Inference Enterprise Modeling for Insider Threat Detection Systems

Shou Matsumoto, Edward Huang, Kathryn B. Laskey
Systems Engineering & Operations Research Department,
George Mason University

David Brown, Sean Vermillion
Innovative Decision, Inc.

Research reported here was supported under IARPA contract 2016-16031400006. The content is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. Government.

Agenda

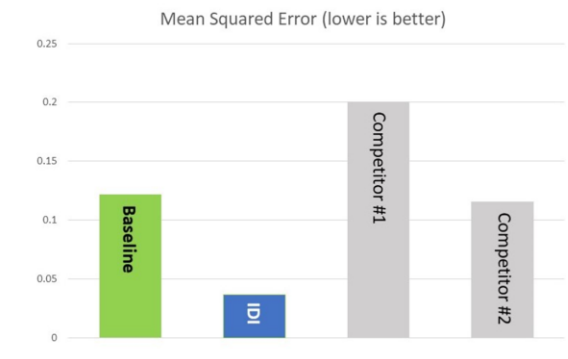
- Introduction
- Insider threat detection systems
- Multi-modeling approach for inference enterprise modeling
- Model integration and sensitivity analysis using ModelCenter®
- Conclusion

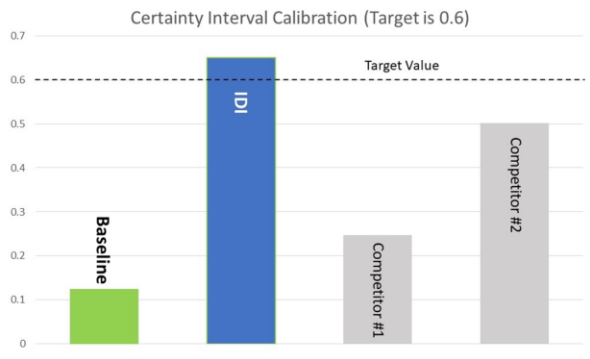
Scientific advances to Continuous Insider Threat Evaluation (SCITE)

- SCITE is an Intelligence Advanced Research Project Agency (IARPA) sponsored research program that seeks to advance the science and practice of insider threat detection
- The GMU track is specifically focused at research into modeling and forecasting the performance of existing and proposed insider threat detection enterprises
- The first phase of the research consisted of three competing teams that were evaluated on their ability to solve a series of increasingly complex challenge problems created by MIT Lincoln Labs
  - GMU integrated our multi-model approach using Phoenix Integration ModelCenter®

SCITE Competition Results

GMU was selected as part of the single transition team due to superior performance on a set of increasingly complex challenge problems





Phoenix Integration ModelCenter® was a major contributor to our success

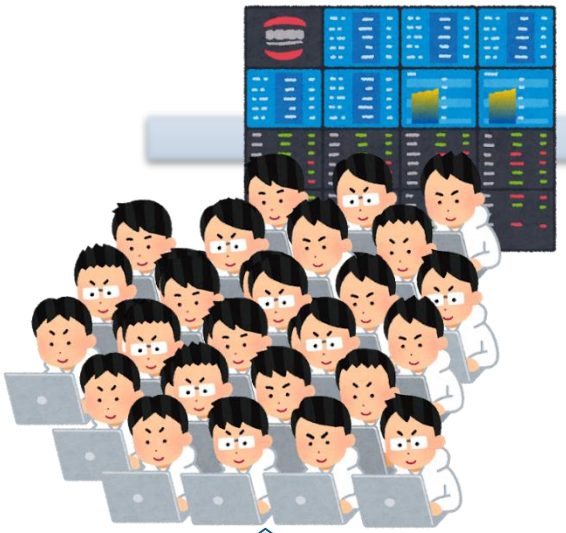
Definition

Insider threat:

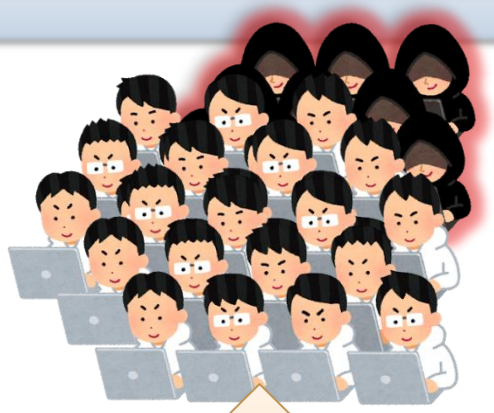
- An individual (or individuals) who….
  - is a current or former employee, contractor, or other business partner
  - has or had authorized access to an organization's network, system, or data
  - intentionally (or unintentionally) exceeds or misuses that access to negatively affect the confidentiality, integrity, or availability of the organization's information or information systems

[CERT, 2012]

Background story: current practice of insider threat detection system



Monitor activities of large, diverse population (workers)

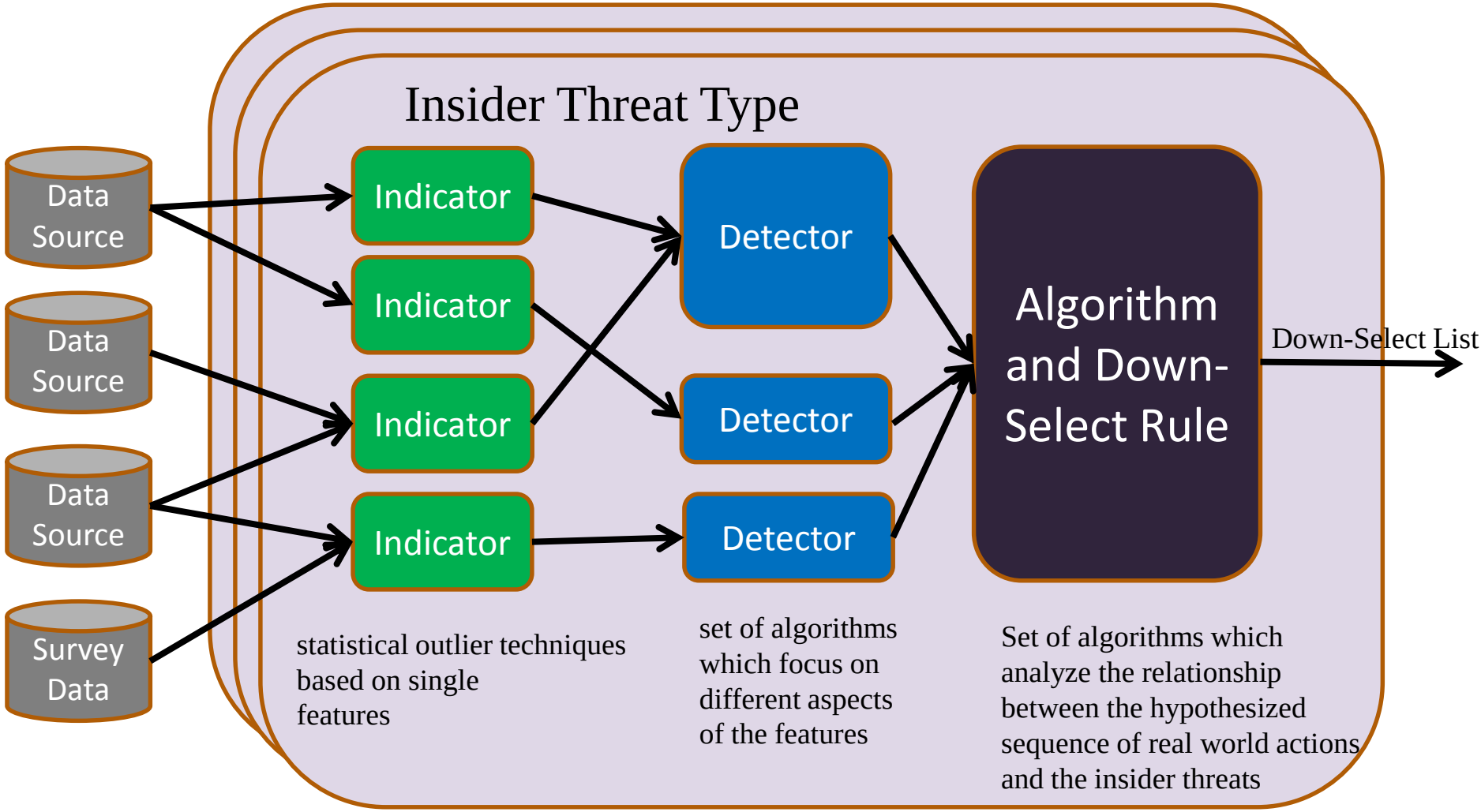


Apply algorithms which identify and use indicators and detectors associated with some potential threat

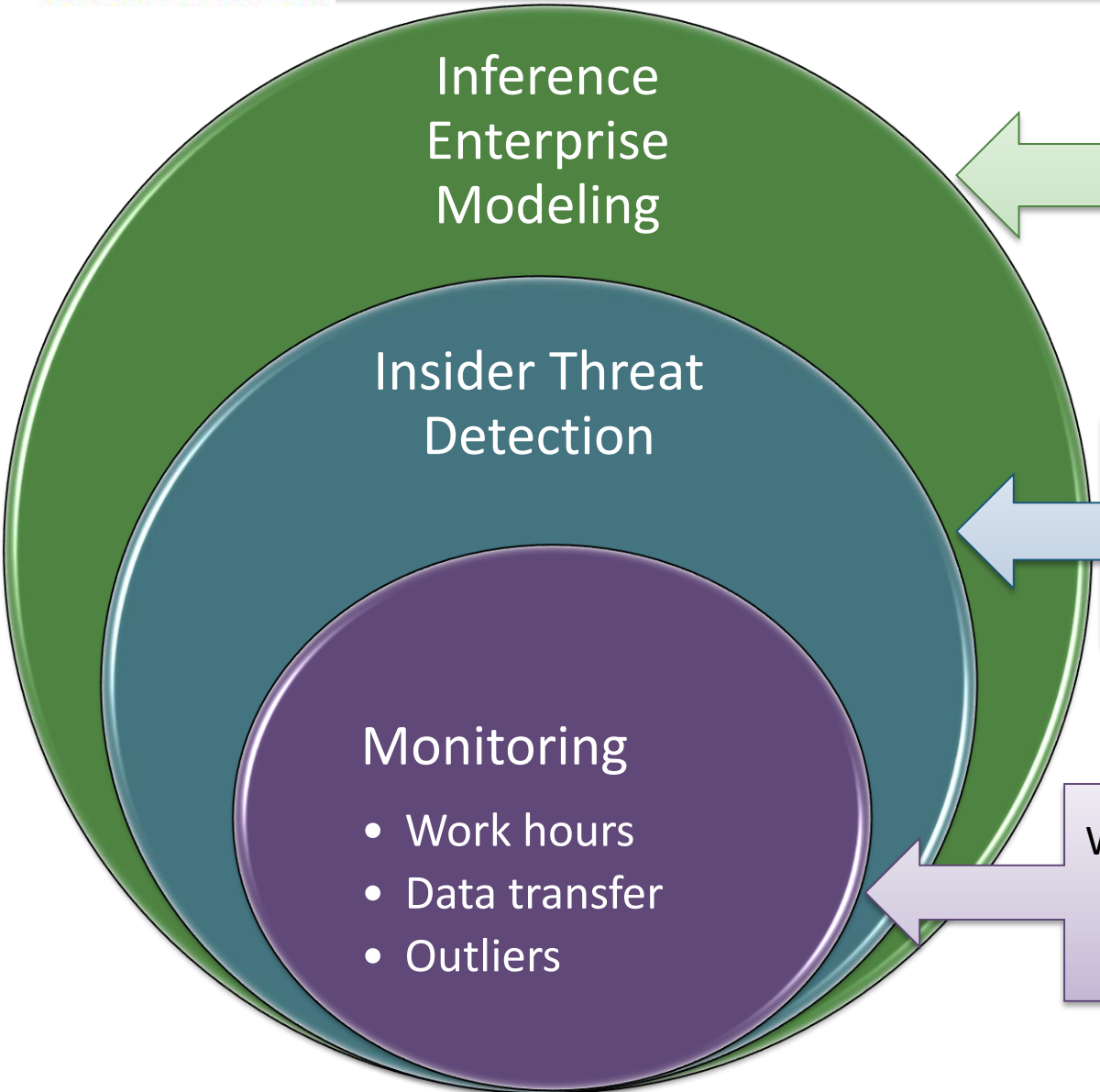
Apply algorithms to down-select "persons of interest" for further investigation

Task: build a model to represent/evaluate/predict performance of this system in identifying threats.

Inference Enterprise



Defining the domain and narrowing down the scope



Modeling computational representations of the enterprise (*e.g.* structure, process/activities, goals…)

Detecting individuals who work in organization and exceed/misuse access to resources

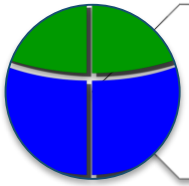
We'll focus on those Insider Threat Detection systems which monitor individuals' activities

Examples of target behaviors (potential threats)

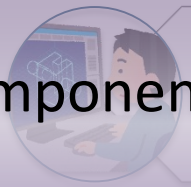
1. Individuals who uses work-owned machine outside normal work hours
   - Data: logs of anti-virus updates, VPN connections, logins, emails, website access outside working hours

2. Outlying changes in web use
   - Data: history of proxy log entries and download logs

Technical problems and solutions

Problem	Approach	Adopted solution
1. Imprecise and/or conflicting data	• Search for joint distribution which is as close as possible to all data	 Stochastic optimization and/or Monte-Carlo simulation
2.Incomplete data (*e.g.* no data about target)	• Ask Subject Matter Experts judgments/beliefs • Simulate based on available data	 Reduced to problem 1
3. Large dimension	• Exploit conditional (in)dependence	 "Factored" Stochastic Optimization with Markov nets
4. Too many potential answers/assumptions to explore	• Quick prototyping, integration, test, evaluation and analysis of multiple answers/solutions	 Perform sensitivity analysis using ModelCenter®

Technical problems and solutions

Problem	Approach	Adopted solution
1. Imprecise and/or conflicting data	• Search for joint distribution which is as close as possible to all data	 Stochastic optimization and/or Monte-Carlo simulation
2.Incomplete data (*e.g.* no data about target)	• Ask Subject Matter Experts • Simulate based on available data	 Reduced to problem 1
3. Large dimension	• Exploit conditional (in)dependence	 "Factored" Stochastic Optimization with Markov nets
4. Too many potential answers (due to multiple assumptions, uncertainty)	• Quick prototyping, integration, test, evaluation and analysis of multiple answers/solutions	 Perform sensitivity analysis using ModelCenter®

Construction of models from data

Quick integration of models/components

Research Objectives

- The goal of this research is
  - To evaluate how well the enterprise's automated threat detection system performs at detecting threats,
  - To understand the reasons for its performance, and
  - To identify ways to improve performance
- Specific objectives:
- Use Phoenix Integration ModelCenter® as an experimental test-bed where existing and proposed inference enterprises can be tested and vetted.
- Use empirical studies to develop an understanding of the performance of an inference enterprises. Develop a quantitative measure of fitness of an inference enterprises for a given organization's needs.
- Provide a capability for risk and cost-benefit analysis for the alternative solutions

Redacted Data

		Indicator 1	
		TRUE	FALSE
Threat	TRUE	211	63
	FALSE	1	3529

		Indicator 2	
		TRUE	FALSE
Threat	TRUE	211	63
	FALSE	1	3529

		Indicator 2	
		TRUE	FALSE
Indicator 1	TRUE	31	184
	FALSE	183	3869

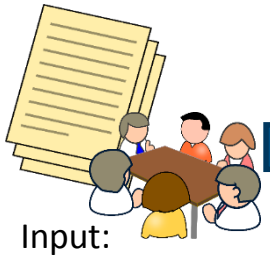
		Detector 1	
		TRUE	FALSE
Indicator 1	TRUE	11	3
	FALSE	5	3785

		Detector 2	
		TRUE	FALSE
Indicator 2	TRUE	200	61
	FALSE	1	3542

		Detector 3	
		TRUE	FALSE
Indicator 3	TRUE	31	84
	FALSE	83	4069

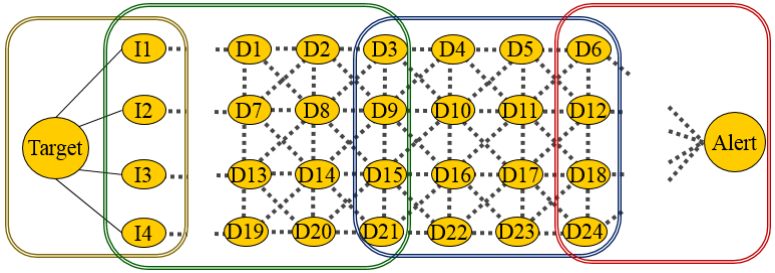
* Data is not real and just for illustration purposes.

Multi-Modeling Approach



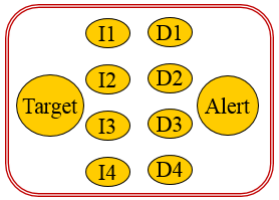
Input: contingency tables of pairs of variables (from data or beliefs)

1



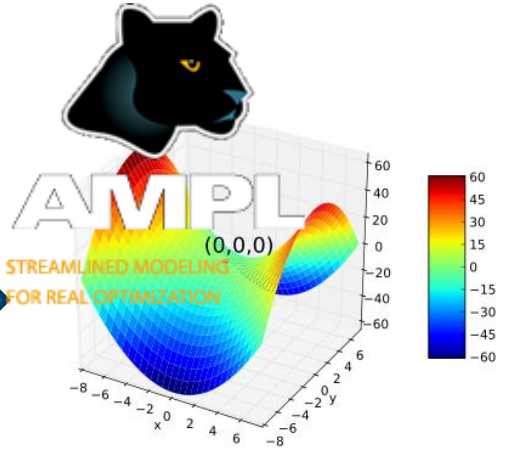
OR

1



If number of variables is high, build Markov net to explore conditional (in)dependence for factorization

2



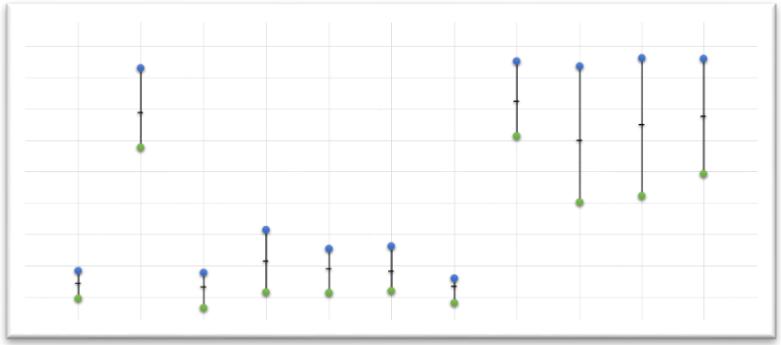
Use non-linear program solver to find joint distribution which matches w/ data (*i.e.* maximizes entropy or log-likelihood)

3

Org1	Detectors	Indicators	...	Alert?	Target?
Person1	0	0	...	no	no
Person2	1	1	...	no	no
⋮	⋮	⋮	⋮	⋮	⋮
PersonN	1	0	...	yes	yes

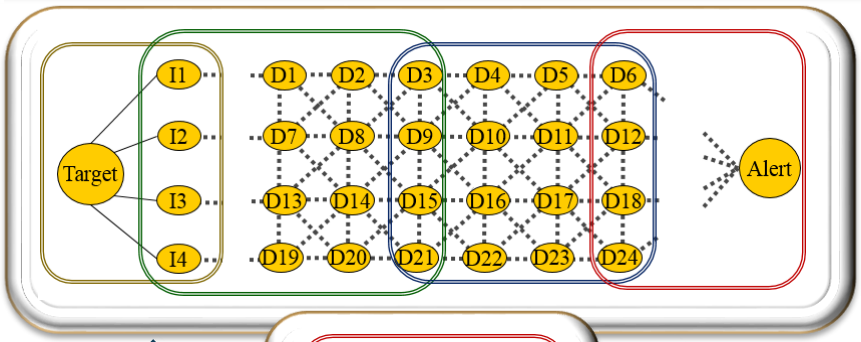
Use Dirichlet-multinomial sampler to simulate populations

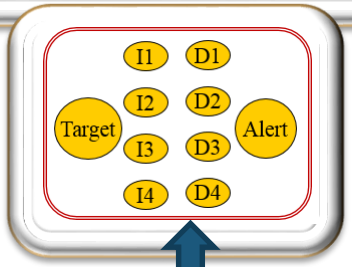
4

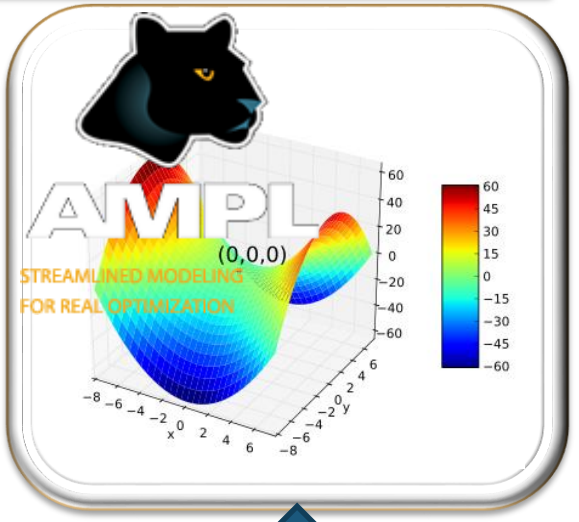


Calculate recalls, precision, statistics…

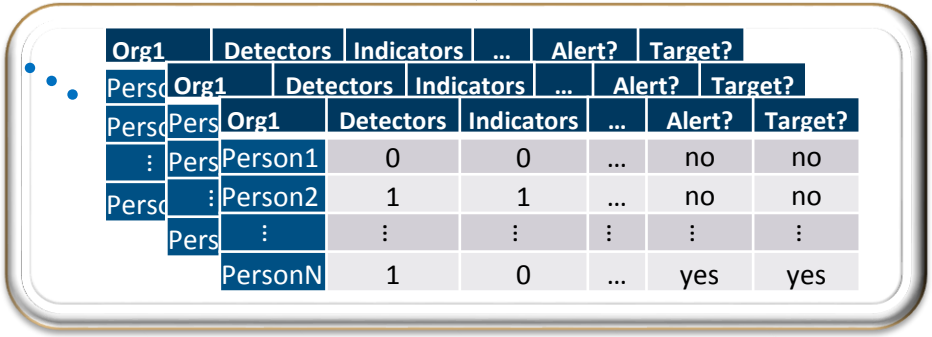
Model-based prototyping, integration, simulation, test, evaluation and analysis



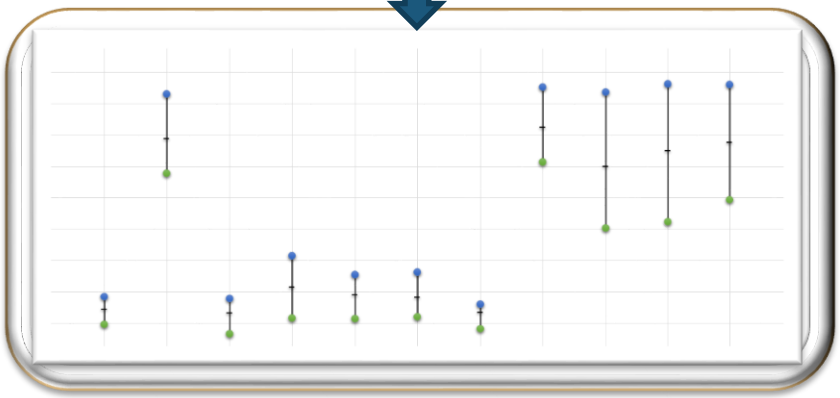




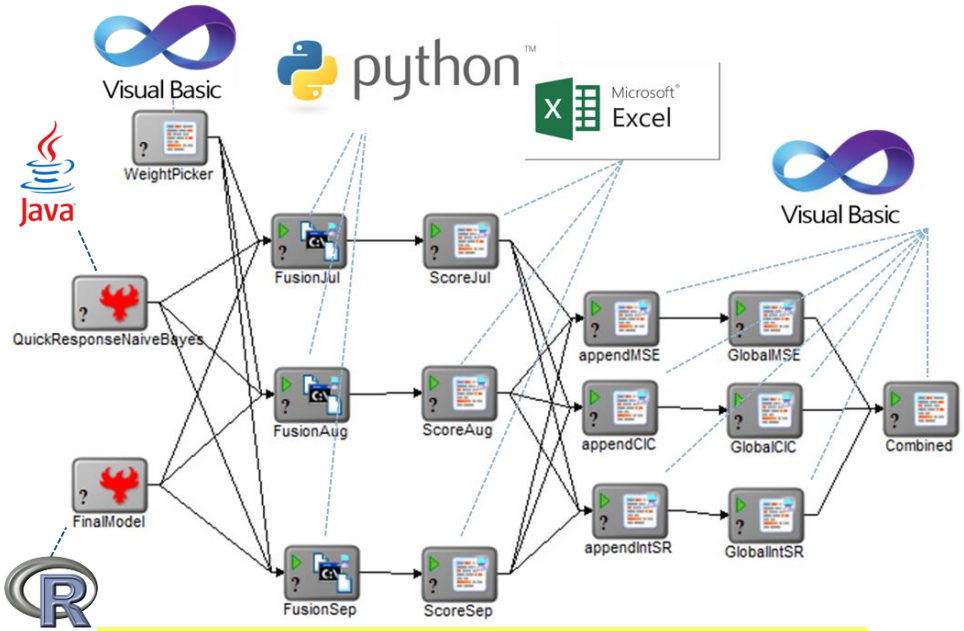
Phoenix Integration ModelCenter®



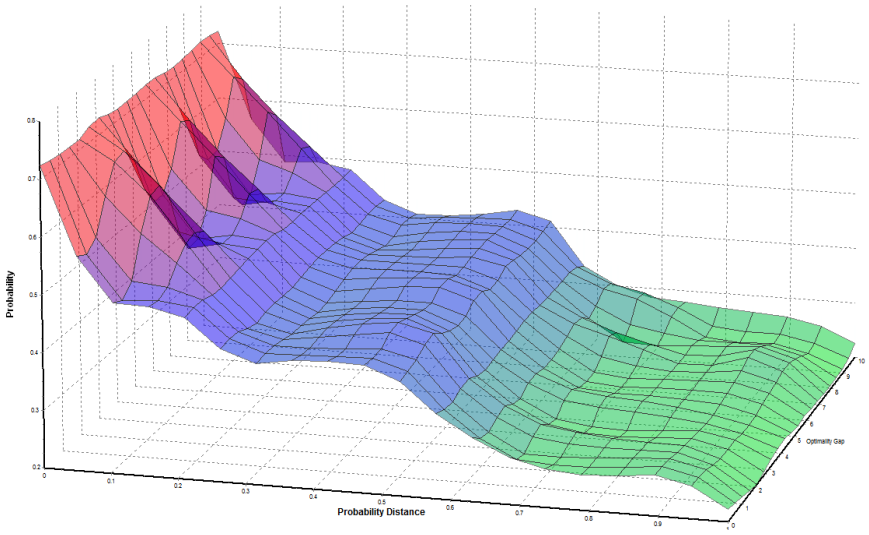
Org1	Detectors	Indicators	...	Alert?	Target?
Person1	0	0	...	no	no
Person2	1	1	...	no	no
⋮	⋮	⋮	⋮	⋮	⋮
PersonN	1	0	...	yes	yes



Capabilities of ModelCenter®



Multi-Model Integration Workflow



Sensitivity Analysis of Whole System Performance

Parallel/Distributed Execution

Wrapper Implementations

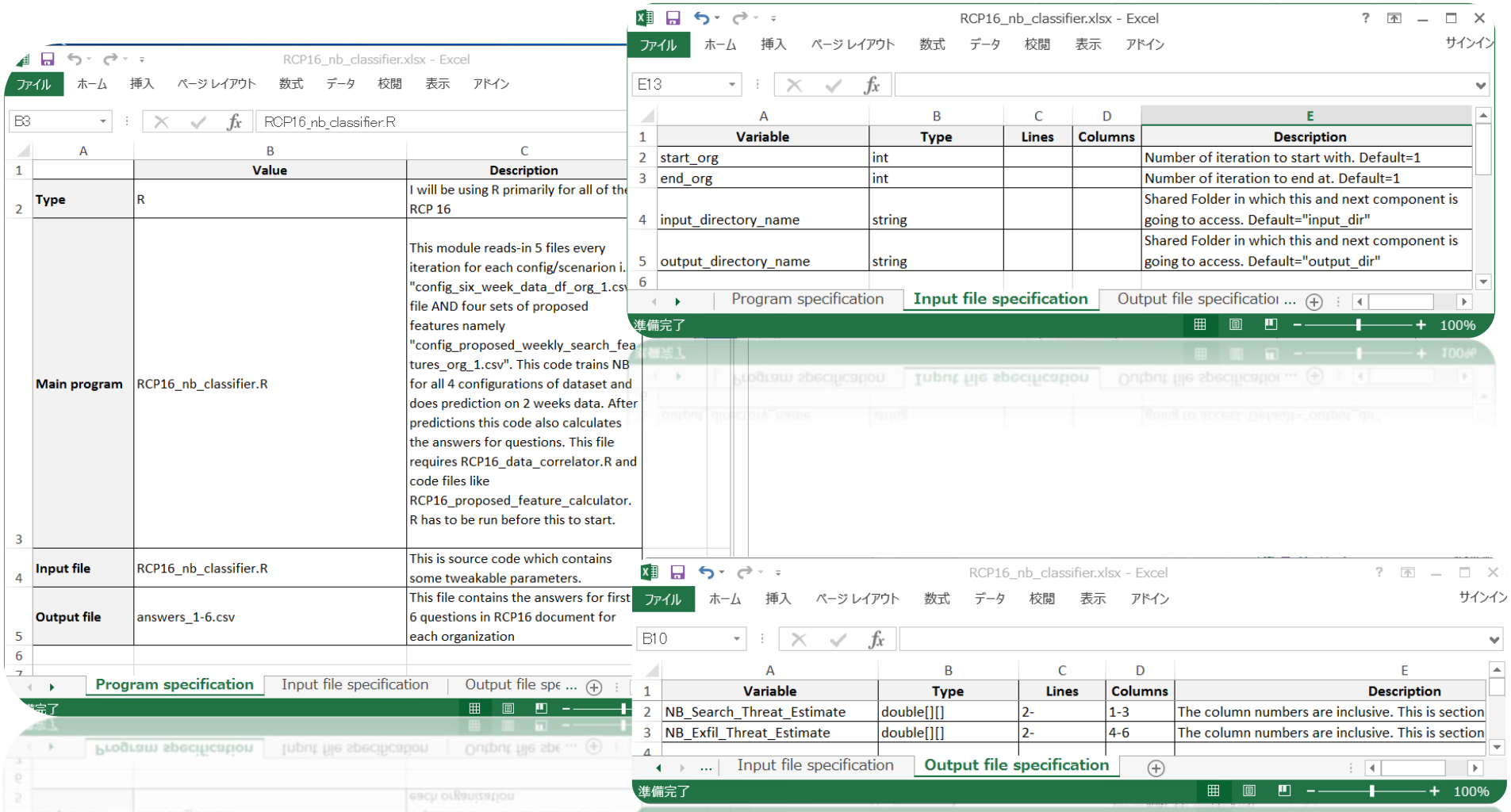
- Support Matlab, MS Excel, R, Python, C# and Java programming languages
- Support Bayesian Network, Optimization, and Simulation models implemented in Netica™, AMPL™, or ExtendSim™
- Support for UnBBayes Noisy-OR/MAX models
- Support for Python scikit-learn decision tree models
- Support for Tree-Augmented Naïve Bayes models
- Support for Neural Network models

Wrapper Implementations

	Full Automation (No coding required)	Manual Setting
R		X
Python		X
C#		X
Java		X
Netica	X	O
AMPL	X	O
ExtendSim	O	X
UnBBayes	O	X

X: Developed wrappers
O: Evaluated as technically feasible, but not implemented yet

Interface Specification Form



RCP16_nb_classifier.xlsx - Excel

B3: RCP16_nb_classifier.R

	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. "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

Program specification | Input file specification | Output file spe ...

RCP16_nb_classifier.xlsx - Excel

E13

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"

Program specification | Input file specification | Output file specificatio ...

RCP16_nb_classifier.xlsx - Excel

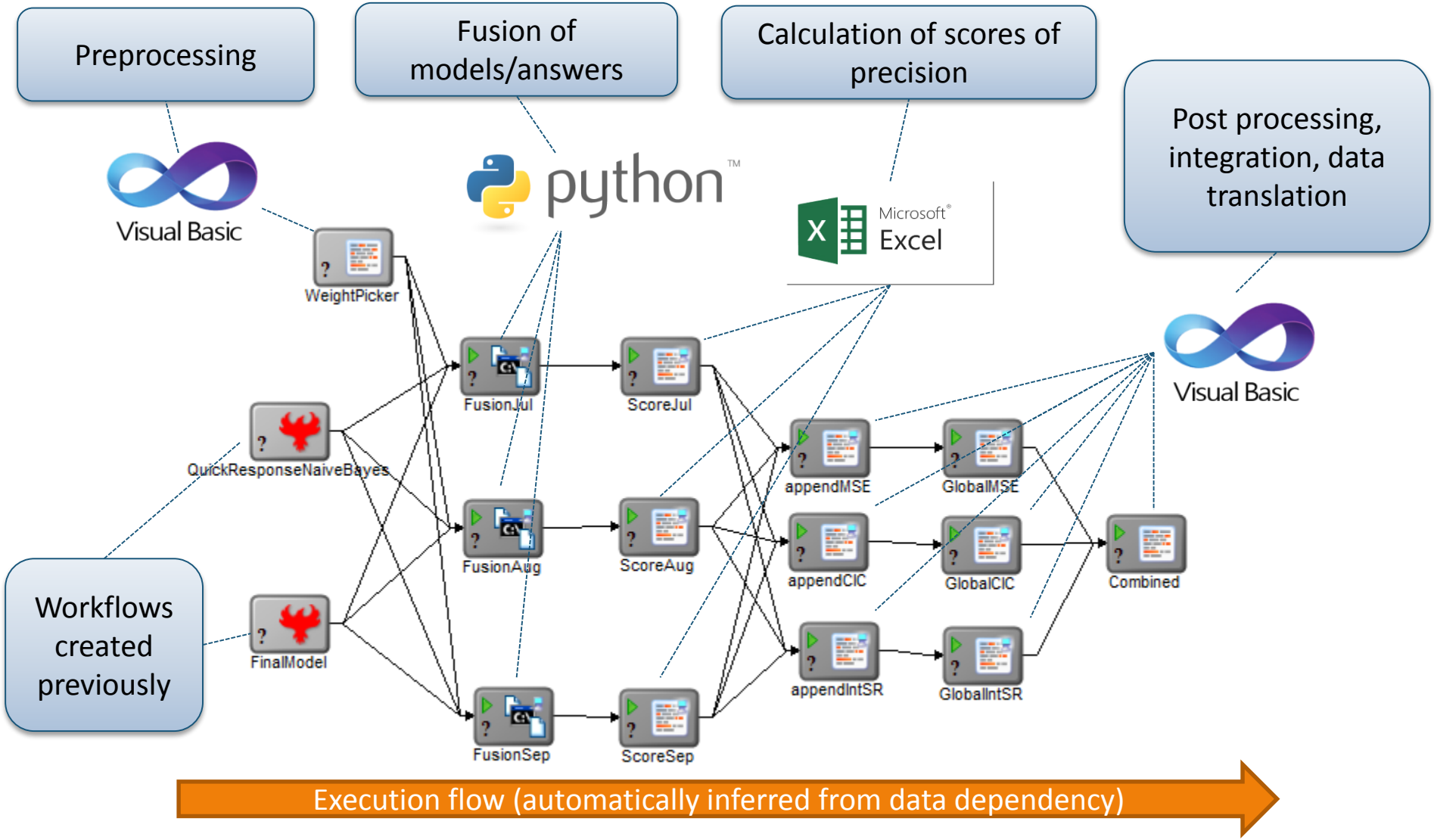
B10

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

... | Input file specification | Output file specification

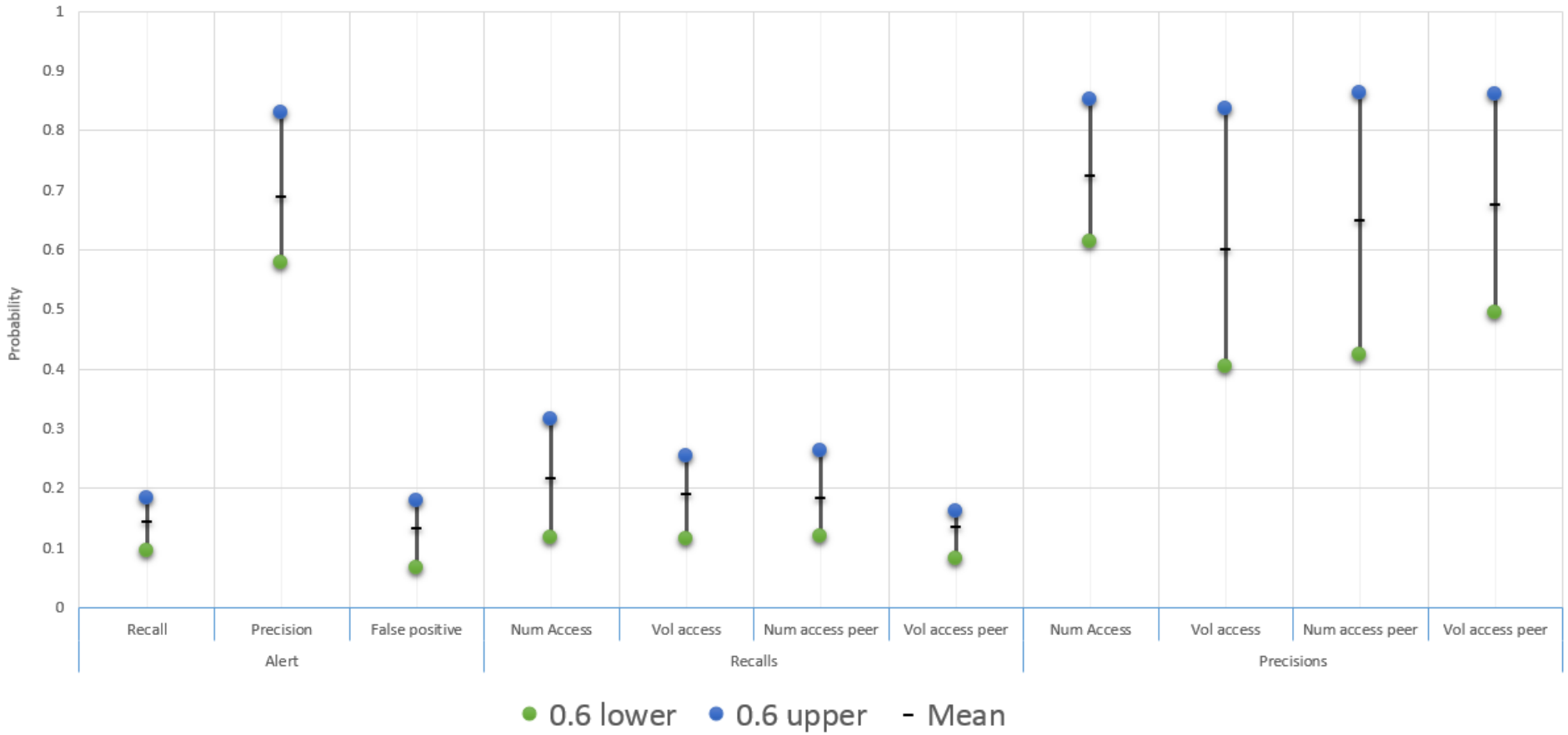
Information in this form is used for generating file wrappers.

Multi-Modeling Computational Process



Some Results

Predicting performance of Insider Threat Detection system

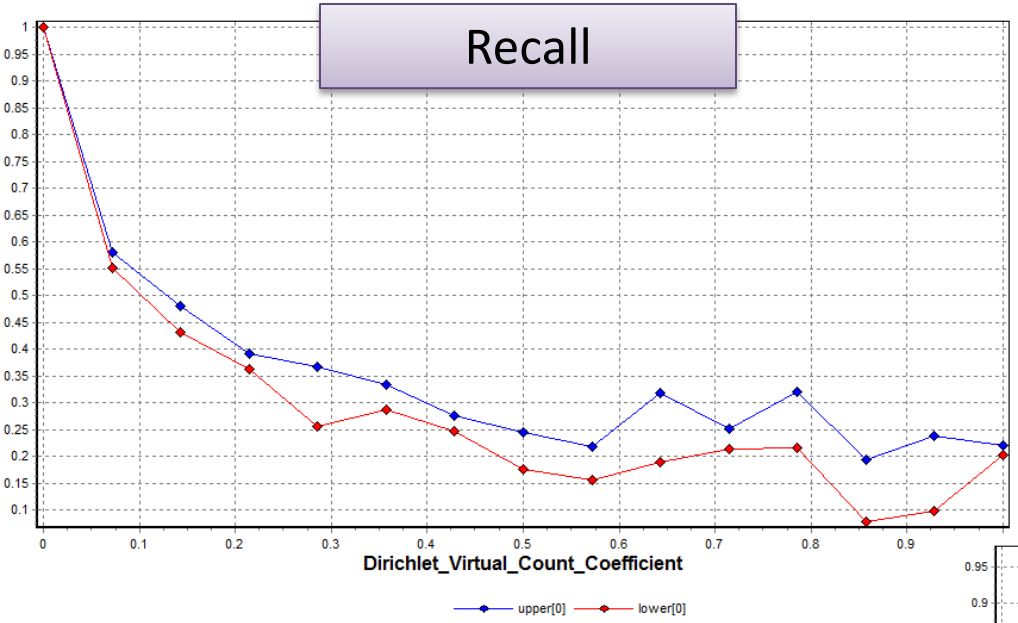


* Data is not real and just for illustration purposes.

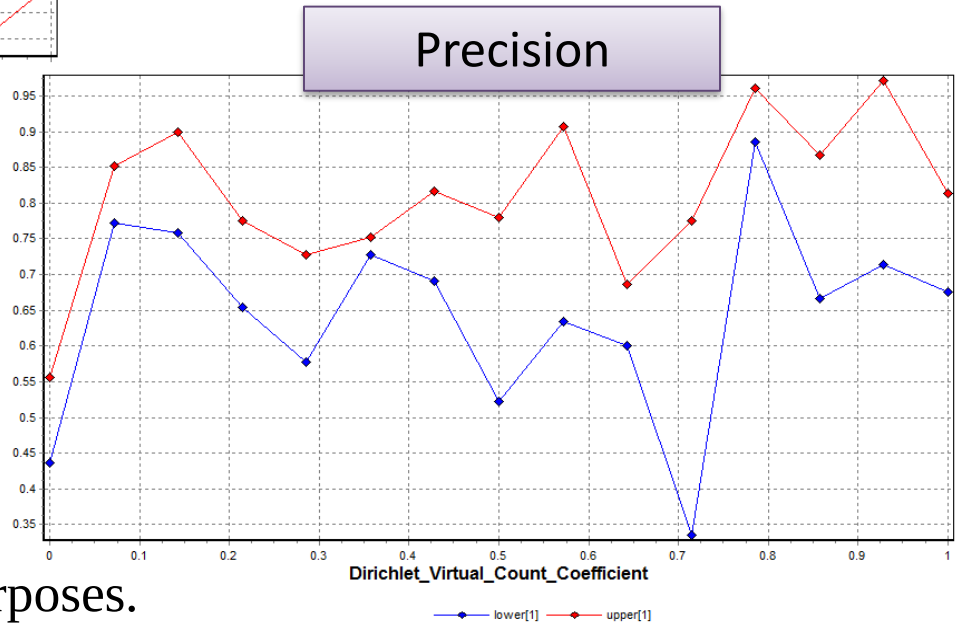
Some Results

How recall/precision changes when virtual counts of Dirichlet distribution are changed

Recall



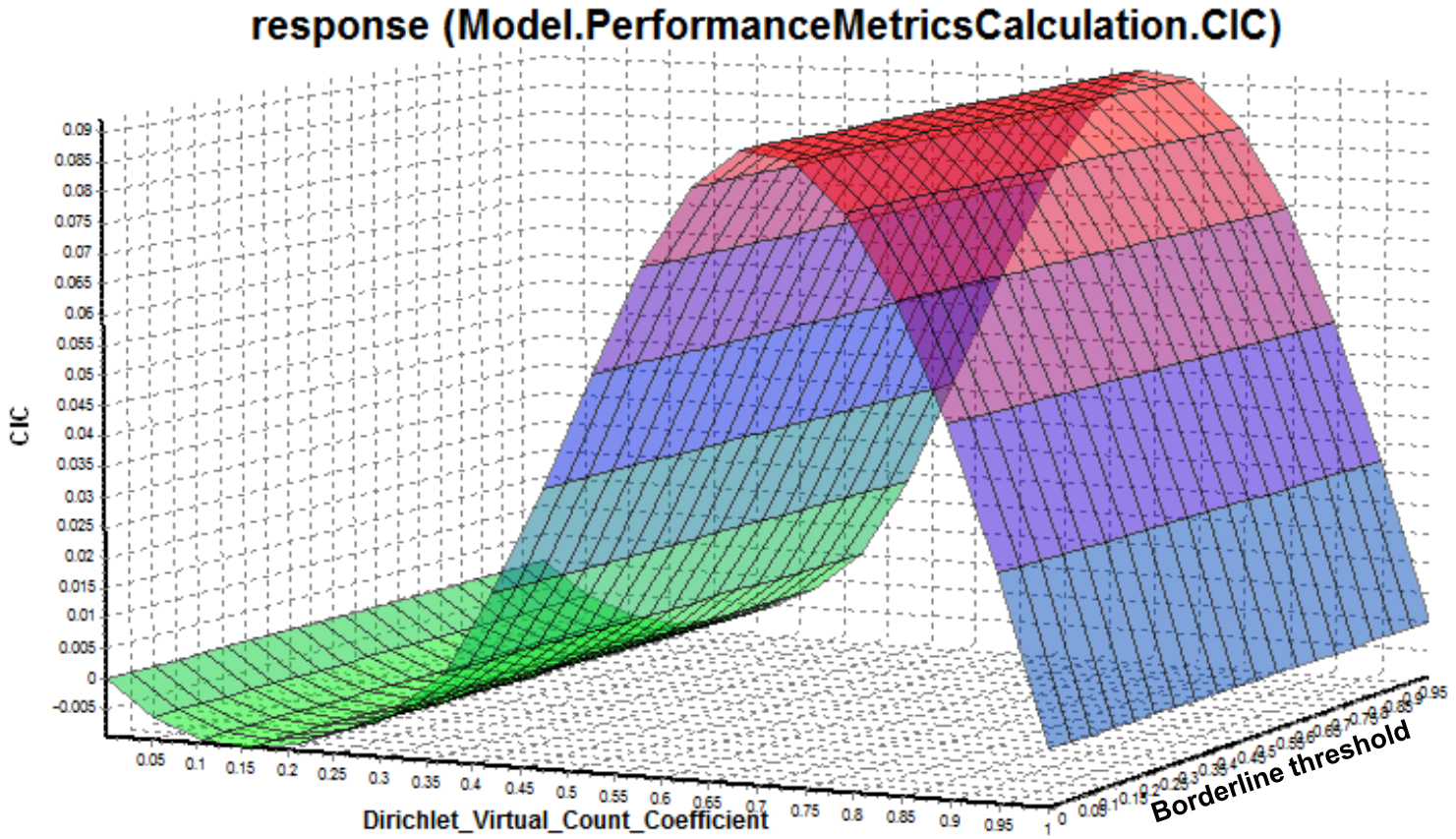
Precision



* Data is not real and just for illustration purposes.

Some Results

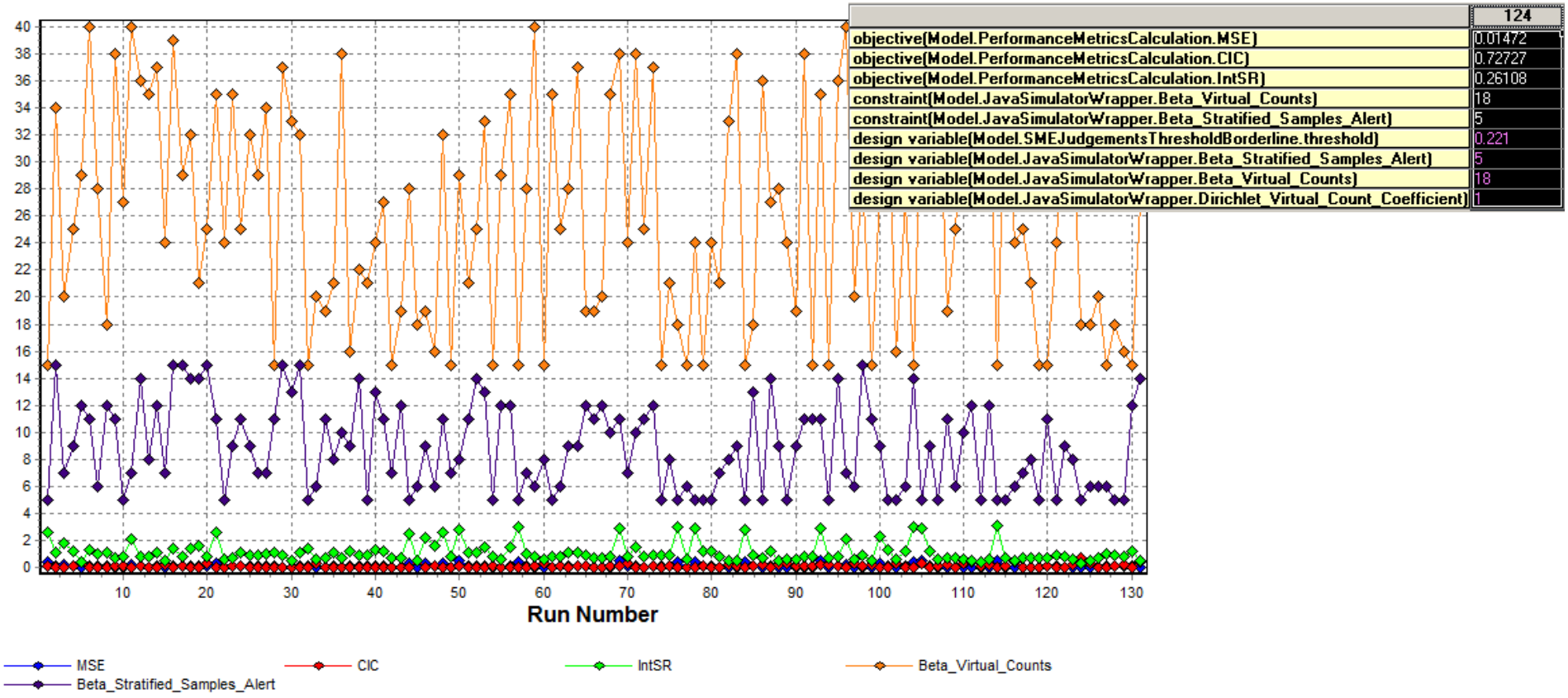
How Coverage/Certainty Interval Calibration (CIC) changes when virtual counts of Dirichlet distribution and threshold of "maybe" judgements in SME data are changed



* Data is not real and just for illustration purposes.

Some Results

Used ModelCenter's simulation/optimization tool to find configuration with best performance metrics.



* Data is not real and just for illustration purposes.

Concluding Remarks

•Purposes of multi-modeling in inference enterprise modeling:

– (1) evaluate how well the enterprise's automated threat detection system performs at detecting threats,

– (2) understand the reasons for its performance, and

– (3) identify ways to improve performance

•Model integration:

- Used Phoenix Integration ModelCenter® as an experimental test-bed where existing and proposed inference enterprises can be tested and vetted.
- Used empirical studies to develop an understanding of the performance of an inference enterprises. Develop a quantitative measure of fitness of an inference enterprises for a given organization's needs.
- Provided a capability for risk and cost-benefit analysis for the alternative solutions