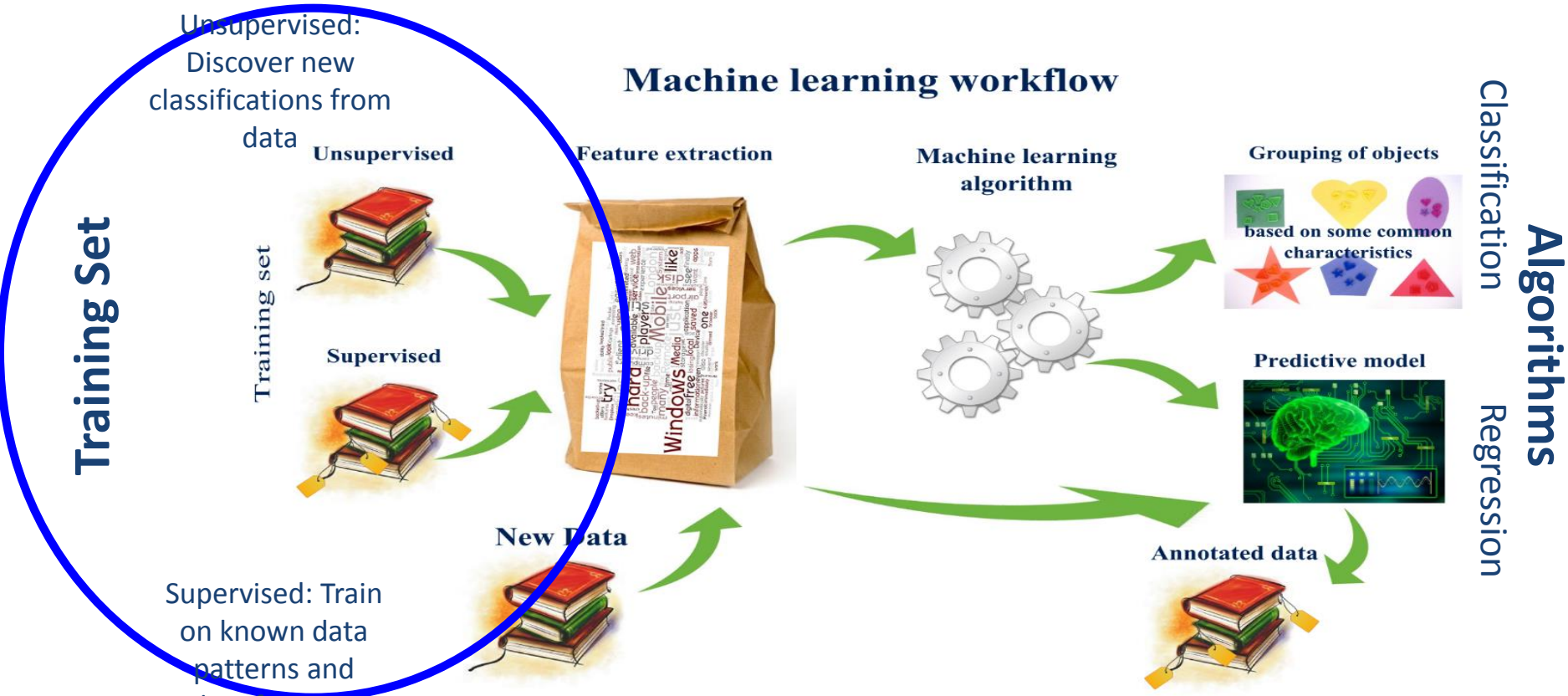
# Machine Learning Workflow

Unsupervised: Discover new classifications from data




Supervised: Train on known data patterns and classifications

# Machine Learning Workflow

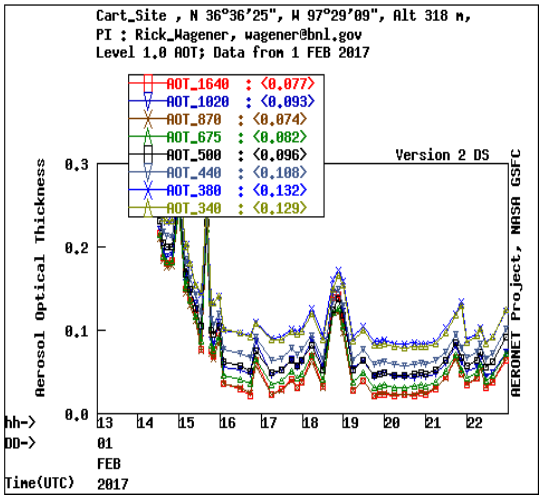


# Identify training sets for Supervised Learning

## Identify Patterns in the Data.

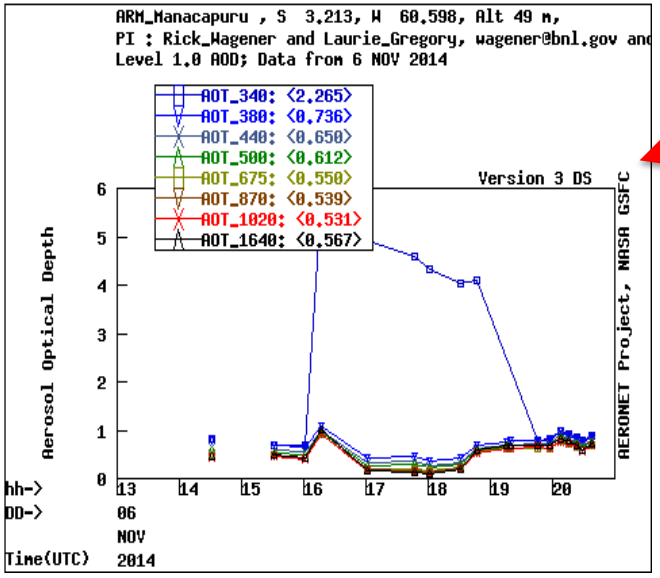
Here we looked for patterns that are associated with known DQ issues. We **train the model** to distinguish between data where instrument is operating normally and data when there are problems.

### Operating Normally



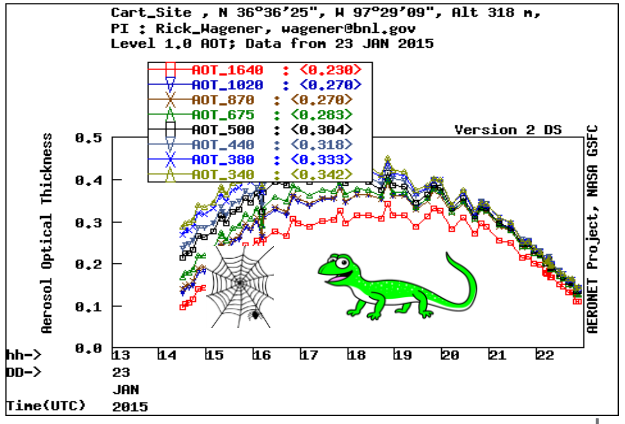

AERONET website: AOD plot

### Filter Degradation




AERONET website: AOD plot

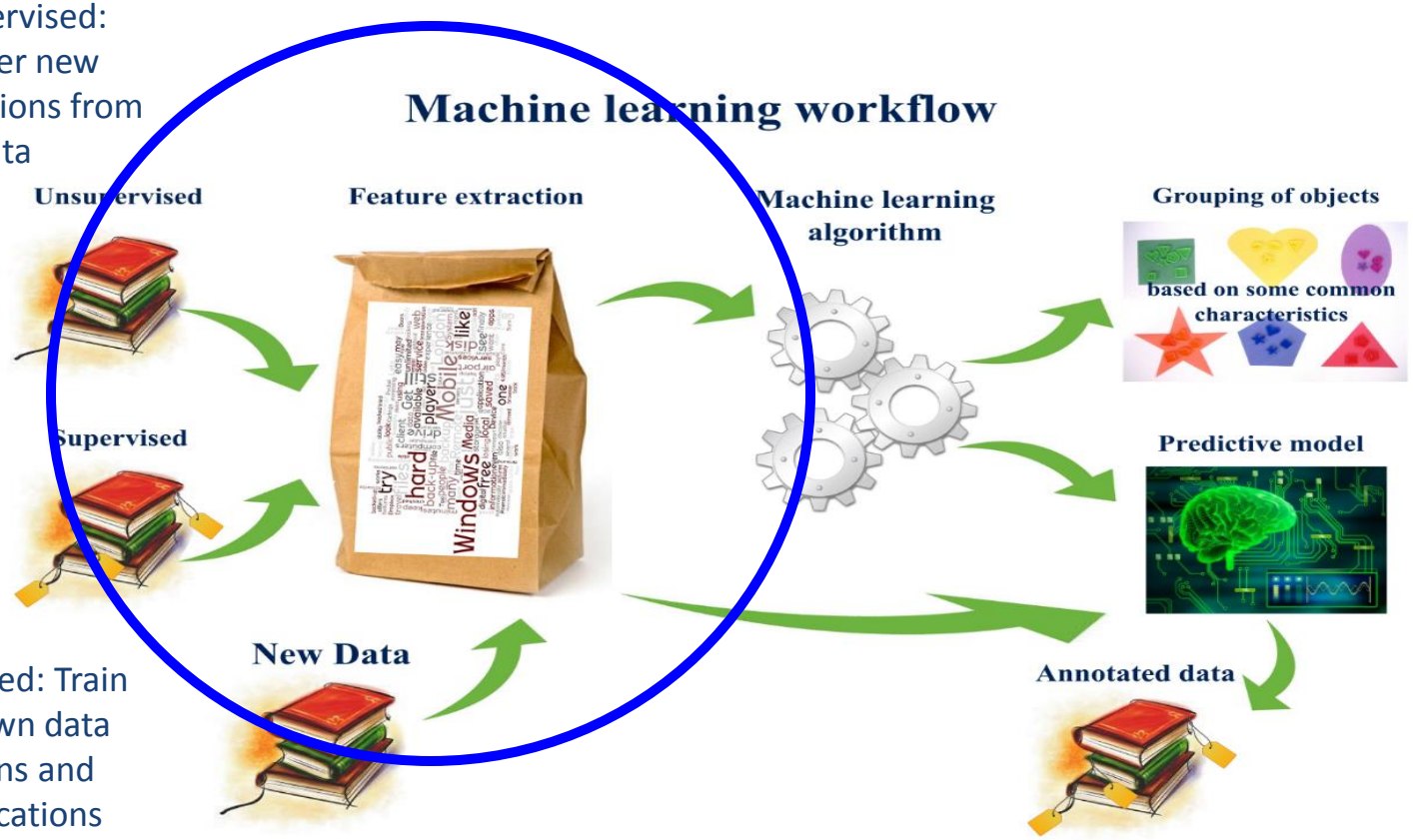
### Tube Obstruction




AERONET website: AOD plot

# Machine Learning Workflow

Unsupervised: Discover new classifications from data




Supervised: Train on known data patterns and classifications

Training Set

Algorithms

Classification

Regression

# Identifying Features

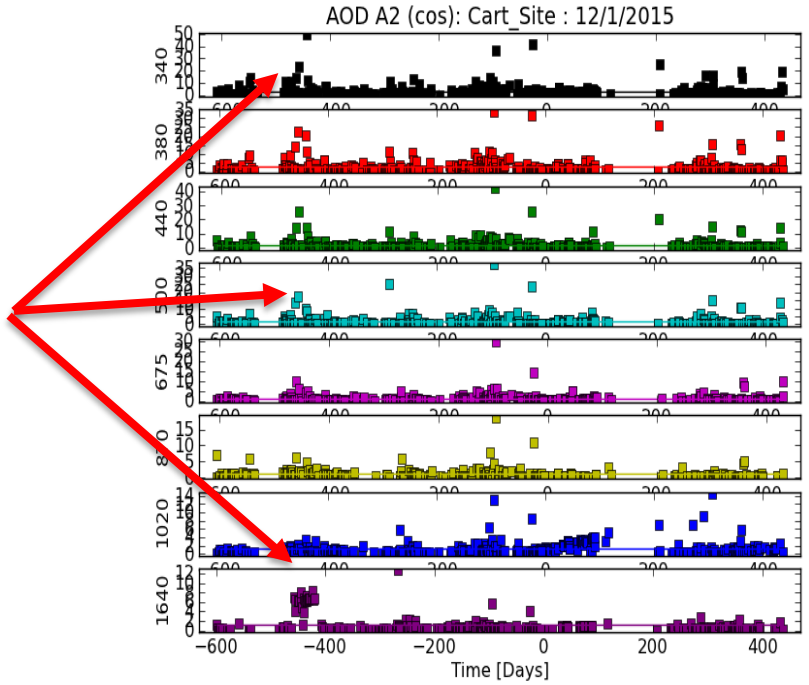
**Identify features that describe the data patterns that we see (e.g. a curve in the data indicating an obstruction) in mathematical terms for the training set.**

- A total of 57 features are extracted and monitored as a time series, such as
  - Angstrom exponents which are sensitive to deviations between channels (which could indicate a channel is out of alignment)
  - Coefficients from a fit to a curve of the data (indicating daily curvature)
- Data affected by clouds are screened and removed from the training set
- The time series for each feature is plotted. Plots can be viewed interactively, saved to a file, or both
- At the end, a summary of days with feature deviations is produced.

# Identifying Features

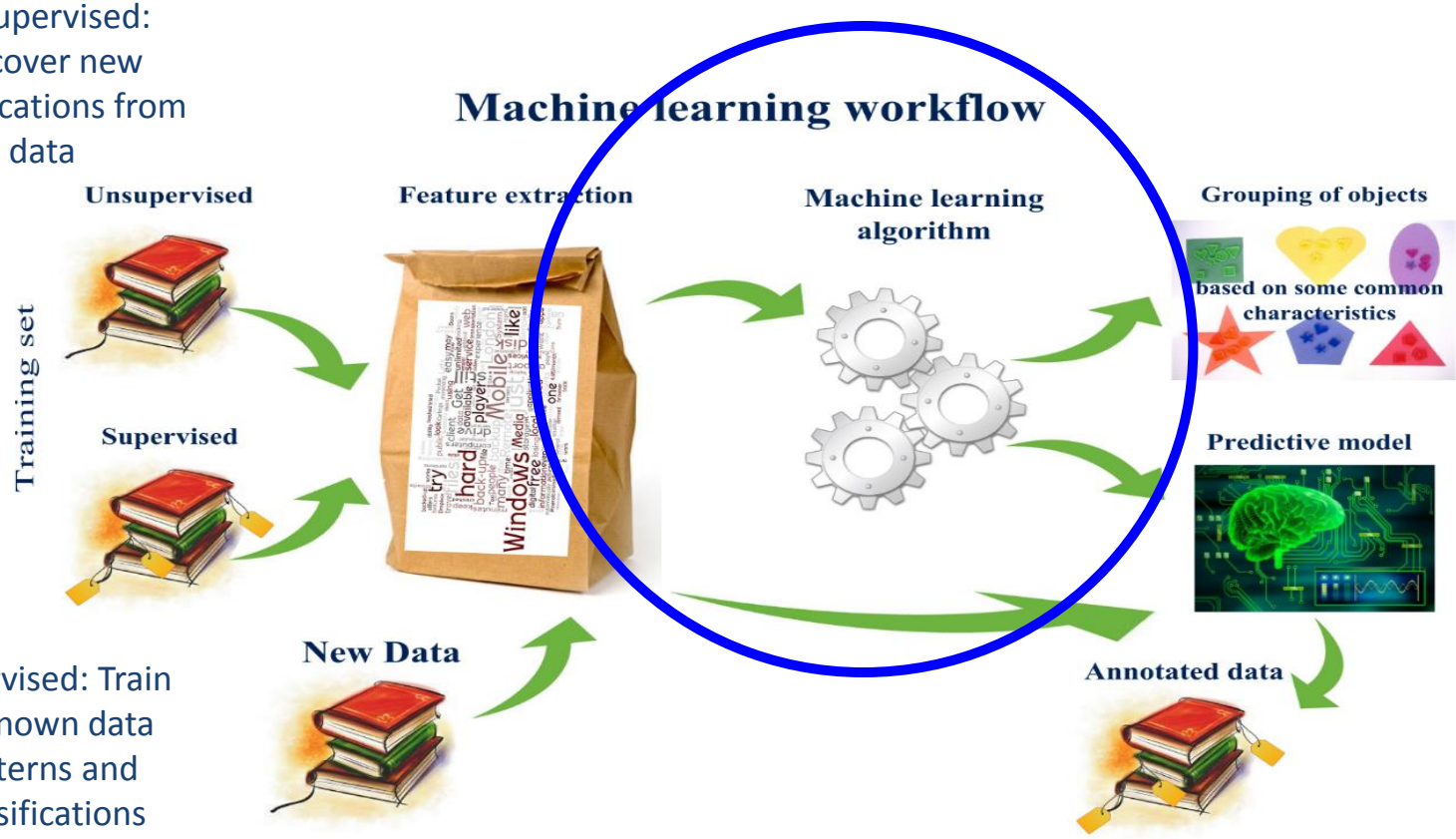
Example of how features can be used to identify problems, like a spider web

**Days where spider web was obstructing the instrument.**



# Machine Learning Workflow

Unsupervised: Discover new classifications from data




Supervised: Train on known data patterns and classifications

Training Set

Algorithms

Classification

Regression

# Choosing an Algorithm - The Random Forest Model

- 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.

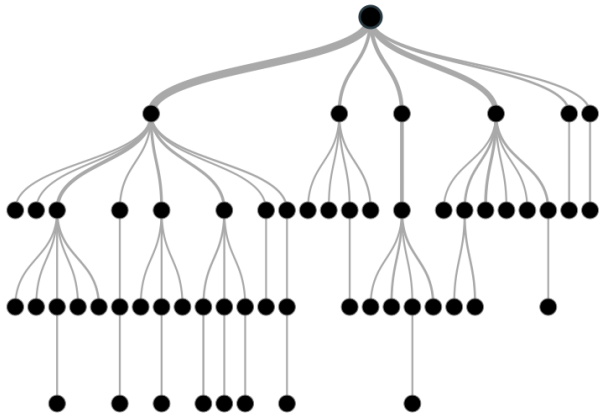
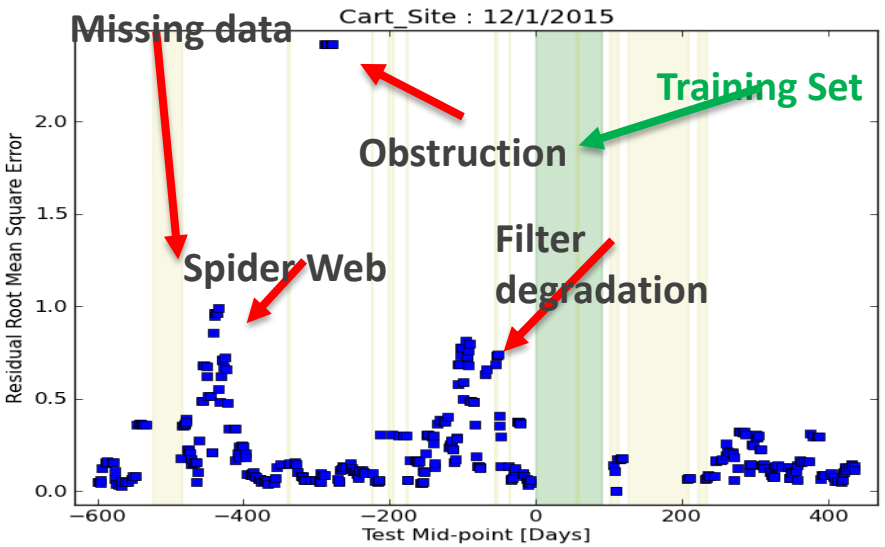
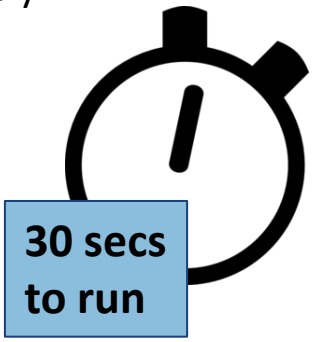


Image showing decision tree from www.analyticsvidhya.com\

# Impressive Machine Learning Results

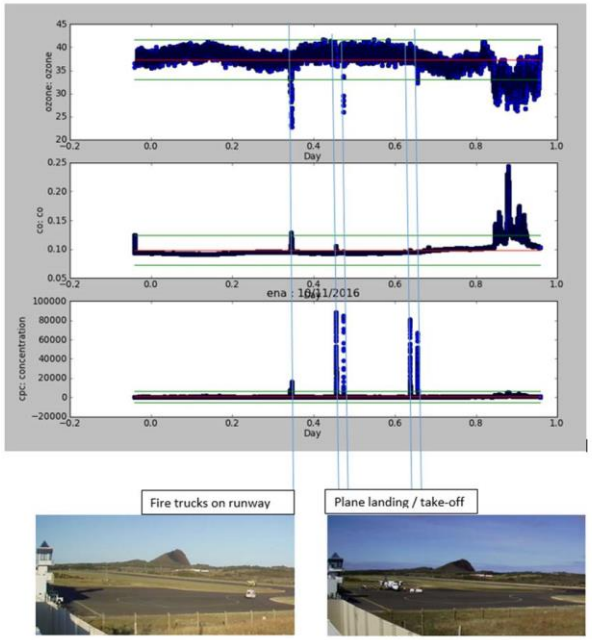



- Model incorporates many features of instrument key measurements simultaneously
- **"Trained" for periods when instrument operated normally**
- Deviations against training fit indicate anomalies
- Validated using existing Data Quality Reports
- **Faster, more sensitive than human eyes, and automated**
- **One two-year run took only 30 seconds**
- Currently testing model in operational mode and running model at other sites



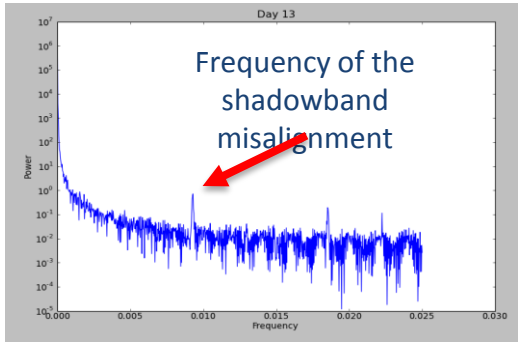
# A Coordinated ARM-wide Effort

Working with AOS Mentors at BNL: combining images data with data from multiple instruments (Poster #101)




Working with Xiao Chen at LLNL on Uncertainty Quantification (see talk at breakout session)

Working with Laura Riihimaki and Connor Flynn to apply the algorithms on other instruments – MFRSR shadowband misalignment problem (Poster #136)



# Future Directions

- Continue applying anomaly detection to other ARM instruments
- Automatically create data quality reports
- Could create open source Python application through GitHub for scientists
- Generalized data quality tools for DQ office
- Integrate into the ARM Development Environment (ADI)

# Summary

- Machine Learning is currently being used in ARM.
- Efficient and effective
- Coordinated effort across ARM
- Can provide more automation in the ARM Facility


## Acknowledgements

- I would like to acknowledge Alice Cialella for her support

# For More Details…

## Come visit our posters

Tuesday night, A2, Poster #101

*Identifying the Influence of Local Source Emissions on the Regional Representativeness of AOS Measurements using Machine Learning*

Wednesday night, B2, Poster #136

*Anomaly Detection for ARM Radiometers using Machine Learning Algorithms*

## Breakout session on Wednesday 1:30 – 3:30

*Beyond QA: Using New Techniques to Quantify Retrieval Quality and Uncertainty (Potomac) Conveners: Laura Riihimaki, Evgueni Kassianov, Connor Flynn, Laurie Gregory*

# References

## References:

Adams, B. , L. Gregory, R. Wagener, “Automatically detecting typical failure signatures in Cimel Sun-photometer data to improve data quality”, Poster presented at New York Scientific Data Summit, NYU ,New York, August 2-5, 2015

Alexandrov, M.D, et al, “Optical depth measurements by shadow-band radiometers and their uncertainties”, M.D. Alexandrov et al., APPLIED OPTICS _ Vol. 46, No. 33 _ 20 November 2007.

## Image References:

Machine Learning workflow images: http://www.datascienceassn.org/content/machine-learning-workflow.

Instrument and application images: www.arm.gov

Random Forest Image: https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python