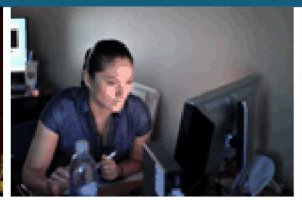
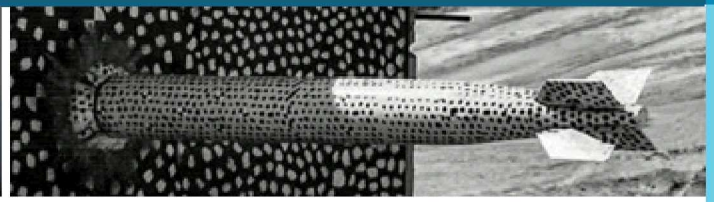
SAND2020-8055PE

DARPA Ground Truth Program Results





Asmeret Naugle, Test & Evaluation Team

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia Test & Evaluation Team


Asmeret Naugle


Kiran Lakkaraju


Laura Swiler


Christy Warrender


Dan Krofcheck


Steve Verzi


Jaimie Murdock


Ben Emery


Mike Bernard


Vicente Romero

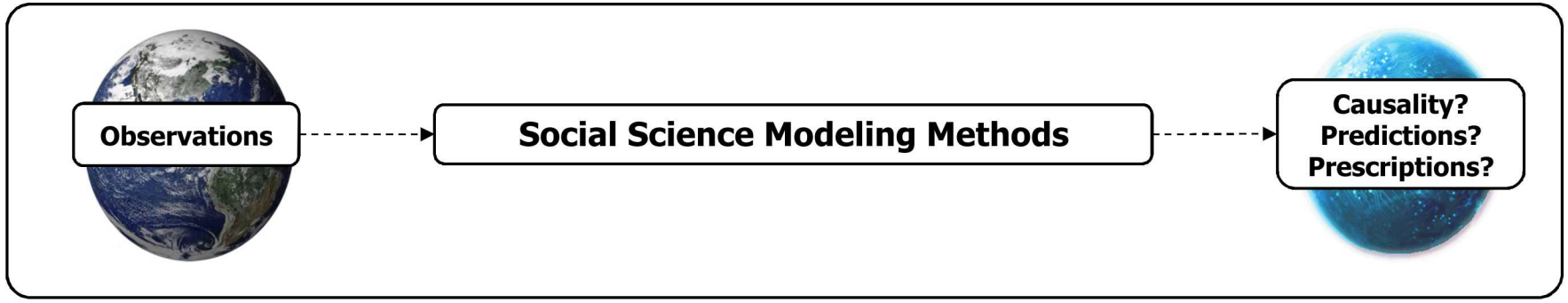
Vamshi Balanaga

Michael Livesay

DARPA Ground Truth: Motivation

Social science is hard

- Can't test validity without ground truth
- Can't freely experiment
- Biases in data and how we gather it
- Difficult to compare methods



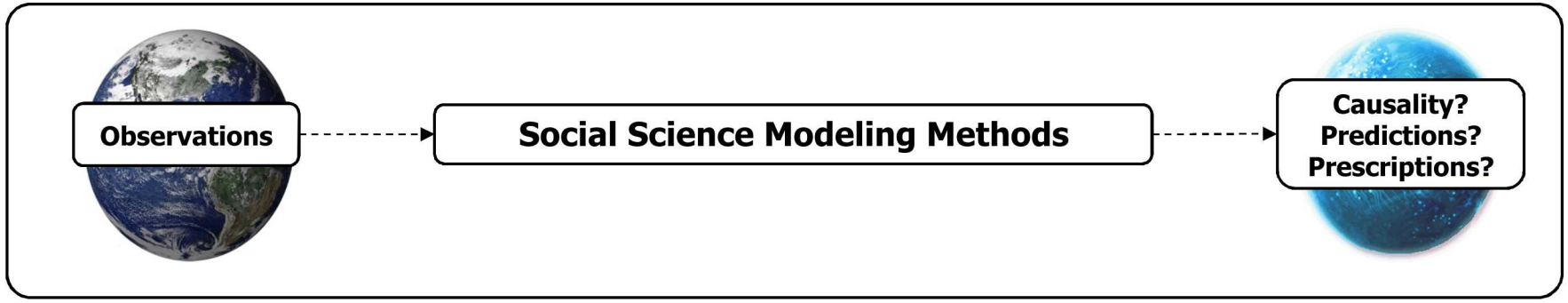
DARPA Ground Truth: Motivation

Social science is hard

- Can't test validity without ground truth
- Can't freely experiment
- Biases in data and how we gather it
- Difficult to compare methods

Ground Truth Program

- Known ground truth
- Simulations enabled experimentation
- TA2 teams collected their own data
- Methods tested on the same systems



Evaluating the TA1 Simulations: The Tests

DARPA Ground Truth: Evaluating the TA1 Simulations 1/1

1. **Simulation accessibility**
   - Definition: Ability to accommodate a range of social science research methods
   - Evaluation: Demonstrate accessibility to a negotiated list of methods

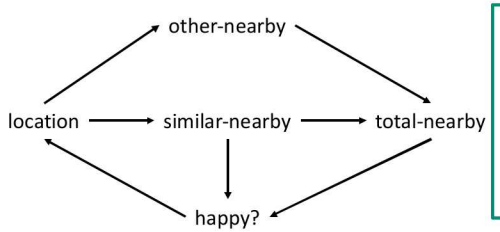
2. **Verifiability of ground truth**
   - Definition: Utility of the simulation as a test bed for inferring ground truth
   - Evaluation: Compare code to causal graph ("ground truth"), verification tests

3. **Plausibility**
   - Definition: Ability to provide non-trivial results without requiring external intervention
   - Evaluation: Entropy, variance

Can the simulations handle social science data collection methods?

Observational data	Event journals
Interviews	Passive data collection
Surveys	Randomized trial
Ethnographic observations	Experiments
Laboratory experiments	Proxy experiments…

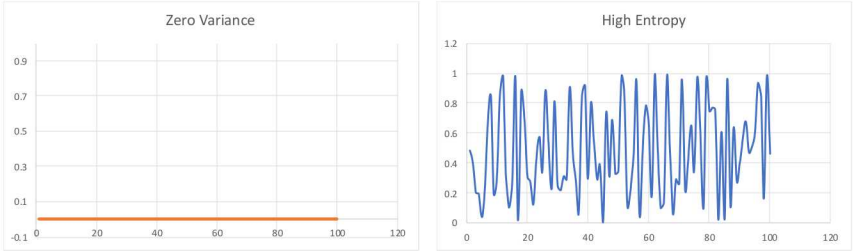
Does the ground truth accurately represent the simulation?




```
if all turtles are happy then stop
 for each turtle
    if unhappy, randomly move to new unoccupied patch
    similar-nearby count =
      number of neighbors with color = turtle's color
    other-nearby count =
      number of neighbors with color != turtle's color
    total-nearby = similar-nearby + other-nearby
    happy? = yes if
      similar-nearby >= (%-similar-wanted * total-nearby/100)
```

adapted from Wilensky (1997)

Is the simulation a self sustaining virtual world?



DARPA Ground Truth: Evaluating the TA1 Simulations 1/1

1. **Simulation accessibility**
   - Definition: Ability to accommodate a range of social science research methods
   - Evaluation: Demonstrate accessibility to a negotiated list of methods

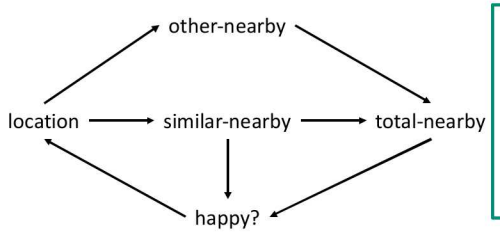
2. **Verifiability of ground truth**
   - Definition: Utility of the simulation as a test bed for inferring ground truth
   - Evaluation: Compare model to causal graph ("ground truth"), verification tests

3. **Plausibility**
   - Definition: Ability to provide non-trivial results without requiring external intervention
   - Evaluation: Entropy, variance

Can the simulations handle social science data collection methods?

Observational data	Event journals
Interviews	Passive data collection
Surveys	Randomized trial
Ethnographic observations	Experiments
Laboratory experiments	Proxy experiments…

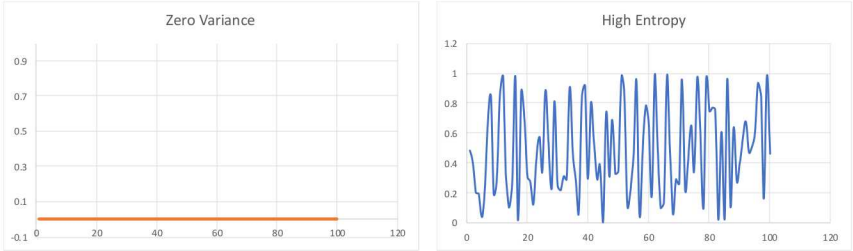
Does the ground truth accurately represent the simulation?




```
if all turtles are happy then stop
 for each turtle
    if unhappy, randomly move to new unoccupied patch
    similar-nearby count =
      number of neighbors with color = turtle's color
    other-nearby count =
      number of neighbors with color != turtle's color
    total-nearby = similar-nearby + other-nearby
    happy? = yes if
      similar-nearby >= (%-similar-wanted * total-nearby/100)
```

adapted from Wilensky (1997)

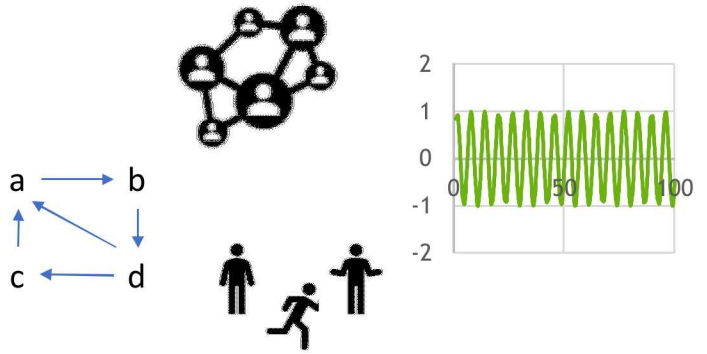
Is the simulation a self sustaining virtual world?



DARPA Ground Truth: Evaluating the TA1 Simulations 2/2

4. **Complexity**
   - Definition: Complexity of the actors, environments, interactions, and outputs of a simulation
     - Two driving questions:
       1. **How hard is the test being posed to the TA2 teams?**
       2. **How representative might the simulation be of the real world?**
   - Evaluation: Later in the presentation…
5. **Flexibility**
   - Definition: Ability to manage and manipulate simulation complexity
   - Evaluation: Fraction of simulation parameters that significantly impact complexity,

How complex is the simulation?



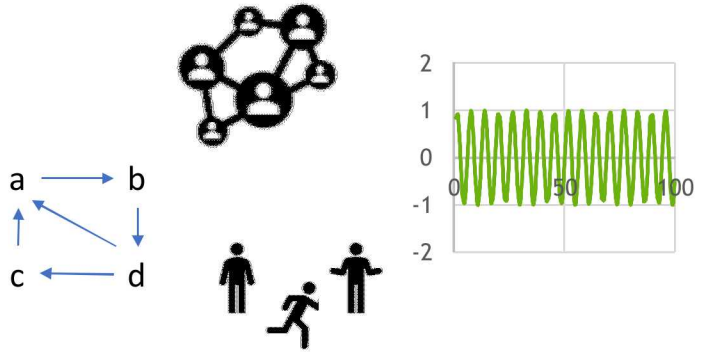
Can the TA1 team manipulate complexity?

DARPA Ground Truth: Evaluating the TA1 Simulations 2/2

4. **Complexity**
   - Definition: Complexity of the actors, environments, interactions, and outputs of a simulation
     - Two driving questions:
       1. **How hard is the test being posed to the TA2 teams?**
       2. **How representative might the simulation be of the real world?**
   - Evaluation: Later in the presentation…

5. **Flexibility**
   - Definition: Ability to manage and manipulate simulation complexity
   - Evaluation: Fraction of simulation parameters that significantly impact complexity,

How complex is the simulation?



Can the TA1 team manipulate complexity?

Complexity & Social Simulation

1. **How hard is the test being posed to the TA2 teams?**
2. **How representative might the simulation be of the real world?**
   - Are the simulations good testbeds (real-world proxies) for testing TA2 research methods?
   - Are the simulations themselves representative of real-world systems?

Challenges

- Many definitions of complexity – how do we capture what is important?
- How to avoid the temptation of focusing on easy measurements (e.g., number of actors represented)?

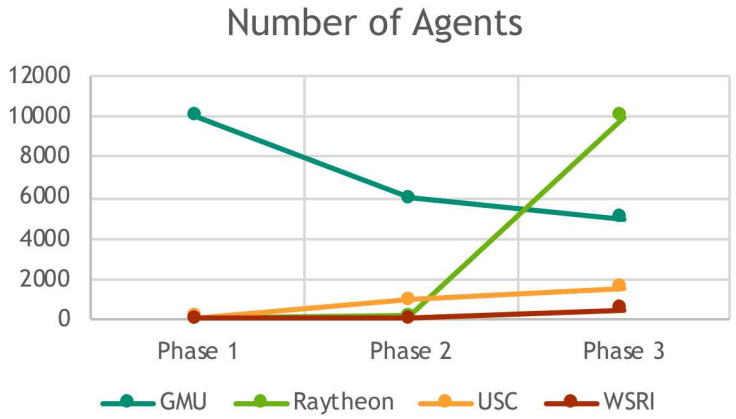
Evolution of Simulation Characteristics

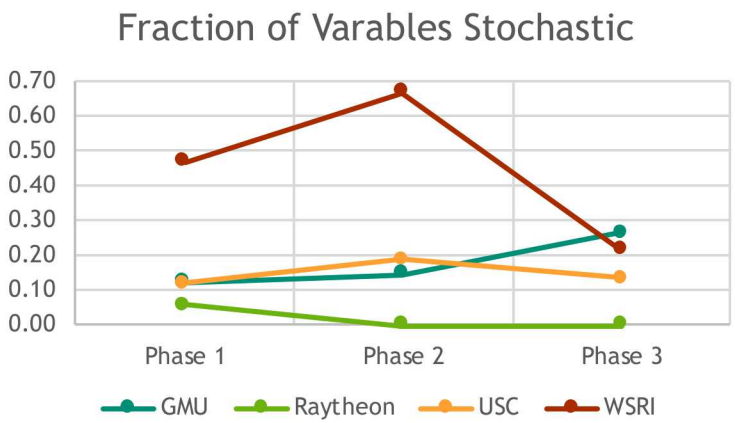
Original plan: Increase complexity over the course of the program

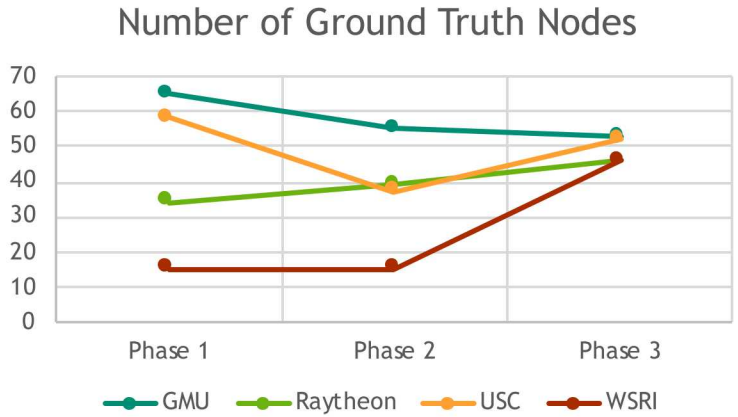
Phase 1: More challenging than expected

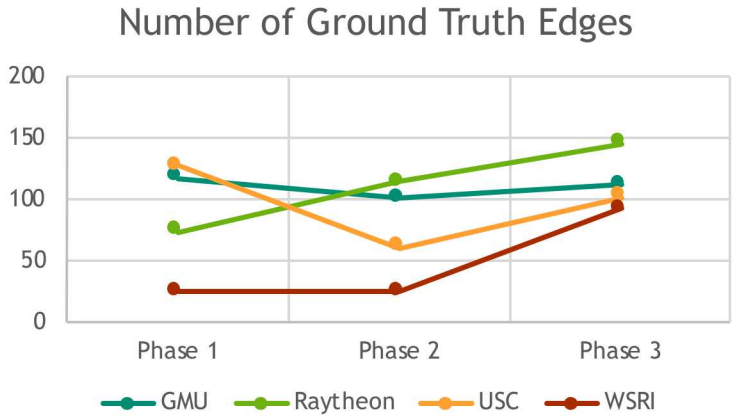
Phase 2: Asked the TA1 teams to keep complexity similar to phase 1

Phase 3: More complex simulations









Organizing Complexity: Two Dimensions

Dimension 1: Tie to social sciences

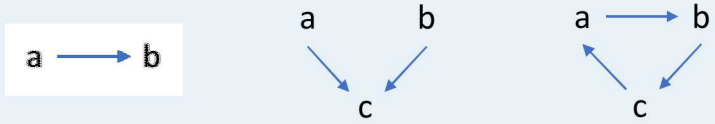
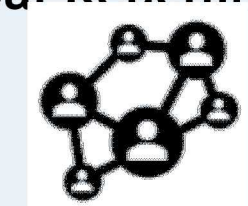
- If metric is inspired by real-world social complexity metrics, there is an obvious tie to rea world systems
- If not, the metric might be more broadly applicable to a variety of application domains

Dimension 2: Knowledge of system's causal structure

- Some dimensions of complexity may be tied to causal structure
- Metrics that don't rely on causal structure might be more broadly applicable
  - For example, to real-world systems

	Not tied to social sciences	Inspired by the social sciences
Requires knowledge of system structure		
Does not require knowledge of system structure		

Ground Truth Complexity Metrics

	Not tied to social sciences	Inspired by the social sciences
Requires knowledge of system structure	**Measures of System Intricacy** *How complicated is the causal structure?* Metric: **Causal Complexity** 	**Behavioral Capacity** *How do interactions and behaviors of actors affect complexity?* Metric: **Number of Differentiated Relationships** 
Does not require knowledge of system structure	**Information-Theoretic Complexity** *What is the information content of the system's behavior?* Metric: **Time-Averaged Normalized Compression Distance** 	**Measures of Social Organization** *How organized are social relationships in the system?* Metric: **Global Reaching Centrality** 

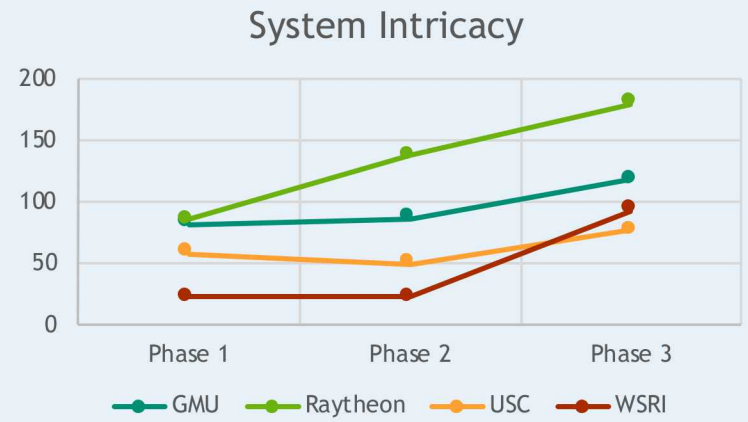
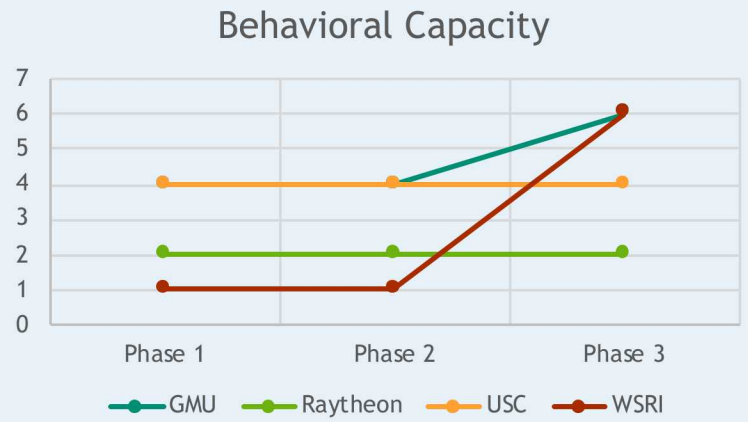
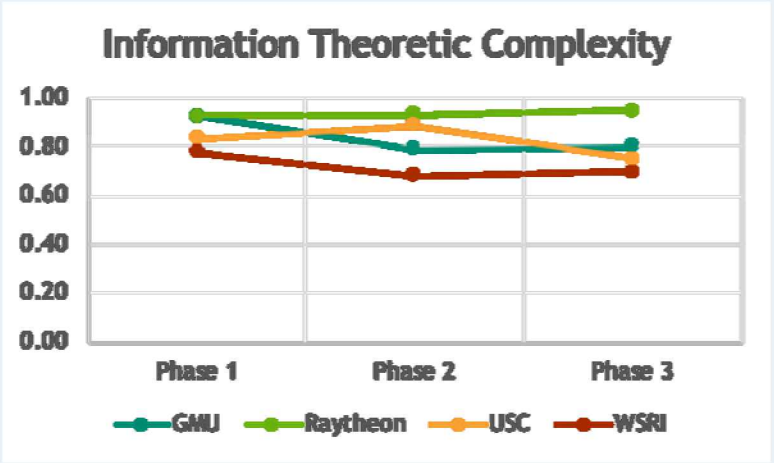
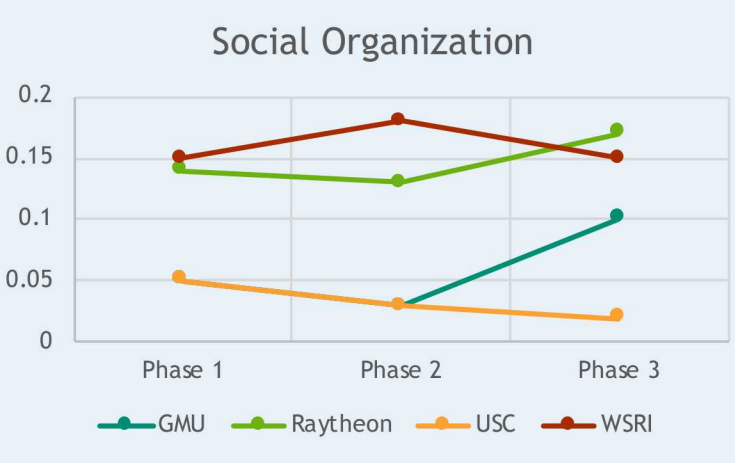
Evolution of Simulation Complexity

Original plan: Increase complexity over the course of the program

Phase 1: More challenging than expected

Phase 2: Asked the TA1 teams to keep complexity similar to phase 1

Phase 3: More complex simulations

	Not tied to social sciences	Inspired by the social sciences
Requires knowledge of system structure		
Does not require knowledge of system structure		

Evaluating the TA2 Research Methods

Explain Tests

TA2 goal: Infer the causal ground truth for each simulation

Evaluation: Compare returned ground truth to actual ground truth

GMU	Raytheon	USC	WSRI

Explain Tests

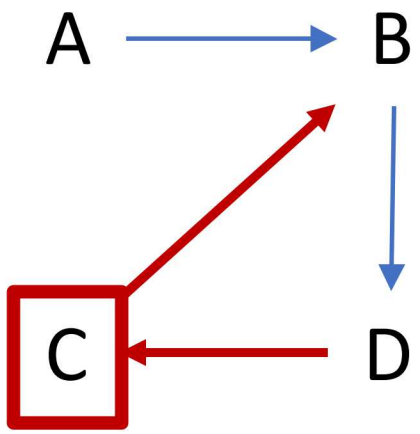
TA2 goal: Infer the causal ground truth for each simulation

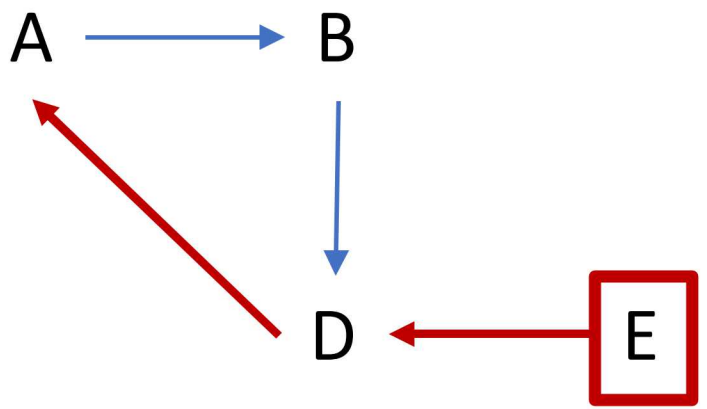
Evaluation: Compare returned ground truth to actual ground truth

Metrics:

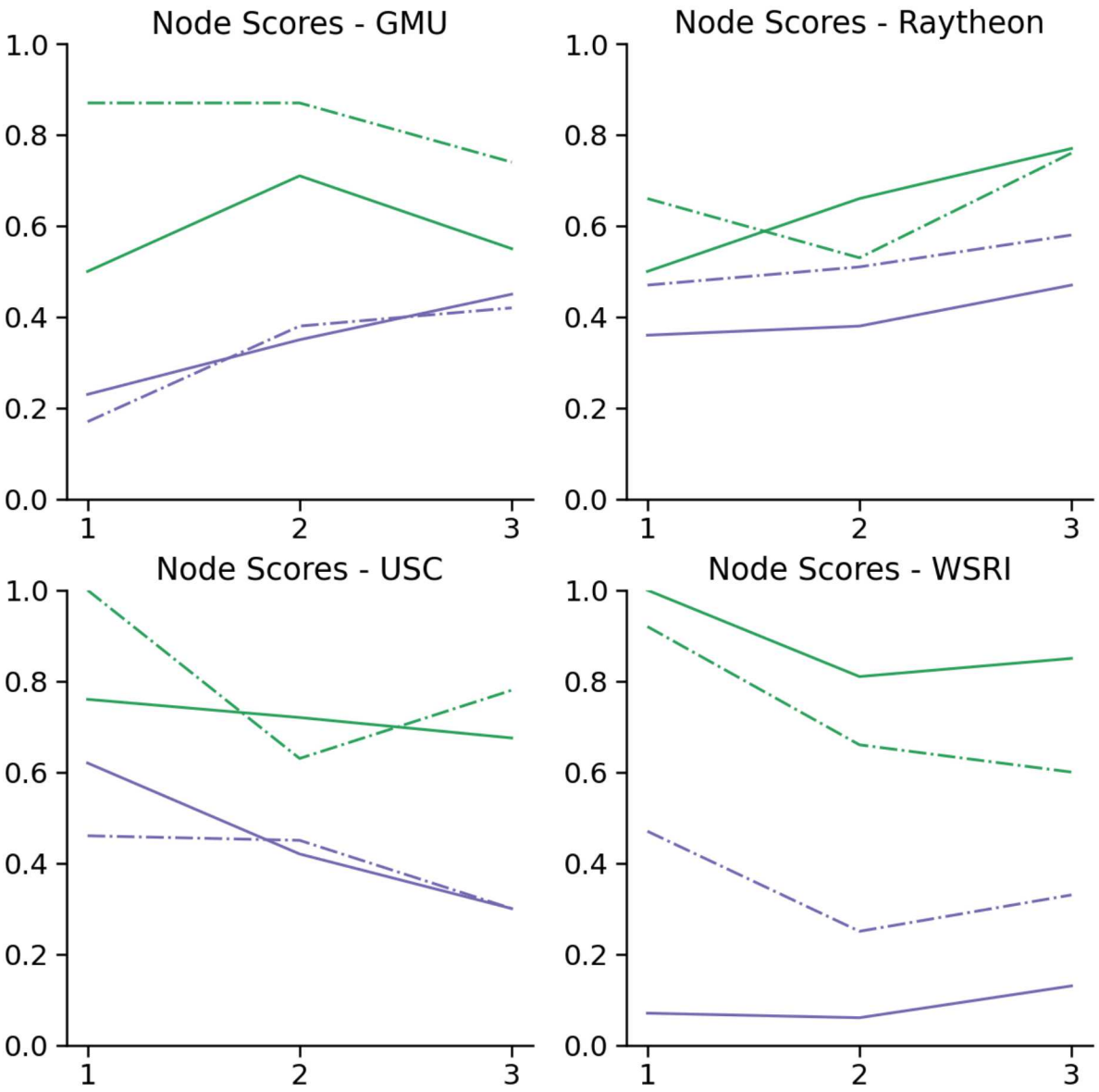
- **Node precision**: how many of the inferred nodes were correct?
- **Node recall**: how many of the actual nodes were inferred?
- **Node F1 score**: combines precision and recall
- **Edge precision**: how many of the inferred edges were correct?
- **Edge recall**: how many of the actual edges were inferred?
- **Edge F1 score**: combines precision and recall

- Partial credit for causal paths





Explain Test Results



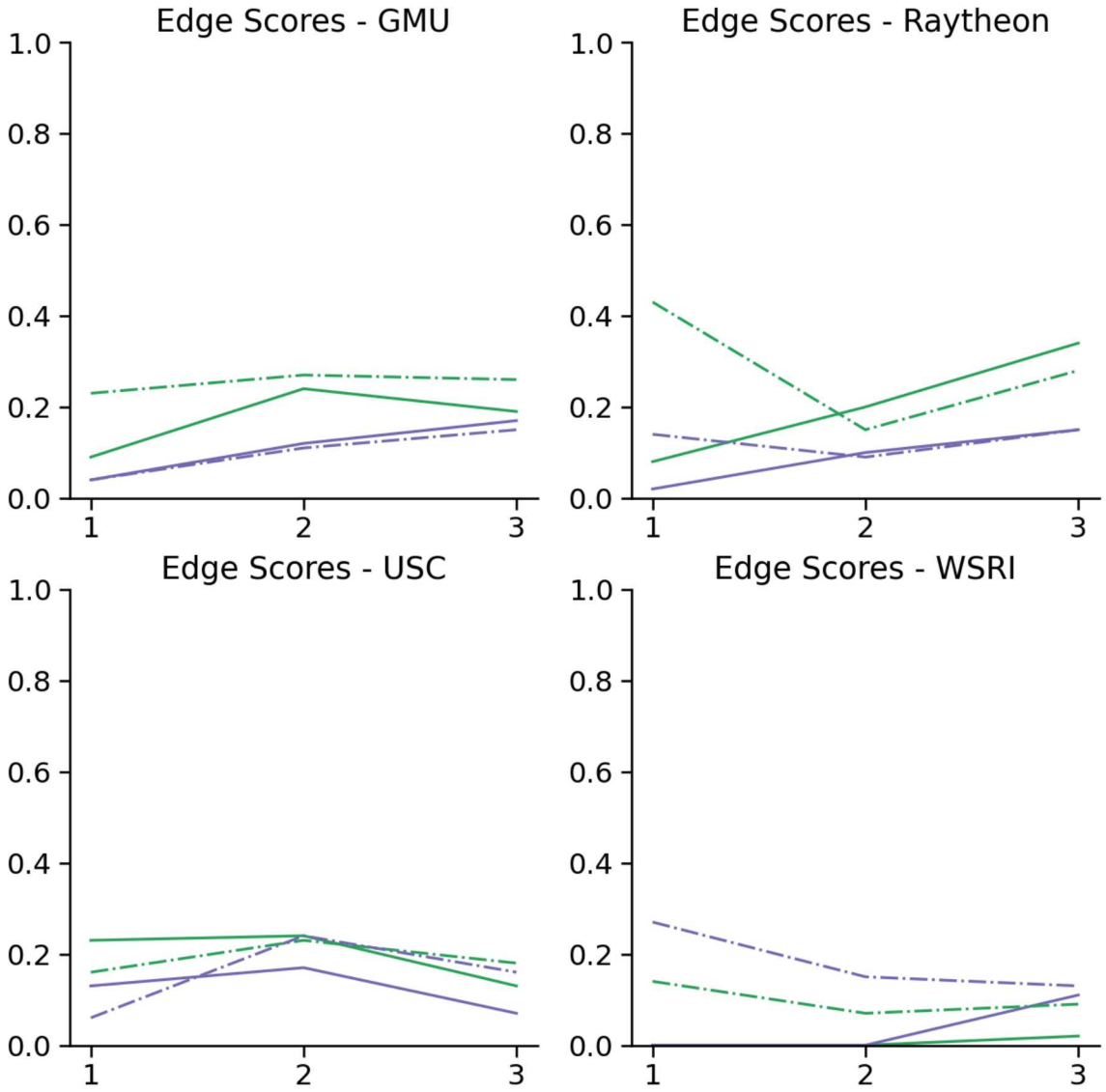
Node scores indicate how successful the TA2 teams were at finding *which variables exist* in the simulations

- Precision: fraction of inferred nodes true
- Recall: fraction of true nodes inferred

Results

- Varied by simulation
- Generally high precision (most inferred nodes were in the simulation ground truth)
- Lower recall (some true nodes were not discovered), usually in middle of range

Explain Test Results



Edge scores indicate how successful the TA2 teams were at finding *causal relationships*

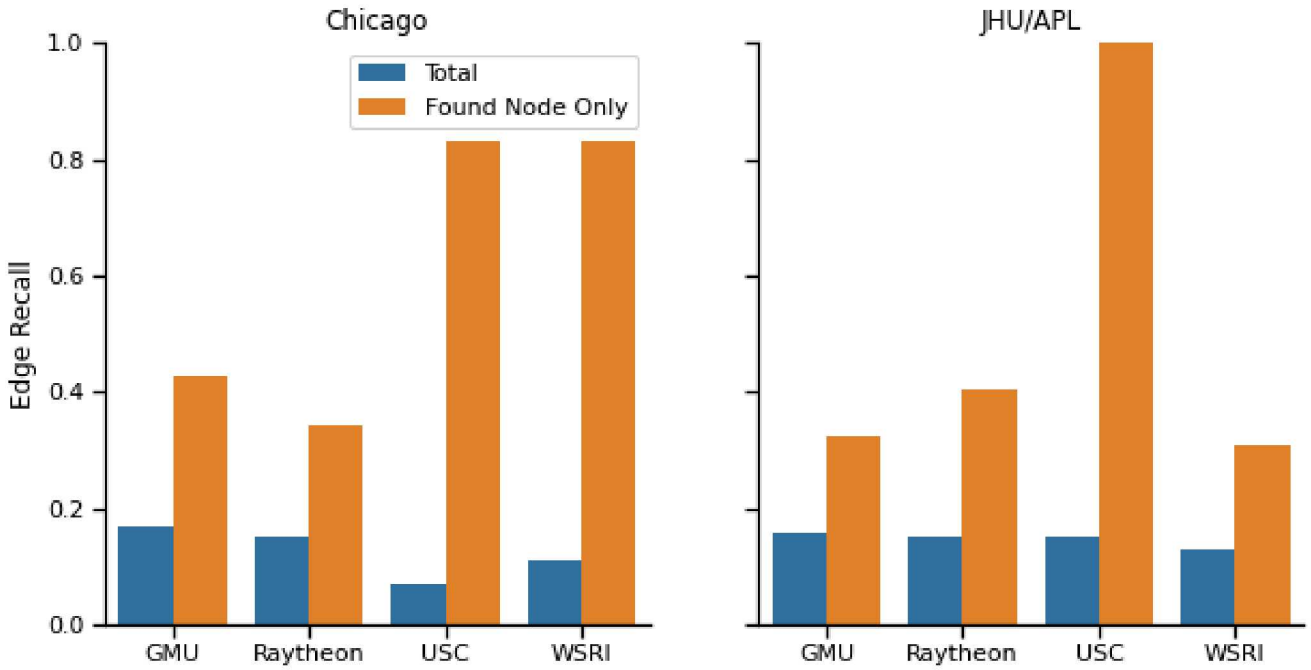
- Precision: fraction of inferred nodes true
- Recall: fraction of true nodes inferred

Results

- Much lower than node scores
- Not independent of node score calculations
- Causality is tough to infer

Explain Test Results

What if we only look at edges connected to nodes found by the TA2 teams?



Edge scores indicate how successful the TA2 teams were at finding *causal relationships*

- Precision: fraction of inferred nodes true
- Recall: fraction of true nodes inferred

Results

- Much lower than node scores
- **Not independent of node score calculations**
- Causality is tough to infer

Predict Tests

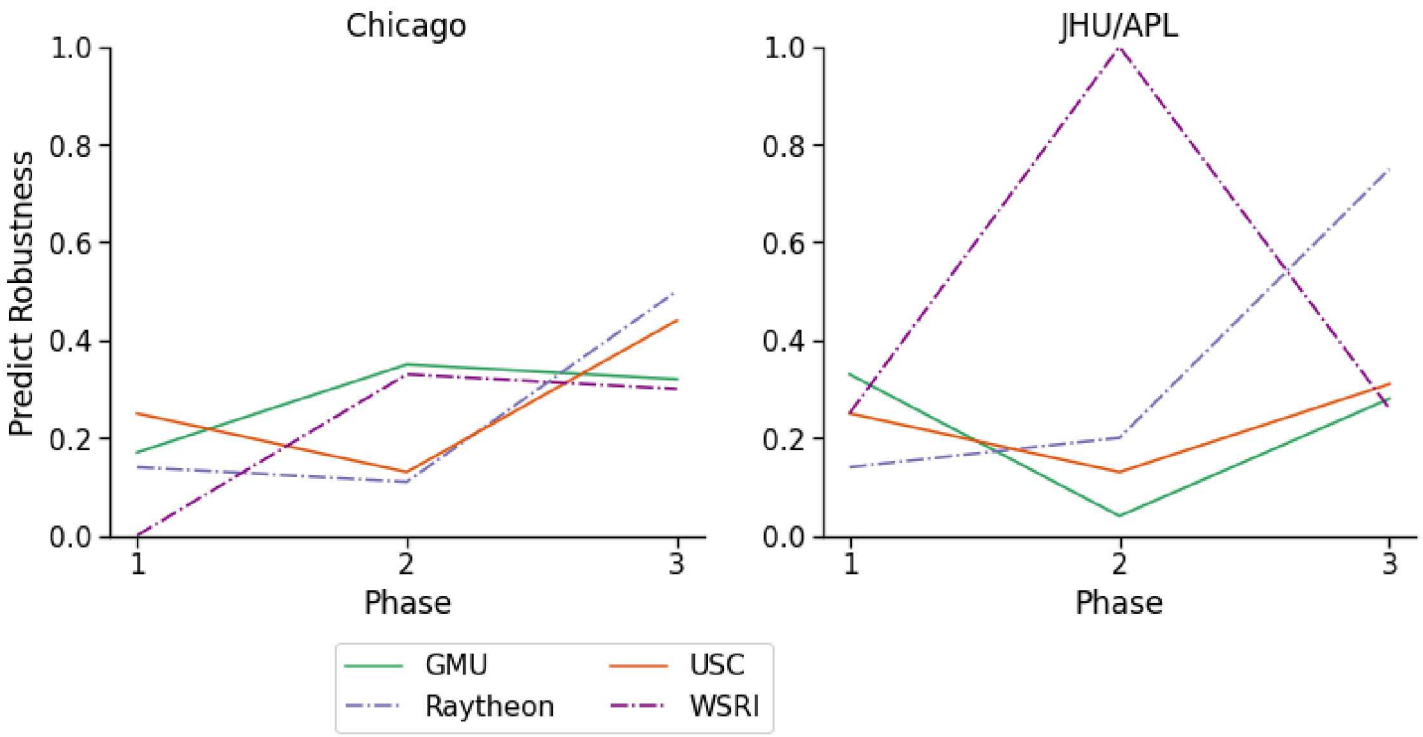
TA2 goal: Predict the behavior of the simulation, based on specific questions posed by the TA1 teams

Evaluation metrics based on questions: differences in values, means, variances, etc.

Sample Predict Questions

- How many individuals will be infected by a disease?
- Which individuals will be infected at a set of specified times?
- Which group will have the highest average happiness?
- If the richest 10% of individuals lose $500, how does popularity change
- Will a certain individual survive the hurricane?
- If certain geographic areas are off-limits, how does average satisfaction change?

Predict Test Results



Robustness: fraction of answers in "acceptable" range as defined by TA1 teams

Further analysis:

How do types of predict questions affect performance?

- Individual versus aggregate focus
- How far out in time the prediction is
- Counterfactuals versus steady state predictions

Prescribe Tests

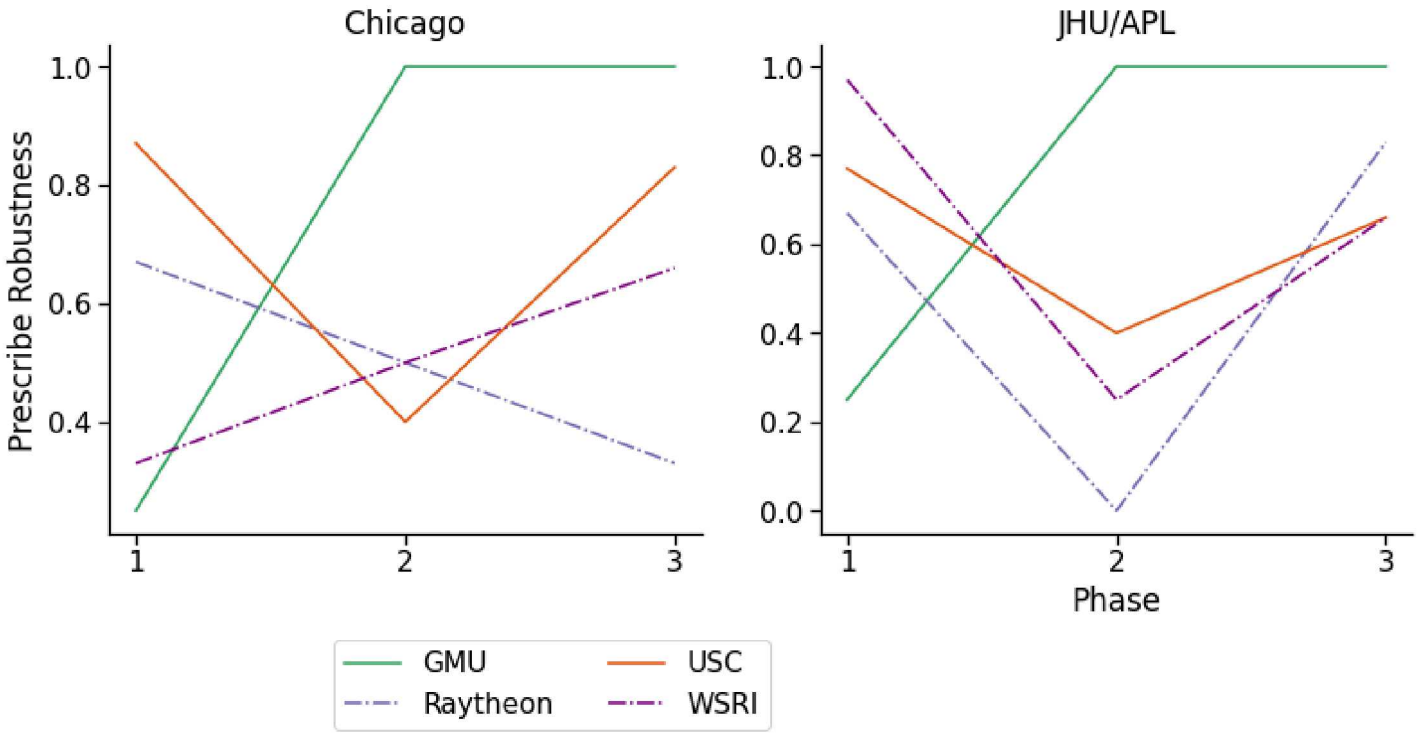
TA2 goal: Prescribe actions to achieve specified goals

Evaluation metrics based on questions: differences in values, means, variances, etc.

Sample Prescribe Tasks

- Minimize reported infections of disease
- Which individuals should groups recruit to maximize the number of contests won?
- Direct government aid by region to minimize hurricane casualties
- Minimize visits to a geographic area by individuals of a certain type

Prescribe Test Results



Robustness: Fraction of prescriptions that moved the key outputs in the desired direction

Further analysis:

How do types of prescribe questions affect performance?

- Individual versus aggregate focus
- How far out in time the prediction is
- Counterfactuals versus steady state predictions

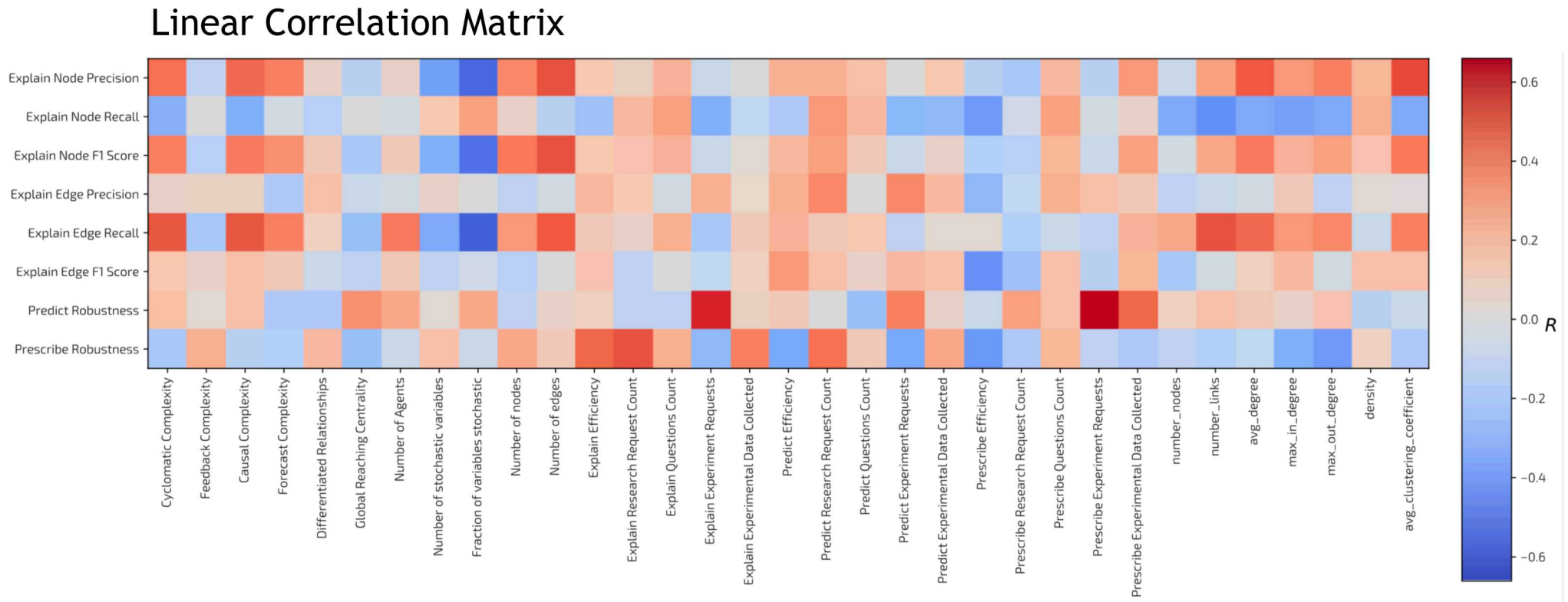
Comparative Analysis (In Progress!)

Comparative Analysis

Collected data over
- Three phases
- Four TA1 simulations
- Two TA2 teams

Allows us to study the overall program

Program design allows high-level comparative analysis
- How did simulation complexity affect explain/predict/prescribe performance?
- Did causal knowledge help with prediction and prescription?
- Did more data lead to better explain/predict/prescribe performance?



Note on Comparative Statistics

Small sample size

Four very different TA1 simulations

Two very different TA2 approaches

Lots of noise

Low significance to statistical relationships

Trend lines shown for description, do not indicate statistical significance

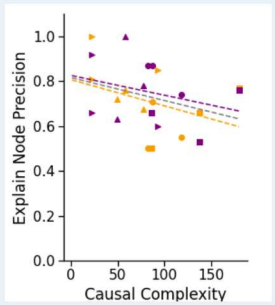
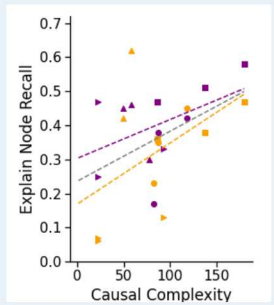
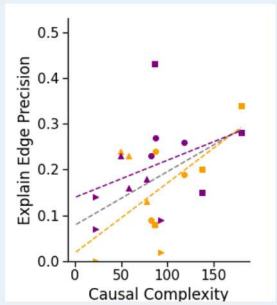
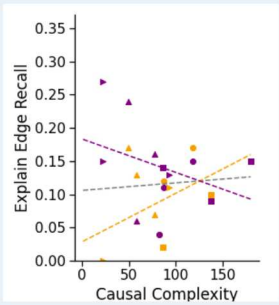
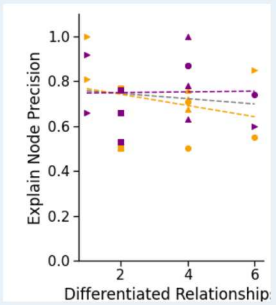
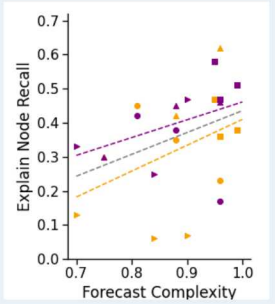
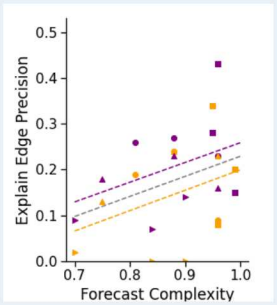
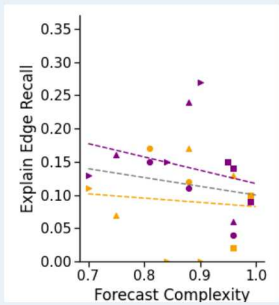
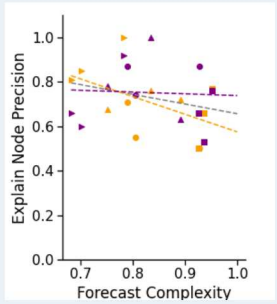
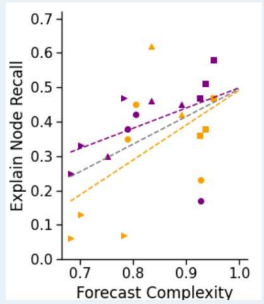
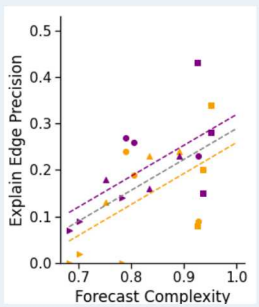
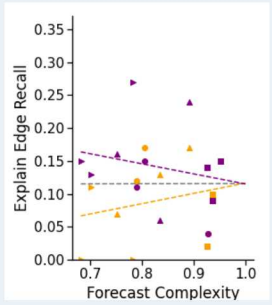
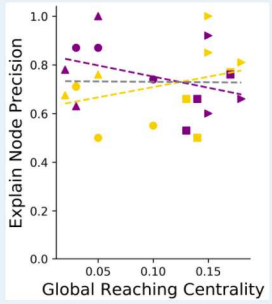
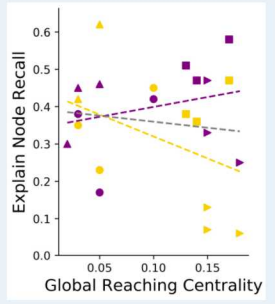
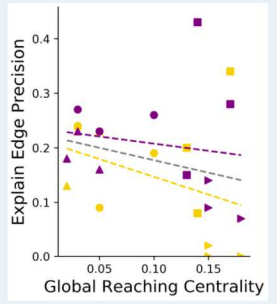
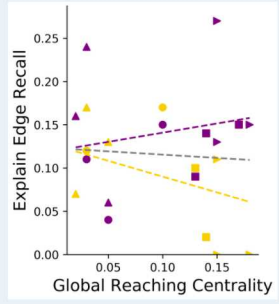
Still interpreting some results

- WSRI predict and prescribe test results may need adjusting (still included here)

How did simulation complexity affect explain/predict/prescribe performance?

Hypothesis: More complex simulations will be more difficult to explain/predict/prescribe

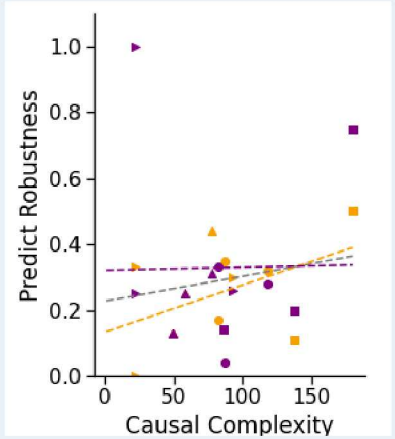
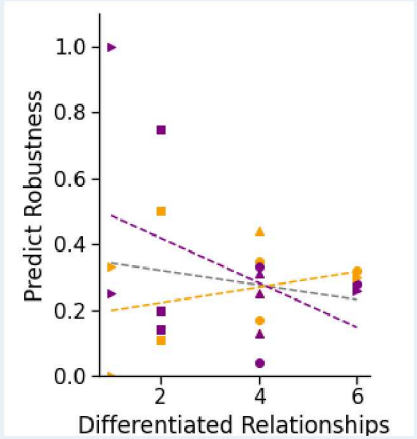
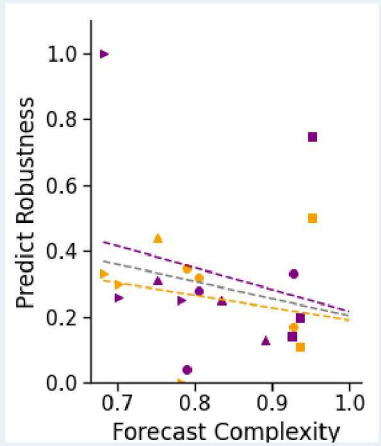
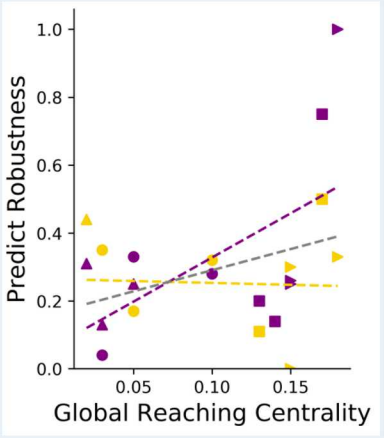
Is it harder to infer causality when the system is more complex?

	Not tied to social sciences	Inspired by the social sciences
Requires knowledge of system structure	   	   
Does not require knowledge of system structure	   	   

Did **simulation complexity** correlate with TA2 **explain test performance**?

- Mixed trends
- More complex simulations also came later in program, when TA2 teams had more practice
- Future analysis: breaking down performance by phase, model sector

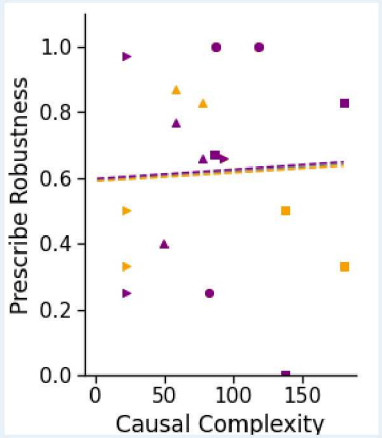
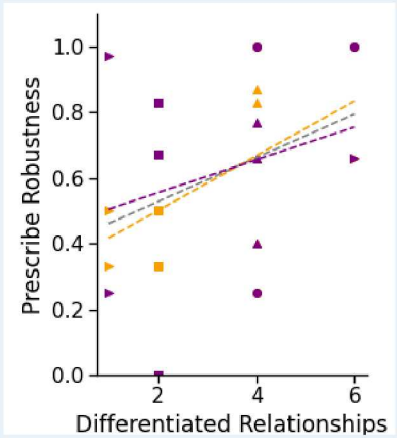
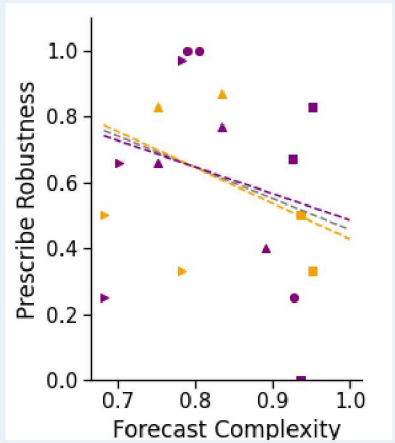
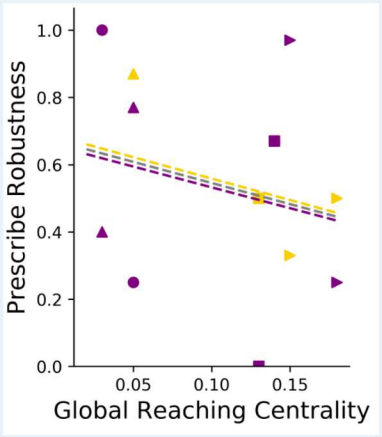
Is it harder to predict when the system is more complex?

	Not tied to social sciences	Inspired by the social sciences
Requires knowledge of system structure		
Does not require knowledge of system structure		

Did **simulation complexity** correlate with TA2 **predict performance**?

Mixed trends

Is it harder to prescribe when the system is more complex?

	Not tied to social sciences	Inspired by the social sciences
Requires knowledge of system structure		
Does not require knowledge of system structure		

Did **simulation complexity** correlate with TA2 **prescribe performance**?

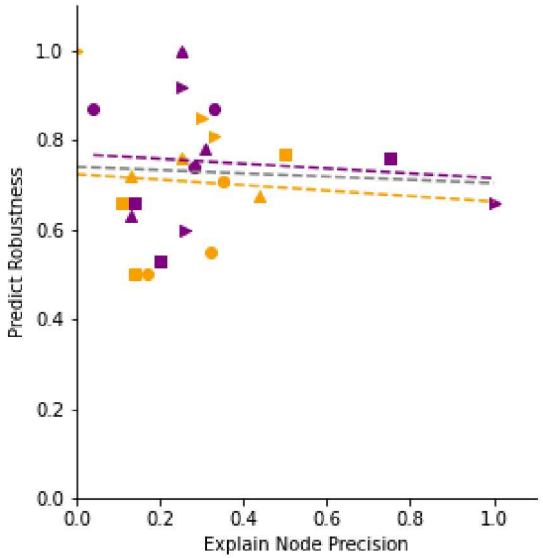
Some stronger trends

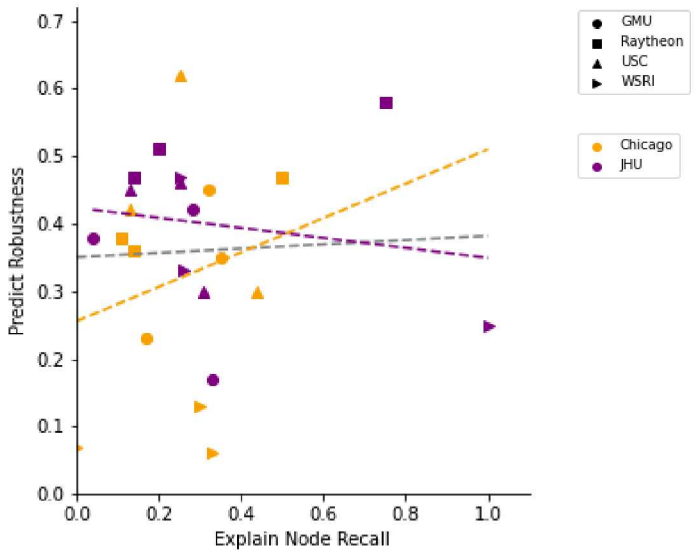
- Positive for system structure complexity metrics
- Negative for system output complexity metrics

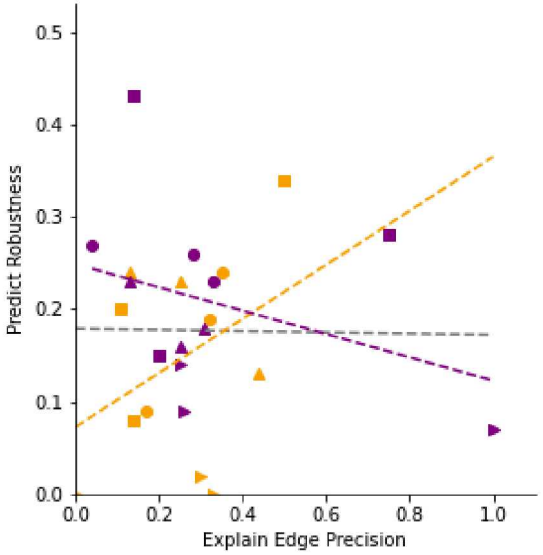
Did causal knowledge help with prediction and prescription?

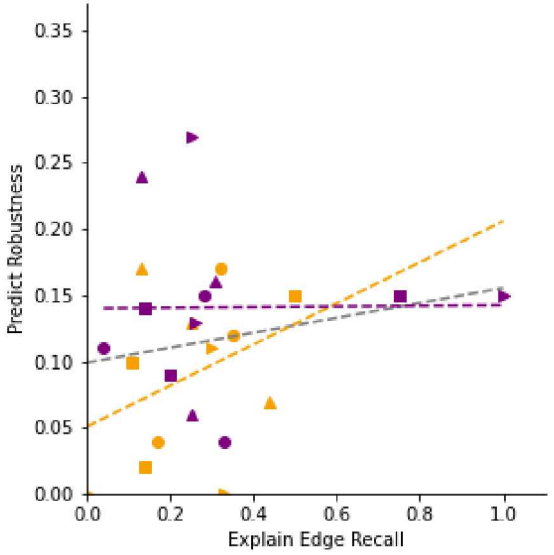
Hypothesis: Better performance on the explain test will lead to better performance on the predict and prescribe tests

Did causal knowledge help with prediction?









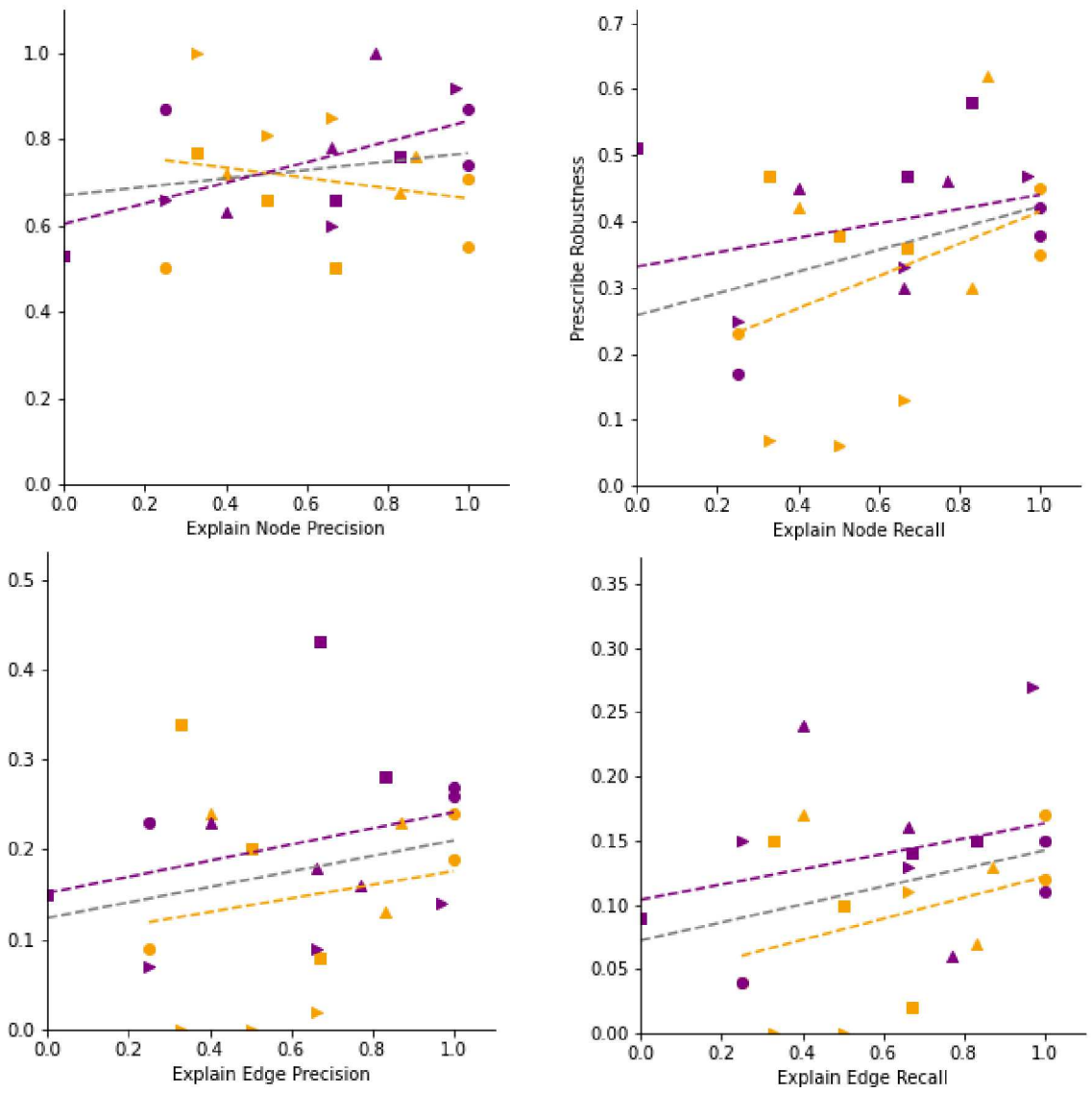
Comparing **explain test performance** to **predict test performance**

No clear pattern

Possible meanings

- May indicate that causal understanding is not helpful
- May indicate that the TA2 teams did not utilize causal understandings

Did causal knowledge help with prescription?



Comparing **explain test performance** to **prescribe test performance**

Mild positive trend in some cases

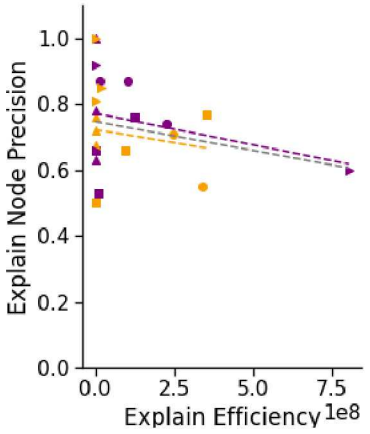
May indicate utility

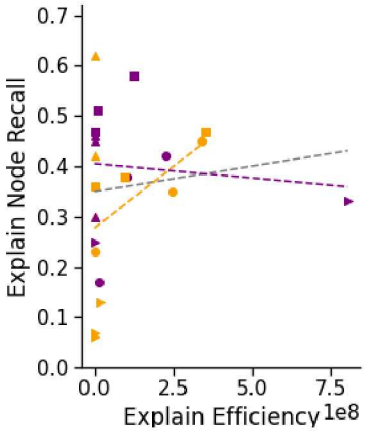
Also depends on TA2 use of causality in determining prescriptions

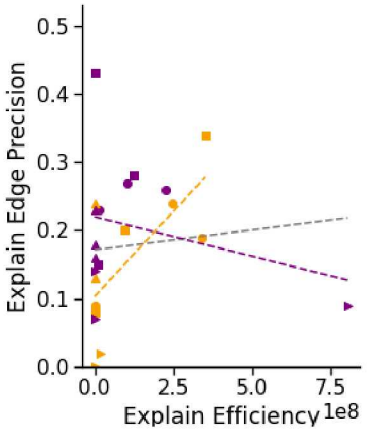
Did more data lead to better explain/predict/prescribe performance?

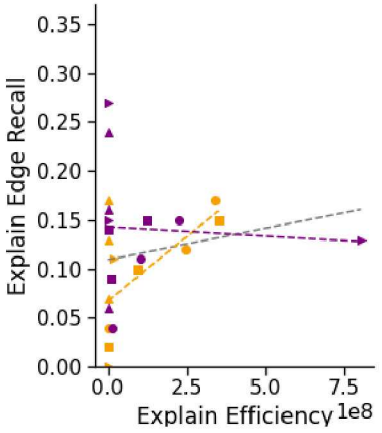
Hypothesis: More data will lead to better performance

Did more data lead to better causal explanations?





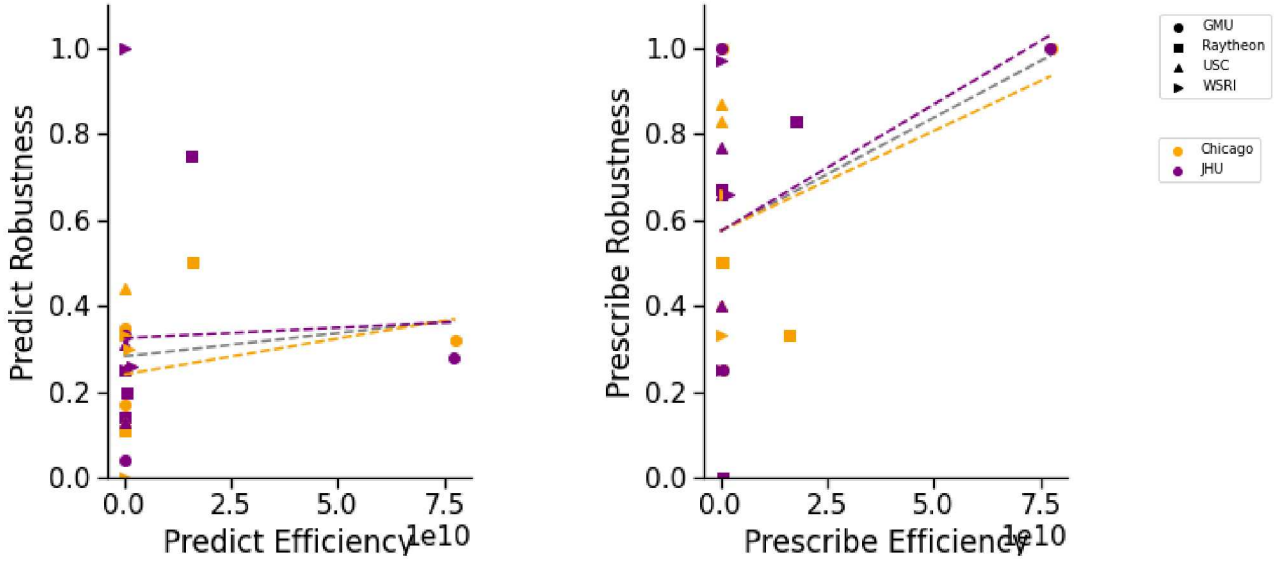




How did **data collection** correlate with **explain test performance**?

Unclear trends

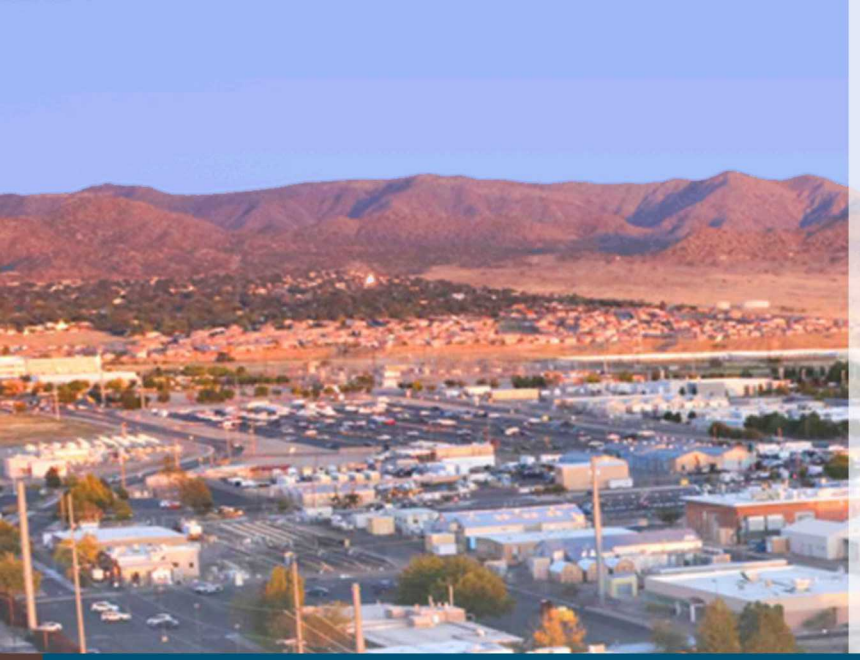
Did more data lead to better predictions and prescriptions?



How did data collection correlate with performance?

Very mild trend for predict

Stronger trend for prescribe



Discussion

Lessons Learned

Pro and con of the program design: *Everyone* was pushed out of their comfort zones

- Different fields have assumptions and habits that are hard to break
  - More realistic test of methods
- TA1 teams
  - All included agent based modelers
  - Had to think of the models as "worlds"
- TA2 teams
  - Had little control over program/test design
  - Asked to do end-to-end research (question development, data collection, analysis)
  - But had more control over what methods they used
- Key lessons
  - Need to put effort into bridging fields

Lessons Learned

Program concept: emulate real-world social science

- Made the TA2 task very difficult
- Phase 1 was challenging
  - Orienting the TA2 teams in the virtual worlds
  - Determining what questions to ask
  - Data collection
- Program design worked, but problem was complex
  - Less comparability than T&E would have liked
  - More realistic

Lessons Learned

Firewalling TA1 and TA2 teams took substantial effort

- Other programs might consider automating more

It's hard to ask the right questions in a domain you don't know well

- Initial datasets were key
- This is one of the things that makes social science hard!

Measuring complexity is difficult

- Many definitions of complexity
- No silver bullet metric
- Especially options that apply to both simulations and the real world

Suggestions for Future Work

Simulations as test beds with a more rigorously controlled research design (less emphasis on emulating the real world)

- Controlled progression of simulations from very simple (3 ground truth nodes) to more complex
- Controlled application of multiple methods, with the same methods applied over all simulations

Exploration of the real-world utility of simulations as test beds

- If methods are tested on simulations, does the measured utility hold up in real-world application?
- Are some real-world systems more amenable (based on complexity, etc.)?

Metascience test bed

- This program required end-to-end social science (data collection, question formulation, analysis, …), and would make a great meta-science test bed

Thank you!

Contact Information:

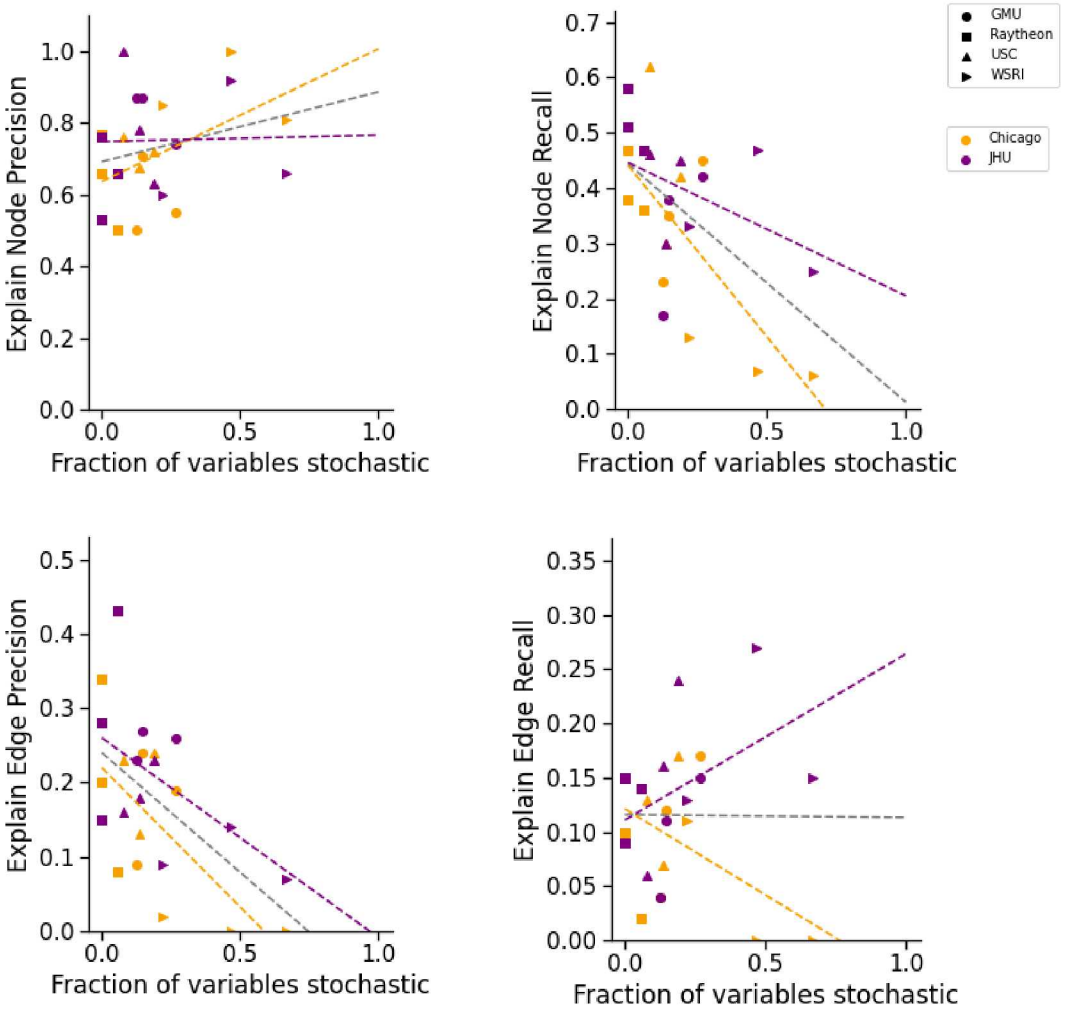
Asmeret Naugle
abier@sandia.gov
(505) 263-1277

Back Up Slides

How did simulation stochasticity affect explain/predict/ prescribe performance?

Hypothesis: Simulations with more stochasticity will be more difficult to explain/predict/prescribe

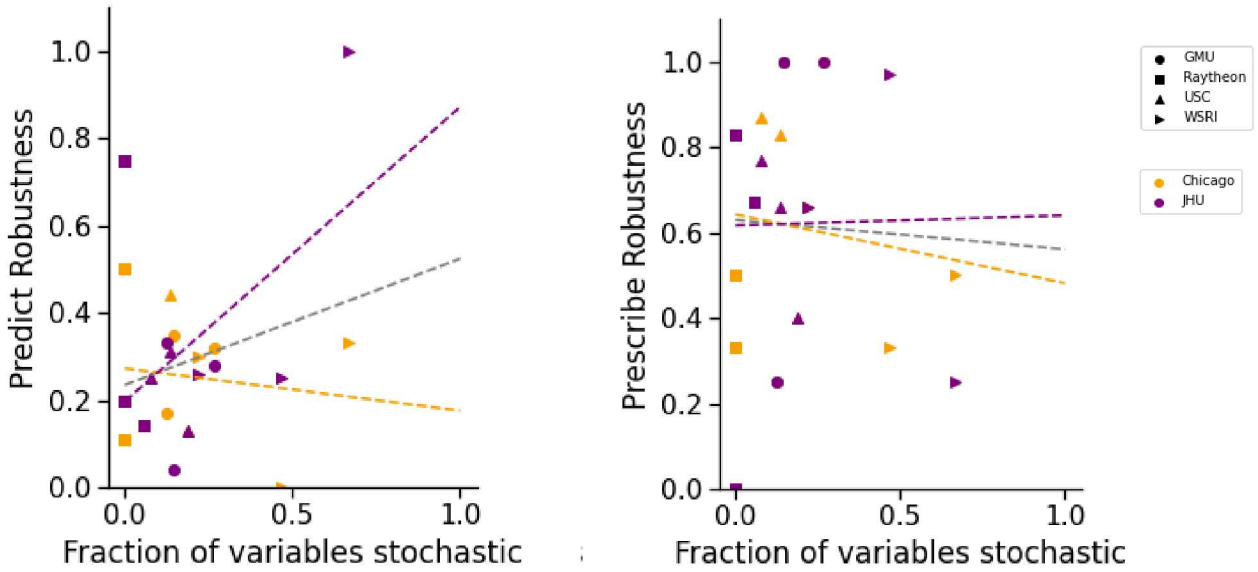
Does stochasticity make causal inference more difficult?



How did **the number of stochastic variables** correlate with **explain test performance**?

Strong negative trend with node recall and edge precision

Does stochasticity make prediction and prescription more difficult?



Comparing **fraction of variables stochastic** with **predict and prescribe robustness**

Mixed trends