Malware and Artificial Immune Systems

Chris Musselle

Bristol Centre for Complexity Sciences (BCCS)
University of Bristol

Supervised by Dave Cliff and Ayalvadi Ganesh

Malware Evolution

- Pre 1990 – Experimental /intellectual pranks. E.g. Morris Worm.
- 1990-1999 – More sophisticated Viruses and Worms e.g. Macro virus, encryption, polymorphic viruses.
- 2000-2003 – Explosion of Worms. CodeRed, Nimda, Slammer etc...
- 2003-present – Increase in malware sophistication, blended threats, countermeasures, updating. e.g. Conficker.

- Shift in motive towards financial gain has driven the increased sophistication and prevalence of malware.
- The Web today provides cyber-criminals with the targets, exploitable weaknesses, and anonymity required for large-scale fraud.

Modern ‘Malware’ Economy

- Cyber-criminals have embraced Web 2.0 technologies, and specialise in various roles.
- Tools of the trade are readily available for purchase, with some malware authors even offering technical support and updates to their products.
- Basic strategy is to host new malicious sites / compromise legitimate ones, and then lure victims to them.
- Shift towards more stealthy and sophisticated malware e.g. Drive by Downloading, large surge in data theft Trojans malware.

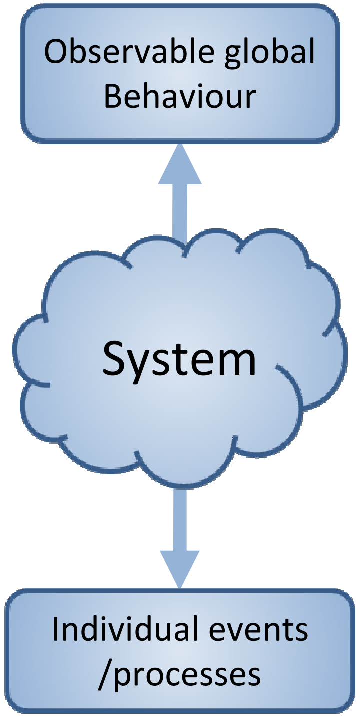
PhD Focus

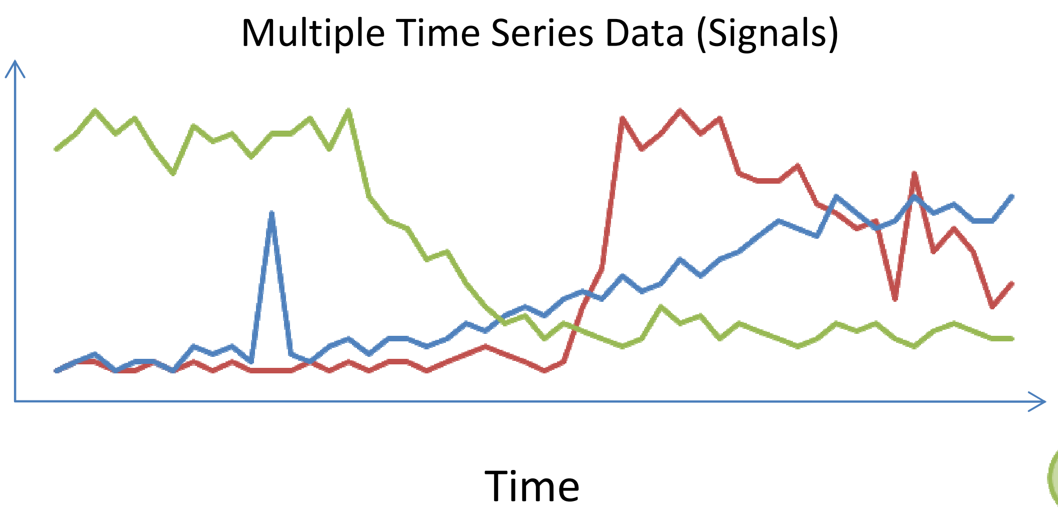
- Anomaly detection techniques to better distinguish between normal and potentially malicious behaviour within a computer system.
- Avenues of investigation
  - Artificial Immune Systems
  - Machine Learning
  - Statistical Techniques

The Dendritic Cell Algorithm (DCA)

- An abstract model of Dendritic Cell behaviour based on the paradigm of Danger Theory.
- Aims to perform anomaly detection by correlating a series of informative signals with a sequence of abstract events (termed `antigens').
- Signals → Multiple time series set to give approximations of normal or anomalous aggregate behaviour (termed either `danger' or `safe').
- Antigens → Symbolic IDs of the individual events.
- The goal is to determine which event is most likely responsible for an observed rise in danger signals .

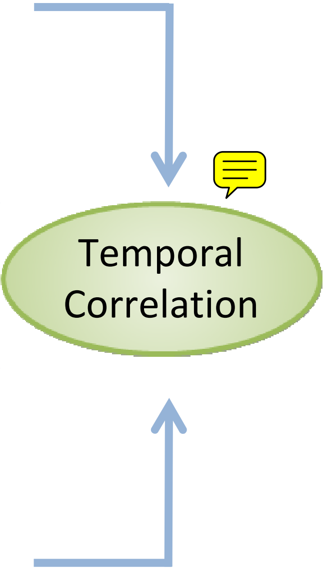
Inputs to the DCA



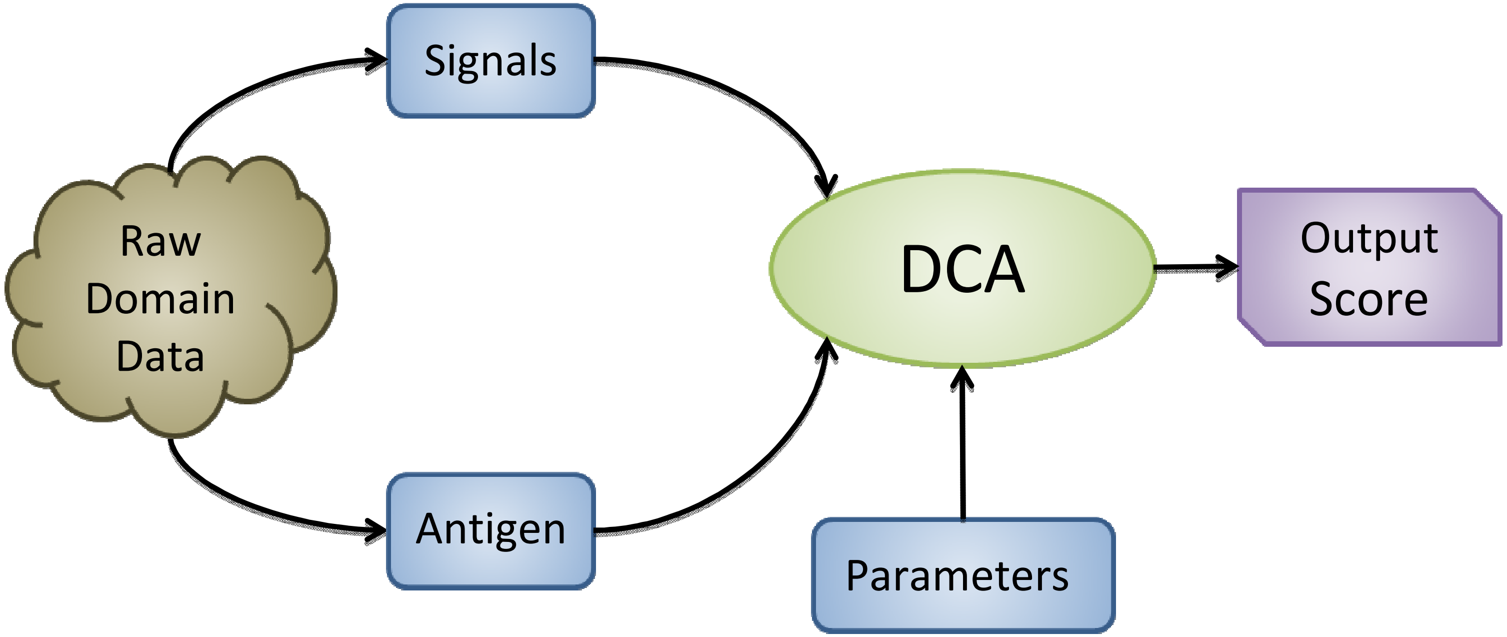


AABCBDECSDESCEADGFEDETEABERSE
RTBD
ABERBJFJK GJUWBGTYOC FGKYBECW
QHD
P ODN FRB T OFOTMNFI SODO FM OPF
K FPF   P U M RJT   DKI MG
P   T   F OI   NBJ OM
P   O   J

Logged Events per Time Interval (Antigens)

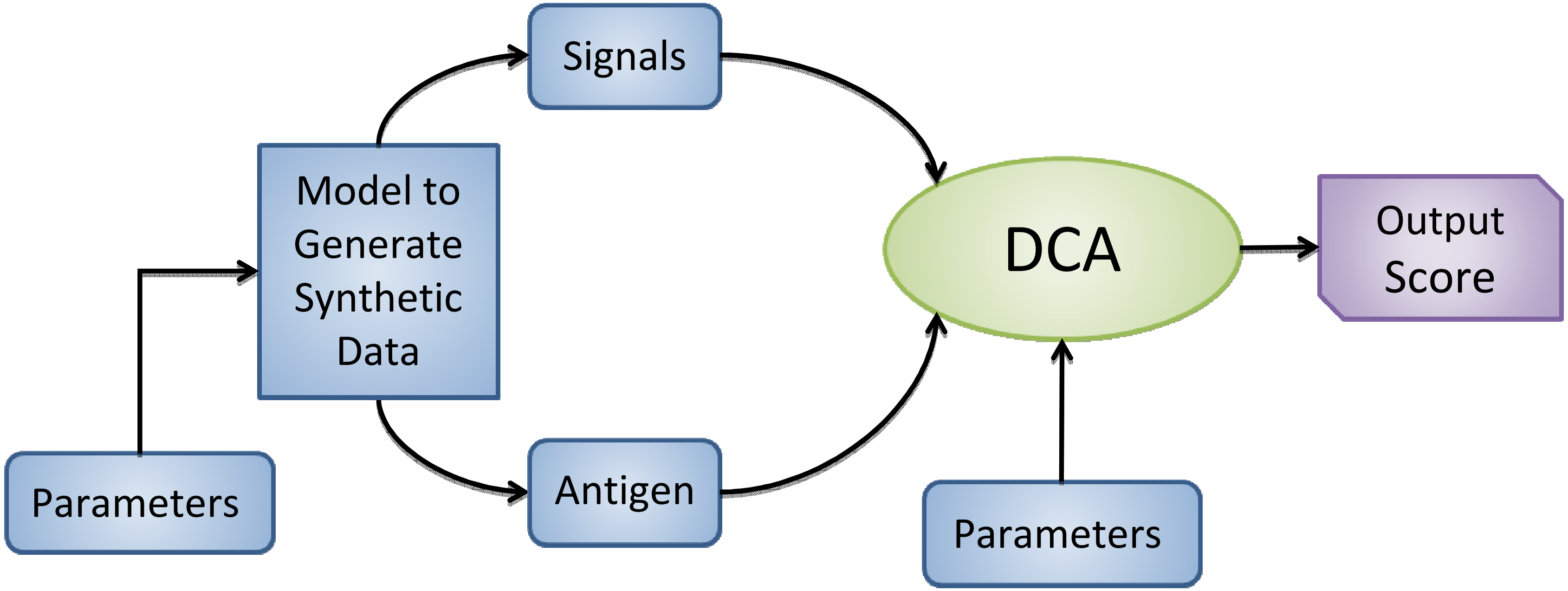


Some Limitations

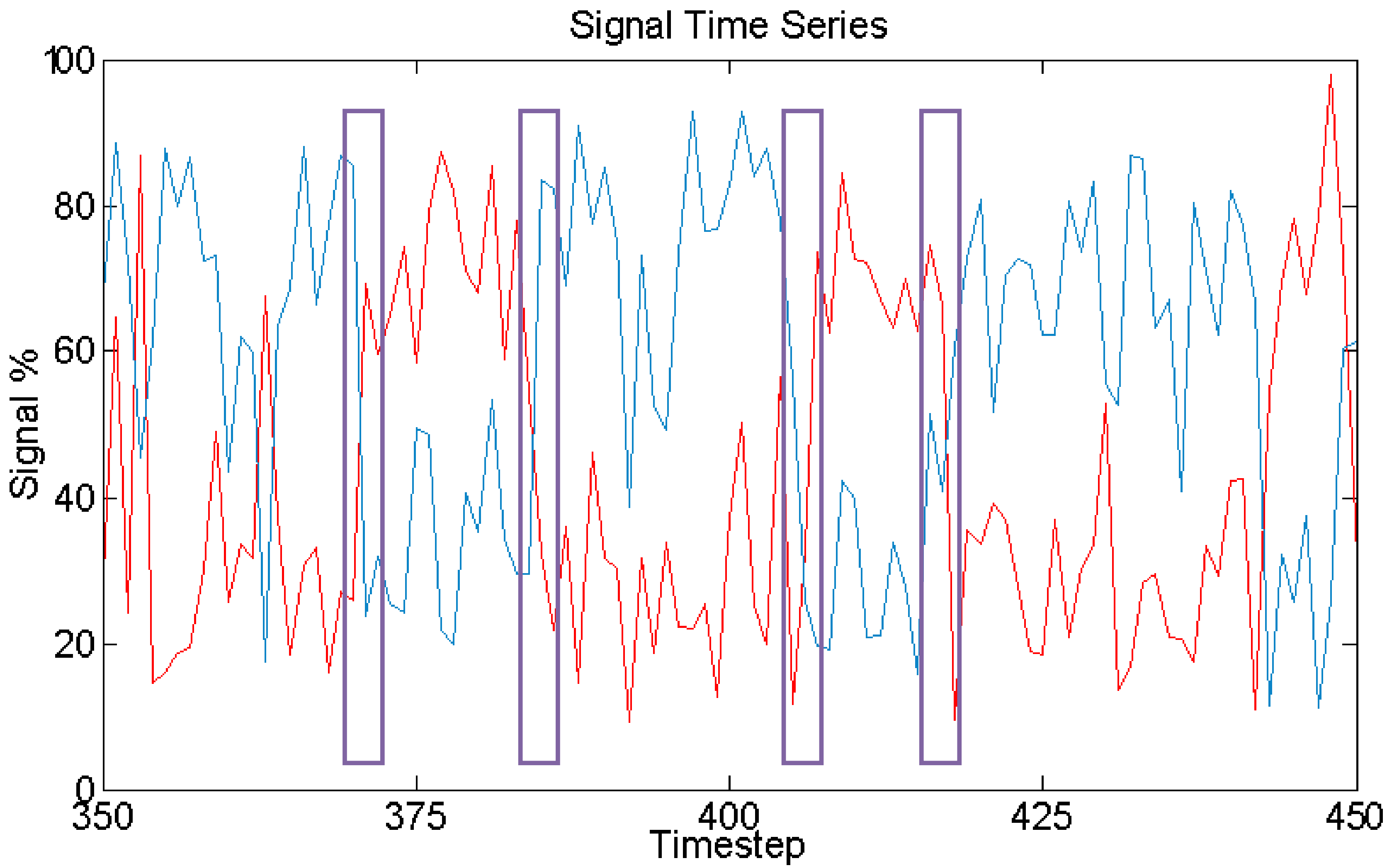


- Reliance on expert knowledge to carry out mapping into the antigen and signal space.
- Can lead to the definition of inputs being quite arbitrary, difficult to compare applications.
- Trial and error in finding appropriate parameters.

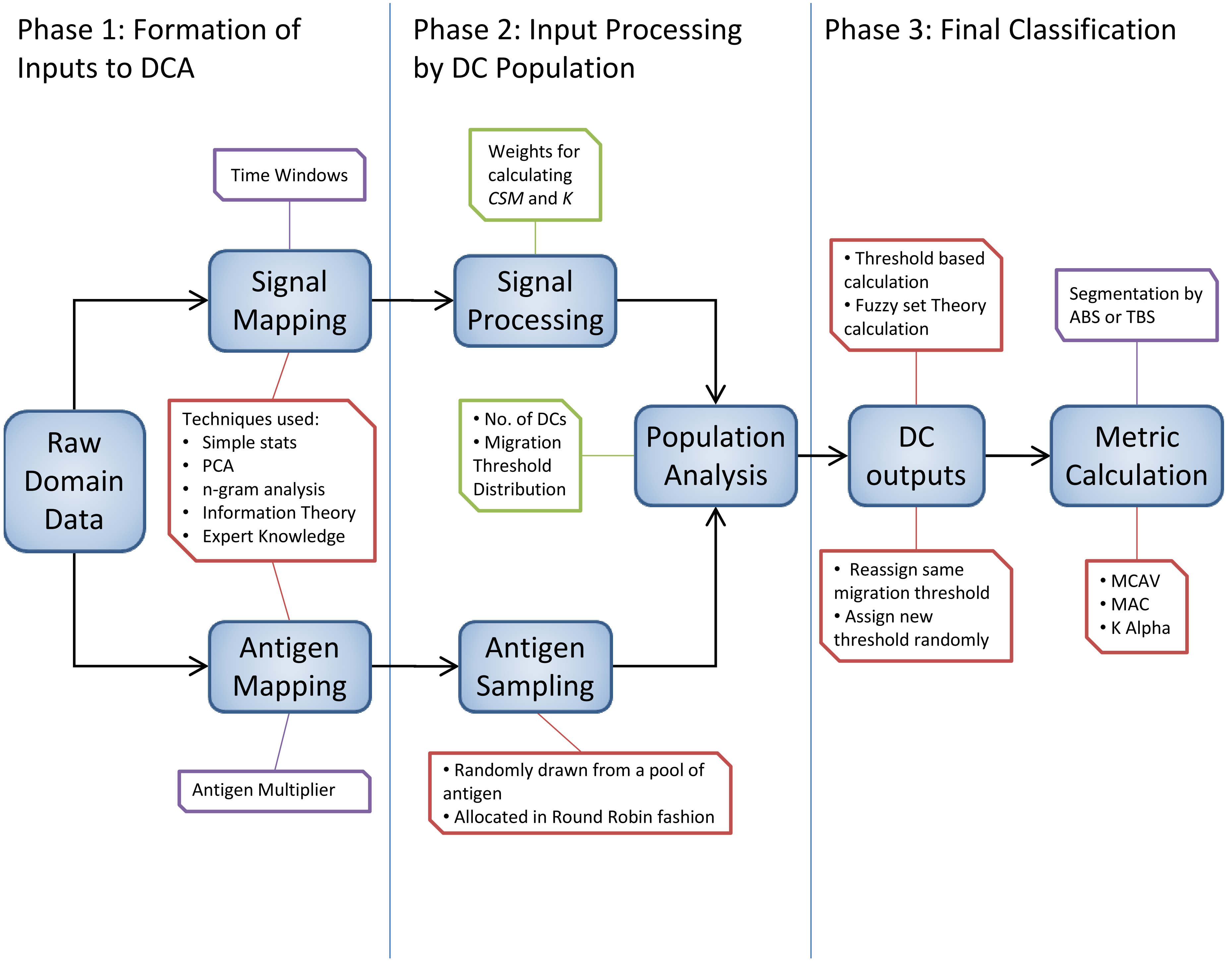
My Approach



- Generate controllable synthetic data using a model.
- Investigate the relationship between inputs, DCA parameters, and algorithm performance.
- Focus on the deterministic DCA (dDCA).



Errors in classification occurred at boundaries


Phase 1: Formation of Inputs to DCA
Phase 2: Input Processing by DC Population
Phase 3: Final Classification
Time Windows
Signal Mapping
Weights for calculating *CSM* and *K*
Signal Processing
Raw Domain Data
Techniques used:
• Simple stats
• PCA
• n-gram analysis
• Information Theory
• Expert Knowledge
• No. of DCs
• Migration Threshold Distribution
Population Analysis
• Threshold based calculation
• Fuzzy set Theory calculation
DC outputs
Segmentation by ABS or TBS
Metric Calculation
• Reassign same migration threshold
• Assign new threshold randomly
• MCAV
• MAC
• K Alpha
Antigen Mapping
Antigen Sampling
Antigen Multiplier
• Randomly drawn from a pool of antigen
• Allocated in Round Robin fashion

Back to Basics

What problem am I really trying to solve?

- Unsupervised classification of previously unseen events, based on cross-referencing multiple heuristic indications of system behaviour. Context based anomaly detection.
- Ideally operating within a sliding window on continual streaming data providing real time detection of anomalies.
- Related to the simpler one of identifying anomalies in streaming data, however:
  - Monitoring multiple time series in parallel.
  - Allowing multiple events to happen at each time step.
- Investigate other approaches to solve the same/similar problems.
  - Time series analysis techniques.
  - ML context based anomaly detection.
  - Rare Event detection.
  - Statistical decision making / Change Point Detection.

Other Approaches

Sliding window Techniques

- Change Point Detection
  - Statistical technique using non-parametric CUSUM.
- Incremental Local Outlier Factor
  - k nearest neighbour.

Multi-time series Analysis Methods

- Multivariate linear regression
  - Relies on relationships between time series as well as the past.
- Multivariate Bayesian Scan Statistic
  - Bayesian Networks, need priors plus complete knowledge of events.

Future Work

- Investigate which techniques are the most effective and incorporate into the danger theory framework.
- Either use these techniques to augment the DCA, or integrate those that prove useful into a new ‘DCA like’ AIS algorithm inspired by Danger theory.
- Test on simulated and real world data sets (hopefully!)

Thanks For Your Attention

Questions?