

Abduction-prediction model of scientific discovery  
reflected in a prototype system for model-based diagnosis

John R. Josephson

The Ohio State University  
Laboratory for Artificial Intelligence Research  
Computer and Information Science Department  
Columbus, Ohio, USA

<http://www.cis.ohio-state.edu/lair>



# *Consider this pattern of reasoning*

D is a collection of data (facts, observations, givens),  
Hypothesis H explains D (would, if true, explain D),  
No other hypothesis explains D as well as H does.

---

Therefore, H is probably correct.

“inference to the best explanation” - G. Harman

“the explanatory inference” - W. Lycan

“abduction” (C.S. Peirce) (sometimes).

*I'll call it "abduction"*

D is a collection of data

Hypothesis H explains D

No other hypothesis explains D as well as H does

---

Therefore, H is probably correct

Distinctive and familiar **pattern of justification.**

Part of "**commonsense logic.**"

# *Abduction*

A pattern of **justification** with:

- related information-processing **goals**, including:
  - generate explanations
  - evaluate
  - decide whether to accept
- related **processes**: to achieve abductive goals

Intelligent agents process information to arrive at judgments that have abductive justifications. (May be explicit or implicit.)

# *Confidence in an abductive conclusion*

D is a collection of data

Hypothesis H explains D

No other hypothesis explains D as well as H does

---

Therefore, H is **probably** correct

The judgment of likelihood in the conclusion should (and typically does) depend on . . . . .

# *confidence depends on*

- how decisively H surpasses the alternatives,
- how good H is by itself, independently of considering the alternatives,
- confidence in the accuracy of the data,
- how thorough was the search for alternative explanations

# *Rejection of rivals*

- **how decisively H surpasses the alternatives,**
  - how good H is by itself
  - confidence in the accuracy of the data,
  - how thorough was the search for alternative explanations

The rejection of rival hypotheses is important for justifying an abductive conclusion.

Failed predictions count against a hypothesis.

Abductions turn negative evidence against some hypotheses into positive evidence for alternative hypotheses.

# *Thoroughness of the search*

- how decisively H surpasses the alternatives,
- how good H is by itself
- confidence in the accuracy of the data,
- **how thorough was the search for alternative explanations**

This condition shows clearly why the logic of justification cannot be neatly separated from the logic of discovery.

Justification depends partly on evaluating the quality of the discovery process.

# *Acceptance also depends on*

pragmatic considerations, including:

- how strong is the need is to come to a conclusion at all?
  - especially considering the possibility of gathering further evidence before deciding

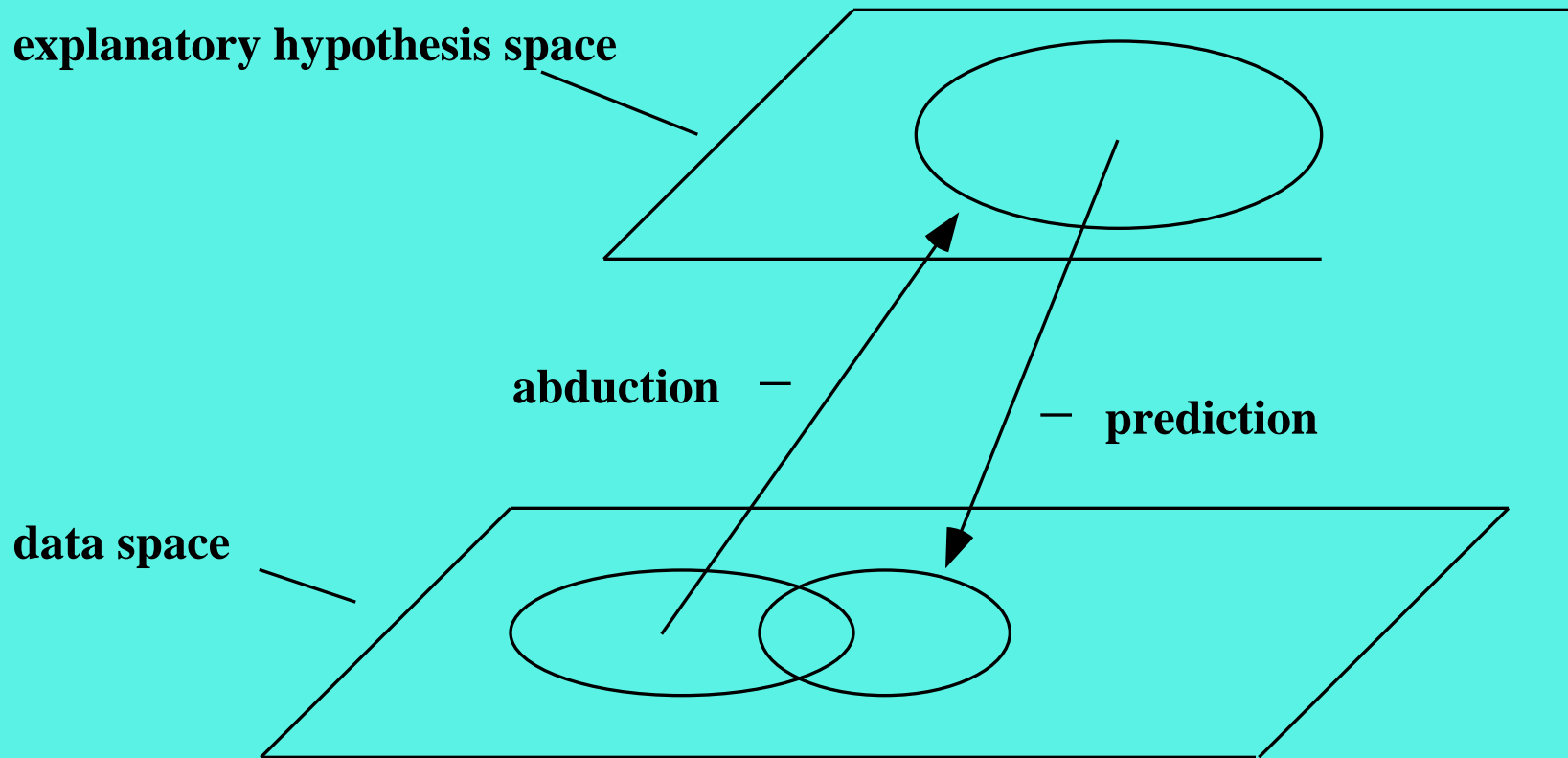
# Abductions are fallible

## *Of Course*

- The conclusion might be false, even if the premises are true. Unlike deductions.
- Yet, sometimes abductive justification is strong, and the conclusion is well justified.

- how decisively H surpasses the alternatives **HIGHLY**
- how good H is by itself **GOOD**
- confidence in the accuracy of the data **HIGH**
- how thorough was the search for alternative explanations **BROAD**

# *Abductions and predictions*



# Abduction-Prediction (A-P) model

*holds that*

- Theory formation and acceptance in science are inferentially well characterized as abduction (inference to the best explanation).
- Alternative explanatory hypotheses are evaluated, in part, based on the success of their predictions.
- A hypothesis gains strength from the weakness of rivals.
  - Also has independent strength.
  - Criteria for evaluation also include simplicity, explanatory power, . . . .

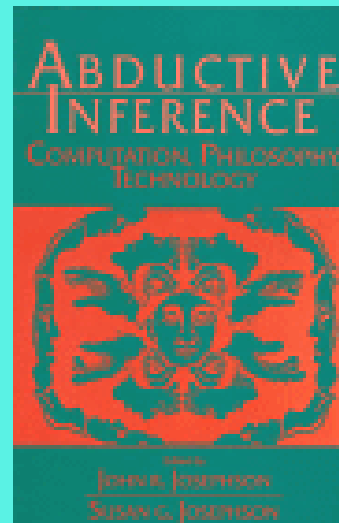
## *A-P model and H-D model*

The A-P model is very similar to the classical hypothetico-deductive model except that the A-P model:

- requires hypotheses to be explanatory,
- denies that predictions are always deductive
- emphasizes the rejection of rival hypotheses.

# *Predictions are not necessarily deductive*

Argued in "Conceptual analysis of abduction"  
in Josephson & Josephson, *Abductive Inference*,  
Cambridge U. Pr. (1994).



*Besides scientific discovery,  
abduction characterizes  
a variety of information-processing tasks*

- diagnosis
  - explain symptoms
- inferring mental state (e.g., intentions)
  - explains behavior
- inferring intended meaning (and other info) from observed speech, writing, gestures, ... .
  - special case of previous

# *Diagnosis & Discovery*

- Diagnosis can be a “laboratory model” of scientific discovery
  - Diagnosis is simpler, more reproducible.
    - relatively closed conceptual world
    - usually more concrete
    - generate theories of individual cases rather than general theory
    - requires lower levels of creativity.
  - There is a practical need for diagnostic systems.
    - Applications pay the bills.
    - Diagnosis has been studied for several years from a computational perspective in AI.
    - Many practical systems actually deployed.

# *Similarities: diagnosis & discovery*

- composite hypotheses (usually)
- need to select a best explanation among alternatives
- evaluate hypotheses by the success and failure of predictions
  - along with other criteria such as simplicity and explanatory power

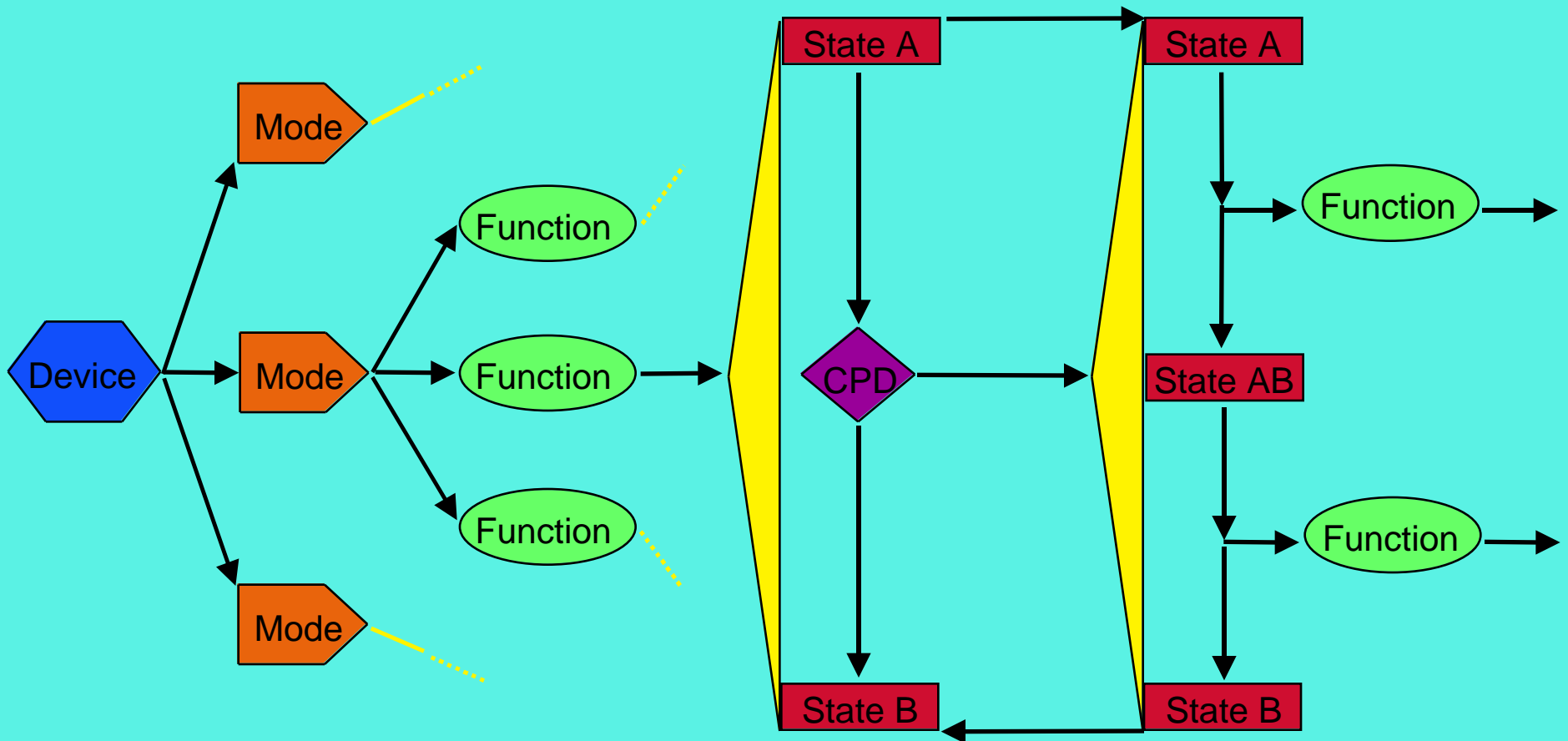
# *Implemented prototype system*

- Details in Hong Wu's Master's thesis:
  - Wu, Hong. *Use of Multi-Purpose Knowledge Database in Simulation and Diagnosis*. The Ohio State University, 1997, (Chemical Engineering).
- To understand the inference strategy, details are not needed about the application domain (manufacturing operations) or about the implementation.

# *Component library*

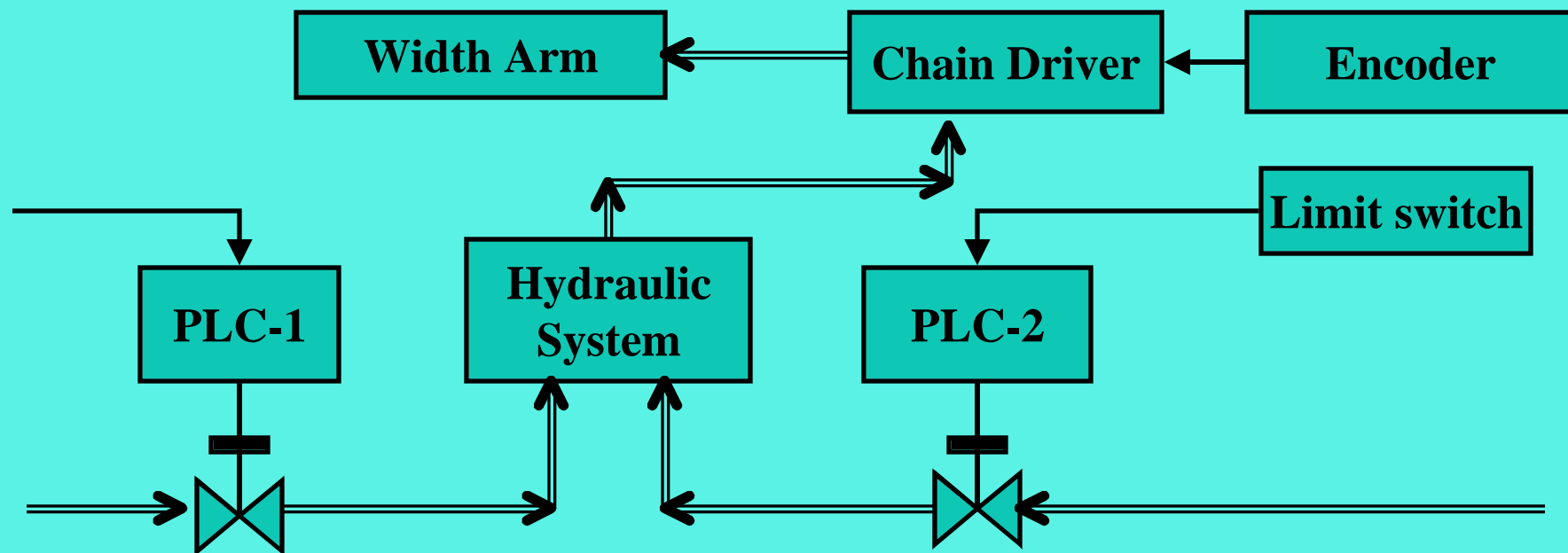
- Devices in a component library are represented as having:
  - input and output **ports**,
  - **modes** (states of a device that effect its behavior)
  - associated with each mode will be one or more **functions** that map inputs to outputs and may change mode.
    - Functions can also typically be read backwards, mapping outputs to possible inputs.
    - Functions will usually specify the time delays that will occur between inputs and corresponding outputs.

# Knowledge Structure



# *Target device*

Represented as a system of components and connections using components from the library



# *Forward and backward simulation*

- Cause-to-effect forward simulation is performed by propagating state descriptions from inputs to outputs across components, and across connections, starting with initial conditions and using functions and assumed modes.
- Similarly, effect-to-cause backward simulation is performed by mapping outputs to inputs, and across connections, considering all possible modes.
- Backward simulation generates branching alternative causal paths.

# *Starting from puzzlement*

- Fault detection, or more generally, detection of deviations from expectations is done by
  - forward simulation from normal or last-updated state,
  - comparing the results in real time with data coming from the target device.
  - significant mismatch implies a deviation from expectations
- A deviation from expectations induces a goal of explaining the deviation.

# *Generating hypotheses*

- Backward simulation is used to generate hypotheses, which are causal stories (causal paths) that potentially explain the deviation from expectation.

# *Removing bad hypotheses*

- As hypotheses are built up by backward simulation, causal paths are cut off, whenever possible, by comparing with device data and eliminating those where mismatch occurs.

# *Evaluating hypotheses*

- Forward simulation from surviving hypotheses is used to produce predictions.
- Predicted observables are matched against data coming from the target device. Mismatching hypotheses are eliminated.

# *Resolving remaining ambiguity by watching and waiting*

- If more than one hypothesis remains, forward simulation is used to continually make predictions from the surviving set of hypotheses
- Predictions are continually compared with new incoming data.
- Mismatching hypotheses are eliminated.

# *Experimental results*

- A small portion of a manufacturing plant was represented.
- A simulated malfunction was used to generate data.
- Detection of deviations of observable values from expected values was performed correctly.
- This triggered backward simulation from the points of deviation, which found 3 possible root causes after cutting off paths that included states that differed from observations.
- Forward simulation from the hypothesized faults and mismatching observations eliminated the 2 incorrect diagnoses, leaving only the correct (simulated) cause. **Right answer!**

# *Interpretation of results*

- The reasoning strategy has thus been demonstrated to work well, at least on one case.
- This constitutes a small-scale validation of the computational strategy for diagnosis.
- It also tends to validate the A-P model of scientific discovery by showing that a version of it can be made precise enough to be implemented and to perform correctly for diagnosis.
- But, the system used a model to generate a causal story to explain the specific case. It didn't generate a new model, as would be needed to reflect theory-forming science.

Discussion?

# *Composable causal-model fragments*

- OSU-CML language under development
  - Ohio State Univ. Compositional Modeling Language
  - Successful alpha testing
- Based on CML (Stanford) and FR (OSU)
- Goal: computer-composable device-situation models
  - supporting multiple types of simulation and analysis.
  - Computer can “playfully” compose models.

# References

- Josephson, J. R., & Josephson, S. G. (Eds.). (1994). *Abductive Inference: Computation, Philosophy, Technology*. New York: Cambridge University Press.
- Wu, H. (1997). *Use of Multi-Purpose Knowledge Database in Simulation and Diagnosis*, Master's Thesis, The Ohio State University.

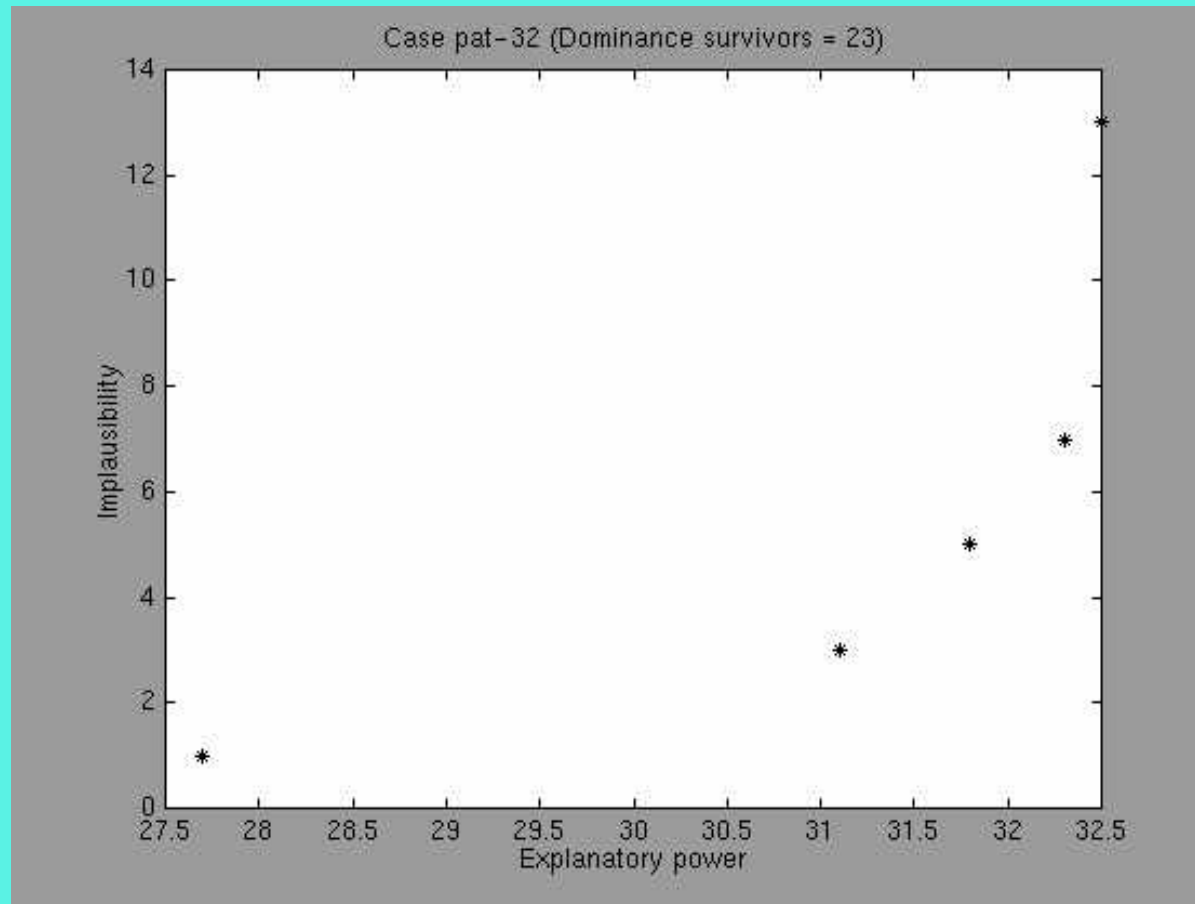
# *Abduction as optimization*

- Multi-criterial optimization
- Experiments using criteria
  - plausibility
  - explanatory power
  - simplicity

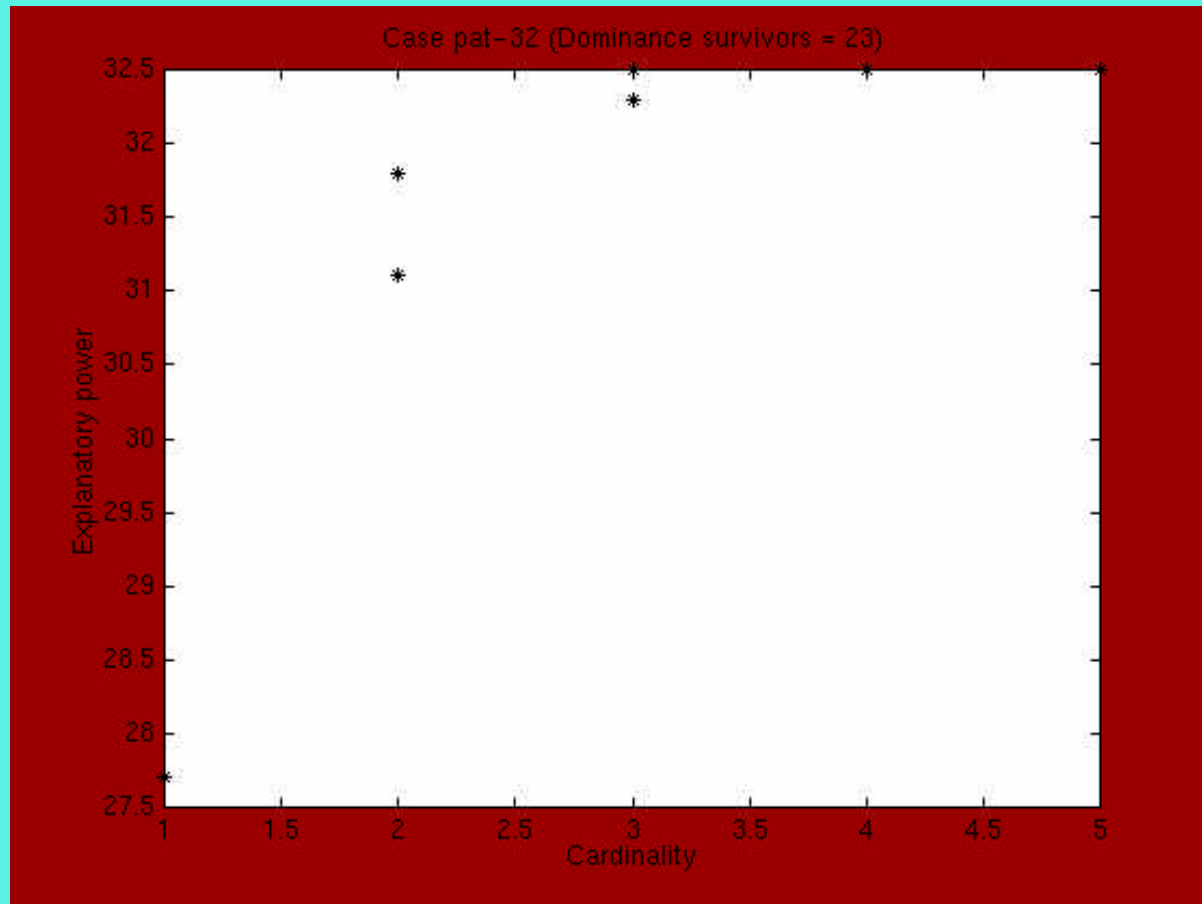
# *Composite Explanatory Hypotheses*

- File case on antibody identification problem
- 65535 composite hypotheses ( $2^{16} - 1$ )
- Dominance filtered using criteria of implausibility, simplicity (subset is simpler), explanatory power.
- Survivors of dominance: 23 !

# *Composite Explanatory Hypotheses*



# *Composite Explanatory Hypotheses*



# *Implications of AI abductive systems*

- Some systems work well even in complicated cases and where evidence is ambiguous.
- Some systems do not follow deductive inference rules nor manipulate numerical probabilities. (See our book.)
- Thus abduction has been formalized (to some degree) beyond the classical formalizations.
  - One dimension of this extension is the computational dimension of **control** of processing

# *Finding the best explanation*

Possible evaluation criteria:

- plausibility
- explanatory power
- simplicity
- specificity
- consistency
- predictive power

# *Best-explanation reasoning*

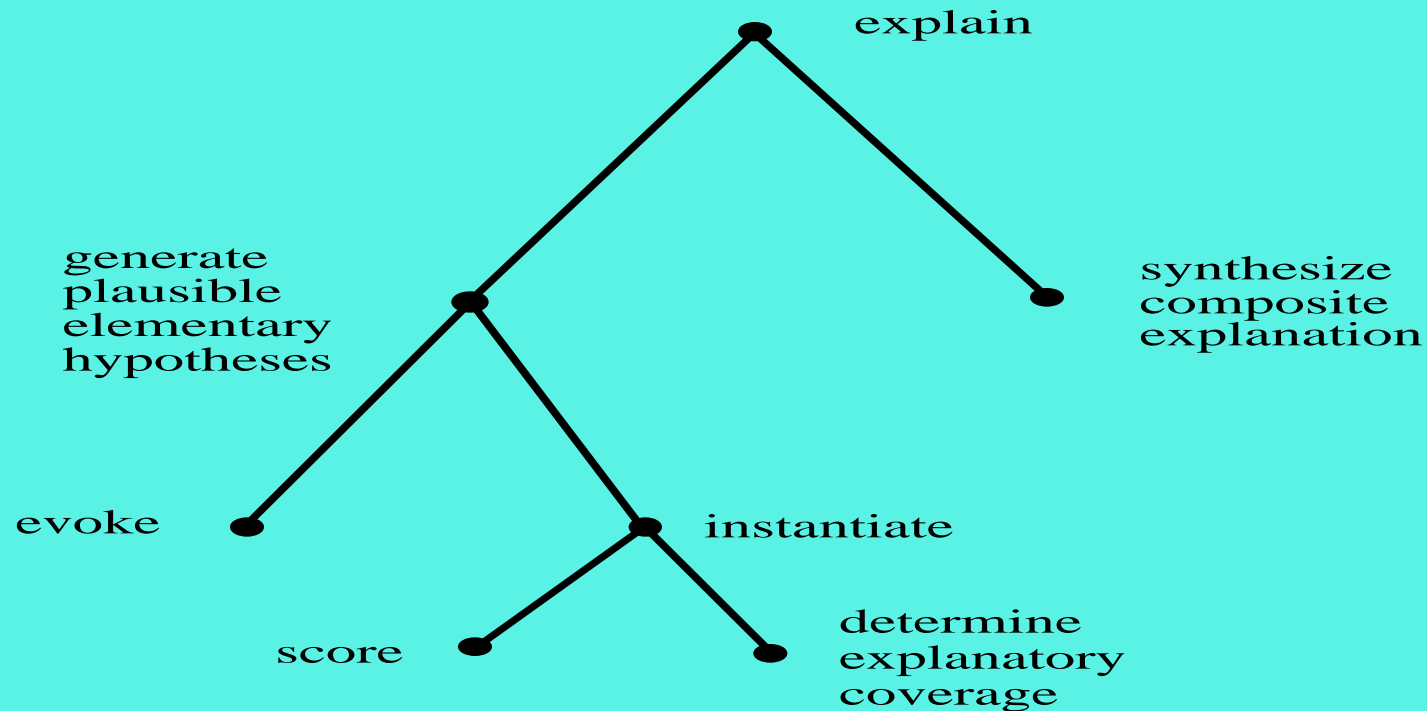
- How can knowledge and control be organized to make it computationally feasible?
- How can an intelligent agent form good composite explanatory hypotheses without getting lost in the large number of potentially applicable concepts, and without getting lost in the numerical vastness of their combinations?

# *Difficulty*

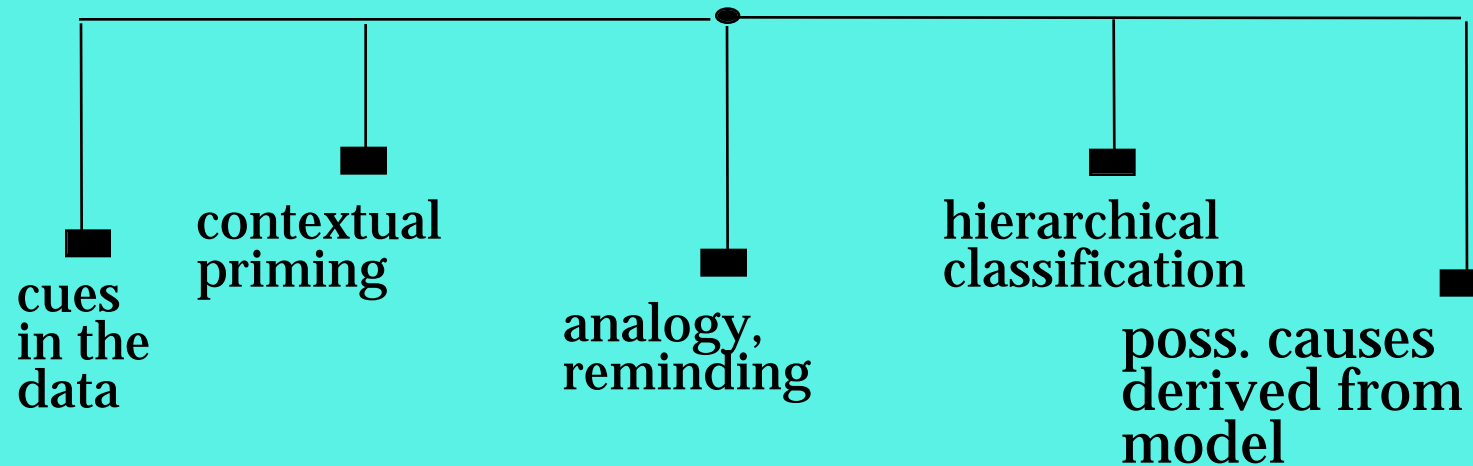
- The problem isn't simply to build the best composite, but to build the best composite *and* compare it with all the other possible composites.
- abductive inference:
  - pervasive
  - prima facie computationally intractable

# *Generating explanations by instantiation and composition*

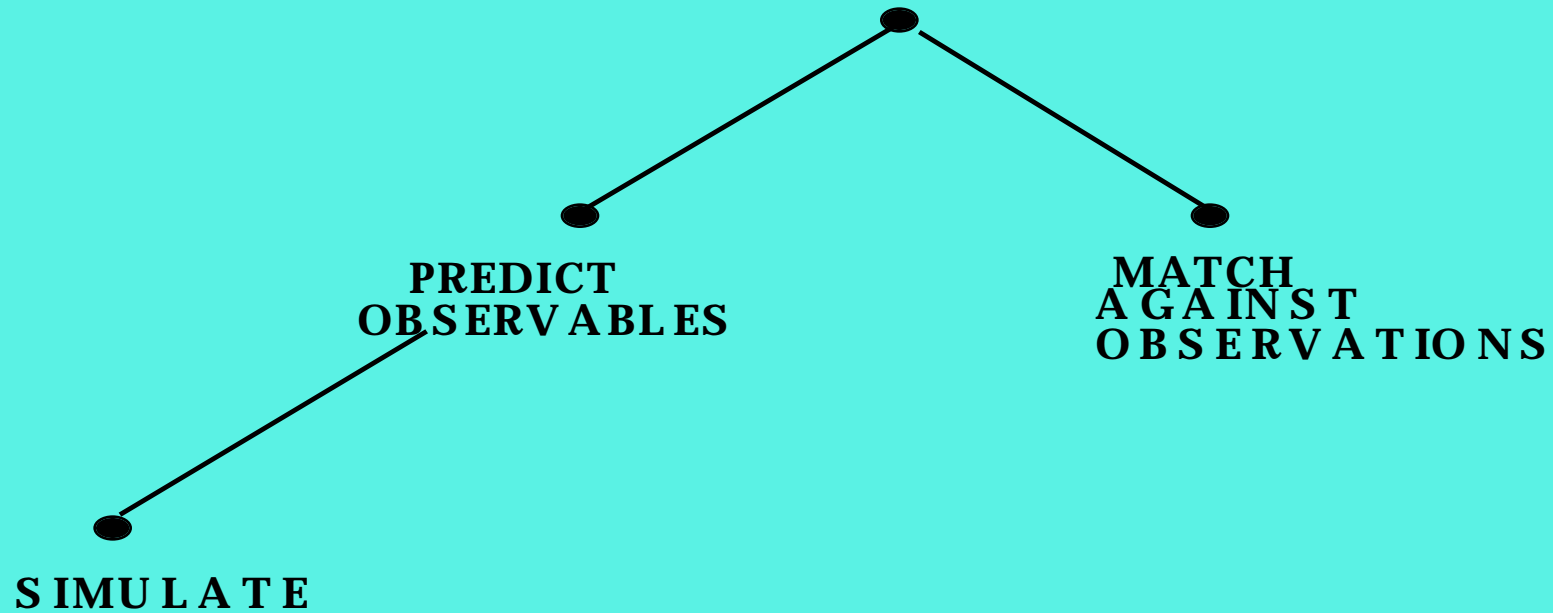
## task-subtask breakdown



# *evolve elementary hypotheses*



# *score hypotheses*



# *Possible hypothesis relationships*

- A and B are mutually incompatible.
- A and B can account for the same data.
- A can account for B.
- A is a refinement of B. (A is more specific.)
- A could be caused by B.
- A implies B.
- A and B have some degree of “sympathy” or “antipathy.”
- A is suggestive of B.
- A and B are mutually compatible and are explanatory alternatives where their explanatory capabilities overlap.

# *Task of hypothesis formation*

Not in general:

- find all possible explanations
- find all good explanations
- find the best complete explanation for the given data

But instead:

- explain as much of the data as can be explained with high confidence; be prepared to guess intelligently beyond that point; and know what to ask to disambiguate further