Goodman's problem

- Why do some hypotheses receive confirmation from examples but not others?
  - "All piece of copper conduct electricity": yes
  - "All men in this room are third sons": no
- Distinguishing *lawlike* hypotheses from *accidental* hypotheses is not easy:
  - "All emeralds are green"
  - "All emeralds are grue", where *grue* means "if observed before *t*, green; else, blue."

Responses to Goodman

- First instinct is a syntactic response:
  - Hypotheses without arbitrary free parameters are more lawlike.
  - Simpler (shorter) hypotheses are more lawlike.

Syntactic levers for induction

- Which hypothesis is better supported by the evidence?
  - "All emeralds are green."
  - "All emeralds are green or less than 1 ft. in diameter."
  - But: "All emeralds are green and less than 1 ft. in diameter"?
- Which curve is best supported by the data?

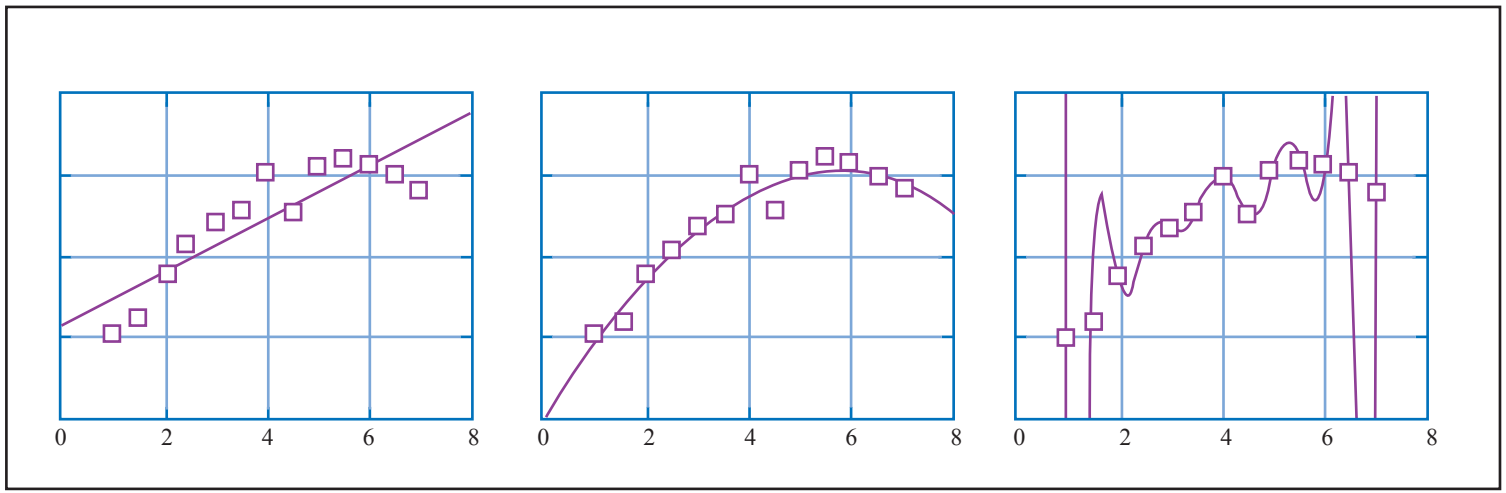


Figure by MIT OCW.

Responses to Goodman

- Hypotheses without arbitrary free parameters are more lawlike.
- Simpler (shorter) hypotheses are more lawlike.
- But “green” and “grue” are logically symmetric:
  - To a Martian who sees *grue* and *bleen*, green just means “if observed before *t*, grue; else, bleen.”

Responses to Goodman

- Hypotheses without arbitrary free parameters are more lawlike.
- Simpler (shorter) hypotheses are more lawlike.
- But “green” and “grue” are logically symmetric.
- *Lawlike* is a semantic (not syntactic) notion, and depends on prior subjective knowledge (not strictly objective world structure).

It's not about time or photoreceptors

- Consider:
  - All emeralds are crystaline.
  - All emeralds are crysquid.
- Crysquid = "if under one foot in diameter, crystaline; else, liquid".
- Liqualine = "if under one foot in diameter, liquid; else, crystaline".
- Then crystaline = "if under one foot in diameter, crysquid; else, liqualine".

The origin of good hypotheses

- Nativism
  - Plato, Kant
  - Chomsky, Fodor
- Empiricism
  - Strong: Watson, Skinner
  - Weak: Bruner, cognitive psychology, statistical machine learning
- Constructivism
  - Goodman, Piaget, Carey, Gopnik
  - AI threads….

Plato

- *Meno*: Where does our knowledge of abstract concepts (e.g., virtue, geometry) come from?
- The puzzle: “A man cannot enquire about that which he does not know, for he does not know the very subject about which he is to enquire.”

Plato

- *Meno*: Where does our knowledge of abstract concepts (e.g., virtue, geometry) come from?
- A theory: Learning as “recollection”.
- The Talmud’s version:

“Before we are born, while in our mother's womb, the Almighty sends an angel to sit beside us and teach us all the wisdom we will ever need to know about living. Then, just before we are born, the angel taps us under the nose (forming the philtrum, the indentation that everyone has under their nose), and we forget everything the angel taught us.”

Plato meets Matlab^{tm}

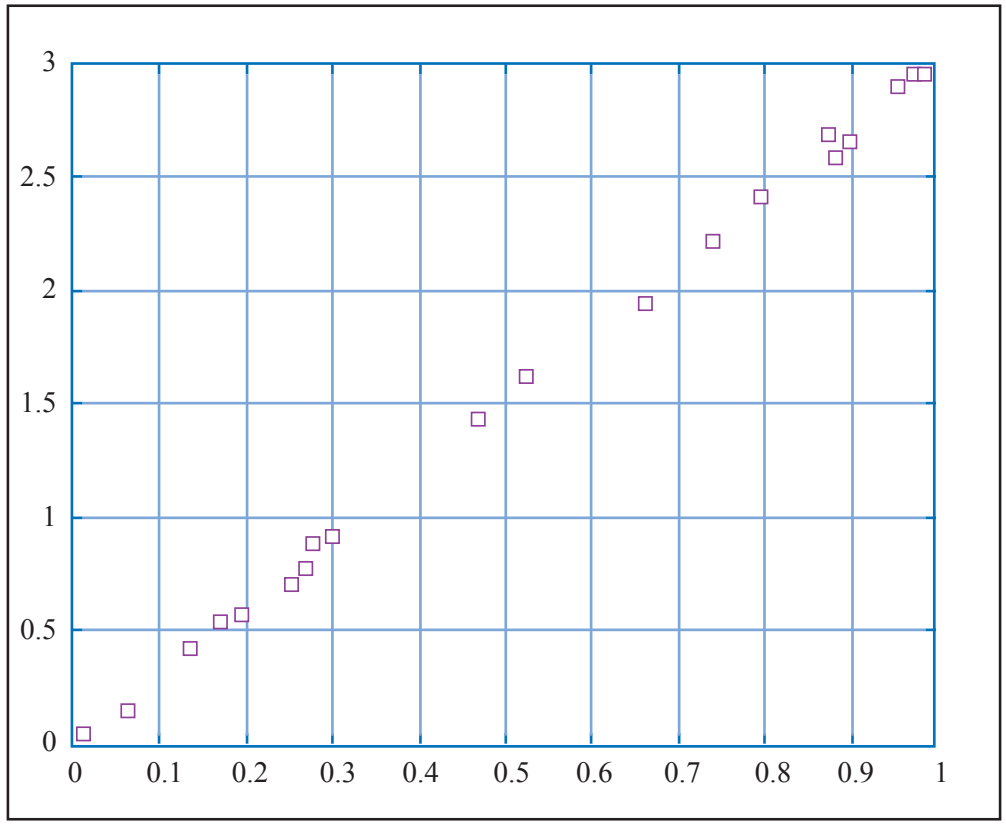


Figure by MIT OCW.

What is the relation between y and x?

Plato meets Matlab^{tm}

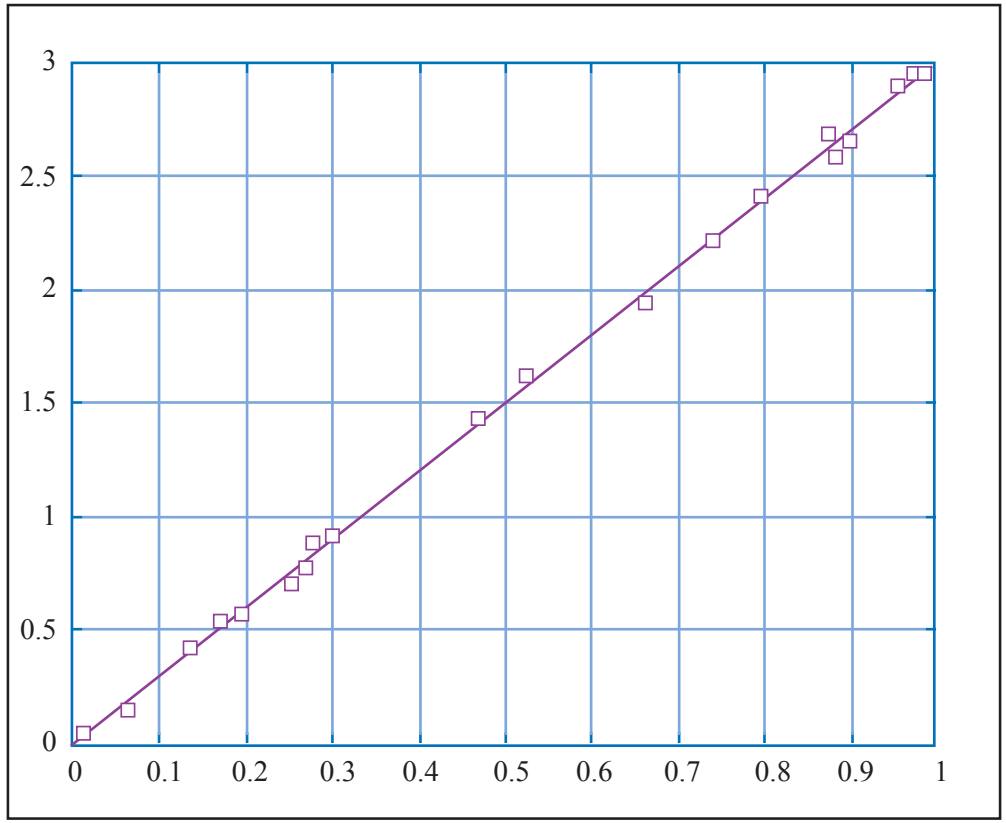


Figure by MIT OCW.

What is the relation between y and x?

Plato meets Matlab^{tm}

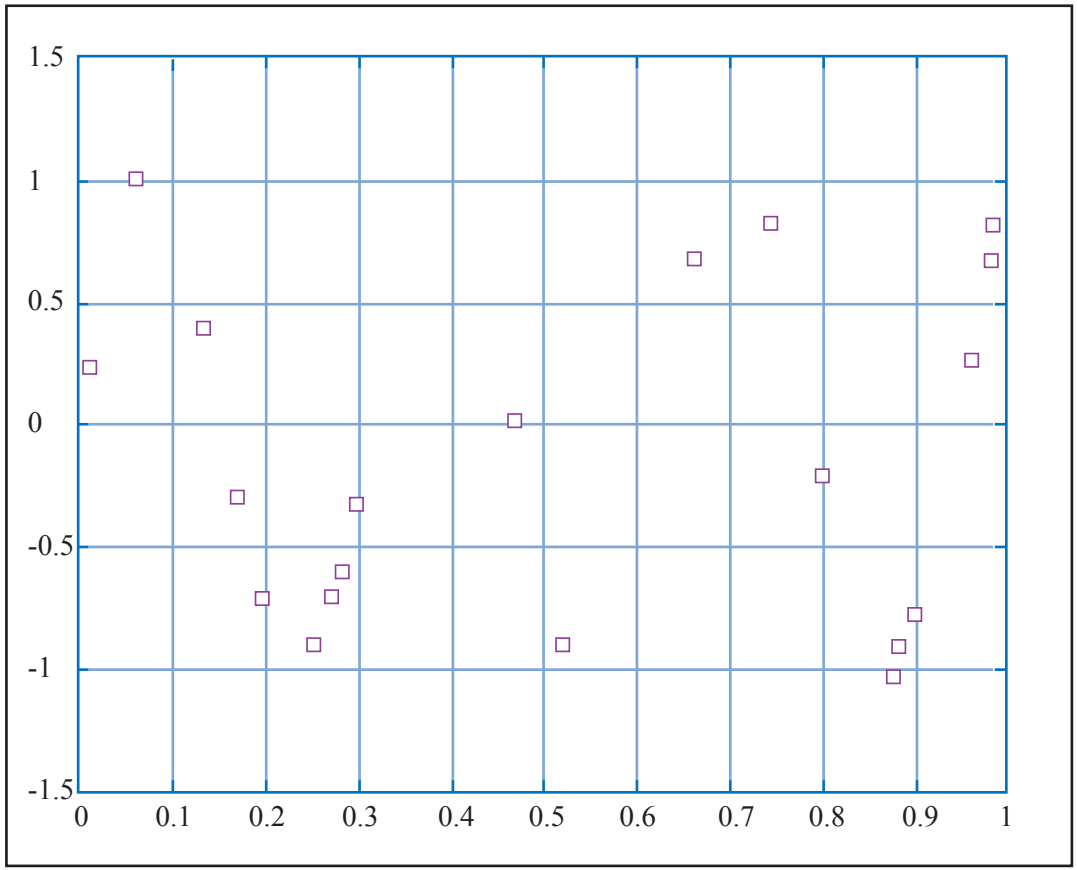


Figure by MIT OCW.

What is the relation between y and x?

Plato meets Matlab^{tm}

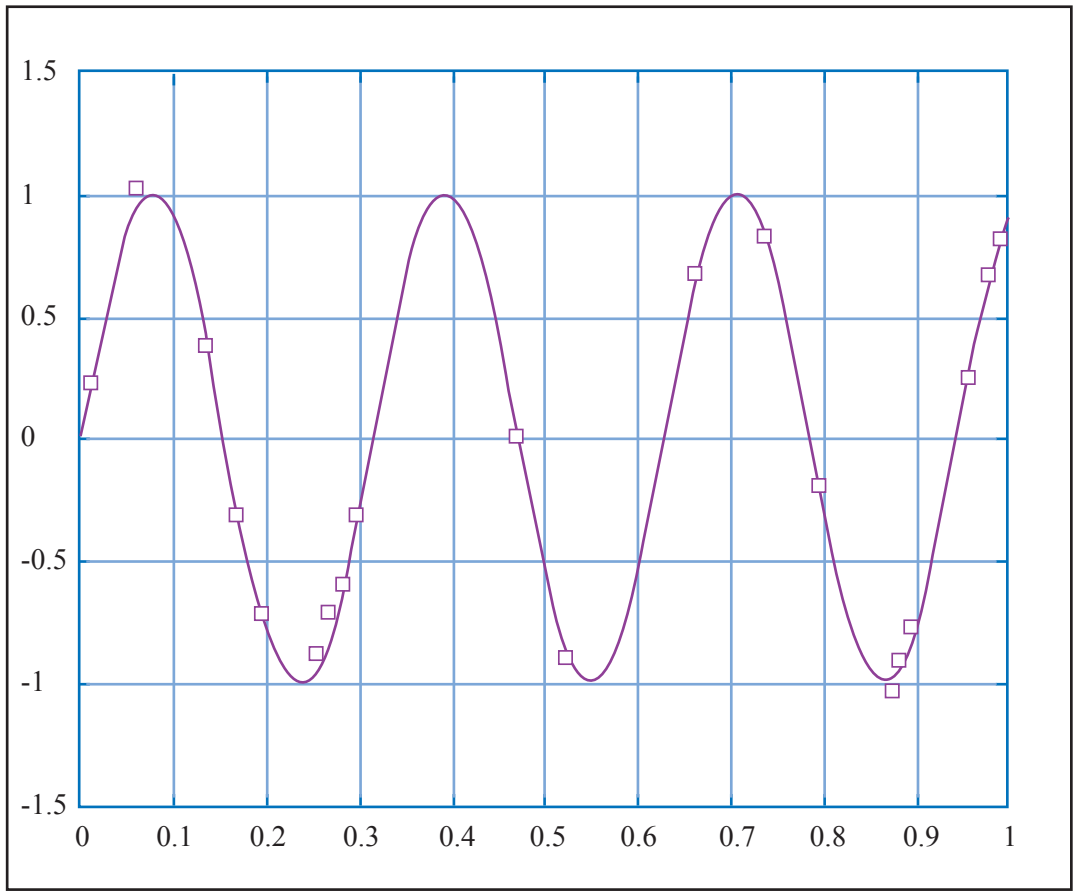


Figure by MIT OCW.

What is the relation between y and x?

The legacy of Plato

- "A man cannot enquire about that which he does not know, for he does not know the very subject about which he is to enquire."
- We can't learn abstractions from data if in some sense we didn't already know what to look for.
  - Chomsky's "poverty of the stimulus" argument for the innateness of language.
  - Fodor's argument for the innateness of all concepts.

The origin of good hypotheses

- Nativism
  - Plato, Kant
  - Chomsky, Fodor
- Empiricists
  - Strong: Watson, Skinner
  - Weak: Bruner, cognitive psychology, statistical machine learning
- Constructivists
  - Goodman, Piaget, Carey, Gopnik
  - AI threads….

Image removed due to copyright considerations. Please see:
Bruner, Jerome S., Jacqueline J. Goodnow, and George Austin. *A Study in Thinking*. Somerset, NJ: Transaction Publishers, 1986. ISBN: 0887386563.

Image removed due to copyright considerations. Please see:
Hull. "Qualitative aspects of the evolution of concepts." *Psychological Monograph* 28, no. 123 (1920).

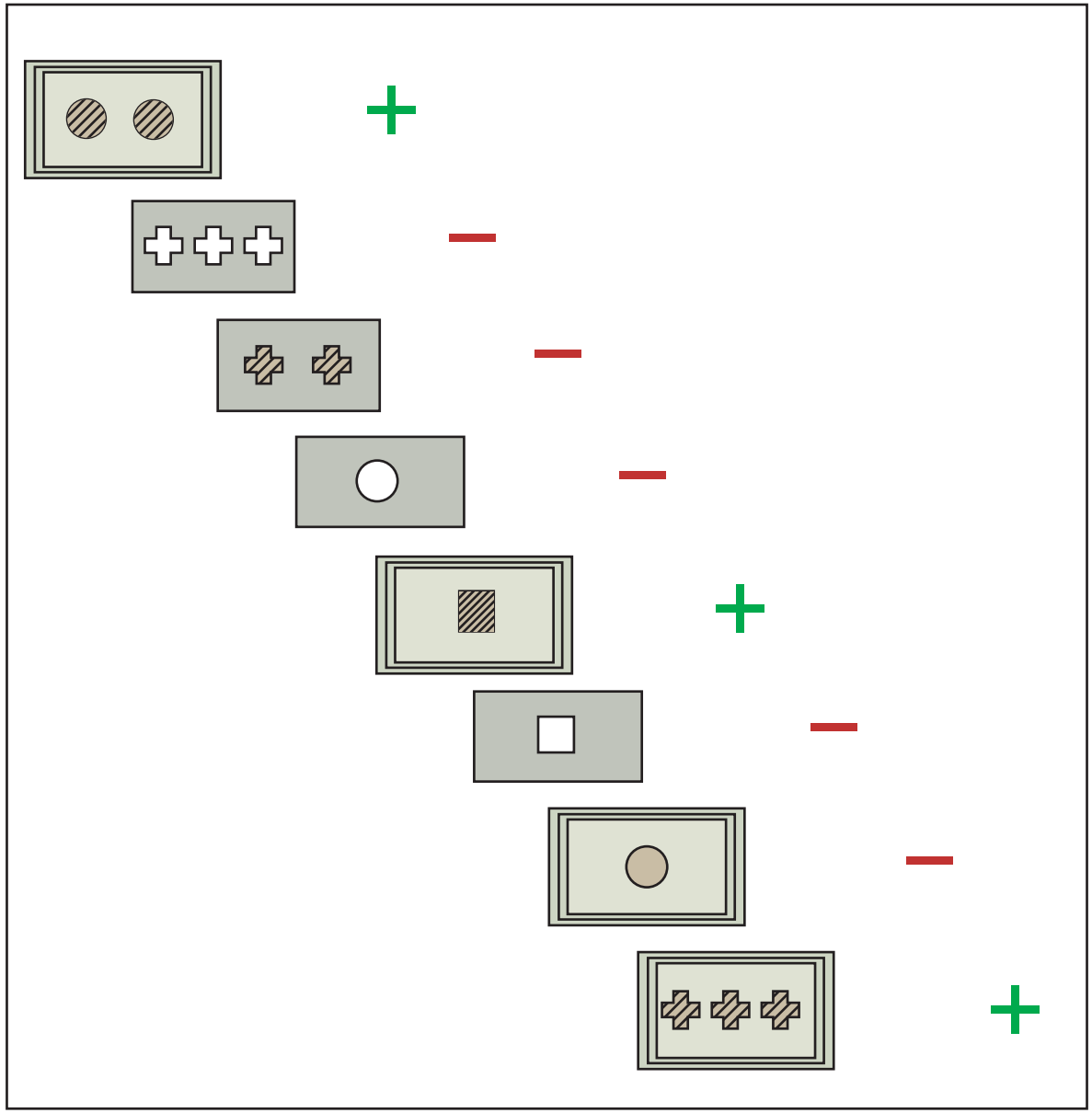
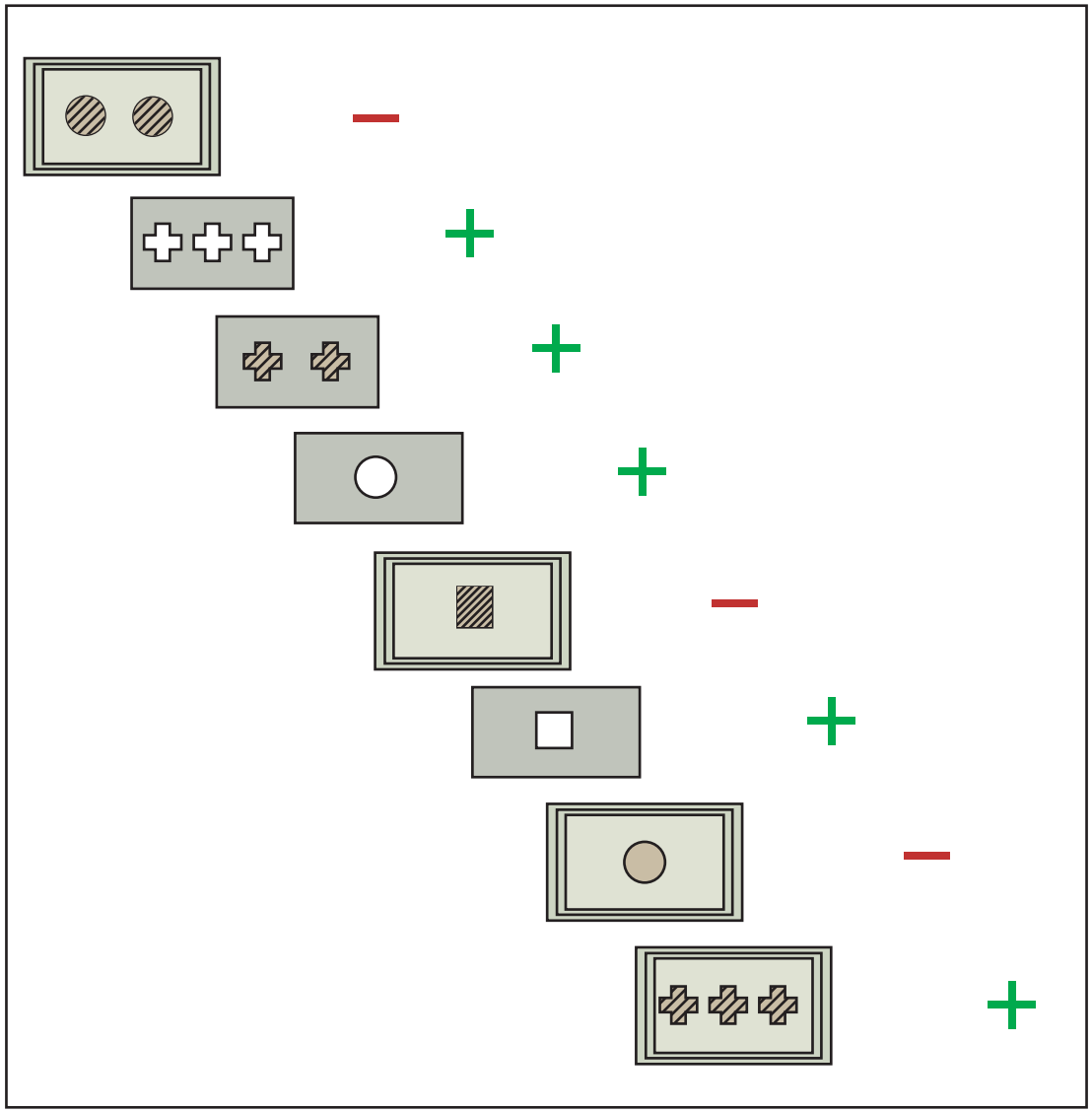


Figure by MIT OCW.

"striped and three borders": *conjunctive* concept



"black or plusses": *disjunctive* concept

Figure by MIT OCW.

Fodor's critique

- This isn't really *concept learning*, it's just *belief fixation*.
  - To learn the rule "striped and three borders", the learner must already have the concepts "striped", "three borders", and "and", and the capacity to put these components together.
  - In other words, the learner already has the concept, and is just forming a new belief about how to respond on this particular task.
- More generally, all inductive learning seems to require the constraints of a hypothesis space -- so the learner must begin life with all the concepts they will ever learn. How depressing.

Fodor's critique

Raises major questions for cognitive development, machine learning, and AI:

- Is it ever possible to learn *truly new* concepts, which were not part of your hypothesis space to begin with?
- What conceptual resources must be innate?
  - Objects?
  - First-order logic?
  - Recursion?
  - Causality?

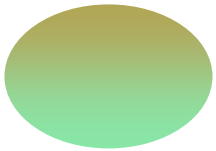
The origin of good hypotheses

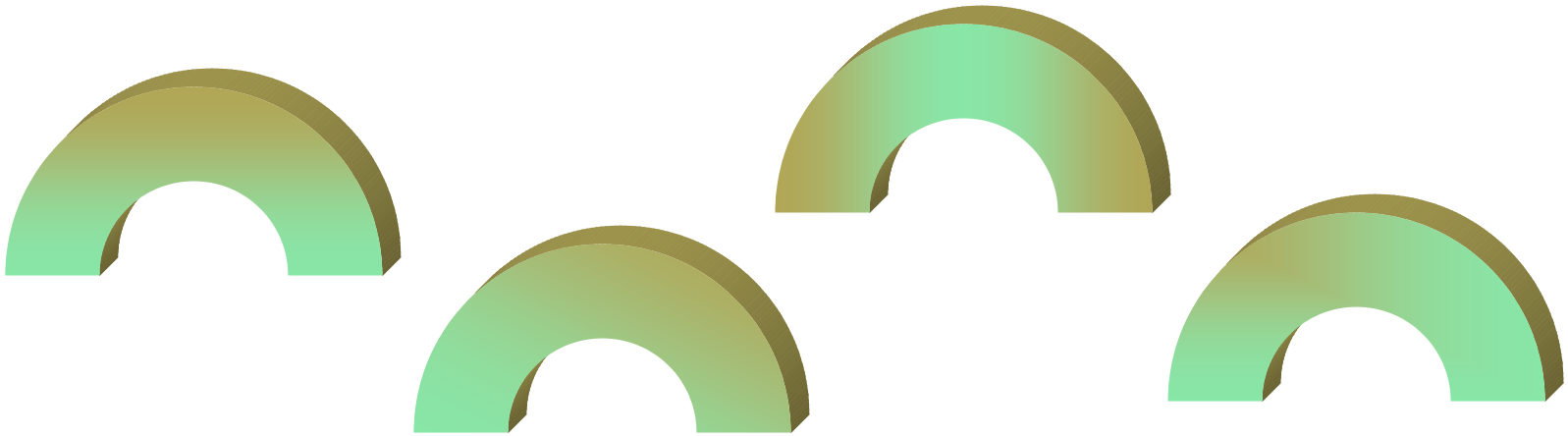
- Nativism
  - Plato, Kant
  - Chomsky, Fodor
- Empiricists
  - Strong: Watson, Skinner
  - Weak: Bruner, cognitive psychology, statistical machine learning
- Constructivists
  - Goodman, Piaget, Carey, Gopnik
  - AI threads….

Goodman's answer to Goodman

- More lawlike hypotheses are based on "entrenched" predicates: *green* is more entrenched than *grue*.
- How does a predicate become entrenched? Is it simple statistics: how often the predicate has supported successful inductions in the past?
- Suppose *grue* means "If observed on Earth, green; if on Mars, blue."
- Entrenchment could come through experience, but could also derive from a causal *theory*. Theory supported by experience seems best.

How do theories work?

- See this look?  It's called "chromium".
- Here are some blickets:



- Which hypothesis is more lawlike?
  - "All blickets are chromium"
  - "All blickets are chromirose", where *chromirose* means "if observed before *t*, chromium; else rose-colored."

How do theories work?

- Theories depend on abstract categories.
  - E.g., *chromium* is a kind of color or material.
  - Emeralds are a kind of mineral.
- Abstract categories depend on theories.
  - E.g., atom, magnetic pole
- Theories support hypotheses for completely novel situations.
- Big open questions:
  - What is a theory, formally?
  - How are theories learned?

Inductive learning as search

Marr's three levels

- Level 1: Computational theory
  - What is the goal of the computation, and what is the logic by which it is carried out?
- Level 2: Representation and algorithm
  - How is information represented and processed to achieve the computational goal?
- Level 3: Hardware implementation
  - How is the computation realized in physical or biological hardware?

- Level 1: Computational theory
  - What is a concept?
  - What does it mean to learn a concept successfully?
  - What kinds of concepts can be learned? Under what conditions? From how much data?
  - What assumptions must be made for learning to succeed?
- Level 2: Representation and algorithm
  - How are objects and concepts represented?
  - How much memory (space) and computation (time) does a learning algorithm require?
  - Is the algorithm online or batch (serial or parallel)?
  - What kinds of concepts are learned most easily (or most reliably) by a particular algorithm?

Level 1: Computational Theory

- What is a concept?
  - A rule that divides the objects into two sets: positive instances and negative instances.
- What does it mean to learn a concept successfully?
  - Given *N* randomly chosen "training" examples (objects labeled positive or negative), and a space *H* of hypotheses (candidate rules), find a hypothesis *h* that is consistent with the training examples and that ….
  - **Identification in the limit:** … is guaranteed to converge to the true rule if in *H*, in the limit that *N* goes to infinity.
  - **Probably Approximately Correct (PAC):** …is likely to perform well on future examples, for a reasonable number of training examples *N*. May be possible even if the true rule is "unrealizable" (not in *H*).

Level 2: Representation and Algorithm

- How are objects and concepts represented?
  - E.g., Binary-feature worlds and conjunctive concepts

Features

	f_1	f_2	f_3	f_4	f_5	f_6	
Objects							
x_1:	1	1	1	0	1	0	
x_2:	1	1	0	0	1	1	
x_3:	0	1	0	1	1	1	
...							
Concepts							
h_1:	*	1	*	*	1	*	= f_2 AND f_5
h_2:	1	1	*	*	1	*	= f_1 AND f_2 AND f_5
h_3:	1	1	1	0	1	0	= f_1 AND f_2 AND f_3 AND -f_4 AND f_5 AND -f_6

Level 2: Representation and Algorithm

- A learning algorithm: subset principle (a/k/a focussing, wholist, find-S)
  - Start with the most specific conjunction (= first positive example) and drop features that are inconsistent with additional positive examples.

Examples

	f_1	f_2	f_3	f_4	f_5	f_6	Label
x_1:	1	1	1	0	1	0	+

Current hypothesis

	f_1	f_2	f_3	f_4	f_5	f_6
h_1:	1	1	1	0	1	0

Level 2: Representation and Algorithm

- A learning algorithm: subset principle (a/k/a focussing, wholist, find-S)
  - Start with the most specific conjunction (= first positive example) and drop features that are inconsistent with additional positive examples.

Examples

	f_1	f_2	f_3	f_4	f_5	f_6	Label
x_1:	1	1	1	0	1	0	+
x_2:	0	0	0	0	0	1	-

Current hypothesis

	f_1	f_2	f_3	f_4	f_5	f_6
h_1:	1	1	1	0	1	0
h_2:	1	1	1	0	1	0

Level 2: Representation and Algorithm

- A learning algorithm: subset principle (a/k/a focussing, wholist, find-S)
  - Start with the most specific conjunction (= first positive example) and drop features that are inconsistent with additional positive examples.

Examples

	f_1	f_2	f_3	f_4	f_5	f_6	Label
x_1:	1	1	1	0	1	0	+
x_2:	0	0	0	0	0	1	-
x_3:	1	1	0	0	1	1	+

Current hypothesis

	f_1	f_2	f_3	f_4	f_5	f_6
h_1:	1	1	1	0	1	0
h_2:	1	1	1	0	1	0

Level 2: Representation and Algorithm

- A learning algorithm: subset principle (a/k/a focussing, wholist, find-S)
  - Start with the most specific conjunction (= first positive example) and drop features that are inconsistent with additional positive examples.

Examples

	f_1	f_2	f_3	f_4	f_5	f_6	Label
x_1:	1	1	1	0	1	0	+
x_2:	0	0	0	0	0	1	-
x_3:	1	1	0	0	1	1	+

Current hypothesis

	f_1	f_2	f_3	f_4	f_5	f_6
h_1:	1	1	1	0	1	0
h_2:	1	1	1	0	1	0
h_3:	1	1	*	0	1	*

Level 2: Representation and Algorithm

- A learning algorithm: subset principle (a/k/a focussing, wholist, find-S)
  - Start with the most specific conjunction (= first positive example) and drop features that are inconsistent with additional positive examples.

Examples

	f_1	f_2	f_3	f_4	f_5	f_6	Label
x_1:	1	1	1	0	1	0	+
x_2:	0	0	0	0	0	1	-
x_3:	1	1	0	0	1	1	+
x_4:	1	1	1	1	1	0	-

Current hypothesis

	f_1	f_2	f_3	f_4	f_5	f_6
h_1:	1	1	1	0	1	0
h_2:	1	1	1	0	1	0
h_3:	1	1	*	0	1	*
h_4:	1	1	*	0	1	*

Level 2: Representation and Algorithm

- A learning algorithm: subset principle (a/k/a focussing, wholist, find-S)
  - Start with the most specific conjunction (= first positive example) and drop features that are inconsistent with additional positive examples.

Examples

	f_1	f_2	f_3	f_4	f_5	f_6	Label
x_1:	1	1	1	0	1	0	+
x_2:	0	0	0	0	0	1	-
x_3:	1	1	0	0	1	1	+
x_4:	1	1	1	1	1	0	-
x_5:	0	1	0	1	1	1	+

Current hypothesis

	f_1	f_2	f_3	f_4	f_5	f_6
h_1:	1	1	1	0	1	0
h_2:	1	1	1	0	1	0
h_3:	1	1	*	0	1	*
h_4:	1	1	*	0	1	*

Level 2: Representation and Algorithm

- A learning algorithm: subset principle (a/k/a focussing, wholist, find-S)
  - Start with the most specific conjunction (= first positive example) and drop features that are inconsistent with additional positive examples.

Examples

	f_1	f_2	f_3	f_4	f_5	f_6	Label
x_1:	1	1	1	0	1	0	+
x_2:	0	0	0	0	0	1	-
x_3:	1	1	0	0	1	1	+
x_4:	1	1	1	1	1	0	-
x_5:	0	1	0	1	1	1	+

Current hypothesis

	f_1	f_2	f_3	f_4	f_5	f_6
h_1:	1	1	1	0	1	0
h_2:	1	1	1	0	1	0
h_3:	1	1	*	0	1	*
h_4:	1	1	*	0	1	*
h_5:	*	1	*	*	1	*

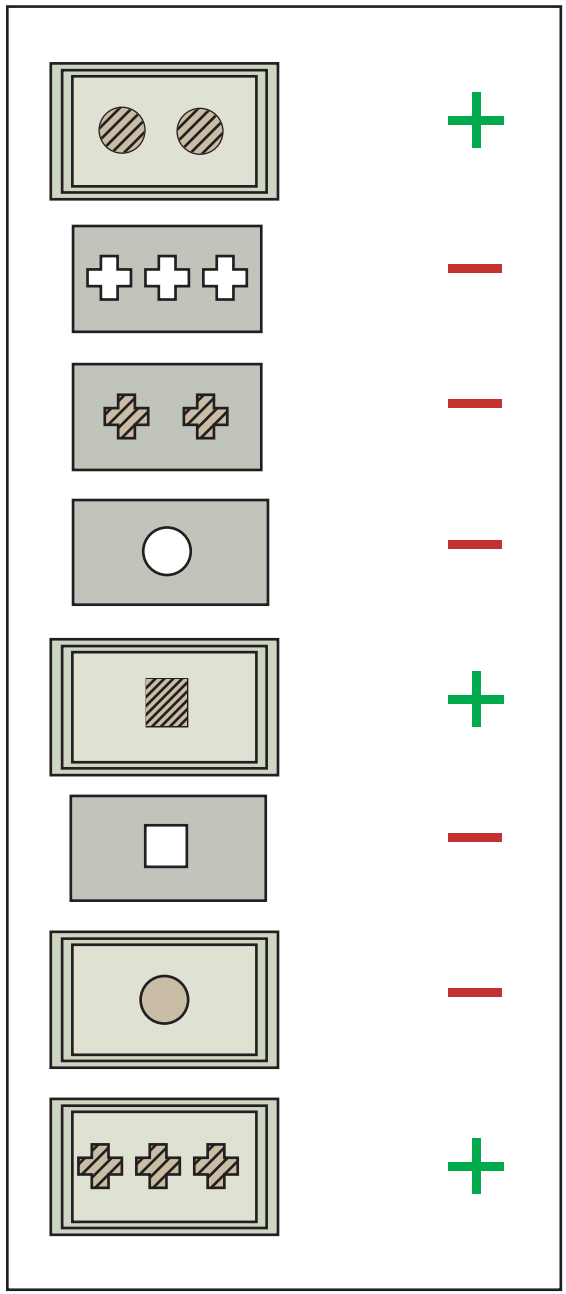


Figure by MIT OCW.

2 striped circles with 3 borders

. . .

. . .

. . .

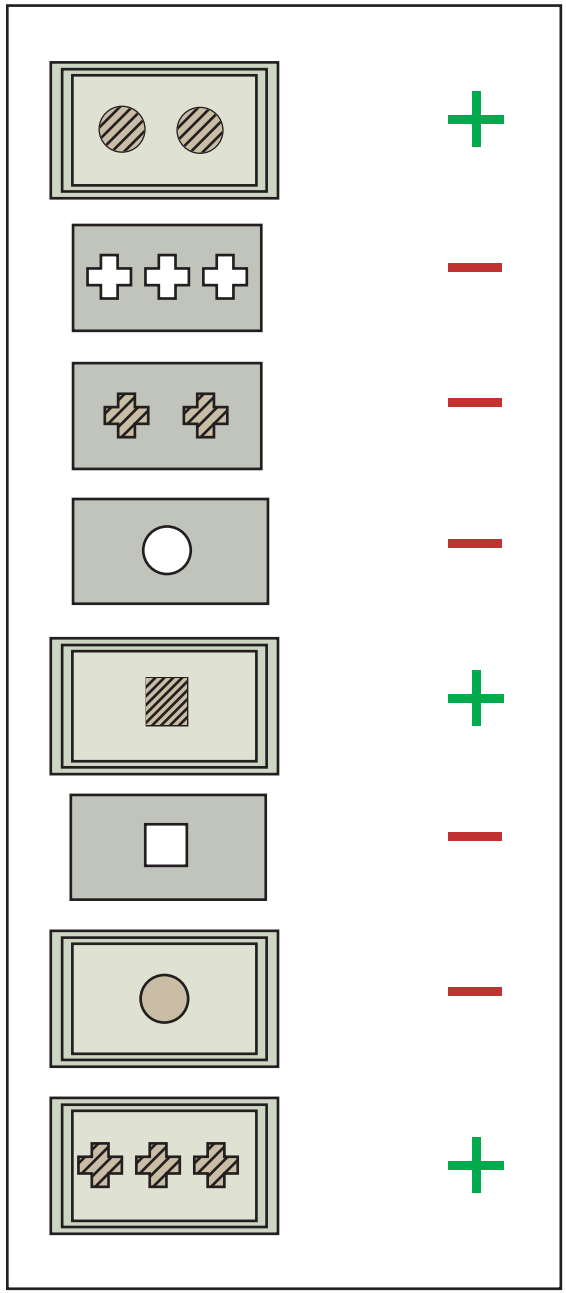


Figure by MIT OCW.

2 striped circles with 3 borders

. . .

. . .

. . .

~~2~~ striped ci~~rc~~les with 3 borders

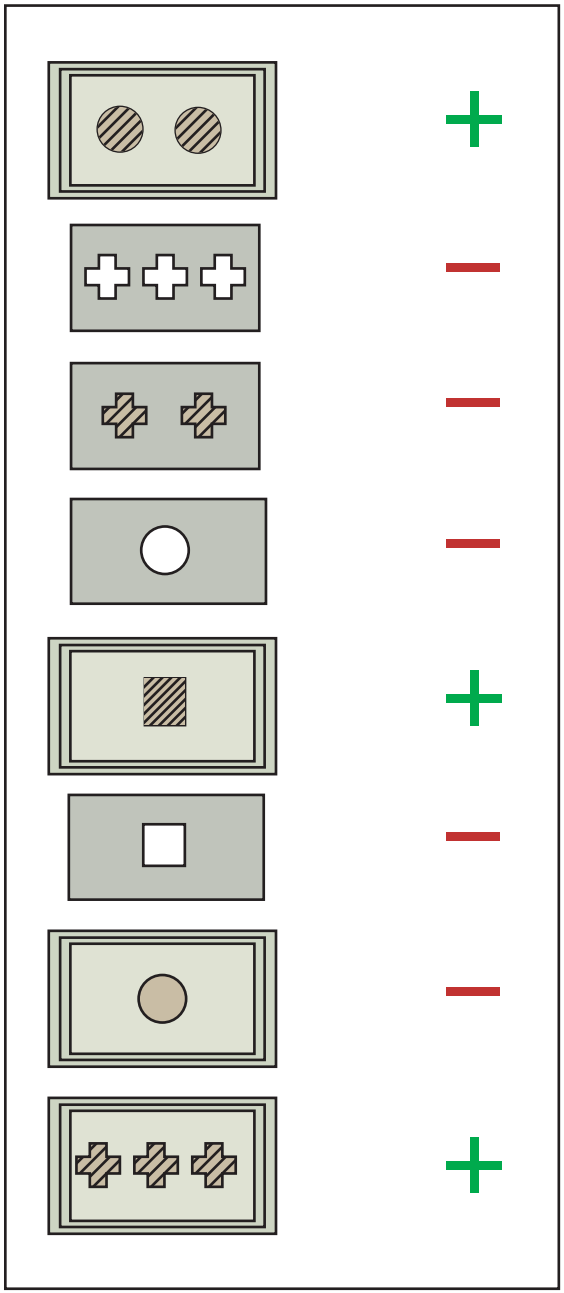


Figure by MIT OCW.

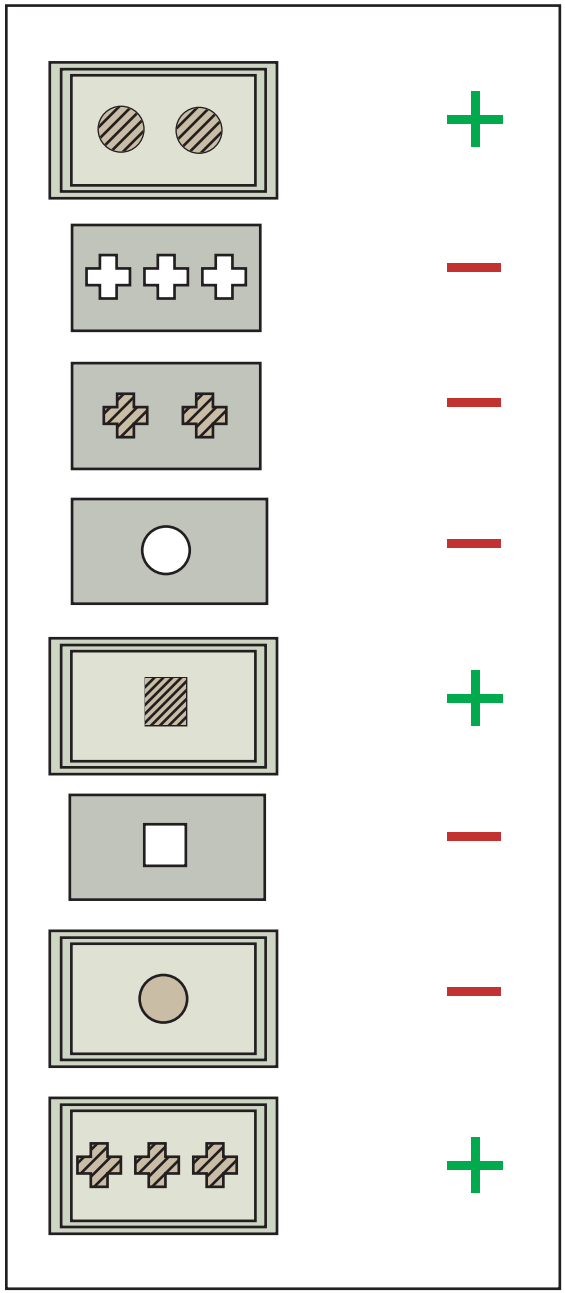
2 striped circles with 3 borders

. . .

. . .

. . .

Striped with 3 borders



2 striped circles with 3 borders

. . .

. . .

. . .

Striped with 3 borders

. . .

. . .

Striped with 3 borders

Figure by MIT OCW.

Computational analysis

- Identification in the limit:
  - Given N randomly chosen training examples and a space H of hypotheses, find a hypothesis h that is consistent with the training examples and that is guaranteed to converge to the true rule if it is in H, in the limit $N \to \infty$.
  - With k features, can make at most k mistakes on positive examples (and none on negatives).
  - Assuming that every example has some probability of occurring, success is certain.
  - Note that only positive examples are required.

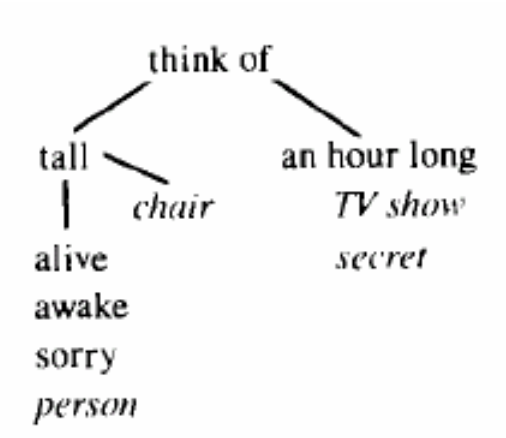
Relevance to human learning

- Bruner, Goodnow and Austin
  - Most people use this strategy in a transparent conjunctive learning task.

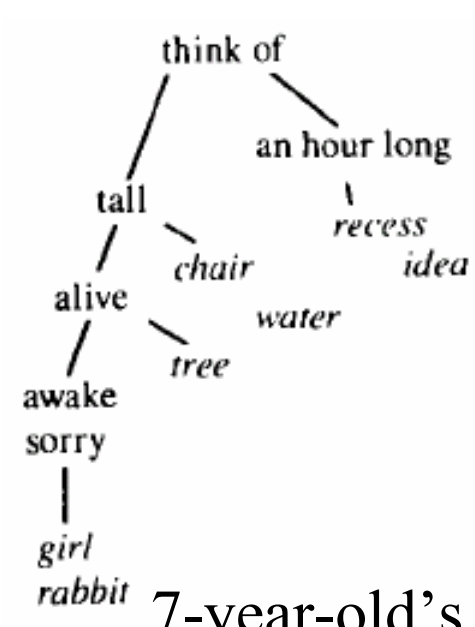
Image removed due to copyright considerations.

Relevance to human learning

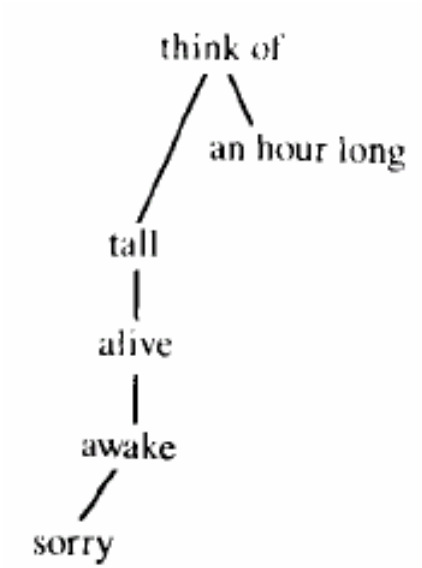
- Berwick
  - Explains development of conceptual hierarchies and syntactic rules.



5-year-old's ontology



7-year-old's ontology



Overly general ontology

Keil, Frank C. *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press, 1989.

Relevance to human learning

- Berwick
  - Explains development of conceptual hierarchies and syntactic rules.

E.g., learning which verbs have optional arguments.

John ate the ice cream cone.

John ate.

John took the ice cream cone.

*John took.

Subset strategy: assume that any verb appearing with an argument *must* take an argument, until it has been observed without an argument.

Computational analysis

- Can learning succeed under weaker conditions? PAC.
  - The true concept is not in the hypothesis space.
  - We are not willing to wait for infinite examples, but we can live with a low error rate.
- What can we say about more complex cases of learning?
  - Richer hypothesis spaces?
  - More powerful learning algorithms?

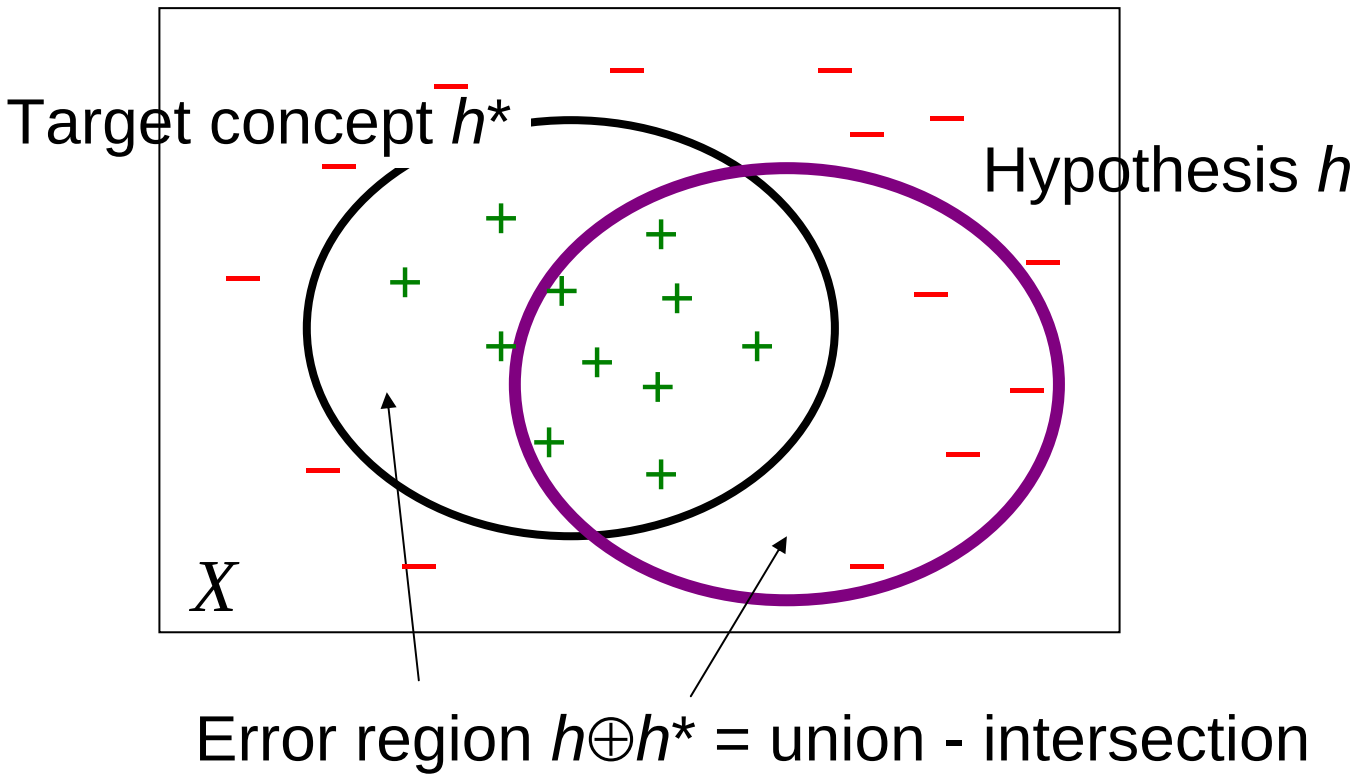
Probably Approximately Correct (PAC)

- The intuition: Want to be confident that a hypothesis which looks good on the training examples (i.e., appears to have zero error rate) in fact has a low true error rate (ε), and thus will generalize well on the test instances.

- Note we do not require that the true rule is in the hypothesis space, or that if it is, we must find it. We are willing to live with a low but nonzero error rate, as long as we can be pretty sure that it is low.

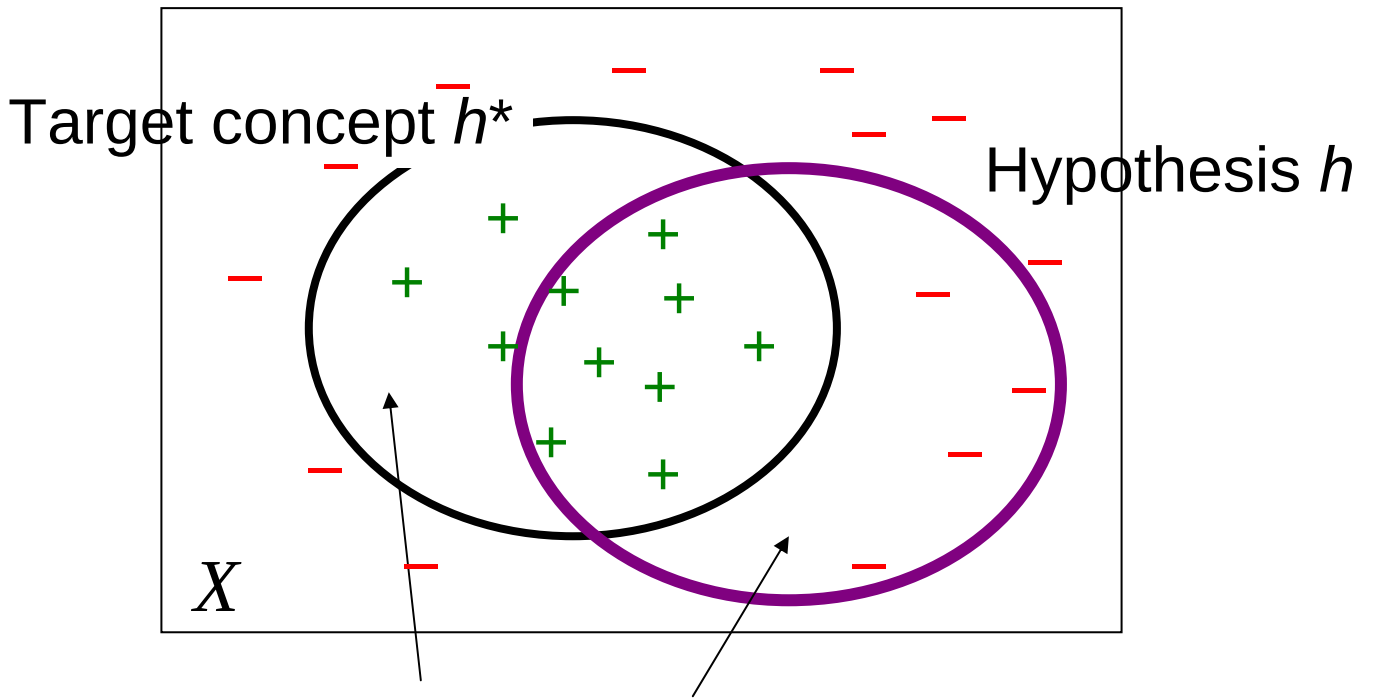
Probably Approximately Correct (PAC)

- Assumption of "uniformity of nature":
  - Training and test instances drawn from some fixed probability distribution on the space *X*.



Probably Approximately Correct (PAC)

- Assumption of "uniformity of nature":
  - Training and test instances drawn from some fixed probability distribution on the space *X*.



Probably Approximately Correct (PAC)

- The intuition: Want to be confident that a hypothesis which looks good on the training examples (i.e., appears to have zero error rate) in fact has a low true error rate (ε), and thus will generalize well on the test instances.

- PAC theorem: With probability 1-δ, a hypothesis consistent with N training examples will have true error rate at most ε whenever

$$N \geq \frac{1}{\varepsilon}\left(\log|H| + \frac{1}{\delta}\right).$$

Probably Approximately Correct (PAC)

- PAC theorem: With probability 1-δ, a hypothesis consistent with *N* training examples will have true error rate at most ε whenever

$$N \geq \frac{1}{\varepsilon}\left(\log|H| + \frac{1}{\delta}\right).$$

- How does *N*, the amount of data required for good generalization, change with problem parameters?
  - As allowable error (ε) decreases, *N* increases.
  - As desired confidence (1-δ) increases, *N* increases.
  - As the size of the hypothesis space (log |*H*|) increases, *N* increases.

Implications for what makes a good hypothesis space or inductive bias.

Probably Approximately Correct (PAC)

$$N \geq \frac{1}{\varepsilon}\left(\log|H| + \frac{1}{\delta}\right).$$

Why does N depend on number of hypotheses, $|H|$?

- Consider the set of "bad" hypotheses, H_{bad}: hypotheses with true error rate greater than or equal to ε.
- We want to be confident that a hypothesis which looks good on N examples is not actually in H_{bad}.
- Each example on average rules out at least a fraction ε of the hypotheses in H_{bad}. The bigger H_{bad} is, the more examples we need to see to be confident that all bad hypotheses have been eliminated.
- The learner doesn't know how big H_{bad} is, but can use $|H|$ as an upper bound on $|H_{\text{bad}}|$.

Demo?

PAC analyses of other hypothesis spaces

- Single features
- Conjunctions
- Disjunctions
- Conjunctions plus *k* exceptions
- Disjunction of *k* conjunctive concepts
- All logically possible Boolean concepts

- Single features: h_1: f_2
  h_2: f_5
- Conjunctions: h_1: f_2 AND f_5
  h_2: f_1 AND f_2 AND f_5
- Disjunctions: h_1: f_2 OR f_5
  h_2: f_1 OR f_2 OR f_5
- Conjunctions plus *k* exceptions:

h_1: (f_1 AND f_2) OR (0 1 0 1 1 0)
h_2: (f_1 AND f_2 AND f_5) OR (0 1 0 1 1 0) OR (1 1 0 0 0 0)

- Disjunction of *k* conjunctive concepts:

h_1: (f_1 AND f_2 AND f_5) OR (f_1 AND f_4)
h_2: (f_1 AND f_2) OR (f_1 AND f_4) OR (f_3)

- All logically possible Boolean concepts:

h_1: (1 1 1 0 0 0), (1 1 1 0 0 1), (1 1 1 0 1 0), ...
h_2: (0 1 0 1 1 0), (1 1 0 0 0 0), (1 0 0 1 1 1), ...

- Single features: h_1: f_2
  h_2: f_5
- Conjunctions: h_1: f_2 AND f_5
  h_2: f_1 AND f_2 AND f_5
- Disjunctions: h_1: f_2 OR f_5
  h_2: f_1 OR f_2 OR f_5
- Conjunctions plus *k* exceptions:

h_1: (f_1 AND f_2) OR (0 1 0 1 1 0)
h_2: (f_1 AND f_2 AND f_5) OR (0 1 0 1 1 0) OR (1 1 0 0 0 0)

- Disjunction of *k* conjunctive concepts:

h_1: (f_1 AND f_2 AND f_5) OR (f_1 AND f_4)
h_2: (f_1 AND f_2) OR (f_1 AND f_4) OR (f_3)

- All logically possible Boolean concepts:

h_1: (1 1 1 0 0 0), (1 1 1 0 0 1), (1 1 1 0 1 0), ...
h_2: (0 1 0 1 1 0), (1 1 0 0 0 0), (1 0 0 1 1 1), ...

- Single features:

  $\log |H| = \log k$ (k = # features)

- Conjunctions:

  $\log |H| = k$

- Disjunctions:

  $\log |H| = k$

$$N \geq \frac{1}{\varepsilon}\left(\log|H| + \frac{1}{\delta}\right).$$

- Conjunctions plus m exceptions:

  $\log |H| \sim km$

- Disjunction of m conjunctive concepts:

  $\log |H| \sim km$

- All logically possible Boolean concepts:

  $\log |H| = 2\text{^}k$ = number of objects in world.

The role of inductive bias

- Inductive bias = constraints on hypotheses.
- Learning with no bias (i.e., H = all possible Boolean concepts) is impossible.
  - PAC result
  - A simpler argument by induction.

The role of inductive bias

- Inductive bias = constraints on hypotheses.
- Relation to Ockham's razor: $N \geq \frac{1}{\varepsilon}\left(\log|H| + \frac{1}{\delta}\right)$.
  - "Given two hypotheses that are both consistent with the data, choose the simpler one."
  - log $|H|$ = number of bits needed to specify each hypothesis h in H. Simpler hypotheses have fewer alternatives, and shorter descriptions.
  - E.g. Avoid disjunctions unless necessary:

    "All emeralds are green and less than 1 ft. in diameter" vs. "All emeralds are green and less than 1 ft. in diameter, or made of cheese".

What this doesn't tell us

- Why conjunctions easier than disjunctions?
  - C.f., Why is "all emeralds are green and less than 1 ft. in diameter" better than "all emeralds are green or less than 1 ft. in diameter"?
  - What are concepts useful for in the real world?
  - What is structure of natural categories?

What this doesn't tell us

- Why conjunctions easier than disjunctions?
- How we choose the appropriate generality of a concept, given one or a few examples?
  - Subset principle says to choose a hypothesis that is as small as possible.
  - Occam's razor says to choose a hypothesis that is as simple as possible.
  - But these are often in conflict, e.g. with conjunctions. People usually choose something in between, particularly with just one example. Consider word learning ….

What this doesn't tell us

- Why conjunctions easier than disjunctions?
- How we choose the appropriate generality of a concept?
- How we should (or do) handle uncertainty?
  - How confident that we have the correct concept?
  - When to stop learning?
  - What would the best example to look at next?
  - What about noise (so that we cannot just look for a consistent hypothesis)?
  - Should we maintain multiple hypotheses? How?

What this doesn't tell us

Compare PAC bounds with typical performance in Bruner's experiments or the real world.

- E.g., need > 200 examples to have 95% confidence that error is < 10%
- Bruner experiments: 5-7 examples
- Children learning words:

Images removed due to copyright considerations.

Other learning algorithms

- Current-best-hypothesis search

Image removed due to copyright considerations.

- Version spaces:

Image removed due to copyright considerations.

Summary: Inductive learning as search

- Rigorous analyses of learnability.
  - Explains when and why learning can work.
  - Shows clearly the need for inductive bias and gives a formal basis for Occam's razor.
- Many open questions for computational models of human learning, and building more human-like machine learning systems.
  - Where do the hypothesis spaces come from? Why are some kinds of concepts more natural than others?
  - How do we handle uncertainty in learning?
  - How do we learn quickly and reliably with very flexible hypothesis spaces?