Themes and Progress in Computational Scientific Discovery

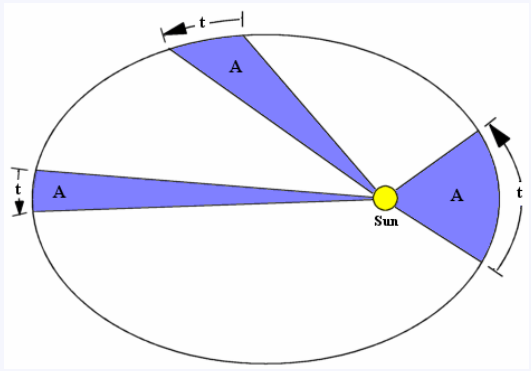
Pat Langley

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

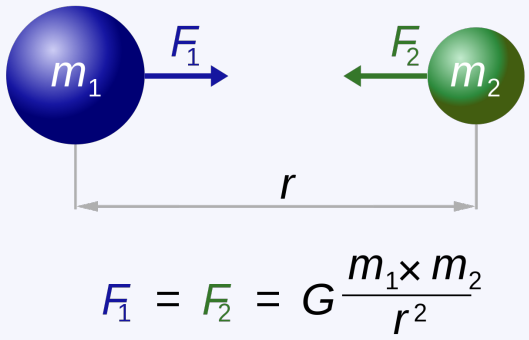
Thanks to G. Bradshaw, W. Bridewell, S. Dzeroski, H. A. Simon, L. Todorovski, R. Valdes-Perez, and J. Zytkow for discussions that led to many of these ideas, and to research funding through ONR Grant No. N00014-11-1-0107.

Examples of Scientific Discoveries

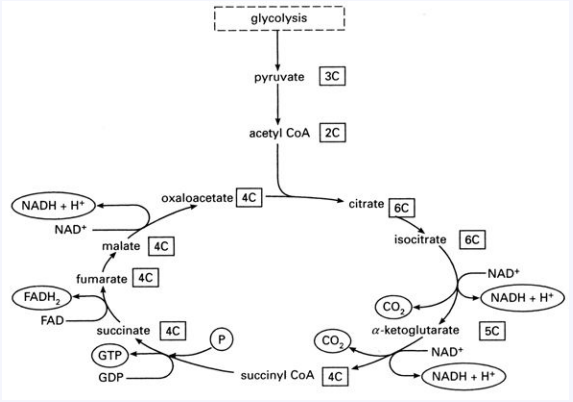
Science is a distinguished by its reliance on formal laws, models, and theories of observed phenomena.



Kepler's laws of planetary motion

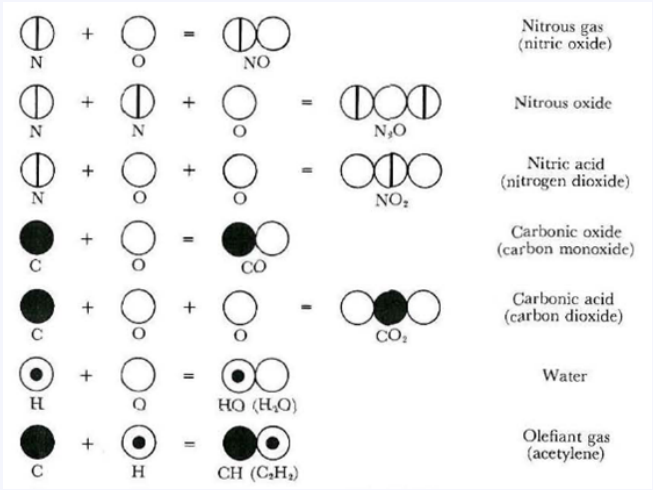


Newton's theory of gravitation



Krebs' citric acid cycle

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



Dalton's atomic theory

Mystical Views of Scientific Discovery

Philosophers of science claimed that scientific discovery could not 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.

However, advances made in two fields – *cognitive psychology* and *artificial intelligence* – in the 1950s suggested otherwise.

Scientific Discovery as Problem Solving

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

- *Search* through a space of *problem states*
- Generated by applying mental *operators*
- Guided by *heuristics* to make it tractable

Heuristic search had been implicated in many cases of human cognition, from proving theorems to playing chess.

This framework offered not only a path to understand scientific discovery, but also ways to *automate* this mysterious process.

Early Progress

One of the first systems that adapted Simon's ideas on discovery was *Bacon* (Langley, 1981), a computer program that:

- Input numeric observations for a number of variables;
- Carried out search in a problem space of theoretical terms;
- Using operators that combined old terms into new ones;
- Guided by heuristics that noted regularities in data; and
- Applied these recursively to formulate higher-level relations.

This approach let it rediscover scientific laws from the history of physics and chemistry.

The system adopted Sir Francis Bacon's proposal that scientific discovery use *data-driven* strategies.

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

Basic algebraic relations:

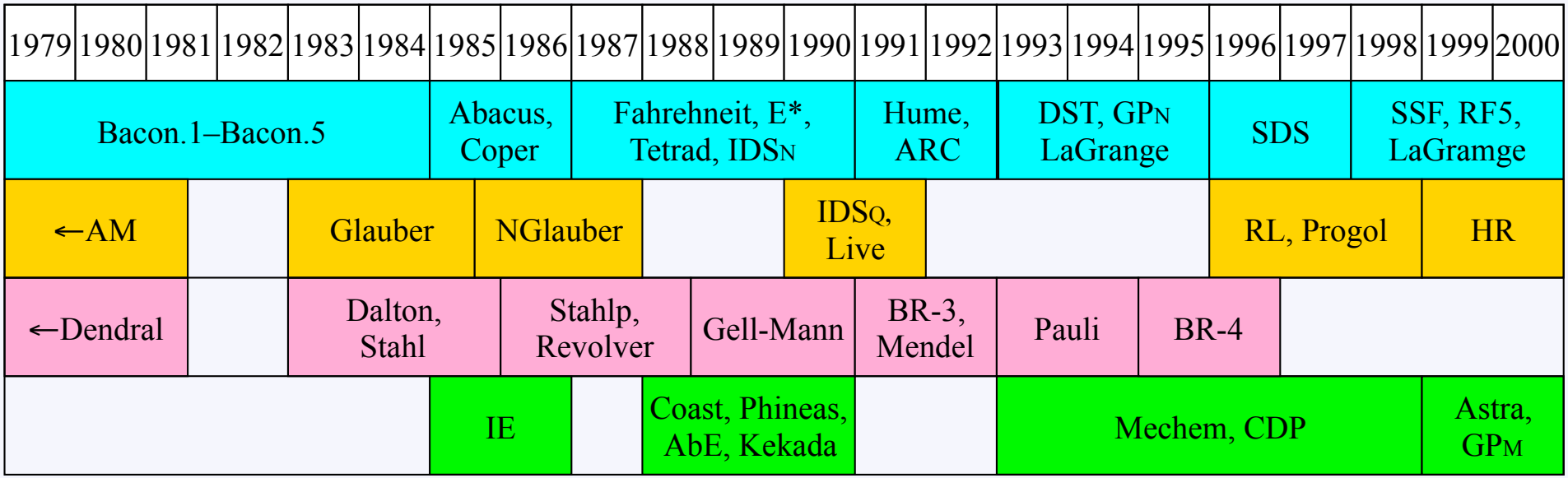
- Ideal gas law $PV = aNT + bN$
- Kepler's third law $D^3 = [(A - k) / t]^2 = j$
- Coulomb's law $FD^2 / Q_1Q_2 = c$
- Ohm's law $TD^2 / (LI - rI) = r$

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

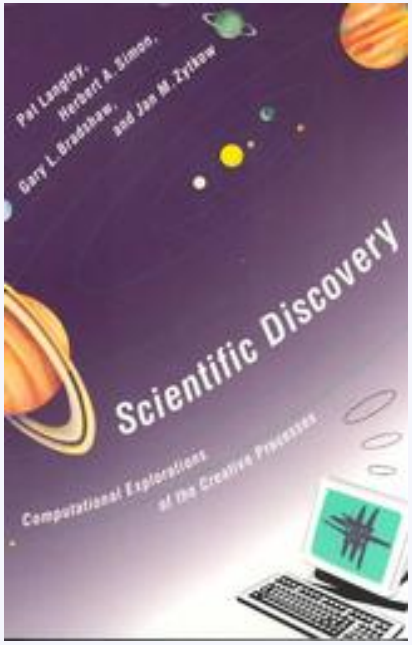
AI systems of this type have helped to discover new knowledge in many scientific fields:

- Reaction pathways in catalytic chemistry (Valdes-Perez, 1994, 1997)
- 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)
- Dynamic laws of ecological behavior (Todorovski et al., 2000)
- Models of gene-influenced metabolism in yeast (King et al., 2009)

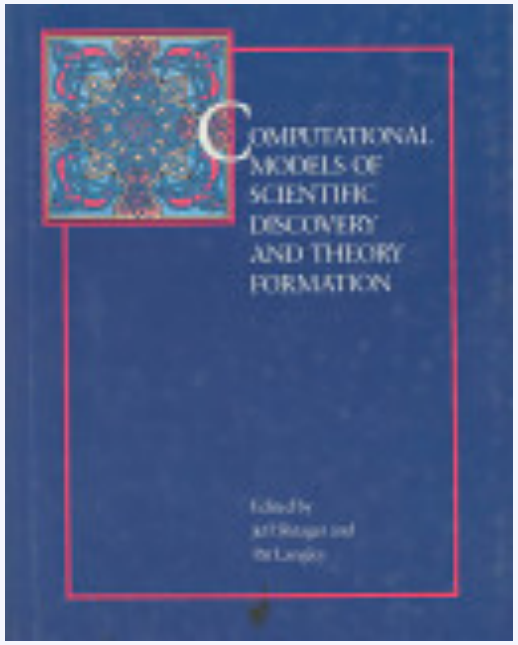
Each of these has led to publications in the *refereed literature of the relevant scientific field* (Langley, 2000).

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:

- Used computational methods to find regularities in the data
- Adopted heuristic search through a space of hypotheses
- Emphasized the availability of large amounts of data
- 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

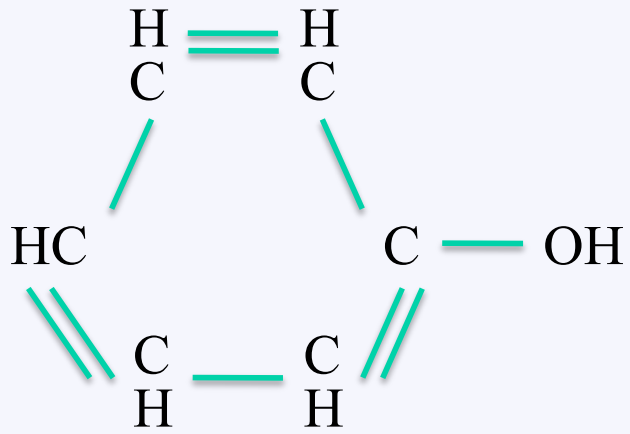
Explanatory models move beyond description to provide deeper accounts linked to theoretical constructs.

Can computational mechanisms 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.

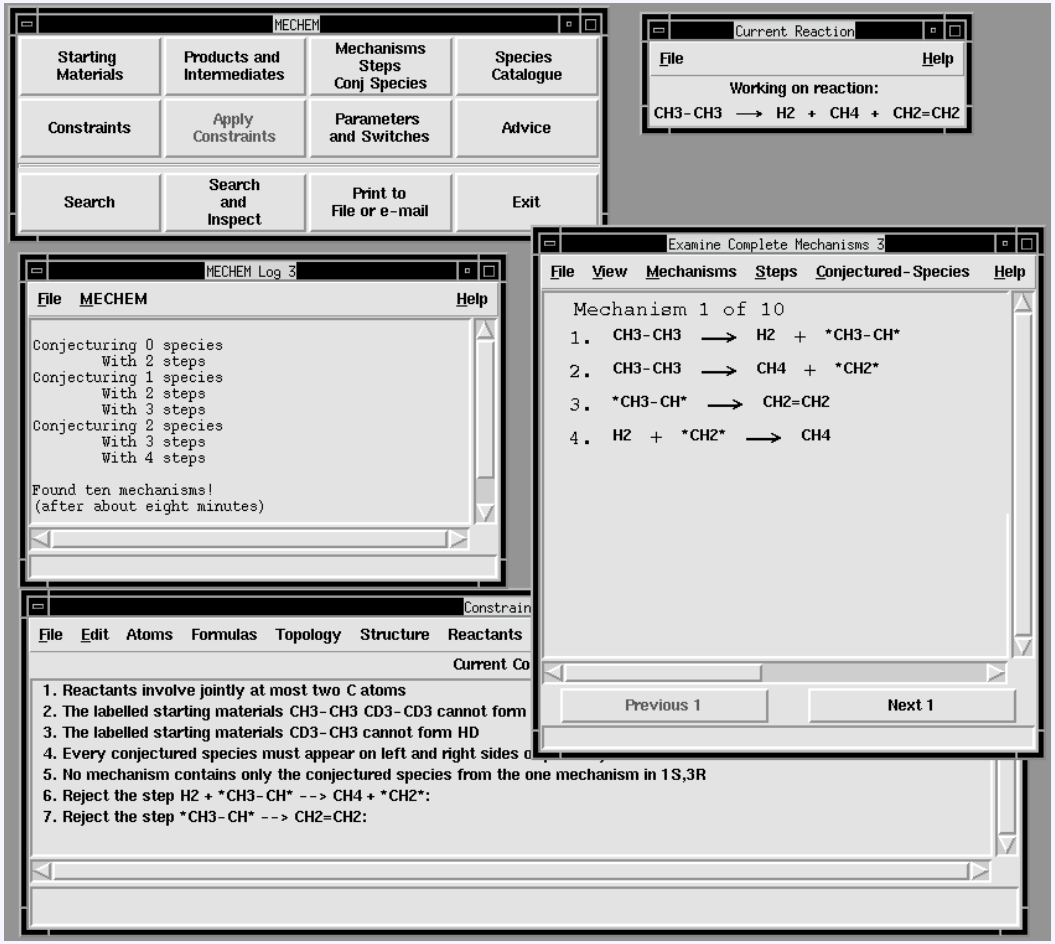
E.g., from the formula C_6H_5OH and other relevant information, the program produced structures like:



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.



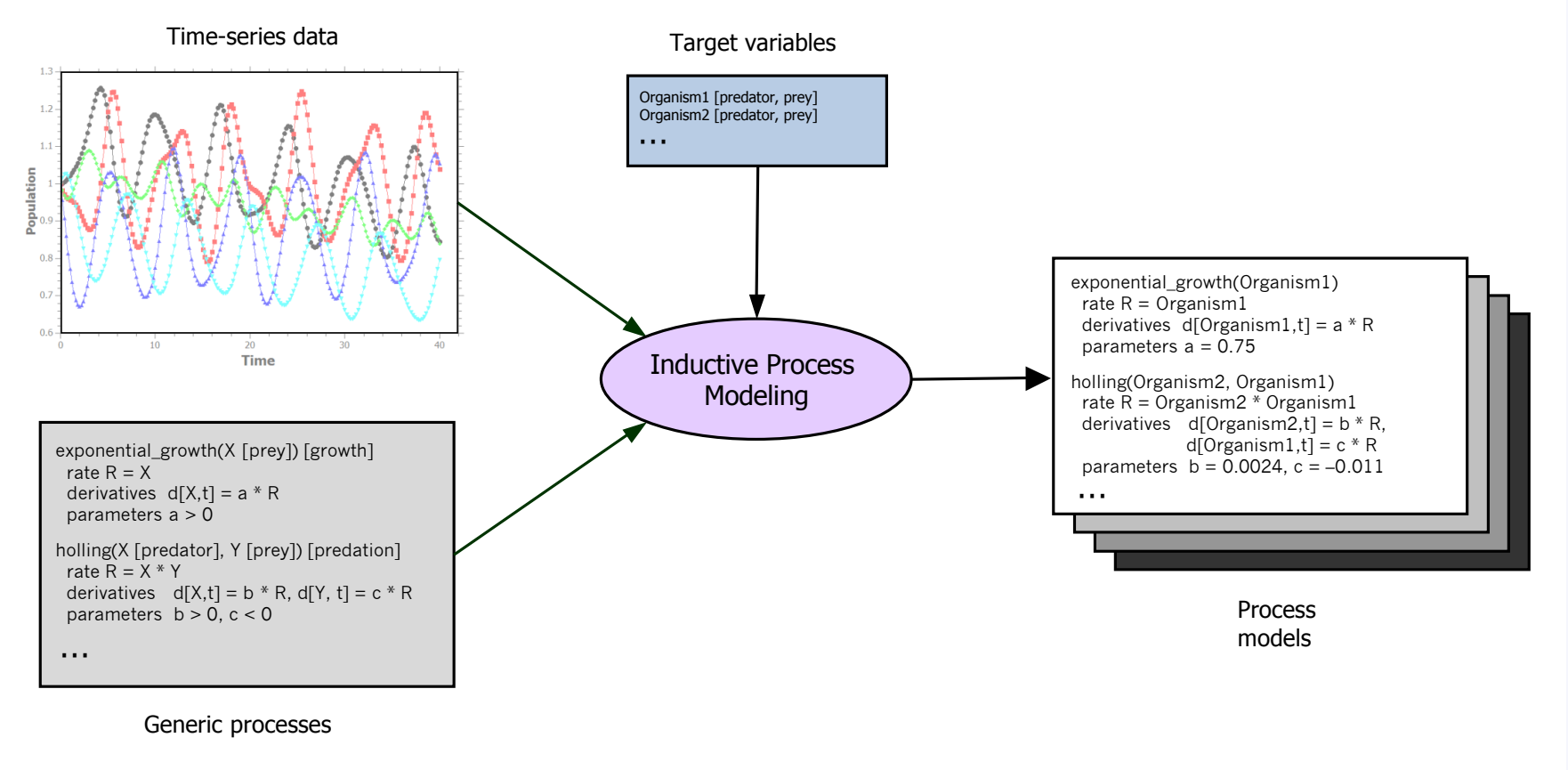
The system used constrained exhaustive search to generate candidate explanations.

Users could select constraints they deemed relevant to the current task.

MECHEM found numerous pathways that led to articles in the chemistry literature.

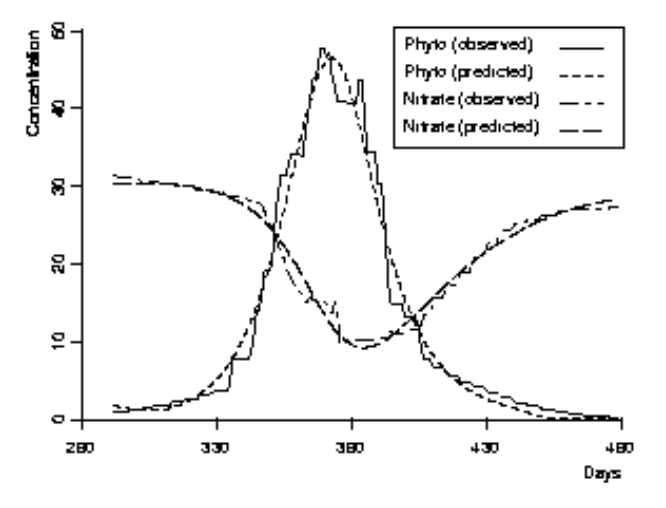
Recent Progress: Inductive Process Modeling

Inductive process modeling constructs explanations of time series from background knowledge (Bridewell et al., *2008*).

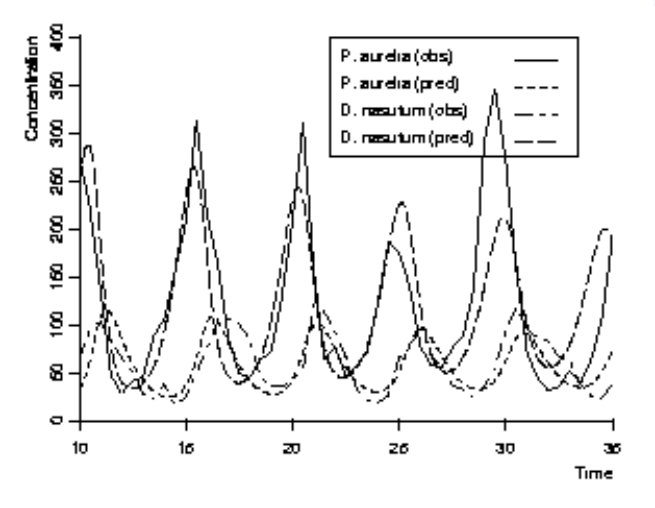


Models are stated as sets of *differential equations* organized into higher-level *processes*.

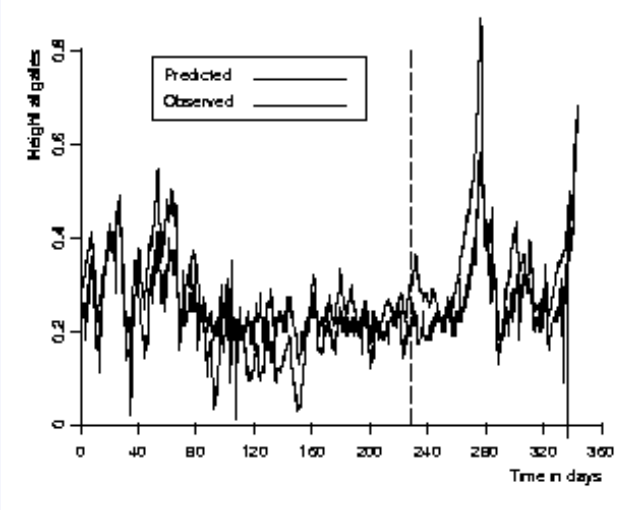
Successes of Inductive Process Modeling



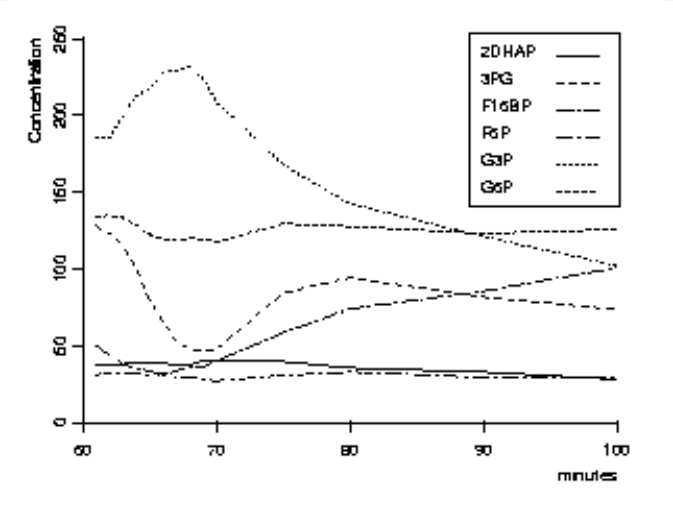
aquatic ecosystems



protist dynamics

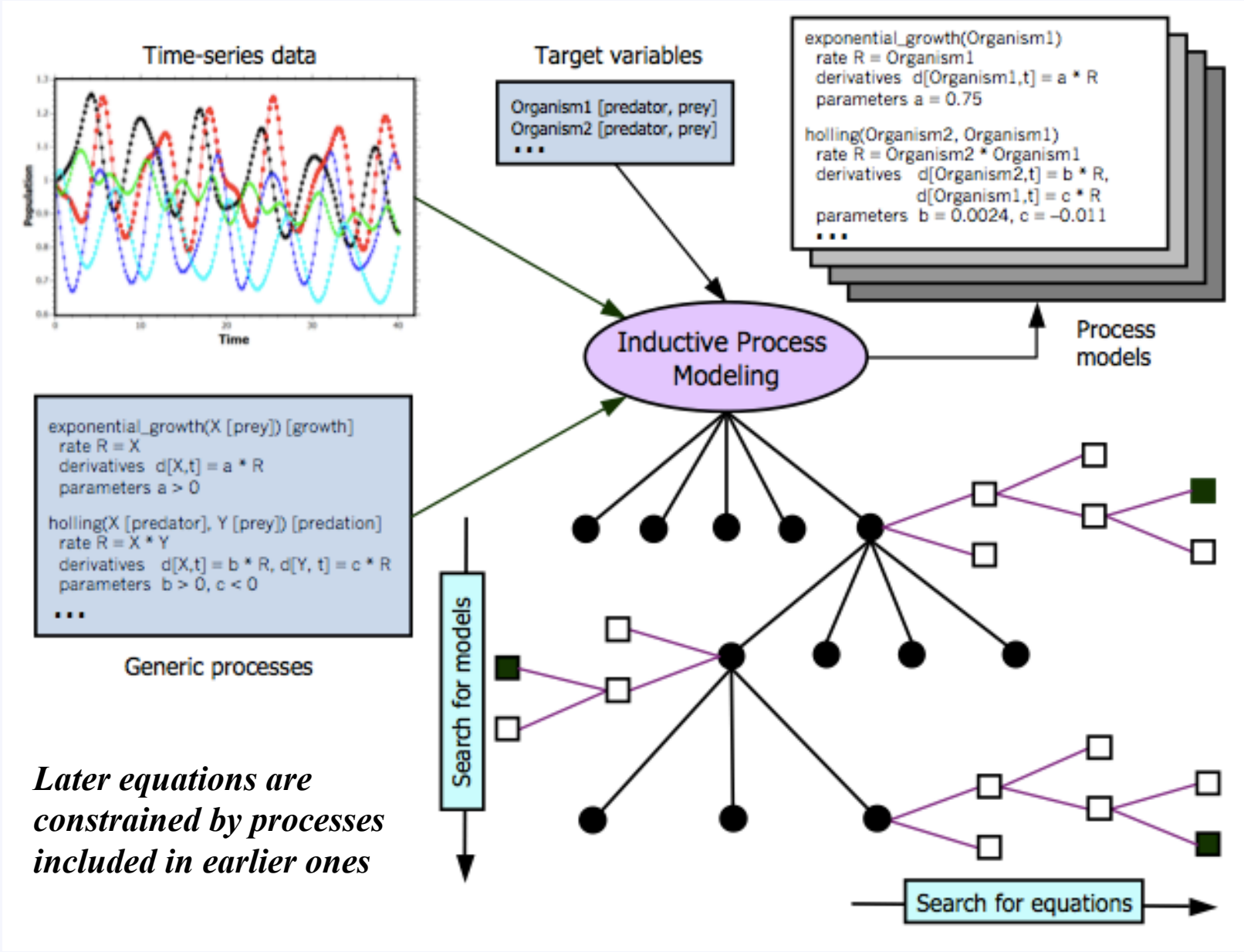


hydrology



biochemical kinetics

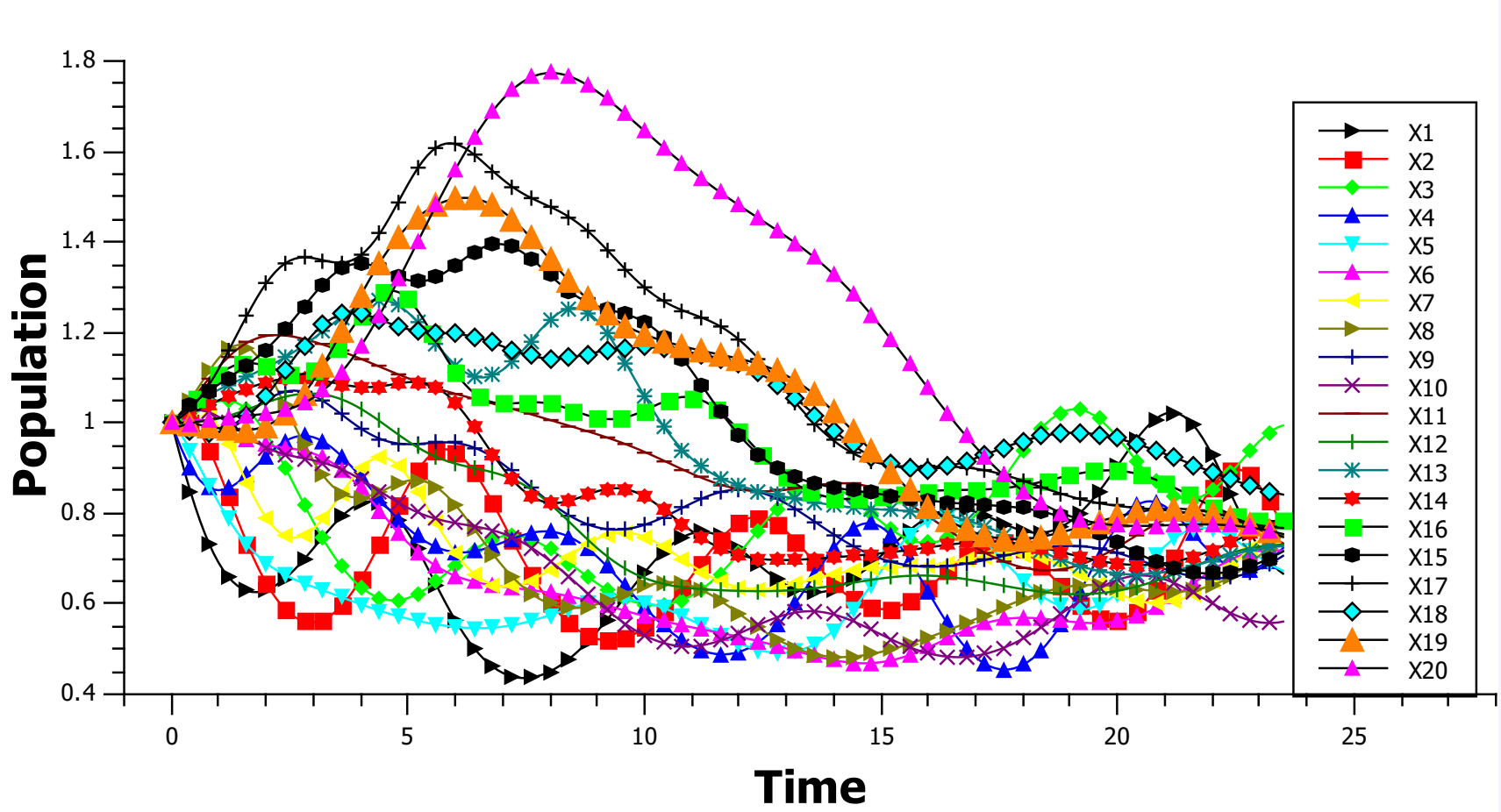
Heuristic Search for Process Models in RPM



Langley and Arvay (2015, 2016)

Scaling to Complex Models

RPM also finds an accurate model for a *20-organism* food chain.



This suggests the system scales well to difficult modeling tasks.

Recent Progress: Cell Biology

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
- Revises its causal model for how genes influence metabolism

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

Their system has found models of metabolic regulation in yeast.

Zytkow et al. (1990) reported a much earlier robot scientist in the field of electrochemistry.

Recent Progress: Food Webs in Ecology

In other recent work, Bohan et al. (2011) have used abductive logic programming to:

- Process data on relative abundances on invertebrates in fields
- Use knowledge about relative size, cooccurence, and predation
- Infer a three-level food web that relates 45 distinct species

Examination of the literature showed that most of these links were consistent with known predatory relations.

However, the system also hypothesized novel predations that ecologists found interesting and important.

Recent Progress: Cosmogenic Dating

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

- Inputs nucleotide densities for rocks from a landform
- Incorporates knowledge about possible geological processes
- Generates process models for how the landform was produced
- Weighs arguments for and against each process explanation

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

Lessons for the Research Community

Research on scientific discovery offers some important lessons:

- Science adopts *explicit formalisms* for theories and models that are communicable to others.
- Scientific research is not entirely data driven; it often uses *existing knowledge* to aid the discovery process.
- Data is not the sole driver of discovery; science is a *closed loop* of model revision and data collection.
- Science is concerned with more than prediction; mature fields insist that observations be *explained* in deeper terms.
- Scientific insights do not require large amounts of data; in many fields, one must work with *sparse samples*.

We need less work on large data sets and more work on scaling to *complex models* and to *large spaces of models*.

Myths about Computers in Science

Finally, we should debunk three damaging myths (PITAC, 2005):

- Computing is changing the basic nature and operation of science.
  - No. Science has *always been a computational endeavor*, and digital computers do not alter its basic steps or their relationships.
- Traditional science stood on two legs – theory and observation – and computing offers a third – simulation – and a fourth – data analysis.
  - No. *Every facet of science is computational*, and we can develop digital aids to make it more efficient and effective.
- Computer-aided science is best pursued with domain-specific tools.
  - No. There are *general principles* of science that apply to many fields, and we can encode them in programming abstractions.

We need less rhetoric on how 'computers will change everything' and more work on how to aid the standard scientific process.

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 emphasized 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.

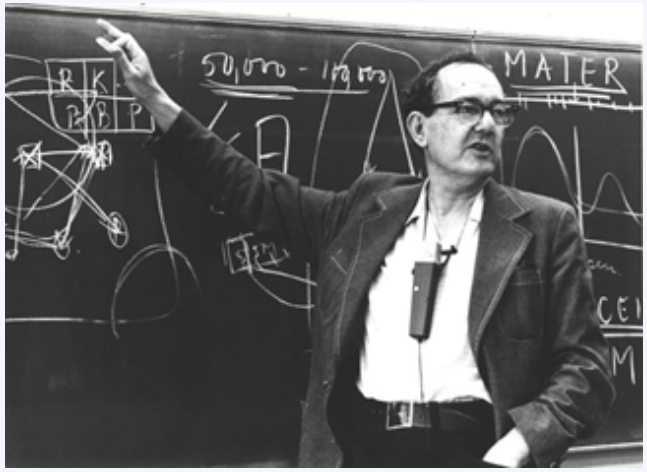
Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

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