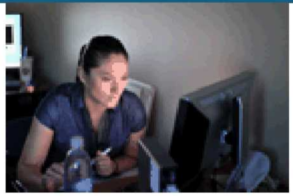
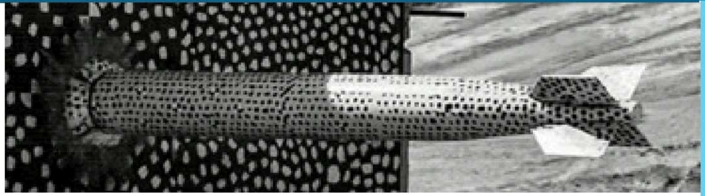
SAND2020-8034PE

DARPA Ground Truth Program Overview





Asmeret Naugle, Test & Evaluation Team

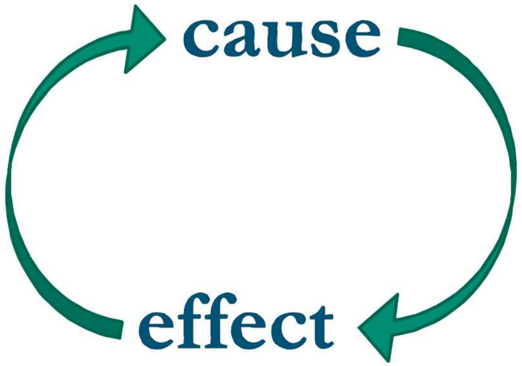
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

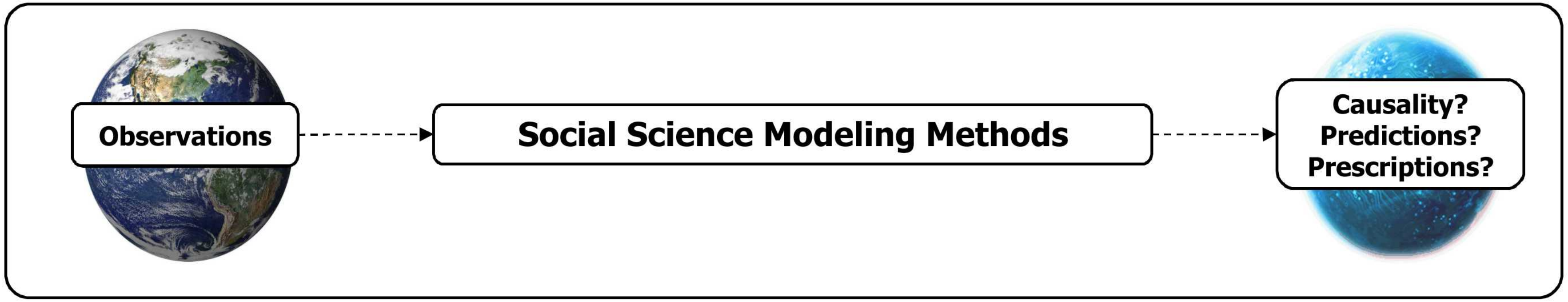
DARPA Ground Truth: Motivation

To improve the way we understand and influence the world

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




Observations

Social Science Modeling Methods

Causality?
Predictions?
Prescriptions?

DARPA Ground Truth: Program Overview

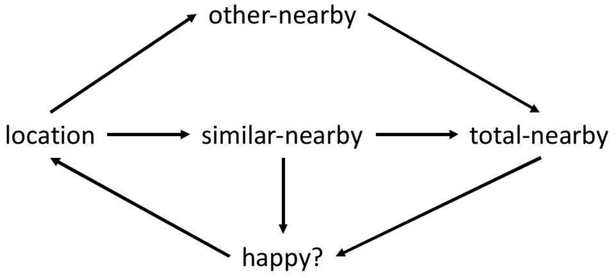
Program goal:

Use artificial but socially-plausible simulations with known causal rules (aka “ground truth”) as testbeds to validate social science modeling methods

Three tests in each of three phases (with increasing complexity):

Explain:

Infer the causality of the simulation



other-nearby

location

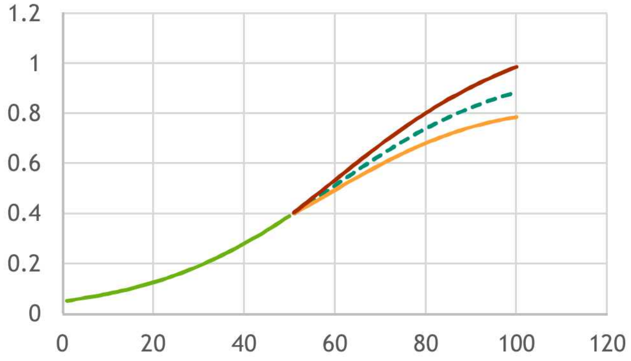
similar-nearby

total-nearby

happy?

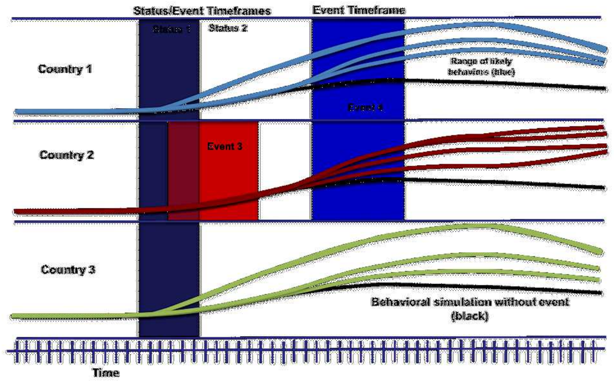
Predict:

Predict what will happen

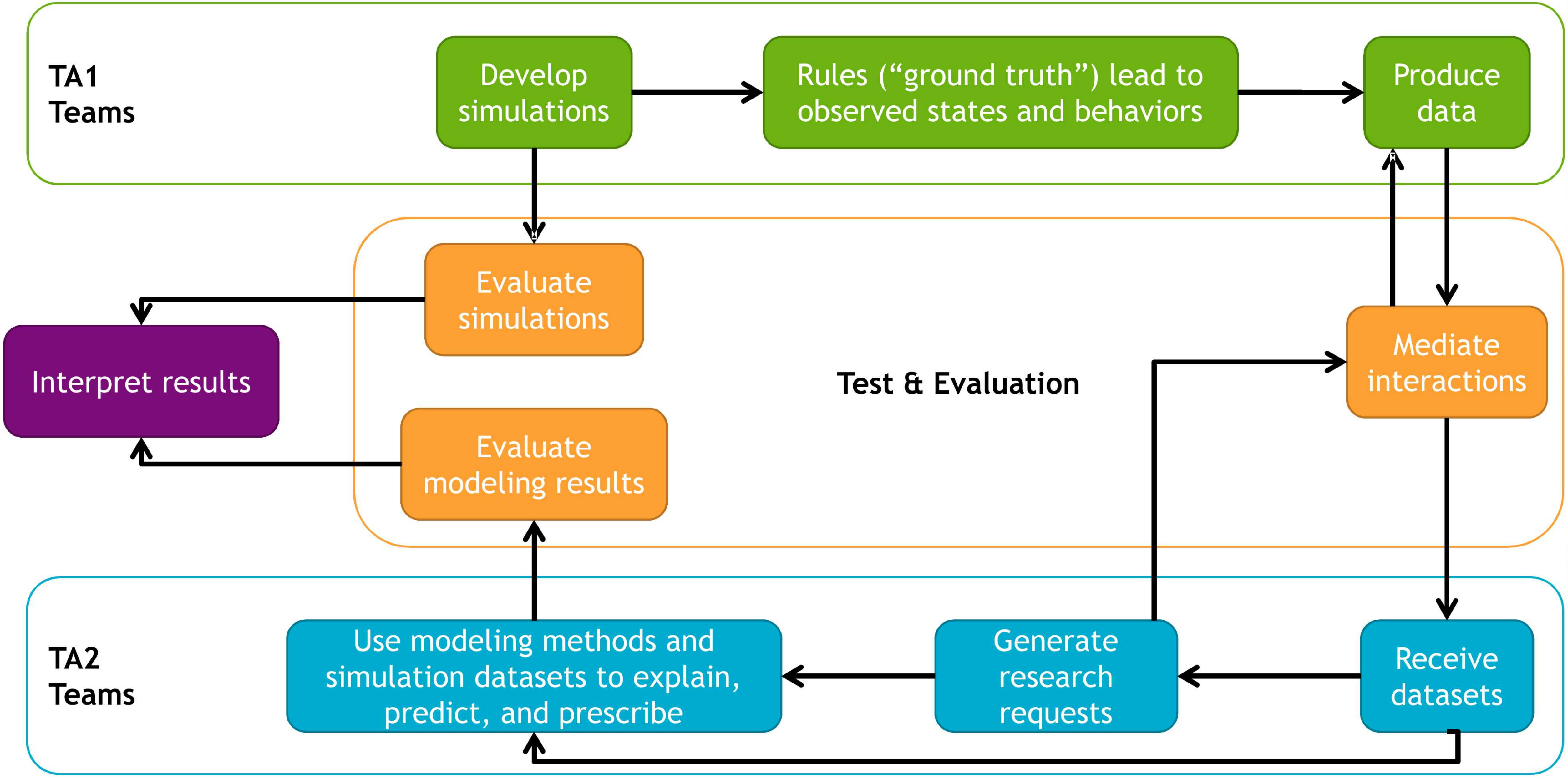


Prescribe:

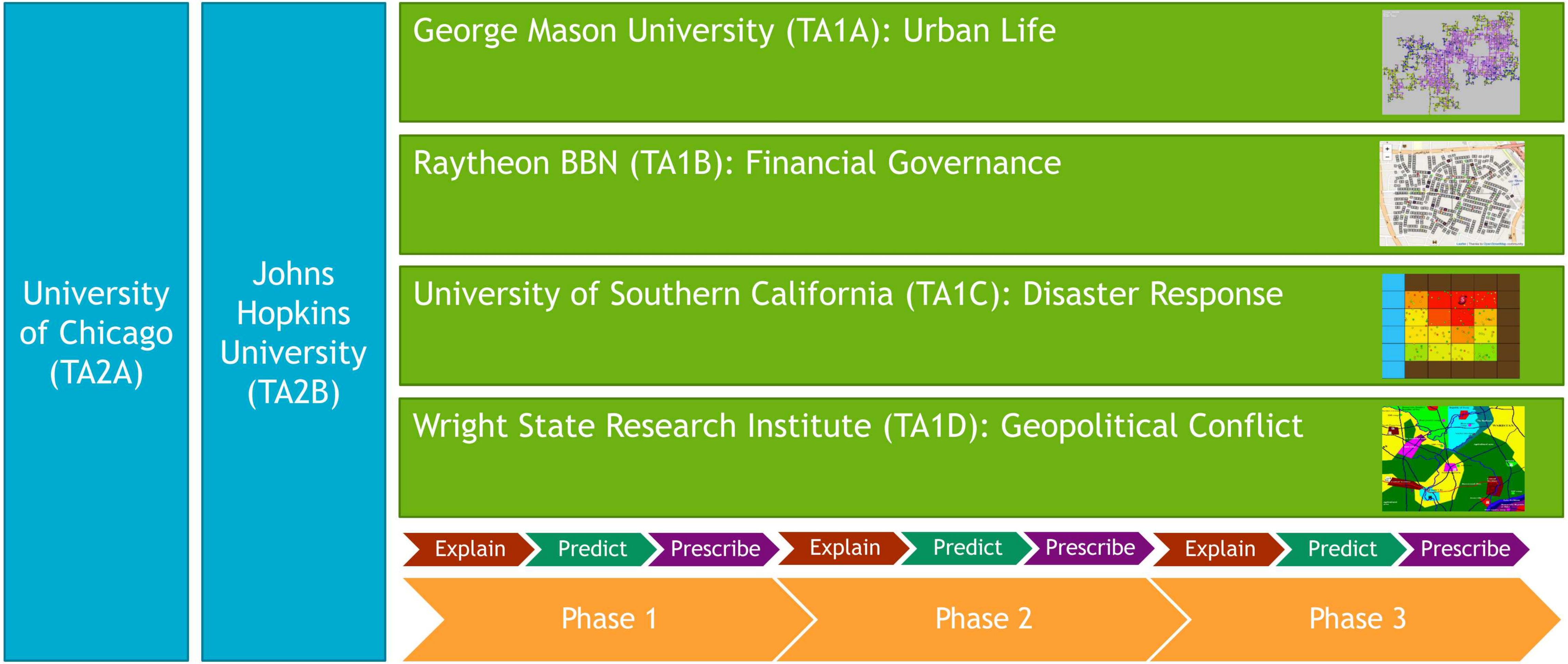
Prescribe actions to achieve goals



DARPA Ground Truth: Program Design


TA1
Teams
Develop simulations
Rules ("ground truth") lead to observed states and behaviors
Produce data
Evaluate simulations
Interpret results
Test & Evaluation
Mediate interactions
Evaluate modeling results
TA2
Teams
Use modeling methods and simulation datasets to explain, predict, and prescribe
Generate research requests
Receive datasets

Performer Teams and Timeline



University of Chicago (TA2A)

Johns Hopkins University (TA2B)

George Mason University (TA1A): Urban Life

Raytheon BBN (TA1B): Financial Governance

University of Southern California (TA1C): Disaster Response

Wright State Research Institute (TA1D): Geopolitical Conflict

Explain → Predict → Prescribe → Explain → Predict → Prescribe → Explain → Predict → Prescribe

Phase 1 → Phase 2 → Phase 3

DARPA Ground Truth: TA1 Simulation Requirements

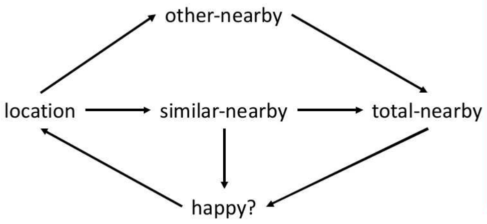
1. **Simulation accessibility:** Can the simulations handle social science data collection methods?

Data Collection Methods

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

2. **Verifiability of ground truth:** Does the ground truth accurately represent the simulation?

Ground Truth Represents Causal Structure

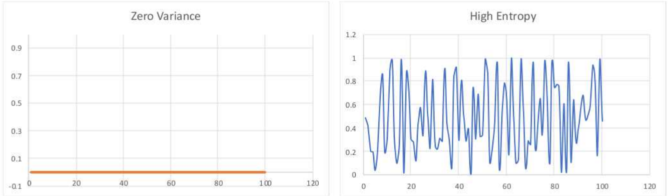


```
if all turtles are happy then stop
 for each turtle
    if unhappy, randomly move to new unoccupied patch
    similar-nearby count =
      number of neighbors with color = turtle's color
    other-nearby count =
      number of neighbors with color != turtle's color
    total-nearby = similar-nearby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

adapted from Wilensky (1997)

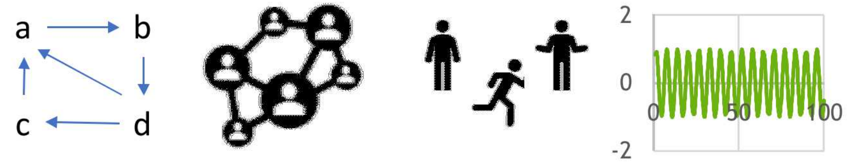
3. **Plausibility:** Is the simulation a self sustaining virtual world?

Simulation-Driven "Interesting" Behavior



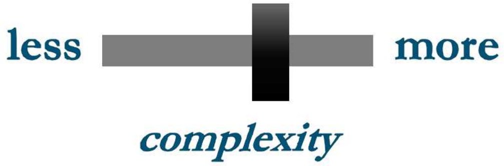
4. **Complexity:** How complex is the simulation?

Multiple Dimensions of Complexity



5. **Flexibility:** Can the TA1 team manipulate complexity?

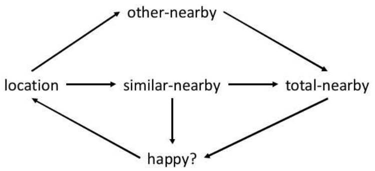
Ground Truth Represents Causal Structure

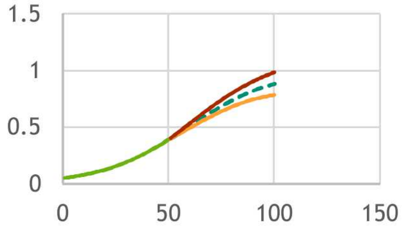


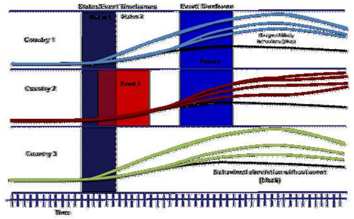
DARPA Ground Truth: Evaluating the TA2 Research Methods

1. **Accuracy**
   1. **Explain** test
      - Definition: Ability to infer the causal processes that serve as ground truth for the simulations
      - Evaluation: Compare returned ground truth to actual ground truth
   2. **Predict** test
      - Definition: Similarity between prediction and simulated outcome for a specific scenario
      - Evaluation: Differences in values, means, variances…
   3. **Prescribe** test
      - Definition: Performance in prescribing simulation settings that result in the simulation attaining some desired state
      - Evaluation: Percentage of distance between baseline and target achieved by prescription
2. **Robustness**
   - Definition: How well a TA2 method performs over a range of applications of the method
   - Evaluation: Average accuracy across simulations
3. **Efficiency**
   - Definition: How much data is required to apply methods
   - Evaluation: Data delivered from simulation

How well did the methods explain/predict/prescribe?







How well did the methods do over a range of tests?

How much data did the methods require?

Program Evolution

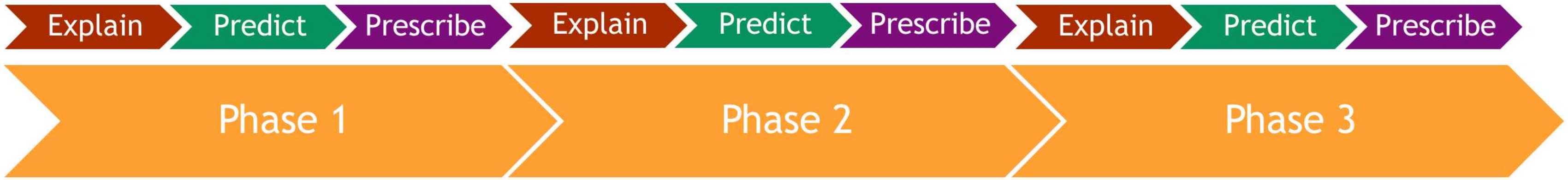
Original plan: Use the simulations as "realistic" proxies for real world systems, with complexity increasing over the course of the program

Phase 1: Tried to emulate real social science research as much as possible

- Included limits on data collection
- Caused substantial frustration
- TA2 accuracy wasn't as high as we had hoped

Phase 2: Kept simulation ground truth almost identical to phase 1, increased data availability substantially

Phase 3: More complex simulations, high data availability, full data for predict & prescribe



What Did We Hope To Learn?

Are simulations useful as test beds for social science research?

What TA2 research methods are most effective?

- How effective are they?

What Did We *Actually* Learn (At Least Partially)?

Are simulations useful as test beds for social science research?

What characteristics of simulations make them better/worse for this purpose?

What difficulties in communication between fields might affect a program like this?

What TA2 research methods are most effective?

What is the accuracy and robustness of the research methods?

What are some of the limitations of the tested research methods?

How might data collection and analysis be integrated to improve social science research?

How does system complexity affect the ability to explain, predict, and prescribe?

What are the limitations of existing complexity metrics?