A Dual-Process Approach for Automated Knowledge Creation

Alan Ransil, † Chhi'mèd Künzang**

* Protocol Labs, †Massachusetts Institute of Technology

Abstract: Scientific knowledge growth combines elements of existing theories into new proposed models, which is combinatorially intractable. **Inspired by dual-system psychological theories, we conceptualize a knowledge creation process in two stages. Stage One** narrows the space of existing computational elements based on contextual queues, supplying components from which a new model will be proposed. It is trained on large datasets but is computationally inexpensive at runtime. **Stage Two** permutes these elements in accordance with their explicit constraints, resulting in a set of proposed computable theories. We have developed a system that implements Stage Two. This system provides robust infrastructure for expressing constraints imposed by scientific theories, supplying a framework relating theory sub-graphs to experimental datasets stored in relational databases. We demonstrate an implementation of this two-stage approach solving materials chemistry problems using experimental datasets.

1. An overview of Dual-Process Theories

Theories across several fields within psychology have identified two distinct types of cognitive process. Type I Processes are automatic, use implicit knowledge, operate quickly, and provide intuitive results based on subconscious cognition. Type II processes are deliberative, use explicit knowledge, operate slowly, and require conscious effort. These processes frequently work synergistically, but evidence of their distinct workings appears in situations where the two offer conflicting solutions. Type I processes frequently respond to complex tasks by using simple heuristics that often contradict deliberative Type II responses. By manipulating priming effects informing these heuristics, competing decision making processes can be assessed (Lucas 2005). Evidence for this dualism has been found in the study of reasoning (Wason 1974, Evans 2006), decision making (Van Gelder 2014, Motro 2018, Dhar 2013), learning and memory (Smith 2000), social psychology (Wilson 2000, Hochman 2015, Strack 2015), and the development of metacognition (Amsel 2008).

In this work we are inspired by the parallels between a generalized Type I psychological process and many existing machine learning approaches (see Table 1). However, corresponding representations of scientific theories analogous to Type II processes are undeveloped. The apparent utility of synergistic processing between the two psychological systems suggests that development of these "Stage II" approaches may allow interoperability between implicit and explicit theories when applying machine learning approaches to scientific datasets.

Type I Psychological Processes	Proposed Stage I	Type II Psychological Processes	Proposed Stage II
U c c e a ce e	Ne a Ne S a ca Me d	C c e a ce e	E c e e e a f e e
E a d	Decade d ec e	E a Rece	Ge e a e e
Fa a a e	Fa a e	S e e a ed b g e	C b a a ed
I d c e He c	C e a a	Ded c e A a c	L g ca
I c La g age de e de	B ac b	E c La g age ba ed	H a eadab e
a ed c e fe e	T a ed g a ge da a e	M de a be ed a e e e be ed	R e a be e c e c ded
S a ed a a	Ne a Ne	D c e a	

Table 1: a comparison of dual-process theories in psychology with the proposed dual-stage automated knowledge generation system.

2. A Dual-Process Computational Approach for Science

In most scientific fields there is a **persistent computability gap** between the theories and models used in experimental design on one hand, and the datasets generated by these experiments on the other (figure 1). Explicit mental models are typically described in lab books and written documents but are rarely translated into a computable format.

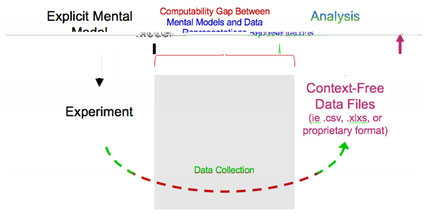

Figure 1: the computability gap in scientific datasets

Developing tools that allow researchers to easily input general mental models of experiments in a computable form has the potential to solve several important research problems by serving as a computable 'Type II' model. These problems include (1) improving the efficiency of data analysis by allowing improved search capabilities, (2) leveraging institutional knowledge across an organization and among distant collaborators, and (3) the potential for closed-loop automated knowledge generation processes that produce **cogent insights** rather than black-box correlational results.

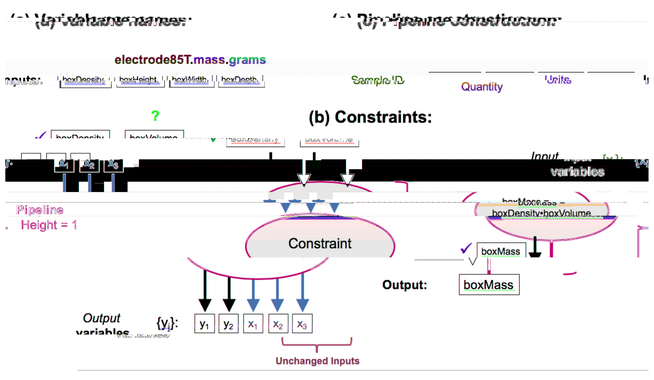

Figure 2: construction of pipelines to solve specific problems by leveraging constraints

3. Constraint-Pipeline Systems for Representing Explicit Knowledge

We have implemented a generalized system for representing scientific knowledge across domains shown in Figure 2. In this system, a **Schema** consists of a set of **Variables** representing quantities or abstract objects and **Constraints** representing relationships between these objects. **Variables** participate in a hierarchical namespace as shown in Figure 2 (a). **Constraints** relate a set of input ***Variables*** to a set of output **Variables** as shown in Figure 2 (b). These **Constraints** may correspond to simple algebraic operations or to any more abstract function. A problem may be posed to the **Schema** in the form of a set of input and output **Variables**. The function schema.plan() will then attempt to find a **Pipeline**, or a sequence of **Constraints**, that solves for the output **Variables** as a function of input **Variables** as illustrated in Figure 2 (c). **A list of successful pipelines representing this Schema's solutions to the posed problem, are exported as outputs. All of these objects can be exported and imported using JSON serialization**, in order to support efficient sharing and transfer between different solver implementations.

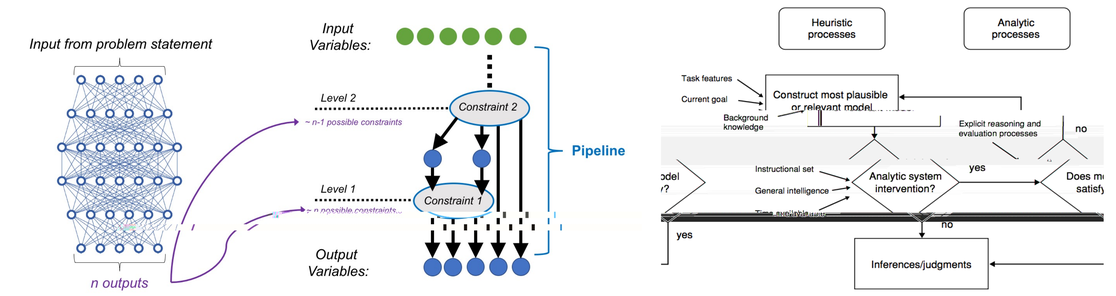

Figure 3 (above): (a-b) A comparison of the first and second stages in the proposed dual-stage process. (c) A comparison of this process to a heuristic-analytic dual process model from the literature (Evans, 2006).