

The Need for a Social Science Ontology to Support Data-Driven Social Simulations

Catriona Kennedy, Georgios Theodoropolous

School of Computer Science,
The University of Birmingham, UK

<http://www.cs.bham.ac.uk/~cmk/>

(in collaboration with Peter Lee, Ed Ferrari and Chris Skelcher,

School of Public Policy, University of Birmingham

<http://www.publicpolicy.bham.ac.uk/>).

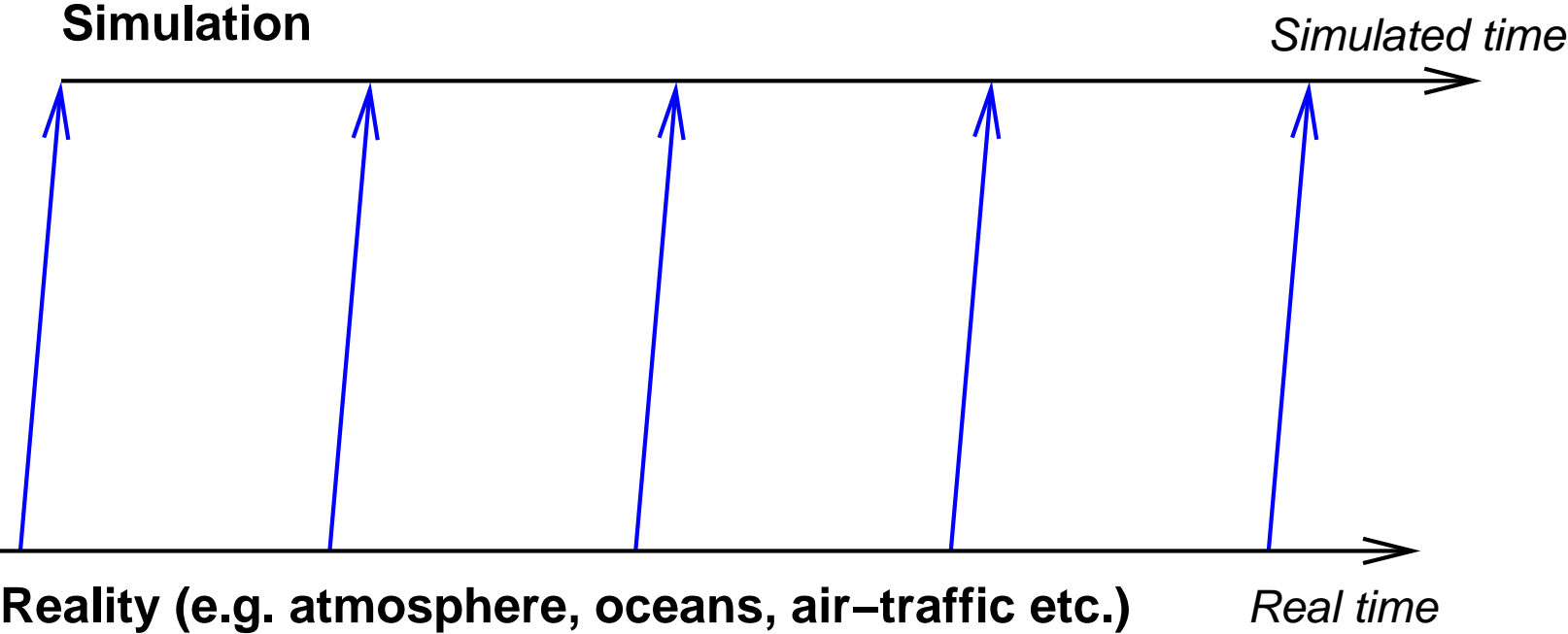
Overview

- **AIMSS project: background**
- **Data Driven Simulation: background**
- **Applying Data Driven Simulation to Social Science**
- **Housing case study**
- **Requirement for domain ontology**

Adaptive Intelligent Model-Building for the Social Sciences (AIMSS): Background

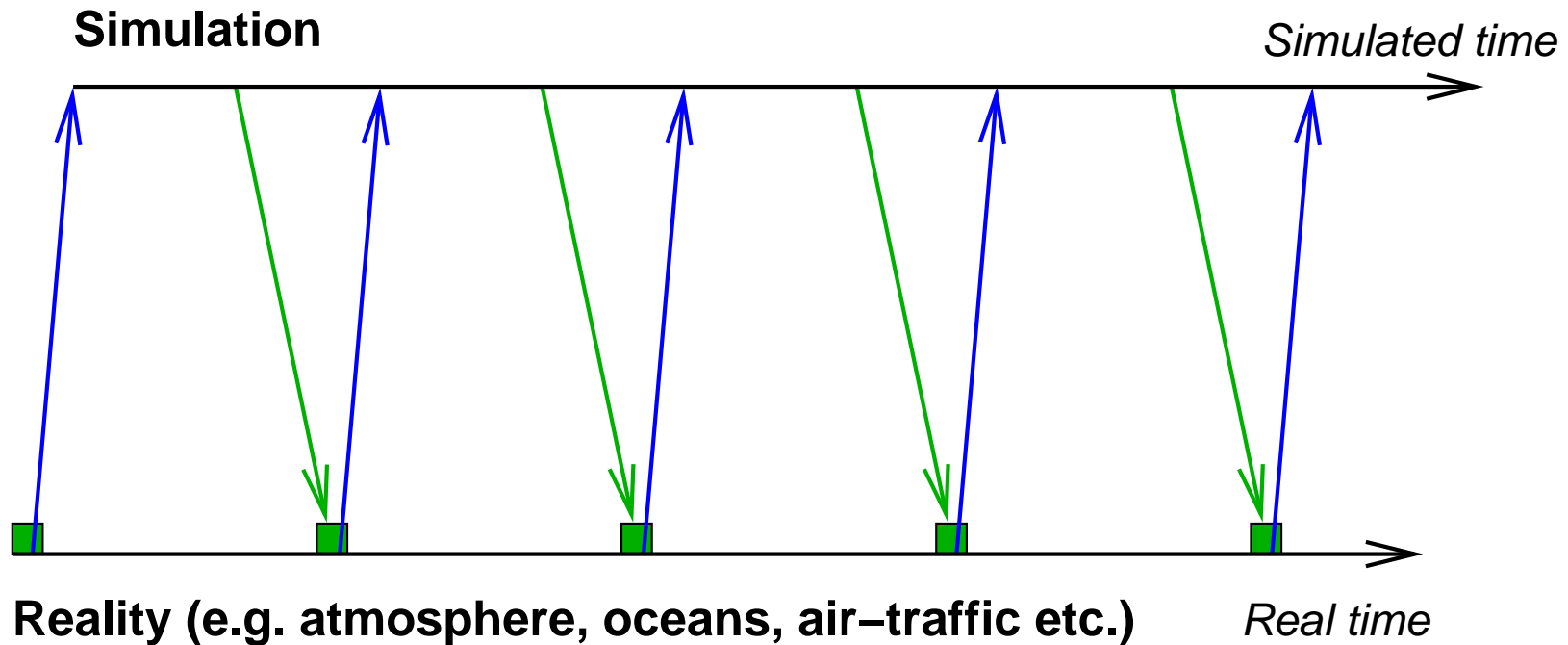
- **Policy decision making** requires understanding of a complex system;
- **Data-driven simulation** can assist with **model building and revision**;
- **Agent-based simulations** can predict **future states, given current state** - or more about present state, given partial state;
- Also possible to run **“what-if” scenarios** for policy actions: start with a **hypothetical state**;
- Simulated “agents” can represent individuals, groups, organisations etc.
- In AIMSS, there are TWO kinds of “agent”: a software agent that assists with **building and testing the model** (by data-driven simulation) and the agents **in the world that are BEING MODELLED**.

Simplest data-driven simulation



 **Data inputs**

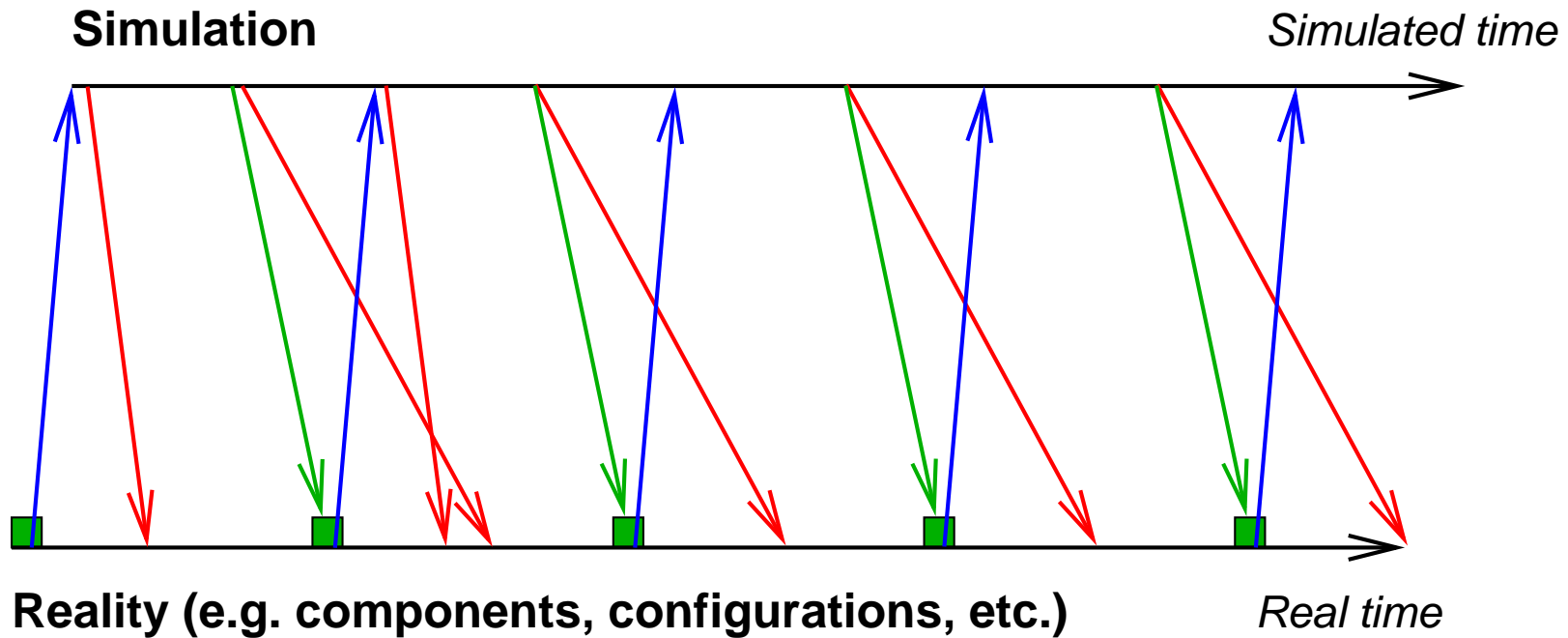
Data driven simulation with sensor control



 Data inputs  Sensors  Sensor control

Predicted state points to a need for more information of a certain type

Symbiotic Simulation



 Data inputs  Sensors  Sensor control  Actions

Note: Different timings of actions: some reactive, some deliberative.

What is being Simulated?

- **An actual observed system (an instance):**
 - simulation states are **expected measurements** of the real system;
 - **direct update** from measurements to simulation;
 - simulation may run concurrently with observed system.
- **A class of systems with similar properties (more likely in social sciences):**
 - simulation states are **abstract states** applicable to this class;
 - model may be revised as a result of **generalisations** from data;
 - data may be collected from **multiple instances** of similar systems.

How to apply to the Social Sciences?

Two processes:

1. Data-driven adaptation of the model:

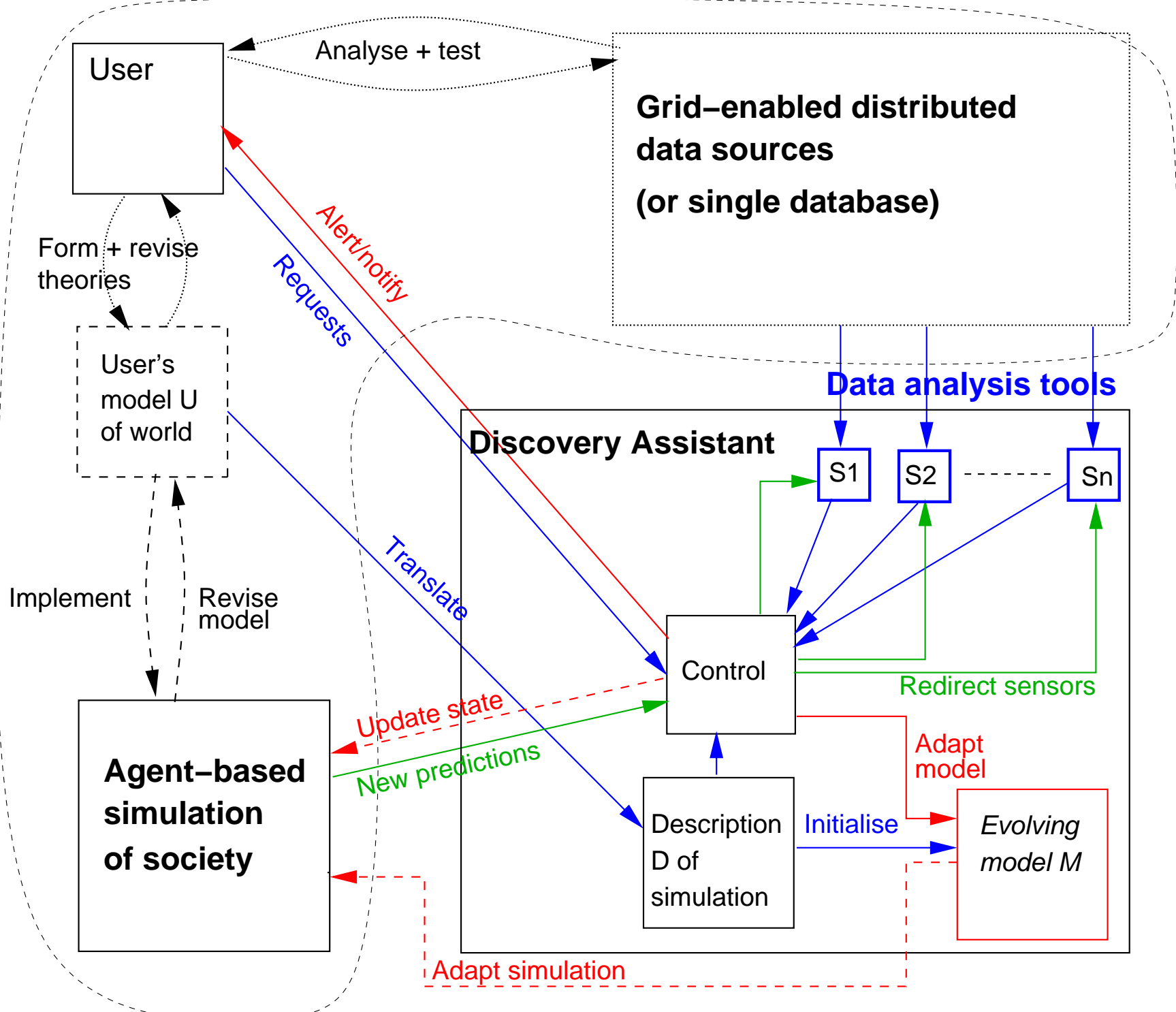
- Simulation **generates hypotheses** about future or current states, given an initial state and a model;
- if the predicted state should already have happened, **test the hypotheses** by querying of data analysis tools.
- if the simulation is about a particular instance of a system (e.g. a particular supermarket, not a typical supermarket), **update its current state** with the acquired information from the data.
- if there is an **inconsistency** between prediction and actual state, **adapt the model**.

How to apply to the Social Sciences? contd.

2. Model-driven adaptation of data selection:

- Hypotheses generated by the simulation **directs and focuses** the data queries;
- the “sensors” that are being directed are the data analysis tools.
- data may be **integrated** from multiple sources in **unexpected ways**.

Social Science Assistance Architecture



Case Study: Housing Policy

- *Problem 1:*

- Current housing market models are **too simplistic**;
- assumptions may not hold in all scenarios;
- need modelling of micro-level behaviour;

- *Problem 2:*

- Understanding micro-level behaviour is a **multi-dimensional problem**;
- incomplete data;
- expensive data acquisition;
- need assistance in determining **what kinds of micro-level data are significant** for policy goals.

Proposed Solution

- Use **agent-based simulation** to represent residents making decisions on whether to move house and where.
- **Predicted states** of the social simulation can be tested by analysing the data;
- The “predictions” are states that **would be expected to exist now** if the model’s assumptions are true;
- **Data analysis and mining tools** are used to inquire whether the predicted state is actually true;
- If insufficient data is available, the discovery agent can suggest new kinds of data that are required in future surveys (Problem 1);
- **Persistent discrepancies** between the simulation predictions and the results of the data analysis prompts the discovery agent to **suggest model revisions** (Problem 2);

Work so far: simple agent-based simulation

- Initial exploratory prototype: represent important features of a typical city in a generalised way.
- Agents are households (one or more individuals);
- Agents inhabit a “space” divided into 4 quadrants:
 - Q1: expensive city centre apartments, small, densely populated;
 - Q2: inner city towerblocks, inexpensive, cramped, densely populated;
 - Q3: modest suburb, moderately populated;
 - Q4: wealthy suburb, sparsely populated.
- Simulation parameters:
 - Density of properties in an area;
 - Distribution of large, small, expensive, cheap etc.
 - Density of occupied properties;
 - Income distribution;
 - Distribution of household size;
- **IMPORTANT**: agents have *needs* and are “happy” or “unhappy”.

Agent and Environment Behaviour

- Start with some assumptions about behaviour that **we know to be incomplete**. This is the initial user model (U); aim to **“discover” something that we already know** - to add to the model.
- Encode the assumptions (e.g. as rules) into the simulated agents' decision-making;
- Example assumptions: an agent wants to move only for the following reasons:
 1. **the current property is not affordable (rent or mortgage too high)**
or
 2. **the current property is overcrowded.**

The first rule overrides the second. As long as an agent is not happy it will keep on attempting to move.
- Given a database on households, properties and moves in the social rented sector, the “discovery assistant” (DA) analyses the available data to **support or refute the assumptions**.

Verifying predictions from the available data

- **Example predictions (from current simulation):**
 - **Small low-income households (1-2 persons) in inner city areas are happy and tend not to move.**
 - **Larger low income households in the inner city tend to move more frequently;**
- **Questions to be asked (from database):**
 - **Which addresses are in areas that correspond to “inner city”?**
 - **How many moves within or from these areas are large households? How many are small households?**
- **If it finds most moves are small households, this particular data refutes the prediction.**
- **Assuming this data is “typical”, the model needs revision.**

How to Revise the Model

- To revise the model, we ask the question:

**What additional behavioural rules might explain the actual data?
(Predictive data mining).**

- How to look for **relevant attributes** in the available data? For example, “Reasons for moving” is relevant.
- BUT this is only an answer to a survey question (with codes 1, 2, 3 etc.)
- Need to connect this to **other relevant data** (e.g. on employment, schools etc.) and direct the search accordingly.
- **Therefore, we need an ontology that connects up these concepts**
- Informal ontology exists (in Java code).
- Need semantic interoperability.

AIMSS: Desired Outcomes

- Exploratory prototype;
- Use **example ontology** to interpret simulation states and data;
- Specify what is **important** for policy goals;
- Show how we can use the experience of PolicyGrid and MoSeS nodes;
- Show that it is worth pursuing as a longer project.
- AIMSS website:
<http://www.cs.bham.ac.uk/research/projects/aimss/>
- Data driven simulation website:
<http://www.nsf.gov/cise/cns/dddas/>