Fourth Paradigm GIScience?
Prospects for Automated Discovery and Explanation from Data

Mark Gahegan
Computer Science, the University of Auckland, New Zealand

This paper is the pre-submission draft of a paper that will appear in IJGIS sometime soon

Abstract
The article discusses the prospects for automated discovery of process models directly from geospatial data. Rather than taking an approach based on machine learning, which generally leads to models that cannot be understood by humans or related to domain theory, the approach described here suggests we can instead construct process models from fragments of domain understanding—such as commonly encountered equation forms, known constants and laws—resulting in a discovered model that can both be understood by humans and directly compared with known theory.

We examine how progress in scientific computing and discovery informatics can lead to the automated production of explanatory models in geography, based on current research work. We then propose a conceptual model of the discovery process by which the various stages and components of discovery and explanation work together to learn models from data. The approach described weaves together ideas for describing models from Harvey's book 'Explanation in Geography' with current thinking on how explanatory models might be 'discovered' from data from Inductive Process modelling.

On the way, we also highlight: (i) why it is important to have models that explain as well as predict, (ii) how such an approach contrasts with—and goes beyond—current work in deep learning, (iii) how the task of model discovery might be tackled computationally and (iv) how computational model discovery can play a valuable role in creating geographical explanations.

Keywords: *computational model discovery*, *inductive process modelling, explanation, geographical information science, big data, data driven science.*

1 Introduction: moving beyond models that predict, but do not describe

As described by Hey *et al*. (2008), science has evolved three main paradigms thus far. The first is *Experimentation*, characterised by observation and measurement. A good example is determining the relationship between the length of string and periodicity of a pendulum. Experimentation became possible with the invention of reliable ways to measure physical quantities, such as time, weight and length. Accurate measurement allowed observations to be standardised and compared, and so generalisations could be sought.

The second paradigm is *Analytical Theory*, which searches for the theory that might explain some system. Note that the system does not need to be measurable, or even real, it can be hypothesised, so does not necessarily require data. Einstein's famous equations describing relativity fall into this category, and were motivated by thought experiments. The data to establish the validity of special relativity was not even available until after his death.

The third paradigm is *Numerical Simulation*, characterised by the extensive application of computing power to model the physical world, often in great detail, and using forecasting techniques to predict future states. A good example is climate forecasting, combining simulation, historical data and prediction methods. Another is the search for nano-materials with interesting properties, using computational chemistry. Many newer branches of science now are home for research communities who work entirely on such *in-silico* experiments.

The Fourth Paradigm, as proposed by Hey *et al*. (2008), is *Data-Driven Science*, and it goes further by suggesting that data itself can drive the discovery of new knowledge. Of all the current hype surrounding Big Data, this is perhaps the most intriguing aspect, with some authors even heralding: "*The end of theory.*" [Anderson, 2008]. The argument goes that, with the availability of very rich data that comprehensively describes a given situation, it becomes possible to discover theoretically-grounded explanations that make sense of, or explain, our observations. This goes far beyond the established methods of data mining and knowledge discovery (e.g. Miller & Han 2009), which are limited to identifying far simpler descriptions of data artefacts and trends. But note that this article is not about deep learning, a topic on which much has been written of late, including in the GIScience literature (e.g. Chen & Zipf, 2017; Mou *et al*., 2017; Yang *et al*., 2018). Deep learning, whilst remarkable in terms of the progress made towards highly effective pattern-based inference, is still very much operating as an *inductive* (predictive) analytical method—and a 'black box' one at that. By contrast, the focus here is on *explanation*, via methods that strive towards *abductive* reasoning (Gahegan, 2009). The fourth paradigm suggests that processes could be discovered and also explained entirely by a computational system. The conjunction of massive computational power and massive geodata now create the opportunity for us to move beyond automated prediction and towards automated explanation. This article describes how this might happen.

As an aside, GIScience sometimes seems to be stuck in the second paradigm—continually searching for more 'theory', as a means to justify its science credentials. In the light of the four paradigms listed above, this is perhaps a rather narrow view of what a science is. On a similar note, one might characterize the geocomputation community as trying to nudge GIScience on from Analytical Theory to Numerical Simulations.

1.1 Review: Two kinds of (process) models

Before delving further into the possibilities of data-driven science, recall that the analytical models[1] (here called *process models*) used in research can be divided into two kinds. The first is

[1] As opposed to models of data—data models, as used to describe logical relationships amongst the entities being modelled.

descriptive: mathematics, statistics and logic are used to describe the various entities and processes at work, and their inter-relationships. These statistical, mathematical or logical functions represent an abstraction of our understanding of the world, and its outputs can be readily interpreted in this light. For example, we might create a model of water discharge from a catchment by formulating a theoretical relationship between rainfall, soil moisture, topography and time. To change the model, we would change either the values of these entities, or the functions that describe their relationships.

The second kind of process model is predictive: given a set of input values (states) the model is conditioned to produce a desired output, usually by supplying training examples and engaging in some kind of inductive machine learning (Gahegan, 2003). But the model is essentially a black box: exactly how the relationships between input and output are constructed may bear no resemblance to the actual processes operating within the problem domain. Nevertheless, such models can produce extremely accurate results, often better than descriptive models, because they are not limited so much by simplifying assumptions, for example concerning statistical relationships (Brenning, 2005); (Fischer, 2013). To continue with the above example, to create an inductive model of catchment discharge, we need to measure the state of the variables we think play an important role, and associate these values with some known drainage outputs. This forms our training data from which the learning methods automatically construct the functions that describe the relationships between inputs and outputs, using a process of search guided by error minimization. The resulting model may have little or no relationship to known theory—and indeed may be difficult to 'extract' from the methods used. It is usually in the form of a graph of functions describing relationships between input states and outputs, and therefore difficult for humans to interpret.

1.2 Computer Science, Geography, and Explanation

Predictive models are exemplified by much of the work undertaken within the geocomputation community to date (e.g. Cheng *et al*., 2012). One problem that geocomputation has yet to overcome is the resistance among researchers to models that do not have an obvious connection to established theory or known mechanistic steps. This is entirely reasonable, since one can never be sure if predicted outcomes are even possible given the real-world processes involved. The 'black box' provides an answer, but offers no comfort in terms of understanding the inner workings.

Peter Gould (1999) used to refer to the computer as a "high-speed idiot" because it can only work quickly, not smartly. This has certainly been the case, for the larger part of the history of computing. But it is becoming less true. We have examples now of computers outperforming even the most expert of humans at complex games such as chess and go (e.g. Silver *et al*., 2017). The computational systems at work in both cases have a well-developed meta-model of the game: in other words, they have an underlying structure based on the theory of the game and how to reason over the situations within the game. They also have the ability to learn new knowledge from past experience, to expand or correct this meta-model. Granted, these are narrow tasks, but then so are many research-related activities. So perhaps such techniques can be used on science problems?

One of the longstanding goals of artificial intelligence is to create machines that can learn, can reason and can explain. These goals have proven difficult, and are taking far longer to realise than was first envisioned when the potential to solve them was first postulated. In the meantime, AI has fragmented into many sub-disciplines (such as machine vision, deep learning, description logics…) that have addressed aspects of these goals, but concentrated less on the big picture. There has been good success in learning when the problem is very specific and good example data is available. However, complex reasoning and problem-solving is still a major

challenge. Machine explanation, in particular, has achieved less success to date (e.g. Carenini & Moore, 2006).

David Harvey's book: *Explanation in Geography* (1969) provides us with some key insights as to why this is. Specifically, the more abstract cognitive and philosophic aspects of research—such as the methodology and the epistemology—play key roles in explanation and need to be explicitly acknowledged. Without them, the understanding and communication of results is inherently limited to numbers. In a field that was experiencing a computational revolution at the time, Harvey's book stands as a reminder that there is more to analysis than can be handed off to a computer. At the time of his writing, inference and explanation were exclusively the domains of the human expert, but that situation has now changed, as we shall see later.

Researchers have always been in the business of learning from data, this is by no means a new idea! An early example from GIScience of such learning enabled by the power of computers is shown in the work of Stan Openshaw. The Geographical Explanation Machine (GEM), proposed by Openshaw and colleagues (1990) takes us beyond simply finding patterns to helping to suggest possible explanations for them. Having pioneered a method to discover geographical clusters from point data (the Geographical Analysis Machine or GAM: Openshaw *et al*., 1987), Openshaw and his team experimented with whether these clusters showed correlation with the various overlays that could be added to their GIS. In other words, if the data points representing cases are intersected with data in a given coverage (say a geology map), do we find that their distribution is random among the spatial units in the map, or is it biased towards a specific region, or range of values? If the latter, it becomes possible to 'explain' some of the pattern with the map layer. This may suggest a potential hypothesis for the researcher to explore. However, two things are important to note here. Firstly, correlation is not causation: explaining a pattern in terms of correlation or entropy is not the same as explaining it in terms of theory. Secondly, GEM does not have a model of the research process and does not itself make 'discoveries', it plays a supporting role in suggesting possible explanations that a user might supply.

1.3 Bayesian Belief Networks and automated model construction

A useful step on the way to building models that both predict and explain appeared in the 1990s as a development of Bayesian Belief Networks (BBNs). BBNs allow us to predict likely states for particular variables given a chain of supporting evidence. A BBN is essentially a Directed Acyclic Graph that represents a set of variables or nodes and their conditional dependencies, the edges (Pearl, 1986). Some of these variables may be unobserved: that is, they must be inferred. Within the field of GIScience, Stassopoulou *et al*. (1998) show how BBNs can improve the accuracy of landcover change, by modelling the interactions of gathered variables such as slope, soil depth and erosion. The network learns the conditional dependencies that are needed to match the observed outcomes and in doing so, it improves the predictability of more complex kinds of change.

BBN learning (Lam & Bacchus, 1994) goes further by providing a method to discover the underlying graph structure of such networks directly from the data. There are several ways this can be achieved, but they boil down to a search for graph fragments that 'fit' the observed variables, which are then progressively combined, edited or augmented to improve their predictive power. At each step, we can think of the algorithm as posing the basic question: If a link is added between these two nodes, does the model get stronger? There are three possible kinds of link that can be inserted:

1. Chain or pipeline, where a linked set of nodes form a directed path such that variable *a* influences *b* and *b* influences *c*, and so on: $a \rightarrow b \rightarrow c$. If *a* is observed, it may influence *b*, and so forth. (Also, if *b* is observed, then *a* does not influence *c.*)

2. Diverging or fork, where all edges are directed away from a node: a $\rightarrow b \leftarrow c$. If b is an observed variable, then any influence between a and c will be blocked by b.

3. Converging or collider, where all edges are directed towards a node: a $\leftarrow b \rightarrow c$. If b is unobserved, then a and c cannot influence each other. If b is observed, then a and c may influence each other.

BBNs emphasise prediction over explanation. Nevertheless, the learned graph is indeed a kind of model that can be interpreted from the standpoint of theory—it describes to what extent, and in what direction, observed variables are conditionally dependent on each other. However, BBNs do not model processes, but rather the conditional dependencies caused by processes, and any relationship that these dependencies have to actual processes remains entirely hidden, i.e. it is still a statistical model, not a process model.

1.4 Progress on creating autonomous systems for discovery

In recent times, progress has been made on developing systems to aid with discovery. Two examples of discovery systems are provided below, the first from remote sensing, the second from bio-chemistry.

1.4.1 A wildfire discovery system

The Geostationary Operational Environmental Satellite system (GOES) provides imagery every 20 minutes and for several years this has been used to help identify new wildfire outbreaks. A fire detection workflow (called GOES WF-ABBA) has been coupled to the satellite data stream that automates: image pre-processing, fire detection methods and results validation. The resulting accuracy and timeliness of the workflow has also been validated against historical fire incident data (Koltunov *et al.*, 2012). The system can rightly claim to have 'discovered' many fire outbreaks before they were reported by humans, but it cannot explain the discoveries it makes, nor does it learn new models for the process of fire ignition. It is an example of workflow automation, where the workflow describes the steps a known discovery process.

1.4.2 A drug discovery system

Eve (Williams et al, 2015) is a bench robot designed to aid in drug discovery. Eve is based on an automated batch-testing, bench robot, the kind used in larger wet-labs to test possible new drug candidates against a library of known compounds, to see how they behave. Thousands of tests can be carried out efficiently in batches by a machine that simply mixes the two chemicals together then can test for various kinds of reaction. But in addition to the standard robotics, Eve also has:

— An updatable knowledge base describing all the experimental evidence gathered to date
— A model of the science process to determine the next steps to take
— Constraints based on theory describing what is already known, such as what is likely or unlikely to be successful if tried and where the gaps are in terms of drug candidates and diseases.

Equipped in this way, Eve is capable of constructing and testing a useful hypothesis, conducting the related experiments and learning from the results obtained. The use of Eve has led directly to several published discoveries—including that certain cancer-fighting drugs are also effective against rare tropical diseases. It requires only a little imagination to see that Eve could also write portions of a research article describing the experiments she conducts[2].

[2] Interestingly, Eve herself has yet to be listed as an author on any of the papers describing discoveries made, which perhaps hints at a raft of ethical questions concerning AI that we are yet to face.

Certainly, Eve does not discover new process models that describe how specific drugs work, the discoveries made are more like black-box predictions. Nevertheless, Eve represents an artificial intelligence capable of operationalising large portions of the research life-cycle independently, based on a meta-model of the research process itself. This is important progress in our journey towards autonomous discovery.

2 Data driven discovery of models

In many respects, creating explanations is the holy grail of Artificial Intelligence research (Russell & Norvig, 2003), one that up to now has seen comparatively little obvious progress: it remains far easier to inductively learn predictive models that produce reliable outcomes from given input variables. As we have seen above, such predictive models contain no explicit domain insights. Moving from prediction to explanation, specifically explanations built upon feasible processes and their interactions, would represent a massive step forward for the sciences. Models are then not only useful in terms of their predictive power, but also useful to explain the likely processes and mechanisms that give rise to their predictions. Progress here also offers the tantalizing possibility of discovering entirely new process models based on data alone. And, of course, if explanatory models can be inferred, then this opens up the possibility of uncovering a host of unknown connections and processes that may be operating in a given situation, thereby increasing our understanding.

Using computational systems to aid scientific discovery and explanation is not a new idea, but it is a big idea, requiring research progress on many fronts: from scaling up computational search and optimization, through theorem-proving, to the conceptual modelling and computational representation of the research process itself. Valdez-Perez (1996) provides a useful summary of some of these challenges and the early history of the field. The major challenge, then and now, is to create a single integrated system that represents and empowers discovery science, rather than the many fragmented systems for representation and analysis that we currently use.

An excellent volume of collected papers edited by Shrager & Langley (1990) provides some examples of the depth of research needed to represent the scientific discovery process. Bridewell et al. (2008) show how more complex process-based models with several dependent variables can be learned from data. Džeroski & Todorovski (2007) also provide useful coverage of a wide range of approaches to automated discovery and related challenges.

2.1 The challenge of linking discovery with explanation

Inductive machine learning has served us well, and deep learning implementations now show even more promise in terms of reliability and ease of training (Schmidhuber, 2015). But we need to raise out sights a little. What if, instead of using hyperplanes and decision rules, predictive models could draw on existing theory for their basic primitives, and assemble predictive models from these pieces of theory that could be more easily understood and explained?

Using equations describing actual processes that are known to occur to learn a descriptive model is called *inductive process modelling*. A process model is a chain of processes that link observable variables with each other via causal mechanisms, expressed as equations. An *inductive* process model is simply one that is inductively learned. It makes explanations from a kind of theory and data soup: the soup contains a set of theory fragments, rules, constraints, and a lot of observations drawn from the system under investigation. Using these building blocks, it is possible to iteratively assemble pieces together, guided by heuristic search and feedback on the quality of the model produced. Since the model produced is a set of equations with known links to current theory, then it becomes possible to 'read' the model; it is essentially a

mathematical description of the process under investigation. See Schwabacher and Langley, (2001) and Asgharbeygi *et al*. (2006) for more details. When coupled with the descriptive richness of large amounts of training data, this approach can also pick up subtleties that are currently missed or excluded from a traditional descriptive model. It can also start with an existing theory-based model, and improve upon it.

To substantiate these claims, the inductive process modelling community has had good success in constructing models from complex data that describe various spatial and temporal processes. As Langley showed nearly forty years ago, it is possible to 'discover' some well-known physical laws from data, such as acceleration due to gravity (Langley, 1981). In this case, the model is simple: just one equation with one rate term, and it can indeed be learned autonomously from experimental data. But more recent work includes discovering ecological models that can describe: complex predator-prey relationships (Bridewell et al, 2008), the occurrence of large algal blooms (Džeroski *et al*., 2007) and complex reaction networks in systems biology (Džeroski & Todorovski, 2008). These models are every bit as good as those produced by expert scientists, and sometimes better.

Recent work by Arvey (2018) describes experiments learning of ecological food webs, building on the earlier work of Langley, Bridewell and colleagues and colleagues. Figure 1 below shows the accuracy of a learned model describing predation in a two organism system, compared with observed values. The trajectories correspond to an R^2 score of 0.84 for *Paramecium Aurelia* and 0.71 for *Didinium Nasutum*. Working back from the equations used, the process model also supports the following explanation: that *Nasutum* is a predator species with an exponential death rate, and *Aurelia* is a prey species with an exponential growth.

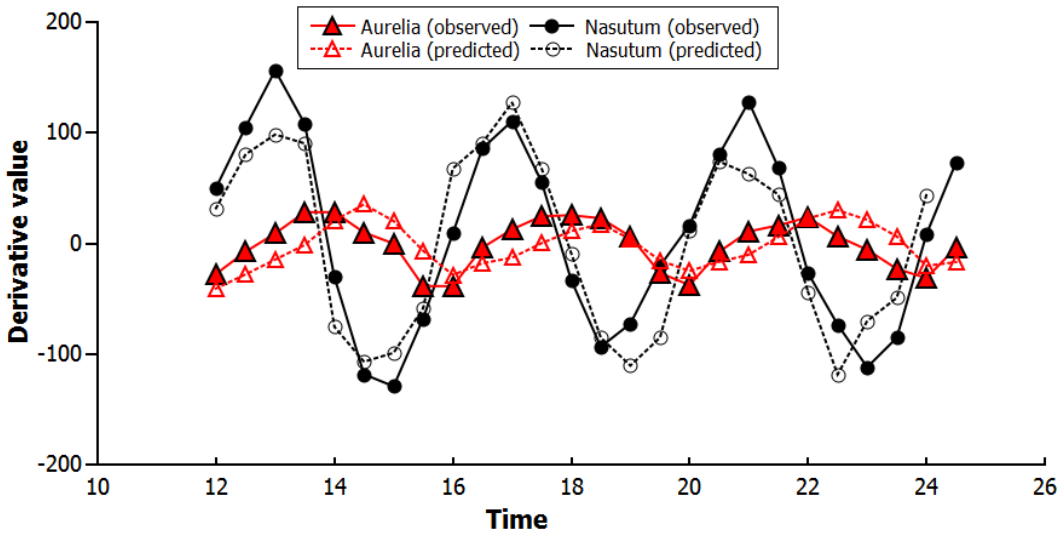


Figure 1. The output of learned process model describing predation between two species of microbe: *Paramecium Aurelia* and *Didinium Nasutum,* compared to observed values. Source: Arvay, 2018 (used with permission).

Perhaps even more remarkable, the system used to generate Figure 1 can scale to very complex models, with 15 or more different interacting species, successfully learned from data describing the observed counts of individual organisms. That's a very complex model, with an intricate and lagged set of co-dependencies, as shown in Figure 2 below. Such a model would be very difficult to construct manually.

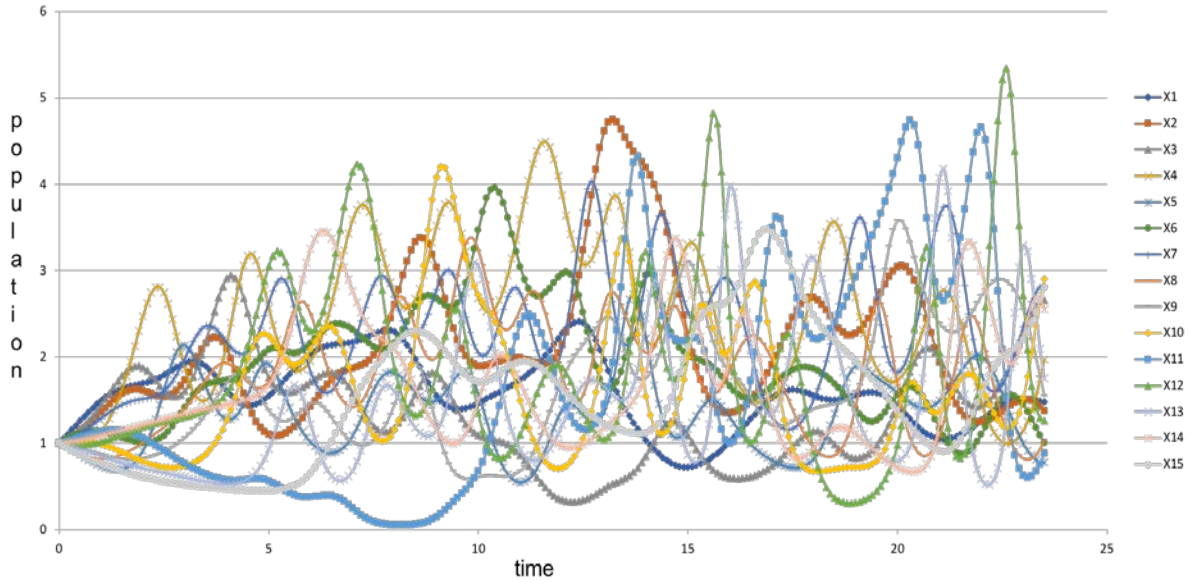


Figure 2. Graph showing the output produced by a discovered food web process model, representing the relationship between 15 species through time. Source: Arvay, 2018 (used with permission).

Just as with inductive machine learning, in order to use computational methods to discover scientific models, we need to restate the problem as one of search—a search over a truly massive space that represents all the models that could exist to find those that describe the variation of the data in ways that are both simple (parsimony) and which minimise error (accuracy). Graphs are constructed to represent the assembly of a model from the 'soup' of theory fragments. The process here is analogous to that described above for building Bayesian Belief Networks from data (Section 1.3) except that here there are a lot more constructs in play so the space of possibilities is very much larger. By utilising learning and optimisation strategies available from AI and inferential statistics, this task has recently become computationally tractable, even for complex models as shown in Figure 2—the model in this case takes of the order of minutes to create.

To give a specifically geographical example of model discovery, imagine that we wish to understand the mechanisms that drive overland water drainage and soil erosion. Some simple examples of equations to add to the process library might be: exponential decay, simple geometric calculations of upslope area, water accumulation, soil erosion coefficients and equations governing fluid flow and saturation. To guide the model discovery process we need a measure of success, which can be straightforwardly defined as a reduction of error in the model's ability to correctly predict the observed data (say the outflow of the basin or the amount of soil eroded over time). This provides the feedback needed to guide and refine the model's development. We will also need observation data that describes the variables that appear in the equations in the process library.

3 Inductive Process Modelling in detail

Inductive process modelling is perhaps the strongest current candidate for automated model building, whilst still meeting the goal of creating models that can be interpreted in terms of theory, by humans. As noted above, the basic idea has been around for some time, but it is the recent progress towards more robust, scalable solutions that now make it viable as a discovery tool. This section takes us through the modelling process in detail, from initial setup to final evaluation, and briefly describes each step along the way. Following this, Section 4 then takes a

broader perspective on the research process itself, and asks what else is needed to truly create a fourth-paradigm discovery environment.

Inductive process modelling comprises the following steps:

1. *Setup*
    - Training data
    - Process Library of domain knowledge
    - Rules and constraints
2. *Structure search*
    - Combine equations together, using an evaluation strategy to structure the search and constraints to limit the possibilities
3. *Parameter search*
    - For a given process model, fit parameters and evaluate
4. *Model Refinement*
    - Work to improve the model
5. *Report Results*
    - List of models ranked by score

The descriptions of each stage that follow draw from a more comprehensive account provided by Bridewell et al. (2008) with some additional examples and detail added in to suit our purposes here.

3.1 Setup/Input

Setup consists of three activities, as follows.

3.1.1 Training data

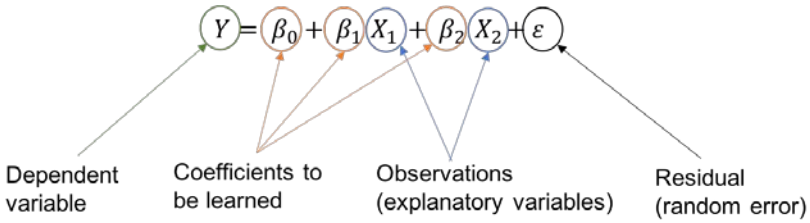
The input data are various observations (quantitative measurements) of the system under study. In most model discovery research to date, the observations are time-dependent rates. They need not be independent, indeed, for time-series data we would expect to see temporal dependence within the data. Observations must be typed—i.e. we need to know their statistical scale (nominal, ordinal, integer, real etc).

3.1.2 Library of domain knowledge

The specific process instances woven together in a model are drawn from instantiations of known, generic processes. These generic processes are essentially little fragments of theory. For example, we may have a generic process for predation, such as those described by Holling (1959), that contains a variable rate term to be instantiated (via learning) for a given predator/prey interaction. Such terms are essentially unobserved variables, and the equations helpfully describe where they should appear in the process model. Other kinds of functions might be included too, such as standard statistical tests, spatial analysis functions, clustering methods and so forth. The form of the equations used is an open question, but most research so far has used differential equations for representing dynamic processes.

Taking a common equation such as regression as an example, we can view the function (shown here with two explanatory variables) as follows[3]:

[3] Inspired by: http://desktop.arcgis.com/en/arcmap/latest/tools/spatial-statistics-toolbox/regression-analysis-basics.htm.



Generic process:	**linear regression** (with two dependent variables)
dependent variable:	*Y*: regression_score [float]
observations:	X_1, X_2
unobserved variables:	$\beta_0, \beta_1, \beta_2, \varepsilon$
equation:	$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \varepsilon$
constraints:	X_1, X_2 must be integer or real
	X_1, X_2 must be independent (VIF score < 7.5)

In selecting and parameterising a particular generic equation, the system is in fact hypothesising a process as the causal mechanism that links together observations. The description of an equation shown above defines a pattern that determines where in a workflow such an equation might fit[4]. And by recording the selected equations, an explanation for these observations is created. Such a commentary can be given as part of the results, provided each generic equation is provided with a meaningful (theory-based) name or description. For example, an equation that explains the combined behaviour of population variables *A* and *B* over time might represent a *Migration* mechanism with *A* and *B* being inflow and outflow measures respectively. If this process fits the observations well, then we may hypothesise that migration is the mechanistic process by which the data are best explained.

3.1.3 Rules and constraints

The generic equations in the domain library can have rules associated with them that restrict how they can be used. The simplest case is restricting the statistical scale (real, integer, ordinal, nominal) of the input variables that can be 'plugged in' to them; again, with the effect of reducing the possibility space. Additionally, a constraint may govern the range of permissible values that a term may take, for example a power term might only take an integer value from [0..2] or [0..3]. Other rules and constraints that are known to hold within a field of research can also be expressed; for example, ensuring that the methods used are consistent with a single representational model of geography (e.g. objects vs. fields).

3.2 Structure search

In this step, the system builds a chain (graph) of theory fragments into a model, guided by an error measure of the fitness of the model produced so far. Each process defines an interaction between actors in a dynamic system, and most dynamic systems will have multiple processes operating concurrently, thus the dynamics of a system emerges from the combination of multiple processes. Searching through all possible equations for each of the interactions in a model would be an impossible task, and unhelpful, since established domain knowledge already suggests equations that are likely to appear. The generic theory fragments help by both limiting the size of the model possibility space and by suggesting where unobserved variables are needed.

---

[4] The reason *Y* is strongly typed as a regression_score is so that the method will not be placed in a workflow unless some other method requires a regression to be calculated (i.e. has a regression score as one of its unobserved variables). Otherwise this pattern would 'fit' anywhere there are two observed variables.

Specific instances of individual processes are often not predictive by themselves, it us usually the case that all or most relevant processes must be present in a solution for it to be useful and ranked highly. This creates a problem at the initial stages of model building, where the difference between one choice and another may be negligible. It is only once the model begins to take shape that the value of various possible additions or changes can be easily distinguished. See section 3.4 on Model Refinement for strategies to overcome this problem.

3.3 Parameter search

Each equation fragment used that contains any free variables—such as coefficients and factors—will need to have these parameters instantiated. In current approaches to model building, this must occur each time a new generic equation is added to the graph, as it is needed before a new fitness score can be computed.

3.4 Model Refinement

There are many ways that the search can be improved in terms of reliability and error minimisation. One obvious way is to start the model from multiple points, to ensure that early choices (when error measures are most unreliable) do not lead to a poor solution, where the model gets trapped in a local minima. Another useful strategy is to take a promising solution and try to improve it by swapping out individual equations or changing their parameters. Of course, in today's computational environments, the model construction and improvement can make use of thread parallelisation and highly efficient stochastic search methods (e.g. Hans *et al.*, 2007).

3.5 Report Results

Each completed model is given a fitness score, so all models can be ranked. Other measures, such as the stability of the solution (similar model variants repeatedly score the highest) and confidence in the solution (how much effort was expended to explore the solution space of models) can also be used along with a fitness score to offer a sense of confidence in the solution.

At this final stage, the model can also be used to create a human-readable account describing which generic equations were used, and how they were connected together, as a kind of paraphrasing of the mathematics of the model.

3.6 Discussion on inductive process modelling

Inductive process modelling can successfully 'discover' complex models from complex data. The fragments of domain knowledge used do indeed lead to solutions that explain the data, rather than simple predict likely outcomes, so the model produced can be interpreted by humans who are already familiar with the theory of a domain. Or putting it another way, we get equations that can make sense to humans, rather than hyperplanes and weight matrices that do not. And with recent improvements in the optimisation of model creation, inductive process modelling has become practical for complex systems involving interaction between many processes, as shown in Section 2.1.

Note that a strategy is needed to govern how the system moves through the various steps outlined above, based on making adequate progress, or abandoning a partially-developed model and moving back to an earlier stage if needed, similar to the Eve drug discovery system described in Section 1.4.2). However, there is more to the discovery process than connecting equations together. One future goal is to build discovery systems that more strongly reflect the *process* of discovery, say as it is understood from the philosophy of science, or how it is practiced in a specific research community. To have a system that genuinely and autonomously discovers new theory, we need to equip it with a roadmap of the discovery process, as practiced

by the relevant discipline. This involves encoding some of the epistemological and ontological principles that guide what a discovery is, and what is required in order to back up a claim that something new has indeed been 'discovered'. The next section reviews what might be needed to address this bigger and most ambitious challenge.

4 Building a richer model of the discovery process

Broadly, we understand how the process of scientific discovery occurs (e.g. Popper, 1959). An excellent account is given by Souzou *et al*. (2017) and an account focussed on GIScience is provided by Gahegan (2005). Under what Kuhn (1962) would call "normal science" we posit a hypothesis, build an analytical model, run experiments, evaluate the results, iterate with modifications and refinements, then when we are confident in our findings, we share them.

Within the sciences in general (also true for GIScience) we have concentrated our attention on supporting only parts of this cycle computationally, focussing on the data and analytical methods, whilst other aspects usually go unrecorded. For example, the hypotheses, the workflows, the ontology and the epistemology are usually not specified but assumed, and where these are recorded, it is not in the same system used for analysis and modelling, nor even connected with it. Making an effective system for scientific discovery requires that we join all of these pieces back together so that they function as one. In essence, computational model discovery ultimately requires a computational representation (model) of the discovery process itself, which describes the interplay of theory, representation and analysis that occurs during discovery. We will call this a *meta-model* here (a model by which analytical models are constructed), to avoid confusion with the analytical models it may create. For our purposes, this meta-model needs the capacity to propose, synthesise and evaluate plausible discovery workflows, which we can think of as the basic steps by which discoveries might be made. Then we light the blue touch paper, stand back and cross our fingers!

One useful consequence of building systems that represent the science process is that they can provide a detailed description of exactly what has been done, and why, during any analytical activity. They therefore have a very important role to play in fostering reproducibility and reuse, because they can aid in the transfer of conceptual understanding, as well as procedural understanding (e.g. Konkol *et al*., 2019). Practically speaking, a meta-model also helps to constrain the kinds of analytical models that can exist, so it is also useful to restrict the truly massive search space that discovery systems must contend with.

In order to build such a system, we must first find a meta-model that is consistent with how we understand the process of research. There are many useful accounts of the process of geographical research, for example: Casetti (1972), Abler et al. (1992), Pratt (1995) and Silverman (2013). One that has stood the test of time and still has much to say is Harvey's (1969) book: '*Explanation in Geography*'. In it, Harvey works through the various steps and conceptual apparatus used in geographical research, showing how theory, philosophy, representation and analytical methods work together to support explanation. He furthermore recognises the very important—but often unrecorded—role that these conceptual aspects play. In the remainder of this section, we will work through how a computational explanation system might be constructed by combining Harvey's conceptual ideas with the more practical workflow used in inductive process modelling (Section 3—introduction).

Harvey (1969) describes explanation as weaving together the following threads:

- ***methodological frameworks****: the nature of investigation and*
- ***philosophy****: the research process and its various conceptual artefacts (includes ontology),*

— *which determine* ***representation:*** *how we abstract and represent the world and*
— ***analysis:*** *how we model and analyse the world*
— *through to* ***explanation****: which uses theory to describe what our analysis reveals.*

Granted, it may not be the only valid science model, nor is there any reason to insist that it should be since in theory a computational system could support many different meta-models of science, or more broadly, of research.

Our understanding of the themes touched upon by Harvey has been remarkably stable over the past 50 years; what has changed dramatically is our ability to support them using computational methods on a truly massive scale. Harvey's arguments (from 1969) represent a time and approach where theory and methodology were considered as first class citizens and it serves as a useful reminder to us that our current approach of ignoring them or separating them from analysis in our computational systems is fundamentally flawed. In expanding Harvey's ideas to accommodate discovery of theory from data, it is necessary to include four additional components: the data, associated data structures, the processes (theory fragments) and constraints operating on these relations. Placing the process of inductive process modelling in the context of Harvey's approach to explanation, and merging the common terms, gives the following structure, extended from Bridewell et al. (2008):

Given:

— a ***methodological framework*** for research and
— a ***meta-model*** for the process of research and
— a *set of* ***representational models*** (domain data models) and associated data structures,
— and ***observations*** for a set of variables as they change over time and / or space;
— a *set of* ***entities*** that may participate in the model (such as map features, categories) and
— a *library of* ***processes*** that specify relationships among entities and
— a *set of* ***constraints*** describing plausible relationships and values among processes and entities;

Find:

— a ***specific process model and associated parameterization*** that not only explains the observed data but also predicts unseen data accurately.

Such a process model would be anchored by the ***methodological framework*** and ***meta-model*** adopted, and could be interrogated to explain its inner workings both mathematically and conceptually in terms of domain theory. A key question for us, then, is: *What is our state of maturity for each of the above items?*

Beginning with the more concrete items, there are many mature ***domain data models*** specifically developed for geography and GIScience. Examples are the Open Geospatial Consortium's *Simple Feature Model*[5] *and* the *WaterML* data model developed by Valentine *et al*. (2010). Both are implemented in existing systems and work well. A more comprehensive (and ambitious) model, which integrates the field and object view of geographical information, is proposed by Goodchild *et al*. (2007). These domain models also give us the structure for any ***entity*** instances that the analytical processes need. It is clear too, that we have amassed a huge and growing body of ***observations***, including decades of remote sensing and census data, to give but two examples.

[5] https://www.opengeospatial.org/standards/sfa

A typical GISystem contains literally hundreds of methods for conditioning, selecting, analysing and mapping geospatial data. Many of these methods have been collected together into open-source libraries, such as *Pysal and GeoDA* (Anselin, 2014), or the spatial methods added to the statistical package, *R* (Diggle and Giorgi, 2019). These are our candidate ***process*** libraries. A useful overview of both form and process in GIScience research is provided by *Goodchild (2004).*

As for ***constraints***, we are less well organised, I believe. Theory provides some useful constraints on the range of values that some equation terms can take, but we know far less about how methods are used in practice and how they are combined into workflows. Put another way: if we could mine the millions of GIS projects conducted to date, what might we learn about how analytical methods are combined in practise? This information could be valuable to constrain the search space for discovery.

The more conceptual aspects of discovery have also received significant attention, but generally in isolation from analytical systems. ***Methodological frameworks*** and ***meta-models*** are difficult to disentangle, and the meta-models we currently have implicitly assume a methodological framework based around one or more of the first three science paradigms described in the introduction.

A formal model of conceptual change as it applies in scientific discovery was first proposed by Thagard (1990) and followed by an example of the model used to represent changes in geological concepts following the acceptance of the theory of plate tectonics (Thagard and Nowack, 1990). This earlier work was not implemented in a computational system at the time, but it could be now; De Jong & Rip (1997) propose how this might be achieved. Brodaric *et al*. (2011) describe a rich, formal model of scientific knowledge that connects data with assertions, propositions, laws and inference activities, all packaged as an implemented ontology called *SKI*. Dos Santos and Travassos (2016) provide a very useful review of the state of the art in scientific knowledge engineering, and contrast the many approaches to date. Gupta & Gahegan (*submitted*) describe an example of such a system, that captures and represents the evolution of categories within a combined analytical and knowledge environment—that is, containing a meta-model of how categories evolve during use.

While it seems clear that usable solutions are weakest for methodological frameworks and meta-models of discovery, one clear advantage that GIScience has over many other disciplines is that we have actively encouraged work in this area, and there is already a body of work on representing our semantics and pragmatics. Indeed, all the requisite pieces described above have attracted much research effort over the years, but seldom are they considered within the context as a single system; the task remaining for us is to recognise how they fit together and synthesize them into a complete whole—a system of systems—rather than pursuing them as independent threads. This is by no means a trivial task, in fact it is a grand challenge.

5 Discussion and Conclusions

At the core of the computational model discovery movement is the idea that not only can data be turned into knowledge, but that this knowledge can in turn be expressed as theory. When the data mining or machine learning community refers to knowledge discovery, it typically means uncovering the patterns and categories that characterize trends or outliers in data; such 'knowledge' is usually expressed in the language of data (statistics) or machine learning method parameterisations and *not* the language of human explanation (domain theory). But in the case of computational model discovery, the discovered knowledge is a mathematical model (e.g. a set of Partial Differential Equations) that characterises some phenomena of interest and that can be translated (via these equations) into rules and relations between data—for example as

processes and causal links. This is closer to theory, and can be interpreted from a theoretical perspective in many cases. Over time, as additional layers of the science process are adequately represented in computational systems, these discovery systems will become capable of producing rich and detailed descriptions of solutions and their explanations, couched in relevant domain theory and the process of research.

Computational model discovery is a big step forward in terms of inference, but like all techniques, if it is used without understanding its limitations, it is likely to lead to errors. Firstly, there is still no guarantee that a model (of whatever kind) that is learned from data is valid, no matter how that model is constructed. Careful validation is always required. Secondly, there is danger too if we ever forget that correlation in data does not necessarily indicate a causal relationship in nature. However, since the model in this case can be interpreted by humans, a non-valid model is easier to recognise, reject or modify. Finally, and in common with all inductive methods (including inferential statistics) the model is, at best, as good as the data. But as datasets and associated explanatory models become more complex and difficult to fathom, there is also the tantalizing possibility that hitherto unrecognised and valid theory may be indeed be automatically generated.

It is a somewhat frightening and confronting prospect that discovery might happen in the absence of domain experts, it challenges some long-held beliefs about the valid roles computers can play in supporting research. It is also potentially exciting and helpful in our quest to better understand the complexities of the world around us. But to revisit the opening section, there seems to be no reason why computational discovery should herald the "*end of theory*" (Anderson, 2008) since we have shown how a learned explanation can be couched in theory, and be stronger for it.

If we wish to continue to exploit big data and the power of machine learning for geographical analysis, we need to move beyond prediction and towards explanation. Explanations are debatable, verifiable and provable and we need science that we can rely on and back up if we are to base policy arguments around it. The proposed framework for geographical explanation (Section 4), using inductive process modelling for model discovery (Section 3) appears to be a good place to begin. Obviously, the basis for explanation may need to change as we move from one problem domain to another: laws and equations that apply in hydrology may not make sense in demography or ecology. It may also be the case that the research process and methodology change as we move from subject to subject, requiring deeper changes to the meta-model of the discovery process itself. But there is the tantalising prospect that we might also find some equation fragments or forms that appear in many geographical models. For example: Does Tobler's Law show up in some form in learned geographical process models? Do other equations or relationships appear with regularity? Are there in fact other geographical 'laws' that can be shown to apply to a large number of models? Computational model discovery provides a mechanism by which we might uncover more of this underpinning theory.

Acknowledgements

Many thanks to Dr. Adam Arvay for providing figures 1 and 2 to use as examples.

References

Abler, R. F., Marcus, M. G. and Olson, J. M. (eds.) (1992). *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography.* Rutgers University Press.

Anderson, C. (2008). The end of theory. *Wired Magazine* 16.

Anselin, L. (2014). *Modern spatial econometrics in practice: A guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press.

Arvay, A. (2018). Computational scientific discovery using rate-based process models. Ph.D. thesis available electronically from: https://researchspace.auckland.ac.nz/handle/2292/37637.

Arvay, A. and Langley, P. (2016). Selective induction of rate-based process models. *Advances in Cognitive Systems*, 4, 1-15.

Asgharbeygi, N., Langley, P., Bay, S. and Arrigo, K. (2006). Inductive revision of quantitative process models, *Ecological Modelling* 194 (1-3), 70-79.

Borrett, S. R., Bridewell, W., Langley, P. and Arrigo, K. R. (2007). A method for representing and developing process models. *Ecological Complexity*, 4, 1-12.

Brenning, A. (2005). Spatial prediction models for landslide hazards: review, comparison and evaluation, *Natural Hazards Earth System Science*, 5, 853-862.

Bridewell, W., Langley, P., Todorovski, L., and Dzeroski, S. (2008). Inductive process modeling. *Machine learning,* 71(1), 1-32.

Brodaric, B., Reitsma, F. and Qiang, Y. (2008). SKIing with DOLCE: Toward an e-Science knowledge infrastructure. *Frontiers in Artificial Intelligence and Applications*, 183(1), 208-219.

Carenini, G., and Moore, J. D. (2006). Generating and evaluating evaluative arguments. *Artificial Intelligence*, *170*(11), 925-952.

Casetti, E. (1972). Generating models by the expansion method: applications to geographical research. *Geographical analysis*, *4*(1), 81-91.

Chen, J., and Zipf, A. (2017). DeepVGI: Deep learning with volunteered geographic information. In: *Proceedings of the 26th International Conference on World Wide Web Companion.* International World Wide Web Conferences Steering Committee, (pp. 771-772).

Cheng, T., Haworth, J. and Manley E. (2012). Advances in geocomputation (1996–2011). *Computers, Environment and Urban Systems*, 36 (2012), 481–487.

De Jong, H., and Rip, A. (1997). The computer revolution in science: Steps towards the realization of computer-supported discovery environments. *Artificial intelligence*, *91*(2), 225-256.

Dos Santos, P. S. M., and Travassos, G. H. (2016). Scientific Knowledge Engineering: a conceptual delineation and overview of the state of the art. *The Knowledge Engineering Review*, *31*(2), 167-199.

Diggle, P. J., and Giorgi, E. (2019). *Model-based Geostatistics for Global Public Health: Methods and Applications*. Chapman and Hall/CRC Press.

Džeroski, S., Langley, P., and Todorovski, L. (2007). Computational discovery of scientific knowledge. In S. Dzeroski and L. Todorovski (Eds.), *Computational discovery of communicable scientific knowledge*. Berlin: Springer.

Džeroski, S. D. and Todorovski, L. (Eds.) (2007). *Computational discovery of 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*(4), 360-368.

Fischer, M. M. (2013): Neural spatial interaction models. Network training, model complexity and generalization performance. In Murgante, B., Misra, S., Carlini, M., Torre, C.M., Nguyen, H.-Q., Taniar, D., Apduhan, B.O., and Gervasi, O. (Eds.): *Computational Science and Its Applications - ICCSA*, Part IV, Lecture Notes in Computer Science, vol. 7974, Springer, Heidelberg, pp. 1-16.

Gahegan, M. N. 2003. Is inductive machine learning just another wild goose (or might it lay the golden egg?). *International Journal of Geographical Information Science* 17, 69-92.

Gahegan, M. (2005). Beyond tools: visual support for the entire process of GIScience. In: Dykes, J., MacEachren, A. M. and Kraak, M.-J. (Ed.) *Exploring geovisualization* , Elsevier, pp. 83-99.

Gahegan, M. (2009). Visual exploration and explanation in geography: analysis with light. In: (Miller, H and Han, J., Eds) *Geographic data mining and knowledge discovery*, Ch 11, pp. 291-315.

Goodchild, M. F. (2004). GIScience, geography, form, and process. *Annals of the Association of American Geographers*, 94(4), 2004, pp. 709–714

Goodchild, M. F., Yuan, M. and Cova, T. J. (2007). Towards a general theory of geographic representation in GIS, *International Journal of Geographical Information Science*, 21(3), 239-260, DOI: 10.1080/13658810600965271

Gould, P. (1999). *Becoming a Geographer*. Syracuse University Press.

Gupta, P. and Gahegan, M. (submitted). Categories are in flux but their computational representations are fixed. That's a problem. Under review with *Transactions in GIScience*.

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, *102*(478), 507-516.

Harvey, D. (1969). *Explanation in Geography*, Arnold: London.

Hey, T., Tansley, S. and, Tolle, K. (Eds.) (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, WA, USA.

Holling, C. S. (July 1959). Some characteristics of simple types of predation and parasitism. *The Canadian Entomologist.* 91 (7) pp. 385–98. doi:10.4039/Ent91385-7.

Horacek, H. (2017). Requirements for Conceptual Representations of Explanations and How Reasoning Systems Can Serve Them. In *Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017)*.

Koltunov, A., Ustin, S. L., and Prins, E. M. (2012). On timeliness and accuracy of wildfire detection by the GOES WF-ABBA algorithm over California during the 2006 fire season. *Remote sensing of environment*, *127*, 194-209.

Konkol, M., Kray, C., and Pfeiffer, M. (2019). Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study. *International Journal of Geographical Information Science*, *33*(2), 408-429.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lam, W. and Bacchus, F. (1994). Learning Bayesian Belief Networks: An Approach Based on the MDL Principle. *Computational Intelligence* 10: 269-294.

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

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

Miller, H. and Han, J. (Eds.). (2009). *Geographic Data Mining and Knowledge Discovery*. Boca Raton: CRC Press.

Mou, L., Ghamisi, P., and Zhu, X. X. (2017). Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(7), 3639-3655.

Openshaw, S., Charlton, M., Wymer, C. and Craft, A. (1987) A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets, International Journal of Geographical Information System, 1:4, 335-358.

Openshaw, S. (1995). Developing automated and smart spatial pattern exploration tools for geographical information system applications. *The Statistician*, 44, No 1, pp 3-16.

Openshaw, S., Cross, A. and Charlton, M.E. (1990). Building a prototype geographical correlates exploration machine. *International Journal of Geographical Information Systems*, 3: 297–312.

Pazzani, M. J., Mani, S., and Shankle, W. R. (2001). Acceptance by medical experts of rules generated by machine learning. *Methods of Information in Medicine*, 40, 380-385.

Pearl, J., (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29, 241-288.

Popper, K. (1959). *The Logic of Scientific Discovery*. Abingdon-on-Thames: Routledge.

Pratt, A. C. (1995). Putting critical realism to work: the practical implications for geographical research. *Progress in Human Geography*, *19*(1), 61-74.

Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

Schwabacher, M., and Langley, P. (2001). Discovering communicable scientific knowledge from spatio-temporal data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 489-496). Williamstown, MA: Morgan Kaufmann.

Shrager, J. and Langley, P. (editors) (1990). *Computational Models of Scientific Discovery and Theory Formation*, Morgan Kaufman.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. and Chen, Y. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), p.354-359.

Silverman, D. (2013). *Doing qualitative research: A practical handbook*. SAGE Publications Limited.

Sozou P.D., Lane P.C., Addis M. and Gobet F. (2017). Computational Scientific Discovery. In: Magnani L., Bertolotti T. (eds.) *Springer Handbook of Model-Based Science*. Springer Handbooks. Springer, Cham.

Stassopoulou, A., Petrou, M. and Kittler, J. (1998). Application of a Bayesian network in a GIS based decision making system, *International Journal of Geographical Information Science*, 12:1, 23-46

Thagard, P. and Nowak, G. (1990). The conceptual structure of the geological revolution. In: Shrager, J. and Langley, P. (editors), *Computational Models of Scientific Discovery and Theory Formation*, pp. 27—172, Morgan Kaufman.

Valdez-Perez, R. E. (1996). Computer science research on scientific discovery. *Knowledge Engineering Review*, 11, 51–66.

Valentine, D., Taylor, P., and Zaslavsky, I. (2012, April). WaterML, an information standard for the exchange of in-situ hydrological observations. In *EGU General Assembly Conference Abstracts* (Vol. 14, p. 13275). See also: https://www.opengeospatial.org/standards/waterml

Wang, S., Wilkins-Diehr, N. R., and Nyerges, T. L. (2012). CyberGIS-Toward synergistic advancement of cyberinfrastructure and GIScience: A workshop summary. *Journal of Spatial Information Science*, (4), 125-148.

Williams, K., Bilsland, E., Sparkes, A., Aubrey, W., Young, M., Soldatova, L.N., Grave, K.D., Ramon, J., Clare, M.D., Sirawaraporn, W., Oliver, S.G., & King, R.D. (2015). Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases.

Journal of the Royal Society: Interface, 12: (104) (online publication, no page numbers): https://royalsocietypublishing.org/doi/full/10.1098/rsif.2014.1289.

Yang, L., MacEachren, A., Mitra, P., and Onorati, T. (2018). Visually-enabled active deep learning for (geo) text and image classification: a review. *ISPRS International Journal of Geo-Information*, 7(2), 65.