15

Theory-Interpretable, Data-Driven Agent-Based Modeling

William Rand

Department of Marketing, Poole College of Management, North Carolina State University, Raleigh NC 27695, USA

The Beauty and Challenge of Big Data

The beauty of big data is all around us and has the potential to be a huge boon to the social sciences. Right now billions of people are carrying cell phones, posting content on social media, driving smart cars, recording their steps, and even using smart refrigerators (Riggins and Wamba 2015). Data is being generated at an extraordinarily fast rate, and many of our traditional methods of analyzing this data face challenges.

The hope and promise for the social sciences is that these vast data sets will give us new insight into the basic way that people interact and behave. There are many different types of data that are available as part of the big data revolution. First, there is the digitization of traditional administrative data (Kitchin 2014). This includes everything from income tax records to parking violation data. This data has always been collected (though not always stored) but now is much more amenable to analysis due to the digitization of the data and the development of ways to get access to this data. For instance, the open government data movements (Ubaldi 2013) have the stated goal of making and increasing the amount of data available to be analyzed by private citizens. Hackathons have been one way to take advantage of this data and to create apps that the average individual can use to make better decisions using this data (Matheus et al. 2014). Though the use and the analysis of this data may be new and different, the basic data has existed for a long time, and so this chapter will not dwell too much on this type of data.

Another form of big data is trace data. Trace data is the data left by an individual as they move through the world sometimes purposefully recording data about themselves, e.g. social media data, and sometimes being unaware that they are creating the data, e.g. the GPS traces stored on a user's cell phone.

Social-Behavioral Modeling for Complex Systems, First Edition.
Edited by Paul K. Davis, Angela O'Mahony, and Jonathan Pfautz.

Companion website: www.wiley.com/go/Davis_Social-Behavioralmodeling

Some researchers have derided this type of data as digital exhaust (Watts 2013), but this data is more than that since it can sometimes give us insights into what people are doing (Eagle and Pentland 2006) or even their political beliefs (Golbeck and Hansen 2011). However, even if this data is nothing more than the social equivalent of the exhaust from a car's tailpipe, that still would potentially give us new insights into human behavior. After all a car's exhaust if we could track it would tell us where the car is and how well the car was running. In the same way, tweets from a user on Twitter may tell us where that user is (Backstrom 2010) and whether or not they are mentally healthy (De Choudhury et al. 2013). Regardless, this data is clearly valuable, since it was essentially trace data in the form of liking of pages that enabled Cambridge Analytica to target users with political campaign ads and potentially affect the 2016 presidential election (Grassegger and Krogerus 2017). Since this form of data is much newer in the social science toolkit, I will concentrate on this form of data throughout this chapter. However, many of the principles discussed apply equally well to administrative data and trace data.

The fact that we have millions of traces of individual-level data at a time resolution as precise as a tenth of a second or even higher is beautiful, but it is also a challenge. Never before in the history of social science have researchers had access to such a wealth of data, and as a result the academic community does not really know how to process such data. There have already been a few missteps. The most classic case might be Google Flu Trends, which attempted to predict the prevalence of flu based on examining hundreds of millions of Google searches (Cook et al. 2011). Flu Trends worked well for a while but then started wildly over- and underpredicting the prevalence of flu. The Google Flu Trends approach, and many other approaches to big data, uses an averaging approach of some sort, but the true beauty of big data is the richness of individual-level data. Traditionally social scientists have had to depend on samples or surveys of data (Bertrand and Mullainathan 2001), but the presence of big data means that we have a more complete picture of individual behavior than ever before. Though there is probably an optimal level of data resolution for any given problem, for many questions that are determined by individual-level heterogeneity, it would be useful to have a methodology that could automatically account for the differences between individuals as recorded in big data.

We want computer models that can be built or constructed from data. Many theory-driven models are never really compared to data. In other words, many models have been inspired and generated from theory or speculation, and therefore are somehow taken to be correct representations of the phenomenon at hand, but have never been shown to apply to data, or potentially have just been tested on some data set that also inspired the original theories, i.e. they have not been tested in an out-of-sample situation. Cathy O'Neil in her book, *Weapons of Math Destruction*, warns about the

reliance on models (O'Neil 2017), for instance, that make predictions about teacher performance based solely on theories about how teachers should add value to student learning. However, if these models have never been validated (Wilensky and Rand 2015), i.e. showing that they do in fact predict some inherent aspect of teacher performance, then these models are not valid tools of reasoning. As a community, we need to be able to create valid models from big data.

A modeling approach has been used in the social sciences that does give the researcher the ability to account for all of the individual-level heterogeneity that might be present. This approach is called agent-based modeling (ABM), and it is defined by creating a computational representation of every agent or individual (Wilensky and Rand 2015). This gives an agent-based model (ABM) the power to capture a rich variety of heterogeneity in the underlying system. However, historically many social science ABMs have been developed purely from theory, and not from data. Some of these theories have been grounded in pure speculation, based on casual observations and intuitions, and have often focused on one magical cause. Other theories have been driven by more in-depth thinking with at least some consideration of interactions. Still other theories have initially been grounded by examining data, but no serious effort has been made to test them outside of the original data. Instead, I would urge the development of data-driven ABMs. There have been cases of data-driven ABMs in the past in the social sciences (Axtell et al. 2002), but they are the exception rather than the rule.

What if it was possible to derive an ABM directly from data? Theoretically, such a system would not suffer from the problem of being dependent on averages, since the agents would manifest the rich heterogeneity of the underlying individuals. Moreover, if the system was trained using individual-level data, then it would not suffer the problems that arise when a system is based solely on averages. In order to achieve this goal, I will begin by discussing a framework that will guide the construction of big data, social science models that relies on the idea that a good model needs theory, and then building on this framework examine two methods for the creation of large-scale, data-driven ABMs. The first method is parameter optimization (PO), which involves taking an ABM and altering its parameters until the output of the model fits the data, and the second method is rule induction (RI), which involves inducing rules from big data that are directly used to create an ABM.

In order to temper expectations, it is necessary to realize that the vision presented in this chapter is not complete, and we do not yet have a large-scale, dynamically rich ABM that has been derived directly from data. However, the components are being put together, and progress is being made. The goal of this chapter is not to provide the perfect solution for big data, social science models, but rather to start to discuss how such models should be created and what some first steps toward their creation currently look like. Thus, this chapter is not the

end of a small research project by a small team of investigators, but rather the beginning of a community-wide project over years.

A Proposed Unifying Principle for Big Data and Social Science

Many of the problems of big data are often related to the fact that the computer and statistical big data models are not connected to theory, meaning that there is no explanation as to why the inputs are connected to the outputs beyond correlation. It is worth noting that in many disciplines the words theory and model are often considered equivalent words, but in some of the social sciences, especially the managerial sciences, there is a distinction that is often drawn between a theory, i.e. a causal explanation, and a model, i.e. a statistical or computational model (Shmueli 2010). I will use this distinction throughout this chapter. Thus, model in this chapter can be read as statistical or computational model relating inputs to outputs, and theory can be read as explanatory framework or causal theory, where such a framework could be anything from a simple one-variable causal hypothesis to a multivariate causal theory to a procedural description that explains behavior. Regardless of its format, the theory should provide some explanation of the phenomenon not just a black-box relationship between inputs and outputs.

Given this distinction the problem with many big data approaches is that without a theory we have not created a grounded model, but rather just a particular prediction of the future. For instance, with regards to Google Flu Trends, many arguments have been advanced as to why it stopped working, but it is hard to assess the cause of the inaccuracies because there was never a theory as to why it was working in the first place beyond the idea that somehow a user's searches on different subjects may be correlated with whether or not they have the flu. Without a theory it is hard to justify an explanation as to why something went wrong, since the theory provides the reason why it might work in the first place. With a theory, if something stops working, then it can be explored whether (i) the theory is wrong, i.e. it has been falsified, or (ii) the assumptions of the theory have been violated, i.e. the theory no longer applies.

This problem is particularly profound in the social sciences, where the basic elements of the systems, i.e. humans and human-created organizations, can change and adapt over time, meaning that we truly need to understand how the system works and not just what the end result of the current set of inputs will be. This does not mean that we need to necessarily know exactly what theory best fits the data before we start building models. Sometimes the goal of modeling is to generate a bunch of different models based on different theories and compare and contrast based on how well they fit the data or predict future outcomes. In fact, we may generate millions of different models, many of which

we are just going to discard based on fits to data, but if we have a few that actually match well with data, then we need the ability to explore those models from a theoretical perspective. A model that is constructed in such a way that it is essentially a black-box model, such as many of the deep learning models that are popular right now (LeCun et al. 2015), cannot ever really be compared to theory since there is no easy way to tell why the model is generating its results[1] (Yosinski et al. 2015).

Sometimes the distinction here is referred to as black box vs. white box. Black-box models just take inputs and give outputs, while white-box models allow inspection to determine why they are generating the outputs. Fundamentally, this means that black-box models are usually not very useful from a social science perspective, because it is hard to gather additional knowledge about human processes when it is not possible to explain how the inputs are related to the outputs. Moreover, a black-box model that may not be useful outside the data is trained on, because it is not possible to understand under what assumptions it works. This makes it difficult to determine that a black-box model that works now will also work next year, which seems to have been the case in the Google Flu Trends. This is somewhat related to the Lucas critique of macroeconomics (Lucas 1976), which states that no macro-model is useful for predicting what would happen as a result of a change of macroeconomic policy since the model was trained on data where that policy was not the case. In much the same way, the application of a black-box model is not useful if the goal is to understand how society will respond to a change in a structural element of society, since the model was not developed in a world where the structure had changed.

The result of all this is that we need social science models that are at least potentially amenable to a theoretical analysis. However, we also need theories that are useful as well. Though the focus of this chapter is on models and not theories, it is important to remember that a theory that does not lend itself to be interpreted from the perspective of a modeling framework is also problematic, since the theory is not testable. If the theory is not testable, then it has the same problems that a big data-driven correlational analysis does since there is no way to determine why the theory does not explain a particular set of data. Given these concerns, I propose the following principle for all theories and models that use big data in the social sciences.

Principle. *Theory-Interpretable Models and Model-Interpretable Theory (TIMMIT) Principle: All social science models using big data should be interpretable from a theoretical perspective, i.e. they should not be black box, and all social science theories applied to big data should be interpretable from a modeling perspective, i.e. they should be falsifiable.*

1 There are recent efforts to make deep learning neural nets more understandable. This agenda is sometimes referred to as explainable AI.

As mentioned, this does not require that the model be built from first principles, i.e. by hand, but it does require that it is potentially possible to inspect the model and then compare the model behavior to theory. Let us examine a quick example. Imagine that a researcher is trying to understand how information spreads on social media and they are examining Twitter. In particular, they want an individual-level model that predicts whether or not a focal user will retweet a particular piece of content. They could accomplish this by gathering a bunch of features about the user, including all of the time series of all the users that the focal users follow, and feeding this into a large wide and deep neural network. The resulting model might be very predictive, but if there is no way to understand why it is making the decisions, i.e. it is not theory interpretable, that is potentially problematic. On the other hand, the researcher could take the same inputs and create a Markovian-type model from the data. Though this model is not necessarily developed from theory, it is possible to inspect the model and potentially explore why the model is producing the outputs that it is by examining the states and how the user transitions through those states as time goes on. Even though there was no causal theory that was used to generate this model, the model was created using a framework that enables the researcher to explore each and every action that the agent in the model takes from a mechanistic point of view. These mechanisms can then be compared to theory, and the researcher can determine if the model makes sense from a causal theory point of view. In this particular example, it turns out that many traditional theories of information diffusion, such as the threshold (Granovetter and Soong 1983) and cascade theories (Goldenberg et al. 2001), can be written in this same modeling form. This means it may be possible to compare the model directly to theory and see how much the fully data-driven model gains you versus a potentially more restricted theory-based model. This is what I mean by a potentially theory-interpretable model.

Data-Driven Agent-Based Modeling

So now that we have a principle to guide our creation of social science models, generated from big data, the next question is how do we implement this principle. In the rest of this chapter, we will explore one particular solution that involves using machine learning to take an ABM and fit it to data. There may very well be other approaches that work and still meet the goals of the *TIMMIT* principle, but we will focus on this particular solution here. The ABM solution has a number of important benefits. First, ABM is flexible, i.e. it can incorporate many different modeling frameworks for the individual agents. Second, ABM is interdisciplinary, i.e. it applies equally well across the spectrum of the social sciences and has been used in just about all of them. Third, ABM potentially provides an answer to the Lucas critique. If we can understand low-level

decision rules that drive human behavior, then theoretically we should be able to explore how changes in their environment or policies affect their behavior. In macroeconomics, the Lucas critique resulted in a shift of research effort within that field to a better understanding of micro-foundations for the same reason. If you can understand the micro-principles, then even if you change the world, the model should adapt appropriately. This does not guarantee that the model will be correct, since it may still be the case that some hidden variables were not included in the model, but it is closer to building a correct model.

Within the space of machine learning and ABM, we will examine two different approaches to developing an ABM that fits to data and is theoretically interpretable: (i) PO and (ii) RI. PO starts with a theoretically grounded model and then modifies the parameters of the model until the output of the model matches some empirical data set (Bonabeau 2002). RI, on the other hand, attempts to induce rules of behavior directly and then asks whether the combined output of all of the agents acting together matches the empirical data. We will explore each of these approaches in turn.

Parameter Optimization

PO has been around since at least the first computer simulations, and even the use of machine learning to optimize parameters (Weinberg and Berkus 1971) has been studied for some time. The basic method for PO is that the researcher first constructs an ABM based on first principles and theory and then does their best to manually match the parameters to measured data about the world based on previous research, a process, which is sometimes called input validation (Rand and Rust 2011). However, it is usually the case that there is some uncertainty or unknown aspects to these parameters (Smith and Rand 2018). Given that, and if the goal of the simulation enterprise is to create a model that is descriptive and potentially predictive of the real world, one way to choose a final set of parameters is to compare the output of the model to an empirical data set and then tune the parameters until the model output and the real-world data are in close alignment.

In previous work, with Forrest Stonedahl, we described this more formally (Stonedahl and Rand 2014). We start by assuming a real-world data set, R, and a model, $M(P, E)$, with parameters, P, and environmental variables, E. The parameters, P, are the input values to the model that we do not know and are trying to optimize. Though we discuss this in the context of finding particular point values for P, it could also be that P is a set of distributions or even a choice of behaviors. The environmental variables, E, are the input variables that vary from particular context to particular context, and are not being optimized. For instance, they could be the input variables that control the geography the model is operating in. Given this notation, we can formalize PO in the following way. We split R into data sets: R_{train} that is the data set we

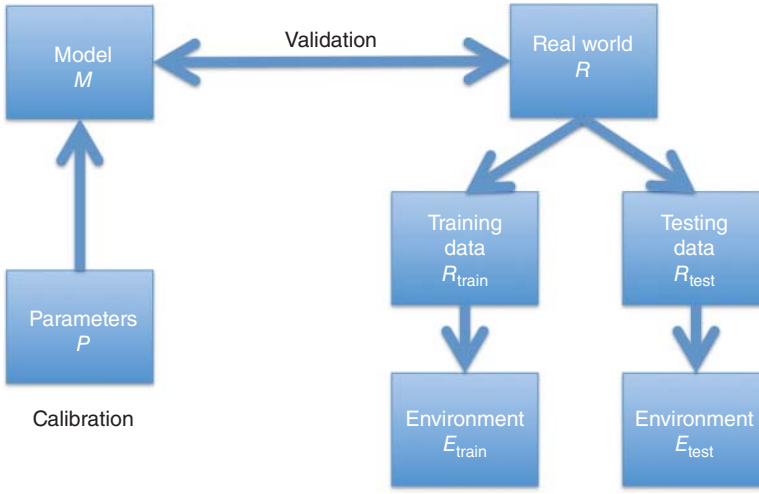

Figure 15.1 An illustration of the parameter optimization process.

will optimize the model with and R_{test} that is the data set that we are using to test the optimization. The environmental variables may differ between these two data sets, so we can denote them E_{train} and E_{test}. Given this notation, we can start the PO process by calibrating on the training data, which is simply a matter of identifying a set of parameter, P^*, such that some error measure $\epsilon(R_{train}, M(P^*, E_{train}))$.[2] This becomes a search problem to some extent, and any number of machine learning methods can be employed to optimize the system. Once the model has been calibrated, we can then determine if the model is valid by computing $\epsilon(R_{test}, M(P^*, E_{test}))$. If this value is less than some threshold T, then we can say that model has been validated. This approach is illustrated in Figure 15.1.

This provides a general framework by which an ABM that is driven by theory can be compared to and calibrated by big data. One way to use this framework would be to exactly find the best P values to match the underlying data as well as possible. However, it would also be possible to adjust what was within P and how ϵ was defined to instead explore both an estimate of the mean values and the uncertainty ranges around those values. Moreover, this framework enables the comparison of multiple models that have been generated from theory. For instance, if we have two models constructed from competing theories, say, M_1 and M_2, we can use the above procedure to calibrate each model to data and then examine the resulting parameter sets, P_1^* and P_2^*, as well as the minimal error that is achievable given the model and the error measure. If one of

2 In our original paper, we showed that the choice of ϵ is critical since the error measures directly affect the way in which the model is calibrated to data.

the models fits better to the data than another, then that can help adjudicate between the theories (Claeskens and Hjort 2008).

Alternatively, if both models fit the data fairly well, then the parameter sets can be investigated, and if it is determined that one of the parameter sets is more realistic than another, then that can also help to adjudicate between the models. It could also be the case that one of the parameter sets is very different from the other one, but not enough is known about the real world to adjudicate the distinction. This could be used as fodder for additional research into this area and helps isolate the distinction between the models.

Finally, it is possible that both models fit the data very well, and there is not much difference between the parameter sets. In this case, there are a number of possibilities: (i) the training and testing data is not sufficiently large enough to explore the differences in the model, and more data is needed; (ii) the environmental parameters might not be sufficiently varied to illustrate the differences between the models, and additional circumstances should be examined; or (iii) the models are both potentially good explanations of the underlying phenomenon, and it is necessary to reassess the differences between the models and see if the differences really are that significant.

This is an abstract explanation of how to use big data to drive the calibration and construction of an ABM. In the next two subsections, we will explore two specific examples. The first is in the context of examining a news consumption model, and the second is in the context of calibrating an ABM of urgent diffusion on social media.

News Consumption

This example is primarily drawn from previous work that developed the PO framework (Stonedahl and Rand 2014). The Internet is quickly and dramatically changing a number of different industries. One industry that has been particularly hard-hit by the rampant growth of the Internet is the news industry. However, news is considered more vital than ever to a well-functioning democracy. Well before the current #fakenews cycle, we investigated this phenomenon using an ABM. There is a desire by a number of individuals to figure out a way to remonetize news (Schmidt 2009), but to do so we need an understanding of what kinds of revenue platforms would work online. Fortunately, the takeover of the news industry by the Internet has resulted in huge, large, and at times nearly unmanageable trace of how individuals are consuming news. This data is eminently trackable, but researchers do not really have a good understanding of how people consume news online, which prevents the analysis of revenue models. In order to start to solve this problem, it would be useful to first construct an ABM that represents user behavior.

The goal of this project was therefore to create a model that matches the data patterns as close as possible. In this case the data was represented by clickstream data, i.e. for a panel set of the same users we could observe over time,

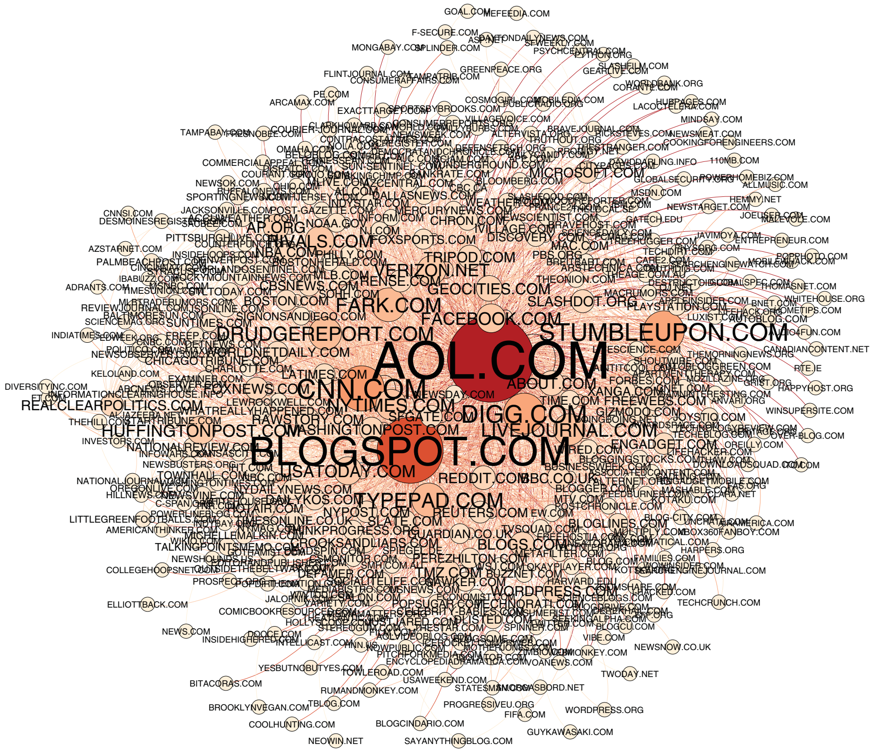

Figure 15.2 The eventual news network generated from the underlying data. The size of the node indicates the amount of traffic that node received. Image Credit: Forrest Stonedahl.

which links the user clicked on. This data contained 90 000 users during the entire year of 2007. In order to organize the data in a proper framework, we took one month of data (January) to use for training and one month of data (December) to use for testing. We also filtered the data to only contain clicks on news-oriented websites, e.g. cnn.com and nytimes.com. We also recorded all of the incoming traffic to these websites. We created a network by linking nodes that individuals had clicked from and to. This resulted in a network of 422 nodes in January and 417 nodes in December. The network is visualized in Figure 15.2.

We then built an ABM in NetLogo (Wilensky 1999) where we constructed a representative average Internet surfer who started at a random node in the network and then would decide which node to click on or to end its Internet session. The hypothesis is that users decide to click on different nodes based on their reputation and that the reputation of the node can be proxied by various network properties of the node. Therefore, the way the user decides which node to click on was decided based on a utility function where the

utility of a node was determined by (i) randomness – a pure random element, (ii) in-degree – number of nodes that directly link to this node, (iii) out-degree – number of nodes that this node directly links to, (iv) in-component – number of nodes that can reach this node, (v) out-component – number of nodes reachable from this node, (vi) PageRank score (Page et al. 1999) – the original Google ranking algorithm, (vii) Hits-Hubs score (Kleinberg 1999) – a measure of how much this node serves as a hub, (viii) Hits-Authorities score (Kleinberg 1999) – a measure of how much this node serves as an authority, (xi) clustering – how clustered the connected nodes are, (x) betweenness – the betweenness centrality of the node, and (xi) eigenvector centrality – the eigenvector centrality measure of the node. In this case the answer to this question would provide us with a potential explanation of the underlying user choices observed even if it did not provide us with a causal theory.

The weights that were associated with each of these 11 measures were the variables of concern. In addition to the weights of these 11 utility components, there were two additional parameters. The first parameter, *random-restart*, controlled how often the model restarted, i.e. it was a probability that the user restarted their search at any node. The second parameter, *no-backtrack*, was a Boolean flag that controlled whether the user was allowed to go back to the node they had just visited. Essentially, the model was run with a set of 13 different parameters for each of these variables for a number of user visits equivalent to the number of user visits observed in the actual data, and then the number of clicks on each website node was recorded. This distribution of website traffic was then compared with the actual website traffic using a number of different error measures, including correlation and a number of different L-norms, which included a right or wrong distance measure, Euclidean distance, Manhattan distance, and the L^∞-norm (Stonedahl and Rand 2014). The model was trained on the January data, and then the results were evaluated on the December data.

This can be placed into the PO framework. The real-world data set R is the whole set of clickstream data. The model, $M(P, E)$, is the ABM, where P is the set of 11 parameters describing the agent's utility function and the two additional parameters and E is the number of clicks observed in the corresponding set of data; this varies slightly between the January and December data. R_{train} is the January data set, and R_{test} is the December data set. The error measure, ϵ, is the corresponding set of measures, corr, L^0, L^1, L^2, and L^∞. The model was then calibrated on the January data (R_{train}) using the BehaviorSearch package in NetLogo (Stonedahl and Wilensky 2010) using a genetic algorithm (Holland 1975).

In this discussion rather than exploring the actual model implications for news consumption, since those results were largely inconclusive, let us concentrate on the methodological results, which directly bear on this chapter's goal of exploring the utility of PO. One central question was whether one of the error functions that was used to train the model would be superior to the other

error functions. For instance, would using the correlation-based error measure also optimize the L^1 (Manhattan) error measure or some other error measure? Three clear findings stood out. First, all of the error measures achieved similar performance on the testing data for the L^0 (the number of matching elements) error measure, which means if this measure is a goal, the choice of how you train the model is largely irrelevant. Second, the correlation error measure was pretty good at optimizing its own measures on the testing data set, and did okay on the other error measures, but none of the other error measures did a very good job of achieving high performance on the correlation measure in the testing data. Finally, the L^2 (Euclidean) measure actually achieved high performance on the L^1, L^2, and L^∞ measures, meaning that it is a potentially robust error measure that should be considered more in the future, but none of the measures was strictly dominant.

Urgent Diffusion

The study of news consumption essentially resulted in a finding that was methodologically relevant, but did not have much practitioner value. In a different study, we focused more on the practical applications of calibrating an ABM to data (Yoo et al. 2016). In this study, we examined the role of social media in spreading information about disasters. An area of information diffusion is called urgent diffusion (Rand et al. 2015), since unlike traditional information diffusion, there is a time-critical nature to the diffusion process. The fundamental question of interest was: how can humanitarian organizations use social media to spread information better in urgent diffusion scenarios, such as those that occur during disasters?

We did this by first modeling the process of diffusion using a standard implementation of an agent-based information diffusion model, namely, the independent cascade model (Goldenberg et al. 2001). This model has two parameters, an internal influence parameter q (governing how quickly information spreads within the social media platform) and an external influence parameter p (governing how external sources of information influence the spread of information). We also constructed a social media network that was based on the longest observed cascade (i.e. series of retweets) observed in the actual data. Once we had constructed this model, we then calibrated the model to a number of different diffusions that had been observed in social media data. However, unlike the news consumption model, we were not interested in creating a predictive model, but rather in creating the best descriptive model that we could from the data. As a result, we did not separate into training and testing data, but instead merely fit the model as best as possible to all of the data. This is because we were actually going to use inferred values from the model as the dependent variable for another model, and so validation of the predictive capability of this model was not relevant (Shmueli 2010).

As a result, though the purpose is slightly different, we can still place this model in the PO framework. The real-world data set R is the set of different diffusion patterns on Twitter, essentially being a set of time series where each point of time was the number of new people who had tweeted about the disaster. The model, $M(P, E)$, is the ABM, where P is the two parameters p and q, describing the agent's influence function, and E was the observed social media network. The error measure, ϵ, was mean average precision error, or MAPE, which was a measure of how different the observed time series was from the model time series. The model was then calibrated using all of the data in R using the BehaviorSearch package in NetLogo (Stonedahl and Wilensky 2010) using a simulated annealing algorithm (Kirkpatrick et al. 1983).

In the end, the goal with this chapter was to identify the p and the q values that were most likely have given rise to the observed diffusion patterns. Once these were identified the ratio of the two values $\frac{q}{p}$ was used to express how much social media affected the diffusion. We could then regress properties of the diffusion process to identify factors that led to high rates of social media diffusion. Our results indicated that diffusion events were more likely to spread quickly on social media if (i) they were started by influential individuals, (ii) they were posted earlier in the overall timeline of the disaster, and (iii) the original information was posted repeatedly over time. These are all actionable items from the perspective of a humanitarian organization. However, interestingly we also found that fake news was more likely to spread on social media and that diffusion events containing content that promoted situational awareness (i.e. information about what is going on in the environment) did not significantly affect the spread of news.

Rule Induction

RI is not as old an idea as PO, but it has been explored in the past. The basic idea is that rather than trying to construct the rules that an agent will follow from scratch, the modeler tries to infer rules of behavior directly from data. To some extent this could be used as a large-scale form of PO. Where PO attempts to alter and tweak the parameters of a system to get it to fit better, the goal of RI is to identify the best rule that fits the data. However, there is a significant qualitative difference. The focus in most PO is altering micro-level, meso-level, or even macro-level parameters of the system to get macro-level model outputs that correspond well with actual data. However, in the RI context the goal is to start with empirical data at the micro level and then develop the best agent rules that match that real-world data. In other words, PO starts with altering micro-parameters to generate macro-patterns that match macro-level data. RI works by starting with micro-level data and then generating rules that create macro-level patterns. As a result, both the types of data (micro vs. macro) are

different, and the rules/parameters being generated are also often at different levels of modeling complexity.

Commuting Patterns

One example of RI that came about almost 10 years ago was work by Lu et al. (2008) to investigate the use of public transportation by commuters in the Chicago area. In this context, they were interested in investigating how different public policies would affect the decision by commuters to use one particular mode of transit over another. In particular, is it possible to reverse the trend toward car dependence in a large city?

In order to carry out this investigation, they created a synthetic population based on statistics of Chicago's actual population. All agents then had to make two decisions: (i) transit mode choice, i.e. do they use public transit or drive a private car?, and (ii) are they happy with their current residential location, or should they move? For the purposes of this chapter, we will focus on the first decision, because they used a machine learning approach known as (class association rules) CAR to derive the rules of behavior for agents with regard to transit mode choice. They attempted to identify one of five modes of transportation that an agent might use to get to work: (i) walking, (ii) driving, (iii) passenger in a private car, (iv) transit that had to be driven to, and (v) transit that did not have to be driven to. The CAR then used a number of different features to make this decision: household income, household size, number of vehicles in the household, age, gender, employment status, total travel time, access time to public transit, number of public transit transfers required, distance from work, and whether work was located in the central business district. All of this data was derived from actual data about real commuters that were part of the Chicago Area Transportation Survey. This resulted in 404 rules that were then embedded into each agent. The overall structure of this approach is illustrated in Figure 15.3.

It should be noted that this process is almost a mini-calibration exercise in that the rules of behavior were calibrated to real-life behavior, and in fact a training and testing approach was used in order to determine the best rules to use.

The interesting result of this approach is that agents can now make decisions contingent upon their situation. For instance, in this work, Lu et al. took the fully calibrated model and then asked what would happen if you changed the distribution of residential locations and the location of transit options. Since the agent rules were calibrated at the agent level and not at the aggregate level, it was possible to observe how changes in the underlying population and urban form affected agent decisions without having to actually build the city where those patterns existed, creating the possibility for a policy flight simulator (Holland 1992; Sterman 2001). The results showed that though mode choice is

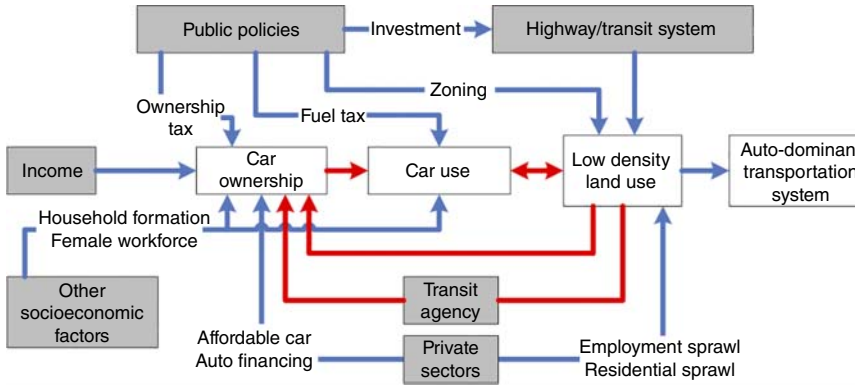

Figure 15.3 A description of the framework in the Lu model. Source: Courtesy of Yandan Lu.

sticky, i.e. people do not switch often, it is possible to design a city that has a high transit share.

Social Media Activity

In the commuting mode choice example, the decision rules were determined for a large population and then embedded contingently in each agent, and the rules for every agent were essentially the same, but the actions differed based on the context the agents found themselves in, i.e. location, number of children, income, etc. However, sometimes this level of heterogeneity is not enough to adequately describe an individual's behavior. In some cases, it may be necessary to create not just different actions for each agent to take, but also different rules of behavior that govern those actions. In these cases it may make sense to learn the behavior of every individual directly from the observable data. This approach would require a large amount of data for every individual and so may not be practical in many cases. For instance, in the commuting mode choice example, it would take a long time to accumulate enough data to determine how each individual chooses different modes of commuting and, in some cases, may not even be practical since you cannot easily alter the income an individual has and then observe how that affects their behavior.

However, in other circumstances it may well be that it is possible to get large amounts of data about the way an individual makes decisions in a limited context. In fact, in the case of the explosion of big data, that is often what we have. We have large-scale trace of individuals using apps, devices connected to the Internet of things, social media, and websites. In these situations we can build models of the individual's interaction with these platforms at the individual level, and then we can embed these individual-level models in an ABM and observe the interactions between them.

We did this recently (Darmon et al. 2013; Harada et al. 2015; Ariyaratne 2016) using a technique known as causal state modeling (CSM). For the purposes of this discussion, the CSM approach, also called computational mechanics or ϵ-machine approach, creates a hidden Markov model-like representation of a time series that is minimally complex and maximally predictive, i.e. has the fewest number of states necessary to predict the time series as well as possible (Shalizi and Crutchfield 2001). In each of the chapters where we used this approach, we took Twitter data and inferred a model of behavior for the individual users of Twitter. The complexity of these models varied substantially. In one case, 12.8% of the inferred models had one state, 58.8% had two states, 4.4% had three states, 3.3% had four states, and 20.7% had more than four states (Darmon et al. 2013). This indicates that the heterogeneity of the underlying population was substantial. The four most common structures that we observed are illustrated in Figure 15.4.

Once we built these models, we were able to predict whether or not a user would tweet in the near future fairly accurately based on these models. We compared the use of CSMs with echo state networks, a recurrent neural network architecture, and found that the model performed fairly similarly at predicting the behavior of an individual (Darmon et al. 2013). Moreover, we

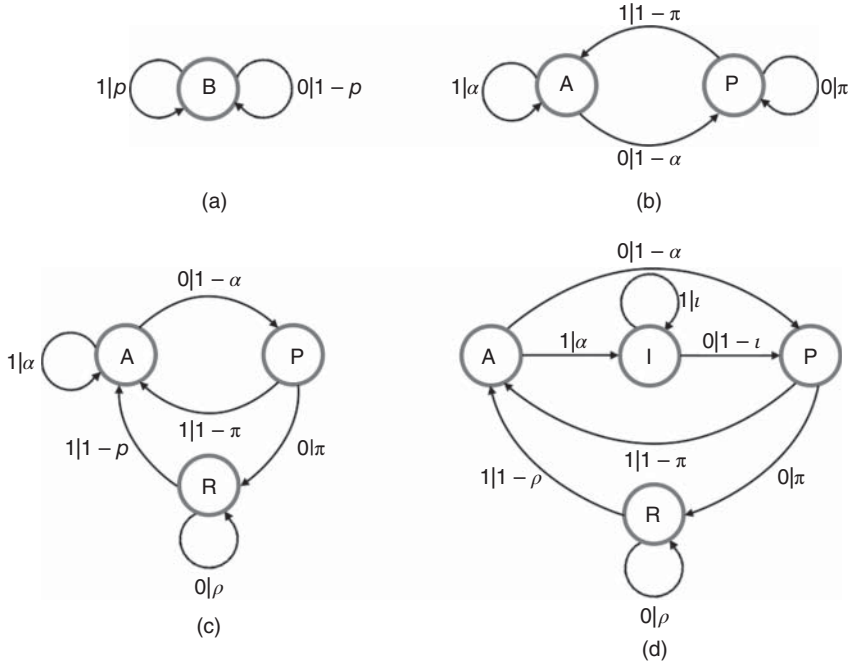

Figure 15.4 An illustration of the four most common CSM structures in the Twitter data set, ranging from the simplest (a) to the most complex (d). Source: Reproduced with permission of David Darmon.

showed that when the behavior of these models was aggregated across large groups of individuals, we were also able to accurately predict the aggregate patterns of behavior of a group of Twitter users. In fact, we performed as well as an aggregate-level autoregressive moving average model and outperformed a pure seasonality model. This last result is interesting, because our model was never trained to perform well at the aggregate level, but rather it was trained at the individual level, and the aggregate-level results are a by-product of the agent-based approach (Harada et al. 2015).

One of the reasons why we have chosen to use CSMs as opposed to neural network architectures or some other data mining approaches is that causal state models are potentially interpretable. By that we mean that once the CSM is learned, a practitioner could interpret the states of the model and label them. This enables comparison to social science theory about how users will behave. In one case it was pretty clear that over half of the users of Twitter had an active and passive state, which corresponds to the idea that users are either actively engaged with Twitter or have stepped away from their phone or the app. Moreover, we can also retroactively generate explanations for why the model makes the prediction that it does, which is contrary to a black-box model where it is difficult to interrogate the model. This transparency of the model gives it a quality that is necessary when pursuing social science explanations and not just curve fitting.

However, there is still work to do in this space. So far the models we have constructed that have been successful have been stand-alone agent models where there was minimal to no interaction between the agents. When we attempted to construct a model where there was social interaction between the agents, we found that the system performed adequately for a few minutes but then stopped predicting the overall state of the system accurately (Ariyaratne 2016). We feel that this is because over long periods of time, seasonality effects have a large role in interactions, and the framework that we constructed did not account for this. Nonetheless, we feel that the CSM approach can be adapted to take a clock as an input that would account for seasonality.

Conclusion and the Vision

ABM and large-scale individual-level data are a great match. The data can provide the insight we need to generate individual-level models of human behavior. This should never be an exercise in curve fitting, since if it is, the individual-level models will probably do no better than the aggregate models. Instead, big data and ABMs should be combined under the TIMMIT principle, and the models should be theory interpretable, and the theories should be model interpretable. This enables the creation of models that are constantly being validated by comparing them with real-world data. If done well then these models that are specified at the level of the individual and not at the

level of the aggregate pattern of data may well be able to overcome the Lucas critique (Lucas 1976), since the models of the individual specify an individual's beliefs, goals, and actions, and therefore these models should be able to adapt to changes in the macro-level policies governing individual actions. Thus, the promise of data-driven ABM provides one of the few approaches that may actually work for examining individual responses to new incentive structures imposed by macroscopic policy changes.

In this chapter, I have presented two basic concepts of how this could be done. One approach involves taking a model completely developed from theory and then calibrating it to fit real-world data, i.e. PO. The other approach involves taking a set of data and inferring rules of behavior for that data, often informed by theory development and compared with theoretical models, i.e. RI. I have highlighted two examples of each approach that seem to work very well, but there are many other ways that these processes can be carried out using additional machine learning or statistical inference approaches. Many of them have been explored in previous work (Zhang et al. 2016).

One vision for this work would be the eventual creation of an ABM directly from data. Especially, in the context of the RI approach, it might be possible to specify some basic features with relation to the basic form of an ABM and then give a set of data to a computational data processing pipeline and have that pipeline automatically spit out a fully realized ABM in a common modeling language, such as NetLogo (Wilensky 1999) or Repast (North et al. 2006). Moreover, this model could be updated on a continuous basis without much additional effort by continuing to feed new data into it, creating a dynamically created realistic model of a complex system.

For instance, imagine an ABM of a city or urban landscape that was constantly fed data from social media, e.g. Twitter, administrative data feeds, e.g. police feeds, and satellite or aerial imagery. This model might adequately represent the focal city at enough level of detail that it could then be used as a policy flight simulator (Holland 1992; Sterman 2001) and used to assess the effect of street closures, new crime policies, and changes in the zoning laws. In fact, if carried to the extreme, and developed at the right resolution, such a model might enable detailed predictions about the future state of our fictional urban center. This could lead to the creation of a new science, such as Asimov's psychohistory, where history, psychology, and mathematical statistics (or in this case ABM) are combined to make predictions about how large populations will act in the future (Asimov 1951).

Acknowledgments

I would like to thank all of my coauthors and the researchers who worked on the projects showcased above, especially, Forrest Stonedahl, Uri Wilensky, Eunae

Yoo, Jeffrey Herrmann, Brandon Schein, Neza Vodopivec, Yandan Lu, Moira Zellner, Kazuya Kawamura, David Darmon, Jared Sylvester, and Michelle Girvan. Most of these colleagues also gave invaluable feedback on early versions of this chapter. I also want to thank Paul Davis and Angela O'Mahoney for making this opportunity possible and for providing advice on how to clarify this chapter.

References

Ariyaratne, A. (2016). Modeling agent behavior through past actions: simulating twitter users. Master's thesis. College Park, MD: University of Maryland.

Axtell, R.L., Epstein, J.M., Dean, J.S. et al. (2002). Population growth and collapse in a multiagent model of the Kayenta Anasazi in long house valley. *Proceedings of the National Academy of Sciences of the United States of America* 99 (Suppl. 3): 7275–7279.

Backstrom, L., Sun, E., and Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In: *Proceedings of the 19th International Conference on World Wide Web*, 61–70. ACM.

Bertrand, M. and Mullainathan, S. (2001). Do people mean what they say? Implications for subjective survey data. *American Economic Review* 91 (2): 67–72.

Bonabeau, E. (2002). Agent-based modeling: methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America* 99 (Suppl. 3): 7280–7287.

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*. Cambridge Books.

Cook, S., Conrad, C., Fowlkes, A.L., and Mohebbi, M.H. (2011). Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS ONE* 6 (8): e23610.

Darmon, D., Sylvester, J., Girvan, M., and Rand, W. (2013). Predictability of user behavior in social media: bottom-up v. top-down modeling. In: *2013 International Conference on Social Computing (SocialCom)*, 102–107. IEEE.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. *ICWSM* 13: 1–10.

Eagle, N. and Pentland, A.S. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* 10 (4): 255–268.

Golbeck, J. and Hansen, D. (2011). Computing political preference among twitter followers. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1105–1108. ACM.

Goldenberg, J., Libai, B., and Muller, E. (2001). Talk of the network: a complex systems look at the underlying process of word-of-mouth. *Marketing Letters* 12 (3): 211–223.

Granovetter, M. and Soong, R. (1983). Threshold models of diffusion and collective behavior. *Journal of Mathematical Sociology* 9 (3): 165–179.

Grassegger, H. and Krogerus, M. (2017). The data that turned the world upside down. Vice Magazine (30 January). https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win.

Harada, J., Darmon, D., Girvan, M., and Rand, W. (2015). Forecasting high tide: predicting times of elevated activity in online social media. In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 504–507. ACM.

Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Application to Biology, Control, and Artificial Intelligence*, 439–444. Ann Arbor, MI: University of Michigan Press.

Holland, J.H. (1992). Complex adaptive systems. *Daedalus* 121 (1): 17–30.

Asimov, I. (1951). *Foundation*. New York: Gnome Press.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science* 220 (4598): 671–680.

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal* 79 (1): 1–14.

Kleinberg, J.M. (1999). Hubs, authorities, and communities. *ACM Computing Surveys (CSUR)* 31 (4es): 5.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553): 436.

Lu, Y., Kawamura, K., and Zellner, M.L. (2008). Exploring the influence of urban form on work travel behavior with agent-based modeling. *Transportation Research Record: Journal of the Transportation Research Board* 2082 (1): 132–140.

Lucas, R.E. Jr. (1976). Econometric policy evaluation: a critique. In: *Carnegie-Rochester Conference Series on Public Policy*, vol. 1, 19–46. Elsevier.

Matheus, R., Vaz, J.C., and Ribeiro, M.M. (2014). Open government data and the data usage for improvement of public services in the Rio de Janeiro city. In: *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*, 338–341. ACM.

North, M.J., Collier, N.T., and Vos, J.R. (2006). Experiences creating three implementations of the repast agent modeling toolkit. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 16 (1): 1–25.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The Pagerank Citation Ranking: Bringing Order to the Web. Tech. Rep. 422. Stanford InfoLab.

Rand, W. and Rust, R.T. (2011). Agent-based modeling in marketing: guidelines for rigor. *International Journal of Research in Marketing* 28 (3): 181–193.

Rand, W., Herrmann, J., Schein, B., and Vodopivec, N. (2015). An agent-based model of urgent diffusion in social media. *Journal of Artificial Societies and Social Simulation* 18 (2): 1.

Riggins, F.J. and Wamba, S.F. (2015). Research directions on the adoption, usage, and impact of the internet of things through the use of big data analytics. In: *2015 48th Hawaii International Conference on System Sciences (HICSS)*, 1531–1540. IEEE.

Schmidt, E. (2009). How Google can help newspapers. The Wall Street Journal (1 December). https://www.wsj.com/articles/SB10001424052748704107104574569570797550520.

Shalizi, C.R. and Crutchfield, J.P. (2001). Computational mechanics: pattern and prediction, structure and simplicity. *Journal of Statistical Physics* 104 (3–4): 817–879.

Shmueli, G. (2010). To explain or to predict? *Statistical Science* 25 (3): 289–310.

Smith, E.B. and Rand, W. (2018). Simulating macro-level effects from micro-level observations. *Management Science*. In press.

Sterman, J.D. (2001). System dynamics modeling: tools for learning in a complex world. *California Management Review* 43 (4): 8–25.

Stonedahl, F. and Rand, W. (2014). When does simulated data match real data? In: *Advances in Computational Social Science*, 297–313. Springer.

Stonedahl, F. and Wilensky, U. (2010). Behaviorsearch [computer software]. In: *Center for Connected Learning and Computer Based Modeling*. Evanston, IL: Northwestern University. http://www.behaviorsearch.org.

Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives. OECD Working Papers on Public Governance, (22):0_1.

Watts, D.J. (2013). Computational social science: exciting progress and future directions. *The Bridge on Frontiers of Engineering* 43 (4): 5–10.

Weinberg, R. and Berkus, M. (1971). Computer simulation of a living cell: Part I. *International Journal of Bio-Medical Computing* 2 (2): 95–120.

Wilensky, U. (1999). *NetLogo: Center for Connected Learning and Computer-Based Modeling*. Evanston, IL: Northwestern University.

Wilensky, U. and Rand, W. (2015). *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. MIT Press.

Yoo, E., Rand, W., Eftekhar, M., and Rabinovich, E. (2016). Evaluating information diffusion speed and its determinants in social media networks during humanitarian crises. *Journal of Operations Management* 45, 123–133.

Yosinski, J., Clune, J., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. In: *ICML Workshop on Deep Learning*.

Zhang, H., Vorobeychik, Y., Letchford, J. and Lakkaraju, K. (2016). Data-driven agent-based modeling, with application to rooftop solar adoption. *Autonomous Agents and Multi-Agent Systems* 30 (6): 1023–1049.