The Observational-Inductive Framework for Science

Timothy E. Eastman

Plasmas International, 1225 Edgevale Road, Silver Spring, MD 20910

Abstract. A new observational-inductive framework for science is emerging due to recent developments in sensors, data systems, computers and knowledge discovery techniques. This new framework complements the standard hypothetical-deductive model that has sometimes been held up as the standard of what is meant by "science." The hypothetical-deductive/inductive schemas were developed before the massive growth (by orders of magnitude) in the volume of observational data and power of high performance computing. The strength of the observational-inductive model is its firm foundation on both of these revolutionary developments in the history of science.

Keywords: cosmology, data-driven method, data mining, eScience, falsification, Grid systems, hypothetical-deductive, induction, knowledge discovery, KDD, methodology, observational tests, philosophy of science, theory-dependence, virtual observatory.
PACS: 01.70.+w; 01.65.+g; 89.20.Hh; 95.75.-z

INTRODUCTION

For the first time in the 400 years since Francis Bacon introduced induction, a confluence of new technologies is enabling an *observational-inductive* approach to scientific inference that is complementary to the standard hypothetical-deductive approach. This standard approach has been marvelously successful in high-energy physics and certain other fields where quantitative theories can provide well-defined, falsifiable predictions that can be directly tested in controlled, laboratory experiments. The hypothetical deductive framework was developed prior to, and has not significantly changed since the massive growth (by orders of magnitude) in the volume of observational data and power of high performance computing techniques. Further, for complex, interrelated multi-scale systems like ecology, space physics or cosmology, especially without a focus on in-principle falsifiability of hypotheses, it can lead to circumstances in which a particular research program becomes an arbiter of acceptable data [1]. Such conditions can excessively shield a preferred theoretical framework from falsification (as happened with the uniformitarian doctrine in geology prior to plate tectonics). This tendency towards theory-dependence is a key weakness of the hypothetical-deductive approach, a weakness that can be offset by the observations-driven approach of the observational-inductive framework.

FRAMEWORKS FOR KNOWLEDGE DISCOVERY

There have been many ways to represent the scientific method (hypothesis formation, experimental preparation, test, model refinement…). Most scientists agree that there is no one single method and that simplistic reference to the scientific method is insufficient; however, it remains unclear what these methods are when more fully evaluated and expressed. One thoughtful attempt to capture the essence of the scientific process led to seeing the "scientific method as information-seeking by questioning" and "problem-solving power [keeping in mind] the basic theoretical presupposition… of one's questioning procedure" [2]. An inductive logic of theories remains incomplete in philosophy of science.

Table 1 outlines the three basic frameworks of scientific method and their characteristics in terms of theory to observation level; emphasis on logical versus causal implication; and principal driver (theory vs. observation). Brief explanations and examples are provided for both the hypothetical-deductive and hypothetical-inductive frameworks. The following three sections provide more background, detail and finally four specific examples of the observational-inductive framework.

TABLE 1. Frameworks of scientific method.

	HYPOTHETICAL-DEDUCTIVE	HYPOTHETICAL-INDUCTIVE	OBSERVATIONAL-INDUCTIVE
LEVELS	top-down	interplay of levels	bottom-up
FOCUS	logical implication	causal implication	causal implication
DRIVER	theory	theory/observation balance	observations

Hypothetical-deductive framework: The standard hypothetical-deductive methodological framework for science, which focuses on logical implication, derives its strength from the consistency, coherence, and testability of deduced consequences resulting from initial hypotheses. Its first clear formulation as a methodological framework was carried out by Karl Popper in the 1930s [3]. Hypotheses in this framework are, in part, inspired by observations but may be highly dependent on prior theory as, for example, research on dark matter or dark energy. When controlled, laboratory experiments are routinely available, the hypothetical-deductive framework, with its top-down strategy and focus on logical implication, has proven to be very robust in fields such as atomic physics or high-energy physics.

For example, in 1957, two competing theories (hypotheses) of weak interactions had two very different deduced consequences – one that mirror-reflection or parity symmetry is conserved and the other for which parity is not conserved. A crucial experiment was carried out that year by C. S. Wu and collaborators demonstrating that parity symmetry was not conserved, which clearly falsified the theory requiring parity symmetry (details of this episode are provided in [4]).

As noted in Table 1, the hypothetical-deductive framework tends to be theory-driven and top-down (from creatively-inferred hypotheses to deduced consequences) with a focus on logical implication.

Hypothetical-inductive framework: Until the 1970s, early problems with the concept of induction contributed to a nearly exclusive focus on the hypothetical-deductive framework in philosophy of science circles. Recent work has recognized fundamental limitations with this standard account of scientific process and has introduced hypothetical-inductive inference in addition to hypothetical-deductive inference. In particular, *Niiniluoto and Tuomela* show how inductive and deductive inference remain as irreducible elements of the scientific process [5], and this recognition has led to new research in inductive inference [e.g., 6, 7]. The hypothetical-inductive framework adequately addresses scientific practice in many fields that lack controlled experiments but retain some balance between theory and observation.

For example, many quantitative space plasma studies employ a combination of plasma and field observations and single-particle, kinetic plasma or magnetohydrodynamic (MHD) simulations, which are applied iteratively in a theory-model-observation trialogue. Recent examples include the following: (1) Nonadiabatic acceleration of ion beams in the plasma sheet boundary layer have been demonstrated using four-point in situ Cluster spacecraft observations and single-particle model calculations [8]; (2) Successful correlations have recently been made of observed changes in Earth's ionospheric polar cap in response to solar wind input parameters by comparisons with global MHD simulations of Earth's magnetosphere [9].

As noted in Table 1, the hypothetical-inductive framework maintains a rough balance of theory and observation with a focus on causal implication.

Observational-inductive framework: See next three sections – four examples are provided in the section on the observational-inductive framework.

TAKING DATA TO KNOWLEDGE

Space science research has faced many challenges within the past few decades: high-data-rate sensors and the data explosion [10], the subtleties of plasmas and multiscale physical systems [11], and the complexities of nonlinear systems [12]. In response, new technologies have emerged that promise to meet these profound challenges: Grid systems and virtual observatories, broadband linkage of distributed data systems, and advanced visualization, among others [13, 14]. These new technologies can be represented by the Data-Sensor-HPC-Model linkages illustrated in Figure 1. Visualizing this as a tetrahedron with Data at the center (or top) emphasizes the importance of data and new data grids for meeting the data explosion; turning it over on another side places Sensor in the middle and points to new Sensor Webs being developed in Earth systems science; putting high performance

computing (HPC) in the center indicates the power of Grid computing; and placing Model in the middle can be associated with virtual modeling centers [15].

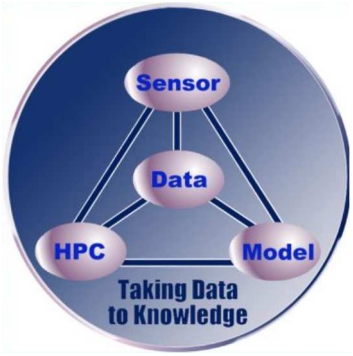

Figure 1. Data-Sensor-HPC-Model as a unifying concept for Grid systems, virtual observatories, and related developments.

In the late 20th century, breakthroughs in nonlinear dynamics emerged from the HPC-Model linkage with the advent of new supercomputer resources. Considered within the broader perspective of Data-Sensor-HPC-Model linkages, Grid systems and virtual observatories may have a similar transformative impact well beyond their initial role to expand access to data and computing resources and to enhance analysis tools across distributed databases worldwide.

KNOWLEDGE DISCOVERY IN DATABASES

In parallel with the increased synergism of Data-Sensor-HPC-Model, there have been major advances in data mining, neural networks, pattern recognition, clustering, principal component analysis, Bayesion networks, Markov models and other tools, which are here referred to collectively as Knowledge Discovery in Databases (KDD). KDD is particularly useful for the discovery of hidden relationships in large, complex databases that can exceed the limits of human pattern recognition or even model application. Knowledge discovery denotes “the nontrivial extraction of implicit, previously unknown, and potentially useful information” [16].

Data selection, automating access through registries, translation and formatting are just a few of the many data preparation steps that are essential for successful KDD applications, which can take up to 80% of a data-mining project [17]. With such preparation and with sufficiently robust data sets, however, previously hidden facts can be discovered such as specific rare events, anomaly detection, patterns, correlations, linkages, complex multi-variable interdependencies and more [18]. Emergence of the International Virtual Observatory (IVO) [19] and other new venues for data access provide important new opportunities for applying KDD tools.

OBSERVATIONAL-INDUCTIVE FRAMEWORK

The observational-inductive framework is emerging from the confluence of both KDD and Data-Sensor-HPC-Model linkages, as described above. This framework is especially needed in those fields such as geophysics and space science where direct testing of certain initial conditions or core hypotheses is difficult, if not impossible, but where gigabyte to petabyte datasets are rapidly expanding. As noted in Table 1, the observational-inductive framework is observations-driven and bottom-up (from observations to inductively-inferred hypotheses, with testing via deduced consequences) with a focus on causal implication.

Four examples of the Observational-Inductive Framework: (1) A KDD study using spatial-temporal Earth science data across multiple domains with multiple time lags has discovered correlations and unexpected event associations in human activity, the rise of atmospheric carbon dioxide, decreases in global leaf cover, and natural disasters [20]. **(2)** Another study using association mining discovered patterns in spatial-temporal data that correctly predicts El Nino events [21]. In these two KDD studies, both by the same NASA Ames research team, the correlations and associations discovered came out of applying KDD directly to the data and not from a specific test about some previously predicted effect (as in the hypothetical-deductive approach) or from parameter searches linked to known hypotheses (as for the hypothetical-inductive approach), except for some data preparation such as making the data “deseasonalized.” **(3)** Through extensive dataset preparations for diagnostics, classification, and spectra, extensive searches of large Two Micron All Sky Survey (2MASS) datasets have been carried out leading to

the discovery of T dwarf stars. Search criteria emerged iteratively from KDD analyses and only partially from model-inspired parameter ranges (as for the hypothetical-inductive approach), and not by a focus on identifying particular stars with specific characteristics predicted by theory (as for the hypothetical-deductive approach). Using searches based on spectral correlations, this data mining procedure has now yielded more than 50 of these stars, which are the coldest and most intrinsically faint brown dwarfs [22]. **(4)** An iterative data mining method substantially reduces the number of calculations needed to reach a given predictive accuracy in *ab initio* quantum mechanical calculations for inferring properties of broad classes of materials. This example utilizes the hypothetical-deductive approach with respect to particular *ab initio* quantum calculations, but focuses on applying data mining methods to order candidate structures for new alloy possibilities. Such KDD-boosted data analysis decreased the number of required calculations by a factor of four in obtaining successful crystal structure prediction for binary alloys [23]. These four cases represent nascent examples of the observational-inductive framework because they are, at least in part, observations-driven, bottom-up, and focused on causal implication (see Table 1) through the application of both KDD tools and linkages of Data-Sensor-HPC-Model.

All three frameworks of scientific methodology discussed here benefit from the best ideals of the scientific process, which include systematic examination of presuppositions, framing of testable hypotheses (falsifiable in principle), model development (preferably quantitative), and careful design of observational tests. Though KDD embodies certain assumptions with regard to data relevance, etc., these are transparent. Unlike theory-based assumptions imbedded in many research activities, KDD assumptions must be fully explicated in order to design and use KDD tools. Many KDD tools can help to reduce this theory dependence. The hypothetical-deductive/inductive and observational-inductive frameworks are complementary and synergistic; however, reduction in theory dependence through applying observational-inductive inference may be especially valuable in resolving scientific controversies in fields such as cosmology. Datasets providing for new tests of cosmological theories are becoming available, such as the Sloan Digital Sky Survey and large redshift Hubble datasets. In addition, new computer and data-intensive Grid systems are bringing these datasets to researchers worldwide [19].

ACKNOWLEDGMENTS

I am thankful for information on KDD provided by Dr. Kirk Borne of George Mason University, and acknowledge Dr. Carolyn T. Brown of the Library of Congress for invaluable discussions, and Dr. Farzad Mahootian for critical insights and references in philosophy of science. Further, I am especially gratefully for constructive reviewer criticism that led to a major, and very helpful, rewrite of this paper.

REFERENCES

1. Lakatos, Imre, *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge, UK, 1970.
2. Hintikka, Jaakko, True and false logics of scientific discovery, in *Logic of Discovery and Logic of Discourse*, edited by J. Hintikka and F. Vandamme, Plenum Press, New York, 1985.
3. Popper, Karl. *Logik der Forschung*, Springer Verlag, Vienna, 1935; *The Logic of Scientific Discovery*, Basic Books, New York, 1959.
4. Franklin, Allan. *The Neglect of Experiment*, Cambridge University Press, Cambridge, UK, 1986.
5. Niiniluoto, Ilkka, and Raimo Tuomela, *Theoretical Concepts and Hypothetico-Inductive Inference*, D. Reidel, Dordrecht, 1973.
6. Han, J., Y. Cai, and N. Cercone, Data-driven discovery of quantitative rules in relational databases, *IEEE Trans. On Knowledge and Data Engineering* **5**, No. 1, 29-40 (1993).
7. Computational Scientific Discovery. Available from http://www.isle.org/~langley/discovery.html.
8. Keiling, A., G. Parks, H. Reme, I. Dandouras, J. Bosqued, M. Wilber, et al., Bouncing ion clusters in the plasma sheet boundary layer observed by Cluster-CIS, J. Geophys. Res. **110**, A09207, doi:10.1029/2004JA010497 (2005).
9. Rastätter, L., M. Hesse, M. Kuznetsova, J. B. Sigwarth, J. Raeder, and T. I. Gombosi, Polar cap size during 14–16 July 2000 (Bastille Day) solar coronal mass ejection event: MHD modeling and satellite imager observations, *J. Geophys. Res.*, **110**, A07212, doi:10.1029/2004JA010672 (2005).
10. Lee, J., Data explosion: Challenges your disaster recovery plans, *Disaster Recover Journal* **17**, Issue 4 (2004).
11. Goedbloed, J., and S. Poedts, *Principles of Magnetohydrodynamics: With Applications to Laboratory and Astrophysical Plasmas*, Cambridge University Press, Cambridge, UK, 2004.
12. Strogatz, Steven H., *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*, Addison-Wesley Pub., Reading, Mass., 1994.
13. *Grid Computing: Making the Global Infrastructure a Reality*, edited by Fran Berman, G. Fox, and T. Hey, John Wiley & Sons, Ltd, Chichester, UK, 2003.

14. *The Grid: Blueprint for a New Computing Infrastructure*, edited by Ian Foster and Carl Kesselman, *2nd ed.*, Elsevier, Amsterdam, 2004.
15. Eastman, T., K. Borne, J. Green, E. Grayzeck, R. McGuire, and D. Sawyer, eScience and Archiving for Space Science, *Data Science Journal* **4**, 67-76 (September1, 2005) Available from http://www.datasciencejournal.org.
16. Frawley, William J., G. Piatetsky-Shapiro, and C. Matheus, Knowledge discovery in databases: An overview, in *Knowledge Discovery in Databases*, ed. G. Piatetsky-Shapiro and W. Frawley, Menlo Park, CA: AAAI Press, 1991.
17. Pyle, Dorian, *Data Preparation for Data Mining*, Morgan Kaufmann Publishers, San Francisco, 1999.
18. Borne, K. D., Distributed data mining in the National Virtual Observatory, in *Data Mining and Knowledge Discovery: Theory, Tools, and Technology V*, edited by Belur Dasarathy, *Proc. SPIE* **5098**, 211-218 (2003).
19. International Virtual Observatory. Available from http://www.ivoa.net.
20. Potter, Christopher, P.-N. Tan, M. Steinbach, S. Klooster, V. Kumar, R. Myneni, and V. Genovese, Major disturbance events in terrestrial ecosystems detected using global satellite data sets, *Global Change Biology* **9**, 1005-1021 (2003).
21. Tan, P.-N., C. Potter, M. Steinbach, S. Klooster, V. Kumar, and A. Torregrosa, Spatio-temporal patterns in Earth science data, in *KDD Workshop on Temporal Data Mining*, August, 2001.
22. Burgasser, Adam, M. McElwain, J. Kirkpatrick, K. Cruz, C. Tinney, and I. Reid, The 2MASS wide-field T dwarf search. III. Seven new T dwarfs and other cool dwarf discoveries, *AJ* **127**, 2856-2870 (2004).
23. Morgan, Dane, Gerbrand Ceder, and Stefano Curtarolo, High-throughput and data mining *ab initio* methods, *Meas. Sci. Technol.* **16**, 296-301 (2005).