Integrated Systems for Computational Scientific Discovery

Pat Langley

Institute for the Study of
Learning and Expertise
Palo Alto, California

Thirty-Eighth AAAI Conference on Artificial Intelligence
Vancouver, BC, February 20–27, 2024

Three Main Points

This Senior Member Track presentation has three distinct but complementary messages:

- The discovery of new scientific knowledge is a complicated and *multi-faceted* activity.
- Since the 1970s, AI researchers have been automating *individual facets* of discovery in many disciplines.
- A remaining challenge for the field is to combine these elements into ***integrated discovery systems***.

For this reason, it combines some features of a survey talk with aspects of a blue sky talk.

A Brief History of Chemistry

One of the earliest scientific disciplines, chemistry, illustrates different facets of discovery:

- *Taxonomic classification of many substances (~800 and earlier)*
- *Many qualitative chemical reactions (~800 to ~1300 and later)*
- *Laws of definite proportions (1797), combining volumes (1809)*
- *Phlogiston (1731) and oxygen (1774) models of combustion*
- *Structures of molecules, inorganic (1808) and organic (~1860)*
- *Biochemical processes of cellular metabolism (early 1900s)*

Later stages built on earlier ones, progressing from descriptive summaries to deeper understanding.

The Task of Scientific Discovery

We can specify the generic problem of scientific discovery in terms of inputs and outputs:

- Given: *Scientific data or phenomena to be described or explained*
- Given: *Knowledge and heuristics about the scientific domain*
- Given: *A space of candidate laws, hypotheses, or models*
- Find: *Laws or models that describe or explain the observations*

The results should not only generalize well; they should also be stated in an ***established scientific formalism***.

Thus, we can formulate discovery as ***heuristic search*** through a space of interpretable candidates.

Scientific Discovery vs. Data Mining

Computational scientific discovery has some similarities to data mining, but they are not the same:

Data Mining	Scientific Discovery
Use computational methods	Use computational methods
Search space of laws / models	Search space of laws / models
Commercial applications	***Scientific disciplines***
Large to giant data sets	***Small to moderate data sets***
Computer science notations	***Scientific formalisms***

Data-mining methods can be applied to scientific ***data***, but they seldom produce scientific ***knowledge***.

Five Types of Scientific Discovery

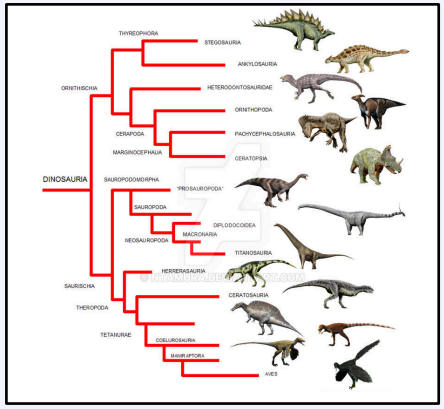
We can partition scientific discovery into five broad classes of component activities:

- *Forming taxonomic hierarchies*
- *Finding qualitative laws*
- *Inducing numeric laws / equations*
- *Formulating structural models*
- *Generating process models*

We must understand scientific discovery's facets before talking about how we might combine them.

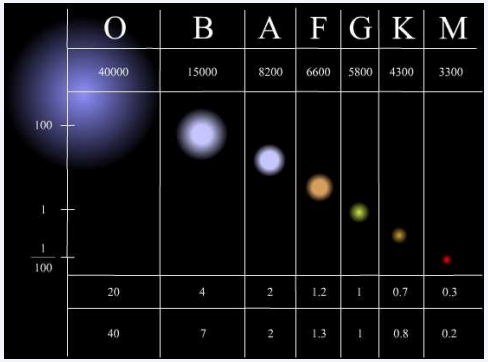
Forming Taxonomies

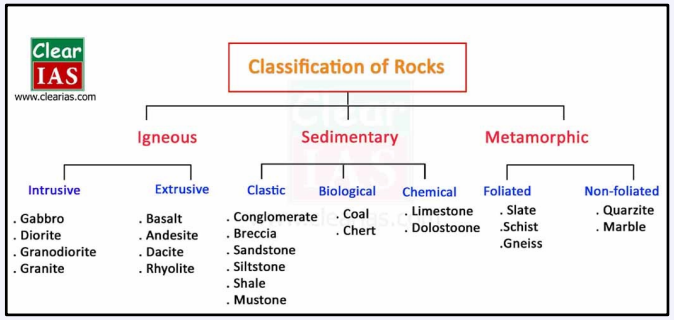
Given a set of observed entities, find a taxonomy that organizes them into classes with associated descriptions.



Dinosaurs

Stars





Rocks

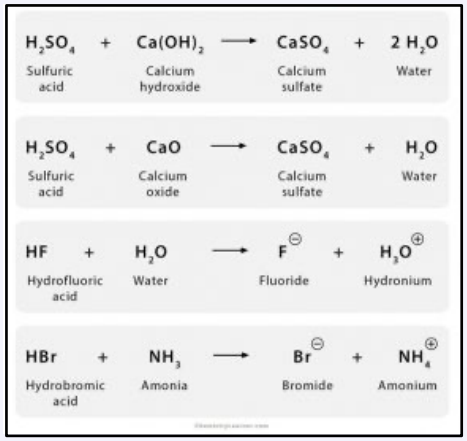
Examples of computational systems:

- Numerical taxonomy (Sokal & Sneath, 1963) – Biology
- AutoClass (Cheeseman et al, 1988) – Astronomy
- Computational phylogenetics (Warnow, 2018) – Biology

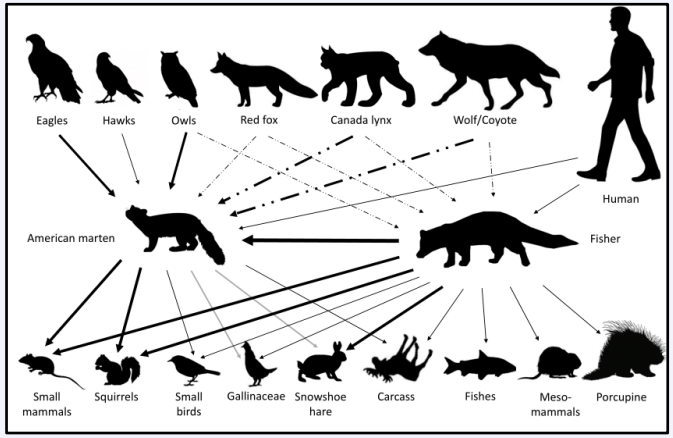
Finding Qualitative Laws

Given observed entities, their features, and relations, find a set of qualitative laws that describe them.

Chemical Reactions



Predator-Prey Relations

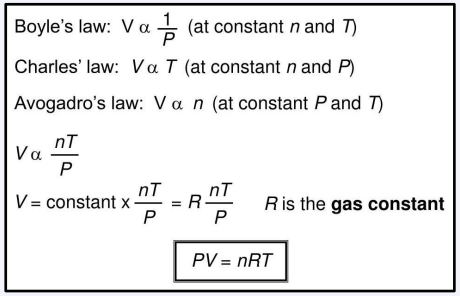


Examples of computational systems:

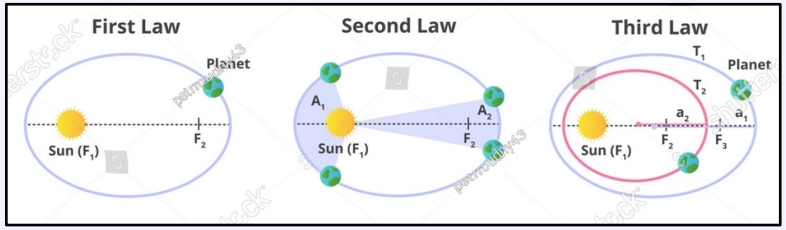
- Glauber (Langley et al., 1987) – Reactions of acids and alkalis
- RL (Lee et al., 1998) – Respiratory syndromes, carcinogens
- PROGOL (King et al., 1996) – Mutagenic chemical structures

Inducing Numeric Laws

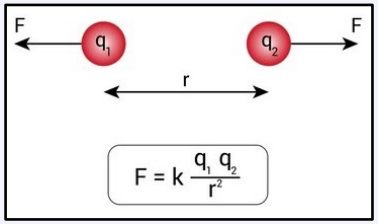
Given a set of observed entities with numeric descriptors, find one or more equations that describe these observations.



Gas Laws



Kepler's Laws



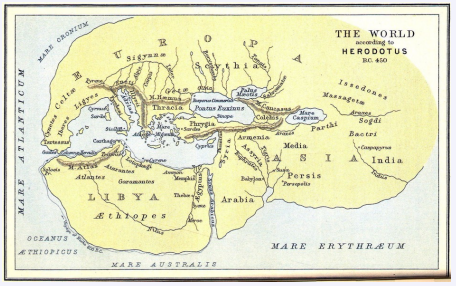
Coulomb's Law

Examples of computational systems:

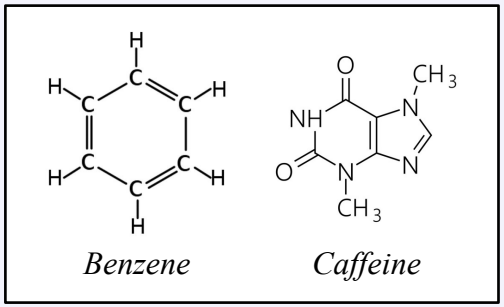
- Bacon (Langley et al., 1980, 1983) – Laws of physics and chemistry
- Fahrenheit (Zytkow et al., 1990) – Laws of electrochemistry
- LaGrange (Dzeroski & Todorovski, 1994) – Ecological dynamics
- Others – Eureqa (Schmidt & Lipson, 2009), SINDy (Brunton, 2016)

Formulating Structural Models

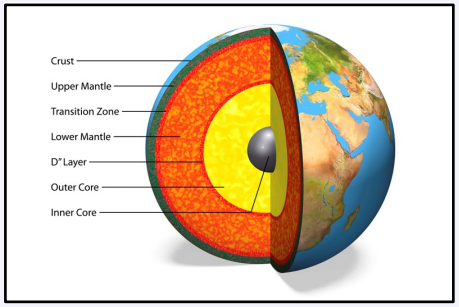
Given a set of observed entities with descriptors, find structural models with inferred components that explain them.



Cartographic Maps



Chemical Structures



Planetary Layers

Examples of computational systems:

- Dalton (Langley et al, 1987) – Inorganic chemical structures
- Gell-Mann (Zytkow & Fischer, 1990) – Elementary particles
- DENDRAL (Lindsay et al., 1980) – Organic chemical structures
- AlphaFold (Jumper et al., 2021) – Protein structures

Generating Process Models

Given entities described at different points in time, postulate a set of interacting processes that explain their behavior.

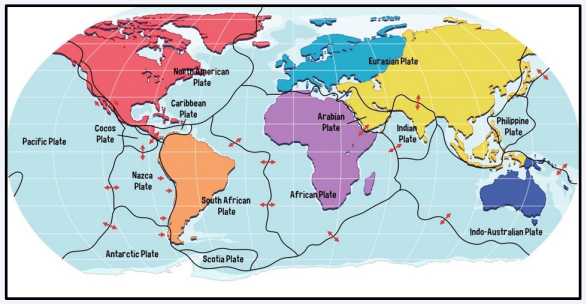
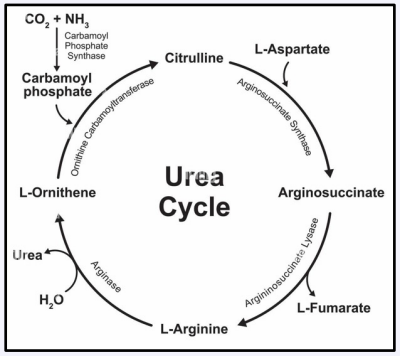
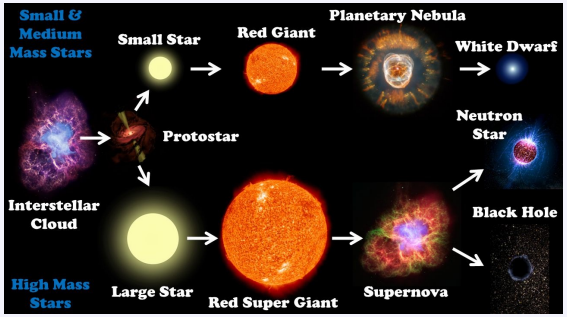


Plate Tectonics



Metabolic Pathways



Stellar Transitions

Examples of computational systems:

- MECHEM (Valdes-Perez, 1994) – Chemical reaction pathways
- ACE (Anderson et al., 2014) – Creation of geological landforms
- ALP (Bohan et al., 2011) – Invertebrate predation networks
- LaGramge (Atanasova et al., 2008) – Aquatic ecosystem models

Challenges for Integrated Discovery

Despite steady progress on these elements in isolation, we need research on integrated systems that:

- *Generate scientific context*
- *Revise laws and models*
- *Combine experimentation with discovery*
- *Identify and measure variables*
- *Interact with human scientists*

There have been some efforts on each topic, but each deserves far more attention than it has received.

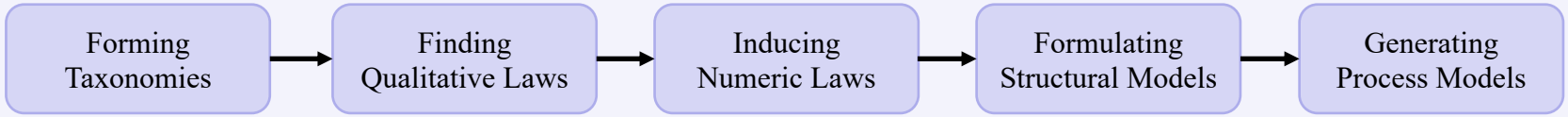
Challenge: Generating Scientific Context

Scientific discovery always occurs in some ***context*** that takes on a diverse set of forms:

- E.g., methods for law induction assume an existing taxonomy.
- E.g., methods for process modeling build on law-like elements.

Isolated systems depend on *humans* to provide this context, but integrated ones must generate their own.

Response: *Cumulative systems can use the output from some modules as input to others.*



A basic design would be a simple pipeline architecture, although feedback loops can be important.

Challenge: Revising Laws and Models

New observations become available over time, which makes science an ***on-line*** activity and means that:

- Batch processing will not suffice for extended operation
- The discovery process must support *revision* of law and models

This poses a challenge to cumulative approaches, as the context for previous discoveries can change.

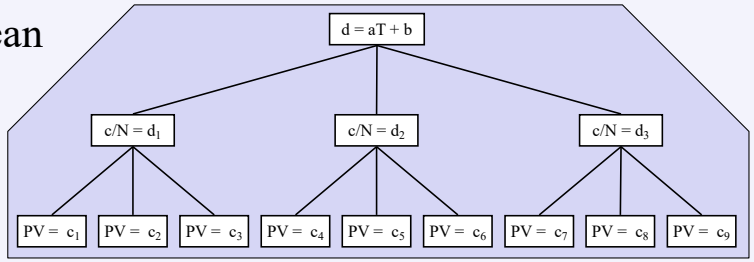
Response: *An integrated system can record these dependencies, identify where revisions are needed, and make local updates.*

This is akin to classical techniques for *truth maintenance* that support belief revision (e.g., Doyle, 1979).

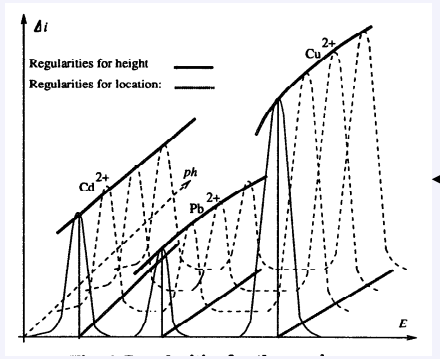
Challenge: Closed-Loop Discovery

A few AI systems have merged experiment design with discovery.

Knowledge-Lean Experiments



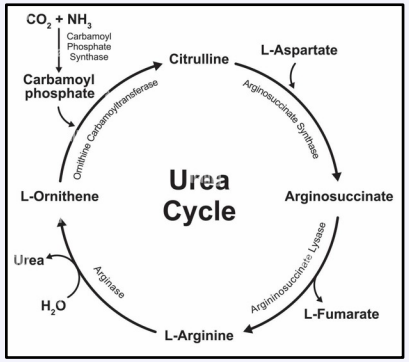
Bacon (Langley, 1981)



From AAAI 1990

Fahrenheit (Zytkow et al., 1990)

Knowledge-Rich Experiments



Kekada (Kulkarni & Simon, 1988)

Adam (King et al., 2009)

Robot Scientists

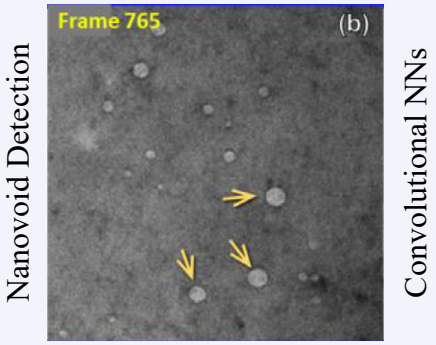
"Self-driving labs" are popular, but few find interpretable models.

Challenge: Measuring / Identifying Variables

Discovery systems can also measure and identify new variables.

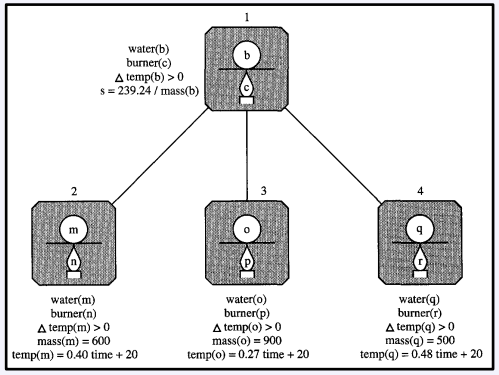
Construct *virtual measuring devices* for user-given terms

SKICAT (Fayyad et al., 1993)

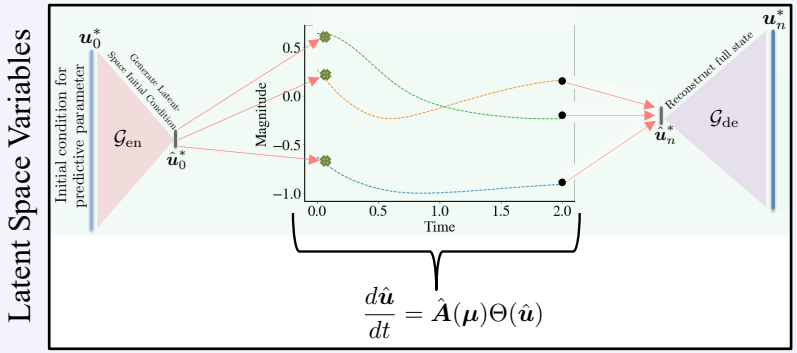


Nasim et al. (1993)

Introduce new variables during equation discovery



IDS (Nordhausen & Langley, 1993)



LASDI + SINDy (Fries et al., 2022)

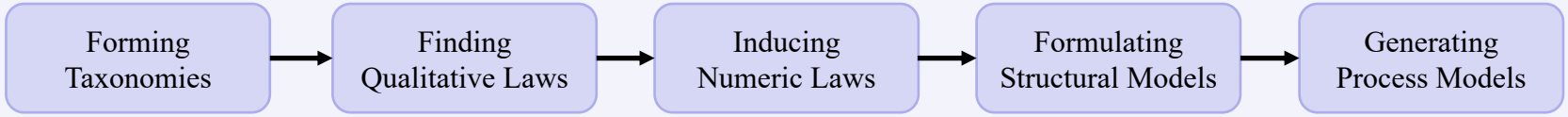
The second variety is rare but is a form of integrated discovery.

Challenge: Interacting with Human Scientists

Autonomous discovery is not the only target; AI systems should also interact and collaborate with human scientists.

- There have been some examples in discovering taxonomies, causal models, and process accounts, but we need more.

The first step is a *cognitive task analysis* that identifies elements of scientific discovery, which we have done.



Response: *Choose which discovery elements / subelements to automate and which to reserve for humans.*

We can base decisions on factors like subtask difficulty, effort involved, and human preferences.

Challenge: Evaluating Integrated Discovery

We must also identify testbeds and criteria to evaluate integrated discovery systems, including:

- Synthetic environments that obey known principles:
  - E.g., AI2's *ScienceWorld* (Wang et al., 2022)
  - Includes laws of chemistry, electricity, thermodynamics
  - Can compare discovered knowledge to known targets
- Natural domains that support integrated discovery:
  - E.g., astronomy, materials science, gut microflora
  - Must measure predictive accuracy without known targets
  - But humans can also rate understandability, plausibility

Any viable testbed should involve all facets of discovery and provide ready sources of data.

Summary Remarks

This talk made three points about the computational discovery of scientific knowledge:

- Discovery has many facets – forming taxonomies, inducing descriptive laws, finding explanatory models.
- The past 50 years have seen major progress on automating each of these scientific tasks.
- Integrating these abilities, and combining them with others, remains a key challenge for the field.

We need more research in the spirit of early AI, which pursued audacious visions like integrated discovery systems.

References on Scientific Discovery

Addis, M., Lane, P. C. R., Sozou, P. D., & Gobet, F. (Eds.). (2019). *Scientific discovery in the social sciences*. Cham, Switzerland: Springer.

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

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. San Diego, CA: Academic Press.

Langley, P. (2000). The computational support of scientific discovery. *International Journal of Human-Computer Studies*, *53*, 393–410.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Lindsay, R., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1980). *Applications of artificial intelligence for organic chemistry: The Dendral Project*. New York: McGraw.

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

Todorovski, L. (2011). Equation discovery. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning*. Boston, MA: Springer.

Valdés-Pérez, R. E. (1996). Computer science research on scientific discovery. *Knowledge Engineering Review*, *11*, 57–66.