BIG DATA, AGENTS, AND MACHINE LEARNING: TOWARDS A DATA-DRIVEN AGENT-BASED MODELING APPROACH

Hamdi Kavak
Modeling Simulation Visualization Eng. Dept.
Old Dominion University
1300 Engineering & Computational Sciences Bldg.
Norfolk, VA, USA
hkava001@odu.edu

Jose J. Padilla
Virginia Modeling Analysis and Simulation Center
Old Dominion University
1030 University Boulevard
Suffolk, VA, USA
jpadilla@odu.edu

Christopher J. Lynch
Virginia Modeling Analysis and Simulation Center
Old Dominion University
1030 University Boulevard
Suffolk, VA, USA
cjlynch@odu.edu

Saikou Y. Diallo
Virginia Modeling Analysis and Simulation Center
Old Dominion University
1030 University Boulevard
Suffolk, VA, USA
sdiallo@odu.edu

ABSTRACT

We have recently witnessed the proliferation of large-scale behavioral data that can be used to empirically develop agent-based models (ABMs). Despite this opportunity, the literature has neglected to offer a structured agent-based modeling approach to produce agents or its parts directly from data. In this paper, we present initial steps towards an agent-based modeling approach that focuses on individual-level data to generate agent behavioral rules and initialize agent attribute values. We present a structured way to integrate Big Data and machine learning techniques at the individual agent-level. We also describe a conceptual use-case study of an urban mobility simulation driven by millions of geo-tagged Twitter social media messages. We believe our approach will advance the-state-of-the-art in developing empirical ABMs and conducting their validation. Further work is needed to assess data suitability, to compare with other approaches, to standardize data collection, and to serve all these features in near-real time.

Keywords: agent-based simulation, data-driven modeling, big data, machine learning.

1 INTRODUCTION

Agent-based modeling has been increasingly used in the past two decades to model systems based on their constituent units that interact with and within an environment (Bonabeau 2002). This bottom-up representation in ABMs makes it possible to have one-to-one ontological correspondence between the model and the real-world (Gilbert 2008) and capture parallel processes, non-linear interactions, and heterogeneity (Bruch and Atwell 2015) of social systems which are challenging to capture with other modeling approaches (Gilbert 1999). Ideally, modelers develop ABMs to answer specific research questions that help identify the level of detail and realism needed for the model (Bruch and Atwell 2015). However, this process is not always straightforward.

SpringSim-ANSS, 2018 April 15-18, Baltimore, MD, USA; ©2018 Society for Modeling and Simulation (SCS) International

For cases in which the research question implies a highly accurate prediction mechanism, the model needs to incorporate realistically represented agent behaviors and an agent environment. This situation is particularly relevant for models that are developed to guide policy-making (e.g., urban models that encompass policy interventions). These type of models are often named *empirical models* to reflect the basis of how agents are modeled (Gilbert 2008).

Alternatively, the research question may entail the investigation of existing theories or the creation of new ones. For these cases, ABMs are often conceptualized at highly abstract levels with the goal to find mechanisms that generate particular aggregate-level patterns. Tipping point models such as Schelling's racial segregation model is one of the prime examples in this category (Schelling 1971). These models are often called *abstract models* (Gilbert 2008). Especially a significant number of ABMs developed in the 90s and the early 2000s fall into this category.

While fully empirical and fully abstract models represent two extremes, there are a significant and increasing number of models in-between the two. Models closer to the abstract side can still be used to explore the implications of empirical research or formal theories. These models can have some dimensions empirically grounded while others remain abstract (Bruch and Atwell 2015). Models closer to the empirical side are often more generic versions of empirical models aiming to create a typical model of real-world systems (Gilbert 2008).

Notwithstanding the need for empirical grounding in models that seek complete or partial realism, agent-based modeling approaches incorporating data-driven practices are not common. This may result from a lack of behavioral data availability (Kennedy 2012) or a lack of quality data at suitable abstraction levels (Bruch and Atwell 2015). However, even when that is not the case, researchers often develop their models without the systematic inclusion of empirical data. To date, only a handful of studies have scratched the surface of this issue and developed methodological approaches to incorporate data in modeling agents. Yet, these approaches integrate data obtained through qualitative methods like surveys which are small-scale, costly to conduct, and need to be measured periodically (Bruch and Atwell 2015, Padilla et al. 2014).

Modelers can make use of newly established data sources as the world shifts from a data-poor era into a new data-rich era (Helbing and Balietti 2011). Communication technologies have become ubiquitous in the latest decades and inspire new ways to collect big data from a variety of sources. Big data means that the data is provided with high velocity (i.e., near real-time) and exhaustively covers the entire population (Kitchin and McArdle 2016). Ranging from cell phone traces to social media, big data carries a variety of signals that reflect human actions and preferences. When made available for ABMs, big data can beneficially reveal hidden insights and behavioral patterns of humans, can cover a large percentage of a population, can be collected longitudinally at relatively low-cost (Padilla et al. 2014).

The purpose of this paper is to provide the initial steps towards creating a new agent-based modeling approach that incorporates big data at the individual-level to generate agent behavior rules and initialize agent attributes. Section 2 presents a review of agent-based modeling approaches that support a data-driven model development and synthesizes a gap that these studies do not address. Section 3 proposes a data-driven modeling approach to bridge the gap and Section 4 provides a conceptual use-case on how the approach could be used. Finally, Section 5 wraps up the paper with conclusions and future work.

2 AGENT-BASED MODELING AND DATA

In this section, we delve into the data-driven agent-based modeling realm. Note that data-driven means that the approach should prioritize the use of empirical data in all modeling steps in the context of ABMs (Collado 2009). To provide a comparative review of data-driven approaches, we use the ABMs in the following subsection to develop a characterization of data usage. The subsequent subsection reviews data-driven modeling approaches that provide modelers guidance for designing simulation models. Lastly, we examine big data studies that have the potential to provide insights into data-driven ABMs.

2.1 Characterizing Data Usage in ABMs

We characterize data usages in ABMs with respect to four categories that covers both the data-focused properties and model focused implications. Each category is summarized below.

- ***Data type:*** There are two main types of data used in ABMs: *qualitative* and *quantitative*. Qualitative data contains information represented using text/words and does not allow undistorted conversion of information (Yang and Gilbert 2008). Interviews, focus groups, and ethnographic studies are among well-known methodologies for eliciting qualitative data (Smajgl et al. 2011). Quantitative data contains information that can be represented as a number, such as a person's age. This data type can be further classified as *discrete* (e.g., integers) or *continuous* (e.g., π=3.1415..). Official public datasets (e.g., the US Census) contain a wealth of both qualitative and quantitative data about populations. These data types are both used across theoretical and empirical ABMs. The appropriate type is dependent on the specific case being modeled and the availability and accessibility of the data.

- ***Repeated measurements:*** Whether qualitative or quantitative, it is important to consider the repetition of the measurements (data collection) as this determines if it is possible to capture temporal changes in empirical data. We identify *one time* (a snapshot) and *repeated* (multiple snapshots) as measurement repetition types. Snapshot data may be challenging to use in behavioral ABMs, especially representing human behaviors, since behaviors vary and change over time. This type of data could be better utilized for initializing ABMs. On the other hand, repeated data collection indicates multiple measurements at different times which can be broken down as *longitudinal* (same subjects measured) and *cross-sectional* (different subjects measured). Using repeated data collection makes it possible to capture individual-level and population-level behavioral patterns over time.

- ***Impact on model:*** The impact of data on the model shows the changes caused dependent upon how the data is used. We identify three types of impact: *initialization*, *structure modifications*, and *structure generation*. For *initialization*, the model stays the same but the initial values of agent attributes or environment variables change. For *structure modification*, a trackable modification occurs to the model, such as modifying a formula for assigning yearly agent income using data fit to a linear regression model. For *structure generation*, a complete or partial model structure is created from scratch. Thus, the change cannot be tracked. Data's impact on the model increases greatly from initialization to structure generation.

- ***Agent consideration:*** Agent consideration refers to how the data applies to the agents within an ABM, either at the *population-level* or the *individual-level*. Population-level considerations usually involve aggregate-level data, such as agent types, probability distributions, or percentages describing the value of an agent property or the likelihood of events/actions. These quantities are assigned or distributed the same across the agent population. Individual-level considerations provide longitudinal data fed into each agent uniquely. Data used in each agent is based on a pre-specified set of steps or algorithms. Population-level considerations are common due to the availability of aggregate level data (Bruch and Atwell 2015); however, communication technology advances and the resulting data generated are making it increasingly plausible to incorporate individual-level considerations in modeling agents.

2.2 Data-Driven Agent-Based Modeling Approaches

Despite the fact that the use of data in ABMs has been common, the literature reveals few general-purpose data-driven agent-based modeling approaches. In fact, the majority of approaches reviewed here describe their data-driven modeling approach based on a particular use-case while the approach is generalizable to other problem areas.

The first data-driven methodological effort we identify goes back to a 2007 study by Kennedy et al. (2007) who developed a simulation assistance system called Adaptive Intelligent Model-building for the Social Sciences (AIMSS). In AIMSS, an ABM is developed as a classical rule-based theoretical model while there is a simultaneous data collection effort from the real-world. Once a simulation model is developed and simulated, the real-world data is processed to gather high-level patterns seen in the data. These patterns are then compared to the simulation output. If these quantities are off by a margin, agent behavioral rules, attributes, and initial values are reconfigured to match the real-world patterns. The description reads like an advanced calibration effort where initial values and agent model is modified. Nevertheless, the data in their approach does not directly impact the model and offers insufficient insight into data-driven modeling. Another challenge is that it is unclear how to update the model or its parameters when the simulation's results do not match with the aggregate level data descriptions.

Collado (2009) proposes a broad data-driven agent-based modeling approach and emphasizes using data to improve empirical grounding wherever possible. The method suggests gathering data from the real-world system and using it to guide the abstraction process (e.g., mathematical equations) and to initialize the model parameters. Separate data captured from the real-world is used for operational validation of the simulation. Data sources considered in Collado (2009)'s approach include surveys, panels, interviews, and official documents involving both quantitative and qualitative data. The use-case uses a panel survey that has longitudinal measurements. The impact of data in this approach is not substantial and is reduced to the initialization of values (as statistical distributions) and the modification of existing model structures. Finally, while longitudinal data is used, the primary agent consideration of the model is at the population-level.

Smajgl et al. (2011) present an agent-based modeling approach that focuses on empirical characterization of agent behaviors and attributes. While the main premise is very similar to Collado (2009), Smajgl et al. (2011)'s case provides specific methodologies that can be used in modeling steps. Smajgl et al. (2011) embrace the use of qualitative and quantitative data sources including expert knowledge, participant observation, surveys, interviews, census data, field/lab experiments, and role-playing games. However, there is no mention of measure repetition. Regarding the impact on the model, the approach covers initialization and the structure modification with a population-level focus.

Sajjad et al. (2016) offer a data-driven agent-based modeling approach for modeling family formation that uses three decades of Korean census data (collected in 1990, 2000, and 2010). The approach uses the 2010 census data to initialize certain parameters (as probabilities and ranges) in the population model. The evolution of family formation-related output of the simulation is tracked over time and the results are compared with the consequent census data (2000 and 2010). Unlike the previous two studies, this approach only uses data for simulation initialization and validation.

Jensen and Chappin (2017) propose an algorithmic approach for modeling ABMs using data. This approach takes initial agent attribute values, network connections between agents, a model properties list, and expected patterns to be developed by the model. Then, it selects a decision model from a repository of models and checks the suitability of model outputs concerning expected patterns. The model selection process continues until the desired outputs are gathered – similar to AIMSS (Kennedy et al. 2007). The improvement in this approach is in the detailed and structured way to update the models to match the desired outputs and is showcased with several examples. As the focus of this approach is on automation, it does not offer novelty for using data in ABMs. The use of data type, measurement repetition, its impact on the model is not discussed while the agent consideration is at the population-level.

We compare all data-driven modeling approaches in **Error! Reference source not found.** concerning the characterization from Section 2.1. The majority of the reviewed studies focus on survey data and utilize repeated measurements. When it comes to impact on the model, a majority provided simple initialization cases while only two cases use data to modify structures (e.g., equations). Collectively, however, agent consideration only occur at the population-level and none consider the generation of model structures directly from data. To sum up, we found no approach that generates a model structure from data considering an individual-level representation of agents. In the next subsection, we show examples from big data analytics studies that provide insights into individual-level agent consideration.

Table 1. Summary of the reviewed data-driven modeling approaches (N/R: not reported).

Data type		**Repeated measurement**	**Impact on model**			**Agent consideration**		**Reference**
Qual.	Quant.		Init.	Structure mod.	Structure gen.	Pop.	Ind.	
Yes	Yes	No	No	No	No	Yes	No	Kennedy et al. (2007)
Yes	No	Yes	Yes	Yes	No	Yes	No	Collado (2009)
Yes	Yes	N/R	Yes	Yes	No	Yes	No	Smajgl et al. (2011)
Yes	No	Yes	Yes	No	No	Yes	No	Sajjad et al. (2016)
N/R	N/R	N/R	Yes	N/R	N/R	Yes	No	Jensen and Chappin (2017)

2.3 Big Data and Its Implications for Agents

Big Data analytics deals with the process of gathering useful insights from large-scale and often real-time data sources to develop an understanding of real-world phenomena and help make better decisions (Ankam 2016). In particular, the field of human dynamics offers a wealth of studies with the potential to deliver insights into agent-based modeling. We are interested in studies analyzing human characteristics and behavior at the individual-level which manifest in ABMs as agent attributes and behavioral rules.

In terms of human characteristics, the literature reveals examples that aim to identify demographics and personality attributes of people from Big Data. The examples below predominantly use social media data from Twitter. The data is cleaned and organized before being used to determine human characteristics and a computational algorithm is proposed to extract the attribute values.

Mislove et al. (2011) capture hometown, gender, and race of Twitter users. Hometown is based on peoples' self-reported location property within their profiles checked using a commercial geo-coding API. Gender is identified based on Twitter users' first names checked against a name-gender database. A similar approach is used for race identification. Chen et al. (2010) examine personality traits from social media text. They focus on the personality traits of openness and neuroticism derived from social media message history. The method uses the extraction of lexicon-based features. The study's focus is on individuals' potential interaction with advertisements, but the idea of deriving personality traits from a text is shown to be attainable. This work is later extended to capture personality values (Hsieh et al. 2014). Ryoo and Moon (2014) investigate Twitter users' home locations. They model the geographic distribution of the words used in a geographically explicit social media content. In this respect, they use the center of a word based on its overall shares in different tweets and calculate its dispersion. There are other approaches that estimate the home location of users with accuracies ranging from hundreds of kilometers (Mahmud, Nichols, and Drews 2014) to tens of meters (Kavak, Vernon-Bido, and Padilla 2018).

Regarding human behavioral rules, a wealth of human mobility studies use call detail records (CDR) and social media entries. These studies commonly collect large-scale location footprints of people and clean them to identify a statistically suitable subset for understanding human mobility. Then, they use a variety of techniques (e.g., statistical modeling, network analysis) to identify movement patterns.

González, Hidalgo, and Barabási (2008) develop a statistical model to understand human movements in an area using a large scale (100k users) CDR dataset. The model uses individuals' movement distances between locations as a proxy to characterize people's typical movement distance (i.e., the radius of gyration) and investigates how this distance is distributed among a population. This model provides one of the first comprehensive studies of human mobility using large-scale granular data. Schneider et al. (2013) propose a more granular model that investigates both CDRs and mobility surveys to capture daily movement patterns. Unlike González, Hidalgo, and Barabási (2008), this study captures movement as sequential moves between places (nodes) called daily motifs. This information forms a directed graph with explicitly identified movement sequences. They show that 90% of movement patterns can be explained using 17 unique motifs. While not building an ABM, they attempt to create an individual-based model to represent the movement of a person. Jiang, Ferreira, and González (2015) aim to create a transportation planner using the daily motifs idea of (Schneider et al. 2013). The methodology describes all steps starting with preparing the CDR data and ending with trip generation. Motifs are captured as activities and are extrapolated to cover the population. Lastly, Malleson and Birkin (2012) propose the idea of using publicly available geo-tagged Tweets to understand human movement behavior within a city. The approach aims to estimate people's home locations and to understand the transitions between home and other locations. While a decent initial attempt, few practical examples are given to demonstrate the feasibility of the idea.

3 TOWARDS A DATA-DRIVEN MODELING APPROACH

As highlighted, the absence of a methodological approach at the individual agent-level to create behavioral rules and to initialize attributes from data produces a gap. We follow a top-down style focused on the data flow to create and describe the initial design of our data-driven approach. To this end, we first describe the main elements of our approach at the contextual-level in Section 3.1. We then go one-level down in Section 3.2 and explain some of the high-level details of the data-driven agent generation system. Finally, inner workings of the individual-level agent creation are described in Section 3.3.

3.1 Contextual View

At the contextual-level, our data-driven approach is composed of four elements depicted in Figure 1. *Conceptual model* describes the details of a model in some form like text or diagrams (McKenzie 2010). A *data source* is a system/repository that generates/contains data to be used for agents. *Data-driven agent generation system* is the core of the approach that takes the data and the conceptual model and creates data-driven agents. *Simulation engine* runs pre-specified scenarios using the generated agents.

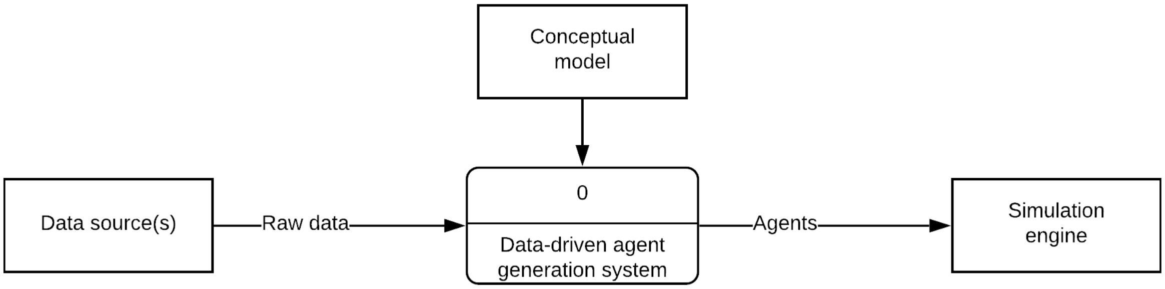

Figure 1: Contextual Data Flow Diagram of the proposed data-driven approach.

We assume that the modeler has a conceptual model of the system to be studied, including at least: a *purpose statement* describing the goal of the model; the *agent types* with their *attribute* and *behavior signatures*; and the *environment type* and *variables*. The conceptual model here is similar to what one can abstract for a standard ABM (Gilbert 2008).

Our approach uses the conceptual model (mainly agent attribute and behavior signatures) in the identification and selection of data sources suitable to be used in generating behavioral rules and initializing agent attributes. Example data sources include online repositories, on-demand streams (e.g., APIs), and data generated by physical devices or software. For both behaviors and attributes, the data source has to provide the necessary information at the granularity of an individual agent. Typically, behavior rule generation requires historical data to include the real-world entity's behavioral actions (e.g., interactions with other entities over time). Attribute initialization data could have direct values of attributes like in Kavak et al. (2017) or may provide the information needed to infer the values through a process. Having a conceptual model and data sources, the data-driven agent generation system uses them to create the agents semi-automatically. We provide a high-level view of this process in the next subsection.

3.2 High-Level View

Figure 2 describes the data-driven agent generation system and its data flow by representing the inter-relationships between the components. The system contains three main processes namely *data preparation*, *attribute model training/fit*, and *agent behavior creation and attribute initialization*. This subsection describes the first two components; the next one focuses exclusively on the third.

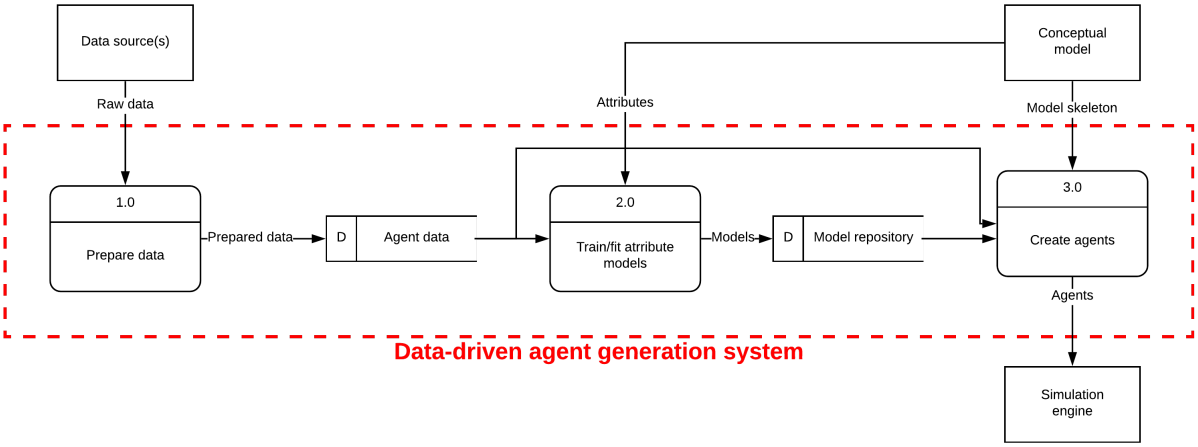

Figure 2: Level 1 Data Flow Diagram of the agent generation system.

The data preparation process covers the data lifecycle from data acquisition to its use in the later steps of the approach. During data acquisition, computer software consumes data from the selected data sources. Depending on the data source (e.g., online repository, streams, etc.), data acquisition may be as simple as downloading files from an online location or as complex as authenticating to an online service and continuously downloading live data. Continuous downloads are mostly present when dealing with Big Data. The preprocessing step makes use of the KDD Process, a well-respected approach for inferring novel patterns from data (Fayyad, Piatetsky-Shapiro, and Smyth 1996). The primary operations here involve removing noisy entries, handling missing values, transforming data fields, and removing unusable entries for behavior creation and attribute value inference. Lastly, the agent data is stored for use in further steps.

The attribute training/fit process is an optional step used when at least one agent attribute needs to be initialized based its dispersion among the real-world population. The process requires having a sample of the targeted attribute value gathered from the population (ground truth) using additional data collection. Then, this ground truth data is used to train a machine learning model or to fit a statistical model. Then this trained/fit model is captured in a repository to be used in attribute initialization.

3.3 Individual Agent-Level View

Depicted in Figure 3, the agent creation process is the last significant step in our approach to creating data-driven agents. The initial process is the filtering of the agent data at the individual-level and then separating it into *attribute data* and *behavioral data*. From this point on, the process continues along different paths where the attribute data is used in initializing attributes whereas the behavioral data is used in creating behavioral rules.

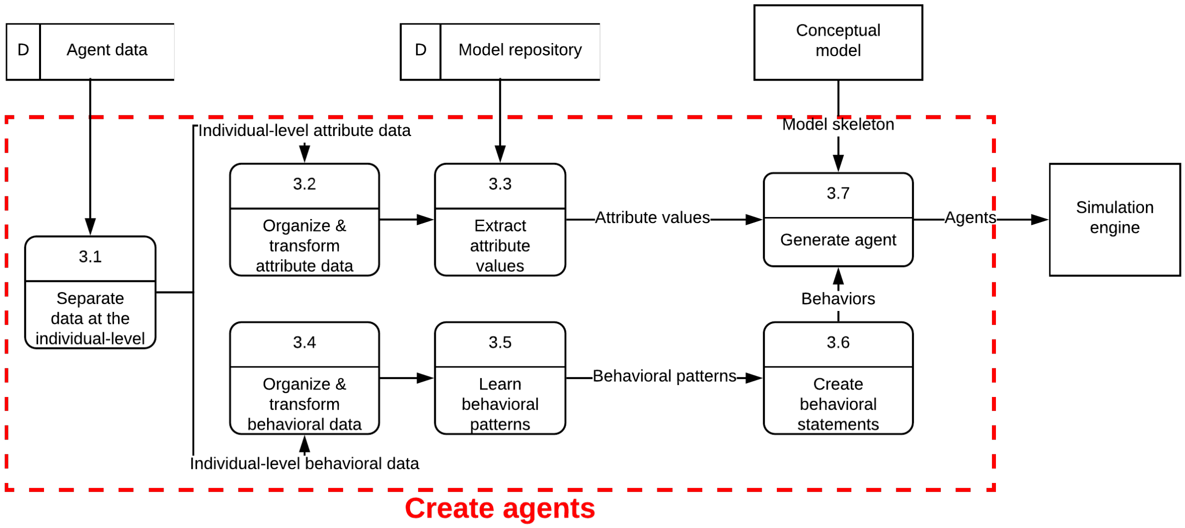
D
Agent data
D
Model repository
Conceptual model
Model skeleton
Individual-level attribute data
3.1
Separate data at the individual-level
3.2
Organize & transform attribute data
3.3
Extract attribute values
Attribute values
3.7
Generate agent
Agents
Simulation engine
Behaviors
3.4
Organize & transform behavioral data
3.5
Learn behavioral patterns
Behavioral patterns
3.6
Create behavioral statements
Individual-level behavioral data
Create agents

Figure 3: Level 2 Data Flow Diagram of the agent generation system focused on agent behavioral rule creation and attribute initialization.

In the attribute initialization case, processes 3.2 and 3.3 cover three different cases. *Direct initialization* passes values from both processes without modification (e.g., agent age). *Look-up-based initialization* transforms attribute data into specific values in process 3.2 using a look-up table (e.g., 'age' transforms to 'age range') and that value is not changed in process 3.3. *Model-based initialization* organizes and transforms data in process 3.2 into an input that is suitable to be used in the attribute value inference model (machine learning or statistical). This process is different from earlier works like Padilla et al. (2016) who focused on agent population-level parameters.

In the behavioral rule generation case, individual-level behavioral data provides signals that represent human actions (e.g., purchase). In the first step (process 3.4), behavioral data is organized and transformed in a way that each action and related parameters are captured as a single record. These records are then used in training a machine learning model (process 3.5). In the training, the expected output of the model is given as the action parameter whereas other parameters are provided as input. In this way, behavioral patterns are captured through a trained machine learning model. This model is then encapsulated as a function and turned into agent programming language statements (process 3.6) that represent generated agent behavioral rules. We note that performance of the machine learning model needs to be evaluated through cross-validation.

In the last step (process 3.7), an actual agent program is created where the model skeleton identified at the conceptual model is used as the blueprint of the agent. Attributes are initialized with the values come from process 3.3. For behaviors, there is a likelihood to have theoretical behavioral statements that come from the conceptual model as well as data-driven behavioral statements that come from process 3.6. The modeler codifies these statements according to the precedence provided in the conceptual model. While trained machine learning models of data-driven behavior rules can be automatically loaded, the combination process is considered to be manually coded at this moment. It is important to note that the processes starting with the separation of data to agent generation is repeated for every single agent. We describe our approach regarding data flow elements to show the lifecycle of data and let the modeler choose a particular implementation solution fit their needs.

4 USE CASE: A CONCEPTUAL HUMAN MOBILITY SIMULATION MODEL

In this section, we describe how the proposed approach can be applied to an urban simulation use-case. Urban areas have densely situated population and numerous sensory inputs (e.g., traffic sensors, cell phone reception, transportation ticketing systems) that can provide individual-level data for use in agent-based simulations. Here, we focus on human mobility as a particular type of urban simulation. The main elements of the use-case are described below while noting that the use-case is still being developed.

4.1 Conceptual Model

The purpose of this model is to simulate the movement of people (mobility) in urban areas. Human mobility is a significant research area with implications on daily life from sitting in the traffic to the spread of infectious diseases (González, Hidalgo, and Barabási 2008). The ultimate goal is to extend the model later to tackle these daily life problems.

Figure 4 depicts a conceptual model representing the agent skeleton as a class diagram and a visualization of how an agent can move. *HumanAgent* is an entity situated in a geographical map environment and has static (value stays the same) and dynamic attributes (value changes during simulation). Static attributes are *name*, *gender*, and *home location*. Dynamic attributes are *current location* and *visited locations*. All these attributes are initialized at the beginning of a simulation run. The agent is set to an initial *current location* and it moves to other locations based on the movement patterns captured in the movement data.

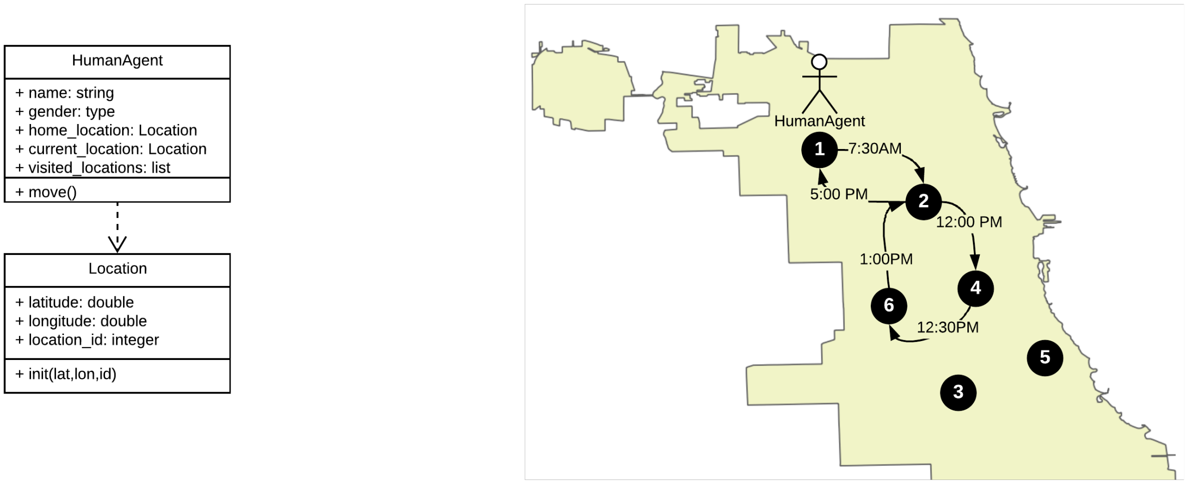

Figure 4: A part of the conceptual model. The class diagram on the left side shows basic attribute and behavior skeleton of the *HumanAgent*. The map on the right side conceptualizes the movement of the agent.

4.2 Data and Preprocessing

To model the movement of people in the simulation, we capture geo-tagged social media entries from Twitter using Twitter's Streaming API by identifying a bounding box (as latitude-longitude pairs) that covers the conterminous U.S. We collected this data between May-16-2014 and April-27=2015 (343 days – 4 days missing due to interruptions) with the total records of 826,021,868 messages. The Twitter social media platform includes regular users, but it is also a popular platform for companies and automated software (bots). Our preprocessing step performed a cleaning process to remove non-human originating messages identified by the number of messages shared on a daily basis and percentage of links appear in those messages. Preprocessing reduced the data to 716,553,502 records from 6,375,210 users. To keep our model at a reasonable size, we focused on the use-case city – Chicago, Illinois, USA. When we focus on users with active social media usage with more than five geo-located messages, our final dataset resulted with ≈7.78 million messages from 92,296 users.

4.3 Agent Generation

We follow the process described in the proposed approach and focus on attribute initialization and behavior rule generation for reporting purposes. For the attribute initialization, we prepare the Twitter data to initialize an agent's name, gender, and home location.

We gather the most recent tweet of the agent, initialize the name attribute by directly using the Twitter name, and initialize gender using a name-gender look-up table following Mislove et al. (2011). For the home location attribute, Twitter does not provide this information, so we infer it using the location visit tweet history of the person. From the entire dataset, we identify users through a keyword search in their Twitter messages that are potentially related to 'being at home' and create a crowdsourcing platform that lets us label whether those expressions meant being home. In this way, we identify a subset of users whose home location is known with confidence. We then create an SVM-based home location classifier trained several mobility features that capture a person's visit characteristics. After cross-validation, we confirm that our classifier performs up to ≈0.87 accuracy for predicting people's home location. Details are described in (Kavak, Vernon-Bido, and Padilla 2018).

To capture the movement rules, we organize individual-level data with location footprints sorted by time ascending. We then use DBSCAN (Ester et al. 1996) algorithm to mitigate GPS inaccuracies by clustering visits that are made to closely located places. Next, we create an hourly visit list history for the agent. Our next step, which is still ongoing, is to test different machine learning algorithms (e.g., Recurrent Neural Networks, Markov Chains, and Decision Trees) to capture the movement dynamic and report on the cross-validation performances.

5 DISCUSSION AND CONCLUSION

In this paper, we argue that the literature lacks a structured modeling approach to generate agent behavioral rules and initialize agent attributes from data at the individual-level. We review the data-driven agent-based modeling literature to pinpoint this gap and show the reasonability of the idea from the Big Data Analytics literature. We propose a data-driven agent-based modeling approach for agents that bridges the mentioned gap. From a top-down perspective, we describe our approach in detail and provide a conceptual use-case application of human mobility using publicly available social media footprints.

We believe this is the first structured approach that focuses on individual-level data-driven agents. While the goal is to have all agent parts driven by data, our experience shows that, based on the modeling problem, the modeler will likely need theory-driven parts of the model. To allow for this, we include a conceptual model as an input to the agent generation system. One of the unique advantages of the proposed individual-level data-driven modeling is to have one extra level of validation at the individual level which can be assigned to an accuracy number based on cross-validation score.

As the next step, we will complete the movement behavioral rule creation in the use-case model and report our results. There remain numerous aspects of this approach still to be addressed, including (1) development of data suitability assessment metrics; (2) experimental design to compare the performance of our approach with other agent-based modeling approaches; (3) development of standardized data collection and use mechanisms using ontologies; and (4) extension of the work to be used in near-real time Big Data.

ACKNOWLEDGMENTS AND DISCLAIMER

This material is based on research sponsored by the Office of the Assistant Secretary of Defense for Research and Engineering (OASD(R&E)) under agreement number FAB750-15-2-0120. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Office of the Assistant Secretary of Defense for Research and Engineering (OASD(R&E)) or the U.S. Government.

REFERENCES

Ankam, V. 2016. *Big Data Analytics*: Packt Publishing.

Bonabeau, Eric. 2002. "Agent-based modeling: Methods and techniques for simulating human systems." *Proceedings of the National Academy of Sciences* 99:7280-7287. doi: 10.1073/pnas.082080899.

Bruch, Elizabeth, and Jon Atwell. 2015. "Agent-Based Models in Empirical Social Research." *Sociological methods & research* 44 (2):186-221. doi: 10.1177/0049124113506405.

Chen, Jilin, Eben Haber, Ruogu Kang, Gary Hsieh, and Jalal Mahmud. 2010. "Making Use of Derived Personality : The Case of Social Media Ad Targeting."

Collado, Samer Hassan. 2009. "Towards a Data-driven Approach for Agent-Based Modelling: Simulating Spanish Postmodernisation."

Ester, Martin, Hans-Peter Krieger, Jörg Sander, and Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." Portland, Oregon, 1996.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. "From data mining to knowledge discovery in databases." *AI magazine*:37-54. doi: 10.1145/240455.240463.

Gilbert, Nigel. 1999. "Simulation A new way of doing science." *American Behavioral Scientist* 42 (10):1485-1487.

Gilbert, Nigel. 2008. *Agent-Based Models*. Vol. 153. Thousand Oaks, CA, USA: SAGE Publications.

González, Marta C., César A. Hidalgo, and Albert-László Barabási. 2008. "Understanding individual human mobility patterns." *Nature* 453 (7196):779-782. doi: 10.1038/nature06958.

Helbing, Dirk, and Stefano Balietti. 2011. How to Do Agent-Based Simulations in the Future: From Modeling Social Mechanisms to Emergent Phenomena and Interactive Systems Design. Santa Fe Institute.

Hsieh, Gary, Jilin Chen, Jalal Mahmud, and Jeffrey Nichols. 2014. "You Read What You Value : Understanding Personal Values and Reading Interests." *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* April:983-986. doi: 10.1145/2556288.2557201.

Jensen, Thorben, and Émile J. L. Chappin. 2017. "Automating agent-based modeling: Data-driven generation and application of innovation diffusion models." *Environmental Modelling & Software* 92 (Supplement C):261-268. doi: https://doi.org/10.1016/j.envsoft.2017.02.018.

Jiang, Shan, Joseph Ferreira, and Marta C. González. 2015. "Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data : A Case Study of Singapore." *ACM KDD UrbComp'15*.

Kavak, Hamdi, Jose J. Padilla, Daniele Vernon-Bido, Ross J. Gore, and Saikou Y. Diallo. 2017. "The Spread of Wi-Fi Router Malware Revisited." Spring Simulation Multi-Conference, Virginia Beach, VA, USA, 2017.

Kavak, Hamdi, Daniele Vernon-Bido, and Jose Padilla. 2018. "Fine-Scale Prediction of People's Home Location using Social Media Footprints." SBP-BRIMS, Washington, D.C., USA, In Press.

Kennedy, Catriona, Georgios Theodoropoulos, Volker Sorge, Edward Ferrari, Peter Lee, and Chris Skelcher. 2007. "AIMSS: An architecture for data driven simulations in the social sciences." 2007.

Kennedy, William G. 2012. "Modelling Human Behaviour in Agent-Based Models." In, edited by A. J. Heppenstall, A. T. Crooks, L. M. See and M. Batty, 167-179. Springer.

Kitchin, R., and G. McArdle. 2016. "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets." *Big Data & Society* 3 (1):1-10. doi: 10.1177/2053951716631130.

Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews. 2014. "Home Location Identification of Twitter Users." *arXiv preprint arXiv:1403.2345* xx (xx). doi: 10.1145/2528548.

Malleson, Nick, and Mark H. Birkin. 2012. "Estimating Individual Behaviour from Massive Social Data for An Urban Agent-Based Model." *Geosimulation: Modeling Social Phenomena in Spatial Context* (September):23-29.

McKenzie, Frederic D. 2010. "Systems modeling: analysis and operations research." In, 147-180. John Wiley & Sons, Inc.

Mislove, Alan, Sune Lehmann, Yong-yeol Ahn, Jukka-pekka Onnela, and J. Niels Rosenquist. 2011. "Understanding the Demographics of Twitter Users." *Artificial Intelligence*:554-557.

Padilla, Jose J., Saikou Y. Diallo, Hamdi Kavak, Olcay Sahin, and Brit Nicholson. 2014. "Leveraging Social Media Data in Agent-based Simulations." Tampa, FL, 2014.

Padilla, Jose J., Saikou Y. Diallo, Hamdi Kavak, Olcay Sahin, John S. Sokolowski, and Ross J. Gore. 2016. "Semi-automated initialization of simulations: an application to healthcare." *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 13 (2):171-182. doi: 10.1177/1548512914565503.

Ryoo, Kyoungmin, and Sue Moon. 2014. "Inferring Twitter user locations with 10 km accuracy." *Proceedings of the companion publication of the 23rd international conference on World Wide Web*:643-648. doi: 10.1145/2567948.2579236.

Sajjad, Mazhar, Karandeep Singh, Euihyun Paik, and Chang-Won Ahn. 2016. "A Data-Driven Approach for Agent-Based Modeling: Simulating the Dynamics of Family Formation." *Journal of Artificial Societies and Social Simulation* 19 (1):9. doi: 10.18564/jasss.2988.

Schelling, Thomas C. 1971. "Dynamic models of segregation." *The Journal of Mathematical Sociology* 1 (2):143-186. doi: 10.1080/0022250X.1971.9989794.

Schneider, Christian M., Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C. González. 2013. "Unravelling daily human mobility motifs." *Journal of the Royal Society, Interface / the Royal Society* 10 (84):20130246-20130246. doi: 10.1098/rsif.2013.0246.

Smajgl, Alex, Daniel G. Brown, Diego Valbuena, and Marco G. A. Huigen. 2011. "Empirical characterisation of agent behaviours in socio-ecological systems." *Environmental Modelling & Software* 26 (7):837-844. doi: 10.1016/j.envsoft.2011.02.011.

Yang, L. U., and Nigel Gilbert. 2008. "Getting Away From Numbers: Using Qualitative Observation for Agent-Based Modeling." *Advances in Complex Systems* 11 (02):175-185. doi: 10.1142/S0219525908001556.

AUTHOR BIOGRAPHIES

HAMDI KAVAK is a Ph.D. candidate in Modeling and Simulation (M&S) at Old Dominion University (ODU) and a graduate research assistant at Virginia Modeling Analysis and Simulation Center (VMASC). He received his MS in M&S from ODU. His research focuses on data-driven agent-based simulations with special emphasis on human mobility and cyber security. His email address is hkava001@odu.edu and his website address is http://www.hamdikavak.com.

JOSE J. PADILLA is Research Assistant Professor at VMASC at ODU and leads the Simulate to Create research group. His research focuses on the modeling of problem situations; methodological development for M&S of human behavior using social media data and agents; simulating users, insider threats, and hackers for cybersecurity; and creating platforms for simulation development and data capture. His email address is jpadilla@odu.edu and his web page is https://www.odu.edu/directory/people/j/jpadilla.

CHRISTOPHER J. LYNCH is a Senior Project Scientist at VMASC. He is currently pursuing a Ph.D. in M&S from ODU where he also received his MS in M&S in 2012 and a BS in Electrical Engineering in 2011. His research interests include multi-paradigm modeling and verification of simulation models. His email address is cjlynch@odu.edu and his web page is http://www.odu.edu/directory/people/c/clync008.

SAIKOU Y. DIALLO is Research Associate Professor at VMASC and adjunct Professor of Modeling, Simulation, and Visualization Engineering at ODU. He received his MS and Ph.D. degrees in Modeling and Simulation from ODU. His email address is sdiallo@odu.edu and his web page is https://www.odu.edu/directory/people/s/sdiallo.