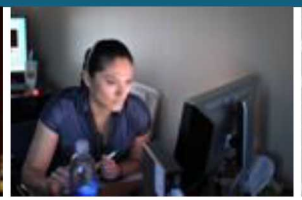
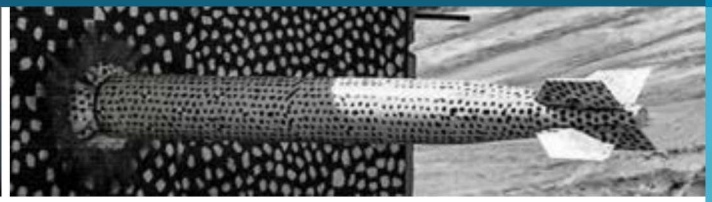
SAND2019-8843PE

Data-Driven Analysis of PV Failures from O&M Records





PRESENTED BY

Thushara Gunda & Birk Jones

Renewables O&M Innovation Workshop

July 31, 2019

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Introductions

CMMS

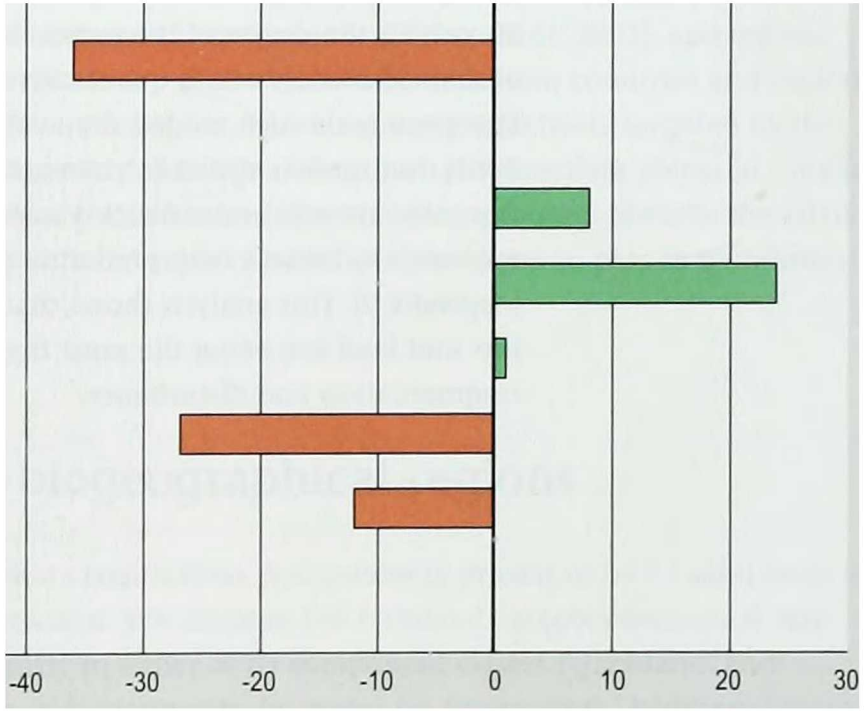
Computerized Maintenance Management Systems (CMMS) significantly improve workflow by helping to track and prioritize needed maintenance activities to ensure ongoing operations of systems

Source: Manufacturing.Net

More than a To Do List!

Collect key pieces of information, including timing, descriptions, and actions associated with an event.

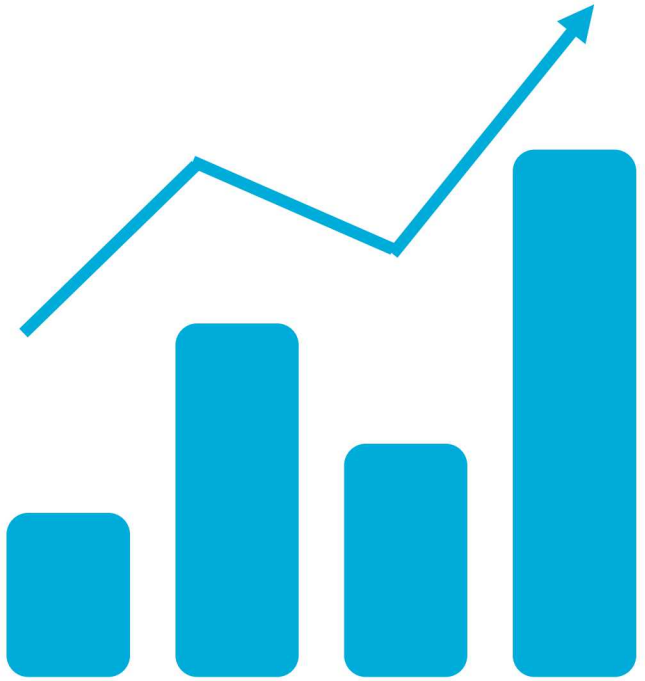
Event data can be characterized and processed to reveal valuable insights into system-level and portfolio-level activities.

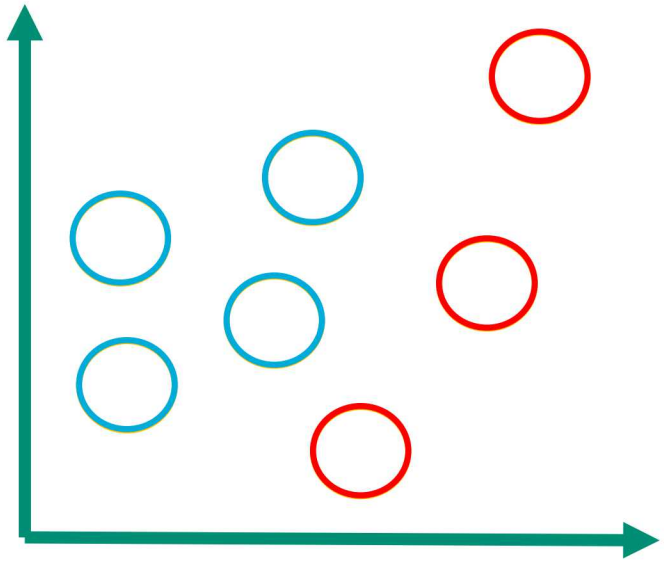


Analytical Techniques

Analytical Techniques

1. Data Visualizations
    1. Word Clouds
    2. Pareto Charts
    3. Time Series Charts
2. Machine Learning
    1. Unsupervised Techniques
    2. Supervised Techniques
3. Distribution Analyses





Data Visualizations

Strongly informed by the data type

Continuous event data:

- energy generated, weather data, etc.
- consistent time intervals
- numerical entries

Discrete event data:

- specific CM or PM actions
- capture dates
- mostly text

LocalPeriodEnd	TotalGeneration	ExpectedEnergy	Insolation	MedianPOAIrradiance	MedianBackPanelTemp
4/1/2018 8:00	119.551	112.359296	0.0623677	59.325	13.218
4/1/2018 9:00	420.737	347.789764	0.1973222	177.9455	19.9415
4/1/2018 10:00	1226.775	1249.959302	0.5720632	602.071	33.4555
4/1/2018 11:00	1653.231	2000	0.7606871	754.2395	39.679
4/1/2018 12:00	1961.264	2000	0.9062962	910.563	44.845
4/1/2018 13:00	2014.994	2000	0.9155277	916.731	47.7575
4/1/2018 14:00	1511.998	1229.058135	0.6348016	576.8555	39.513
4/1/2018 15:00	1272.255	1132.119992	0.5242663	534.0695	38.311
4/1/2018 16:00	1323.599	1232.318724	0.5549863	579.366	39.1125
4/1/2018 17:00	1184.033	971.127381	0.4834061	458.972	37.875
4/1/2018 18:00	778.03	630.441741	0.3064237	304.9295	32.5
4/2/2018 8:00	122.806	178.716181	0.1028118	97.016	7.459
4/2/2018 9:00	670.866	680.513562	0.3470578	347.727	20.226
4/2/2018 10:00	1335.555	1268.045971	0.6175628	619.3055	30.281
4/2/2018 11:00	1867.516	2000	0.8430331	843.744	37.0285
4/2/2018 12:00	2067.981	2000	1.0048543	1009.629	42.071
4/2/2018 13:00	2067.2	2000	1.0883391	1091.2615	43.952
4/2/2018 14:00	1975.894	2000	0.9593882	976.453	43.781
4/2/2018 15:00	1717.104	2000	0.7590282	737.6225	38.231
4/2/2018 16:00	1611.148	1337.451264	0.6801357	634.7125	36.9085

Incident Description	Corrective Action	Additional Comments	Occurred
Tilt sensor wire ripped on tracker controller #1.	Replaced with a new tilt sensor.		2/24/12 20:00
Broken module with impact hole in the middle of the module on tracker controller #9.	Replaced with a new module and scrapped the old one.	Impacted module replaced completely.	2/24/12 20:00
Controller #18 stuck east. Suspect PLC issue.	Replaced existing PLC not sending signal to VFD.		3/5/12 20:00

Data Visualizations

Strongly informed by the data type

Continuous event data:

- energy generated, weather data, etc.
- consistent time intervals
- numerical entries

LocalPeriodEnd	TotalGeneration	ExpectedEnergy	Insolation	MedianPOAIrradiance	MedianBackPanelTemp
4/1/2018 8:00	119.551	112.359296	0.0623677	59.325	13.218
4/1/2018 9:00	420.737	347.789764	0.1973222	177.9455	19.9415
4/1/2018 10:00	1226.775	1249.959302	0.5720632	602.071	33.4555
4/1/2018 11:00	1653.231	2000	0.7606871	754.2395	39.679
4/1/2018 12:00	1961.264	2000	0.9062962	910.563	44.845
4/1/2018 13:00	2014.994	2000	0.9155277	916.731	47.7575
4/1/2018 14:00	1511.998	1229.058135	0.6348016	576.8555	39.513
4/1/2018 15:00	1272.255	1132.119992	0.5242663	534.0695	38.311
4/1/2018 16:00	1323.599	1232.318724	0.5549863	579.366	39.1125
4/1/2018 17:00	1184.033	971.127381	0.4834061	458.972	37.875
4/1/2018 18:00	778.03	630.441741	0.3064237	304.9295	32.5
4/2/2018 8:00	122.806	178.716181	0.1028118	97.016	7.459
4/2/2018 9:00	670.866	680.513562	0.3470578	347.727	20.226
4/2/2018 10:00	1335.555	1268.045971	0.6175628	619.3055	30.281
4/2/2018 11:00	1867.516	2000	0.8430331	843.744	37.0285
4/2/2018 12:00	2067.981	2000	1.0048543	1009.629	42.071
4/2/2018 13:00	2067.2	2000	1.0883391	1091.2615	43.952
4/2/2018 14:00	1975.894	2000	0.9593882	976.453	43.781
4/2/2018 15:00	1717.104	2000	0.7590282	737.6225	38.231
4/2/2018 16:00	1611.148	1337.451264	0.6801357	634.7125	36.9085

Discrete event data:

- specific CM or PM actions
- capture dates
- mostly text

Incident Description	Corrective Action	Additional Comments	Occurred
Tilt sensor wire ripped on tracker controller #1.	Replaced with a new tilt sensor.		2/24/12 20:00
Broken module with impact hole in the middle of the module on tracker controller #9.	Replaced with a new module and scrapped the old one.	Impacted module replaced completely.	2/24/12 20:00
Controller #18 stuck east. Suspect PLC issue.	Replaced existing PLC not sending signal to VFD.		3/5/12 20:00

Word Clouds

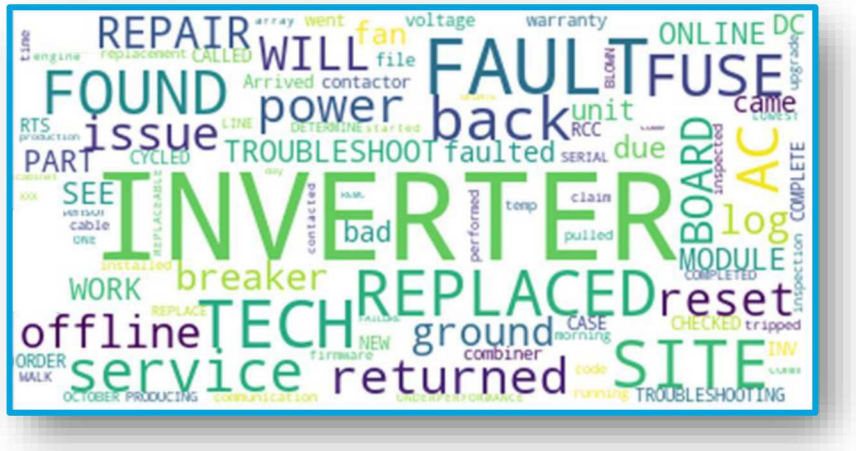
Visualize common words

Word size normalized by frequency of occurrence

Word Clouds

Visualize common words

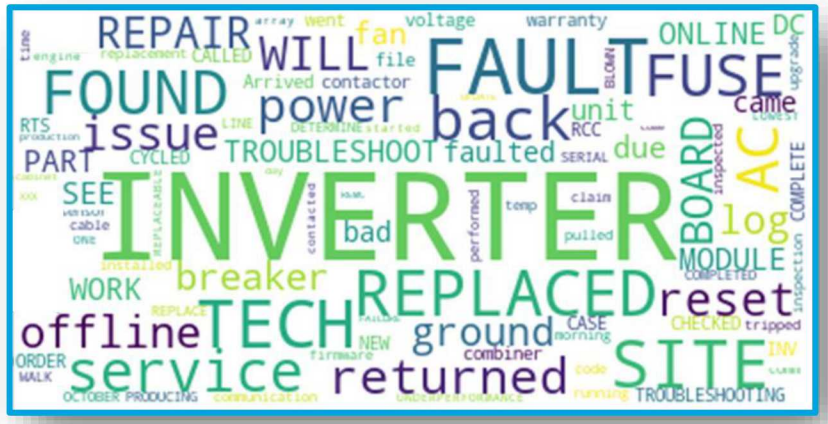
Word size normalized by frequency of occurrence



Word Clouds

Visualize common words

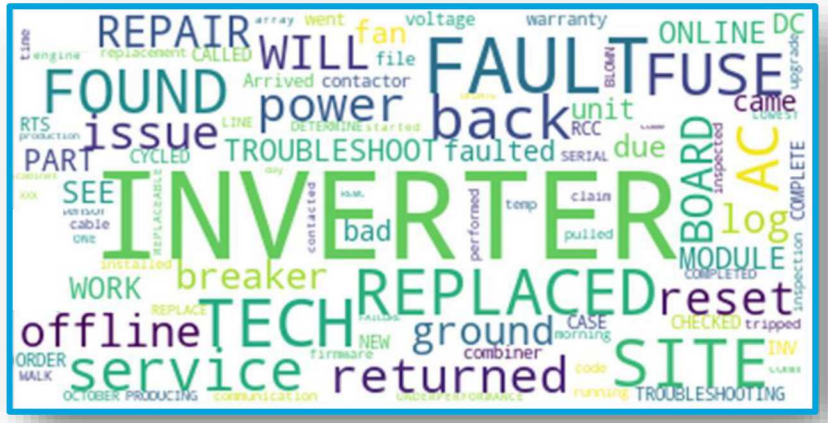
Word size normalized by frequency of occurrence

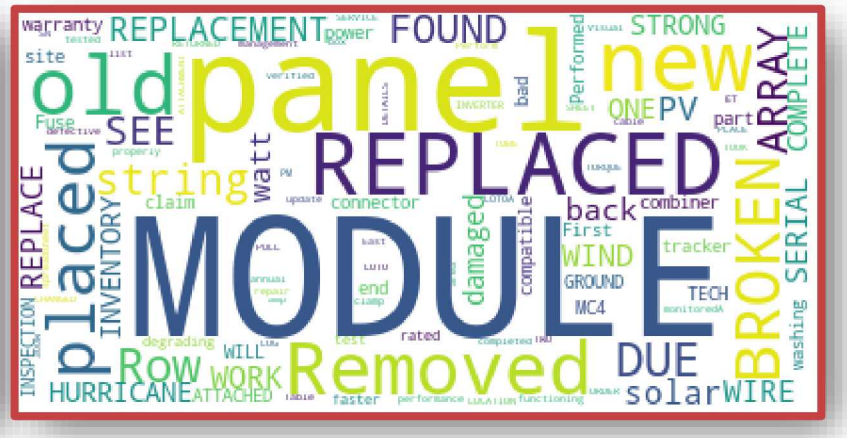


Word Clouds

Visualize common words

Word size normalized by frequency of occurrence

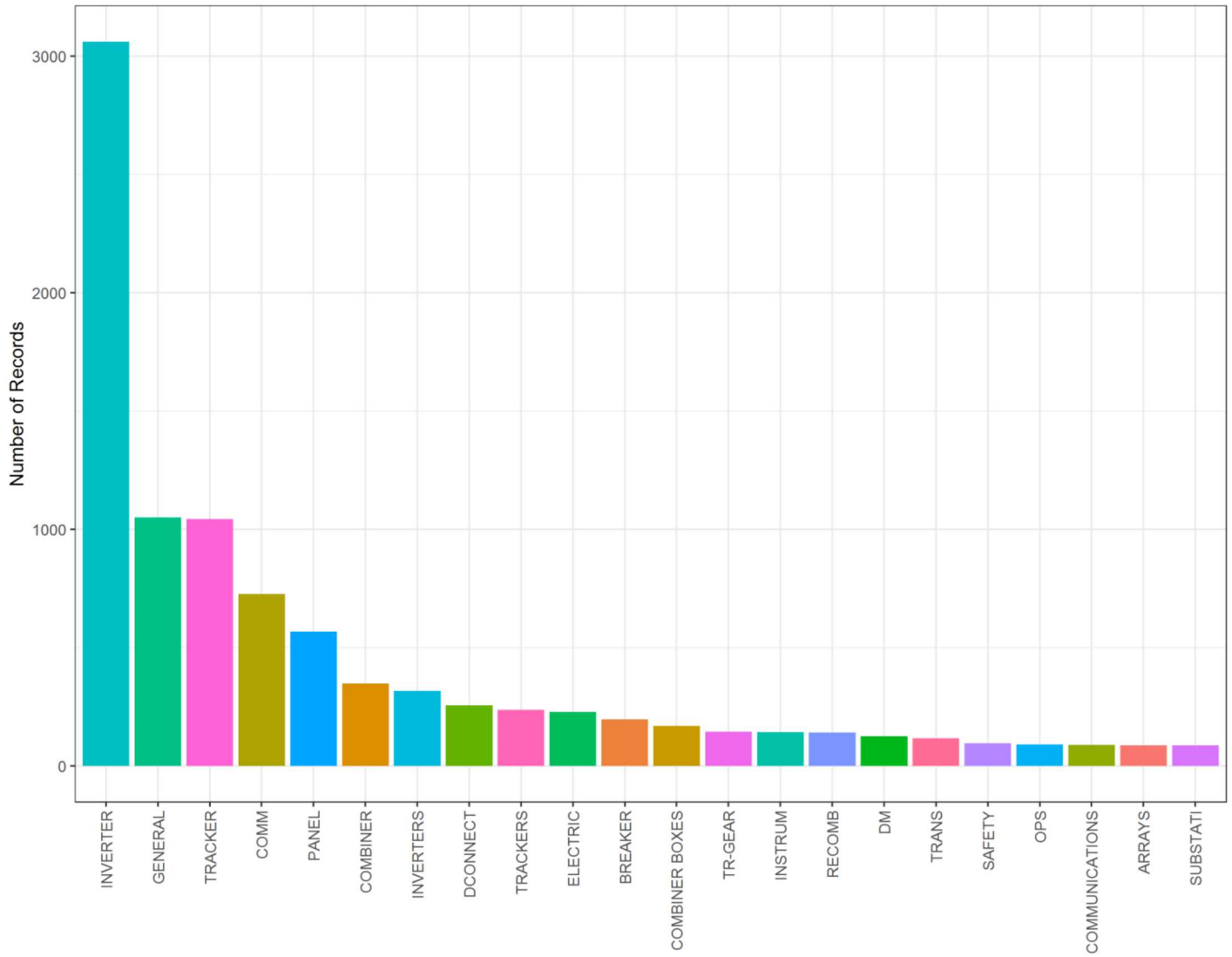




Pareto Charts

Rank order items based on frequency

Often considered a valuable reliable tool for identifying vital problems (Barringer)



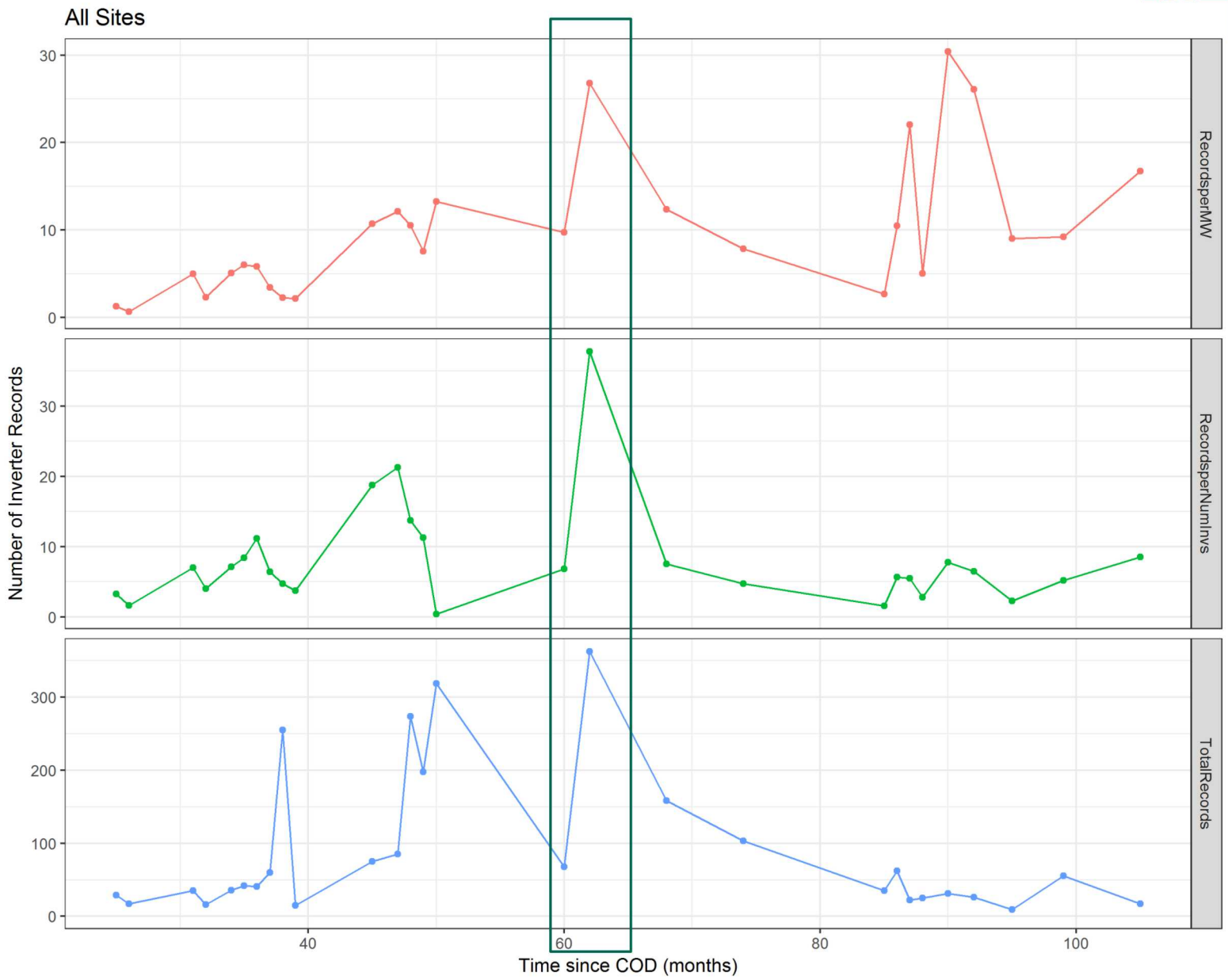
Time Series

Visualize patterns over time

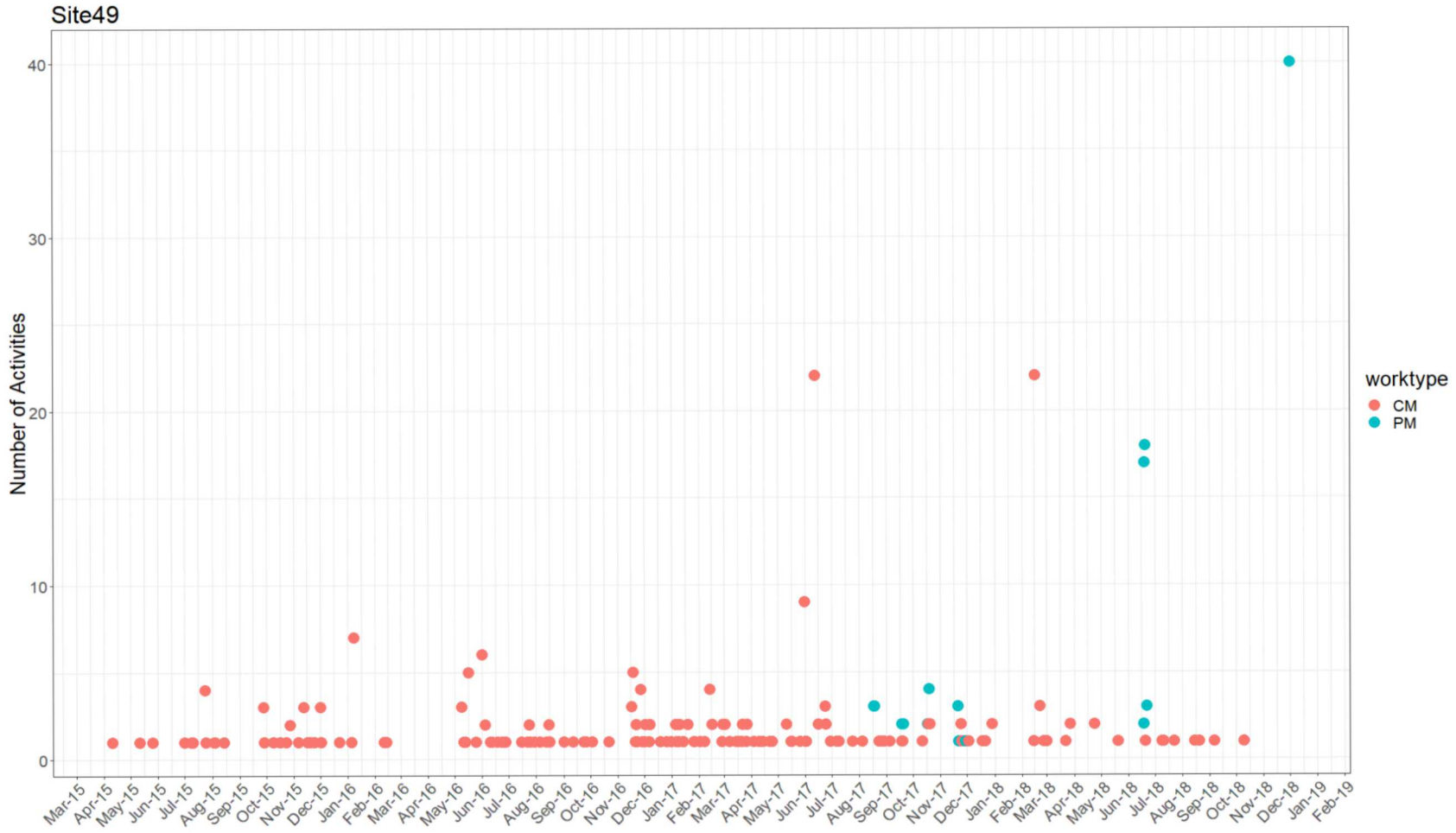
Site-specific characteristics become relevant

Ensure patterns are robust by normalizing across key features

- All 3 visualization indicate a high number of inverter issues ~60 months
- Generally see an increasing trend in issues based on Records per MW until 60 months then see a notable dip

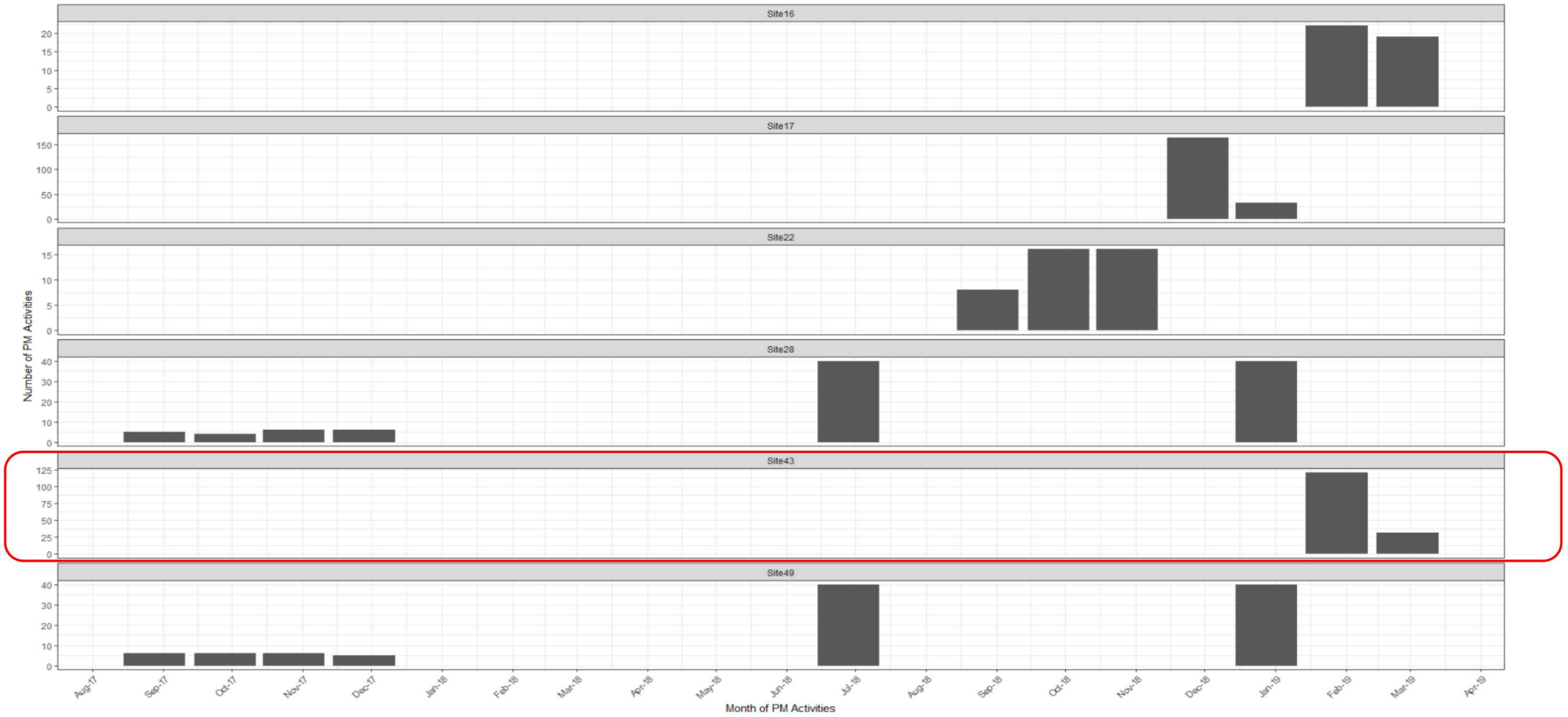


Time Series



- "What" is being visualized is also strongly dependent on data availability
- Only 15 months of overlap in CM and PM datasets…
- With larger dataset, could try to understand timing of PM and CM activities

Data Quality Considerations



Common challenge:

Does lack of data = absence of activity or just lack of data capture?

No Inverter-Related PM Activities
from Aug 2017 to Jan 2019
Site is ~9 years old.
Potential data gap?

Machine Learning 101

Unsupervised

Algorithm groups "like" records

We have to impose meaning/relevance

Supervised

Train algorithm with "correct" answer

Test algorithm with "new" data

Machine Learning 101

Unsupervised

Algorithm groups "like" records

We have to impose meaning/relevance

Supervised

Train algorithm with "correct" answer

Test algorithm with "new" data

Machine Learning 101

Unsupervised

Algorithm groups "like" records

We have to impose meaning/relevance

Supervised

Train algorithm with "correct" answer

Test algorithm with "new" data

Unsupervised Technique

Machine Learning 101

Unsupervised

Algorithm groups "like" records

We have to impose meaning/relevance

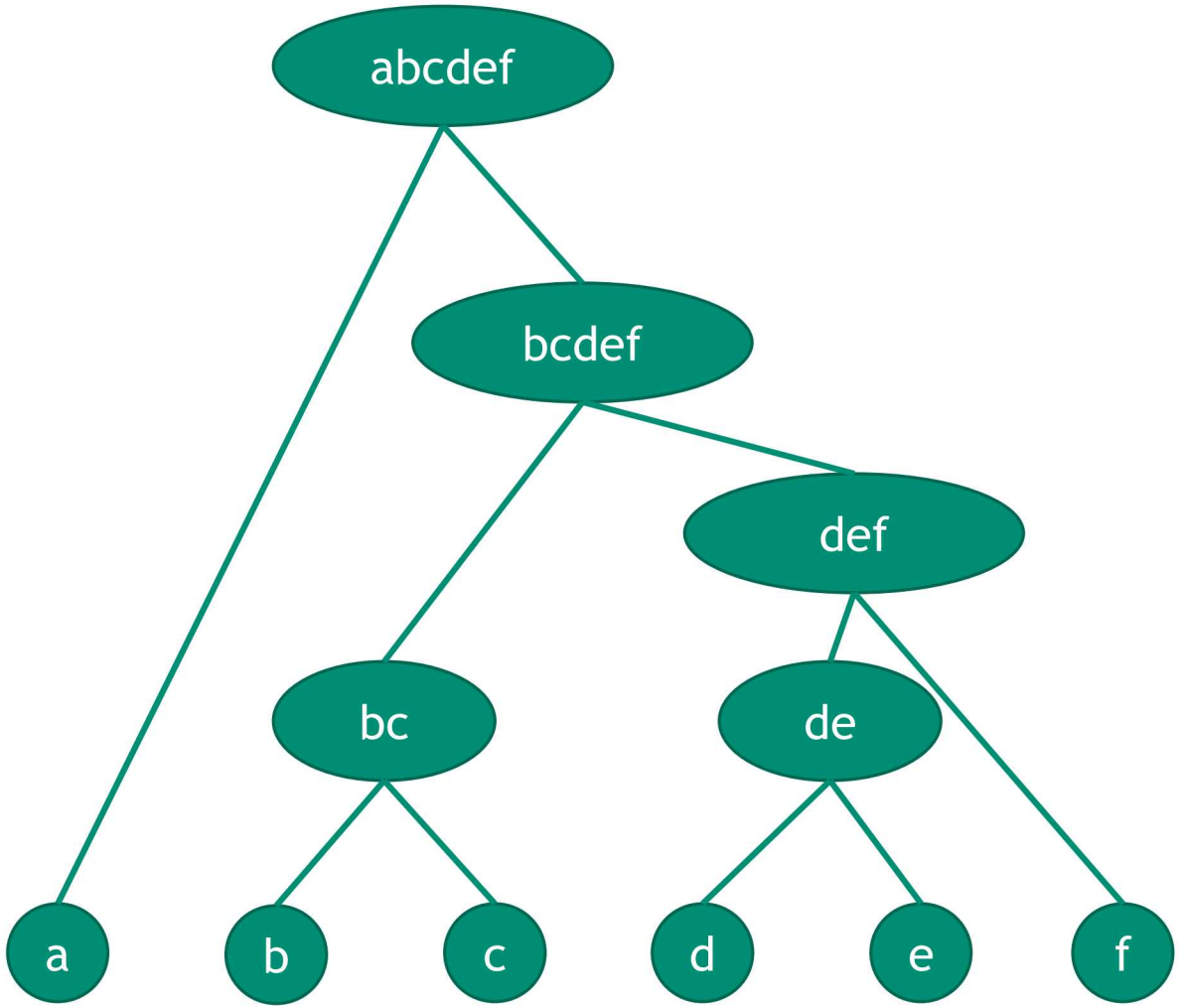
Supervised

Train algorithm with "correct" answer

Test algorithm with "new" data

Supervised Technique

Unsupervised: Hierarchical Clustering



Can either start at the bottom (agglomerative) or top (divisive)

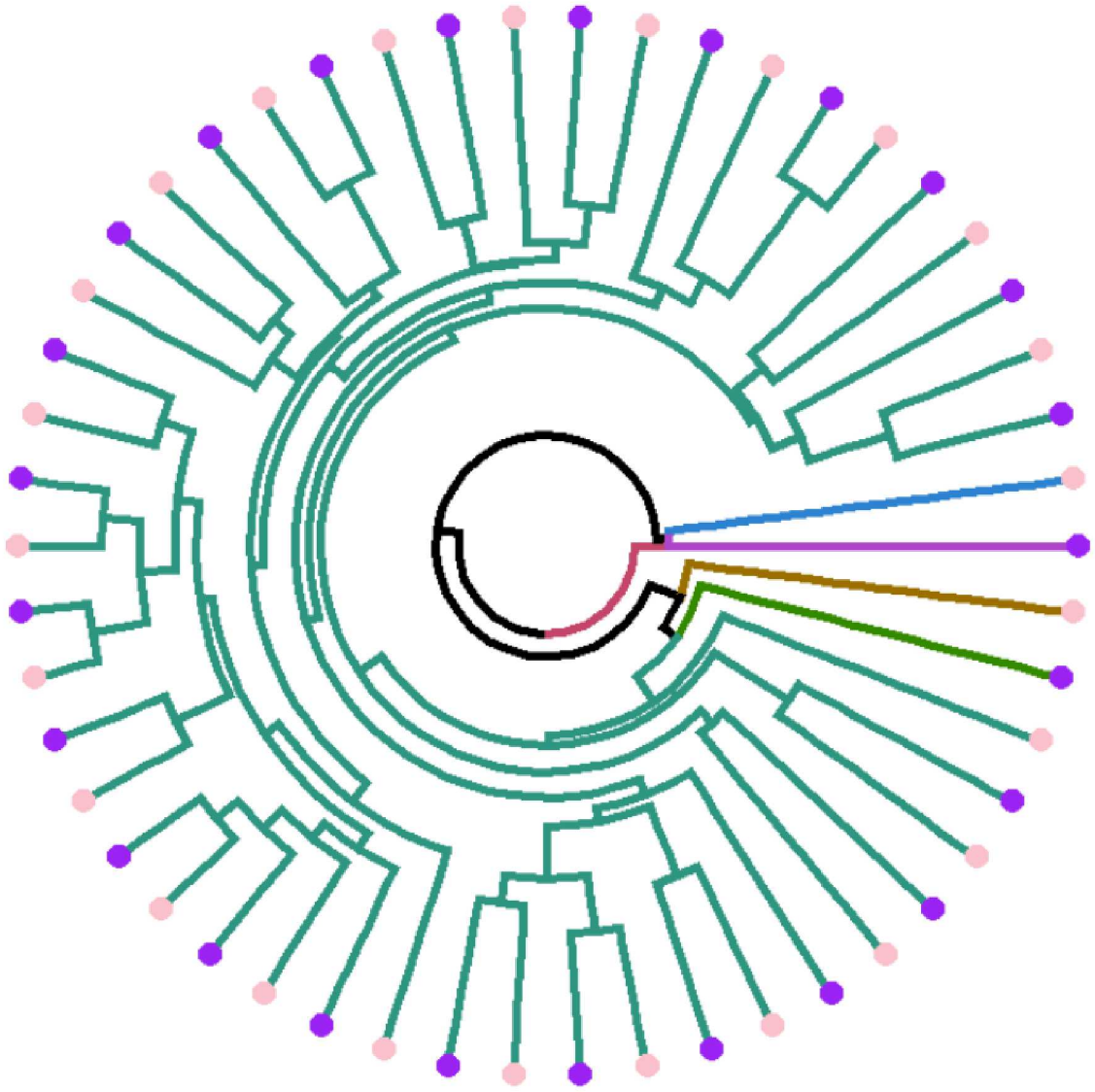
Distances can be calculated in multiple ways

- ***Euclidean***
- Manhattan
- Maximum

Proximity can also be calculated in multiple ways

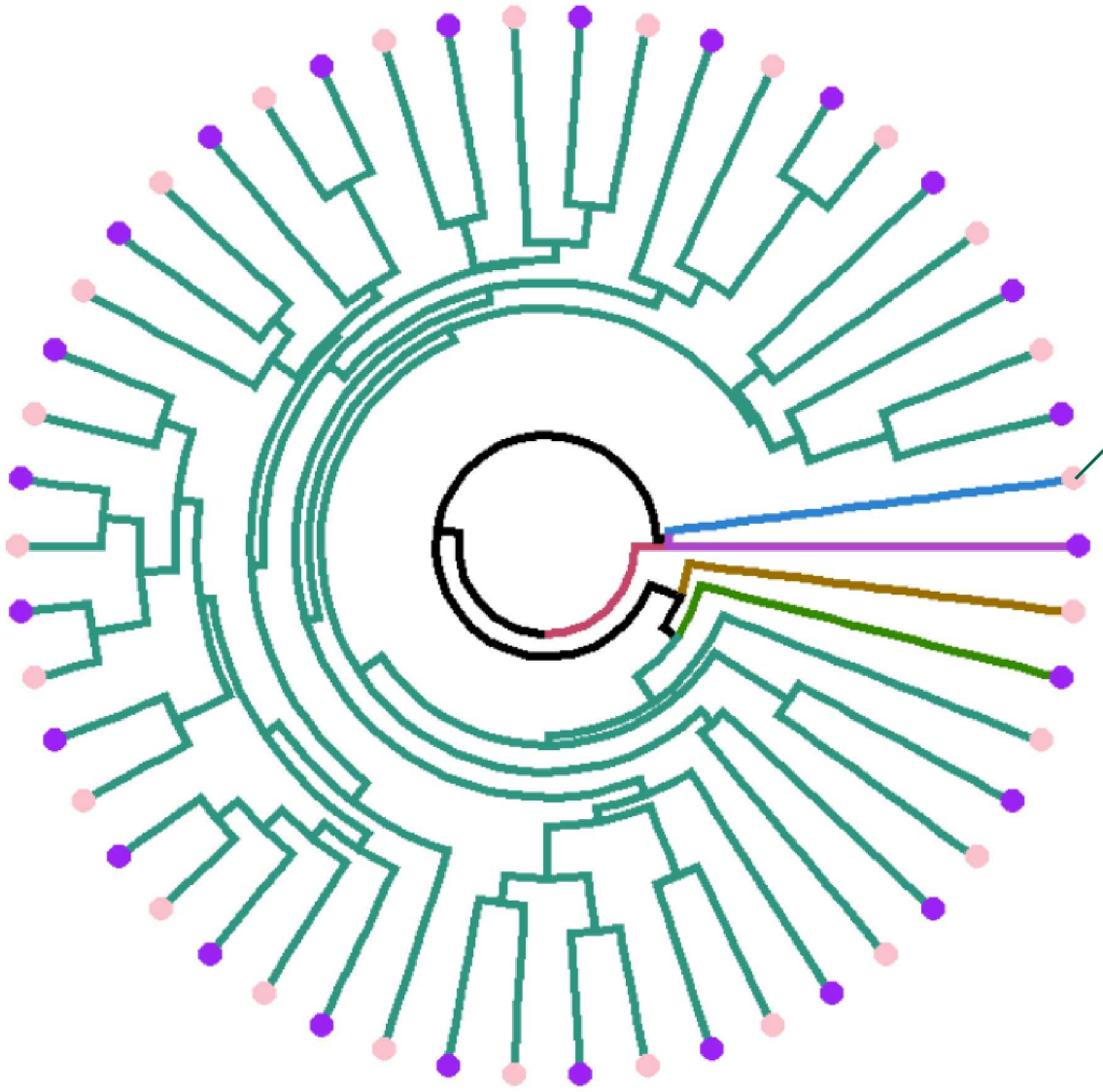
- Complete
- Minimum
- Mean pairwise
- Centroid
- Ward

Implementation of Hierarchical Clustering



Example of Euclidean distance and complete proximity

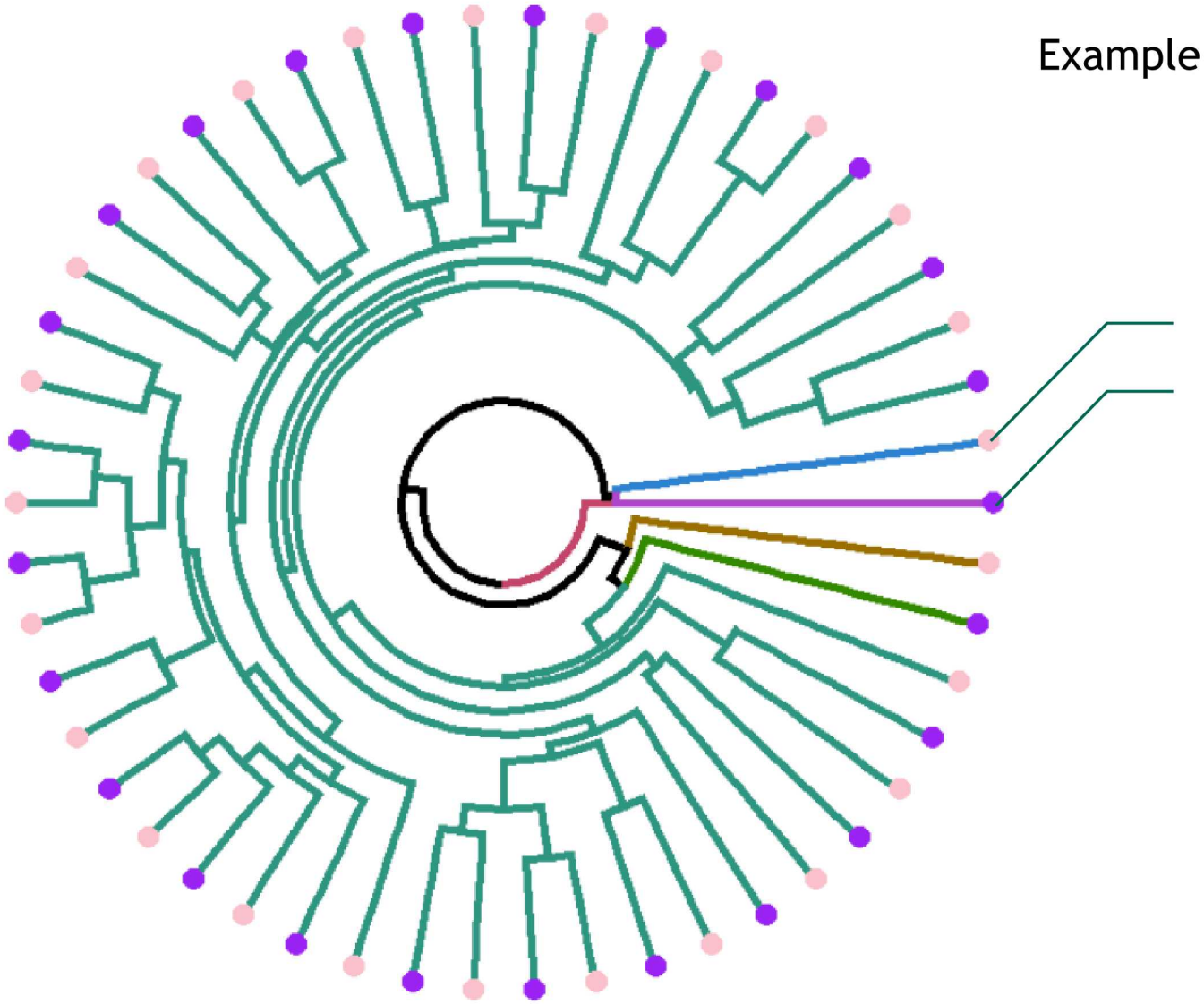
Implementation of Hierarchical Clustering



Example of Euclidean distance and complete proximity

Furthest south of all sites

Implementation of Hierarchical Clustering

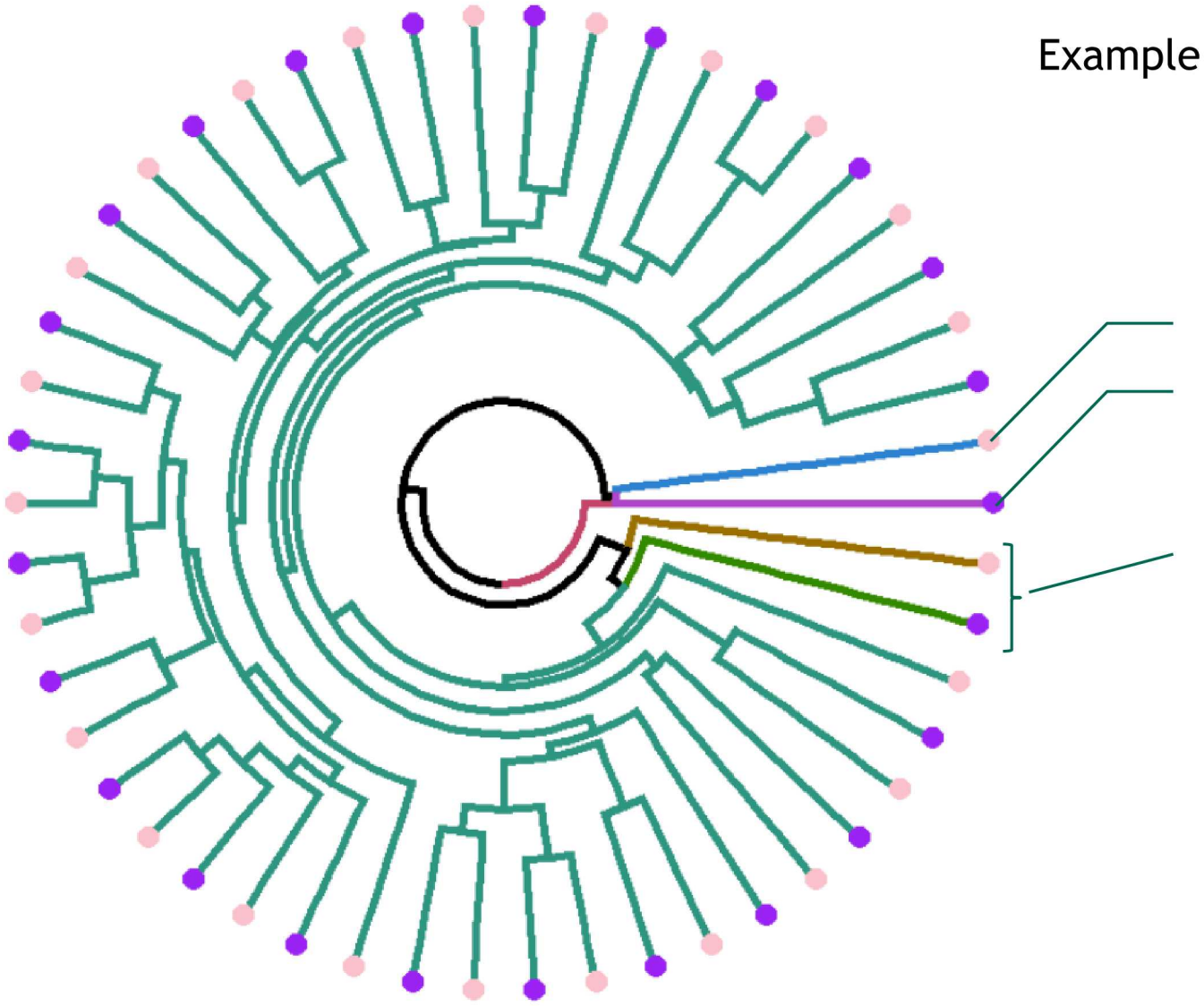


Example of Euclidean distance and complete proximity

Furthest south of all sites

Largest of all sites, by 100s of acres

Implementation of Hierarchical Clustering



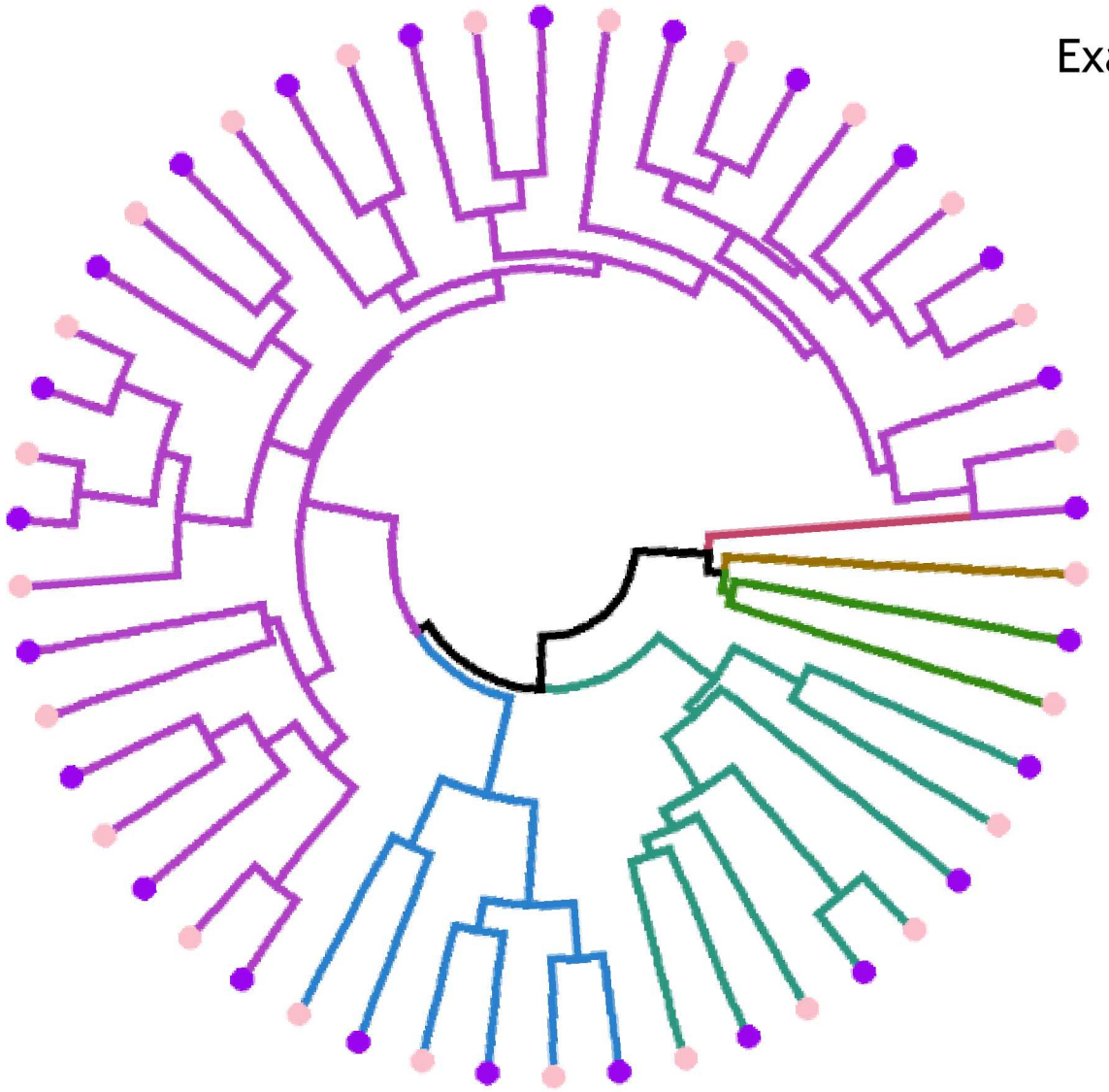
Example of Euclidean distance and complete proximity

Furthest south of all sites

Largest of all sites, by 100s of acres

Climatological region: Dry/Arid

Implementation of Hierarchical Clustering



Example of Squared Euclidean distance and Ward proximity

Still see the distinctive parsing of the 4 individual sites

But larger grouping is now parsed into three distinctive groups

Whether this breakdown is significant, depends on subsequent analysis to understand underlying interpretations of clusters

Supervised Techniques: Text Data

Problem: Not all records were tagged with relevant equipment

Description	Equipment
CHECKED FOR DEAD STRINGS, FUSES, TESTED INLINE...	COMBINER BOXES
TROUBLESHOOTING - Changed the pressure switch ...	TRACKER
COMPLETED ALL SKIDS ABOARD WILDWOOD I, PICTURE...	INVERTERS
Noticed trackers 12 and 18 were out of alignme...	nan

Opportunity: Can convert text data to a vector representation for use in Machine Learning algorithms

- Vectors of word frequency (ngrams), or
- Vectors of word relatedness (word2vec)

Overview of TF-IDF

Term Frequency: ratio of each word to total words in a "document"

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

Inverse Data Frequency: weights of words based on relative frequency (rare words have higher IDF)

$$idf(w) = \log\left(\frac{N}{df_i}\right)$$

TF-IDF: Score weighting words in a "document" with relative frequency of word presence in overall "corpus"

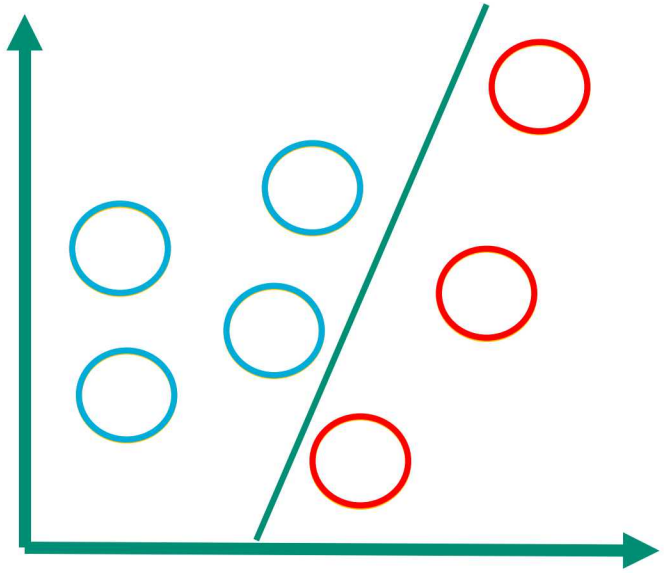
$$w_{i,j} = tf_{i,j} \times idf(w)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

TF-IDF & SVM

Inverter	Fault	…	Troubleshoot		Equipment
0	0.05	…	0.35		Tracker
0	0	…	0.1	~	Combiner
0.24	0.11	…	0		Inverter
0.13	0.09	…	0		Inverter

Support vector machine algorithm, which aims to distinguish between classes



Implementation of TF-IDF & SVM

Trained algorithm is able to effectively predict 76% of test set → indicative of "consistency" within data

Identified some data quality issues that could present opportunities for improvement/training:

-
-
-

Predicted_Equipment	Listed_Equipment
combinerboxes	combinerboxes
trgear	tracker
combiner	panel
inverter	inverter
inverter	inverter
dconnect	dconnect
comm	instrum
panel	array
comm	comm

Implementation of TF-IDF & SVM

Trained algorithm is able to effectively predict 76% of test set → indicative of "consistency" within data

Identified some data quality issues that could present opportunities for improvement/training:

- Original tags were recorded incorrectly
-
-

Predicted_Equipment	Listed_Equipment
combinerboxes	combinerboxes
trgear	tracker
combiner	panel
inverter	inverter
inverter	inverter
dconnect	dconnect
comm	instrum
panel	array
comm	comm

"ta not tracking properly. bad pulse counter. waiting on new pulse counter"

"recycled power to scada enclosure, inverters comms came back up"

Implementation of TF-IDF & SVM

Trained algorithm is able to effectively predict 76% of test set → indicative of "consistency" within data

Identified some data quality issues that could present opportunities for improvement/training:

- Original tags were recorded incorrectly
- Original descriptions were incomplete
-

Predicted_Equipment	Listed_Equipment
combinerboxes	combinerboxes
trgear	tracker
combiner	panel
inverter	inverter
inverter	inverter
dconnect	dconnect
comm	instrum
panel	array
comm	comm

"ta not tracking properly. bad pulse counter. waiting on new pulse counter"

"nan"

"recycled power to scada enclosure, inverters comms came back up"

Implementation of TF-IDF & SVM

Trained algorithm is able to effectively predict 76% of test set → indicative of "consistency" within data

Identified some data quality issues that could present opportunities for improvement/training:

- Original tags were recorded incorrectly
- Original descriptions were incomplete
- Inconsistencies in tags used

Predicted_Equipment	Listed_Equipment
combinerboxes	combinerboxes
trgear	tracker
combiner	panel
inverter	inverter
inverter	inverter
dconnect	dconnect
comm	instrum
panel	array
comm	comm

"ta not tracking properly. bad pulse counter. waiting on new pulse counter"

"nan"

"recycled power to scada enclosure, inverters comms came back up"

"locate defective modules on tracker #3.12/13/18 - decided to use new drone procedure for a faster and more accurate way to locate the defective modules"

Value of Distributions

Another valuable reliability tool

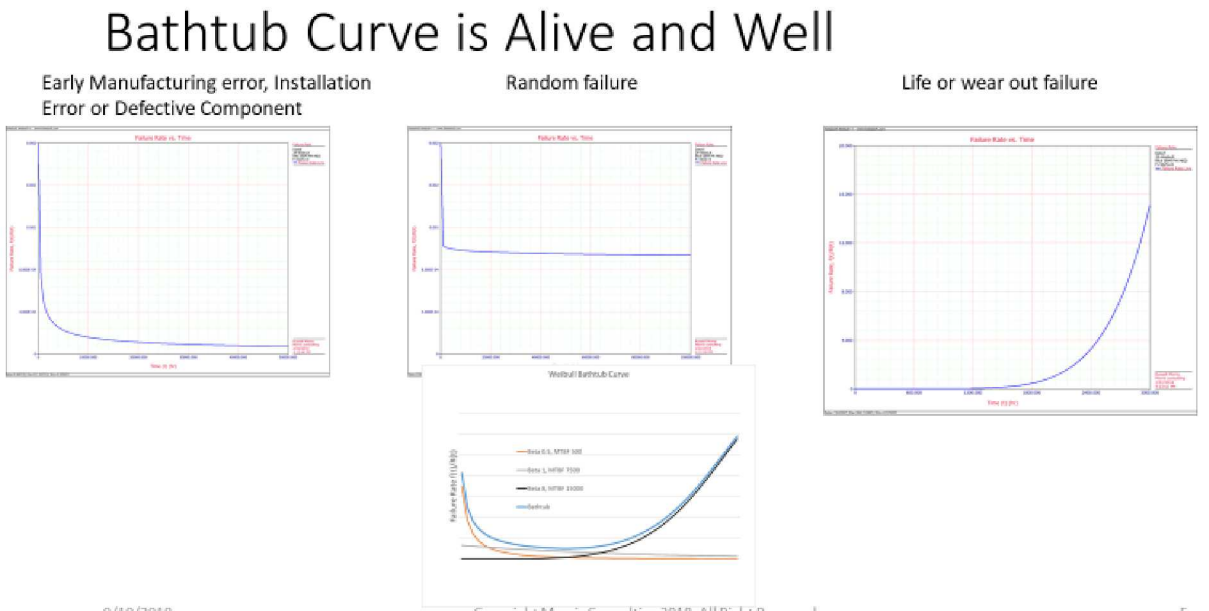
- "When adversity arises, Weibull database distribution is everyone's prized possession with proprietary info" (Barringer)

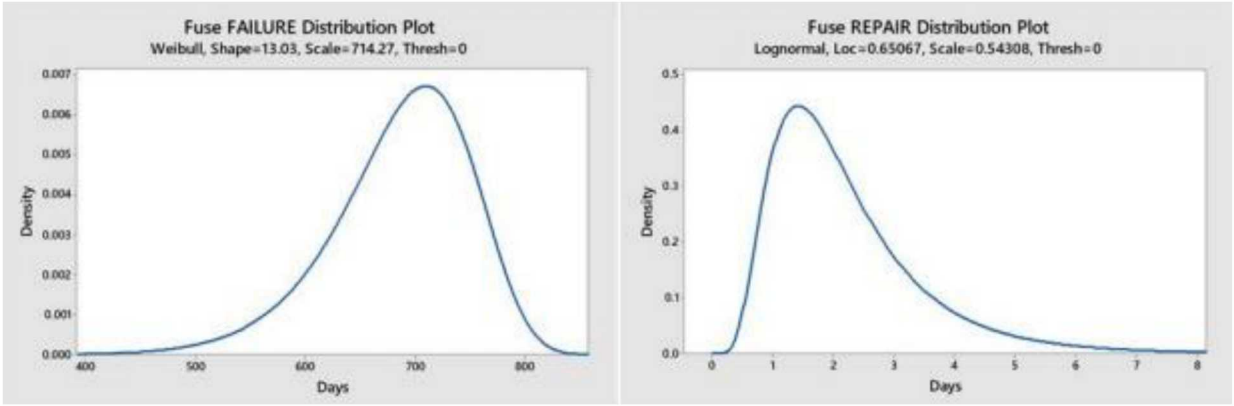
Characterize systemic issues in performance. E.g.,

- Infant mortality vs random
- Lags between failures and response rates

Inform performance models

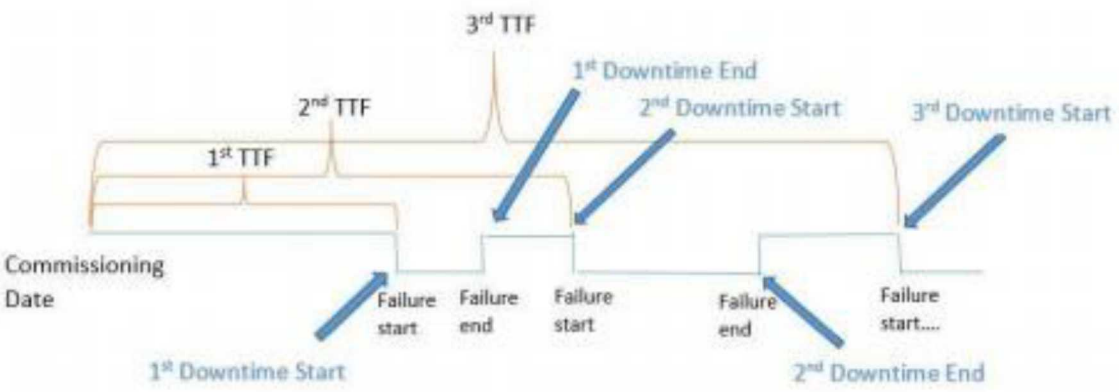
Inputs for O&M cost modeling



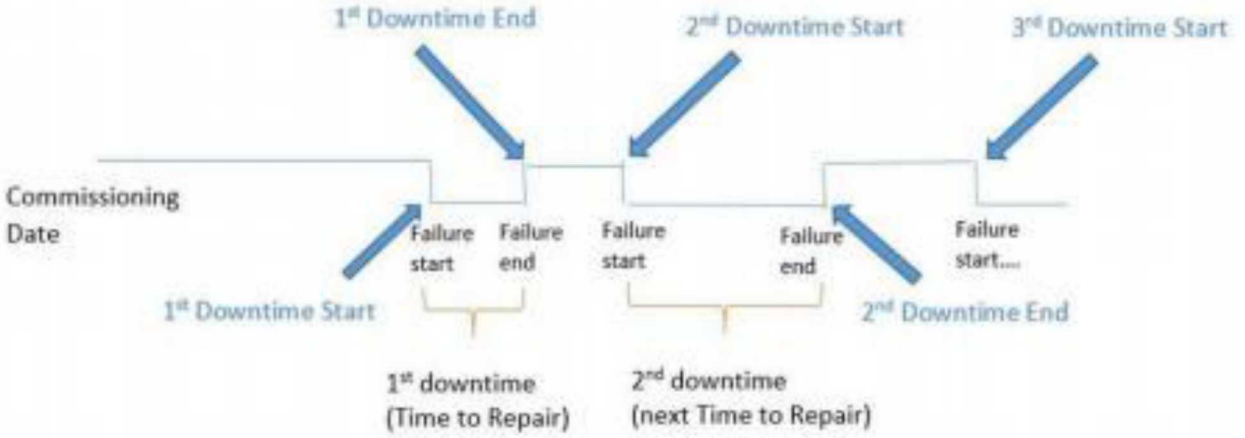


Time to Failures & Repairs

Time to Failure



Time to Repair



Event	Inverter Commissioning Date	Downtime Start	Downtime End	TTF (days) = Downtime Start – Commissioning Date	TTR(days) = Downtime End – Downtime Start
Fan failure	6/15/2016 0:00	6/30/2016 14:05	7/1/2016 23:59	=6/30/2016 14:05 - 6/15/2016 0:00 = 15.586	= 7/1/2016 23:59 - 6/30/2016 14:05 = 1.412
Fan failure		7/13/2016 13:15	7/13/2016 15:05	=7/13/2016 13:15 - 6/15/2016 0:00 = 28.552	= 7/13/2016 15:05 - 7/13/2016 13:15 = 0.076

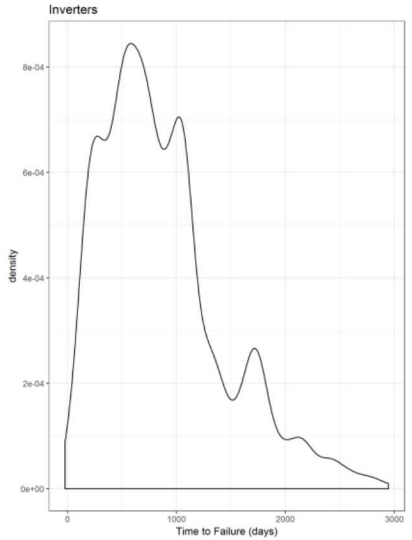
Distribution Fits

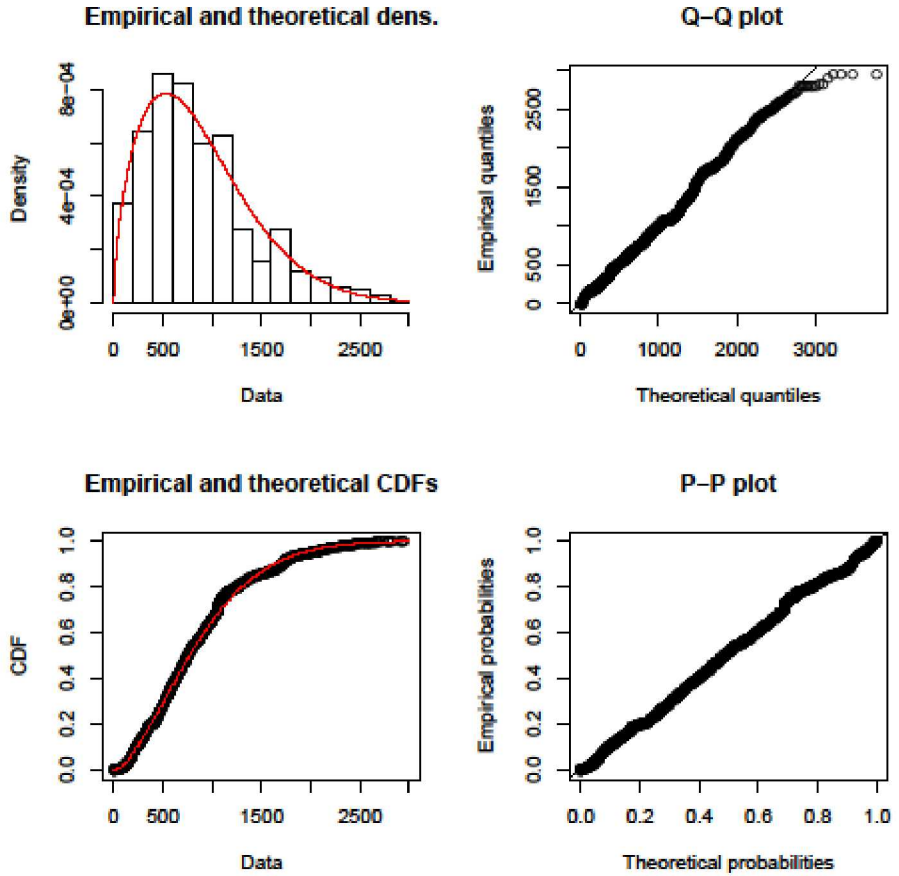
Objective: develop a cumulative distribution plot of TTF or TTR

Various programs exist for fitting different distributions

Goodness of Fit Tests

- Visual inspection
- Anderson-darling (AD) statistics (<1)
- P-value (<0.05)





Inverter TTF data follow a Weibull distribution

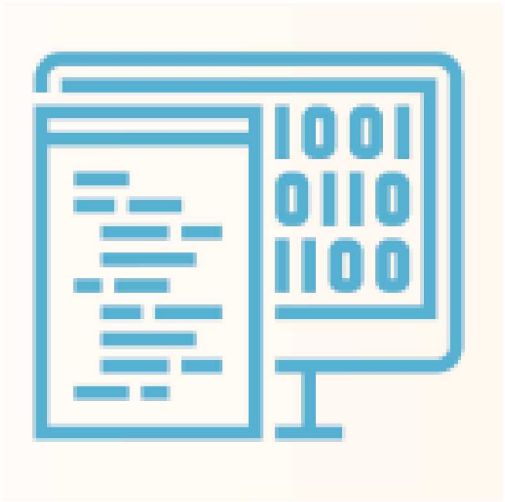
Key Takeaways

A number of techniques exist to leverage insights from O&M data

- Can be implemented for diverse datasets (both text and numerical)
- Can be used to extract different types of patterns (including training opportunities)

Standardization of data collection and associated language is critical

- "Best data is your own" but gaps and inconsistencies are inevitable
- Large amounts of data are required to parse noise from signal



Thank you for your time!

Thushara Gunda

tgunda@sandia.gov