Do Not Circulate or Cite Without Author's Permission

THE DANGERS OF RUSHING TO DATA: CONSTRAINTS ON DATA TYPES AND TARGETS IN COMPUTATIONAL SOCIAL MODELING AND SIMULATION

Jessica Glicken Turnley, Ph.D.

Galisteo Consulting Group, Inc. and Joint Special Operations University, USSOCOM
Albuquerque, NM 87110
505-889-3927
jgturnley@aol.com

This paper is a draft prepared under contract to The National Academies for the Committee on Unifying Social Frameworks. Points of view or opinion in this document are those of the author(s) and do not necessarily represent the official position or policies of The National Academies or the Committee on Unifying Social Frameworks.

Bibliographical Note:

Dr. Jessica Glicken Turnleyis President of Galisteo Consulting Group, Inc., a consulting firm in Albuquerque, NM. She also holds an appointment as Senior Fellow, Joint Special Operations University, USSOCOM. Through her work with Galisteo Consulting Group, Dr. Turnley has provided services in policy analysis and national security, strategic business planning, organizational development, corporate culture change, and economic development to a wide variety of clients in the public and private sector. As a Senior Fellow, she provides research, analysis and concept development of selected special operations issues, initiatives and emerging concepts, focusing on those related to organizational and cultural topics pertinent to our own forces and national security organizations, as well as the adversary.

Abstract

By the time most modeling projects address data, the project team has made significant decisions in the course of the project that determine the type of data they need and constrain the part of a 'comprehensive picture' they can provide. In fact, I will argue here that it is not possible to create, *a priori* with data, a 'comprehensive picture' of some area of interest.

To have a model, we require three elements: a thing in the world or in some target domain; a re-presentation of that thing (the model itself); and the relationship between the two that says that the re-presentation is of a particular type and of that particular thing and not some other. A model is not*all* things and *all* relations in the target domain but a selection from them. That selection is made by the modeler: the person (or team) who constructs the model. By exercising this selection process, the modeler acts as sort of a prism, controlling which part of the target domain we see and how we see it. The model as artifact, once it is constructed, embodies this prism.

A theory is a set of phenomena (objects, relations, and dynamics) selected from a target domain and given a particular structure. A model is an illustration or expression of this abstraction. This is called the 'semantic view of models, because the model actually illustrates a set of (mathematical or natural language) sentences, not a thing in the real world. So all models are both true and false: true because the model's idealized structures must connect through theory some part of the target domain so that it is a model of that thing and not of some other, and false because the model does not include all phenomena from the target domain.

So how do we assess the 'goodness' of a model? Verification is a means for determining whether or not the code does what the code-builders wanted it to do. It is usually operationalized as an assessment of internal consistency. Validation, on the other hand, is more problematic. There is no agreement on a precise definition of validation, particularly in the realm of social phenomena. Here, we have questions as basic as the nature of the target domain, e.g., is it behavior or heuristics that drive behavior? However, all seem to agree that the process of modeling, in general, is the process of constructing something 'like' a target domain. The question lies in how we define and determine 'likeness.'

Analogies are one type of statement of likeness and have figured prominently in literature on the philosophy of science (which speaks of models and theories). Analogies and metaphors (which are strong analogies) are attempts to understand the unknown in terms of the known. An analogy is created by examining a target domain, abstracting properties of interest, and mapping those properties to a new domain. Many argue that this is what theories do, thus establishing a close relationship among theories, analogies, and models.

Metaphors and analogies both represent and create similarities between domains, not unlike Clifford Geertz's 'models of, models for' description of symbolic systems. If we call the social dimension 'human terrain,' we think about the target system (the social dimension) in different ways than we would if we called it 'theater.' We give different aspects of it primacy and ascribe to it different types of dynamics. The similarities that

cause us to call it 'terrain' in one case and 'theater' in another are not inherent in the social dimension: they are perceived and communicated by the analogy (model)-maker.

This gives great power to the people involved in the modeling process. I have parsed that process into six different social roles, each of which contributes differently. They are such: the questioner, who poses the question that initiates the process and establishes the model's purpose; the user, who exercises the model in a particular socio-technical environment, is a disciplinary or theoretical expert who identifies the elements to include in the model and the relationships among them; the data provider; and the model builder who translates relevant theory and data into the chosen medium. In some environments, there may be a sixth role, the funder, who may be behaviorally distinct from the questioner and the model user. All the roles are present in every modeling exercise. In some instances, a single individual may occupy more than one role. In such a case, his qualifications for all the roles he occupies must be evaluated and weighed in the context of the overall project.

The exercise of these social roles helps build a particular analogy or model and not another. The questioner, user and funder begin the process of delimiting the target domain. The theoretical expert identifies which portions of that domain are of interest. The data provider offers an instantiation of that abstract structure in a particular time and space. And the model builder captures it all in some presentation medium.

The choice of a presentation medium is an important decision in the modeling process. Computational media put significant constraints on data, particularly on the type of context-rich data often found in the socio-cultural domain. Modelers often address this problem by using surrogates for that data type. These same constraints give primacy to certain theoretical approaches such as grounded theory over others like sensemaking, simply because the medium can better accommodate them. Although a decision may be made in favor of computational media because of the benefits it provides, this should be a decision that explicitly recognizes and weighs the costs in terms of both data and theory.

Finally, the modeling approach (e.g., social networks, agent-based models, or systems dynamics) should be driven by the theory, not determined*a priori*. Since the approach is theory-dependent, it will also drive data selection and collection.

A model is much more than an artifact or bucket into which data can be dumped. It actually is a process of creating a particular way of looking at the world. It is like Karl Weick's sensemaking, a process that 'structures the unknown': using theory to choose elements of the target domain that are relevant to a particular problem, models complete an analogy by representing an instantiation of that selection logic for a particular place and time. So while 'we don't got no stinkin' data' is a legitimate complaint from a modeling team, rushing too quickly to the data question is likely to lead the team to the dangerous and impossible request to 'collect everything' or to collect the 'wrong' things. And finally, by definition, no model will provide us with a 'comprehensive picture' of anything. In fact, the creative power of models may actually cause us to revise our picture through the very act of constructing the analytic tool.

THE DANGERS OF RUSHING TO DATA: CONSTRAINTS ON DATA TYPES AND TARGETS IN COMPUTATIONAL SOCIAL MODELING AND SIMULATION[1]

Introduction

Computational social modelers have been heard to argue that their models may fall short in certain analytic applications because "we don't got no stinkin' data". This paper argues that it is a spurious explanation. Data acquisition and integration is only one small part of a model-building process. As I will show in this paper, there are many critical decisions that affect the nature of any model which must be made before data is incorporated into it. Thus, the model's applicability, and subsequent utility, for some given problem may be compromised long before the data question arises.

While data can certainly contribute in many important ways to the utility of the model (particularly in tactically-focused models), there are many ways in which models can be used independent of data. As Thomas Karas points out, "providing a framework for discussion can become the primary function of a model" (2004, p.12). Participatory or companion modeling is an extreme version of this. In this type of model application, the process uses a role-playing game to help elicit revealed preferences from stakeholders in key decision-making contexts to make explicit their strategic goals and to illuminate political positioning (Mayer and de Jong, 2004). These goals and positions are then input into a computational model. Running simulations then allows the players to see the consequences of positions and goals on certain types of decisions. In this type of model, it is the engagement in the construction and exercising of the model by the model users that is useful to them. It forces them to explicitly declare many subtle elements of the decision-making process that are usually left implicit. This allows them to see how those elements may be manipulated to get different results.

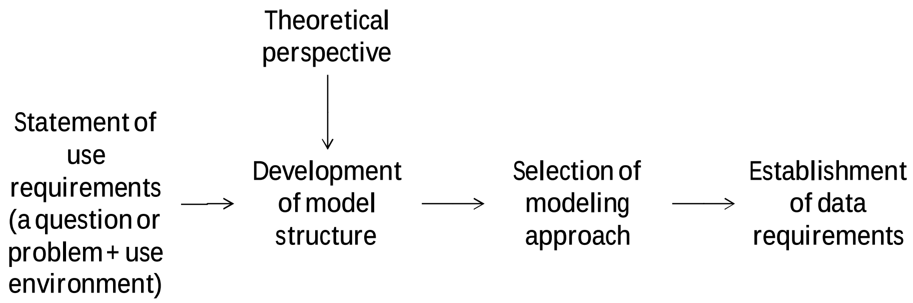
I would like to follow this route by making explicit some aspects of the model development process which are often left implicit. These hidden assumptions and processes seem to be a particularly acute problem in the world of computational social science models: making them explicit allows us to better understand how they impact the nature of the model, its relative goodness, and its utility. In turn, this allows us to manipulate these dimensions to improve the quality of our products.

I will turn the data lament on its head and argue that if we do not have a clearly defined problem, a clear theoretical means for parsing that problem, and a proposed means for manipulating (analyzing) the data we collect, we do not even know which data to gather. I will elaborate on this by addressing the way in which we develop these data requests, i.e., the nature of the mental (and, subsequently, computational) models that need to be populated. In brief, I will argue that data needs are ultimately driven by use requirements. Use requirements are generally framed in terms of a question, sometimes but not always posed by the ultimate model user, and constrained by the use environment. Use requirements drive the development of the model structure which is informed by, and

[1] This paper draws heavily on Turnley and Perls, 2008. The author also wishes to thank Jennifer Perry of DTRA/ASCO for her thoughtful contributions and ideas.

expresses, a theoretical perspective. This, in turn, drives the modeling approach. Finally, the modeling approach drives data requirements, which is illustrated in Figure Error! Reference source not found.. Although the figure makes the process appear linear, as we shall see later, there are many opportunities for iteration in this progression.

Figure 1: Development of data requirements


Theoretical perspective
Statement of use requirements (a question or problem + use environment)
Development of model structure
Selection of modeling approach
Establishment of data requirements

In order to make my argument, I will discuss the nature of a model *qua* model, and then identify various social roles at play in the modeling process. I will conclude by illustrating how the implicit nature of these roles in our engagement with models often drives us prematurely to the data question.

What is a model?

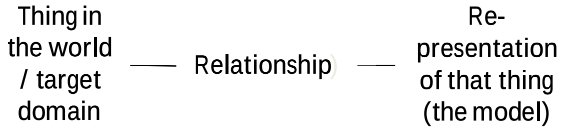
First, for clarification, we need to bound what we mean when we speak of a model. Let me note here that unless I specifically state otherwise, when I speak of a model, I mean a model presented in any one of a variety of media, from mental models to computational models. I am referring to a model in its most abstract sense, as I will define it in this section. If I am speaking of a computational model, I will say so specifically.

Models as prisms

Karas provides us with six different definitions of a model, with sources ranging from the dictionary through glossaries of the Environmental Protection Agency (EPA) and the General Accounting Office *(sic)* (GAO). Interestingly, the Department of Defense (DoD), which arguably uses models as much or more than either the EPA or the GAO, does not provide a definition of a model in its Dictionary of Military and Associated Terms (U.S. Department of Defense, 2010a). All the definitions Karas identified, as well as most others in the literature, suggest that a model – any model – is a re-presentation of things in the world and it is not the thing itself. While this seems obvious, it is important to state, because it raises the question of the nature of the relationship between the two.

To have a model, we require three elements: a thing in the world or in some target domain (a thing which Mary Hesse and others in the philosophy of science characterize as the *explanandum*—the thing to be explained [Hesse, 1966, p.161-162]); a representation of that thing (the model itself); and the relationship between the two that says that the representation is of a particular type and of that particular thing and not some other (Giere, 2004). It is the interaction among these three elements, as presented in Figure 2, that allows models to serve as vehicles for creative perception and, ultimately, for new understanding of the target domain.

Figure 2: Elements necessary for a model


Thing in the world / target domain
Relationship
Re-presentation of that thing (the model)

To say that the model is a representation of the thing in the target domain and not the thing itself is to say that the model is not *all* things and *all* relations in the target domain but a selection from them. (On the data dimension, we might say that it is not all data in the world but a selection from it.) The modeler thus creates, as Margaret Morrison and Mary Morgan put it, a model which is "a partial representation that either abstracts from, or translates into another form, the real nature of the system..." (Morrison and Morgan, 1999, p.27). Those involved in the modeling process, interested in addressing a problem through the creation of a representation, must pick part of the world to represent. This implies some logic of selection which is a critical part of the 'relationship' between the target domain and the model or representation. As Mary Morgan noted, "…modeling requires making certain choices…" (Morgan, 1999, p.386). Ronald Giere emphasized that these choices are not inherent in the model (representation) itself but are exercised by those who construct that representation: "It is not the model that is doing the representing; it is the scientist using the model who is doing the representing" (2004, p.747)

It is the logic of selection that will drive what is included in the model and thus what data is required and ultimately what we 'know' about the target domain. To return to Giere, "scientists do this by picking some specific features of the model that are then claimed to be similar to features of the designated real system to some (perhaps fairly loosely indicated) degree of fit" (Giere, 2004, p.747-748). This gives tremendous power to those who make that choice of what to include in the representation (model). Those involved in the modeling process, through the construction of a statement of similarity, shape what part of the 'real world' (the target domain) we do and do not see. Yet, despite its importance, this logic of selection is rarely made explicit in discussions of models.

So to review what we have on the table so far: a model is not all things in its target domain but a selection of things from it. That selection is made by the modeler, the person (or team) who constructs the model. By exercising this selection process, the modeler acts as sort of a prism controlling what we see and how we see it. The model as artifact embodies this prism once it is constructed.

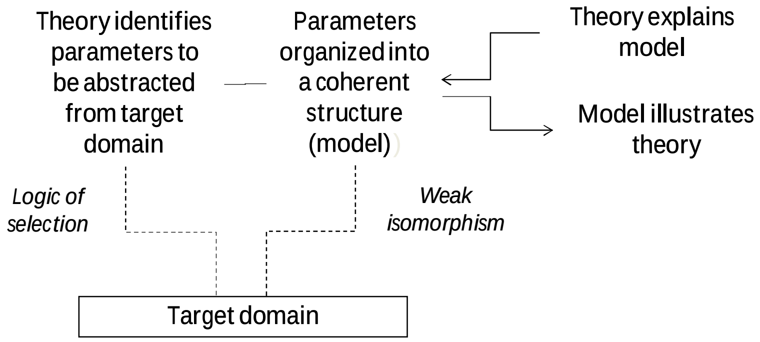
This argument extends the definition of a model from a simple isomorphism (identical reproduction) to a technique for the creation of new knowledge through the logic of selection. It is here that the nature of the relationship between the model and its target domain becomes interesting.

Models, analogies and theories

The body of work in the philosophy of science known as the 'semantic view of models'(see for example, Suppes, 1967; McKelvey, 1999; Beth, 1961; Beatty, 1981, Suppe, 1977, and Liu, 1997) sees models as deriving directly from the nature of theories (in which we may include, for this purpose, heuristics). As Frederick Suppe writes, a theory "does not attempt to describe all aspects of the phenomena in its intended scope; rather it abstracts certain parameters from the phenomena and attempts to describe the phenomena in terms of just these abstracted parameters" (Suppe, 1977, p.223). The set of abstracted phenomena (objects, relations, and dynamics) is represented as a model. The theory thus explains the model and the model illustrates the theory. Hence, the moniker, the 'semantic view of models' because the model actually illustrates a set of sentences (which can be in mathematical or natural language), not a thing in the world; this is illustrated in Figure 3.

.

Figure 3: The semantic view of models


Theory identifies parameters to be abstracted from target domain
Parameters organized into a coherent structure (model)
Theory explains model
Model illustrates theory
Logic of selection
Weak isomorphism
Target domain

The semantic view does posit some isomorphism between the model and the real world, although it may be weak. Bill McKelvey suggests that "ontological adequacy [of a model] is tested by comparing the isomorphism of the model's idealized structures/processes against that portion of the total relevant 'real-world' phenomena defined as 'within the scope of the theory'" (1999, p.17). In this way, "all models are true and false" as Levins says (1966, p.430). They are true because the model's idealized structures must demonstrate *some* degree of fitness with the 'real world' so that it is a model of that thing and not some other. They are false because they do not include all structures in the real world.

Assessing model goodness

This raises interesting questions of the assessment of goodness of models, particularly of computational models. Validation and verification are two activities that are often utilized in this regard. Definitions of verification converge; definitions of validation are more varied, with important implications for computational social models (for further discussion see Turnley, 2005).

Most agree that verification refers to the performance of the code or model script itself. It is defined as follows in the DoD dictionary: "In computer modeling and simulation, [verification is] the process of determining that a model or simulation implementation accurately represents the developer's conceptual description and specifications" (U.S. Department of Defense, 2010a). Or, as IEEE defines it, "Verification is the process of evaluating a system or component to determine whether the products of a given development phase satisfy the conditions imposed at the start of that phase" (IEEE 1990). It is usually operationalized as an assessment of internal consistency ('debugging'): a measurement of whether the code performed against requirements as the developers intended (Gilbert and Troitzsch, 1999, p.17). A suite of statistical tools has been developed to determine the goodness of fit of the code to performance as measured through output.

The validation exercise is more interesting to us here. There are many, many definitions of validation in the literature but all revolve around some assessment of the goodness of fit between the model and the target domain. For our purposes, we will start with the DoD definition: "In computer modeling and simulation, [validation is] the process of determining the degree to which a model or simulation is an accurate representation of the real world from the perspective of the intended uses of the model or simulation" (U.S. Department of Defense, 2010a). Clearly, this begs all kinds of questions, given our early discussion on the definition of a model. For example, how do you measure 'accuracy' of representation if every model is an abstraction from the whole? If the world being modeled is one of motivations or beliefs, how does one establish the 'reality' of that world? Are we modeling the 'reality' of observed behavior, or are we modeling the 'rules' or heuristics that generate that behavior (Dreyfus, 1999)? Is the model focused on behavior or on the meaning placed on the behavior? (And, if not on the meaning, we might ask: what makes a social model different from a model of the movement of inanimate objects in space?)

Note that the DoD definition also includes a statement about the utility of the model. This underscores the importance of a clear statement of the question which initiated the model development process. It also reminds us of George Boc's famous statement, "Models, of course, are never true but, fortunately, it is only necessary that they be useful" (GEP Box, 1979, p.2). But how, then, do we measure utility in this context?

Nigel Gilbert and Martin Troitzsch's definition of validation specifically targets social simulations. They write that "A model which can be relied on to reflect the behavior of the target is 'valid'. Validity can be ascertained by comparing the output of the simulation with data collected from the target" (Gilbert and Troitzsch, 1999, p.20). This still leaves us with questions about how much of the target's behavior must be represented for a model to be valid. We also still need to ask which part of the behavior is relevant: do decision rules or acts of belief count? And how do we assess the goodness of 'toy' models such as Sugarscape, the very simple artificial society developed by Joshua Epstein and Robert Axtell (1996)? Although extremely simple in structure, it has provided some very interesting and useful insights into human dynamics.

The creative side of models

All seem to agree that the process of modeling is the process of constructing something 'like' a target domain. The question lies in how we define and determine 'likeness.'

Analogies are one type of statement of likeness and have figured prominently in literature on the philosophy of science which speaks to models and theories. An analogy is a statement such as "a man is like a wolf" or "the social dimension is like physical terrain [i.e. is 'human terrain']." An analogy – and metaphors, which are strong analogies (a man *is* a wolf; the social dimension *is* terrain) – is an attempt to understand the unknown in terms of the known. Mary Hesse, one of the earliest proponents of the models-as-analogy approach, says that an analogy "may be said to exist between two objects in virtue of their common properties" (1966, p.58). One thus creates an analogy by examining a target domain, determining what properties are of interest, and mapping those properties to a new domain (Hesse, 1966, p.157). Peter Godfrey-Smith writes, "The modeler's strategy is to gain understanding of a complex real-world system via an understanding of simpler, hypothetical system that resembles it in relevant aspects" (2006, p.726); that simpler, hypothetical system is, of course, a model.

To construct a model/analogy, the target domain must be perceived as having certain properties, some of which are relevant and others not. As Roman Frigg writes in his theoretical discussion of models, the target domains themselves have no inherent structure (Frigg, 2002, p.2). That structure is ascribed by those who construct the model through the imposition of the logic of selection. George Lakoff and Mark Johnson take a similar position when they argue in their work on metaphor that similarities are not inherent in objects themselves: "[T]he only similarities relevant to metaphor are *similarities as experienced by people"* (emphasis in the original; Lakoff and Johnson, 2003, p.154).

Metaphors (and analogies) simultaneously represent and construct these similarities. This is not unlike Clifford Geertz's description of symbolic systems as both "models of" and "models for" action. "Culture patterns have an intrinsic double aspect: they give meaning, i.e., objective conceptual form, to social and psychological reality, both by shaping themselves to it and by shaping it to themselves" (Geertz, 1973, p.93). Theodore Brown extends this argument directly into scientific work using the metaphor of 'protein folding' to show how highlighting certain similarities drove experimentation in particular directions (Brown, 2003).

This is similar to Hesse's concept of a 'neutral analogy'.*Positive analogies*, she says, are properties that we believe belong to both parts of the analogy, the target domain (the explanandum) and the model. Properties which are significant to one element but not the other constitute *negative analogies*. Finally, properties of which we are unsure Hesse calls '*neutral analogies*'. A model is the representation of positive and neutral analogies (1966, p.8). Scientists extend theory by identifying neutral analogies and testing them to see whether they are positive (meaningful) or negative (Rentetzi, 2005, p.382).

Hesse calls the creative power of the 'interaction view of metaphor" (1966, p.158ff) and links it back to her notion of the neutral analogy. Because we speak of 'protein folding', we imagine what things (proteins) can do that we otherwise might not have conceived. In this way "theoretical explanation [is] metaphoric redescription of the domain of the explanandum" (Hesse, 1966, p.157). In our national security domain, think of the implications of calling the computer-focused dimension of national security cyber*space.* That metaphorical statement puts the cyber domain in a similar conceptual space as air, land, sea and outer space. It takes on what may be spurious physical characteristics and the characteristics that are unique to the cyber domain may be lost to analytical view. If we speak of the 'human terrain', to take another example, we engage with the social dimension much differently than if we think of it as theater. As terrain, the social dimension has identifiable attributes, is relatively stable over time (after all, a road is always a road – it does not 'renegotiate its identity' with you the next time you travel it), and we can engage with it without significantly affecting it. As theater, every performance or social engagement is different even if the play is the same; the actors themselves change and change others as they engage.

'What is a model?' summarized

So a model is not all things in the target domain but a selection from them. That selection is driven by theory and expressed analogically. The choice of theory and construction of the analogies is made by the modeler (modeling team). This choice and construction constrains what we 'see' of the target domain. It is not and cannot be a 'comprehensive picture' of it, for then it would not be a model but would be the domain itself. The model is validated in some sense not by assessing its isomorphism with the real world, but by ascertaining how well it represents the theory – the logic of selection.

The modeling process can also help us create new knowledge, not just manipulate knowledge we already possess (Morgan, 1999). This is a "view of science not only as a set of evolving theories but as a complex process that includes an interrelation of theories, experiments, and instrument making".

Social roles: the power of the people

In the discussion about the nature of models, I made it clear that those who construct the models have creative power. They define what we 'see' through the model by the choice of elements from the target system and their subsequent incorporation of a particular structure in the model. In this section, I will parse the social roles involved in constructing any model and, in particular, a computational model. I will suggest how each of those roles contributes to the nature of the prism that the model affords and how each interacts with the others. Finally, I will argue strongly that the product of the process is a function of the quality (as in qualifications) of the individuals chosen to fill those roles. The old computer adage still holds: garbage in, garbage out.

I argue that there are six social roles involved in any modeling process, where a social role is a set of expected behaviors.[2] These roles are the *funder* or *project supporter*, the *questioner*, the *user*, the *disciplinary or theoretical expert*, the *data provider*, and the *model builder*. Keep in mind as we go through this explanation that the same individual may occupy more than one role, but each role is exercised in every model-building process.

A *funder* or *project supporter* provides the resources for the project. This role is exercised even if the project is 'self-funded'. The way in which this role is exercised provides constraints upon the project in terms of both resources and general direction. A funder may be primarily (or partially) interested in a product because it will establish his credibility in a particular community or allow him access to a specific social space, as well as for its contribution to the stated need of a project.

The *questioner* poses the question which establishes the model's purpose. This role is often conflated with the model *user*. Again, while they may be exercised by the same individual, the behavioral expectations are different. In practice, conversations between the questioner and user can establish the required precision or accuracy of the model. The user will be familiar with the use environment while the questioner may not; the questioner may set requirements of theoretical rigor and justification because of the demands placed upon him for explanation and accountability that may be beyond those that would be useful to the user (see Levins, 1966, p.421-431 for a much-cited discussion of the tradeoffs between realism, generality and precision).

The *disciplinary or theoretical expert* and the *data provider* are often confused in practice, although their roles are quite different. It is the disciplinary expert who determines the structure and dynamics of the model by invocation of theory in the context of the question. If the question revolves around changing cultural narratives, the disciplinary expert might bring into play theory about the ways in which social relationships are exercised, ideological structures constructed and communicated, and theories of social and cultural change. Since social science theory is contested in ways that theories about the functioning of the physical world are not, the selection of the theoretical expert will have a significant impact on the structure and dynamics of the model. Recall that the theory—and its embodied associated analogies—embody some but not all elements of the target domain. The theoretical expert is a major player in the construction of the prism and the constraints on perception a model imposes. As the theory drives the data collection requirements, this is an important point for our discussion.

The *data provider* provides data on the articulation of the model structure and dynamics in some specific area. In some cases, particularly in the early boom of computational social modeling in the national security world, we found the model builders themselves developing the model structure, identifying elements and relationships they believed were important. Social scientists served as minor members of the team, usually only in a data provider role. In this case, the model builders were filling the role of theoretical expert and so driving the call for data. In practice, we find that the

[2] While I found a great deal of literature on the conceptual construction of models, including literature on the construction process as a communication process, a conceptual development process and the like, I could find nothing that addressed the social roles at play in the process.

data provider in the form of a regional expert is often also used as a theoretical expert and identified generally as a subject matter expert (SME). While the data provider may, in fact, be a theoretical expert, he also may know only of a particular instantiation of the attributes in the world. This pulls the model towards a representational rather than an analogic model. The model is creating a map of a particular part of the world rather than applying theory about social relationships to a particular case.

Finally, there is the *model builder*. The model builder expresses the theory in some presentation format, e.g., computational or textual. Ideally, this expression is one that will accommodate the data provided by the data provider. In the case of conceptual or mental models, the model builder usually is a theoretical expert who expresses his theories in natural language. In the case of computational models where the model is expressed in code, with some notable exceptions, the model builder is a different individual than the theoretical expert. We summarized these roles in Table 1 (Turnley and Perls, 2008, p.7).

Table 1: Roles in the modeling process.

Role descriptor	*Expected behaviors*
Funder	Provides resources to fulfill a purpose
Questioner	Establishes model purpose
User	Utilizes model for intended purpose
Disciplinary or theoretical expert	Provides theoretical knowledge
Data provider	Provides data relative to a specific instantiation of the theory
Model builder	Translates theory into the chosen presentation format

Every one of these roles is exercised every time a model is built. One individual may (and often does) perform several roles, but qualifications that suit him for performance in each should be separately evaluated. In practice, this rarely occurs.

'Data' as a problematic concept

I would like to address one final element of the dangers of the rush to data. This element has to do with the relationship of the presentation medium and associated modeling approach to data. We will address first the constraints placed on data by the medium and then those placed by the approach.

The presentation medium is the way in which the structure and the data are manipulated and usually (although not always) presented to the user. A computational model, for example, will manipulate the data according to computationally programmed structures and processes but the results can be presented textually or graphically. However, the medium chosen for the model puts constraints on and provides opportunities for the manipulation of data. In the computational realm, particularly when addressing socio-cultural data, the benefits are often highly touted but the constraints and limitations are generally not presented.

Computational models require quantitative data, or (to put it another way), data that can be manipulated quantitatively. Much of the data collected about sociocultural phenomena is in narrative form. Furthermore, many of the targets of interest are abstract phenomena such as beliefs, motivations, and the affective dimensions of behavior. If we are to help develop 'alternative cultural narratives' for applications such as irregular warfare (U.S. Department of Defense, 2010b, p.30),[3] then we must understand those narratives. And if our primary analytic tools are computational, it is unclear how we can get relevant 'data' about narratives into those tools.

As Daniel Dohan and Martin Sánchez-Jankowski point out, "Computer assistance is not free – theoretically or methodologically" (p.484). While they are speaking specifically of tools to analyze ethnographic data, the statement holds for the kinds of computational models about which we are talking here. Dohan and Sánchez-Jankowski argue that computational ethnographic analysis tools push the researcher towards a grounded theory approach by virtue of the structure and requirement of the tools (p.479). We find that computational social models exert the same push away from context-dependent data used by theoretical approaches, such as symbolic interactionism and other sensemaking.

This is not necessarily bad but it should be recognized, for it has consequences. What has happened in practice with these computational models is that the context-sensitive ethnographic data is being converted into computationally manipulable data through the use of surrogates which strip it of context. As a rather simplistic example, 'religious belief' is, in many cases, a complex phenomenon at the intersection of such dimensions as the importance of community participation, a need for belonging, the manipulation of local power structures as well as beliefs in the transcendent. Religious belief which is*manifest* through behavior may be *reduced* to that behavior in order to be computationally manipulated. Such behavior (e.g., the number of times one prays per day or attends church per week) is observable and quantifiable. Intensity of belief, thus, is represented by different types and quantities of performance. This does not take into account the fact that belief may endure although the performance may change.[4]

In the construction of most computational social models, the selection of a computational format for data manipulation is driven almost entirely by problem type and use environment. It generally is assumed that data will be available in the appropriate format.

[3] "The second principle [of irregular warfare] is to work with the host nation or local partner to …assist it in crafting alternative narratives that are culturally authentic and at least as compelling as the adversaries'."

[4] This really may not be such a simplistic example. For example, we have heard tell of 'rice bowl Christians' but never of 'rice bowl Muslims.' Is there, perhaps, some analog to rice bowl Christians in more impoverished areas of interest where the Taliban have made headway by providing certain social services along with religious instruction that we are missing in some instances because of an inappropriately constrained focus?

I want to make it clear here that at issue is *not* that the use environment and the benefits of computational modeling tools are trumping the costs of data conversion; this tradeoff may be correct and appropriately made. The issue I have is threefold:

- the costs of data conversion are not recognized and factored into any assessment of the usefulness of the model;
- the theoretical constraints the computational tools impose on analysis are not acknowledged; and
- the attention given to the development of new types of computational tools that can manipulate qualitative and textual data, and do this in qualitatively different ways than we do it now, is almost invisible in the funding and program landscape.

There are other issues related to collecting, storing, and managing sociocultural data that we will not explore in this paper. Social and cultural data change over time. Kinship relationships change as people marry. Economic relationships begin and end. People adopt or discard ideologies, religions, and political beliefs. Data can be collected in multiple formats (audio, visual, and textual), each of which will provide different information on the same 'event'. The theories that drive the selection of data collected also will change over time as our questions change.

And how is the data stored so it can be retrieved? The creation of the data taxonomy (or ontology) is a non-trivial and hugely important task.[5] And such ontologies are driven, to a certain extent, by determinations of relevance to a central question. When questions change, data that was previously irrelevant may become relevant.[6] It is the questions we want answered that drive the selection of theories that ultimately determine what we deem important in the target domain.

Selection of a modeling approach

So this brings us to modeling approach. The modeling approach is the particular type of presentation selected from the entire class of possible presentations in a particular medium. Social networks, agent-based models, and systems dynamics are all computational modeling approaches. Each one of these approaches expresses a different theoretical model. If the theoretical model says that relationships between individuals are the key to answering the question posed to the modeling team, then clearly social networks are the approach of choice. If we are interested in why certain social communities form, an agent-based approach might serve us well. If we would like to know the effect of a particular intervention on a system, systems dynamics can provide

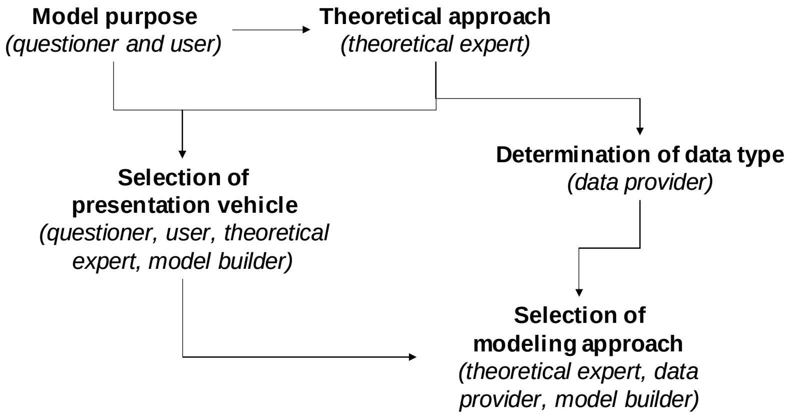
[5] See the Outline of Cultural Materials for the Human Relations Area Files for an example taxonomy. http://www.yale.edu/hraf/collections.htm

[6] As I was searching for sociocultural data on a particular denied area for a model I was helping develop, I was told by a Special Forces operator that they had been collecting this type of data for a couple of years and holding it in theater. When I asked how I could get access to it, he said it was collected for a lot of different reasons so it was of a lot of different kinds, and it was collected via a variety of mechanisms. It was all then just dumped onto a server because they did not know how to structure or organize it. So there was no way to access or analyze it.

some useful insights. Some questions are best answered by a combination of approaches, but the question and the relevant theory should drive the selection of approach. A modeling project should never begin with “I need a social network model of…”.

Each modeling approach requires a different *kind* of data set, not just different data. The data on relationships collected to construct a network-based model of recruiting will not answer questions about what causes people to join that network. Therefore, the modeling approach should be selected by the questioner, the theoretical expert, and the model builder. The questioner bounds the problem. The theoretical expert characterizes the target domain by selecting from it the elements, relationships, and dynamics he believes to be of explanatory importance given a particular question. The model builder will then select an approach that can most effectively manipulate those elements, relationships, and dynamics. The modeling approach thus serves as an additional lens to focus the model user’s attention on certain parts of the target domain and not others. The shape and dimension of that focus should be determined by the theoretical expert and guided by the questioner. The relationships among these roles are shown in Figure 4 (Turnley and Perls, 2008, p.28).

Figure 4: Participants in selected aspects of the modeling process.


Model purpose
(questioner and user)
Theoretical approach
(theoretical expert)
Determination of data type
(data provider)
Selection of presentation vehicle
(questioner, user, theoretical expert, model builder)
Selection of modeling approach
(theoretical expert, data provider, model builder)

Back to data

I would be remiss if I did not bring us back to the data question and I will do that by returning to the discussion of validation. Recall that I have argued that a model is actually validated against theoretical statements, that what is assessed is the fidelity of the representation (the model) to the selection logic. How does data come into play?

Suppose we have developed and constructed a computational model of recruitment into violent religious extremist groups. We have identified certain theories about how and why individuals might join such groups – need for belonging, economic concerns, belief in a particular ideology, and so on. We then constructed a structure that can express that theory. This structure tells us we need data of a certain type. We then populate that model with data from some part of the world, data that corresponds to the theoretical space we have defined. (In reality, these two steps of constructing a structure and populating it are generally performed iteratively but, for discussion purposes, we shall describe them sequentially.) If we posit that joining such groups is a function of a need to belong, data on economic circumstances may not be required except as it plays into the development of this need; data on jewelry or height and weight probably is not relevant either.

We input the data and run the model. We then compare what the model tells us with what we see in the 'real world'. If it does not match, we may assume that (a) we have incorrectly identified one or more elements of importance, or (b) our data was 'bad,' i.e., inaccurate or incomplete. If our answer is (a), we have disproved the theory we are using through a validation exercise. If the answer is (b), we have performed a type of verification exercise which may have identified issues regarding data collection, or storage and retrieval that are outside the scope of this paper. However, neither of these answers means that the model is useless and so completely 'bad'. Even if we directly disprove the theory (answer (a)), then we still may learn something. Discovering which aspects of 'folding' do not apply to proteins may still tell us something very useful about proteins.

Summary and conclusion

I have shown here that a model is much more than an artifact or bucket into which data can be dumped. It actually is a process of creating a particular way of looking at the world. It is like Karl Weick's sensemaking, a process that 'structures the unknown' (Weick, 1995). Using theory to choose elements of the target domain that are relevant to a particular problem, models complete an analogy by representing an instantiation of that selection logic for a particular place and time.

This discussion has shown that by the time most modeling projects address data, they have made significant decisions in the course of the project that determine the type of data they need and constrain the part of that 'comprehensive picture' they can provide. The way the challenge question for the model is phrased (generally couched in language around the use of the model) and the theoretical predilections of the modeling team will determine the theoretical framework for the project. This will determine, in turn, the modeling approach. Now, and only now, should the modeling team request data. Some of the data required may not be amenable to inclusion in the model due to the manipulation requirements placed on data by, for example, a computational medium. Some data, therefore, while required for the model by theory, may be data that come at a cost that is difficult to calculate. So while "we don't got no stinkin' data" is a legitimate complaint from a modeling team, rushing too quickly to the data question is likely to lead the team to the dangerous and impossible request to 'collect everything'. And finally, by definition, no model will provide us with a 'comprehensive picture' of anything. In fact, the creative power of models may actually cause us to revise our picture through the very act of constructing the analytic tool.

Bibliography

Beatty, J. (1981). On Behalf of the Semantic View.*Biology and Philosophy*, 2, 17-23.

Beth, E. (1961). Semantics of Physical Theories. In H. Freudenthan, ed.,*The Concept and Role of the Model in Mathematics and Natural and Social Sciences*. Dordrecht, Germany: Reidel. P.48-51.

Brown, T.L. (2003). *Making Truth: Metaphor in Science.* Urbana-Champaign, IL: University of Illinois Press. p. 122-145

Dreyfus, H.L. (1999). *What Computers Still Can't Do: A critique of artificial reason.* Cambridge, MA: MIT Press.

Epstein, J.M., and R. Axtell. (1996). *Growing Artificial Societies Social Science From the Bottom Up*. Brookings Institution Press and MIT Press.

Frigg, R. (2002). Models and Representations: Why Structures are not Enough. Center for Philosophy and Natural Sciences, Measurement in Physics and Economics Technical Paper 25/02, London School of Economics. London, England.

Geertz, C. (1973). Religion as a Cultural System. In Clifford Geertz,*The Interpretation of Cultures.* New York, NY: Basic Books. (pp.87-125.).

Quoted in G.E.P. Box. (1979). Some Problems of Statistics and Everyday Life.*Journal of the American Statistical Association*, 74(365), 2.

Giere, R.N. (2004). How Models Are Used to Represent Reality. *Philosophy of Science,* 71 (Dec. 2004), 742-752.

Gilbert, N. and K.G. Troitzsch. (1999). *Simulation for the Social Scientist*. 1st edition. Philadelphia, PA: Open University Press.

Godfrey-Smith, P. (2006). The Strategy of Model-based Science. *Biological Science*, 21, 725-740.

Hesse, M.B. (1966). *Models and Analogies in Science*. Notre Dame, IN: University of Notre Dame Press. Notre Dame, IN. (pp. 161-162).

IEEE Standard Glossary of Software Engineering Terminology. IEEE Std. 610.12-1990

Karas, Thomas. (2004). *Modelers and Policymakers: Improving the Relationship.* SAND2004-2888. Sandia National Laboratories, U.S. Department of Energy, Albuquerque, NM.

Lakoff, G., and M. Johnson.(2003). *Metaphors We Live By.* Chicago, IL: University of Chicago Press.

Levins, R. (1966). The Strategy of Model Building in Population Biology. *American Scientist*, 54, 421-431.

Liu, C. (1997). Models and theories I: The semantic view revisited.*International Studies in the Philosophy of Science,* 11(2), 147-164.

Mayer, I., and M. de Jong. (2004). Combining GDSS and Gaming for Decision Support. *Group Decision and Negotiation*, 13, 223-241.

McKelvey, B. (1999). Complexity Theory in Organization Science: Seizing the Promise or Becoming a Fad? *Emergence*, 1, 5-32.

Morgan, M.S. (1999). Learning from Models. In Mary S. Morgan and Margaret Morrison, eds., *Models as Mediators*. Cambridge, UK: Cambridge University Press. pp.347-388.

Morrison, M., and M.S. Morgan. (1999). Models as mediating instruments. In Mary S. Morgan and Margaret Morrison, eds., *Models as Mediators*, Cambridge, UK: Cambridge University Press, pp.10□37.

Morgan, M.S., and M. Morrison, eds. (1999). *Models as Mediators*. Cambridge, U.K.: Cambridge University Press.

Rentetzi, M. (2005). The Metaphorical Conception of Scientific Explanation: Rereading Mary Hesse. *Journal for General Philosophy of Science* / Zeitschrift für allgemeine Wissenschaftstheorie, 36(2), 377-391.

Suppe, F. (1977). *The Structure of Scientific Theories* (2nd end), Chicago, IL: University of Chicago Press. P.223 quoted in McKelvey, op.cit. P.15

Suppes, P. (1967). What is Scientific Theory? In S. Morgenbesser, ed.,*Philosophy of Science Today.* New York: Meridian Books. P.55-67.

Turnley, J.G. (2005). *Validation Issues in Computational Social Simulation.* Working paper prepared for 3rd Lake Arrowhead Conference on Human Complex System. May 18-22, 2005. http://hcs.ucla.edu/lake-arrowhead-2005/HCS2005_JessicaTurnley2.pdf

Turnley, J.G., and A.S. Perls. (2008). What is a Computational Social Model Anyway?: A discussion of definitions, a consideration of challenges and an Explication of Process. Defense Threat Reduction Agency, Advanced Systems and Concepts Office, Report No. ASCO 2008-0013.

United States Department of Defense. (2010). Joint Publication 1-02, DOD Dictionary of Military and Associated Terms, 12 April 2001, as amended through April 2010. Accessed online at http://www.dtic.mil/doctrine/dod_dictionary/29 July 2010.

U.S. Department of Defense. (2010b). Irregular Warfare: Countering Irregular Threats. Joint Operating Concept. Version 2.0. 17 May 2010. Washington, D.C.

Weick, K.E. (1995). *Sensemaking in Organizations*. Thousand Oaks, California: Sage Publications. See especially chapter 1, p. 1-16