Validation

Kathleen M. Carley

kathleen.carley@cs.cmu.edu

Carnegie Mellon

Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/

Key Questions

- Should Every Model be Validated?
- How much validation is needed?
- What is the difference between validation and curve fitting?
- What is validation?

<Your Name>

VV&A

- Verification
  - Is the programming correct
  - *Did I build the thing right?*
  - *The process of determining that a model implementation and its associated data accurately represent the developer's conceptual description and specifications*
    - ***Note – in the management literature verification is used to mean validation!!!***
- Validation
  - Does the model match reality at a level sufficient for the model's purpose
  - *Did I build the right thing?*
  - *The process of determining the degree to which a model and its associated data provide an accurate representation of the real world from the perspective of the intended uses of the model*
- Accreditation
  - What systems can this model be used on and in what context
  - *Should it be used?*
  - *The official certification that a model, simulation, or federation of models and simulations and its associated data is acceptable for use for a specific purpose*

Verification

- Two Meanings depending on literature
- Engineering
  - Accuracy of code
- Economics and Management and Operations Management
  - Validation and verification used interchangeably
  - Verification used in the "philosophical" sense of verified against the real world

Documentation is Critical

- Note – for the DOD requires specific types of documentation - http://dvdt.nmso.navy.mil/
- VV&A is performed
  - RISK: when the potential risk of making an incorrect decision based on a simulation
  - COST: outweighs the time and cost of performing VV&A to ensure that simulation can produce results that are sufficiently accurate and reliable
- CAVEAT – ALL of VV&A is relative to the “purpose” of the model

Specification vs Validation

- Specification:
  - Static
  - Architecture of software and structure of code
  - Verifications done through formal methods, which are insufficient for multi-agent social-network and other simulations (e.g., lacking emergent properties)
  - Can be formalized using UML or OWL
- Validation:
  - Dynamic
  - Simulation outputs, inputs, and happenings
  - Empirical data and knowledge
  - Inference based

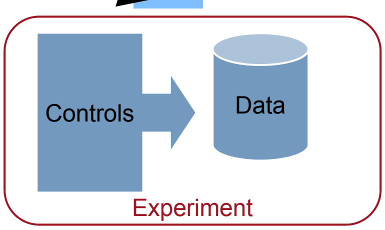
Credibility – *Should it be trusted?*

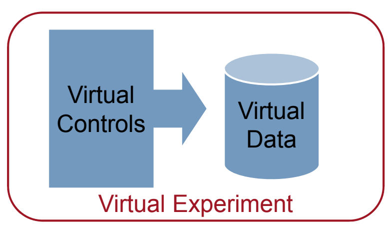
Orthogonal Concept --- Credibility

- depends on ***correctness***
  - the level of confidence that its data and algorithms are sound and robust and properly implemented,
  - the accuracy of the simulation results will not substantially and unexpectedly deviate from the expected degree of accuracy
- Depends on ***usability***
  - the training and experience of those who operate it,
  - the quality and appropriateness of the data used in its application
  - the configuration control procedures applied to it
  - The ease of entering and extracting data

The Theory

Idea





Assumptions:
Experiment is repeatable
Results can be replicated
Data is complete
Controls are stable and well defined

Assumptions:
experiment assumptions +
Code is accurate
All factors have been modeled
Standard statistical tests can be used

Carnegie Mellon
isr Institute for Software Research

BUT: With Simulations of Complex Socio-Technical Systems

- *Validation Assumptions do not hold !!*
- Experiment is repeatable
  - Lab experiments with people are only partially repeatable
  - Field studies can not be repeated
- Results can be replicated
  - Results may not be replicated precisely
- Data is complete
  - For field data not all information is available at same granularity
- Controls are stable and well defined
  - Some “controls” may not be obvious
- Code is accurate
  - YES!
- All factors have been modeled
  - No – only the critical subset
- Standard statistical tests can be used
  - Maybe not - non parametric as distribution not known

CASOS

Carnegie Mellon
isr Institute for Software Research

Consequently …

- *New Approach*
- Experiment is repeatable
  - Trade repeatability for large quantities of data
- Results can be replicated
  - Treat results as a probability distribution
- Data is complete
  - Data fusion and validation by parts
- Controls are stable and well defined
  - Search for critical or dominant factors
  - Discuss boundary conditions
- Code is accurate
  - YES!
- All factors have been modeled
  - No – only the critical subset
  - Decision support not decision making
- Standard statistical tests can be used
  - Many diverse approaches
  - Forecast not prediction

The Space of Possibilities!

CASOS

CASOS

Carnegie Mellon
isr Institute for Software Research

Definitions

- Validation – a set of techniques for determining whether or not a model is valid. Used for both internal validity, matching with other models, and matching with non-computational data.

Special forms of Validation

- Calibration – a set of techniques for tuning a model to fit detailed non-computational data.
- Training – procedures for supplying data and feedback to computational models that can learn.
- Docking – a set of techniques for determining the level of comparability or equivalence of two models.

Carnegie Mellon
isr Institute for Software Research

Traditional Validation Approaches

- Requirements engineering and formal methods
- Evolutionary verification and validation
- Docking, including against math and system dynamics models
- Statistical methods alone
- Expert systems (not usually done because simulations are usually numerical)
- Domain experts (human experts/subject matter experts)
- Response Surface Methodology (validation against empirical data)

Carnegie Mellon
ISRI Institute for Software Research

Methods and Levels for Validating a Computational Model

- Validation techniques vary in
  - Method
  - Level
  - Intensity
  - Purpose
- Similar approaches can be used for
  - Any form of validation
  - Calibration
  - Training
  - Docking

CASOS

Carnegie Mellon
ISRI Institute for Software Research

Validation Levels

- Internal validity aka Verification - error free code
- Parameter validity - parameters match
- Process validity - processes fits
- Face validity - right type of things
- Pattern validity - pattern matches observed
  - Stylized Facts
  - Statistical Patterns
- Value validity - values match
- Theoretical validity - theory fits

Validation in Parts

CASOS

CASOS

Carnegie Mellon
isr Institute for Software Research

Validation is Difficult

- Models are a subset of reality; model assumptions may not match the reality
- Cognitive bias of human modelers/validators
- Validation is knowledge intensive
- Complexity and stochasticity of social agents
- Agent history (non-Markovian), starting conditions, etc.
- Validation consumes a significant amount of time and resources
- Quality and quantity of empirical data vary
- Least developed area of Multi-Agent Social-Network computational modeling

CASOS
October 2016 15

Carnegie Mellon
isr Institute for Software Research

A Caveat

- Computational modeling is sufficiently complex that a single individual in a single research period (e.G. 6 months to a year) can not build, analyze, and validate a computational model.
- Most models take multi-person years to build and analyze.
- Data collection and analysis from a virtual experiment often takes as long as a human experiment and requires statistical training comparable to that required for human experiments.

CASOS
2016 16

Validation In Parts

- Input
  - Match with real world
  - Actual real data
- Internal Processes
  - Algorithms derived from real data
    - Statistical - ERGM Model or model created by machine learning
    - Mathematical – model describing experiment data
    - Logical description
- Output
  - Predictive forecasting
    - WARNING WARNING WARNING – Don't over fit

Face Validity

- Is the model a reasonable simplification of reality?
- Techniques to increase face validity:
  - Set parameters based on real data
  - Model a specific organizational or inter-organizational process
  - Show that others have made similar assumptions
  - Discuss model limits and how left out factors may or may not affect results
  - Don't over-claim model applicability

Carnegie Mellon
isr Institute for Software Research

Illustration: Model & Reality

Simulated Annealing	Organizational Strategic Adaptation
system	organization's CEO or central unit
state	organizational design
current state	current organizational design
temperature	risk aversion
accepting a cost increasing move	taking a risk
high temperature means accepting many cost increasing moves	liability of newness
move set	re-design strategies
heuristic optimization process	satisficing & BR process
minimize cost	maximize performance
cooling schedule	approach to becoming risk averse
proposed state	proposed new design
evaluation of proposed state	limited lookahead, anticipation of future
state evaluation	observed performance

CASOS

Carnegie Mellon
isr Institute for Software Research

(Social) Turing Test

- The model does the task it seeks to explain.
  1. substitutability.
  2. Turing test.
  3. Social Turing test.
- Construct a collection of social agents according to the hypotheses and put them in a social situation, as defined by the hypotheses. Then recognizably social behavior should result.
- Aspects not specified by the hypotheses, of which there will be many, can be determined at will.
- The behavior of the system can vary widely with such specification, but it should remain recognizably social.

CASOS

Carnegie Mellon

Stylized Facts

- Simple techniques for seeing of model results are reasonable.
- Techniques to demonstrate validity:
  - Are there stereotypical facts about the problem that this model generates; E.G., Models of organizational evolution should predict liability of newness.
  - Are there behaviors that any model of this ilk should generate; E.G., All diffusion models should generate an s-shaped adoption curve, all neural networks should take a long time to train.
- These are non-surprising findings but if model can't generate them then it is not valid.

Carnegie Mellon

Empirical Validation (remember in some areas this is called verification)

Carnegie Mellon

Validation to see if model is Credible vis Reality

- Involves testing a computational model's predictions given a set of non-computational data
- Have available the results of a virtual experiment
- Have available non-computational results
  - May be archival, survey, experimental
- Is sometimes done on uncalibrated models
- Demonstrates that model's predictions match non-computational data

Carnegie Mellon

Types of Validation

- Level of validation:
  - Pattern - same trends are observed
  - Value - same values are observed
- For multi-agent models:
  - Group or organizational level - matches overall behavior of collection of agents
  - Agent level - matches specific entities behavior
- For stochastic models:
  - Point - on average behavior is the same
  - Distribution - distribution of results is the same
  - Detail match - one entire run is the same

Carnegie Mellon
isr Institute for Software Research

Data for Validation

- Type: anything
  - May be archival, survey, experimental, subject matter expert (though least accepted statistically)
- Quantity: high
  - Sufficient for statistical analysis
- Level of detail: low
  - Do not need the same level of process data that is needed for calibration

CASOS

Carnegie Mellon
isr Institute for Software Research

Calibration

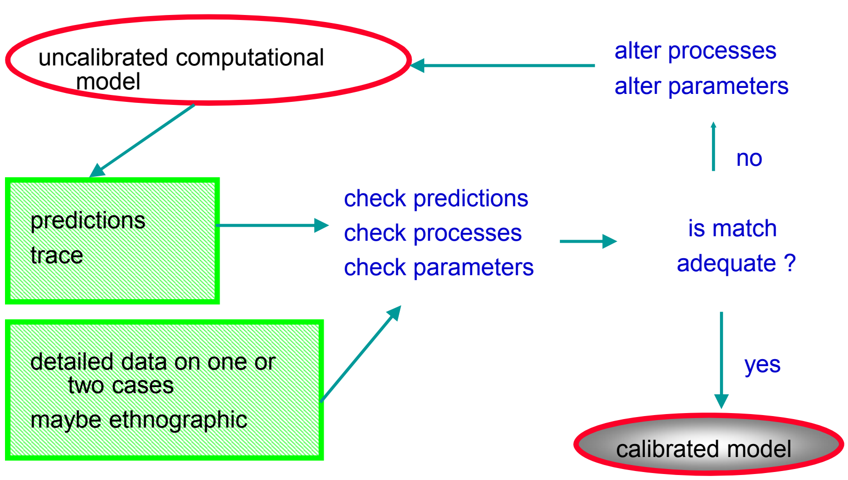
CASOS

CASOS

Calibration

- Calibration involves fitting a computational model to a set of data
  - May require programming (adding modules or new processes)
  - May require parameter setting
- Have available detailed data on one or more cases
- Calibration is often the only validation step carried out for emulations
- Calibration demonstrates that model can match non-computational data

Calibration Cont.



Carnegie Mellon
isr Institute for Software Research

Locating Cases

- Ideally:
  - Use a set of cases that span the key categories you are concerned with
- Next best:
  - Choose 2-4 cases that represent typical behavior and 1 or 2 that represent atypical behavior
- In practice:
  - Most intellective models are not calibrated
  - Lucky to have even one case with sufficient detail
  - Often detailed case is a matter of opportunity
- Sources:
  - Archival data, ethnographies, participant observation, subject matter expert (SME)

Carnegie Mellon
isr Institute for Software Research

What if using SME Data: Multi-expert Problem

- What if experts or cases disagree?
- Typical solution: average the two cases
- Alternative: put in both cases as options with a certain probability of occurring
  - Probability:
  - Equally weighted
  - Weight can reflect degree of agreement

NOTE: For rule based models detailed cases may be the opinions of experts

Carnegie Mellon
isr Institute for Software Research

Training

CASOS

Carnegie Mellon
isr Institute for Software Research

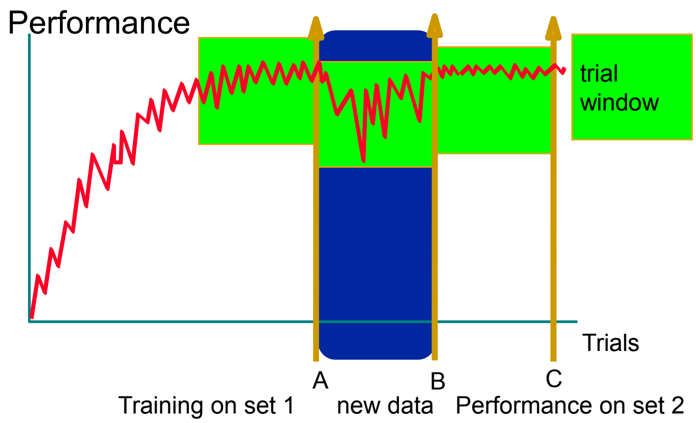
Model Training

- Use this when:
  - You have models that learn
  - You want to test the "goodness" of what they have learned
- Approach
  - Divide non-computational data into two sets
  - Train the model on first set
  - Generate model predictions to second set
  - Test model against second set

CASOS

CASOS

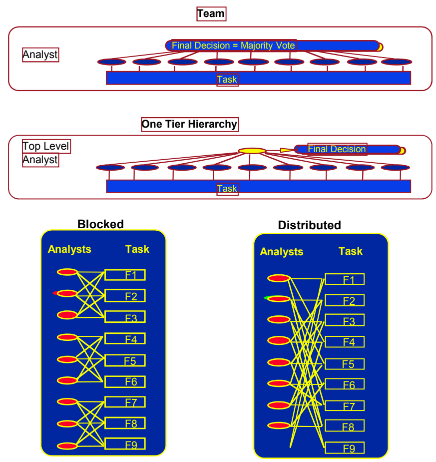
Impact of Training



Matched Analysis

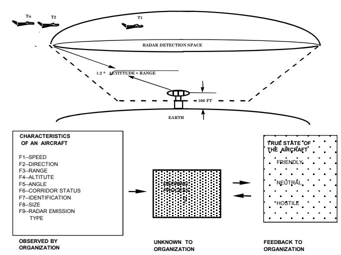
Carnegie Mellon
ISRI Institute for Software Research

Corp (A predecessor to ORGAHEAD)



Used a subset that matched human experiment

Task



CASOS

Carnegie Mellon
ISRI Institute for Software Research

Matched Analysis

Simulation

- Vary organizational design
- Vary task environment
- Measure performance as accuracy
- Monte Carlo 19683 cases
- Estimate of performance on average

Corporate Data

- Vary organizational design
- Vary task environment
- Measure performance as actual/potential severity
- General performance
- 69 cases, technological disasters

Matched Set

- Predict performance
- What if analysis: if organization did/did not shift what would be impact

CASOS

CASOS

CarnegieMellon

Archival Match

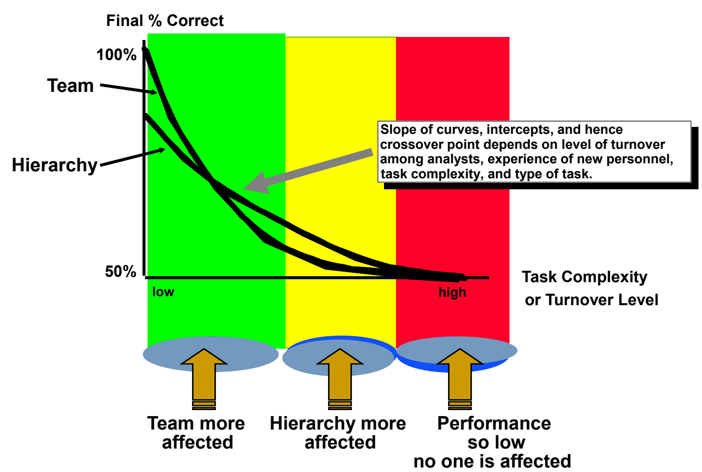
Training	Prediction Performance in General	Prediction Performance During Crisis
Experiential	2.10(21,0.18)	2.38(21,0.11)
Operational	1.83(48,0.10)	1.42(48,0.07)

Training	Observation Performance in General	Observation Performance During Crisis
Experiential	1.86(21,0.17)	2.38(21,0.13)
Operational	1.83(48,0.09)	1.46(48,0.08)

Note: Number of cases and standard errors are in parentheses

CarnegieMellon

Hierarchies More Robust



Carnegie Mellon
isr Institute for Software Research

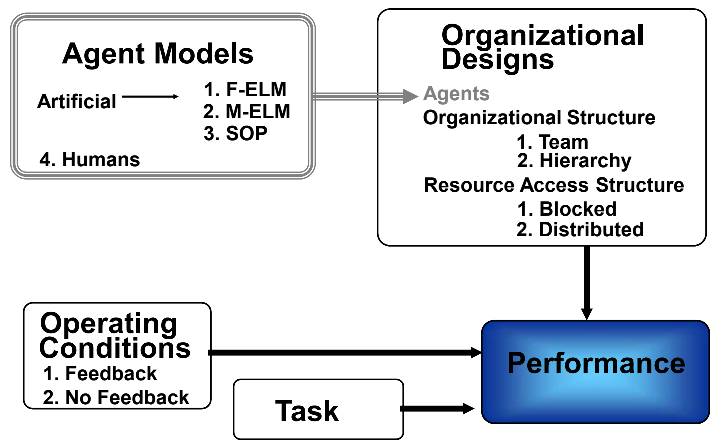
Reality: Teams Better

	TEAM		HIERARCHY		SEGREG		NON-SEG	
	exp	sop	exp	sop	exp	sop	exp	sop
Simulated	3.00	1.50	2.35	1.41	2.30	1.40	2.45	1.60
Human	3.00	1.50	2.35	1.46	2.10	1.42	2.64	1.80
	(1)	(4)	(20)	(44)	(10)	(43)	(11)	(5)

Carnegie Mellon
isr Institute for Software Research

Micro –versus- Macro Validation Conundrum

Illustration: CORP & Human Experiment



Human Experiment

Two stages -

Stage 1) Analysts / Operators Decisions based on raw data

Stage 2) Commander/mid-level Decisions based on analyst's decisions

---

Each analyst sees 120 problems of form

l,m,h on 3 characteristics

speed is high, altitude is high, radar type is weapons emissions

set 1) 30 problems with characteristics a,b,c with feedback

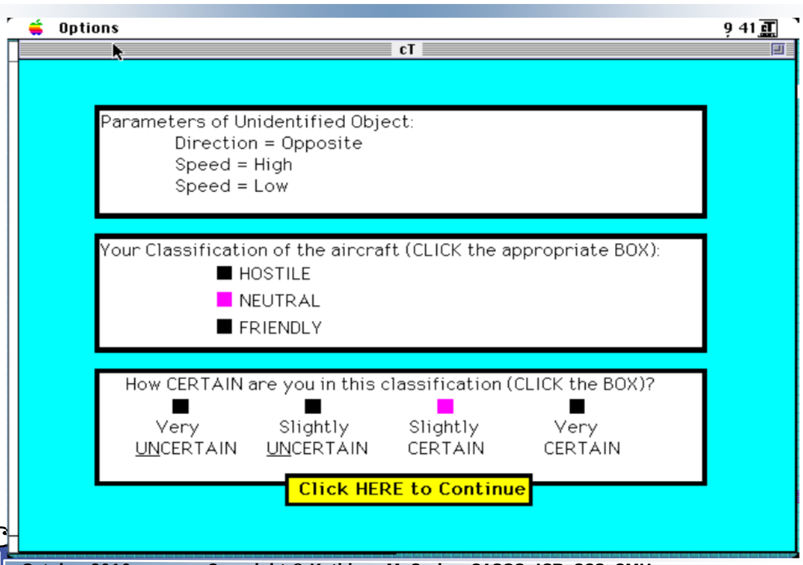
set 2) 30 problems with characteristics a,b,c without feedback

set 3) 30 problems with characteristics d,e,f with feedback

set 4) 30 problems with characteristics d,e,f without feedback and with equipment failures

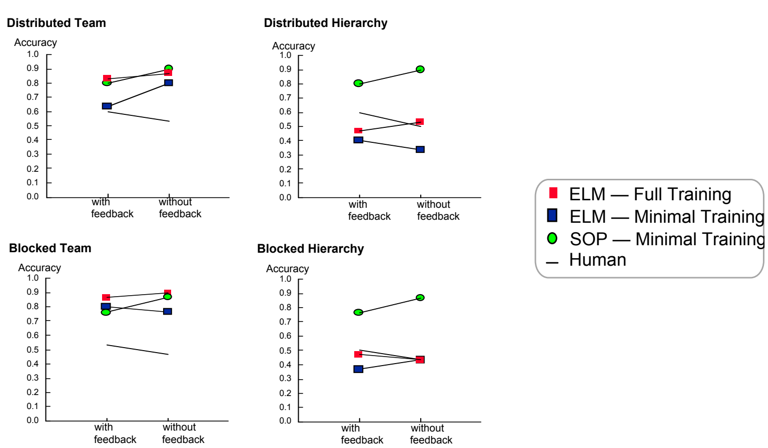
Carnegie Mellon

Details of Experiment



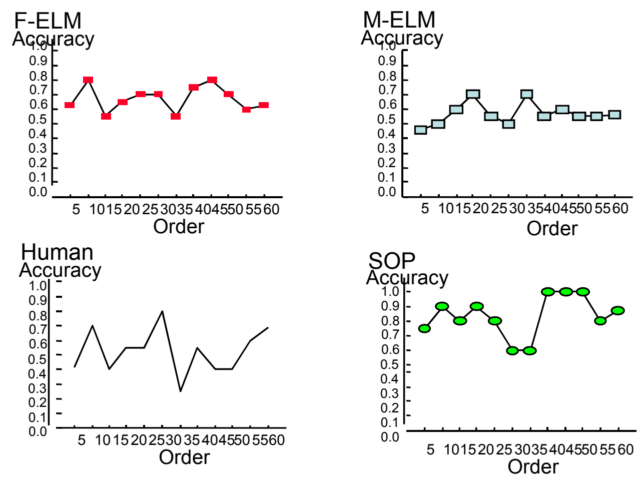
Carnegie Mellon

Verification by Pattern



<Your Name>

Detailed Pattern



Comparison by Value

Organization's Accuracy by Agent Model and Organizational Design

Agent Training	Team		Hierarchy	
	Blocked	Distributed	Blocked	Distributed
ELM full	88.3%	85.0%	45.0%	50.0%
ELM min	78.3%	71.7%	40.0%	36.7%
SOP full/min	81.7%	85.0%	81.7%	85.0%
Human	50.0%	56.7%	46.7%	55.0%

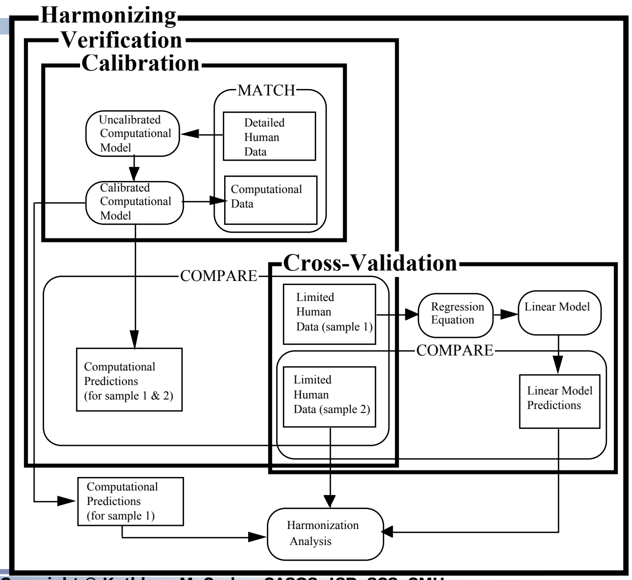
Harmonization
Evidence for Gain from Complexity

Harmonization

- Assessment of theoretical adequacy by comparison with a cross-validated linear model
- Harmonization involves contrasting the predictions of a computational model and a linear model
  - Requires enough cases that you have two samples large enough for statistical analysis
  - Requires that there is a reasonable linear model
- Harmonization can locate areas of the model where the embodied theory is inadequate

statistical comparison process

Harmonizing



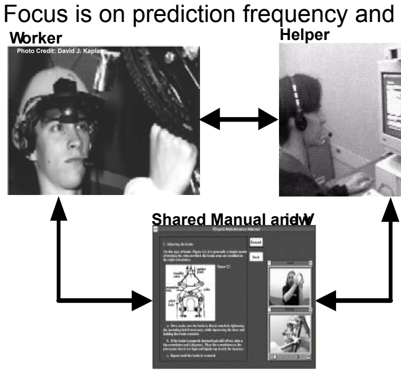
Locating the Linear Model

- Many sources for such a model:
  - Linear model of inputs
  - Easily collected data that might be used by management to make the same prediction the computational model is making
  - Model presented in the literature

Carnegie Mellon
isr Institute for Software Research

Illustration & Harmonization

COMIT - impact of technology on task performance
Human data on a bicycle repair task using various technology (including vide an non-video)
Focus is on prediction frequency and order of communication actions



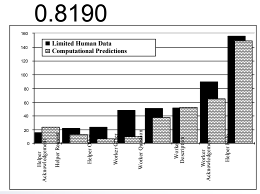
- Coded audio soundtracks
- Took half of subjects and created linear model and verified COMIT
- Linear model based on data accessible by managers
- Generated predictions for remaining subjects for COMIT and linear model
- Calculated R2 across subjects
- Data - 4 people, COMIT 2 people

Carnegie Mellon
isr Institute for Software Research

COMIT Results

Action	Linear Model	COMIT	Correlation
Helper Acknowledgment	0.230	0.653	0.8083
Helper Request	0.498	0.141	0.7488
Helper Other	0.118	0.323	-0.3306
Worker Other	0.307	0.368	0.6886
Worker Question	0.215	0.011	0.4442
Worker Description	0.086	0.182	0.8148
Worker Acknowledgment	0.044	0.473	0.2986
Helper Help	0.310	0.130	0.8190

positive correlation means both models doing well/poorly in same areas, negative correlation means models tend to have opposite predictions



What Happens When Models are "HUGE"

BioWar

Kathleen M. Carley
kathleen.carley@cs.cmu.edu

<Your Name>

BioWar

- City-level Multi-Agent Dynamic-Social-Network model of population response to loss-of-life events – bio-warfare, epidemiological, chemical
- High fidelity
- US Centric
- 5 cities/ 6 cities
  - Pittsburgh, San Diego, San Francisco, Washington DC, Norfolk/Hampton Roads
- 62 diseases
  - Weaponized – Smallpox, anthrax,
  - Background – influenza
  - Epidemic – pandemic influenza, SARS
- Sub-Model for military bases for use in planning, training and assessment of force vulnerability
- GenCity: Network generation: Census based imputation tool for realistic network generation that is context sensitive

The Model …

- Input
  - Military Bases
  - Census data – social, economic, occupation
  - School district data
  - Worksite and entertainment locat
  - Hospitals and clinics locations &
  - Social Network characteristics
  - IT communication procedures
  - Wind characteristics
  - Spatial layout
  - Disease models (symptom level)
  - OTC and Prescription drug info
  - Attack or event
  - Interventions
- Illustrative Output
  - OTC & Prescription drug sales
  - Insurance claim reports (Dr. visi
  - Emergency room reports
  - Absenteeism (school and work)
  - Infected, Contagious, Mortality

Agents move in networks which influence what they do, where, with whom, and what they know, what dise they get, when, how they respond m, etc. Major difference in netwo l disease nd age.

Vast quantities of data
Real and Virtual
Format of virtual to match real
But real data is …
Incomplete
Diverse sources
Inconsistent
Different levels of granularity

Model Objectives

- Automated tools for:
  - Evaluating response policies, data efficacy, attack severity, and detection tools related to weaponized biological attacks *in the presence of background diseases such as flu*
  - Generating high fidelity virtual data for testing detection and fusion algorithms, and exploring potential impact of never before seen events
- Systematically reason about:
  - The rate and spread of disease at the symptom level with high degree of realism
  - Early presentation of diseases as seen in secondary data streams
  - Potential response scenarios, such as inoculation
  - Policy design for potential problems
- Push the frontier of high dimensionality social simulation models (fidelity, precision, speed, comprehensiveness, etc.)

Illustrative Respiratory and GI Diseases Modeled

Disease	ICD9Code	ICD9Name
ACUTE_NASOPHARYNGITIS	460	Acute nasopharyngitis [common cold]
INFLUENZA	487.1	FLU W RESP MANIFEST NEC
INFLUENZA_PNEUMONIA	480.2	PARINFLUENZA VIRAL PNEUM
SEVERE_ACUTE_RESPIRATORY_SYNDROME	480.9	Viral pneumonia, unspecified
BACTERIAL_PHARYNGITIS_ACUTE_NON_STREPTOCOCCAL_NON_GONO	462	ACUTE PHARYNGITIS
GRAM_NEGATIVE_PNEUMONIA_NON_KLEBSIELLA	482.1	PSEUDOMONAL PNEUMONIA
MYCOPLASMA_PNEUMONIA	31	PULMONARY MYCOBACTERIA
PNEUMOCOCCAL_PNEUMONIA	481	PNEUMOCOCCAL PNEUMONIA
PULMONARY_LEGIONELLOSIS	482.89	PNEUMONIA OTH SPCF BACT
STAPHYLOCOCCAL_PNEUMONIA	482.4	STAPHYLOCOCCAL PNEUMONIA
STREPTOCOCCAL_PHARYNGITIS_ACUTE	34	STREP SORE THROAT
STREPTOCOCCUS_PYOGENES_PNEUMONIA	482.3	STREPTOCOCCAL PNEUMONIA*
TUBERCULOSIS_CHRONIC_PULMONARY	11.9	PULMONARY TB NOS*
TUBERCULOSIS_DISSEMINATED	18.9	MILIARY TUBERCULOSIS NOS*
VARICELLA_PNEUMONIA	52.1	VARICELLA PNEUMONITIS
VIRAL_PHARYNGITIS_ACUTE_NON_HERPETIC	79.3	RHINOVIRUS INFECT NOS
BRONCHIAL_ASTHMA	493.1	INT ASTHMA W/O STAT ASTH
BRONCHITIS_CHRONIC_SIMPLE	491	SIMPLE CHR BRONCHITIS
PULMONARY_EMPHYSEMA	492	Emphysema, NOS
PLAGUE_PNEUMONIA	20.2	SEPTICEMIC PLAGUE
ANTHRAX_INHALATIONAL	22.1	Respiratory anthrax
Gastro-intestinal:		
STAPHYLOCOCCAL_GASTROENTERITIS_FOOD_POISONING	8.41	STAPHYLOCOCC ENTERITIS
BOTULISM	5.1	BOTULISM
CAMPYLOBACTER_ENTERITIS	8.43	Intestinal infection due to campylobacter
GIARDIASIS_INTESTINAL	7.1	Infection by Giardia lamblia
SALMONELLA_ENTEROCOLITIS_NON_TYPHI	3	SALMONELLA ENTERITIS
VIRAL_GASTROENTERITIS	8.8	VIRAL ENTERITIS NOS

Carnegie Mellon

Contagious Diseases

- Modeled as either attacks or outbreaks

Characteristic	Influenza	Avian Flu
Start tick	402	402
End tick	940	940
Scale	200	20
Number of [illegible]	[illegible]	1
Number of [illegible]	[illegible]	1
Low basera[illegible]	[illegible]	.0001
High baserate	.0004, .0008	.0012
Mean baserate	.000496, .000496	.00045
Low transmisivity	.02, .03	.01
High transmisivity	.03, .06	.08
Mean transmisivity	.033, .033	.063

The expert's don't agree
Do sensitivity analyses

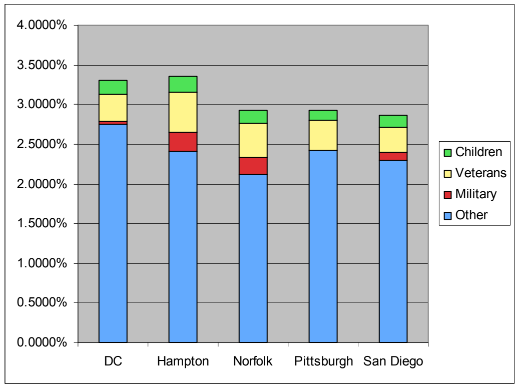
Carnegie Mellon

Symptom-based Behavior

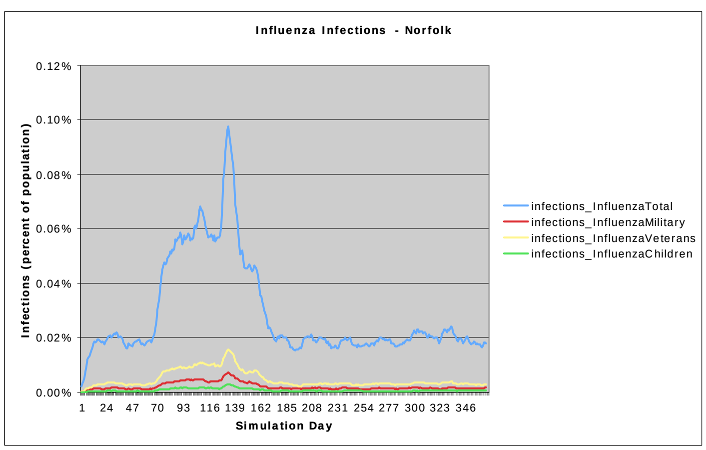
- People who contract an anthrax infection and display fevers/chills may consider their symptoms to stem from influenza/cold, and not significantly alter their behavior. However, if they began having shortness of breath, chest pains, or other symptoms suggestive of a serious problem, they would likely stay home from work, go to doctor, or go to an emergency room.
- A set of symptom severity thresholds guides an agent's decisi[illegible] to visit a me[illegible] erities of ob[illegible] fying a be[illegible]
  - Low severity - no effect
  - Mild severity - go to the pharmacy
  - High severity - go to the doctor
  - Extreme severity - go to the emergency department
- If alerted, individuals will lower their threshold to seek more advanced care.

High flexibility
Many behavioral sub-models
Facilitates modeling the disease that has never been seen

Projected Mortality per City



Illustrative Over Time Projection



Carnegie Mellon
isr Institute for Software Research

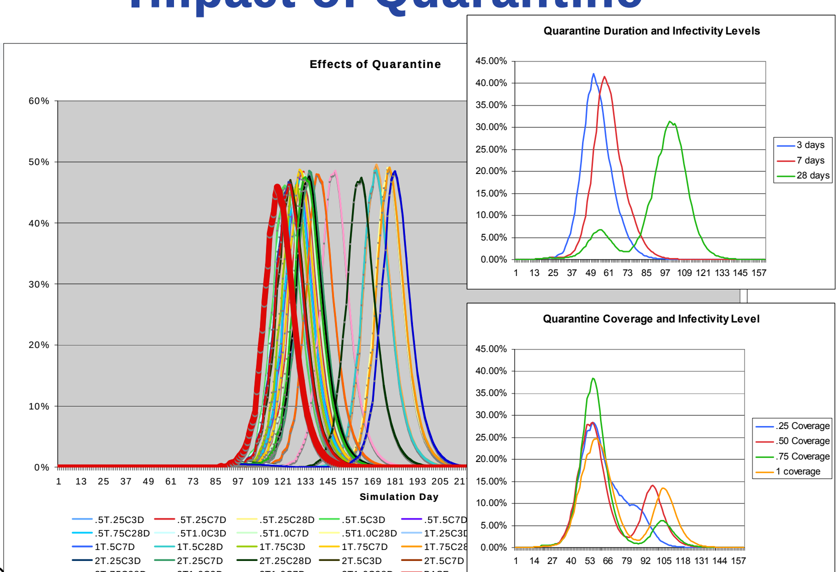
Interventions

- Quarantine
  - Fraction of infected quarantined - .25, .5, .75, 1
  - “reduced contact” - 90% reduction
  - Length – 3 7 28 days
  - When – .05% infected, 1% infected, 2% infected
- School Closures
  - Historic evidence that even when schools close, children are in contact
  - “another Saturday”
  - Length – 3 7 28 days
  - When – .05% infected, 1% infected, 2% infected
- Vaccinations
  - May not be 100% effective - .25, .5, .75
  - May not be 100% distributed - .25, .5, .75

CASOS
October 2016

Carnegie Mellon
isr Institute for Software Research

Impact of Quarantine

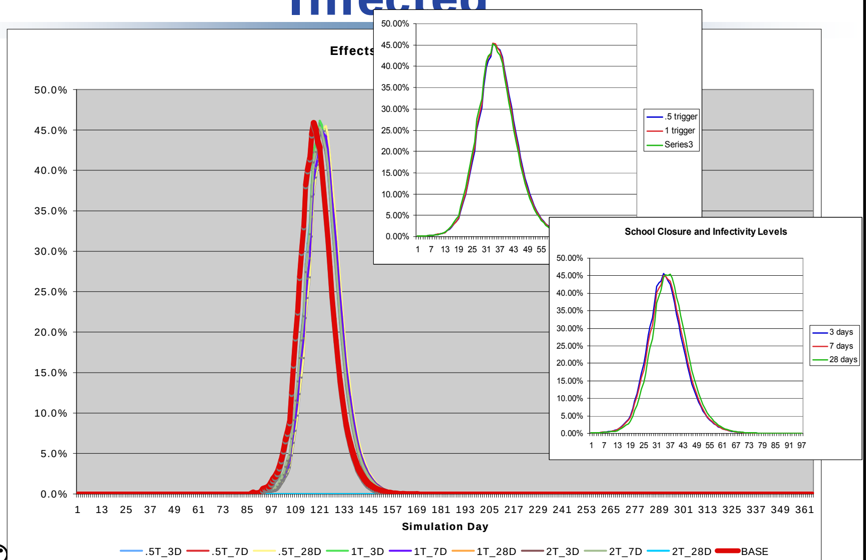


CASOS
October 2016

CASOS

Carnegie Mellon
isr Institute for Software Research

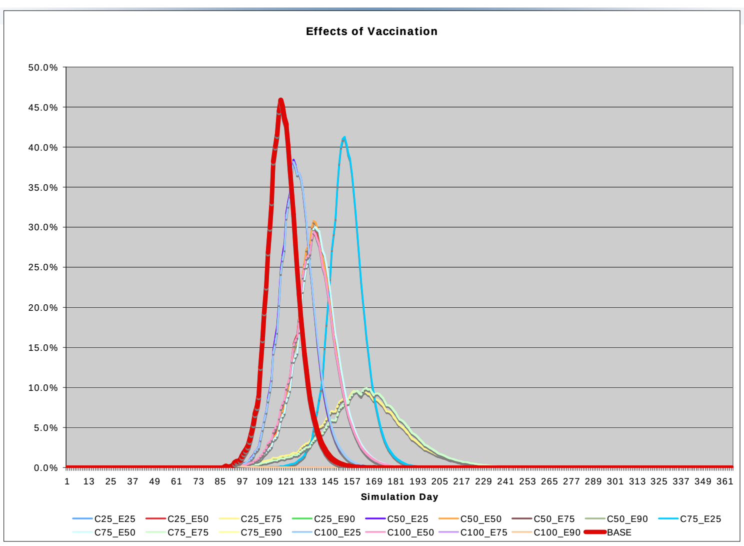
Impact of School Closure on Infected



CASOS

Carnegie Mellon
isr Institute for Software Research

Impact of Vaccination on Infected



CASOS

CASOS

Validated Features

- Anthrax attack & disease model ("docking" or computational model alignment with IPF, Incubation-Prodromal-Fulminant – a revised SIR -- Model)
- Smallpox attack & disease model, docking with SIR
- School absence
- Work absence
- Doctor visit
- ER visit

Validation over Time (C5=last "Challenge", C1..4=previous "Challenges")

Type	C1	C2	C3	C4	C5
Docking: Comparison against another model				✓	✓
Generic Pattern: Showing pattern for each generated data stream matches observed patterns	✓	✓	✓	✓	✓
Characteristic Matching: Showing for each generated output data stream that it has correct seasonal or daily pattern		✓	✓	✓	✓
Relative Timing of Peaks: Showing time between peaks for dif. data streams matches observed dif.		✓	✓	✓	✓
Empirical Pattern: Showing pattern for each generated data stream matches empirical pattern – best for input streams			✓	✓	✓
Within Bounds: Showing for each generated output data stream that the mean of simulated stream falls within min/max of that stream for real data			✓	✓	✓
First moments: Showing for each generated output data stream that mean is not statistically different than real data – yearly, monthly or daily				✓	✓

Carnegie Mellon

isr Institute for Software Research

Verification & Validation

- Internal Tuning
  - Existing data sets to parameterize
    - Reporting delays
    - Disease profiles
    - Agent social networks
    - Age, race, gender, economic differences on behavior and susceptibility
    - Variation in behavior by time of day, day of week, month, season
    - Usage of IT
  - Sources
    - Behavioral surveys
    - Nursing studies
    - CDC reports
    - Communication studies
    - OTC purchases
  - City profiling
    - Census data
    - School district
    - Maps
- Validation – emergent behavior compared to real data
    - Death reports
  - General behavior
    - Disease replication for historic cases
    - Pharmacy purchases
    - Cold shelf and influenza spike
  - Influenza
    - Grade School Absenteeism
    - ER reports
    - OTC purchases
  - Level
    - General pattern
    - Mean, std
    - Variation in disease reports by day of week, month, season, local

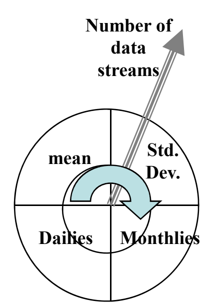
CASOS

Carnegie Mellon

isr Institute for Software Research

What Data Streams is Validation Done On

Data Stream	C2	C3
Work absenteeism	**Yes**	**Yes**
School absenteeism	**No**	**Yes**
ER visits	**Yes**	**Yes**
Doctor visits	**Yes**	**Yes**
OTC drug purchase	**No**	**Yes**
Sentinel trace	**No**	**No**
Network distribution	**No**	**Yes**



CASOS

CASOS

What validation or tuning has been done

- Work absenteeism within the lower & higher empirical bounds
- School absenteeism within the lower & higher empirical bounds
- Doctor visits within the lower & higher empirical bounds
- ER visits within the lower & higher empirical bounds
- Drug sales per group is near the empirical mean
- Face validation of a sentinel population trace
- Automated output check

Sources of Data for Validation

- NCES Indicator 17 & Indicator 42-1, for calculating school absenteeism
- CDC Advance Data, from Vital and Health Statistics, no. 326, 2002, for calculating ER visits
- CDC Advance Data, from Vital and Health Statistics, no. 328, 2002, for calculating doctor visits
- 1997 US Employee Absences by Industry Ranked (http://publicpurpose.com/lm-97absr.htm) for determining work absenteeism
- OTC Sales by Category from AC Nielsen (http://www.chpa-info.org/statistics/otc_sales_by_category.asp) and PSC's FRED data for pharmacy OTC drug sales

Empirical School Absenteeism Bounds

- Data from NCES Indicator 17 & Indicator 42-1
- NCES Indicator 42-1 gives total absenteeism rate of 4.9% for 8th graders in urban fringe/large town
- NCES Indicator 17 gives the absenteeism reasons of illness of 53.1%, skipping 9.0%, others 37.9%.
- For 10th graders, the corresponding total absenteeism rate is 6.2%, absenteeism due to illness of 45.4%, skipping 15.6%, others 39.0%
- For 12th graders, the corresponding total absenteeism rate is 8.6%, portion of it due to illness is 34.2%, skipping 26.1%, others 39.7%
- As we don't have reasons other than illness or skipping in C3, the lower bound for all schools is 3.04%, with the upper bound of 5.18% absenteeism rate

School Absenteeism

City, percent of simulated population	Empirical lower bound	Empirical higher bound	No Attack (mean)	Anthrax (mean)	Smallpox (mean)
Norfolk, 20%	3.04%	5.18%	3.45%	3.75%	3.55%
Pittsburgh, 20%	3.04%	5.18%	3.52%	4.67%	4.46%
San Diego, 20%	3.04%	5.18%	3.78%	3.81%	5.57%
Veridian Norfolk, 20%	3.04%	5.18%	3.73%	4.05%	4.31%

Empirical Work Absenteeism Bound

- Data from the 1997 US Employee Absences by Industry Ranked
- As we don't yet have the specifics of workplace types in C3, we take the lower bound to be the lowest absence rate of any industry type, the higher bound to be the highest.
- So, from the data, we have the lower bound of 2.3% and the higher bound of 4.7% absenteeism rate.

Work Absenteeism

City, percent of simulated population	Empirical lower bound	Empirical higher bound	No Attack (mean)	Anthrax (mean)	Smallpox (mean)
Norfolk, 20%	2.30%	4.79%	2.72%	4.65%	2.82%
Pittsburgh, 20%	2.30%	4.79%	2.77%	5.79%	3.99%
San Diego, 20%	2.30%	4.79%	3.26%	4.99%	5.78%
Veridian Norfolk, 20%	2.30%	4.79%	3.16%	5.50%	3.81%

CarnegieMellon

Empirical Doctor Visits Bound

- Data from CDC Advance Data, Vital & Health Statistics, No. 328, 2002
- Table 1 of the report shows MSAs (metropolitan areas) have 294.6 visits per 100 persons per year
- The lower bound is based on major disease categories, while the higher bound is based on all disease categories in the simulation
- Table 11 of the report gives 14.1% of all the causes of visits to fall within major disease categories of infectious & respiratory diseases, and 54.7% for all disease categories in the simulation
- This gives us the lower bound of 0.415 visits per person per year and the higher bound of 1.611 visits per person per year

CarnegieMellon

Doctor Visit (visit per person per year)

City, percent of simulated population	Empirical lower bound	Empirical higher bound	No Attack (mean)	Anthrax (mean)	Smallpox (mean)
Norfolk, 20%	0.415	1.611	0.499	0.476	0.499
Pittsburgh, 20%	0.415	1.611	0.493	0.485	0.573
San Diego, 20%	0.415	1.611	0.726	0.753	0.796
Veridian Norfolk, 20%	0.415	1.611	0.707	0.821	0.738

Carnegie Mellon
isr Institute for Software Research

Empirical ER Visits Bound

- Data from CDC Advance Data, Vital & Health Statistics, No. 326, 2002
- Table 1 of the report shows MSAs have 37.6 visits per 100 persons per year
- The lower bound is based on major disease categories, the higher bound on all disease categories in the simulation
- Table 7 in the report gives us 14.8% of all causes tp fall within major disease categories of infectious & respiratory illness, and 77.7% of all disease categories of the 62 disease present in the simulation
- So the lower bound is 0.056 visits per person per year, the higher bound 0.232 visits per person per year

Carnegie Mellon
isr Institute for Software Research

ER Visit (visit per person per year)

City, percent of simulated population	Empirical lower bound	Empirical higher bound	No Attack (mean)	Anthrax (mean)	Smallpox (mean)
Norfolk, 20%	0.056	0.232	0.112	0.108	0.112
Pittsburgh, 20%	0.056	0.232	0.109	0.106	0.129
San Diego, 20%	0.056	0.232	0.149	0.159	0.188
Veridian Norfolk, 20%	0.056	0.232	0.161	0.187	0.168

Carnegie Mellon
isr Institute for Software Research

BioWar Validation with Smallpox SIR in Details

CASOS

Carnegie Mellon
isr Institute for Software Research

Docking BioWar with the SIR model

- Our method
  - Aligning the complex BioWar model with the SIR model
- The Susceptible-Infected-Recovered (SIR) model
  - Widely used to model the spread of a contagious disease in epidemiology literature
  - A simpler and well-understood model
- Our goal
  - Validate BioWar
  - Demonstrate that it is at least able to produce fairly similar results to the accepted standard epidemiological model
  - Obtain a sense of validity needed to develop the new model
- Challenges
  - Their radically different structures and assumptions

CASOS

The two models

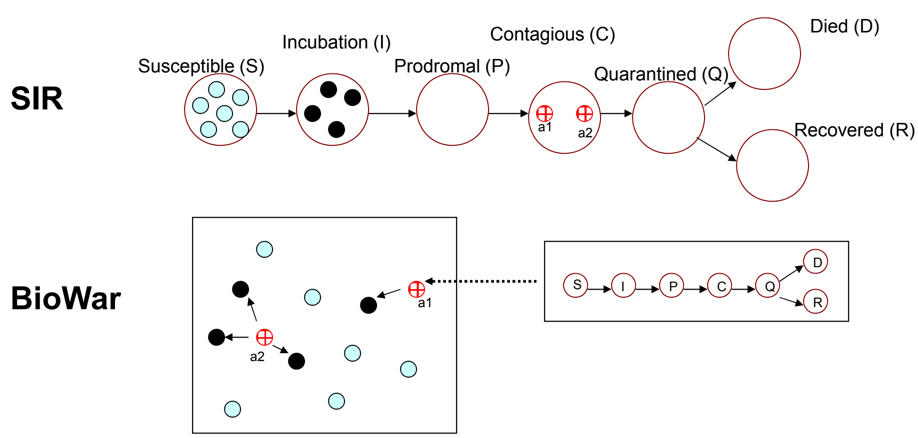
Carnegie Mellon
isr Institute for Software Research

Disease stages of smallpox infection

- Incubation
  - No symptoms, around 12-14 days
- Prodrome (early-symptomatic)
  - Sometimes contagious
  - Non-specific symptoms
    - High fever, head and body aches, and possibly vomiting
  - around 2-4 days
- Fulminant (late-symptomatic)
  - Contagious
  - Specific symptoms
    - Early rash (about 4 days)
    - Pustular rash (about 5 days)
    - Pustules and scabs (about 5 days)
    - Resolving scabs (about 6 days)

Carnegie Mellon
isr Institute for Software Research

The two models of smallpox outbreaks

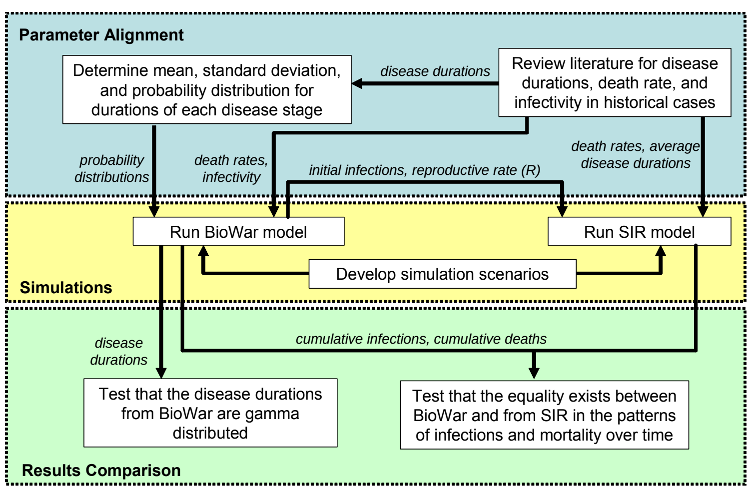


Qualitative differences

- Population assumptions
  - heterogeneous vs homogeneous
- Disease model design
  - Micro vs macro
- Computational process
  - BioWar needs higher computational power than SIR
- Initialization
  - Individual level info. vs population level info.
- Parameterization
  - Parameterized attack scenarios vs parameterized infection info.

The two models

The processes of model alignment



The Process of Docking

Carnegie Mellon
isr Institute for Software Research

Parameter alignment

- Disease stage durations
  - The duration that a infected person stay in a certain disease stage
  - BioWar: a gamma distributed probability functions for each of the disease stage
  - SIR: the ratio in transition probability moving from one stage to another is set to the mean value of the probability function
- Reproduction rate
  - The number of secondary cases infected by one infected person
  - BioWar: generate the number of infected and reproduction rate based given an attack scenario
  - SIR: use the number of infected and the reproduction rate generated from BioWar as input parameters

CASOS

The Process of Docking

Carnegie Mellon
isr Institute for Software Research

Simulation assumptions

- Washington, DC area, scaled down to 10% of its original size
- The total population after scaling is about 55,900
- Assume that the attack goes undetected
- Assume that no public health responses or warnings occur after the attack

CASOS

Simulations

Carnegie Mellon
isr Institute for Software Research

Attack scenarios

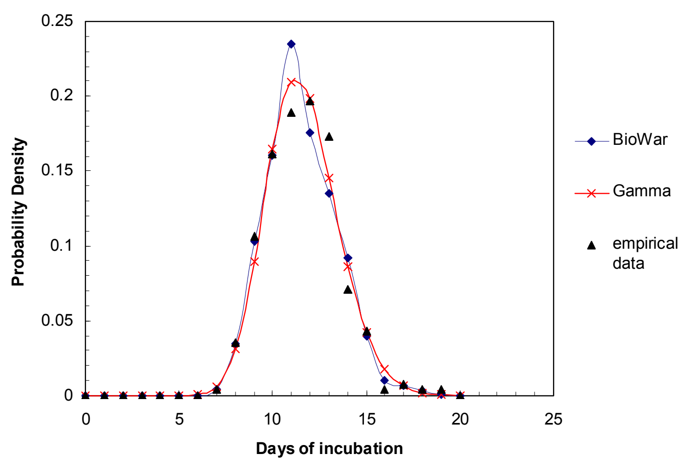
Scenarios	Residual immunity (% of total population)	Fresh vaccination (% of total population)	Is infected population quarantined?
base	0%	0%	no
vaccination	46%	50%	no
quarantine	46%	0%	yes (on average, 2 day after the onset of rash)

CASOS

Simulations

Carnegie Mellon
isr Institute for Software Research

Stage durations are proportional to the gamma distributions



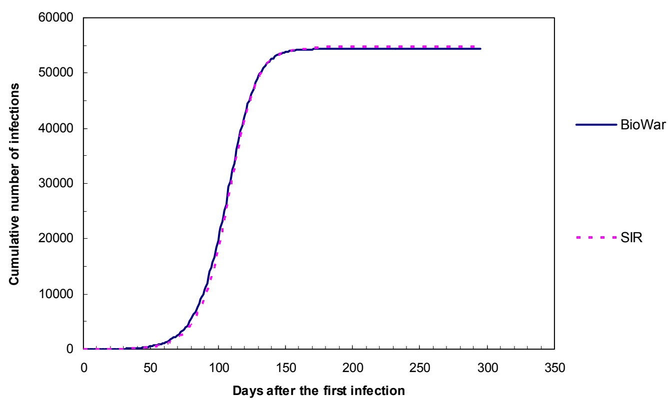
CASOS

Results

CASOS

Carnegie Mellon
isr Institute for Software Research

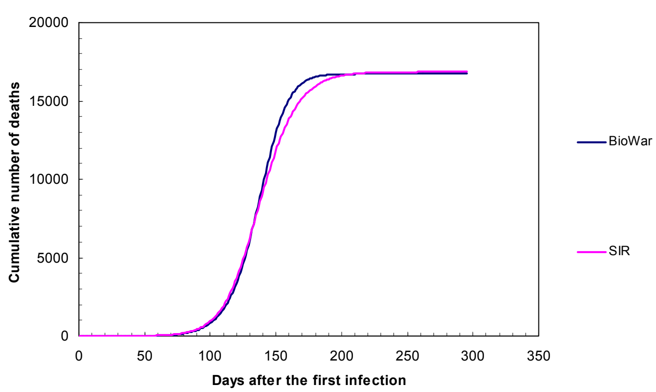
Cumulative infections from the two models are comparable



Results

Carnegie Mellon
isr Institute for Software Research

Mortality from the two models are comparable



Results

Carnegie Mellon
isr Institute for Software Research

Discussions

- On a gross level the two models give approximately the same results but subtle differences exist because of the differences in mixing assumptions
- The differences were most evident in the scenarios involving vaccination and quarantine
  - the agent-level complexity required for such scenarios is easily accommodated by BioWar, but not by SIR
- BioWar provides a way to manage these model parameters in order to represent the heterogeneous properties of individuals
  - The emergent properties of agent-based models which cannot be generated from the SIR model
  - E.g., the reproductive rate is actually partly the result of interactions between individuals and these interactions

CASOS
Conclusions

Carnegie Mellon
isr Institute for Software Research

Contributions

- Develop a methodology to partially validate a complex agent-based model
  - We provided a method to align a multi-agent model of weaponized biological attacks, BioWar, with the classical susceptible-infected-recovered (SIR) model
- Identify the differences in model inputs and model assumptions of smallpox simulations
  - It is important for policy makers to understand the differences and similarities between agent-based models and the SIR model before making decisions based on any one model

CASOS
Conclusions

Carnegie Mellon
isr Institute for Software Research

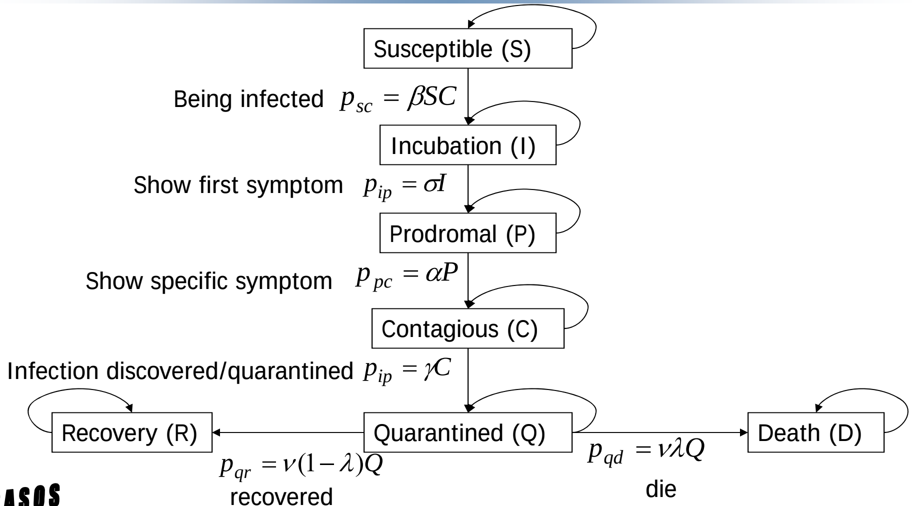
Model Alignment

- Also referred to as “docking”
- The comparison of two computational models to see if they can produce equivalent results
- Uncover the differences and similarities between models
- Reveal the relationships between the different models’ parameters, structures, and assumptions

CASOS

Carnegie Mellon
isr Institute for Software Research

The revised SIR model for comparison



The two models

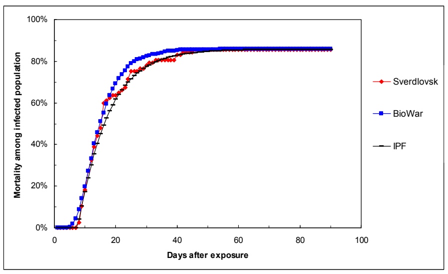
CASOS

CASOS

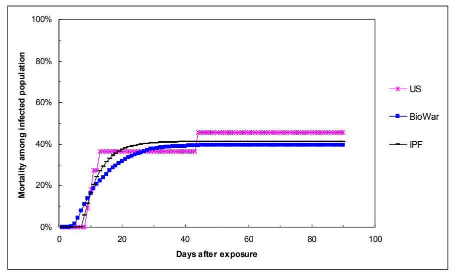
Carnegie Mellon
isr Institute for Software Research

Anthrax: BioWar vs. IPF based on Time to Death

Sverdlovsk

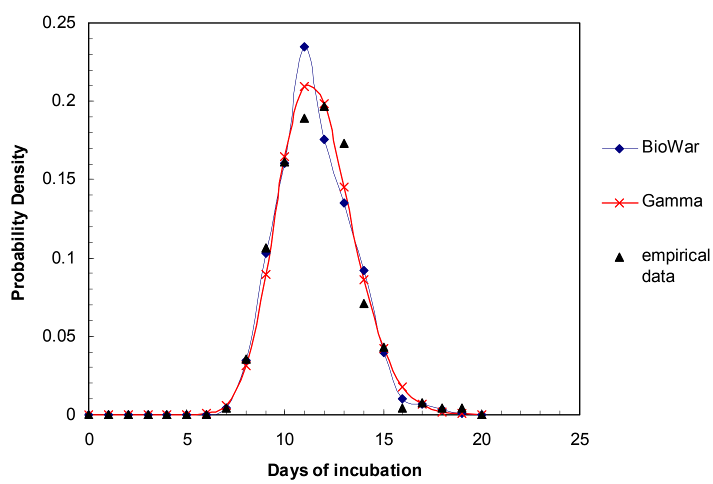


US Mail



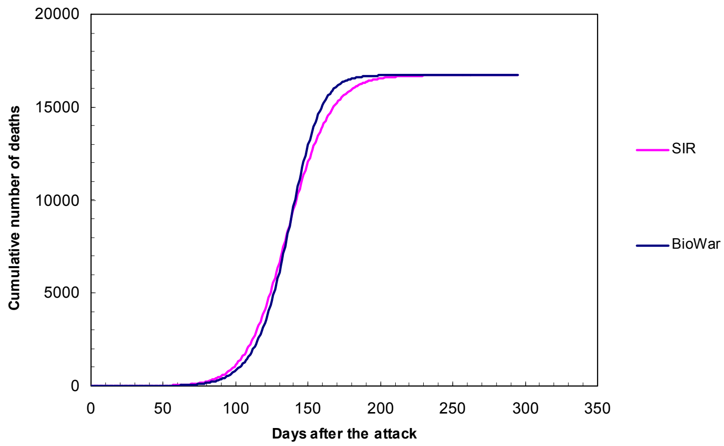
Carnegie Mellon
isr Institute for Software Research

Smallpox Incubation



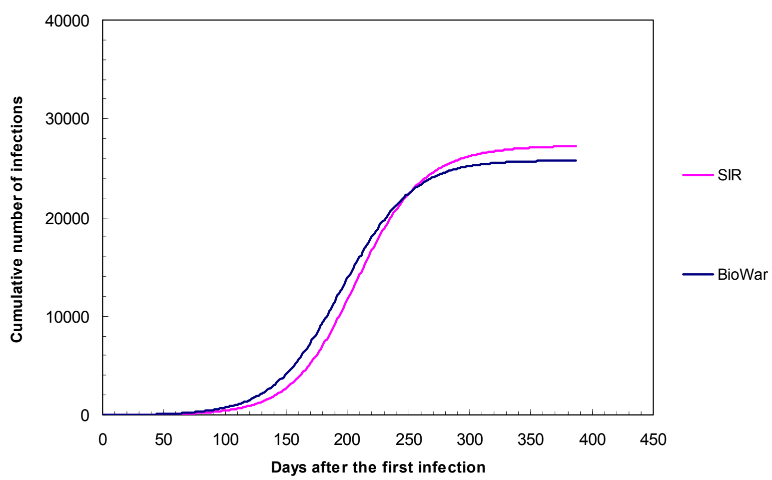
Carnegie Mellon

Smallpox Death: SIR vs. BioWar



Carnegie Mellon

Smallpox Infections: SIR vs. BioWar



Carnegie Mellon
isr Institute for Software Research

Limitations

- Actual transmission rates during pandemic flu are unknown. By definition these will be more than normal influenza strain but we don't know by how much.
- The model does not include behavior such as panic. We don't know how people's behavior will change and that may affect transmission and spread of disease.
- Since we don't know what strain will lead to the pandemic, we don't know the effects of vaccination and anti-virals
- The population is assumed to be static. Population influx and emigration are not included.
- The interplay between co-morbidity and influenza may not be clear during a pandemic, but will affect mortality rates, hospitalization, health care resource use. Mortality and transmission rates will be different but are unknown.
- There may be secondary spreads within pandemic.

- Note: these are limitations to most if not all pandemic models.

Carnegie Mellon
isr Institute for Software Research

The Need for Automated Validation

Carnegie Mellon

Why is There The Need for Validation Automation?

- To effectively use simulations, human analysts require to have justifiable and measurable confidence in them
- Real-time revalidation of BioWar to changing real world situations is especially important
- Validation is difficult to do manually due to model complexity and variable interactions
- Scaling BioWar up to take in more models – local models and diverse secondary data streams – would increase the code size and the complexity of validation
- Simulation assumptions are numerous and often implicit.
- Knowledge underlying simulations is NOT usually codified and operable. Codified knowledge is critical.

Carnegie Mellon

Manual Parametric Study Experience

- Parametric study of 3-4 parameters for BioWar which has hundreds of semi-constrained parameters
- Lessons learned:
  - Intelligent analysis and response techniques are needed to optimize the search in parameter space
  - Automated tools to create and execute parametric studies (called virtual experiments) are needed
- What matters is elucidating cause-and-effect relationships using simulations, not just patterns or statistical associations.

<Your Name>

Carnegie Mellon
isr Institute for Software Research

WIZER Description

- WIZER (What-If AnalyZER):
  - a coupled inference and simulation engine
  - that can deal with high dimensional, symbolic, stochastic, emergent, and dynamic nature of complex multi-agent systems
  - by performing knowledge-intensive data-driven search steps via an inference engine constrained by simulation
  - and by explaining the reasoning behind inferences using both the simulation and the inference engine
- WIZER views simulation systems as knowledge systems
- WIZER controls the simulator and also the design and execution of experiments based on the Scientific Method
- It exploits the knowledge intensity of social systems to bound search through model space

CASOS

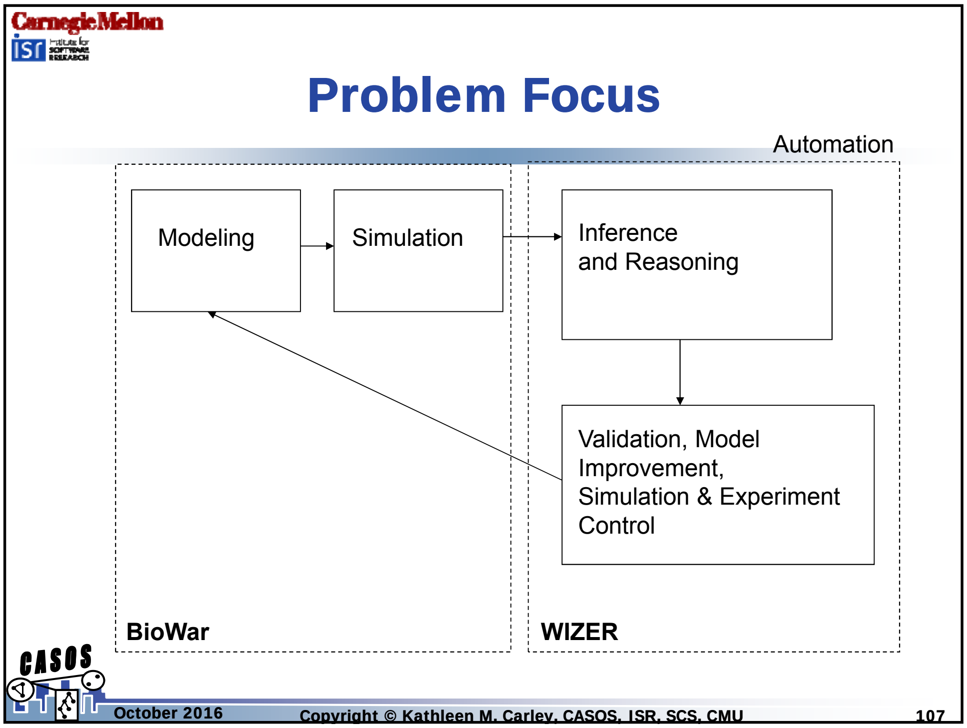
Carnegie Mellon
isr Institute for Software Research

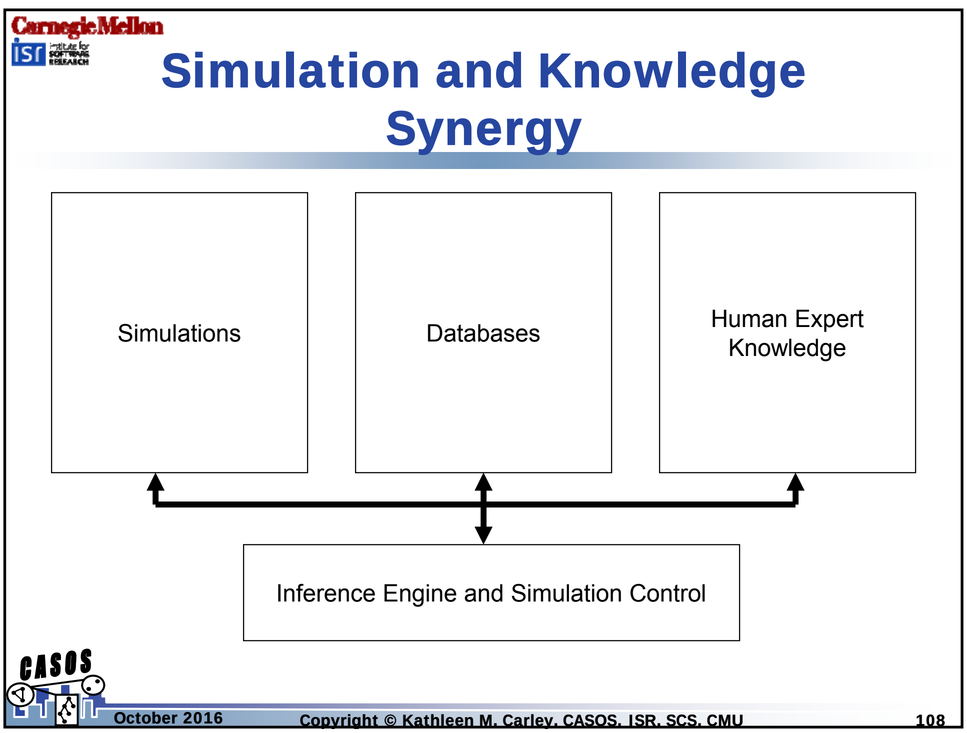
WIZER Proposed Capabilities

- WIZER provides automation of validation and model-improvement
- Human analyst can see his/her domain knowledge in action in simulations
- WIZER would perform inferences and narrow search through parameter and meta-model space, cutting down the analysis time
- Making simulation assumptions explicit and operable, facilitating multiple experts collaboration and objective assessment of expert opinions

CASOS

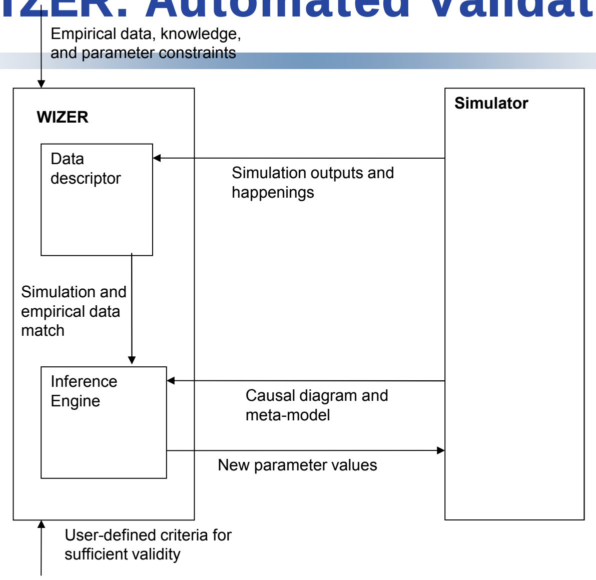
CASOS


Carnegie Mellon
Problem Focus
Automation
Modeling
Simulation
Inference
and Reasoning
Validation, Model
Improvement,
Simulation & Experiment
Control
BioWar
WIZER
October 2016
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU
107


Carnegie Mellon
Simulation and Knowledge
Synergy
Simulations
Databases
Human Expert
Knowledge
Inference Engine and Simulation Control
October 2016
Copyright © Kathleen M. Carley, CASOS, ISR, SCS, CMU
108

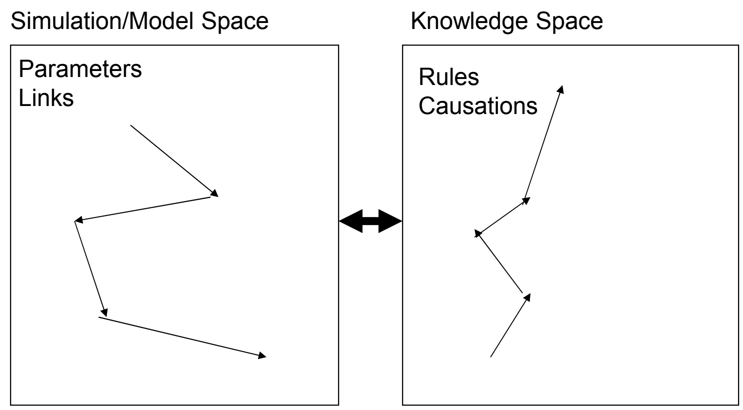
Carnegie Mellon
isr Institute for Software Research

WIZER: Automated Validation



Carnegie Mellon
isr Institute for Software Research

Search Spaces



Assumption-based Causal Model

- WIZER consists of an assumption-based causal model, which is a quintuple
  - *M = (A, V, L, S, U)*
- where
- A = variables (propositional symbols) denoting
- assumptions, which are exogenous variables
- *V* = endogenous variables (propositional symbols)
- *L* = logical and/or causal mapping
- *S* = simulation (mechanism) or virtual experiment
- *U* = uncertainty measure, for example, in the
- form of the degrees of support and plausibility

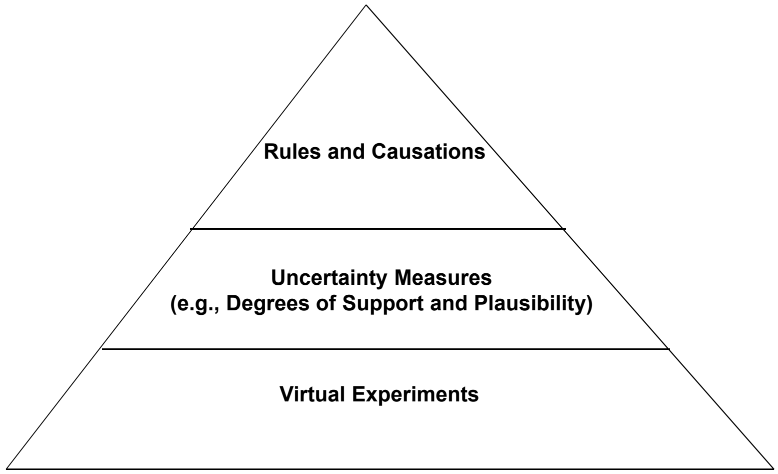
WIZER Inference Engine

The inference engine is based on "if-then" and "causal" rules

Type of Knowledge	Logical Representation	Meaning
A fact	P1	P1 is true
A rule	P1 => P2	P1 implies P2
An uncertain fact	a1 => P1	If assumption a1 is true, then P1 is true
An uncertain rule	a2 => (P1 => P2) or equivalently P1 ∧ a2 => P2	If assumption a2 is true, then P1 implies P2

Type of Knowledge	Logical Representation	Meaning
A fact	P1	P1 is true
A causation	P1 caused P2	P1 caused P2
An uncertain fact	a1 => P1	If assumption a1 is true, then P1 is true
An uncertain causation	a2 => (P1 caused P2)	If assumption a2 is true, then P1 caused P2

Layers



First WIZER Iteration

- Four runs of 100% Hampton with no attacks.
- WIZER Data Descriptor checks:
  - ER registration is above the empirical bound of 0.232 visits per person per year
    - edregistration-yearly-avg,2.24856,0.056,0.232,above the bounds
  - Doctor visit is above the empirical bound of 1.611 visits per person per year
    - insuranceclaim-yearly-avg,3.16572,0.415,1.611,above the bounds
  - School absenteeism is within the empirical bound of absence rate (in percentage)
    - school-yearly-avg,3.62525,3.04,5.18,within bounds
- WIZER Inference Engine outputs:
  - Increase the behavior threshold for going to ER
  - Increase the behavior threshold for going to doctor office
  - Leave the behavior threshold related to school going behavior

Carnegie Mellon
isr Institute for Software Research

Conclusion

- WIZER would enable the human analyst to have a better confidence interval in a simulation system
  - The human analyst can see his/her domain knowledge in action in simulations
  - Automated validation and explanation
- WIZER would perform inferences and narrow search based on knowledge and simulations with empirical values, assumptions, rules, and causations, cutting down the analysis time
- Making simulation assumptions explicit and operable

CASOS