Themes and Progress in Computational Scientific Discovery

Pat Langley

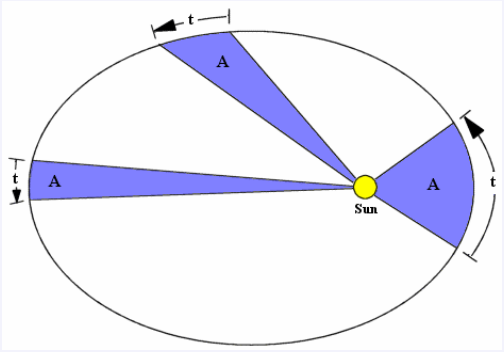
Department of Computer Science
University of Auckland

Silicon Valley Campus
Carnegie Mellon University

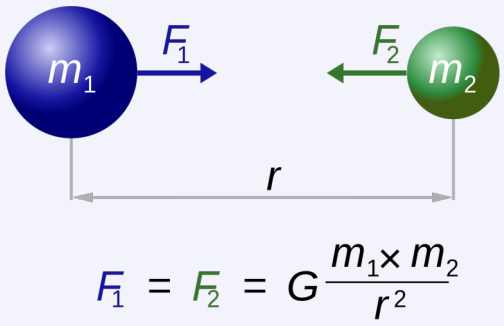
Thanks to G. Bradshaw, W. Bridewell, S. Dzeroski, H. A. Simon, L. Todorovski, R. Valdes-Perez, and J. Zytkow for discussions that led to many of these ideas, which was partly funded by ONR Grant No. N00014-11-1-0107.

Examples of Scientific Discoveries

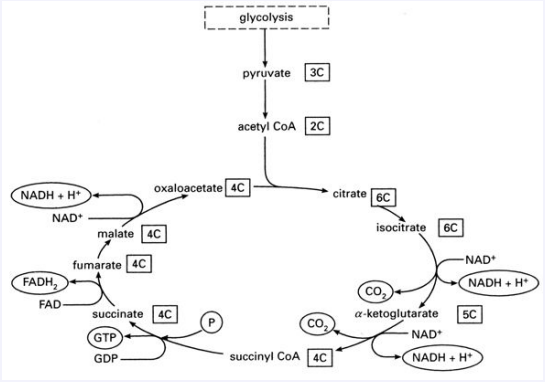
Science is a distinguished by its reliance on formal laws, models, and theories of observed phenomena.



Kepler's laws of planetary motion

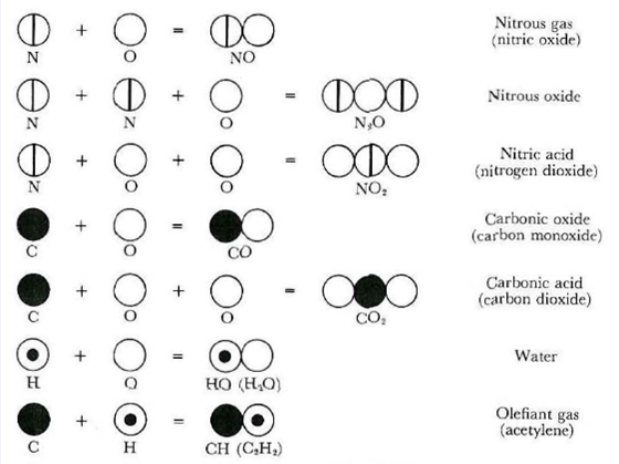


Newton's theory of gravitation



Krebs' citric acid cycle

We often refer to the process of finding such accounts as *scientific discovery*.



Dalton's atomic theory

Philosophy of Science

The *philosophy of science* has studied science since the 19th Century, focusing on the:

- character of scientific observations and experiments;
- structure of scientific theories, laws, and models;
- nature of scientific explanations and predictions;
- evaluation of scientific theories, models, and laws.

However, philosophers of science typically avoided the topic of scientific discovery.

Mystical Views of Scientific Discovery

Many have claimed that scientific discovery cannot be analyzed in rational terms. Popper (1934) wrote:

> *The initial stage, the act of conceiving or inventing a theory, seems to me neither to call for logical analysis nor to be susceptible of it … My view may be expressed by saying that every discovery contains an 'irrational element', or 'a creative intuition'…*

He was not alone. Hempel and many others believed discovery was inherently irrational and beyond understanding.

Scientific Discovery as Problem Solving

Simon (1966) offered another view – that scientific discovery is a variety of *problem solving* that involves:

- *Search* through a space of connected *problem states;*
- Generated from earlier states by mental *operators;*
- Guided by *heuristics* that keep the search tractable.

His ideas provided a powerful new approach to understanding the nature of scientific discovery.

Moreover, it offered ways to *automate* this mysterious process.

The Task of Scientific Discovery

We can state the discovery task in terms of the inputs provided and the outputs produced:

- *Given:* A set of scientific data or phenomena to be modeled;
- *Given:* A space of candidate laws, hypotheses, or models stated in an *established scientific formalism*;
- *Given:* Knowledge and heuristics for the scientific domain;
- *Find:* Laws or models that describe or explain the data or phenomena (and that generalize well).

We can develop AI systems that carry out search through this space of candidate accounts.

Some Laws Discovered by Bacon
(Langley et al., 1983)

Basic algebraic relations:

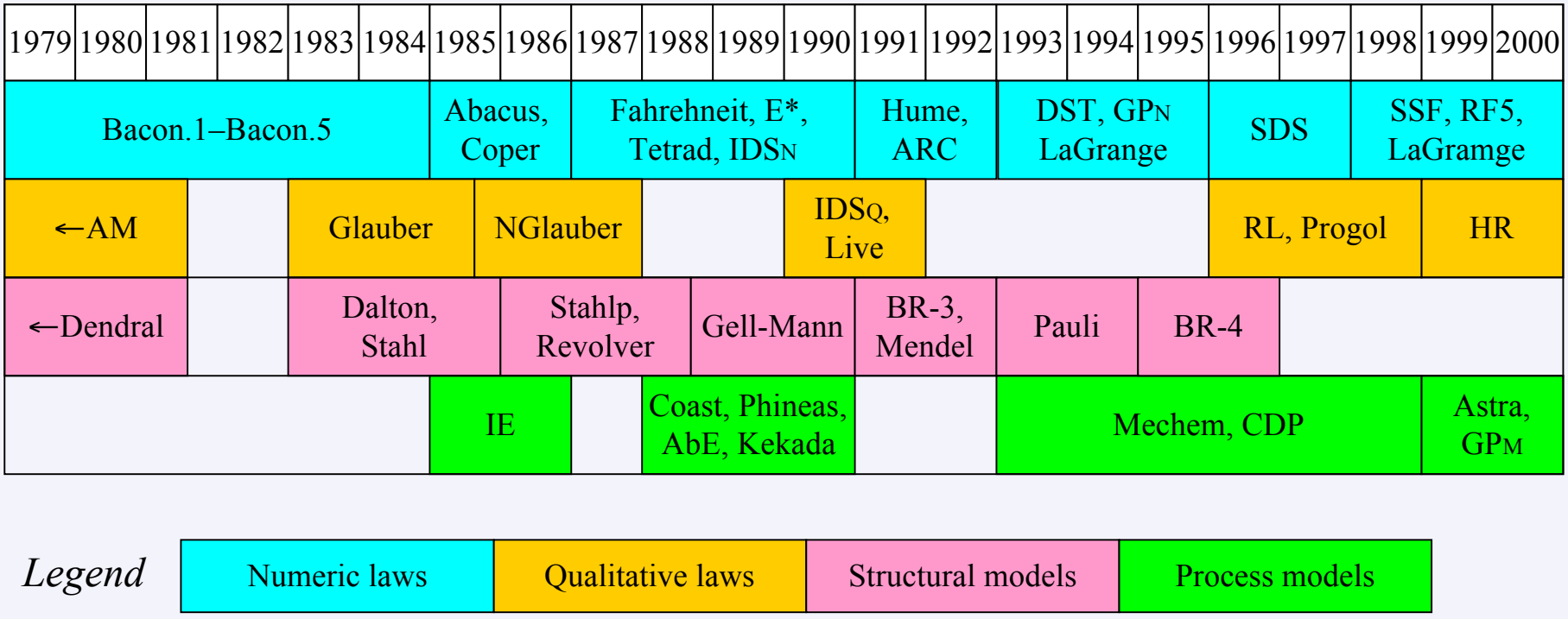
- Ideal gas law $PV = aNT + bN$
- Kepler's third law $D^3 = [(A - k) / t]^2 = j$
- Coulomb's law $FD^2 / Q_1Q_2 = c$
- Ohm's law $TD^2 / (LI - rI) = r$

Relations with *intrinsic properties*:

- Snell's law of refraction $\sin I / \sin R = n_1 / n_2$
- Archimedes' law $C = V + i$
- Momentum conservation $m_1V_1 = m_2V_2$
- Black's specific heat law $c_1m_1T_1 + c_2m_2T_2 = (c_1m_1 + c_2m_2)T_f$

Early Progress in Scientific Discovery

Research on computational scientific discovery covers many forms of laws and models.



Most early work focused on historical examples, but more recent efforts have aided the scientific enterprise.

Successes of Computational Scientific Discovery

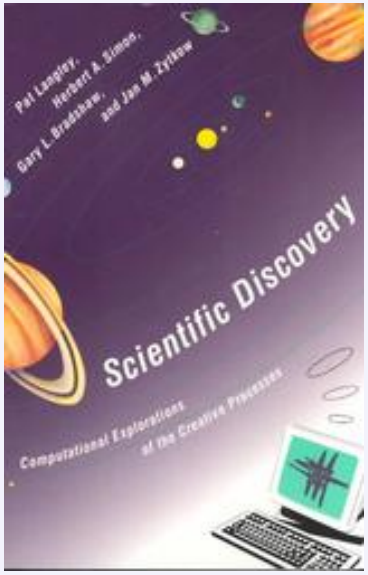
AI systems of this type have helped to discover new knowledge in many scientific fields:

- Qualitative chemical factors in mutagenesis (King et al., 1996)
- Quantitative laws of metallic behavior (Sleeman et al., 1997)
- Quantitative conjectures in graph theory (Fajtlowicz et al., 1988)
- Qualitative conjectures in number theory (Colton et al., 2000)
- Temporal laws of ecological behavior (Todorovski et al., 2000)
- Reaction pathways in catalytic chemistry (Valdes-Perez, 1994, 1997)

Each of these has led to publications in the *refereed literature of the relevant scientific field* (Langley, 2000).

Books on Scientific Discovery

Research on computational scientific discovery has produced a number of books on the topic.



1987

1990

2007

These further demonstrate the diversity of problems and methods while emphasizing their underlying unity.

The Data Mining Movement

During the 1990s, a new paradigm known as *data mining and knowledge discovery* emerged that:

- Emphasized the availability of large amounts of data;
- Used computational methods to find regularities in the data;
- Adopted heuristic search through a space of hypotheses;
- Initially focused on commercial applications and data sets.

Most work used notations invented by computer scientists, unlike work on scientific discovery, which used *scientific formalisms*.

Data mining has been applied to scientific data, but the results seldom bear a resemblance to scientific *knowledge*.

Discovering Explanatory Models

The early stages of any science focus on *descriptive laws* that *summarize* empirical regularities.

Mature sciences instead emphasize the creation of *models* that *explain* phenomena in terms of:

- Inferred *components* and *structures* of entities;
- Hypothesized *processes* about entities' interactions.

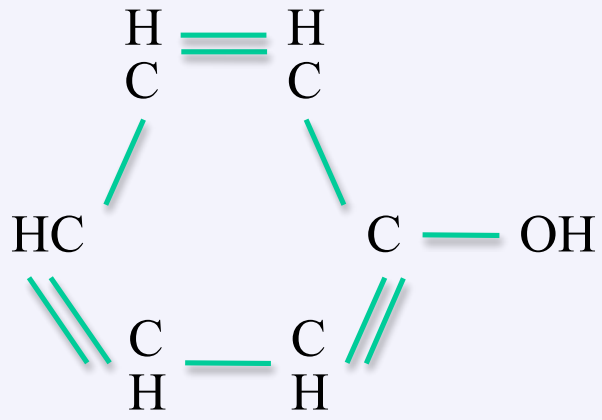
Explanatory models move beyond description to provide deeper accounts linked to theoretical constructs.

Can we develop computational systems that address this more sophisticated side of scientific discovery?

Classic Work: DENDRAL
(Lindsay et al.,1980)

The DENDRAL system inferred a molecule's chemical bonds given its component formula and a mass spectrogram.

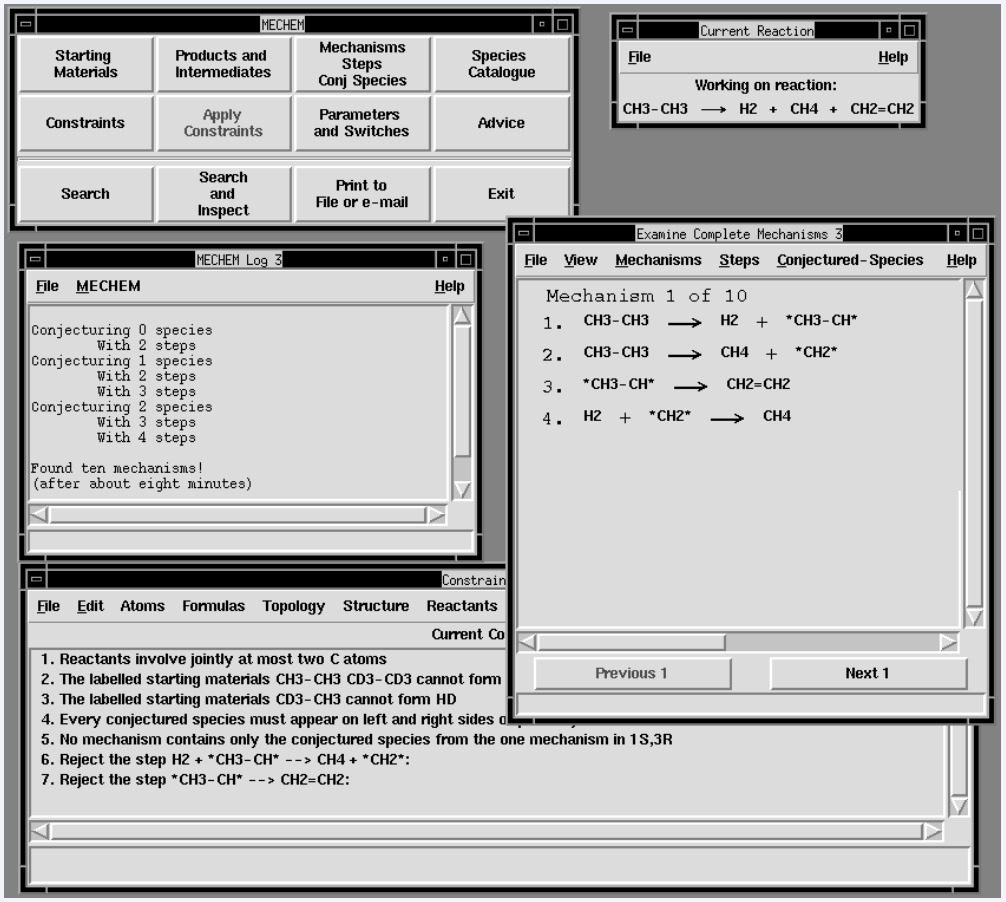
E.g., from the formula C_6H_5OH and other relevant information, the program produced structures like:



DENDRAL relied on heuristic search to infer structural models, using knowledge from 20th Century chemistry as a guide.

Classic Work: MECHEM
(Valdes-Perez, 1994)

MECHEM was a graphical interactive system that generated plausible pathways to explain chemical reactions.



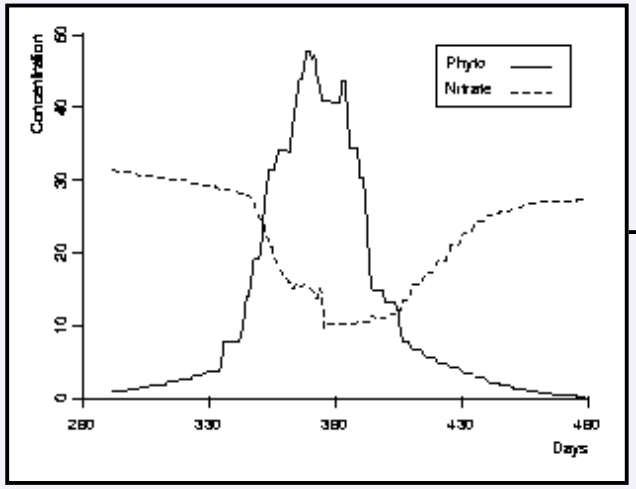
The screenshot shows, clock-wise from the upper left:

- the main menu;
- the current reaction;
- a sample mechanism;
- a set of constraints; and
- the system's output log.

These made MECHEM more accessible to chemists.

Inductive Process Modeling
(Bridewell et al., 2008)

observations



entities

phyto, nitro, zoo,
nutrient_nitro, nutrient_phyto

process model

model AquaticEcosystem

variables: nitro, phyto, zoo, nutrient_nitro, nutrient_phyto
observables: nitro, phyto, zoo

process phyto_exponential_growth
  equations: d[phyto,t] = 0.1 × phyto

process zoo_logistic_growth
  equations: d[zoo,t] = 0.1 × zoo / (1 – zoo / 1.5)

process phyto_nitro_consumption
  equations: d[nitro,t] = –1 × phyto × nutrient_nitro,
  d[phyto,t] = 1 × phyto × nutrient_nitro

process phyto_nitro_no_saturation
  equations: nutrient_nitro = nitro

process zoo_phyto_consumption
  equations: d[phyto,t] = –1 × zoo × nutrient_phyto,
  d[zoo,t] = 1 × zoo × nutrient_phyto

process zoo_phyto_saturation
  equations: nutrient_phyto = phyto / (phyto + 0.5)

Heuristic Search

process exponential_growth
  variables: P {population}
  equations: d[P,t] = [0, 1,∞] × P

process logistic_growth
  variables: P {population}
  equations: d[P,t] = [0, 1, ∞] × P × (1 – P / [0, 1, ∞])

process constant_inflow
  variables: I {inorganic_nutrient}
  equations: d[I,t] = [0, 1, ∞]

process consumption
  variables: P1 {population}, P2 {population},
nutrient_P2
  equations: d[P1,t] = [0, 1, ∞] × P1 × nutrient_P2,
  d[P2,t] = – [0, 1, ∞] × P1 × nutrient_P2

process no_saturation
  variables: P {number}, nutrient_P {number}
  equations: nutrient_P = P

process saturation
  variables: P {number}, nutrient_P {number}
  equations: nutrient_P = P / (P + [0, 1, ∞])

generic processes

constraints

Always-together[growth(P), loss(P)]
Exactly-one[lotka-volterra(P, G), ivlev(P, G), watts(P, G)]
At-most-one[photoinhibition(P, E)]
Necessary[nutrient-mixing(N), remineralization(N, D)]

Recent Progress: Biological Models

King et al. (2009) have constructed an integrated system for biological discovery that:

- Designs auxotrophic growth studies with yeast gene knockouts;
- Runs these experiments using a robotic manipulator;
- Measures the growth rates for each experimental condition; and
- Revises its causal model for how genes influence metabolism.

This closes the loop between experiment design, data collection, and model construction in biology.

But note that Zytkow et al. (AAAI-90) reported an even earlier robot scientist in the field of electrochemistry.

Recent Progress: Cosmogenic Dating

Anderson et al. (2014) report ACE, an AI system for cosmogenic dating that:

- Designs inputs nucleotide densities for rocks from a landform;
- Generates process accounts for how the landform was produced;
- Weighs arguments for and against each process explanation.

ACE has been downloaded ~600 times and is used actively by many geologists to understand their data.

The system is user-extensible and, years after it launch, has led to zero requests for help from computer scientists.

Big Data and Scientific Discovery

Digital collection and storage have led to rapid growth of data in many areas.

The *big data* movement seeks to capitalize on this content, but, in science at least, we must address *three* distinct issues:

- Scaling to large and heterogeneous *data sets*;
- Scaling to large and complex *scientific models;*
- Scaling to large *spaces of candidate models*.

We need far more work on the last two issues, for which methods from computational scientific discovery are well suited.

Summary Remarks

There has been a long history of work on computational scientific discovery, including methods for constructing:

- Descriptive laws stated as numeric equations
- Explanatory models of structures and processes

Recent research has focused on the latter, which is associated with more mature fields of science.

Work in this paradigm discovers knowledge stated in formalisms and concepts that are *familiar to scientists*.

Challenges involve dealing not with ‘big data’, but with *complex models* and *large search spaces*.

Publications on Computational Scientific Discovery

Bridewell, W., & Langley, P. (2010). Two kinds of knowledge in scientific discovery. *Topics in Cognitive Science*, *2*, 36–52.

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

Bridewell, W., Sanchez, J. N., Langley, P., & Billman, D. (2006). An interactive environment for the modeling and discovery of scientific knowledge. *International Journal of Human-Computer Studies*, *64*, 1099-1114.

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

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

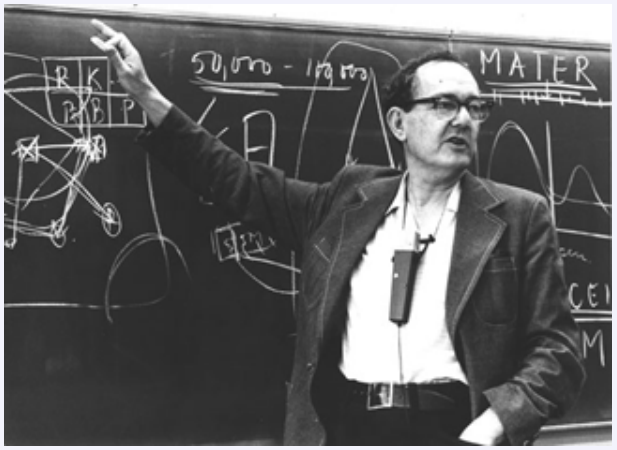
Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Langley, P., & Zytkow, J. M. (1989). Data-driven approaches to empirical discovery. *Artificial Intelligence*, *40*, 283–312.

Todorovski, L., Bridewell, W., & Langley, P. (2012). Discovering constraints for inductive process modeling. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. Toronto: AAAI Press.

In Memoriam

In 2001, the field of computational scientific discovery lost two of its founding fathers.



Herbert A. Simon
(1916 – 2001)

Jan M. Zytkow
(1945 – 2001)

Both were interdisciplinary researchers who published in computer science, psychology, philosophy, and statistics.

Herb Simon and Jan Zytkow were excellent role models for us all.

End of Presentation