Scientific Discovery, Process Models, and the Social Sciences

Pat Langley and Adam Arvay

Abstract In this chapter, we review research on computational approaches to scientific discovery, starting with early work on the induction of numeric laws before turning to the construction of models that explain observations in terms of domain knowledge. We focus especially on inductive process modeling, which involves finding a set of linked differential equations, organized into processes, that reproduce, predict, and explain multivariate time series. We review the notion of quantitative process models, present two approaches to their construction that search through a space of model structures and associated parameters, and report their successful application to the explanation of ecological data. After this, we explore the relevance of process models to the social sciences, including the reasons they seem appropriate and some challenges to discovering them. In closing, we discuss other causal frameworks, including structural equation models and agent-based accounts, that researchers have developed to construct models of social phenomena.

1 Introduction

The scientific enterprise is a diverse collection of activities that is distinguished from other areas of human endeavor by a number of important characteristics. These include the systematic collection and analysis of observations, the formal statement of theories, laws, and models, invoking those theories and models to explain and predict observations, and using those observations in turn to evaluate theories and models. Within this broad endeavor, a central element – *scientific discovery* – is

Pat Langley
Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306 USA,
e-mail: patrick.w.langley@gmail.com

Adam Arvay
Department of Computer Science, University of Auckland, Private Bag 92019, Auckland 1142,
New Zealand, e-mail: aarv914@aucklanduni.ac.nz

widely viewed as one of the highest forms of human achievement. Clearly, computational insights into discovery processes would have important implications, both theoretical and practical.

This holds especially in the *system sciences*, like ecology, that examine many interacting variables, develop complex models with feedback loops, and typically lack experimental control. Such fields are arguably less advanced that disciplines like biology, chemistry, and physics because they are more complicated and present so many difficulties. The latter have developed more precise and complete accounts not because they are 'hard' sciences, but rather because they are 'easy' disciplines that involve relatively simple phenomena and lend themselves to experimental study. In contrast, the system sciences – including those focused on social phenomena – stand to benefit more from a deeper understanding of scientific discovery and tools that might aid them.

In this chapter, we examine the problem of model construction in such disciplines, with special emphasis on the social sciences. We review efforts to understand the discovery process in computational terms, recounting its progress and successes over the past 35 years. Most research in the area has focused on induction of empirical laws, but some work has addressed construction of explanatory models that are more relevant to the social sciences. We examine one paradigm – *inductive process modeling* – in some detail, as it deals with discovery of explanatory models from multivariate time series in the absence of experimental control. Research in this framework has concentrated on ecological phenomena, but we argue that it also lends itself nicely to social science problems, presenting examples to support this claim. In closing, we discuss challenges to the use of inductive process modeling in this arena, along with other approaches to constructing explanatory accounts in the social sciences.

2 Computational Scientific Discovery

There is no question that discovery plays a central in the scientific enterprise. Science could not proceed without using laws and models to generate testable predictions. Neither could it make progress without collecting observations that let it evaluate those laws and models. But discovery brings these two activities together by generating new laws and models, or revising existing ones, in response to anomalies and surprising results. Prizes in science are typically awarded for researchers' role in discovering new laws or mechanisms, not for making predictions or collecting data, even when the latter lead to the former.

However, the philosophy of science, which attempts to understand and formalize scientific structures and processes, has generally avoided addressing issues related to discovery. The standard argument has been that this important activity is immune to logical analysis. For example, Popper (1961) wrote:

> The initial stage, the act of conceiving or inventing a theory, seems to me neither to call for logical analysis nor to be susceptible of it . . . My view may be expressed by saying that every discovery contains an 'irrational element', or 'a creative intuition' . . .

Many others adopted Popper's position in the decades that followed. Hempel (1966) and other scholars also maintained that discovery was inherently irrational and thus beyond understanding, at least in terms that philosophers accepted. This mystical attitude toward discovery hindered progress on the topic for many years.

Yet advances made by two fields – cognitive psychology and artificial intelligence – in the 1950s suggested a path forward. In particular, Simon (1966) proposed that scientific discovery is a variety of problem solving that involves search through a space of problem states – in this case laws, models, or theories. These structures are generated by applying mental operators to previous ones, with search through the space of candidates guided by heuristics to make it tractable. Heuristic search had been implicated in many cases of human cognition, from proving theorems to playing chess. This framework offered not only an approach to understanding scientific discovery, but also ways to automate this mysterious process.

Early research in this area focused on discovery of empirical laws. For instance, BACON (Langley, 1981; Langley, Simon, Bradshaw, & Zytkow, 1987) carried out search through a problem space of algebraic terms, using operators like multiplication and division to generate new terms from old ones. The system was guided by heuristics that noted regularities in data, one of which was defining a new product term, $x \cdot y$, when it found that x increased as y decreased. In many cases, this led BACON toward higher-level terms with constant values, as in its rediscovery of Kepler's third law of planetary motion. The system also applied this idea recursively, treating parameters at lower levels as variables at higher ones, to formulate complex relations, as in its rediscovery of the idea gas law. BACON reconstructed a variety of numeric relations from the history of physics and chemistry, lending evidence to Simon's claim that heuristic search underlies scientific discovery.

Responses to the BACON work were mixed, with some critics viewing it as clarifying important aspects of the discovery process. But others claimed that the 'real' key to discovery instead lay in other activities the system did not address, such as deciding which variables to measure and relate, determining which problem space to search, or selecting which scientific problem to pursue. Others held that BACON 'only did what it was programmed to do', and thus did not really 'discover' anything. The system's developers only claimed that it offered insights into the operation of scientific discovery and removed much of its mystery, but they acknowledged that it offered only a partial account and that much work remained to be done.

Later research took inspiration from the initial BACON results and developed more powerful approaches to discovering numeric laws. Papers by Falkenhainer and Michalski (1986), Moulet (1992), Zytkow, Zhu, and Hussam (1990), Kokar (1986), Schaffer (1990), Nordhausen and Langley (1990), Gordon et al. (1994), Murata, Mizutani, and Shimura (1994), Washio and Motoda (1997), and Saito and Nakano (1997) all reported systems that induced algebraic laws describing relations among quantitative variables, many of them reproducing results from the history of science. Other work – by Džeroski and Todorovski (1995), Bradley, Easley, and Stolle

(2001), Koza et al. (2001), Todorovski and Džeroski (2008), and Schmidt and Lipson (2009) – extended this idea to differential equations for dynamic systems, often focusing on novel scientific data. These relied on different methods but also searched for explicit mathematical laws that matched data. Interest in computational discovery spread to other aspects of science, including induction of qualitative laws and construction of explanatory models (Shrager & Langley, 1990; Džeroski & Todorovski, 2007).

Although early work in this area emphasized reconstructions of discoveries from the history of science, later efforts went on to help to generate new knowledge in many scientific fields. These success stories have included discovery of:

- reaction pathways in catalytic chemistry (Valdes-Perez, 1994; Bruk et al., 1998);
- qualitative chemical factors in mutagenesis (King & Srinivasan, 1996);
- quantitative laws of metallic behavior (Mitchell et al., 1997);
- quantitative conjectures in graph theory (Fajtlowicz, 1988);
- qualitative conjectures in number theory (Colton, Bundy, & Walsh, 2000);
- temporal laws of ecological behavior (Todorovski, Džeroski, & Kompare, 1998);
- models of gene-influenced metabolism in yeast (King et al., 2004).

In each case, the authors published their findings in the refereed literature of the relevant scientific field, showing that experts in the discipline viewed them as relevant and interesting. Langley (2000) has analyzed a number of these successes, identifying elements in each case that were handled by computer and others that were done manually. Together, these results demonstrate convincingly that we can automate important aspects of the scientific process.

Before continuing, we should distinguish between research on computational scientific discovery and another research paradigm – *data mining* – that emerged in the 1990s, which also aims to find regularities in observations and uses heuristic search through a space of hypotheses to this end. Work in this area typically focuses on commercial applications and emphasizes the availability of large data sets. Most data-mining research has adopted notations invented by computer scientists, such as decision trees and Bayesian networks, in marked contrast to computational scientific discovery, which focuses on formalisms used by domain scientists. Data-mining methods have been applied to scientific data (e.g., Fayyad, Haussler, & Stolorz, 1996), but the results seldom bear a resemblance to scientific knowledge.

3 Discovery of Explanatory Process Models

The early stages of science typically focus on finding descriptive laws that summarize empirical regularities. This is understandable, as paucity of knowledge about the area encourages inductive inquiry. For similar reasons, it was natural for most initial research on computatinal discovery to focus on these early stages. In contrast, more mature sciences emphasize the creation of models that explain phenomena (Hempel, 1965) in terms of hypothesized components, structures that relate them,

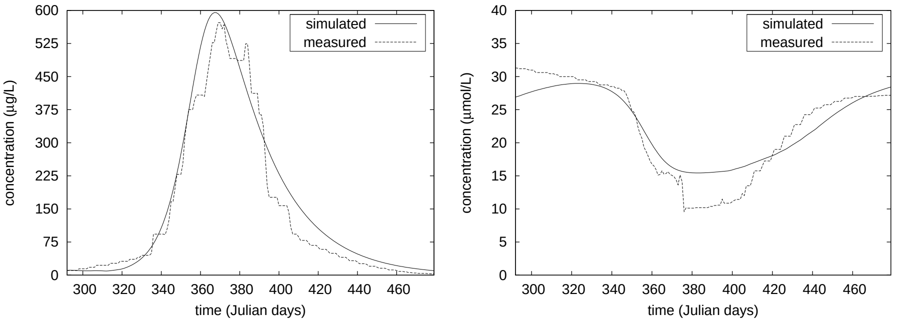


Fig. 1 Observed concentrations for phytoplankton (left) and nitrogen (right) from the Ross Sea, along with simulated values (smooth curves) produced by a quantitative process model, from Bridewell, Langley, Todorovski, and Džeroski (2008).

and interactions among these elements. Such explanatory models move beyond description to provide deeper accounts that are linked to theoretical constructs. Can we also develop computational systems that address this more sophisticated side of scientific discovery?

3.1 Quantitative Process Models

Consider an example from aquatic ecosystems. Figure 1 displays observations for the concentration of phytoplankton and nitrogen, taken at daily intervals, from the Ross Sea in Antarctica. Formal models of ecosystem dynamics are often cast as sets of differential equations, such as those shown in Table 1 (a). This model includes four equations, two for observed variables, *phyto.conc* (phytoplankton concentration) and *nitro.conc* (nitrogen concentration), and two for unobserved but hypothesized variables, *zoo.conc* (zooplantkton) and *detritus.conc*. When simulated, such equations can match observed trajectories with considerable accuracy, as illustrated in the figure.

However, the equations by themselves merely describe the ecosystem's behavior. The text in articles suggests that scientists have a richer story, which might go something like:

> As phytoplankton uptakes nitrogen, its concentration increases and the nitrogen decreases. This continues until the nitrogen is exhausted, which leads to a phytoplankton die off. This produces detritus, which gradually remineralizes to replenish nitrogen. Zooplankton grazes on phytoplankton, which slows the latter's increase and also produces detritus.

Each sentence here refers to some *process* that contributes to the model's equations. The first one focuses on the process of nutrient absorption; this is associated with the *phyto.conc* term in the first equation, which increases the phytoplankton concentration, and with the *phyto.conc* term in the fourth equation, which decreases

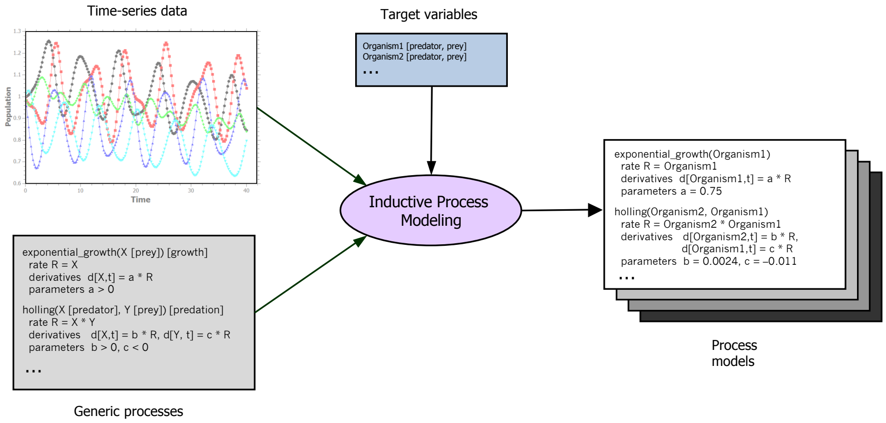
Table 1 (a) A set of linked differential equations for an aquatic ecosystem that relates concentrations of phytoplankton, nitrogen, zooplankton,and detritus to reproduce trajectories like those in Figure 1 and (b) a process model that compiles into the same set of equations.

```
(a) d[phyto.conc,t] = 0.104 · phyto.conc − 0.495 · zoo.conc
    d[nitro.conc,t] = 0.005 · detritus.conc − 0.040 · phyto.conc
    d[zoo.conc,t] = 0.053 · zoo.conc
    d[detritus.conc,t] = 0.307 · phyto.conc + 0.060 · zoo.conc − 0.005 · detritus.conc
```

```
(b) process phyto_loss(phyto, detritus)
        equations: d[phyto.conc,t] = −0.307 · phyto.conc
                   d[detritus.conc,t] = 0.307 · phyto.conc
    process zoo_loss(zoo, detritus)
        equations: d[zoo.conc,t] = −0.251 · zoo.conc
                   d[detritus.conc,t] = 0.251 · zoo.conc
    process zoo_phyto_grazing(zoo, phyto, detritus)
        equations: d[zoo.conc,t] = 0.615 · 0.495 · zoo.conc
                   d[detritus.conc,t] = 0.385 · 0.495 · zoo.conc
                   d[phyto.conc,t] = −0.495 · zoo.conc
    process nitro_uptake(phyto, nitro)
        equations: d[phyto.conc,t] = 0.411 · phyto.conc
                   d[nitro.conc,t] = −0.098 · 0.411 · phyto.conc
    process nitro_remineralization(nitro, detritus)
        equations: d[nitro.conc,t] = 0.005 · detritus.conc
                   d[detritus.conc,t] = −0.005 · detritus.conc
```

the nitrogen concentration. Similarly, the final sentence describes a grazing process; this contributes to *zoo.conc* in the first equation, to the *zoo.conc* term in the second equation, and to *zoo.conc* in the third equation.

We can reformulate such an account by restating it as a quantitative process model, like that shown in Table 1 (b). This includes five distinct processes for phytoplankton loss, zooplankton loss, zooplankton grazing on phytoplankton, phytoplankton uptake of nitrogen, and remineralization of nitrogen from detritus. Each process describes how one or more variables change as a function of current variables' values. For instance, the equation *d[detritus.conc, t] = 0.05 · phyto.conc* states that the derivative of detritus concentration is five percent of the phytoplankton concentration. After combining terms, this model is equivalent to the differential equations shown earlier if we assume that, when two or more processes influence the same derivative, their effects are additive. For example, summing the effects of processes for *d[phyto.conc, t]* from Table 1 (b) produces the expression *−0.307 · phyto.conc + −0.495 · zoo.conc + 0.411 · phyto.conc*, and combining the two *phyto.conc* terms gives the right-hand side of the first equation in Table 1 (a). Yet the elaborated version in Table 1 (b) also states explicit assumptions about the underlying processes, each of which indicates terms in the differential equations that must stand or fall together.



```
Organism1 [predator, prey]
Organism2 [predator, prey]
...
```

```
exponential_growth(X [prey]) [growth]
  rate R = X
  derivatives  d[X,t] = a * R
  parameters a > 0

holling(X [predator], Y [prey]) [predation]
  rate R = X * Y
  derivatives   d[X,t] = b * R, d[Y, t] = c * R
  parameters  b > 0, c < 0
...
```

```
exponential_growth(Organism1)
  rate R = Organism1
  derivatives  d[Organism1,t] = a * R
  parameters a = 0.75

holling(Organism2, Organism1)
  rate R = Organism2 * Organism1
  derivatives   d[Organism2,t] = b * R,
                d[Organism1,t] = c * R
  parameters  b = 0.0024, c = –0.011
...
```

Fig. 2 The task of inductive process modeling, which is given a set of variables, time series for their values, and generic knowledge about possible processes. The result is a set of parameterized models that fit the time series and explain its variations in terms of inferred processes.

3.2 Inductive Process Modeling

These observations suggest a class of discovery tasks that involve construction of such process accounts. We can state this problem, which we will refer to as *inductive process modeling* (Langley, Sanchez, Todorovski, & Džeroski, 2002), in terms of its inputs and outputs:

- *Given:* A set of typed entities and observed trajectories for their variables over time;
- *Given:* A subset of these variables whose values one wants to explain;
- *Given:* Background knowledge about process types that might appear in these explanations;
- *Find:* A quantitative process model that reproduces observed trajectories, explains them in terms of known processes, and predicts future values.

For instance, given a multivariate time series, like that shown in Figure 1, and background knowledge about candidate processes, inductive process modeling constructs one or more models like the one in Table 1 (b). Besides reproducing the observed values, this structure explains the trajectories in terms of unobserved but plausible processes. Figure 2 depicts the overall task graphically.

Background knowledge takes the form of *generic* processes that specify possible interactions among entities. These have roughly the same syntax as concrete processes, except that they refer to types of entities (e.g., organisms or nutrients) rather than specific ones (e.g., *phyto*) and they contain functional forms that relate attributes of these generic entities rather than specific equations. Functions do not include particular coefficients but instead refer to parameters with bounds on their possible values (e.g., between zero and one). The same parameter may occur in more than one of a process's equations, thus reducing dimensionality of the parameter space. Generic processes serve as the building blocks for constructing quantitative process models.

Traditional methods for inductive process modeling (e.g., Bridewell, Langley, Todorovski, & Džeroski, 2008) carry out search through a two-level space. The first works in the space of model structures, carrying out exhaustive depth-first search to find all ways to instantiate the known generic processes with specific entities; these become the elements for candidate models. The procedure starts with an empty model and, on each step, adds a new instantiated process if the new structure would not exceed user-specified size limitations.[1] For each such model structure, it then carries out search through a second parameter space. Starting with values sampled randomly from within each parameter's boundaries, this uses a conjugate gradient descent method to find parameters. The objective function that directs search is the squared error of a model's simulated trajectory compared to observations. Because this can halt at local optima, parameter estimation typically includes ten or more restarts from different initial values. This two-level strategy produces a list of parameterized model structures ranked by their error on the training data.

Table 2 presents pseudocode for this two-level approach to process model induction, which has been implemented and applied successfully to a variety of modeling tasks (Todorovski et al., 2005; Bridewell et al., 2008). These comprise fields as diverse as ecology, hydrology, and biochemistry (Asgharbeygi et al., 2006; Bridewell et al., 2008; Langley et al., 2006). Papers on the topic have also reported a number of extensions that increase the robustness and coverage of the modeling framework. These include revision of existing quantitative process models (Asgharbeygi et al., 2006), hierarchical generic processes to constrain search (Todorovski, Bridewell, Shiran, & Langley, 2005), an ensemble-like method that mitigates overfitting effects (Bridewell, Bani Asadi, Langley, & Todorovski, 2005), and an EM-like method that estimates missing observations (Bridewell, Langley, Racunas, & Borrett, 2006). Another extension has expanded the framework to models that involve partial differential equations (Park, Bridewell, & Langley, 2010), which take into account changes over both time and space. These results suggest that the basic framework is sound, and many encouraging empirical studies have buttressed this impression.

3.3 Discovering Rate-Based Process Models

Despite these successes, the initial methods for inductive process modeling have suffered from four drawbacks. They generate and evaluate full model structures exhaustively, which means the number of candidate structures grows exponentially with both the number of variables and the number of generic processes. Moreover,

[1] Some variants (e.g., Bridewell & Langley, 2010) ensure that candidate structures are consistent with constraints on relations among processes, say that an organism cannot take part in two distinct growth elements. These offer another form of theoretical knowledge about the domain.

Table 2 Pseudocode for IPM (Inductive Process Modeler), a system for inducing quantitative process models (Bridewell et al., 2008), which combines exhaustive search through the space of model structures with gradient descent search during parameter estimation.

```
Generate all process instances consistent with type constraints.
Combine these processes into candidate model structures up to a maximum size.
For each structure S,
    Select random parameters values for S within ranges of generic processes.
    Simulate parameterized S to produce a multivariate trajectory.
    Calculate simulation error E by comparing to observations.
    Use conjugate gradient descent to determine new parameter values.
    Repeat until there is no reduction in error E.
    Fix the parameters for S to this values and store E with S.
Return a list of parameterized model structures ranked by error scores.
```

parameter estimation requires repeated simulation of models to calculate an error score for gradient descent through the parameter space. Even so, this technique often halts at local optima, which can require many random restarts to find acceptable parameter values. Finally, despite these steps, the approach can still find models that fit the observations poorly. In summary, it does not scale well to complex modeling tasks, incurs high computational costs even for moderate tasks, and it is not reliable.

In recent research, Langley and Arvay (2015) have reported a new approach that addresses these problems. In their modeling formalism, each process *P* must include a *rate* that denotes *P's* speed or activation on a given time step. This rate is determined by an algebraic equation that is a parameter-free function of known variables. The process also specifies one or more *derivatives* that are proportional to P's rate, with negative coeffients for inputs and positive ones for outputs. Table 3 presents an example in this formalism that is equivalent to the one seen earlier in Table 1 (a). The notation has an important property that assists discovery substantially: a process model compiles into a set of differential equations that are linear combinations of rate terms. It also comes closer to Forbus' (1984) notion of qualitative processes, which inspired early work in the area.

Langley and Arvay also described RPM (Rate-based Process Modeler), a system that takes advantage of this more constrained formalism. Like its predecessors, the program estimates the derivative for each variable on each time step by calculating successive differences of observed values. It also generates a set of process instances by binding generic processes with variables in all possible ways consistent with their type constraints, but at this point it diverges. Because RPM assumes that each variable is observed, it can calculate the rate of each candidate process instance on each time step. As described in Table 4, it uses these derived values to carries out greedy search through the space of process models. For each target derivative, the system invokes multiple linear regression to find an equation that predicts it as a linear combination of process rates. Later differential equations must include processes that are consistent with those in earlier ones and vice versa. For instance,

Table 3 A process model that is equivalent to the one in Table 1 (b) but that separates algebraic rate expressions from equation fragments. Such a model always compiles into a set of differential equations that are linear combinations of rate expressions.

```
process phyto_loss(phyto, detritus)
   rate:        r = phyto.conc
   equations: d[phyto.conc,t] = −0.307 · r
                d[detritus.conc,t] = 0.307 · r
process zoo_loss(zoo, detritus)
   rate:        r = zoo.conc
   equations: d[zoo.conc,t] = −0.251 · r
                d[detritus.conc,t] = 0.251 · r
process zoo_phyto_grazing(zoo, phyto, detritus)
   rate:        r = 0.495 · zoo.conc
   equations: d[zoo.conc,t] = 0.615 · r
                d[detritus.conc,t] = 0.385 · r
                d[phyto.conc,t] = −1.0 · r
process nitro_uptake(phyto, nitro)
   rate:        r = 0.411 · phyto.conc
   equations: d[phyto.conc,t] = 1.0 · r
                d[nitro.conc,t] = −0.098 · r
process nitro_remineralization(nitro, detritus)
   rate:        r = detritus.conc
   equations: d[nitro.conc,t] = 0.005 · r
                d[detritus.conc,t ] = −0.005 · r
```

if an equation RPM selected for $d[x]$ includes a process P that influences not only $d[x]$ but also $d[y]$, any equation it considers for $d[y]$ must also include P's rate in its right-hand side. This approach factors the model construction task into a number of tractable components, combining the efficiency and robustness of linear regression with the use of domain knowledge to ensure explanations are internally consistent.

As expected, experiments revealed that RPM was not only vastly more efficient than SC-IPM (Bridewell & Langley, 2010), an earlier system for process model induction, running 83,000 times faster on even simple tasks, but that it found both the structure and parameters of target models much more reliably. Heuristic search through the structure space led to reasonable scaling with increases in the number of variables and generic processes, and simple smoothing over trajectories let it deal well with noise. However, some important drawbacks remained, including an assumption that all variables were observed and reliance on exhaustive search through the equation space.

More recently, Arvay and Langley (2016) have presented another system, SPM, that combines sampling of rate terms with backward elimination to induce equations in the presence of many irrelevant processes. The new program also replaces greedy search through the space of model structures with a two-stage strategy. First, the system finds multiple differential equations for each variable, after which it uses depth-first search to find sets of equations that incorporate consistent processes.

Table 4 Pseudocode for the RPM system for inducing rate-based process models. Rather than generating complete model structures, it carries out heuristic search by inducing one equation at a time, using process knowledge to ensure consistency across equations.

```
Generate all process instances consistent with type constraints.
For each process $P$, calculate the rate for $P$ on each time step.
For each target variable $X$,
    Estimate $d[X,t]$ on each time step with center differencing.
    For each subset of processes with up to $k$ elements,
        Find a regression equation for $d[X,t]$ in terms of process rates.
        If the equation's $r^2$ is high enough, retain for consideration.
    Add the equation with the highest $r^2$ to the process model.
```

Experiments with synthetic time series suggest that this approach scales better to equation complexity and identifies consistent process models better in domains like chemistry, where many different equatons predict each variable's derivative accurately, but where only a few combinations of these equations are consistent.

4 Extension to the Social Sciences

In the introduction, we mentioned our concern with computational discovery in the social sciences, and we are finally ready to return to that topic. In this section, we discuss briefly the relevance of process models to this area, consider some examples of social processes, outline how automated construction of social models might occur, and discuss some challenges that we must overcome to support the general form of this vision.

4.1 Relevance to Social Science

The social sciences share a number of features with ecology that suggest quantitative process models, and computational methods for their discovery, will prove as useful in the former disciplines as they have for the latter. These similarities concern both the nature of data that arise in these fields and the form of models that researchers often develop to explain them.

First, most data sets in the social sciences are observational rather than experimental in character. Not only would social experiments that involve large groups of people be prohibitively expensive, but many societies would view them as ethically unacceptable. The majority of data sets for ecological systems are also observational, for the same reasons, so the two fields must work with similar forms of nonexperimental evidence. They must also deal with dynamic behavior that requires explanation of changes over time.

Second, although social phenomena are composed of interactions among individuals, there are typically measured at the aggregate level. This produces values for a set of quantitative variables that describe summary features of the entire group, rather than those of their constituents. Most data about ecological phenomena also have this characteristic. As a result, many models in both fields postulate relations among aggregate, quantitative variables, as do the process models we have discussed in earlier sections.

Finally, many dynamic social phenomena appear to involve interactions among many variables. Endogeous terms both influence other variables and are influenced by them, often through feedback loops, although exogenous variables that only influence others also play a role. Thus, human behavior at the aggregate level is a complex system in the same sense as ecological networks. In other words, both are instances of system sciences that study interactions among entities.

Taken together, these similarities indicate that inductive process modeling offers a promising approach to discovery in the social sciences. At first glance, an important difference is that rate-based processes in ecology are described naturally in terms of inputs and outputs, often with an assumption that they involve variables that are conserved over time. This seems a less likely premise in social settings, but it is not a strict requirement on process models, so it does not prevent their application in this new arena.

4.2 Social Process Models

We can clarify the notion of social process models with some examples. First, consider some types of entities that might appear in a model. These might include social groups (e.g., different ethnicities, professions, or political parties) and physical resources (e.g., food, water, or power). Each entity will have one or more attributes that specify quantities on some dimension. For instance, a particular social group might have a membership count, an average annual income, a status score, and numbers who live in different neighborhoods. Consider two urban gangs, the *Jets* and the *Sharks*, with *Jets.count* and *Sharks.count* denoting their respective numbers, and with *Jets.status* and *Sharks.status* specifying their status scores.

Social processes produce changes in one or more such attributes. For example, a *migration* process might lead to an increase in the number of a group's members that live in one neighborhood and decrease those in another. Similarly, a *consumption* process might describe reduction in food or water when a group uses these resources. Or consider a *conversion* process that transfers some members of one group to another group:

```
conversion[Jets, Sharks]
  rate:      $r = \text{Jets.count} \cdot (\text{Sharks.status} - \text{Jets.status})$
  equations: $d[\text{Jets.count},t] = -1.0 \cdot \text{rate}$
             $d[\text{Sharks.count},t] = 1.0 \cdot \text{rate}$
```

The coefficient for *Jets* is -1.0, indicating that the process causes *Jets* membership to decrease, while *Sharks* is 1.0, stating that it leads *groupB* membership to grow. The two constants have the same absolute value because every person who leaves the first group must joint the second. The rate of transfer depends on three factors: the status of *Jets*, the status of *Sharks*, and the number of people in *Jets*. The rate increases with larger differences in status and with more people in the low-status group. Of course, different rate expressions are possible that depend on other attributes, but this example should clarify the notion of a social process.

A social process model would include a number of such elements, each of which describes change in one or more attributes and the rate at which they jointly occur. As we saw earlier, one can automatically compile these processes into a set of linked differential equations and then simulate the compiled model to generate multivariate time series. To the extent that the resulting trajectories match the observed social behavior, they provide an explanation of that behavior in terms of unobserved but still plausible processes.

4.3 Discovering Social Process Models

As in ecology, we can automate discovery of social process models using heuristic search through a space of candidate accounts. This requires that we provide a set of generic processes, instances of which might plausibly occur in the target setting. Each of these templates will specify the types of entities involved (e.g., a group or location) and associated attributes (e.g., count or status), whether the process causes the latter to increase or decrease, and the algebraic form of the expression that determines its rate. This set of generic processes may include different versions of the same process type. For instance, there may be variants of the conversion process with the same input and output relations but that differ in their rate expressions or even in the variables that influence them. These provide the building blocks for generation of candidate process models.

Given observations about dynamic social behavior and a set of target variables, a computational system like SPM (Arvay & Langley, 2016) can search the space of social process models defined by these generic elements. The program would induce one or more differential equations, stated as linear combinations of process rates, for each target variable. The system would then find combinations of equations that make consistent assumptions about which processes actually determine social dynamics. As before, search would occur for individual equations (which rate terms to include in the right-hand side) and model structure (which equations and associated processes to incorporate). Heuristics at both levels would limit the effective branching factor, making tractable the discovery of complex social process models that explain observed behavior.

4.4 Challenges in Social Process Modeling

The approach we have just described should apply, in principle, to any dynamic social setting for which multivariate time series are available. However, in practice, modeling tasks are likely to introduce challenges that require additional research on computational methods for discovering process accounts. The most basic is current techniques' reliance on a library of generic processes that specify not only attributes that interact but the functional forms that determine their changes over time. These may not be as obvious for social systems as for biological, chemical, or physical ones, in that they are typically less well understood. Automating the discovery of candidate generic processes is one response. Langley and Arvay (in press) have reported progress on this front, describing a system that introduces new generic processes by combining conceptual relations with algebraic rate templates, but we need more work on this problem.

Another challenge is that social science observations are more difficult to obtain than data in many other disciplines. This means that time series may have relatively sparse sampling rates; for instance, census data are only collected every four years. This is not an issue for social systems that change at rates slower than samples are taken, or even if signs of derivatives remain the same during unsampled periods. Moreover, for similar reasons, some social variables may not be measured on any time steps. Jia (personal communication, 2016) has identified certain conditions under which one can induce rate-based process models when some terms are unmeasured, but they must each participate in the same processes as observed variables. We need more research on this topic before our discovery methods can contribute fully to the social sciences.

A more basic limitation of process model induction is that it assumes the entities remain fixed over time. An organism's population or a group's count can decrease to zero, which can mimic the disappearance of an entity, but a model has no means for creating new entities during a simulation run. This would make it difficult to explain events such as the fissioning of a group into two splinters or the merging of two groups into a new composite. We might expand the process formalism to allow discrete events of this sort that occur under certain conditions (e.g., a group splitting when it grows too large), and extend simulation methods to predict their behavior over time, but inducing models that include both continuous and discrete processes would take us into new territory for computational scientific discovery.

5 Related Work on Social Model Construction

We are not the first to propose using computational methods to construct accounts of social phenomena. In fact, this idea has a long history, and in this section we examine its various threads. In each case, we describe the formalism for representing models and the extent to which their construction has been automated.

Structural equation models have been widely used in the social sciences (Goldberger, 1972). Such a model relates a set of quantitative variables, $X_1, \ldots X_n$, using a set of linear equations. The first variable, X_1, is a function of X_2 through X_n, the next one, X_2, is a function of X_3 through X_n, and so forth, with the final variable, X_n, being exogenous. This means one can display the model as a directed acyclic graph or as a coefficient matrix in which half of the nondiagonal entries are zero. Traditionally, a human would specify the graphical structure of such a linear causal model and invoke computational methods to estimate coefficients for each equation. Social scientists have applied this approach to observational data in many different settings, but standard approaches are limited nondynamical models with no feedback loops.

More recently, automated methods have emerged for inducing the structure of such causal models. The earliest example was TETRAD (Glymour et al., 1987; Spirtes et al., 1993), which used relations among partial correlations (e.g., the product of two partials equals the product of two others) to identify constraints on causal links. The system carried out search through a space of model structures, eliminating links that were inconsistent with these inferred constraints. TETRAD's developers applied it to a variety of observational data sets from the social sciences, but their approach did not support the discovery of dynamical models that explained change over time. More recent efforts, including related work in inducing the structure of Bayesian networks, have typically adopted greedy search through the space of model stuctures, based on degree of fit. However, Maier, Taylor, Oktay, and Jensen (2010) report an extension of the constraint-based approach that handles discrete relations among entities, rather than linear influences among continuous variables.

An alternative paradigm for explaining social behavior involves *agent-based* models (Epstein & Axtell, 1996). These are stated as collections of individual entities that interact over time in simulated environments. Members of an agent category follow the same decision-making rules, but they may differ in their parameters. As with differential equation models, simulation occurs in discrete time steps, but each agent responds separately to its situation. Activities are aggregated into summary statistics for groups or the entire population, which in turn are compared to measurements made at the aggregate level. Simulations that involve millions of agents are not uncommon. Researchers in this tradition develop their agent-based models by hand, but one can imagine automating their discovery by providing elements of agent programs and using computational methods similar to inductive process modeling to search the space of candidates.

6 Concluding Remarks

In the preceding pages, we reviewed computational advances in our understanding of scientific discovery. We recounted early research in this area, which treated the induction of numeric equations as heuristic search through a space of candidate laws. We then examined more recent work on inductive process modeling, which finds sets of processes and their associated differential equations that reproduce and explain multivariate time series. Contemporary efforts on this problem partition each process into a rate that is an algebraic function of known variables and a set of derivatives that are proportional to this rate. This assumption allows the use of multiple linear regression to find constituent differential equations, with process knowledge constraining their combination into models. This provides reliable, efficient, and scalable methods for discovering explanatory process models of dynamical observations, as repeated experiments have demonstrated.

We also discussed applications of process model induction to the social sciences. We argued that these disciplines have much in common with ecology, which has served as the main testbed for work on process modeling. These included a focus on nonexperimental data, quantitative measurements at the aggregate level, and dynamic interactions among variables. We provided examples of processes that might appear in explanations of social phenomena and how computational discovery might construct them using heuristic search. We also noted challenges that the social sciences pose to process model induction, such as fewer insights into candidate processes, difficulty in obtaining time-series data, and discrete events that remove entities or introduce new ones. Despite these challenges, inductive process modeling offers a promising approach to automating the discovery of explanatory models in the social sciences.

Acknowledgements

The research reported in this chapter was supported by Grant No. N00014-11-1-0107 from the US Office of Naval Research, which is not responsible for its contents. We thank Will Bridewell, Sašo Džeroski, Ruolin Jia, and Ljupčo Todorovski for useful discussions that led to the results reported.

References

Alberdi, E., & Sleeman, D. (1997). RETAX: A step in the automation of taxonomic revision. *Artificial Intelligence*, *91*, 257–279.

Asgharbeygi, N., Bay, S., Langley, P., & Arrigo, K. (2006). Inductive revision of quantitative process models. *Ecological Modelling*, *194*, 70–79.

Arvay, A., & Langley, P. (2016). Selective induction of rate-based process models. *Proceedings of the Fourth Annual Conference on Cognitive Systems*. Evanston, IL.

Arvay, A., & Langley, P. (2016). Heuristic adaptation of quantitative process models. *Advances in Cognitive Systems*, *4*, 207–226.

Bradley, E., Easley, M., & Stolle, R. (2001). Reasoning about nonlinear system identification. *Artificial Intelligence 133*, 139–188.

Bridewell, W., Bani Asadi, N., Langley, P., & Todorovski, L. (2005). Reducing overfitting in process model induction. *Proceedings of the Twenty-Second International Conference on Machine Learning* (pp. 81–88). Bonn, Germany.

Bridewell, W., Langley P., Racunas, S., & Borrett, S. R. (2006). Learning process models with missing data. *Proceedings of the Seventeenth European Conference on Machine Learning* (pp. 557–565). Berlin: Springer.

Bridewell, W. & Langley, P. (2010). Two kinds of knowledge in scientific discovery. *Topics in Cognitive Science*, *2*, 36–52.

Bridewell, W., Langley, P., Todorovski, L., & Džeroski, S. (2008). Inductive process modeling. *Machine Learning*, *71*, 1–32.

Bruk, L. G., Gorodskii, S. N., Zeigarnik, A. V., Valdés-Pérez, R. E., & Temkin, O. N. (1998). Oxidative carbonylation of phenylacetylene catalyzed by Pd(II) and Cu(I): Experimental tests of forty-one computer-generated mechanistic hypotheses. *Journal of Molecular Catalysis A: Chemical*, *130*, 29–40.

Colton, S., Bundy, A., & Walsh, T. (2000). Automatic identification of mathematical concepts. *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 183–190). Stanford, CA: Morgan Kaufmann.

Džeroski, S., & Todorovski, L. (1995). Discovering dynamics: From inductive logic programming to machine discovery. *Journal of Intelligent Information Systems*, *4*, 89–108.

Džeroski, S., & Todorovski, L. (Eds.). (2007). *Computational discovery of communicable scientific knowledge*. Berlin: Springer.

Džeroski, S., & Todorovski, L. (2008). Equation discovery for systems biology: Finding the structure and dynamics of biological networks from time course data. *Current Opinion in Biotechnology*, *19*, 360–368.

Epstein, J. M., & R Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Cambridge, MA: MIT Press.

Fajtlowicz, S. (1988). On conjectures of GRAFFITI. *Discrete Mathematics*, *72*, 113–118.

Falkenhainer, B. C., & Michalski, R. S. (1986). Integrating quantitative and qualitative discovery: The ABACUS system. *Machine Learning*, *1*, 367–401

Fayyad, U., Haussler, D., & Stolorz, P. (1996). KDD for science data analysis: Issues and examples. *Proceedings of the Second International Conference of Knowledge Discovery and Data Mining* (pp. 50–56). Portland, OR: AAAI Press.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence*, *24*, 85–168.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling. San Diego: Academic Press.

Goldberger, A. S. (1972). Structural equation models in the social sciences. *Econometrica*, *40*, 979–1001.

Gordon, A., Edwards, P., Sleeman, D., & Kodratoff, Y. (1994). Scientific discovery in a space of structural models: An example from the history of solution chemistry. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Atlanta: Lawrence Erlbaum (pp. 381–386).

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays*. New York: Free Press.

Hempel, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs, NJ: Prentice-Hall.

King, R. D., & Srinivasan, A. (1996). Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. *Environmental Health Perspectives*, *104* (Supplement 5), 1031–1040.

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B., & Oliver, S. G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist, *Nature*, *427*, 247–252.

Kokar, M. M. (1986). Determining arguments of invariant functional descriptions. *Machine Learning*, *1*, 403–422.

Koza, J. R., Mydlowec, W., Lanza, G., Yu, J., & Keane, M. A. (2001). Reverse engineering of metabolic pathways from observed data using genetic programming. *Pacific Symposium on Biocomputing*, *6*, 434–445.

Langley, P. (1981). Data-driven discovery of physical laws. *Cognitive Science*, *5*, 31–54.

Langley, P., & Arvay, A. (2015). Heuristic induction of rate-based process models. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. Austin, TX: AAAI Press.

Langley, P., & Arvay, A. (in press). Flexible model induction through heuristic process discovery. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. San Francisco: AAAI Press.

Langley, P., Simon, H. A., Bradshaw, G. L., & Żytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Langley, P., Sanchez, J., Todorovski, L., & Džeroski, S. (2002). Inducing process models from continuous data. *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 347–354). Sydney: Morgan Kaufmann.

Langley, P., Shiran, O., Shrager, J., Todorovski, L., & Pohorille, A. (2006). Constructing explanatory process models from biological data and knowledge. *Artificial Intelligence in Medicine*, *37*, 191–201.

Langley, P., & Żytkow, J. M. (1989). Data-driven approaches to empirical discovery. *Artificial Intelligence*, *40*, 283–312.

Maier, M., Taylor, B., Oktay, H., & Jensen, D. (2010). Learning causal models of relational domains. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (pp. 531–538). Atlanta: AAAI Press.

Mitchell, F., Sleeman, D., Duffy, J. A., Ingram, M. D., & Young, R. W. (1997). Optical basicity of metallurgical slags: A new computer-based system for data visualisation and analysis. *Ironmaking and Steelmaking*, *24*, 306–320.

Moulet, M. (1992). ARC.2: Linear regression in ABACUS. *Proceedings of the ML 92 Workshop on Machine Discovery* (pp. 137–146). Aberdeen, Scotland.

Murata, Mizutani, & Shimura (1994). A discovery system for trigonometric functions. *Proceedings of the Twelfth National Conference on Artificial Intelligence* (pp. 645–650). Seattle: AAAI Press.

Nordhausen, B., & Langley, P. (1990). A robust approach to numeric discovery. *Proceedings of the Seventh International Conference on Machine Learning* (pp. 411–418). Austin, TX: Morgan Kaufmann.

Park, C., Bridewell, W., & Langley, P. (2010). Integrated systems for inducing spatio-temporal process models. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. Atlanta: AAAI Press.

Popper, K. R. (1961). *The logic of scientific discovery*. New York: Science Editions.

Saito, K., & Nakano, R. (1997). Law discovery using neural networks. *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 1078–1083). Yokohama: Morgan Kaufmann.

Schaffer, C. (1990). Bivariate scientific function finding in a sampled, real-data testbed. *Machine Learning*, *12*, 167–183.

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, *324*, 81–85.

Shrager, J., & Langley, P. (Eds.) (1990). *Computational models of scientific discovery and theory formation*. San Francisco: Morgan Kaufmann.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.

Todorovski, L., Bridewell, W., Shiran, O., & Langley, P. (2005). Inducing hierarchical process models in dynamic domains. *Proceedings of the Twentieth National Conference on Artificial Intelligence* (pp. 892–897). Pittsburgh, PA: AAAI Press.

Todorovski, L., Džeroski, S., & Kompare, B. (1998). Modeling and prediction of phytoplankton growth with equation discovery. *Ecological Modelling*, *113*, 71–81.

Valdés-Pérez, R. E. (1994). Human/computer interactive elucidation of reaction mechanisms: Application to catalyzed hydrogenolysis of ethane. *Catalysis Letters*, *28*, 79–87.

Washio, T., & Motoda, H. (1997). Discoverin admissable models of complex systems based on scale types and identity constraints. *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 810–817). Yokohama: Morgan Kaufmann.

Żytkow, J. M. Zhu, J., & Hussam, A. (1990). Automated discovery in a chemistry laboratory. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 889–894). Boston: AAAI Press.