The Representation of Uncertainty for Validation and Analysis of Social Simulations

Deborah Duong
Augustine Consulting, Inc.
Naval Postgraduate School
Monterey, CA
dvduong@nps.edu

David Makovoz
Impact Computing
Silver Spring, MD
david.makovoz@impact-computing.com

Hyam Singer
Impact Computing
Silver Spring, MD
hyam.singer@impact-computing.com

(Authors listed in alphabetical order)

Keywords:
HSCB Modeling, Computational Social Science, Integrative Model Framework, Probabilistic Ontology, Uncertainty, Bayesian Networks

ABSTRACT: An iterative process of social science theory improvement through computational social science includes theory, computation, and then readjustment of theory. The key to iterative improvement is the ability to make use of the partially correct enough to improve upon it. The techniques of soft computation can help us to make full use of partially correct and inconsistent data. Probabilistic ontologies serve to repurpose data for use by drawing probabilistic correspondence. Probability theory can accommodate all types of uncertainty, from credibility in the theories or the data, to intrinsic uncertainty, to uncertainty of match. Probabilistic ontologies can process this data, whether the processing involves Bayesian data generation, the computation of a scalar value of match for validation, the representation of dynamic in a Markov Process, or expressing the data with gradient so that data mining may be used to help readjust theory.

1. Improving Social Theory through Computational Social Science

The scientific method has helped us to accumulate knowledge about, navigate, and control the physical world through an iterative process of theory and statistical hypothesis testing. Scientific empirical methods tease out cause by holding all else the same but a single independent variable, to see if the results behave according to the theory. The computer has helped to make non-empirical advances in science, in the form of elaborate thought experiments, called simulations. Computational science is able to take the building blocks of consistent relations that were established during empirical experimentation and put them together, thinking out the complexities of interrelations between the parts that our minds are not capable of. Computational science is becoming more important than empirical science in physics and in chemistry, sciences in which our theories of fundamental relations are accurate. For example, simulation is the only way that nuclear testing is done anymore.

However, computational science is not as trustworthy as empiricism in every science. In the hard sciences, computational science and simulation have predictive accuracy because the hard sciences are well understood. Computational social science, on the other hand, has less predictive accuracy because social science is less well understood. It is difficult to make controlled experiments where “all else is the same” in social sciences, and so it is difficult to tease out accurate fundamental relations. More importantly, the social world is a complex adaptive system in which the results come from complicated interactions that could not be teased out by keeping all else the same.
even if there were no moral considerations to prevent experimentation. Methodologies are needed to improve the scientific thought experiment of computer simulation, in order to improve theory, so that theoretical cause may be teased out through augmented observation and reasoning, despite the fact that it cannot be empirically manipulated. Our goal is to find ways to understand complex adaptive systems, in which many things cause each other through feedback, on the basis of observation as opposed to manipulation. Since social experiments cannot be adequately controlled for, we need the ability to understand social systems in the absence of as full a capacity to experiment as we have in the hard sciences. A science of society needs methodologies of gaining knowledge given many observations with inexact conditions as opposed to the few observations under exact conditions of physical science.

Natural phenomena appear random to the extent that the phenomena are misunderstood, and sound theories direct us in ways to see the world so that patterns appear and sense is made out of the apparent randomness. With methodologies to improve social theories so that knowledge can accumulate, regularities could be understood and controlled in the social and economic world as they are in the physical world now. This paper proposes an iterative methodology for improving scientific theory by computer, so that the computer may be a mental prosthesis for the thought experiment, to help guide us to regularities and make sense of data from multiple perspectives. Our methodology emphasizes the repurposing of data: data that is observed from many different theories and perspectives because no theory or perspective about the social world has given us a satisfactory amount of control over our social policies, data that does not tease out truth in lower quantities, but helps us to approach truth in higher quantities. Our purpose is to suggest processes through which we may iteratively make sense of data and recognize when we are successful at it. One good thing about our situation is that we know what success looks like: the more accurate our theories are, the more they will reduce uncertainty in a general manner.

An iterative process of theory improvement needs to make use of and improve upon that which is incorrect, inconsistent, and uncertain, and draw up structured correspondence to make sense of it. The techniques of soft computing provide the robustness needed, and the techniques of inference provide the structured correspondence. In the iterative process, first, simulation would compute out the implications of the assumptions in a simulation experiment (thought experiment). Then, we compare the result to results in the real world (validation), and finally, we adjust the categorization of the data and its relations so that the results more generally match the real world (theory discovery), and start over again. Human in the loop is important in all stages of this process, especially in the discovery of new theoretical categorizations and relations stage. Our goal is to suggest ways to assist the decision-maker in all stages of this process as much as possible, even the discovery stage.

2. Uncertainty in Computational Social Science

There is more uncertainty in social science data than there is in the data of the hard sciences. The primary source of this uncertainty is the fact that social scientists themselves do not agree on the appropriate categorizations and theoretical relations of their domain. People’s representations of knowledge about people tend to be very subjective: even the same label may mean different things to different people, and this bleeds over into the social sciences. Because data is categorized differently in different theories, there is less confidence in the veracity of data generated by one theory when used in another. Social data is developed under a particular perspective and set of assumptions, and repurposing it to a different perspective and set of assumptions is problematic.

Add to this the uncertainty of confidence in the source: one subject matter expert may hold less expertise than another does, even if they are in the same theoretical school and agree upon the appropriate categorization of social phenomena and relations between categories. Uncertainty of theory and source are epistemic, meaning that they have to do with what we don’t know about the system rather than the system itself. But, even if we had accurate theory that teased out the patterns from the random, there would still be a lot of randomness and unpredictability in social systems due to intrinsic uncertainty, that is, uncertainty from the system itself. Social systems are Complex Adaptive Systems (CAS), and are thus path dependent, meaning that there are strong arbitrary components to how societies evolve. Added to the fact is that human beings are “the symbolic species” and symbols are arbitrary by definition. Fortunately, all of the kinds of uncertainty in the social sciences can be dealt with in the framework of probability theory, which can combine all sorts of uncertainty to create a variety of datasets, correspondence measures, and confidence measures.

3. Making Use of Uncertain Data

Uncertain data can be used to come to a consensus as is needed to paint a coherent picture of the environment, as a surrogate when data that does not fit exactly is found, and to compare with more trusted data for validation purposes. These uses all involve scalar measures of correspondence, of how much one set of data matches another set.
Painting a coherent picture out of disparate data involves a concept of more or less coherence, which involves translations from one theory to a set of possibilities in another. Together, many theories may create a coherent set that is more than the sum of its parts. Surrogating data involves straightforward matching, for the purposes of replacement, while validating involves a straightforward matching for the purposes of comparison with trusted data.

Painting a coherent picture has an important place in theory discovery and iterative theory improvement. The key to iterative theory improvement with a computer is the capacity to make use of the concept of “partially correct.” Theories that people believe and find useful for their purposes are approximations that usually have some truth to them; within some set of conditions, although we do not always know what, they have helped to navigate some aspects of social problems. In effect, we are trying to put less correct data together to get more correct data. As an analogy, we are trying to find the elephant by interviewing the blind men in the Indian legend who thought the trunk was a snake and the leg was a tree trunk. Each identified correctly from their perspective, but had they talked to each other they would have expanded their individual perspectives into a coherent whole. The only data we have are data taken from many different perspectives and for many purposes. Even if it is not already in the ideal categorization scheme that helps us to understand underlying processes, it is what we have to work with, and a starting point for finding better categorizations to use in better, more explanatory theories. This goes not only for data that is already created for another purpose and seen through another perspective, but also data that is generated from a computer program that sees the world from another theoretical perspective. These data are tailored to particular usages, for which the theoretical perspective is “correct enough.”

Often, what we lack in quality of data for our own particular usage, we can make up for in quantity. Larger quantities of data help us to see what parts of the data are statistically significant, and what parts are spurious: more data is more significant data. Voting systems in Natural Language Processing, for example, increase their accuracy by taking advantage of the fact that there are more ways to be wrong than there are to be right, through voting. The problem with voting systems in the case of simulations is expressing the data in such a way that a correspondence may be drawn to a similar event. To make use of the partially correct is to make use of the partially matching. Social science data, because it is expressed in partially incorrect theories, is also expressed in a variety of different categorization schemes that do not have an exact correspondence to each other. The most valuable parts of the data are those that are the most general, or consistent across perspectives, that is, that which is predicted by both perspectives and co-occurs across perspectives. For example, a model of the economy may be able to predict a recession based on indicators quite accurately, and a sociological model may be able to predict an increase in anomie accurately, and these two social phenomena may be highly correlated across models, though not exactly correlated. The co-occurrence of anomie and recession across models may help to suggest a broader theory that explains patterns of co-occurrence and exclusion in both models. For example, if the sociological model says “anomie,” and the economic model says “recession,” and we know that these match because they co-occur in the real world often, and a third psychological model predicts something that does not co-occur with the other two, such as “well-being,” we can compose a coherent larger picture of the environment with “anomie” and “recession” and perhaps disregard the “well-being” data. In this example, the different systems have voted for “an elephant,” a coherent story from all of the models, because we are able to draw correlative correspondence between them. This statistical coherence may suggest a logical coherence in the scientists mind, to aid in theory discovery.

More straightforward usages of uncertain data include surrogate data when not enough is available, and data for validation. The anomie data mentioned above could be used as a surrogate for the recession data because they co-occur, and if the anomie data is augmented by other data which lends mutual support because of its co-occurrence, it becomes all the more valuable as a surrogate. Furthermore, if a model outputs recession data, the mutual support lent by the co-occurrence of the anomie and other supporting data output by other models is a measurable way to confirm and validate the recession data. The use of uncertain data is important because if we continue to build our software according to crisp rules which need exact data, we are likely to end up with very little data that can be used for either scenario or to validate output data against. However, if we make use of partially matching, inexact data, and can take advantage of large quantities of this data to form patterns we have greater confidence in, we can start talking about how theories expressed in different simulations compare to each other, and have more to work with in order to improve the theories. If we start out with credibility values on data, we can use comparison to assign more credibility values. If we work with what is partially correct and contradictory, we can improve upon it, which is preferable to breaking down when data is incomplete. The techniques of soft computing help keep computations robust and working with contradictory, partially matching data of varying confidence.
4. Integrative Model Frameworks

An integrative framework is essential to implement the painting of a coherent picture between model outputs, and convenient for making partially matching data available for scenario and validation. Concerning painting a coherent picture, theoretical uncertainty in social simulation may be taken into account in simulation results in the same way any other type of uncertainty is taken into account, through repeated runs that try combinations of different possibilities. We want a federation of models to paint a coherent picture together, but we also want to have federations of models that represent all reasonable combinations of different theories. However, in taking into account theoretical uncertainty, instead of trying those different possibilities through drawing different random variates from internal distributions, different simulation models representing different theories would be switched in and out. In order to do justice to all theories and find the policy decisions and courses of action that are robust with respect to the theories, an integration framework would do the switching in and out of theories. Too often, policy decisions are made without all schools of social thought represented, and an integrative model framework could not only find robustness of courses of action against the uncertainty of social theory, but combine this uncertainty with all of the other epistemic and intrinsic uncertainties to create a more accurate depiction of the state space.

With different federations will come different amounts of agreement and disagreement. The social science literature is comprised of two types of studies, both of which can be used in a federation to paint a coherent picture. The first is social theory studies, that can be expressed in a computer simulation. These studies and their computer simulations address cause. Theoretical simulations are designed to forecast states from first principles in different combination. Typically, these simulations are deterministic to the extent that they model a theory, but also use probabilistic draws to model what is outside their theory. The second type of study is a correlational study, that finds out what occurs with what, but does not address why. These studies address correlation and may be represented probabilistically, in the form of prescriptions for how often phenomena should co-occur or appear in sequence. One would expect many correlations to occur between simulation of different aspects of social phenomena, because the different fields in social science are really different perspectives on the same phenomena. Figure one illustrates the large amount of correlation expected between models from the different fields of social sciences. The prescription that social correlative studies give for what should occur together can be used to calculate a numerical measure of how much simulations match and to create consensus states to be fed back into all the simulations of a federation as an agreed upon checkpoint. See figure two for an illustration of drawing correspondence between causal simulations based on social correlational studies. One of these simulations might be, as in the previous example, the sociological simulation of anomie, and another might be the economic simulation of recession, and the correlation between them may have come from social correlative studies that show that anomie and recession are correlated.

If the simulations themselves are adaptive, then changes that cause correlations with other simulations to occur in accordance with social correlational literature would be selected for. Some evolving agent simulations, such as the Nexus network learner[1], are designed to adapt to and in effect, explain, correlative data. There are many different ways that the correlations could be used to define consensus states, give numerical measures of fit of data, and even generate data of the correct correlational structure to feed other simulations. For example, a constraint satisfaction fuzzy/neural network could find a consensus state that seeks to maximize consistency. Instead of simple voting, as in the previous example with the sociological, economic, and psychological model where the psychological model’s data is thrown out, a more complex constraint satisfaction algorithm can be used to find a compromise between the partially conflicting data of all the models. Alternatively, a Bayesian network could give a measure of the consistency of the entire system, or generate a population of agents for the next round that is consistent with the consensus. Bayesian networks have the capability of keeping track of how often different states should co-occur, and have the property of being able to generate a population with attributes with the correct mutual correlations. Given all the correlations that should be amongst phenomena, the Bayesian network can both generate likely example combinations that can be a consensus state to re-seed models with.

Figure 1. Modules in a federation of social science simulations should have a high degree of correlation, overlapping with similar and co-occurring events, as illustrated on the left, rather than running independent of each other as on the right.
5. Conceptual Models

Compositions of simulations have the same problems with drawing correspondence between data of multiple perspectives and resolutions for variable trading as there are between static scenario data and simulations. Just as data from different data bases are categorized differently and without exact translations so are data in different simulations. However, the problem of retaining meaning across translations is worse in the case of the simulation, because meaning is more often inexplicit. The standard practice of databases is to have a schema that describes the data explicitly, and this tends to be enforced because the categorizations of the schema are needed to access the data. On the other hand, although it is standard practice in software engineering to describe programs functionally, in practice the functional design does not keep up with the software changes as often, perhaps because it is not needed to actually use the software as a database schema is needed to use a database. The problem is even more difficult with simulations, because as a scientific methodology it is important to know just what constitutes the theory that is being put to the test, and what is merely implementation details. The conceptual model of a simulation study is what is subject to validation and scientific refutation, but, as in software design, often times this is not as explicited or as maintained as it should be. For example, it may be that we are simulating a type of tank with six wheels, and that part should have fidelity to the real world, but we may not in fact know that is what is simulated until we go through a “for loop,” buried in the code, six times. Separation of the data model from the business logic is a principle for all software engineering, but it is especially important in simulation because the conceptual model defines what should draw correspondence with the real world and what is open to scientific refutation, in this case the important features of a certain type of tank.

Since it is so important to scientific analysis to make the conceptual model explicit, it is worthy of the creation of a standard for this explication for simulation, if not as a universal practice than as a department of defense or policy requirement. Regardless of whether it is a requirement or not, it is easier to enforce the explication of the conceptual model if it is required to use the software, just as it is easier to keep up database schema, because they are put to use.

Something more than UML is needed to describe the conceptual model of a simulation because UML is not in a form that can be used in machine inference. It does not have the richness of description necessary to exactly define what puts two instances in the same category and exactly how they are related, and so is not used in machine inference. However, ontologies have exact enough definitions that can fully specify a functional description of categories and relations. An ontology, in the field of artificial intelligence, is a taxonomic scheme of subsumption, or “isa” relationships, that define what make an instance belong to a category including properties and relations between properties. Particular, lower levels of description come under higher more general categories, as in a taxonomy, but additionally, the relationships between attributes are included. An example of a taxonomic relation is, a car is a type of vehicle, and a subcompact car is a type of car. An ontology would further use properties to define the categorizations, for example, “a subcompact car has no more than four seats.”

Ontologies are open to machine inference, which can check consistency of the conceptual model and keep track of how inconsistent it is. Ontology technology is one of the only ways to explicate a conceptual model that can be an active and necessary part of the entire analysis process, from checking consistency of the conceptual model, to the automation of implementation of the conceptual model for its enforcement, the drawing of correspondence to data whether it be scenario, variable trading data, or data to check validations against. Ontologies can represent a theory through structured descriptions that are unambiguous, or what is known in AI as decidable. Their unambiguous and rich nature put them in a form amenable to both automated implementation, which would enforce a consistent picture of the environment, and any kind of refutation, especially automated refutation, which is essential to the scientific method.

6. Ontologies to Implement Model Composition

Composing models of multiple resolutions is another problem that ontology technology is good at solving. The subsumption hierarchy of ontologies make them appropriate for multiresolutional descriptions and translations. For example, a high resolution tactical level model can generate data described at a tactical level, but would need to be translated into a more general level to be used in a model at the strategic level. An ontology would contain the knowledge with which to make this translation.
Theoretically, an exact ontology is enough to make any transition. However in practice, we may not know enough to draw a correspondence between concepts. Theoretically, if the logic in our ontology was exact, we could automatically go from a lower level of resolution to a higher level of resolution with certainty. Theoretically, to transfer data from a higher level model to a lower level model, the particular lower level description should not matter and could be drawn randomly. For example, a “vehicle” in a higher level model can be a “subcompact car” or a “truck” in a lower level model. However, in practice, not everyone knows or thinks out their logic so carefully, and models are bound to have inconsistencies. This is true not only for model composition, but for any multiresolutional data composition. Probability theory can cover for us where our knowledge ends, in the form of the probabilistic ontology. For example, we may not know all the rules by which one thing should be categorized with another according to theory. When we do not know, we can use probabilities. For example, social theory may say that “anomie comes with a recession if and only if the unemployed identify with their work, and have an identity crisis when unemployed” in which case we would not need a correlative model. However, if the causes are not so crisp, we have greater degrees of uncertainty as reflected in lower correlation coefficients, and settle for relations like, “sixty percent of the time, anomie co-occurs with recession” without needing to explain why.

7. Probabilistic Ontologies

It is important to use the probabilistic ontology when drawing correspondence in the social sciences because our knowledge is very limited, and often we have only statistical studies to draw on. Limited knowledge means many possible perspectives as well, and these have inexact correspondences. If we cannot make use of and draw upon inexact correspondence of data, then we have nothing to give us ideas that lead to theories of unambiguous relations between data born of truer perspectives. Some examples of the use of probabilistic ontologies include:

- Probabilistic correspondence. The exact way to perform correspondence would be to describe variables at their lowest most primitive levels, and recompose them back up into categories of a different perspective. However, we do not always have that information, so sometimes the matching is probabilistic. For example, there is a 20% chance that what a child identifies as a bug is what an entomologist would.

- Properties. Instead of defining an ontological category as either having a property or not, give a likelihood of that property for a category, so a density based membership value of instances may be made. For example, there is a 70% chance that a chair has four legs.

- Subsumed relations. Give a likelihood that any particular value in a higher level category would be a lower level category. Correct frequencies make for better simulations. For example, there is a 90% chance that a murder weapon is a gun.

- Social Correlative Studies. These studies define how often one phenomena should co-occur, or occur in sequence, with another social phenomena. Correlative mappings can be used along with identity mappings to measure how close the properties of one set of data are to another, for the purposes of consensus and validation. For example, 10% of the time inflation occurs, unemployment is also high.

- Dynamic descriptions. A Markov Process can describe the dynamic structures of a simulation, and the probabilistic ontology can be used to describe the probabilities that one set of states will lead to another. For example, there is over 90% chance that a state with factionalization and with a partial democracy will become unstable.

- Causal descriptions. Causal descriptions are a subset of correlative descriptions that are capable of generality. With correct causal descriptions low resolution populations can be generated with Bayesian methods, to be sent to higher resolution models. The results of these models may be read back again into models of lower resolution, integrating a hybrid resolution system. For example, a population is created out of the correlations described in a Bayesian network on the population level, read into a micro theory level model in which actions change those relations, and correlations are once again measured and put into population level Bayesian network.

- Confidence factors. One source may not be as trusted as another source, and we may want to use this in our estimation of what we believe will happen. For example, we may believe the New York Times is wrong twenty percent of the time and the National Enquirer is wrong eighty percent of the time.

- Feasible parameter sets. The feasible inputs to a simulation can be seen as a population of attributes that have correlations. Stochastic runs may be generated with the knowledge of these correlations. For example, ninety-five percent of the time that the visibility is low, the speed of the boat is low.
The nice thing about all these forms of uncertainty is that they can all be used together to generate data through straightforward probabilistic manipulation. For example, we may want to generate data from a variety of uncertain models that thoroughly describes the state space of possible outcomes. Using probability theory, we take into account all types of uncertainty in our results, including possible parameters, intrinsic uncertainty, possible theories, and possible matches between data. We have a single frame of reference, implemented in an integrative framework with probabilistic ontologies, to include and express this state space accurately.

The following are three concrete examples from regular OWL ontologies that are modified for use in probabilistic ontologies:

1. Regular ontology rdfs:subClassOf will have probabilities added to those values:
   \{Gun subClassOf MurderWeapon\}, \{Axe subClassOf MurderWeapon\}, \{Knife subClassOf MurderWeapon\}
   will turn into
   \{Gun subClassOf MurderWeapon; hasProbability 70%\}, \{Axe subClassOf MurderWeapon; subclass hasProbability 10%\}, \{Knife subClassOf MurderWeapon; subclass hasProbability 5%\}

2. Regular ontology owl:someValuesFrom will have probabilities added to those values:
   \{Table hasLegs someValuesFrom (lessThan3legs, 3legs, 4legs, moreThan4legs)\} will turn into
   \{Table hasLegs someValuesFrom (lessThan3legs, 3legs, 4legs, moreThan4legs); lessThan3legs hasProbability 1%; 3legs hasProbability 10%; 4legs hasProbability 70%, moreThan4legs hasProbability 19%\}

3. Regular ontology owl:equivalentProperty will turn into
   \{Pashtu_Region isEquivalent Pashtu_Tribe; Equivalence hasProbability 90%\}
   \{Pashtu_Region isEquivalent Mashtu_Tribe; Equivalence hasProbability 5%\}

8. Measures of Fit to Data for Validation

Probabilistic Ontologies enable a measure of closeness of fit. Once you know how much you expect a match to occur, you can easily calculate how likely or how unusual it is, and probabilistically, what is expected fits better. If you have an unusual scenario, in which things that should correspond don’t, then we can say that the scenario is inconsistent with the data that says that it should. This measure of likelihood or unusualness, this degree of inconsistency, as opposed to the crisp determinations of consistency of traditional ontologies, is what gives a measure of fit to data. Whether this data is in the form of clean causal relationships or messy correlations that double count, soft computational techniques can still derive a measure of fit to data. This is precisely the validation measure that we are looking for, because validation is a comparison of a more trusted data set to a less trusted one. For example, say we have a probabilistic ontology that includes many correlational relations between phenomena, that does not attempt to model cause. For example, it may say that people slip on banana peels in the summer, and then also that many people die in diving accidents. This data does not contain a theory of why people die in diving accidents: for example, it does not say that they die because they slip on banana peels. However, we can still see if summer co-occurs with greater frequencies of slipping on banana peels and also with more diving accidents in our models, matching theory outputs to correlative studies. Better models or federations of models will be correlated with similar correlation coefficients to those found in trusted correlative studies.

The ontology gives us a way to draw correspondences between states in one simulation to those in another, and the probabilistic ontology gives us a way to represent a degree of correspondence. However, those states should also be represented both temporally and stochastically, because in order to measure how close a stochastic simulation is to another stochastic simulation or a time series data set, it is necessary not only to draw correspondence between events, but between probabilistic sequences of events. Markov processes capture probabilistic sequences of events of stochastic simulations well. Graphically, a Markov Process is a series of nodes that represent states of the simulation, states that can be assigned according to sets of variable values within the simulation. The links indicate the probabilities of going from one node to another (see figure 3). It is necessary to run the simulation many times to hit all the plausible outcomes. A degree of probabilistic match between two Markov Processes may be calculated using probability theory, to provide a total picture of probabilistic match between dynamic systems. A Markov process may be created from static data, as long as the data is in a time series. For simulation data, it is necessary to run the simulation many times to hit all the plausible outcomes. If the indicators relevant to the variables of the study are chosen for the states, then a degree of match between the simulation and the time series data may be found based on those indicators. Once in the form of the Markov process, it is irrelevant whether the process came from a simulation or from time series
real world data, and so any combination may be measured for fit. Aaron Bramson used Markov processes to formalize tipping points and path dependencies in the data, which could be points of comparison in addition to a single scalar measure of fit to data, since tipping points are of primary interest in policy making and course of action analysis[2].

Figure 3. A Markov Process for representing the dynamics of a simulation. The letters within nodes represent variable values, and the scalar values on the links represent the probability of transition. From Bramson, Aaron. “Measures of Tipping Points, Robustness and Path Dependence.” AAAI Fall Symposium, 2009.

The reason that Markov Processes are relevant to the comparison of model output to data is that they are a common framework to put all static data and model outputs into, in a way that expresses their dynamics. Once in a common format, a measure of distance between Markov Processes may be calculated. For example, if one state is more likely given another state in another model, we can quantify this difference in likelihood. All Markov Processes that are compared would be using the same model parameters, chosen because they are deemed relevant.

9. Feasible Combinations of Parameters for Data Farming

Finding feasible combinations of parameters to simulations in a federation is really the whole point of using probabilistic ontologies, because that is what in the end gives the commander or policy maker knowledge of the likely outcomes of plausible courses of action. Probabilistic ontologies can help us not only with the description of the data in a Markov Process and the computation of a fit to data, but also with the run of the simulation experiment. To make the Markov process we need to run many times but in a realistic fashion, covering the space of feasible parameter sets. However, simulations usually require a human being to enter the parameters in feasible combinations. In order to automate the process, a way of representing the analyst’s rules for the feasible combinations of parameter sets is needed. The ontology, as a cognitive wrapper for repurposing data, can serve this function. Through an ontology, the input of the simulation may be repurposed for a scientific exploration of the space of possibilities. The simulation programs implement a conceptual model, with mediation ontologies making sense of combinations of parameters. A probabilistic ontology can hold correlations between parameters as seen in the real world, without logic to tell why, or it can include cognitive rules that use logic to choose feasible parameter sets. For example, if a simulation has parameters for boat speed and visibility, the probabilistic ontology could hold the rule that when visibility is low, speed is rarely high, and so this combination of parameters would not become part of the analysis and would never get into the Markov Process.

Strategic data farming is one example of the use of strategic rules to find feasible sets of parameters[3]. Strategic data farming uses game tree techniques from combinatorial game theory to choose which combinations of move parameters to put into scripted simulations. It emulates the analyst’s choice of parameters, including the moves and countermoves that agents would do in a two-sided conflict. The rules about the decision points, branches and sequels, and goal states that the analyst has in mind may be represented in an ontology, which forms a “cognitive wrapper” around a simulation program, that would perform automatically the task of choosing feasible parameter sets for multiple runs. Where rules don’t exist, probabilities can cover for us. For example, probabilities that simply relate states in the environment to possible next moves, that do not include mental models of the opponent. If we have domain knowledge about the opponent’s mental models, it facilitates more realistic parameter sets, but statistical machine learning may be adequate short of that knowledge.

10. Data Mining as a method of Knowledge Discovery in Science

Once we run the experiment with feasible parameter sets, and have come up with a measure of fit to data for validation, we can readjust the categories that we view and the relations between them, so that the simulation fits the desired data more, in our automated iterative process of computational social science. Finding patterns in data that was originally collected for other purposes was the first application of “data mining,” and data mining is a fundamental component of the theory discovery stage of computational social science. However, data mining is controversial in the social sciences. Naysayers argue that data is always seen through theories and therefore cannot discover new theories. One data mining program that
purports to simulate discovery, the BACON program, has come under fire[4]. BACON is given input values to scientific equations and finds the equation. For example, if it were given values for E, M and C, it could find “E=MC^2.” Of course, there is more to Einstein’s discovery than that: the genius of discovery occurred in doing the “simulation” thought experiments that found that C is a relevant variable in the first place. In other words, theory is more about accurately categorizing phenomena than it is about finding relations between the correct categories. Social scientists and artificial intelligence researchers alike agree that knowledge representation is most of the solution. However, the artificial intelligence researchers believe the knowledge representation gap is a gap which their technology can overcome, while the naysayers among the social scientists who are not familiar with promising technological advances in data mining may tend to dismiss the field prematurely for not showing hard evidence of filling the gaps. It is only natural not to trust or believe in what you do not understand: while AI researchers may see programs like BACON as a first step towards a promising future, naysayers see broken promises. However, it remains true that experts in a field are more able to see promising paths to answers that have not yet materialized.

The technology in artificial intelligence that has promise to bridge the knowledge representation and discovery gap is the ontology. Technologies such as ontologies and methodologies such as ontology mediation can help with the repurposing of data and the recategorization of variables.

Furthermore, naysayers argue that data mining picks up spurious (coincidental) patterns. For example, an accepted alpha value for statistical significance is .05. This means that, if an experiment is run, there is a one in twenty chance that we would see a pattern when there was no pattern. However, the availability of adequate amounts of data to separate the testing set from the training set should mitigate the problem. Data mining deserves a place in social science as a technology that can make use of richer descriptions than social experiments that try to tease out cause based on holding all else the same, which are often criticized for being thin descriptions that do not take enough data into account.

Probabilistic ontologies can be redefined with the relevant variables found through data mining, and they can provide gradient for data mining techniques so that new, more relevant patterns in the data may be more found. For example, we may have thought that a terrorist attack was a relevant variable, and so we put it in our Markov Processes, and found it did not lead to the political riots that we thought it would. However, data mining found that only particular types of attacks, the IEDs, reliably lead to riots. The ontology provides different level of description so that the most relevant variables will be found and the more certain the paths in the Markov Processes. Data mining can help to form tight, causally based Markov Processes for comparison by suggesting the relevant variables to include for comparison, and the gradient provided by the ontologies allow for incremental improvement of the model.

11. Summary

An iterative process for incremental improvement in the social sciences includes computation from assumptions, comparison to real world data, and readjustment of assumptions. Uncertainty must be taken into account throughout this process, by covering all cases, in the analysis. Forms of uncertainty, such as epistemic uncertainty in theory, correlations from social studies, and simulation programs themselves are all expressed from different points of view, calling for representation in ontologies. Probabilistic ontologies allow comparisons to data to be drawn for the validation stage, and facilitate discovery of ways to readjust the data representation, in order to suggest new theoretical relations.

12. References


Acknowledgements

The authors acknowledge OUSD(I) and TRAC-Monterey for their support and funding of this project. Opinions expressed in this paper are those of the authors and do not necessarily represent the position of Augustine Consulting, Inc., Impact Computing, OUSD(I), TRAC-Monterey, or SISO.
Authors' Biographies

DEBORAH DUONG is a Senior OR Analyst at Augustine Consulting, Inc. on contract at US Army TRAC-Monterey. Dr. Duong has twenty years experience in Computational Social Science, including invention of the first agent based cognitive social model in 1991. Her experience includes OSD HBR modeling on the Joint Warfare Simulation (JWARS) project and for the OSD SAC IW team, and US Army INSCOM work on data representation. She authored the Nexus, Oz and Indra programs, and is now working on US Army TRAC’s Cultural Geography model. Her email address is <dvduong@nps.edu>.

DAVID MAKOVÖZ possesses over 20 years of experience in applied research and development. With a background firmly rooted in physical science, Dr. Makovoz has been responsible for a number of innovative initiatives across a wide array of complex domains including machine learning, pattern recognition, image processing, signal processing, information retrieval, data mining, natural language processing, and high performance parallel computing. Dr. Makovoz currently serves as Director of Research and Development for Impact Computing Corporation. He possesses a Ph.D. in Physics from the University of Washington. His email address is <david.makovoz@impact-computing.com>.

HYAM SINGER possesses over 25 years of experience in the software industry, with a strong track record spearheading technical innovation, from concept development through deployment in the field. Mr. Singer's broad range of technical expertise spans the areas of visualization, user interface design, geographic information systems (GIS), web technologies, and complex systems architecture. Mr. Singer is President and founder of Impact Computing Corporation. He possesses a B.S. in Computer Science from the University of Maryland. His email address is <hyam.singer@impact-computing.com>. 