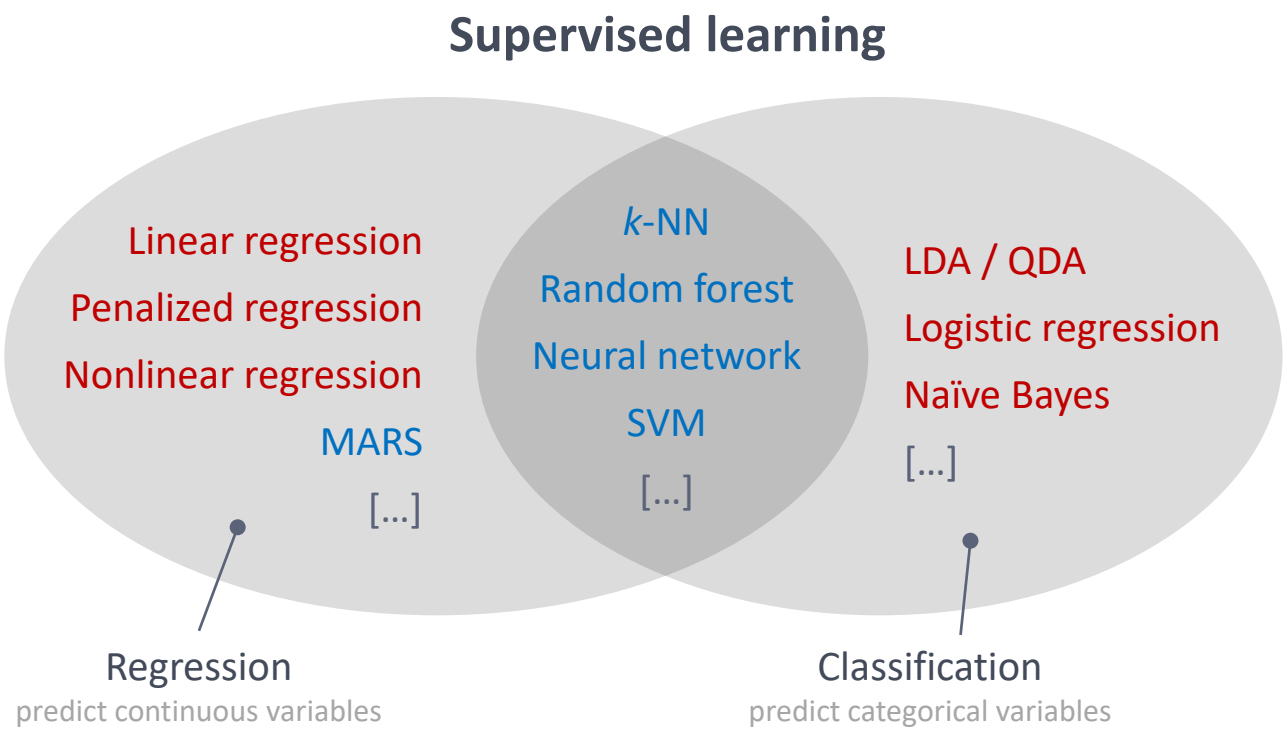
Case Studies of Machine Learning Techniques Applied to Cost Research

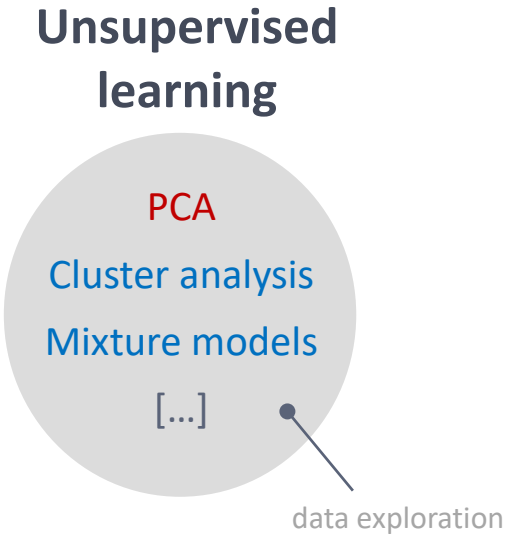
Michael Schiavoni

ICEAA SoCal Chapter Workshop

March 20, 2019

A Top-level View of Machine Learning





Legend

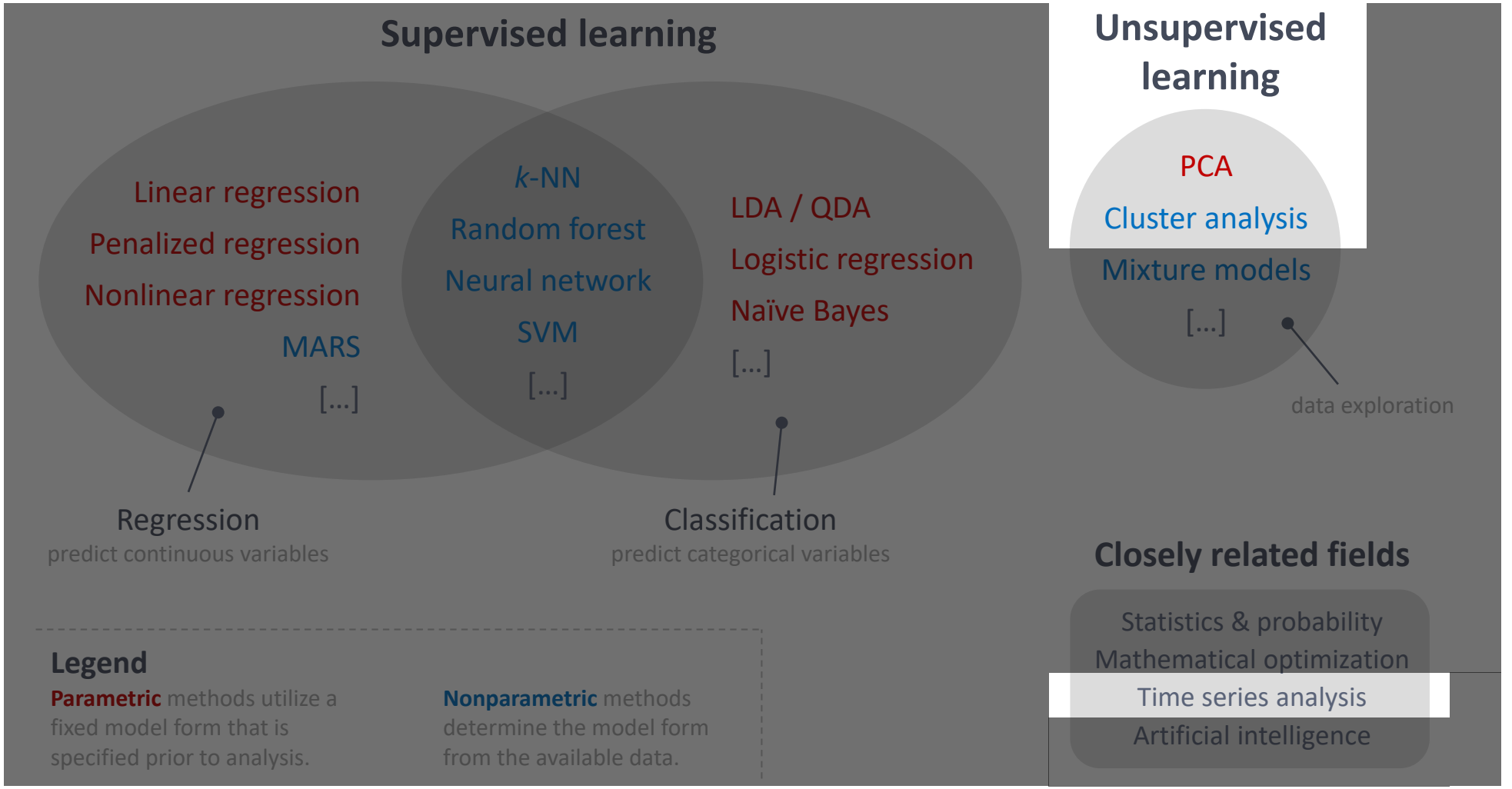
Parametric methods utilize a fixed model form that is specified prior to analysis.

Nonparametric methods determine the model form from the available data.

Closely related fields

Statistics & probability
Mathematical optimization
Time series analysis
Artificial intelligence

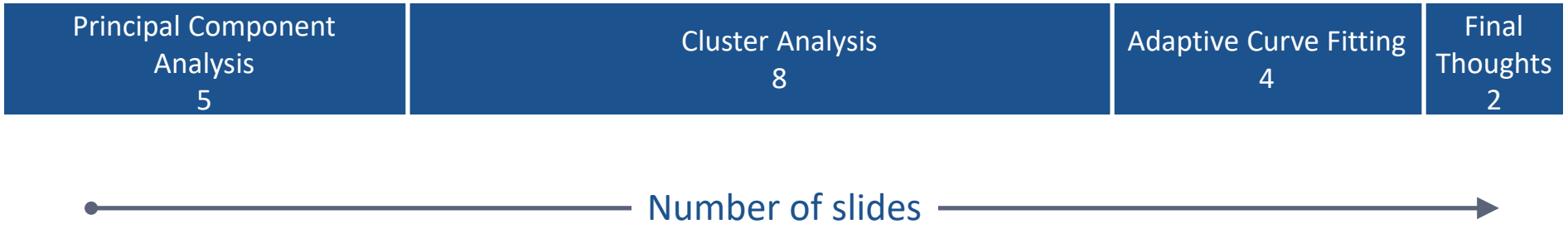
We will highlight two unsupervised methods and a custom time series forecasting algorithm



Outline

Goals:

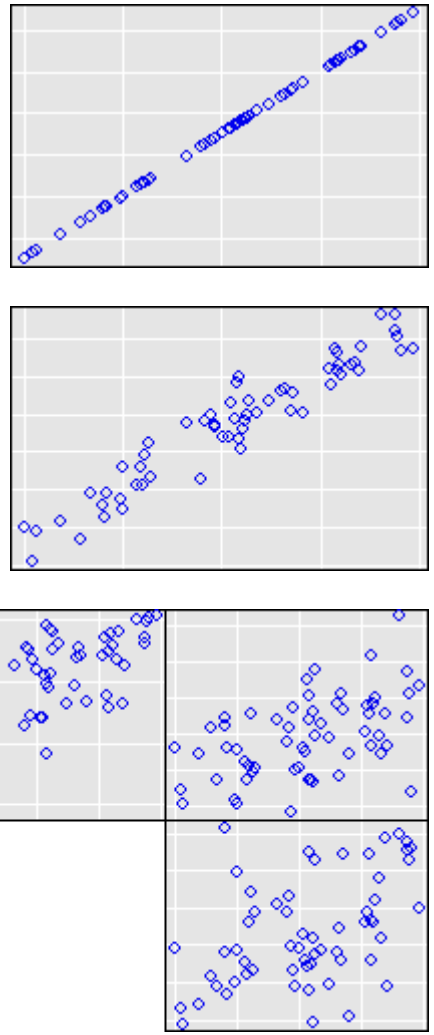
1) To introduce two unsupervised learning methods
2) To briefly highlight real-world use cases
3) To hype my presentation at the ICEAA national conference in Tampa ☺



Principal Component Analysis (PCA)

Background

- Imagine you are performing regression or some other analysis on a dataset
  - You would not include a variable twice with different units in your investigation (e.g. weight in lb. *and* weight in kg)
    - They provide the same exact information (correlation = 1.0)
  - What if there are two different variables that are highly correlated (ρ ≈ 0.9)?
    - Depending on the analysis, you could probably just discard one of them without much detriment
  - What if there are several variables that are all moderately correlated (ρ ≈ 0.5 – 0.8)?
    - They contain some duplicate information, but you might not want to discard any one of them

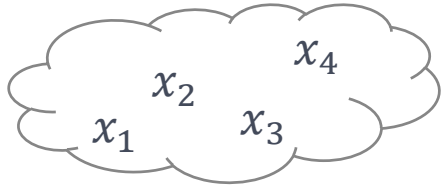


PCA can be useful in this last scenario

What it is

- PCA utilizes matrix algebra to generate uncorrelated linear combinations of continuous variables

variables of interest:



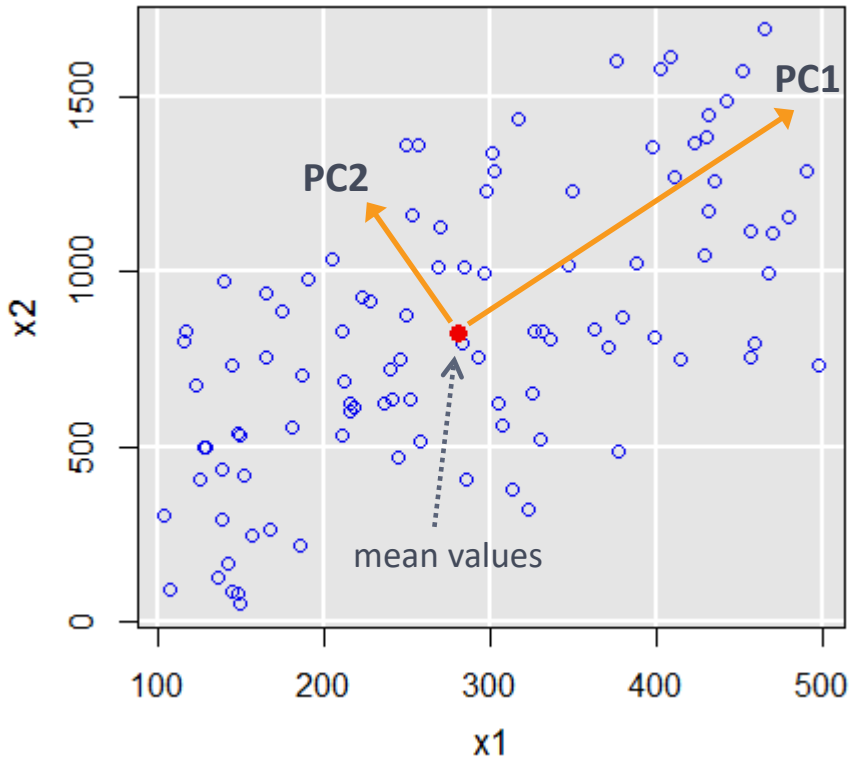
linear combinations:

$$PC1 = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4$$
$$PC2 = b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$$
$$PC3 = c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4$$

- The first principal component (PC1) is the linear combination that explains the maximum amount of variance in the variables
  - PC2 explains the maximum amount of the remaining variance while having zero correlation with PC1
  - PC3 explains the maximum amount of the remaining variance while having zero correlation with PC1 *and* PC2, and so on…
- This is helpful because:
  - It allows for **dimension reduction** (often the first 2 or 3 PC's explain a significant majority of variance)
  - PC's can be used in linear regression to **eliminate multicollinearity** and increase degrees of freedom
  - **PC's can be inputs** for other analyses, e.g. clustering or artificial neural networks

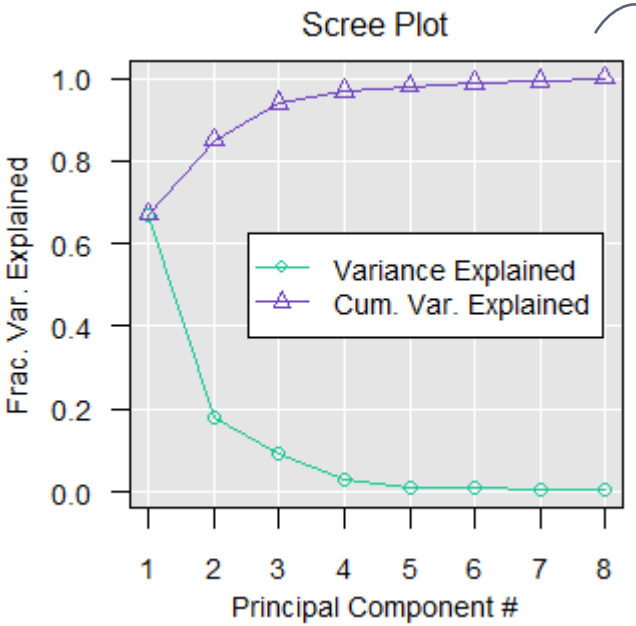
Simplest Case (Two Dimensions)

- The variables to the right display moderate positive correlation
- PC1 is in the direction of correlation
  - It is equivalent to the **orthogonal regression** line (a.k.a. total least squares)
- PC2 is perpendicular to PC1
- This can be extended to many dimensions
  - Typically applied to datasets with 4 – 40 variables
  - Can be performed with >100 variables



PCA Use Case: Acquisition Complexity Score

- Customer wanted to generate complexity scores for its satellite subsystems
  - Purpose: incorporate into existing CERs to improve regression statistics and explain additional variation not captured by the primary drivers
  - 8 complexity-related indicators were identified (e.g. technology readiness level, # of CDRLs, amount of oversight, etc.)



PCA results from one CER

PC #	Cumulative
1	67%
2	85%
3	94%

Outcome

- PC's added as predictors (1, 2, or 3, depending on the CER)
- Some CERs were improved, others were not
- Customer revisited scoring scheme for some of the subjective indicators
- Overall viewed as a successful effort

Cluster Analysis

Background

- Terminology: an observation is one or more measurements pertaining to an item of interest

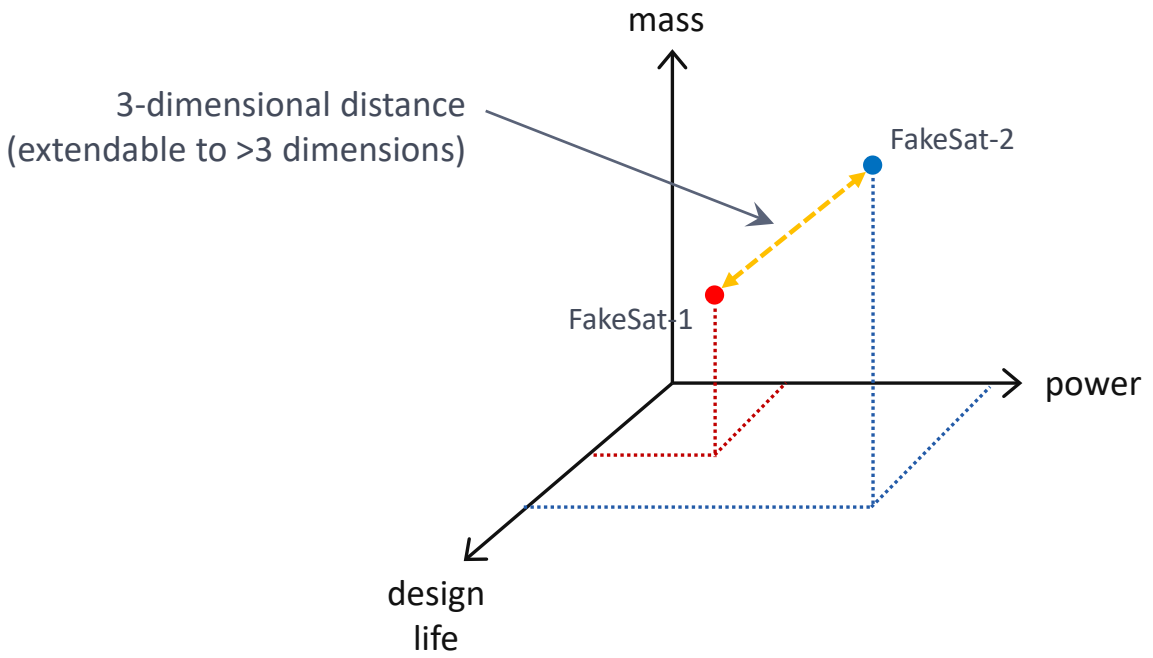
	Mass (lb)	Power (W)	Design Life (mo)	# Payloads
FakeSat-1	2000	620	24	3
FakeSat-2	3752	1740	36	2

Two observations of four variables

- Clustering can be used to divide a group of observations into multiple subgroups (clusters) in an objective manner
  - Observations within the same cluster are more similar to each other than those between different clusters
  - The measure of similarity can differ based on the method and user-defined settings
- Methods
  - *k*-means and *k*-medoids
  - Hierarchical (agglomerative and divisive)
  - Distribution-based
  - Density-based
  - ANN techniques

Similarity Measures

- The most commonly used measure of similarity between observations is the Euclidean distance
  - This is the ordinary straight-line distance in multidimensional space
  - Other similarity measure alternatives also exist



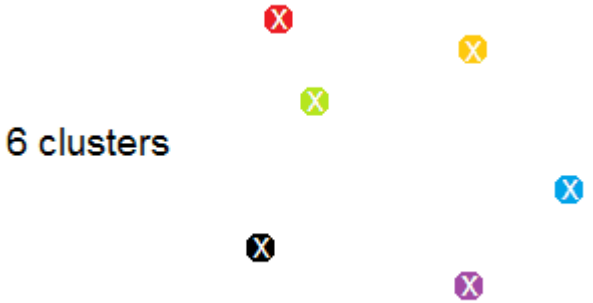
Hierarchical Cluster Analysis: How it works

- Features/limitations:
  - Generates a complete tree diagram (dendrogram) containing all cluster linkages
  - Not subject to randomness; provides same result every time
  - More computationally expensive than *k*-means; not suitable for very large datasets
- Agglomerative (bottom-up) approach: each observation begins as its own cluster
  - Pairs of clusters are repeatedly merged, moving up the hierarchy until there is only one cluster
- Centroid linkage and Ward's Method are two variants (others exist)
  - Centroid linkage: at each step the two clusters with minimum centroid distance metric are merged
  - Ward's: an ANOVA approach, at each step the two clusters are merged that result in the minimum increase in combined SSE

Hierarchical Cluster Analysis: How it works

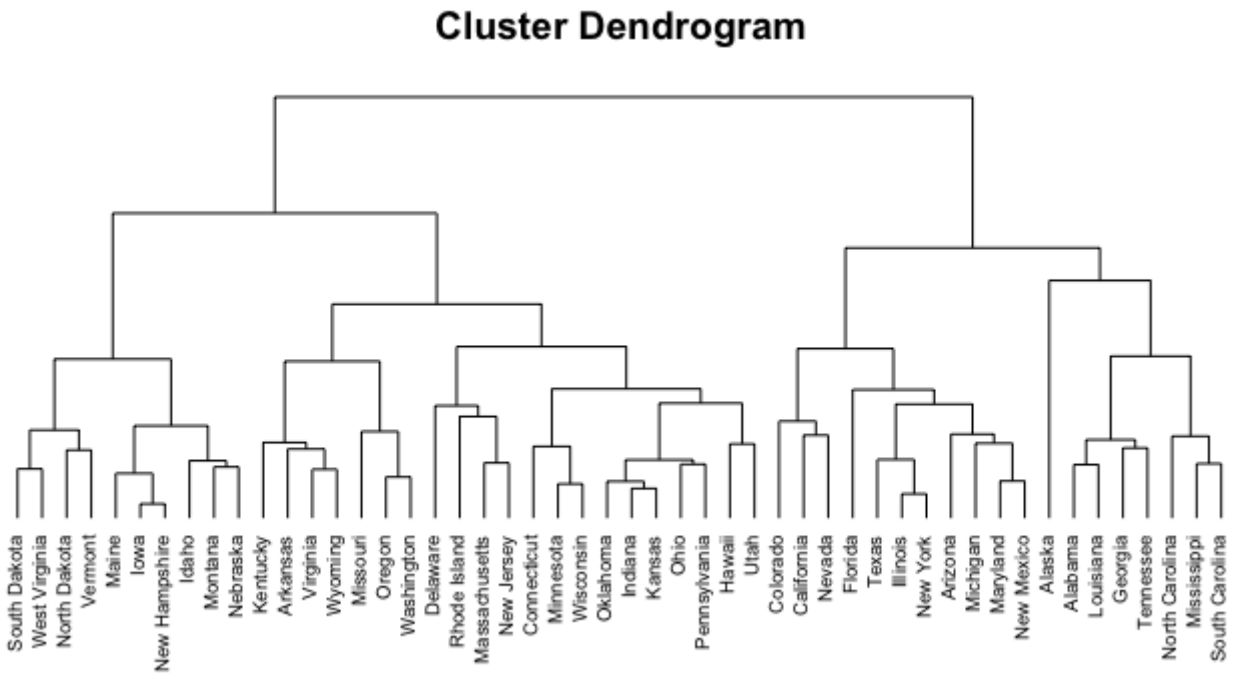
- Features/limitations:
  - Generates a complete tree diagram (dendrogram) containing all cluster linkages
  - Not subject to randomness; provides same result every time
  - More computationally expensive than *k*-means; not suitable for very large datasets
- Agglomerative (bottom-up) approach: each observation begins as its own cluster
  - Pairs of clusters are repeatedly merged, moving up the hierarchy until there is only one cluster
- Centroid linkage and Ward's Method are two variants (others exist)
  - Centroid linkage: at each step the two clusters with minimum centroid distance metric are merged
  - Ward's: an ANOVA approach, at each step the two clusters are merged that result in the minimum increase in combined SSE

Agglomerative centroid linkage in action with six observations:



[view in slideshow mode for animation]

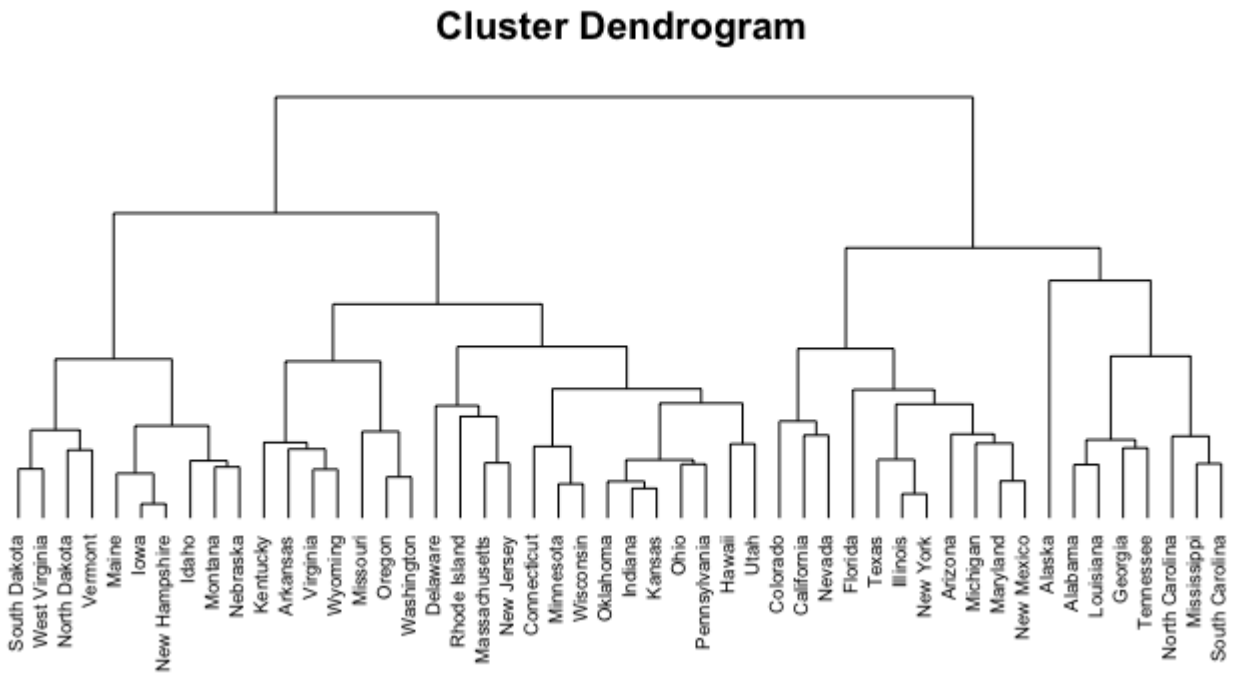
Interpreting a Hierarchical Tree Diagram



By making one continuous cut perpendicular to the branches, the tree is divided into clusters of maximum separation.
By varying the location of the cut, different numbers of clusters are formed.
This allows for exploration of the linkage of the data without having to assume a specific number of clusters.

https://uc-r.github.io/hc_clustering

Interpreting a Hierarchical Tree Diagram



By making one continuous cut perpendicular to the branches, the tree is divided into clusters of maximum separation.
By varying the location of the cut, different numbers of clusters are formed.
This allows for exploration of the linkage of the data without having to assume a specific number of clusters.

[view in slideshow mode for animation]

https://uc-r.github.io/hc_clustering

“The Robuster Cluster Tool” Homegrown App

The Robuster Cluster Tool

Choose file to upload

Browse... clust_demo.csv

Upload complete

Separator

◉ Comma ○ Semicolon ○ Tab

☑ Header ☑ Plot Labels

Cluster on Principal Components

Impute Missing Values

Data | Agglomerative | Divisive | k-means | k-medoids | ANN

Agglomerative hierarchical clustering via Ward's method of minimum variance or Euclidean distance linkage.

Method

◉ Ward's (default)
○ Complete linkage
○ Average linkage

Number of Clusters

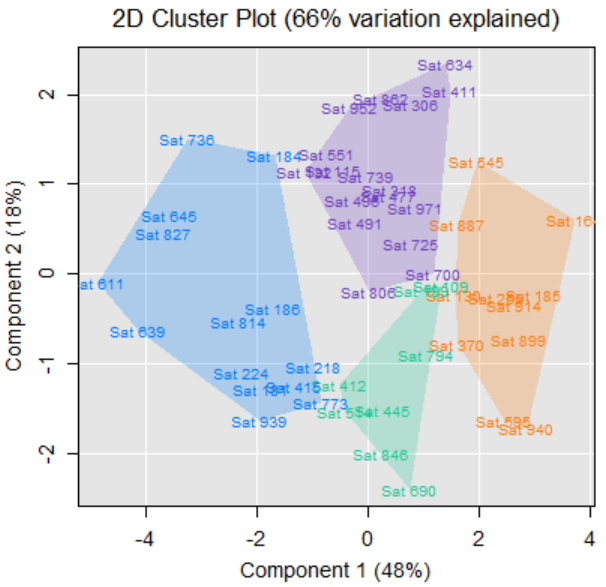
2 [4] 10

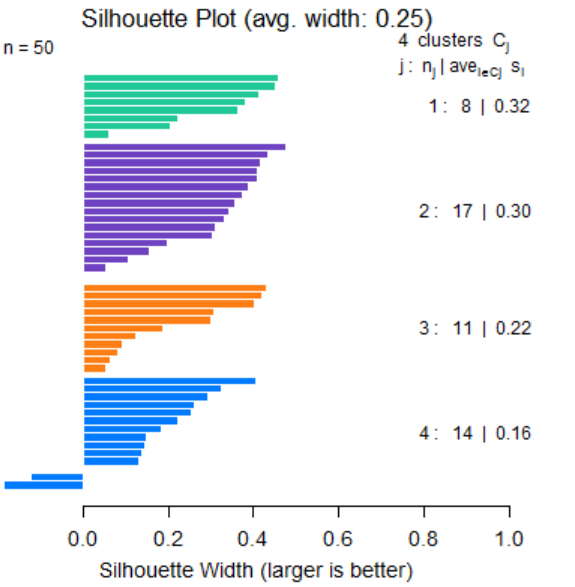
2 3 4 5 6 7 8 9 10

Dendrogram Type

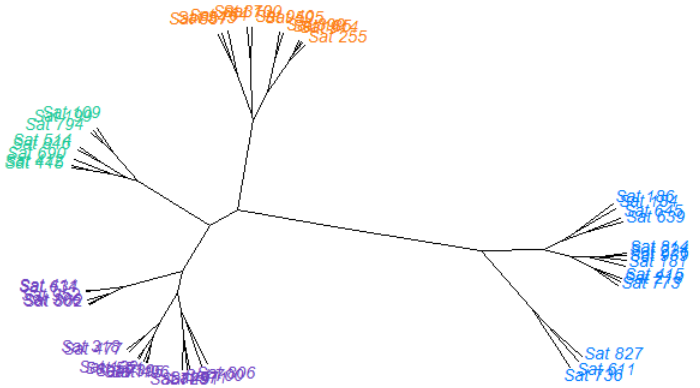
○ Standard
◉ Unrooted

Export Results

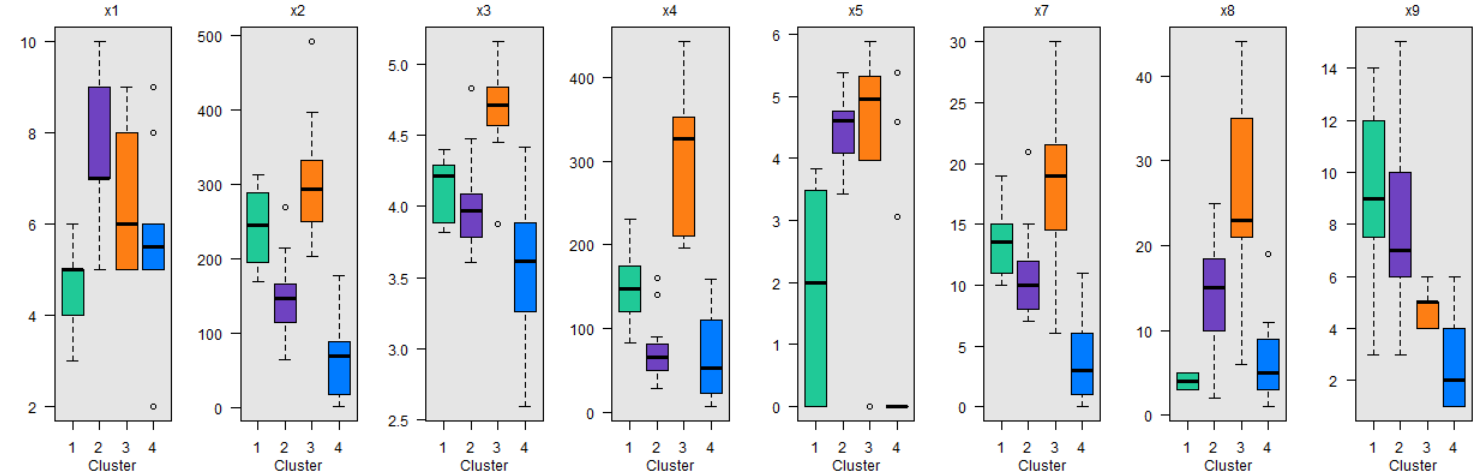




“The Robuster Cluster Tool” Homegrown App



Cluster distributions by variable:



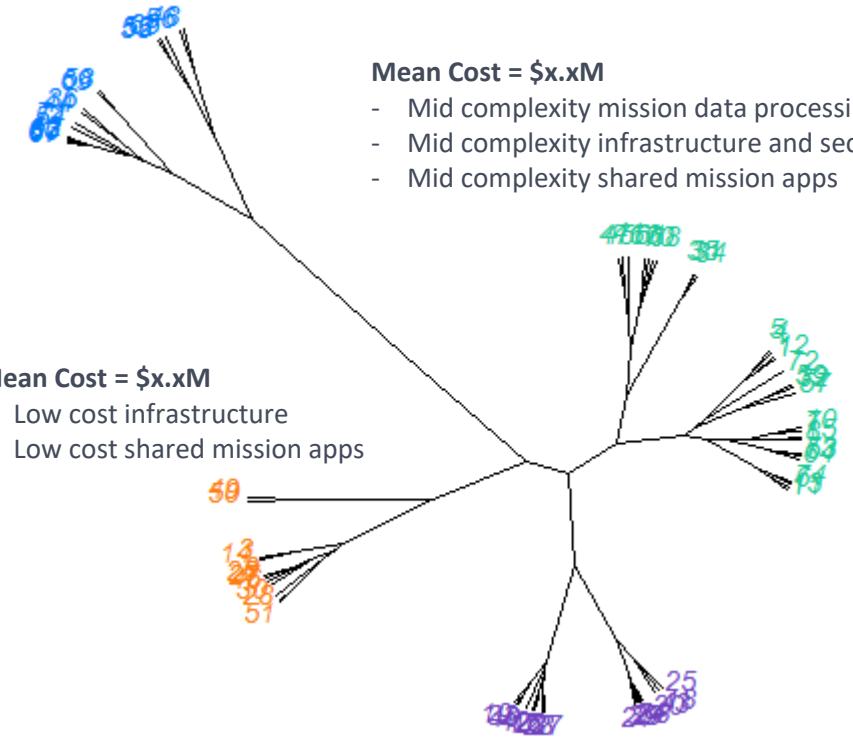
Clustering Use Case: Ground Systems Study

- Hierarchical clustering used to separate 74 CSCIs/apps into groups in an objective and automated fashion
  - Agglomerative clustering using Ward's method provided the best group separation
  - 10 variables used for clustering (legacy DSLOC, # requirements, integration ESLOC, etc.)
- Utility:
  - Segregates disparate apps into meaningful groupings without potential SME bias
  - SME still required to interpret the groupings
  - Allows for quick turn estimates using basic information when little or no technical data is available
  - Now that groups have been defined, one could train a classifier if desired

Mean Cost = $x.xM
- Mission unique software
- Mission mgt/planning/scheduling

Mean Cost = $x.xM
- Mid complexity mission data processing
- Mid complexity infrastructure and security
- Mid complexity shared mission apps

Mean Cost = $x.xM
- Low cost infrastructure
- Low cost shared mission apps



Mean Cost = $x.xM
- High complexity mission data processing apps
- Higher cost infrastructure and security apps

Adaptive Curve Fitting (a teaser)

Adaptive Curve Fitting (ACF): What It Does

Inputs from in-progress program

- monthly expenditures (required)
- range for total cost (optional)
- range for total duration (optional)

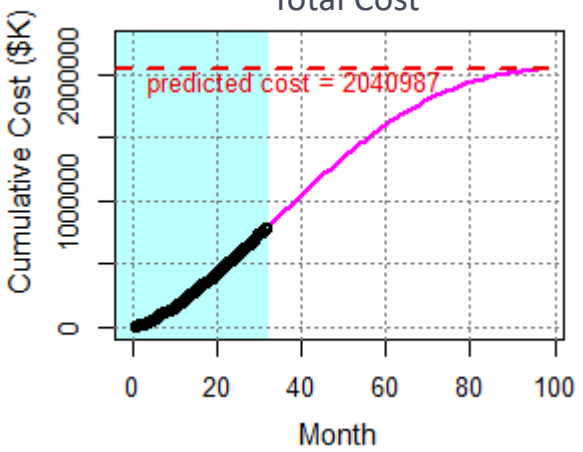
Adaptive Curve-Fitting Algorithm

Employs a variety of mathematical techniques, including:

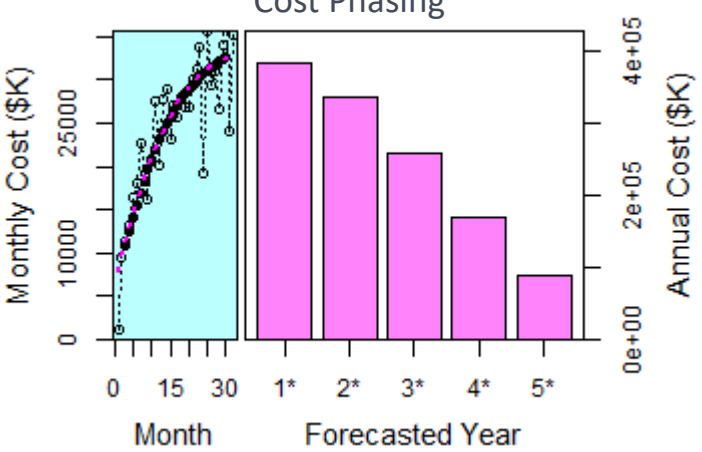
- Smoothing filters
- Calculus-based curve projection
- Nonlinear least-squares optimization
- Rayleigh/Weibull/Beta/Normal distributions

Outputs

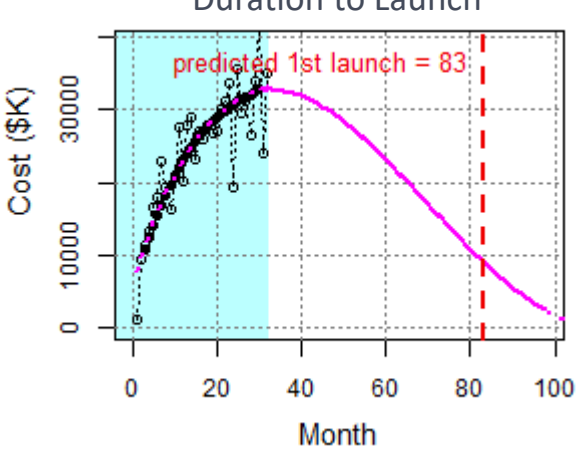
Total Cost



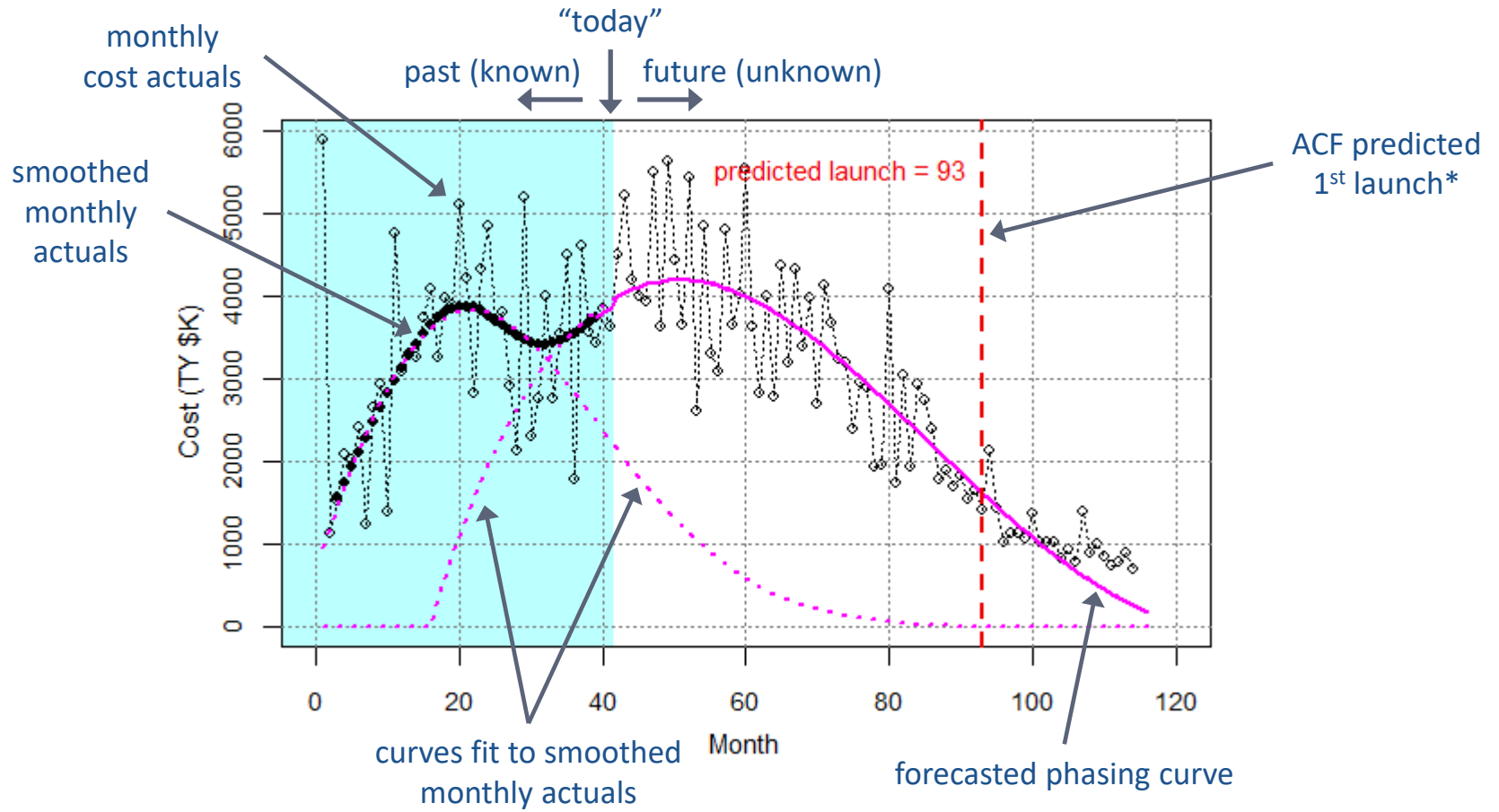
Cost Phasing



Duration to Launch

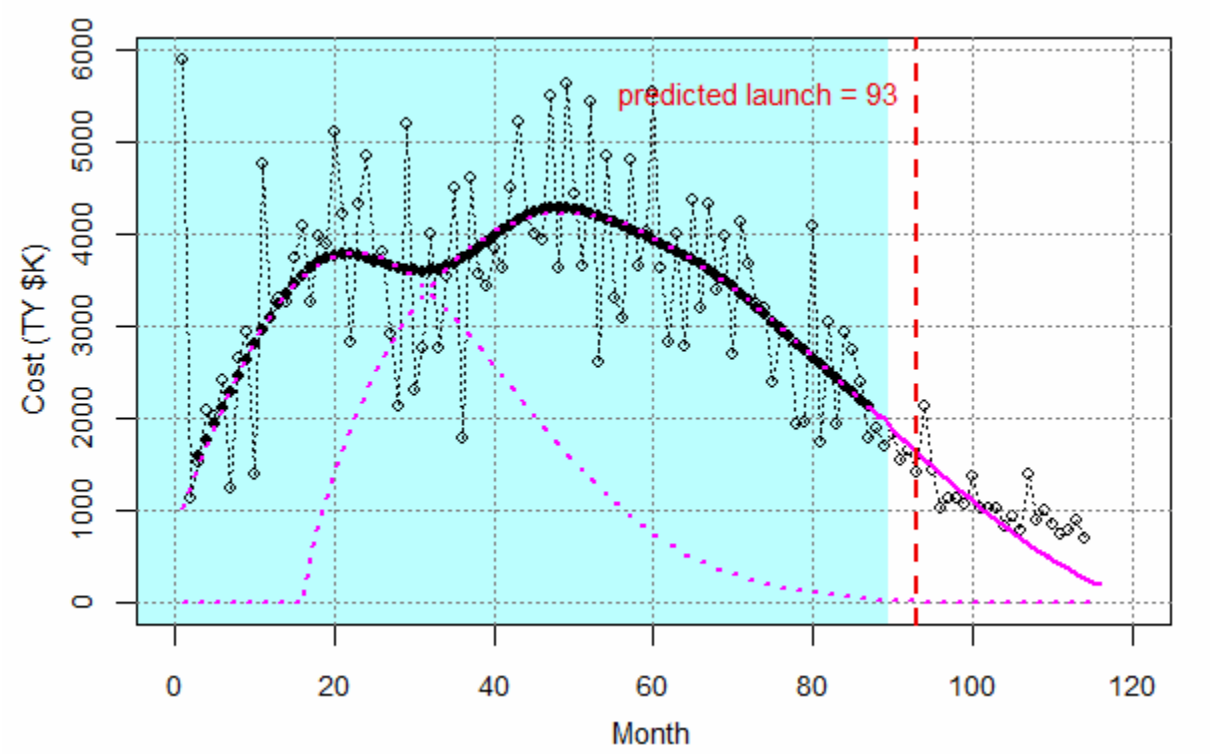


ACF Example



*Launch is predicted by determining the month along the forecasted phasing curve at which a %-spent metric is met.

Animated Example over Time



[view in slide show mode for animation]

Closing Thoughts

- ML is widely used in tech, intel, finance, retail, transportation, healthcare, etc.
- Advanced ML techniques seem to still be gaining steam in the government cost/schedule/EVM/PM communities
- Tecolote is actively infusing nonstandard methodologies into our research and estimating efforts
- Considerations:
  - Each technique has strengths & weaknesses; there is no silver bullet
  - Beware of over-fitting, especially with small samples and nonparametric methods
  - Education is a hurdle; customers might not get excited if they don't understand it
    - Interpretability of nonparametric methods can be a challenge
  - Data quality still matters… "garbage in, garbage out"
  - Different methods will yield different answers to the same problem

Contact:

Michael Schiavoni
mschiavoni@tecolote.com

Ben Kwok
bkwok@tecolote.com

For more information on ACF, please check out the paper/presentation, *Adaptive Curve Fitting: An Algorithm in a Sea of Models*, at the national ICEAA workshop in May 2019.