

# A primer on data-driven modeling of complex social systems

Alexandria Volkening

**ABSTRACT.** Traffic jams on roadways, echo chambers on social media, crowds of moving pedestrians, and opinion dynamics during elections are all complex social systems. These applications may seem disparate, but some of the questions that they motivate are similar from a mathematical perspective. Across these examples, researchers seek to uncover how individual agents—whether drivers, Twitter accounts, pedestrians, or voters—are interacting. By better understanding these interactions, mathematical modelers can make predictions about the group-level features that will emerge when agents alter their behavior. In this tutorial, which is based on the lecture that I gave at the 2021 American Mathematical Society Short Course, I introduce some of the terms, methods, and choices that arise when building such data-driven models. I discuss the differences between models that are statistical or mathematical, static or dynamic, spatial or non-spatial, discrete or continuous, and phenomenological or mechanistic. For concreteness, I also describe models of two complex systems, election dynamics and pedestrian-crowd movement, in more detail. With a conceptual approach, I broadly highlight some of the challenges that arise when building and calibrating models, choosing complexity, and working with quantitative and qualitative data.

*A complex system might be defined as a system  
for which no single model is appropriate.*

and

*As Picasso said of art, a good model  
“is a lie that helps us see the truth.”*

(Lee A. Segel and Leah Edelstein-Keshet [SEK13])

---

*Key words and phrases.* Complex social systems, complex systems, mathematical modeling, data-driven modeling, election forecasting, pedestrian movement.

In putting together my lecture for the 2021 AMS Short Course, I acknowledge Heather Zinn Brooks, Jonathan Desponds, Simon Freedman, Brian Hsu, Kara Maki, Niall Mangan, and Bridget Torsey for helpful examples in their earlier talks or pointers to references. Special thanks to Jeffrey Humpherys, Rachel Levy, and Thomas Witelski for their minitutorial [HLW16] on modeling courses at the 2016 SIAM Annual Meeting; I drew from their minitutorial for several of the concepts in Section 3. Thanks to my Short Course co-organizers Heather, Mason, and Michelle for their encouragement and for being a terrific team, and to Mason for introducing me to the term “data-driven modeling of complex systems” in the first place.

## 1. Introduction

Traffic jams on roads [SFK<sup>+</sup>08, SCDM<sup>+</sup>18, BD11, JHZ<sup>+</sup>14, BHN<sup>+</sup>95], pedestrian crowds [BCD18, HM95], swarming locusts [AA15, BCME<sup>+</sup>20], animal aggregations [DAB<sup>+</sup>20, CKJ<sup>+</sup>02, PEK99, LLEK10, BCC<sup>+</sup>08, KTI<sup>+</sup>11], collections of cells [BEK20, Vol20b, GBKM20, GG93], and echo chambers [SCP<sup>+</sup>21, EF18, CDFMG<sup>+</sup>21, CFPSS19] are examples of complex systems. In each of these cases, rich, group-level dynamics emerge from the interactions of smaller components—e.g., drivers, people, locusts, animals, or cells—with one another and with their environment [DBC<sup>+</sup>19, Bro22]. The interdisciplinary field of *complex systems* [New11] centers on the questions that arise from these emergent dynamics. Complementing experimental approaches to complex systems, mathematicians develop methods in dynamical systems, topology, network science, numerical analysis, probability, partial differential equations, and many other areas. Here I focus on data-driven mathematical modeling, mainly for complex social systems. My goal for this tutorial chapter is to help provide a starting point for folks who are new to this area, and I reflect on some broad questions and choices that emerge when combining models and data.

Figure 1 highlights several complex social systems, ranging from traffic flow [SFK<sup>+</sup>08, BD11, NS92] to Brexit voter dynamics [SHP16]; I also recommend the supplementary material of [SSS17, SFK<sup>+</sup>08] and the websites [Loc, PMS17] for related animations. Across these applications, one interesting feature is the common challenges that they raise from a modeling perspective. For example, in each of the images in Figure 1, a researcher may want to characterize alignment. This can be physical alignment, with pedestrians, locusts, or drivers adjusting how they move in response to other individuals or obstacles in their environment. A different type of alignment is present in Figure 1(d)–(e): people are forming opinions and may be influenced to align with (or against) the beliefs of others. Another thread in complex systems is heterogeneity [MP07]: each person, animal, or social-media account in Figure 1 is unique. Guided by the data available, each modeler must choose how much detail to include. Should we model voters as having a binary opinion (e.g., “for Brexit” or “against Brexit”) or allow opinions to live on a spectrum? Changes in behavior are also present: for example, in evacuation conditions, an emotional contagion can propagate through a crowd, changing how pedestrians act [BDM<sup>+</sup>09, BHK<sup>+</sup>11, TFB<sup>+</sup>11, BRSW15].

Higher-order interactions are widespread in complex systems: peer influence and social reinforcement from multiple friends may cause someone to change their opinion or adopt a new technology, when an isolated or pairwise interaction might not [OT12, BR06, IPBL19, GBC18, Sch73]. In a related vein, the presence of short- and long-range interactions in complex systems leads to rich dynamics. In Figure 1(c), drivers are interacting locally, basing their acceleration on the cars near them. The addition of autonomous vehicles allows for long-range dynamics. Stern *et al.* [SCDM<sup>+</sup>18] have shown that judiciously modulating the speed of one autonomous vehicle can result in the disruption of phantom traffic jams and improved fuel usage in some experiments. (Phantom traffic jams are jams that appear to emerge from drivers, rather than through external forces [JHZ<sup>+</sup>14, SCDM<sup>+</sup>18].)

Modeling complex social systems stems from and leads to questions that are of societal and mathematical interest. From an applied perspective, in the case of traffic flow, we might want to shed light on what driver behaviors cause jams



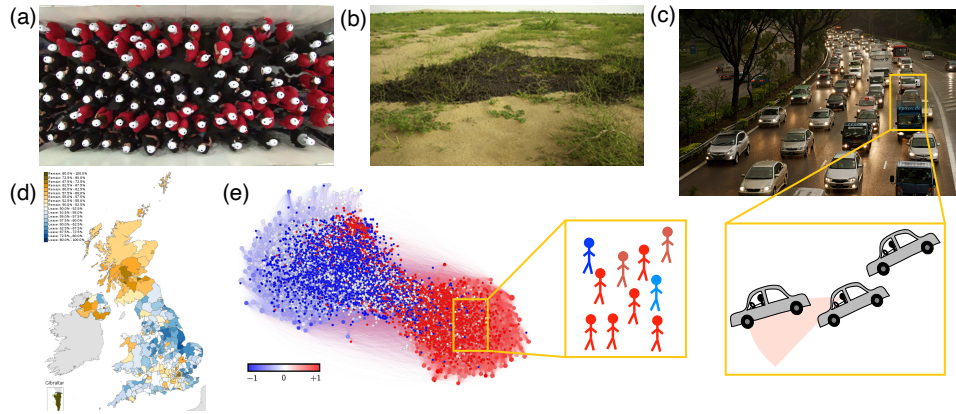


FIGURE 1. Examples of complex social systems. (a) Lanes can emerge from the interactions of pedestrians moving to the right (in black) and left (in red) in a corridor [SSS17, ZKSS12]; see Section 5.2. (b) Locusts form bands as they move over the ground, destroying crops [Cre16]. (c) Drivers react to one another and external signals. In an experiment on a circular road [SFK<sup>+</sup>08], Sugiyama *et al.* instructed originally equidistant drivers to drive normally. Despite the lack of external signals, a phantom traffic jam formed. This jam or pulse of high density traveled backward relative to the direction that the cars moved; see the supplementary material of [SFK<sup>+</sup>08] for an animation. (d) Election outcomes [SHP16, MBN<sup>+</sup>16] and (e) echo chambers [CFPSS19] may emerge from conversations, news coverage, interactions on social media, or other factors. Image (a) adapted (cropped) from [SSS17] and licensed under CC-BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>); image (b) reproduced from [Cre16] with permission from Elsevier, Copyright (2016) Elsevier Inc.; image (c) reproduced from [epS11] and licensed under CC-BY 2.0 (<https://creativecommons.org/licenses/by/2.0/>); image (d) reproduced from [MBN<sup>+</sup>16] and licensed under CC BY-SA 3.0 (<https://creativecommons.org/licenses/by-sa/3.0/deed.en>); image (e) reproduced from [CFPSS19] and licensed under CC-BY 4.0 (<http://creativecommons.org/licenses/by/4.0/>); I added the boxes and cartoons with detail in (c) and (e).

or suggest how to use external controls—e.g., time-dependent gating at ramps—to improve traffic. Models can also provide insight into how echo chambers form or suggest interventions to help dissipate divisions. These goals fall into the framework of seeking to understand normal and altered agent interactions, and to predict resulting group-level features. From a mathematical perspective, modeling complex systems can be a starting point to drive the development of new methods and inspire researchers to combine subfields in novel ways.

Motivated by the breadth of complex social systems, my tutorial lecture “Data-driven modeling” kicked off the 2021 American Mathematical Society (AMS) Short

Course on Mathematical and Computational Methods for Complex Social Systems, and this chapter is an offshoot of that talk. Many of the figures in this tutorial are related to slides in my presentation; these slides and my talk recording are available at [Vol21]. Following the structure of my Short Course presentation, this chapter has three main parts and takes a conceptual approach throughout. First, in Section 2, I highlight some resources, including those that I drew on when preparing my lecture. Second, in Sections 3–4, I overview mathematical modeling and discuss some of the approaches, challenges, and choices that can arise when working with data. Third, in Section 5, I discuss two case studies—election forecasting and pedestrian movement—in more depth.

Mathematical modeling is a big field, and data-driven modeling can be defined in different ways. The array of approaches that modelers can choose from is a strength, since different perspectives contribute in complementary ways to our understanding of complex systems. As a central theme, I want to acknowledge these choices and use the quotations from Segel and Edelstein-Keshet [SEK13] at the start of this chapter as a guide. The abundance of modeling approaches to complex systems, coupled with their multidisciplinary nature, also means that communication is more challenging; researchers may not mean the same thing when they say the same term. With this in mind, I discuss some of the things that I—from my perspective as an applied mathematician and math biologist—consider when I think about modeling complex systems. There are many, many perspectives on modeling, and this tutorial represents one, informed by the references herein.

## 2. Some Resources on Modeling

I point out some resources below, including the materials that I drew on for my Short Course lecture [Vol21].

**2.1. Free Online Resources.** The websites [Bro22, DBC<sup>+</sup>19] provide dynamic examples of research in complex systems and are an excellent place to gain intuition and explore this field. The Society for Industrial and Applied Mathematics (SIAM) hosts two modeling handbooks [BKGL18, BFG14]; and SIAM and the Consortium for Mathematics and its Applications provide guidelines on teaching mathematical modeling [GAI19]. Humpherys, Levy, and Witelski organized a very useful minitutorial discussing graduate and undergraduate education in modeling at the 2016 SIAM Annual Meeting; both their slides and a recording of their presentation are available online [HLW16]. Kutz and Brunton have posted a rich collection of videos [Bru, Kut] on YouTube, discussing topics including data-driven model discovery. For a demonstration of how to go from a biological paper to making simplifications to building different models, my tutorial lecture [Vol20a] for a broad audience may be of interest. Also geared toward a biological audience, the course “What do Your Data Say?” [MJ] includes a large collection of video lectures with a statistical, data-driven perspective. To see examples of research talks related to modeling complex systems, I highlight some of the BIRS workshop videos [BIR] (this collection from the University of British Columbia library contains a wider selection of topics than just modeling), as well as videos in the virtual SIAM Data Science minisymposium “Topological Techniques and Data-Driven Modeling in Complex Systems” organized by Brooks and Porter [BP20b].

**2.2. Books.** I found the books [KBBP16, Kut13, BK19] to be especially helpful as I developed my Short Course lecture, and the book [SEK13] by Segel and Edelstein-Keshet provides the quotations that open this chapter. Additional books related to complex systems and modeling include [Mit09, THK18, Boc10, MP07].

**2.3. Publicly Available Data.** Accessing data can be a challenge in complex-systems research. As a starting point, I highlight some publicly available data for a few specific applications here. For studies on elections and political opinions in the United States, I recommend the breadth of polling data aggregated by FiveThirtyEight [BBG<sup>+</sup>22]. HuffPost Pollster also curates a broad collection of public polls, with a search bar for finding data [Huf22a, Huf22b]. At a finer scale, the 2016 presidential election results in California are available at the precinct level from the *Los Angeles Times* [SFK16]. Ciocanel, Topaz, and other researchers through the Institute for the Quantitative Study of Inclusion, Diversity, and Equity (QSIDE) [QSI] developed a large-scale database (called JUSTFAIR, for Judicial System Transparency through Federal Archive Inferred Records) holding over 500,000 federal district court records [CTS<sup>+</sup>20]. Data from the social-media platform Twitter, as well as tutorials, are available from sources including [AAA<sup>+</sup>21, Sto, KS20, Twi].

### 3. Some Perspectives on Data and Models

Because terminology can vary across fields, I survey some terms for describing models (Section 3.1) and data (Section 3.2), and then define data-driven modeling for the purposes of this chapter (Section 3.3). If you are coming to this tutorial with an applied question that you want to address, I encourage you to keep your complex system in mind as you read—what are the parameters in your system, what data could you use to constrain your model, and at what scale do you want to make predictions or describe the system? If you are a mathematician new to modeling, what mathematical challenges does thinking from the perspective of complex systems raise? If you are from a different disciplinary background than mine, how does what we mean by “data-driven modeling” differ from and complement each other? And, if you happen to be a modeler who—like me—was introduced to modeling through research, it might be interesting to reflect on how we teach modeling.

**3.1. Types of Models.** The term “model” means different things in different fields. In the life sciences, “model” may refer to a model organism (e.g., zebrafish, fruit flies, or worms) [HLW16] or a schematic hypothesizing the relationship between things. In mathematics, we may think of models that take the form of differential equations or stochastic rules, for example. Mathematical models are described using many terms, and I include a few in Figure 2. Figure 2 also highlights some of the initial choices that modelers face, often constrained by their data. Importantly, the distinctions in Figure 2 are not sharp: models often fall on a spectrum and this can depend heavily on the perspective that one takes.

Models can be described as deterministic or stochastic; stochastic models include variability. Depending on their goals and data, researchers must choose whether to build models that are static (time-independent) or dynamic. Similarly, scientists are faced with the choice of building models that are spatial or non-spatial. Do we need to understand where individual cars are located on a road,

Models can be:

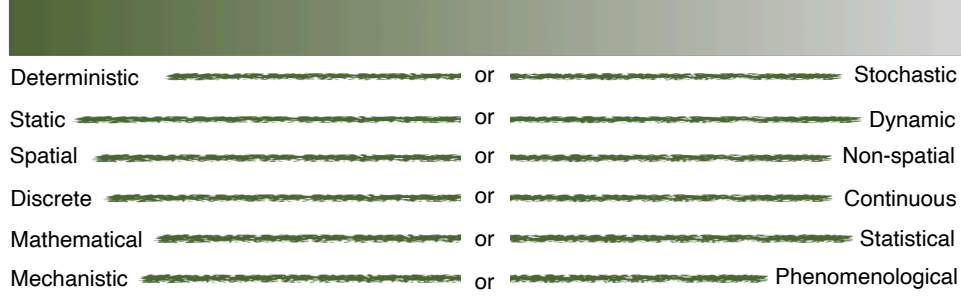


FIGURE 2. Example modeling perspectives [Vol21]. Many models fall somewhere in the middle of each of these scales. For example, a model may have both stochastic components (e.g., stochastic rules for when new pedestrians enter the corridor in Figure 1(a)) and deterministic components (e.g., differential equations for pedestrian movement). Models can be discrete or continuous in many ways: they can be discrete in terms of types of opinions (e.g., Republican or Democratic voting opinions in the United States), physical space, or time, for example. The distinction between phenomenological and mechanistic models is difficult, and folks have different opinions on what this means, as I discuss in Section 3.1.

or is it sufficient to know how the number of cars evolves? Multiscale approaches are also possible, and I provide an example for the case of pedestrian movement in Section 5.2. We can think of models as being discrete or continuous in time or in space (e.g., so-called “on-lattice” or “off-lattice” microscopic models; see Section 5.2), but models can also be discrete in terms of types of agents; for example, do we assume voters live on an ideological spectrum or assign them a binary opinion? Whereas the choice of making a model discrete or continuous in space and time is often a choice of mathematical and computational implementation, the choice of modeling agents as having discrete or continuous features can be particularly meaningful from the perspective of the application. Understanding how choices of implementation impact model predictions is an important area of research (e.g., [KBF17]), as is uncovering how different modeling approaches—such as microscopic and macroscopic (see Section 5.2)—are related (e.g., [BT11, CP21, BV05]).

Some researchers distinguish between mathematical and statistical models, and others see statistical models as a type of mathematical model. A related categorization is phenomenological or mechanistic. These are difficult distinctions, and, in my opinion, scientists use the terms “phenomenological” and “mechanistic” in different ways. Mechanistic models of complex systems get at the mechanism underlying agent behavior. For example, the drivers in Figure 1(c) want to avoid running into one another, and we could model this by specifying repulsive forces between cars. This model can be seen as phenomenological since it describes the affect (e.g., drivers avoid one another) without getting at the mechanism of how the repulsion occurs. If we modeled the physics of the vehicles, the vision cone of individual drivers, and each driver’s internal decision process, this would be more

mechanistic. However, what are the variables in a model of how people make decisions? This is in some sense a phenomenological model as well, raising further questions that involve neuroscience. I suggest that the meanings of “mechanistic” and “phenomenological” depend on the question that we want to answer and the perspective from which we are studying an application. In my opinion, many models are mechanistic at one scale, and phenomenological as soon as we step deeper into the complex system.

**3.2. Types of Data.** The methods that modelers use to build predictive models that balance model and data complexity look different depending on the form of their data. However, the core concepts are the same when building and validating data-driven models if we look more closely, and, for this reason, I overview some types of data here. For example, data may be quantitative (e.g., the speed of the  $i$ th car in Figure 1(c)) or qualitative (e.g., the presence of lanes emerging from pedestrian behavior in Figure 1(a)). See Section 5.1 for an example of modeling with quantitative data, and Section 5.2 for a discussion of the challenges that qualitative data introduce to the modeling process. Textual data also emerge from many complex social systems (e.g., [AAA<sup>+</sup>21, MCM<sup>+</sup>22]).

Sometimes we find ourselves with so much data that we cannot open the files, and other times there is nearly no data. In the first case, the “black-box” modeling approaches that I discuss in Section 3.3 may be useful; for example, if we are working with a huge set of tweets, we could complete some data analysis to identify meaningful categories of accounts. If we are working with large sets of qualitative data (e.g., many images), this may motivate the development of new computational and mathematical approaches for extracting quantitative information from our data. On the other hand, if we have nearly no data, it can be challenging to know where to start. In this case, it is a matter of making many simplifications (and being actively aware of the choices that we make in this process), so that the number of assumptions that we build into our model is balanced with the small amount of data available.

On a related note to amount, data for some complex systems describe rare events. For example, a model may be fit to measurements of average traffic flow, but how do we account for events that are relative outliers, like car accidents? In the case of election forecasting, we might judge a model as wrong if, despite giving Candidate  $A$  a 75% chance of winning and Candidate  $B$  a 25% chance of winning, Candidate  $B$  wins. The reality is that we do not have enough information to determine whether the model is good or bad. Forecasts are more meaningfully judged in aggregate across many elections, but limited polling data are available. Like models, data can also be time-independent (e.g., a Twitter followership network at one snapshot in time) or dynamic (e.g., the timeline of tweets from a given account) and spatial or non-spatial. The initial form of data is often messy, and in some cases a large portion of the time that researchers spend modeling complex systems is focused on cleaning [BKGL18], gathering, and tracking down the oddities in their data.

All forms of data can have bias and require human choices, particularly in the case of complex social systems. I point the reader to the chapter [Por22] by Porter and references therein in this volume for a discussion of data ethics. Importantly, just because data exist does not mean they should be used, and as the author mentions in [Por22], determining when to use or not use data is a critical step

in research on complex social systems. Modelers need to be actively aware of the choices that they make when handling data, and of the presence of any choices made prior to the time that they gained access to the data. For example, if we are interested in understanding the online conversation about a recent event, we might start by downloading a large set of tweets using hashtags associated with that event. There are multiple choices wrapped up in this process, and I name a few here [CY16, MPLC14, Tuf14, TECP20]. First, we chose one of many social-media platforms, so our analysis will be specific to the groups that use Twitter [Tuf14]. Second, we had to select what hashtags to search for and how we would identify tweets “associated with” our recent event [MPLC14, TECP20]. Third, while the Twitter API provides a rich sample of tweets, it is not fully clear how this selection is made [MPLC14]. All of these choices will affect the results of our model.

**3.3. Perspectives on Modeling with Data.** In their 2016 SIAM Annual Meeting minitutorial [HLW16], Humpherys, Levy, and Witelski discussed a useful classification of models based on “shades of model uncertainty”. As I highlight in Figure 3, black-box, gray-box, and white-box models have different levels of dependency on data [HLW16], and their parameters mean different things. According to the classification system in [HLW16], black-box models are based heavily on data and can be thought of as maps between inputs and outputs; these models include regression, classification, and machine learning. For example, Tien *et al.* applied [TECP20] principal component analysis to Twitter data (the input) to distinguish groups of accounts (the output) based on their media followership. The parameters in black-box models may be internal or hidden, and it is the model output—rather than the model structure—that is often of most interest. On the other hand, white-box models are based on first principles; these include equations from physics, such as those describing fluid dynamics or optics [HLW16]. The parameters in white-box models are measurable, and examples are viscosity and conductivity. Gray-box models depend on a combination of data, first principles, and domain expertise. For example, an ordinary differential equation (ODE) model for driver movement could include equations for velocity and acceleration that are based on phenomenological descriptions of repulsion and attraction between cars (i.e., domain expertise) and measurements of speeds (i.e., data).

The distinctions in Figure 3, like the distinctions in Figure 2, are not perfect. For example, equation-learning and model-selection approaches (e.g., [MKBP17, BPK16, NBSF21, KBT<sup>+</sup>22, MBPK16]) might be thought of as “dark gray”. It is also important to keep in mind that there are choices present and domain expertise needed across the spectrum in Figure 3. This is especially true when working with data from complex social systems, since even the data that are selected for training black-box models rely on a modeler’s choice to use those data [Por22]. For the purposes of this tutorial, I thus think of data-driven modeling as being mathematical modeling that is driven by data, motivated by a given question, and combined with domain expertise. This encompasses developing predictive, mechanistic models based on data; equation learning and model selection<sup>1</sup>; machine learning, regression, or classification to understand data; and using models to raise questions and drive further data collection. Both black-box and gray-box models

---

<sup>1</sup>Equation learning and model selection—sometimes referred to as “data-driven modeling”—are outside the scope of this survey. See, for example, [MKBP17, BPK16, NBSF21, KBT<sup>+</sup>22, MBPK16] for more discussion of these topics.

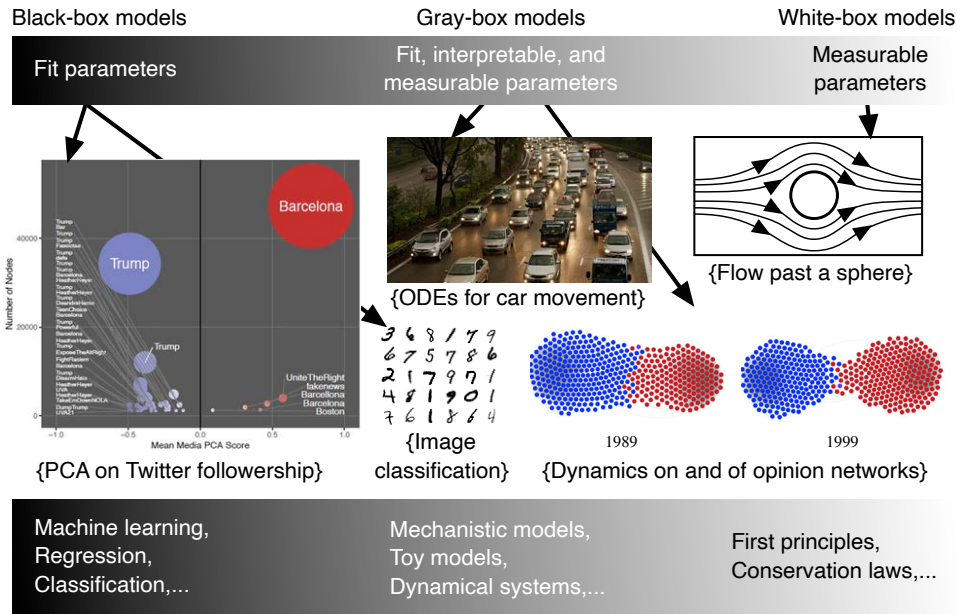


FIGURE 3. Shades of modeling with data [HLW16]. Black-box, gray-box, and white-box models depend on data to varying degrees and have different relationships with parameters. Black-box modeling approaches rely on data and often have internal parameters, while white-box models are largely dictated by first principles and have measurable parameters (e.g., conductivity of a material). Gray-box modeling involves visible, interpretable parameters that are fit, specified, or measured using data. All of these approaches require domain expertise. As some examples, I highlight principal component analysis (PCA) applied to media followership on Twitter [TECP20], and recognizing handwritten numbers [LBBH98] (black-box modeling); deterministic models of traffic flow and game-theoretic models of opinion dynamics on networks [EF18] (gray-box modeling); and fluid flow past a sphere (white-box modeling). PCA–Twitter image and opinion-network images reproduced from [TECP20] and [EF18], respectively, and licensed under CC-BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>); image-classification image reproduced from [LBBH98] with permission, Copyright (1998) IEEE; traffic image reproduced from [epS11] and licensed under CC-BY 2.0 (<https://creativecommons.org/licenses/by/2.0/>).

fit this description, but I predominantly focus on gray-box models in this survey, though again I stress that the distinctions are not sharp.

#### 4. Challenges, Choices, and Creativity in Data-Driven Modeling

Data-driven modeling involves creativity and choices, informed by the modeler’s driving question, data, and domain expertise. In Section 4.1, I provide an example

modeling process and highlight some of the places where modelers make choices. In Section 4.2, I then discuss challenges related to data and model calibration. See Sections 5.1 and 5.2 for illustrations of these topics for two specific applications. I take a conceptual approach throughout.

**4.1. Building Data-Driven Mathematical Models.** As an example data-driven, gray-box modeling process, we might follow the steps below [GAI19, BFG14, BKGL18]:

- (1) formulate our broad motivation and specific goals
  - get to know the application area or talk to domain experts
  - search for data (qualitative or quantitative) and prior work
  - identify hypotheses to be tested or proposed and questions to be “answered” or raised
- (2) come up with a plan for building and evaluating our model
  - determine baseline assumptions and simplify where possible
  - identify our variables, parameter names, timescales, and units
  - specify the values of measurable parameters and determine what parameters need to be fit
  - handle formatting, cleaning, and quantifying our data as needed
  - break our data into sets for fitting parameters, testing, and predicting
- (3) simulate, analyze, and use our model
  - identify remaining parameter values using data for fitting
  - validate our model on test data
  - perform a sensitivity analysis or bifurcation analysis if possible
  - use our model to gain intuition, raise questions, and make predictions
  - communicate results to an interdisciplinary audience
  - iterate to improve

These steps are not necessarily linear and data-driven modeling is iterative [GAI19, BFG14, BKGL18]. The starting point may be data, domain expertise, or questions, and Step (1) involves research to begin filling in gaps in our knowledge of these three areas and to formalize our goals. I often review literature in Step (1) with Step (2) in mind, tagging papers with quantitative data that I can use later for parameter fitting and noting studies that show alternative experimental conditions that could be used for model testing. Steps (2) and (3) then treat complementary parts of model building.

In Step (2), we select our overall approach and the variables, parameters, group dynamics, and agent behaviors in which we are most interested. This means making choices related to the concepts in Figures 2 and 3: for example, if we are studying traffic flow on a stretch of roadway, will we track the number  $N(t)$  of cars on the road in time, or the position  $\mathbf{x}_i(t)$  and velocity  $\mathbf{v}_i(t)$  of each vehicle  $i$  in time? If we are accounting for driver differences, will we assume that each driver’s phenomenological “level of cautiousness” is time-dependent or static? It is important to make these choices in a way that accounts for the complexity of the problem and our data, so Step (2) involves making a plan for how we will use data to develop (or train, or fit) our model and later test (or validate) our model, as I discuss in Section 4.2. At the end of Step (2), our model is written down (e.g., as a system of differential equations on paper or as a set of stochastic rules in code).



In Step (3), we turn to filling in parameter names with parameter values, setting initial conditions, and determining our boundary conditions, as needed. Step (3) involves validating our model to test its predictive value and performing various analyses to check how sensitive our model is to uncertainty in parameter values, initial conditions, boundary conditions, or data. Depending on the form of our model, we may be able to perform a bifurcation analysis to understand how changes in parameters influence our results. We may also brainstorm alternative ways of judging model output and comparing this with data, since how we choose to describe model output can impact how we interpret our results. At the end of Step (3), we use our model to gain understanding and, if possible, suggest new experiments, resulting in model-driven data collection.

More broadly, Step (1) is where we realize that a model can help us accomplish our goals, Step (2) is the place where we build the model structure, and Step (3) is where we test and prod this structure. Data enter the picture in Step (1) as motivation. In Steps (2) and (3), we work closely with data to build, test, and use our model in a way that balances model and data complexity to accomplish our goals. In the remainder of this tutorial, I focus primarily on the later parts of Step (2) and broadly discuss the early parts of Step (3). To learn more about some of the analyses and computational approaches possible in Step (3), I suggest the books [Smi14, Str15, Kut13, SEK13].

**4.2. Balancing Model and Data Complexity.** While data-driven models take many forms and scientists use a range of methods to understand them, the overarching theme of balancing model and data complexity is present throughout. Depending on our goals and data, what modeling approach do we choose? How do we build a data-driven, *data-appropriate* model? If we have access to a wealth of domain expertise and a rich set of data, it may make sense to build a complex model, since, in this case, the majority of the model will be purely descriptive, framing known agent interactions in a mathematical way. The new hypothesis that we are testing, along with its few parameters, enters the picture as our assumption. On the other hand, if we are leading the way to model a poorly understood complex system, our model needs to be very simple, again so that the assumptions and hypotheses that we introduce match the amount of data available.

In either case, it is helpful to break our data up into sets for model training (or development) and testing. Training/development data is the data that we use to build our model, specifying parameter values as well as the form of model rules and terms as appropriate. After this, we take a step back and test whether or not our model behaves well on the data that we withheld—our testing data. If the model does well on the testing set, we can use our model to predict future dynamics or shed light on poorly understood dynamics. If the model does not do well on the testing set, we need to return to model development. As a guiding principle, the more parameters and assumptions that we build into a model, the more that it needs to be able to reproduce in order to have predictive value<sup>2</sup>.

Figure 4(c)–(d) highlights two concepts that are related to balancing model and data complexity: underfitting and overfitting. For illustrative purposes, I consider the example of population growth of some organism in time, given some (imperfect)

---

<sup>2</sup>It is worth noting that data-driven models can be used for many purposes, including describing, explaining, or predicting, and Shmueli discusses these goals from a statistical perspective in [Shm10].

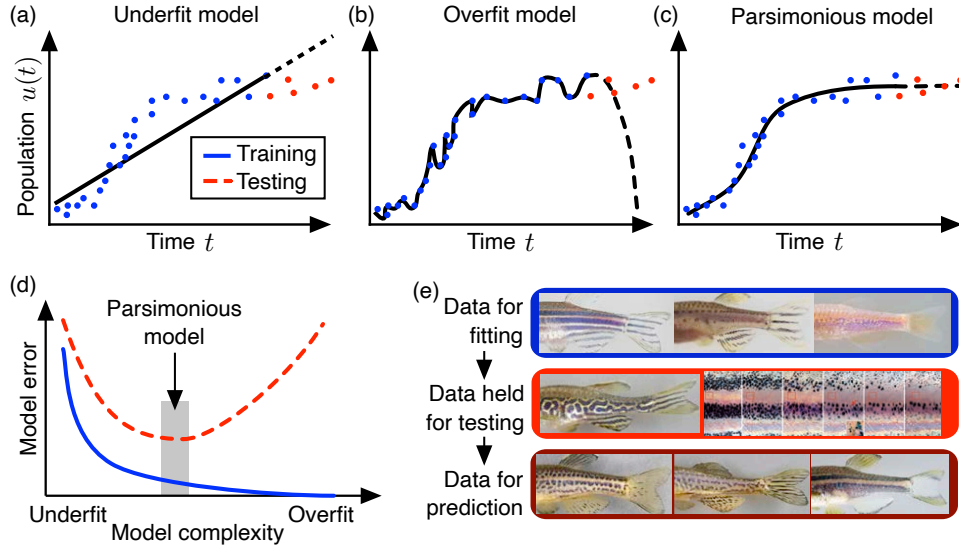


FIGURE 4. Balancing model and data complexity. (a) Underfit models miss meaningful features in data, (b) overfit models include too many assumptions, and (c) parsimonious models balance model and data complexity. In (a)–(c), blue points denote data that we use to develop our model and fit parameters, and red points denote our testing set. (d) Underfit models agree poorly with both our training and testing data, while overfit models represent our training data well and our testing data poorly [BK19, MKBP17]. Parsimonious models perform well on both sets of data. (e) Creating a plan for model training/development and testing is key to data-driven modeling. This involves breaking data into sets for training and testing, a process that depends on our complex system. For example, if our goal is to understand how cells interact to form patterns in fish skin, this could mean breaking our (qualitative) data into images of fish that are well understood (and involve setting parameters in a model to zero in a clear way), more images of fish that are well understood (and involve changing the values of nonzero parameters in a clear way), and images of poorly understood fish (and involve changing parameters in unknown ways) [VS18, Vol17]. The first set is used for model development, the second for testing, and the third as a place where we can make predictions [VS18, Vol17]. Image (d) is based partly on [BK19, MKBP17]; first row, third row, and left image in second row of (e) adapted from [SNV15] with permission from Elsevier, Copyright (2015) Elsevier Ltd.; remaining images in second row of (e) reproduced from [PT03] with permission of The Company of Biologists, Ltd.

measurements of the number of organisms at discrete time points. At one extreme, I could assume a linear relationship between population size and time, fitting a

line to the data. This involves few parameters, and the difference between the model and training data is high. At the other extreme, I could draw a curve that goes through every single data point [BKGL18]—this would mean introducing many parameters. In terms of these models’ ability to approximate population size at some new time in our testing set, neither will do well [MKBP17]. A better model lies somewhere in between these two extremes. What we are after is a “parsimonious” model [MKBP17, BK19]: a model that it is supported by our data and no more complex than it needs to be.

Building predictive, data-appropriate models that avoid overfitting and have strong predictive value looks different based on the problem and relies on domain expertise. (See e.g., [Smi14, BK19] for a more detailed discussion of methods—I focus on broad concepts here.) If our goal is to understand social-media engagement in time, for example, we might build a gray-box model driven by some data  $\{w_i\}_{i=1,\dots,T}$ , where  $w_i$  is the number of accounts on a social-media platform on day  $t = i$ . As one approach, we could split the data into a training set  $\{w_i\}_{i=1,\dots,\tilde{T}}$  and a testing set  $\{w_i\}_{i=\tilde{T}+1,\dots,T}$  with  $\tilde{T} < T$ . We could develop our model and specify its parameters using the training set and then run our model until  $t = T$  to evaluate how well it does on the testing set. If our model does well in testing, we could use it to predict future social-media engagement.

When working with qualitative data, the process of balancing model and data complexity looks different, but it is the same at its core. In Figure 4(e), I highlight the complex biological system that most of my work is on: pattern formation in zebrafish skin [VS15, VS18]. Wild-type and mutant zebrafish feature different patterns, which form through the interactions of pigment cells [SNV15, PT03]. Although there are some quantitative data (e.g., cell speeds), most take the form of images of fish. To build the model [VS18], we broke these qualitative data into three sets. The first set of images contains patterns that correspond to setting specific parameters to zero in a mathematical model (e.g., setting the birth rate of black cells to zero). The second set holds some fish patterns that are relatively well understood; in this case, we know simulating them means changing parameters in a clear way (e.g., slowing domain growth). The final set contains mutant patterns that are poorly understood, patterns that form due to cell interactions that are altered in unknown ways. The first set serves as a natural model development/training set, and once we identified a model that could reproduce these fish patterns, the next step was to step back and break it down, checking if there were any ways that we could simplify the model and still maintain consistency [VS18]. “Minimal” model in hand, we used the second set of images for testing, asking whether or not the model could reproduce data that we did not build into it. And, finally, the tested model now serves as a predictive tool to understand the fish in the third set: at this stage, we change parameters in the model with the goal of identifying altered cell interactions that may lead to mutant patterns [Vol17].

In order to further improve predictive value and avoid overfitting, there are a wealth of other approaches modelers can take. We can test how uncertainty in our parameters, boundary conditions, or initial conditions affect our results, and we can explore whether other modeling approaches lead to the same conclusions. We can set parameters in our models to zero or remove rules, checking to see if our models can be made simpler without losing agreement with the training set. We can also ask questions about whether the methods that we use to judge our

models influence our results: what alternative methods for measuring agreement between model output and data can we test? Throughout this process, the goal is to critically investigate our modeling assumptions as we build a parsimonious—or minimally complex—model based in our data.

## 5. Illustrative Case Studies

In the remainder of this tutorial, I turn to two case studies of complex social systems: opinion dynamics during elections (Section 5.1) and pedestrian movement in crowds (Section 5.2). These examples illustrate some of the types of models and data from Section 3 in the broader framework of the challenges and choices introduced in Section 4. I highlight the benefits and drawbacks of different modeling choices, with the quotations from Segel and Edelstein-Keshet [SEK13] at the beginning of this chapter as a guide.

**5.1. Forecasting Elections.** Political opinion dynamics are a complex social system, and here I focus on the goal of forecasting elections in the United States. Election forecasting is highly interdisciplinary, drawing on probability, geometry, dynamical systems, topology, and statistics, as well as political science, history, economics, computer science, and sociology more broadly. It naturally involves communication and public science, and different forms of data (Section 5.1.1). Framed by this interdisciplinarity, I illustrate a statistical, static modeling approach to elections in Section 5.1.2 and a dynamic, mathematical model in Section 5.1.2.

Many other models and methods for incorporating data into forecasts exist beyond the scope of this survey (e.g. data-assimilation techniques [LSZ15]). Election forecasting raises questions at many different scales; for example, using a compartmental model, Restrepo *et al.* [RRH09] investigated how polling data affect whether potential voters decide to vote, and Biondo *et al.* [BPR18] developed an agent-based model to better understand how surveys influence opinions. Election forecasting is related to the broader field of opinion dynamics [CFL09, PG16], which includes the formation and dynamics of echo chambers (e.g., [SCP<sup>+</sup>21, EF18, CDFMG<sup>+</sup>21]) and polarization (e.g., [SMA20, YAKM20]). There are many approaches to opinion formation, such as voter models [FGSR<sup>+</sup>14, BdA17] and threshold models [LYY18].

Because elections receive attention so widely and forecasts have the potential to impact turnout, the example of election forecasting highlights a place in complex-systems research where carefully presenting the results of data-driven models is especially important. Communicating probabilistic forecasts in a tangible, interpretable way itself leads to questions, and I suggest [GHWM20, FPS<sup>+</sup>21] for further discussion about visualizing and communicating uncertainty. Election forecasting also presents interesting challenges when it comes to evaluating model success and forecast accuracy [GHWM20], as I mentioned in Section 3.2.

**5.1.1. Election Data.** In terms of Step (1) in Section 4.1, as a starting point, the data used to build election-forecasting models include historical results, approval ratings, economic indicators, information about incumbency, and polls [HR14, Sil12, Abr08]. Analysts often separate these data into two types: polls and “fundamental data” (or “fundamentals”). Fundamental data are the data from which voters may form their opinions and determine how they will vote [GK93]; for example, economic data fall into the fundamentals category. Regardless of the type, all data come with challenges: data may not go back in time as far as we would

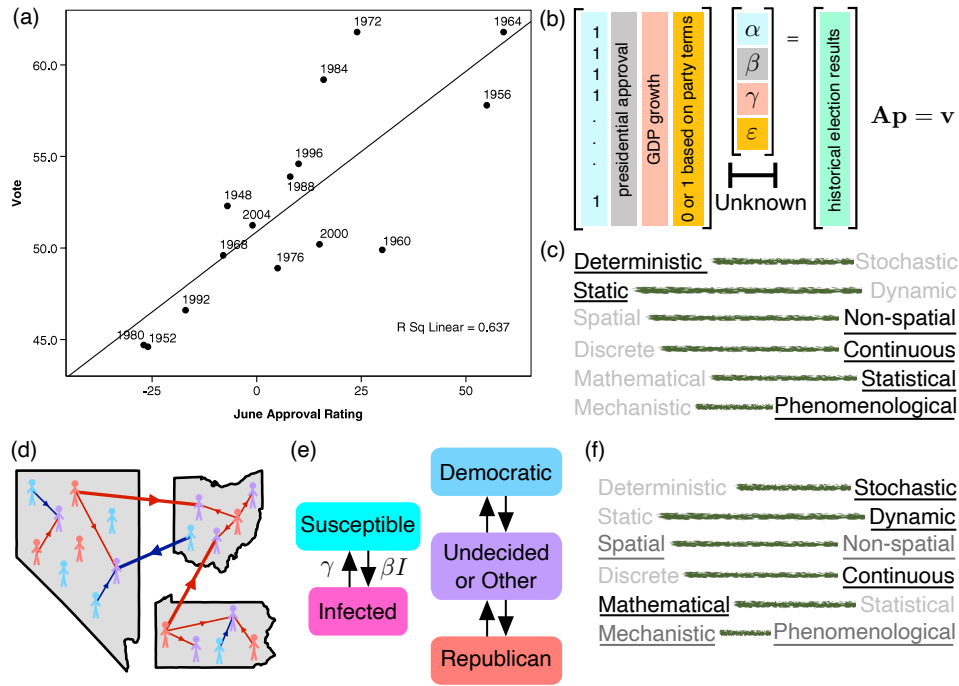


FIGURE 5. Example approaches to forecasting U.S. elections. (a) The president's June approval rating and the percentage of the national vote for their party tend to be related [Abr08]. (b) Abramowitz's [Abr08] model is driven by fundamental data. (c) This statistical model is deterministic and continuous in the sense that, once the parameter values are set, the result is one prediction of the national vote for the incumbent party (a continuous number between 0 and 100). (d) In a network model, one could investigate the interactions between undecided (purple), Republican (red), and Democratic (blue) voters, as I illustrate in this cartoon. Because networks [PG16, Str01, New10] are the focus of another chapter in this volume, I do not cover them; I suggest the lectures by Brooks and DeFord in our Short Course [Bro21, DeF21]. (e) Compartmental modeling [Het00, DH00, BCC12] involves grouping individuals into categories and investigating how folks change compartments. (f) The compartmental model [VLPR20] is a stochastic mathematical approach to forecasting elections. It is spatial in the sense that it produces state-level forecasts, and non-spatial in the sense that it does not track the locations of individual voters. Image (a) reproduced from [Abr08] with permission, published by Cambridge University Press, and Copyright (2008) The American Political Science Association; images (d)–(e) adapted from [VLPR20].

hope or may not be as fine-scale as we would like (e.g., data at the national or state level, rather than the district level).

For forecasts that depend on historical data, one assumption is that the past and the future will behave similarly. Modeling with fundamental data allows forecasters to produce early predictions, prior to when accurate polls may be available [HR14, Lin13]. However, opinions are dynamic—both across years and within the same election year—and past elections may not be representative of how voters will behave in the future. On the other hand, it is not always clear whether shifts in polls in a given year represent real shifts in opinion or just differences in pollster methods [GK93, WE02, Jac05]. Moreover, polling data are often biased [Jac05] and adjusted in proprietary ways; for example, pollsters make decisions such as how to define “likely voters”. Polling data can be spotty, with some states being polled more frequently than others [Lin13]. Adding another layer of complexity, pollster herding is a phenomenon in which polling organizations adjust their results when their data do not align with other polls [Sil14, CR13, GHWM20].

5.1.2. *Example Statistical Approach.* In Figure 5(a), I reproduce a plot from [Abr08] of net presidential approval ratings<sup>3</sup> in June versus the percentage of the vote that went for the incumbent president’s party in November of the same year. This motivates a statistical modeling approach to forecasting U.S. elections that is driven by fundamental, historical data. As an example of such an approach, I highlight some of the ideas in Abramowitz’s “time-for-change” model [Abr08, Abr88]:

$$(5.1) \quad \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_m \end{bmatrix} = \begin{bmatrix} 1 & a_1 & g_1 & c_1 \\ 1 & a_2 & g_2 & c_2 \\ 1 & a_3 & g_3 & c_3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_m & g_m & c_m \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \varepsilon \end{bmatrix},$$

where  $v_i$  is the percentage of the national vote that went for the presidential candidate from the incumbent party in the  $i$ th election in the data set;  $m$  is the number of years for which data is available;  $a_i$  is a measurement of presidential approval before the  $i$ th election;  $g_i$  includes information about economic growth in the year leading up to the  $i$ th election; and  $c_i$  is a variable related to incumbency. Once the parameters  $\alpha, \beta, \gamma$ , and  $\varepsilon$  are determined from historical data (e.g., using regression), the time-for-change model [Abr08, Abr88] can predict an election  $m + 1$  by computing  $v_{m+1} = \alpha + \beta a_{m+1} + \gamma g_{m+1} + \varepsilon c_{m+1}$ .

Equation (5.1) has the general form  $\mathbf{v} = \mathbf{A}\mathbf{p}$ , where  $\mathbf{p}$  corresponds to parameters,  $\mathbf{A}$  contains fundamental data, and  $\mathbf{v}$  holds  $m$  past election outcomes. If we were to introduce more types of historical data, the number of parameters  $n$  would grow. With more parameters, we would expect to get a better fit between the model predictions and past election results. As Figure 4 highlights, however, this does not necessarily correspond to better predictions of future elections, since allowing  $n$  to become too large can lead to overfitting. This raises questions about model complexity. How many kinds of fundamental data should a modeler include? How many terms in the model is the “right” number of terms?

---

<sup>3</sup>This is defined as approval minus disapproval (see [Abr08] for details), so it can be negative.

To address these questions, we need to define what a good model means and choose how to measure error. For example, consider the function [BK19]:

$$(5.2) \quad E(\mathbf{p}) = \underbrace{\|\mathbf{A}\mathbf{p} - \mathbf{v}\|_2}_{\text{least-squares term}} + \underbrace{\lambda_1 \|\mathbf{p}\|_1}_{\text{LASSO term}} + \underbrace{\lambda_2 \|\mathbf{p}\|_2}_{\text{ridge-regression term}}.$$

We can minimize  $E(\hat{\mathbf{p}})$  to find the parameter values most consistent with our data:

$$\mathbf{p} = \underset{\hat{\mathbf{p}}}{\operatorname{argmin}} E(\hat{\mathbf{p}}).$$

When  $\lambda_1 = \lambda_2 = 0$  in Equation (5.2),  $E(\mathbf{p})$  is the least-squares difference between the model's predictions and the election outcomes under the parameters  $\mathbf{p}$ . This method for measuring goodness-of-fit is sensitive to variability [BK19]. If  $\lambda_1 > 0$  and  $\lambda_2 = 0$ , we instead implement LASSO regression [Tib96], which selects sparse models and helps prevent overfitting by forcing some parameters to zero [BK19]. When  $\lambda_1 > 0$  and  $\lambda_2 > 0$ , Equation (5.2) corresponds to elastic-net regularization.

Importantly,  $\lambda_1$  and  $\lambda_2$  provide a means of calibrating model complexity. We can choose to minimize Equation (5.2) for different values of the hyper-parameters  $\lambda_1$  and  $\lambda_2$ , resulting in different models (in the form of the parameter values  $\mathbf{p}$ ) for each choice. Information criteria, such as Akaike information criteria (AIC) and Bayes information criteria (BIC), can come in handy to select the best model from among these alternatives [MKBP17, Aka98, Aka74, Sch78]. *The Economist's* 2020 forecasts [eGH20], for example, depend in part on a statistical model of the form  $\mathbf{A}\mathbf{p} = \mathbf{v}$  with a matrix  $\mathbf{A}$  that contains many types of fundamental data. To help prevent overfitting, *The Economist* [eGH20] team combines leave-1-out cross validation [BK19] and elastic-net regularization with a range of  $\lambda_1$  and  $\lambda_2$ .

Broadly, leave- $k$ -out cross validation is a means of breaking data into training and validation sets. To implement this method, one removes  $k$  samples of the training data; the removed data then becomes the validation set, and the remaining data is used for training [BK19]. For example, if  $k = 1$  in the presidential election setting and the available data are for the years 2004, 2008, 2012, and 2016, one first removes one year of data (e.g., the 2012 data). The next step is determining the parameter values  $\mathbf{p}_{2012}$  that result from fitting based on the data for the remaining years (2004, 2008, and 2016, in this example). Repeating this for the other years leads to four sets of parameter values. One option is to define the final parameter values  $\mathbf{p}$  as the mean of these four sets of parameters. Other approaches to testing and validation include  $k$ -fold cross validation [BK19].

In Figure 5(b)–(c), the statistical, phenomenological approach of this section has benefits and drawbacks, like all models do. Because it is driven by fundamental data, the time-for-change model [Abr08, Abr88] is not dependent on noisy polling data; instead, it is able to generate forecasts as early as approval, economic, and incumbency data are available. Moreover, this model is simple and has few parameters. On the other hand, the model [Abr08, Abr88] in Figure 5(b) is static, and it does not add mechanistic understanding of what causes opinions to change in time during an election year.

5.1.3. *Example Dynamical-Systems Approach.* As a more mechanistic approach, one example is the mathematical model [VLPR20] that my collaborators and I developed for forecasting U.S. elections. This model, driven by polling data [Huf22a, Huf22b, Rea22, BBG<sup>+</sup>22], has a compartmental Susceptible–Infected–Susceptible

(SIS) model at its core. Compartmental modeling is a widely used method for describing disease dynamics (e.g., [KM27, KM32, KM33, Het00, DH00, BCC12]), and it has also been applied to social contagions (e.g., [BCAKCC06, BGBD<sup>+</sup>18]). The central concept is that the population of interest can be grouped into compartments<sup>4</sup>. In the SIS setting (Figure 5(e)), there are two compartments: susceptible and infected. Susceptible individuals become infected through interactions with infected folks (i.e., transmission), and infected individuals recover, becoming susceptible. If we track the fraction of the population that is susceptible or infected in time, the result is a gray-box model in the form of differential equations.

In the approach [VLPR20], we adapt the traditional SIS compartmental model by introducing two “contagions” (Democratic and Republican voting inclinations) and replacing susceptible individual with undecided or other voters. For each state or region  $i$ , we track the fraction of undecided  $S^i(t)$ , Democratic  $I_D^i(t)$ , and Republican  $I_R^i$  voters in time according to the stochastic ordinary differential equations:

$$(5.3) \quad dI_D^i(t) = \underbrace{-\gamma_D^i I_D^i}_{\text{Dem. recovery}} dt + \underbrace{\sum_{j=1}^M \beta_D^{ij} \frac{N^j}{N} S^i I_D^j}_{\text{Dem. transmission}} dt + \underbrace{\sigma dW_D^i(t)}_{\text{uncertainty}},$$

$$(5.4) \quad dI_R^i(t) = \underbrace{-\gamma_R^i I_R^i}_{\text{Rep. recovery}} dt + \underbrace{\sum_{j=1}^M \beta_R^{ij} \frac{N^j}{N} S^i I_R^j}_{\text{Rep. transmission}} dt + \underbrace{\sigma dW_R^i(t)}_{\text{uncertainty}},$$

where we use that  $S^i(t) = 1 - I_D^i(t) - I_R^i(t)$  to reduce the number of equations. Here  $I_D^i, I_R^i$ , and  $S^i$  are stochastic processes;  $W_D^i$  and  $W_R^i$  are Wiener processes;  $M$  is the number of states or regions;  $N^j$  is the number of voting-age individuals in state  $j$ ; and  $N$  is the total number of voting-age individuals across our  $M$  regions. This model involves the simplifying assumption that we can bin voters as Democratic, Republican, or undecided. Bounded-confidence and related models (e.g., [WPCG<sup>+</sup>14, DNAW00, HK02, BP20a]) account for opinions existing on a continuous spectrum.

The parameters in Equations (5.3)–(5.4) call for special attention. There are  $2 \times M$  parameters  $\{\gamma_D^i, \gamma_R^i\}_{i=1, \dots, M}$  that describe the rates at which committed voters become undecided. There are also  $2 \times M^2$  parameters  $\{\beta_D^{ij}, \beta_R^{ij}\}_{i,j=1, \dots, M}$  for the rates at which Democratic (Republican) voters in state  $j$  “infect” undecided voters. To find the values of these parameters, we [VLPR20] relied on polling data. For the ODEs associated with Equations (5.3)–(5.4) (with  $\sigma = 0$ ), we minimized the least-squares difference between our model output under parameters  $\mathbf{p}$ :

$$\mathbf{X}(t_k; \mathbf{p}) = [I_R^1(t_k; \mathbf{p}), \dots, I_R^M(t_k; \mathbf{p}), I_D^1(t_k; \mathbf{p}), \dots, I_D^M(t_k; \mathbf{p}), S^1(t_k; \mathbf{p}), \dots, S^M(t_k; \mathbf{p})],$$

<sup>4</sup>This general structure is very flexible: for example, in Figure 6(d), I highlight one way that compartmental modeling could be used to describe pedestrian dynamics. Here the compartments are leading pedestrians moving to the right, following pedestrians moving to the right, leading pedestrians moving to the left, and following pedestrians moving to the left. If we are mainly interested in understanding how many leaders and followers are present, this approach could suffice. We might consider the transition of left-moving leaders to left-moving followers as dependent on interactions with other leaders.



and the averaged state- or region-level polling data:

$$\mathbf{x}(t_k) = [R^1(t_k), \dots, R^M(t_k), D^1(t_k), \dots, I_D^M(t_k), S^1(t_k), \dots, S^M(t_k)],$$

where  $k = 1, \dots, T$  with  $T$  months of polling data considered. The parameter values are different for each election year and race, depending on the associated polls.

The goal of forecasting elections provides a natural means of building and testing a model. By using only the polling data (but not the election results) for past races, we can test Equations (5.3)–(5.4) by retroactively forecasting past elections [VLPR20]. For the statistical model in Figure 5(b), one of the challenges is selecting what types of fundamental data to include in the model, and this comes down to determining what parameters are zero or nonzero. In contrast, for the mathematical model here, it is more the format of the differential equations and the assumptions of an SIS-style model, rather than the values of the parameters, that we want to evaluate. Because the parameters in Equations (5.3)–(5.4) depend only on the polls for a given election year, this model can be tested by applying it to forecast previous elections, one at a time. This step in some sense combines model training and validation together. In terms of predictions, there is also a natural—and high-stakes—opportunity: the model can be used to forecast upcoming elections.

One of the benefits of the continuous, stochastic mathematical model in Equations (5.3)–(5.4) is that it includes some mechanistic hypotheses about opinion dynamics. The model [VLPR20] is also dynamic in time; see Figure 5(f). Once polls become available, Equations (5.3)–(5.4) can forecast a new U.S. election with parameters that are specific to that election. However, opinion dynamics are not the same as biological disease transmission. Instead, we might think of the transmission terms in Equations (5.3)–(5.4) as capturing interactions between committed voters in state  $j$  and undecided voters in state  $i$  in a phenomenological way. These interactions could be direct (e.g., via conversations between a committed voter in one state and an undecided voter in another state) or indirect (e.g., through news coverage). As another drawback, the model [VLPR20] has many more parameters than the statistical approach [Abr88, Abr08].

**5.2. Modeling Pedestrian Movement.** Crowds of people exhibit rich collective behavior, including lane formation and oscillating flows [HM95, HBJW05, HJ09, SSS17]. For example, as I show in Figure 6(a), pedestrians may form lanes when two groups walk in opposing directions in a narrow corridor. Like the application of election forecasting in Section 5.1, studying the dynamics of crowds touches on many fields, including engineering, sociology, psychology, physics, computer science, and mathematics [BCG<sup>+</sup>16, SSS17, BR19, HJ09]. This interdisciplinarity stems from the goals that can motivate models of pedestrian movement. Researchers may be interested in designing functional buildings, testing how guidelines influence disease transmission in a crowd, developing methods to improve evacuation in emergency settings, or something else. Here I focus on the goal of understanding under what conditions lanes emerge from pedestrian interactions, and I assume accounting for the spatial organization of individuals in time is important.

For this tutorial, I use crowd movement as a venue for discussing approaches to modeling agent behavior in space, and highlighting some challenges associated with qualitative data (Section 5.2.1). Pedestrian movement provides an opportunity to illustrate a range of gray-box, spatial models, including continuum models, cellular-automaton perspectives, and agent-based approaches (Section 5.2.2). There are

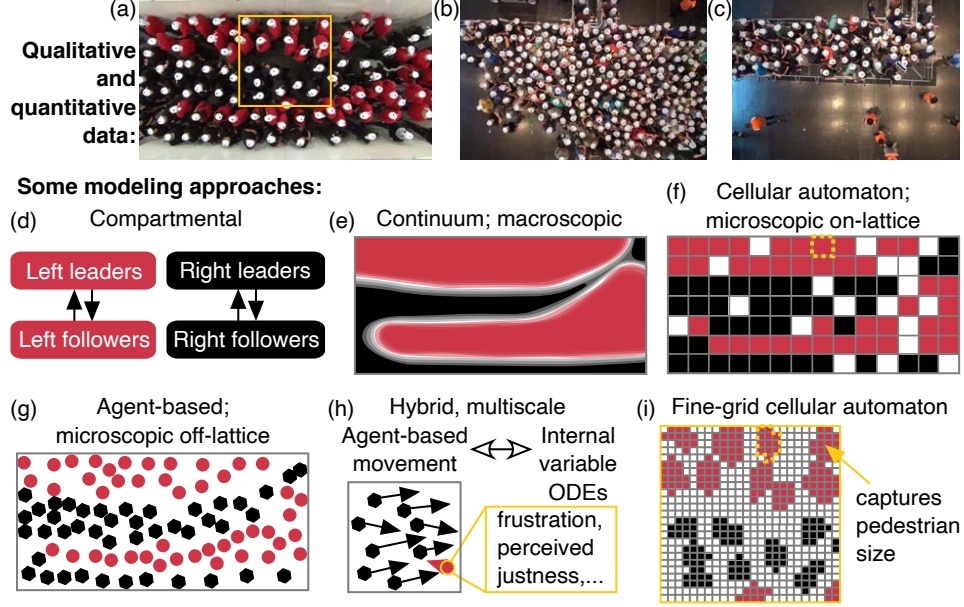


FIGURE 6. Example modeling approaches to pedestrian movement. Experiments in settings such as (a) bidirectional movement in corridors [ZKSS12], (b) movement through an entrance without broader spatial barriers [SSS17], and (c) movement through an entrance with spatial constraints [SSS17] produce both quantitative and qualitative data. There are many approaches that we could take to describe lane formation in (a). In (a) and (d)–(i), red denotes pedestrians moving to the left and black denotes study participants moving to the right. (d) Compartmental models, discussed in Section 5.1.3 could, for example, track the fraction of study participants who are following others or leading lines; this approach is non-spatial. (e) Macroscopic models track pedestrian density and generally take the form of PDEs. (f) Microscopic on-lattice models consider the positions of individuals in discrete space and involve stochastic, computational rules. (g) Microscopic on-lattice models track the positions of individuals in continuous space through coupled differential equations. (h) Hybrid, multiscale approaches come in many forms; for example, we could couple an agent-based model of pedestrian movement with a compartmental model describing the feelings within each pedestrian, which, in turn, influence their movement. (i) Fine-grid cellular automaton models use multiple grid squares to represent each pedestrian, providing a more detailed perspective on pedestrian position. Images (a)–(c) adapted (cropped) from [SSS17] and licensed under CC-BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>).

other data-driven approaches to crowd dynamics, and I highlight [BR19] for a review of statistical models. From the perspective of building simplified models (in

particular, models that do not include concepts from social psychology [SSS17]), similar challenges and approaches arise in diverse examples of pattern formation and self-organization, including migrating cells (e.g., [BEK20, Vol20b, GBKM20, HRM17, GG93]), animal aggregations (e.g., [CKJ<sup>+</sup>02, PEK99, LLEK10]), swarming locusts (e.g., [AA15, BCME<sup>+</sup>20, BT11]), and more general agents interacting in space (e.g., [CDM<sup>+</sup>07, VCBJ<sup>+</sup>95, LRC01, DCBC06, MEK99, TBL06, CMW16]).

5.2.1. *Pedestrian Data.* Data on pedestrian movement comes in quantitative and qualitative forms, including measurements of velocity [ZKSS12], questionnaires about pedestrian experience [SSS17], and images of crowds [BHK<sup>+</sup>11]. This information may stem from observations in the field or in controlled lab settings. For example, Zhang *et al.* [ZKSS12] performed a series of experiments in which study participants were instructed to move through corridors of different widths. As I show in Figure 6(a), participants in red were asked to move to the left through the corridor, and pedestrians wearing black were asked to move to the right. Lanes—visible as red and black stripes in Figure 6(a)—emerged from the interactions of the pedestrians in some settings [ZKSS12]. In addition to this qualitative data, the experiments [ZKSS12] produced trajectories of each participant’s position, along with measurements of velocity and density.

As another example, Sieben *et al.* [SSS17] performed a series of experiments to better understand how pedestrians respond to different barriers as they seek to pass through an entrance. The setups [SSS17] in Figure 6(b)–(c) are meant to represent what might happen when people are entering a concert venue. After extracting the positions of the white caps worn by pedestrians, Sieben *et al.* [SSS17] collected trajectories of individuals. The authors [SSS17] also asked the study participants questions about their experience of walking through the entrance before and after watching a video of the experiment. This survey [SSS17] produced data on how comfortable the heterogeneous participants reported feeling and how just they felt the entrance process was, among other things.

When we view Figure 6(a), the presence of stripes is striking; while it is not as visible in Figure 6(c), the trajectories of pedestrian movement that Sieben *et al.* [SSS17] extracted from these experiments also show lanes in some cases. This highlights one of the challenges associated with spatial complex systems: many of the features in Figure 6(a)–(c) are qualitative. We may see stripes, but how do we define these stripes objectively and quantitatively in large sets of images? At different timepoints in the experiment (see the videos in the supplementary material of [SSS17]), the stripes are not as clear and do not extend across the full length of the corridor. How do we define stripe width or the time when bands start or end along the length of the corridor? The qualitative nature of data in spatial complex systems presents new challenges when fitting and testing models.

5.2.2. *Example Spatial Modeling Approaches.* Figure 6 shows some approaches to spatial modeling of complex systems, including crowd movement, at different levels of detail. Here I focus on introducing some broad gray-box, mathematical modeling approaches that we could take to study lane formation, rather than discussing specific references. (See the reviews [BCG<sup>+</sup>16, SCS<sup>+</sup>18, DDH13] and references therein for more information about crowd dynamics.) These approaches—namely macroscopic, microscopic on-lattice, microscopic off-lattice, and hybrid (e.g., [KHB13]) models—are used to study a wealth of spatial dynamics. There are

also many perspectives that I do not discuss, including mesoscopic (e.g., [FTW18, BBK13]) and game-theoretic (e.g., [Dog10, LW11, BCD18]) approaches.

Macroscopic, continuum models of pedestrian movement often take the form of partial differential equations (PDEs). As I show in Figure 6(e), this approach stems from a zoomed-out perspective: instead of tracking the locations of individuals in Figure 6(a), continuum models describe the evolution of density in time. If we make the assumption that there are two populations in our corridor example, a continuum model would track the density  $r(\mathbf{x}, t)$  of “red-shirt-wearing” and  $b(\mathbf{x}, t)$  “black-shirt-wearing” pedestrians in space  $\mathbf{x}$  and time  $t$ . One benefit of macroscopic models is that they are often analytically tractable, and they provide a broad perspective on overarching features that may be at work in a complex system. These models often have few parameters, and researchers can perform bifurcation analysis to understand how these parameters influence group dynamics. The drawback is that PDE approaches may simplify the complex dynamics of heterogeneous pedestrians significantly, and it can be challenging to relate the few parameters in these models to specific agent behaviors.

In contrast to macroscopic models, microscopic approaches focus on the positions or features of individuals, and two prominent frameworks are on-lattice and off-lattice models. These models provide more detailed perspectives at the scale of individual agents, which comes at the cost of more parameters. Spatial modeling is a place where vocabulary differs some between fields, particularly in the case of microscopic models. Depending on one’s perspective, the microscopic models in Figure 6(f)–(e) may be described as individual- or agent-based models (IBMs or ABMs), since these models track changes in the positions of agents. The term “agent-based” also refers to more detailed models such as [BHK<sup>+</sup>11, BDM<sup>+</sup>09, TFB<sup>+</sup>11]. Miller and Page [MP07] describe agent-based models as “bottom-up” approaches, because the starting point is interactions of individuals. In interdisciplinary—or even within-discipline—conversations, I suggest asking questions to clarify what folks mean by ABMs and IBMs in their setting.

Microscopic on-lattice (cellular automaton) models consider space as a lattice, and pedestrians can either occupy or not occupy positions on a grid (e.g., [BKSZ01, VCM<sup>+</sup>07, BA01]); see Figure 6(f). Movement, as well as arrival and exit, takes the form of stochastic, computational rules. Notationally, we could denote whether the grid square in row  $i$  and column  $j$  at time  $t_k$  is red (i.e., containing a pedestrian moving to the left in Figure 6(a)), black (i.e., holding a right-moving pedestrian), or white (empty) by:

$$x_{i,j}(t_k) = \begin{cases} -1 & \text{if grid square is red} \\ 0 & \text{if grid square is empty} \\ 1 & \text{if grid square is black.} \end{cases}$$

For example, to model right-traveling pedestrians stepping to the side to avoid collisions with left-moving study participants, we might select a grid square  $(i, j)$  uniformly at random from Figure 6(a) and implement the rule:

$$(5.5) \quad \begin{array}{l} \text{if } \underbrace{x_{i,j}(t_k) = -1 \text{ and } x_{i,j+1}(t_k) = 1}_{\text{conditions for a head-on collision}} \text{ and } \underbrace{x_{i+1,j}(t_k) = 0}_{\text{space available}}, \\ \text{then } \underbrace{x_{i,j}(t_{k+1}) = 0 \text{ and } x_{i+1,j}(t_{k+1}) = 1}_{\text{pedestrian may step to the side}} \text{ with probability } p. \end{array}$$

In one time step, we could iterate through a random perturbation of all of the grid squares, implementing this and other model rules. There are many choices and parameters in Rule (5.5), including the choice of probability  $p$  and the choice of neighborhoods considered (e.g., why should the pedestrian at space  $(i, j)$  only look one grid step ahead to space  $(i, j + 1)$ ? Maybe  $(i, j + 2)$  is more appropriate?).

Microscopic off-lattice models (e.g., [HM95, HBJW05]), in comparison, assume that individuals move continuously in space; see Figure 6(g). In this case, movement is modeled through coupled ordinary or stochastic differential equations, for example, of the form:

$$(5.6) \quad \frac{d\mathbf{V}_i}{dt} = \underbrace{\mathbf{g}(\mathbf{X}_i, \mathbf{V}_i)}_{\text{pedestrian } i\text{'s inherent goals}} + \underbrace{\sum_{j=1}^N \mathbf{f}(\mathbf{X}_i, \mathbf{X}_j, \mathbf{V}_i, \mathbf{V}_j)}_{\text{interactions between pedestrians}}$$

$$(5.7) \quad \frac{d\mathbf{X}_i}{dt} = \mathbf{V}_i,$$

where  $\mathbf{X}_i(t)$  is the position of the  $i$ th pedestrian (e.g., a point mass marking the  $(x, y)$  coordinates of the pedestrian's center of mass) and  $\mathbf{V}_i(t)$  is that pedestrian's velocity. So called "social-force" models are a prominent off-lattice microscopic approach to pedestrian dynamics [HBJW05, HM95]. In both on-lattice and off-lattice models, arrival and exit of pedestrians from either side of the corridor in Figure 6(a) could take the form of stochastic rules. Computationally, we might assume that a new pedestrian enters the corridor at a randomly selected  $(x, y)$  position near the left or right edge of Figure 6(a) with probability  $\alpha \Delta t$ , where  $\Delta t$  is the time step of our simulations.

While microscopic models offer detailed perspectives on the behavior of individuals and can make experimentally testable predictions, they have many more parameters than macroscopic models do. In order to avoid overfitting and improve predictive value, it is thus important to break our data into separate sets for model development and testing. For example, we could fit the parameters in the functions in Equations (5.6)–(5.7) based on measurements of pedestrian–pedestrian distances and pedestrian velocities. We could specify the rates at which pedestrians enter the corridor based on empirical data, and we could use lane width to determine any unmeasurable parameters or guide the form of model rules. To test our model, we could set aside certain experiments (e.g., experiments with wider corridors) to simulate with our final model. We could, for example, use our validated model to predict how the dynamics will change when a pushier agent is introduced or when the structure of the barriers and walls in Figure 6(a)–(c) is changed.

Adding further difficulty, microscopic models are often stochastic and not analytically tractable, and they face some of the same challenges as qualitative data: how do we define and quantitatively describe the stripes in Figure 6(f)–(g) in an automated, objective way? To help address this challenge, topological techniques, especially persistent homology [OPT<sup>+</sup>17, EH08, Car20], have recently been combined with modeling to study complex systems, including aggregation [UZI19, TZH15]. Additional examples of topological data analysis applied to biological and social complex systems include [BMM<sup>+</sup>19, BCT20, MVS20, CJDM21, AQO<sup>+</sup>20, NSF<sup>+</sup>21] and [FHP22, HJJ<sup>+</sup>22], respectively. Pair correlation functions [DBG18, JC19, TSB<sup>+</sup>14] are another method for quantifying spatial data.

Depending on our goals and what our data suggest, building a hybrid, multiscale model that accounts for dynamics within pedestrians may be appropriate; see Figure 6(h). For example, in the off-lattice microscopic setting, we could introduce a variable  $P_i(t)$  that tracks how frustrated each individual is based on their perceived justness of the crowd dynamics around them. We could define  $P_i(t)$  by comparing the distance that pedestrian  $i$  has moved toward their goal in some time interval to the estimated distance that the individuals in a local neighborhood around  $i$  are moving. There are many other ways that we could define  $P_i(t)$ , and we could include feedback between  $P_i(t)$  and how pushy pedestrians choose to be, influencing our ODEs for movement in an associated agent-based model.

As a last example, similar to cellular Potts models in biology [GG93, HRM17], fine-grid cellular automaton represent each individual with a collection of grid squares (e.g., [SHT10]); see Figure 6(i). These detailed approaches are appropriate when folks are interested in the spatial extent of agents. Representing each pedestrian with  $N > 1$  grid squares, instead of just one as in Figure 6(f), increases the number of parameters and the time that it will take to simulate the model. This means fine-grid cellular automaton may make more sense when the goal is to describe the behavior of a few pedestrians in a detailed way; as we consider a larger crowd, agent-based or cellular automaton models become more appropriate; and, as we zoom out further into very large, densely packed crowds, macroscopic models are especially helpful. In crowd dynamics, as for other complex social systems, there are many useful modeling approaches that we could take, and it is a matter of choosing one that is parsimonious and appropriate for our goals.

## 6. Conclusions

I conclude with the best piece of advice that I have been given as a modeler: don't be afraid to be wrong. In particular, developing a model that correctly describes all of the unknown, intricate details of a complex social system would come down to sheer luck, since the space of possible models is huge. This can be discouraging. Instead, I have found it freeing to recognize that all of my models have been and will continue to be "wrong" in some sense. What matters is getting it "wrong" in a meaningful way. By building a parsimonious model, balancing our assumptions with the amount of data available, and designing a clear method for testing the model, we can make a meaningful contribution and generate new insights despite being inevitably "wrong" (or "right" in a simplified way). If the first model of a complex system does not cross disciplinary boundaries, it can lay the groundwork for a bridge that brings disciplines together in the future.

Whether our starting point is a rich data set or a nearly blank space, modeling complex systems is an iterative, creative, and interdisciplinary process. It involves being aware of the choices that we are making to simplify the problem, choosing model complexity based on our data, carefully considering the bias in the data and model, and identifying a plan for model building and validation. Through data collection, model development, prediction, communication, and generating new questions, we can push the field forward, help address societal challenges, develop mathematical approaches, and bring disciplines together in new ways.

## References

- [AA15] G Ariel and A Ayali, *Locust collective motion and its modeling*, PLOS Comput Biol **11** (2015), no. 12, e1004522.
- [AAA<sup>+</sup>21] T Alshaabi, J L Adams, M V Arnold, J R Minot, D R Dewhurst, A J Reagan, C M Danforth, and P S Dodds, *Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter*, Sci Adv **7** (2021), no. 29, eabe6534.
- [Abr88] A I Abramowitz, *An improved model for predicting presidential election outcomes*, PS Political Sci Politics **21** (1988), no. 4, 843–847.
- [Abr08] ———, *Forecasting the 2008 Presidential Election with the Time-for-Change Model*, PS Political Sci Politics **41** (2008), no. 4, 691–695.
- [Aka74] H Akaike, *A new look at the statistical model identification*, IEEE Trans Automat Contr **19** (1974), no. 6, 716–723.
- [Aka98] ———, *Information theory and an extension of the maximum likelihood principle*, Selected Papers of Hirotugu Akaike (E Parzen, K Tanabe, and G Kitagawa, eds.), Springer, New York, 1998, pp. 199–213.
- [AQO<sup>+</sup>20] E J Amézquita, M Y Quigley, T Ophelders, E Munch, and D H Chitwood, *The shape of things to come: Topological Data Analysis and biology, from molecules to organisms*, Dev Dyn **249** (2020), no. 7, 816–833.
- [BA01] V J Blue and J L Adler, *Cellular automata microsimulation for modeling bi-directional pedestrian walkways*, Transp Res B: Methodol **35** (2001), no. 3, 293–312.
- [BBG<sup>+</sup>22] R Best, A Bycoffe, C Groskopf, R King, E Koeze, D Mehta, J Mithani, M Radcliffe, A Wiederkehr, J Wolfe, A Jones-Rooy, N Rakich, D Shan, S Frostenson, J Mason, A Mangan, and C Yee, *FiveThirtyEight: Latest Polls*, <https://projects.fivethirtyeight.com/polls/>, 2022, Last accessed: 06-30-2022.
- [BBK13] N Bellomo, A Bellouquid, and D Knopoff, *From the microscale to collective crowd dynamics*, Multiscale Model Simul **11** (2013), no. 3, 943–963.
- [BCAKCC06] L M A Bettencourt, A Cintrón-Arias, D I Kaiser, and C Castillo-Chavéz, *The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models*, Physica A **364** (2006), 513–536.
- [BCC<sup>+</sup>08] M Ballerini, N Cabibbo, R Candelier, A Cavagna, E Cisbani, I Giardina, A Orlandi, G Parisi, A Procaccini, M Viale, and V Zdravkovic, *Empirical investigation of starling flocks: a benchmark study in collective animal behaviour*, Anim Behav **76** (2008), no. 1, 201–215.
- [BCC12] R Brauer and C Castillo-Chavez, *Mathematical Models in Population Biology and Epidemiology*, 2nd ed., Springer-Verlag, Heidelberg, Germany, 2012.
- [BCD18] R Bailo, J A Carrillo, and P Degond, *Pedestrian models based on rational behaviour*, Crowd Dynamics, Volume 1: Theory, Models, and Safety Problems (L Gibelli and N Bellomo, eds.), Springer, Cham, 2018, pp. 259–292.
- [BCG<sup>+</sup>16] N Bellomo, D Clarke, L Gibelli, P Townsend, and B J Vreugdenhil, *Human behaviours in evacuation crowd dynamics: From modelling to “big data” toward crisis management*, Phys Life Rev **18** (2016), 1–21.
- [BCME<sup>+</sup>20] A J Bernoff, M Culshaw-Maurer, R A Everett, M E Hohn, W C Strickland, and J Weinburd, *Agent-based and continuous models of hopper bands for the Australian plague locust: How resource consumption mediates pulse formation and geometry*, PLOS Comput Biol **16** (2020), no. 5, e1007820.
- [BCT20] L L Bonilla, A Carpio, and C Trenado, *Tracking collective cell motion by topological data analysis*, PLOS Comp Biol **16** (2020), no. 12, e1008407.
- [BD11] N Bellomo and C Dogbe, *On the modeling of traffic and crowds: A survey of models, speculations, and perspectives*, SIAM Rev **53** (2011), no. 3, 409–463.
- [BdA17] D Braha and M A M de Aguiar, *Voting contagion: Modeling and analysis of a century of U.S. presidential elections*, PLOS ONE **12** (2017), no. 5, e0177970.
- [BDM<sup>+</sup>09] T Bosse, R Duell, Z A Memon, J Treur, C N van der Wal, J Otamendi, A Bargiela, J L Montes, and L M D Peera, *A Multi-Agent Model for Mutual Absorption of*

- Emotions*, Proceedings of the 23rd European Conference on Modelling and Simulation (ECMS'09), European Council on Modeling and Simulation, 2009, pp. 212–218.
- [BEK20] A Buttenschön and L Edelstein-Keshet, *Bridging from single to collective cell migration: A review of models and links to experiments*, PLOS Comput Biol **16** (2020), no. 12, e1008411.
- [BFG14] K M Bliss, K R Fowler, and B J Galluzzo, *Math Modeling: Getting Started and Getting Solutions*, SIAM, Philadelphia, <https://m3challenge.siam.org/resources/modeling-handbook>, 2014.
- [BGBD<sup>+</sup>18] L Bonnasse-Gahot, H Berestycki, M-A Depuiset, M B Gordon, S Roché, N Rodriguez, and J-P Nadal, *Epidemiological modelling of the 2005 French riots: a spreading wave and the role of contagion*, Sci Rep **8** (2018), no. 107.
- [BHK<sup>+</sup>11] T Bosse, M Hoogendoorn, M C A Klein, J Treur, and C N van der Wal, *Agent-based analysis of patterns in crowd behaviour involving contagion of mental states*, Lecture Notes in Computer Science, vol. 6704, Springer, Berlin Heidelberg, 2011, pp. 566–577.
- [BHN<sup>+</sup>95] M Bando, K Hasebe, A Nakayama, A Shibata, and Y Sugiyama, *Dynamical model of traffic congestion and numerical simulation*, Phys Rev E **51** (1995), no. 2, 1035–1042.
- [BIR] *The University of British Columbia: BIRS Workshop Lecture Videos*, <https://open.library.ubc.ca/cIRcle/collections/48630>, Last accessed: 06-30-2022.
- [BK19] S L Brunton and J N Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, Cambridge University Press, Cambridge, United Kingdom, 2019.
- [BKGL18] K M Bliss, K F Kavanagh, B J Galluzzo, and R Levy, *Math Modeling: Computing and Communicating*, SIAM, Philadelphia, <https://m3challenge.siam.org/resources/modeling-handbook>, 2018.
- [BKSZ01] C Burstedde, K Klauck, A Schadschneider, and J Zittartz, *Simulation of pedestrian dynamics using a two-dimensional cellular automaton*, Physica A **295** (2001), no. 3, 507–525.
- [BMM<sup>+</sup>19] D Bhaskar, A Manhart, J Milzman, J T Nardini, K M Storey, C M Topaz, and L Ziegelmeier, *Analyzing collective motion with machine learning and topology*, Chaos **29** (2019), 123125.
- [Boc10] N Boccara, *Modeling complex systems*, 2nd ed. ed., Graduate Texts in Physics, Springer, New York, NY, 2010.
- [BP20a] H Z Brooks and M A Porter, *A model for the influence of media on the ideology of content in online social networks*, Phys Rev Research **2** (2020), no. 2, 023041.
- [BP20b] ———, *Topological Data Analysis and Data-Driven Modeling in Complex Systems: A Minisymposium in the 2020 SIAM Conference on Mathematics of Data Science*, [https://www.youtube.com/playlist?list=PLnzqyg\\_akFM1U4KVL0E5IlhyVjAsvJ20d](https://www.youtube.com/playlist?list=PLnzqyg_akFM1U4KVL0E5IlhyVjAsvJ20d), 2020, Last accessed: 06-30-2022.
- [BPK16] S L Brunton, J L Proctor, and J N Kutz, *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*, Proc Natl Acad Sci USA **113** (2016), no. 15, 3932–3937.
- [BPR18] A E Biondo, A Pluchino, and A Rapisarda, *Modeling surveys effects in political competitions*, Phys A: Stat Mech Appl **503** (2018), 714–726.
- [BR06] O Bandiera and I Rasul, *Social networks and technology adoption in northern Mozambique*, Econ J **116** (2006), no. 514, 869–902.
- [BR19] N W F Bode and E Ronchi, *Statistical model fitting and model selection in pedestrian dynamics research*, Collective Dynamics **4** (2019), 1–32.
- [Bro21] H Z Brooks, *Networks in Social Systems*, [https://zerodivzero.com/short\\_course/aaac8c66007a4d23a7aa14857a3b778c/title/5dd029b5e02146d1926c17d5184d8b63](https://zerodivzero.com/short_course/aaac8c66007a4d23a7aa14857a3b778c/title/5dd029b5e02146d1926c17d5184d8b63), 2021, Last accessed: 09-09-2022.
- [Bro22] D Brockmann, *Complexity Explorables*, <https://www.complexity-explorables.org>, 2022, Last accessed 06-17-2022.
- [BRSW15] A L Bertozzi, J Rosado, M B Short, and L Wang, *Contagion shocks in one dimension*, J Stat Phys **158** (2015), no. 3, 647–664.



- [Bru] S Brunton, *Data-Driven Dynamical Systems with Machine Learning*, <https://www.youtube.com/playlist?list=PLMrJAKhIeNNR6DzT17-MM1GHLkuYVjhyt>, Last accessed: 06-30-2022.
- [BT11] A J Bernoff and C M Topaz, *A primer of swarm equilibria*, SIAM J Appl Dyn Sys **10** (2011), no. 1, 212–250.
- [BV05] M Bodnar and J J L Velazquez, *Derivation of macroscopic equations for individual cell-based models: a formal approach*, Math Methods Appl Sci **28** (2005), no. 15, 1757–1779.
- [Car20] G Carlsson, *Topological methods for data modelling*, Nat Rev Phys **2** (2020), 697–708.
- [CDFMG<sup>+</sup>21] M Cinelli, G De Francisci Morales, A Galeazzi, W Quattrociocchi, and M Starnini, *The echo chamber effect on social media*, Proc Natl Acad Sci USA **118** (2021), no. 9, e2023301118.
- [CDM<sup>+</sup>07] Y Chuang, M R D’Orsogna, D Marthaler, A L Bertozzi, and L S Chayes, *State transitions and the continuum limit for a 2D interacting, self-propelled particle system*, Physica D **232** (2007), 33–47.
- [CFL09] C Castellano, S Fortunato, and V Loreto, *Statistical physics of social dynamics*, Rev Mod Phys **81** (2009), no. 2, 591–646.
- [CFPSS19] W Cota, S C Ferreira, R Pastor-Satorras, and M Starnini, *Quantifying echo chamber effects in information spreading over political communication networks*, EPJ Data Science **8** (2019), no. 35, 1–13.
- [CJDM21] M-V Ciocanel, R Juenemann, A T Dawes, and S A McKinley, *Topological data analysis approaches to uncovering the timing of ring structure onset in filamentous networks*, Bull Math Biol **83** (2021), no. 21.
- [CKJ<sup>+</sup>02] I D Couzin, J Krause, R James, G D Ruxton, and N R Franks, *Collective memory and spatial sorting in animal groups*, J Theor Biol **218** (2002), no. 1, 1–11.
- [CMW16] J A Carrillo, S Martin, and M-T Wolfram, *An improved version of the Hughes model for pedestrian flow*, Math Models Methods Appl Sci **26** (2016), no. 04, 671–697.
- [CP21] J A Carrillo and Choi Y P, *Mean-field limits: From particle descriptions to macroscopic equations*, Arch Ration Mech Anal **241** (2021), no. 3, 1529–1573.
- [CR13] J D Clinton and S Rogers, *Robo-polls: Taking cues from traditional sources?*, PS Political Sci Politics **46** (2013), no. 2, 333–337.
- [Cre16] K Cressman, *Chapter 4.2 - desert locust*, Biological and Environmental Hazards, Risks, and Disasters (J F Shroder and R Sivanpillai, eds.), Academic Press, Boston, 2016, pp. 87–105.
- [CTS<sup>+</sup>20] M-V Ciocanel, C M Topaz, R Santorella, S Sen, C M Smith, and A Hufstetler, *JUSTFAIR: Judicial System Transparency through Federal Archive Inferred Records*, PLOS ONE **15** (2020), no. 10, e0241381.
- [CY16] P Cihon and T Yasseri, *A Biased Review of Biases in Twitter Studies on Political Collective Action*, Front Phys **4** (2016).
- [DAB<sup>+</sup>20] S Dodson, B Abrahms, S J Bograd, J Fiechter, and E L Hazen, *Disentangling the biotic and abiotic drivers of emergent migratory behavior using individual-based models*, Ecol Model **432** (2020), 109225.
- [DBC<sup>+</sup>19] M De Domenico, D Brockmann, C Camargo, C Gershenson, D Goldsmith, S Jeschonnek, L Kay, S Nichele, J R Nicolás, T Schmickl, M Stella, J Brandoff, A J Martínez Salinas, and H Sayama, *Complexity Explained*, <https://complexityexplained.github.io>, 2019, Last accessed 10-15-2022.
- [DBG18] S Dini, B J Binder, and J E F Green, *Understanding interactions between populations: Individual based modelling and quantification using pair correlation functions*, J Theor Biol **439** (2018), 50–64.
- [DCBC06] M R D’Orsogna, Y L Chuang, A L Bertozzi, and L S Chayes, *Self-propelled particles with soft-core interactions: Patterns, stability, and collapse*, Phys Rev Lett **96** (2006), 104302.
- [DDH13] D C Duives, W Daamen, and S P Hoogendoorn, *State-of-the-art crowd motion simulation models*, Transp Res C: Emerg Technol **37** (2013), 193–209.

- [DeF21] D R DeFord, *Python Tutorial on Networks*, [https://zerodivzero.com/short\\_course/aaac8c66007a4d23a7aa14857a3b778c/title/628602c8994746e491872a9380676b62](https://zerodivzero.com/short_course/aaac8c66007a4d23a7aa14857a3b778c/title/628602c8994746e491872a9380676b62), 2021, Last accessed: 09-09-2022.
- [DH00] O Diekmann and J A P Heesterbeek, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*, John Wiley & Sons, Inc., New York City, NY, USA, 2000.
- [DNAW00] G Deffuant, D Neau, F Amblard, and G Weisbuch, *Mixing beliefs among interacting agents*, Adv Complex Syst **3** (2000), no. 1n04, 87–98.
- [Dog10] C Dogbé, *Modeling crowd dynamics by the mean-field limit approach*, Math Comput Model **52** (2010), no. 9, 1506–1520.
- [EF18] T Evans and F Fu, *Opinion formation on dynamic networks: identifying conditions for the emergence of partisan echo chambers*, Royal Soc Open Sci **5** (2018), 181122.
- [eGH20] *The Economist*, A Gelman, and M Heidemanns, *Forecasting the US elections: How The Economist presidential forecast works*, <https://projects.economist.com/us-2020-forecast/president/how-this-works>, 2020, Last accessed: 09-07-2022.
- [EH08] H Edelsbrunner and J Harer, *Persistent homology – a survey*, Contemp Math **453** (2008), 257–282.
- [epS11] epSos.de, *Driving cars in a traffic jam*, Available at Wikimedia Commons: [https://commons.wikimedia.org/wiki/File:Driving\\_Cars\\_in\\_a\\_Traffic\\_Jam.jpg](https://commons.wikimedia.org/wiki/File:Driving_Cars_in_a_Traffic_Jam.jpg), 2011, Accessed 10-12-2022.
- [FGSR<sup>+</sup>14] J Fernández-Gracia, K Suchecki, J J Ramasco, M San Miguel, and V M Eguiluz, *Is the Voter Model a Model for Voters?*, Phys Rev Lett **112** (2014), 158701.
- [FHP22] M Feng, A Hickok, and M A Porter, *Topological data analysis of spatial systems*, Higher-Order Systems: Understanding Complex Systems (F Battiston and G Petri, eds.), Springer, Cham, 2022, pp. 389–399.
- [FPS<sup>+</sup>21] S L Franconeri, L M Padilla, P Shah, J M Zacks, and J Hullman, *The science of visual data communication: What works*, Psychol Sci Public Interest **22** (2021), no. 3, 110–161.
- [FTW18] A Festa, A Tosin, and M Wolfram, *Kinetic description of collision avoidance in pedestrian crowds by sidestepping*, Kinet Relat Models **11** (2018), no. 3, 491.
- [GAI19] *GAIMME: Guidelines for Assessment and Instruction in Mathematical Modeling Education, Second Edition*, S Garfunkel and M Montgomery (eds.), COMAP and SIAM, Philadelphia, <https://m3challenge.siam.org/resources/teaching-modeling>, 2019.
- [GBC18] D Guilbeault, J Becker, and D Centola, *Complex contagions: A decade in review*, Complex Spreading Phenomena in Social Systems. Computational Social Sciences (S Lehmann and Y Y Ahn, eds.), Springer, Cham, 2018, pp. 3–25.
- [GBKM20] R Giniūnaitė, R E Baker, P M Kulesa, and P K Maini, *Modelling collective cell migration: neural crest as a model paradigm*, J Math Biol **80** (2020), 481–504.
- [GG93] J A Glazier and F Graner, *Simulation of the differential adhesion driven rearrangement of biological cells*, Phys Rev E **47** (1993), 2128–2154.
- [GHWM20] A Gelman, J Hullman, C Wlezien, and G E Morris, *Information, incentives, and goals in election forecasts*, Judgm Decis Mak **15** (2020), no. 5, 863–880.
- [GK93] A Gelman and G King, *Why are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable?*, Br J Political Sci **23** (1993), 409–451.
- [HB JW05] D Helbing, L Buzna, A Johansson, and T Werner, *Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions*, Transp Sci **39** (2005), no. 1, 1–24.
- [Het00] H W Hethcote, *The Mathematics of Infectious Diseases*, SIAM Rev **42** (2000), no. 4, 599–653.
- [HJ09] D Helbing and A Johansson, *Pedestrian, crowd and evacuation dynamics*, Encyclopedia of Complexity and Systems Science (R A Meyers, ed.), Springer New York, 2009, pp. 6476–6495.
- [HJJ<sup>+</sup>22] A Hickok, B Jarman, M Johnson, J Luo, and M A Porter, *Persistent homology for resource coverage: A case study of access to polling sites*, <https://arxiv.org/abs/2206.04834>, 2022.

- [HK02] R Hegselmann and U Krause, *Opinion dynamics and bounded confidence: Models, analysis and simulation*, J Artif Soc Soc Simul **5** (2002), no. 3.
- [HLW16] J Humpherys, R Levy, and T Witelski, *Directions for Graduate and Undergraduate Modeling Courses*, <https://www.pathlms.com/siam/courses/3028/sections/4132>, 2016, Last accessed 10-13-2022.
- [HM95] D Helbing and P Molnár, *Social force model for pedestrian dynamics*, Phys Rev E **51** (1995), 4282–4286.
- [HR14] P Hummel and D Rothschild, *Fundamental models for forecasting elections at the state level*, Elect Stud **35** (2014), 123–139.
- [HRM17] T Hirashima, E G Rens, and R M H Merks, *Cellular Potts modeling of complex multicellular behaviors in tissue morphogenesis*, Dev Growth Differ **59** (2017), no. 5, 329–339.
- [Huf22a] *HuffPost Pollster*, <https://elections.huffingtonpost.com/pollster>, (2022), Last accessed: 10-16-2022.
- [Huf22b] *HuffPost Pollster API v2*, <https://elections.huffingtonpost.com/pollster/api/v2>, 2022, Last accessed: 10-16-2022.
- [IPBL19] I Iacopini, G Petri, A Barrat, and V Latora, *Simplicial models of social contagion*, Nat Commun **10** (2019), no. 1, 2485.
- [Jac05] S Jackman, *Pooling the polls over an election campaign*, Aust J Political Sci **40** (2005), no. 4, 499–517.
- [JC19] S T Johnston and E J Crampin, *Corrected pair correlation functions for environments with obstacles*, Phys Rev E **99** (2019), 032124.
- [JHZ<sup>+</sup>14] R Jiang, M-B Hu, H M Zhang, Z-Y Gao, B Jia, Q-S Wu, B Wang, and M Yang, *Traffic experiment reveals the nature of car-following*, PLOS ONE **9** (2014), no. 4, e94351.
- [KBBP16] J N Kutz, S L Brunton, B W Brunton, and J L Proctor, *Dynamic mode decomposition: data-driven modeling of complex systems*, Other Titles in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 2016.
- [KBF17] J Kursawe, R E Baker, and A G Fletcher, *Impact of implementation choices on quantitative predictions of cell-based computational models*, J Comp Phys **345** (2017), 752–767.
- [KBT<sup>+</sup>22] F P Kemeth, T Bertalan, T Thiem, F Dietrich, S J Moon, C R Laing, and I G Kevrekidis, *Learning emergent partial differential equations in a learned emergent space*, Nat Commun **13** (2022), no. 1, 3318–3318.
- [KHB13] A Kneidl, D Hartmann, and A Borrmann, *A hybrid multi-scale approach for simulation of pedestrian dynamics*, Transp Res C: Emerg Technol **37** (2013), 223–237.
- [KM27] W O Kermack and A G McKendrick, *A contribution to the mathematical theory of epidemics*, Proc R Soc London **115** (1927), no. 772, 700–721.
- [KM32] ———, *Contributions to the mathematical theory of epidemics. II. –The problem of endemicity*, Proc R Soc London **138** (1932), no. 834, 55–83.
- [KM33] ———, *Contributions to the mathematical theory of epidemics. III. –Further studies of the problem of endemicity*, Proc R Soc London **141** (1933), no. 843, 94–122.
- [KS20] M Kalt and D Scott, *Twitter Data Curation Primer: Data Curation Network GitHub Repository*, <https://github.com/DataCurationNetwork/data-primers/blob/master/Twitter%20Data%20Curation%20Primer/twitter-data-curation-primer.md>, 2020, Last accessed: 10-16-2022.
- [KTI<sup>+</sup>11] Y Katz, K Tunstrom, C C Ioannou, C Huepe, and I D Couzin, *Inferring the structure and dynamics of interactions in schooling fish*, Proc Natl Acad Sci USA **108** (2011), no. 46, 18720–18725.
- [Kut] N Kutz, *Nathan Kutz Videos*, <https://www.youtube.com/c/NathanKutzAMATH/videos>, Last accessed: 06-30-2022.
- [Kut13] J N Kutz, *Data-Driven Modeling & Scientific Computation: Methods for Complex Systems & Big Data*, Oxford University Press, Oxford, England, 2013.
- [LBBH98] Y Lecun, L Bottou, Y Bengio, and P Haffner, *Gradient-based learning applied to document recognition*, Proc IEEE **86** (1998), no. 11, 2278–2324.
- [Lin13] D A Linzer, *Dynamic Bayesian Forecasting of Presidential Elections in the States*, J Am Stat Assoc **108** (2013), no. 501, 124–134.

- [LLEK10] R Lukeman, Y-X Li, and L Edelstein-Keshet, *Inferring individual rules from collective behavior*, Proc Natl Acad Sci USA **107** (2010), no. 28, 12576–12580.
- [Loc] *Plague of Locusts Timelapse — Wild Africa — BBC Earth*, <https://www.youtube.com/watch?v=1YNy2R3hg2Q>, Uploaded by BBC Earth on 08-21-2009. Last accessed 06-17-2022.
- [LRC01] H Levine, W J Rappel, and I Cohen, *Self-organization in systems of self-propelled particles*, Phys Rev E **63** (2001), 017101.
- [LSZ15] K Law, A Stuart, and K Zygalakis, *Data Assimilation: A Mathematical Introduction*, Texts in Applied Mathematics, vol. 63, Springer, Cham, 2015.
- [LW11] A Lachapelle and M-T Wolfram, *On a mean field game approach modeling congestion and aversion in pedestrian crowds*, Transp Res B: Methodol **45** (2011), no. 10, 1572–1589.
- [LYY18] S Lehmann and A Yong-Yeol, *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*, Springer, Cham, 2018.
- [MBN<sup>+</sup>16] Mirrorme22, Brythones, Nilfanion, TUBS, and Sting, *United Kingdom EU referendum 2016 area*, Available at Wikimedia Commons: [https://commons.wikimedia.org/wiki/File:United\\_Kingdom\\_EU\\_referendum\\_2016\\_area\\_results.svg](https://commons.wikimedia.org/wiki/File:United_Kingdom_EU_referendum_2016_area_results.svg), 2016, Accessed 10-12-2022.
- [MBPK16] N M Mangan, S L Brunton, J L Proctor, and J N Kutz, *Inferring biological networks by sparse identification of nonlinear dynamics*, IEEE Trans Mol Biol Multi-Scale Commun **2** (2016), no. 1, 52–63.
- [MCM<sup>+</sup>22] J R Minot, N Cheney, M Maier, D C Elbers, C M Danforth, and P S Dodds, *Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance*, ACM Trans Comput Healthcare (2022), Just Accepted.
- [MEK99] A Mogilner and L Edelstein-Keshet, *A non-local model for a swarm*, J Math Biol **38** (1999), 534–570.
- [Mit09] M Mitchell, *Complexity: A Guided Tour*, Oxford University Press, 2009.
- [MJ] M Madhav and E Johnson, *What do Your Data Say? A course to help you better understand your data*, <https://www.whatdoyourdatasay.com>, Last accessed: 06-30-2022.
- [MKBP17] N M Mangan, J N Kutz, S L Brunton, and J L Proctor, *Model selection for dynamical systems via sparse regression and information criteria*, Proc R Soc A **473** (2017), 20170009.
- [MP07] J H Miller and S E Page, *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*, Princeton Studies in Complexity, Princeton University Press, Princeton, 2007.
- [MPLC14] F Morstatter, J Pfeffer, H Liu, and K M Carley, *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose*, Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2014, pp. 400–408.
- [MVS20] M R McGuirl, A Volkening, and B Sandstede, *Topological data analysis of zebrafish patterns*, Proc Natl Acad Sci USA **117** (2020), no. 10, 5113–5124.
- [NBSF21] J T Nardini, R E Baker, M J Simpson, and K B Flores, *Learning differential equation models from stochastic agent-based model simulations*, J R Soc Interface **18** (2021), 20200987.
- [New10] M E J Newman, *Networks: An Introduction*, Oxford University Press, Oxford, 2010.
- [New11] ———, *Complex Systems: A Survey*, Am J Phys **79** (2011), no. 8, 800–810.
- [NS92] K Nagel and M Schreckenberg, *A cellular automaton model for freeway traffic*, J Phys, I **2** (1992), no. 12, 2221–2229.
- [NSF<sup>+</sup>21] J T Nardini, B J Stolz, K B Flores, H A Harrington, and H M Byrne, *Topological data analysis distinguishes parameter regimes in the Anderson-Chaplain model of angiogenesis*, PLOS Comput Biol **17** (2021), e1009094.
- [OPT<sup>+</sup>17] N Otter, M A Porter, U Tillmann, P Grindrod, and H A Harrington, *A roadmap for the computation of persistent homology*, EPJ Data Science **6** (2017), no. 17.
- [OT12] E Oster and R Thornton, *Determinants of technology adoption: peer effects in menstrual cup take-up*, J Eur Econ Assoc **10** (2012), no. 6, 1263–1293.

- [PEK99] J K Parrish and L Edelstein-Keshet, *Complexity, pattern, and evolutionary trade-offs in animal aggregation*, Science **284** (1999), no. 5411, 99–101.
- [PG16] M A Porter and J P Gleeson, *Dynamical Systems on Networks: A Tutorial*, Front Appl Dyn Syst Rev Tutor **4** (2016).
- [PMS17] H Peng, F Menczer, and K Sasahara, *EchoDemo: How echo chambers emerge from social media*, <https://osome.iu.edu/demos/echo/>, 2017, Last accessed 06-17-2022.
- [Por22] M A Porter, *A non-expert’s introduction to data ethics for mathematicians*, <https://arxiv.org/abs/2201.07794>, 2022.
- [PT03] D M Parichy and J M Turner, *Temporal and cellular requirements for Fms signaling during zebrafish adult pigment pattern development*, Development **130** (2003), no. 5, 817–833.
- [QSI] *Institute for the Quantitative Study of Inclusion, Diversity, and Equity*, <https://qsideinstitute.org>, Last accessed: 06-30-2022.
- [Rea22] *RealClearPolitics: Polls*, [https://www.realclearpolitics.com/epolls/latest\\_polls/elections/](https://www.realclearpolitics.com/epolls/latest_polls/elections/), (2022), Last accessed: 10-13-2022.
- [RRH09] J M Restrepo, R C Rael, and J M Hyman, *Modeling the influence of polls on elections: a population dynamics approach*, Public Choice **140** (2009), 395–420.
- [SCDM<sup>+</sup>18] R E Stern, S Cui, M L Delle Monache, R Bhadani, M Bunting, M Churchill, N Hamilton, R Haulcy, H Pohlmann, F Wu, B Piccoli, B Seibold, J Sprinkle, and D B Work, *Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments*, Transp Res Part C Emerg Technol **89** (2018), 205–221.
- [Sch73] T C Schelling, *Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities*, J Conflict Resolut **17** (1973), no. 3, 381–428.
- [Sch78] G Schwarz, *Estimating the dimension of a model*, Ann Stat **6** (1978), no. 2, 461–464.
- [SCP<sup>+</sup>21] K Sasahara, W Chen, H Peng, G L Ciampaglia, A Flammini, and F Menczer, *Social influence and unfollowing accelerate the emergence of echo chambers*, J Comput Soc Sci **4** (2021), 381–402.
- [SCS<sup>+</sup>18] A Schadschneider, M Chraïbi, A Seyfried, A Tordeux, and J Zhang, *Pedestrian dynamics: From empirical results to modeling*, pp. 63–102, Springer, Cham, 2018.
- [SEK13] L A Segel and L Edelstein-Keshet, *A Primer on Mathematical Models in Biology*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.
- [SFK<sup>+</sup>08] Y Sugiyama, M Fukui, M Kikuchi, K Hasebe, A Nakayama, K Nishinari, S Tadaki, and S Yukawa, *Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam*, New J Phys **10** (2008), 033001.
- [SFK16] J Schleuss, J Fox, and P Krishnakumar, *California 2016 election precinct maps*, <https://github.com/datadesk/california-2016-election-precinct-maps>, (2016), Accessed: 03-14-2019.
- [Shm10] G Shmueli, *To Explain or to Predict?*, Statist Sci **25** (2010), no. 3, 289–310.
- [SHP16] B J Stolz, H A Harrington, and M A Porter, *The Topological “Shape” of Brexit*, <https://arxiv.org/abs/1610.00752>, 2016.
- [SHT10] S Sarmady, F Haron, and A Z Talib, *Simulating crowd movements using fine grid cellular automata*, 2010 12th International Conference on Computer Modelling and Simulation, IEEE, 2010, pp. 428–433.
- [Sil12] N Silver, *The signal and the noise: Why so many predictions fail — but some don’t*, Penguin Press, New York City, NY, USA, 2012.
- [Sil14] ———, *FiveThirtyEight: Here’s Proof Some Pollsters Are Putting A Thumb On The Scale*, <https://fivethirtyeight.com/features/heres-proof-some-pollsters-are-putting-a-thumb-on-the-scale/>, 2014, Last accessed 06-17-2022.
- [SMA20] D Sabin-Miller and D M Abrams, *When pull turns to shove: A continuous-time model for opinion dynamics*, Phys Rev Researchf **2** (2020), 043001.
- [Smi14] R C Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*, Computational science & engineering series, Society for Industrial and Applied Mathematics, Philadelphia, 2014.
- [SNV15] A P Singh and C Nüsslein-Volhard, *Zebrafish stripes as a model for vertebrate colour pattern formation*, Curr Biol **25** (2015), R81–R92.

- [SSS17] A Sieben, J Schumann, and A Seyfried, *Collective phenomena in crowds—Where pedestrian dynamics need social psychology*, PLOS ONE **12** (2017), no. 6, e0177328.
- [Sto] *storywrangler: From the University of Vermont Computational Story Lab*, <https://storywrangling.org>, Last accessed: 06-30-2022.
- [Str01] S H Strogatz, *Exploring complex networks*, Nature **410** (2001), no. 6825, 268–276.
- [Str15] ———, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, 2nd ed. ed., CRC Press, 2015.
- [TBL06] C M Topaz, A L Bertozzi, and M A Lewis, *A nonlocal continuum model for biological aggregation*, Bull Math Biol **68** (2006), 1601–1623.
- [TECP20] J H Tien, M C Eisenberg, S T Cherng, and M A Porter, *Online reactions to the 2017 ‘Unite the right’ rally in Charlottesville: measuring polarization in Twitter networks using media followership*, Appl Netw Sci **5** (2020), no. 10.
- [TFB<sup>+</sup>11] J Tsai, N Fridman, E Bowring, M Brown, S Epstein, G Kaminka, S Marsella, A Ogden, I Rika, A Sheel, M Taylor, X Wang, A Zilka, and M Tambe, *ESCAPES - Evacuation simulation with children, authorities, parents, emotions, and social comparison*, vol. 1, 2011, pp. 457–464.
- [THK18] S Thurner, R Hanel, and P Klimek, *Introduction to the Theory of Complex Systems*, Oxford University Press, 2018.
- [Tib96] R Tibshirani, *Regression shrinkage and selection via the lasso*, J R Stat Soc Ser B Methodol **58** (1996), no. 1, 267–288.
- [TSB<sup>+</sup>14] K K Treloar, M J Simpson, B J Binder, D L S McElwain, and R E Baker, *Assessing the role of spatial correlations during collective cell spreading*, Sci Rep **4** (2014), no. 1, 5713.
- [Tuf14] Z Tufekci, *Big questions for social media big data: Representativeness, validity and other methodological pitfalls*, Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 2014, pp. 505–514.
- [Twi] *Twitter Developer Platform: Tutorials*, <https://developer.twitter.com/en/docs/tutorials>, Last accessed: 06-30-2022.
- [TZH15] C M Topaz, L Ziegelmeier, and T Halverson, *Topological data analysis of biological aggregation models*, PLOS ONE **10** (2015), no. 5, e0126383.
- [UZT19] M Ulmer, L Ziegelmeier, and C M Topaz, *A topological approach to selecting models of biological experiments*, PLOS ONE **14** (2019), no. 3, e0213679.
- [VCBJ<sup>+</sup>95] T Vicsek, A Czirók, E Ben-Jacob, I Cohen, I, and O Shochet, *Novel type of phase transition in a system of self-driven particles*, Phys Rev Lett **75** (1995), no. 6, 1226–1229.
- [VCM<sup>+</sup>07] A Varas, M D Cornejo, D Mainemer, B Toledo, J Rogan, V Muñoz, and J A Valdivia, *Cellular automaton model for evacuation process with obstacles*, Physica A **382** (2007), no. 2, 631–642.
- [VLPR20] A Volkening, D F Linder, M A Porter, and G A Rempala, *Forecasting elections using compartmental models of infection*, SIAM Rev **62** (2020), no. 4, 837–865.
- [Vol17] A Volkening, *Modeling pattern formation on zebrafish*, Ph.D. thesis, Brown University, 2017.
- [Vol20a] A Volkening, *Intro to building models*, <https://northwestern.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7d04a874-a292-4ff2-bd66-ac2500daeeaa1>, 2020, Last accessed: 06-30-2022.
- [Vol20b] A Volkening, *Linking genotype, cell behavior, and phenotype: multidisciplinary perspectives with a basis in zebrafish patterns*, Curr Opin Genet Dev **63** (2020), 78–85.
- [Vol21] ———, *Data-Driven Modeling*, [https://zerodivzero.com/short\\_course/aaac8c66007a4d23a7aa14857a3b778c/title/d56faebff3a24f77a76085c1427038d8](https://zerodivzero.com/short_course/aaac8c66007a4d23a7aa14857a3b778c/title/d56faebff3a24f77a76085c1427038d8), 2021, Last accessed: 06-30-2022.
- [VS15] A Volkening and B Sandstede, *Modelling stripe formation in zebrafish: an agent-based approach*, J R Soc Interface **12** (2015), no. 112, 20150812.
- [VS18] ———, *Iridophores as a source of robustness in zebrafish stripes and variability in Danio patterns*, Nat Commun **9** (2018), no. 3231.
- [WE02] C Wlezien and R S Erikson, *The Timeline of Presidential Election Campaigns*, J Politics **64** (2002), no. 4, 969–993.

- [WPCG<sup>+</sup>14] C H Weiss, J Poncela-Casasnovas, J I Glaser, A R Pah, S D Persell, D W Baker, R G Wunderink, and L A Nunes Amaral, *Adoption of a high-impact innovation in a homogeneous population*, Phys Rev X **4** (2014), no. 4, 041008.
- [YAKM20] V C Yang, D M Abrams, G Kernell, and A E Motter, *Why Are U.S. Parties So Polarized? A “Satisficing” Dynamical Model*, SIAM Rev **62** (2020), no. 3, 646–657.
- [ZKSS12] J Zhang, W Klingsch, A Schadschneider, and A Seyfried, *Ordering in bidirectional pedestrian flows and its influence on the fundamental diagram*, J Stat Mech **2012** (2012), no. 02, P02002.

DEPARTMENT OF MATHEMATICS, PURDUE UNIVERSITY, WEST LAFAYETTE, IN 47907 USA  
*Email address:* [avolkening@purdue.edu](mailto:avolkening@purdue.edu)