Concept Formation in Scientific Knowledge Discovery from a Constructivist View

Wei Peng[1] and John S. Gero[2]

[1] Platform Technologies Research Institute, School of Electrical and Computer Engineering, RMIT University
Melbourne, VIC 3001, Australia
w.peng@rmit.edu.au
[2] Krasnow Institute for Advanced Study and Volgenau School of Information Technology and Engineering, George Mason University, USA
john@johngero.com

Abstract. This chapter argues that the computer-aided scientific knowledge discovery tool should facilitate scientific knowledge development through assisting scientists to build first-person knowledge and third-person knowledge. The chapter reviews cognitive theories of human knowledge construction and presents a hydrological modelling scenario as an exemplar of these concepts. A number of challenges for designing such a system have been discussed.

Key words: Scientific Knowledge Discovery, Constructivism, Concept Formation, First-person Knowledge, Third-person Knowledge

1 Introduction

The central goal of scientific knowledge discovery is to learn cause-effect relationships among natural phenomena presented as variables and the consequences their interactions. Scientific knowledge is normally expressed as scientific taxonomies, qualitative and quantitative laws [1]. This type of knowledge represents intrinsic regularities of the observed phenomena that can be used to explain and predict behaviors of the phenomena. It is generalization that is abstracted, externalized from a set of context and applicable to a broader scope. Scientific knowledge is a type of third-person knowledge independent of a specific enquirer. Artificial intelligence approaches, particularly data mining algorithms that are used to identify meaningful patterns from large data sets, are approaches with the aim to facilitate the knowledge discovery process [2]. A broad spectrum of algorithms has been developed in addressing classification, associative learning and clustering problems. However, their linkages to people who use them have not been adequately explored. Issues in relation to supporting the interpretation of the patterns, the application of prior knowledge to the data mining process and addressing user interactions remain prominent challenges for building knowledge discovery tools [3]. As a consequence, scientists rely on their experience to formulate problems, evaluate hypotheses, reason about untraceable factors and derive

new problems. This type of knowledge which they have developed during their career is called "first-person" knowledge. The formation of scientific knowledge (third-person knowledge) is highly biased by the enquirer's first-person knowledge constructs, which is a result of his or her interactions with the environment. There have been attempts to craft automatic knowledge discovery tools but these systems are limited in their capabilities to handle the dynamics of personal experience. There is now trends in developing approaches to assist scientists applying their expertise to model formation, simulation and prediction in various domains [4], [5]. On the other hand, first-person knowledge becomes third-person theory only if it proves general by evidence and is acknowledged by a scientific community. Researchers start to focus on building interactive cooperation platforms [1] to accommodate different views into the knowledge discovery process.

There are some fundamental questions in relation to scientific knowledge development. What are major components for knowledge construction and how do people construct their knowledge? How is this personal construct assimilated and accommodated into a scientific paradigm? How can one design a computational system to facilitate these processes? This chapter does not attempt to answer all these questions but serves as a basis to foster thinking along this line. A literature review about how people develop their knowledge is carried out through a constructivist view. A hydrological modelling scenario is presented to elucidate the vision.

2 Concept Formation from a Constructivist View

Cognitive science is a multi-disciplinary study with the aim to deliver a theory of intelligence. The basic assumption held by many cognitive science researchers is that there is a common set of principles underlying all instances of intelligence [6]. Aristotle attributed to perception and observation essential roles in acquiring scientific knowledge and proposed an empirical method of gathering observations followed by taxonomic classification, interpretation and inference [6].

2.1 Knowledge as Generalization

One aspect of human cognition is to develop experience and use experience to construct a judgment, in which a certain object is distinguished from other objects and is characterized by some concepts. Concepts are bearers of meanings. The enquiry for the notion of concept has long been the focus of research in philosophy. Both the notion of "concept" and our understandings of the way in which concepts are formed have evolved. John Locke [7] described that a general idea corresponds to a description of concept, which is created by abstracting and drawing commonalities from the particulars. The assumption of a concept as a consequence of induction is that the unobserved events conform to regularities in the already known facts. The assumption is not general enough to address possible exceptions in a dynamic complex world. David Hume [8] argued that discovering "necessary connexion" between objects leads to a better

understanding of *a priori* causation relationships around these objects. He also mentioned that relationships between ideas of causation can be derived from our experience [9]. David Hume's theories inspired Immanuel Kant, who later developed the notion of "*a posteriori*" concept. According to Kant, a concept can be further defined as an *a posteriori* concept or an *a priori* concept [10]. *A posteriori* or empirical concepts are abstracted/induced from specific perceived events and are applicable to them. *A priori* concepts are categories that are not abstractions from perception but are applicable to it. Although there is no unified definition of a concept, a concept is often taken to mean a mental representation of a class or a category [11]. The classical view of concept formation as abstraction or "abstract thinking", as outlined by Van Oers [12], emphasizes creating types, a process of generalizing by removing circumstantial aspects of time and place [13]. This view has been manifested in the recent research on concept formation systems in the field of artificial intelligence. Many researchers consider categorization as the essence of a concept and its formation. Concept formation has been regarded as a process of incremental unsupervised acquisition of categories and their intentional descriptions [14]. Based on this view, a broad spectrum of computational models has been developed, including inductive learning methods, explanation-based learning approaches and connectionist algorithms. However, theories of concept formation that merely focus on categorization are not able to address the complexity of the world [15]. A concept lacking an understanding of why and how the object, entity or event has its particular properties is called a protoconcept [15], [16]. The process by which people form any category knowledge biases the knowledge contents formed. Theories, models are people's approximations to the uncertain (and unknown) *a priori* that rule the universe. They are by nature *a posteriori*. For example, the concept of gravitation in modern physics is cognized differently to what was taken for granted in Newton's time.

2.2 Knowledge as Construction

Knowledge is an empirical term therefore inseparate from a subject and the world external to that subject. In describing the relationship between human experience and nature, John Dewey [17] pointed out that there is a union of experience and nature in natural science. Human experience is considered as "a means of continually penetrating the hearts of the reality of nature". The enquirer must resort to his or her experience and use empirical methods if his or her findings are to be treated as genuinely scientific [17]. Referring to the genesis of knowledge, Dewey [17] mentioned that knowledge is a refined experience:

> *"The intrinsic nature of events is revealed in experience as the immediately felt qualities of things. The intimate coordination and even fusion of these qualities with regularities that form the objects of knowledge, in the proper sense of the word 'knowledge', characterizes intelligently directed experience, as distinct from mere casual and uncritical experience".*

Dewey's vision can also be found in other constructivists' works. In reviewing Piaget's conception of knowledge and reality, Von Glasersfeld [18] conceived that the cognitive organism is first and foremost an organizer who interprets experience and shapes it into a structured world. Piaget used the term "schema" to name the basic mental structure, which depicts how perceptual categories are organized. After examining the development of intelligence in children, Piaget [19] concluded that two intertwined processes ("assimilation" and "accommodation") and their coordination enable a child to construct knowledge from his or her experience. Assimilation tends to subordinate the environment to the organism's *a priori* or acquired schemata, whereas accommodation adapts the organism to the successive constraints of the environment by updating (or incorporating) new schemata [19]. Piaget mentioned [18], [20]:

> *"All knowledge is tied to action and knowing an object or an event is to use it by assimilating it to an action scheme ..... this is true on the most elementary sensory-motor level and all the way up to the highest logical-mathematical operations ....."*

Piaget's theory of assimilation can be interpreted as the mapping between the external world to the existing knowledge structures in the internal world of a cognitive organism. The development of new knowledge structures in the internal world of a cognitive agent is through the accommodation mechanism (Fig. 1).

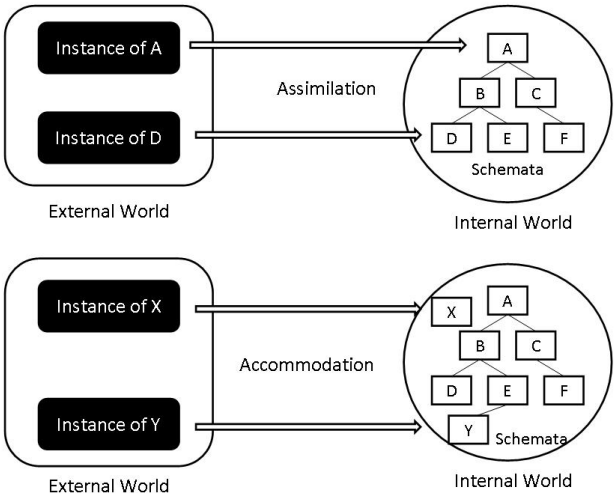

Fig. 1. An interpretation of Piaget's cognition processes of assimilation and accommodation.

Vygotsky [21] introduced the role of "activity" in knowledge construction, defining that activities of the mind cannot be separated from overt behavior, or from the social context in which they occur. Social and mental structures interpenetrate each other [21], [22].

2.3 Experiential Learning

These theories provide insights about ingredients for knowledge construction based on the external events or context, the internal knowledge structures and the interactions between the assimilation and accommodation processes. Kolb [23] developed an experiential learning theory by drawing ideas from Dewey [24], Lewin [25] and Piaget [19]. The major components for learning knowledge consist of concrete experience, observation and reflection, the formation of abstract concepts and testing of implications of concepts in new situations. He represented this idea in the experiential learning circle, which is illustrated in Fig. 2.

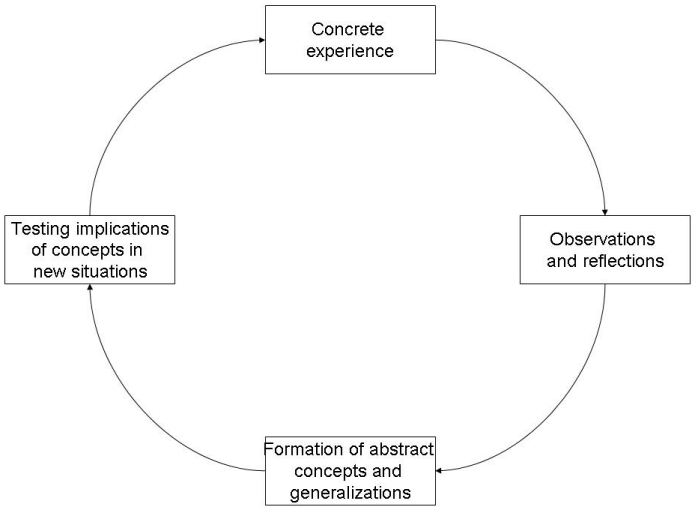

Fig. 2. Experiential learning cycle (adapted from Figure 2.1 of [23]).

Kolb and Fry [26] postulated that learning can commence from any components and operate in a continuous cycle. The observations are assimilated into a "theory" based on the immediate concrete experience. The theory is generalization that can be used to deduce new implications for action. The implications then serve as guides for creating new experience. This model provides operational features for Piaget's notion of "mutual interaction" between assimilation and accommodation. Jarvis [27] enriched the internal process for knowledge construction by introducing flexible behaviors of an intelligent system. He put

experience and memory in the loop. As illustrated in Fig. 3, these behaviors[1] include:

- Non-learning presumption (or non-consideration) (boxes 1 → 2 → 3 → 4) where people react through experience (or choose not to respond to environmental events);
- Non-reflective pre-conscious (boxes 1 → 2 → 3 → 6 → 4 or 9) where people have experience about environmental events but do not attend to that;
- Non-reflective practice (boxes 1 → 2 → 3 → 5 → 8 → 6 → 4 or 9) where people obtain basic skills;
- Non-reflective memorization (boxes 1 → 2 → 3 → 6 → 8 → 4 or 9) where people memorize new things;
- Reflective contemplation (boxes 1 → 2 → 3 → 7 → 8 → 6 → 9) where people reflect upon a situation and make decisions;
- Reflective practice (boxes 1 → 2 → 3 → 5 → 7 → 5 → 8 → 6 → 9) where the individuals reflect and then act upon a situation;
- Reflective experimental learning (1 → 2 → 3 → 7 → 5 → 7 → 8 → 6 → 9) in which people reason about the situation and evaluate their experience.

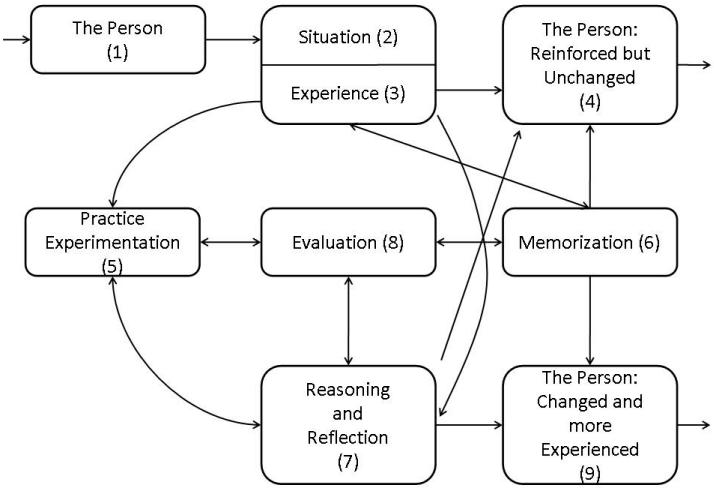

Fig. 3. Experiential learning path from Jarvis [27], [28]. Boxes represent components that consist of the process of learning. These components are general notions that can be either entities or processes. Numbers refer to individual processes described in the text.

[1] The Rejection behavior where the individual refuses to learn from the situation is not list here due to its irrelevance to knowledge construction.

Jarvis' experiential learning model[2] emphasizes the role of experience and cognition in developing new knowledge. Another component of Jarvis' learning model is the pragmatic unit which includes experiment and evaluation processes. A key notion which is implicit in Jarvis' theory but put forward by Kolb is "prediction" (termed as "implication for a concept" or "hypothesis" in [23]). Anticipation plays a key role in human cognition. Human activity of knowing (scientific thoughts and their epistemological interpretations) are the highest form of adaptation, which involves reflecting on past experience, abstracting specific regularities from them, and projecting these as predictions into the future [30]. This behavior of knowing is also described in the "memory prediction framework" [31] and is conjectured as the basic function of our memory system. Our memory serves as a "metaphysical linkage of times past and future" [32]. This means that our memory system constantly uses experience to form invariant representations[3] about the spatial-temporal (and feature) based events, predicts potential occurrences in the environment and biases behaviors of our sensory-motor system accordingly. Based on this understanding, this chapter will present the predominant theory of memory in constructivism and elaborate the concept formation process.

2.4 Concept Formation in Constructive Memory

Memory in cognitive science does not mean a place or device storing descriptions of actions. Neither does it refer to the information that is encoded, stored and retrieved. The view of memory as an encoding-storage-retrieval device cannot account for phenomena such as "false recognition", "intrusion" and "confabulation" [33]. Memory is predominantly conceived as a mental capability for us to make sense, to learn new skills and to compose something new [13]. The constructive view of memory can be traced back to Dewey in the *The Reflex Arc Concept in Psychology* (quoted by [34]):

> *"Sequences of acts are composed such that subsequent experiences categorize and hence give meaning to what experienced before."*

The theory of constructive memory is supported by many cognitive studies [35], [36], [37]. The basic functions of a constructive entity are described by Riegler [33] as:

1. The cognitive apparatus creates a structure;
2. It interacts with other structures (such as the surrounding environment or the older structure of its apparatus);

[2] Jarvis' recent learning model ([28], [29]) extends the experiential learning to lifelong learning theory thus has an emotive component.

[3] Invariant representations are knowledge structures that can be used to classify an object or infer potential acts. For example, the knowledge structures you rely on to recognize a friend's face should always work even if the distance and the orientation of your friend's face vary.

3. It compares the newly created structures with those encountered in step 2;
4. It adapts the structures when needed before returning to step 1.

The cognitive apparatus tends to maintain the fitness of a knowledge structure in relation to the changing environment. For example, if you have never been to Australia, your representation of "swan" is induced from your Northern Hemisphere exposure to a type of elegant white birds from the Anatidae family. Suppose someone tells you that there are black swans in the Southern Hemisphere, you then carry out a literature search. You find that there indeed exist black swans and then realize the concept of swan needs to be adapted to accommodate this new information. Such characteristics of our memory system are in concordance with the reflection and evaluation processes of experiential learning. The essence of the constructive memory is the means of adapting knowledge structures to a dynamic environment. Reigler's constructive memory functions emphasize changes of knowledge structures during a construction process. However, the lack of descriptions of macroscopic behaviors of the involved person leads to speculations about how the structures will change. It is suggested that describing the constructive memory in the context of human experiential learning behavior leads to operational features of a memory system.

The constructive memory process may be viewed as the process of transforming information during interactions. An important notion introduced to illustrate the concept formation process in a constructive memory system is "grounding". Symbolic grounding explores the means by which the semantic interpretation of a formal symbol system can be made intrinsic to that system, rather than relying on the meanings in the head of a third-person interpreter or observer [38]. The grounding problem generally refers to representation grounding [39] or grounding of a concept, in which the concept can be developed through interactive behavior in an environment [40]. A grounding process here is referred to as the evaluation of whether constructed knowledge structures correctly predict environmental changes. The basic information units are "stimuli", "anticipation", "proto-concept", "hypothesis", "concept" and "invariants". Stimuli denote environmental data prior to a constructive memory process. Anticipation is the term for responsive information based on the agent's[4] experience, predicting potential environmental changes. A concept is a result of an interaction process in which meanings are attached to environmental observations. We use the term "proto-concept" to describe the intermediate state of a concept. A proto-concept is a knowledge structure that depicts the agent's interpretations and anticipations about its external and internal environment at a particular time. The term "invariant" has been defined in the previous section as the knowledge structures that an agent uses to identify categories and predict potential acts in the environment. In the scientific knowledge discovery context, this conceptual knowledge structure is composed of sets of categorized abstractions and causal relationships between various observations in various spatial-temporal scales. The term "hypothesis" is associated with agent's explanations for discrepancies between

[4] We use the term agent to represent cognitive apparatus.

its prediction and environmental changes. The grounding of proto-concepts and derived anticipations (or hypotheses) produces a concept. We define concepts as the grounded invariants over the agent's experience. They are abstractions of experience that confer a predictive ability for new situations [41], [42]. On the other hand, a concept contains context-dependent specifications for an abstraction, which are encapsulated in anticipations. The concept formation mechanism (described in Fig. 4) consists of the following processes:

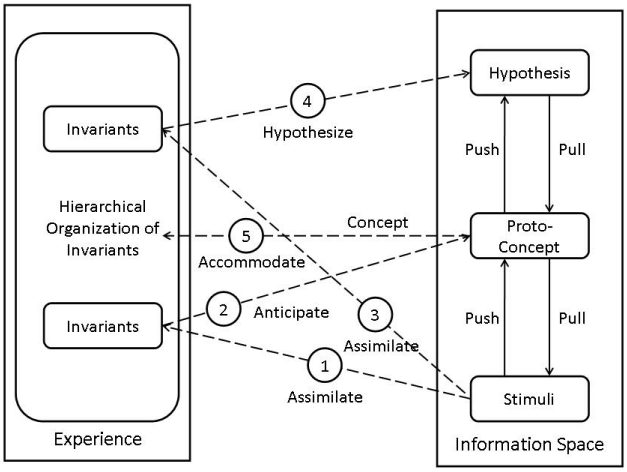

Fig. 4. A view of concept formation in constructive memory

1. Process 1 is where the agent assimilates the encountered stimuli to activate a response from its experience;
2. In Process 2, the agent generates a proto-concept which contains the categorical information and the related prediction;
3. Process 3 is where the agent assimilates the stimuli to create hypotheses for invalid predictions. This process can be called reflection;
4. In Process 4, the agent produces a hypothesis based on the deduced explanations and the re-activated experience;
5. In Process 5, the agent accommodates the concept (validated proto-concept) into experience;
6. The "push" process is a data-driven process where changes in the environment (or autogenous variables of the agent) trigger the transformation of these changes into changes in the experience [43];
7. The "pull" process is an anticipation-driven process where if the process for validating a proto-concept (or a hypothesis) requires the agent to obtain a particular pattern of information from the environment, then the process

is biased in the way that external variables (or autogenous variables of the agent) are filtered out or emphasized [43];
8. The validation process compares the prediction with the pulled data to determine the validity of a proto-concept.

The "push" and "pull" processes are two major components in charge of knowledge structure transformations. The push process uses Processes 1-4 to achieve its goal, which is to transform changes from the environment to the agent's experience to acquire a response. The pull process performs anticipation-driven data acquisition and data construction, which underpins the validation process.

As illustrated in Fig. 4, the agent processes the environmental data in the push process, in which data (stimuli) are transformed into categorical information and predictions based on invariants held in its experience (Processes 1 and 2). The agent subsequently tests the validity of its proto-concepts in a pull process. An ill-formed proto-concept triggers hypothesis-testing processes (Processes 3 and 4) in which the agent reasons and creates explanations for discrepancies between the proto-concepts and observations. The explanations can be used to deduce hypotheses and the associated observation requirements. A well-grounded proto-concept (obtained after processes push and pull) is called a concept and accommodated into the agent's knowledge structures as experience (in Process 5).

2.5 From First Person Construct to Third Person Knowledge

An individual's knowledge can be treated as first-person constructs, which are biased by a person's past experience in his or her interactions with the social context. Scientists build models and theories that represent their individual interpretations of natural phenomena. This type of knowledge is then transformed into a type of generalized knowledge that can apply to non-specific events and is independent of their enquirers. This generalized knowledge once supported by scientific evidence becomes instances of a scientific paradigm that is a form of third-person knowledge. Thomas Kuhn described the notion of paradigm from his philosophical viewpoint of science in *The Structure of Scientific Revolutions* [44]. A scientific paradigm refers to scientific achievements that attract an enduring stream of a scientific community to follow the same rules and standards for their practise. A scientific paradigm centers a space for people to redefine and resolve problems. The development route of "normal science", which is built upon past paradigms, is a non-accumulative multiple trajectories of fact-gathering, paradigm formation and paradigm revolution [44].

According to Kuhn [44], the reconstruction of prior theory and the re-evaluation of prior facts enable a new theory to be assimilated. A normal scientific paradigm tends to reject radically new theories. Normal science is to fit nature into the rigid, incomplete explanatory framework that a paradigm defined. A scientific community is likely to accept and encourage esoteric research works [44]. An

individual's effort is limited and therefore a scientific revolutionary process is seldom achieved by a single person and never overnight [44].

Supporting both individuals and a scientific community to accommodate new and emergent knowledge holds the key for scientific discovery. This requires any computer-based discovery support tool to adopt constructive and interactive approaches. In the next section, scientific concept formation based on a hydrological modelling scenario is presented to exemplify this idea.

3 Concept Formation in a Hydrologic Modelling Scenario

The movement of water through a permeable medium (i.e., soils) has been widely accepted as a complex phenomenon. The soil water behaviors are dynamic in the sense that precipitation, soil texture and profile, the presence of plants, land use and a variety of meteorological variables influence the spatial distribution and temporal evolution of soil moisture [45]. At the same time, water penetration into the soil changes the physical properties of the soil, which can further impact on the soil water behaviors over time. Precise soil moisture measurement to obtain ground truth in this dynamic environment is logically and economically infeasible. Therefore, approximation and simulation become dominant approaches for hydrologists. Hydrological models are simplified representation of the water system within the environment, which can assist hydrologists to understand, explain and predict the behavior of water. Three paradigms of hydrology modelling are empirical, physical and conceptual models [46]. The aim of an empirical modelling approach is to unveil relationships between hydrologic variables that have been observed from the environment without considering the complex hydrologic processes. It is a black box approach based on data. Many data mining and analytic approaches can be used to obtain empirical models. For example, various artificial neural networks have been used in soil classification and rainfall runoff mapping. Physical hydrologic models are based on the knowledge of physical laws for water behavior in different media. They are presented in parameterized equations, which state the understanding of relationships of hydrologic variables. Conceptual models are developed based on hydrologists' endeavors to leverage the data-driven models and their physical interpretations. Hydrologists harness their knowledge of physical laws and apply data mining/analytic tools on observations to learn concepts, which are further formalized as new or validated models.

In this section, a hydrologic modelling scenario based on the slope discharge model (developed by Michel [47] and cited by Baird [48]) is described as an example of scientific knowledge discovery. It is reasonable to consider hill-slope as a single quadratic reservoir rather than attempt to model dynamic flows within the slope. This is because hydrologic flows at a small scale are strongly heterogeneous and currently not possible to model (and validate). However, small-scale complexity can lead to simplicity at a larger scale if certain principal properties exist [48]. As a result, hydrologists always resort to empirical relations obtained from empirical approaches. A simple empirical relation between total volume of

water and discharge from the soil-filled trough can be formulated as a quadratic function [47]:

$$Q = \frac{V^2}{\tau \times V_0} \tag{1}$$

where Q is the total discharge from the slope (with physics dimensions L^3T^{-1}), V is the volume of drainable water remaining in the hill-slope, V_0 is the total volume of water that can drain from the slope (L^3), and τ is the system parameter (T). Empirical lumped reservoir models like Equation 1 can be effective if coupled with real-time observations and calibrations. However, empirical models are criticized as not physically meaningful [49]. In another camp, "bottom-up" hydrologists use their understanding of physical constraints to formulate soil water behavior observed in the laboratory settings. For example, Darcy's Law (cited by Baird [48]) introduces physical relationship between water flux and the hydraulic energy that drives the water movement, describing the flow in saturated soils:

$$Q_f = -K\frac{dh}{dx} \tag{2}$$

where Q_f[5] is the discharge *per unit area* (with physics dimensions $L^3T^{-1}L^{-2}$), K is the hydraulic conductivity (with physics dimensions LT^{-1}), denoting intrinsic permeability of the soil, h is the hydraulic pressure head and x is the distance in the direction of flow (L). The unsaturated version of Darcy's law is Richards' equation [50]. Representative models obtained from the laboratory environment (like Equations 1-2) are of limited usefulness for real heterogeneous slopes [51]. The applicability of a lab-induced model to real-time situations depends on how well one can address uncertainties in initial condition, time-varying inputs and time-invariant parameters. Applying these lab-based models to real-time environment is extremely expertise demanding. The deterministic approach is to use observations to inversely calibrate the parameters of a model (like the Darcy-Richards equation) [52], [53]. However, it is argued that deterministic models are not accurate representations of the real world. There is a growing interest in the inductive paradigm ("top-down" or "data-driven" approaches) in the hydrologic literatures [54], focusing on learning regularities from large volumes of near realtime data obtained from observations. In the mean time, scientific expertise in terms of domain knowledge remains a critical factor in providing physically meaningful interpretations (i.e. causalities) to empirical relationships learned from data.

A hydrologist learns a relationship between the target variable and other variables by investigating their dependencies, Fig. 5. The State Dependent Parameter [55] analytic approach may be used to this end. An assumption for this scenario is that water discharge (Q) appears to be more strongly dependent

[5] We use Q_f to depict the finite element of the discharge therefore the sum-up of Q_f at the bottom of the slope makes Q in Equation 1 that is the total discharge from the slope during a specific time.

on reservoir water (V) than any other variables. The hydrologist uses statistical approaches to identify the structure and order of the model. For example, Akaike Information Criterion [55] (cited by Young, et al. [54]) tends to identify over-parameterized models. The model structure is identified as quadratic in this scenario.

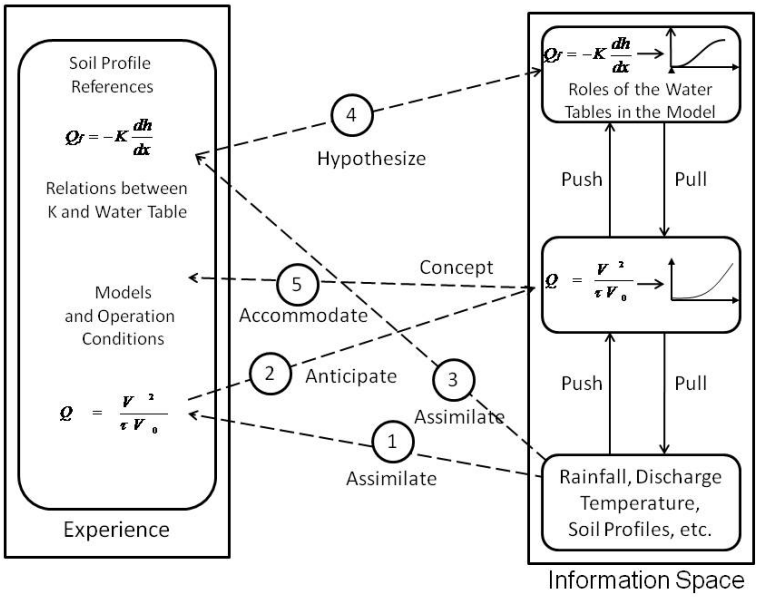

Fig. 5. A scenario of hydrologic concept formation.

The parameters that characterize the model are estimated by optimization techniques such as Maximum Likelihood (cited by Young, et al. [54]). In this way, the hydrologist applies his or her experience and domain theories to construct a proto-concept that is formulated into Equation 1 by assimilating the environmental variables (Process 1). A slope discharge prediction is generated afterward (Process 2). The model is subsequently evaluated using data captured by various sensors in a pull process[6]. Since an empirical model represents a coarse approximation of the related natural phenomenon. Contextual information including operational conditions should be attached to the model.

When complexity cannot be ignored, the modeler relies on domain knowledge to extend this model by introducing more variables in the model. For example, he or she may add new data like soil moisture index and soil profile distribu-

[6] The data used for model validation should be different to those used for model estimation.

tion to compensate for nonlinear errors. After he or she submitted the result to a community for reviewing, he or she was criticized for not using alternative models to validate the result. So the modeler resorted to alternative physical-based bottom-up models like the Darcy-Richards Equation (Equation 2) in a hypothesizing process (Process 3 and 4) to provide cross-validation and physically meaningful explanations. If there is no reasonable model suited to the observations, scientists need to give explanations and re-define model application conditions. For example, as mentioned by Baird [48], the misfit between the Darcy-Richards model and observations can be explained by the heterogeneous water behavior in different media. The initial high rate of discharge from the slope is due to the rapid flow of water in macropores in the unsaturated zone, when these empty, Darcian flow processes in the saturated zone become dominant. However, the phenomenon can also be explained by the roles of the water table. The water table rises after a rainstorm. This causes a general increase in hydraulic conductivity of the soil and as a consequence, outflow from the base of the slope increases rapidly [48]. Scientists need to observe the environment and evaluate the validity of these explanations.

In this way, a grounded model is developed from the convergence of data and domain theory.

4 Challenges for Designing Scientific Knowledge Discovery Tools

A number of specific research challenges are discussed here to foster awareness for data mining approach designers:

- There is an emerging interest in model construction in the scientific knowledge discovery and Knowledge Discovery in Databases (KDD) area [56], [57]. Identifying patterns from data, selecting appropriate models, calibrating and optimizing models have consumed considerable research effort. Scientists use various data analysis, optimization tools and rely on their experience to produce solutions. There are tools which address some aspects of these activities, for example tools for principal component analysis and algorithms for optimizations. Scientists do not yet have available comprehensive platforms aiming at addressing the knowledge discovery process from a holistic point of view. Developments in scientific work-flow processes can be used to model the scientific discovery process. These approaches apply business work-flow techniques (such as control flow modelling and large-scale collaboration) to the science domain to allow scientists to share, analyze and synthesize data [58], [59]. These approaches have not yet attempted to address the sophistication of scientific model construction. Lack of adaptivity of such tools becomes a bottleneck to progress. Many models may provide alternative results under different conditions. Without a model of what models (and algorithms) should be applied to what problems and how they should be used, an individual's effort is constrained by their expertise. The suitability

of various models and their conditional performances should be captured and reusable later. How to learn and use this type of knowledge both effectively and efficiently remains a research question for model construction. Investigating approaches for inclusion of personal discovery experience and the related impacts on science outcome may lead to experience-based scientific discovery-support tools that learn [60];

- Another research issue associated with model construction is that many models may generate quite heterogenous results, how to assist scientists to identify an appropriate model and improve its performance is still a research question. Scientists' causal reasoning processes play important roles in this issue. Foundations for explanatory hypothesis generation have been laid by some researchers, for example Simon [61], Langley et al. [62], [63] and Thagard [64], [65], [66], [67]. Data mining researchers may draw upon these ideas to develop computer-aided scientific discovery tools;

- Observation provides the means for scientists to understand previous unexplainable phenomena. People learn cause-effect relationships based on their ability in coordinating data-driven inductive and knowledge-driven deductive processes. Data-driven induction has been the major approach for KDD in recent years. Machine learning research has not gained much from human learning theories although there exists an expected synergical effect between them [68]. The functionalities of many KDD programs have not targeted assisting scientists in sophisticated reasoning and other intuitive tasks like analogy, mental imagery, hypothesis building and explanation. As we discussed in previous sections, KDD research works have been clustered around quantitative methods. Research ideas drawn from cognitive processes involved in human knowledge construction are not well known in the KDD community. In order to assist scientists' interpretation and evaluation, one needs to understand cognitive processes involved in scientific knowledge discovery. He or she also needs to determine how to build computational models to facilitate these processes?

- Large volumes of domain knowledge are available but they are not structured in a standardized and interoperable way. For example, there has been much effort in building domain-specific ontologies. Comparably fewer resources have been allocated to tackle the interoperability of multi-disciplinary (and inter-disciplinary) knowledge. How to capture and represent this domain knowledge in a cohesive way provides a research topic for knowledge engineers and KDD researchers. How to evolve knowledge structures over time and enhance their effectiveness is another challenge.

- A scientific community provides the means to judge new theories. As mentioned in previous sections, a science paradigm tends to subordinate a discovery. Creativity becomes a concern when a genuine innovation is suppressed in this environment. Developing frameworks and methodologies to improve

scientific cooperation and communication and to ensure unbiased scientific judgments may be a viable route to a scientific revolution.

In summary, scientific knowledge discovery is an interactive process in which scientists construct their first-person and then third-person knowledge. To support this dynamic process, a computer-based scientific knowledge discovery support system not only should facilitate the first-person knowledge construction but also needs to address the transition from first-person knowledge to third-person knowledge. A system could be built on a constructive view that accommodates both *a priori* (existing domain knowledge) and *a posteriori* (learning) approaches. A system should:

- have a model of an individual's experience that contains personalized domain knowledge, applications and operational knowledge of how to apply the knowledge to various problems;
- provide approaches for data-driven induction that can learn patterns and predictions from data;
- provide approaches for model-driven deductive reasoning that can assist scientists' interpretation and explanation of data patterns;
- have mechanisms to coordinate data-driven induction and expectation-driven deduction to evaluate hypotheses;
- have mechanisms to learn from each discovery activity;
- consider multiple interpretations and their impacts on the adoption of the new theory; and
- facilitate the formation of a common ground in the community.

5 Conclusion

In this chapter, scientific knowledge discovery is discussed from a constructivist view. We review cognitive theories about human knowledge construction and relate them to the scientific knowledge discovery process. A hydrologic modelling scenario has been presented to exemplify our view. We argue that there is a need to build scientific discovery support tools based on constructive principles. Challenges for designing such a tool have been identified as:

- including discovery experience and providing adaptive scientific work-flow platforms that enable scientists to construct and optimize models;
- providing means for assisting scientific explorative activities like hypothesis building and testing;
- understanding cognitive processes involved in scientific model construction and building computational models to facilitate these processes;
- producing interoperable multi-disciplinary and inter-disciplinary domain ontologies and addressing the evolution of these knowledge structures; and
- facilitating communication and cooperation in scientific communities to enhance creativity.

References

1. Langley, P.: The Computational Support of Scientific Discovery. Int. J. Human-Computer Studies 53, 393–410 (2000)
2. Muggleton, S.: Scientific Knowledge Discovery using Inductive Logic Programming. Communications of the ACM 42(11), 42–46 (1999)
3. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine 17, 37–54 (1996)
4. Sanchez, J.N., Langley, P.: An Interactive Environment for Scientific Model Construction. In: Proceedings of the Second Int. Conf. on Knowledge Capture, pp. 138–145. ACM Press, Sanibel Island, FL (2003)
5. Bridewell, W., Sanchez, J.N., Langley, P., Billman, D.: An Interactive Environment for the Modelling and Discovery of Scientific Knowledge Source. Int. J. of Human-Computer Studies 64(11), 1099–1114 (2006)
6. Luger, G.F., Johnson, P., Stern, C., Newman, J.E., Yeo, R.: Cognitive Science: The Science of Intelligent Systems. Academic Press (1994)
7. Locke, J.: An Essay Concerning Human Understanding. Collins, London (1690 reprinted in 1964)
8. Hume, D.: An Enquiry Concerning Human Understanding, e-book, `http://ebooks.adelaide.edu.au/h/hume/david/h92e/chapter7.html`
9. Hume, D.: A Treatise of Human Nature, e-book, `http://ebooks.adelaide.edu.au/h/hume/david/h92t/B1.3.6.html`
10. Kant, I.: Critique of Pure Reason. Trans. Smith, N.K., Macmillan Education Limited, London (1781 reprinted in 1956)
11. Medin, D.L., Smith, E.E.: Concepts and Concept Formation. Annual Review of Psychology 35, 113–138 (1984)
12. Van Oers, B.: Contextualisation for Abstraction. Cognitive Science Quarterly 1, 279–305 (2001)
13. Clancey, W.: Is Abstraction a Kind of Idea or How Conceptualization Works. Quarterly 1, 389–421 (2001)
14. Fisher, D.H., Pizzani, M.: Computational Models of Concept Learning. In: Fisher, D.H., Pazzani, M., Langley, P. (eds.), Concept Formation: Knowledge and Experience in Unsupervised Learning, pp. 3–43. Morgan Kaufmann, San Mateo, CA (1991)
15. Bisbey, P.R., Trajkovski, G.P.: Rethinking Concept Formation for Cognitive Agents, Towson University (2005)
16. Vygotsky, L.S.: Thought and Language. MIT Press, Cambridge, MA (1986)
17. Dewey, J.: Experience and Nature. Dover Publications, New York (1929 reprinted in 1958)
18. Von Glasersfeld, E.: An Interpretation of Piaget's Constructivism. Revue Internationale de Philosophie 36(4), 612–635 (1982)
19. Piaget, J.: The Construction of Reality in the Child. Trans. by Cook, M., Basic Books, New York (1954)
20. Piaget, J.: Biologie et connaissance. Paris: Gallimard (1967)
21. Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA (1934 reprinted in 1978)
22. Clancey, W.: A Tutorial on Situated Learning. In: Self, J. (eds.) Proceedings of the Int. Conf. on Computers and Education, pp. 49–70. AACE, Charlottesville, VA (1995)

23. Kolb, D. A.: Experiential Learning: Experience as the Source of Learning and Development. Prentice Hall, Englewood Cliffs, NJ (1984)
24. Dewey, J.: Experience and Education. Kappa Delta Pi, Indiana (1938 reprint in 1998)
25. Lewin, K.: Resolving Social Conflicts: Selected Papers on Group Dynamics. Harper & Row, New York (1948)
26. Kolb. D. A., Fry, R.: Toward an Applied Theory of Experiential Learning. In: Cooper, C. (eds.) Theories of Group Process. John Wiley, London (1975)
27. Jarvis, P.: Adult Learning in the Social Context. Croom Helm, London (1987)
28. Jarvis, P.: Towards a Comprehensive Theory of Human Learning: Lifelong Learning and the Learning Society, volume I. Routledge, London and NY (2006)
29. Jarvis, P.: Towards a philosophy of human learning: An existentialist perspective. In: Jarvis, P., Parker, S. (eds.) Human Learning: An holistic approach, pp. 1-15. Routledge, London and NY (2005)
30. Von Glasersfeld, E.: Anticipation in the Constructivist Theory of Cognition. In: Dubois, D.M. (eds.) Computing Anticipatory Systems, pp. 38–47. American Institute of Physics, Woodbury, NY (1998)
31. Hawkings, J., Blakeslee, S.: On Intelligence: How a New Understanding of the Brain will Lead to the Creation of Truly Intelligent Machines. An Owl Book, Henry Holt and Company, New York (2004)
32. Dudai, Y., Carruthers, M.: The Janus Face of Mnemosyne. Nature 434, 567 (2005)
33. Riegler, A.: Constructive Memory. Kybernetes 34(1/2), 89–104 (2005)
34. Clancey, W.: Situated Cognition: On Human Knowledge and Computer Representations. Cambridge University Press, Cambridge (1997)
35. Bartlett, F.C.: Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge (1932 reprinted in 1977)
36. Von Foerster, H.: Thoughts and Notes on Cognition. In: Von Foerster, H. (eds.) Understanding Understanding, pp. 169–190. Springer, New York (1970 reprinted in 2003)
37. Riegler, A.: Memory Ain't No Fridge: A Constructivist Interpretation of Constructive Memory. In: Kokinov, B. & Hirst, W. (eds.) Constructive Memory, pp. 277–289. NBU Series in Cognitive Science, Sofia (2003)
38. Harnad, S.: The Symbol Grounding Problem. Physica D 42, 335–346 (1990)
39. Chalmers, D.J.: Subsymbolic Computation and the Chinese Room. In: Dinsmore, J. (eds.) The Symbolic and Connectionist Paradigms: Closing the Gap, pp. 25–48. Lawrence Erlbaum, Hillsdale, NJ (1992)
40. Dorffner, G., Prem, E.: Connectionism, Symbol Grounding, and Autonomous Agents. In: Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society, pp. 144–148. Lawrence Erlbaum, Hillsdale, NJ (1993)
41. Rosenstein, M.T., Cohen, P.R.: Concepts from Time Series. In: Proceedings of Fifteenth National Conf. on Artificial Intelligence, pp. 739–745. AAAI Press, Menlo Park, CA (1998)
42. Smith, G., Gero, J.S.: The Autonomous, Rational Design Agent. In: Workshop on Situatedness in Design, Artificial Intelligence in Design'00, pp. 19–23. Worcester, MA (2000)
43. Gero, J.S., Fujii, H.: A Computational Framework for Concept Formation in a Situated Design Agent. Knowledge-Based Systems 13(6), 361–368 (2000)
44. Kuhn, T.S.: The Structure of Scientific Revolutions. The University of Chicago Press, Chicago and London (1962 reprinted in 1996)

45. Margulis, S.A., McLaughlin, D., Entekhabi, D., Dunne, S.: Land Data Assimulation and Estimation of Soil Moisture Using Measurements from the Southern Great Plains 1997 Field Experiment. Water Resource Research 38(12), 1299–1316 (2002)
46. CRC Catchment Hydrology Toolkit, `http://www.toolkit.net.au/pdfs/MC-1.pdf`
47. Michel, C.: Comment on "Can We Distinguish Richards' and Boussinesq's Equations for Hillslopes?: The Coweeta Experiment Revisited" by T. S. Steenhuis et al.. Water Resources Research 35, 3573 (1999)
48. Baird, A.: Soil and Hillslope Hydrology. In: Wainwright, J., Mulligan, M. (eds.) Environmental Modelling: Finding Simplicity in Complexity, pp. 93–106. John Wiley & Sons, Ltd, London (2004)
49. Steenhuis, T.S., Parlange, J.-Y., Sanford, W.E., Heilig, A., Stagnitti, F., Walter, M.F.: Reply to Comment on "Can We Distinguish Richards' and Boussinesq's Equations for Hillslopes?: The Coweeta Experiment Revisited" by Michel, C.. Water Resources Research 35, 3575–3576 (1999)
50. Richards, L.A.: Capillary Conduction of Liquids through Porous Mediums. Physics 1, 318–333 (1931)
51. Binley, A.M., Beven, K.J., Elgy, J.: A Physically-based Model of Heterogeneous Hillslopes II. Effective Hydraulic Conductivities. Water Resources Research 25(6), 1227–1233 (1989)
52. Bitterlich, S., Durner, W., Iden, S. C., Knabner, P.: Inverse Estimation of the Unsaturated Soil Hydraulic Properties from Column Outflow Experiments Using Free-form Parameterizations. Vadose Zone J. 3, 971–981 (2004)
53. Simunek, J., Angulo-Jaramillo, R., Schaap, M.G., Vandervaere, J.-P., Van Genuchten, M.T.: Using an Inverse Method to Estimate the Hydraulic Properties of Crusted Soils from Tension-disc Infiltrometer Data. Geoderma 86, 61–81 (1998)
54. Young, P.C., Chotai, A., Beven, K.J.: Data-based Mechanistic Modelling and the Simplification of Environmental Systems. In: Wainwright, J., Mulligan, M. (eds.) Environmental Modelling: Finding Simplicity in Complexity, pp. 371–388. John Wiley & Sons, Ltd, London (2004)
55. Young, P.C.: Data-based Mechanistic Modelling and Validation of Rainfall-flow Processes. In: Anderson, M.G., Bates, P.D. (eds.) Model Validation: Perspectives in Hydrological Science, pp. 117–161. John Wiley, Chichester (2001)
56. Langley, P.: Lessons for the Computational Discovery of Scientific Knowledge. In: Proceedings of the First Int. Workshop on Data Mining Lessons Learned, pp. 9–12. Sydney (2002)
57. Domingos, P.: Towards Knowledge-rich Data Mining. Data Min. Knowl. Disc. 15, 21–28 (2007)
58. Ellison, A.M., Osterweil, L.J., Hadley, J.L., Wise, A., Boose, E., Clarke, L.A., Foster, D.R., Hanson, A., Jensen, D., Kuzeja, P., Riseman, E., Schultz. H.: Analytic Webs Support the Synthesis of Ecological Data Sets. Ecology 87(6), 1345–1358 (2006)
59. Osterweil, L.J., Wise, A., Clarke, L., Ellison, A.M., Hadley, J.L., Boose, E., David R. Foster: Process Technology to Facilitate the Conduct of Science. In: Software Process Workshop (SPW2005). LNCS, vol. 3840, pp. 403–415. Springer-Verlag, Beijing (2005)
60. Gero, J.S.: Design Tools that Learn: A Possible CAD Future. In: Kumar, B. (eds.) Information Processing in Civil and Structural Design, pp. 17-22. Civil-Comp Press, Edinburgh (1996)
61. Kulkarni D., Simon, H.A.: The Processes of Scientific Discovery: The Strategy of Experimentation, Cognitive Science 12, 139–175 (1988)

62. Langley, P., Shiran, O., Shrager, J., Todorovski, L., Pohorille, A.: Constructing Explanatory Process Models from Biological Data and Knowledge, AI in Medicine 37, 191–201 (2006)
63. Langley, P., Bridewell, W.: Processes and Constraints in Explanatory Scientific Discovery. Proc. of the Thirtieth Annual Meeting of the Cognitive Science Society. Washingon, D.C. (2008 in press)
64. Thagard, P.: Explanatory Coherence. Behaviour and Brain Science 12, 435–467 (1989)
65. Thagard, P.: Probabilistic Networks and Explanatory Coherence. Cognitive Science Quarterly 1, 93–116 (2000)
66. Thagard, P.: Causal Inference in Legal Decision Making: Explanatory Coherence Vs. Bayesian Networks. Applied Artificial Intelligence 18(3), 231–249
67. Thagard, P., Litt, A.: Models of Scientific Explanation. In: Sun, R. (eds.) The Cambridge Handbook of Computational Psychology, pp. 549–564. Cambridge University Press, Cambridge(2008)
68. Mitchell, T.M.: The Discipline of Machine Learning. Machine Learning Department technical report CMU-ML-06-108, Carnegie Mellon University (2006).