28th Annual INCOSE
international symposium

Washington, DC, USA
July 7 - 12, 2018

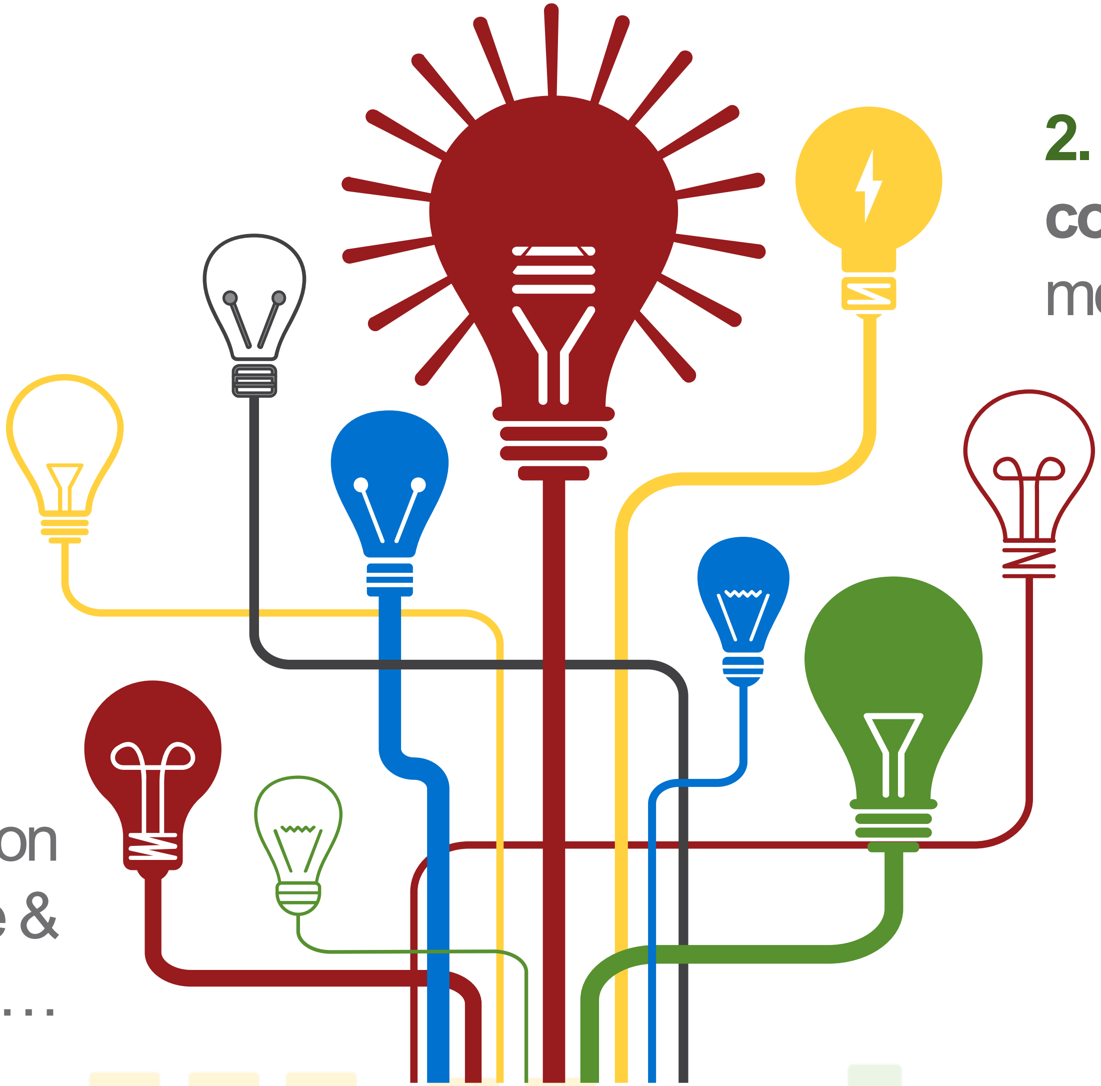
Towards Faster Design, Integration, and Execution of Inference Models

STIEM:

Shou Matsumoto (smatsum2@gmu.edu)
Dr. Edward Huang (chuang10@gmu.edu)
Dr. Kathryn B. Laskey (klaskey@gmu.edu)

C4I and Cyber Center
George Mason University, Fairfax, Virginia

Complex inference problems…



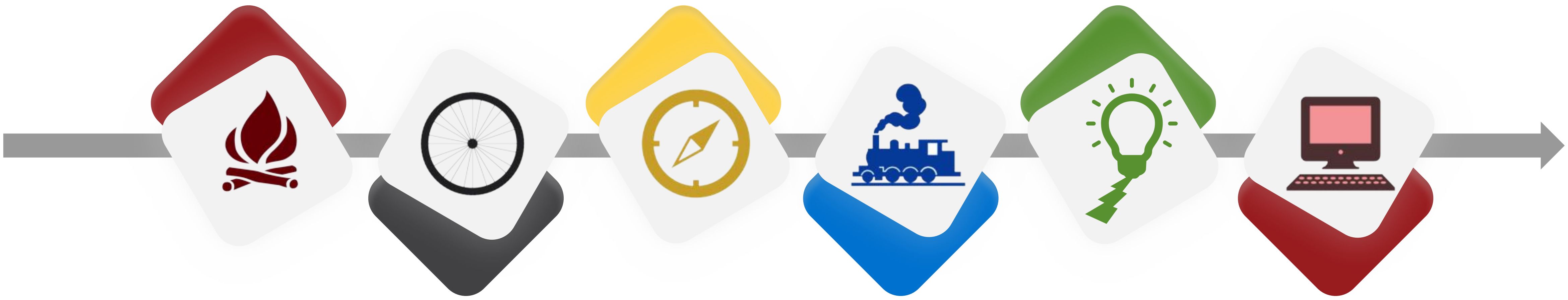
1. Hard to solve with single analytical method…

2. May require **combination** of models/processes, with such models/processes supplementing & complementing each other…

3. Manual integration often cumbersome & error-prone…

Inference = deriving conclusion from facts or premises (Merriam-Webster).

Need tools!



Tools
(and processes/methods)
are needed

to streamline
and support

multi-model
integration…

Therefore,
we developed

Model-Based
Engineering
framework
& tools/methods

to support design,
integration, execution,
and analysis.

Name of the tool: *Semantic Testbed for Inference Enterprise Modeling (STIEM)*

Definitions

Insider Threat:

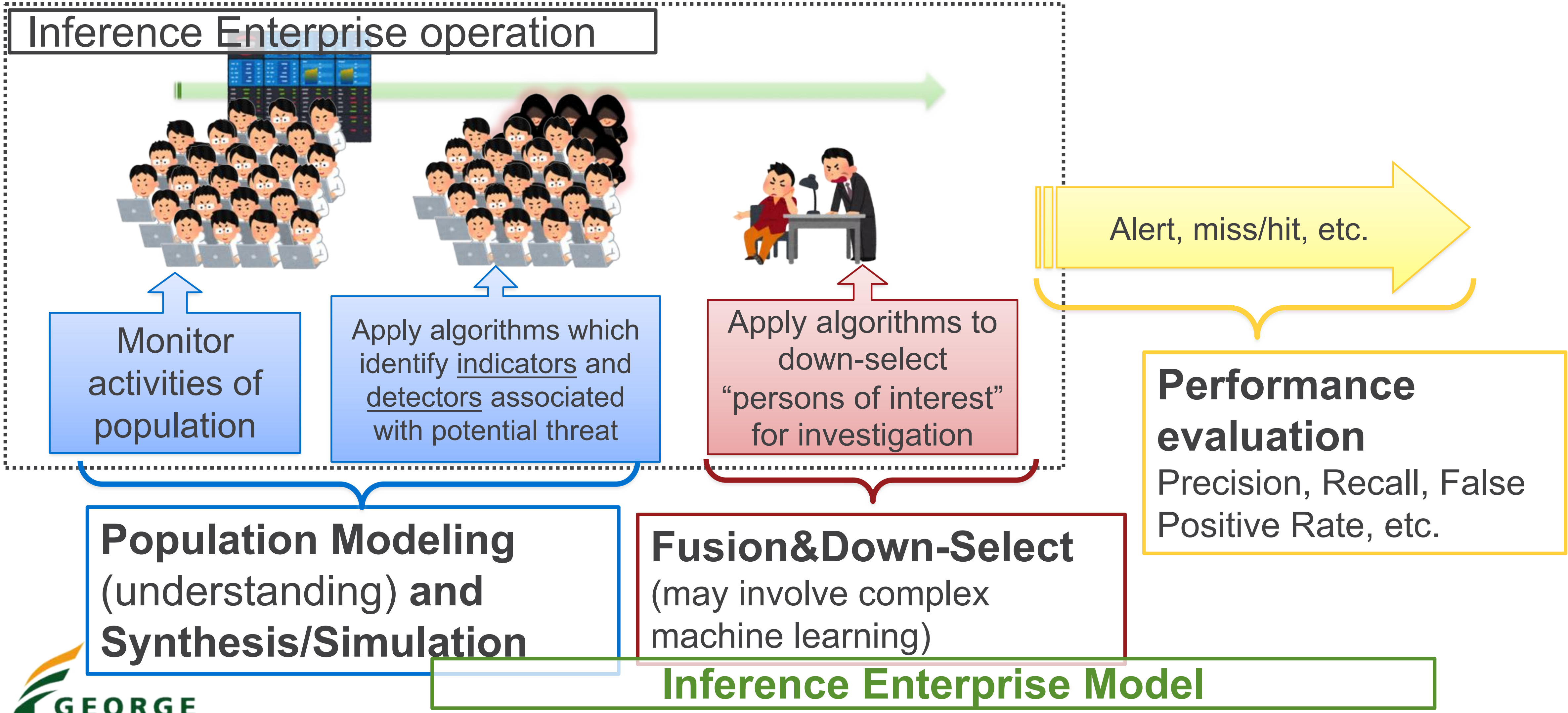
- Individual(s) who….
  - Is current/former employee, contractor, or other business partner
  - Has/had authorized access to organization’s network, system, data
  - Intentionally or unintentionally exceeds/misuses access, to negatively affect confidentiality, integrity, availability of organization’s information or information systems

[CERT, 2012]

Inference Enterprise:

- Collection of…
  - Data, tools, algorithms that organization employs to address some inference task
- We view it as both manual & automated process…
  - But we’re specifically concerned with its automated portion

Inference Enterprise Model



Inference Enterprise operation

Monitor activities of population

Apply algorithms which identify indicators and detectors associated with potential threat

Apply algorithms to down-select "persons of interest" for investigation

Alert, miss/hit, etc.

Population Modeling (understanding) **and Synthesis/Simulation**

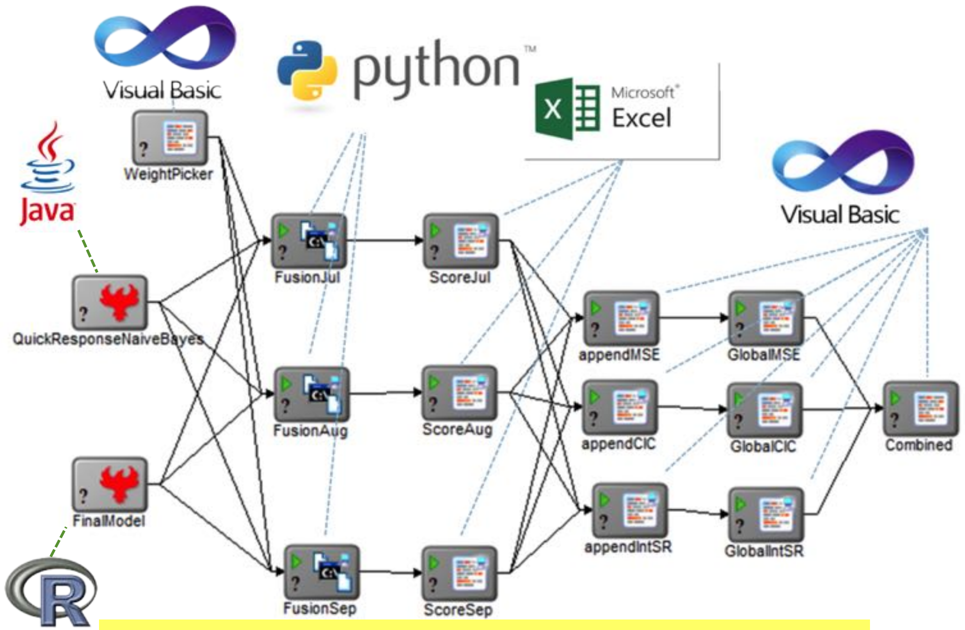
Fusion&Down-Select (may involve complex machine learning)

Performance evaluation Precision, Recall, False Positive Rate, etc.

Inference Enterprise Model

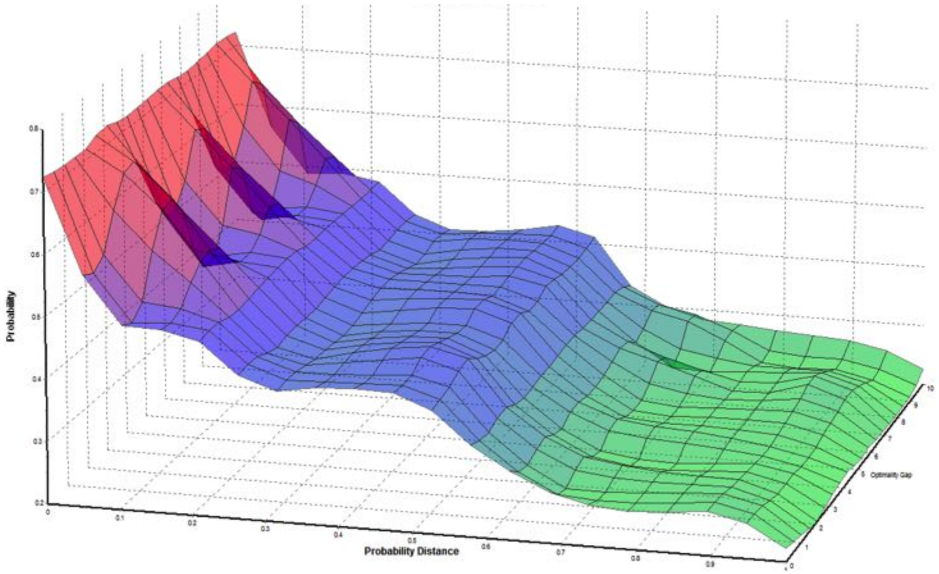
Inference Enterprise Model: represents/evaluates/predicts performance of Inference Enterprise.

Semantic Testbed for Inference Enterprise Modeling



Multi-Model Integration Workflow

PHOENIX INTEGRATION ModelCenter®

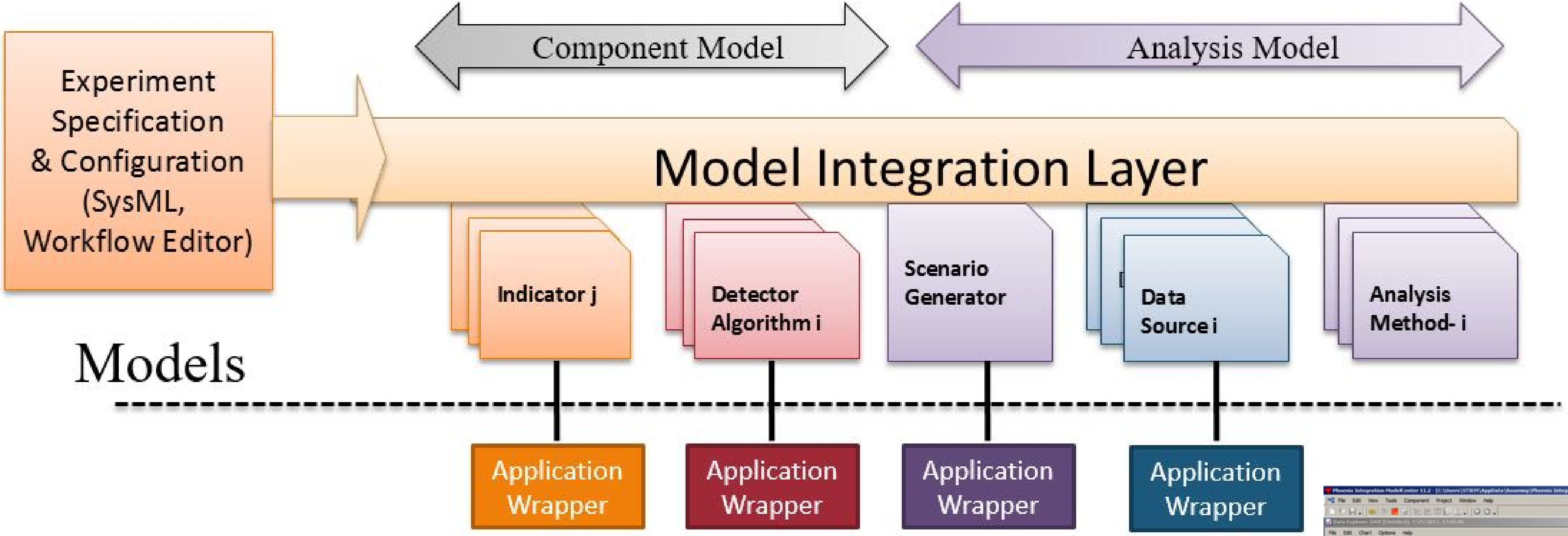


Sensitivity Analysis of Whole System Performance

Parallel/Distributed Execution

Model integration architecture implemented on top of COTS, Phoenix Integration's ModelCenter® (www.phoenix-int.com)

Model Integration Architecture of STIEM



Component Model

Analysis Model

Experiment Specification & Configuration (SysML, Workflow Editor)

Model Integration Layer

Indicator j

Detector Algorithm i

Scenario Generator

Data Source i

Analysis Method- i

Models

Application Wrapper

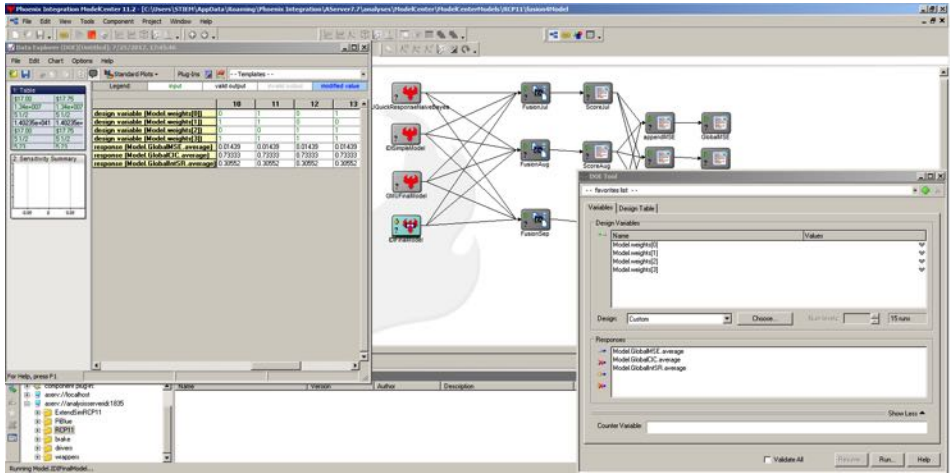
Application Wrapper

Application Wrapper

Application Wrapper

PHOENIX INTEGRATION

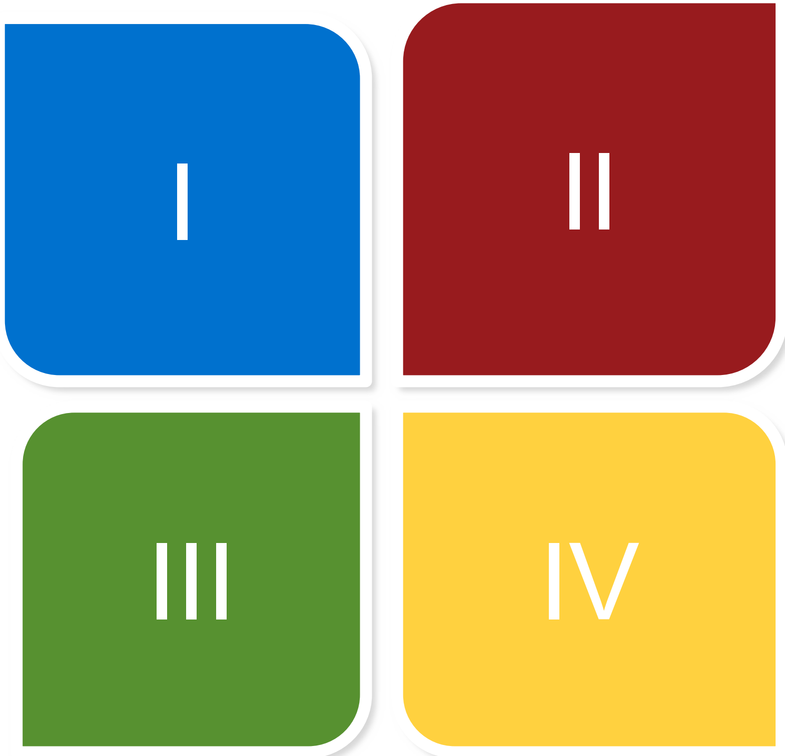
ModelCenter®



MBSE framework for simple integration of components using various software platforms.

Provides variety of tools/methods for analysis/simulation models.

Contributions



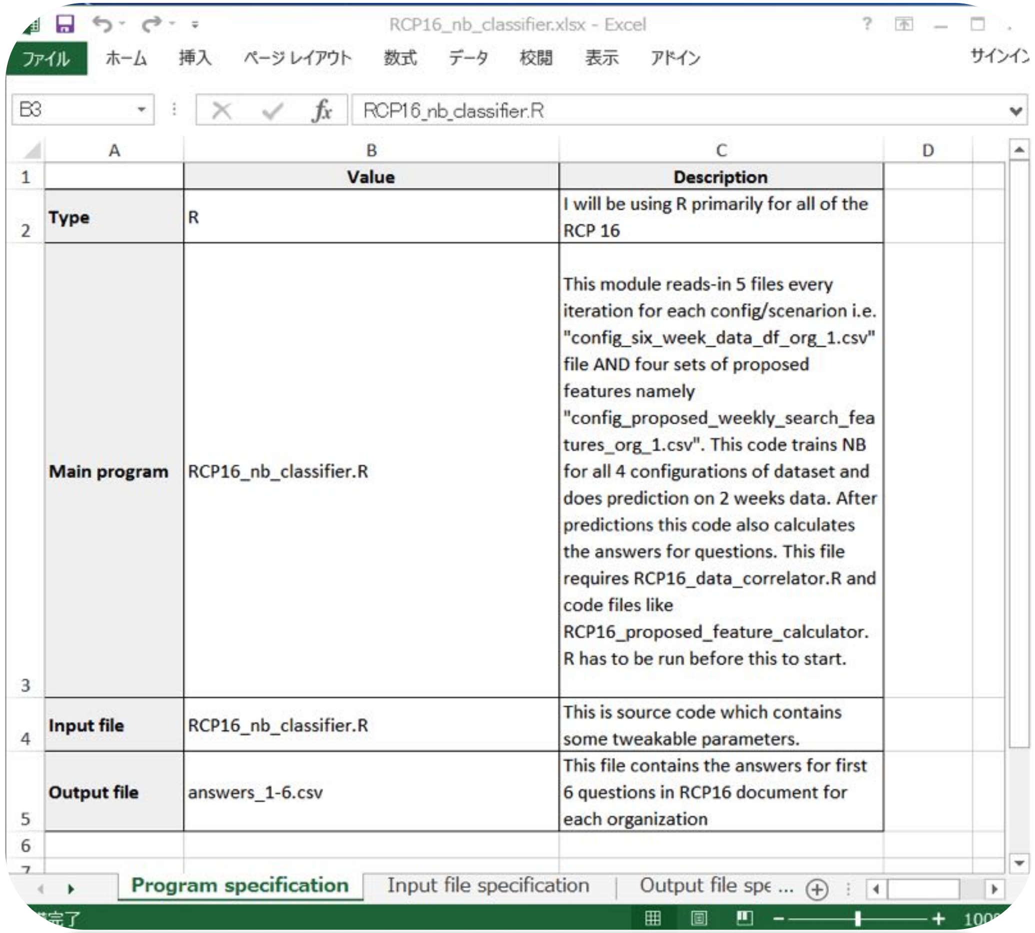
Case study & application of insider threat inference enterprise multi-modeling: workflows & software are managed in STIEM to represent & simulate & analyze different assumptions/scenarios.

Repository of reusable software components: reconstruction of data from summary statistics; machine-learning and inference components like Neural Networks, Decision Trees, Naive Bayes Models, Random Forests, Hidden Markov Models, Support Vector Machines, *etc*.

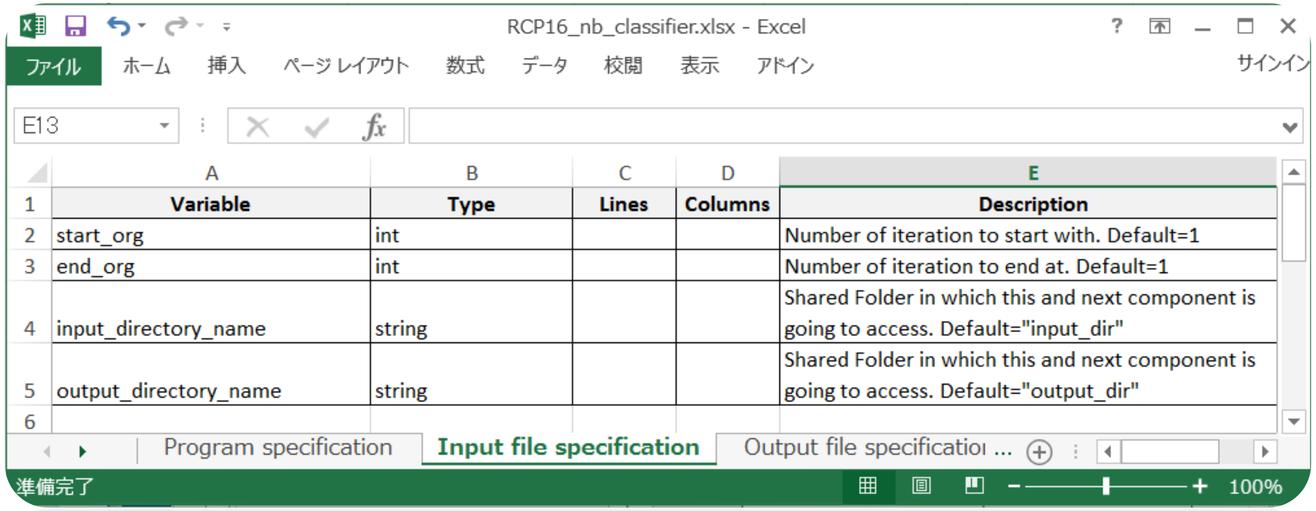
Automatic wrapper generator: translate software interface specification to format compatible with STIEM/ModelCenter®

Platform-independent distributed & asynchronous execution: efficient use of computational resource; reduce overhead of synchronization & process monitoring.

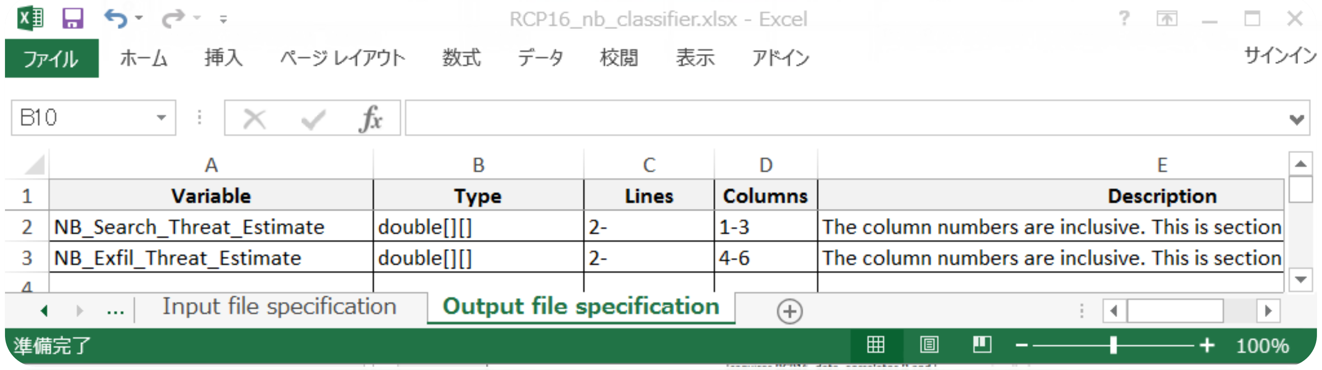
Interface Specification Form



	Value	Description
Type	R	I will be using R primarily for all of the RCP 16
Main program	RCP16_nb_classifier.R	This module reads-in 5 files every iteration for each config/scenarion i.e. "config_six_week_data_df_org_1.csv" file AND four sets of proposed features namely "config_proposed_weekly_search_features_org_1.csv". This code trains NB for all 4 configurations of dataset and does prediction on 2 weeks data. After predictions this code also calculates the answers for questions. This file requires RCP16_data_correlator.R and code files like RCP16_proposed_feature_calculator.R has to be run before this to start.
Input file	RCP16_nb_classifier.R	This is source code which contains some tweakable parameters.
Output file	answers_1-6.csv	This file contains the answers for first 6 questions in RCP16 document for each organization



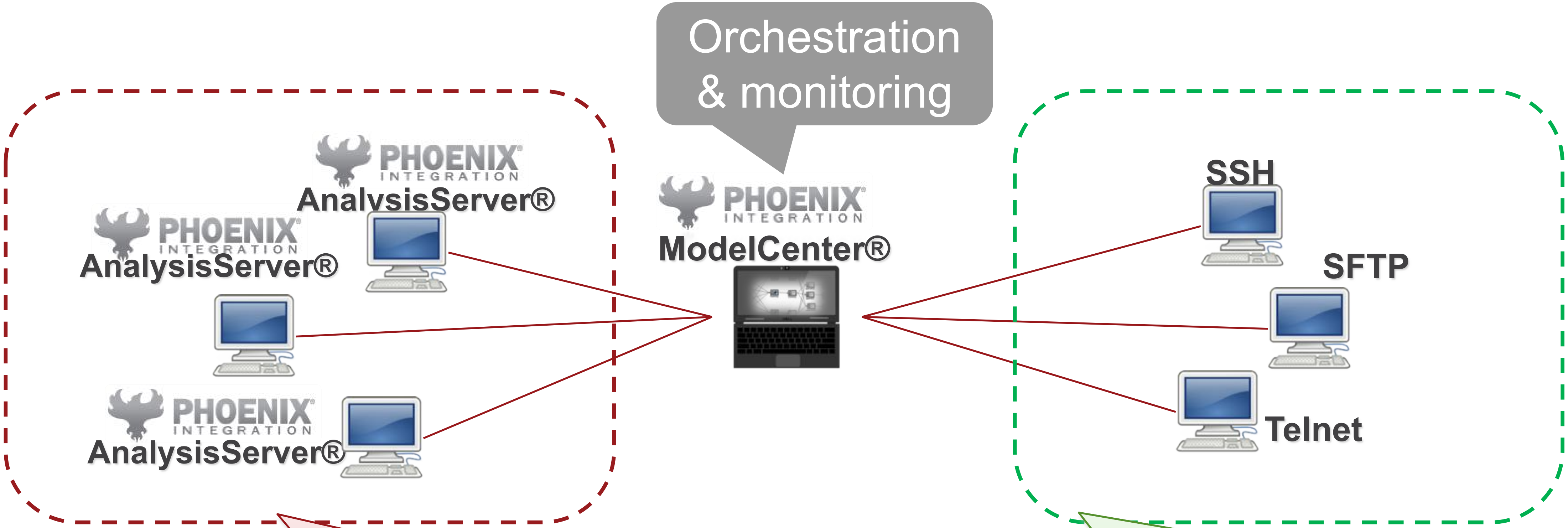
Variable	Type	Lines	Columns	Description
start_org	int			Number of iteration to start with. Default=1
end_org	int			Number of iteration to end at. Default=1
input_directory_name	string			Shared Folder in which this and next component is going to access. Default="input_dir"
output_directory_name	string			Shared Folder in which this and next component is going to access. Default="output_dir"



Variable	Type	Lines	Columns	Description
NB_Search_Threat_Estimate	double[][]	2-	1-3	The column numbers are inclusive. This is section
NB_Exfil_Threat_Estimate	double[][]	2-	4-6	The column numbers are inclusive. This is section

Information in this form is used for generating application wrappers.

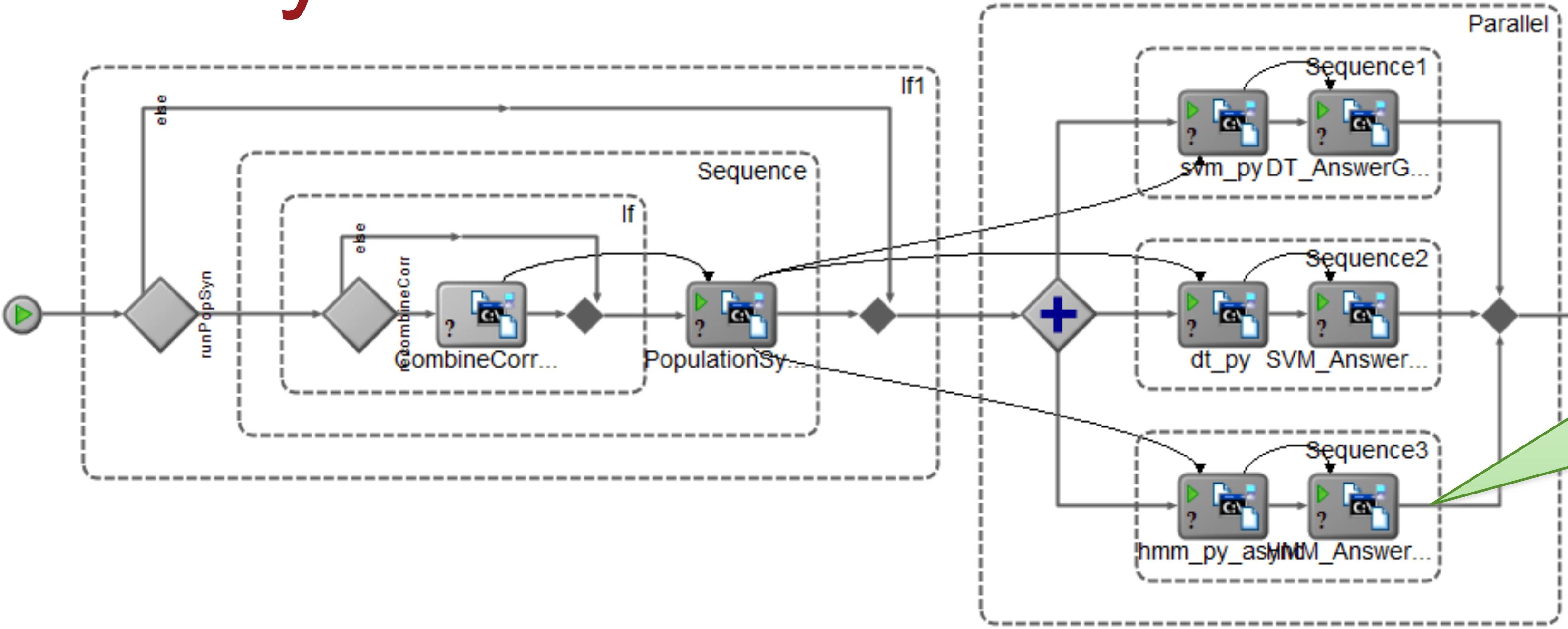
Distributed execution



Distributed execution environment natively supported by ModelCenter® (proprietary).

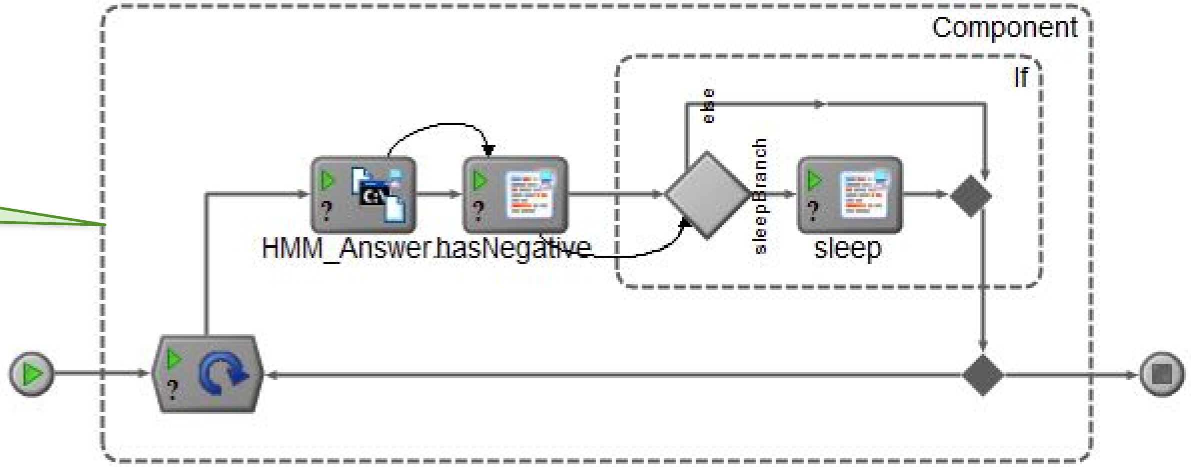
We added quick support for open-standard protocols. Uses same format of interface specification form.

Asynchronous calls



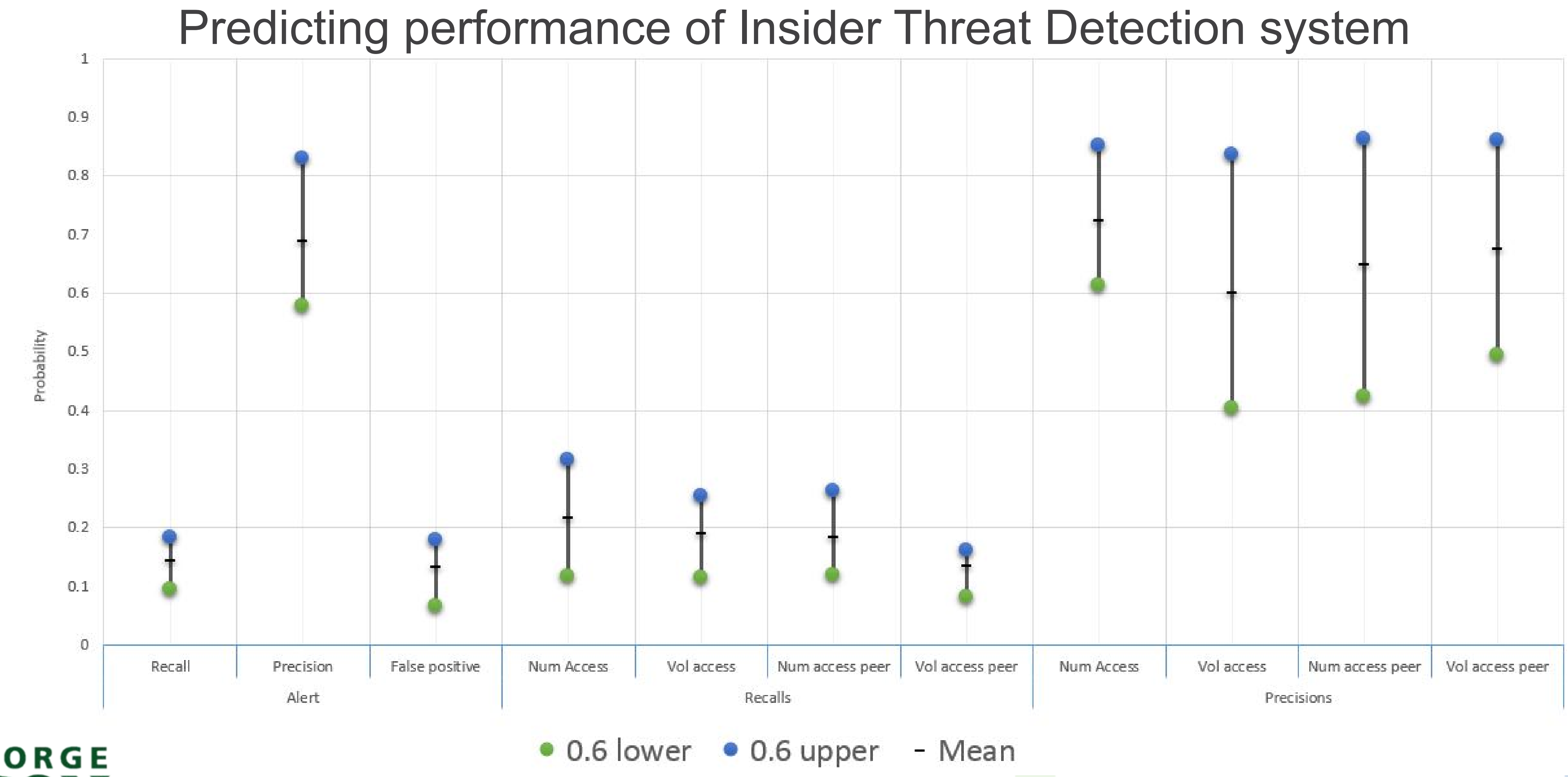
Wrappers written in a way to return immediately (opens background command/process)

Polling workflow *asks* for results later, repeatedly.



Useful when models take long time to finish execution.

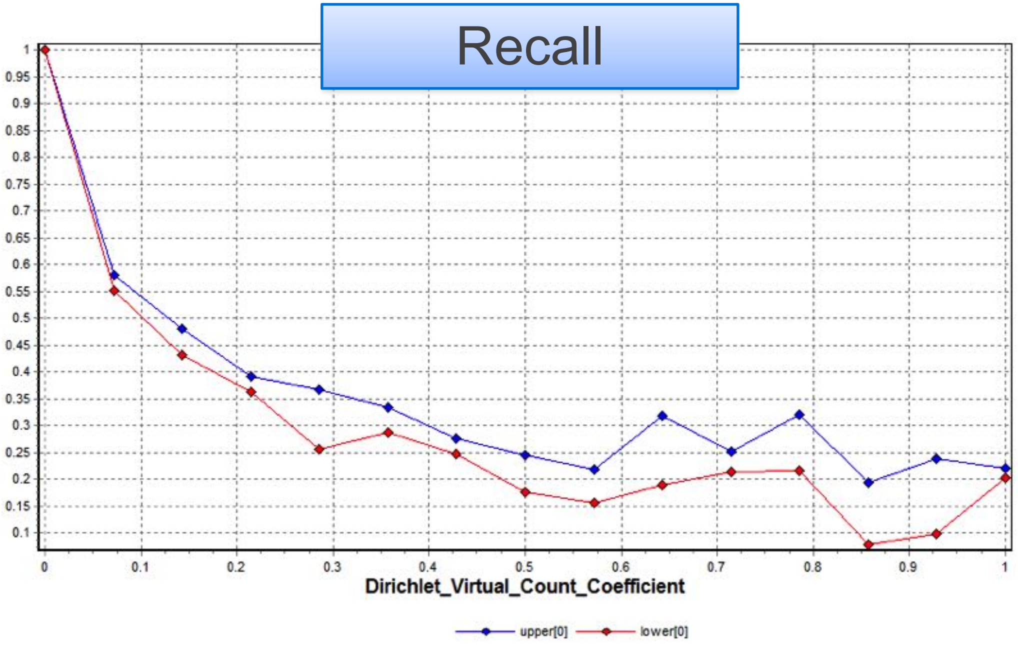
Examples (results)



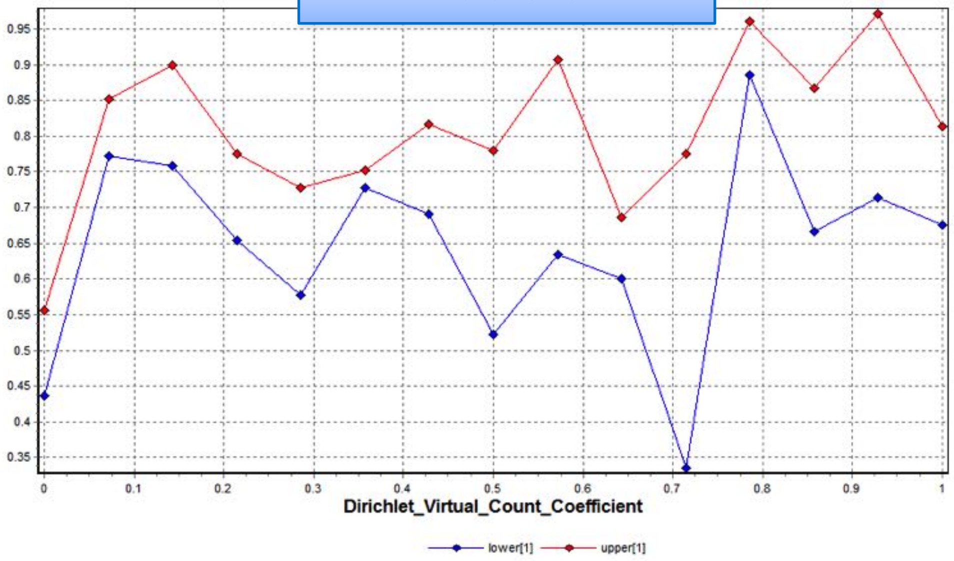
*Data is not real. It's just for illustration purposes.

Examples (results)

Using ModelCenter® built-in plot tool for analyzing how recall/precision changes when virtual count parameters of Dirichlet distribution are changed.



Recall

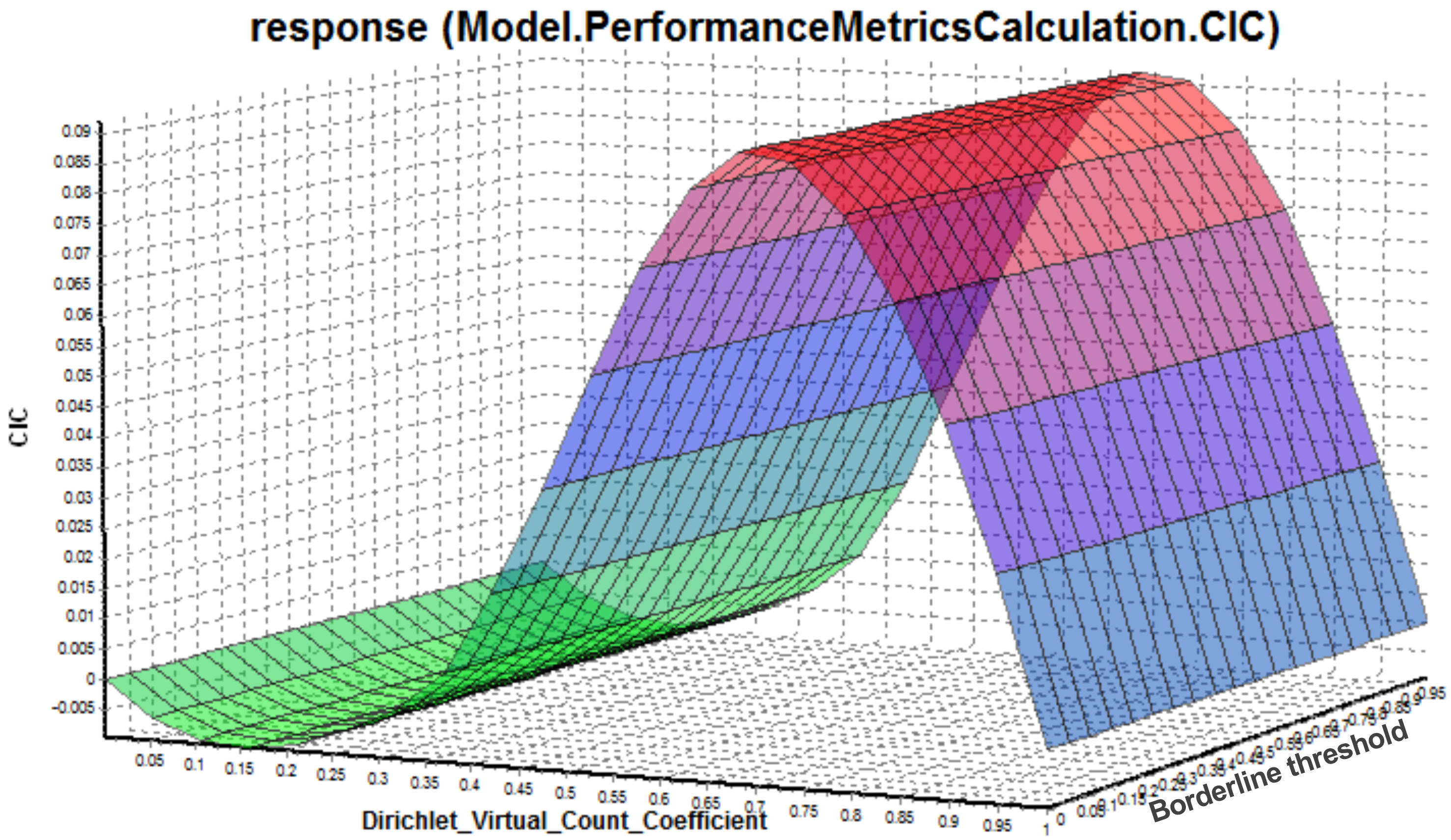


Precision

*Data is not real. It's just for illustration purposes.

Examples (results)

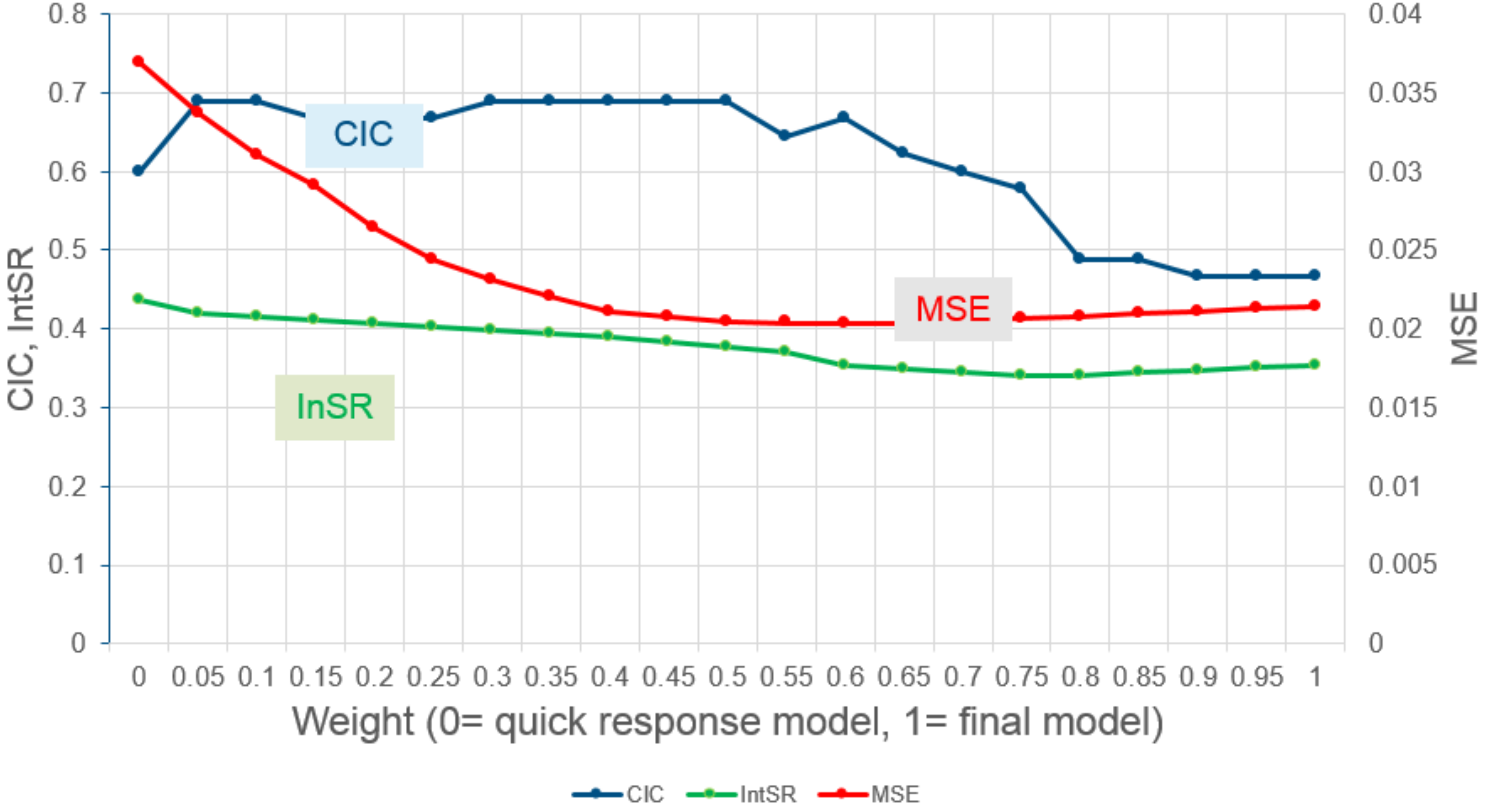
Using ModelCenter® built-in 3D plot tool for analyzing how Coverage (Certainty Interval Calibration—CIC) changes when virtual counts of Dirichlet distribution and threshold of judgements/beliefs in SME data are changed.



*Data is not real. It's just for illustration purposes.

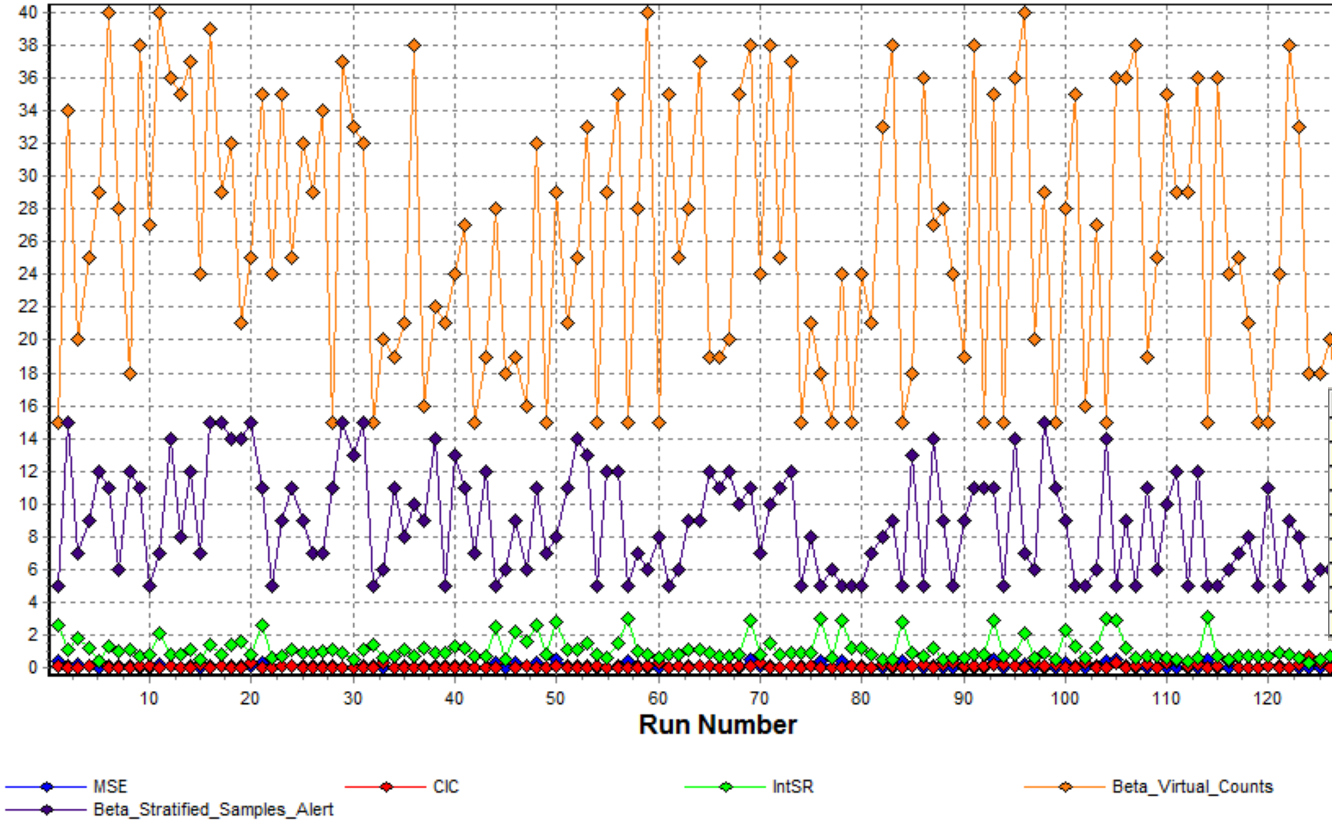
Examples (results)

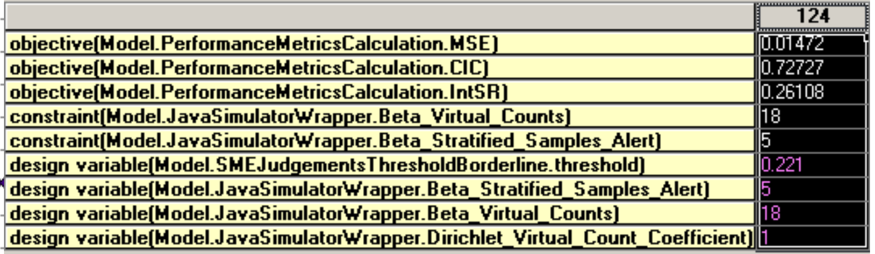
Multi-scale plot of how some performance scores change when merging results from two models and changing weights.



*Data is not real. It's just for illustration purposes.

Examples (results)





	124
objective(Model.PerformanceMetricsCalculation.MSE)	0.01472
objective(Model.PerformanceMetricsCalculation.CIC)	0.72727
objective(Model.PerformanceMetricsCalculation.IntSR)	0.26108
constraint(Model.JavaSimulatorWrapper.Beta_Virtual_Counts)	18
constraint(Model.JavaSimulatorWrapper.Beta_Stratified_Samples_Alert)	5
design variable(Model.SMEJudgementsThresholdBorderline.threshold)	0.221
design variable(Model.JavaSimulatorWrapper.Beta_Stratified_Samples_Alert)	5
design variable(Model.JavaSimulatorWrapper.Beta_Virtual_Counts)	18
design variable(Model.JavaSimulatorWrapper.Dirichlet_Virtual_Count_Coefficient)	1

Using ModelCenter® built-in simulation/optimization tool to find configuration with best performance metrics.

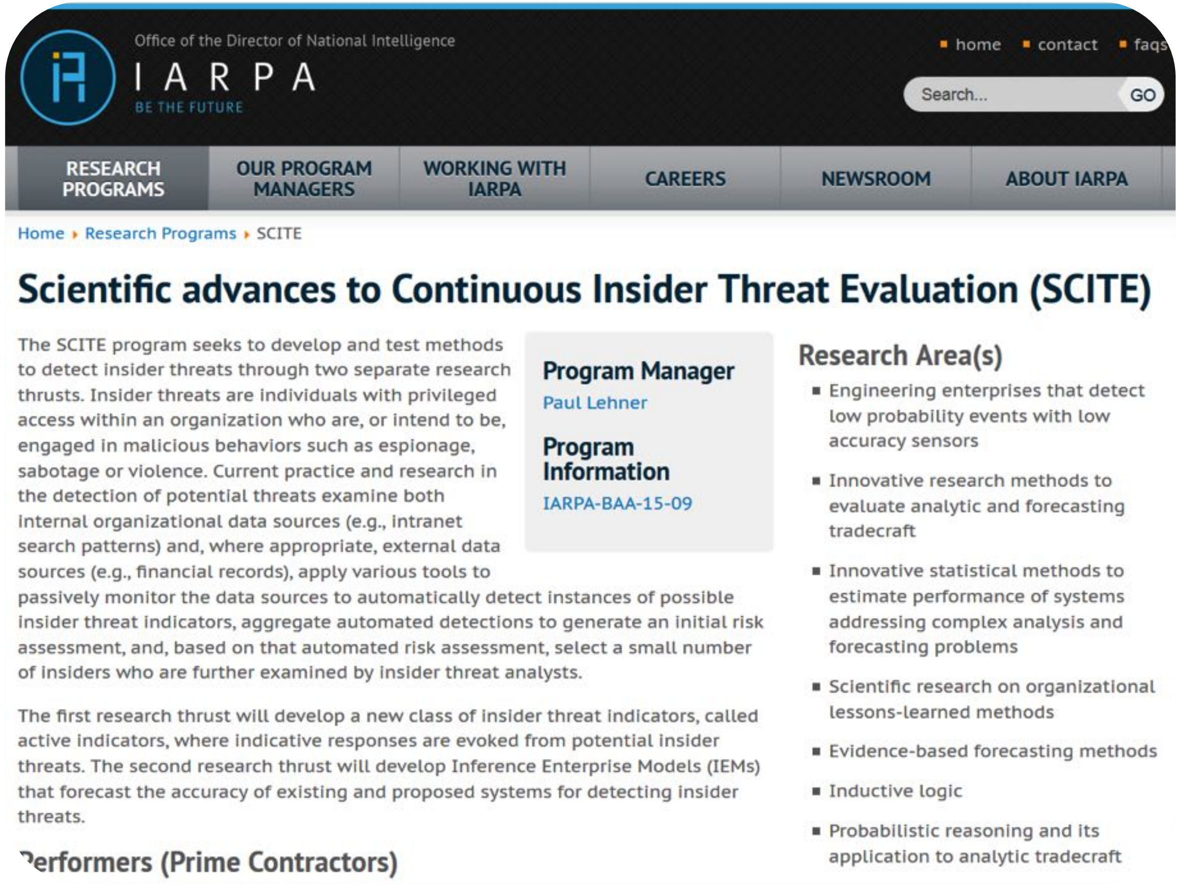
*Data is not real. It's just for illustration purposes.

Concluding remarks

- STIEM is a model-based engineering framework that provides wide variety of tools & methods, storing them as reusable components and creating multi-modeling workflows with different data sources.
  - Suitable for designing, integrating and executing/orchestrating software components with distinct and visible input and output interfaces (*e.g.* input/output files).
- This approach is useful to a broad set of areas that involve multiple analysis models
  - Examples may include defense/national security, homeland security, intelligence analysis, information security, etc.

This work is part of…

- Scientific Advances to Continuous Insider Threat Evaluation (SCITE)
- Objective of the project:
  - evaluate, understand, and improve performance of automated threat detection system



Scientific advances to Continuous Insider Threat Evaluation (SCITE)

The SCITE program seeks to develop and test methods to detect insider threats through two separate research thrusts. Insider threats are individuals with privileged access within an organization who are, or intend to be, engaged in malicious behaviors such as espionage, sabotage or violence. Current practice and research in the detection of potential threats examine both internal organizational data sources (e.g., intranet search patterns) and, where appropriate, external data sources (e.g., financial records), apply various tools to passively monitor the data sources to automatically detect instances of possible insider threat indicators, aggregate automated detections to generate an initial risk assessment, and, based on that automated risk assessment, select a small number of insiders who are further examined by insider threat analysts.

The first research thrust will develop a new class of insider threat indicators, called active indicators, where indicative responses are evoked from potential insider threats. The second research thrust will develop Inference Enterprise Models (IEMs) that forecast the accuracy of existing and proposed systems for detecting insider threats.

Program Manager
Paul Lehner

Program Information
IARPA-BAA-15-09

Research Area(s)

- Engineering enterprises that detect low probability events with low accuracy sensors
- Innovative research methods to evaluate analytic and forecasting tradecraft
- Innovative statistical methods to estimate performance of systems addressing complex analysis and forecasting problems
- Scientific research on organizational lessons-learned methods
- Evidence-based forecasting methods
- Inductive logic
- Probabilistic reasoning and its application to analytic tradecraft

Performers (Prime Contractors)

www.iarpa.gov/index.php/research-programs/scite

Our team

Innovative Decisions, Inc.

- Lead organization;
- Bayesian modeling
- Monte Carlo modeling
- Psychological modeling
- Ensemble model aggregation

George Mason University

- Bayesian reasoning
- Ontology development
- System architectures
- Multi-modeling

Dr. Yung Mei Leong, Consultant

Human Resources Research Organization

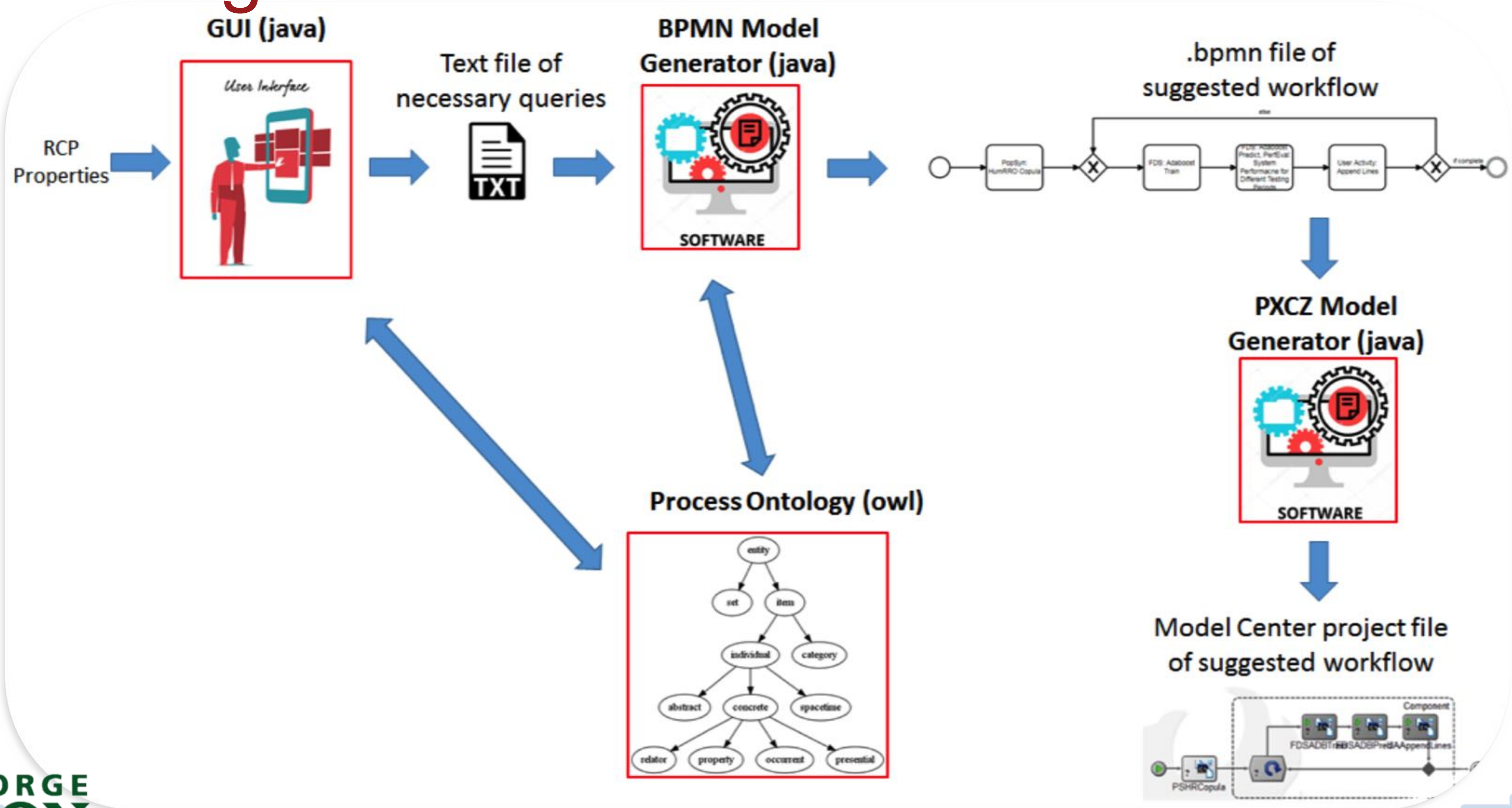
- Bayesian models
- Multi-model Integration
- Quantitative psychology/ statistical model development

PsyberAnalytix LLC

- Insider threat subject matter expertise
- Quantitative psychology and performance evaluation
- Human behavior modeling
- Cybersecurity/Info Analysis R&D

Many thanks to the team.

Ongoing/future work: semi-automatic workflow generation



We thank Mr. James Lee & Dr. Abbas Zaidi (GMU) for assistance with this method.

28th Annual INCOSE international symposium

Washington, DC, USA
July 7 - 12, 2018

www.incose.org/symp2018

Thank you for your attention.