A Sequential Model Approach to Improve Software Assurance

Industry Paper #150

16 November 2009

Michael Gegick and Pete Rotella

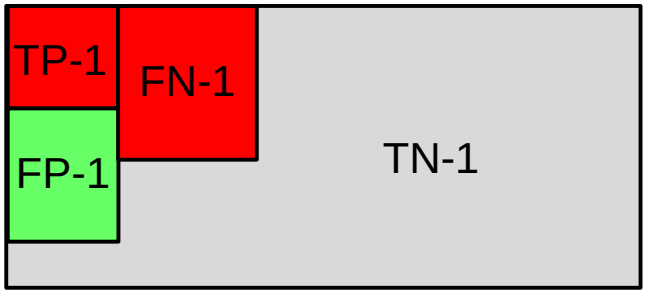
Cisco Systems, Inc.

mcgegick@gmail.com, protella@cisco.com

- Many predictive models have been used to identify fault- and failure-prone software components, however
- Much modeling work results in specifying a single model to identify components most likely to contain faults
- This paper describes an effective way to combine several models to produce results more accurate and useful than if individual models are employed separately.

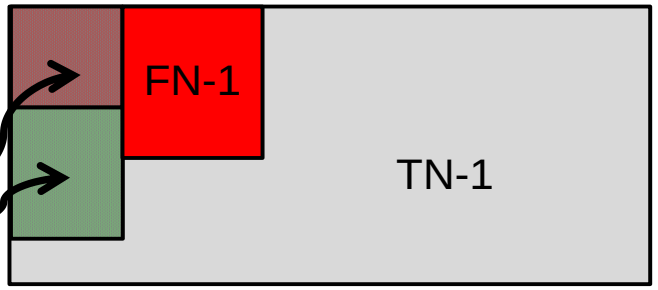
- Sequential models we've developed include:
  - Classification and regression tree (CART) models to predict which software components are most susceptible to attack
  - Sinusoidal and exponential decay models to predict the occurrence and resolution rates of faults in the field
- All models have been developed for large (>10 MLOC) Cisco software systems
- Model results have been validated by Cisco security and reliability teams.

Sequential modeling schematic – example 1:

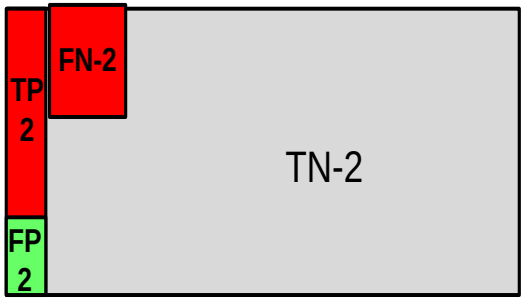


Results of initial model run

Remove TP and FP regions from dataset

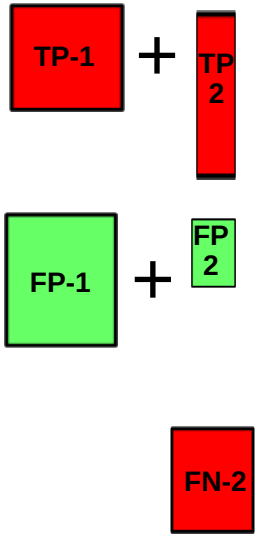


Retrain model and rerun on depleted dataset



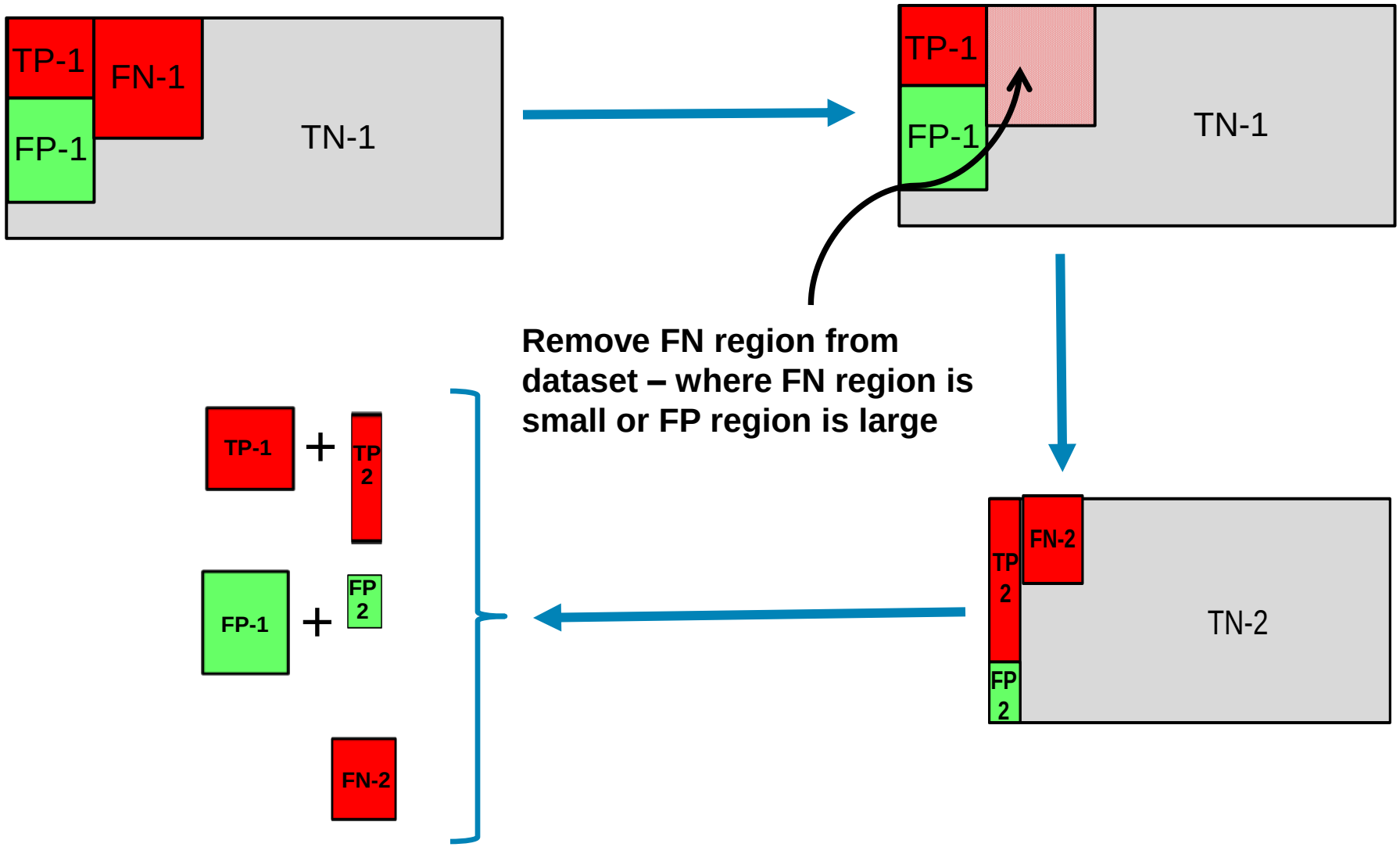
Second (sequential) model run

Combine results:
- large increase of TP
- small increase of FP
- large decrease of FN

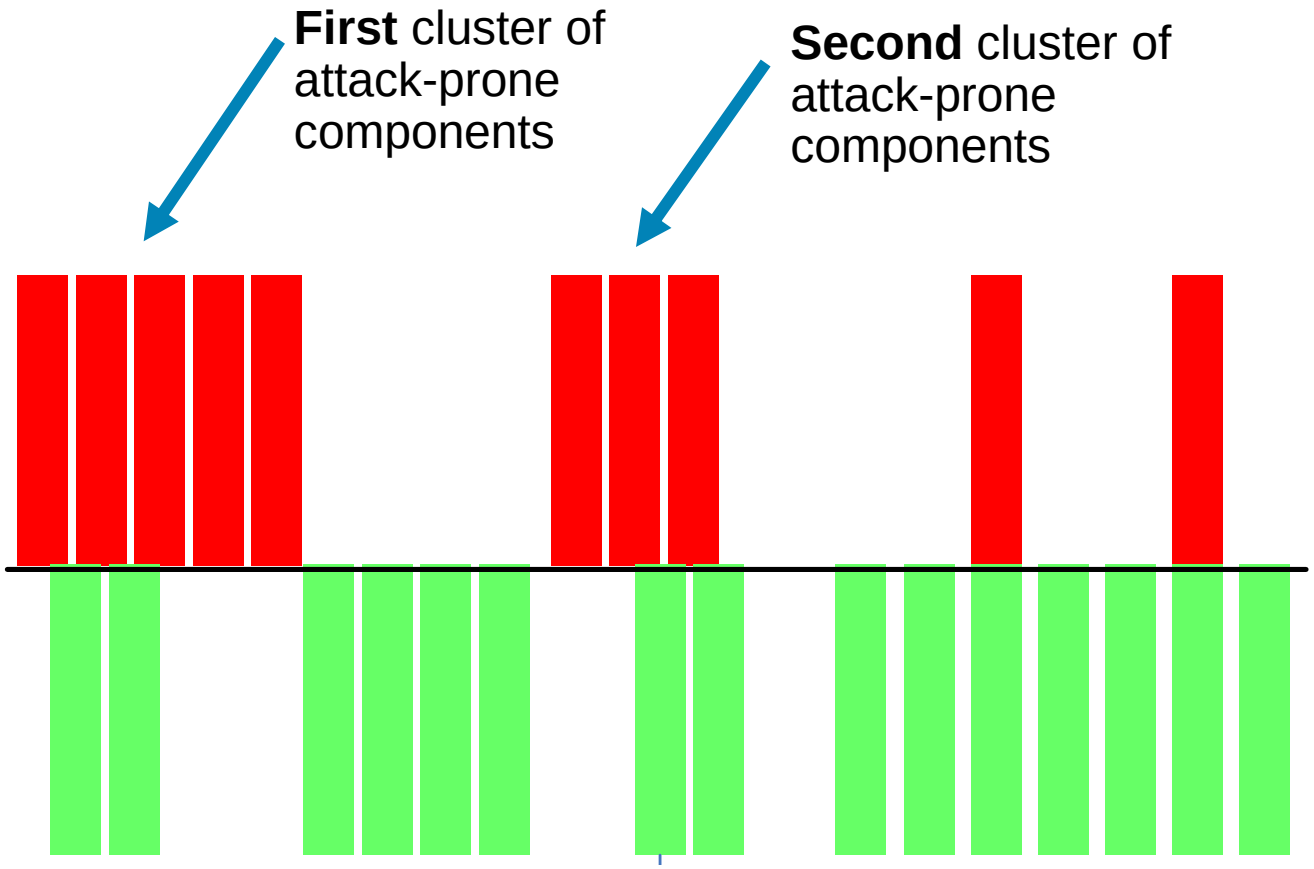


TN (true negatives - correctly classified as not likely)
FN (false negatives- misclassified as not likely)
TP (true positives - correctly classified as likely)
FP (false positives - misclassified as likely)

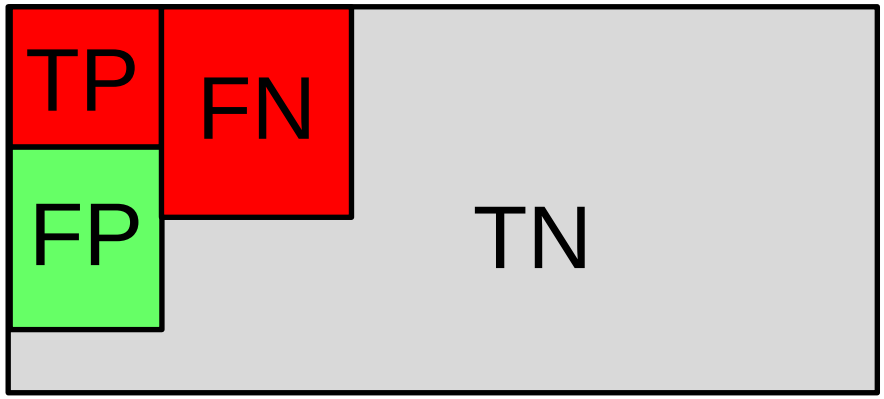
Sequential modeling schematic – example 2:



Component modeling: Initial model run produces two clusters of attack-prone components

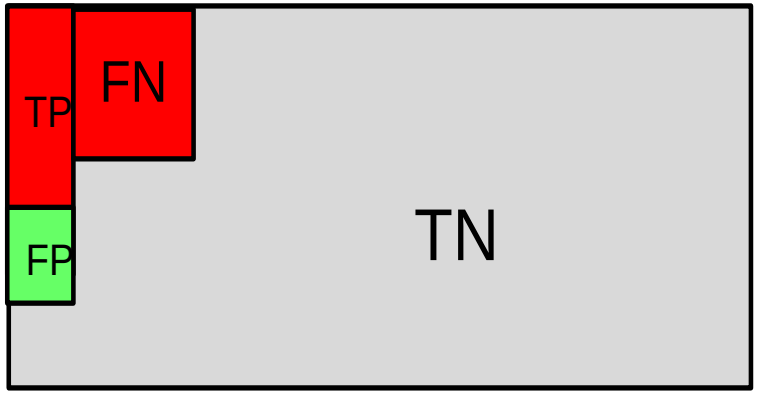


CART model predicts which components are most susceptible to security attack



Results from first run of model

Remove TP+FP; retrain and rerun model



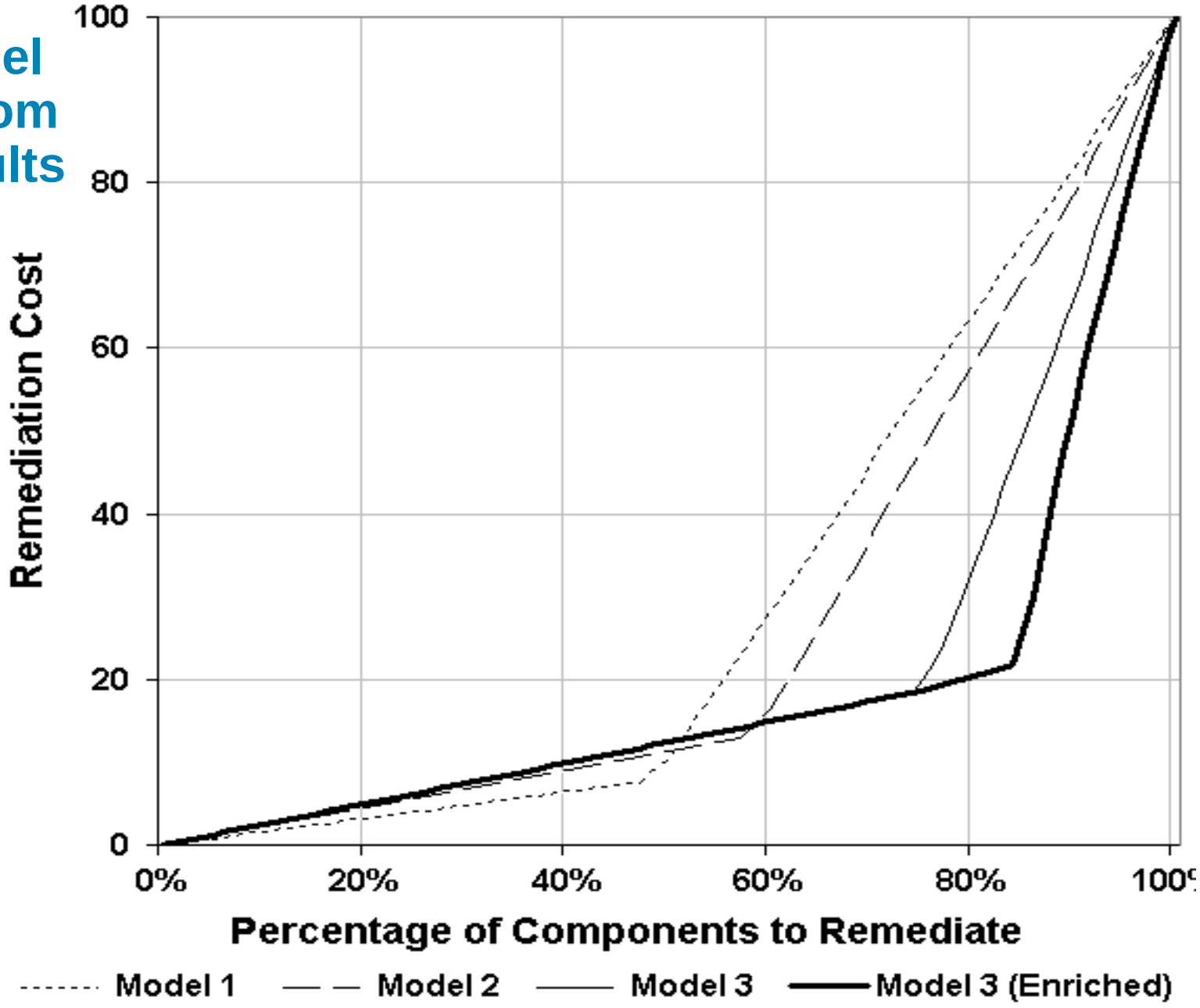
Results from sequential run:

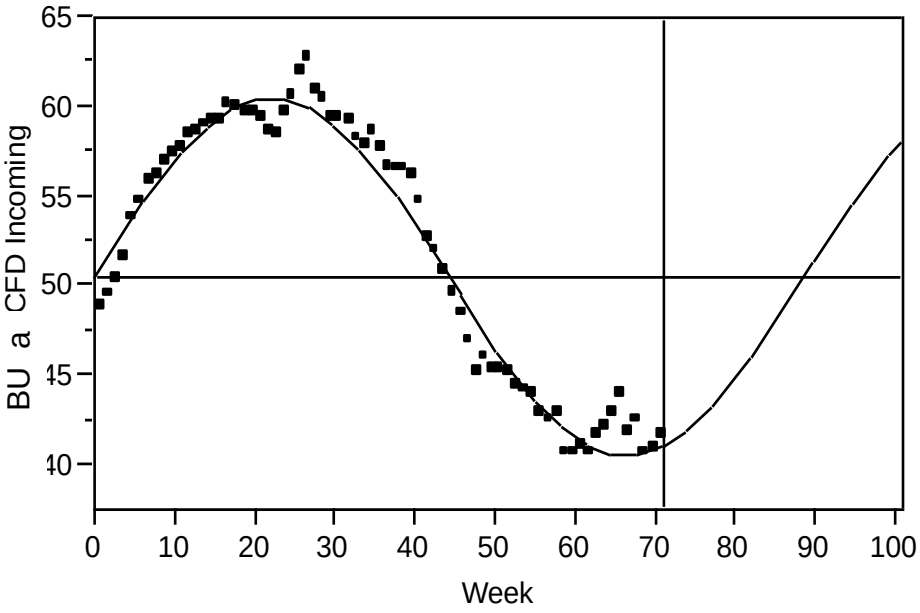
- increases true positives
- small increase in false positives
- decreases false negatives.

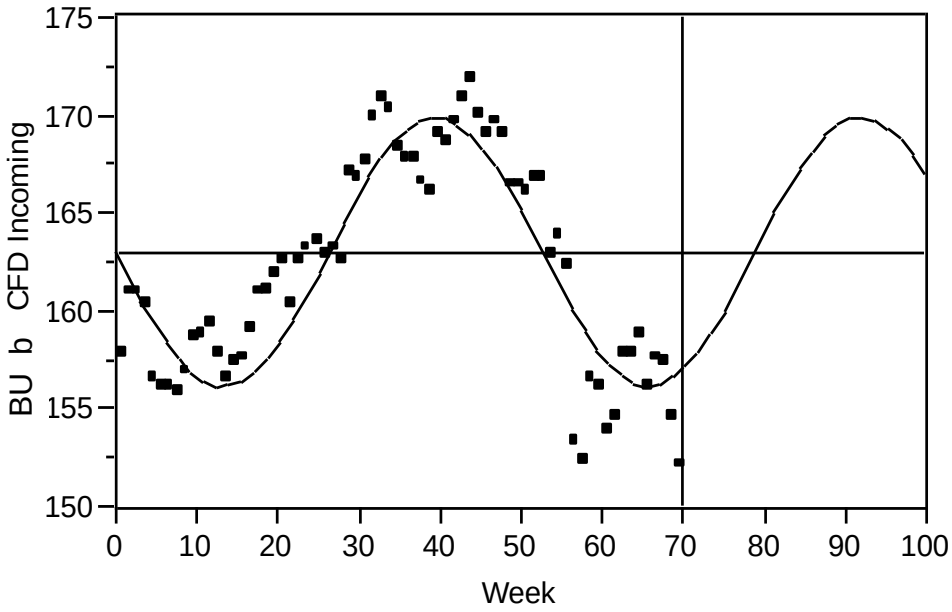
True positive rate increases by 9.8% with sequential model run.

75.6% → 85.4%

Cost model derived from CART results

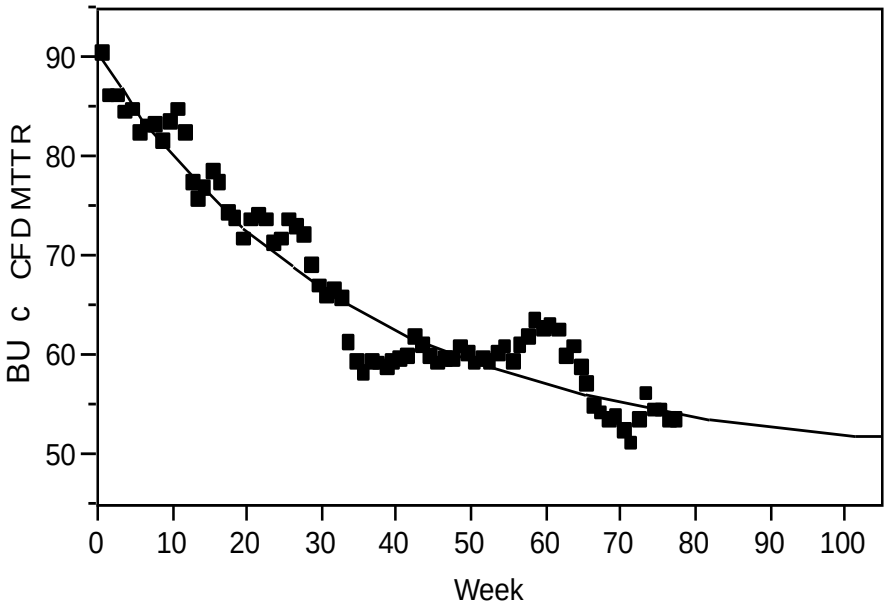


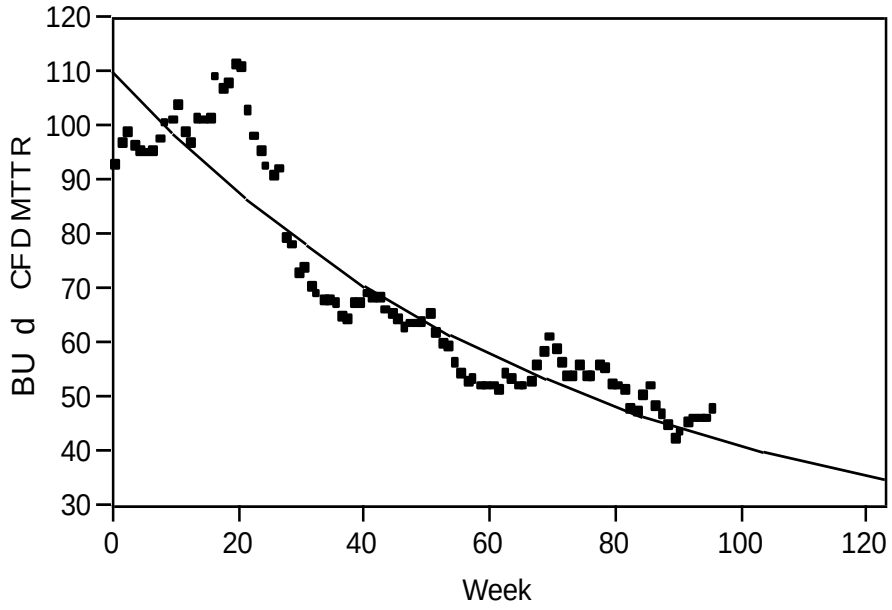




Sinusoidal modeling: Predicting field defect volumes

- Sine wave function ($y = y_0 + b[\sin(\omega t+\varphi)]$) is used in nonlinear regression modeling of incoming customer-found defects (CFDs)
- This model produces false negatives that can be identified by a second run of the model, using different variable coefficients.





Exponential decay modeling: Predicting repair rates of field defects

- Exponential decay function ($y = y_0 e^{-\lambda t}$) is used in nonlinear modeling of CFD MTTR (mean time to repair, defined as backlog divided by average daily close rate)
- This model produces false negatives that can be identified by a second nonlinear (sinusoidal) model, used sequentially with the decay model.

Limitations

- Difficult to quantify the overall probabilities for the aggregate of both model runs
- Large dataset is required to enable the second run of the model
- Limited to models that produce clusters of sought-after observations.

Summary

- Sequential modeling may be useful for scenarios in which a predictive model produces clusters of true positives
- Sequential modeling is useful for CART and several types of nonlinear models
- The identification of more fault-, failure-, vulnerability-, and attack-prone components, code regions, and development behaviors increases software assurance, without much additional effort.

Reference

- M. Gegick, P. Rotella, and L. Williams, "Predicting Attack-prone Components," *Proc of the ICST,* Denver, CO, pp. 181-190, April 2009.