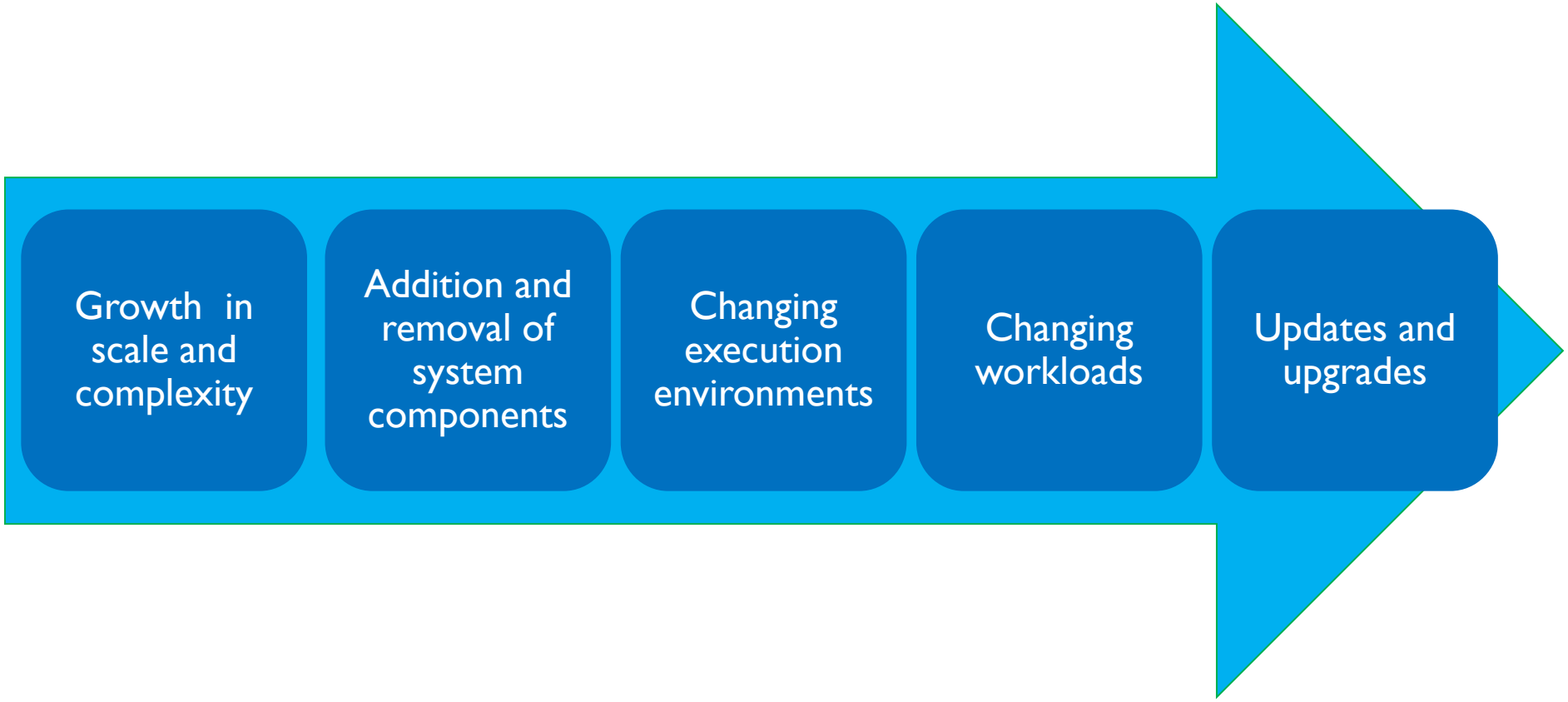
SNIA DSI '16 DATA STORAGE INNOVATION CONFERENCE
Innovation in Storage Products, Services, and Solutions
June 13-15, 2016 | Marriott San Mateo | San Mateo, CA

Machine Learning Based Prescriptive Analytics for Data Center Networks

Hariharan Krishnaswamy
DELL

Modern Data Center Characteristics



Data Center Failure Scenarios

- Network failures
- Device failures
  - Host
  - NIC Team
  - Router, Switch, Firewall
  - Storage array
- Application failures
- Traffic issues
  - Latency/Throughput issues
  - Head of line blocking
  - Microburst

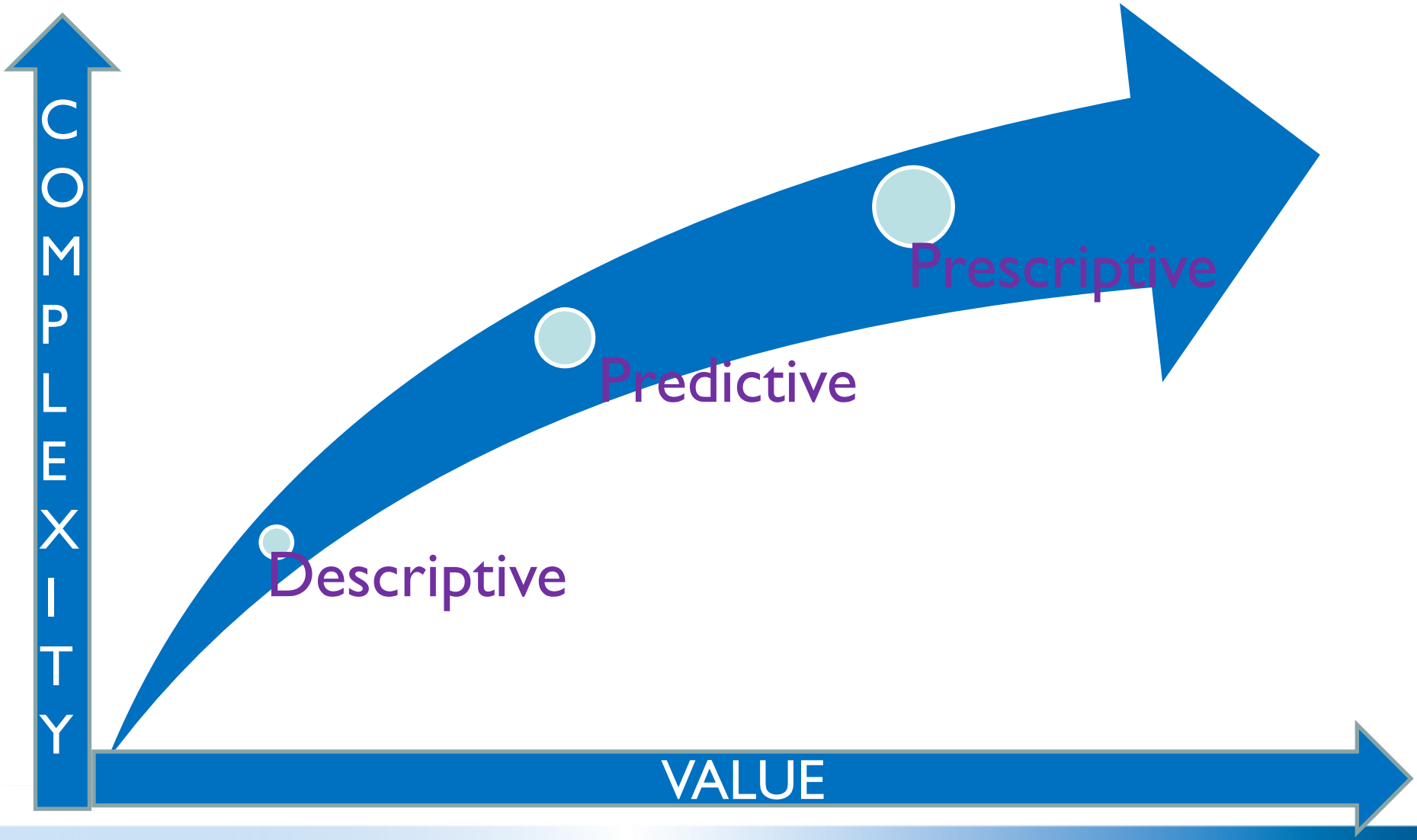
Reasons for Data Center failures

- Hardware failures
  - Some piece of hardware malfunctions/break down
- Software failures
  - Software bugs
- Operating conditions
  - Exceptionally high load is offered to the system/network
  - Environmental conditions
- Operator errors
  - Cabling error, Configuration errors

What an Analytics framework help with ?

- Blend and ingest a variety of structured, semi-structured and unstructured data
- Discover patterns & hidden correlations
- Detect anomalous behavior & Predict failure
- Root cause the anomalous behavior
- Suggest decision options & their implications

Types of analytics



Descriptive & Predictive Analytics

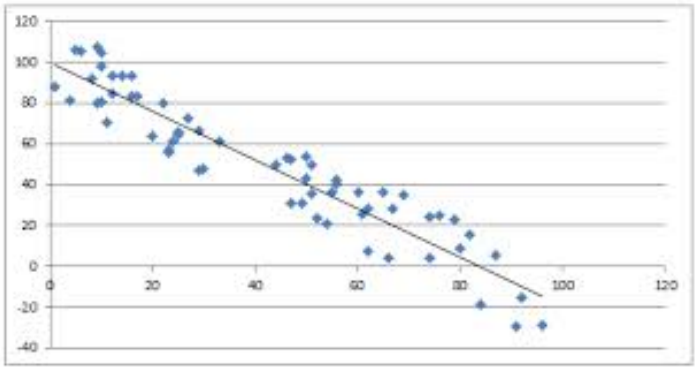
Descriptive Analytics

- Aims to provide insight into what has happened

Predictive Analytics

- Techniques used to make predictions about future events (Statistics, Data mining, Modeling, and so forth)
- What is likely to happen in future?
  - Identify patterns/trends
  - Identify clusters
  - Detect anomalous behavior

Identifying Patterns/Trends

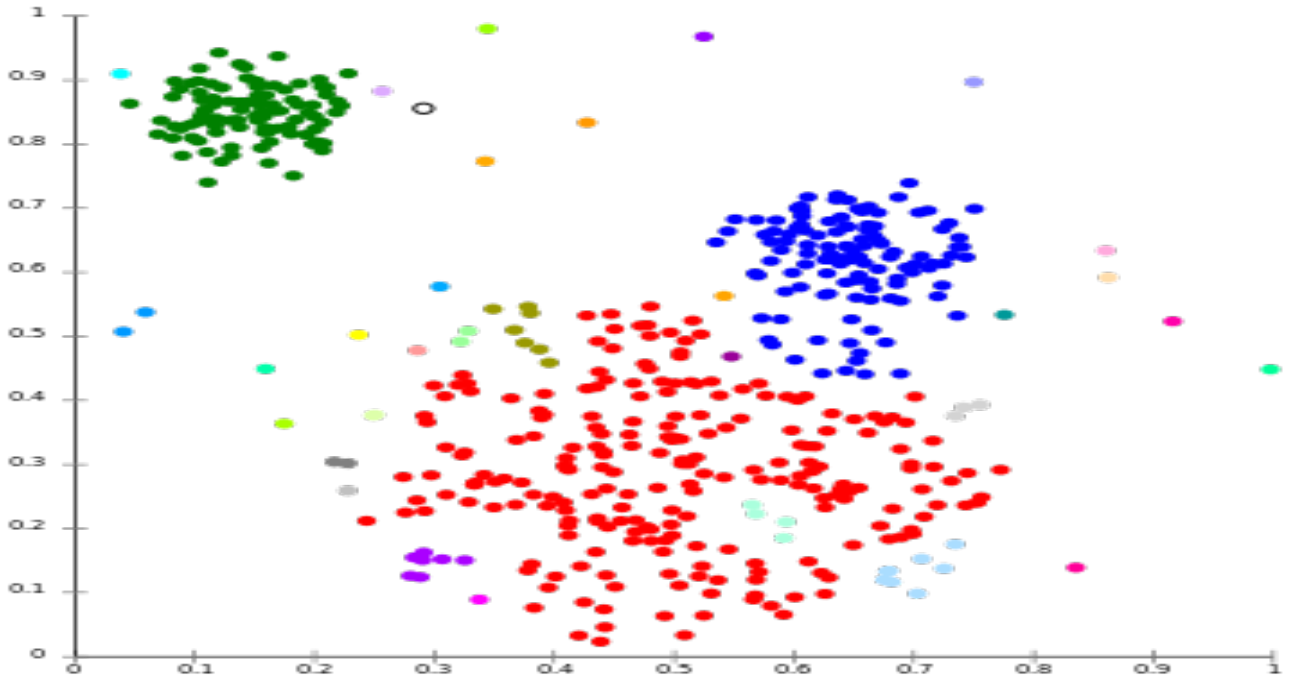


Characteristics of Time Series Data

- Trend over time (Ex: Gradual increase/decrease of activity over time)
- Seasonal trend or cycle (Ex: Traffic increases in the morning hours, peaks in the afternoon and declines late at night)
- Seasonal variability. (Ex: Application requests fluctuate wildly minute by minute during the peak hours of 4-8 pm, but at 1 am application requests hardly vary at all)

These characteristics should be accounted for by the Analytic Model

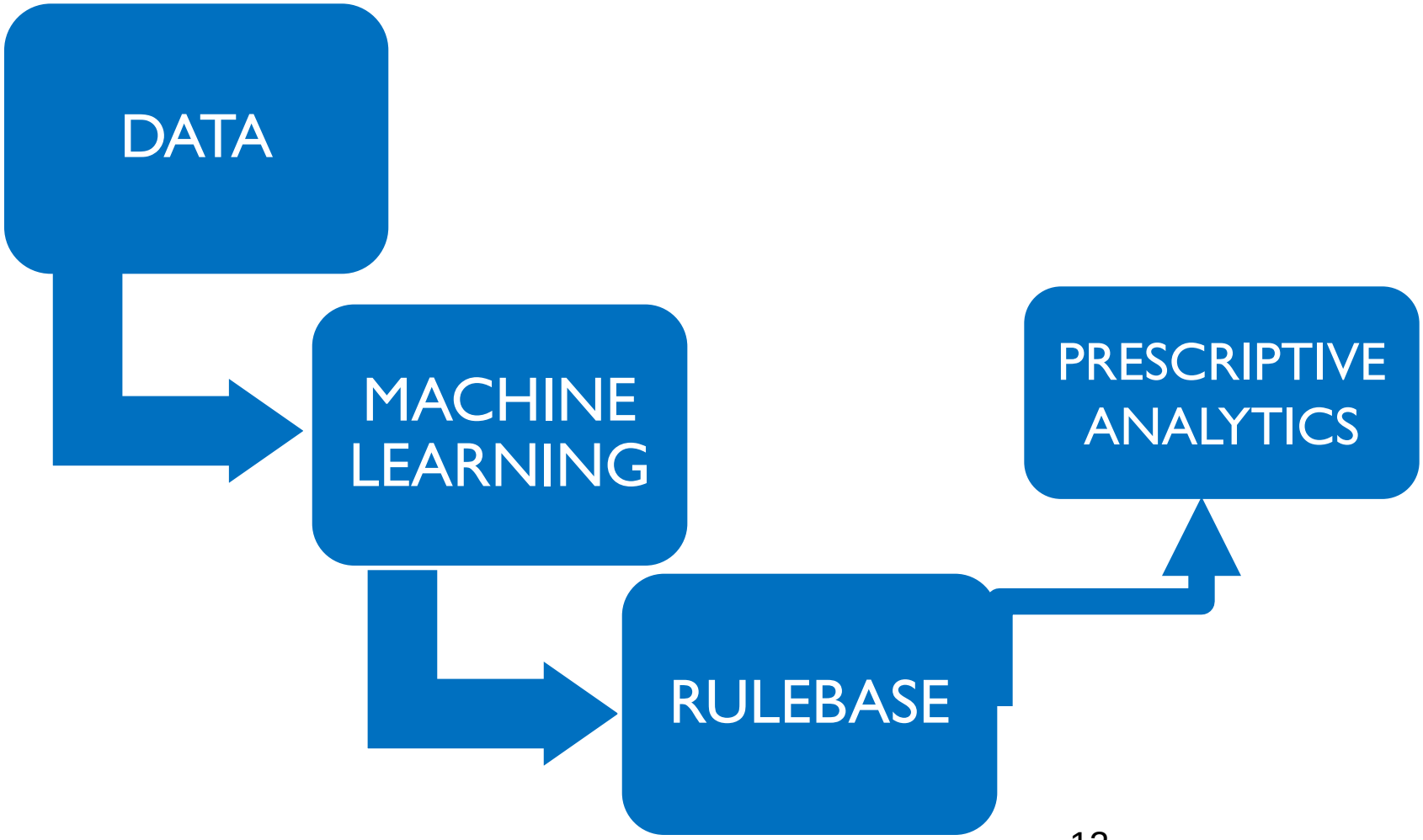
Identifying Clusters



Prescriptive Analytics

- Data mining and predictive analytics precede the prescriptive analytics.
- Prescriptive analytics provides courses of actions and suggests options to specific situations
- Determine the best solution or outcome among various choices
- Can continually take in new data to re-predict and re-prescribe
- Extends beyond predictive analytics by specifying both the actions necessary to achieve predicted outcomes and the interrelated effects of each decision

Approach to Prescriptive Analytics



Data Sources

LOGS

- Time-Stamped
- Semi-structured
- Event Log, Trace Log, Debug Log
- Numerical data and non-numerical data

TRAPS/ALARMS

- Notification Mechanism

IPC

- Inter Process Communication Messages

Data Sources

Network Traffic Statistics

- Packet Level Statistics

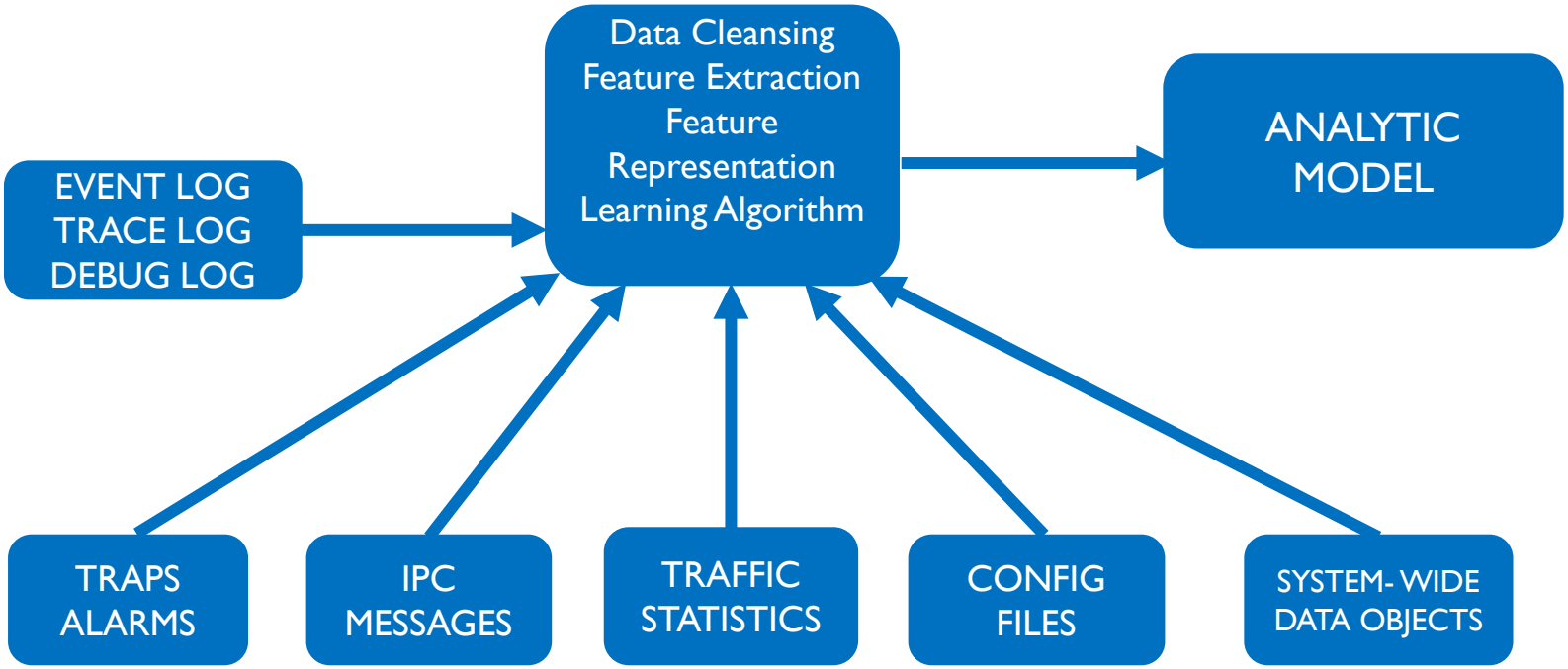
System Configuration

- Config Files

System-wide data objects

- Server Specific
- Switching /Forwarding/Routing Specific
- Application Specific
- Storage Specific

THE ANALYICS ENGINE



Manual Data analysis - Challenges

- Data volume is huge
- Complex and time consuming
- Lack of operational context makes data ambiguous
- Blending/Correlating multiple data sources is difficult

Machine leaning based analytics are more effective !!!

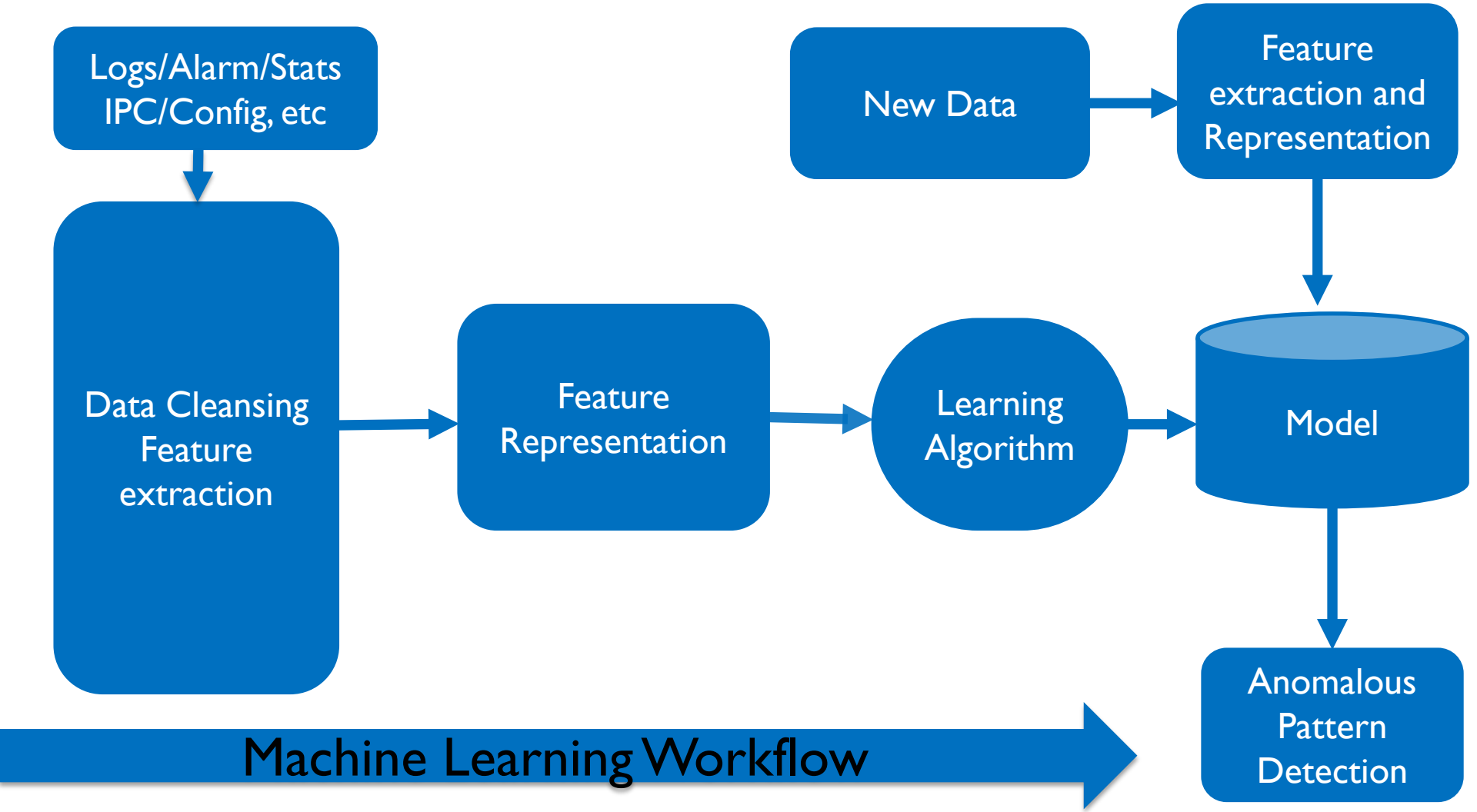
Types of Machine Learning

Unsupervised learning

- The model is not provided with the correct results during the training
- Can be used to cluster the input data in classes on the basis of their statistical properties only

Supervised learning

- Training data includes both the input and the desired results
- Correct results are known and are given in input to the model during the learning process



Machine Learning Algorithms

- Linear Regression
- Logistic Regression
- SVM (Support Vector Machine)
- Naive Bayes
- K- Nearest Neighbors
- K-Means clustering
- Dimensionality Reduction Algorithms

Machine learning workflow

- **Learn patterns on a set of training data**
  - The training dataset is labeled
  - For each data set used to train the model, it is known if it corresponds to a normal execution state or anomalous
- **Apply on a New Data Set**

CLASSIFIER: Support Vector Machine (SVM)

- A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane
- A Supervised learning approach
- Given labeled training data the algorithm outputs an optimal hyperplane which categorizes new Data Sets into appropriate class

Support Vector Machine (SVM)

- SVM uses linear models to implement nonlinear class boundaries.
- SVM transforms the input space using a nonlinear mapping into a new space (F feature space)
- Then a linear model constructed in the new space can represent a nonlinear decision boundary in the original space
- If the Dataset is linearly separable. The maximum margin hyperplane is the one that gives the greatest separation between the classes

Advantages & Disadvantages of SVM

Advantages

- Produce very accurate classifiers
- Less overfitting, robust to noise

Disadvantages

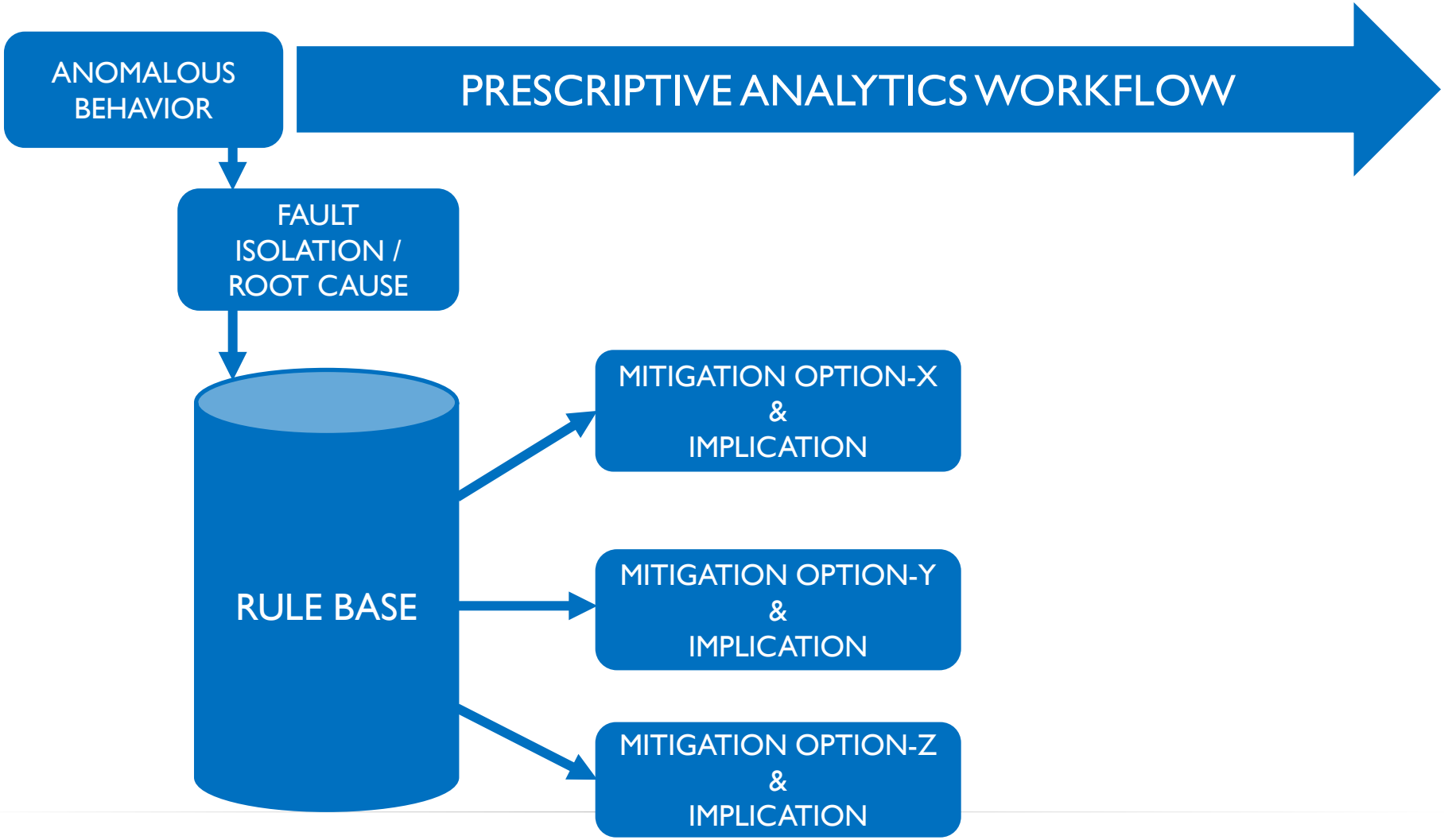
- SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used (one class against all others, for all classes)
- Computationally expensive

Model Development Phase

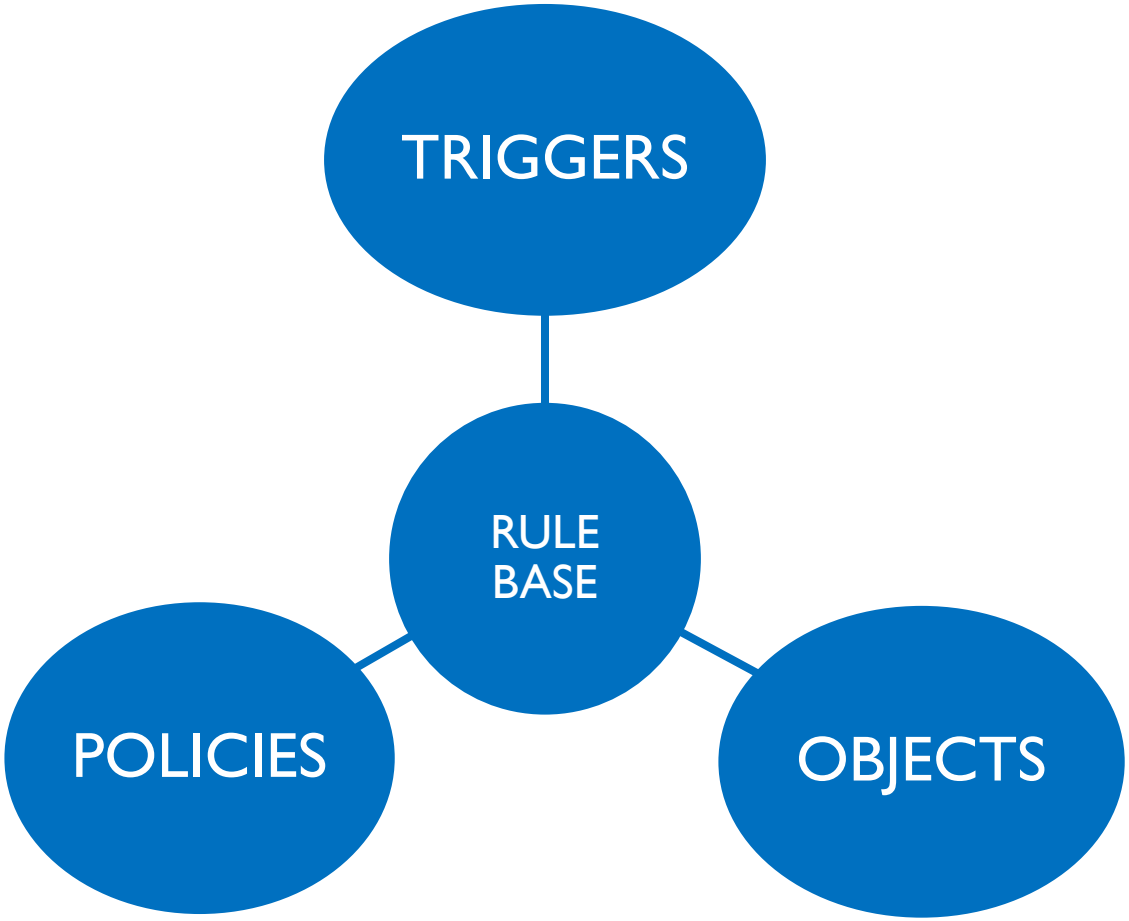
- Certain sequence of Time Series Message data are precursors to certain type of failure
- Time Series Message data are transformed into a multi dimensional vector space representation
- During training phase, the system learns the mapping: $X \rightarrow Y$, where $x \in X$ is some system state and $y \in Y$ is a class label (Normal / Anomalous for example) Model is developed during training phase

Deployment Phase

- In deployment phase, new data is run through the model to detect anomalous condition
- Fault isolation follows Anomalous Condition Detection
- Root cause together with Rule Base yield mitigation options and their implications



RULEBASE



Applications

- Anomaly Prediction & Proactively Avoiding
- Gain insight into the root causes & mitigation of the Anomaly
- Assessment of the impact of a new application
- Long-term trend forecasting
- Capacity Provisioning
  - Memory/processing power to a server
  - Packet buffers of the network switch
  - Storage array upgrades

THANK YOU