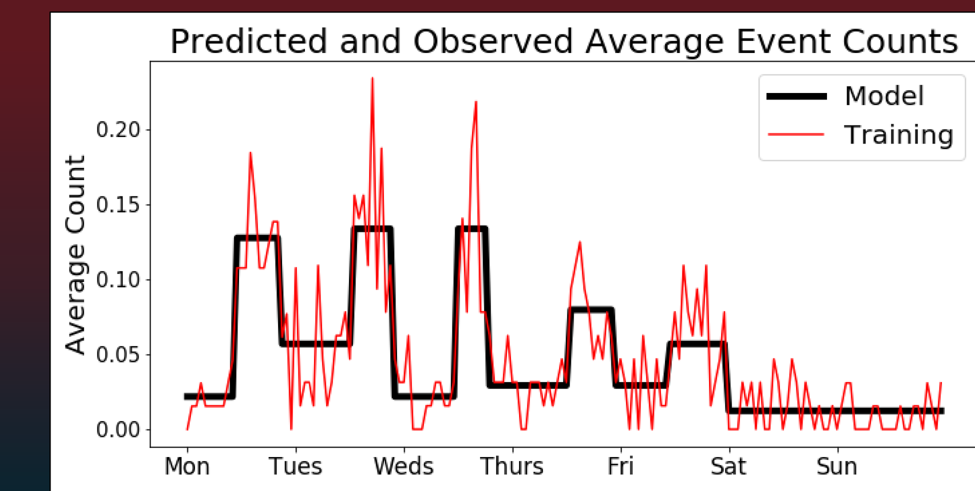
SINAPSE: Seasonal Inhomogeneous Nonconsecutive Arrival Process Search and Evaluation

Kimberly Holmgren, Paul Gibby, Joseph Zipkin – Massachusetts Institute of Technology Lincoln Laboratory

How to best model an arrival process, when:

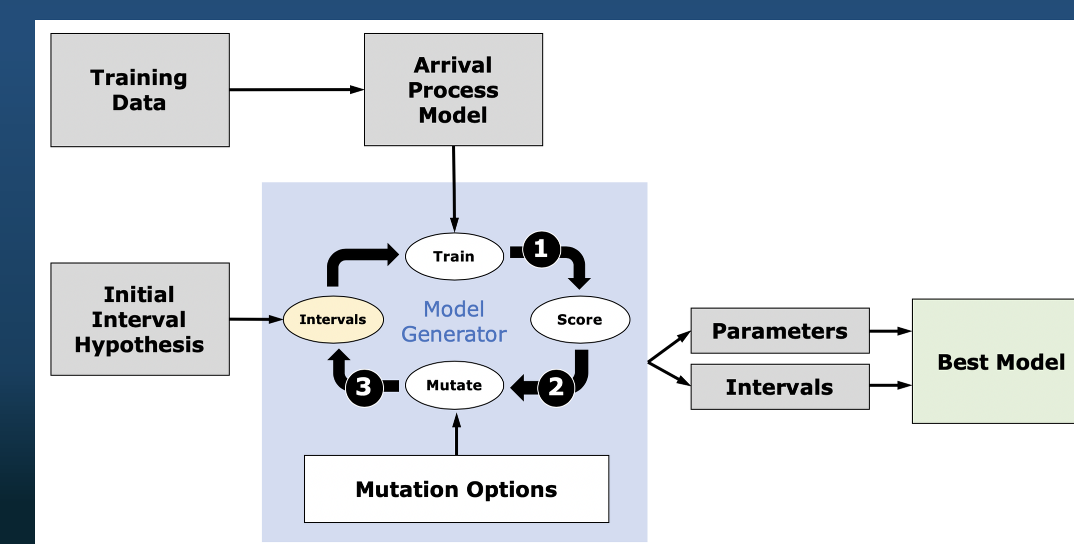
- The dataset is small and/or sparse
- The data exhibit seasonality
- The data includes points at which the parameters of the arrival process sharply change

- For example: malicious emails arriving at a company
- Few events make it difficult to model
- Behavior sharply changes by hour over the course of a week



Algorithm

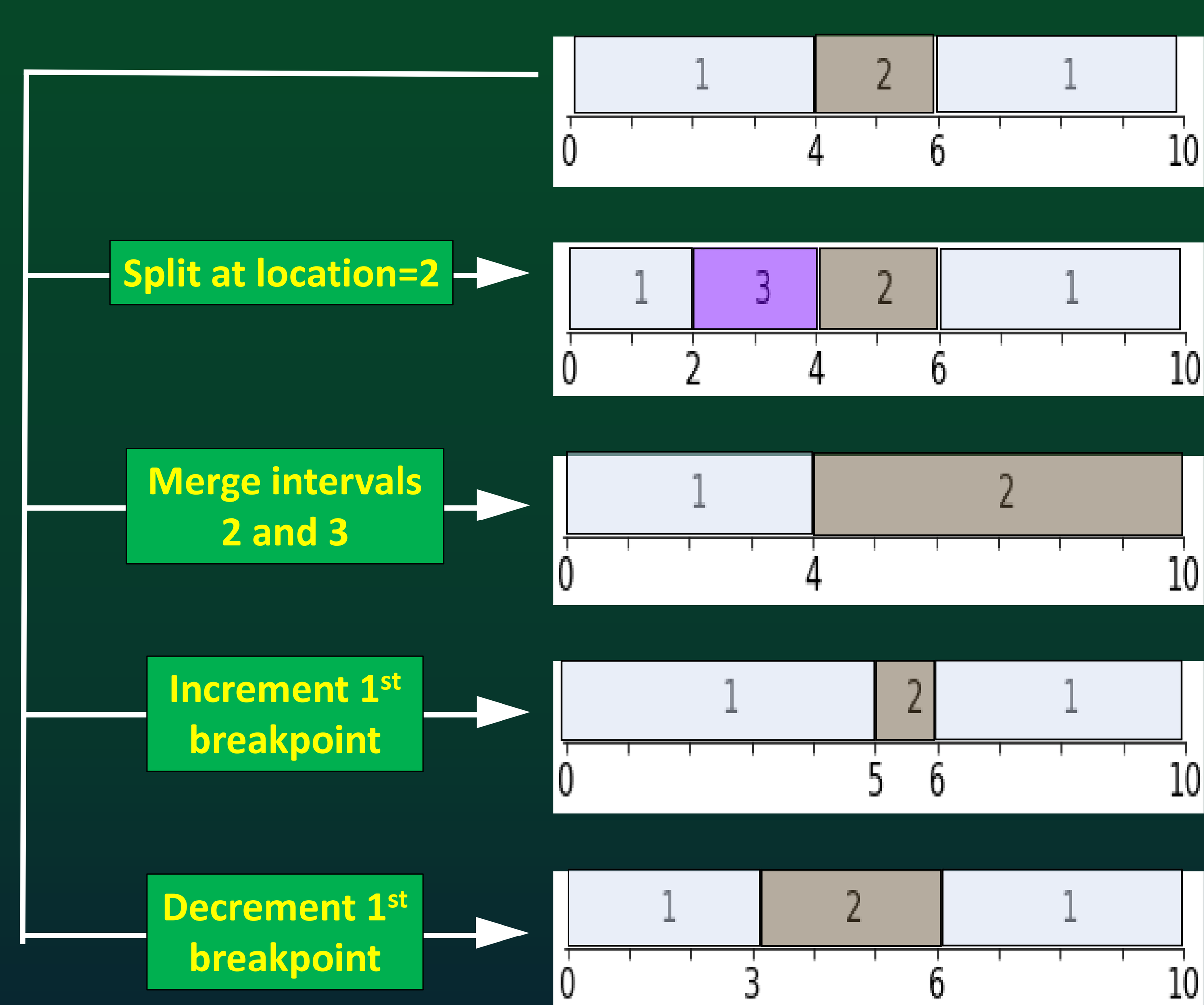
1. Start with a hypothesis division of the time period
2. Fit an arrival process model (Poisson, Negative Binomial, etc) and calculate a score (AICc)
3. Apply mutations by genetic algorithm
4. Repeat until convergence



Definitions

- Arrival process: A collection of points or events randomly placed in time
- Breakpoint: The point at which the distribution of a nonhomogeneous arrival process changes
- Interval: A closed, continuous subset of the timeline with the same set of parameters
- Regime: A collection of intervals, not necessarily contiguous, in which the point process follows a single set of parameters

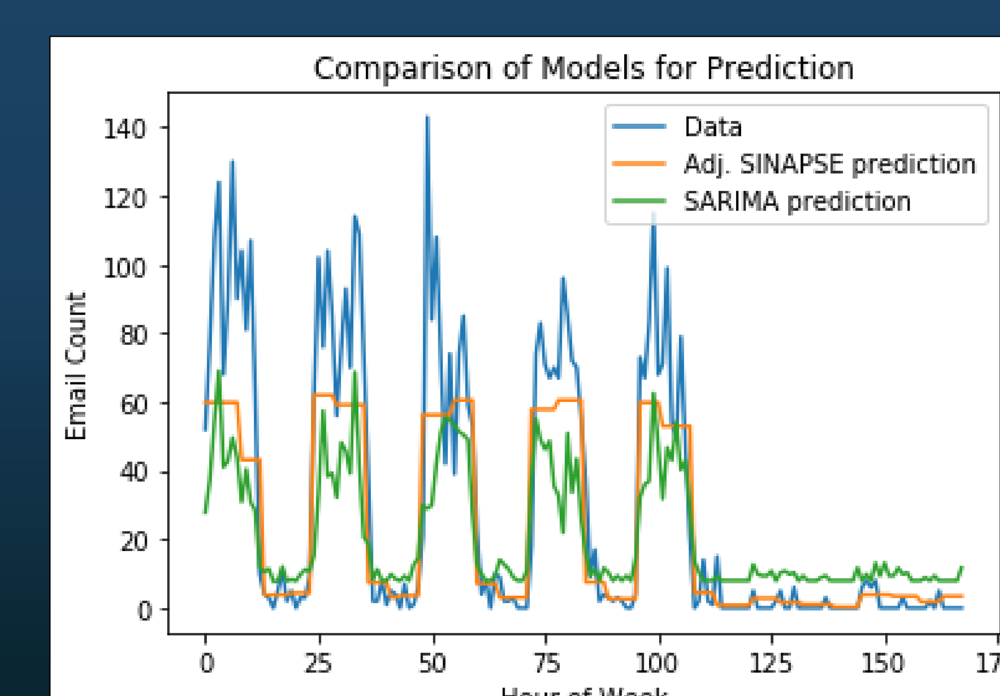
Intervals and Mutation

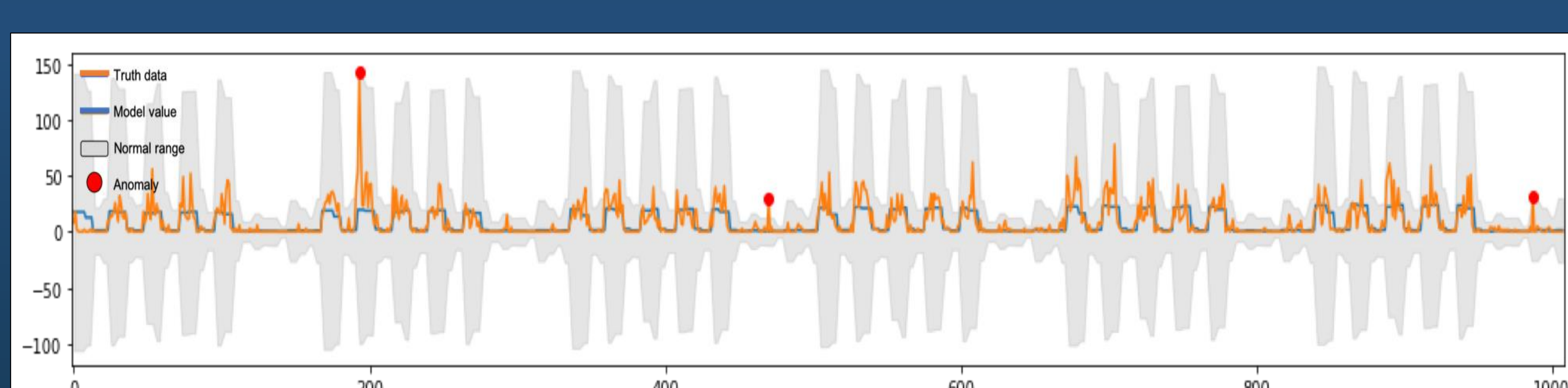


- Define an Interval: Divide the repeated period (often week) into intervals and assign intervals to regimes. Alternatively, start with homogeneous model (one interval and regime)
  - Interval regimes can be discontinuous, allowing a single set of parameters to govern disconnected time periods
- Mutate an Interval: allow for exploration of a space with an indefinite number of dimensions through breakpoint additions, deletions, increments, and decrements
  - Mutation driven by genetic algorithm to optimize a score or information criterion (AICc)
  - Balances modelling accuracy with overfitting by penalizing the number of parameters used

Application to seasonal prediction

- Use of an information criterion and discontinuous regimes leads to more parsimonious model
  - Models are less prone to overfitting
- Automatically group regions with similar occurrence patterns together identifying seasonal components
- Train on small and sparse datasets, automatically adjusting the complexity of the model to suit the data
- Comparison of SINAPSE to SARIMA model fits:





Application to anomaly detection

- SINAPSE-fit models inherently model the variance, which allows threshold-setting
- Example of 6 weeks of trend-adjusted data from the Enron dataset
- Red points correspond to anomalies

Paul Gibby, paul.gibby@ll.mit.edu, Kimberly Holmgren, kimberly.holmgren@ll.mit.edu, Joseph Zipkin, joseph.zipkin@ll.mit.edu

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

© 2020 Massachusetts Institute of Technology.
Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY