Computational Discovery of Quantitative Process Models

Pat Langley

Institute for the Study of
Learning and Expertise

Center for Design Research
Stanford University

AAAI Spring Symposium on
Computational Approaches to Scientific Discovery

Thanks to K. Arrigo, A. Arvay, G. Bradshaw, S. Borrett, W. Bridewell, S. Dzeroski, H. Simon, L. Todorovski, and J. Zytkow for their contributions to this research.

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?

Explanatory Discovery Systems

The answer is *yes*. Discovery researchers have devised systems that address this challenge:

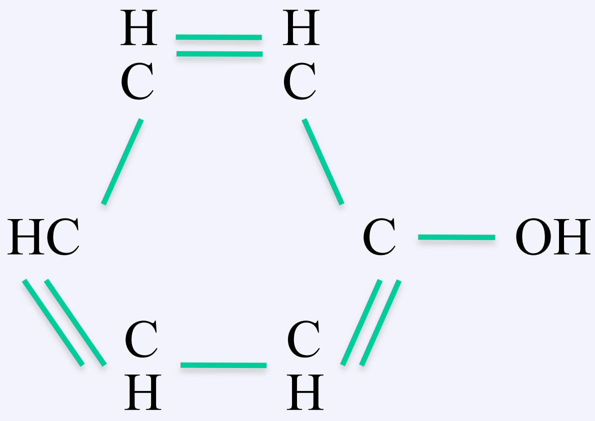
- DENDRAL (Lindsay et al.,1980) infers chemical structure from a formula, a mass spectrogram, and chemical knowledge.
- MECHEM (Valdes-Perez, 1994) generates pathways to explain reactions using chemical knowledge and constrained search.
- Adam (King et al., 2009) combines experimental design, data collection, and causal inference to model yeast metabolism.
- A/ILP (Bohan et al., 2011) uses abductive logic programming to infer a food web for 45 invertebrates from relative abundances.
- ACE (Anderson et al., 2014) uses nucleotide densities of rocks to generate process models for how a landform was produced.

These systems join data with knowledge to guide search, with models offering explanatory accounts of phenomena.

Inferring Chemical Structures

DENDRAL (Lindsay et al., 1980) inferred a molecule's chemical bonds given its component formula and a mass spectrogram.

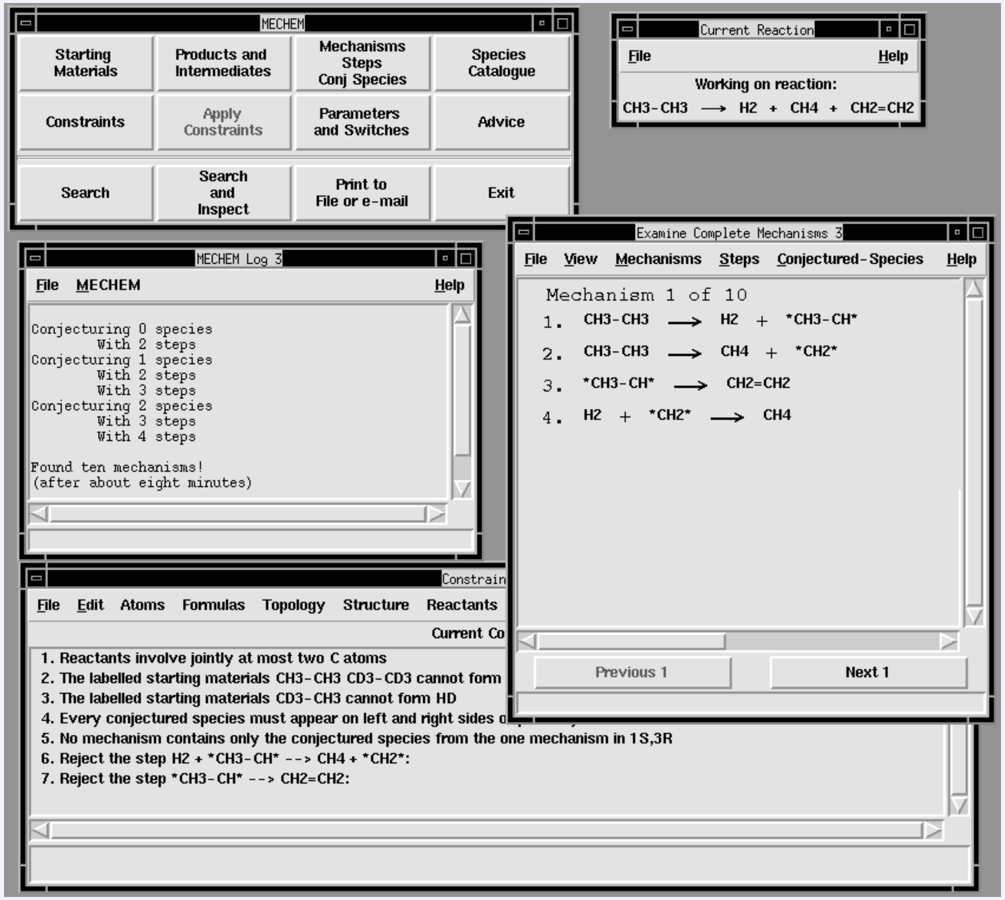
E.g., from the formula C_6H_5OH and other relevant information, the program produced structures like:



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

Discovering Reaction Pathways

MECHEM (Valdes-Perez, 1994) generated plausible pathways to explain chemical reactions.



Working on reaction:

CH3-CH3 ⟶ H2 + CH4 + CH2=CH2

```
Conjecturing 0 species
        With 2 steps
Conjecturing 1 species
        With 2 steps
        With 3 steps
Conjecturing 2 species
        With 3 steps
        With 4 steps

Found ten mechanisms!
(after about eight minutes)
```

```
Mechanism 1 of 10
1.  CH3-CH3   ⟶  H2  +  *CH3-CH*
2.  CH3-CH3   ⟶  CH4  +  *CH2*
3.  *CH3-CH*   ⟶  CH2=CH2
4.  H2  +  *CH2*   ⟶  CH4
```

1. Reactants involve jointly at most two C atoms
2. The labelled starting materials CH3-CH3 CD3-CD3 cannot form
3. The labelled starting materials CD3-CH3 cannot form HD
4. Every conjectured species must appear on left and right sides o
5. No mechanism contains only the conjectured species from the one mechanism in 1S,3R
6. Reject the step H2 + *CH3-CH* --> CH4 + *CH2*:
7. Reject the step *CH3-CH* --> CH2=CH2:

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.

Closed-Loop Discovery in 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.

Proposing Food Webs in Ecology

In other 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.

Cosmogenic Dating

Anderson et al. (2014) reported ACE, a 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 was used actively by many geologists to understand their data.

Quantitative Explanatory Models

The majority of research on computational scientific discovery has focused on either:

- Inducing numeric laws that describe *quantitative* observations
- Abducing structural accounts to explain *qualitative* phenomena

But scientists in advanced fields often combine both activities to create models that:

- Postulate unobserved structural relations among entities
- Incorporate functional forms with numeric parameters

Can we also develop systems that discover such *quantitative explanatory models*?

Early Work on Quantitative Explanations

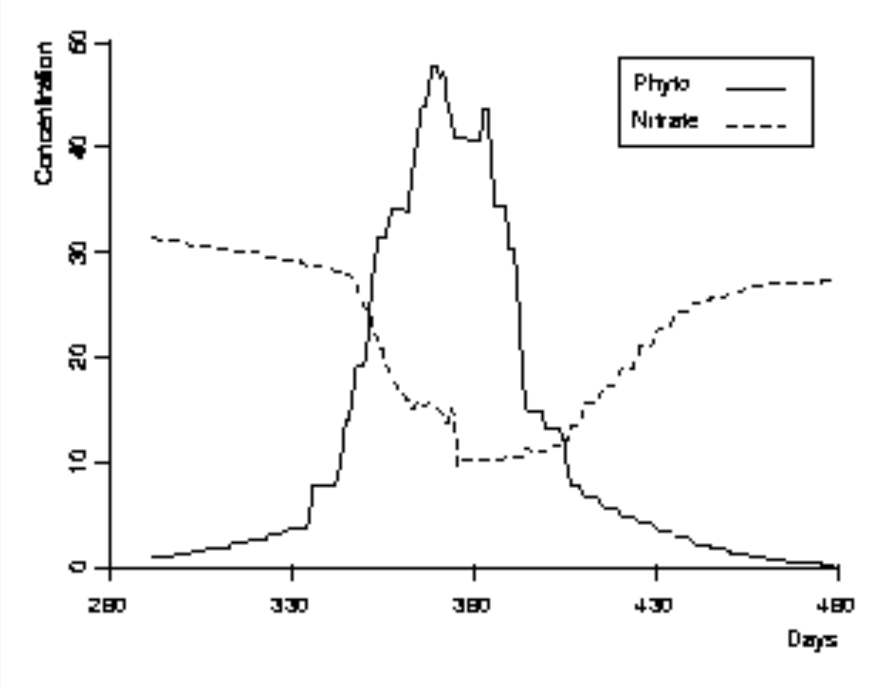
There has been some research on computational discovery of quantitative explanations:

- Inferring *abstract causal models* / structural equation models (Glymour et al., 1987; Spirtes et al., 1993)
- Identifying sets of *linked differential equations* (Dzeroski & Todorovski, 1993; Stolle & Bradley, 1998; Koza et al., 2001)

These combined distinct numeric equations into qualitative structures, but they remained largely descriptive.

Can we also automate the discovery of quantitative models that postulate *unobserved variables and processes*?

Modeling the Ross Sea Ecosystem



Formal accounts of ecosystem dynamics are often cast as sets of differential equations.

Here four equations describe the concentrations of phytoplankton, zooplankton, nitrogen, and detritus in the Ross Sea over time.

Such models can match observed variables with some accuracy.

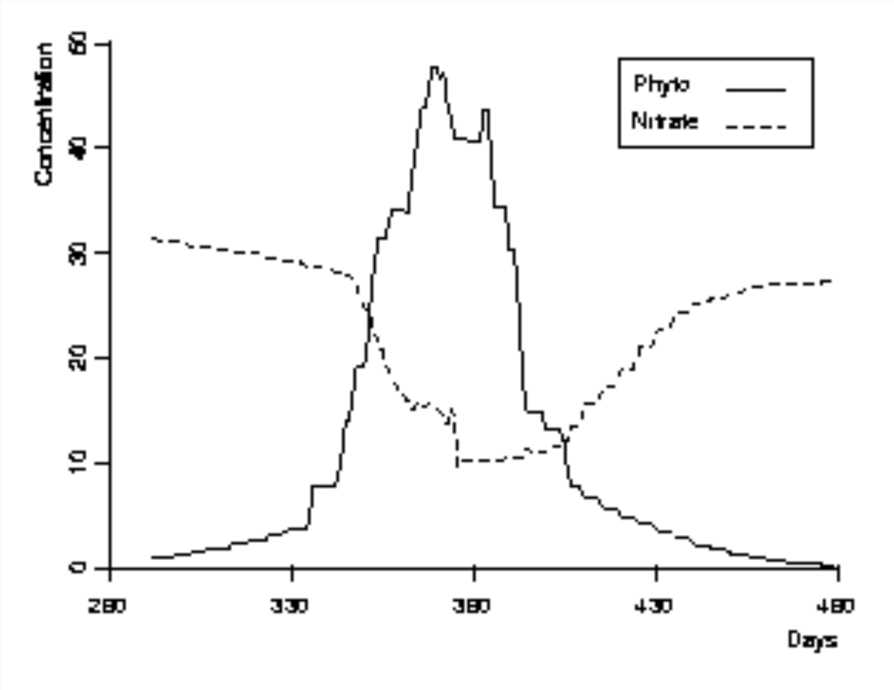
d[phyto,t,1] = – 0.307 × phyto – 0.495 × zoo + 0.411 × phyto

d[zoo,t,1] = – 0.251 × zoo + 0.615 × 0.495 × zoo

d[detritus,t,1] = 0.307 × phyto + 0.251 × zoo + 0.385 × 0.495 × zoo – 0.005 × detritus

d[nitro,t,1] = – 0.098 × 0.411 × phyto + 0.005 × detritus

A Deeper Account of Ross Sea Dynamics



As phytoplankton uptakes nitrogen, its concentration increases and the nitrogen decreases. This continues until the nitrogen is exhausted, which leads to a phytoplankton die off. This produces detritus, which gradually remineralizes to replenish nitrogen. Zooplankton grazes on phytoplankton, slowing the latter's increase and also producing detritus.

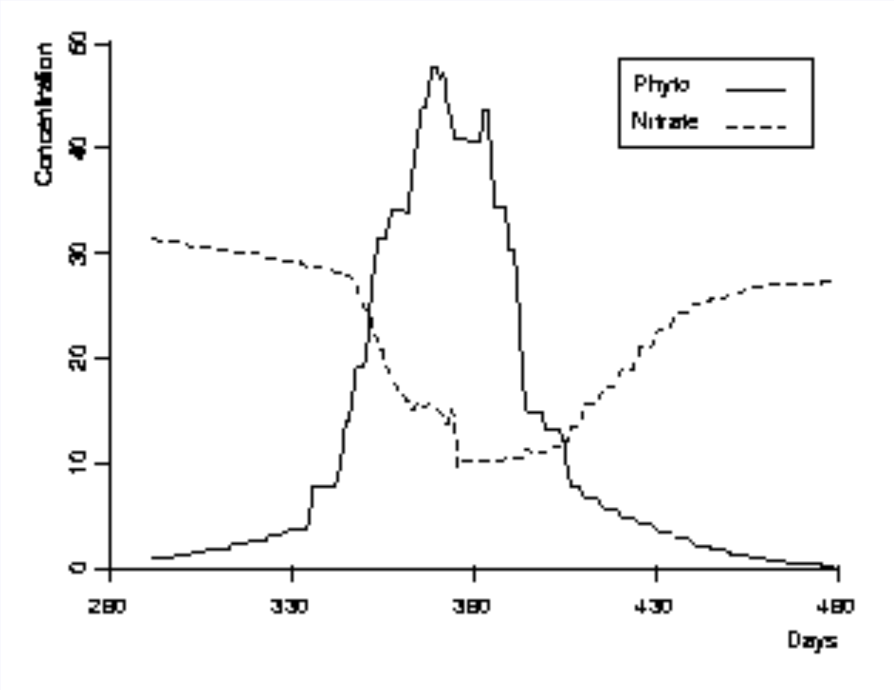
d[phyto,t,1] = – 0.307 × phyto – 0.495 × zoo + 0.411 × phyto

d[zoo,t,1] = – 0.251 × zoo + 0.615 × 0.495 × zoo

d[detritus,t,1] = 0.307 × phyto + 0.251 × zoo + 0.385 × 0.495 × zoo – 0.005 × detritus

d[nitro,t,1] = – 0.098 × 0.411 × phyto + 0.005 × detritus

Processes in Ross Sea Dynamics



As phytoplankton uptakes nitrogen, its concentration increases and the nitrogen decreases. This continues until the nitrogen is exhausted, which leads to a phytoplankton die off. This produces detritus, which gradually remineralizes to replenish nitrogen. Zooplankton grazes on phytoplankton, slowing the latter’s increase and also producing detritus.

$$\boxed{d[\text{phyto},t,1] =} - 0.307 \times \text{phyto} - 0.495 \times \text{zoo} + \boxed{0.411 \times \text{phyto}}$$

$$d[\text{zoo},t,1] = - 0.251 \times \text{zoo} + 0.615 \times 0.495 \times \text{zoo}$$

$$d[\text{detritus},t,1] = 0.307 \times \text{phyto} + 0.251 \times \text{zoo} + 0.385 \times 0.495 \times \text{zoo} - 0.005 \times \text{detritus}$$

$$\boxed{d[\text{nitro},t,1] = - 0.098 \times 0.411 \times \text{phyto}} + 0.005 \times \text{detritus}$$

A Process Model for Ross Sea Dynamics

```
process phyto_loss(phyto, detritus)
  equations:   d[phyto,t,1] = –0.307 × phyto
               d[detritus,t,1] = 0.307 × phyto

process zoo_loss(zoo, detritus)
  equations:   d[zoo,t,1] = –0.251 × zoo
               d[detritus,t,1] = 0.251 × zoo

process zoo_phyto_grazing(zoo, phyto, detritus)
  equations:   d[zoo,t,1] = 0.615 × 0.495 × zoo
               d[detritus,t,1] = 0.385 × 0.495 × zoo
               d[phyto,t,1] = –0.495 × zoo

process nitro_uptake(phyto, nitro)
  equations:   d[phyto,t,1] = 0.411 × phyto
               d[nitro,t,1] = –0.098 × 0.411 × phyto

process nitro_remineralization(nitro, detritus)
  equations:   d[nitro,t,1] = 0.005 × detritus
               d[detritus,t,1 ] = –0.005 × detritus
```

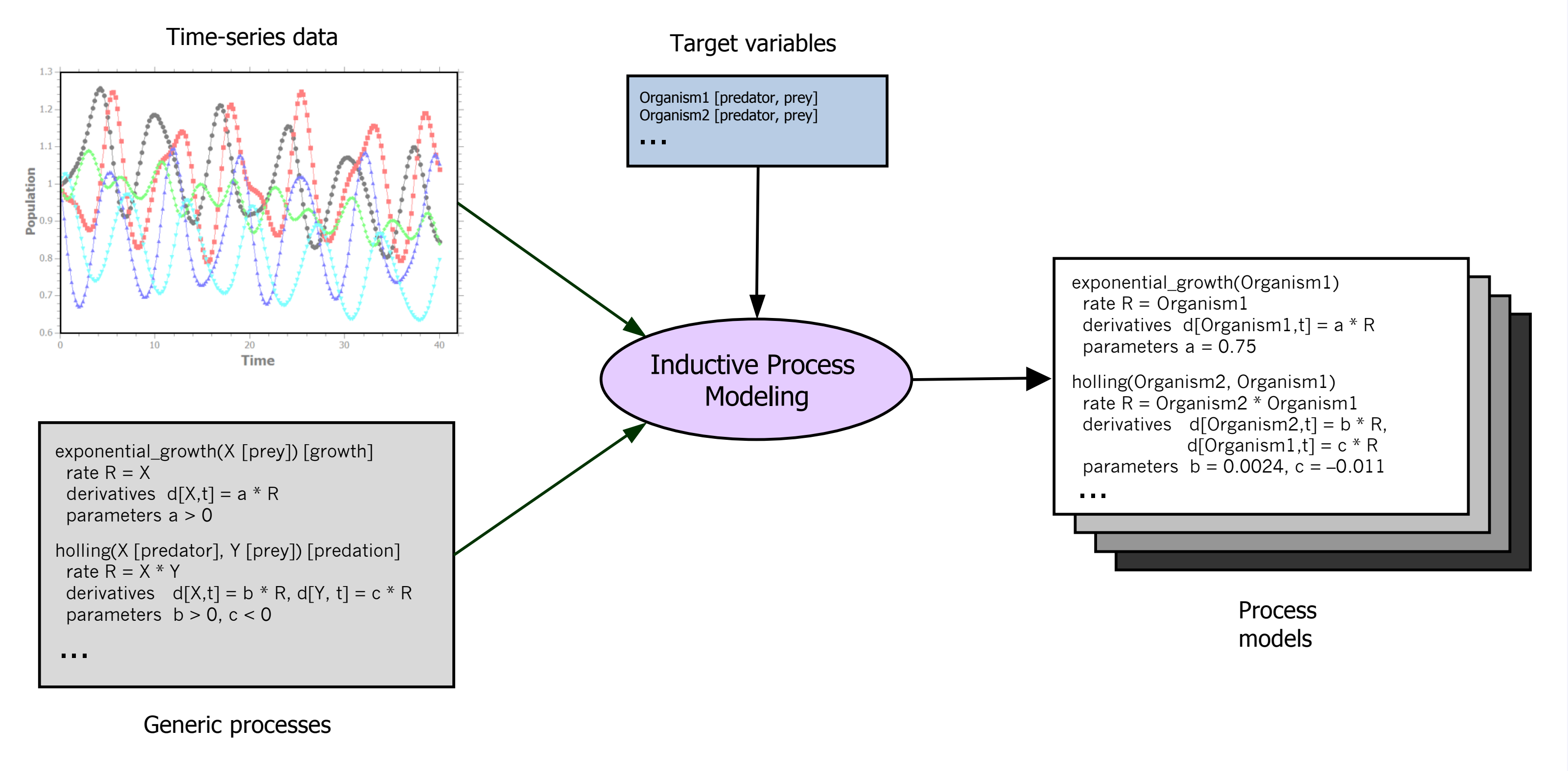
We can reformulate such an account by restating it as a *quantitative process model*.

This maps onto the earlier differential equation model, but it is explicit about the component processes.

Each process indicates that certain terms in equations *must stand or fall together*.

Inductive Process Modeling

Inductive process modeling is the task of constructing such process models from data and knowledge (Langley et al., *ICML-2002*).



Time-series data

Target variables

```
Organism1 [predator, prey]
Organism2 [predator, prey]
…
```

Inductive Process Modeling

```
exponential_growth(X [prey]) [growth]
  rate R = X
  derivatives  d[X,t] = a * R
  parameters a > 0
holling(X [predator], Y [prey]) [predation]
  rate R = X * Y
  derivatives   d[X,t] = b * R, d[Y, t] = c * R
  parameters  b > 0, c < 0
…
```

Generic processes

```
exponential_growth(Organism1)
  rate R = Organism1
  derivatives  d[Organism1,t] = a * R
  parameters a = 0.75
holling(Organism2, Organism1)
  rate R = Organism2 * Organism1
  derivatives   d[Organism2,t] = b * R,
                d[Organism1,t] = c * R
  parameters  b = 0.0024, c = –0.011
…
```

Process models

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

Some Generic Processes

```
process exponential_loss(S, D)
  variables: S{species}, D{detritus}
  parameters: $\alpha$ [0, 1]
  equations:   $d[S, t, 1] = -1 \times \alpha \times S$
               $d[D, t, 1] = \alpha \times S$
```

```
generic process grazing(S1, S2, D)
  variables: S1{species}, S2{species}, D{detritus}
  parameters: $\rho$ [0, 1], $\gamma$ [0, 1]
  equations:   $d[S1, t, 1] = \gamma \times \rho \times S1$
               $d[D ,t, 1] = (1 - \gamma) \times \rho \times S1$
               $d[S2, t, 1] = -1 \times \rho \times S1$
```

```
generic process nutrient_uptake(S, N)
  variables: S{species}, N{nutrient}
  parameters: $\tau$ [0, $\infty$], $\beta$ [0, 1], $\mu$ [0, 1]
  conditions:  $N > \tau$
  equations:   $d[S, t, 1] = \mu \times S$
               $d[N, t, 1] = -1 \times \beta \times \mu \times S$
```

```
process remineralization(N, D)
  variables: N{nutrient}, D{detritus}
  parameters: $\pi$ [0, 1]
  equations:
     $d[N, t, 1] = \pi \times D$
     $d[D, t, 1] = -1 \times \pi \times D$
```

```
process constant_inflow(N)
  variables: N{nutrient}
  parameters: $\nu$ [0, 1]
  equations:   $d[N, t, 1] = \nu$
```

Each generic process specifies variable types, functional forms, and ranges on parameters.

These provide *building blocks* from which to compose models.

Inductive Process Modeling as Search

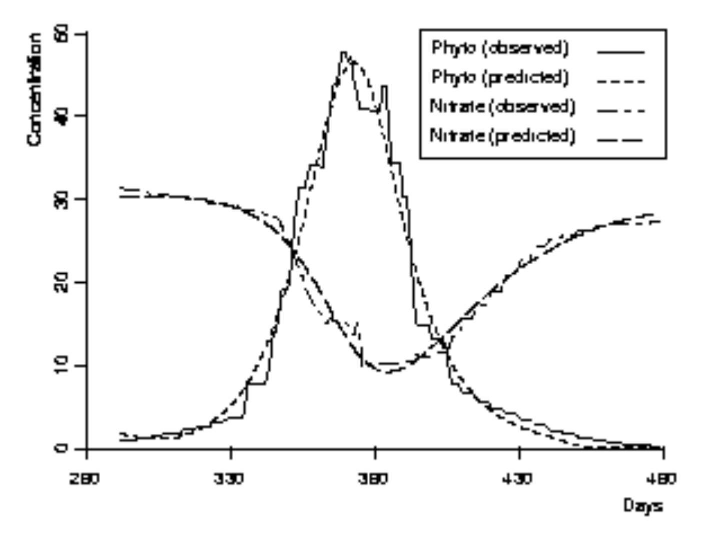
We can view process model induction as constrained search through two distinct but connected spaces:

- A *discrete* space of *model structures*: a set of processes that specify variables and equations that relate them
- A *continuous* space of *numeric parameters* for each model structure, with ranges for possible values.

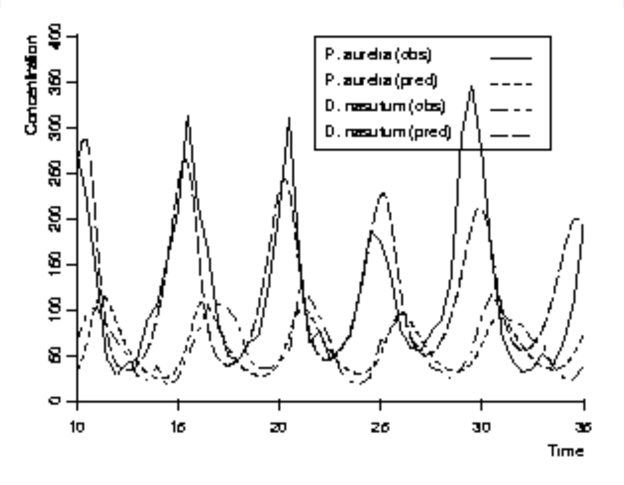
Our early systems carried out *exhaustive* search in the structure space and *gradient descent* to estimate parameters.

This sufficed for models with under ten processes and produced some promising results.

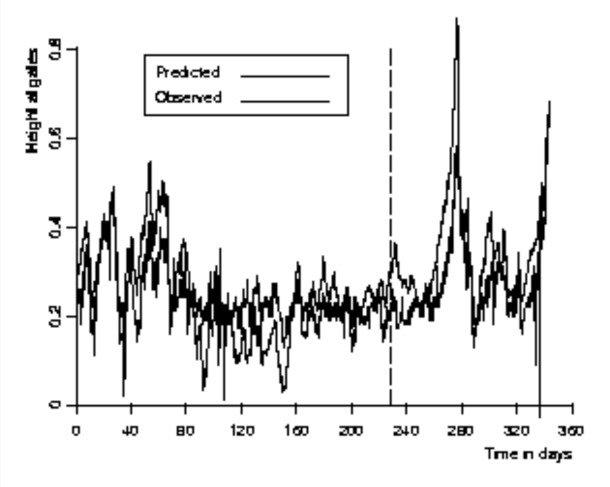
Early Results for Process Modeling



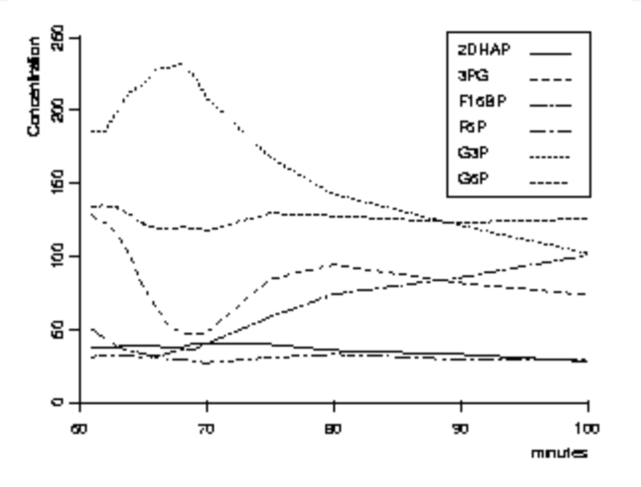
aquatic ecosystems



protist dynamics



hydrology



biochemical kinetics

Extensions to Inductive Process Modeling

In addition, we extended the basic framework to support:

- Inductive revision of process models (*Ecological Modeling*, 2006)
- Hierarchical generic processes to constrain search (*AAAI-2005*)
- Ensembles of processes to mitigate overfitting (*ICML-2005*)
- Iterative optimization for missing observations (*ECML-2006*)
- Induction of spatio-temporal process models *(AAAI-2010)*
- Constraints on processes for plausibility *(Topics in CogSci, 2010)*

These extensions made process model induction more robust along multiple fronts.

Drawbacks of the Approach

Despite these successes, this approach suffers from four key drawbacks, in that it:

- Evaluates *full model structures*, so disallows heuristic search
- Requires *repeated simulation* to estimate model parameters
- Invokes *random restarts* to reduce chances of local optima
- Despite these steps, it can still find poorly-fitting models

99.99 percent of CPU time

As a result, it does not scale well to complex modeling tasks and it is not reliable.

More recent research has *reformulated* the task in ways that avoids these problems (Langley & Arvay, *AAAI-2015*).

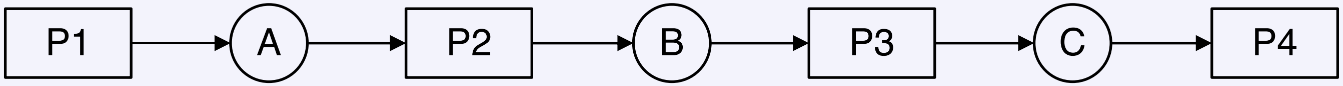
Rate-Based Process Models

The new modeling framework is more constrained in that each process P must include:

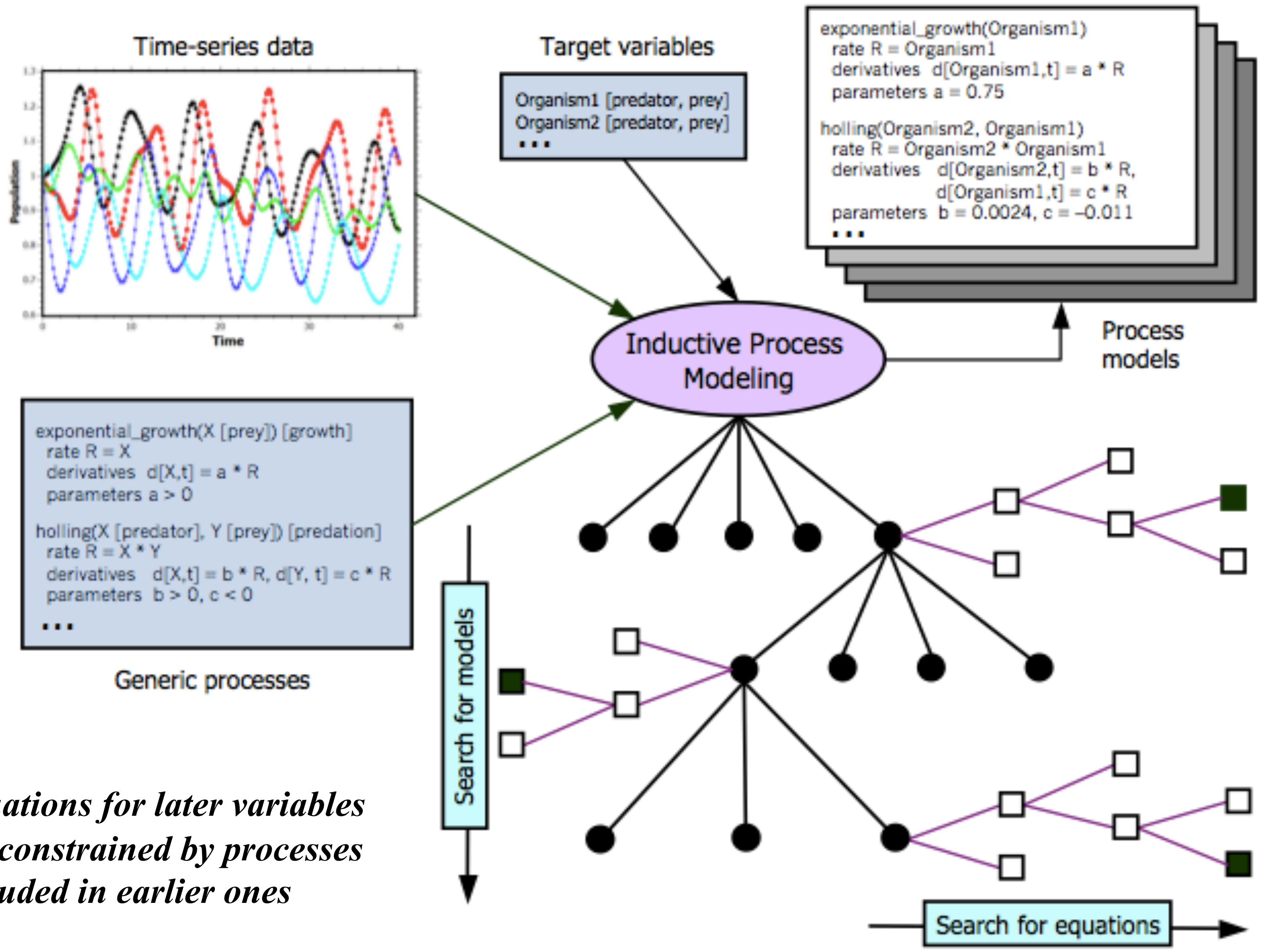
- A *rate* that denotes P's speed / activation on a given time step
- An *algebraic expression* that describes P's rate
- One or more *derivatives* that are proportional to P's rate

Negative derivatives correspond to process *inputs* and positive ones to *outputs*, much as in chemical reactions.

The notation has important mathematical properties that make model induction efficient and robust.



Two-Level Heuristic Search for Process Models



```
Organism1 [predator, prey]
Organism2 [predator, prey]
...
```

```
exponential_growth(Organism1)
  rate R = Organism1
  derivatives  d[Organism1,t] = a * R
  parameters a = 0.75

holling(Organism2, Organism1)
  rate R = Organism2 * Organism1
  derivatives   d[Organism2,t] = b * R,
                d[Organism1,t] = c * R
  parameters  b = 0.0024, c = –0.011
...
```

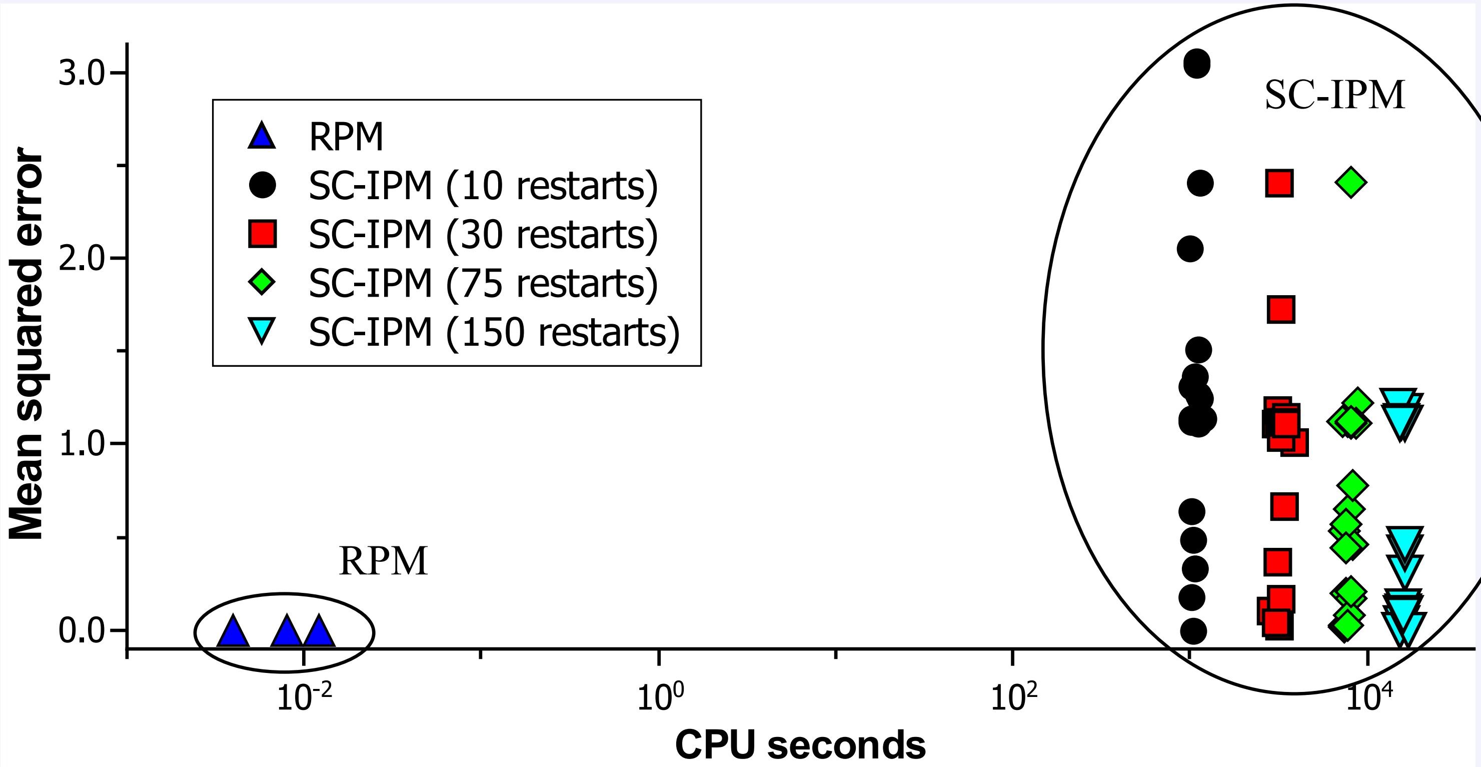
```
exponential_growth(X [prey]) [growth]
  rate R = X
  derivatives  d[X,t] = a * R
  parameters a > 0

holling(X [predator], Y [prey]) [predation]
  rate R = X * Y
  derivatives   d[X,t] = b * R, d[Y, t] = c * R
  parameters  b > 0, c < 0
...
```

Equations for later variables are constrained by processes included in earlier ones

Increased Speed and Robustness

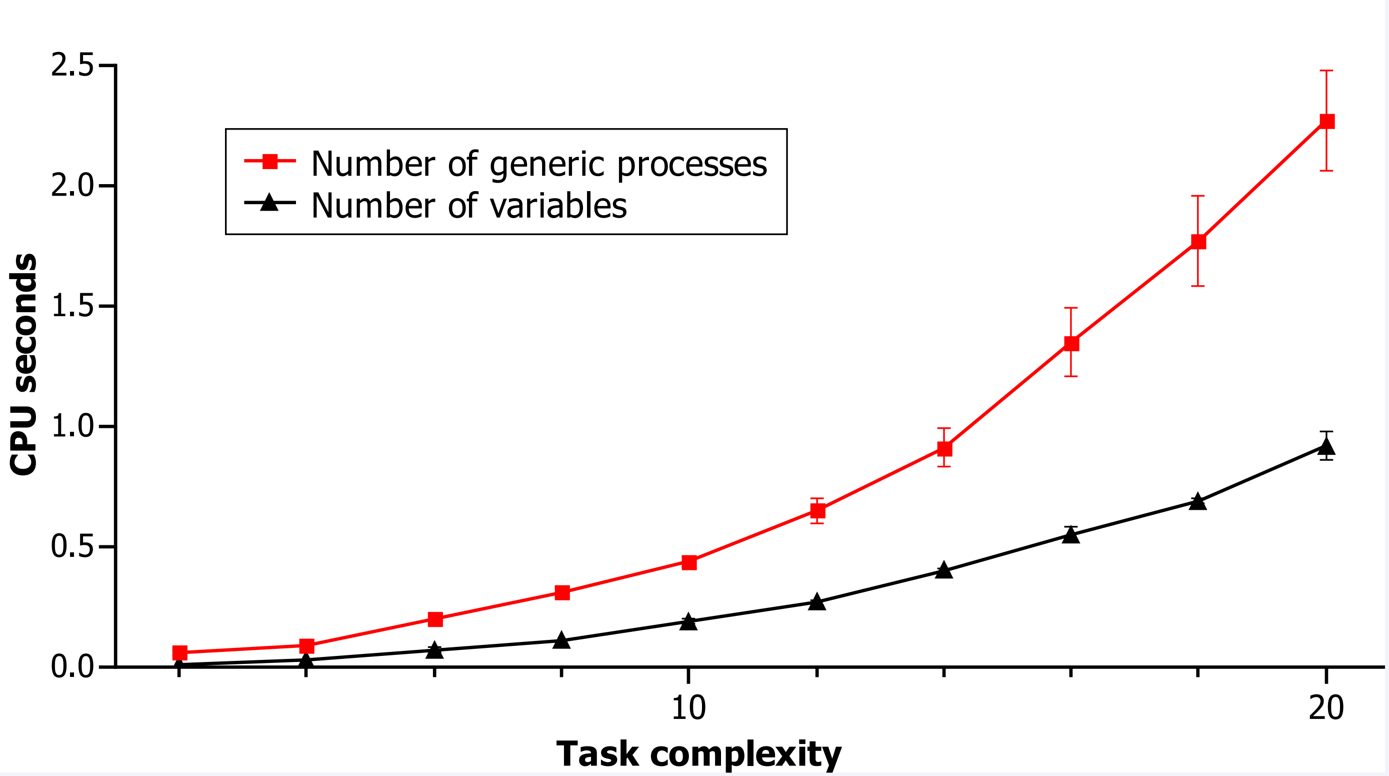
We compared the old and new approach on synthetic data for a three-variable predator-prey ecosystem.



The new system found accurate models far more reliably and ran *800,000 faster* than the earlier one.

Handling Noise and Complexity

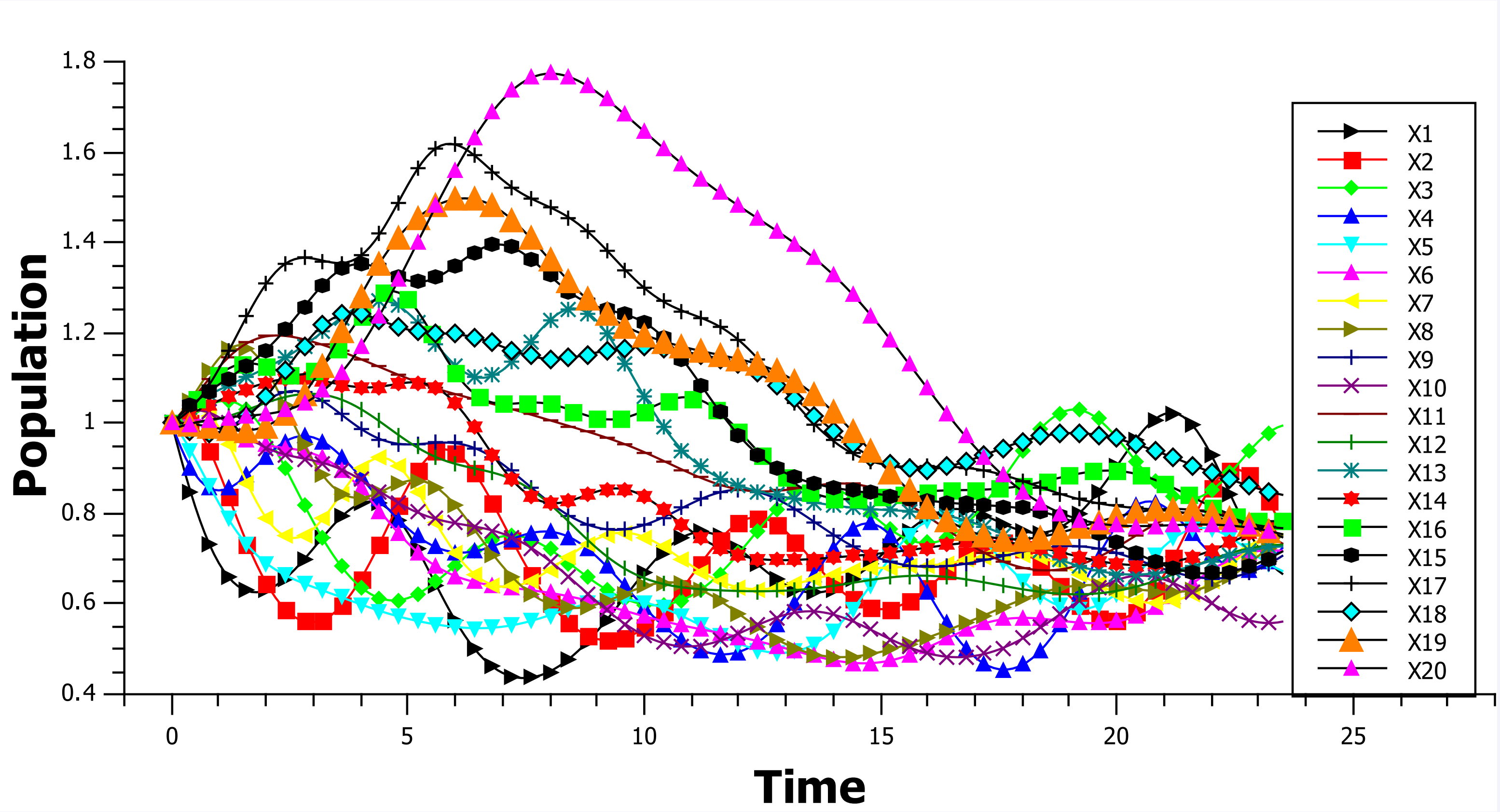
With smoothing, the approach handles 10% noise on synthetic data.



The approach also scales well to increasing numbers of generic processes and variables in the target model.

Behavior on Complex Synthetic Data

The approach finds the correct model for a *20-organism* food chain.



This is more evidence that it scales well to difficult modeling tasks.

Extensions to Rate-Based Modeling

In more recent work, we have augmented the rate-based approach to process model induction to:

- Adapt process models to new settings (Arvay / Langley, ACS 2015)
  - *Detect anomalies, reestimate parameters, revise structure*
- Selective induction of process models (Arvay / Langley, ACS 2016)
  - *Sample processes, delay variable bindings, find multiple equations*
- Posit new processes to break impasses (Langley / Arvay, AAAI-2017)
  - *Combine conceptual relations with algebraic rate templates*

These extensions have provided more coverage, scalability, and reliability than the basic approach.

Related and Future Research

Our approach builds on ideas from earlier research, including:

- Qualitative representations of scientific models (Forbus, 1984)
- Inducing differential equations (Todorovski, 1995; Bradley, 2001)
- Heuristic search and multiple linear regression
- Delayed commitment and feature selection

Our plans for extending the rate-based framework include:

- Devising experiments to discriminate among models
- Discovering forms of entirely new processes
- Finding multi-scale models at different temporal resolutions

These should extend coverage and usefulness even further.

Potential Applications

Scalable methods for process model induction would be useful in many practical settings, including:

- Elucidating new reaction pathways in biochemistry
- Understanding ecological dynamics of human microflora
- Designing reaction pathways for chemical production
- Designing metabolic pathways for synthetic biology

Computational tools for scientific discovery should let us not only interpret observations, but generate new behavior.

Summary Remarks

Inductive process modeling is a promising approach to creating scientific accounts that:

- Incorporates a formalism that is *familiar* to many scientists
- Uses background *knowledge* about the problem domain
- Produces meaningful results from *moderate* amounts of data
- Finds *causal* models that *explain*, not just describe, observations
- *Scales well* both to many processes and complex models

The framework combines search in a space of discrete model structures and a space of continuous model parameters.

For more information, see *http://www.isle.org/process/* .

Papers on Inductive Process Modeling

Arvay, A., & Langley, P. (2016). Heuristic adaptation of quantitative process models. Advances in Cognitive Systems, 4, 207–226.

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.

Langley, P. (2019). Scientific discovery, causal explanation, and process model induction. *Mind & Society*, *18*, 43–56.

Langley, P., & Arvay, A. (2015). Heuristic induction of rate-based process models. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 537–544). Austin, TX: AAAI Press.

Langley, P., Sanchez, J., Todorovski, L., & Dzeroski, S. (2002). Inducing process models from continuous data. *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 347–354). Sydney: Morgan Kaufmann.

Park, C., Bridewell, W., & Langley, P. (2010). Integrated systems for inducing spatio-temporal process models. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (pp. 1555–1560). Atlanta, GA: AAAI Press.