Creating a Digital Twin of an Insider Threat Detection Enterprise using Model Based Systems Engineering

JAMES LEE, GEORGE MASON UNIVERSITY
AHMAD ALGHAMDI, GEORGE MASON UNIVERSITY
DR. ABBAS ZAIDI , GEORGE MASON UNIVERSITY

GEORGE MASON UNIVERSITY

Agenda

Motivation

Benefits of Using a Digital Twin (DT)

Insider Threat Detection Evaluation Challenges

Inference Enterprise Modeling as a DT Technique

Solution Method – Model Based Digital Twin Generation

Discussion

Motivation

Constant threats -> need for continuous monitoring

Operational testing is a challenge

- Lack of ground truth
- Data Privacy
- Blackbox models (COTS)
- Sensitive to disruptions

Solution: Use digital twin techniques to create real-time virtual counterpart of the physical system

Digital Twin Benefits

Provides virtual and dynamic digital representation of the system

Enables real-time monitoring of the system

Reduces cost associated with testing and verification

Integration with MBSE reduces ambiguity through unifying system structures and behaviors

Insider Threat

Malicious threat that comes from people within the organization

Examples:

- 2019 Capital One: $150M
  - Data breach via former employee of vendor
- 2018 Google and Uber: $245M
  - IP theft via former employee

2016 DoD 5220.22-M:

- Insider Threat Program Mandate

Image adopted from: https://securitytoday.com/articles/2019/03/19/-/media/SEC/Security-Products/Images/2019/03/insiderthreat.jpg

Limitations with Insider Threat Detection

RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV

Limited understanding of detection accuracy / performance

Confusion Matrix

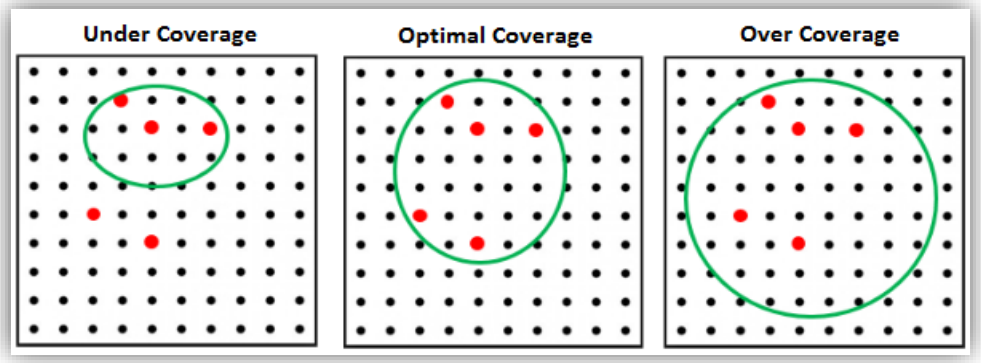
	Positive	negative
Predicted positive	***TP***	***FP***
Predicted negative	***FN***	***TN***

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN}$$

Overwhelming number of false alarms / limited ability to reduce the population to manageable subpopulations for more detailed analysis



Inference Enterprise (IE): enterprises that make inferences based on incomplete information (lack of ground truth)

Methodology Development

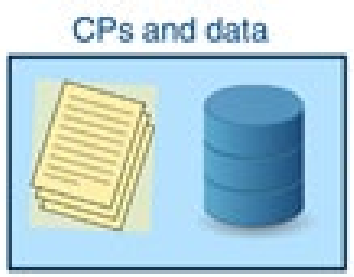
RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV

19 Challenge Problems: description of Inference Enterprise (IE) and data

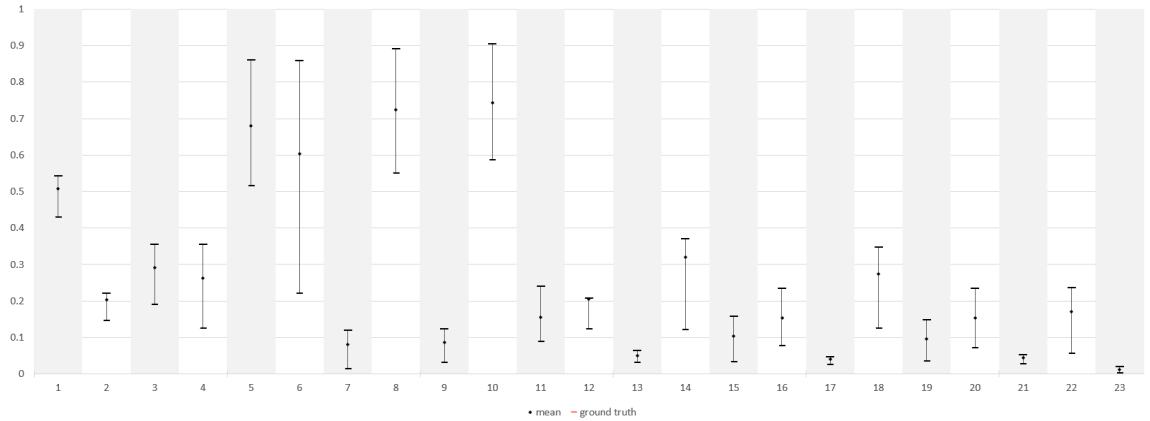
- Range of threat behavior: accidental compromise to deliberate sabotage
- Missing and/or incomplete data
- Down-select algorithms used: classification / clustering algorithms
- Forecast future performance, evaluate hypothetical upgrades to system

Inference Enterprise Modeling (IEM) methodology

- Build models and evaluate performance

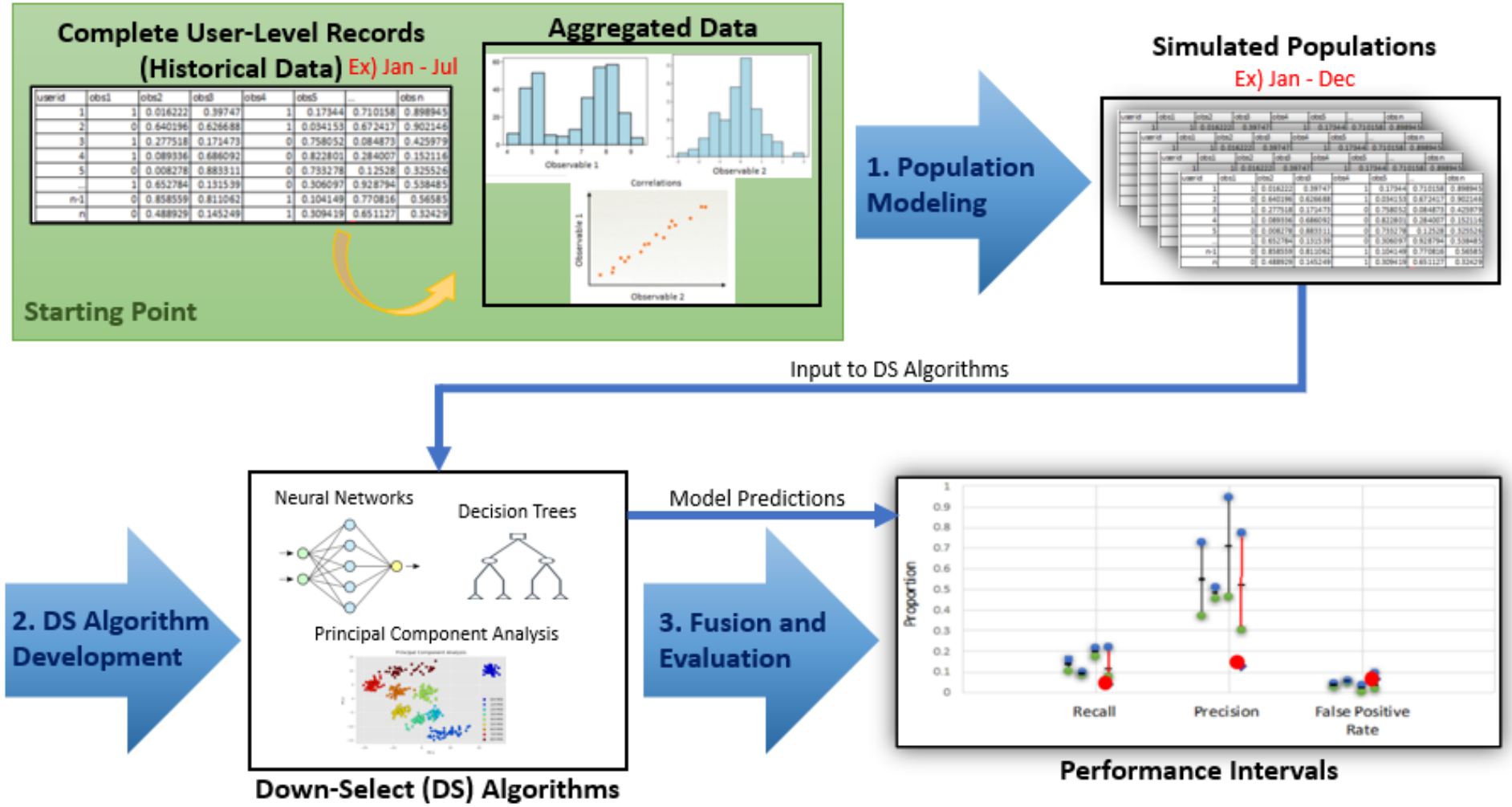


IEM Methodology



Inference Enterprise Modeling (IEM)

RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV



Relevant Methods - Ontology

RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV

Formal description of knowledge as a set of concepts within a domain and the relationships between those concepts

Benefits:

- Provides coherent and easy navigation between one concept to another - classes and instances
- By having essential relationships between concepts, enables automated reasoning
- Easy to connect with other ontologies
- Can represent any data format, enabling smoother data integration

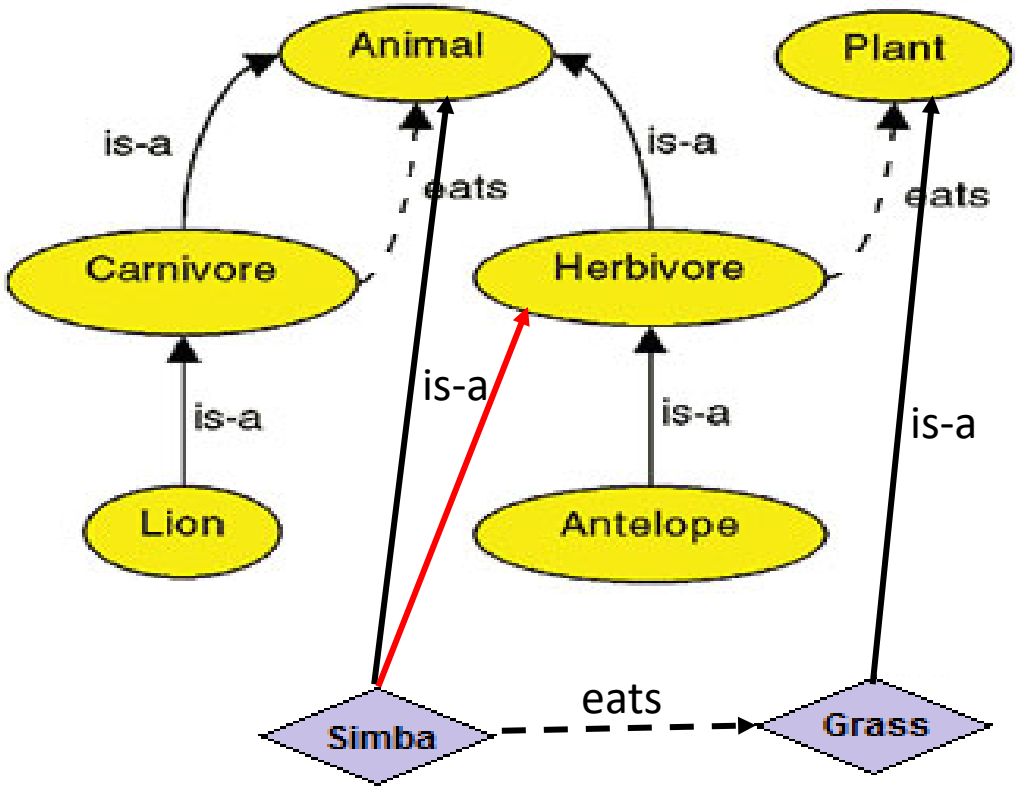


Image adopted from: http://www.differencebetween.info/sites/default/files/images/6/ontology.jpg

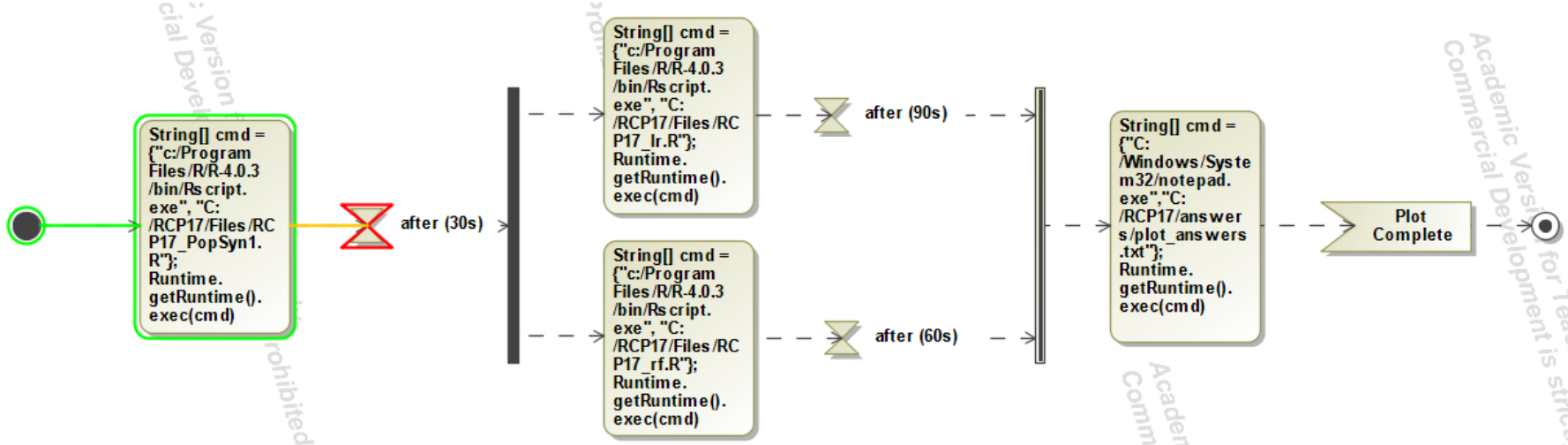
Relevant Methods – Process Modeling

RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV

Graphical representation of business processes or workflows

Business process: collection of tasks an organization performs to create products, reach goals, provide value

Process Modeling Languages: Business Process Modeling Notation (BPMN), Event Driven Process Chain (EPC), UML Activity Diagrams



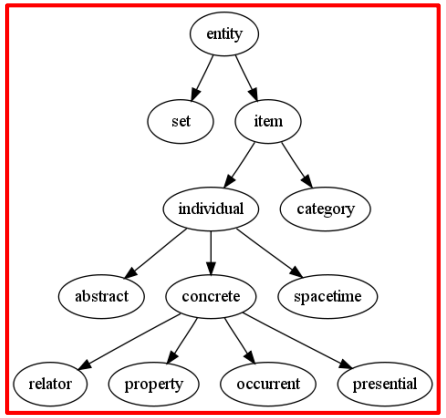
RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV

Solution Components

CP Characteristics: Problem Reqs

Queries

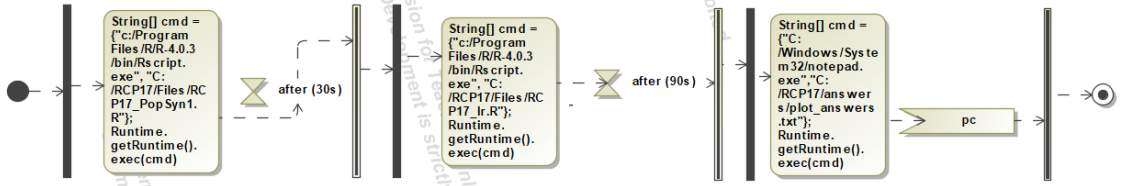
IEM Process Ontology (Protégé)



Activity Diagram elements

AD Model Generator (java)

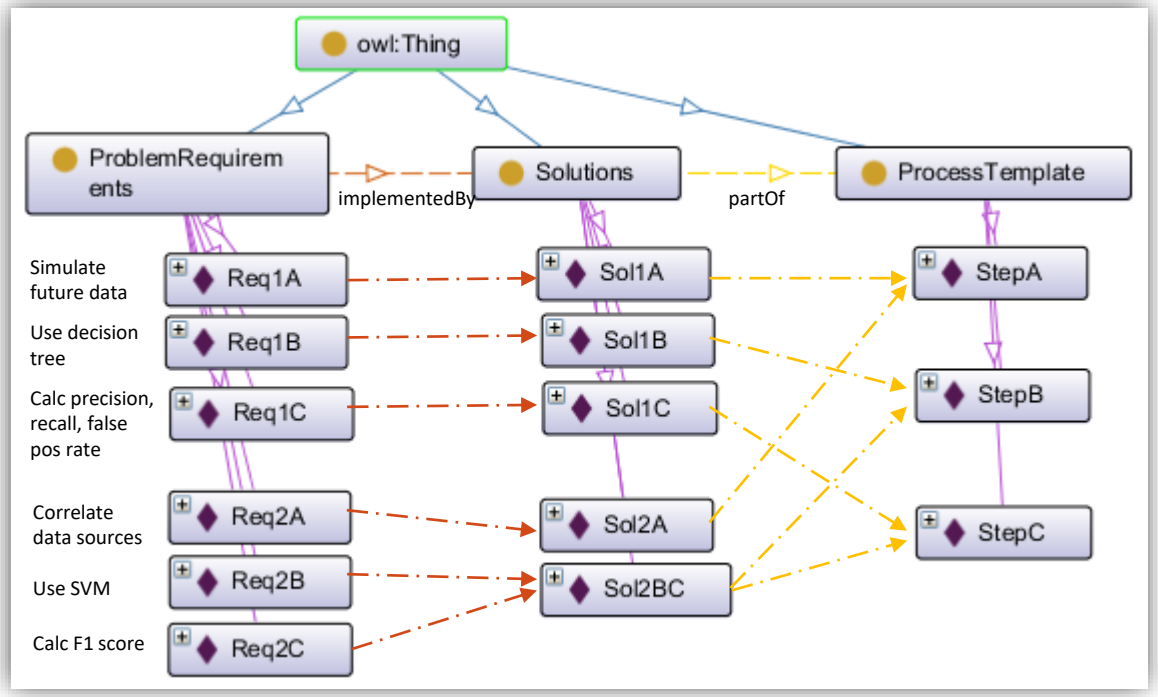
Activity Diagram of suggested workflow



Executable in Cameo

IEM Process Ontology Overview

RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV



ProblemRequirements: characteristics that define problem

- Simulate future data based on data provided
- Use certain ML algorithm to predict
- Calculate precision, recall, and false pos. rate

Solutions: software assets and/or manual activities that fulfill requirements

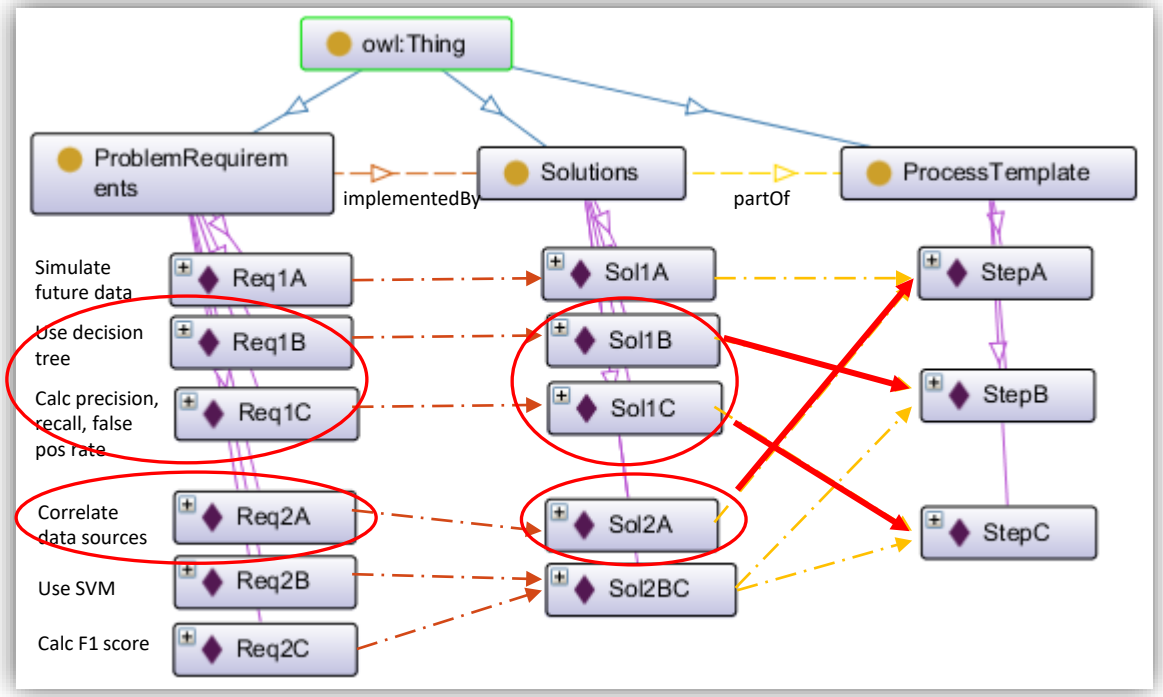
- Population assumptions based on SME input
- Population simulation and/or ML algorithms

ProcessTemplate

- Step A: Simulate Populations
- Step B: Down-select
- Step C: Calculate estimates

Small Example

RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV



Input New Problem:

- Correlate data sources
- Use decision tree to predict
- Calculate precision, recall, and false pos. rate

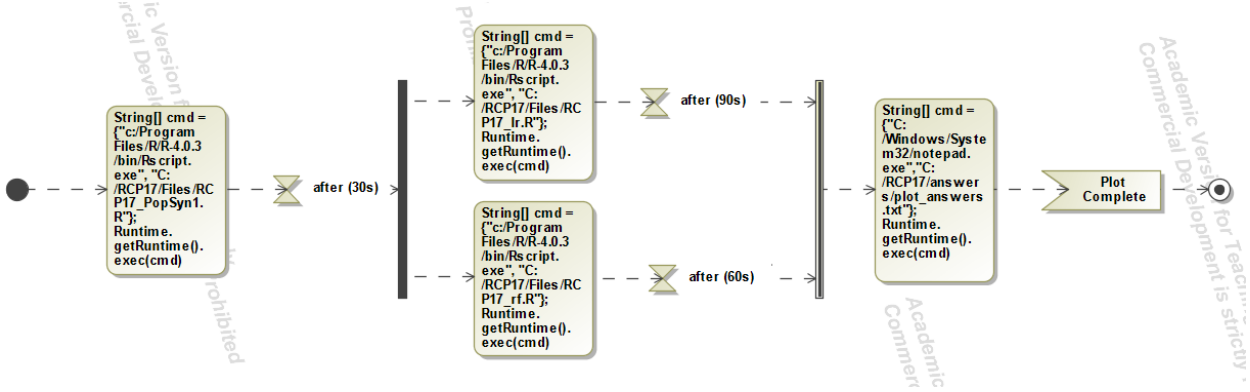
Output ProcessTemplate:

StepA	StepB	StepC
Sol2A	Sol1B	Sol1C

IEM Process Ontology Development

RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV

1. Use formal process modeling to document the solution workflows

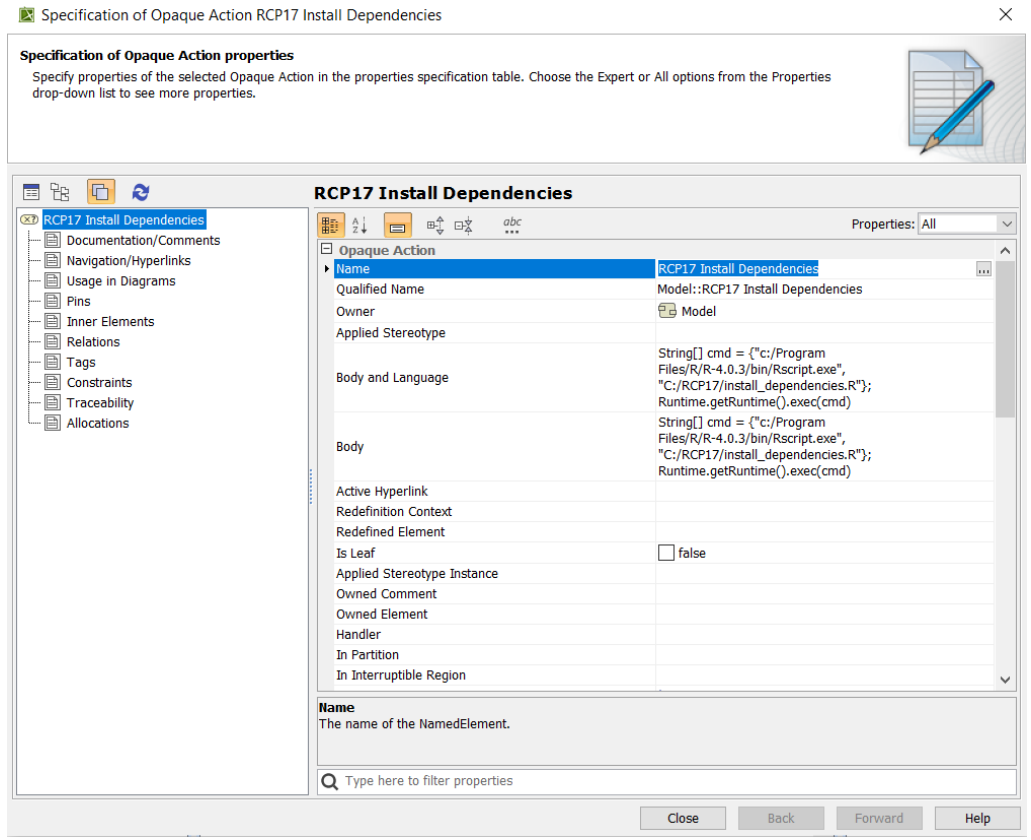


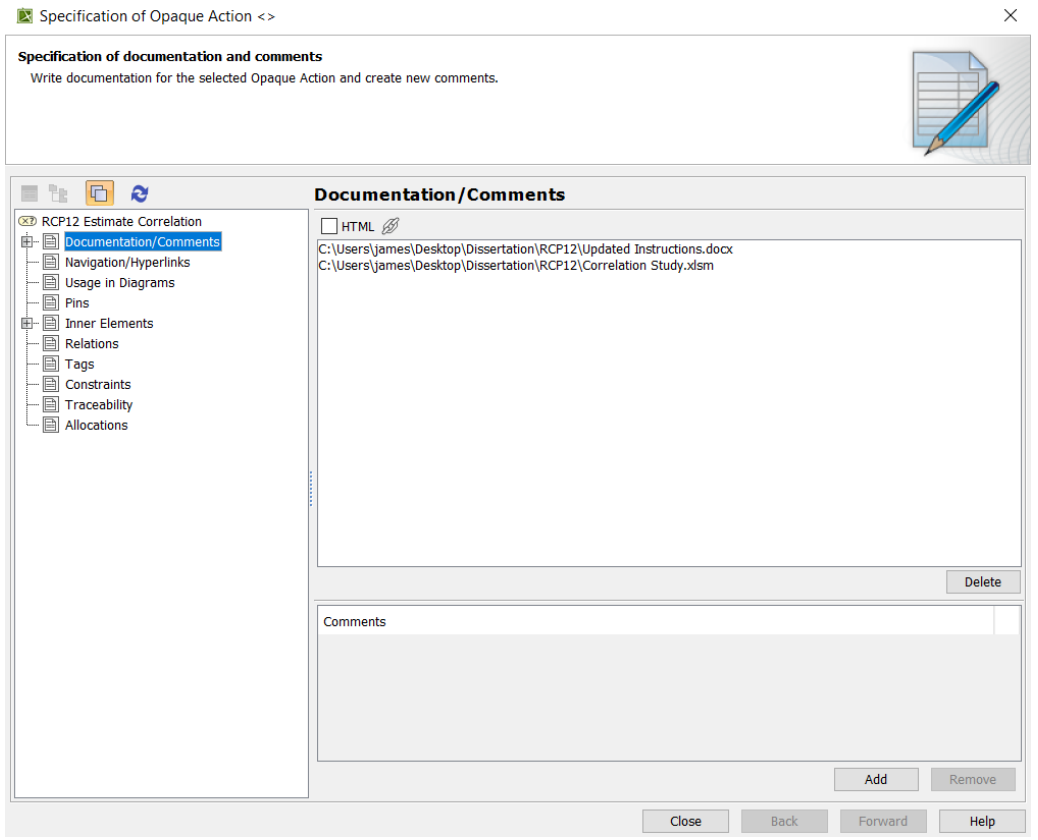
Notation name	Concrete syntax	note
Initial Node	●	Starting point of an activity.
Activity Final Node	◉	Marked the end or completion of an activity.
Fork/Join Node	▮	Fork node has a single incoming edge and many outgoing edges, and it is the reverse for the join node. These elements help modelers in expressing the occurrence of concurrent sequences in an activity.
Decision/Merge Node	◇	Used to show the starting/ending of an alternative sequence in an activity. Usually merge nodes are used in conjunction with decision nodes to model loops within an activity.
Wait Time Action	⧖	Implies wait until certain action occurs.
Opaque Action		Introduced for implementation-specific actions. Allow modeler to implement action by running user defined script in languages such as Java, C++ and others.
Accept Event Action		Action that waits for the occurrence of an Event that meets specified conditions.
Action		An Action is a named element that is the fundamental unit of an executable functionality. Regular actions are used to represent manual tasks in this work.

IEM Process Ontology Development

RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV

1. Use formal process modeling to document the solution workflows

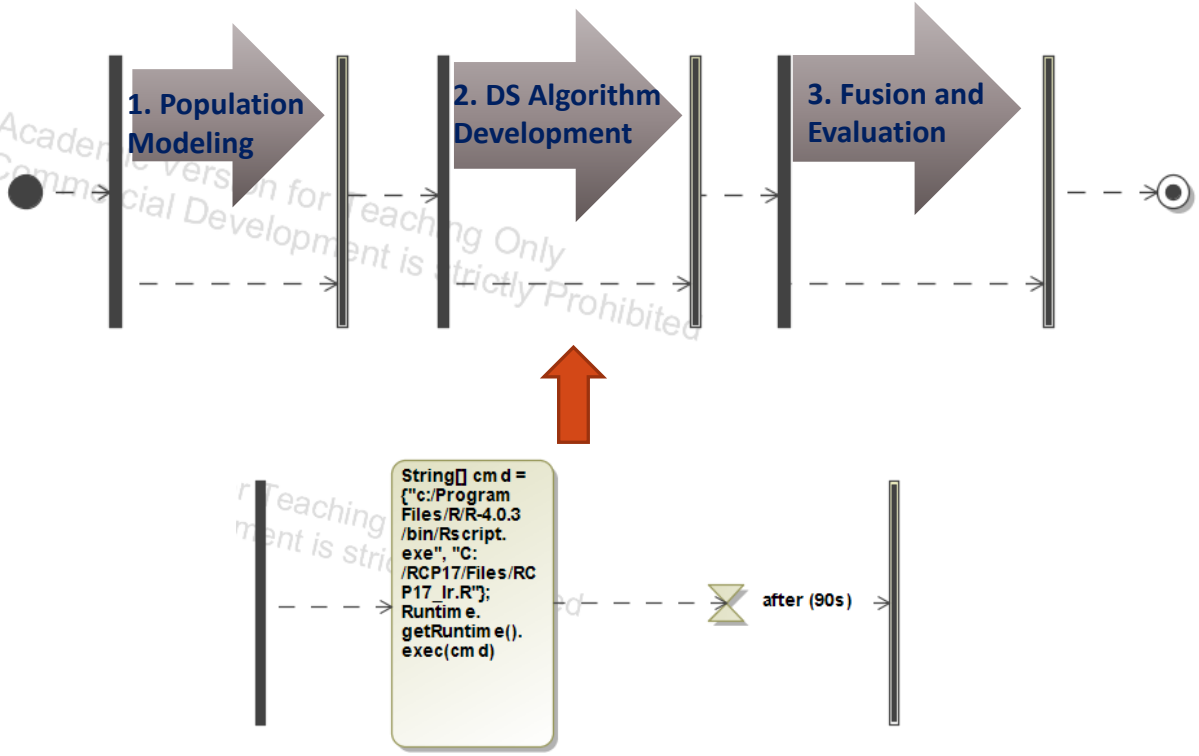


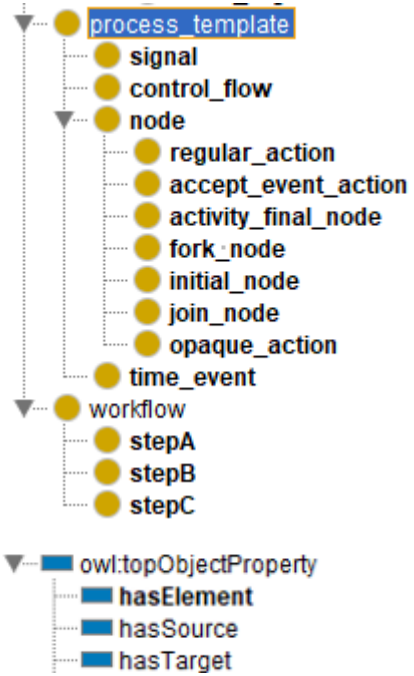


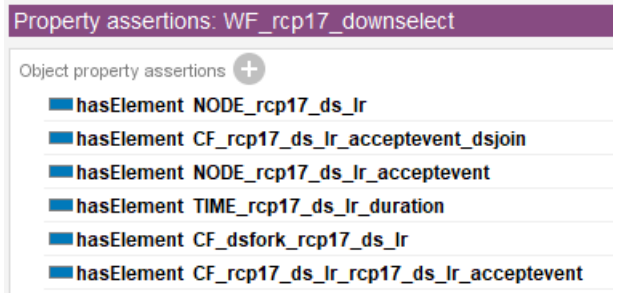
IEM Process Ontology Development

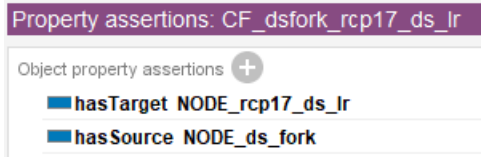
RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV

2. Generalizing solution methods by creating a process template and categorizing each part of the solution as a section of the template





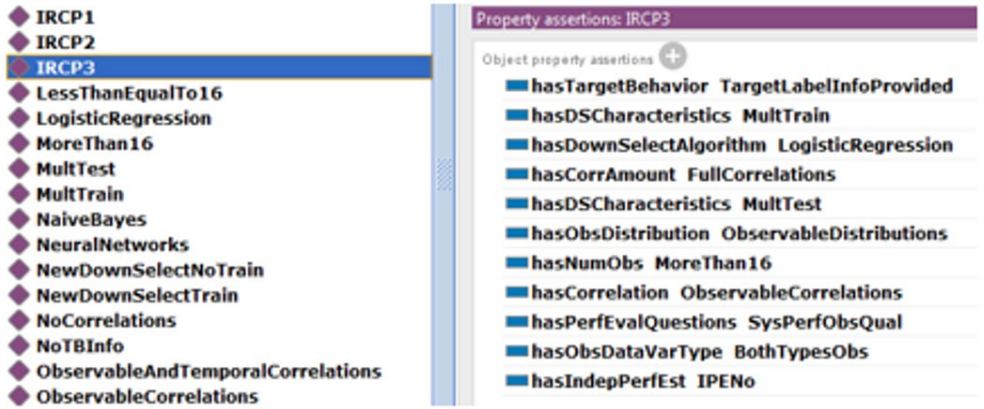




IEM Process Ontology Development

3. Defining traceability links from problem requirements to each solution method – Problem Side

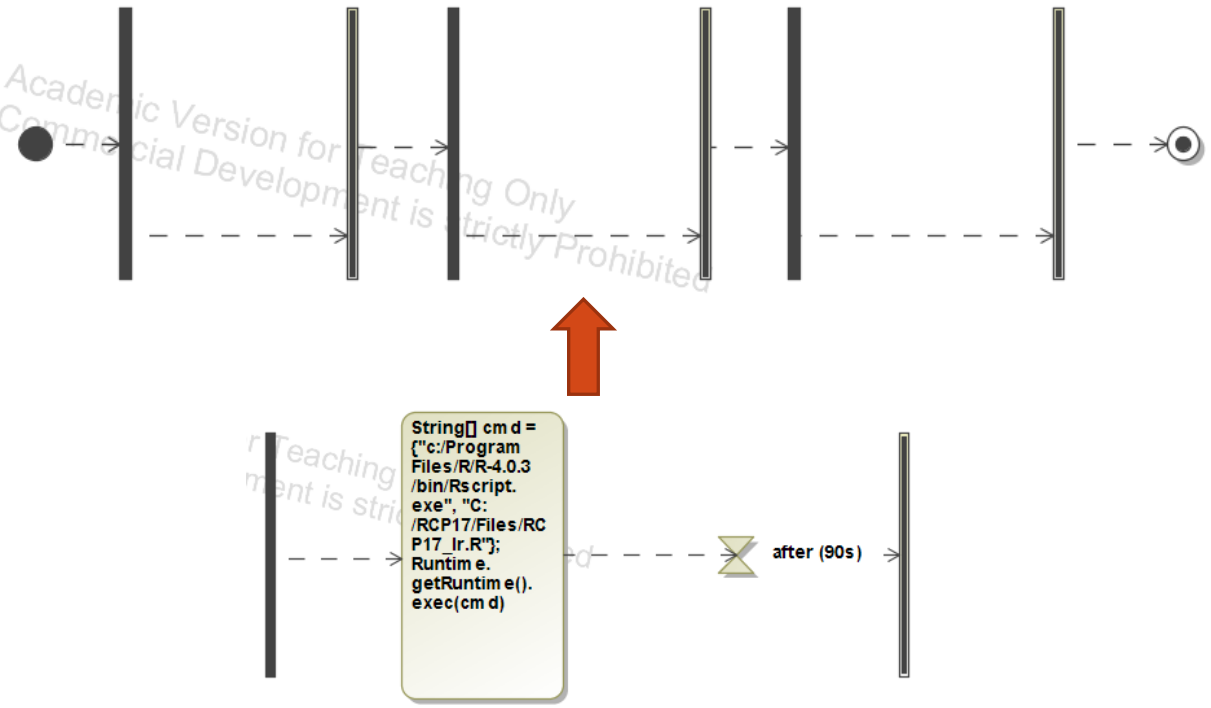
	Available Data								Data Characteristic								Down-Select Algorithm Types												Additional DS Requirements			Performance Evaluation		
	Correlation			IPE		ObsDist	Target Behavior		Correlation Amt		Num Obs		Num TB	Observable Data Types			Algorithms without Training Set						Algorithms with Training Set												
	None	Obs	ObsTime	Yes	No	Yes	Yes	No	Full	Partial	<= 16	> 16	> 2	Both	ContOnly	DiscOnly	Alerts>=2	Cate>=2	Days>=8	PCADBS	NB	RF	DT	LR	GBM	SVM	HMM	NN	HPTune	MultTrain	MultTest	CMR	SP	SPOQ
Q1 (RCP 1, 2, 3)		x			x	x	x		x		x					x	x																	x
RCP 4, 6		x		x		x		x		x	x					x	x																	x
RCP 5		x		x		x		x	x		x				x					x														x
RCP 7		x			x	x		x	x		x				x					x														x
RCP 8		x			x	x		x	x			x				x			x															x
RCP 9		x			x	x		x	x		x					x		x																	x
RCP 10		x			x	x	x		x			x		x									x												x
RCP 11		x			x	x	x		x		x			x									x												x
RCP 12	x				x	x	x				x		x			x							x										x		
Q5 (RCP 13, 14)			x		x	x	x		x			x		x													x								x
RCP 15			x		x	x	x			x		x		x									x			x	x		x	x	x		x		
RCP 16		x			x	x	x		x			x		x							x	x									x		x		
Q7 (RCP 17, 18, 19)			x		x	x	x		x			x		x																	x		x		
IRCP 1		x			x	x		x	x			x			x									x							x	x			x
IRCP 2		x			x	x		x	x			x		x										x							x	x			x
IRCP 3		x			x	x	x		x			x		x										x							x	x			x

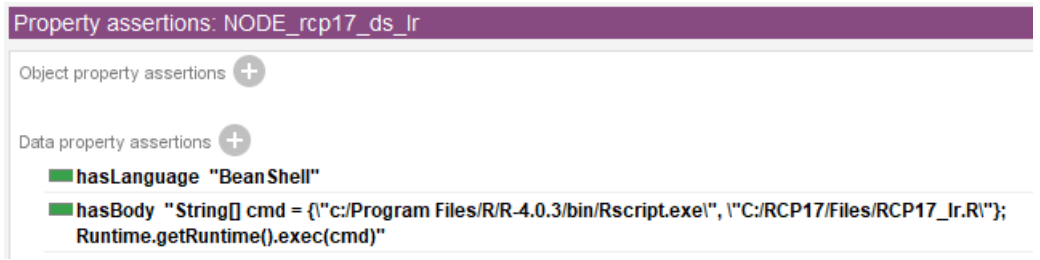


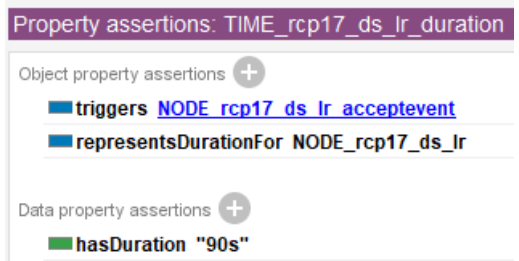
IEM Process Ontology Development

RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV

3. Defining traceability links from problem requirements to each solution method



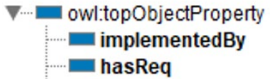


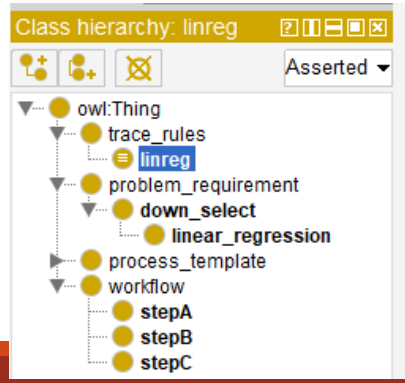


IEM Process Ontology Development

3. Defining traceability links from problem requirements to each solution method – Solution Side

MIEM Problem Specification	Discrete Event Activity Counts Complete	Stochastic Optimization Complete	Discrete Event Activity Counts with IPE	Stochastic Optimization with IPE	Inverse CDF Copula Discrete with SME Correlation	Factorized Stochastic Optimization with DEAC	Inverse CDF Copula	Inverse CDF Copula and Logistic Regression with IPE	Inverse CDF Copula Continuous and Discrete	Rank Correlation Copula	2D Rank Correlation Copula	2D Rank Correlation Copula with SME Partial Correlation	Hierarchical Bayes Net	Tree Augmented Naïve Bayes
no correlations					x									
observable correlations	x	x	x	x			x	x	x	x				
observable and temporal correlations											x	x		
IPE			x	x				x						
no IPE	x	x			x	x	x		x	x	x	x	x	x
observable distributions	x	x	x	x	x	x	x	x	x	x	x	x	x	x
target labels														x
no target label info														
full correlations	x	x					x	x	x	x	x			
partial correlations			x	x								x		
num obs restriction <= 16	x	x	x	x	x									
any observables									x	x	x	x	x	x
continuous observables only							x	x						
discrete observables only	x	x	x	x	x	x								



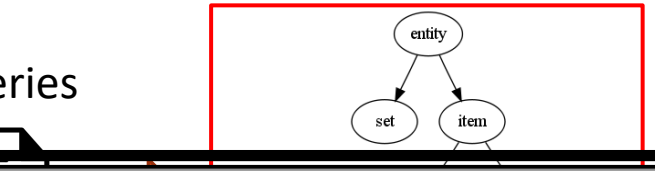


Small Example

Problem Requirements

Queries

IEM Process Ontology



Activity Diagram elements

Activity Diagram Model Generator

DL query:

Query (class expression)

inverse (hasElement) some (inverse (implementedBy) value REQ_linear_regression)

Execute | Add to ontology

Query results

Instances (6 of 6)

- CF_dsfork_rcp17_ds_lr
- CF_rcp17_ds_lr_acceptevent_dsjoin
- CF_rcp17_ds_lr_rcp17_ds_lr_acceptevent
- NODE_rcp17_ds_lr
- NODE_rcp17_ds_lr_acceptevent
- TIME_rcp17_ds_lr_duration

```
<?xml version="1.0" encoding="UTF-8"?>
<xmi:XMI xmlns:uml="http://www.omg.org/spec/UML/20131001"
         xmlns:sysml="http://www.omg.org/spec/SysML/20150709/SysML"
         xmlns:xmi="http://www.omg.org/spec/XMI/20131001"
         xmlns:StandardProfile="http://www.omg.org/spec/UML/20131001/StandardProfile">
   <uml:Model xmi:type="uml:Model" xmi:id="eee_1045467100313_135436_1" name="Model">
      <packagedElement xmi:type="uml:Activity" xmi:id="_19_0_4_61f0211_1618337477763_847249_42601" name="Model">
         <edge xmi:type="uml:ControlFlow" xmi:id="_19_0_4_61f0211_1618359650208_191843_42804"
               source="_19_0_4_61f0211_1618359547156_330700_42747"
               target="_19_0_4_61f0211_1618359633039_885360_42785">
         </edge>
         <edge xmi:type="uml:ControlFlow" xmi:id="_19_0_4_61f0211_1621442112183_794503_42117"
               source="_19_0_4_61f0211_1618337485490_833892_42649"
               target="_19_0_4_61f0211_1618359547156_330700_42747">
         </edge>
         <edge xmi:type="uml:ControlFlow" xmi:id="_19_0_4_61f0211_1621442114791_582857_42122"
               source="_19_0_4_61f0211_1618359633039_885360_42785"
               target="_19_0_4_61f0211_1618337487969_218488_42654">
         </edge>
         <node xmi:type="uml:InitialNode" xmi:id="_19_0_4_61f0211_1618337485490_833892_42649"/>
         <node xmi:type="uml:ActivityFinalNode" xmi:id="_19_0_4_61f0211_1618337487969_218488_42654"/>
         <node xmi:type="uml:OpaqueAction" xmi:id="_19_0_4_61f0211_1618359547156_330700_42747">
            <body>String[] cmd = {"c:/Program Files/R/R-4.0.3/bin/Rscript.exe", "C:/RCP17/Files/RCP17_lr.R"};
Runtime.getRuntime().exec(cmd)</body>
            <language>BeanShell</language>
         </node>
         <node xmi:type="uml:AcceptEventAction" xmi:id="_19_0_4_61f0211_1618359633039_885360_42785">
            <trigger xmi:type="uml:Trigger" xmi:id="_19_0_4_61f0211_1618359633040_700741_42786"
                     event="_19_0_4_61f0211_1618359633041_636077_42787"/>
         </node>
      </packagedElement>
      <packagedElement xmi:type="uml:TimeEvent" xmi:id="_19_0_4_61f0211_1618359633041_636077_42787" isRelative="true">
         <when xmi:type="uml:TimeExpression" xmi:id="_19_0_4_61f0211_1618359633041_831861_42788">
            <expr xmi:type="uml:LiteralString"
                  xmi:id="_19_0_4_61f0211_1618359646328_321972_42800"
                  value="90s"/>
         </when>
      </packagedElement>
   </uml:Model>
</xmi:XMI>
```

Simulation Tool

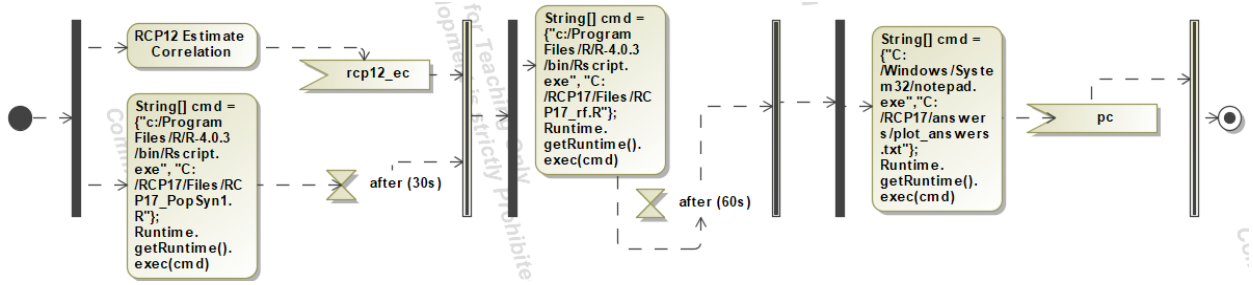
New Challenge Problems

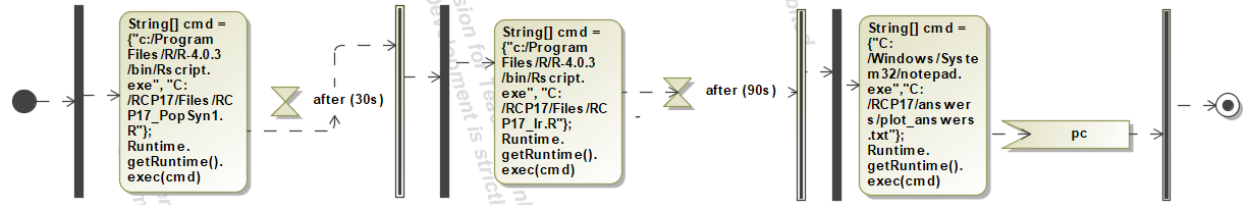
Both problems use the GMU phishing experiment data set

- A: There are two models that use different sets of data to predict susceptibility to phishing. Since the two datasets are disjoint, correlation data between the sets do not exist. Forecast the performance of a model that uses all the data from both models.
  - Has continuous and discrete variables
  - Has binned data
  - Does not have full correlations
  - Has random forest
- B: There is a model that predicts staff members' susceptibility to phishing using staff data. Faculty data is not yet available. Forecast the performance of a model trained on staff data and tested on simulated faculty data.
  - Has full correlations
  - Has continuous variables
  - Has binned data
  - Has logistic regression

Suggested Workflows

RESEARCH AND APPLICATION WORKSHOP
AI for SE & SE for AI
Date: OCT 20-21, 2021
Location: VIRTUAL via ZOOM-GOV





1. Understand the overall flow of the suggested solution
2. Identify the software and documentation location by inspecting the activity elements of the activity diagram
3. Read the documentation, run the software modules as-is, inspect the input and output files, and understand its behavior.
4. Create a copy of the software modules that point to the new dataset of the challenge problems
5. Compare the input data file formats of the new problem to the input and output of the suggested module and identify places were appropriate changes are necessary
6. Make the appropriate code changes to the modules and execute

Research Contributions

Introduced IEM as digital twin technique for insider threat detection enterprises

Developed knowledge base of IEM expertise and model-based solution that can be used to rapidly prototype digital twins of new IEM scenarios

Future Works

Accurate problem requirement generation

Flexible template generation

Full simulation capabilities of systems modeling software

IEM – Ransomware protection

Thank you

Submitted full manuscript at SysCon 2022

Email: jlee194@gmu.edu