Broadening the Agenda for Discovery Informatics

Pat Langley (langley@asu.edu)
Computer Science and Engineering
Arizona State University, Tempe, AZ 85287

Discovery processes are central to the scientific enterprise in that they underlie the creation and extension of knowledge. This makes them a natural target for computational study, and there has been a steady stream of research on this topic for over three decades (e.g., Bradshaw, Langley, & Simon, 1983; Džeroski & Todorovski, 2007; Shrager & Langley, 1990). However, recent years have seen a growing movement in 'data-intensive' science that adopts some questionable assumptions and ignores some key insights from earlier efforts. This essay clarifies how these biases conflict with characteristics of scientific research and argues that discovery informatics should avoid them.

First, the scientific enterprise is distinguished from other pursuits, such as the arts and law, by its emphasis on formal representations of knowledge. Different scientific fields utilize different notations – from mathematical equations in physics to reaction pathways in chemistry to qualitative causal models in biology – but they share a commitment to stating their knowledge explicitly. Moreover, scientists invariably use notations that are *communicable*, in that they can be used to convey content in journals, talks, and other media. Unfortunately, most work on scientific data mining ignores this key idea, with typical systems producing output in formats designed by computer scientists rather than those familiar and accessible to domain scientists. The alternative paradigm of *computational scientific discovery* (Džeroski & Todorovski, 2007; Shrager & Langley, 1990) develops methods that construct knowledge stated in such formalisms, making them more relevant to practicing scientists. Discovery informatics would benefit by adopting this perspective as part of its research agenda.

Second, science does more than produce new knowledge; it also uses existing knowledge to aid the discovery process. Different fields draw on background content to greater or lesser extents, depending on their stage of development. For example, early physicists, chemists, and biologists operated largely in a data-driven manner because they did not yet understand much about their area, but modern branches of these disciplines proceed in a highly knowledge-driven fashion. Unfortunately, most work on data-intensive science ignores this important feature of the scientific enterprise, emphasizing data-intensive techniques that forgo the substantial knowledge available to constrain their operation and, equally important, ensure their results are interpretable by human scientists. In contrast, much work in the paradigm of computational scientific discovery takes background knowledge into account (e.g., Bridewell, Langley, Todorovski, & Džeroski, 2008; Valdés-Pérez, 1995), demonstrating that data-driven methods are not the only option. More research in discovery informatics should incorporate this insight.

Third, although both data and models play key roles in science, neither holds center stage. The starring role is taken by the *relation* between models and data. Science differs from other professions by insisting that its knowledge be supported by observations, and that the latter be consistent witht the former. This concern drives the closed loop of the scientific process, in which models guide the collection of new data and in which these observations influence the revision of models. Unfortunately, most work on scientific data mining and data-intensive science assumes that discovery is a one-way street, emphasizing induction from observations at the cost of ignoring the equally important activity of knowledge-guided data collection and interpretation. In contrast, work in computational scientific discovery has addressed both facets of the research process, and discovery informatics would gain by adopting a similar strategy.

Fourth, science aims for more than empirical generalizations that make accurate predictions. Many scientists, especially in mature fields that have accumulated substantial knowledge, insist that observations be *explained*. These explanations move beyond descriptive summaries to account for empirical results in terms of underlying structures, processes, or mechanisms. Historically, scientists have often rejected observations and even empirical laws for which they could not devise explanations. Unfortunately, the vast majority of work on scientific data mining has focused on inducing shallow empirical descriptions that summarize data without explaining them. In contrast, the paradigm of computational scientific discovery has repeatedly dealt with the construction and revision of models that explain phenomena in familiar terms (e.g., Bridewell et al., 2008; Valdés-Pérez, 1995). The research agenda for discovery informatics should also include this important computational challenge.

Fifth, although scientific progress requires the collection and analysis of data, it does not always depend on processing large *amounts* of data. Despite advances in measuring instruments and infrastructure that have allowed acquisition of very large data repositories in fields like astronomy, the history of science shows that discovery can often be driven by small sets of observations, and this remains the norm in some disciplines. Yet this does not reduce the need for computational methods to aid discovery, as there remain challenges in finding models that accommodate such data (e.g., Lee, Buchanan, & Aronis, 1998). Unfortunately, work on data-intensive computing has emphasized the need for techniques that operate over large data sets, focusing attention on disciplines where they are available and drawing energy away from other fields. In contrast, research on computational scientific discovery has developed methods that operate over data sets of small and medium size (e.g., Bridewell et al., 2008; Valdés-Pérez, 1995), offering a more balanced perspective. The discovery informatics community should devote at least some effort to such settings.

These biases are not the only ones that characterize recent data-intensive approaches to discovery. There has also been an increasing emphasis on statistical representations of induced models, despite the fact that many scientists favor deterministic accounts. The movement has also focused on algorithms that aim to automate component tasks, rather than developing interactive systems that let scientists participate in the computational process. A viable research program in discovery informatics should overcome each of these biases by adopting a broadly based agenda that covers the entire gamut of activities found in science.

References

Bradshaw, G. L., Langley, P., & Simon, H. A. (1983). Studying scientific discovery by computer simulation. *Science*, *222*, 971–975.

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

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

Lee, Y., Buchanan, B. G., Aronis, J. M. (1998). Knowledge-based learning in exploratory science: Learning rules to predict rodent carcinogenicity. *Machine Learning*, *30*, 217–240.

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

Valdés-Pérez, R. E. (1995). Machine discovery in chemistry: New results. *Artificial Intelligence*, *74*, 191–201.