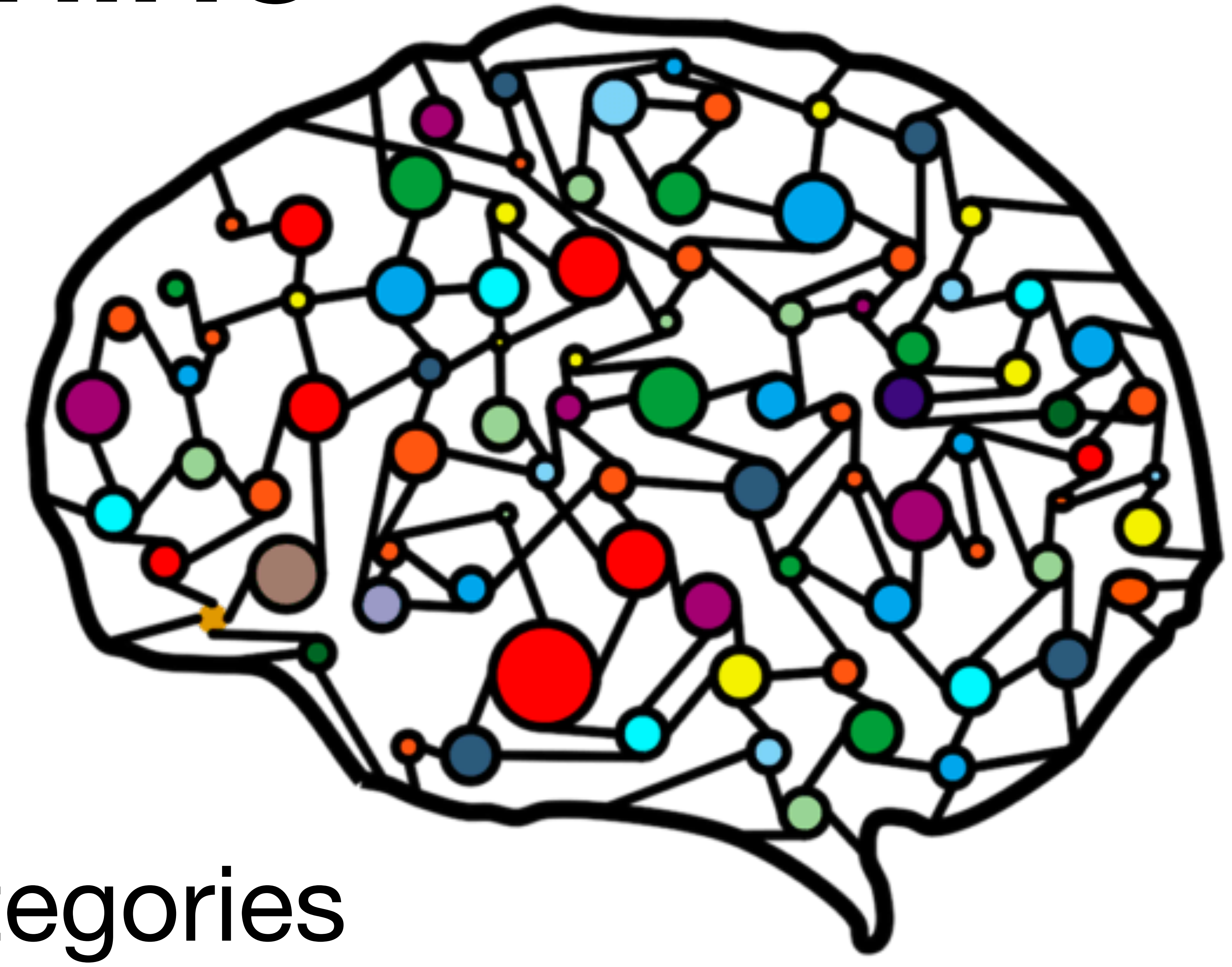
General Principles of Human and Machine Learning



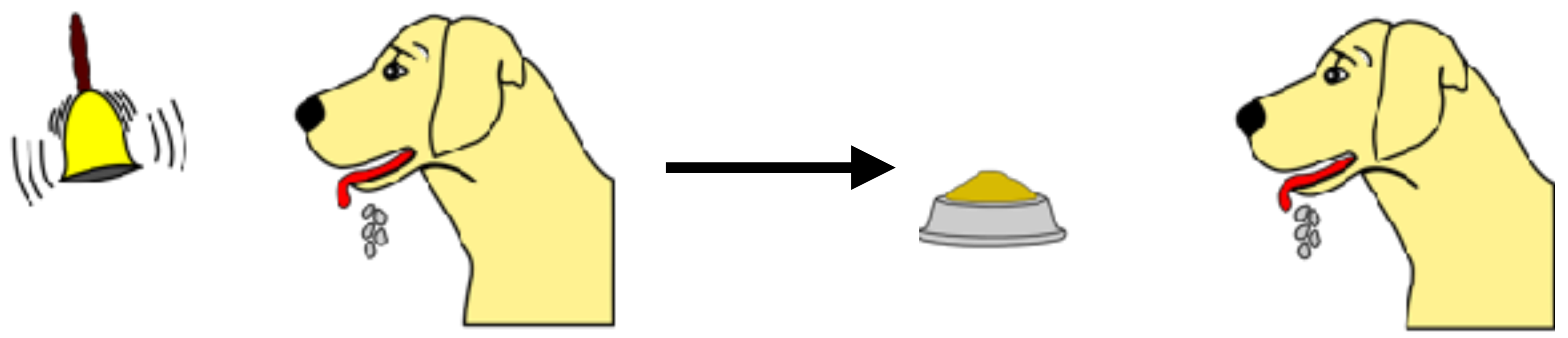
Lecture 6: Concepts and Categories

Dr. Charley Wu

https://hmc-lab.com/GPHML.html

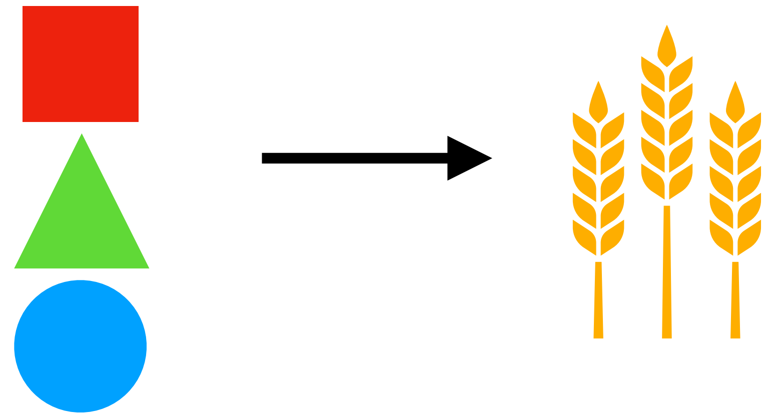
The story so far…

Pavlovian (classical) conditioning



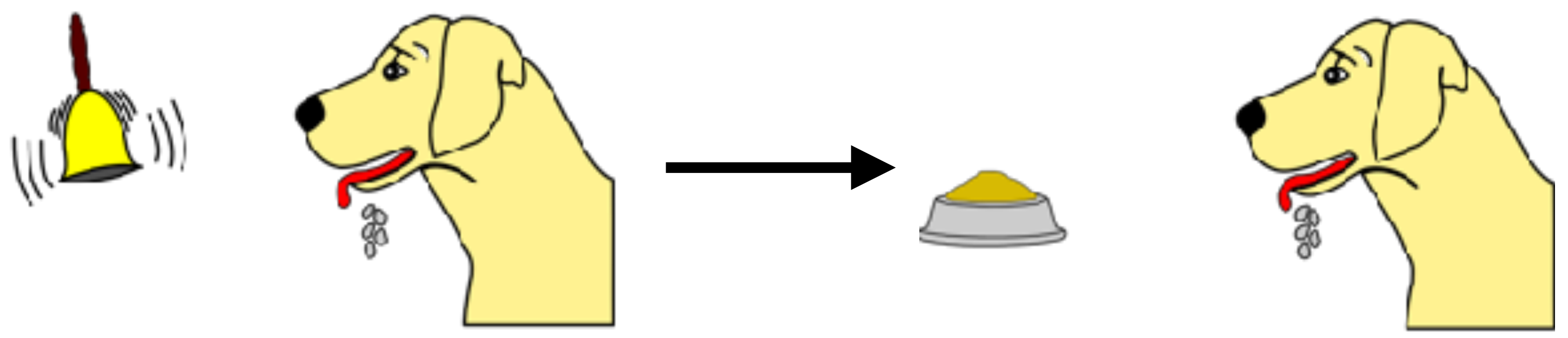
Learn which environmental cues *predict* reward

Operant (instrumental) conditioning

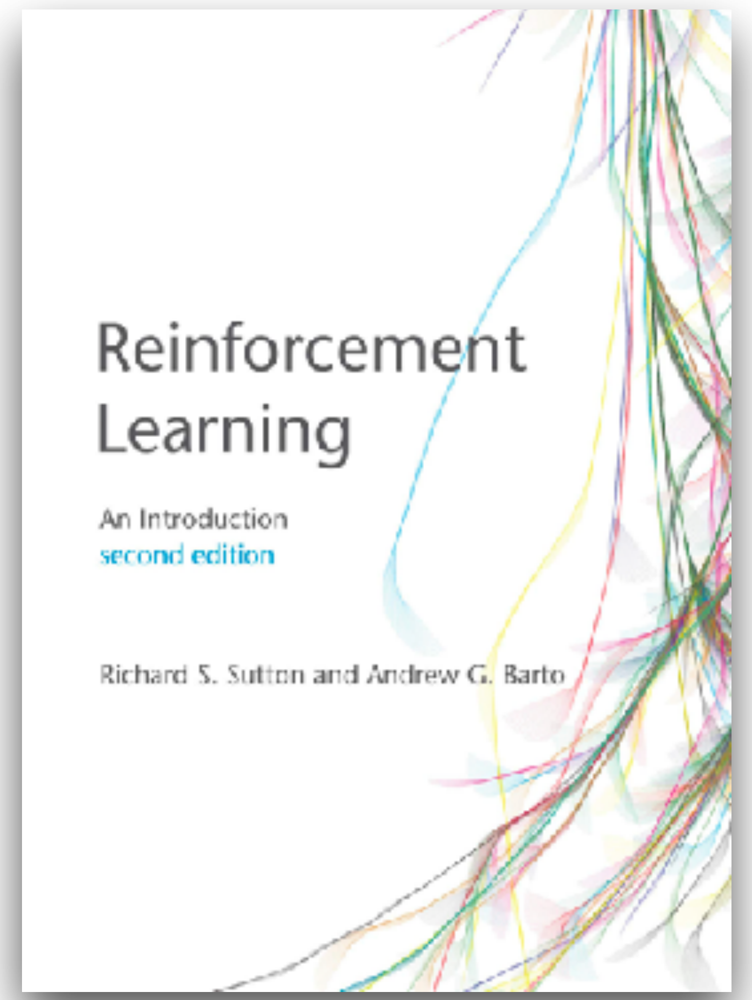


Learn which actions *predict* reward

Pavlovian (classical)
conditioning

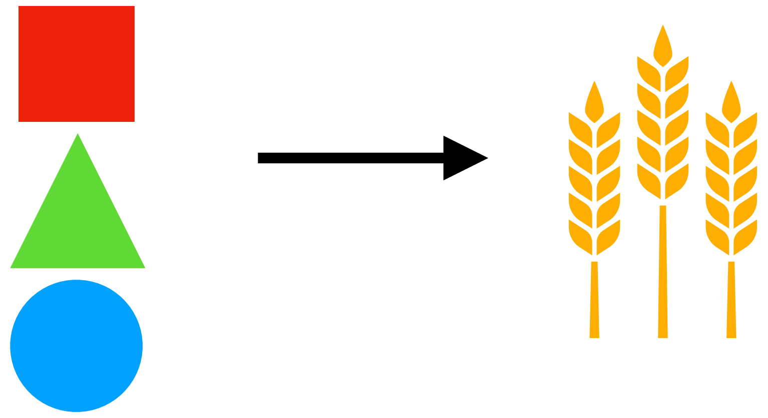


Learn which environmental cues *predict* reward



Reinforcement Learning

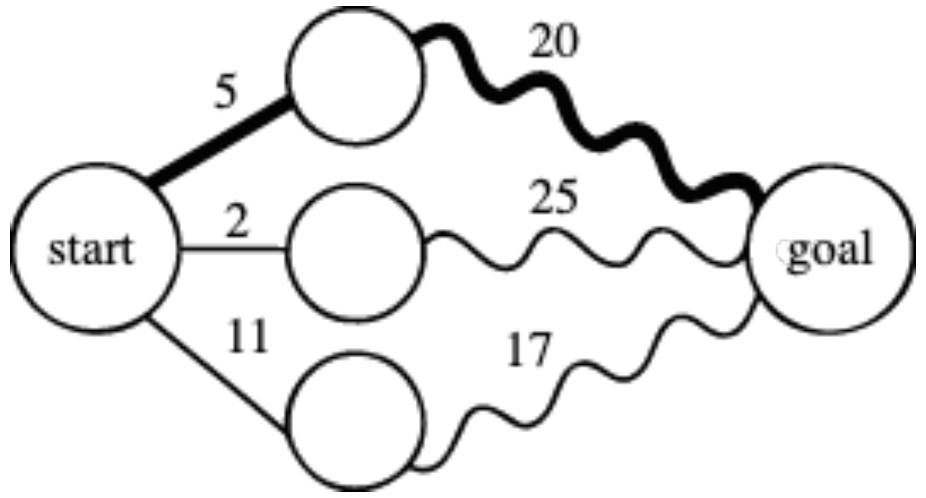
Operant (instrumental)
conditioning



Learn which actions *predict* reward

Neuro-dynamic programing
Bertsekas & Tsitsiklis (1996)

Stochastic approximations to dynamic programing problems



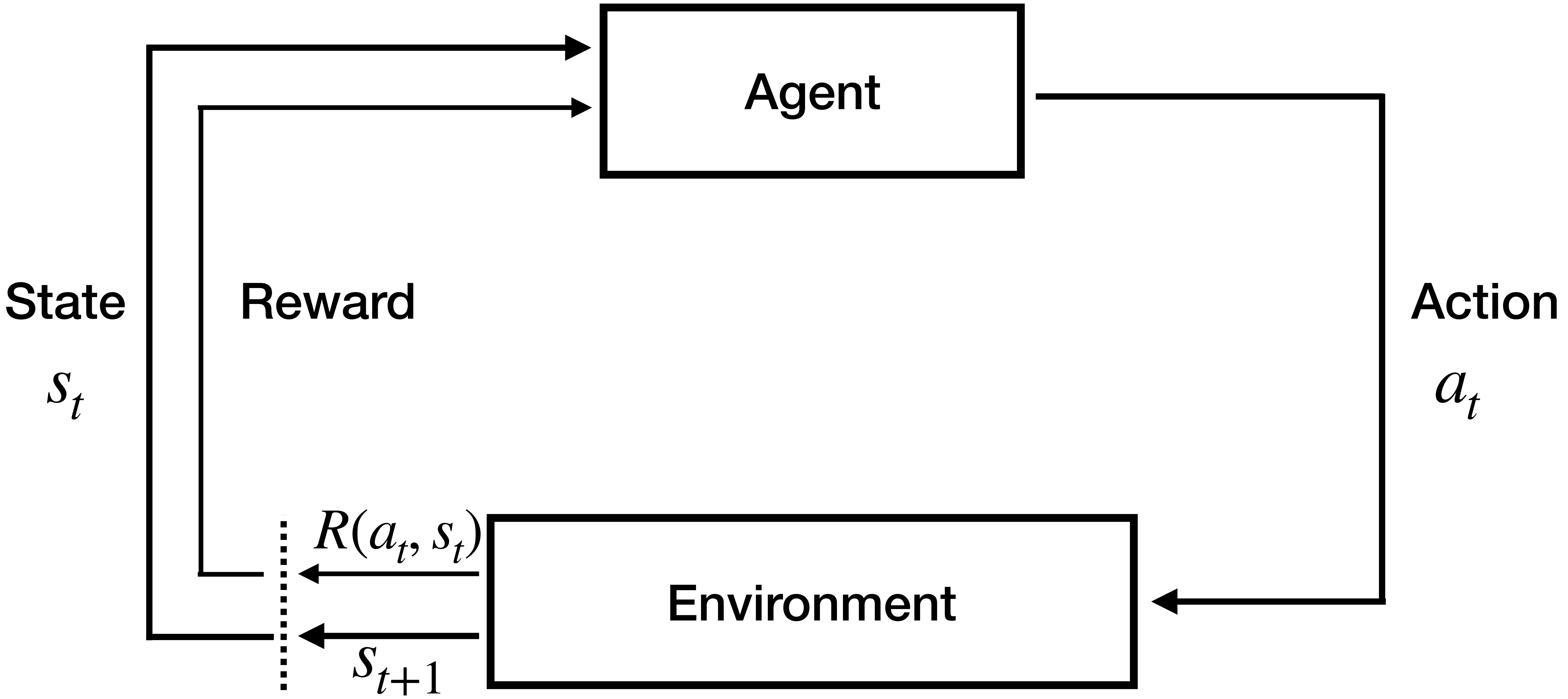
Reinforcement Learning

The Agent:

- Iteratively selects actions a_t based on a policy π
- Receives feedback from the environment in terms of new states s_{t+1} and rewards $R(a_t, s_t)$
- Updates internal representations
  - value $Q(s, a)$ or $V(s)$
  - model of the environment
    - reward function R
    - transitions $T(s' \mid s)$

The Environment:

- governs the transition between states $s_t \rightarrow s_{t+1}$
- provides rewards $R(a_t, s_t)$



Sutton and Barto (2018 [1998])

Reinforcement Learning

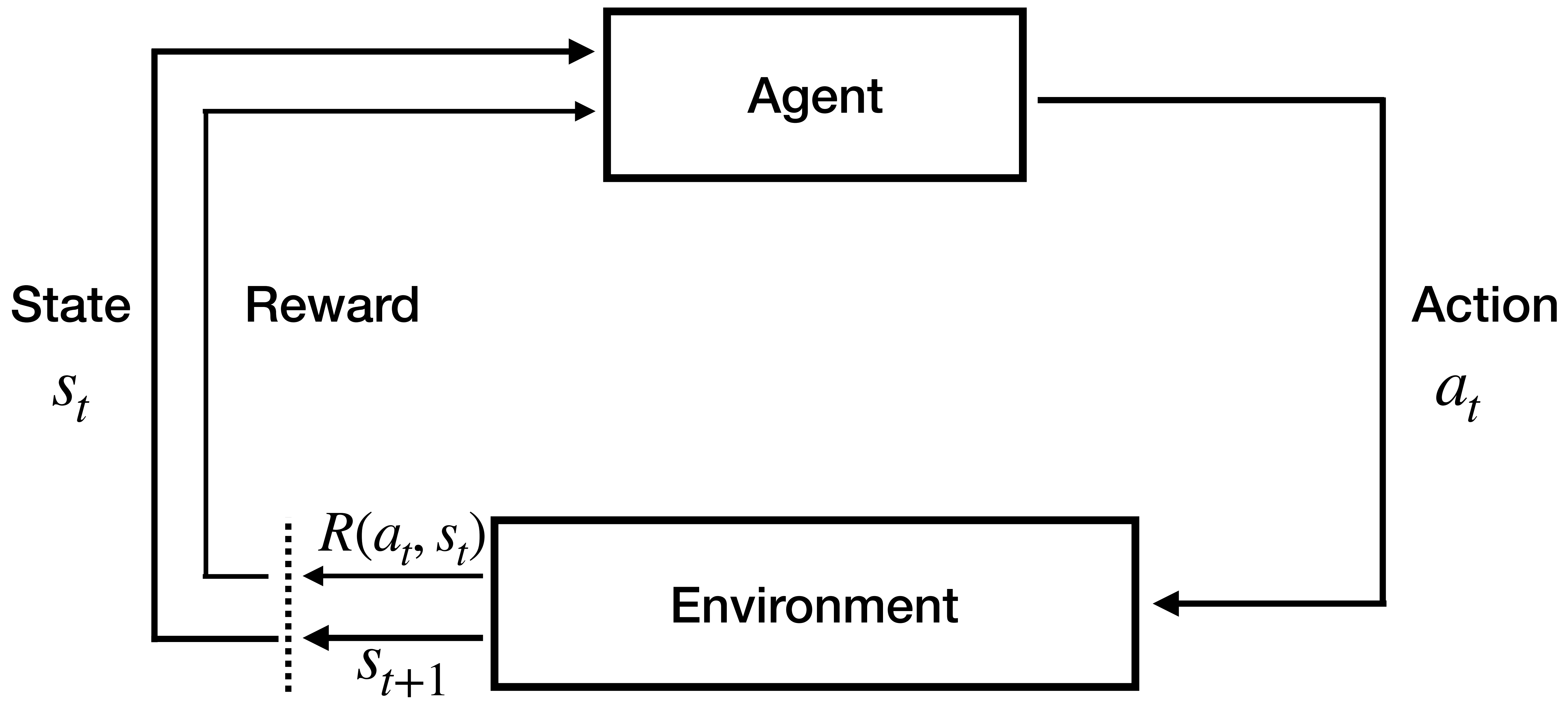
The Agent:

- Iteratively selects actions a_t based on a policy π
- Receives feedback from the environment in terms of new states s_{t+1} and rewards $R(a_t, s_t)$
- Updates internal representations
  - value $Q(s, a)$ or $V(s)$
  - model of the environment
    - reward function R
    - transitions $T(s' \mid s)$

The Environment:

- governs the transition between states $s_t \to s_{t+1}$
- provides rewards $R(a_t, s_t)$

Delta-rule of learning



Sutton and Barto (2018 [1998])

Reinforcement Learning

The Agent:

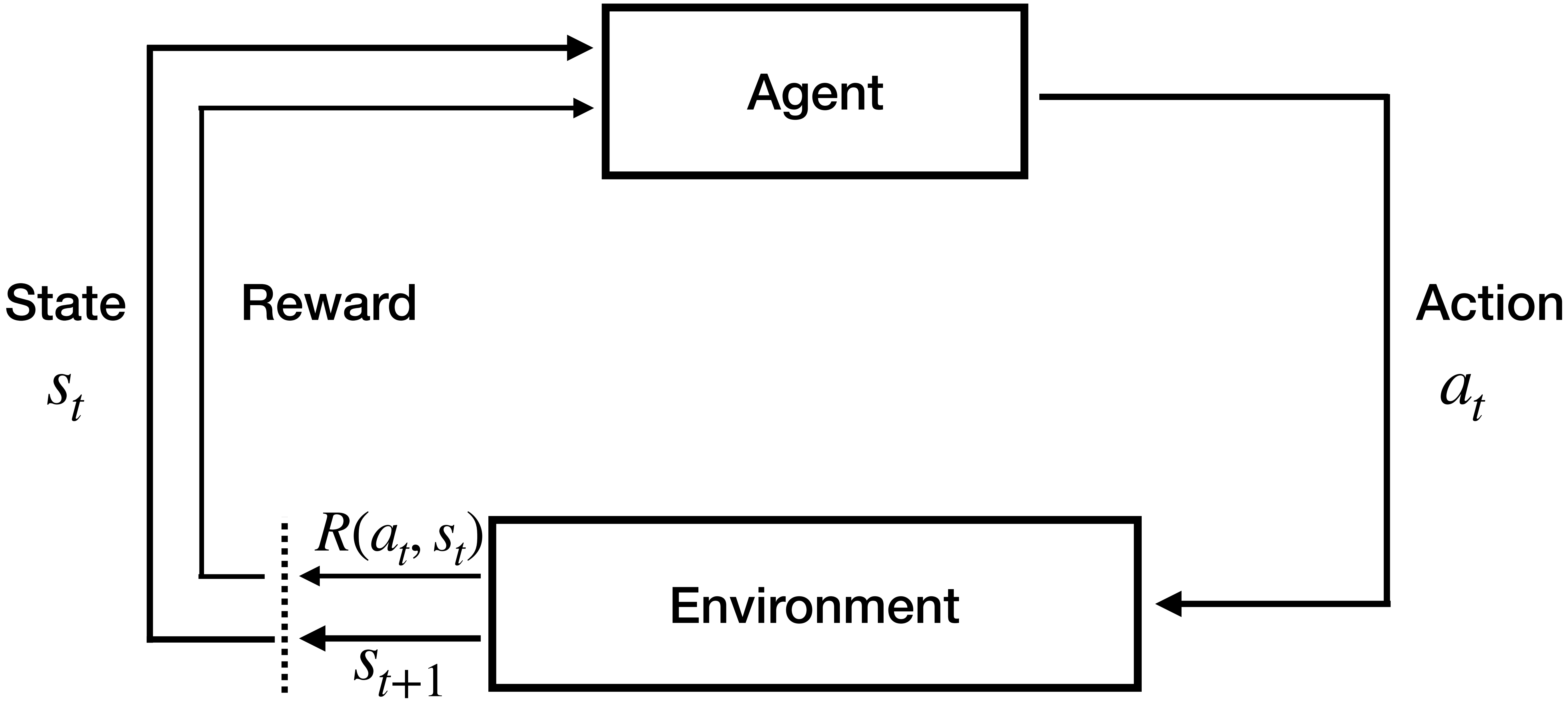
- Iteratively selects actions a_t based on a policy π
- Receives feedback from the environment in terms of new states s_{t+1} and rewards $R(a_t, s_t)$
- Updates internal representations
    - value $Q(s, a)$ or $V(s)$
    - model of the environment
        - reward function R
        - transitions $T(s' \mid s)$

The Environment:

- governs the transition between states $s_t \rightarrow s_{t+1}$
- provides rewards $R(a_t, s_t)$

Delta-rule of learning

Belief-updates are proportional to the magnitude of the reward predition error (RPE)



Sutton and Barto (2018 [1998])

Reinforcement Learning

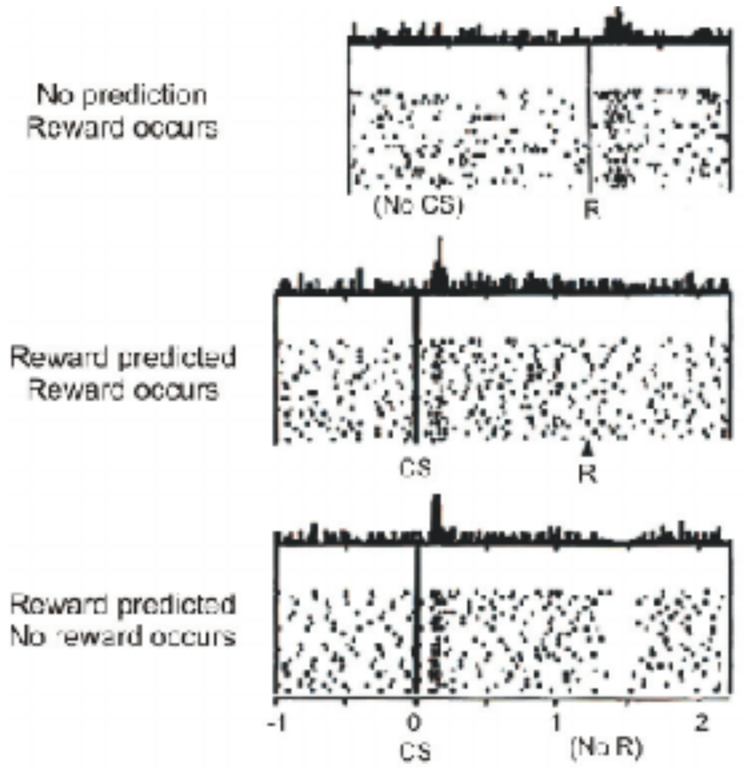
The Agent:

- Iteratively selects actions a_t based on a policy π
- Receives feedback from the environment in terms of new states s_{t+1} and rewards $R(a_t, s_t)$
- Updates internal representations
  - value $Q(s, a)$ or $V(s)$
  - model of the environment
    - reward function R
    - transitions $T(s' \mid s)$

The Environment:

- governs the transition between states $s_t \rightarrow s_{t+1}$
- provides rewards $R(a_t, s_t)$

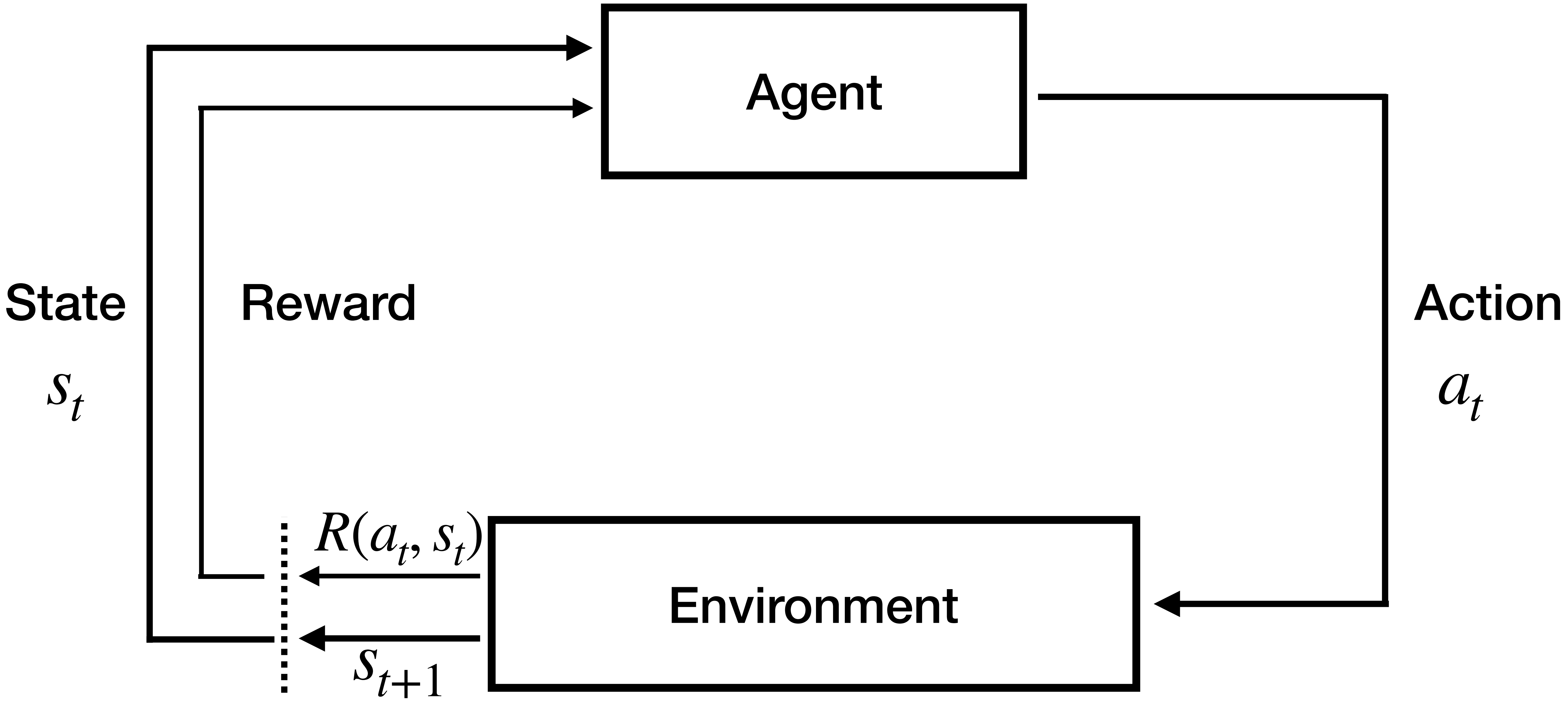
Dopamine RPE Signal



Schultz et al. (1997)

Delta-rule of learning

Belief-updates are proportional to the magnitude of the reward predition error (RPE)

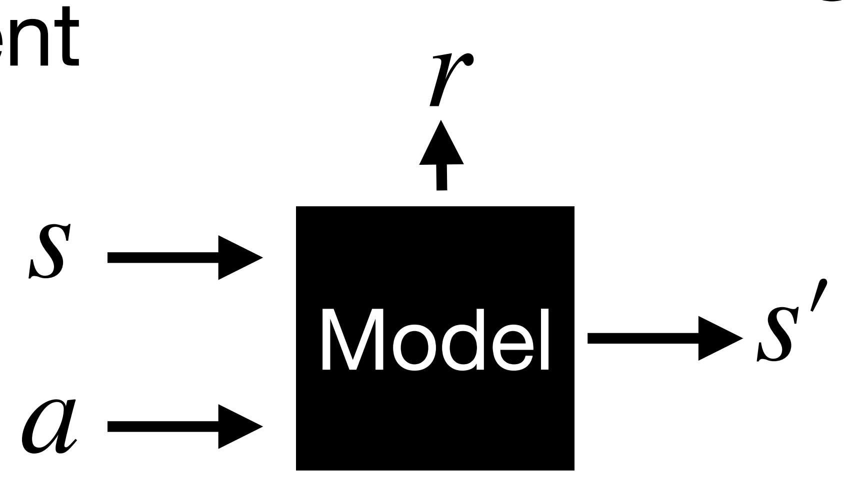


Sutton and Barto (2018 [1998])

Reinforcement Learning

The Agent:

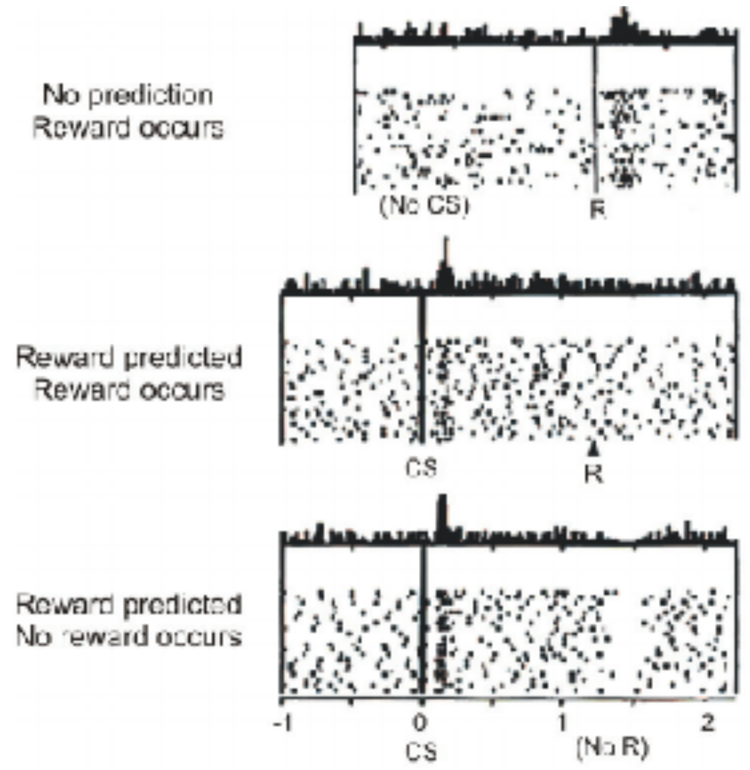
- Iteratively selects actions a_t based on a policy π
- Receives feedback from the environment in terms of new states s_{t+1} and rewards $R(a_t, s_t)$
- Updates internal representations
  - value $Q(s, a)$ or $V(s)$
  - model of the environment
    - reward function R
    - transitions $T(s' \mid s)$



The Environment:

- governs the transition between states $s_t \rightarrow s_{t+1}$
- provides rewards $R(a_t, s_t)$

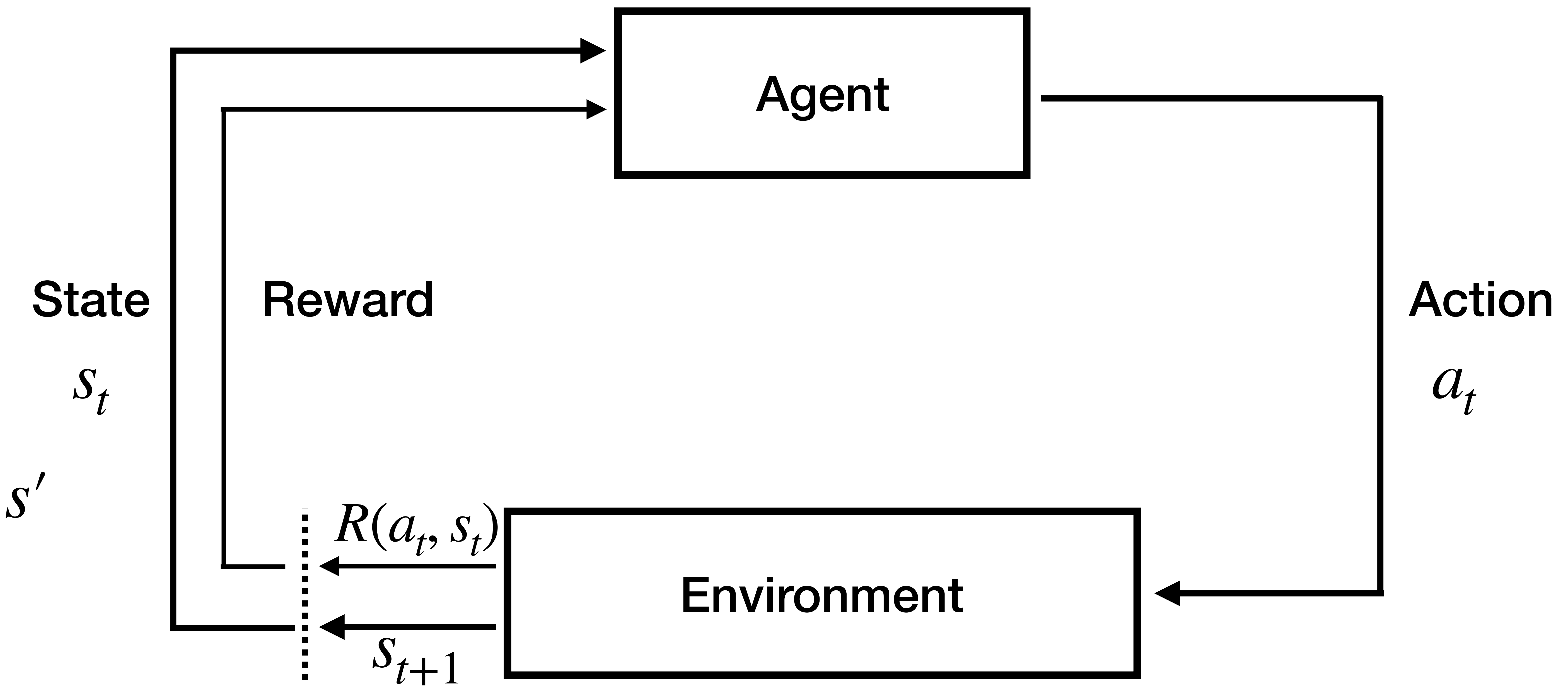
Dopamine RPE Signal



Schultz et al. (1997)

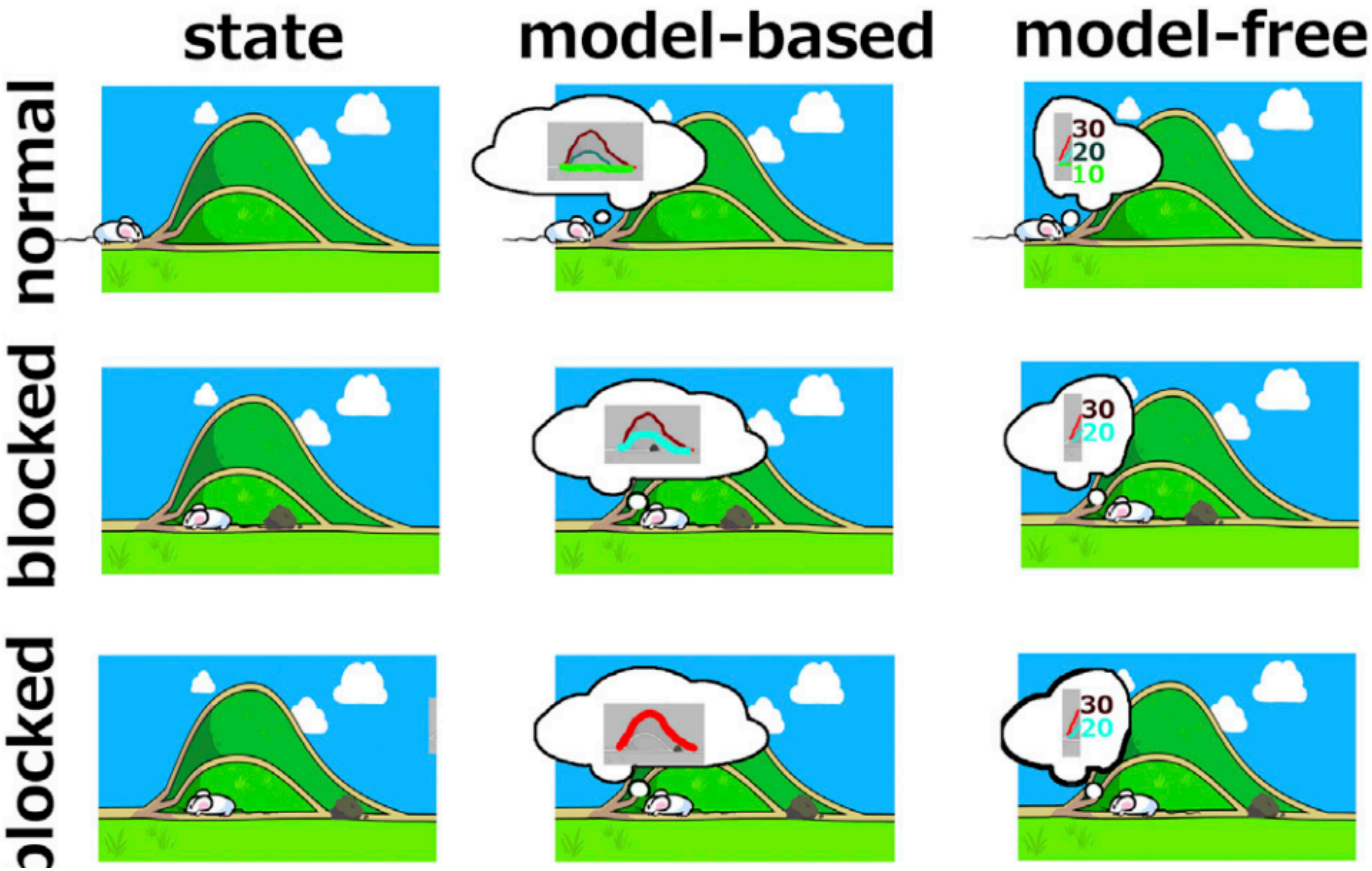
Delta-rule of learning

Belief-updates are proportional to the magnitude of the reward predition error (RPE)



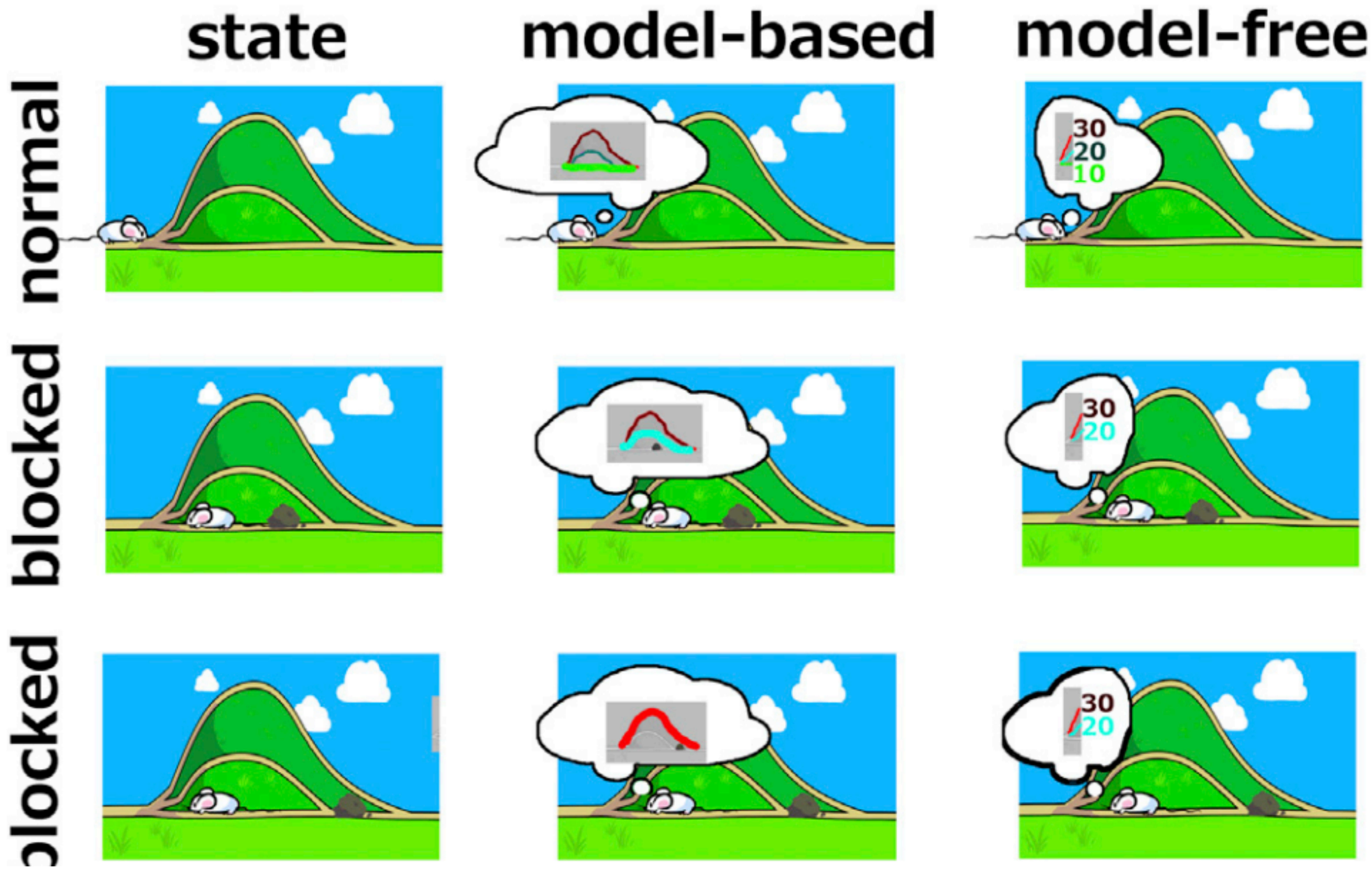
Sutton and Barto (2018 [1998])

Model-free vs. model-based



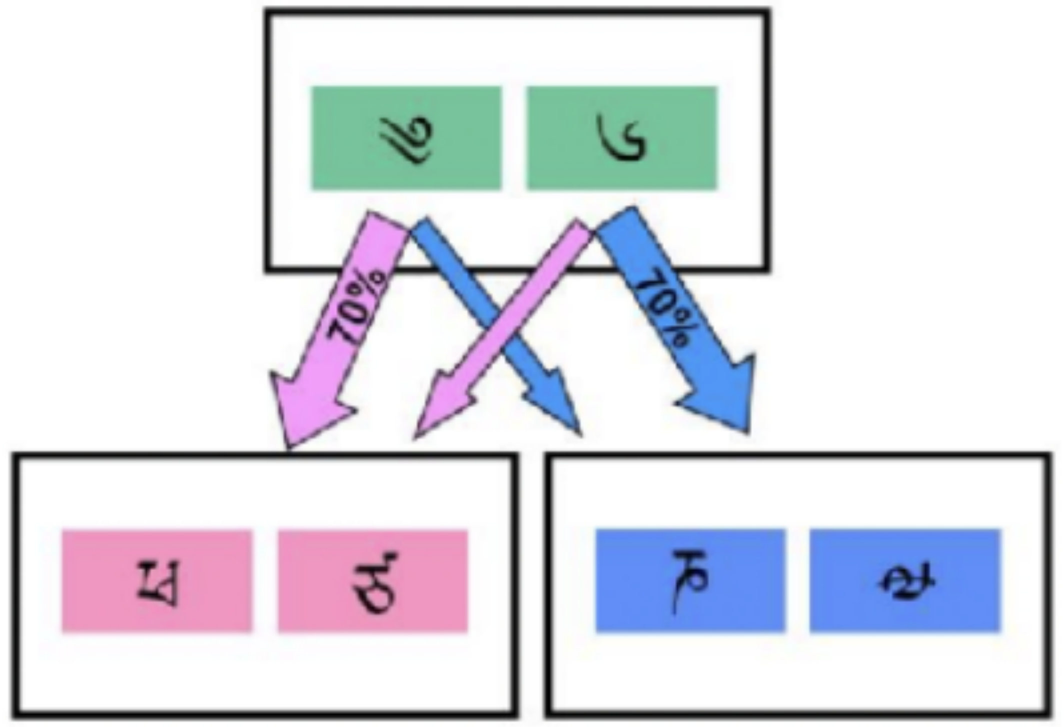
Niv (2009)

Model-free vs. model-based



Niv (2009)

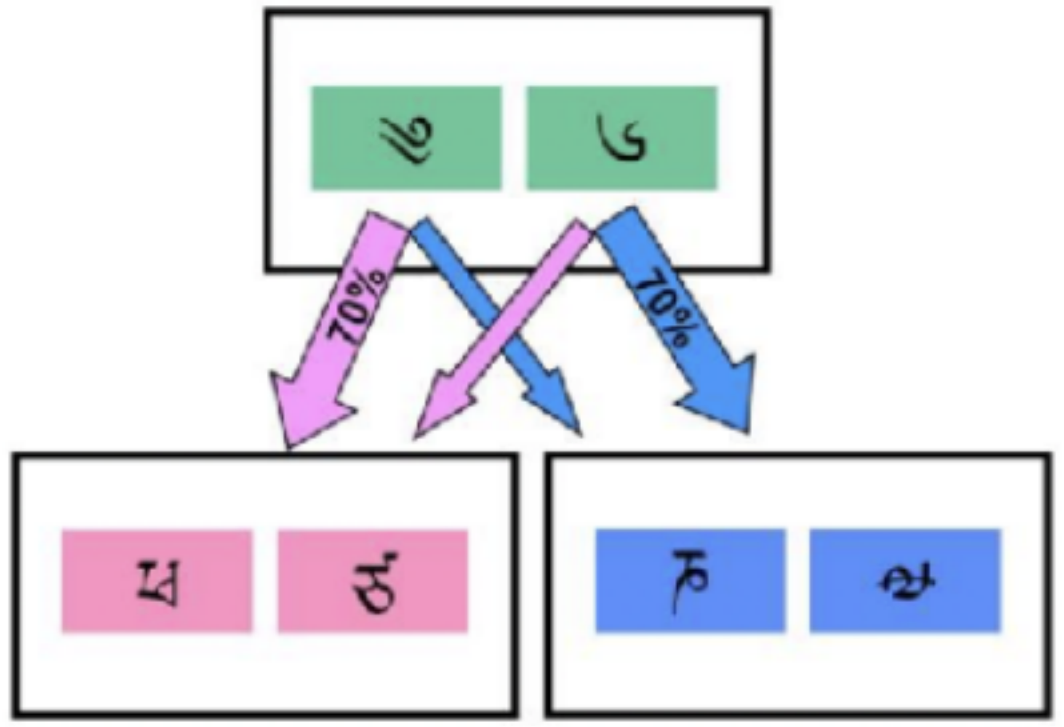
2-step task



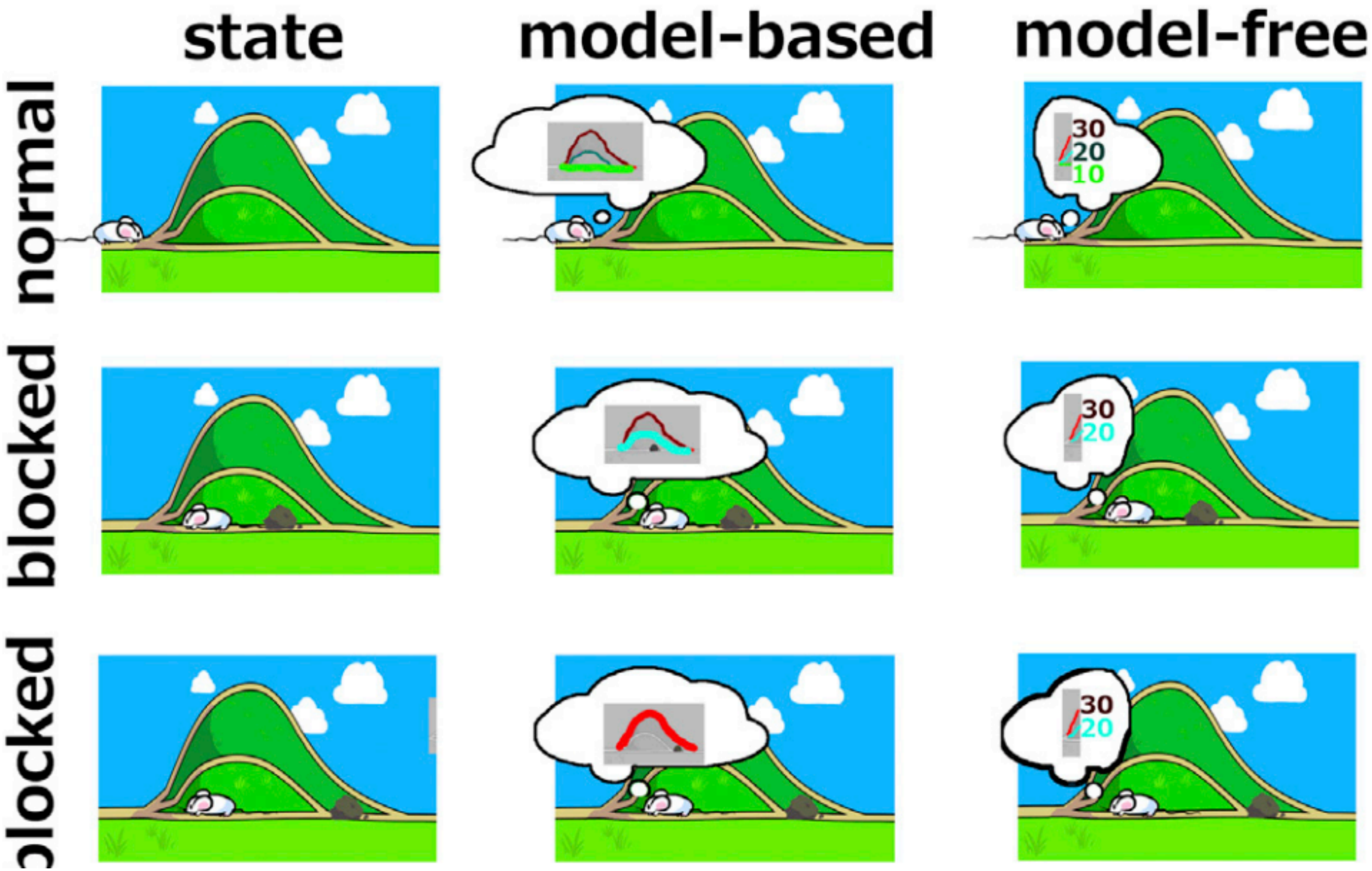
Daw et al., (2011)

Model-free vs. model-based

2-step task

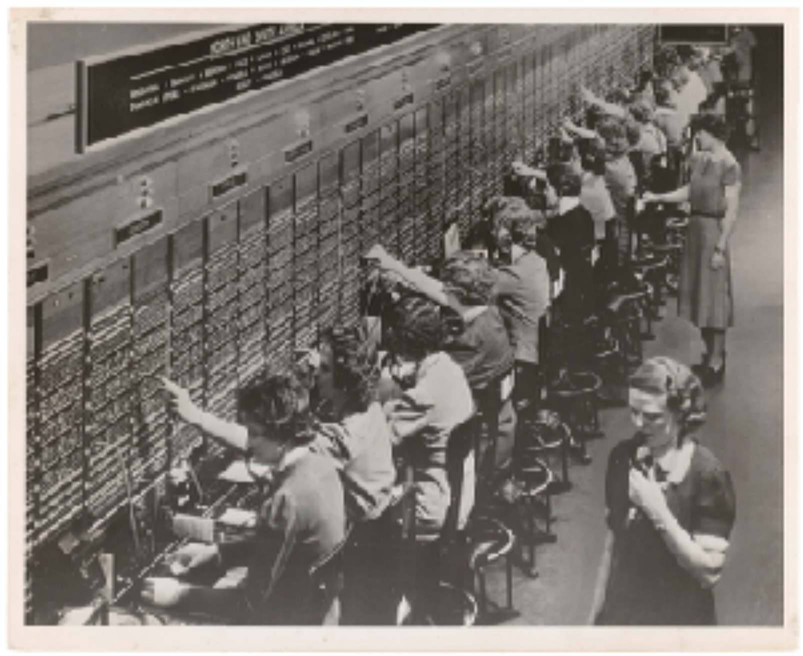


Daw et al., (2011)

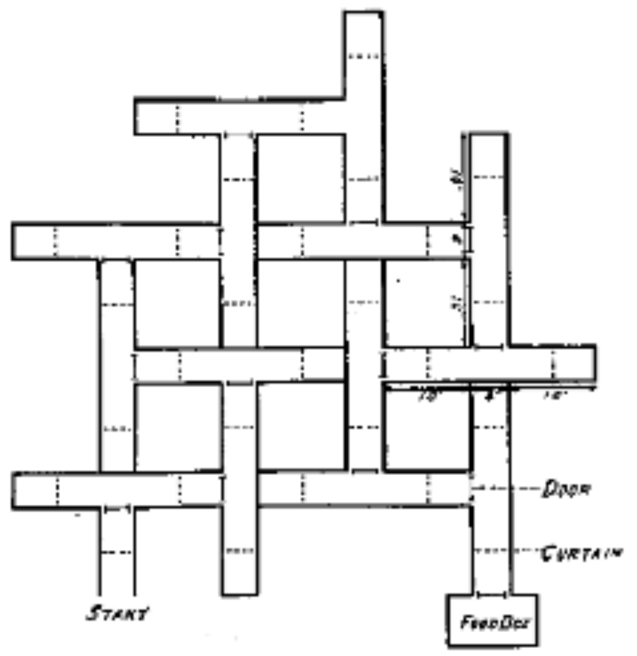


Niv (2009)

S-R learning

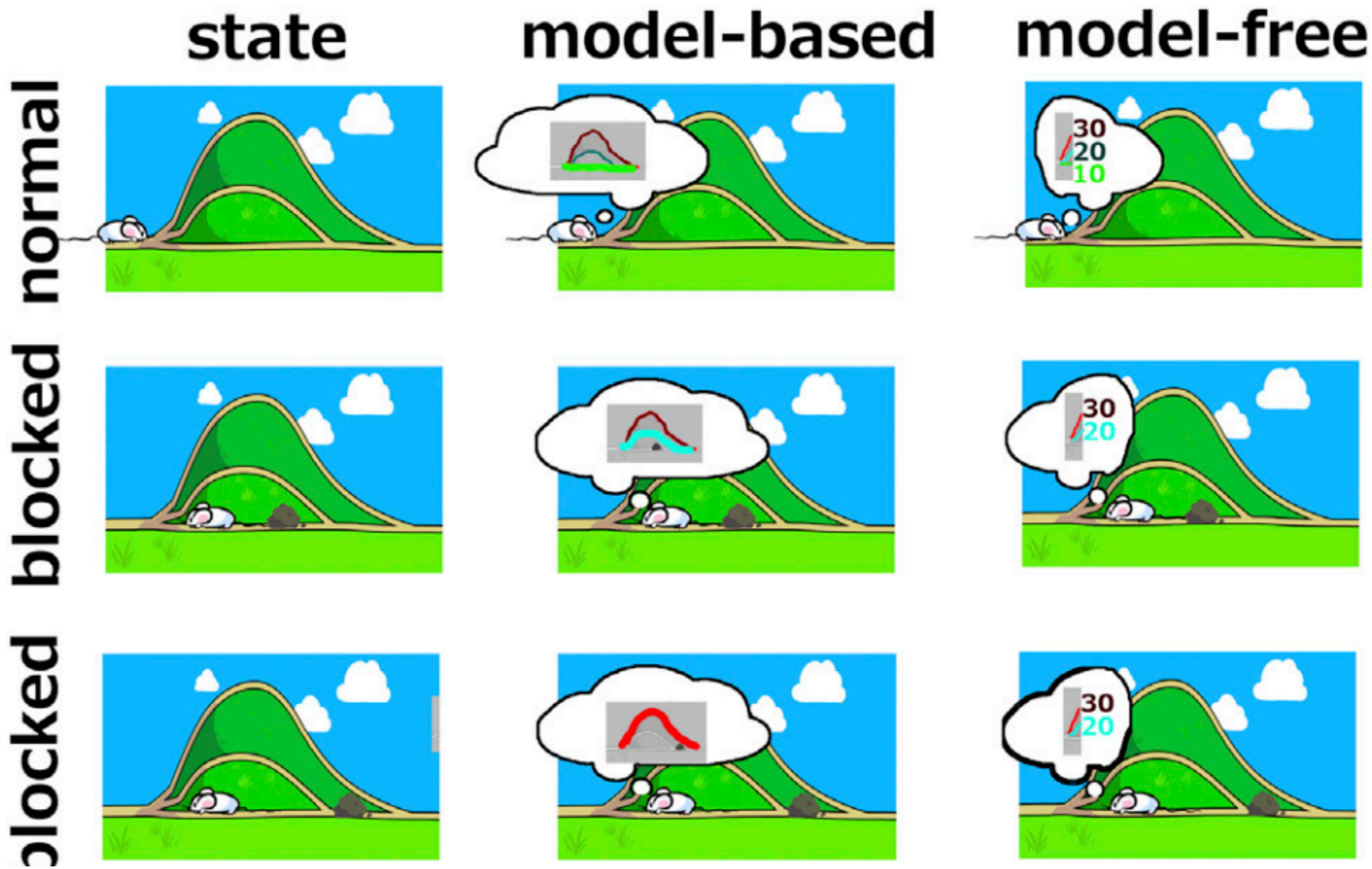


S-S learning



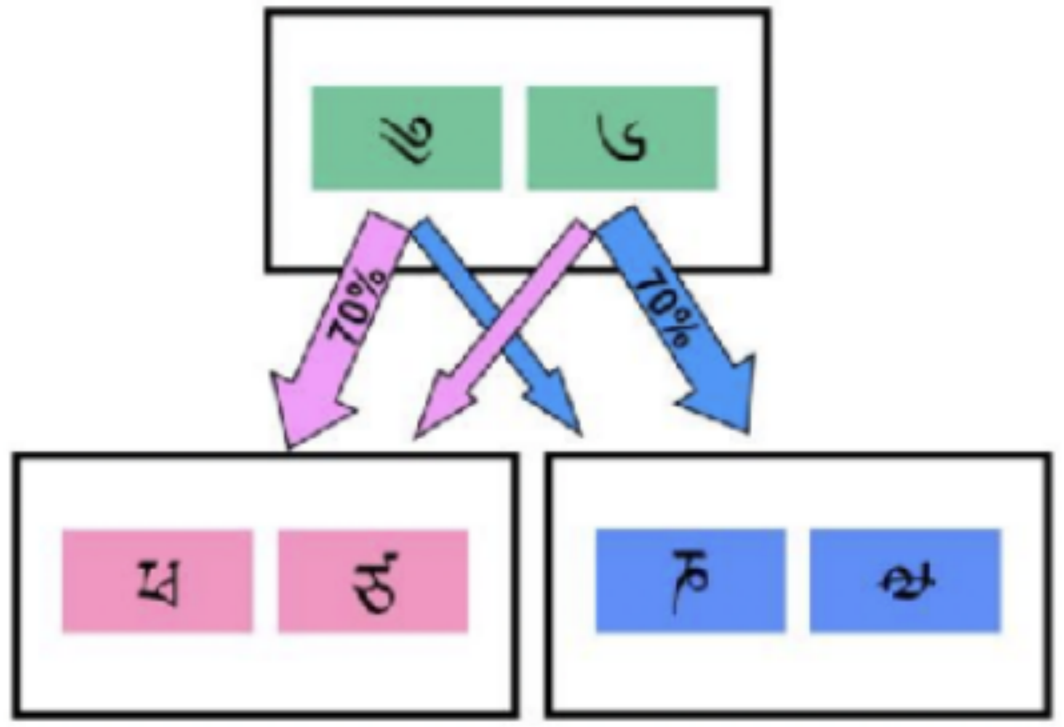
Tolman (1948)

Model-free vs. model-based



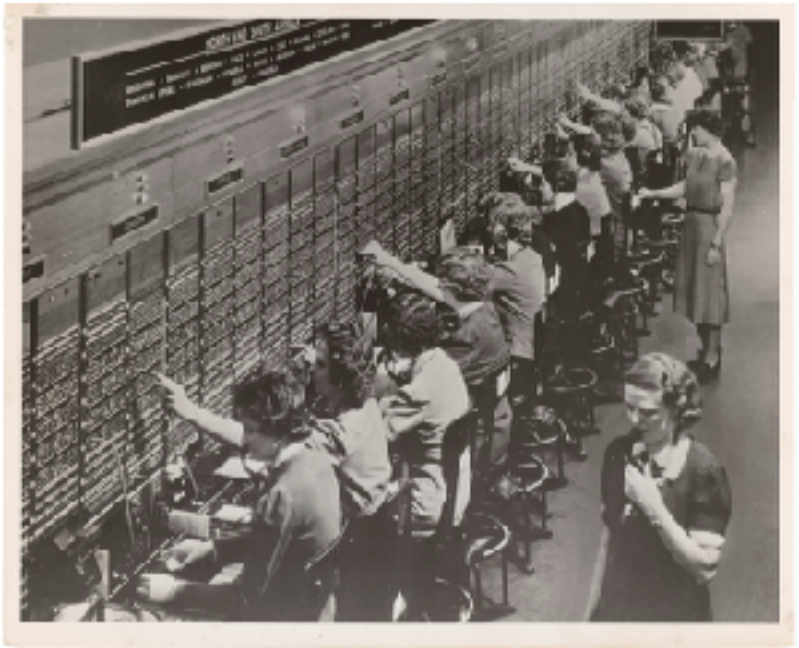
Niv (2009)

2-step task

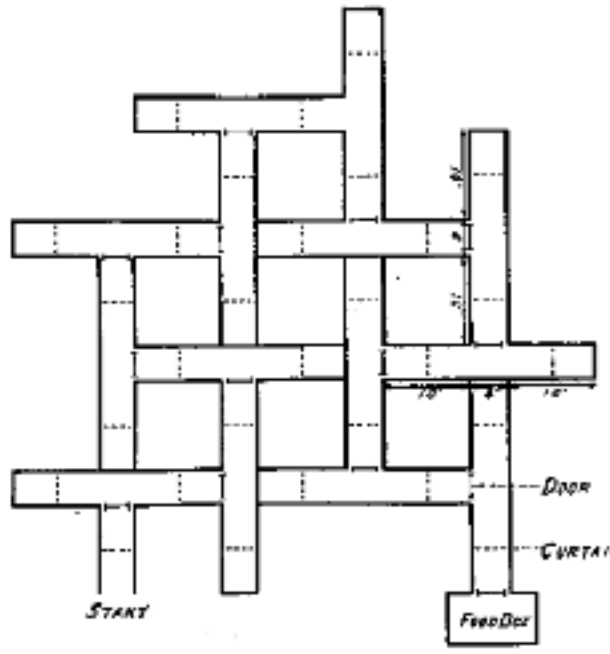


Daw et al., (2011)

(Model-free)
S-R learning



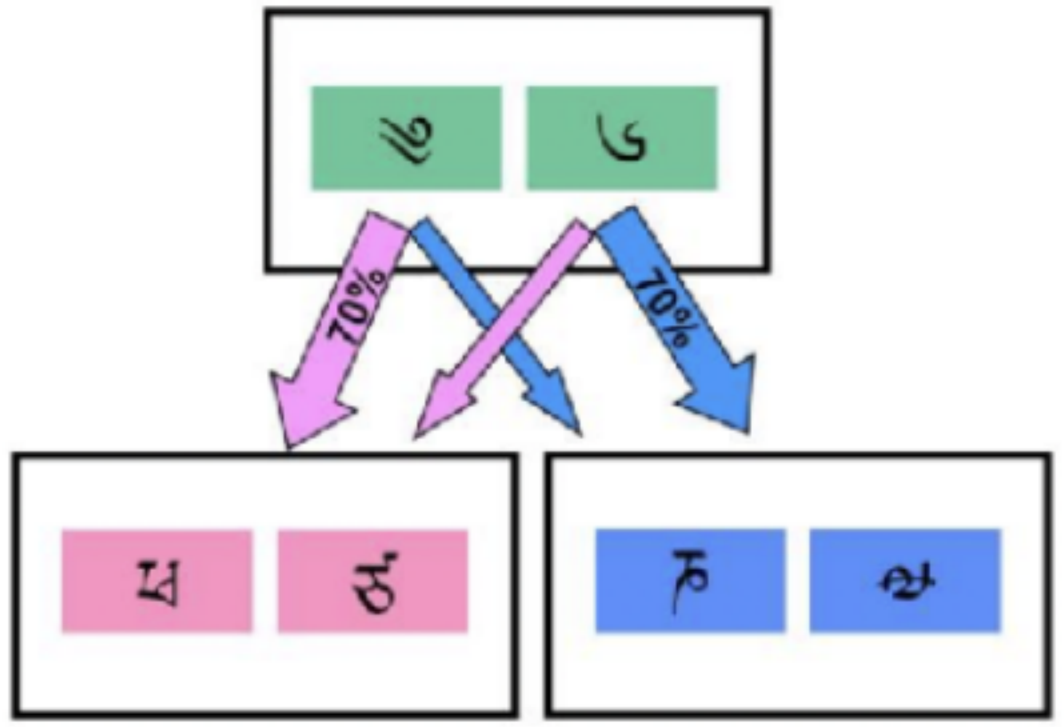
(Model-based)
S-S learning



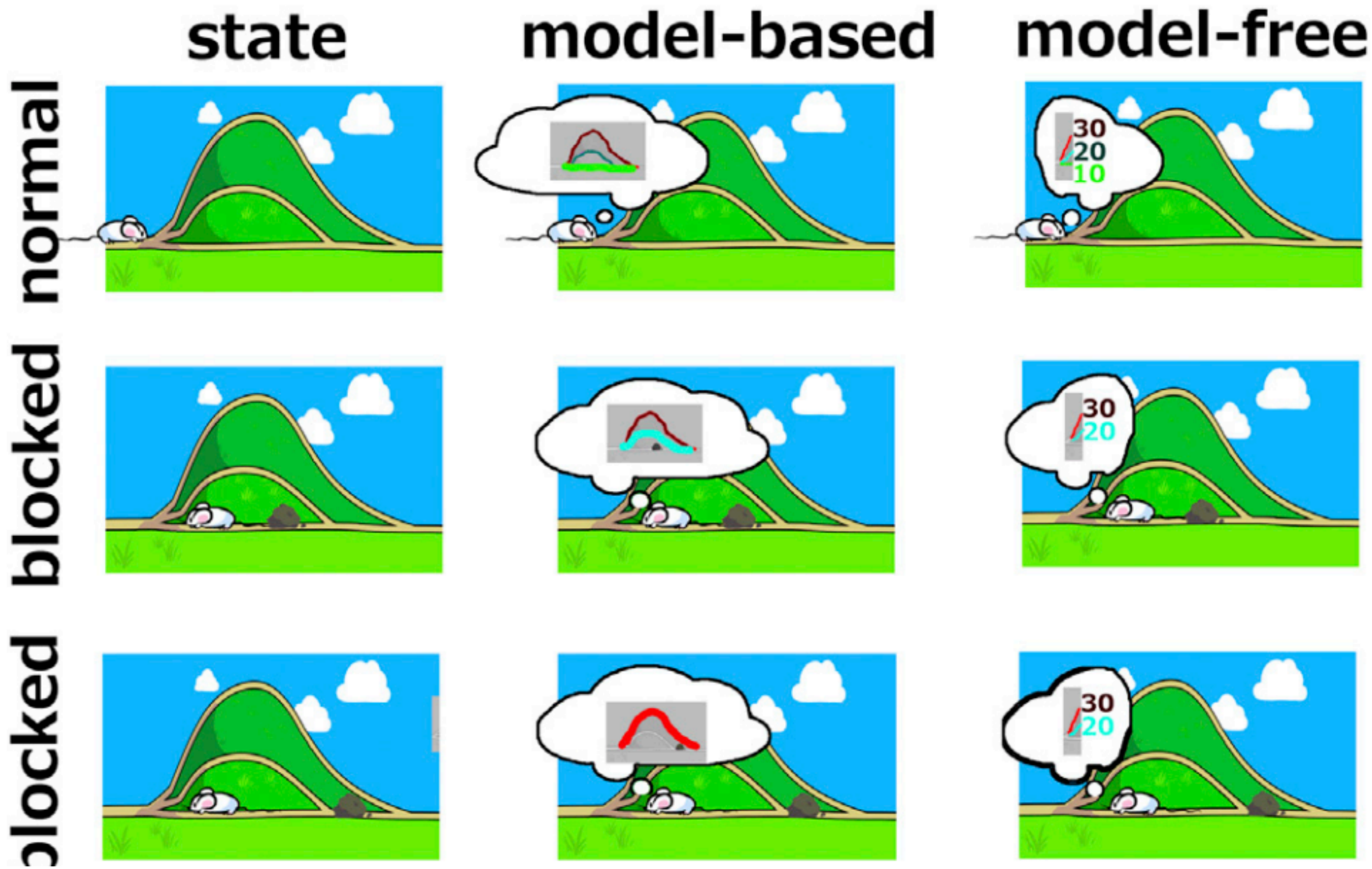
Tolman (1948)

Model-free vs. model-based

2-step task

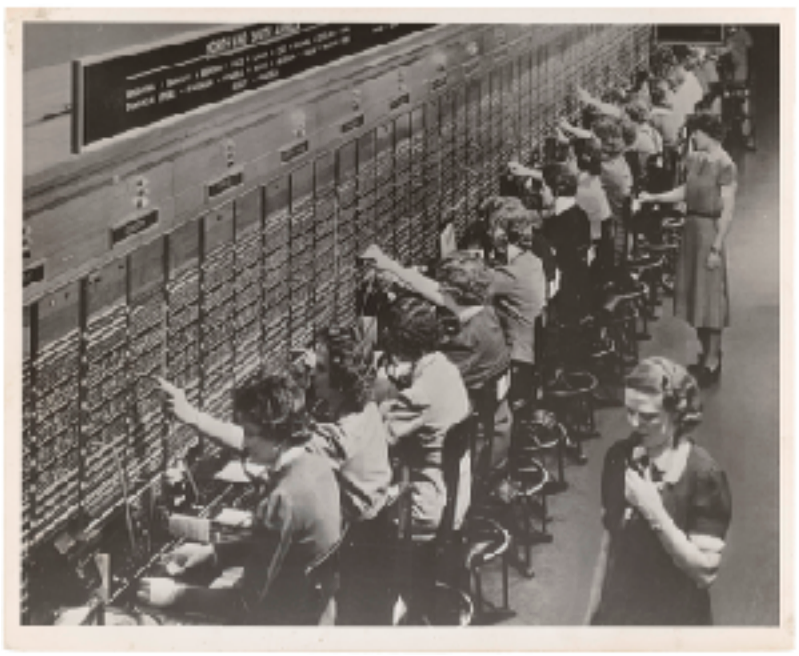


Daw et al., (2011)

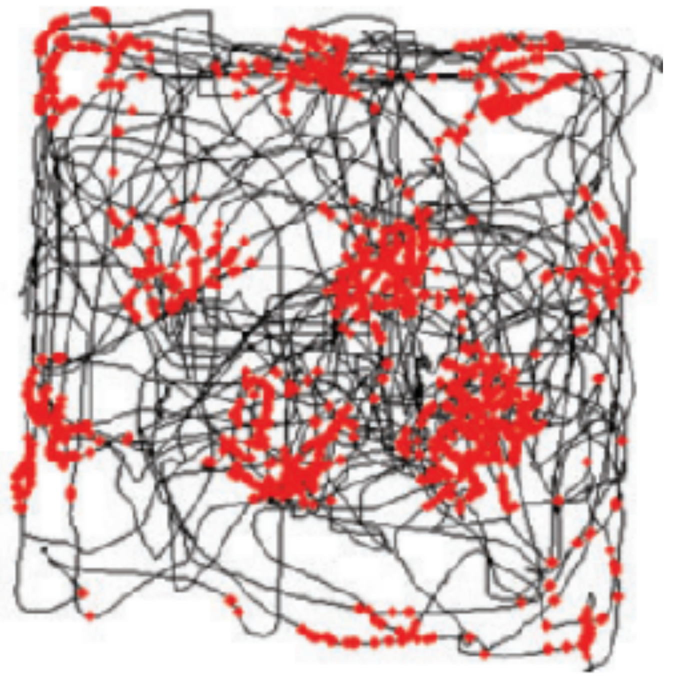


Niv (2009)

(Model-free)
S-R learning



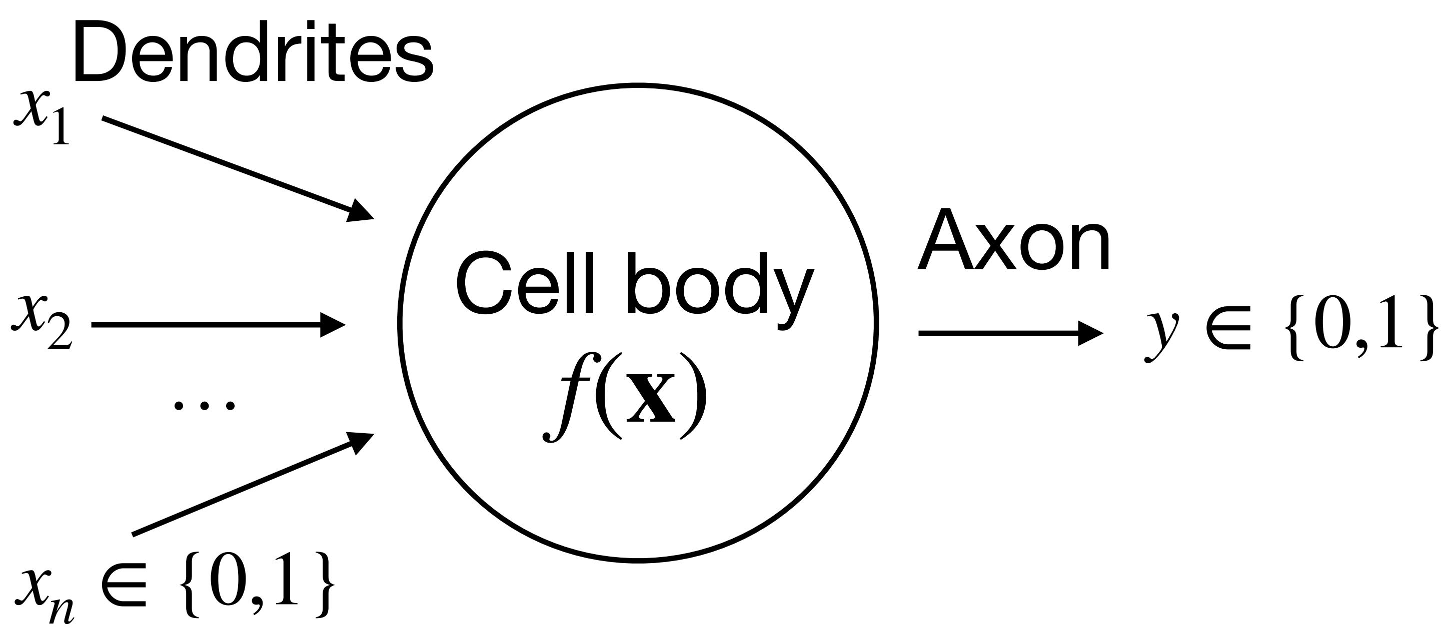
(Model-based)
S-S learning



Tolman (1948)

Symbolic vs. Subsymbolic AI

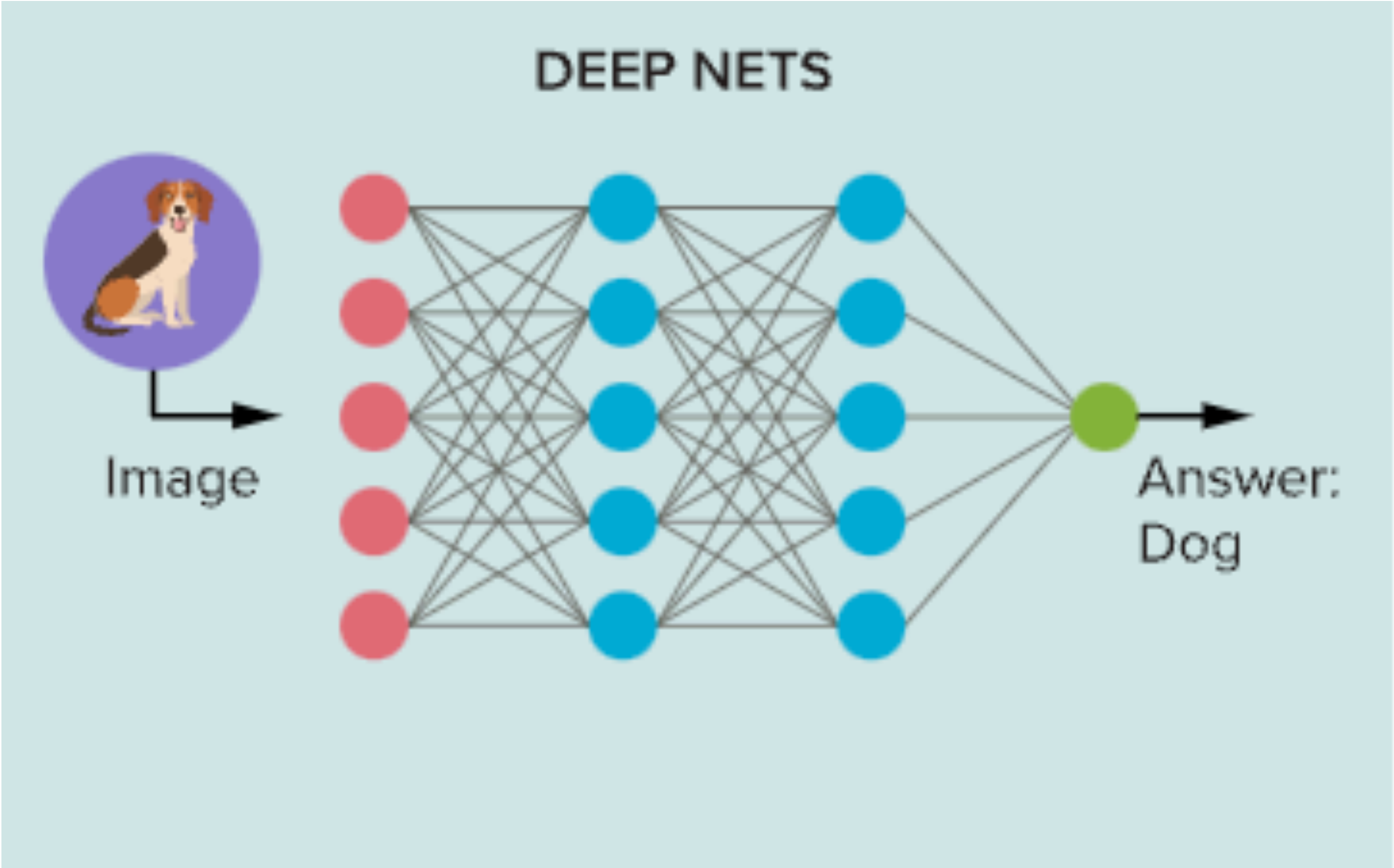
Subsymbolic AI



McCulloch & Pitts (1943)

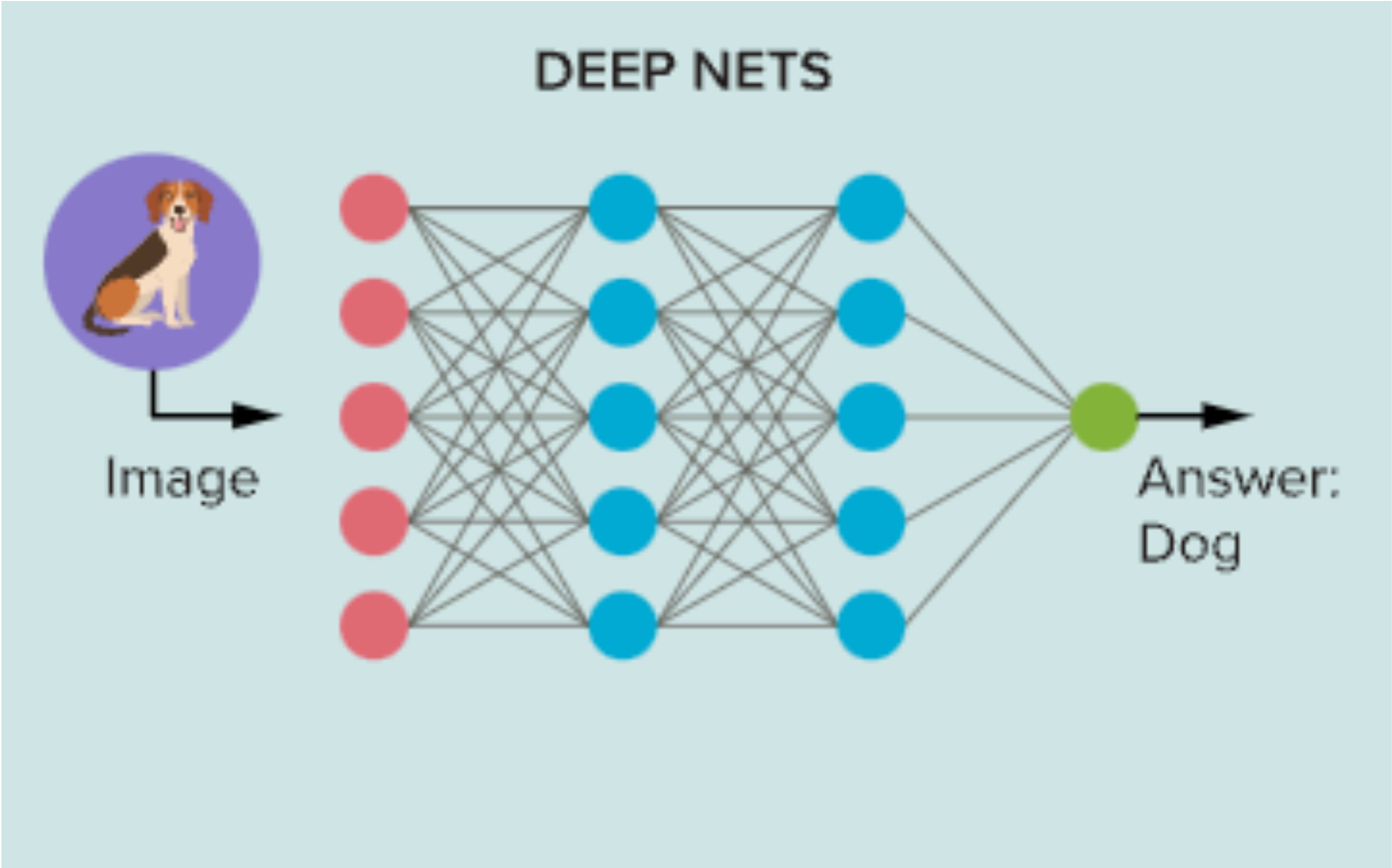
Symbolic vs. Subsymbolic AI

Subsymbolic AI



Symbolic vs. Subsymbolic AI

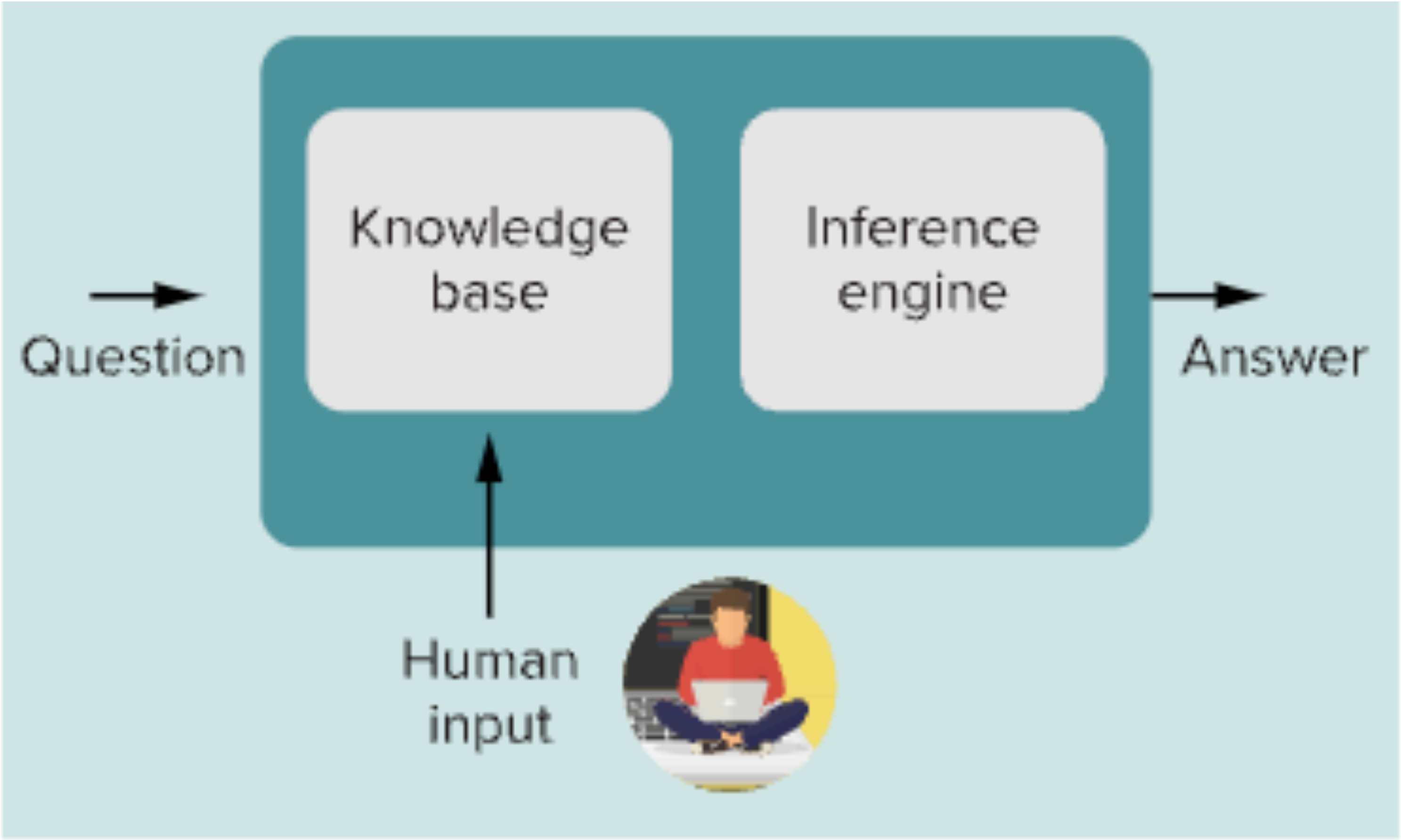
Subsymbolic AI



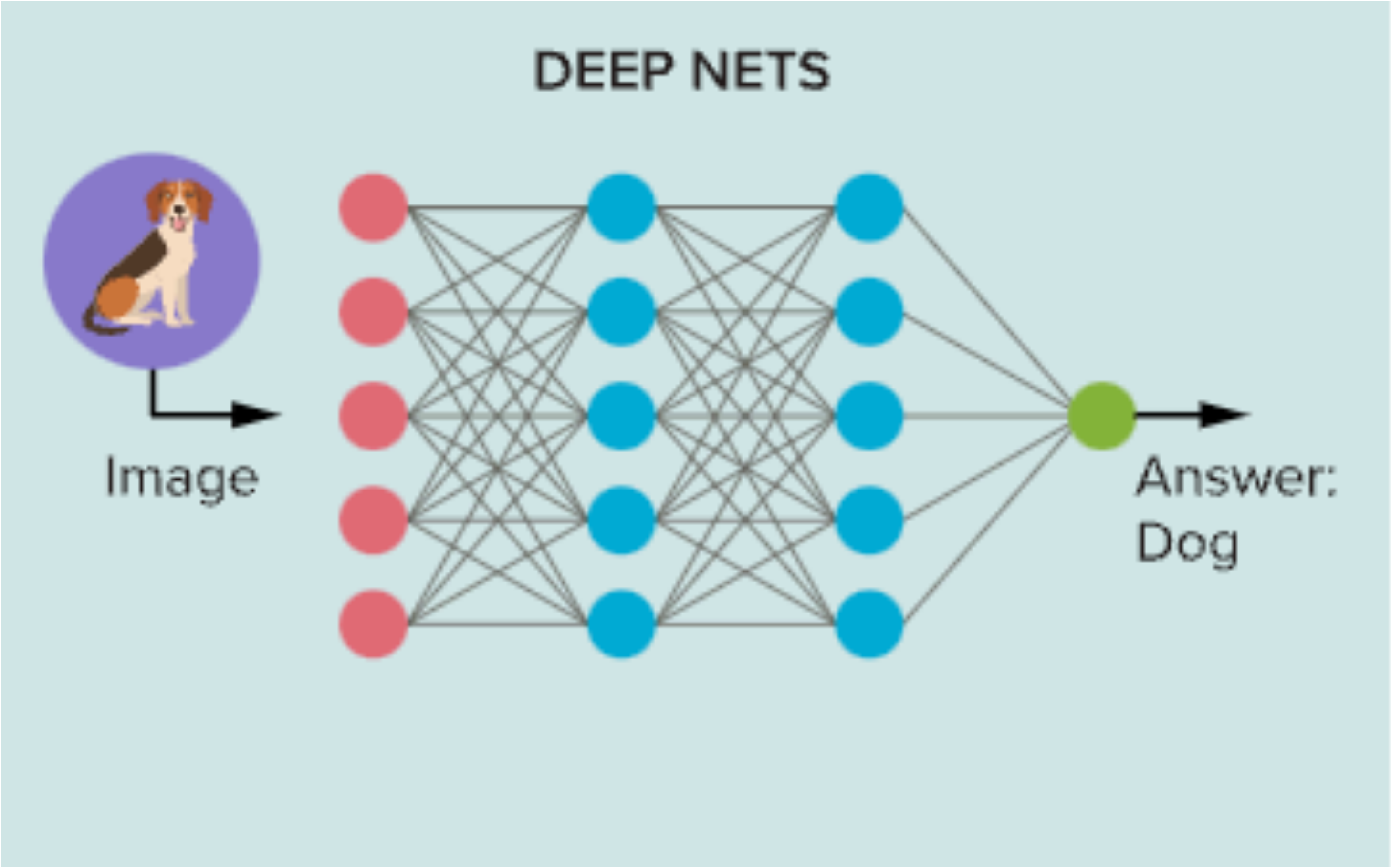
*Gradient descent is analogous to the delta-rule

Symbolic vs. Subsymbolic AI

Symbolic AI



Subsymbolic AI

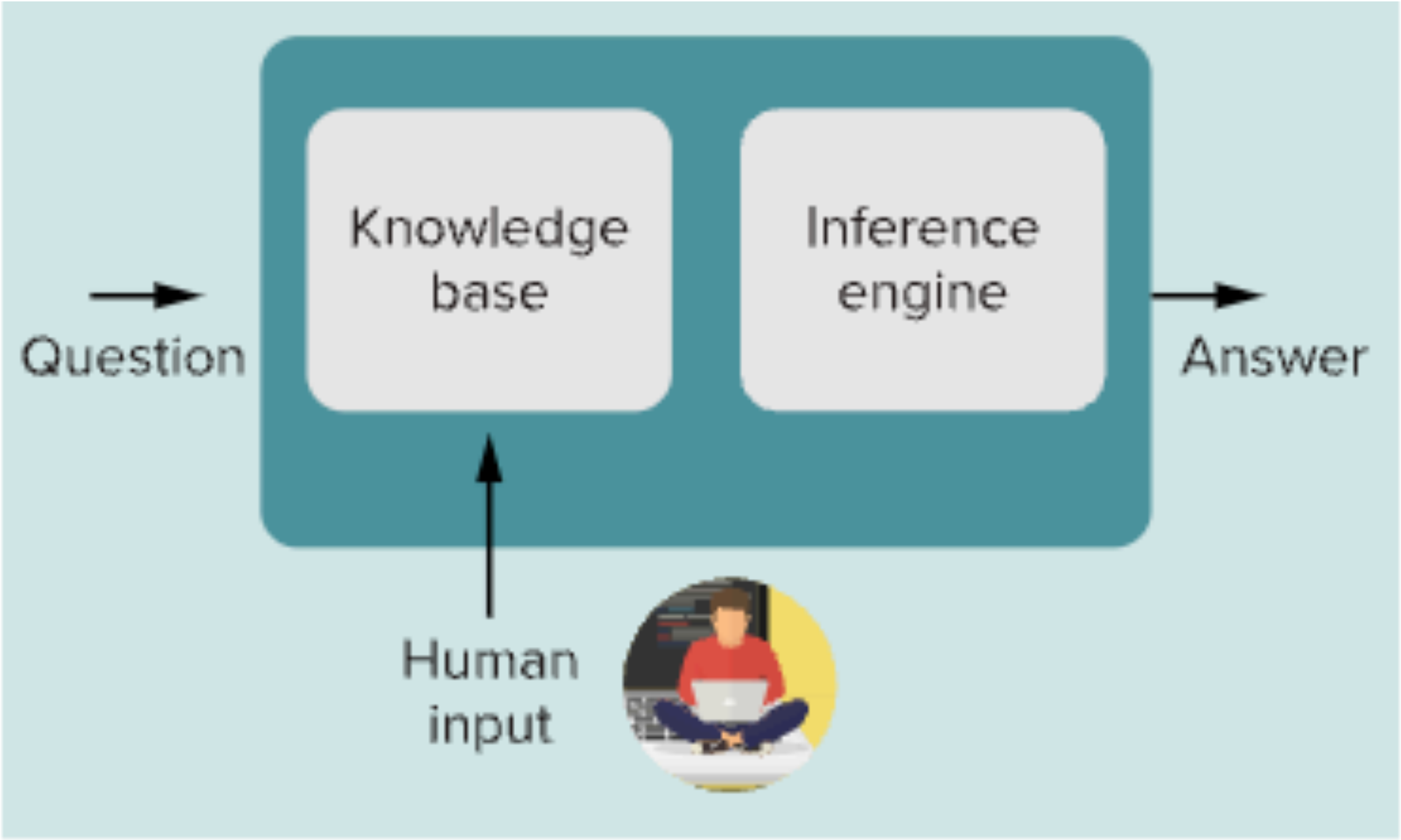


*Gradient descent is analogous to the delta-rule

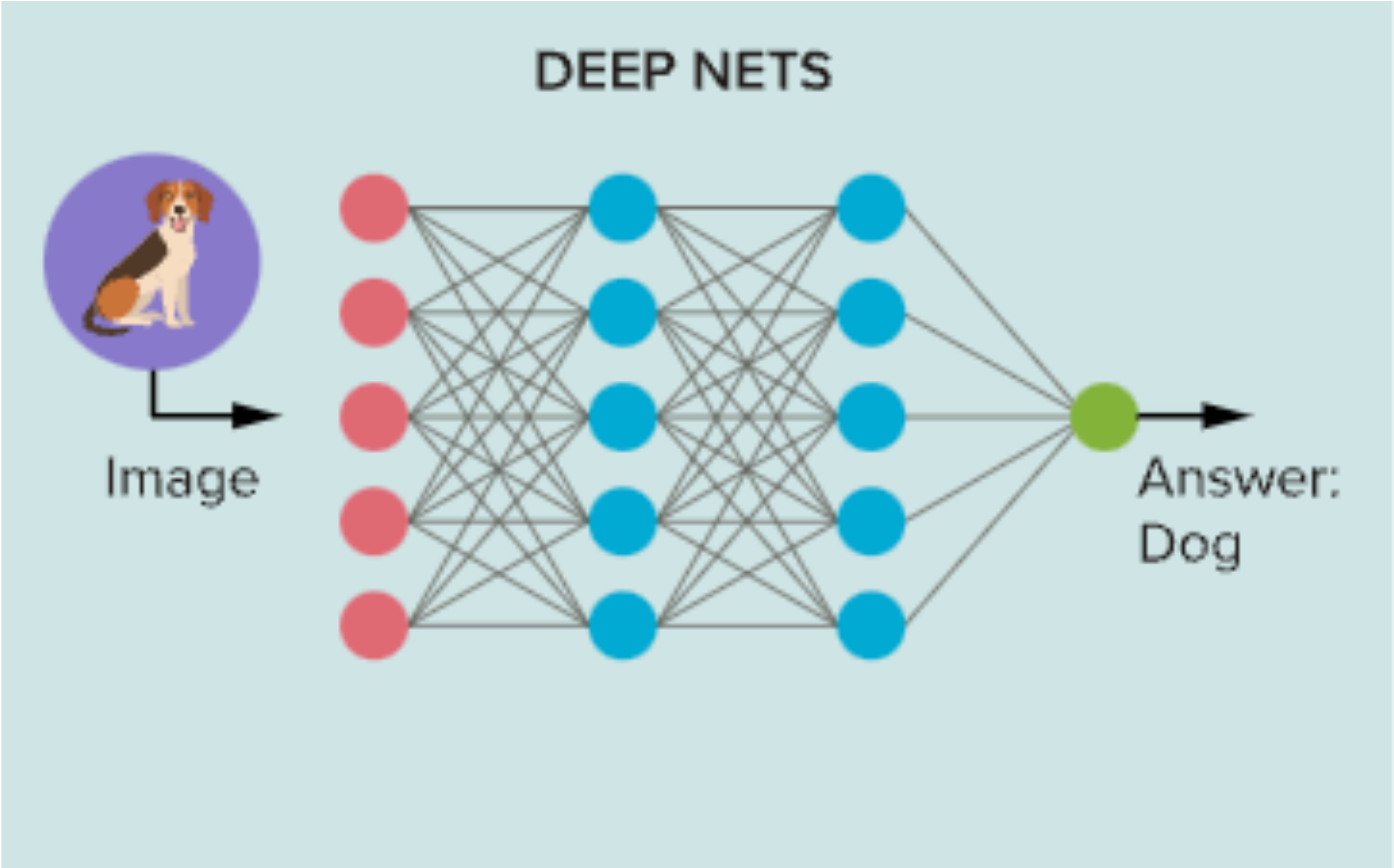
Symbolic vs. Subsymbolic AI

Physical symbol system hypothesis: manipulating **symbols** and **relations**

Symbolic AI



Subsymbolic AI

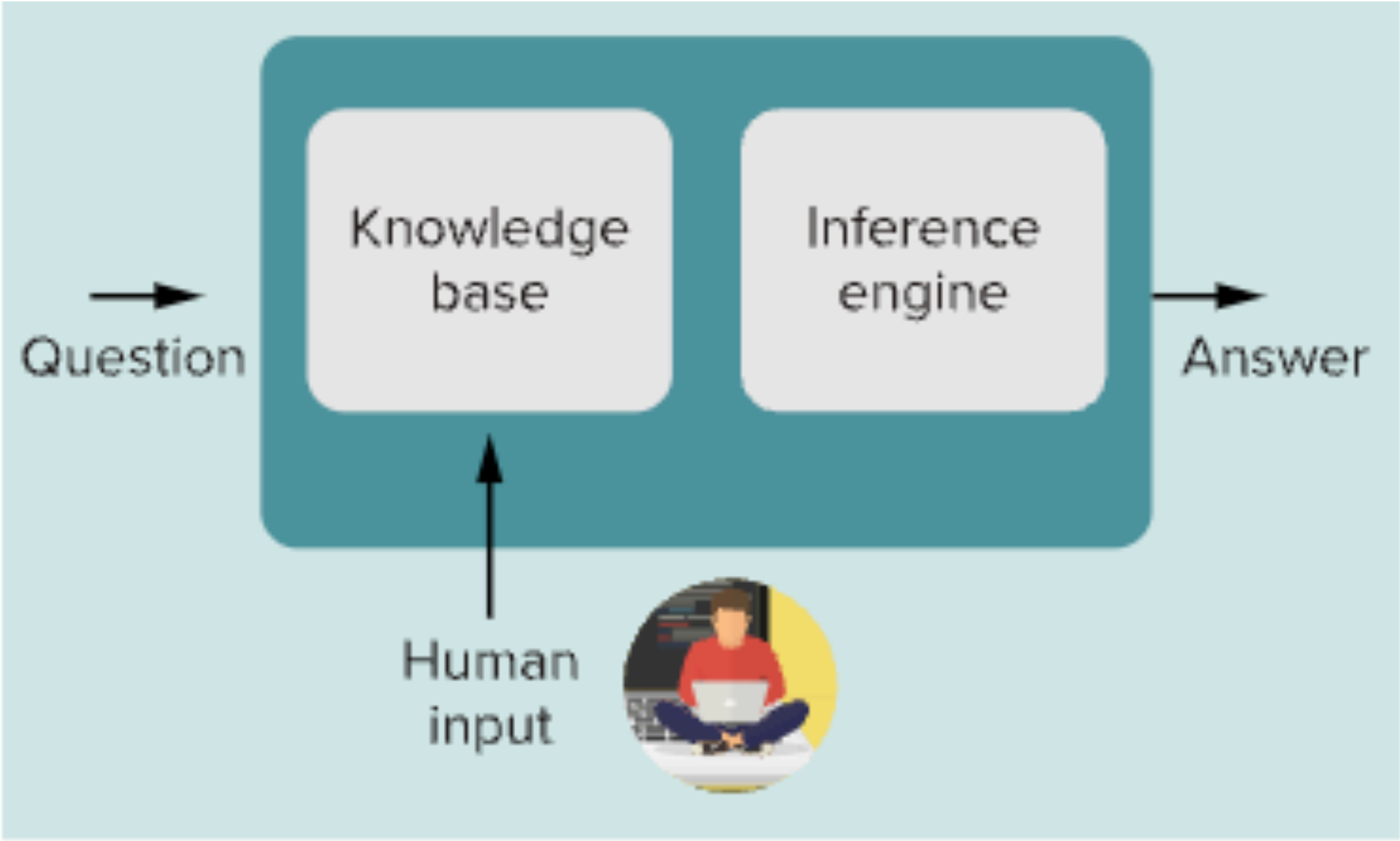


*Gradient descent is analogous to the delta-rule

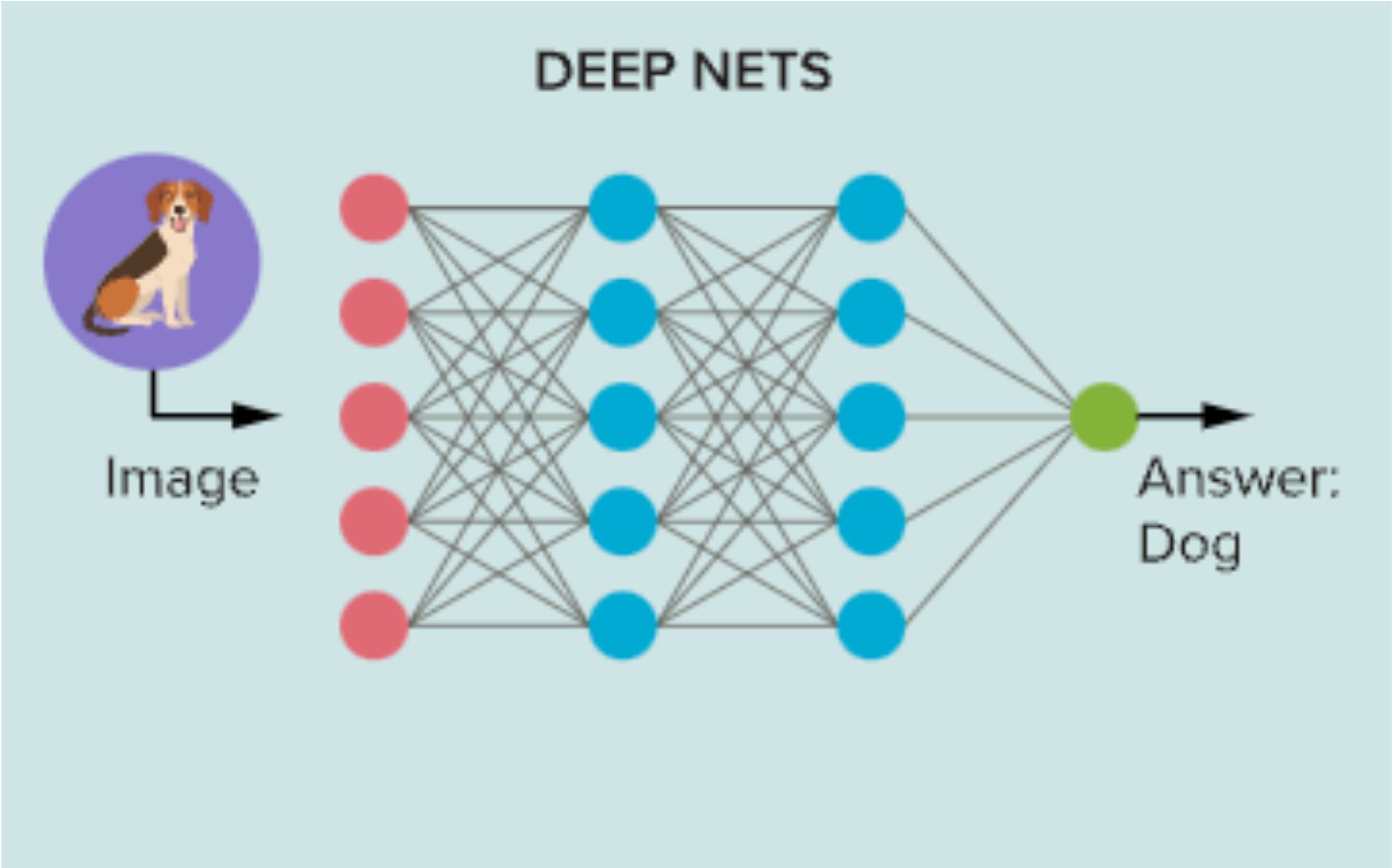
Symbolic vs. Subsymbolic AI

Physical symbol system hypothesis: manipulating **symbols** and **relations**

Symbolic AI

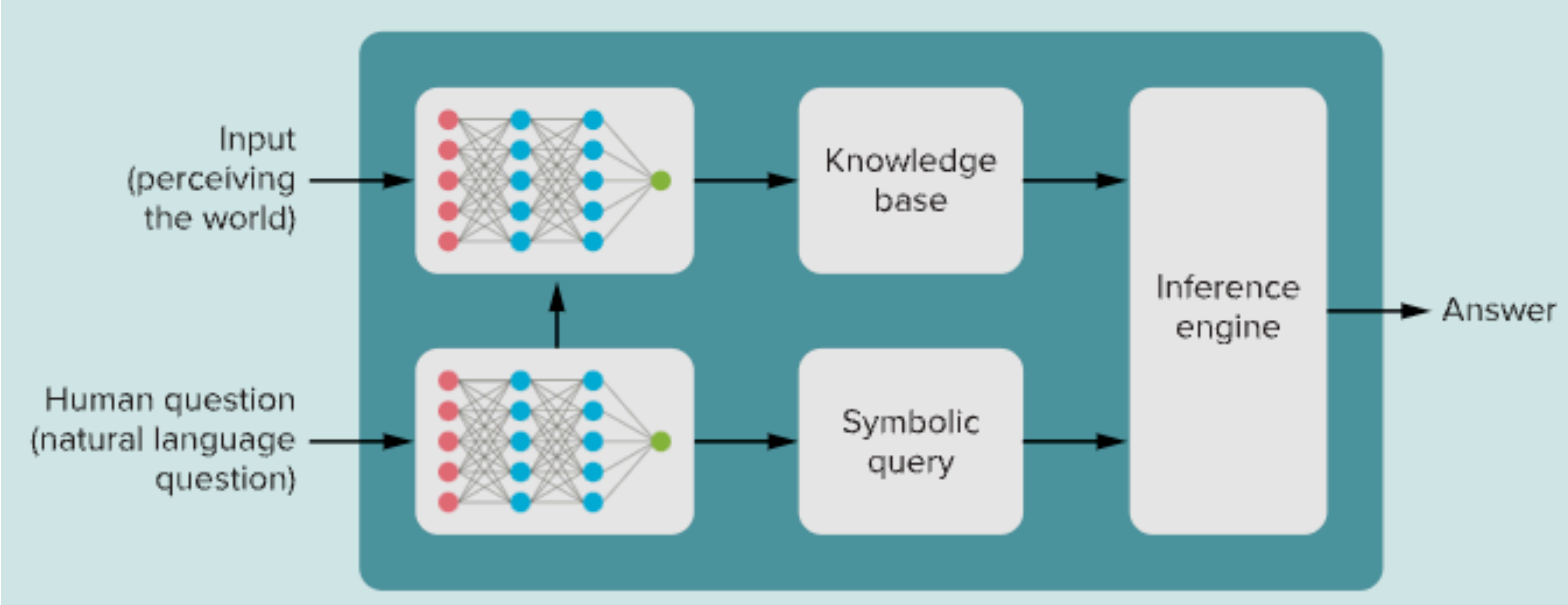


Subsymbolic AI



*Gradient descent is analogous to the delta-rule

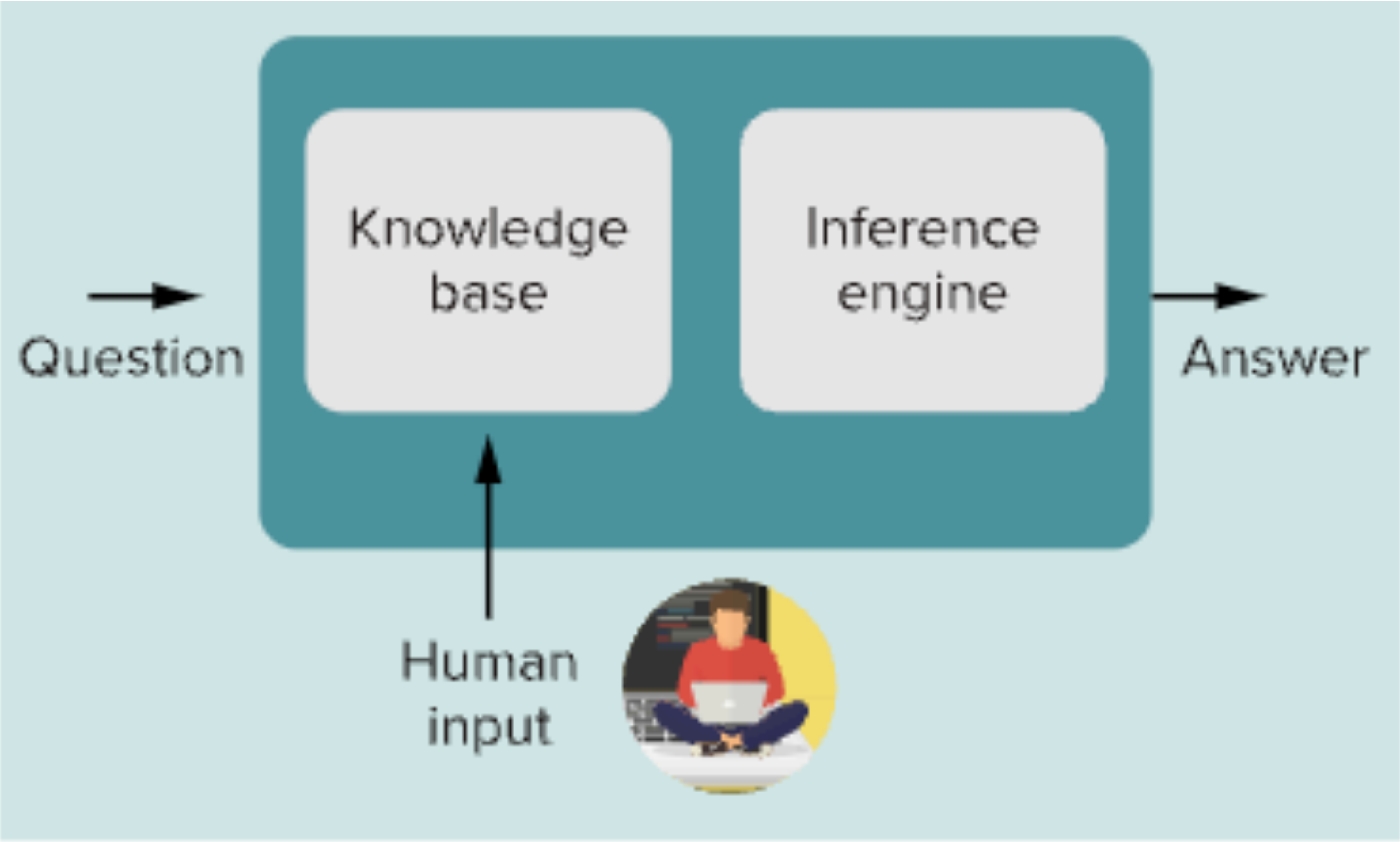
Hybrid systems



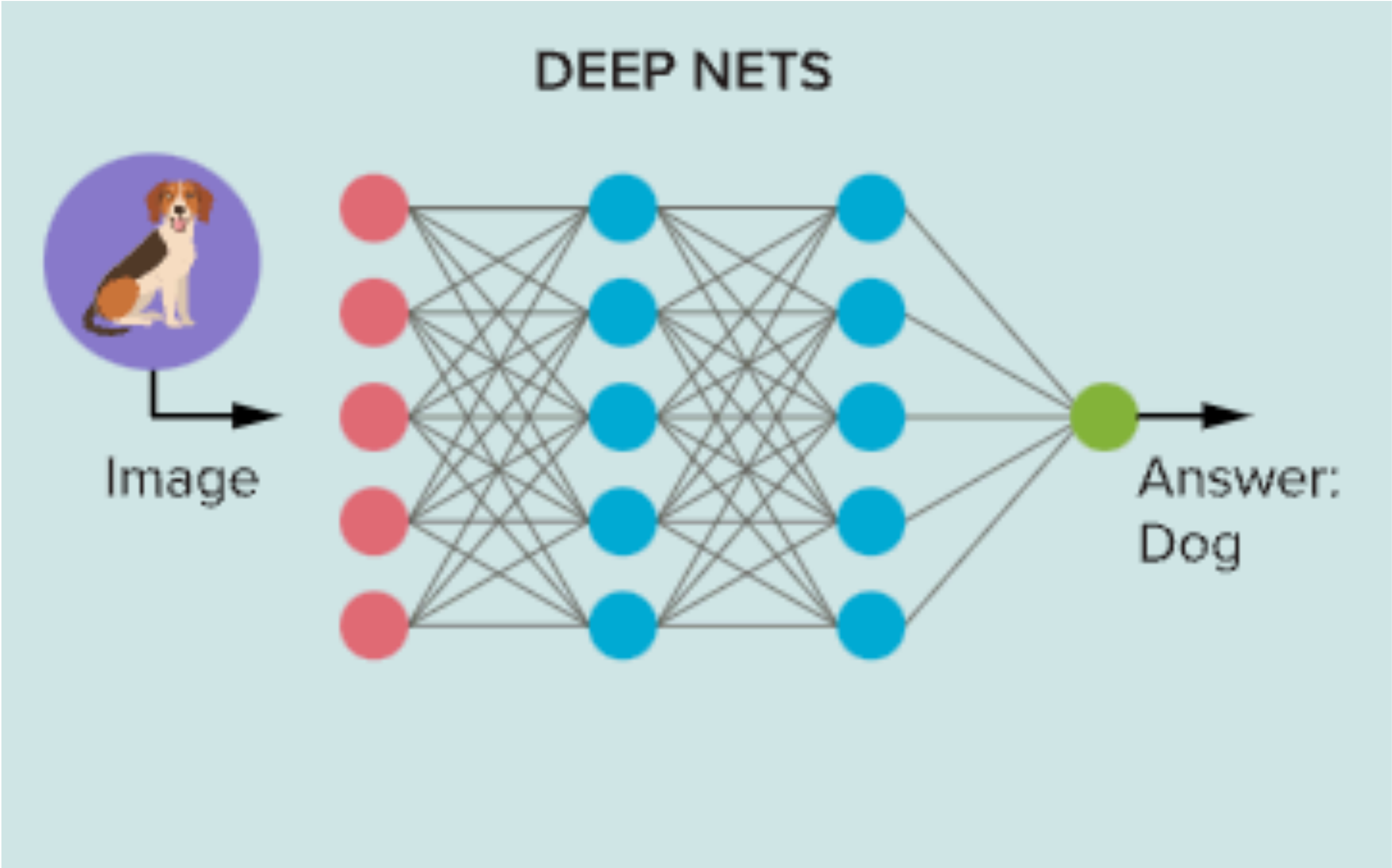
Symbolic vs. Subsymbolic AI

Physical symbol system hypothesis: manipulating **symbols** and **relations**

Symbolic AI

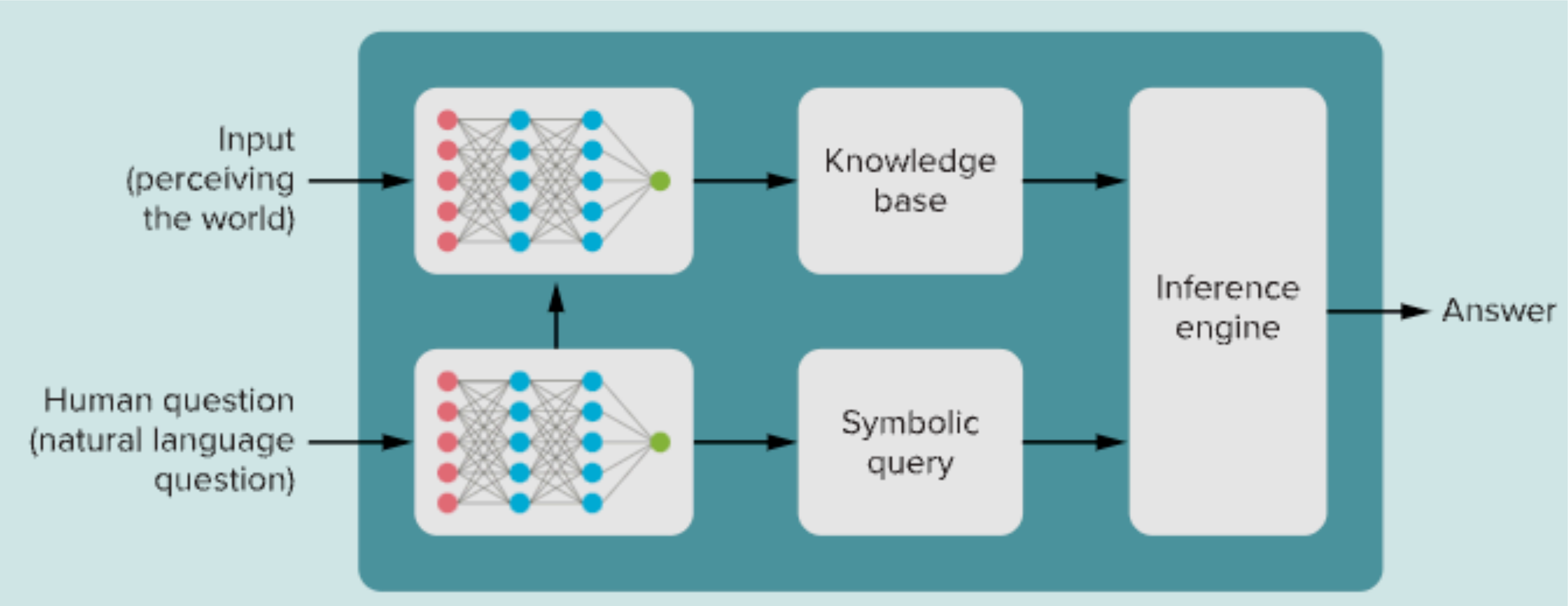


Subsymbolic AI



*Gradient descent is analogous to the delta-rule

Hybrid systems



Agenda for today

1. What is a concept?
2. Rule-based theories
3. Similarity-based theories
4. Hybrid approaches

What is a concept?

What is a concept?

Conceptual art

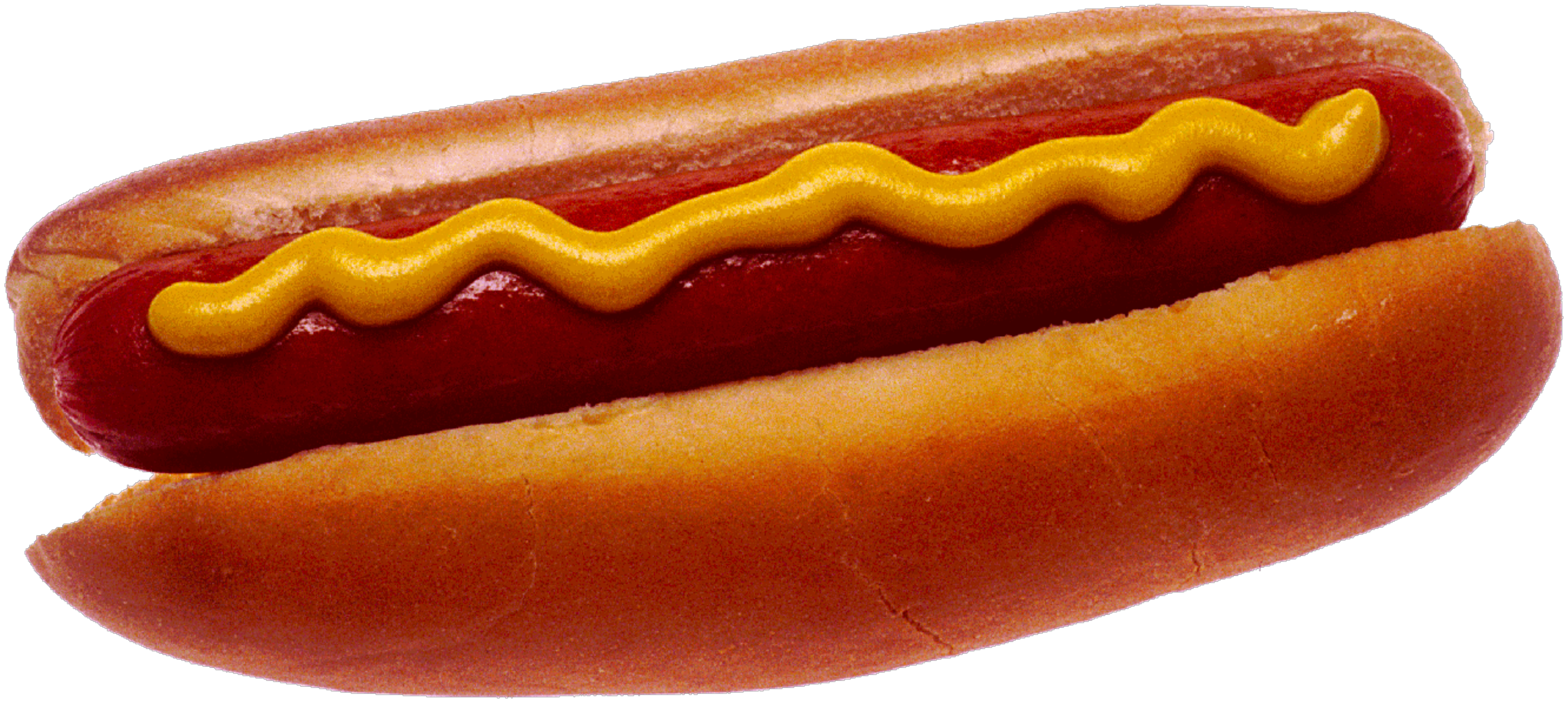
Maurizio Cattelan

Marcel Duchamp

Shutterstock

What is a “Sandwich?”

What is a “Sandwich?”



Is a hotdog a sandwich?

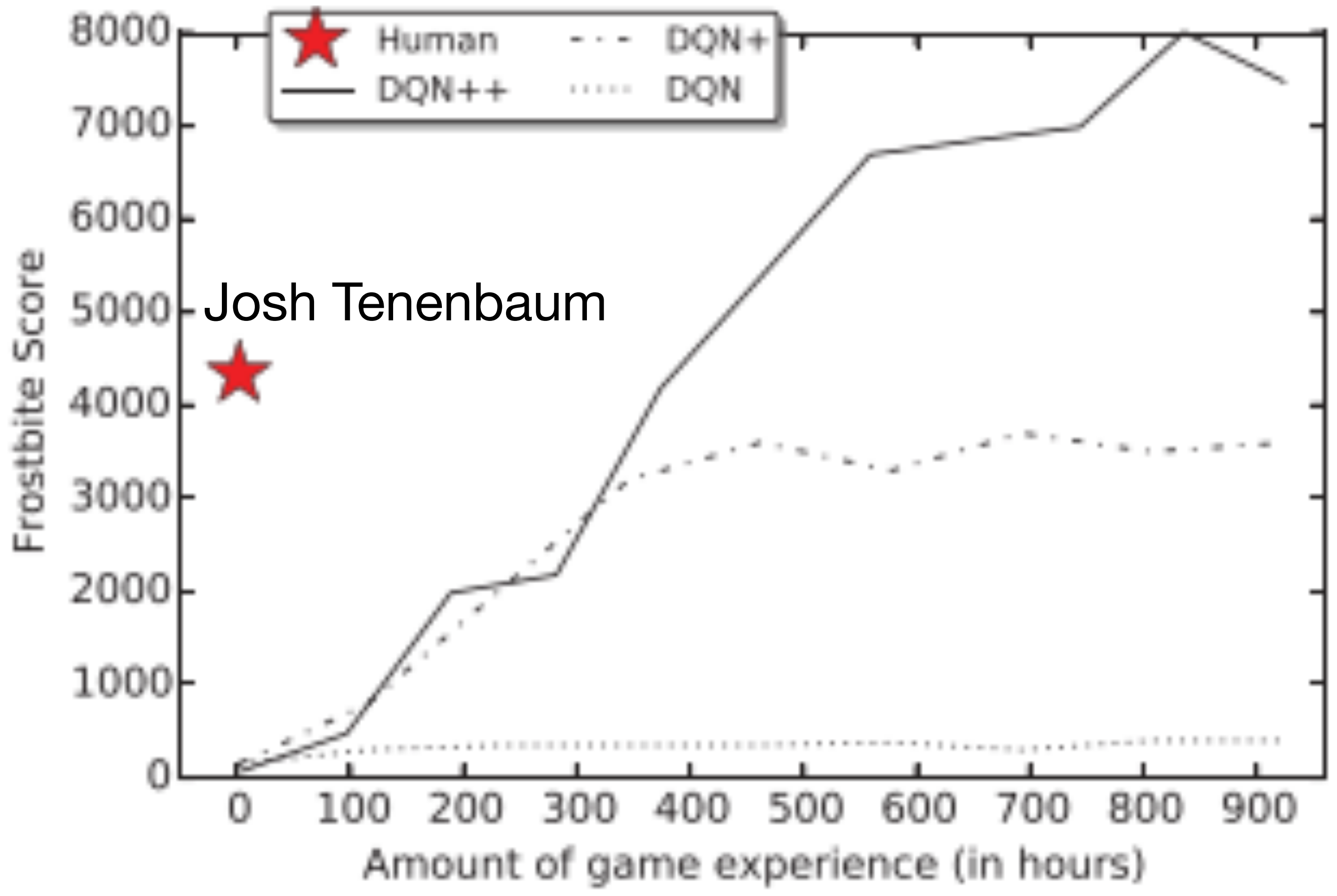
Concept learning is at the heart of many key aspects of intelligence

One-shot generalization

Lake et al., (2015); Lake et al., (2017)

Creative composition

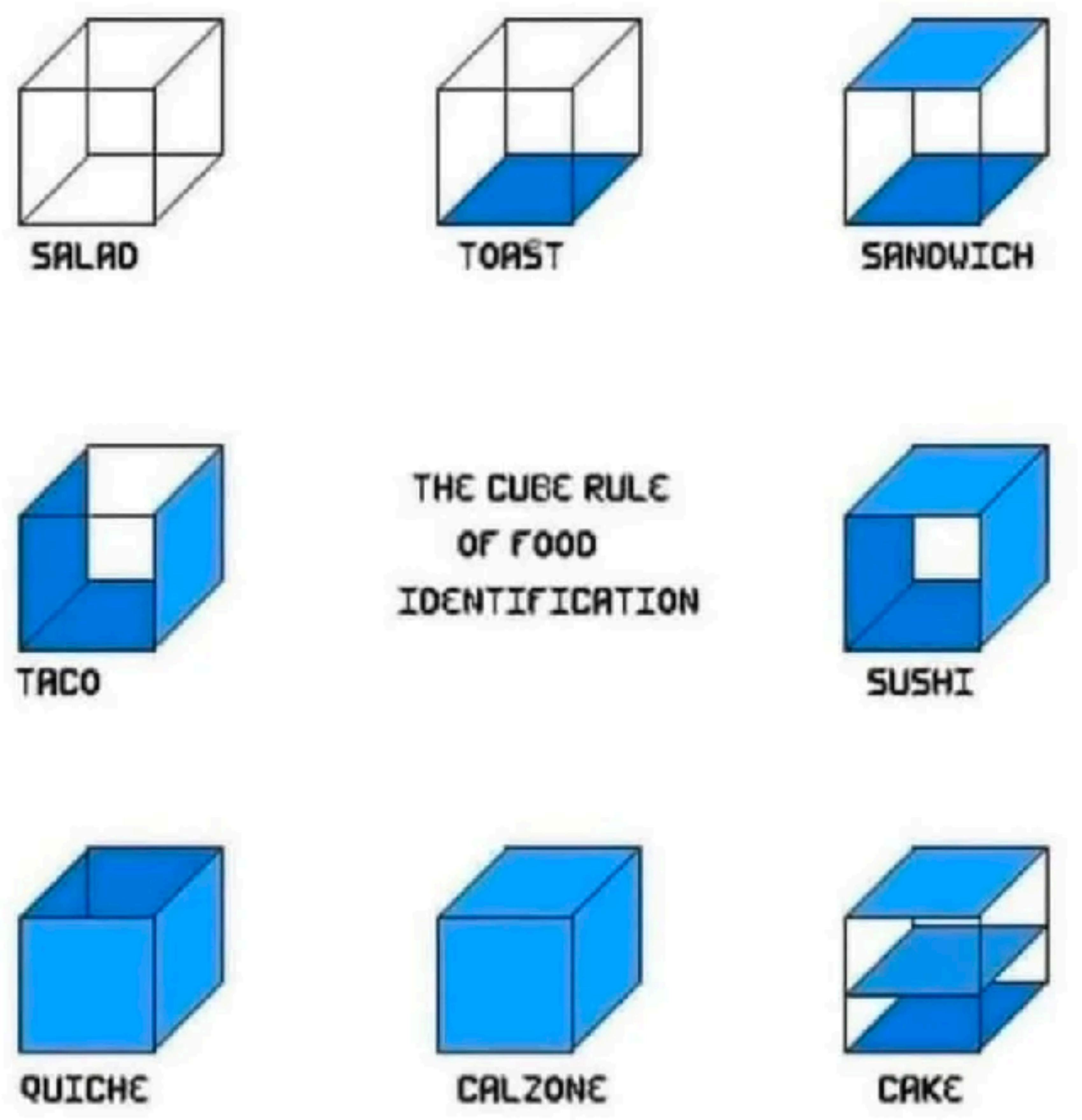
Rapid transfer



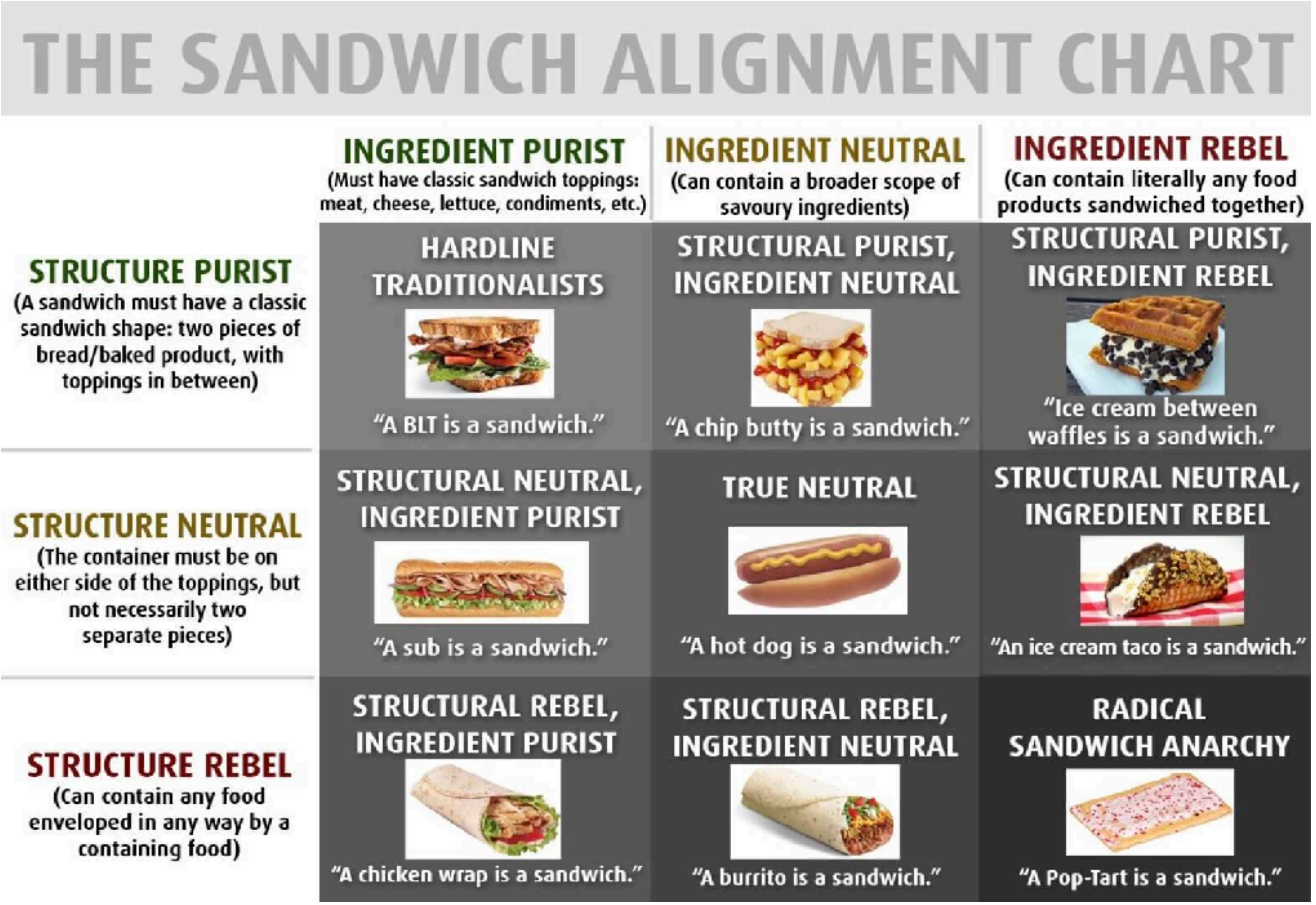
The study of categories and concepts

- A *category* is a set of objects in the world and a *concept* is a mental representation of a category
  - We will use the two interchangeably
- **Classical view** (Bruner et al., 1967):
  1. Concepts can be defined based on necessary and sufficient conditions for category membership
  2. Membership is all-or-nothing. All members are equally good
- This perspective dates to Aristotelian “forms” and Logical positivist philosophy (e.g., Quine, Popper, etc…)
- What are the necessary and sufficient conditions for something to be a sandwich?

Different approaches to defining concepts

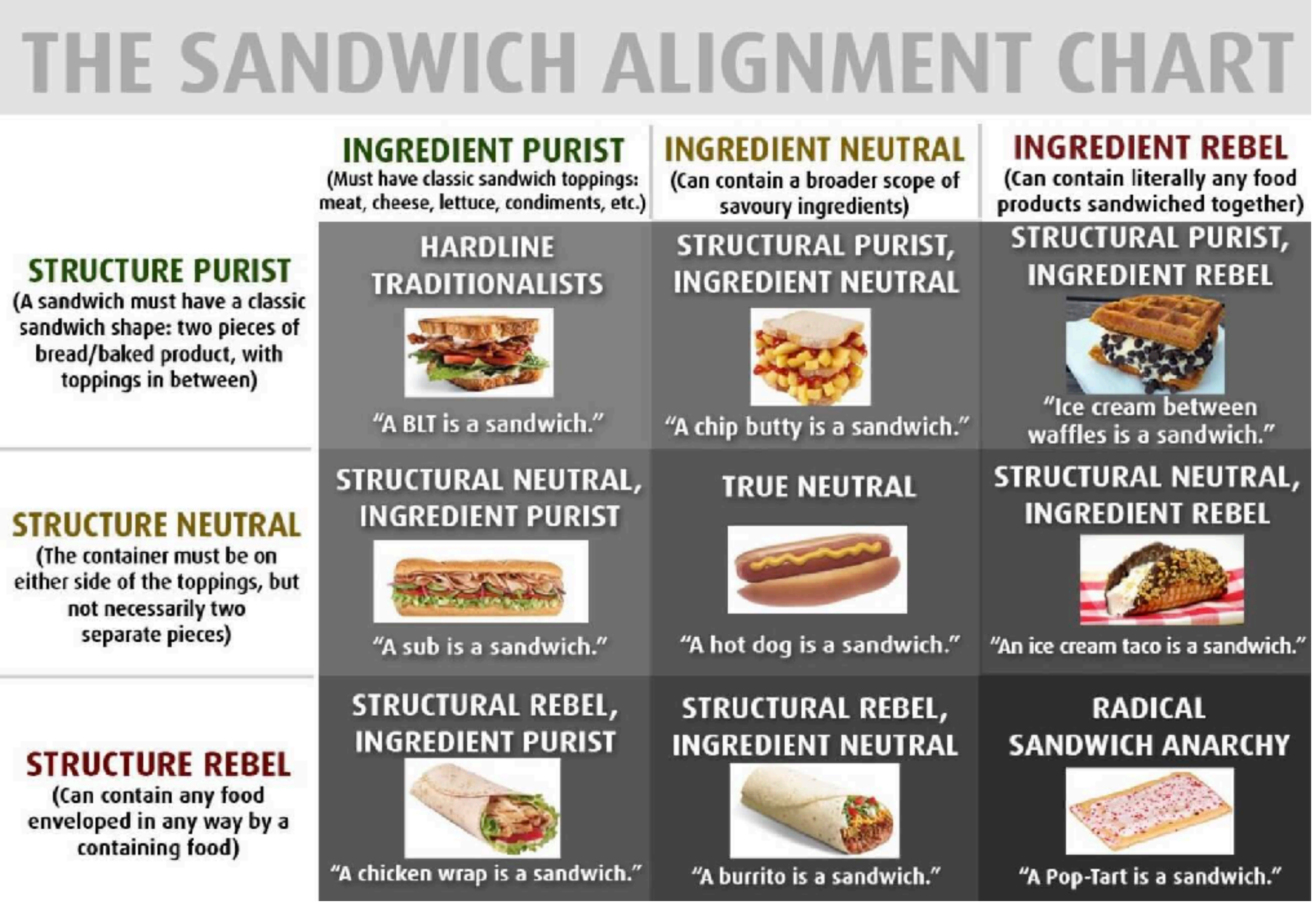


Different approaches to defining concepts



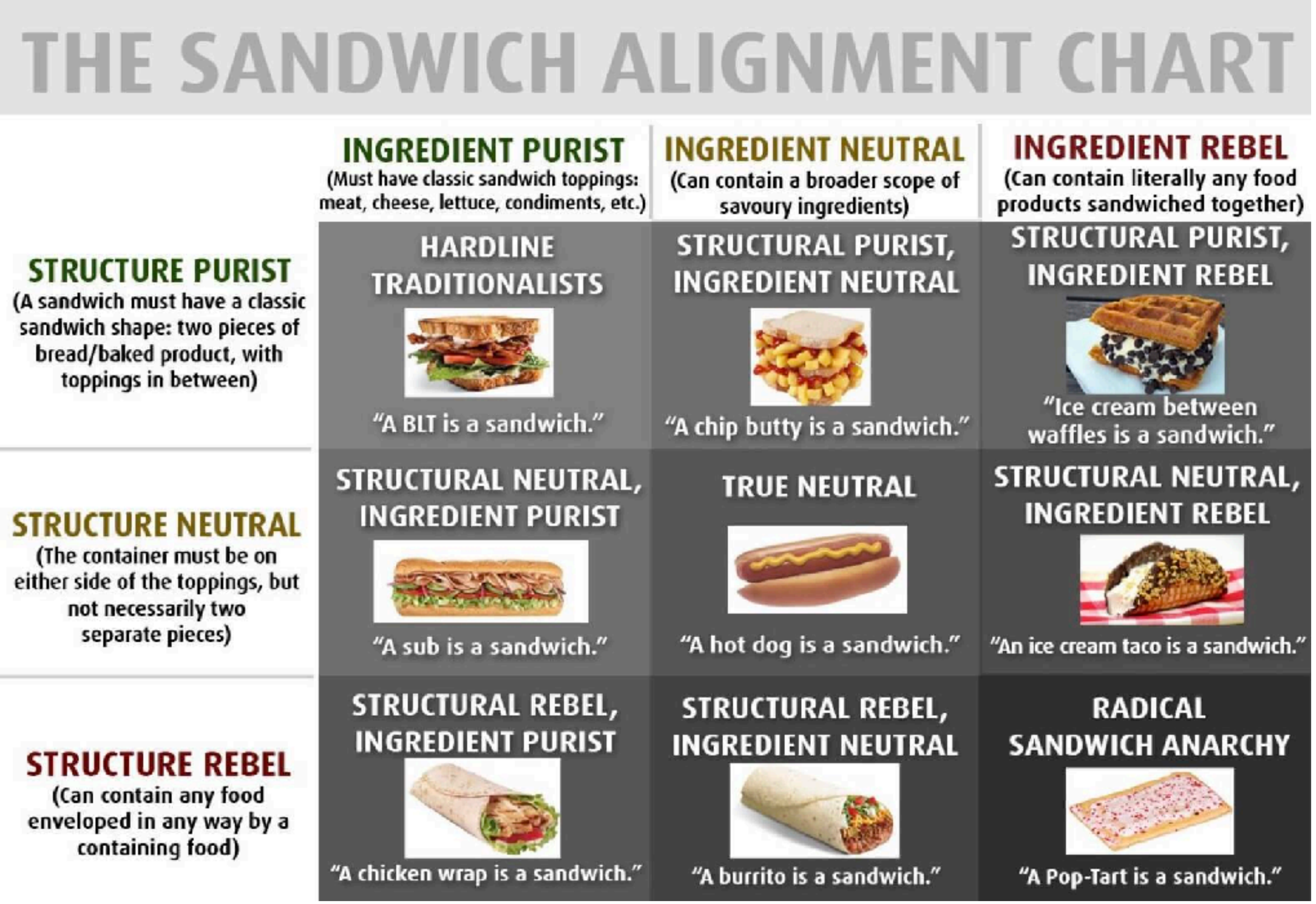
Different approaches to defining concepts

Rule-based approaches



Different approaches to defining concepts

Rule-based approaches

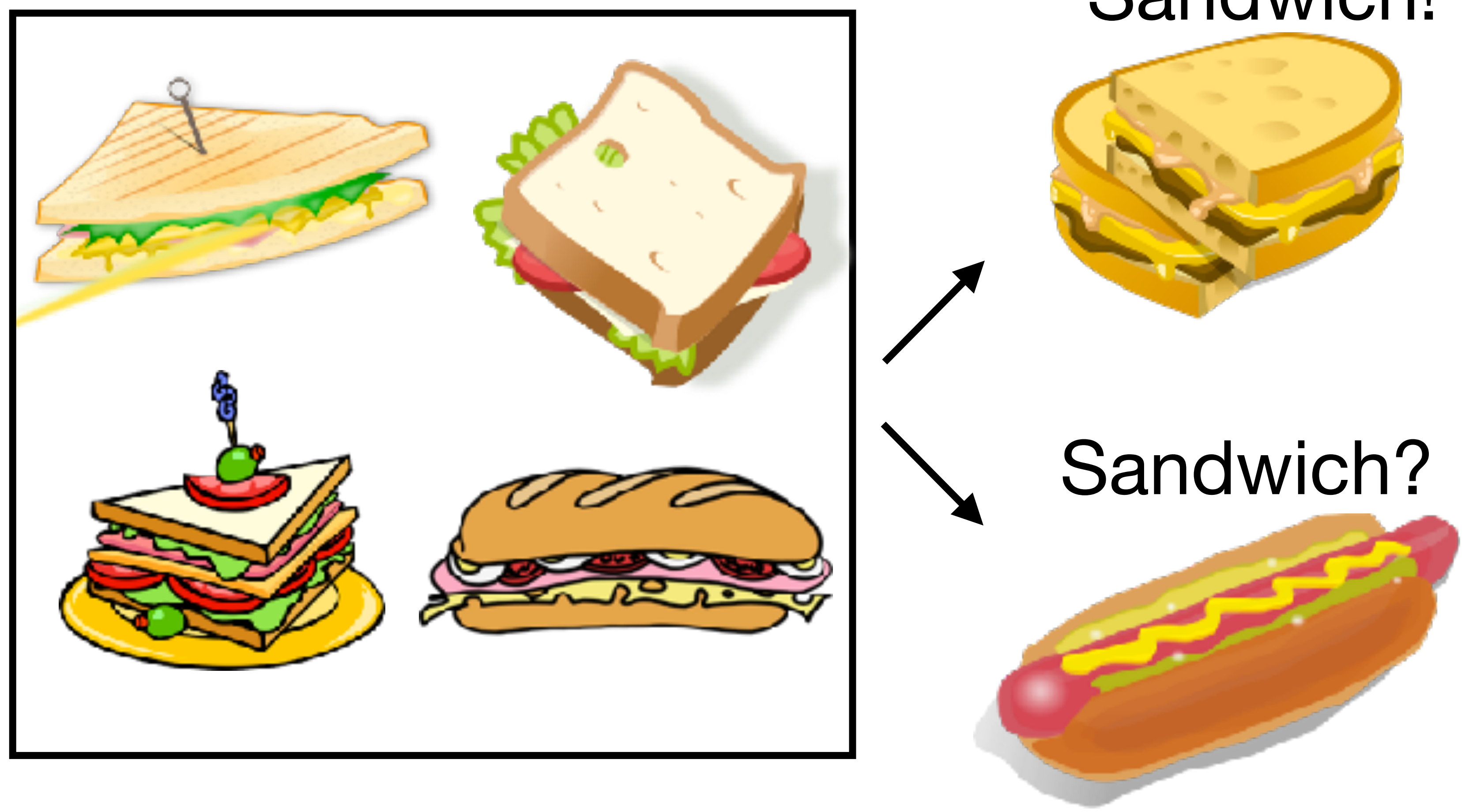


Similarity-bases approaches

Previous Experiences

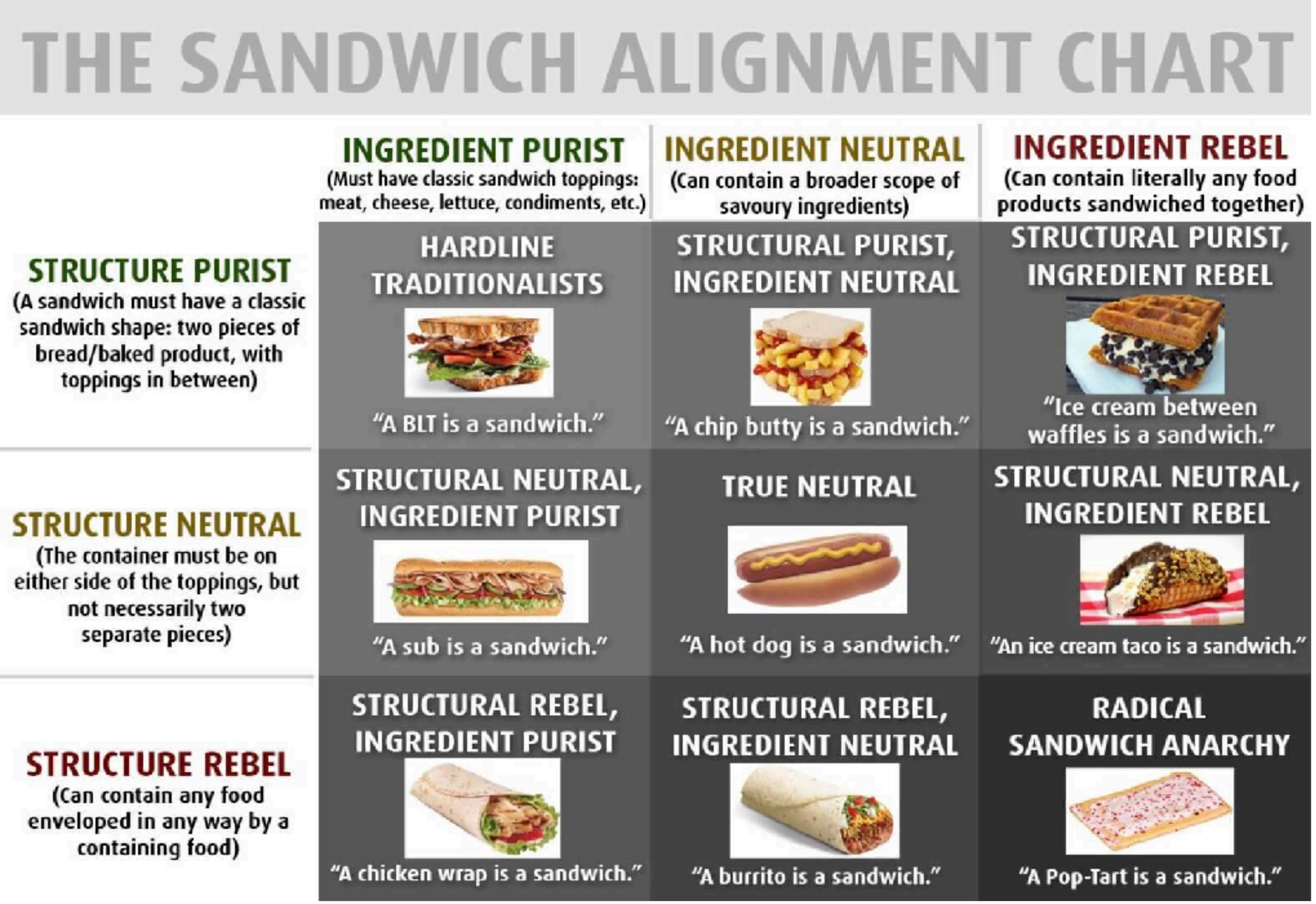
Sandwich!

Sandwich?



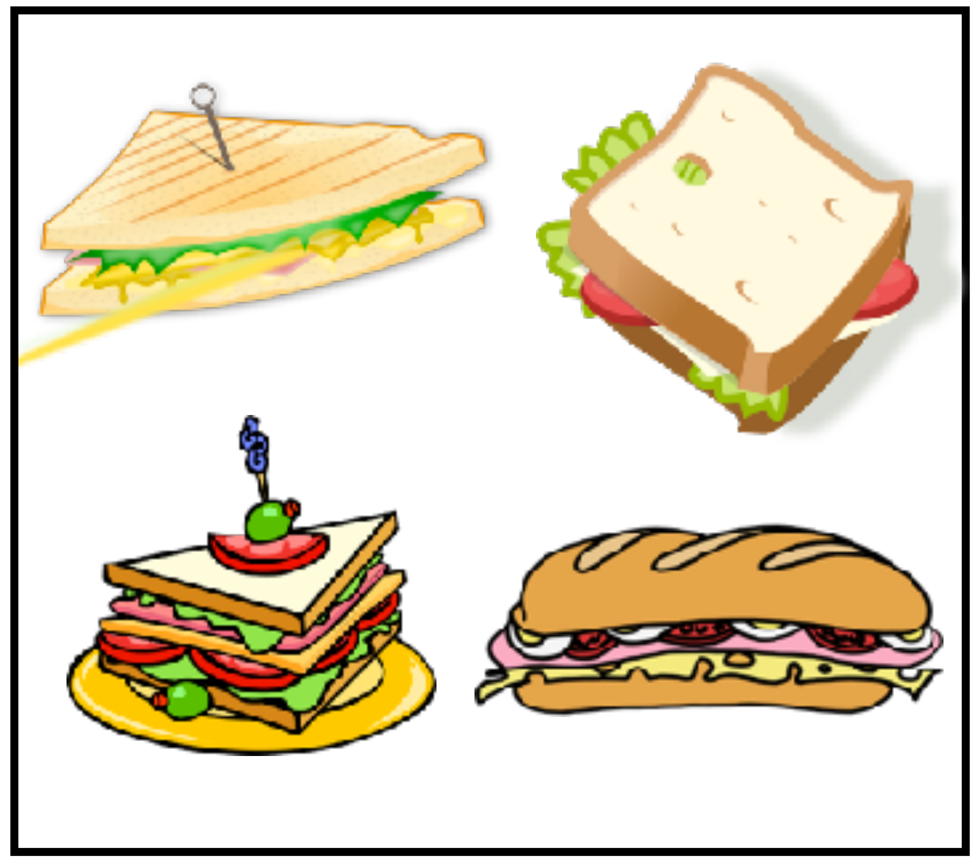
Different approaches to defining concepts

Rule-based approaches

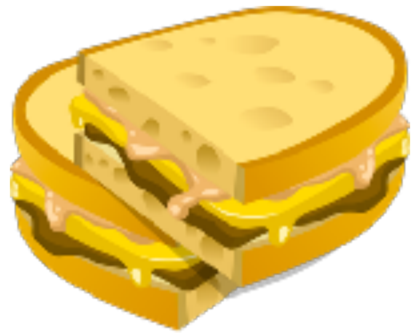


Similarity-bases approaches

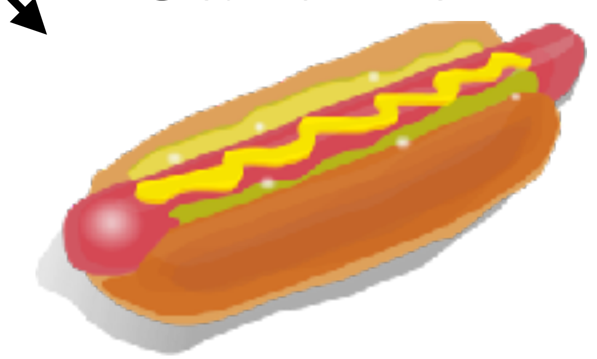
Previous Experiences



Sandwich!



Sandwich?



Rule-based theories

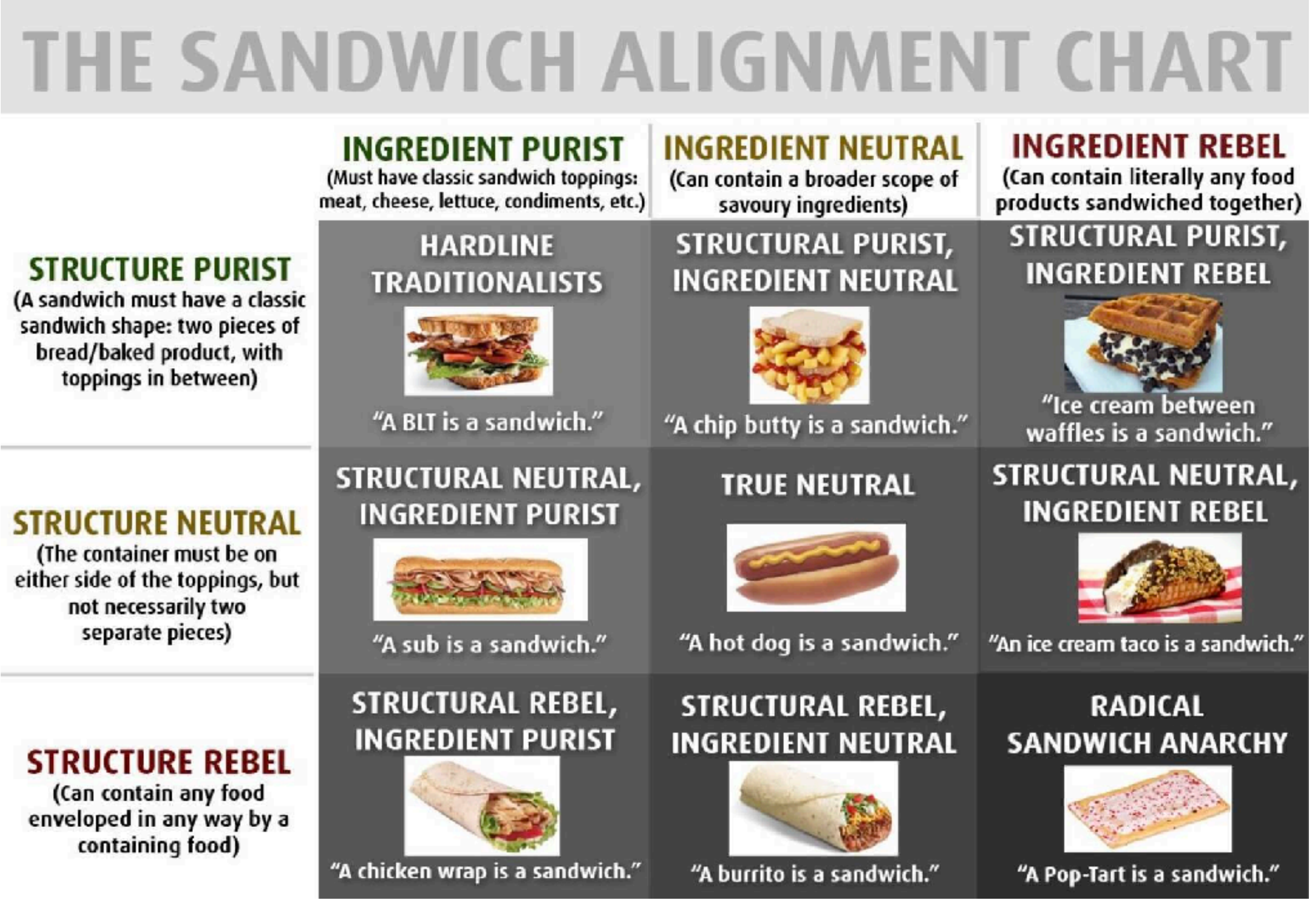
- Explicit boundaries of category membership (Ashby & Gott, 1988)
- *Specificity* facilitates rapid generalization
- Symbolic *compositionality* makes them infinitely productive (Goodman et al., 2008)
- Rigidity makes them inflexible
    - What about root beer? Or open-faced sandwiches?
- Even when accounting for exceptions to rules (Nosofsky et al,. 1994), they perform best when paired with other learning mechanisms (Erickson & Krushke, 1998; Ashby et al,. 1998; Love et al., 2004)

Furthermore, we wish to emphasize that in future in all cities, market-towns and in the country, ***the only ingredients used for the brewing of beer must be Barley, Hops and Water. -*** Reinheitsgebot (1516)

Rule-based theories

- Explicit boundaries of category membership (Ashby & Gott, 1988)
- *Specificity* facilitates rapid generalization
- Symbolic *compositionality* makes them infinitely productive (Goodman et al., 2008)
- Rigidity makes them inflexible
  - What about root beer? Or open-faced sandwiches?
- Even when accounting for exceptions to rules (Nosofsky et al,. 1994), they perform best when paired with other learning mechanisms (Erickson & Krushke, 1998; Ashby et al,. 1998; Love et al., 2004)

Furthermore, we wish to emphasize that in future in all cities, market-towns and in the country, ***the only ingredients used for the brewing of beer must be Barley, Hops and Water. -*** Reinheitsgebot (1516)



Hotdogs as a borderline item

- Early psychological experiments showed that people didn't have well-defined categories (Hampton, 1979; Rosch & Mervis, 1975) and were even inconsistent when labeling the same object twice (McCloskey & Glucksberg, 1978)
- Category boundaries seem to be fuzzy, can shift over time, and can also be sensitive to context

Hotdogs as a borderline item

- Early psychological experiments showed that people didn't have well-defined categories (Hampton, 1979; Rosch & Mervis, 1975) and were even inconsistent when labeling the same object twice (McCloskey & Glucksberg, 1978)
- Category boundaries seem to be fuzzy, can shift over time, and can also be sensitive to context

Is an olive a fruit?

Hotdogs as a borderline item

- Early psychological experiments showed that people didn’t have well-defined categories (Hampton, 1979; Rosch & Mervis, 1975) and were even inconsistent when labeling the same object twice (McCloskey & Glucksberg, 1978)
- Category boundaries seem to be fuzzy, can shift over time, and can also be sensitive to context

Is an olive a fruit?

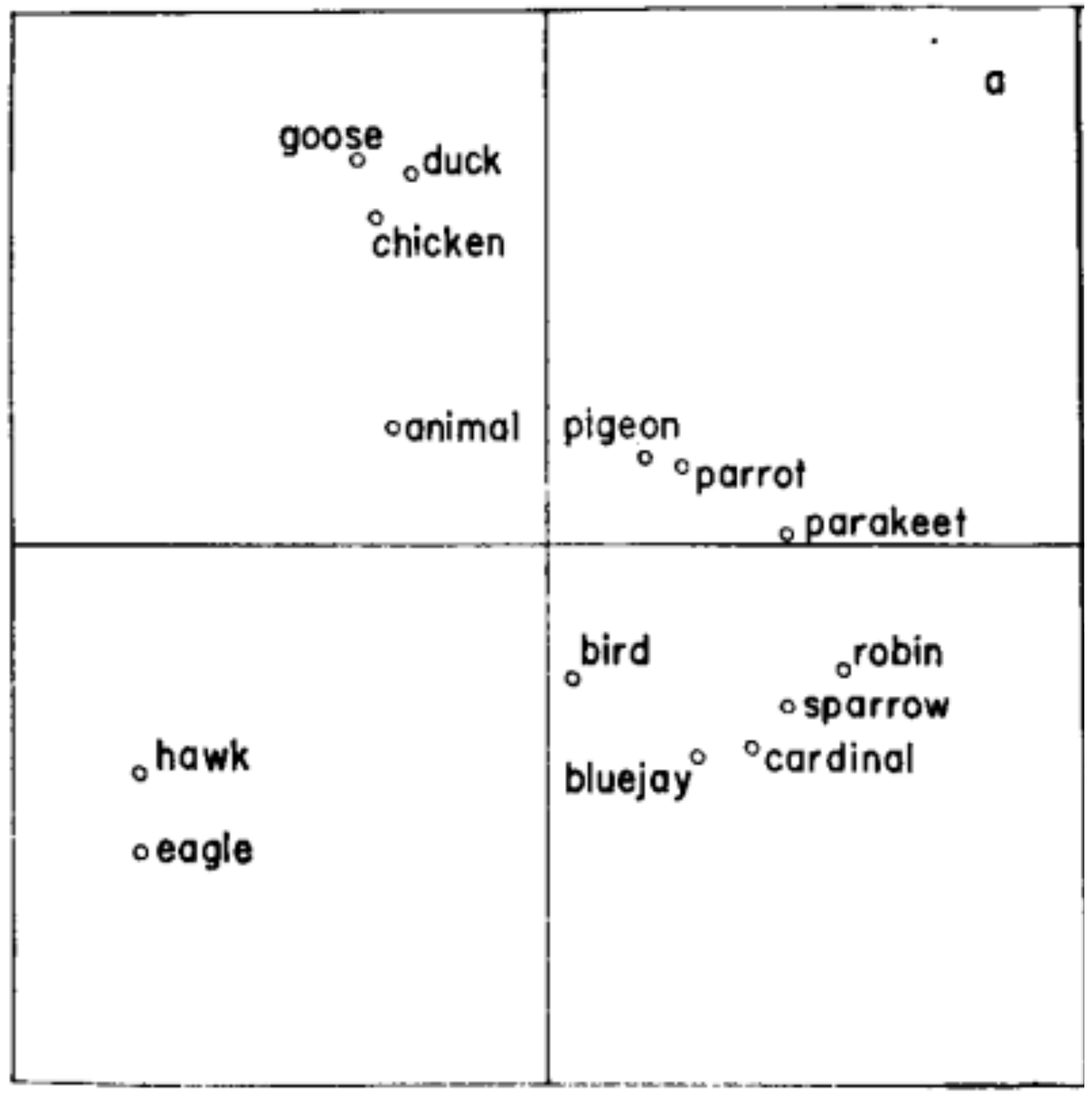
Hotdogs as a borderline item

- Early psychological experiments showed that people didn't have well-defined categories (Hampton, 1979; Rosch & Mervis, 1975) and were even inconsistent when labeling the same object twice (McCloskey & Glucksberg, 1978)
- Category boundaries seem to be fuzzy, can shift over time, and can also be sensitive to context
    - ~~Concepts can be defined based on necessary and sufficient conditions for category membership~~

Is an olive a fruit?

Typicality

- Some objects seem to fit better into categories than others
  - Some are more typical than others
- Frequency of experience has some effect, but not the most important variable (Rosch, Simpson, & Miller 1976)
- *Family resemblance theory* (Rosch & Mervis, 1975):
  - Items are typical if they
    a) have features frequent in the category
    b) don't have features frequent in other categories
- Thus, rather than hard & fast **rules**, **similarity** to typical items matters

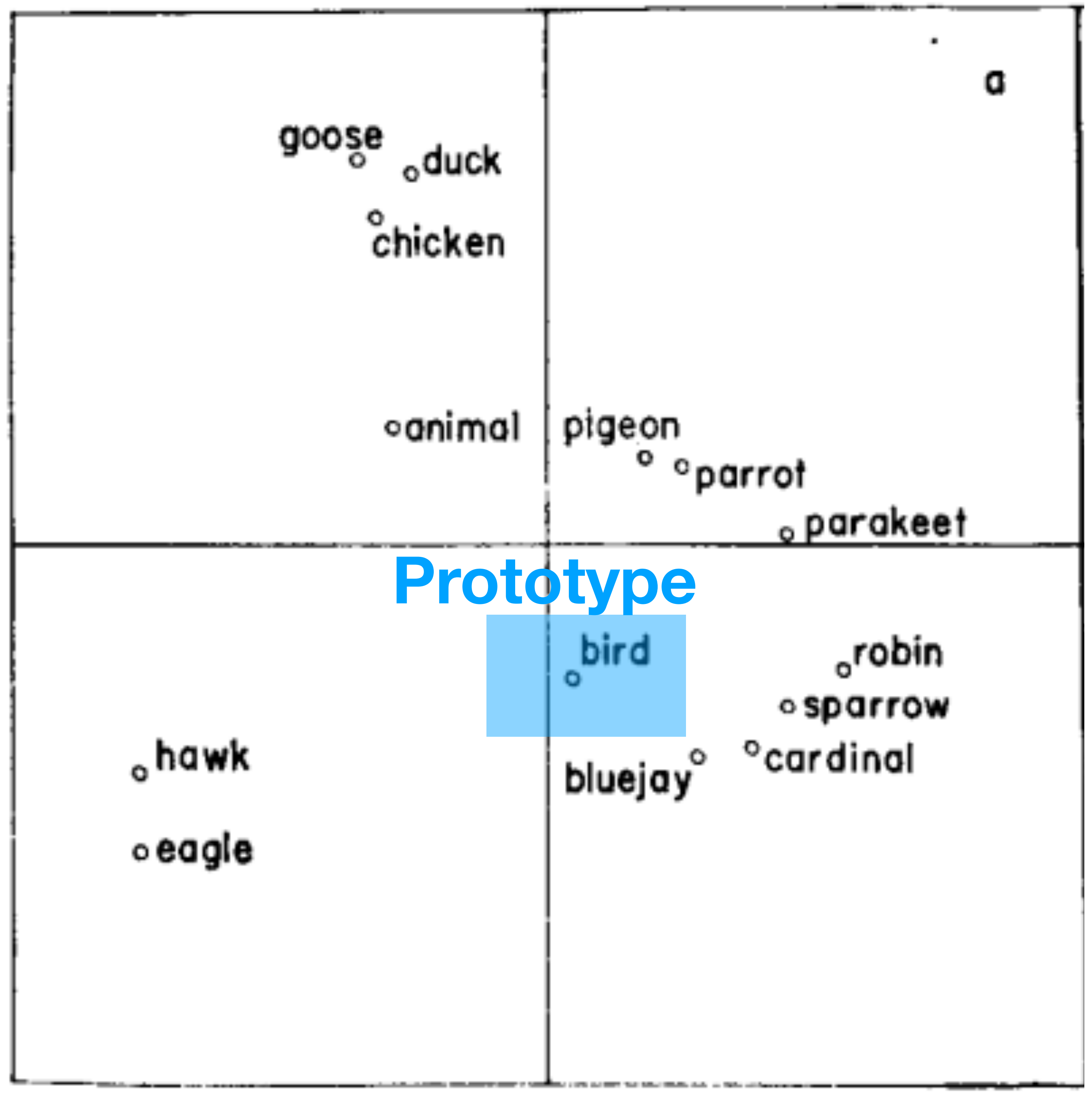


Multi-dimensional scaling of similarity ratings from Rips, Shoben, & Smith (1973)

Armstrong, Gleitman, & Gleitman (1983)

Typicality

- Some objects seem to fit better into categories than others
  - Some are more typical than others
- Frequency of experience has some effect, but not the most important variable (Rosch, Simpson, & Miller 1976)
- *Family resemblance theory* (Rosch & Mervis, 1975):
  - Items are typical if they
    a) have features frequent in the category
    b) don't have features frequent in other categories
- Thus, rather than hard & fast **rules**, **similarity** to typical items matters

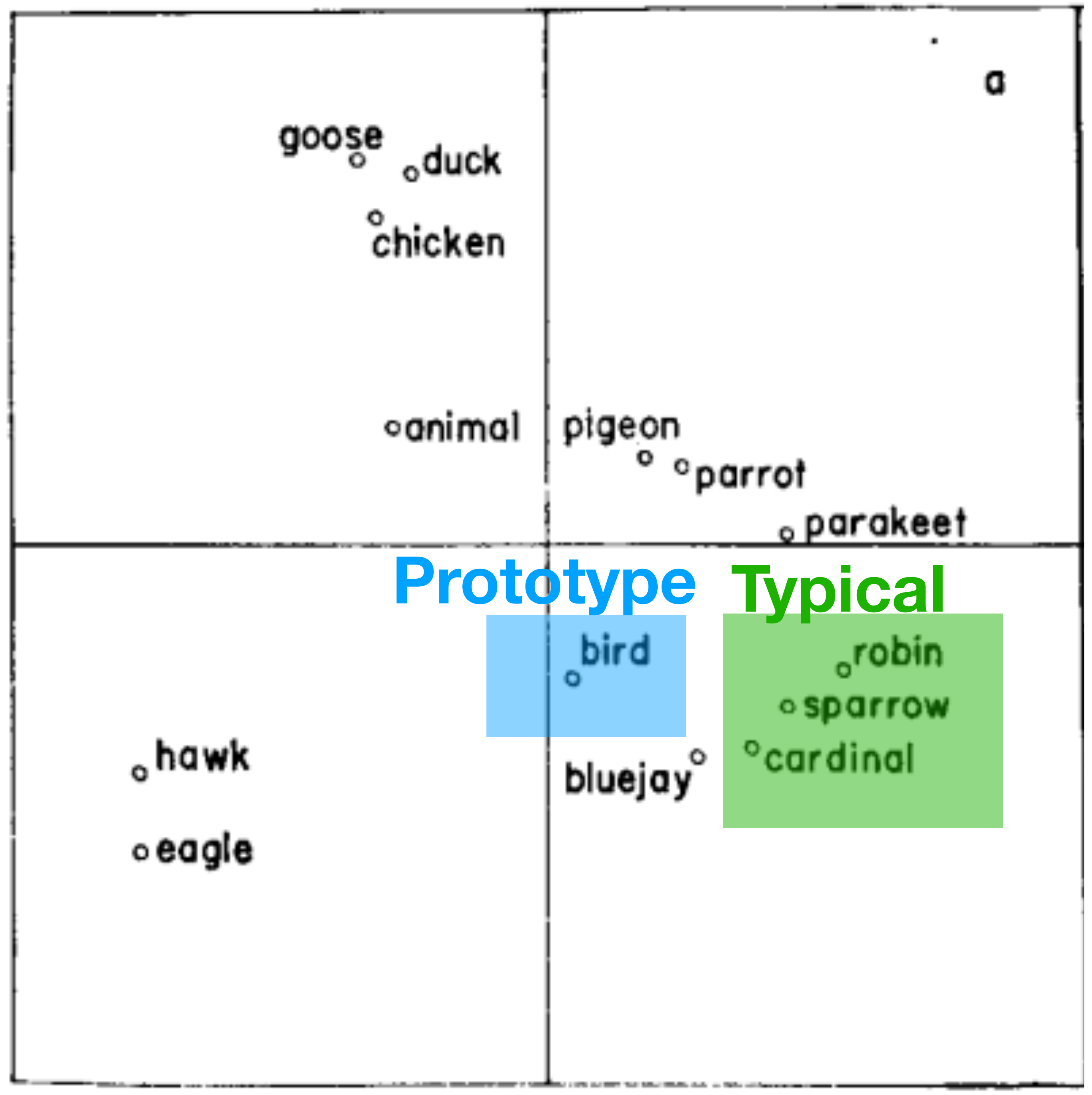


Multi-dimensional scaling of similarity ratings from Rips, Shoben, & Smith (1973)

Armstrong, Gleitman, & Gleitman (1983)

Typicality

- Some objects seem to fit better into categories than others
  - Some are more typical than others
- Frequency of experience has some effect, but not the most important variable (Rosch, Simpson, & Miller 1976)
- *Family resemblance theory* (Rosch & Mervis, 1975):
  - Items are typical if they
    a) have features frequent in the category
    b) don't have features frequent in other categories
- Thus, rather than hard & fast **rules**, **similarity** to typical items matters

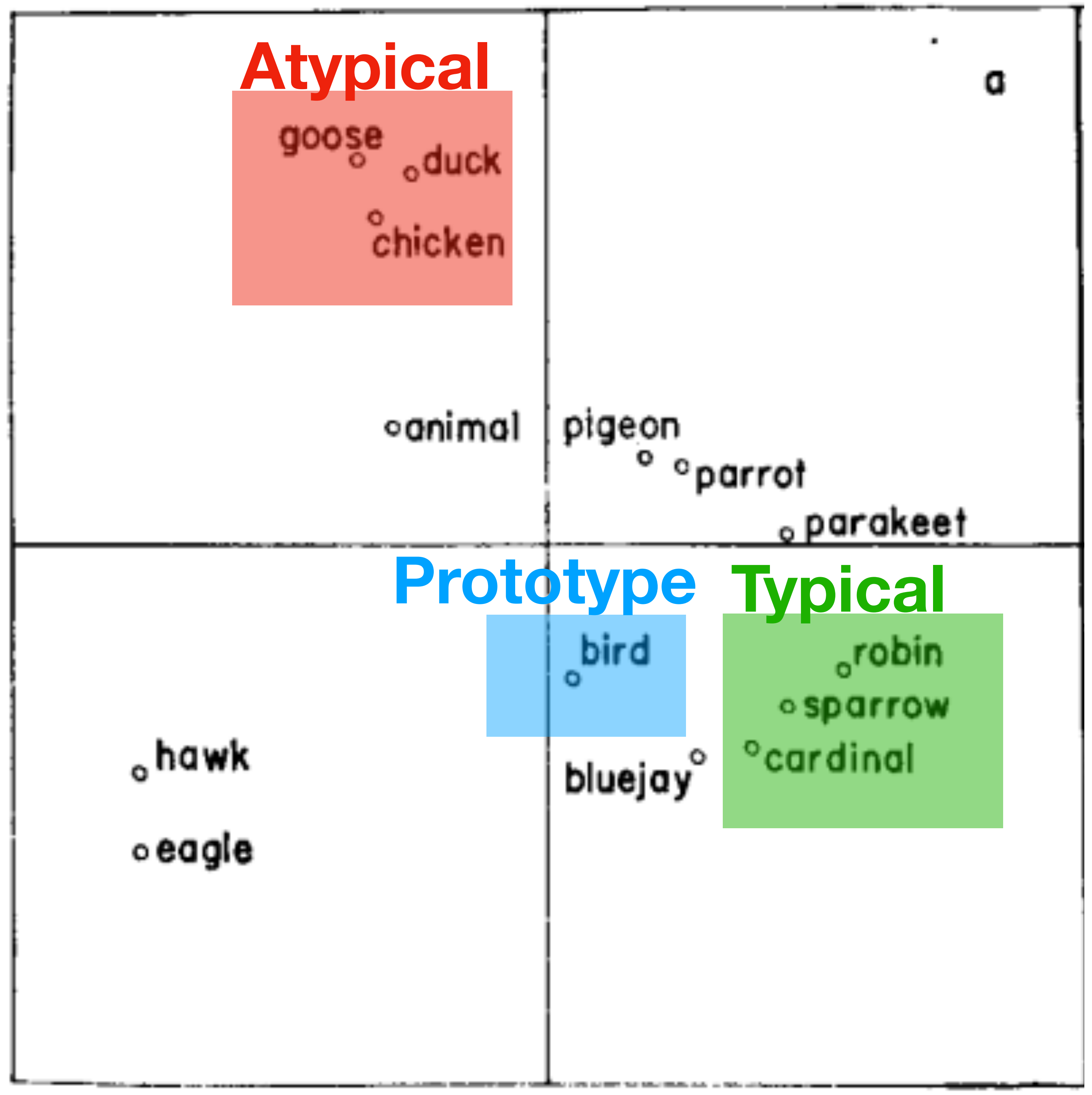


Multi-dimensional scaling of similarity ratings from Rips, Shoben, & Smith (1973)

Armstrong, Gleitman, & Gleitman (1983)

Typicality

- Some objects seem to fit better into categories than others
  - Some are more typical than others
- Frequency of experience has some effect, but not the most important variable (Rosch, Simpson, & Miller 1976)
- *Family resemblance theory* (Rosch & Mervis, 1975):
  - Items are typical if they
    a) have features frequent in the category
    b) don't have features frequent in other categories
- Thus, rather than hard & fast **rules**, **similarity** to typical items matters

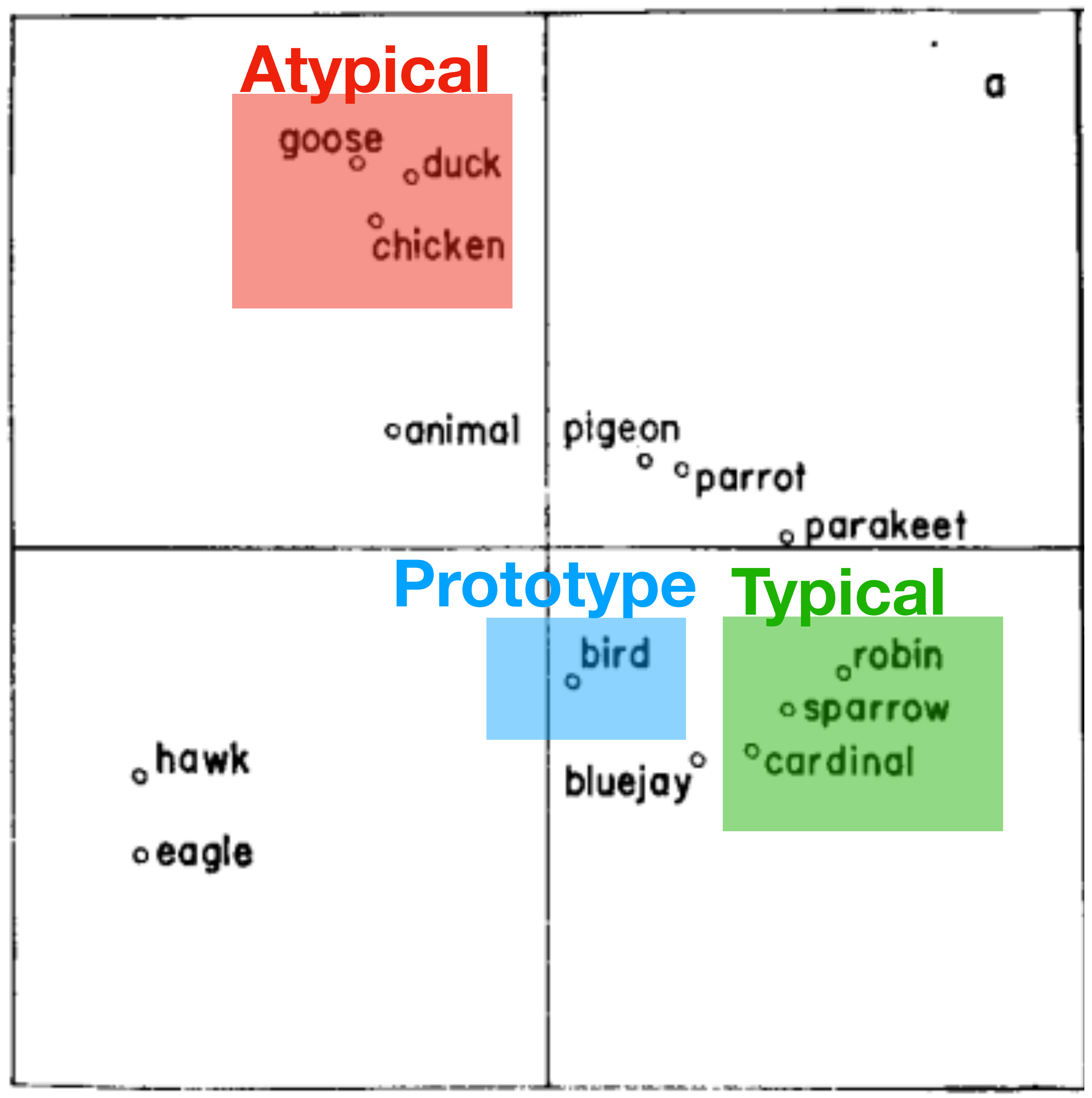


Multi-dimensional scaling of similarity ratings from Rips, Shoben, & Smith (1973)

Armstrong, Gleitman, & Gleitman (1983)

Typicality

- Some objects seem to fit better into categories than others
  - Some are more typical than others
- Frequency of experience has some effect, but not the most important variable (Rosch, Simpson, & Miller 1976)
- *Family resemblance theory* (Rosch & Mervis, 1975):
  - Items are typical if they
    a) have features frequent in the category
    b) don't have features frequent in other categories
- Thus, rather than hard & fast **rules**, **similarity** to typical items matters

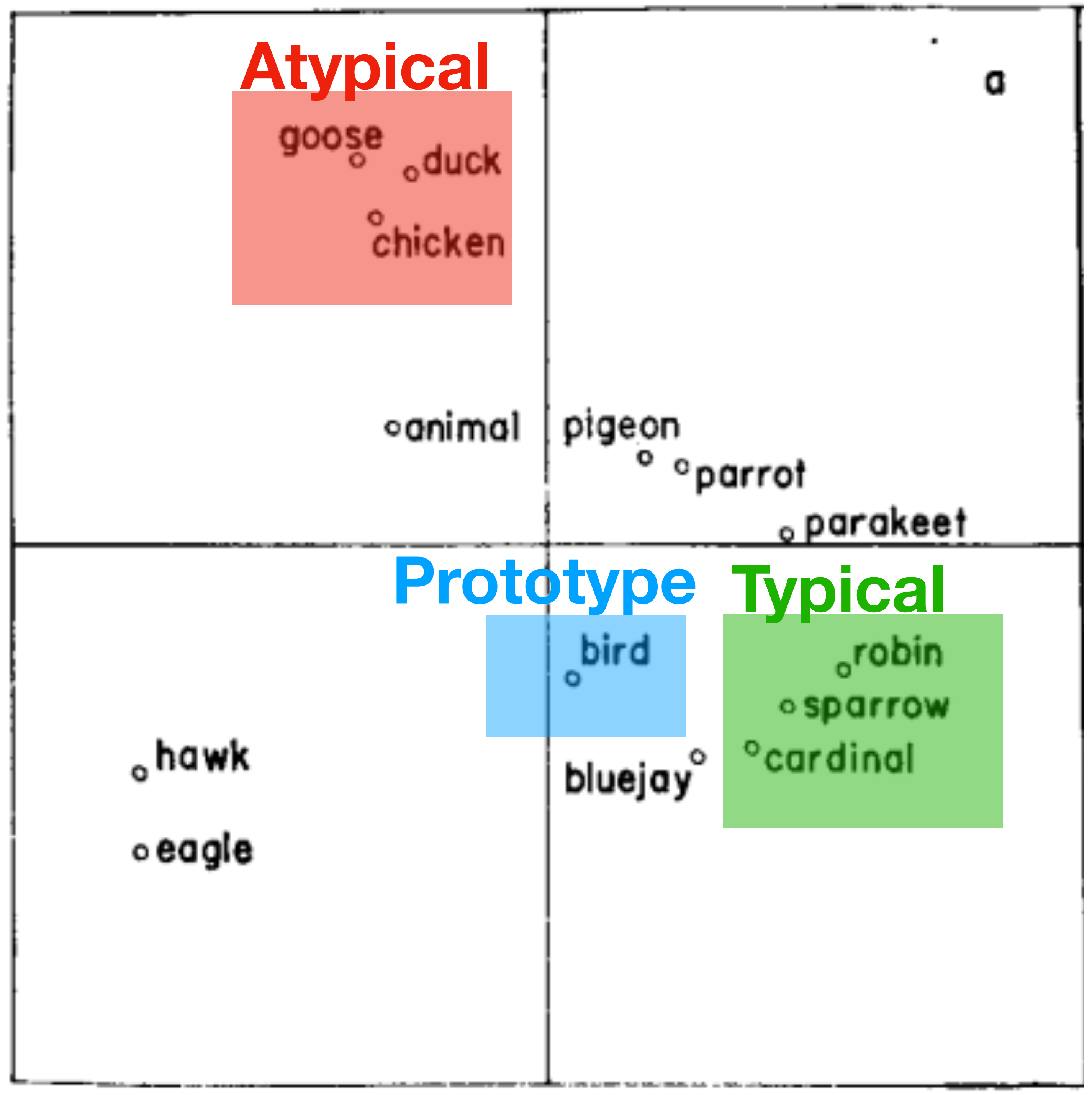


Multi-dimensional scaling of similarity ratings from Rips, Shoben, & Smith (1973)

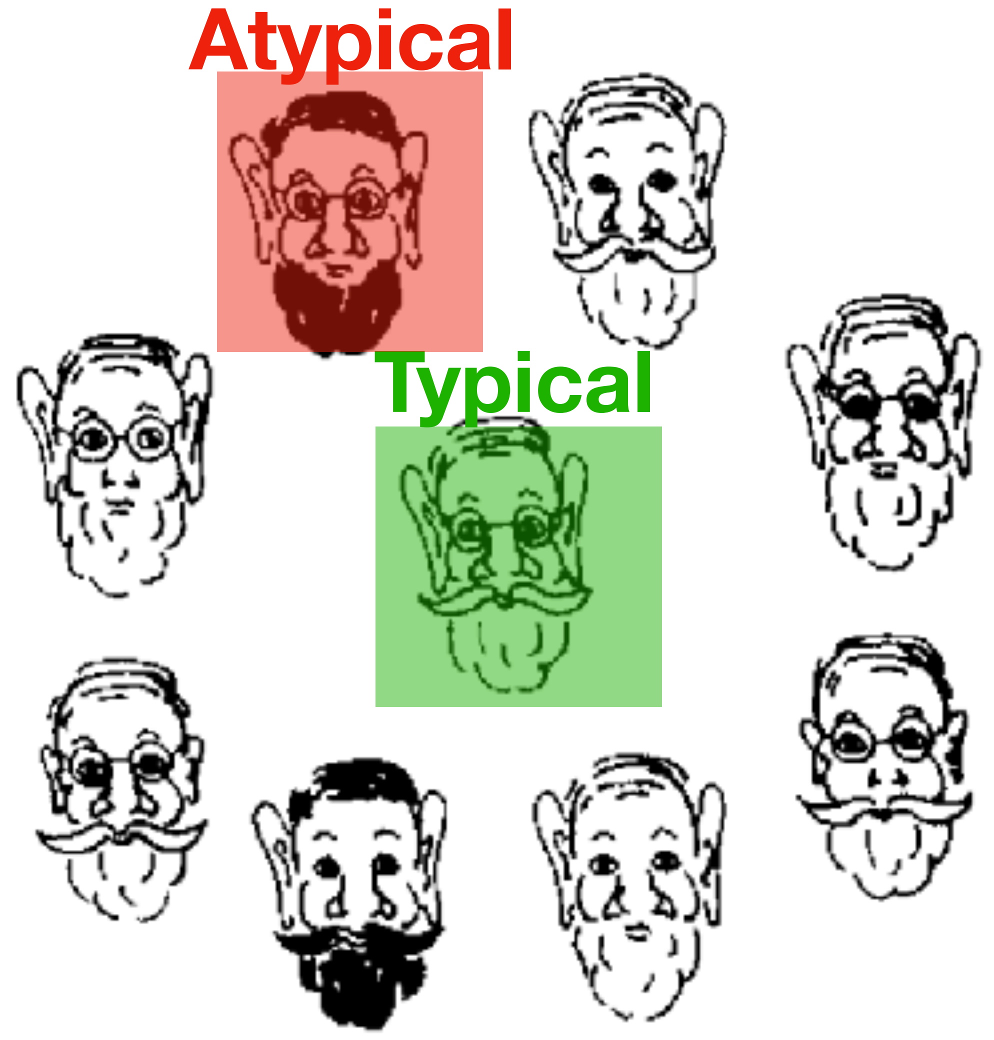
Armstrong, Gleitman, & Gleitman (1983)

Typicality

- Some objects seem to fit better into categories than others
    - Some are more typical than others
- Frequency of experience has some effect, but not the most important variable (Rosch, Simpson, & Miller 1976)
- *Family resemblance theory* (Rosch & Mervis, 1975):
    - Items are typical if they
      a) have features frequent in the category
      b) don't have features frequent in other categories
- Thus, rather than hard & fast **rules**, **similarity** to typical items matters



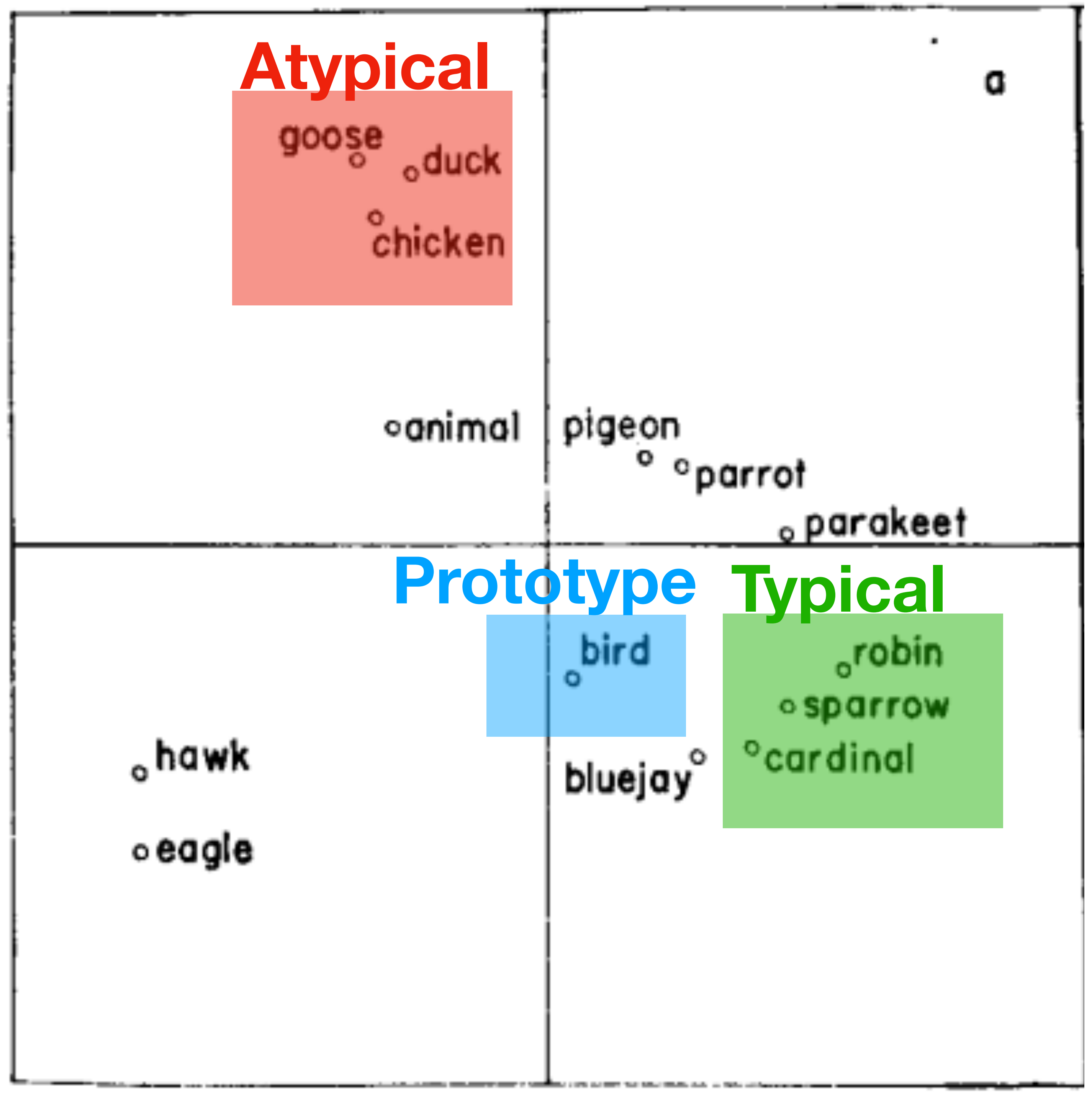
Multi-dimensional scaling of similarity ratings from Rips, Shoben, & Smith (1973)



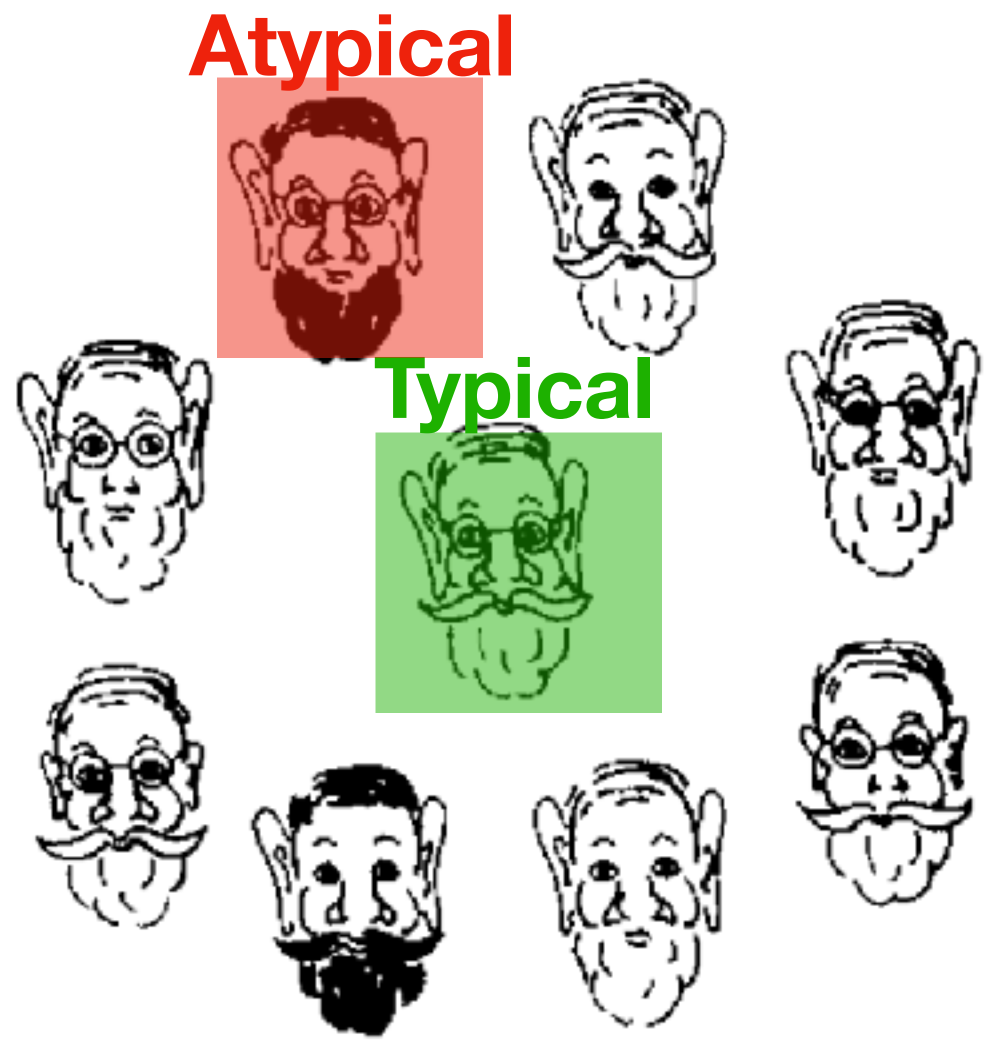
Armstrong, Gleitman, & Gleitman (1983)

Typicality

- Some objects seem to fit better into categories than others
    - Some are more typical than others
- Frequency of experience has some effect, but not the most important variable (Rosch, Simpson, & Miller 1976)
- *Family resemblance theory* (Rosch & Mervis, 1975):
    - Items are typical if they
      a) have features frequent in the category
      b) don't have features frequent in other categories
- Thus, rather than hard & fast **rules**, **similarity** to typical items matters
- ~~Membership is all-or-nothing. All members are equally good~~



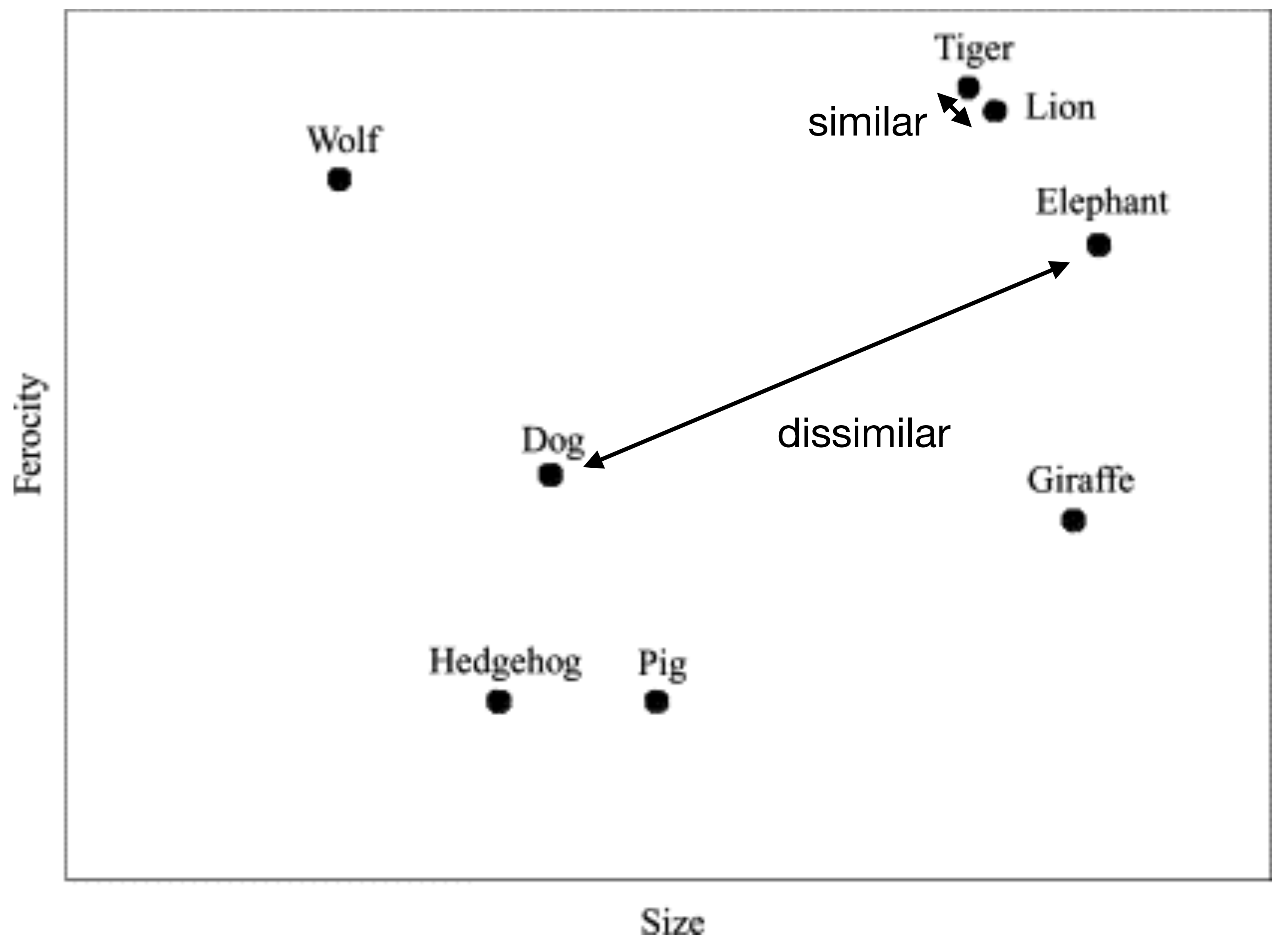
Multi-dimensional scaling of similarity ratings from Rips, Shoben, & Smith (1973)

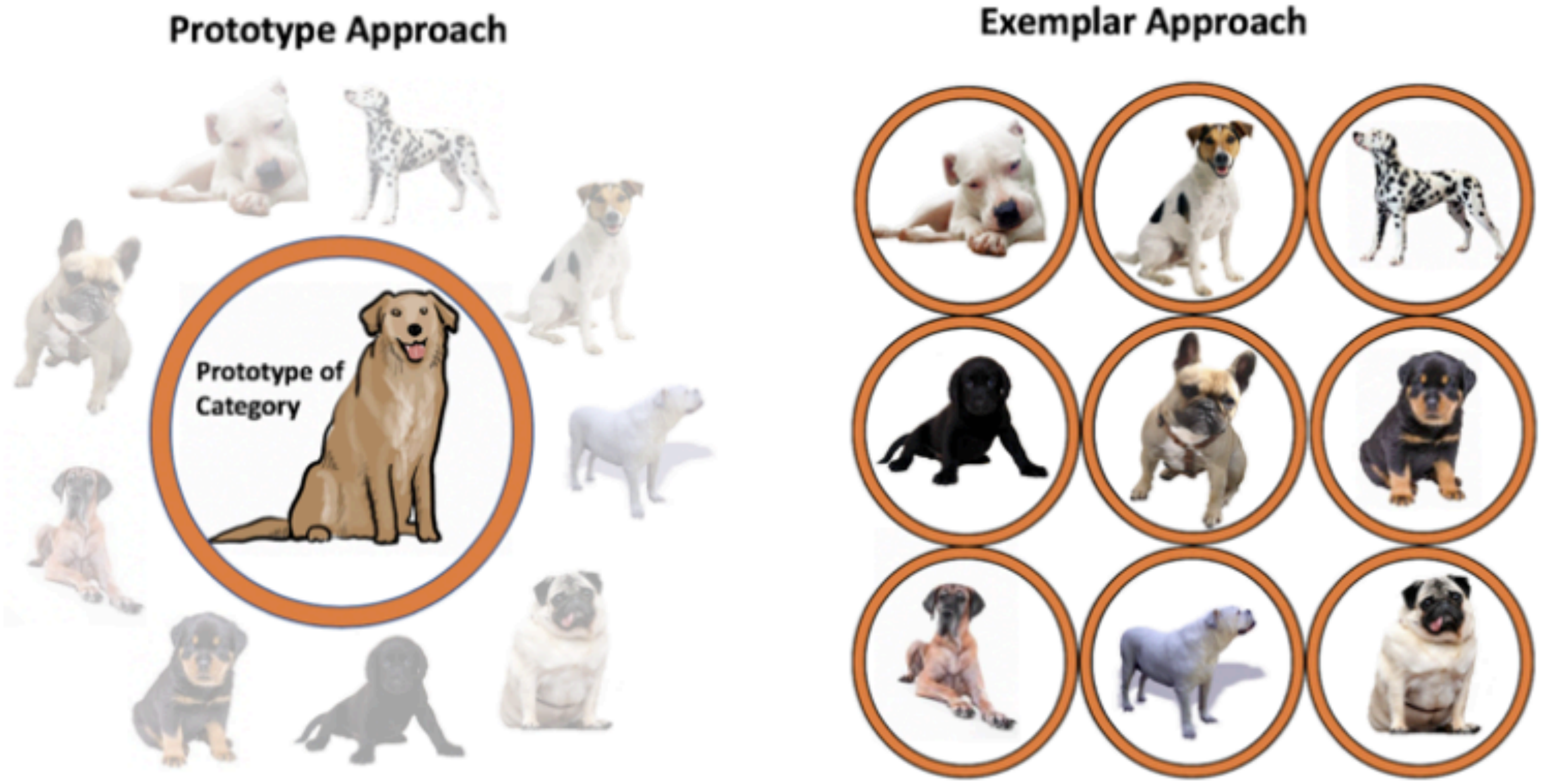


Armstrong, Gleitman, & Gleitman (1983)

Similarity-based theories

- Rather than hard and fast rules, another approach to concept learning proposes that we use similarity comparisons to make on the fly generalizations about new objects
- Stimuli with similar features are more likely to belong to the same category
  - distance in feature space provides a simple quantification of similarity
- Category membership is based on comparisor of any stimuli to previously learned *prototypes* or *exemplars*





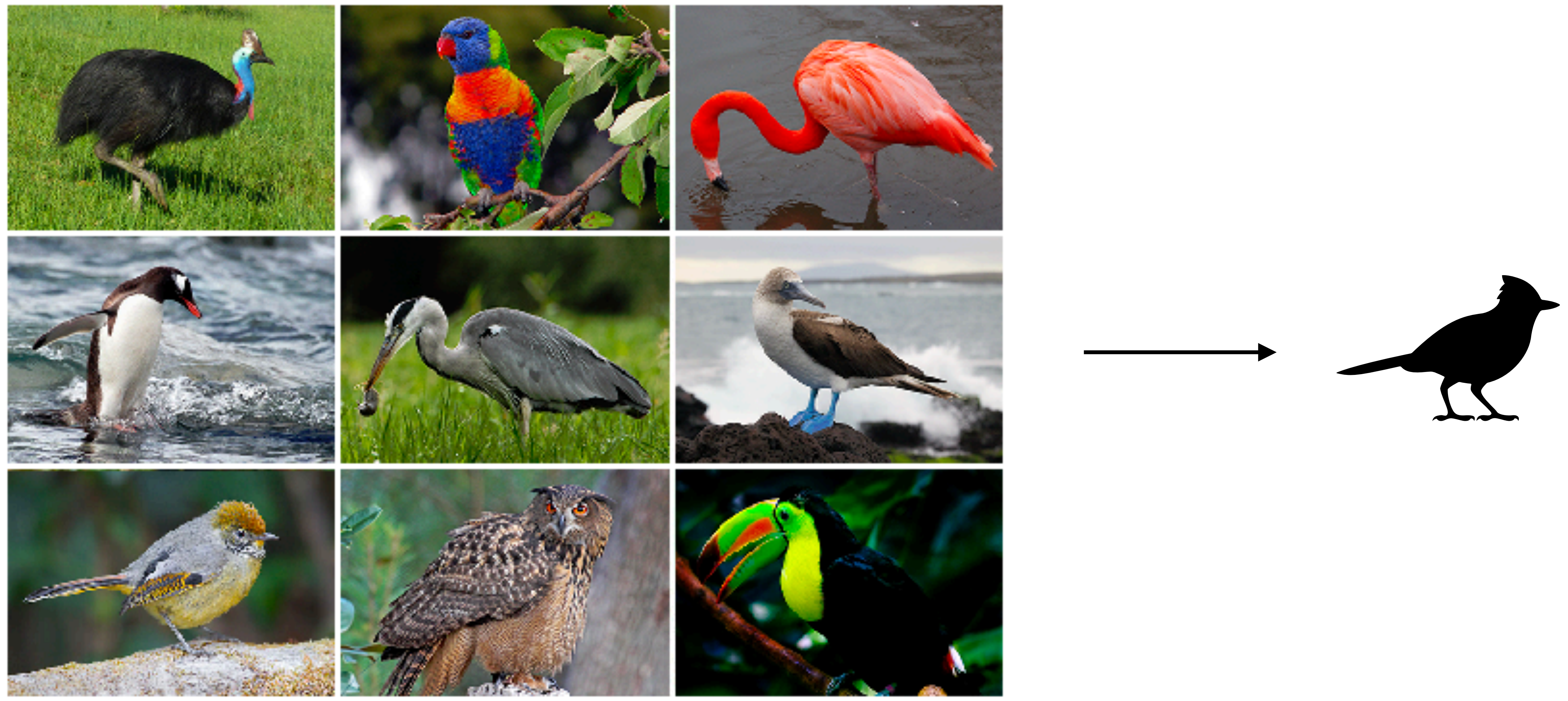
Levering & Kurtz (2019)

Prototype theory

- Prototypes are *summary representations* of a category (Rosch, 1973)
  - Typical can be explained by items being closer to our learned prototype
- Prototypes can be constructed based weighted features (Smith & Medin, 1981)
  - Some features are more important: Birds have wings (1.0), usually fly (0.8), some sing songs (0.3), and a few eat worms (0.1)
- Categories are thus defined by similarity to the prototype

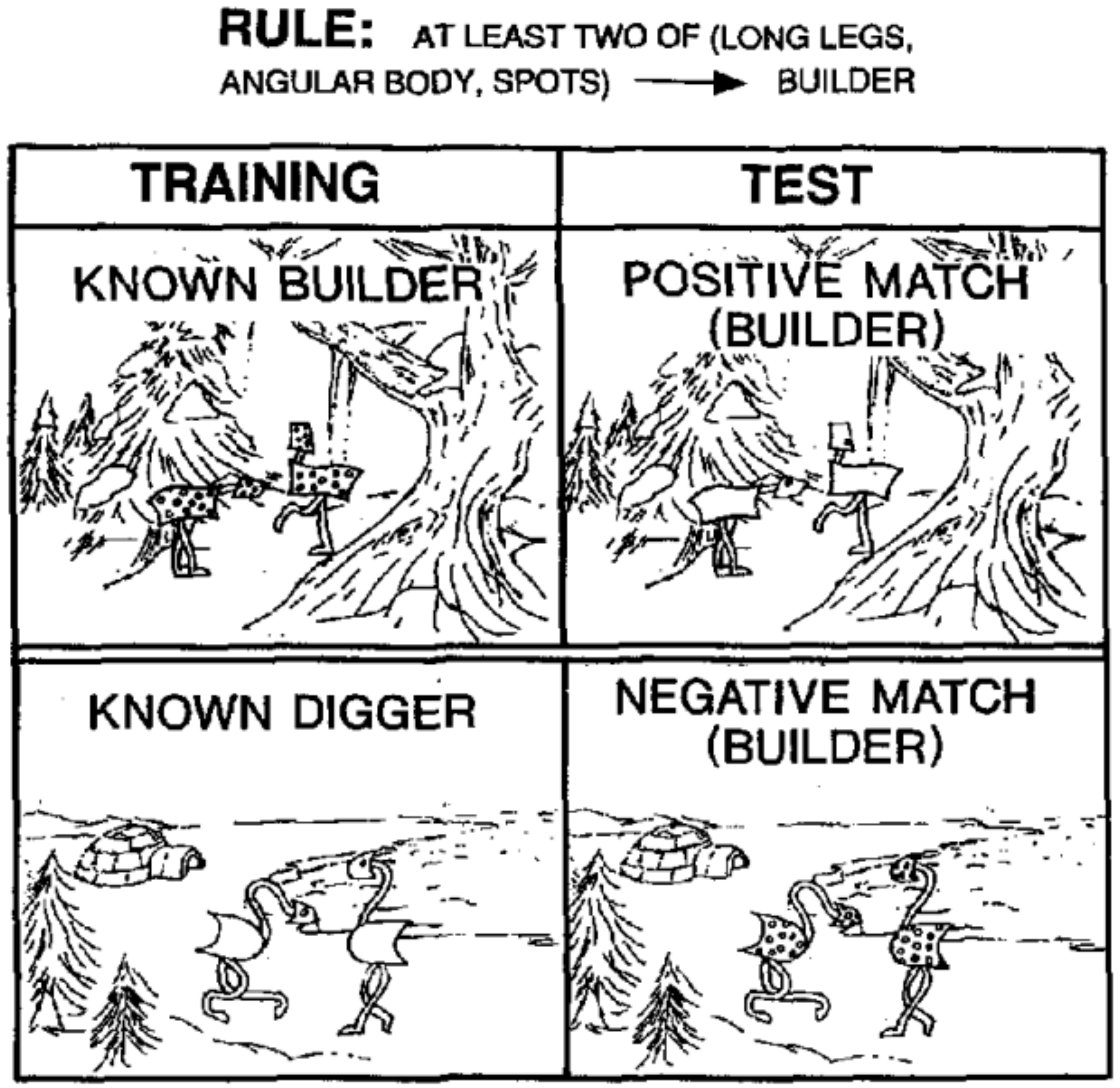
Which is the most prototypical chair?

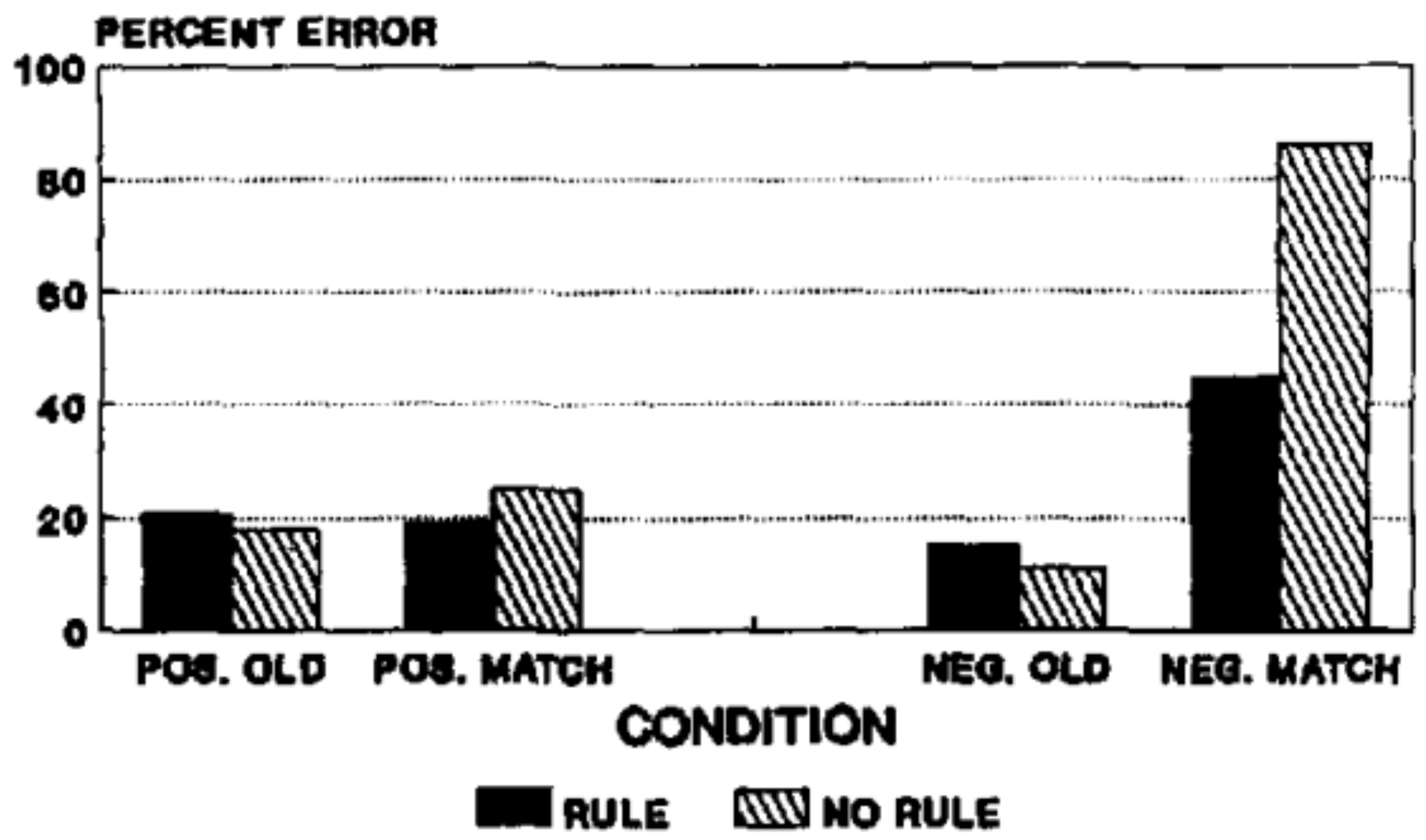
Constructing a prototype by weighing important features



Exemplar theory

- No summary representation
  - We remember each example of a concept, and we compare new instances to these past memories (Medin & Schaffer, 1978)
- Close similarity to well-remembered stimuli has a strong effect on classification:
  - Participants were either told about the rule or not
  - During test, participants were often fooled by the negative match (with spots), even when body and legs didn't match
- Categories are thus defined by similarity to past exemplars

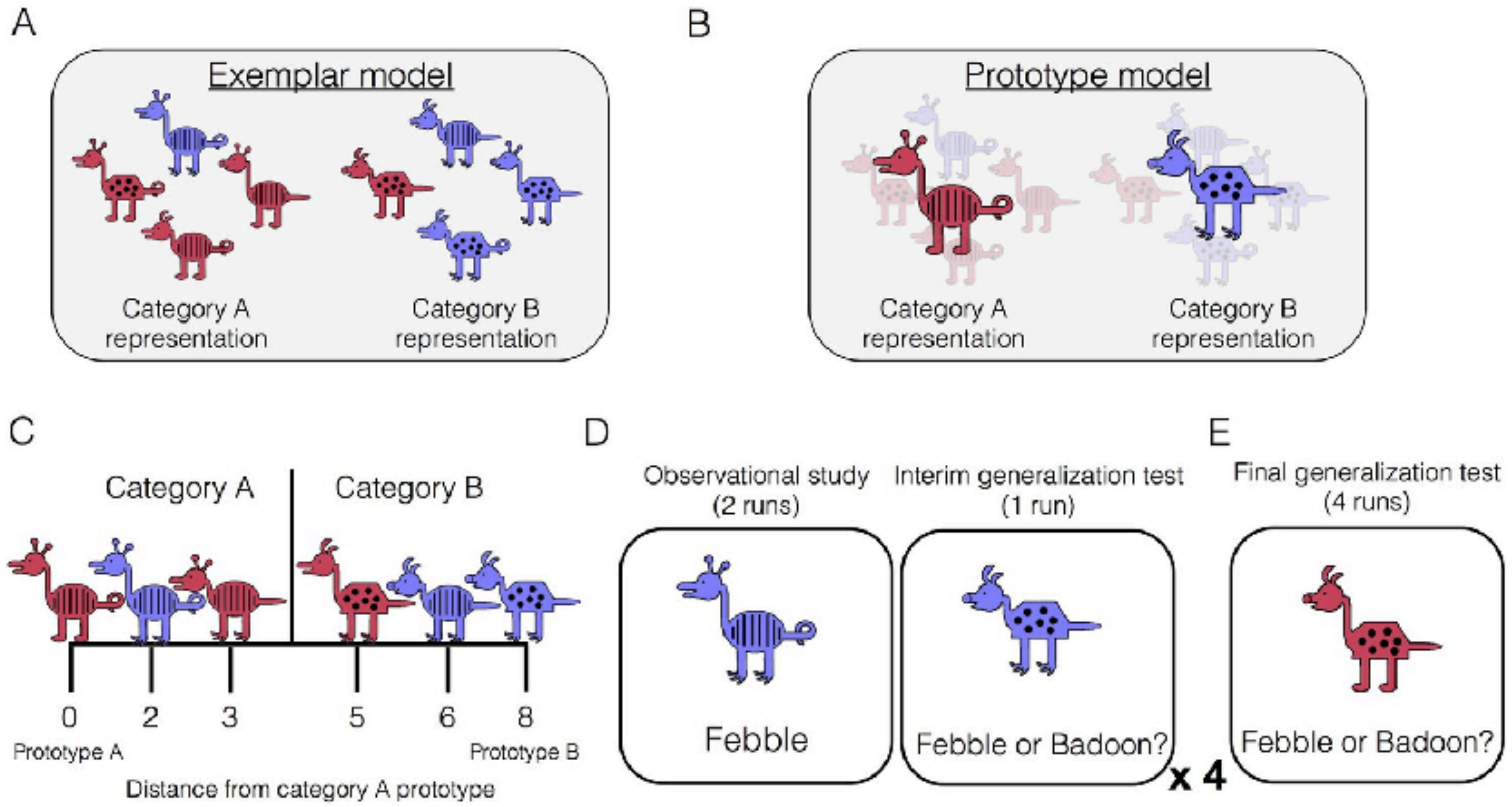


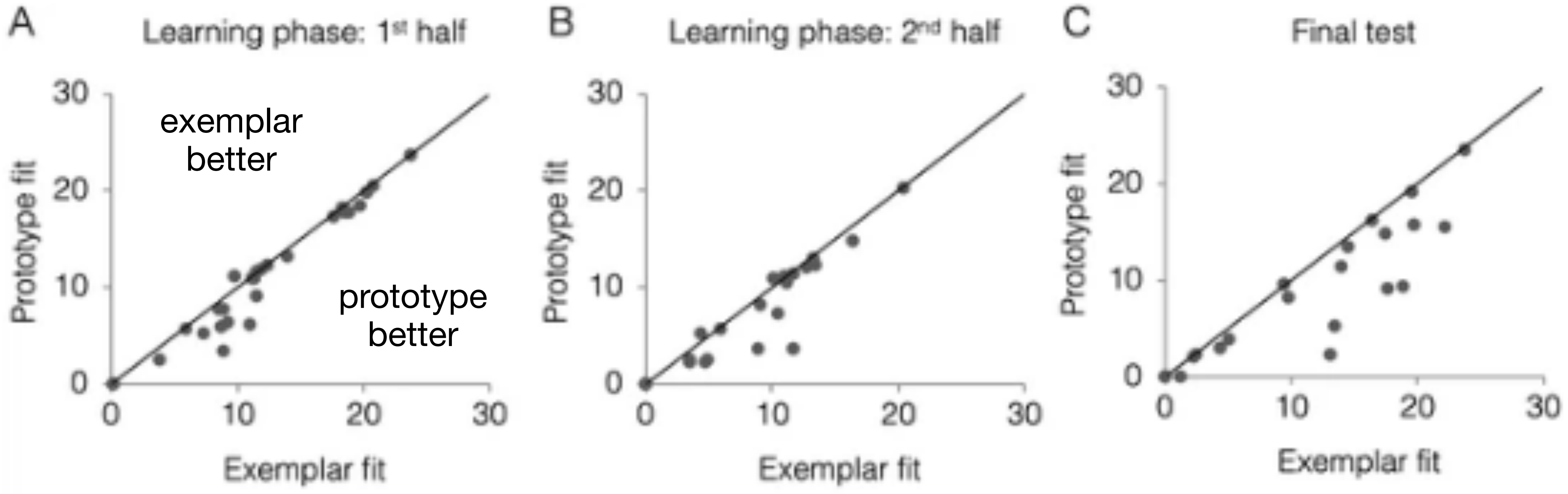


Allen & Brooks (1991)

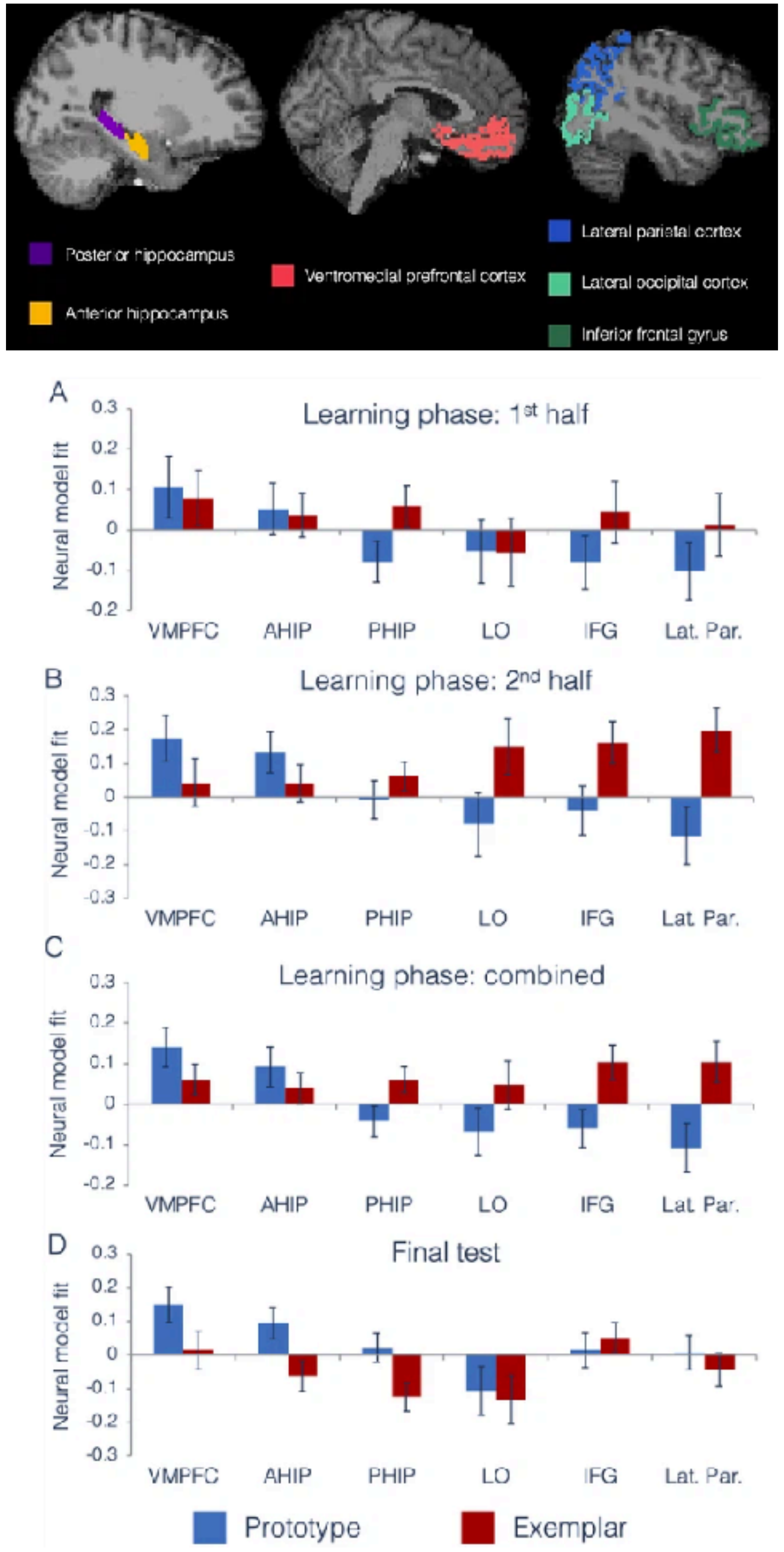
Prototype or exemplar?

- Still an open debate
- Prototype was dominant during the final test
- But neural signatures of both throughout





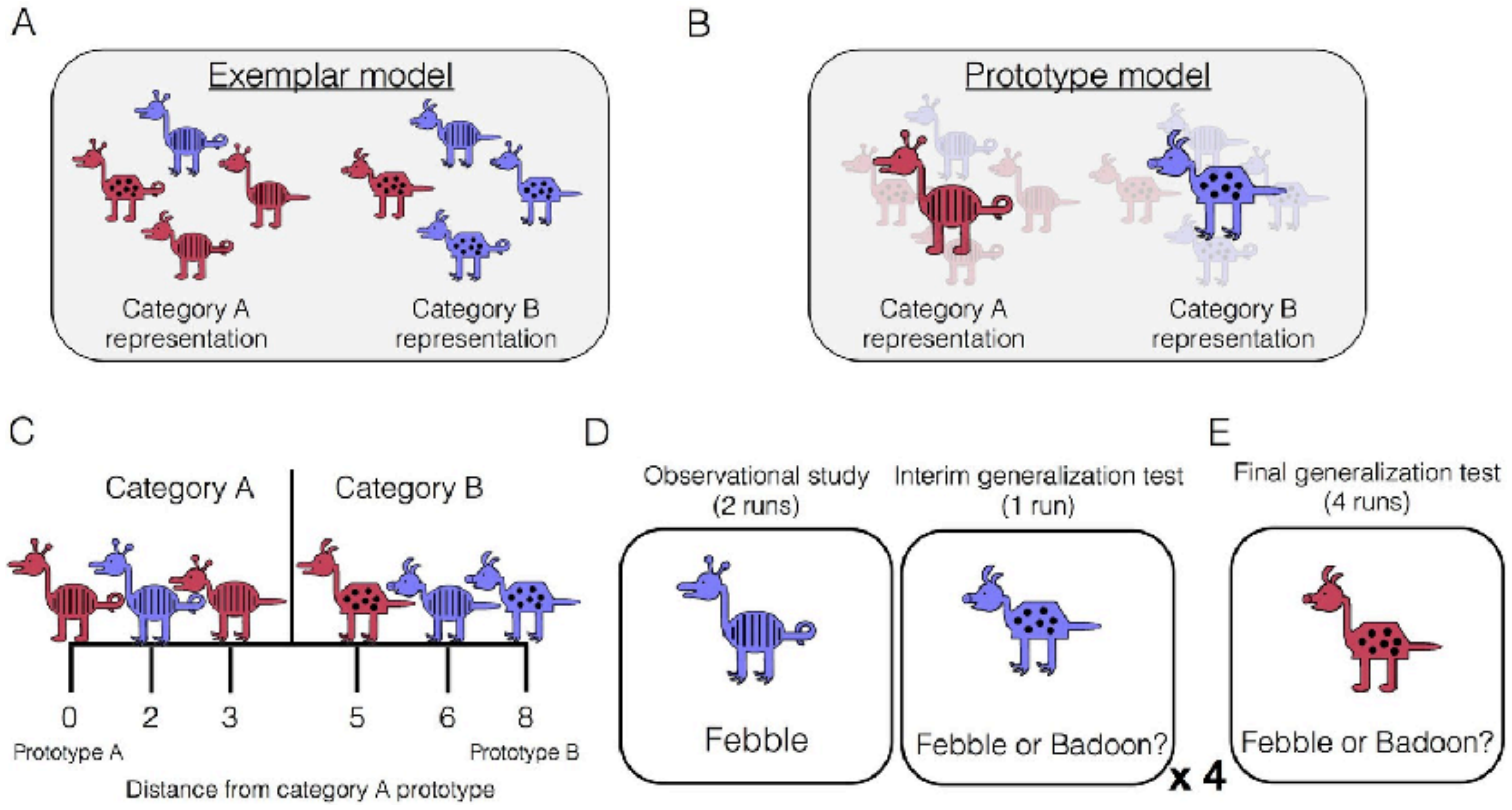
*lower is better

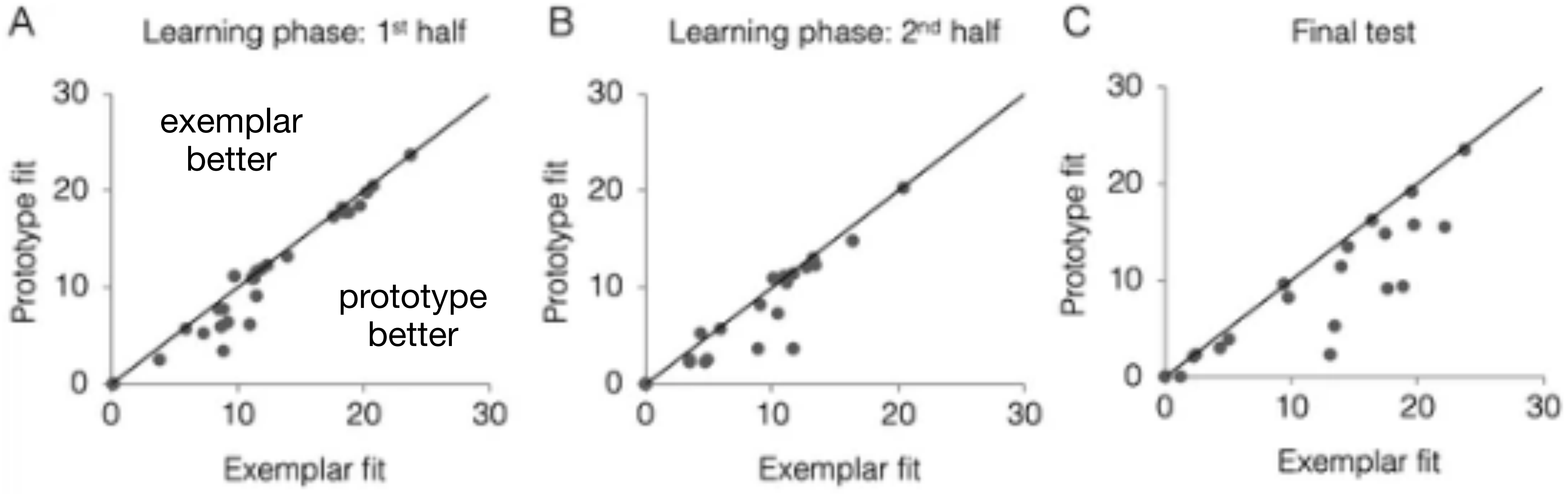


Bowman, Iwashita, & Zeithamova (2020)

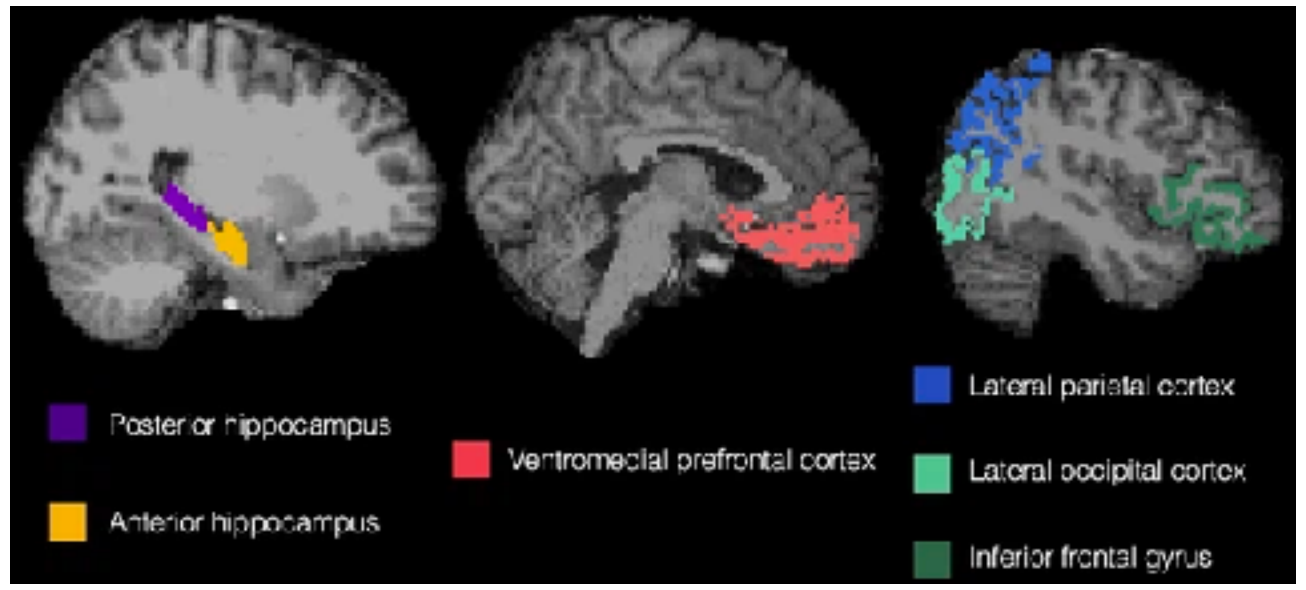
Prototype or exemplar?

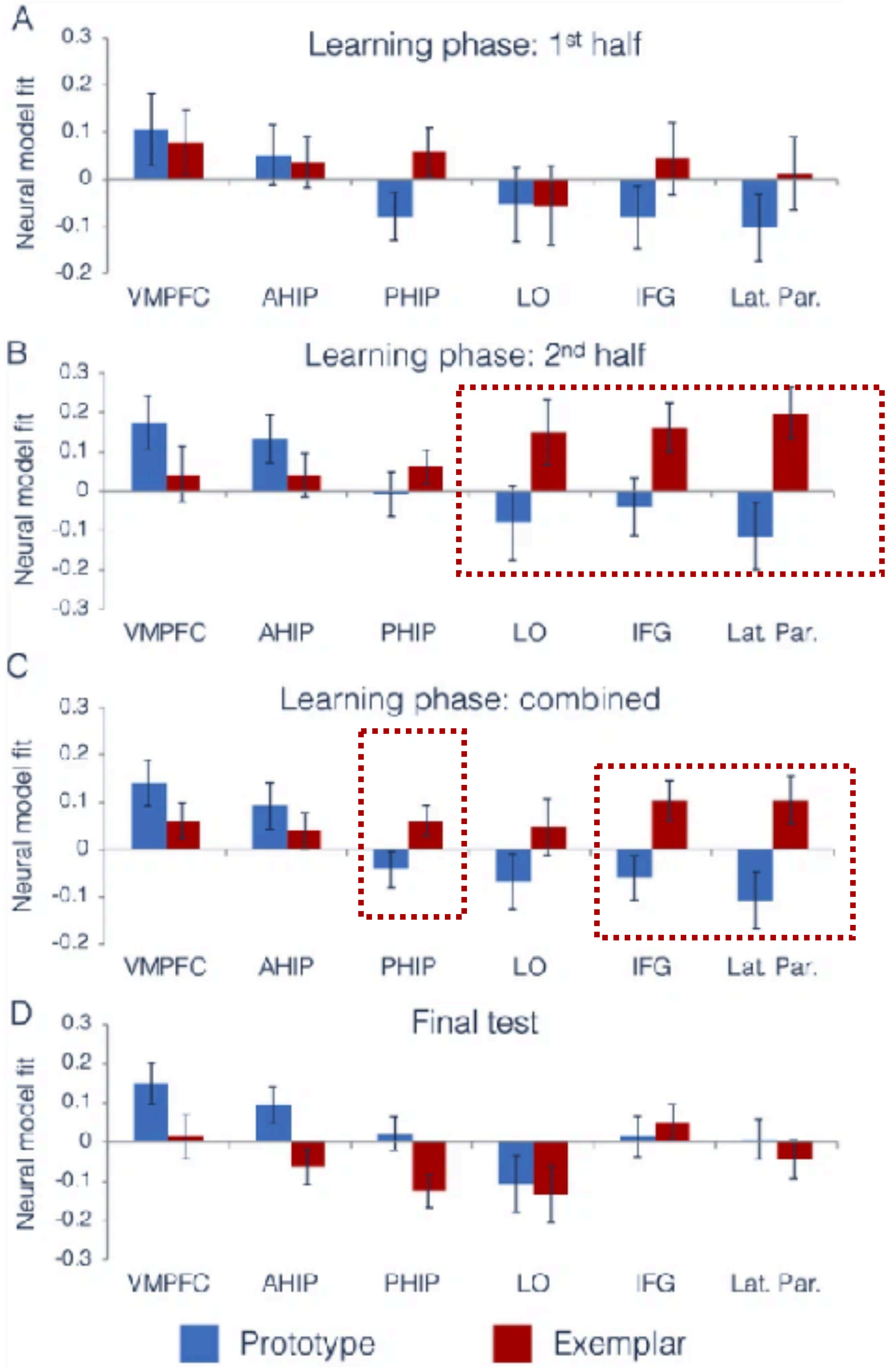
- Still an open debate
- Prototype was dominant during the final test
- But neural signatures of both throughout





*lower is better

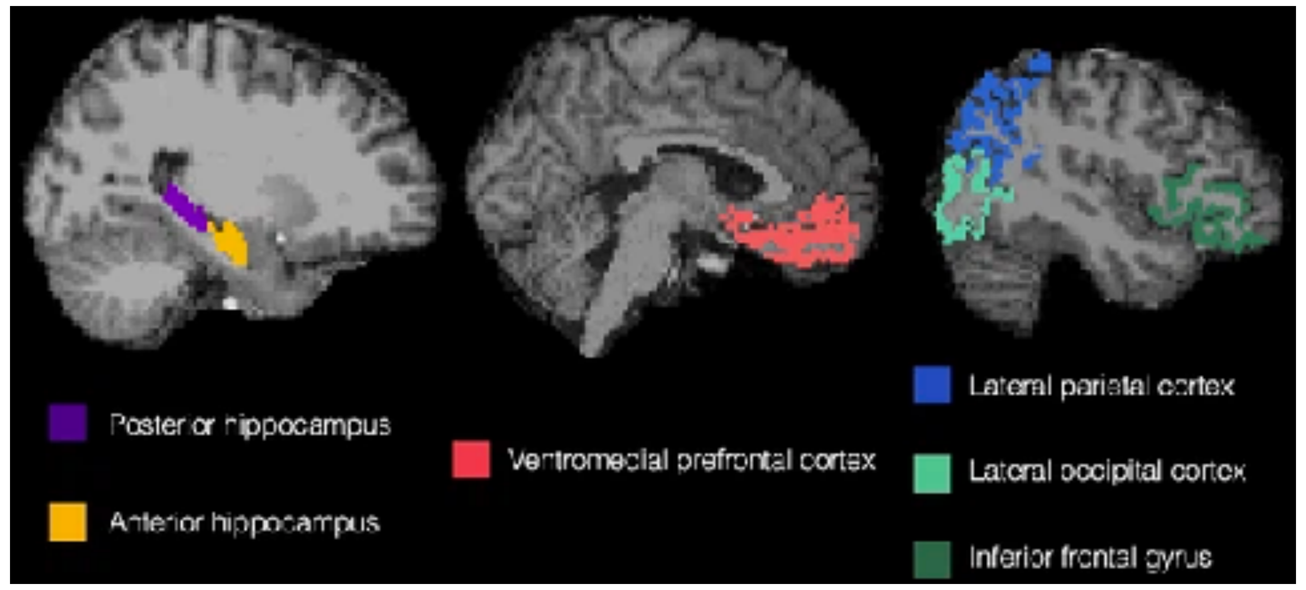


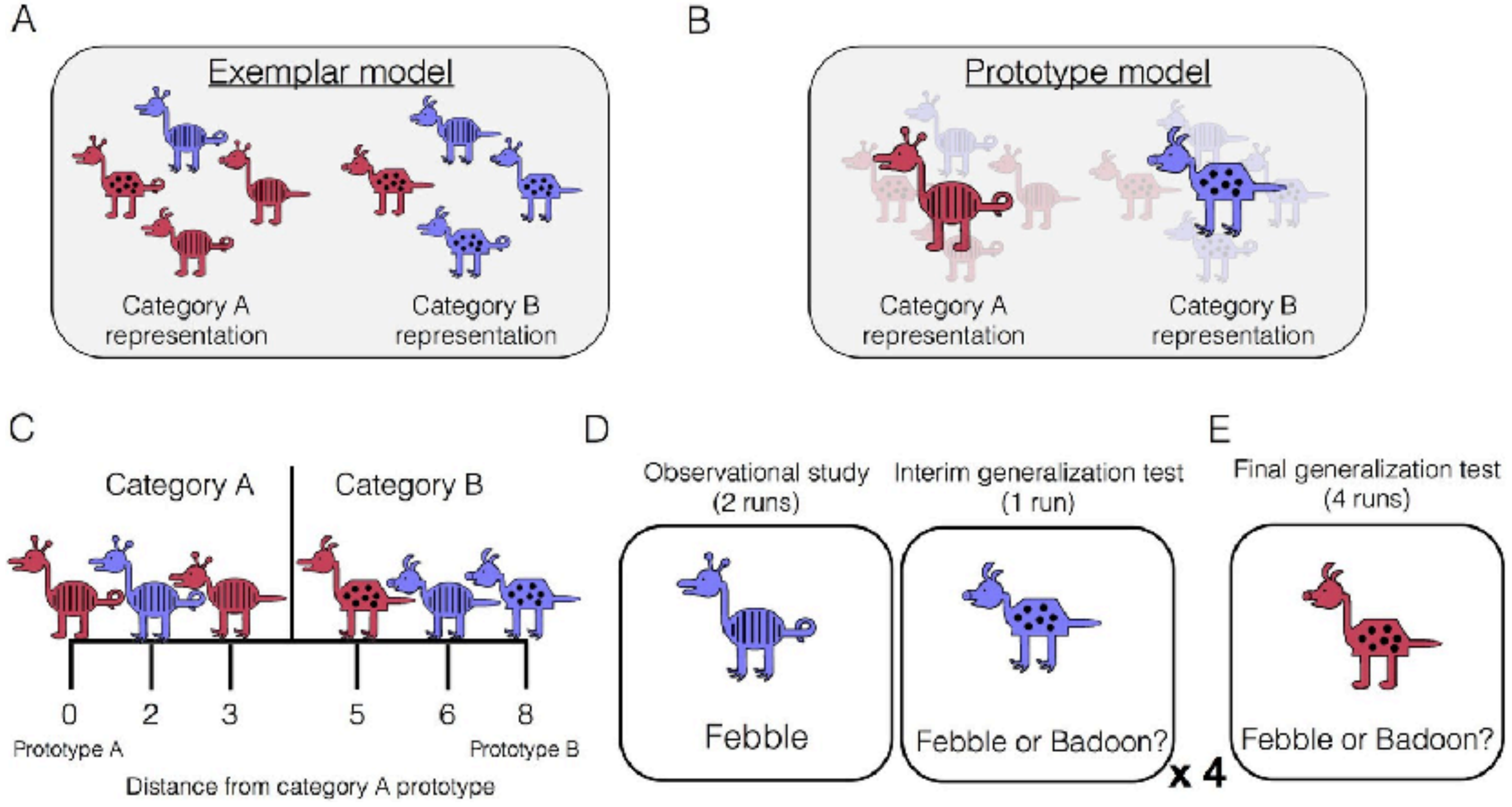


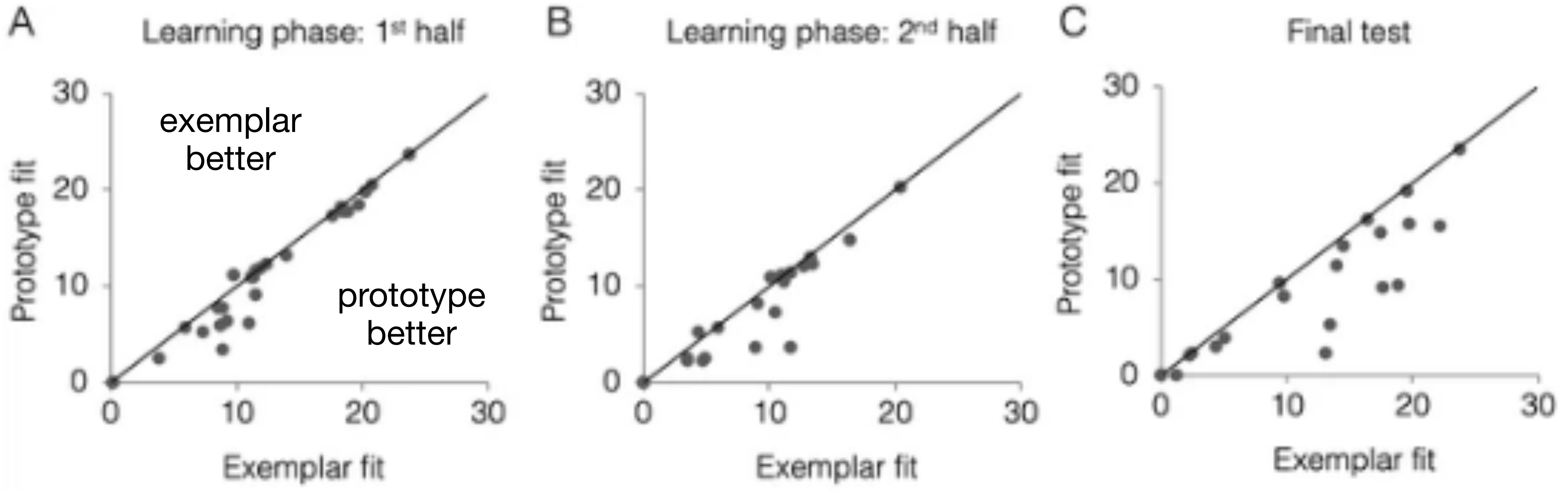
Bowman, Iwashita, & Zeithamova (2020)

Prototype or exemplar?

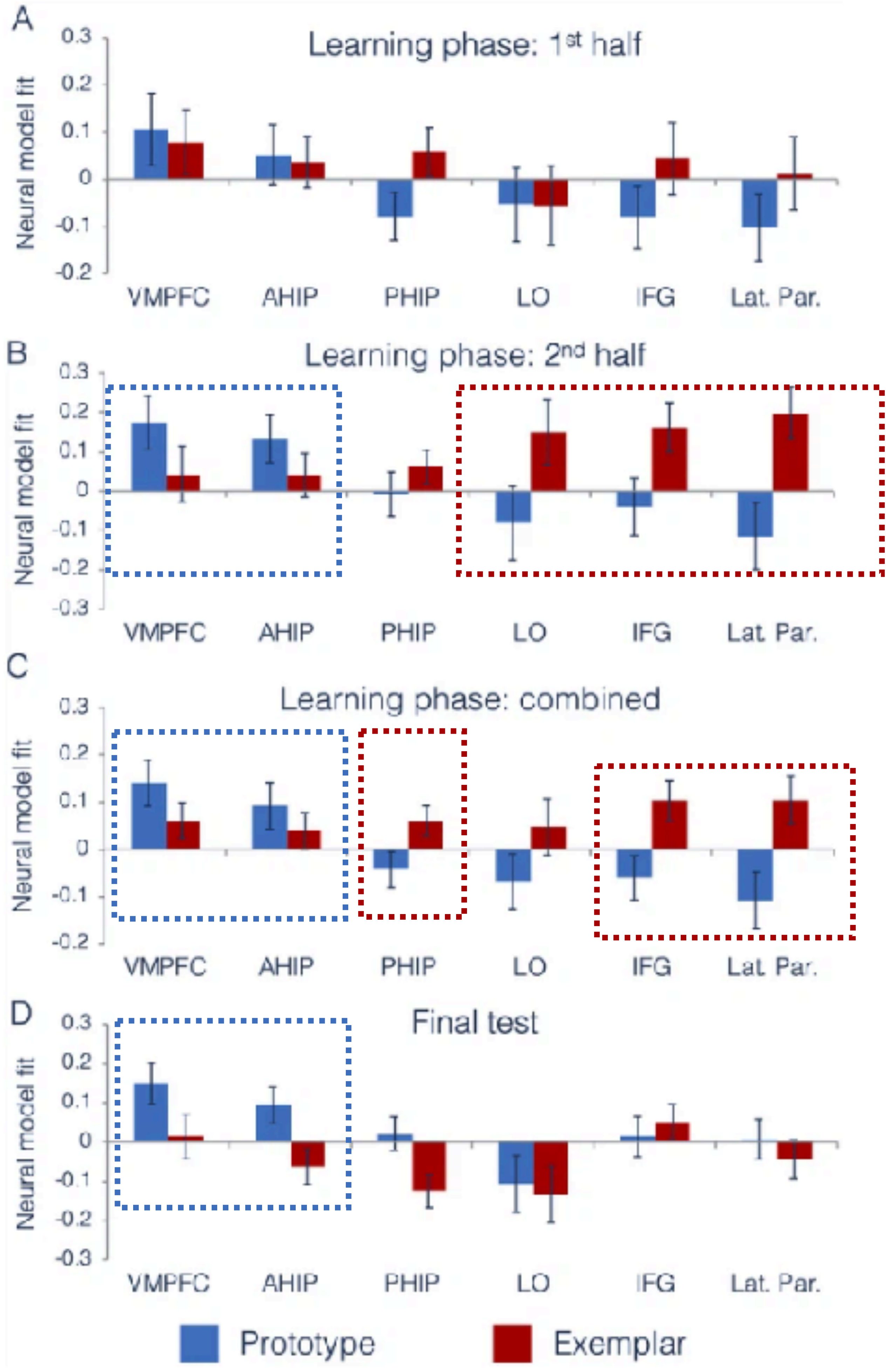
- Still an open debate
- Prototype was dominant during the final test
- But neural signatures of both throughout





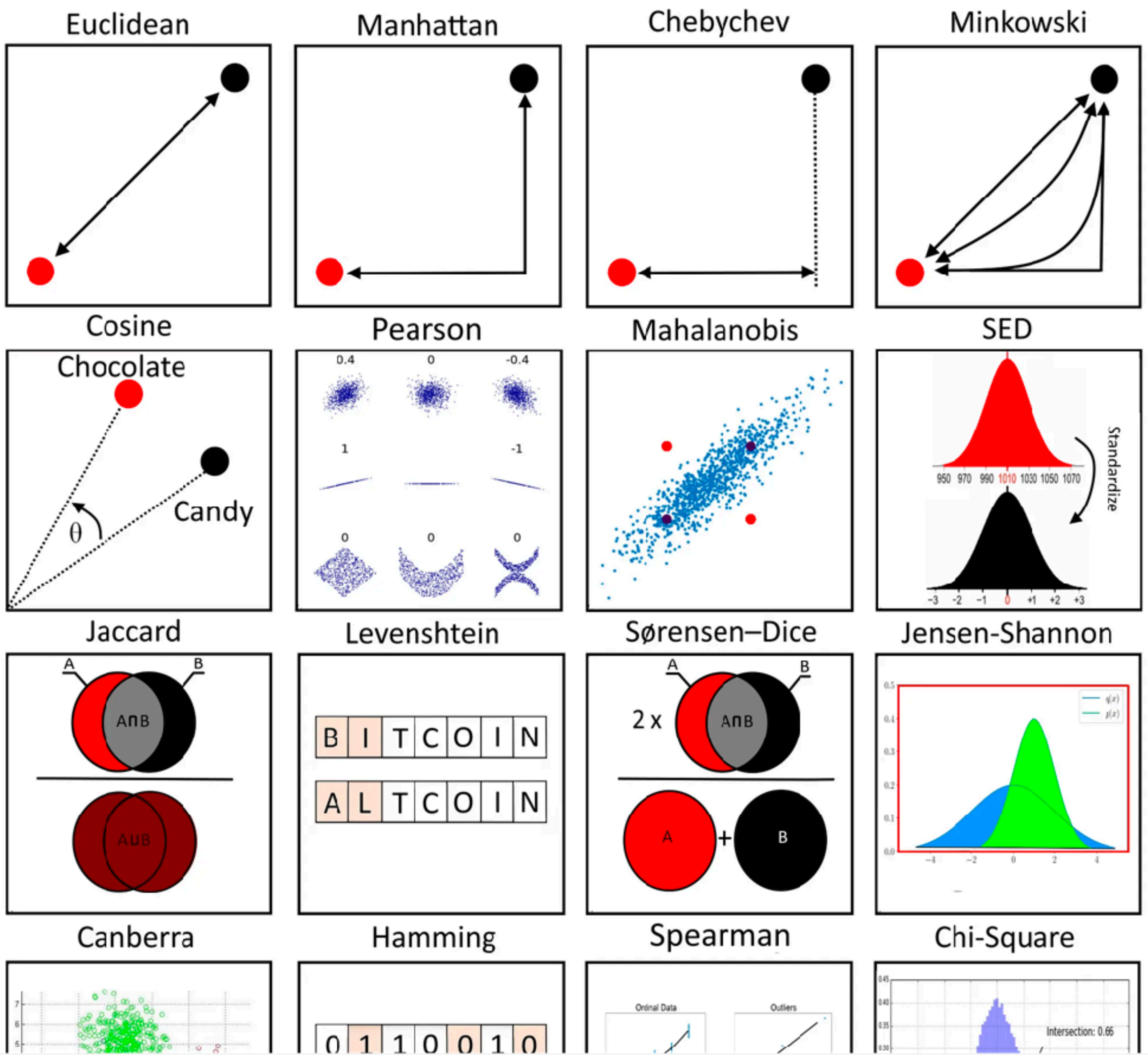


*lower is better



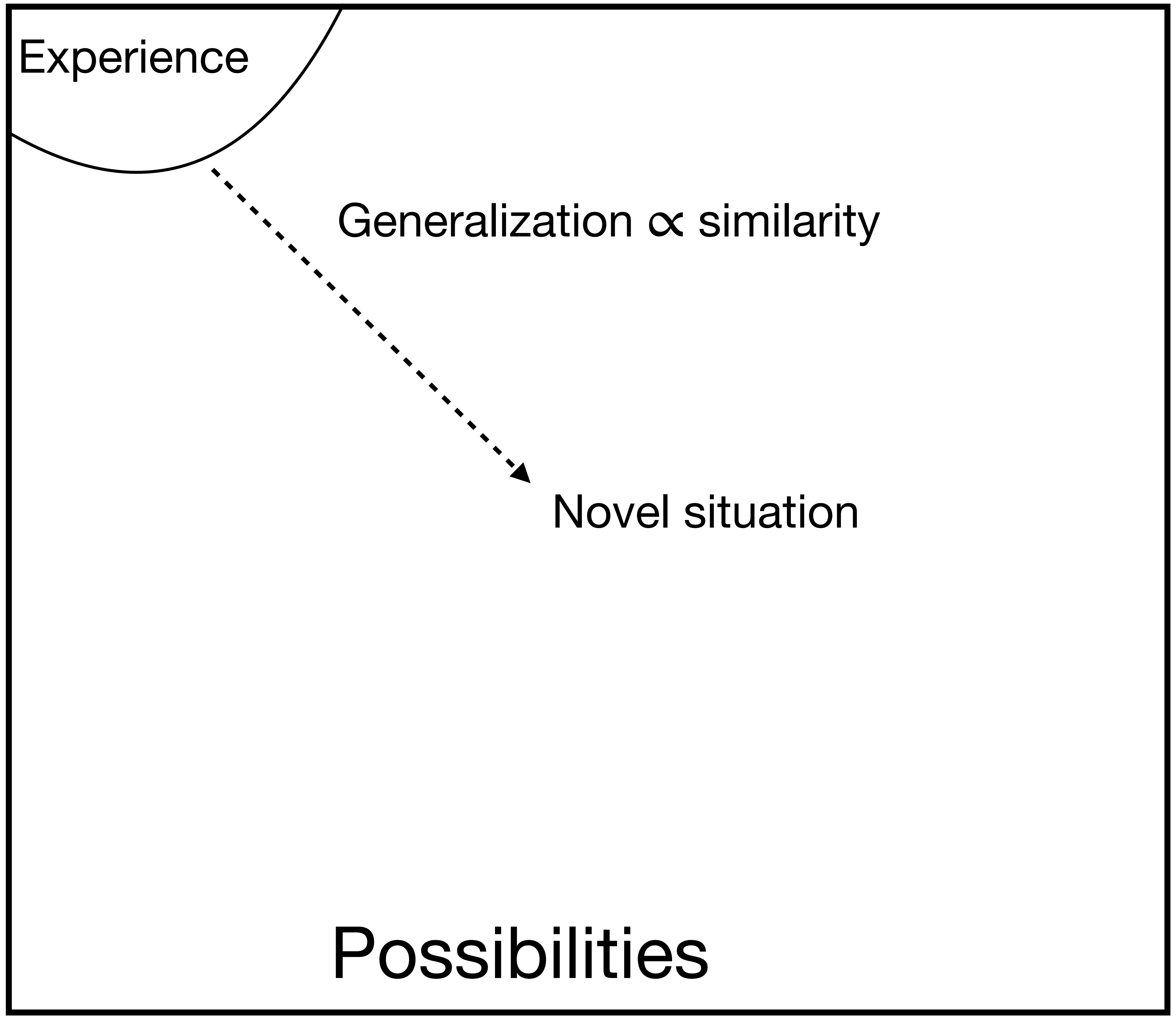
Bowman, Iwashita, & Zeithamova (2020)

How do we define similarity?



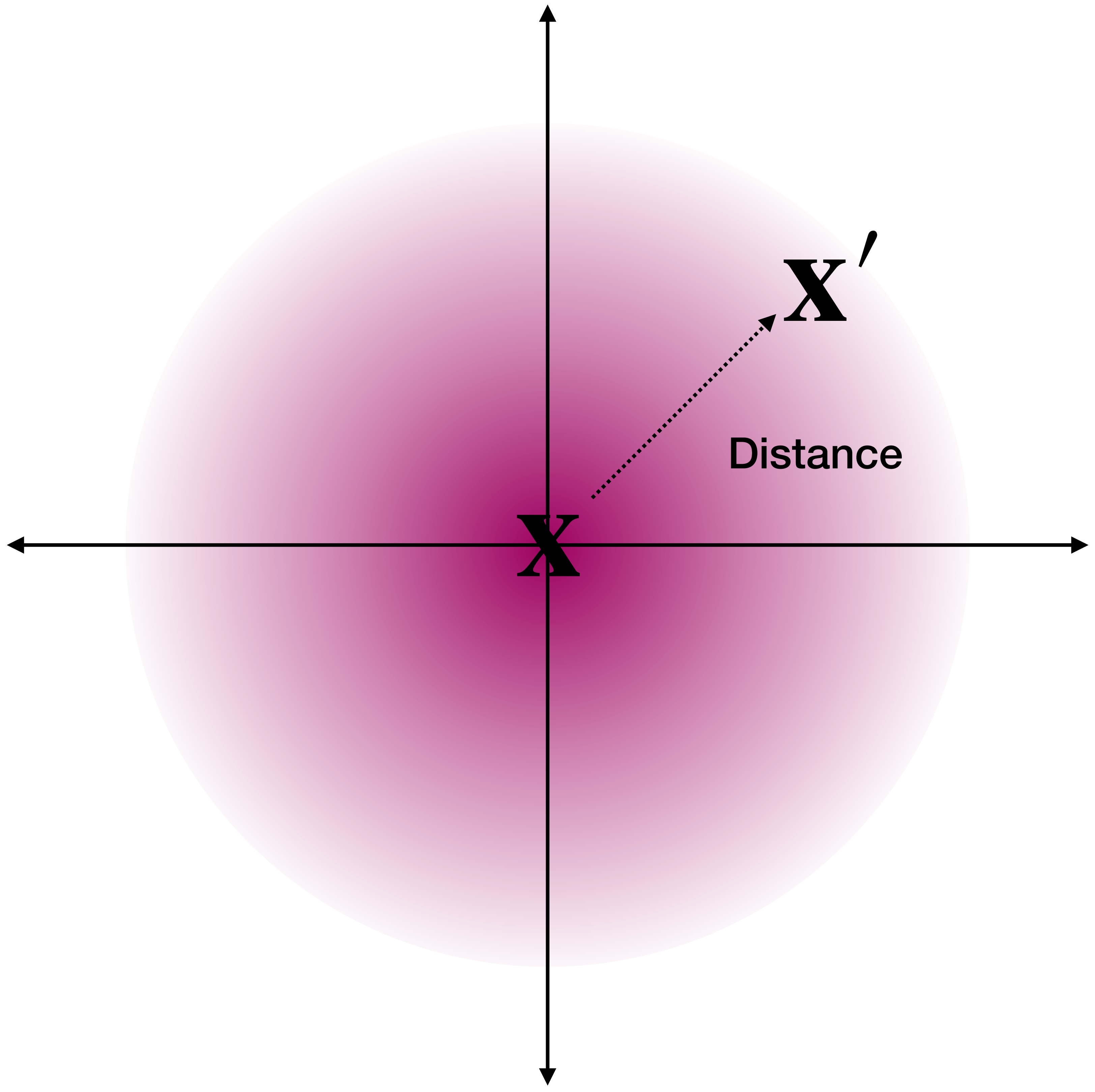
Generalization as a method to test different forms of similarity

- How do we generalize limited experience to novel situations?
  - The degree of generalization should be a function of our latent similarity computations
- The best similarity metric for predicting generalization should also reveal something about how we represent concepts



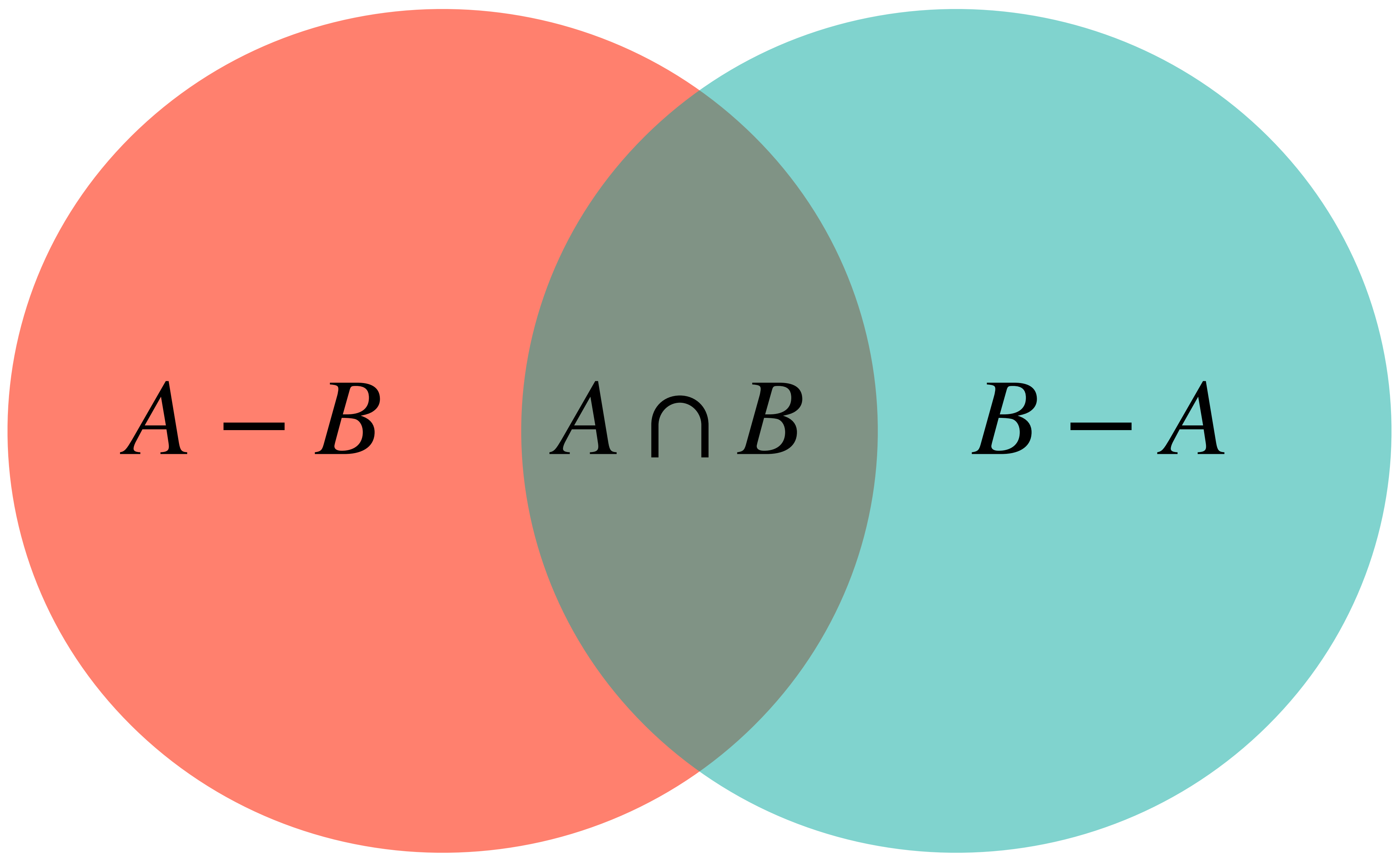
Two main approaches: Metric vs. Set

Metric



Embed data in some vector space and compute similarity as the inverse of distance

Set



Compare which features are jointly shared vs. unique (i.e., disjoint)

Shepard's (1987) Law of Generalization

STIMULUS + RESPONSE = **LEARNING**

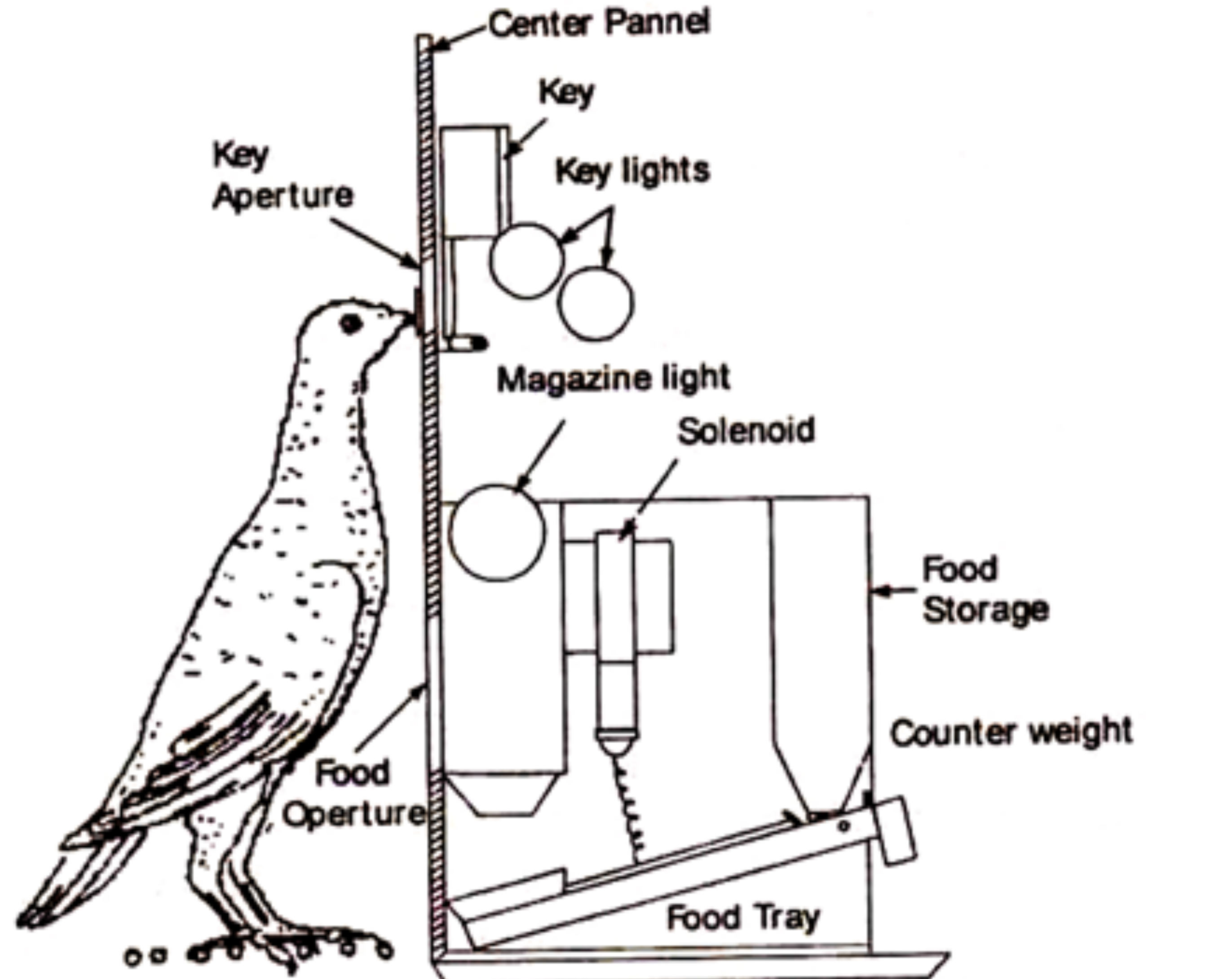
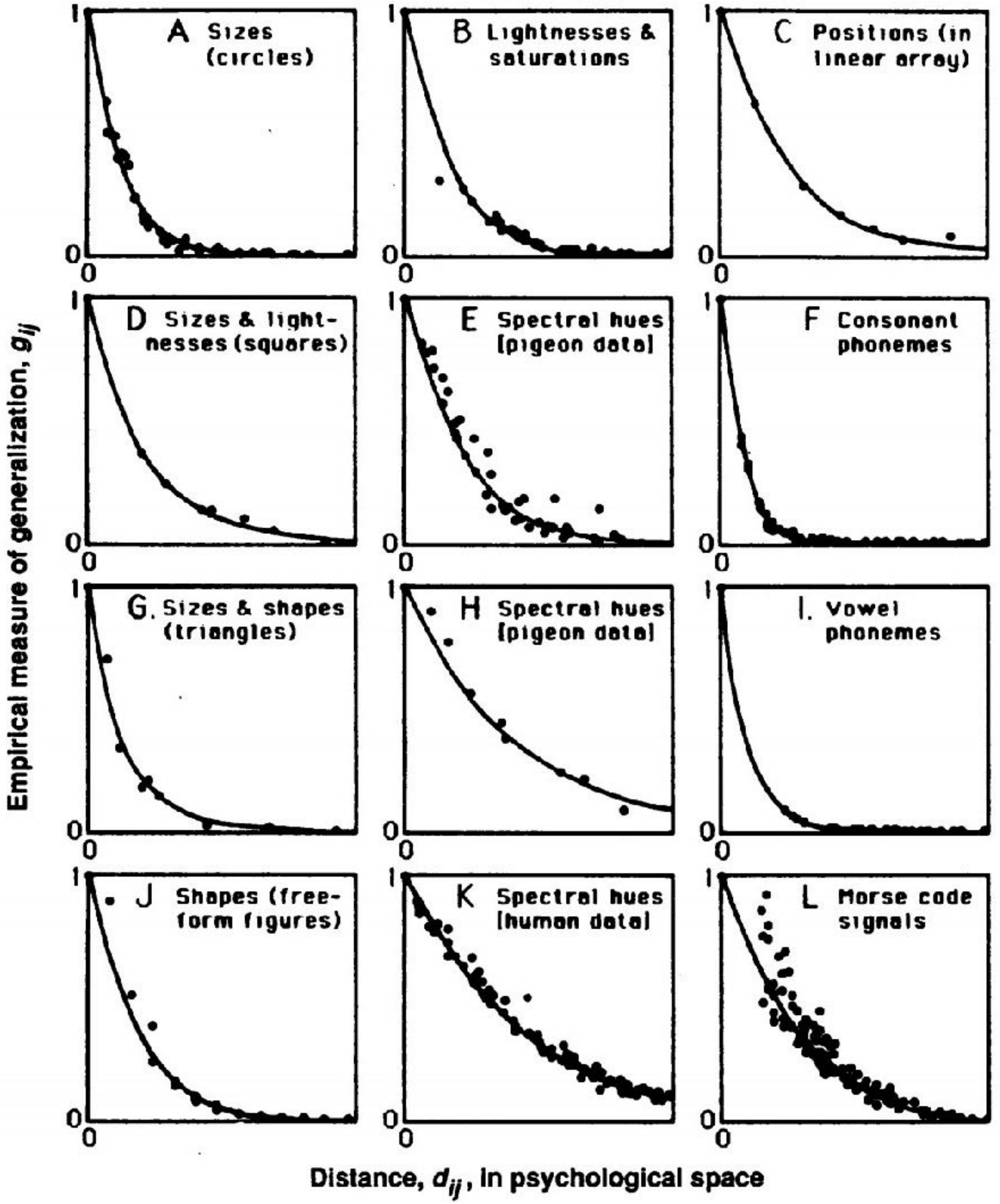


Illustration. Skinner box as adapted for the pigeon.



Shepard's (1987) Law of Generalization

STIMULUS + RESPONSE = **LEARNING**

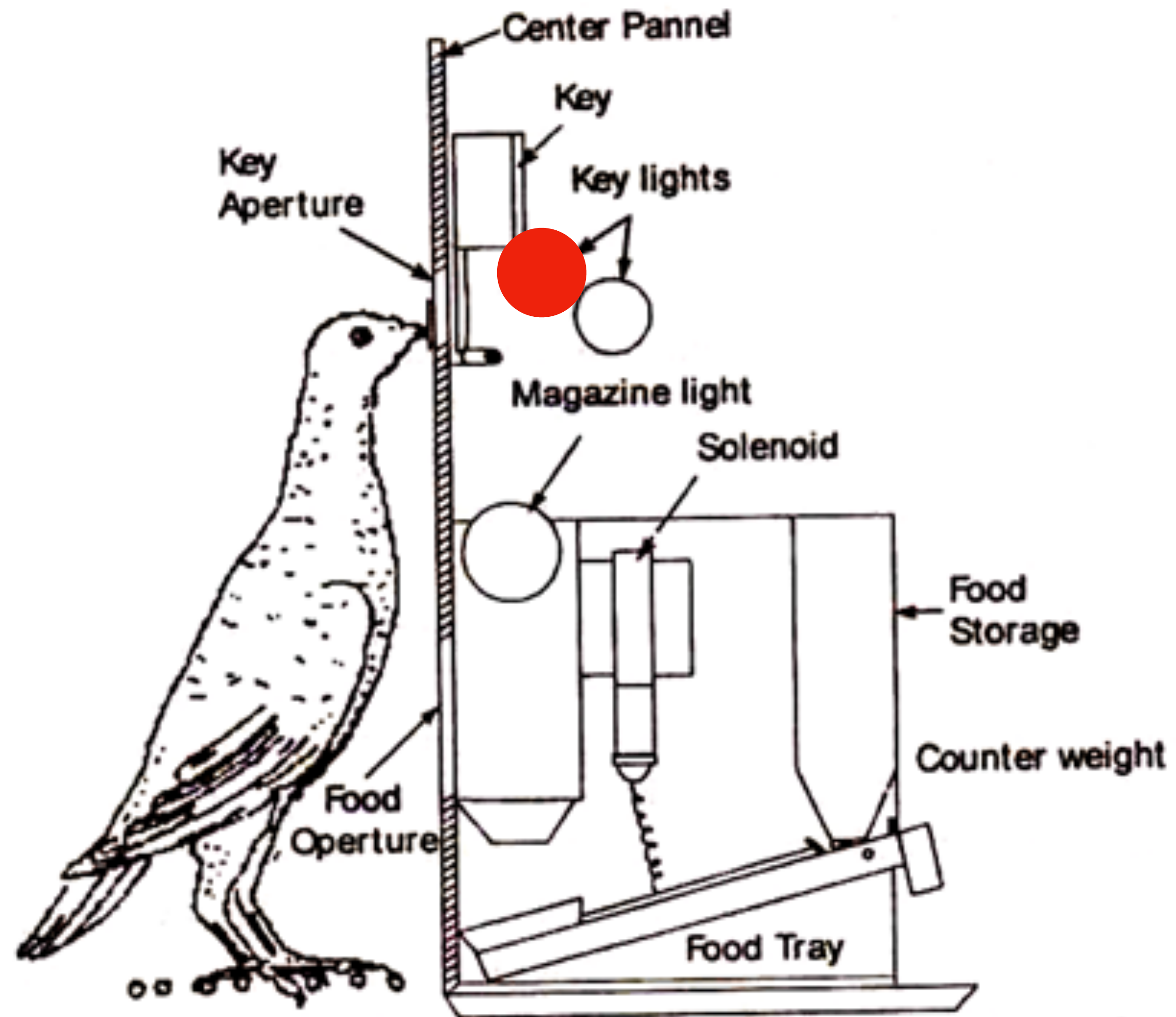
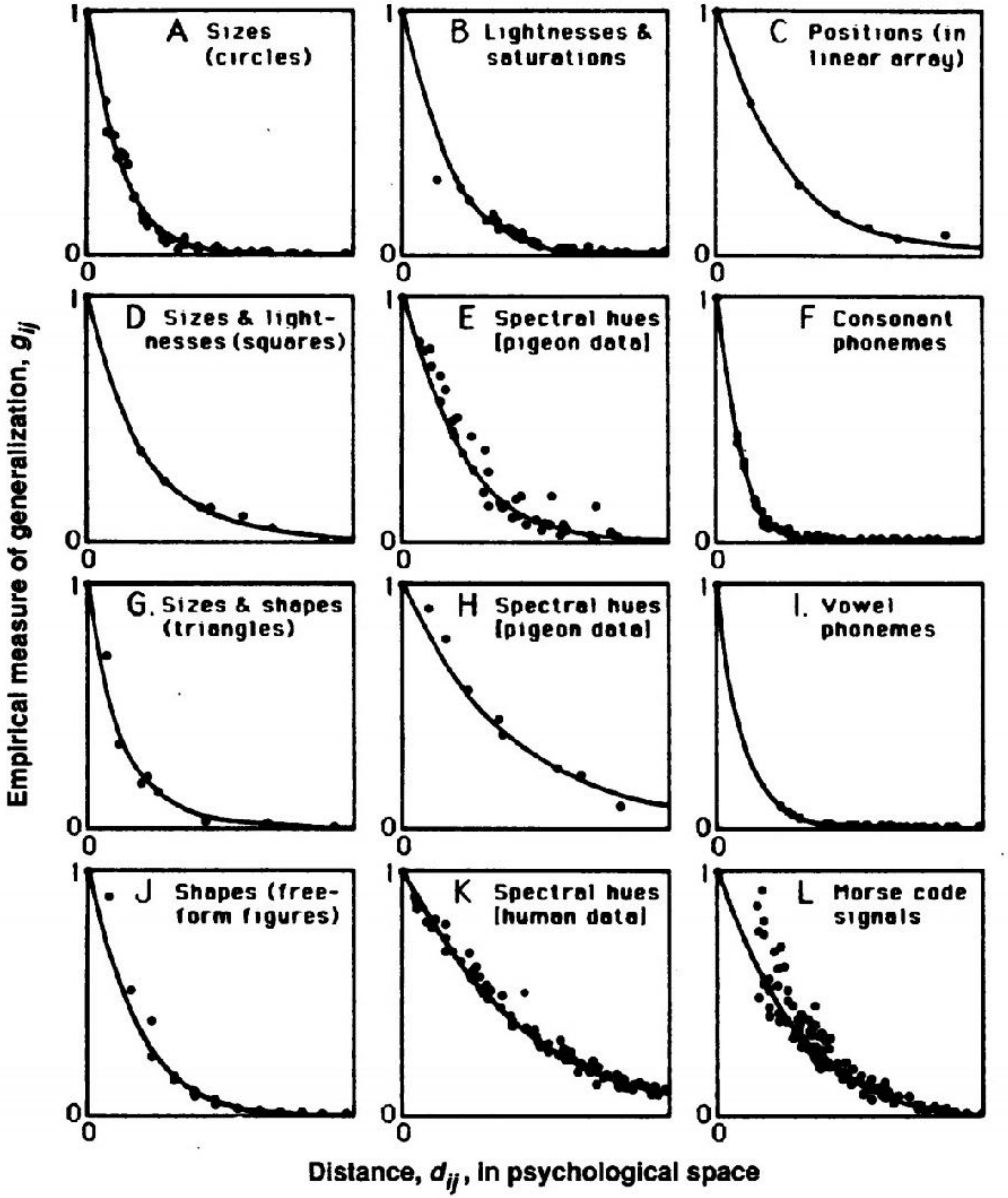
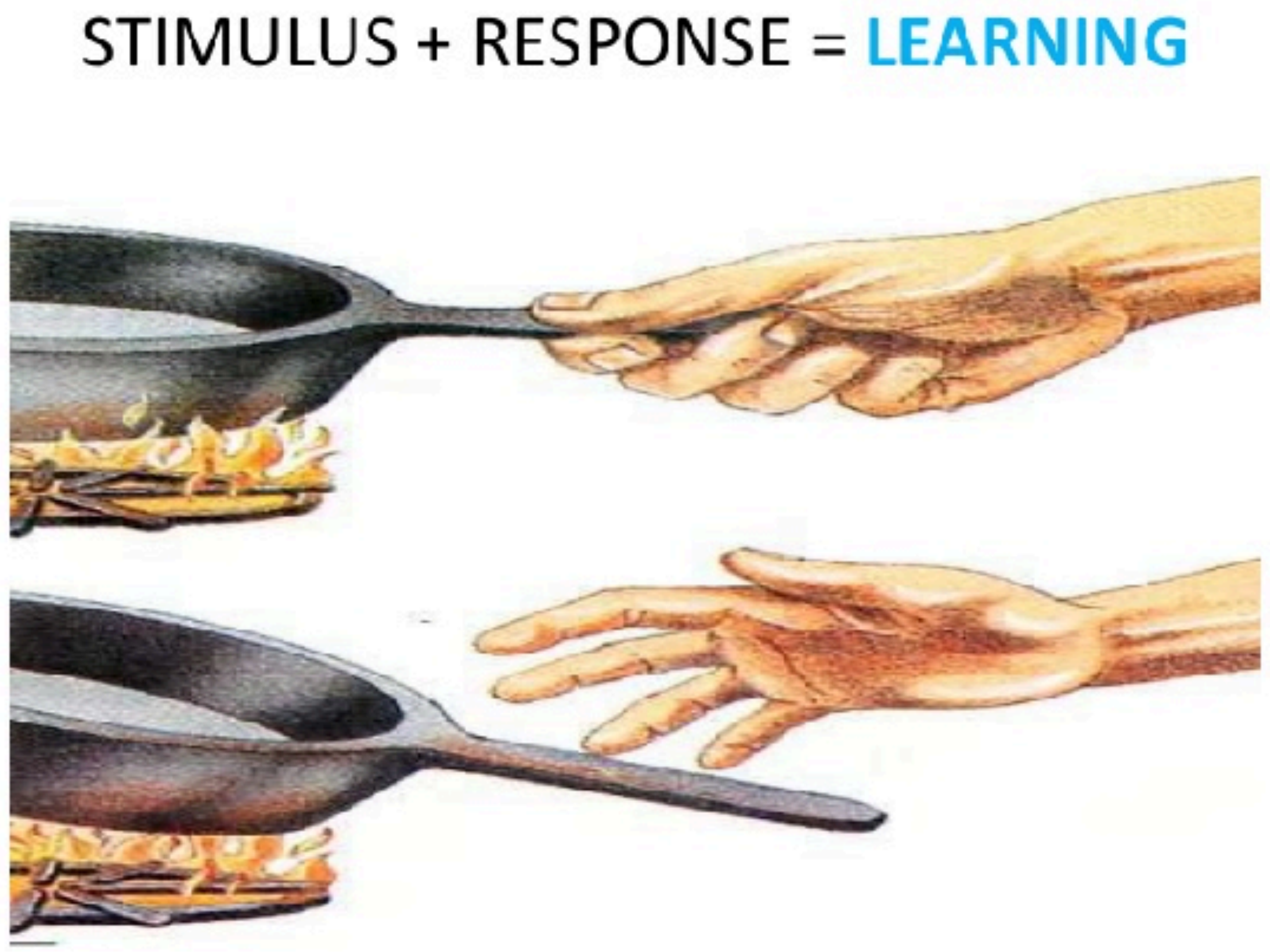


Illustration. Skinner box as adapted for the pigeon.



Shepard's (1987) Law of Generalization



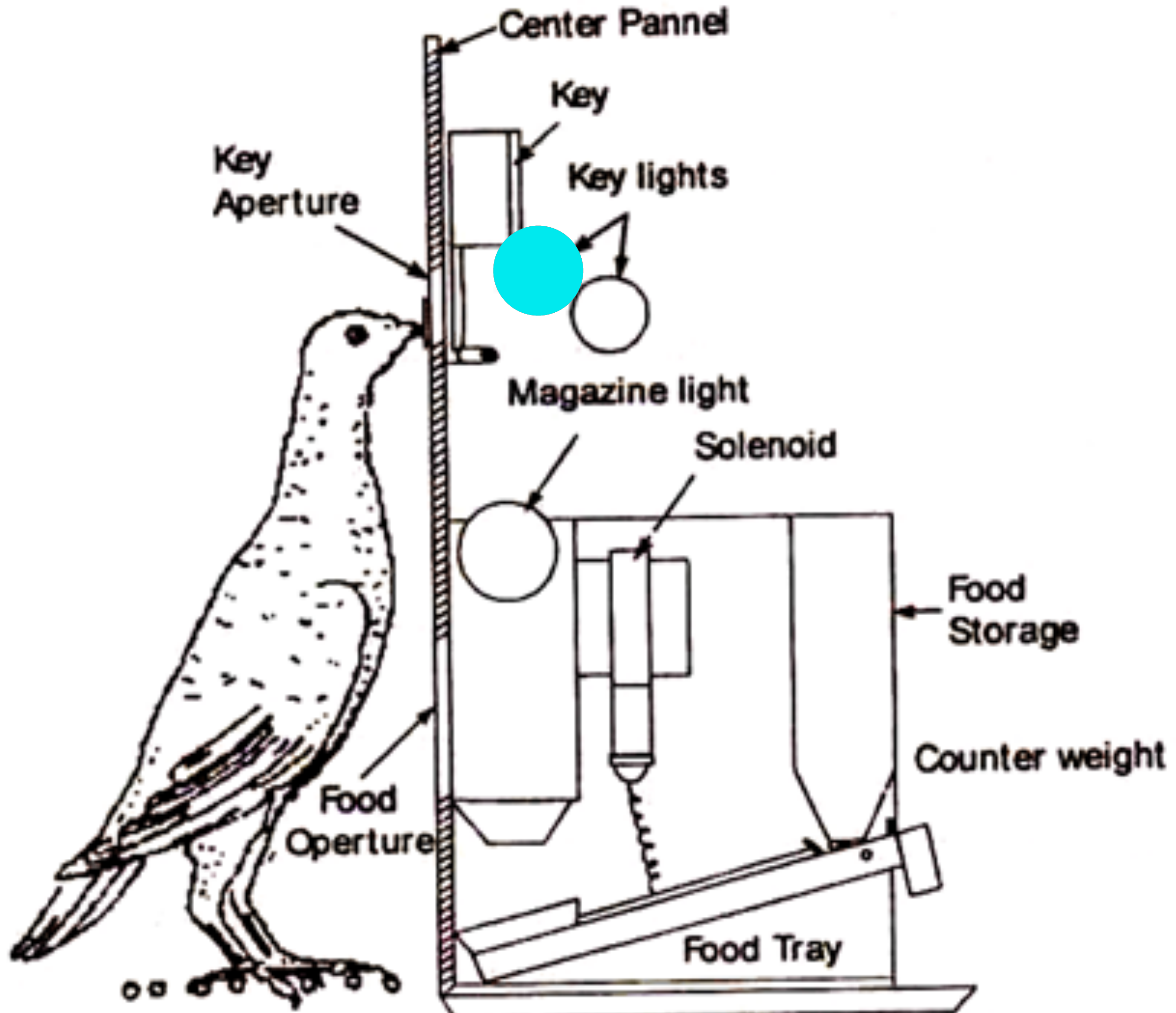
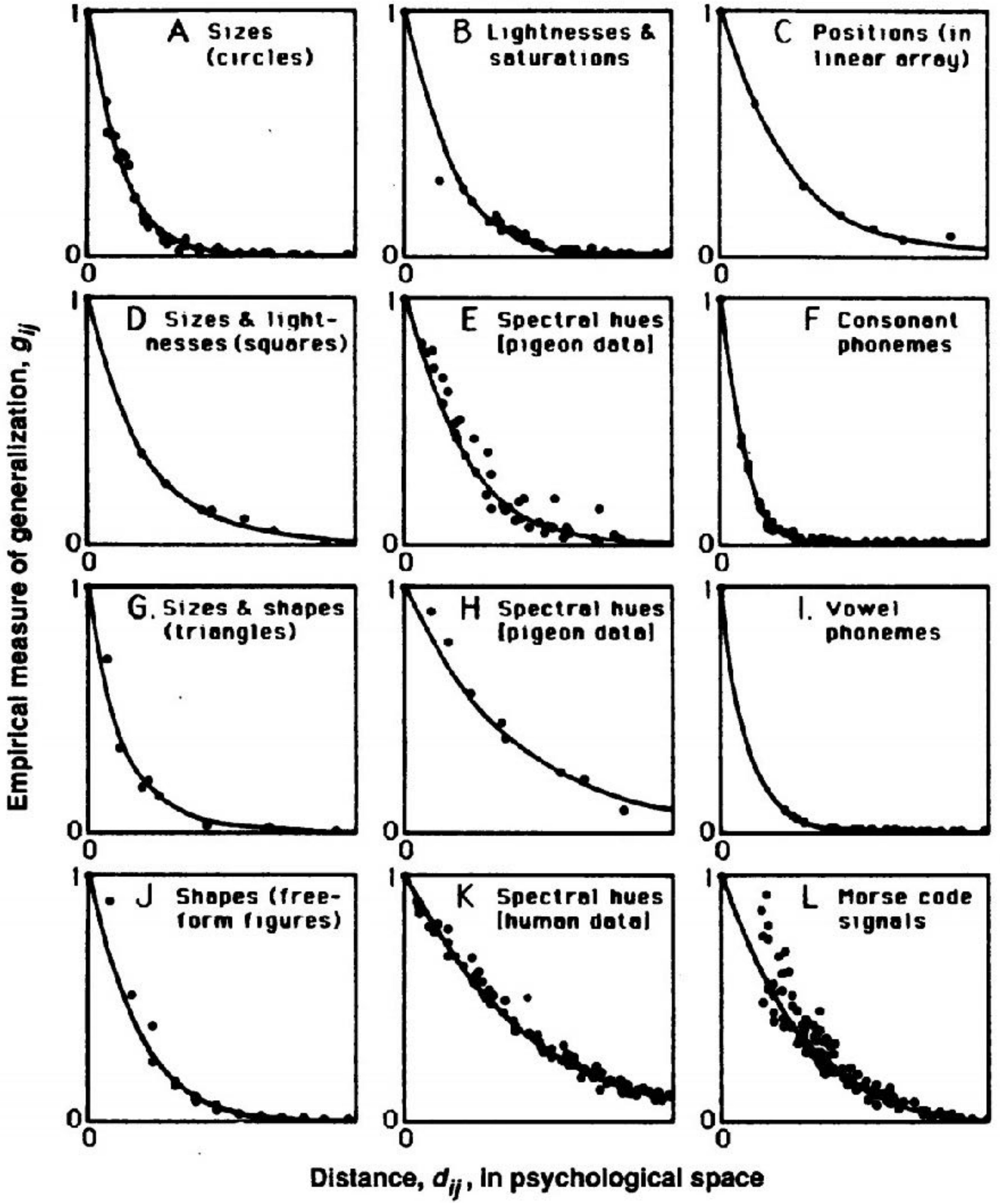


Illustration. Skinner box as adapted for the pigeon.



Shepard's (1987) Law of Generalization

STIMULUS + RESPONSE = **LEARNING**

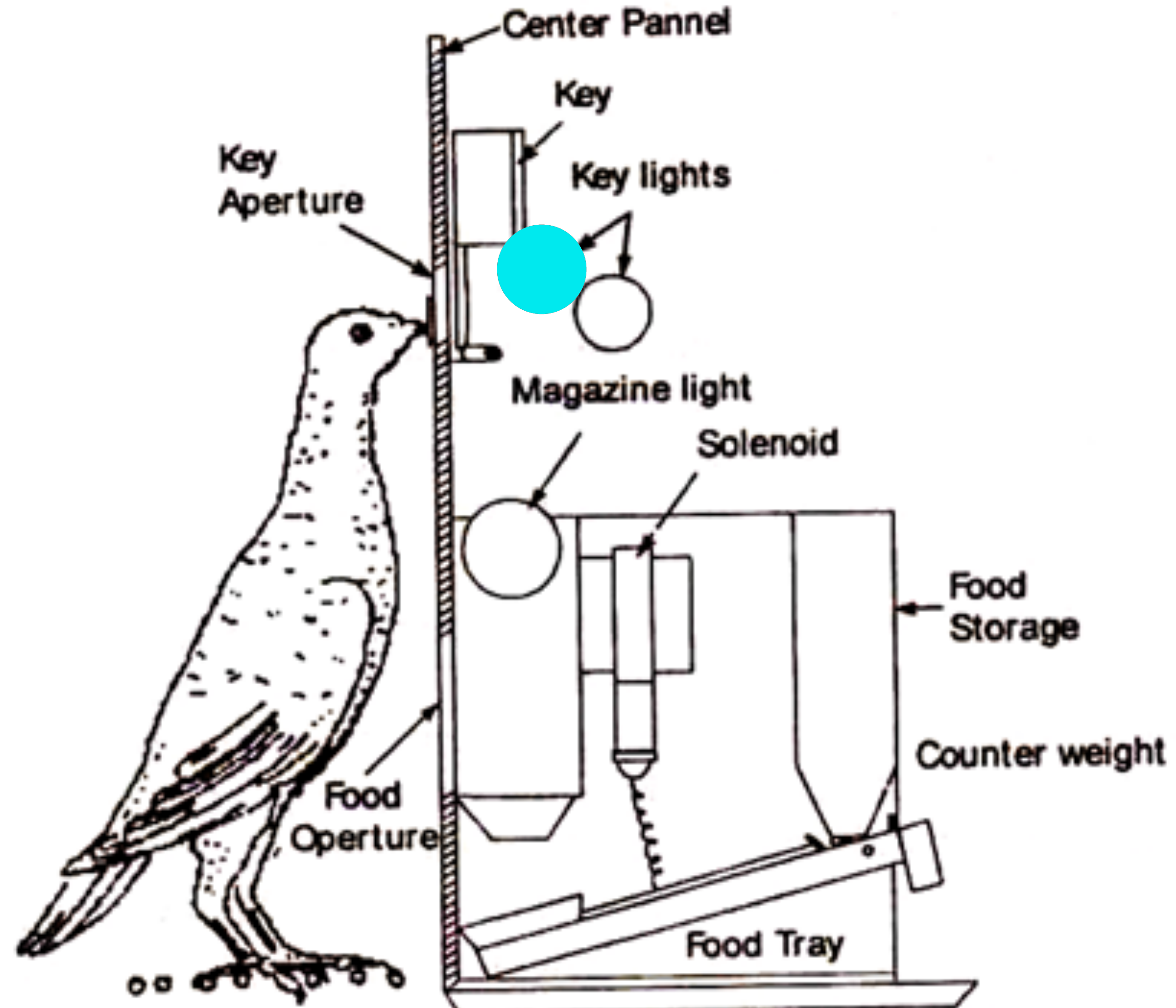
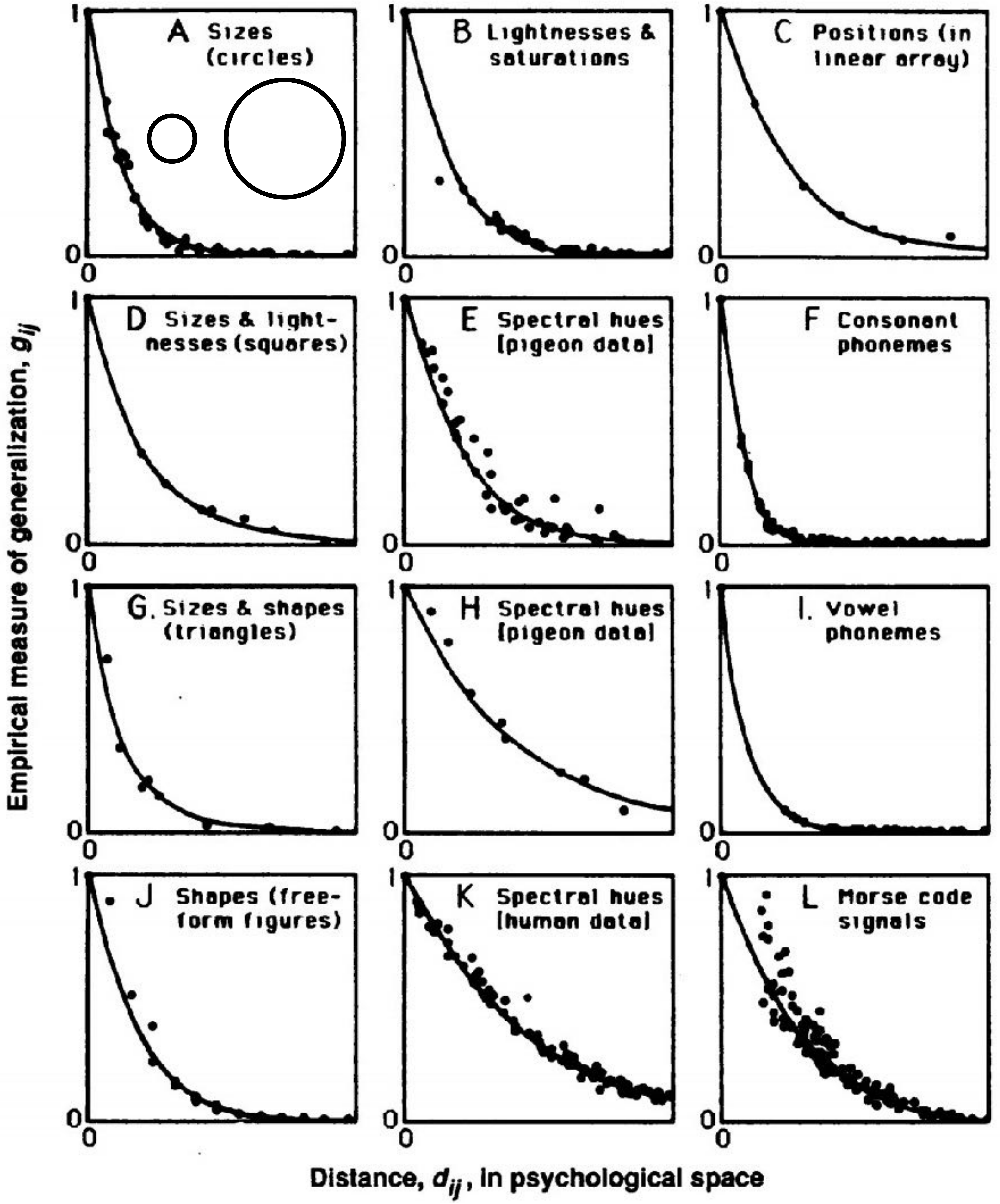


Illustration. Skinner box as adapted for the pigeon.



Shepard's (1987) Law of Generalization

STIMULUS + RESPONSE = **LEARNING**

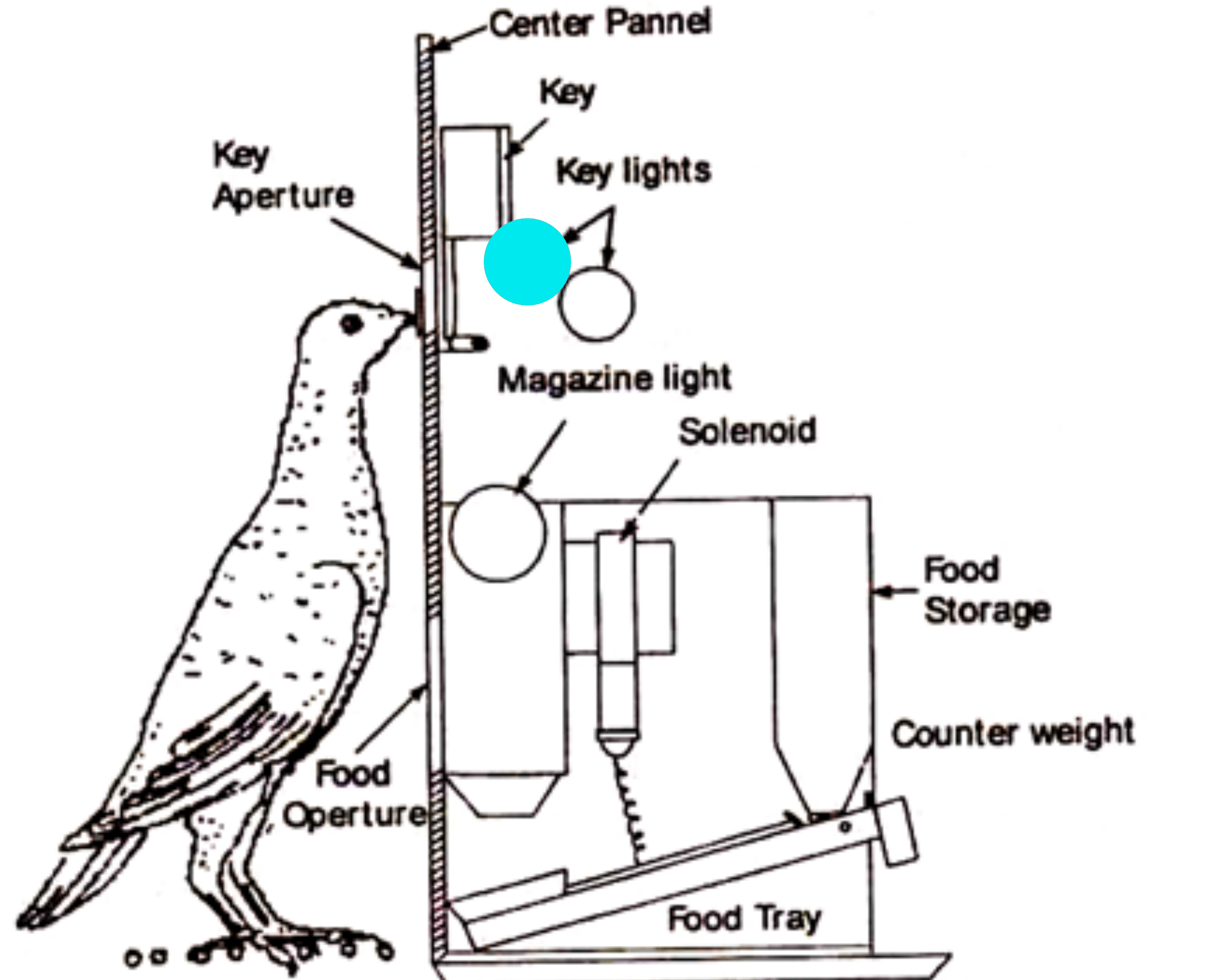
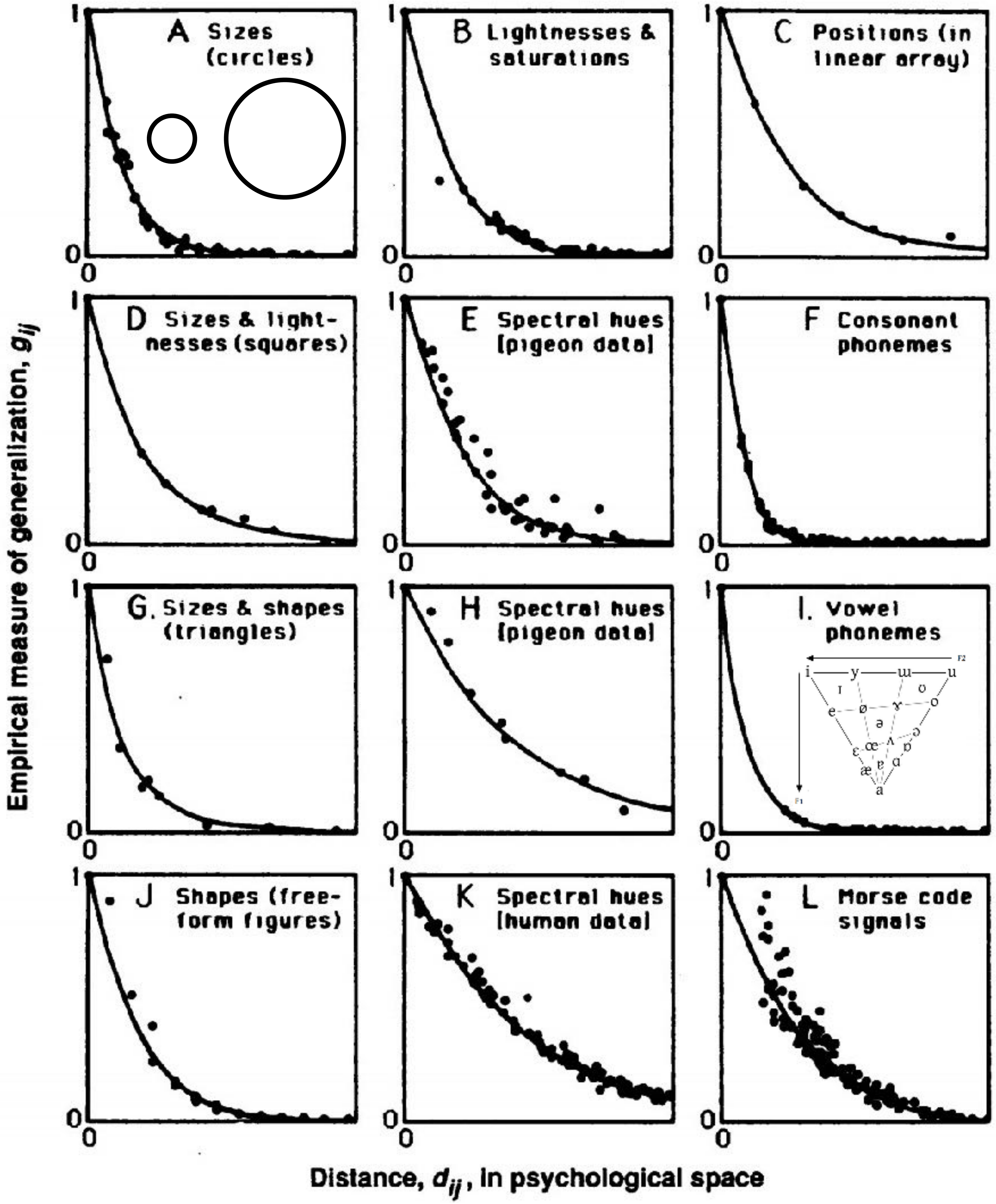
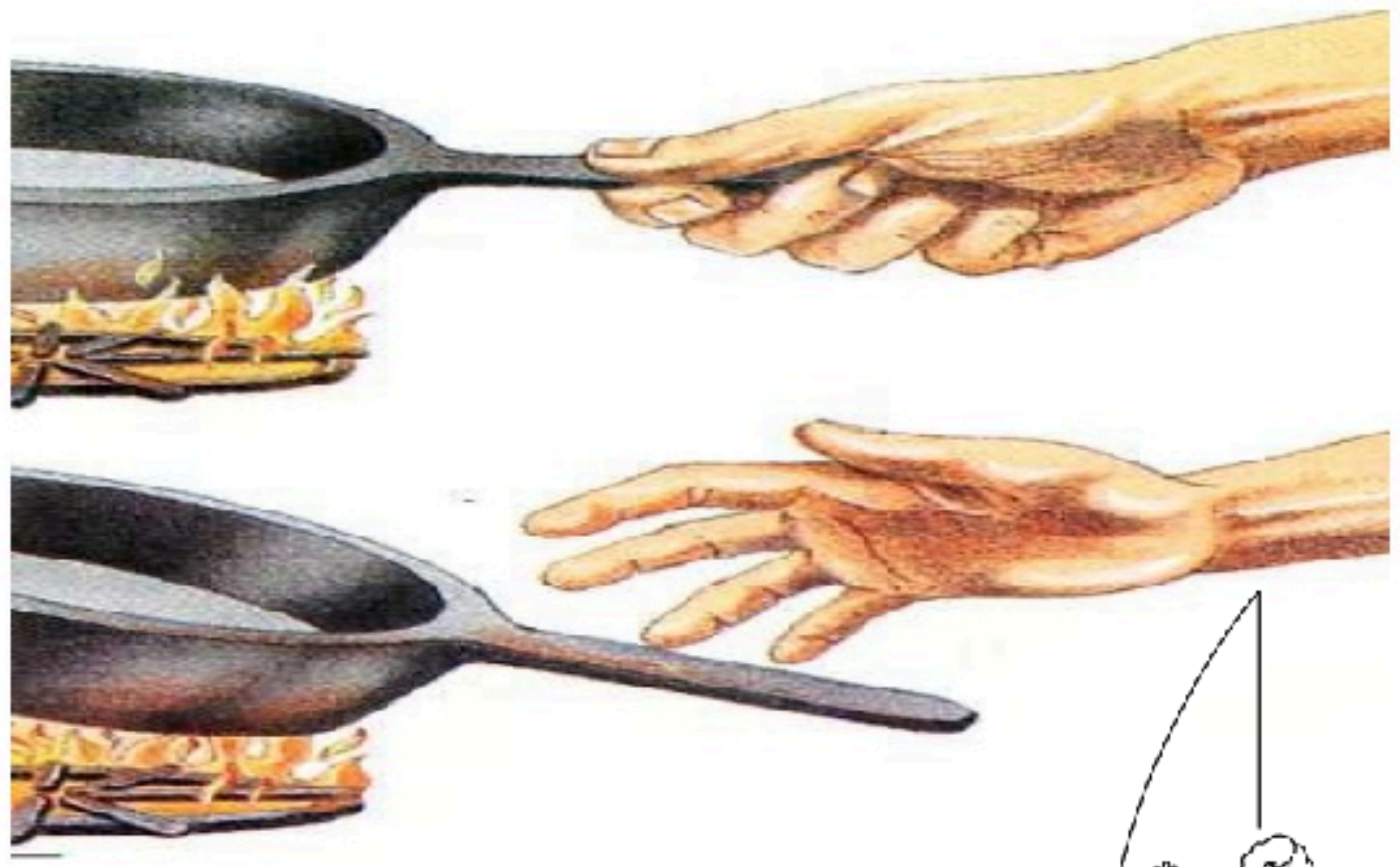


Illustration. Skinner box as adapted for the pigeon.



Shepard's (1987) Law of Generalization

STIMULUS + RESPONSE = LEARNING



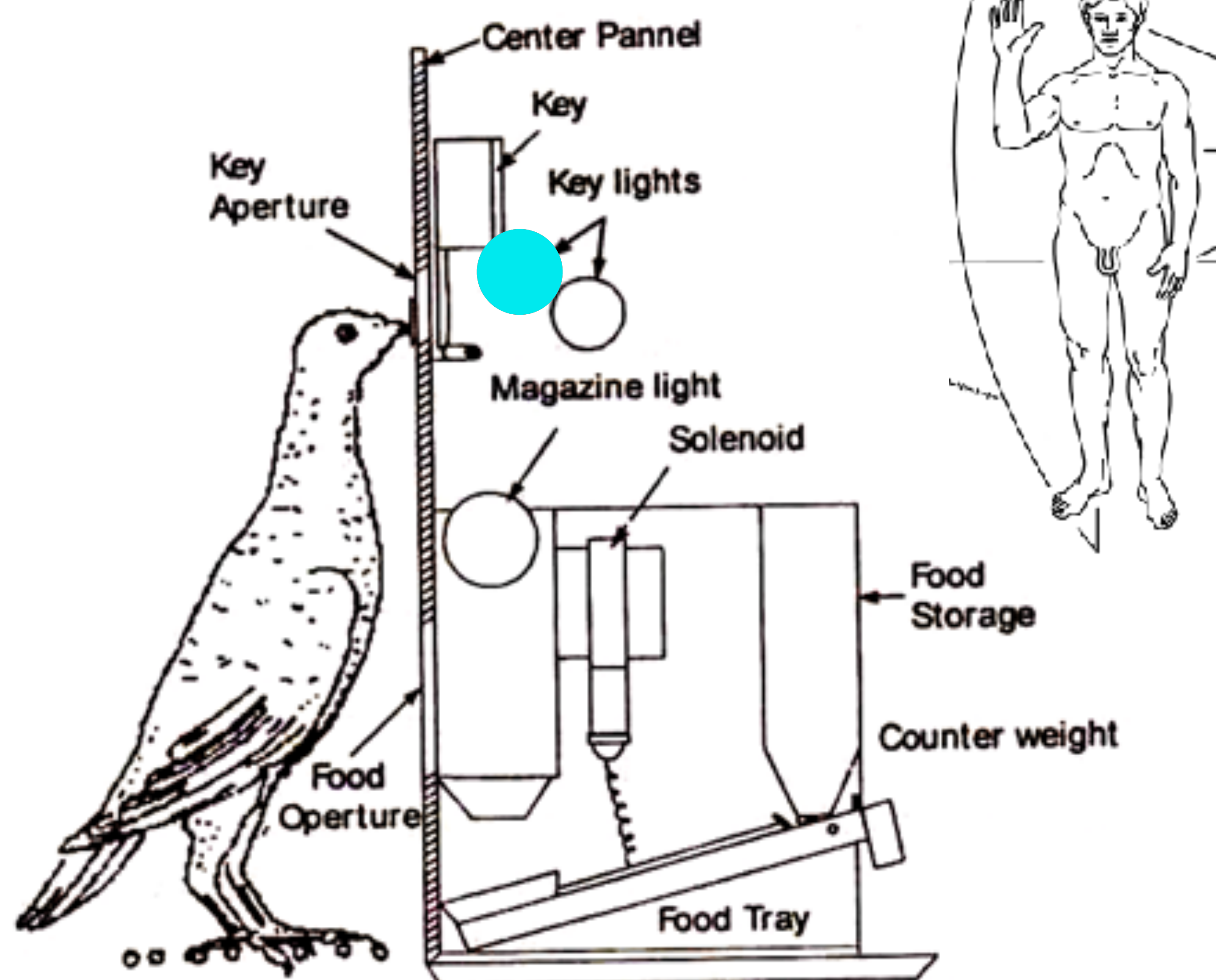
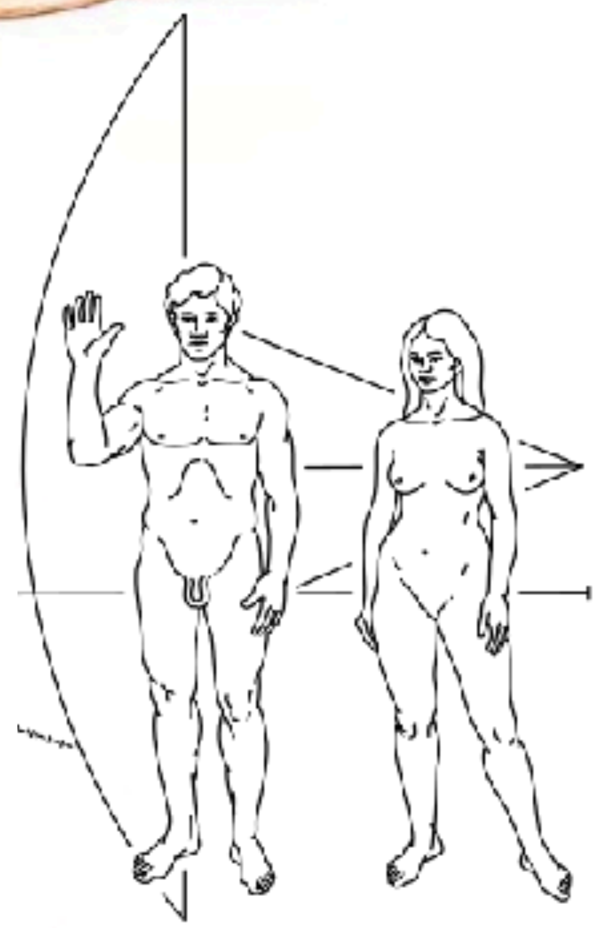
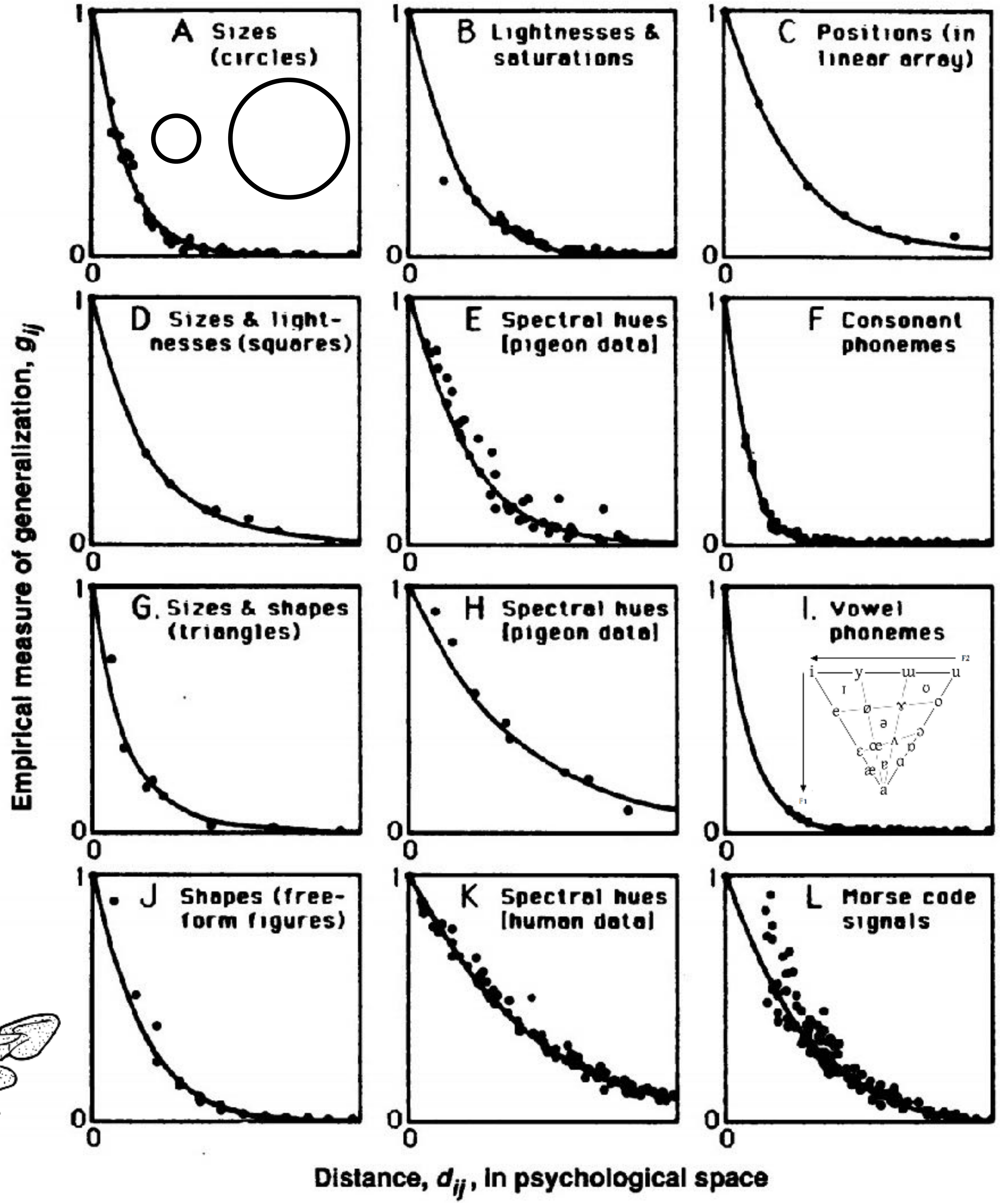


Illustration. Skinner box as adapted for the pigeon.





Generalization in Psychological Space

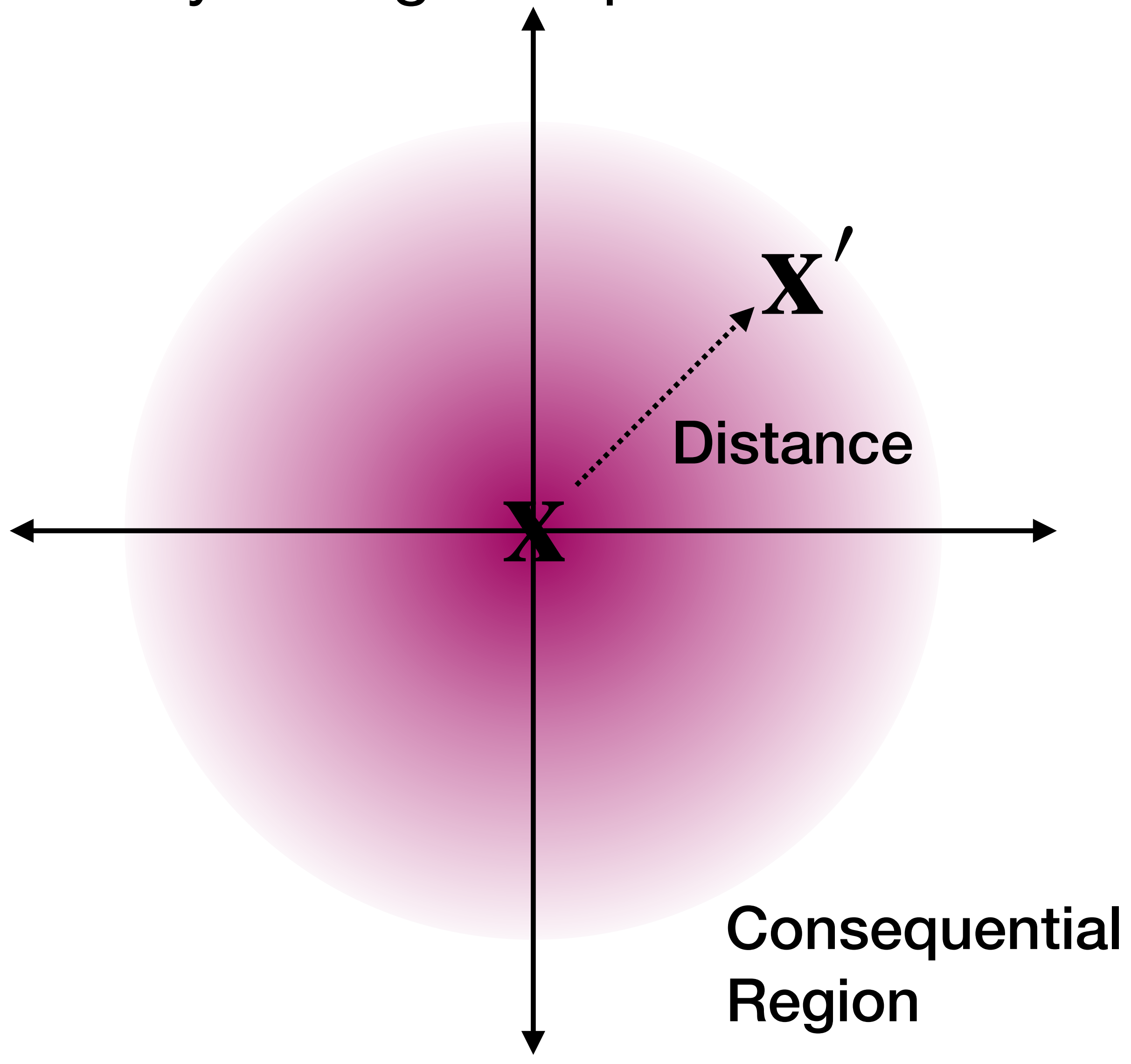
- Shepard (1987) believed that representations about categories or natural kinds correspond to a *consequential region* in psychological space

- Generalization arises from uncertainty about the extent of these regions

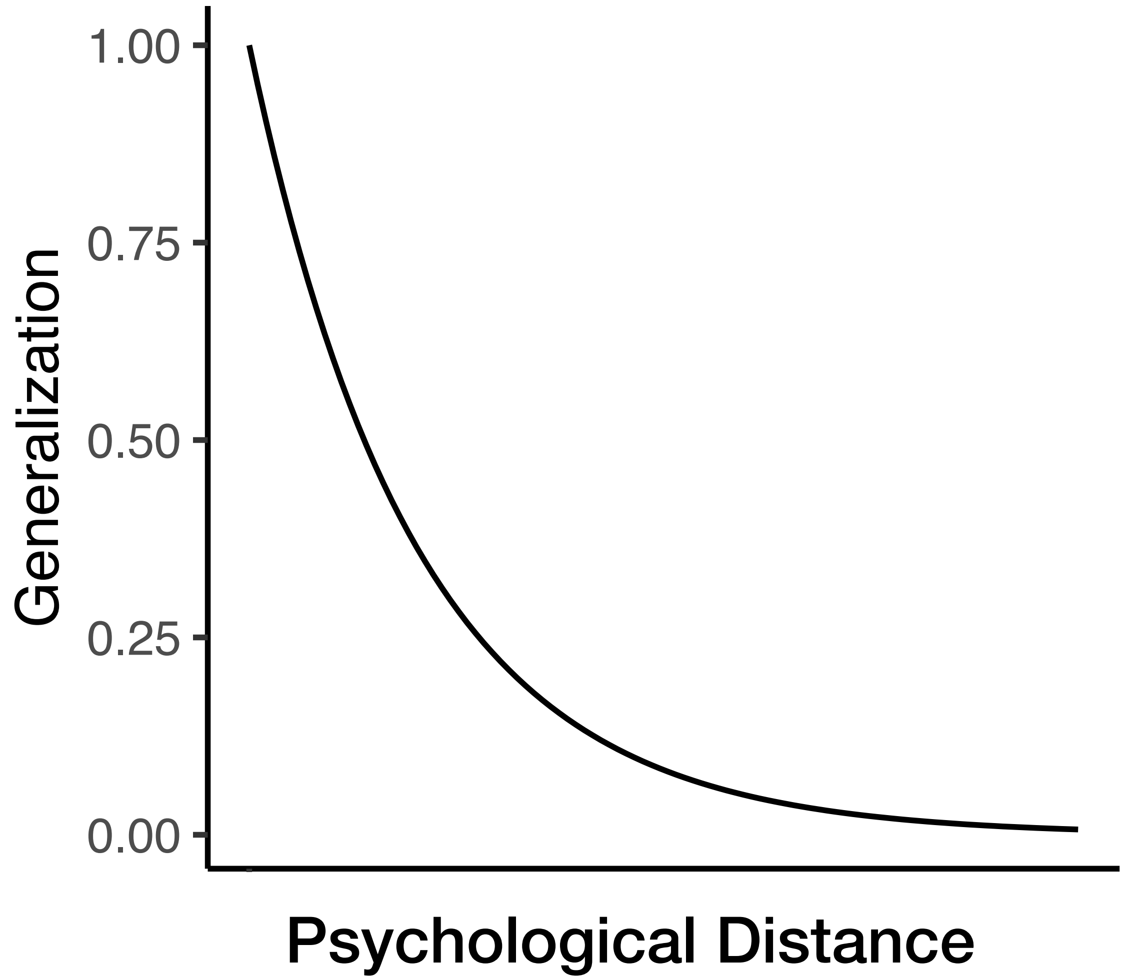
- As representational distance between stimuli **x** and **x**' increases (i.e., become less **similar**), they are less likely to belong to the same region, and thus produce less similar outcomes

- This produces the smooth gradient of generalization

Psychological Space



Generalization gradient



We generalize from one situation to another not because we cannot tell the difference between the two situations but because we judge that they are likely to belong to a set of situations having the same consequence. Generalization, which stems from uncertainty about the distribution of consequential stimuli in psychological space, is thus to be distinguished from failure of discrimination, which stems from uncertainty about the relative locations of individual stimuli in that space.

Shepard (*Science*, 1987)

Measurement invariance explains the universal law of generalization for psychological perception

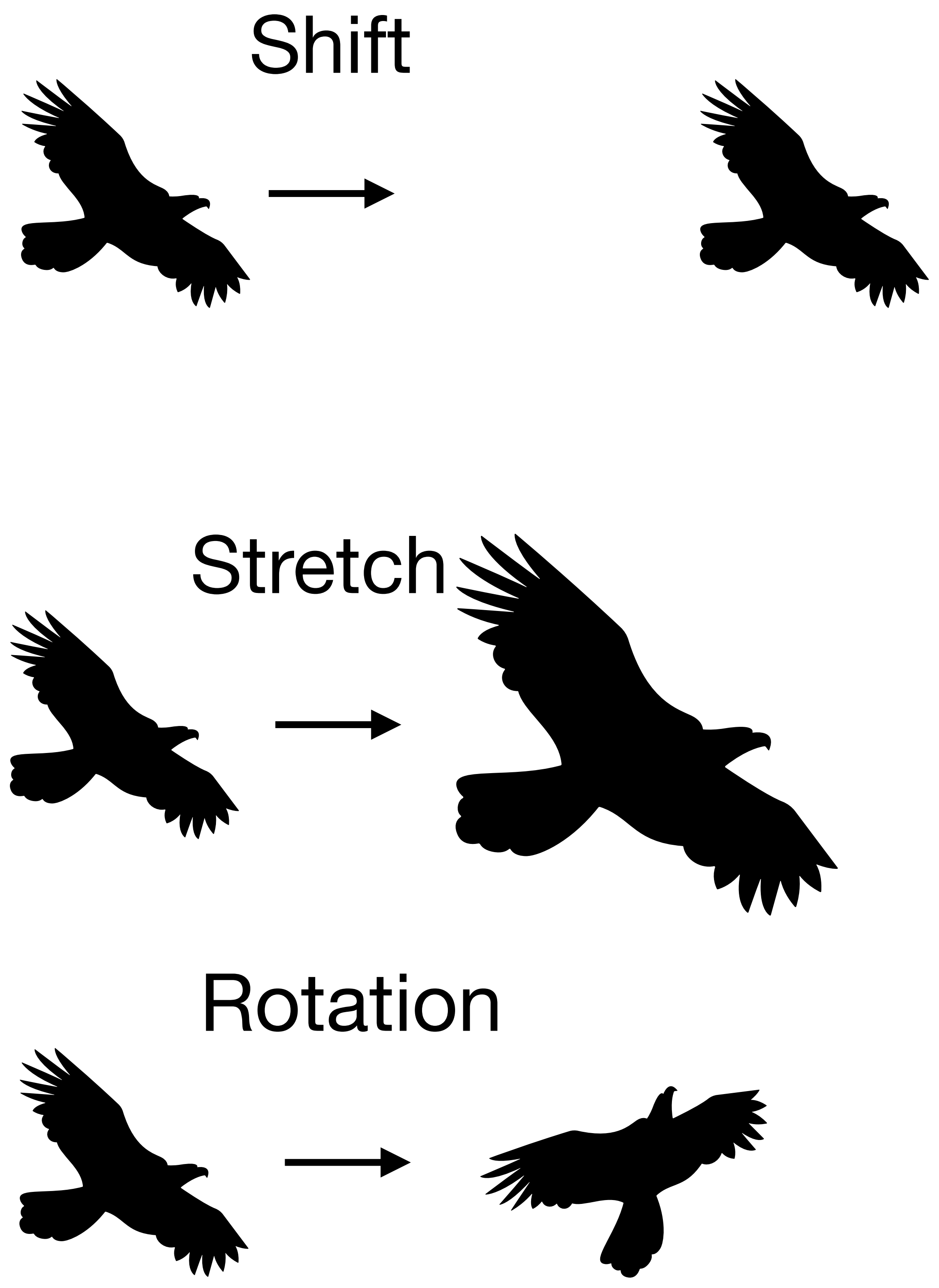
Steven A. Frank[a,1]

[a]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697-2525

Edited by Günter P. Wagner, Yale University, New Haven, CT, and approved August 15, 2018 (received for review June 7, 2018)

The universal law of generalization describes how animals discriminate between alternative sensory stimuli. On an appropriate perceptual scale, the probability that an organism perceives two stimuli as similar typically declines exponentially with the difference on the perceptual scale. Exceptions often follow a Gaussian probability pattern rather than an exponential pattern. Previous explanations have been based on underlying theoretical frameworks such as information theory, Kolmogorov complexity, or empirical multidimensional scaling. This article shows that the few inevitable invariances that must apply to any reasonable perceptual scale provide a sufficient explanation for the universal exponential law of generalization. In particular, reasonable measurement scales of perception must be invariant to shift by a constant value, which by itself leads to the exponential form. Similarly, reasonable measurement scales of perception must be invariant to multiplication, or stretch, by a constant value, which leads to the conservation of the slope of discrimination with perceptual difference. In some cases, an additional assumption about exchangeability or rotation of underlying perceptual dimensions leads to a Gaussian pattern of discrimination, which can be understood as a special case of the more general exponential form. The three measurement invariances of shift, stretch, and rotation provide a sufficient explanation for the universally observed patterns of perceptual generalization. All of the additional assumptions and language associated with information, complexity, and empirical scaling are superfluous with regard to the broad patterns of perception.

scaling patterns | categorization | sensory information | animal behavior | probability theory

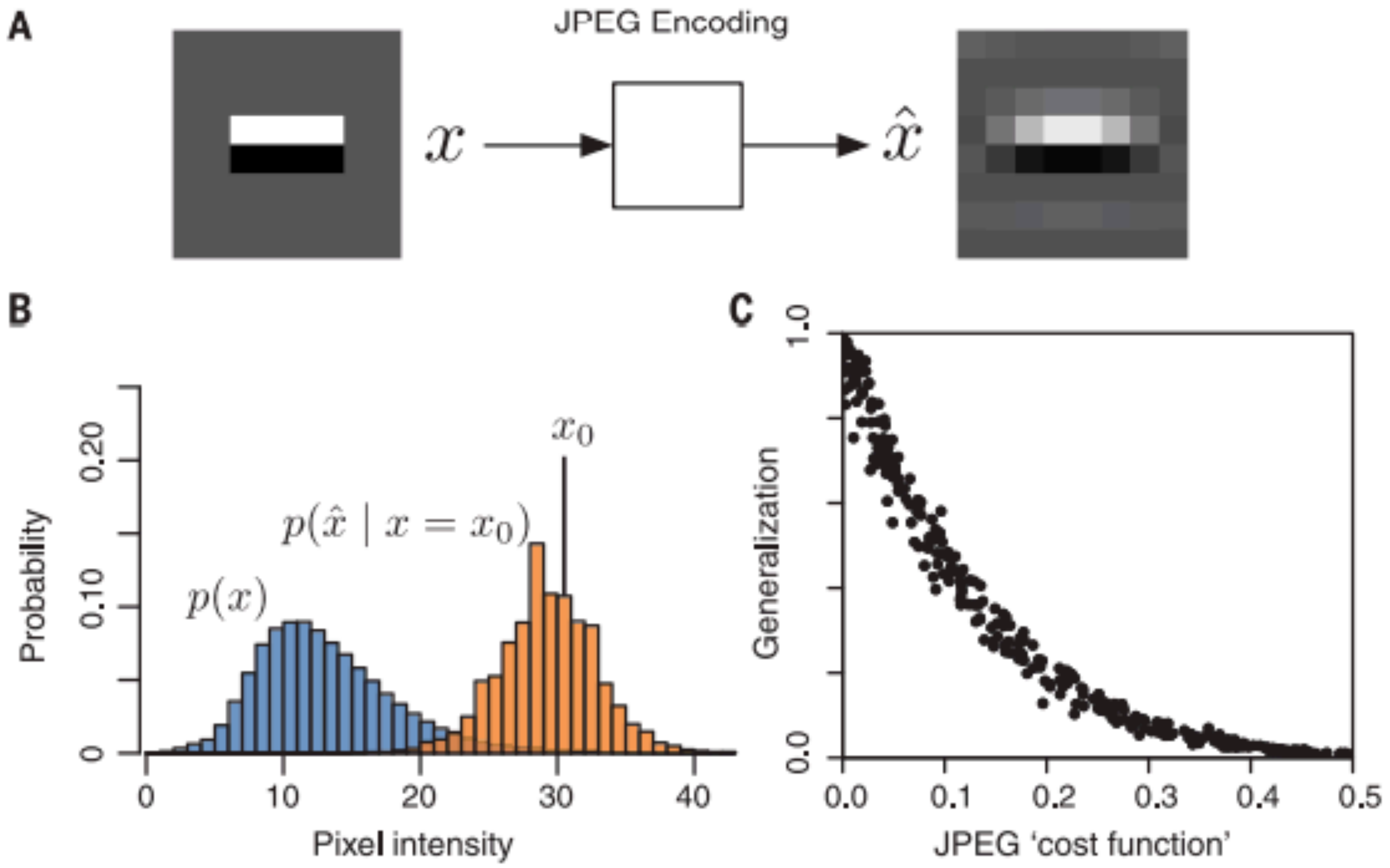


COGNITIVE PSYCHOLOGY

Efficient coding explains the universal law of generalization in human perception

Chris R. Sims*

Perceptual generalization and discrimination are fundamental cognitive abilities. For example, if a bird eats a poisonous butterfly, it will learn to avoid preying on that species again by generalizing its past experience to new perceptual stimuli. In cognitive science, the "universal law of generalization" seeks to explain this ability and states that generalization between stimuli will follow an exponential function of their distance in "psychological space." Here, I challenge existing theoretical explanations for the universal law and offer an alternative account based on the principle of efficient coding. I show that the universal law emerges inevitably from any information processing system (whether biological or artificial) that minimizes the cost of perceptual error subject to constraints on the ability to process or transmit information.



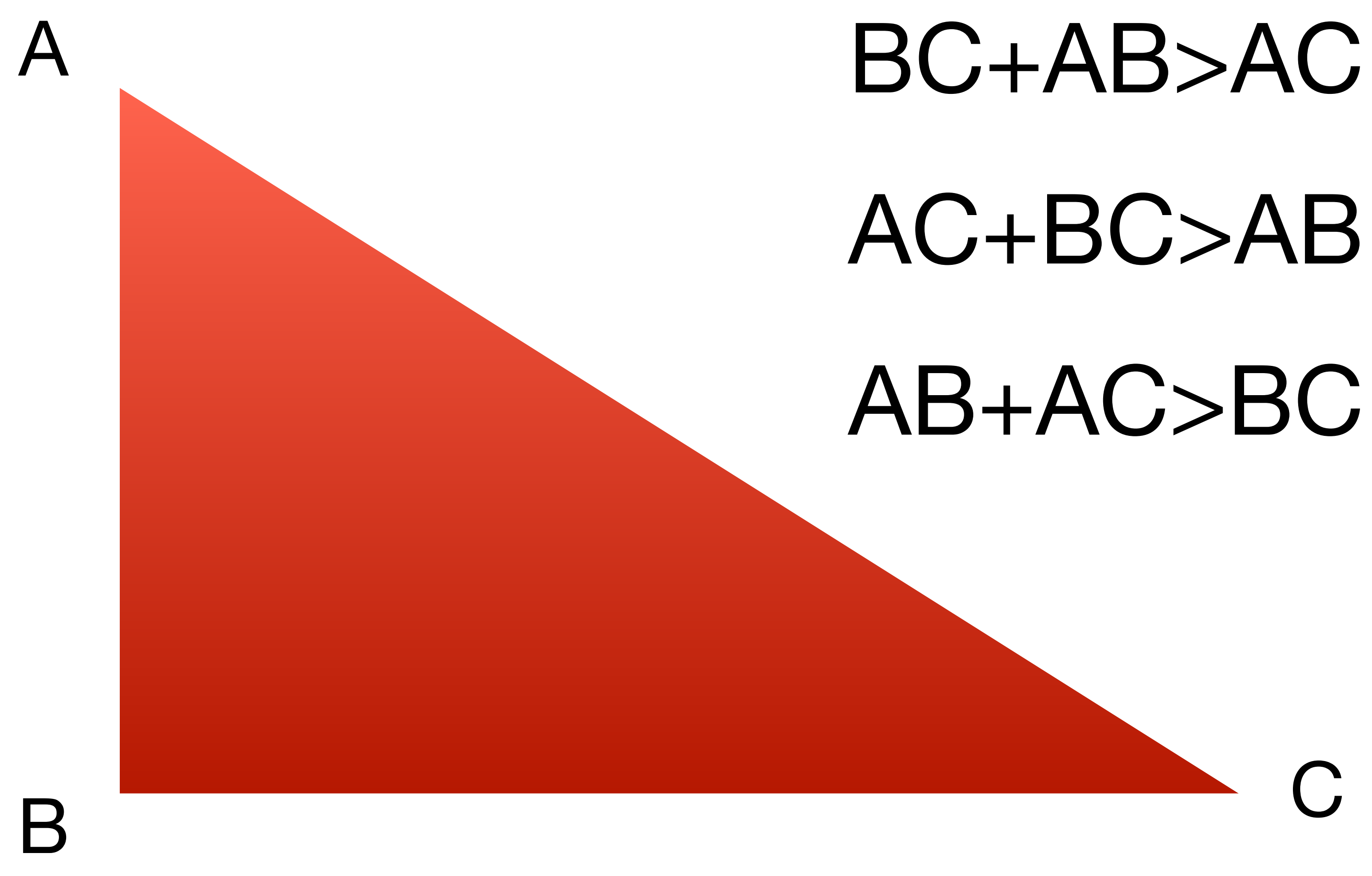
Limitations of metric similarity

- Two definitive properties are *symmetry* and *triangle inequality*
- But they are often violated in human judgments of similarity (Tversky, 1977)

Symmetry

$$d(\mathbf{x}, \mathbf{x}') = d(\mathbf{x}', \mathbf{x})$$

Triangle Inequality



BC+AB>AC

AC+BC>AB

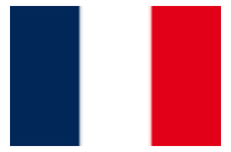
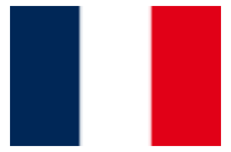
AB+AC>BC

Limitations of metric similarity

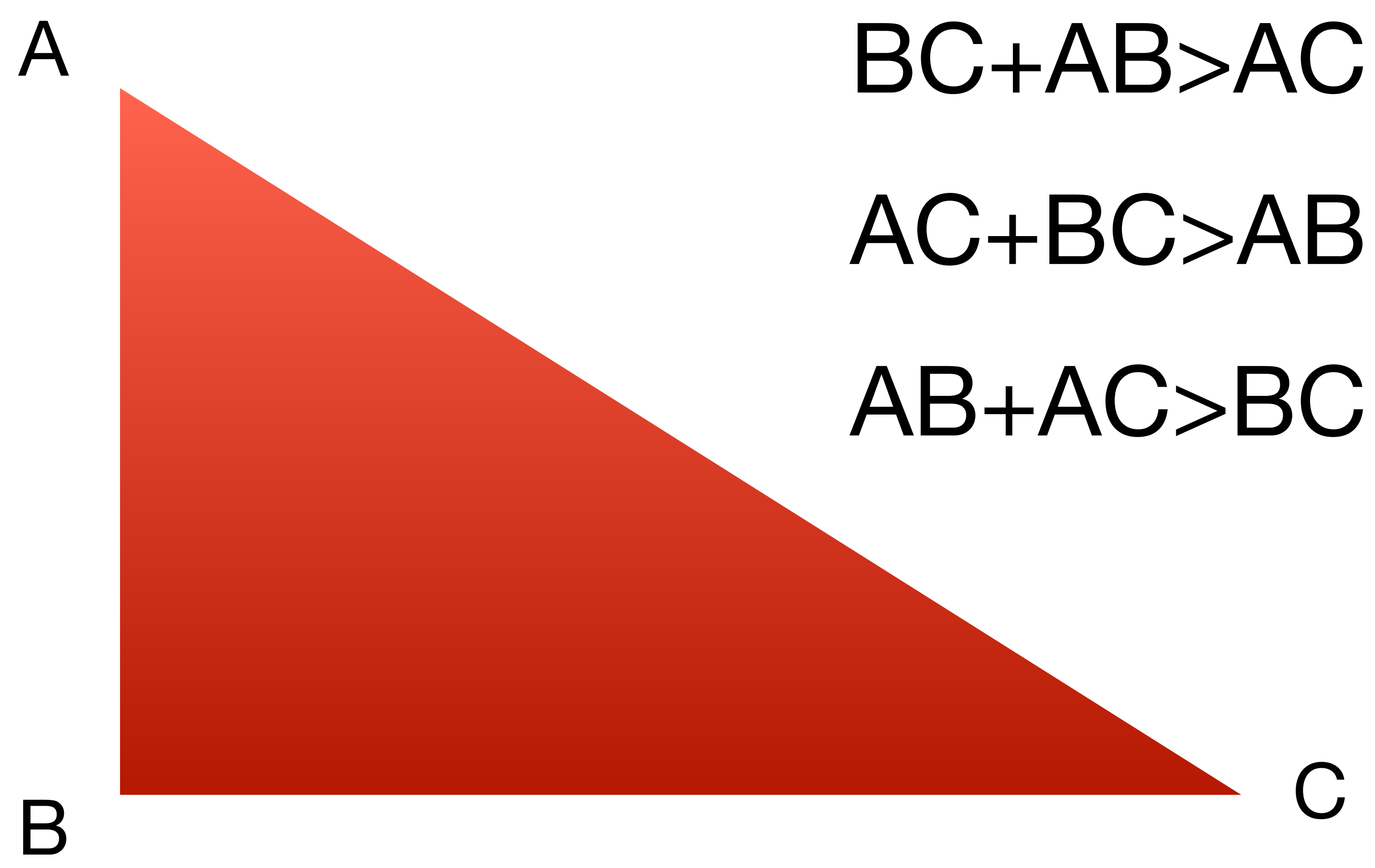
- Two definitive properties are *symmetry* and *triangle inequality*
- But they are often violated in human judgments of similarity (Tversky, 1977)

Symmetry

$$d(\mathbf{x}, \mathbf{x}') = d(\mathbf{x}', \mathbf{x})$$

$d($  ,  $) \neq d($  ,  $)$

Triangle Inequality



BC+AB>AC

AC+BC>AB

AB+AC>BC

Limitations of metric similarity

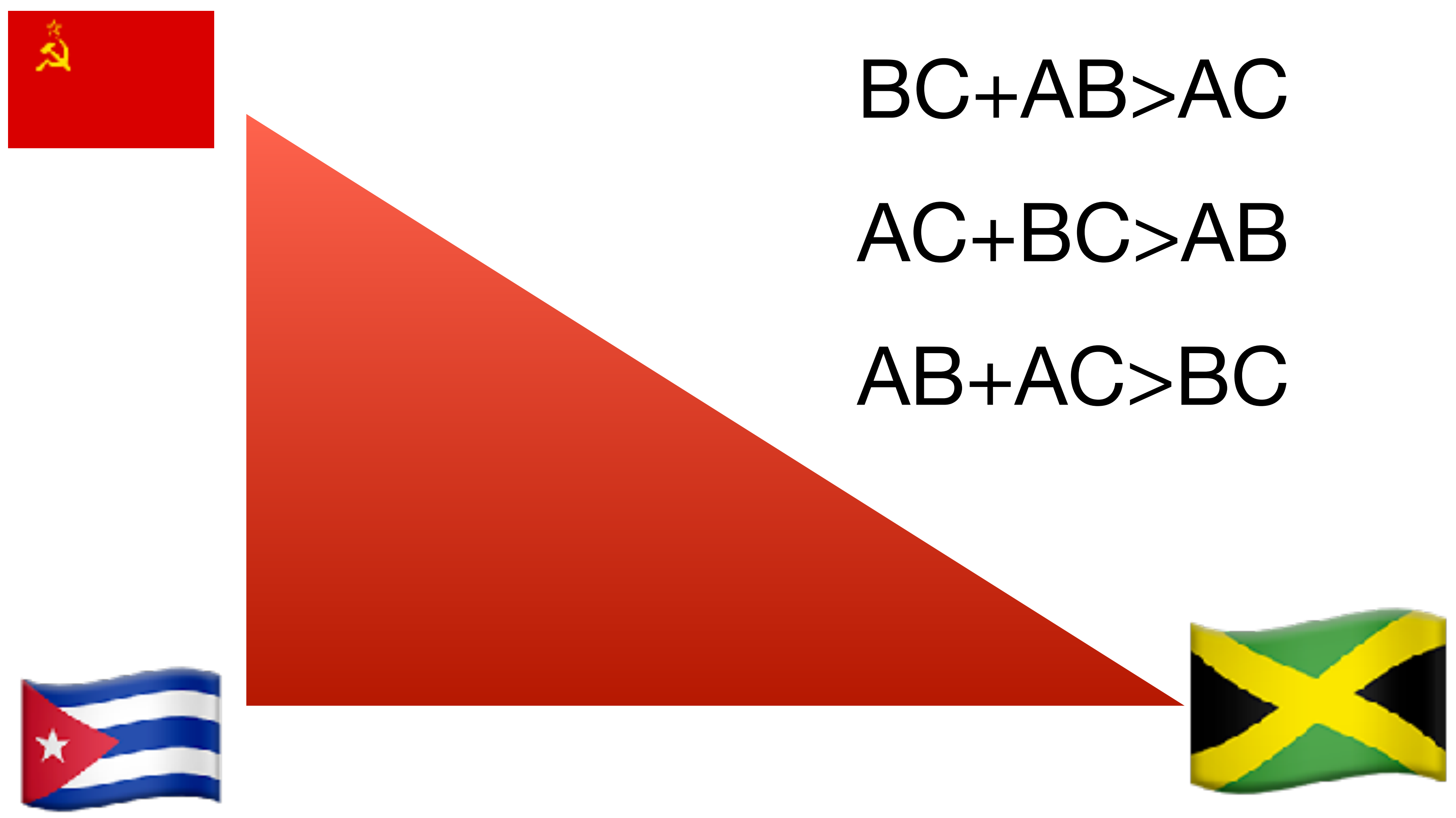
- Two definitive properties are *symmetry* and *triangle inequality*
- But they are often violated in human judgments of similarity (Tversky, 1977)

Symmetry

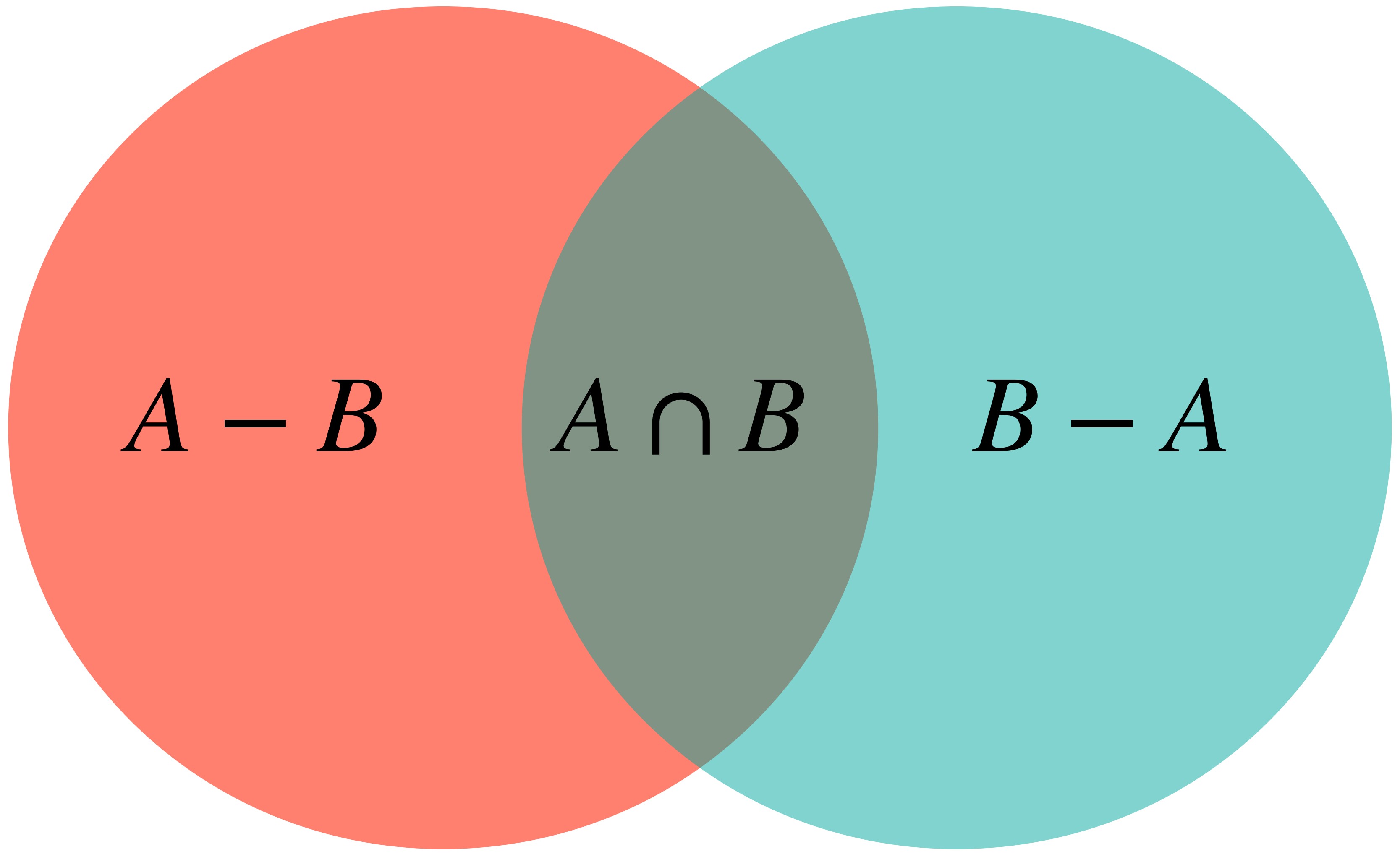
$$d(\mathbf{x}, \mathbf{x}') = d(\mathbf{x}', \mathbf{x})$$

$d($🇩🇪, 🇫🇷$) \neq d($🇫🇷, 🇩🇪$)$

Triangle Inequality



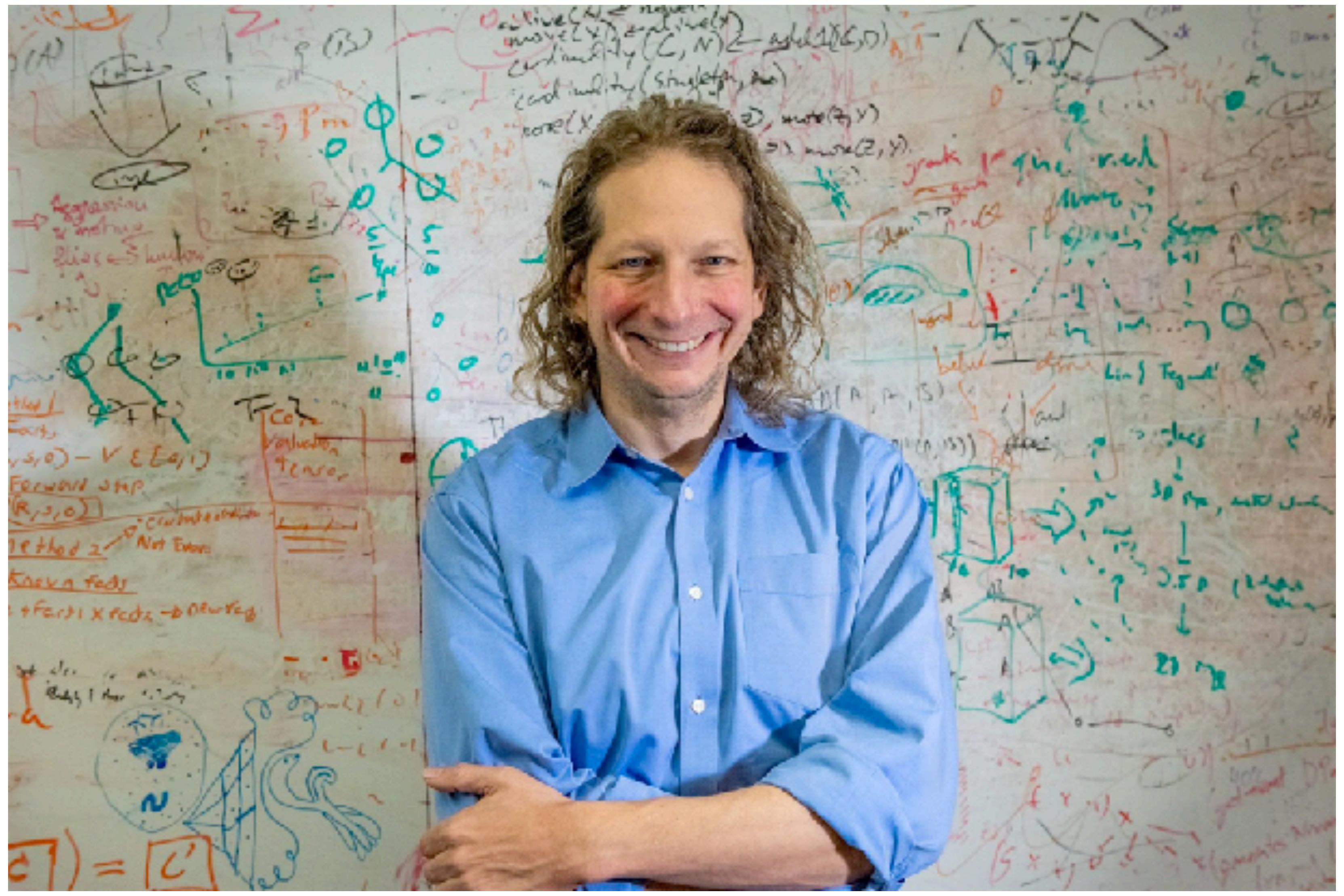
Contrast model



$$\text{sim}(A, B) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$$

- θ, α, β are free parameters
- To translate into Shepard's language, rather than consequential regions in psychological space, concepts are defined based on sets of features
  - Similar to *family resemblance theory* (Rosch & Mervis, 1975)
- Common and disjoint features may be weighted differently
- A more refined similarity theory that allows for asymmetric similarity judgments that can also violate triangle inequality

Bayesian concept learning as a hybrid approach



- Much of modern cognitive science is dominated by Bayesian inference

- Josh Tenenbaum and Tom Griffiths are two individuals who are largely responsible for it's popularity

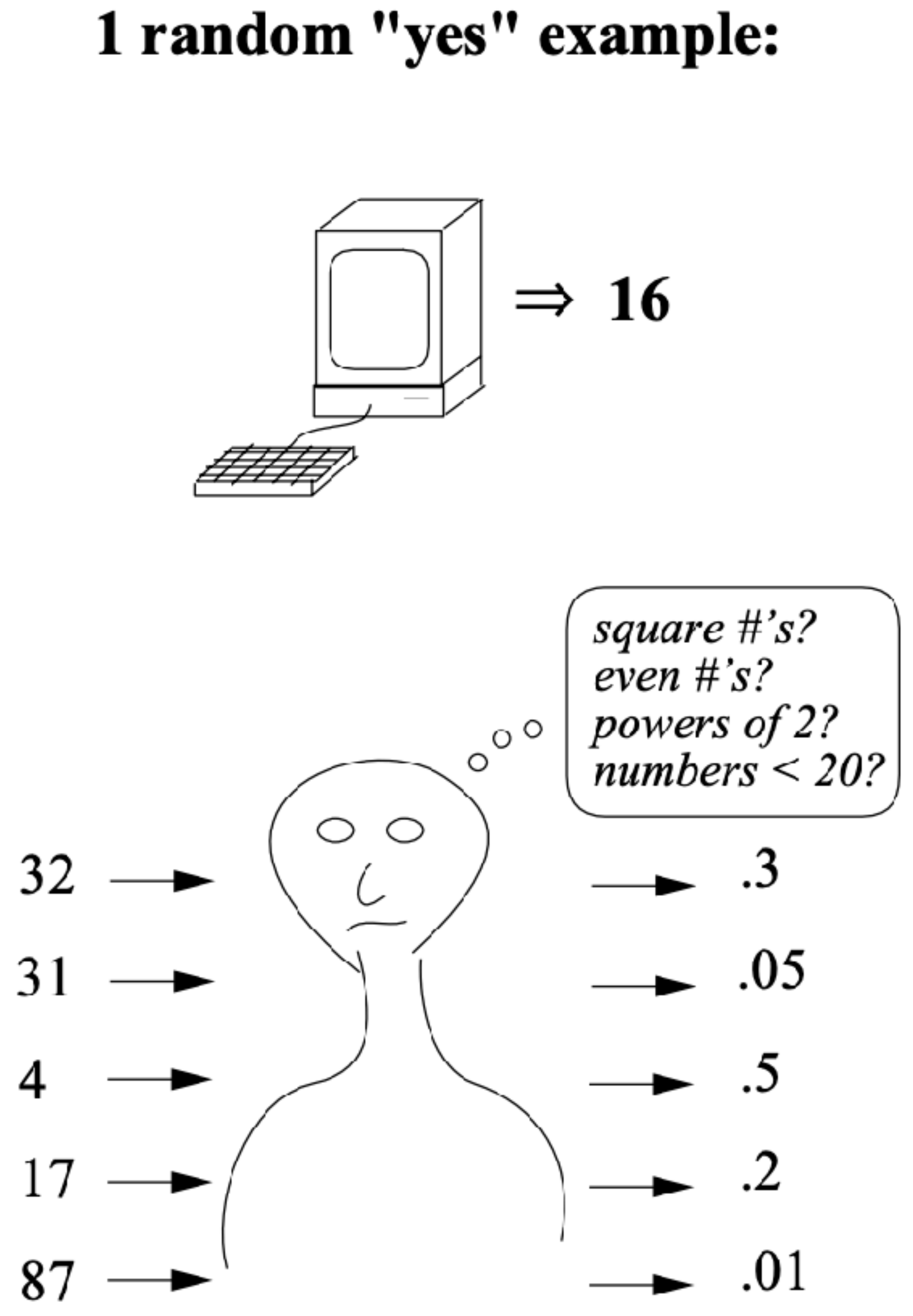
- The same basic concept can explain a huge host of problems, from language acquisition, to structure learning, to program induction

- But it all started with a number game and a model of probablistic rule learning from Josh's PhD thesis

5 min break

Number concepts

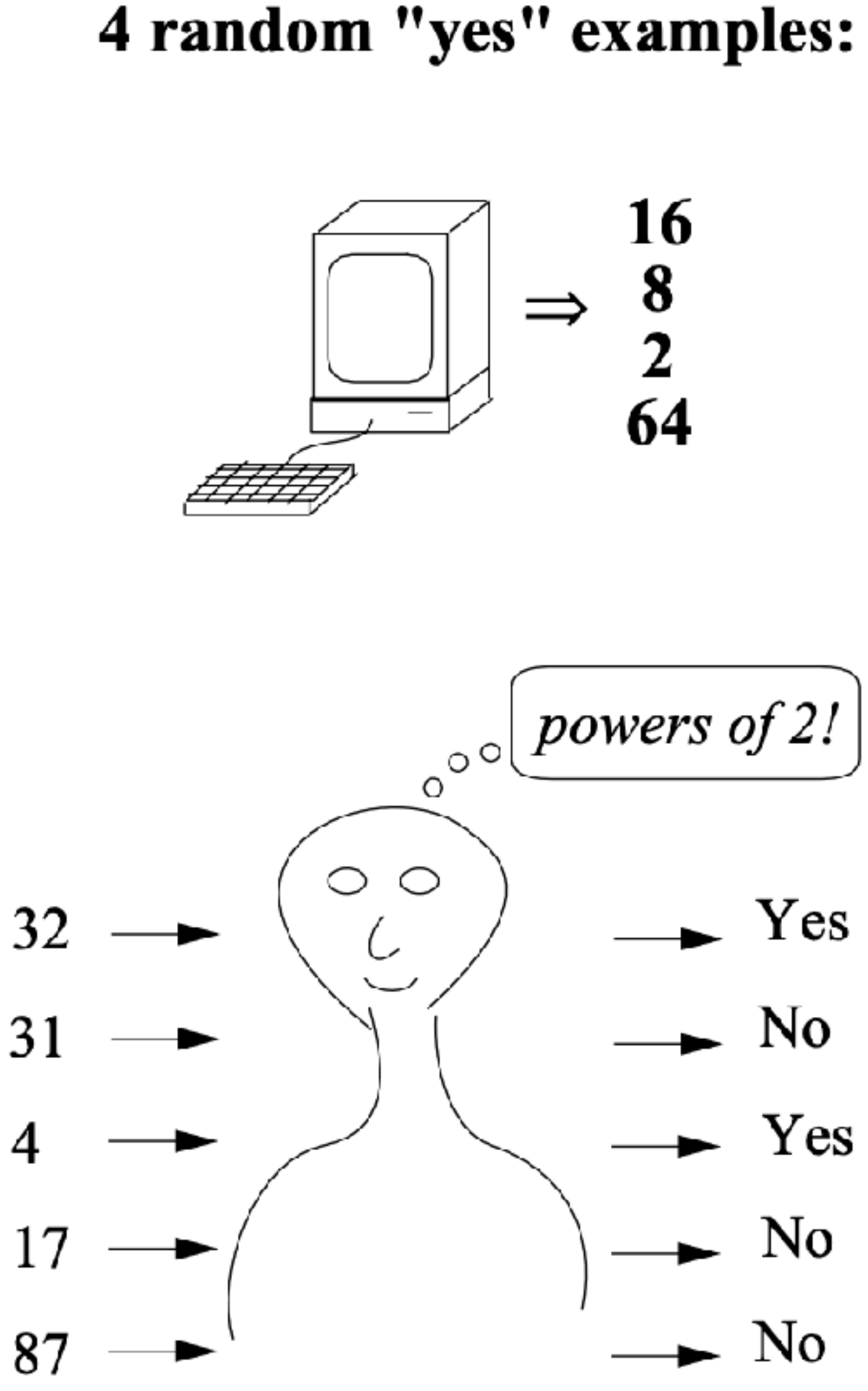
- Examples:
  - X is an even number
  - X is between 30 and 45
  - X is a prime number
- A computer generates a random number from a chosen concept, and you need to guess another number that is likely to fit



Tenenbaum (*PhD thesis* 1999)

Number concepts

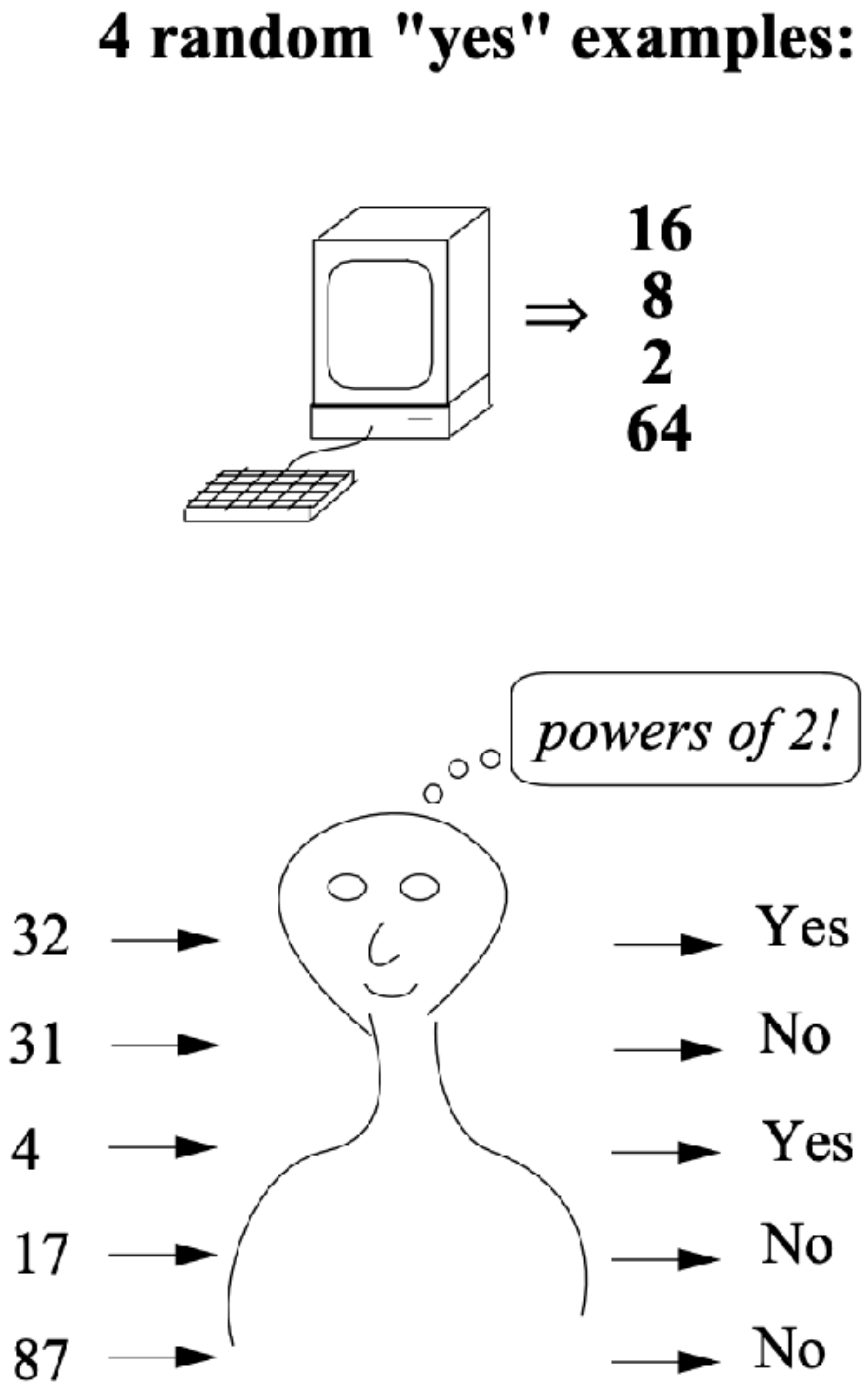
- Examples:
  - X is an even number
  - X is between 30 and 45
  - X is a prime number
- A computer generates a random number from a chosen concept, and you need to guess another number that is likely to fit



Tenenbaum (*PhD thesis* 1999)

Number concepts

- Examples:
  - X is an even number
  - X is between 30 and 45
  - X is a prime number
- A computer generates a random number from a chosen concept, and you need to guess another number that is likely to fit
- *Even restricting the game to natural numbers between 1 and 100, there are more than a billion billion billion subsets of numbers that such a program could possibly have picked out and which are consistent with the observed “yes" examples of 16, 8, 2, and 64*



Tenenbaum (*PhD thesis* 1999)

Bayesian Concept Learning

- Example: The concept of healthy person
- Problem: Given a set of examples (*x*'s in the plot), what is the probablity that some new example y will fall within consequential region *C* defining a healthy person?
- Each *h* is a hypothesis (illustrated as rectangles) about the category boundary

$$p(y \in C|x) = \sum_{h: y \in h} p(h|x).$$

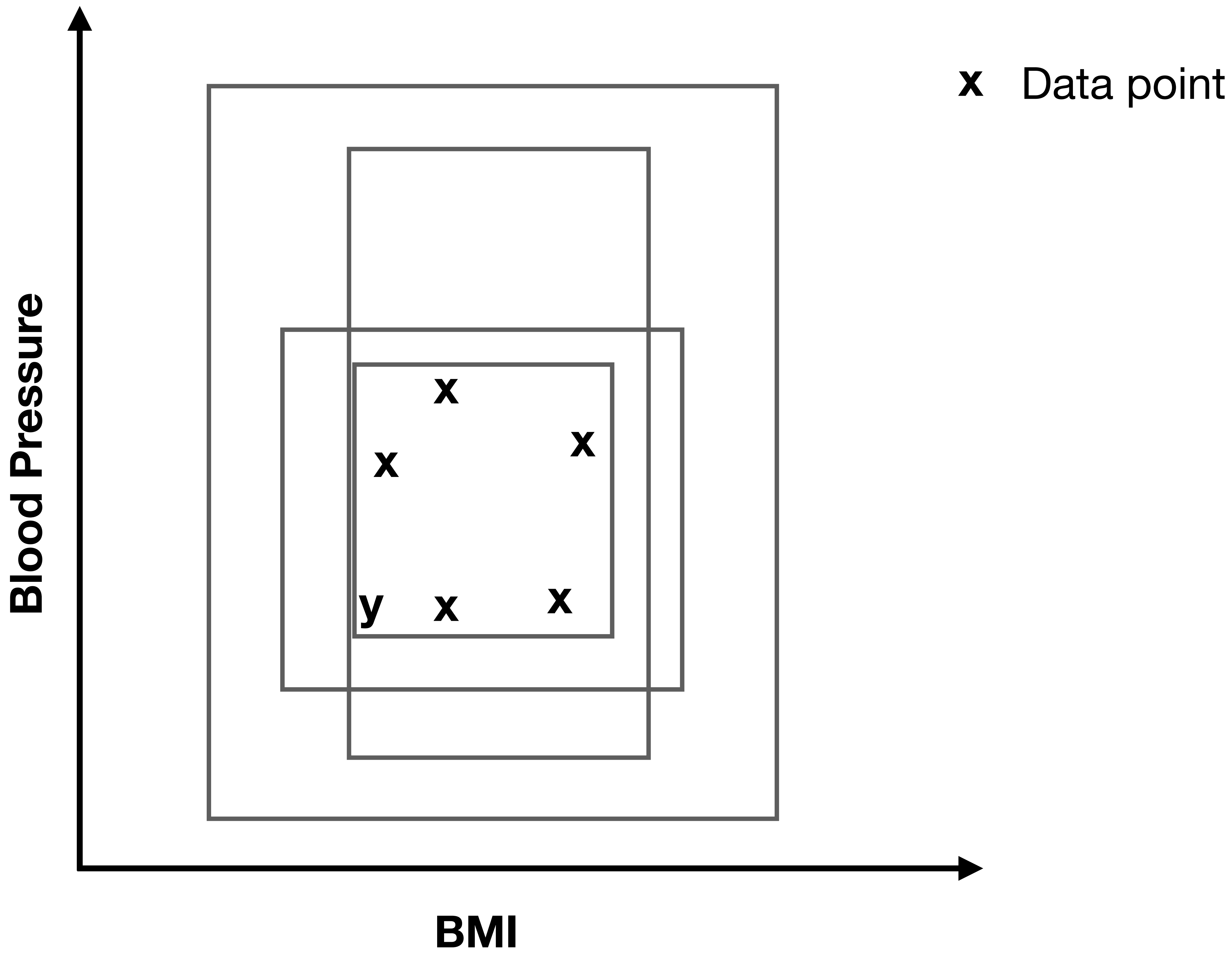
Sum over hypotheses that include y

Bayes' rule

$$p(h|x) = \frac{p(x|h)p(h)}{p(x)}$$

likelihood * prior / evidence

$$= \frac{p(x|h)p(h)}{\sum_{h' \in \mathcal{H}} p(x|h')p(h')}.$$



Tenenbaum (*NIPS* 1999)
Tenenbaum & Griffiths (*BBS* 2001)

Bayesian Concept Learning

Likelihood:

$$p(x|h) = \begin{cases} 1 & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \qquad \text{[weak sampling].}$$

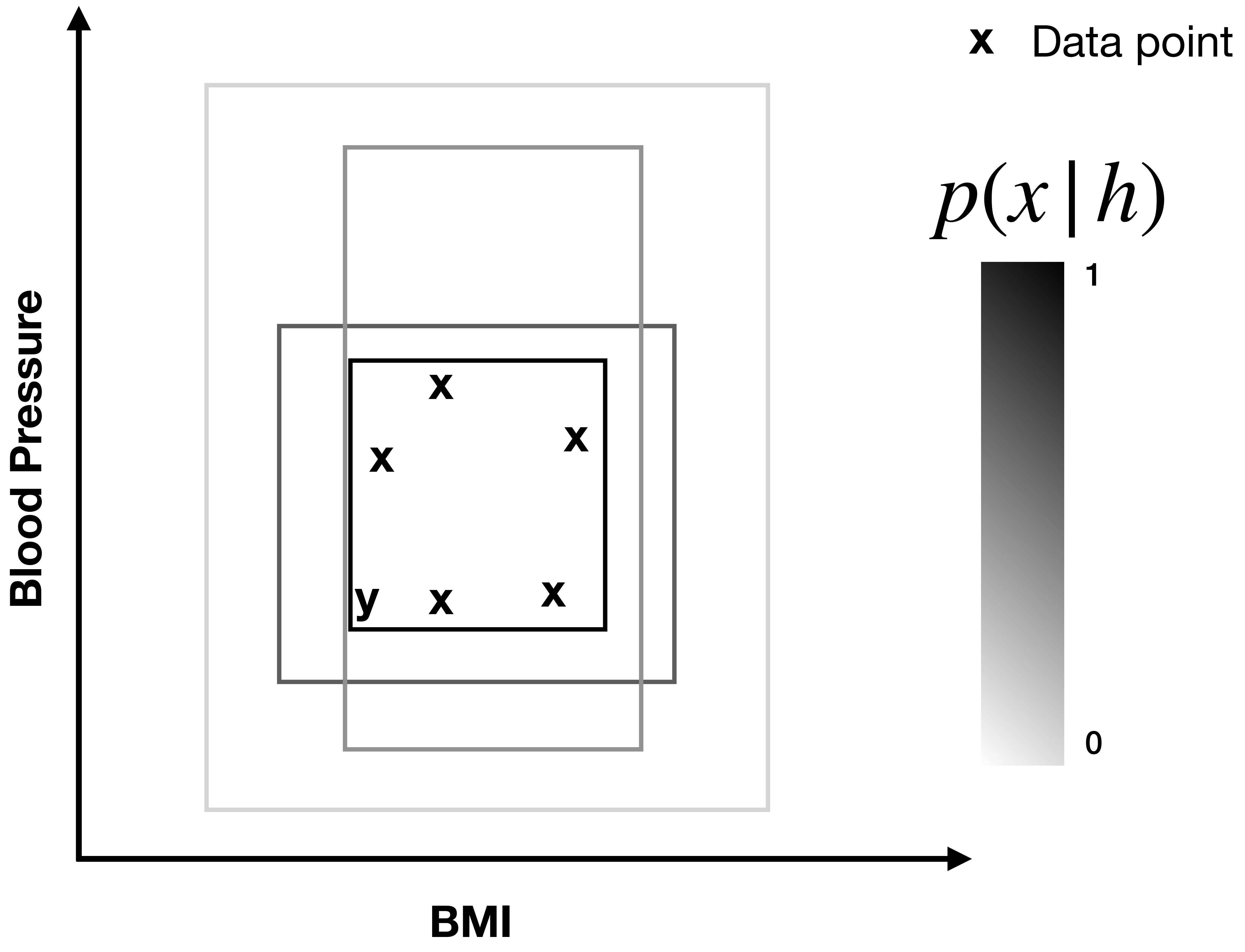
$$p(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \qquad \text{[strong sampling],}$$

Bayesian size principle: under strong sampling, smaller h (consistent with the data) are more likely

Multiple x's with multiple features:

$$p(X|h) = \prod_i p(x_i|h)$$

$$= \begin{cases} \frac{1}{|h|^n} & \text{if } x_1, \ldots, x_n \in h \\ 0 & \text{otherwise} \end{cases}$$



Tenenbaum (*NIPS* 1999)
Tenenbaum & Griffiths (*BBS* 2001)

Bayesian Concept Learning

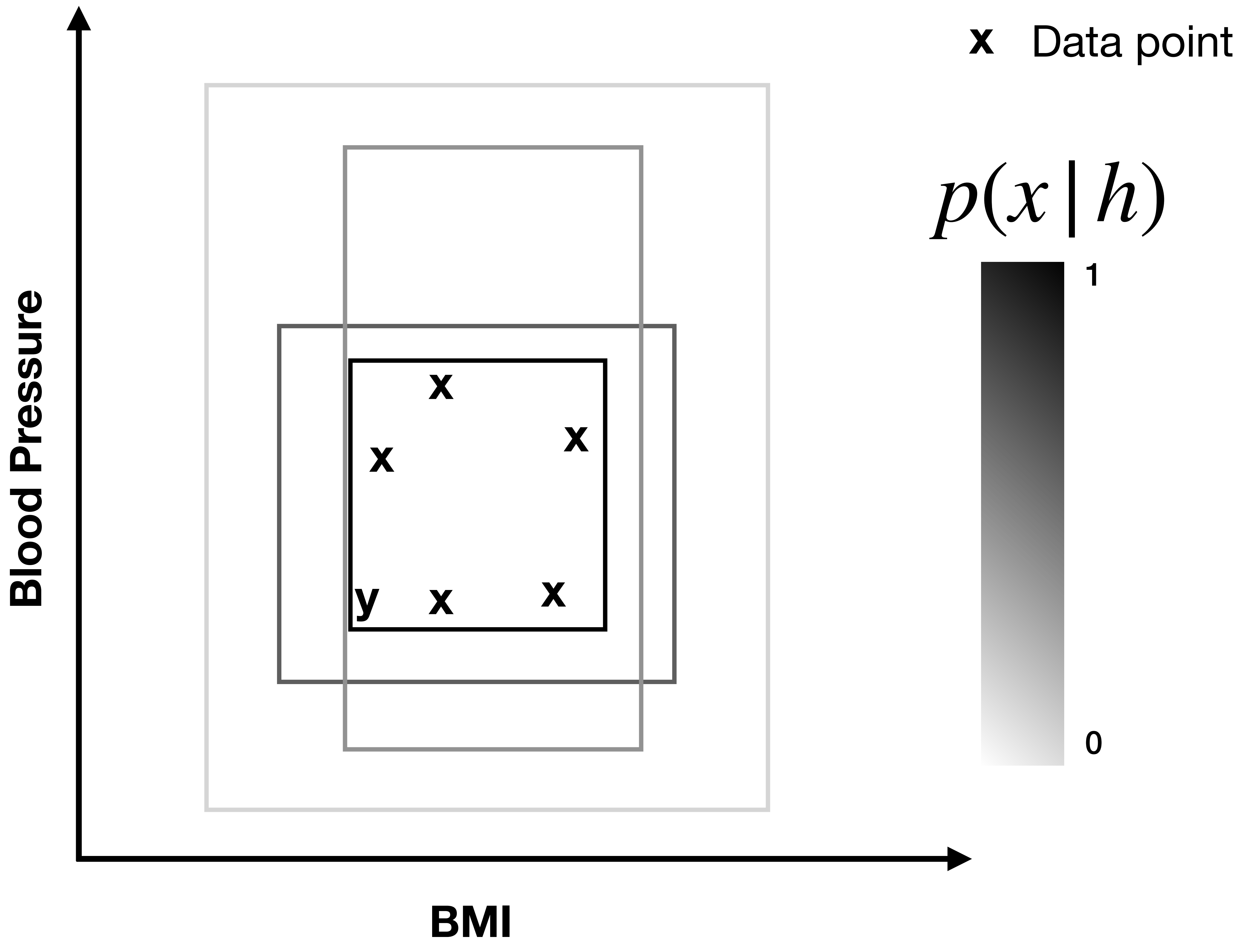
- The probability of y being in the same category of x is thus based on summing over all hypotheses consistent with the data

$$p(y \in C|x) = \sum_{h:y \in h} p(h|x).$$

- Where narrower hypotheses are favored under strong sampling

$$p(h|x) = \frac{p(x|h)p(h)}{p(x)}$$

$$p(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \qquad \text{[strong sampling]},$$



Tenenbaum (*NIPS* 1999)
Tenenbaum & Griffiths (*BBS* 2001)

Hypotheses can capture structured and arbitrary subsets of the data

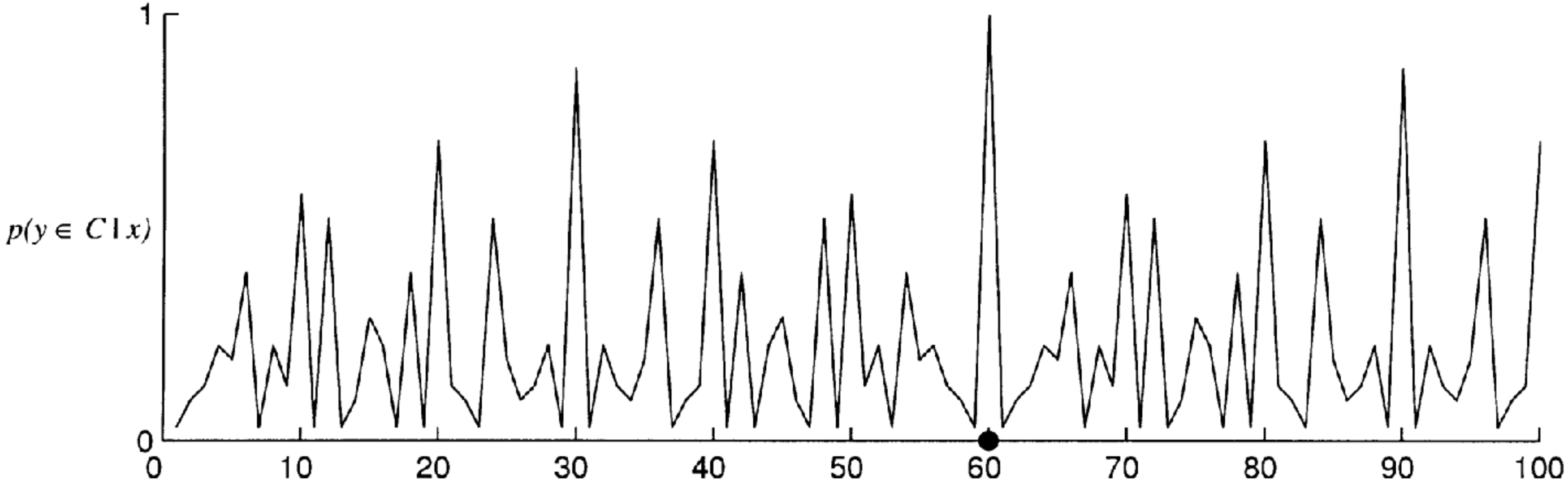
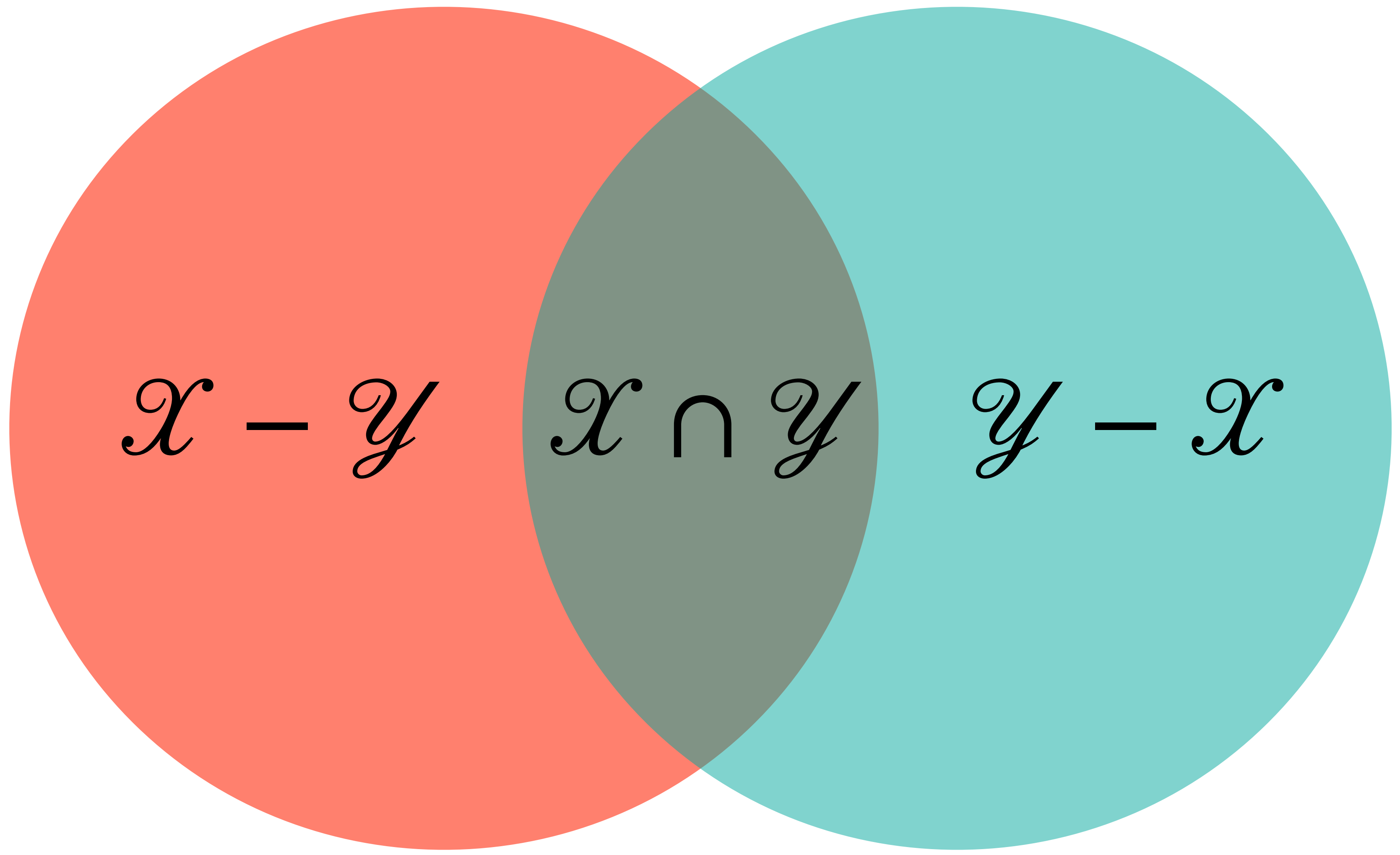
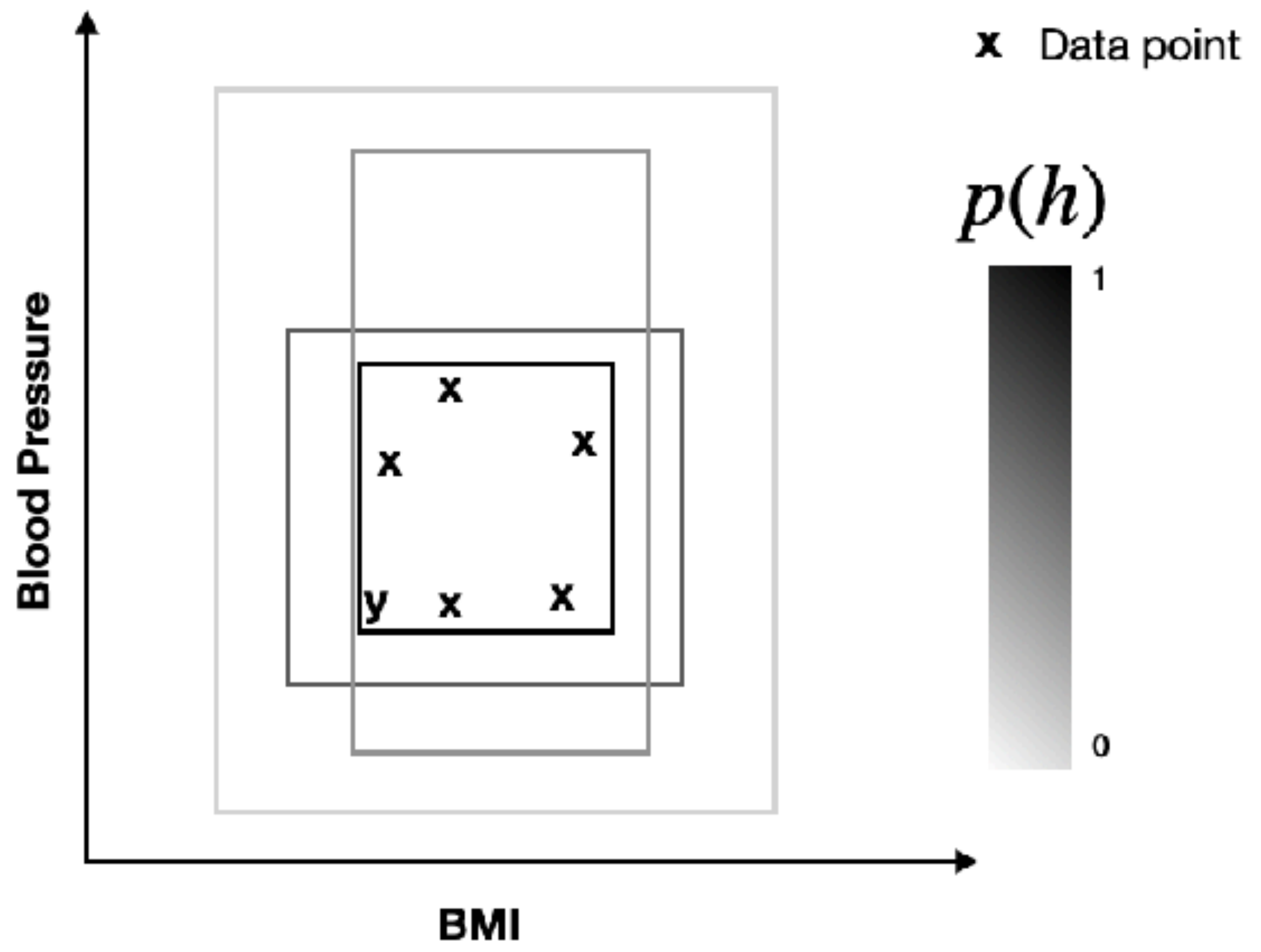


Figure 5. Bayesian generalization in the number game, given one example $x = 60$. The hypothesis space includes 33 mathematically consequential subsets (with equal prior probabilities): even numbers, odd numbers, primes, perfect squares, perfect cubes, multiples of a small number (3–10), powers of a small number (2–10), numbers ending in the same digit (1–9), numbers with both digits equal, and all numbers less than 100.

Bayesian Concept Learning Subsumes Tversky's Contrast Model





Contrast model

$$S(y,x) = \theta f(\mathcal{Y} \cap \mathcal{X}) - \alpha f(\mathcal{Y} - \mathcal{X}) - \beta f(\mathcal{X} - \mathcal{Y}),$$

Ratio model (alternative form)

$$S(y,x) = 1 \Big/ \left[1 + \frac{\alpha f(\mathcal{Y} - \mathcal{X}) + \beta f(\mathcal{X} - \mathcal{Y})}{f(\mathcal{Y} \cap \mathcal{X})} \right].$$

(equivalent when α=0 and β=1)

Bayesian concept learning

$$p(y \in C \mid x) = \sum_{h : y \in h} p(h \mid x).$$

$$= 1 \Big/ \left[1 + \frac{\sum_{h: x \in h, y \notin h} p(h,x)}{\sum_{h: x, y \in h} p(h,x)} \right].$$

Bayesian Concept Learning Extends Shepard's Law of Generalization to Multiple Examples

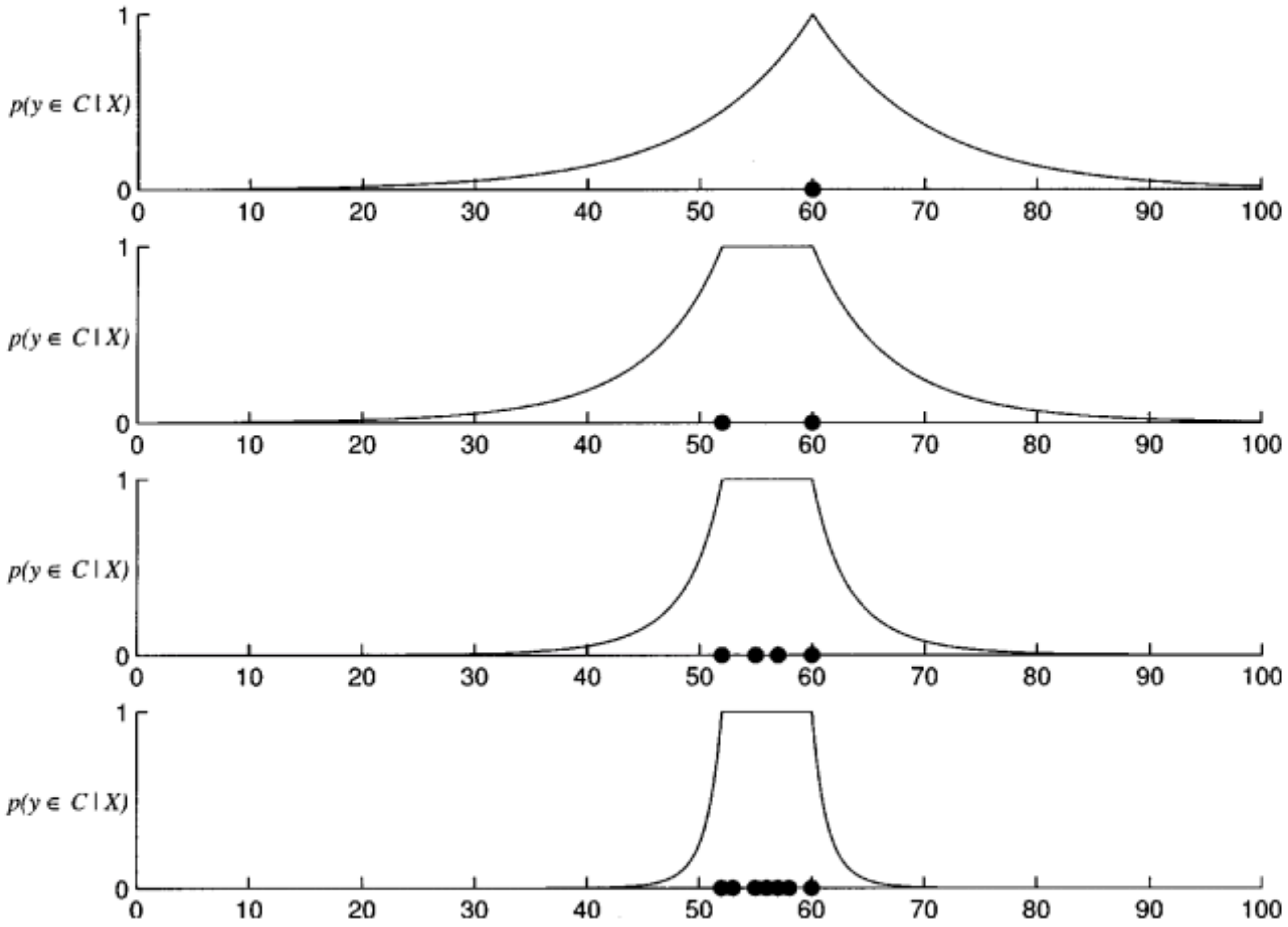
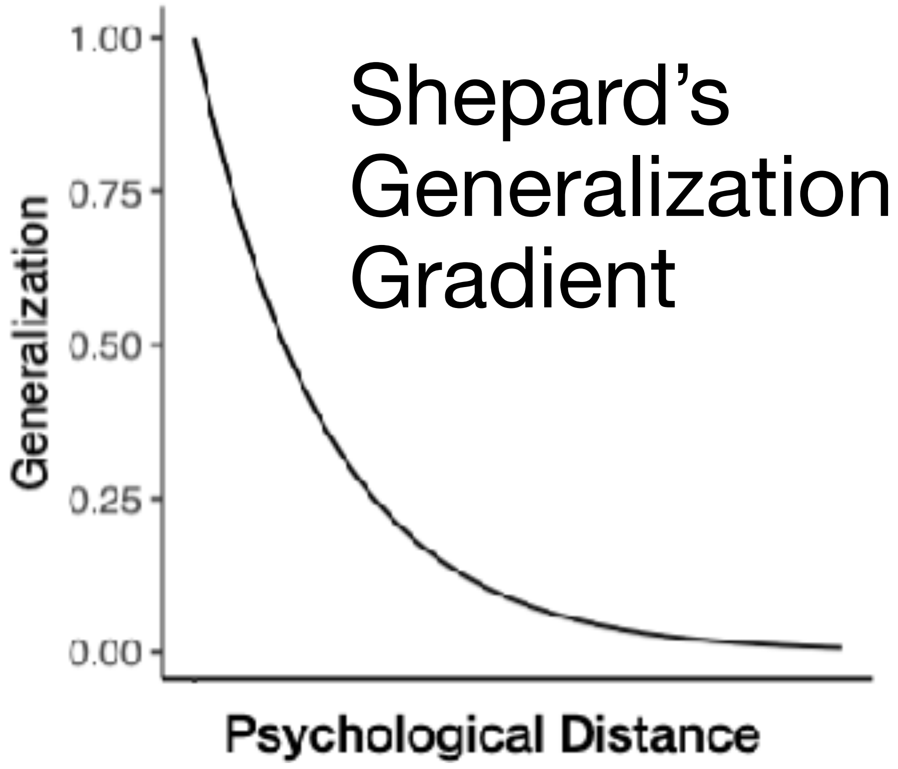


Figure 3. The effect of the number of examples on Bayesian generalization (under the assumptions of strong sampling and an Erlang prior, $\mu = 10$). Filled circles indicate examples. The first curve is the gradient of generalization with a single example, for the purpose of comparison. The remaining graphs show that the range of generalization decreases as a function of the number of examples.



Shepard's Generalization Gradient

Bayesian Concept Learning Extends Shepard's Law of Generalization to Multiple Examples

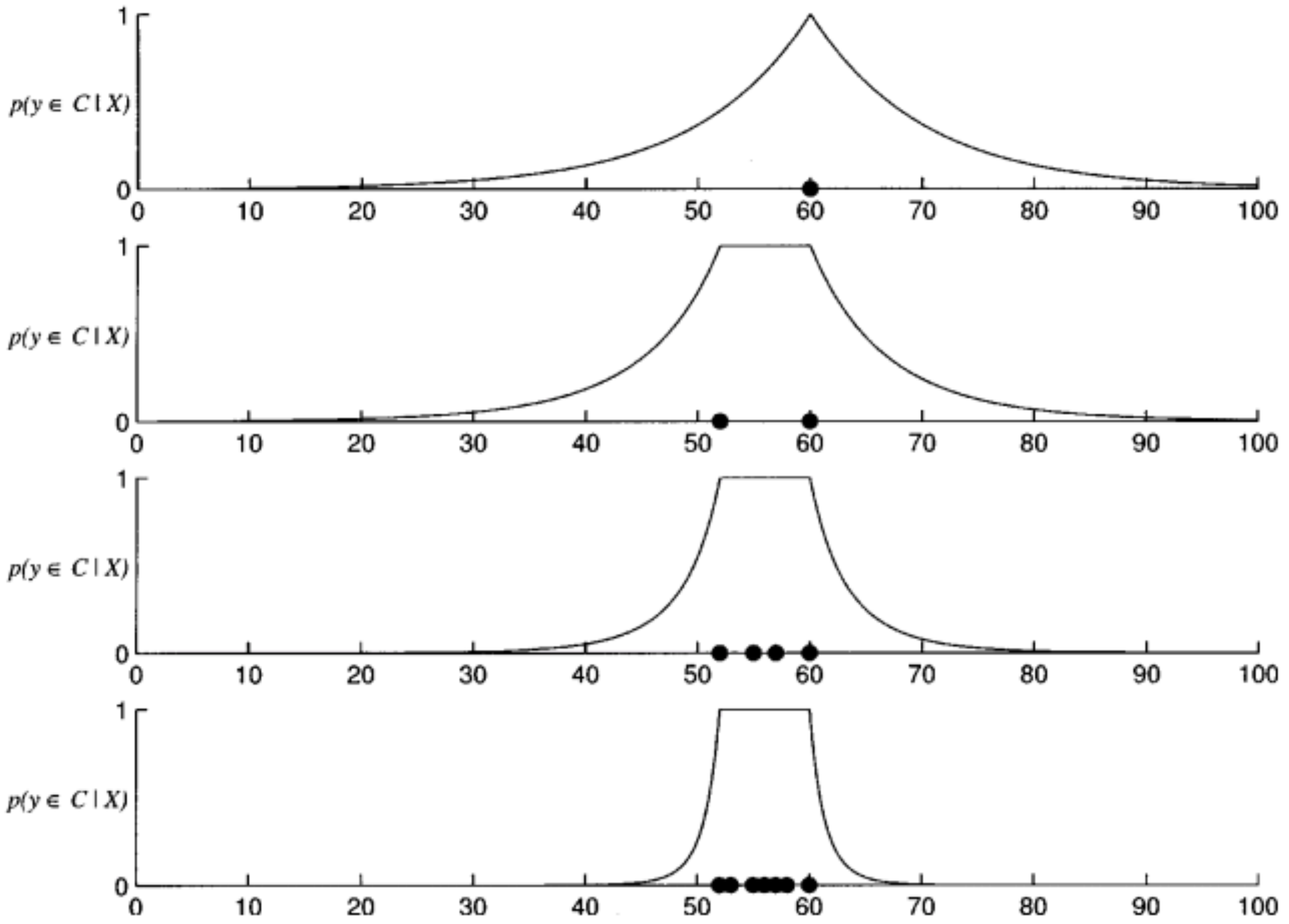
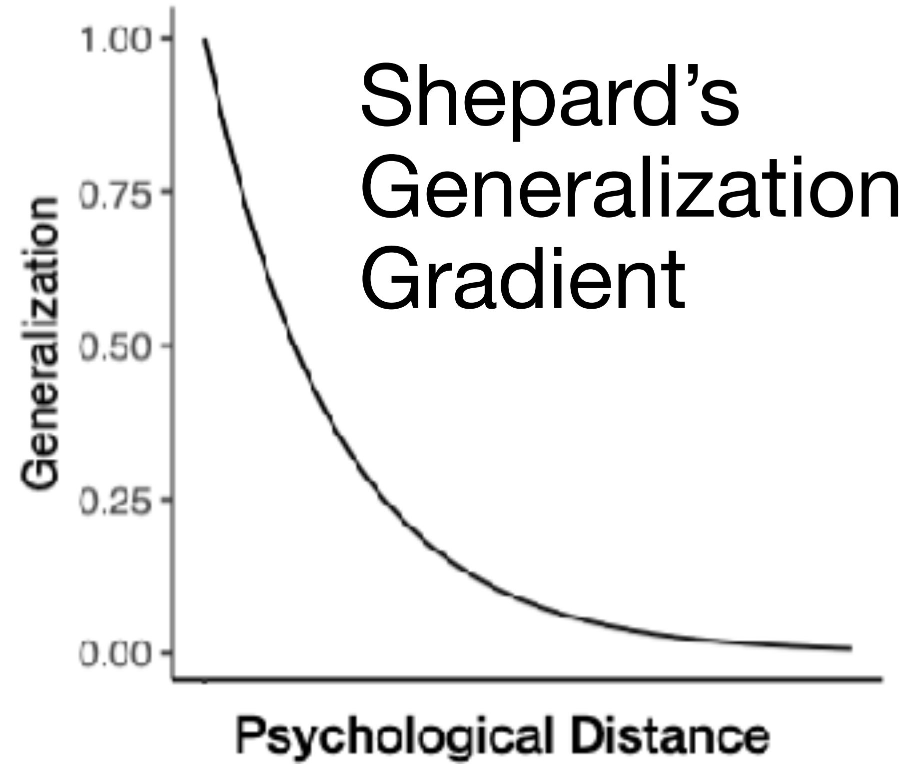


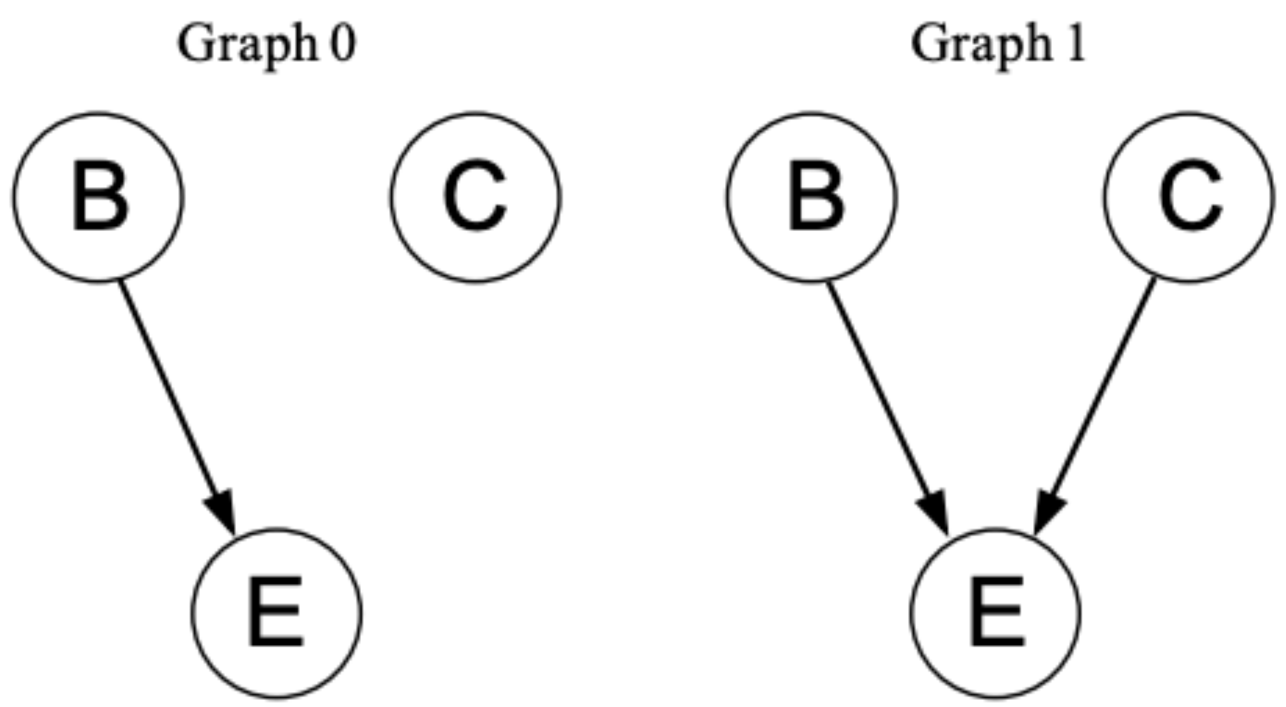
Figure 3. The effect of the number of examples on Bayesian generalization (under the assumptions of strong sampling and an Erlang prior, $\mu = 10$). Filled circles indicate examples. The first curve is the gradient of generalization with a single example, for the purpose of comparison. The remaining graphs show that the range of generalization decreases as a function of the number of examples.



Range of generalization decreases with more examples

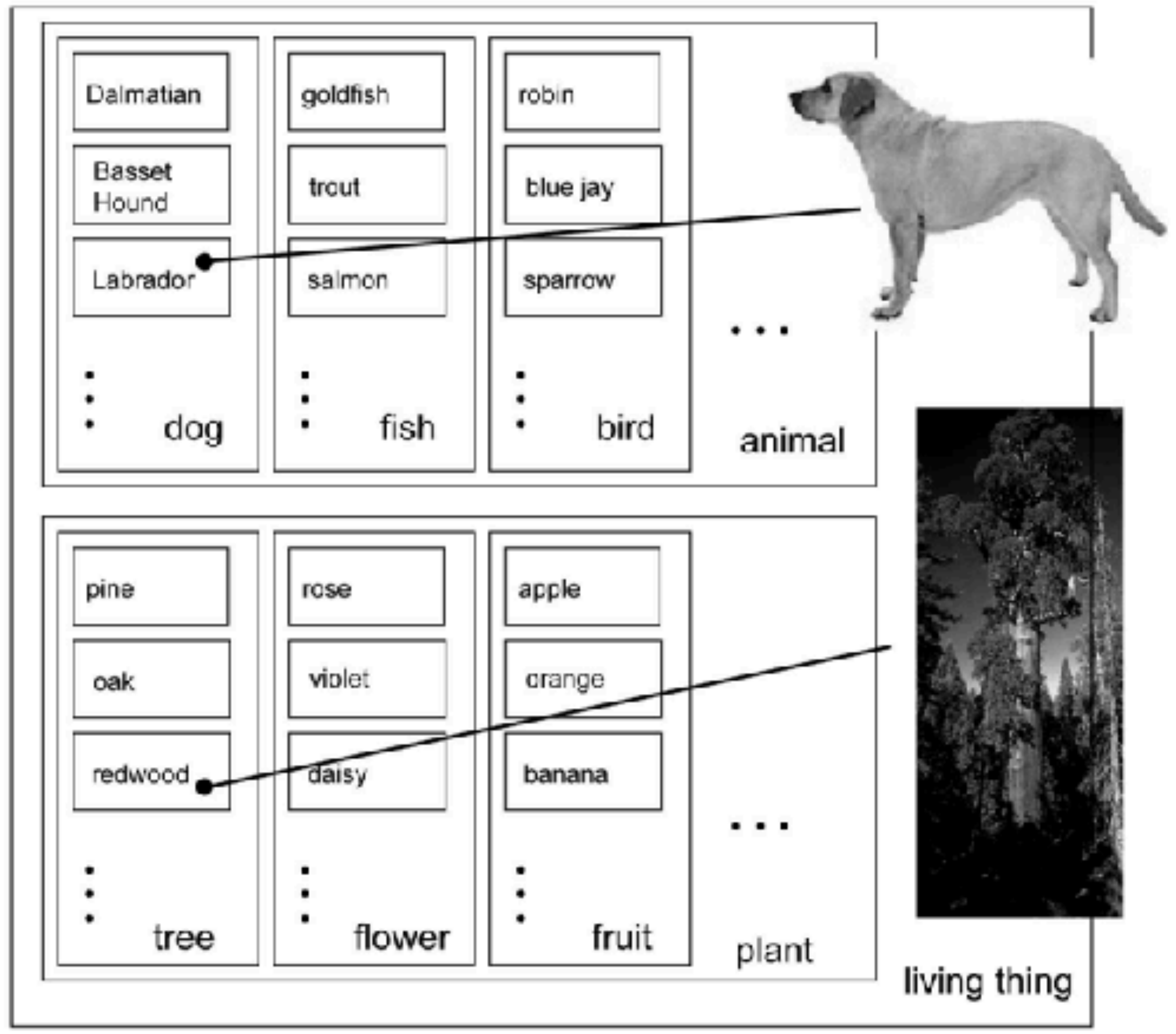
more examples = less uncertainty about the extent of consequential region

Causal learning



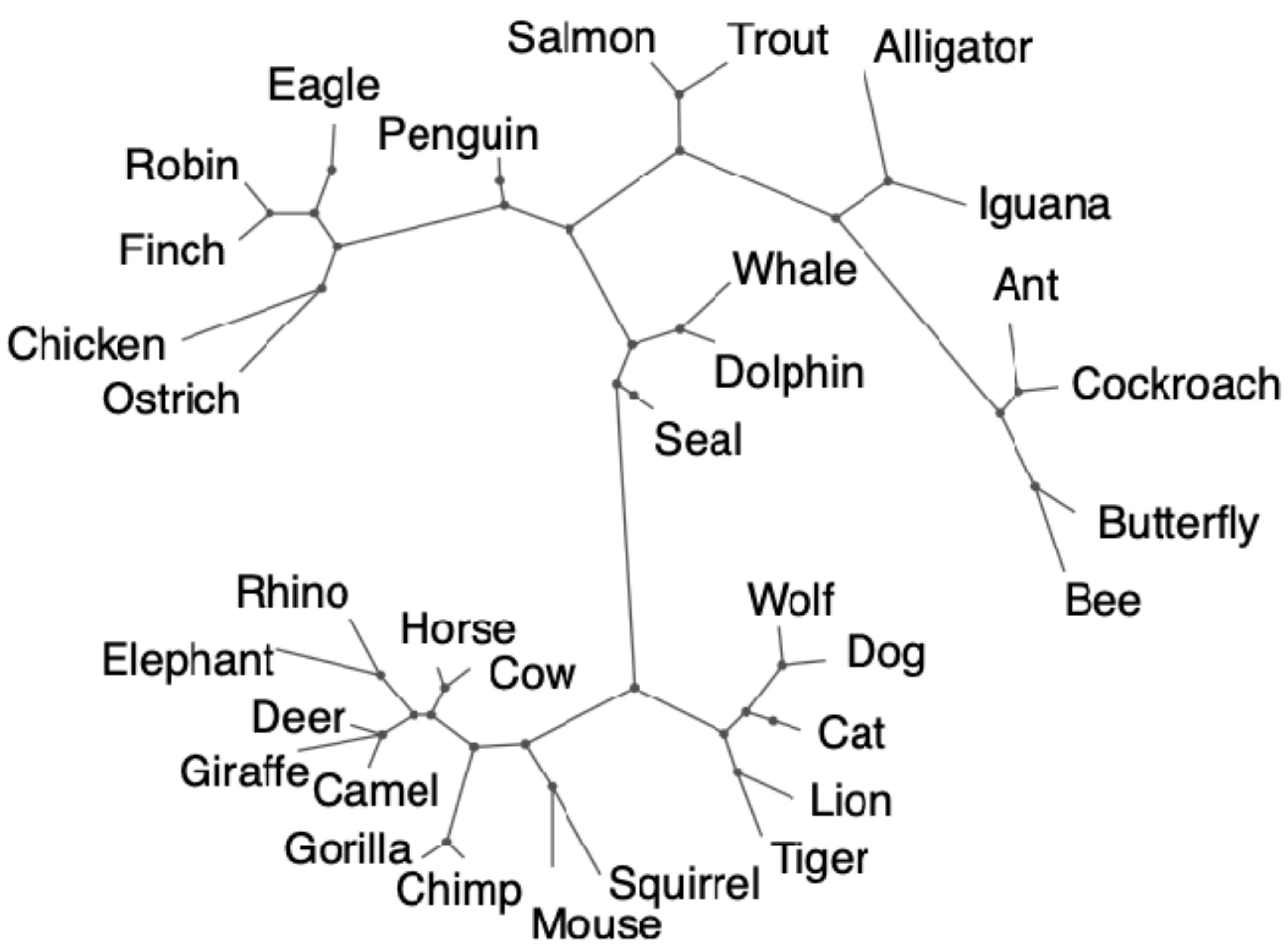
Griffiths & Tenenbaum (2005)

Word learning



Xu & Tenenbaum (2007)

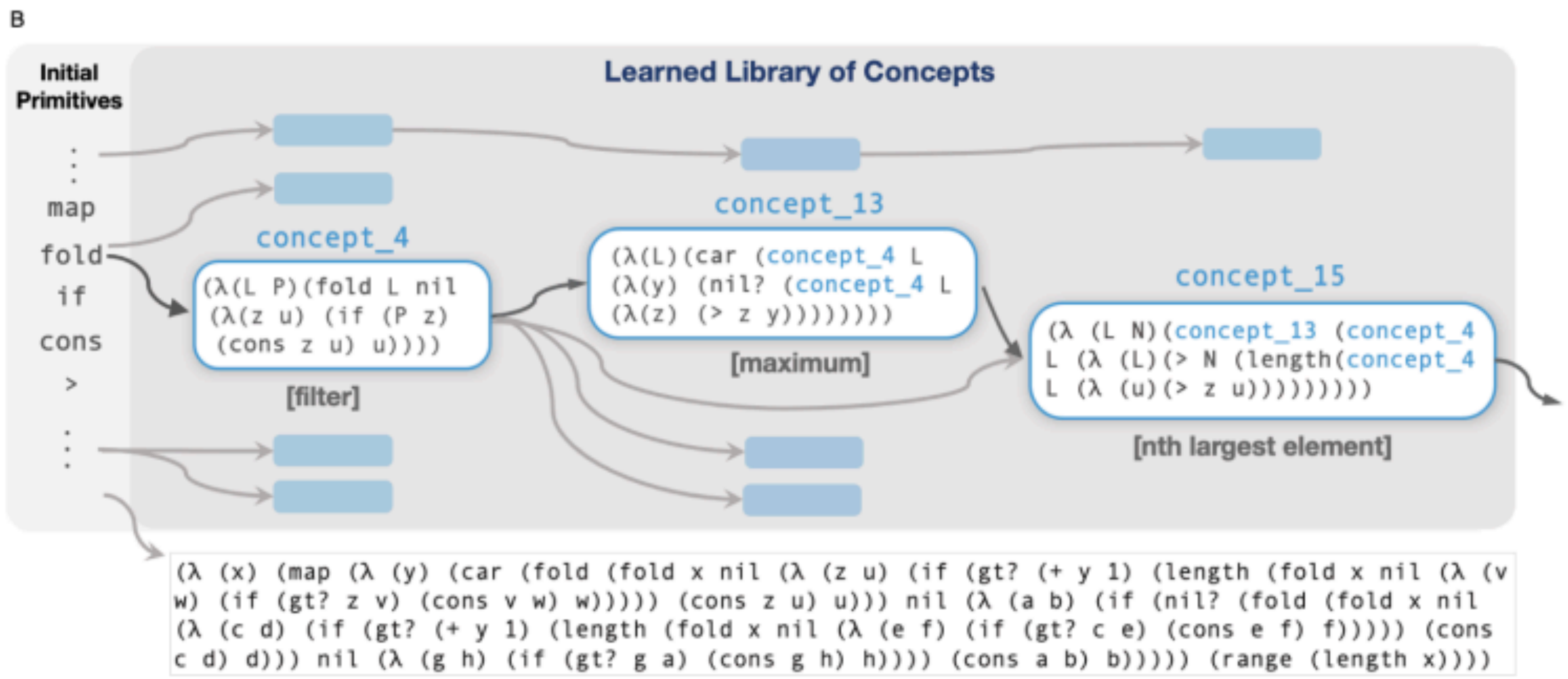
Structure learning



Kemp & Tenenbaum (2008)

Program Induction

Lake, Salakhutdinov, & Tenenbaum (2015)

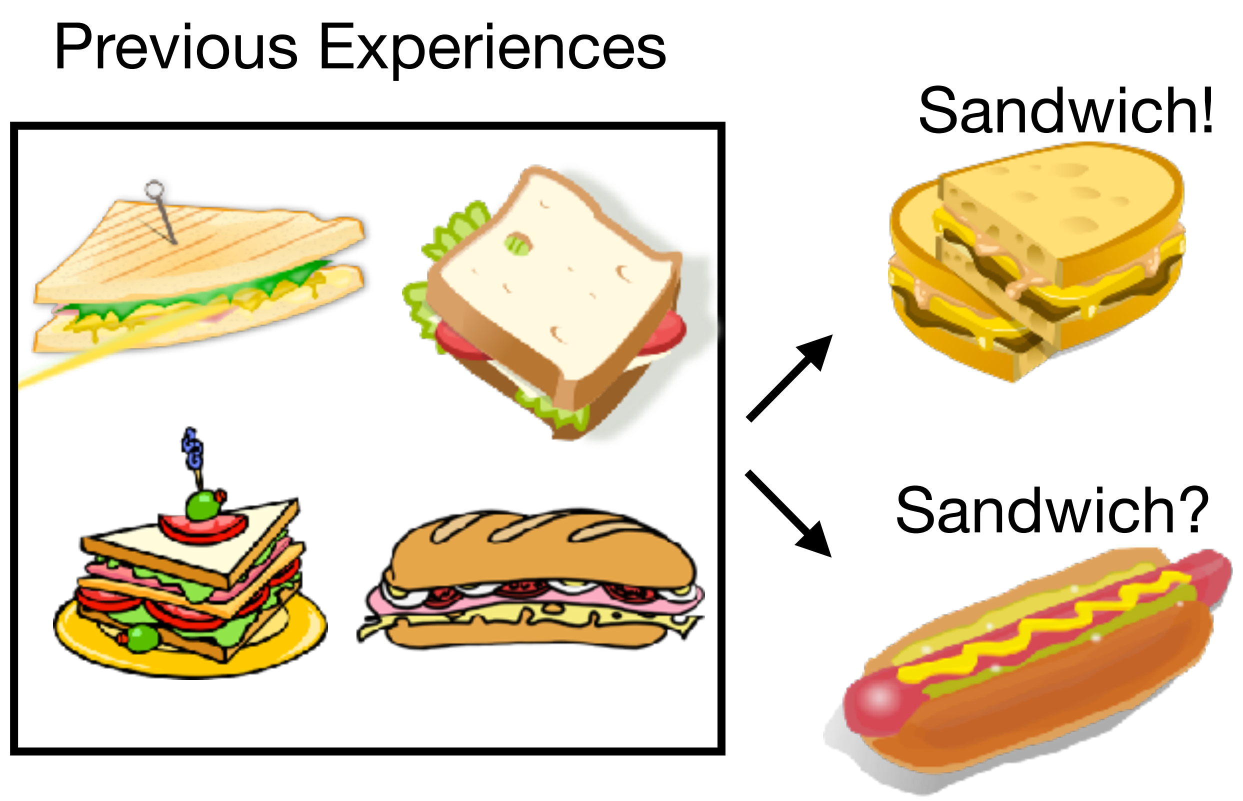


Dreamcoder: Ellis et al,. (2020)

… and many more

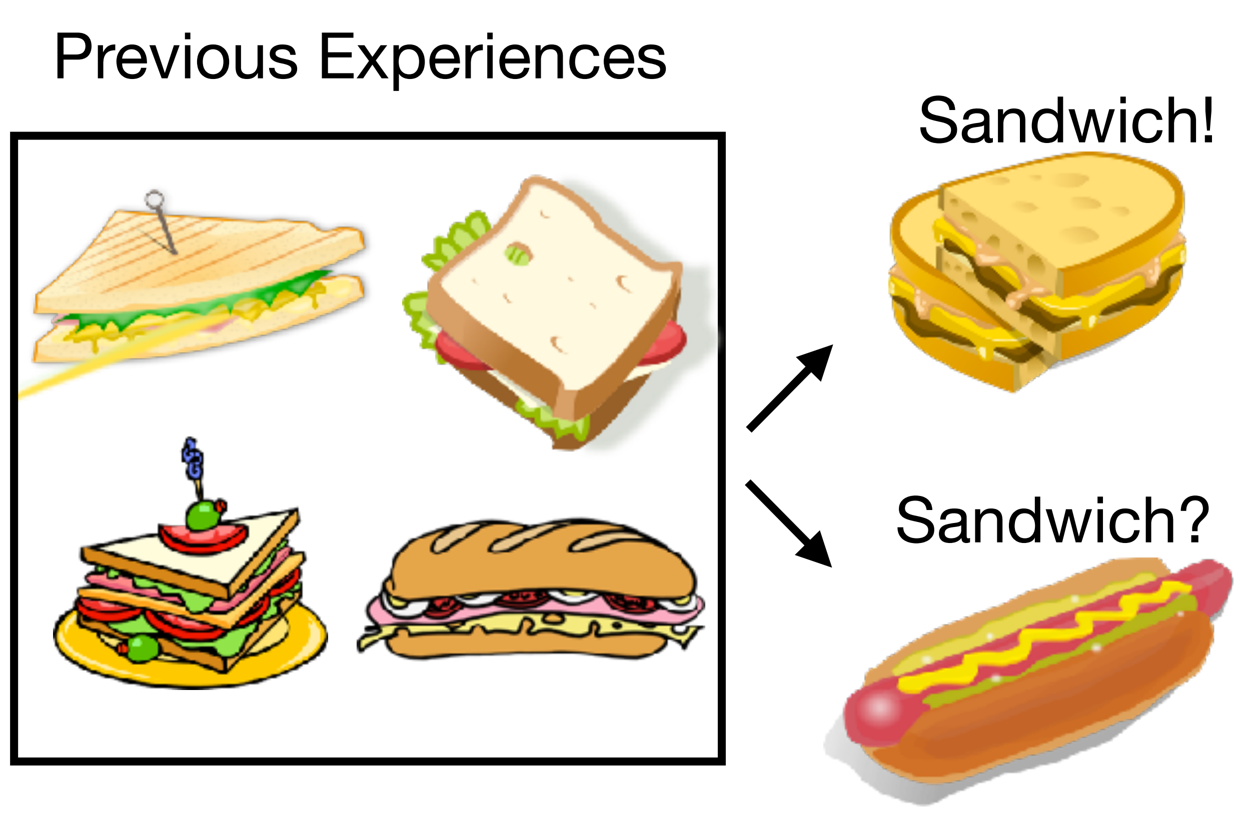
Theories of Concept Learning

Classification task

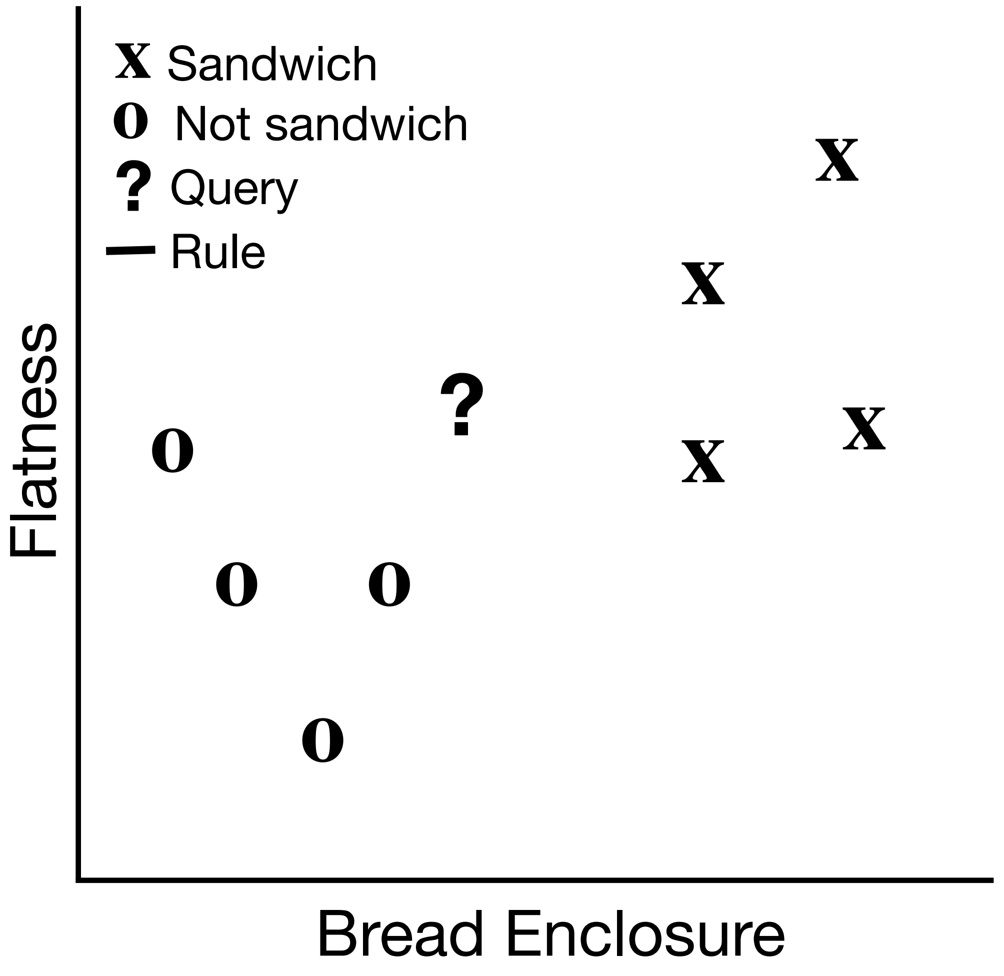


Theories of Concept Learning

Classification task

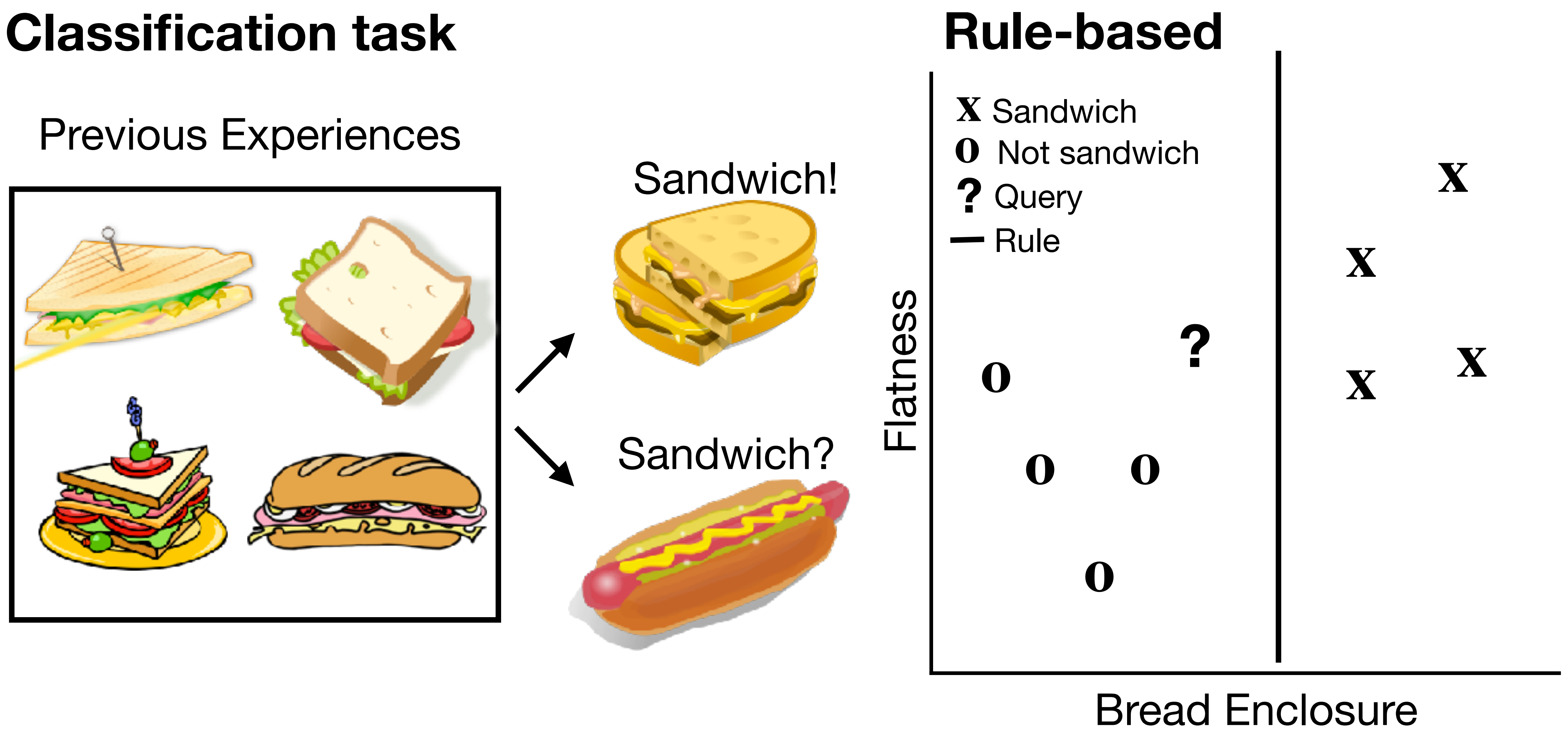


Rule-based



- *Rules* describe the explicit boundaries of category boundaries
  (Smith & Medin, 1981; Ashby & Gott, *JEP:LMC* 1988)

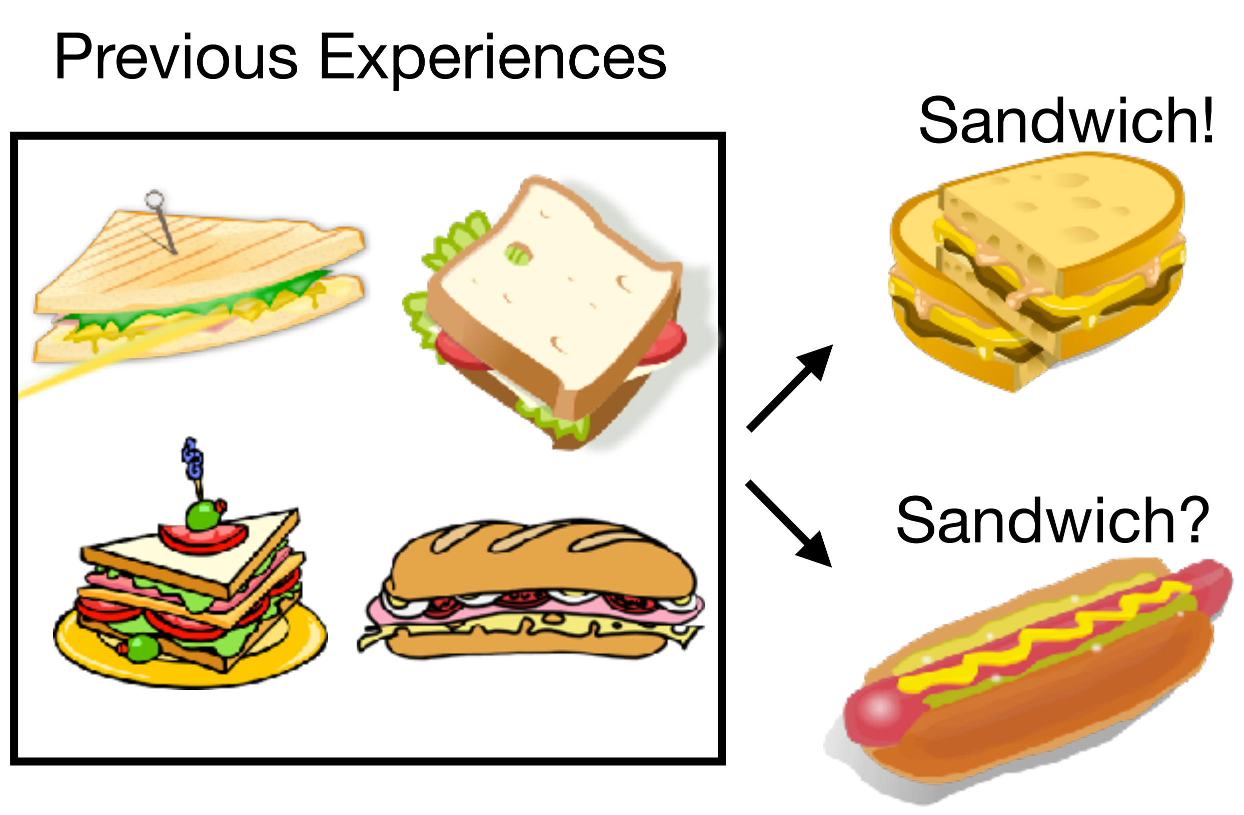
Theories of Concept Learning



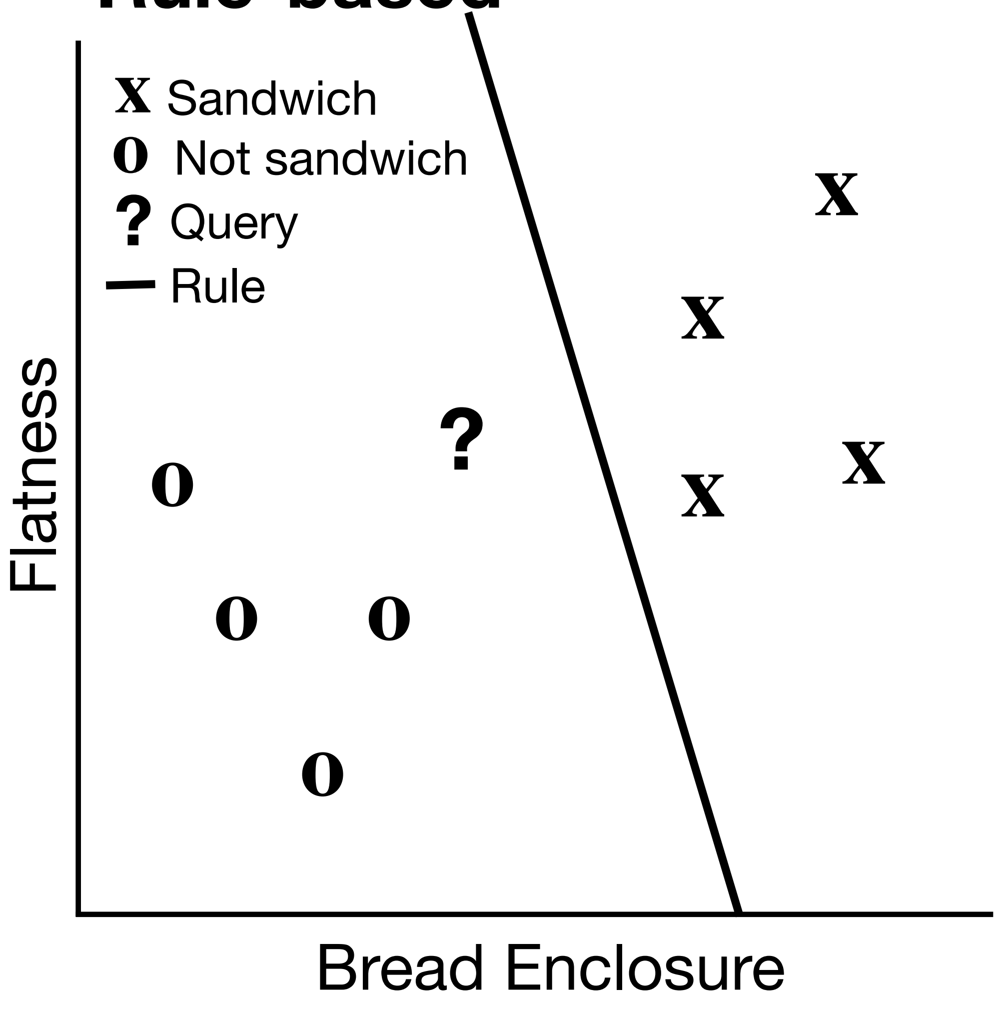
- *Rules* describe the explicit boundaries of category boundaries
  (Smith & Medin, 1981; Ashby & Gott, *JEP:LMC* 1988)

Theories of Concept Learning

Classification task

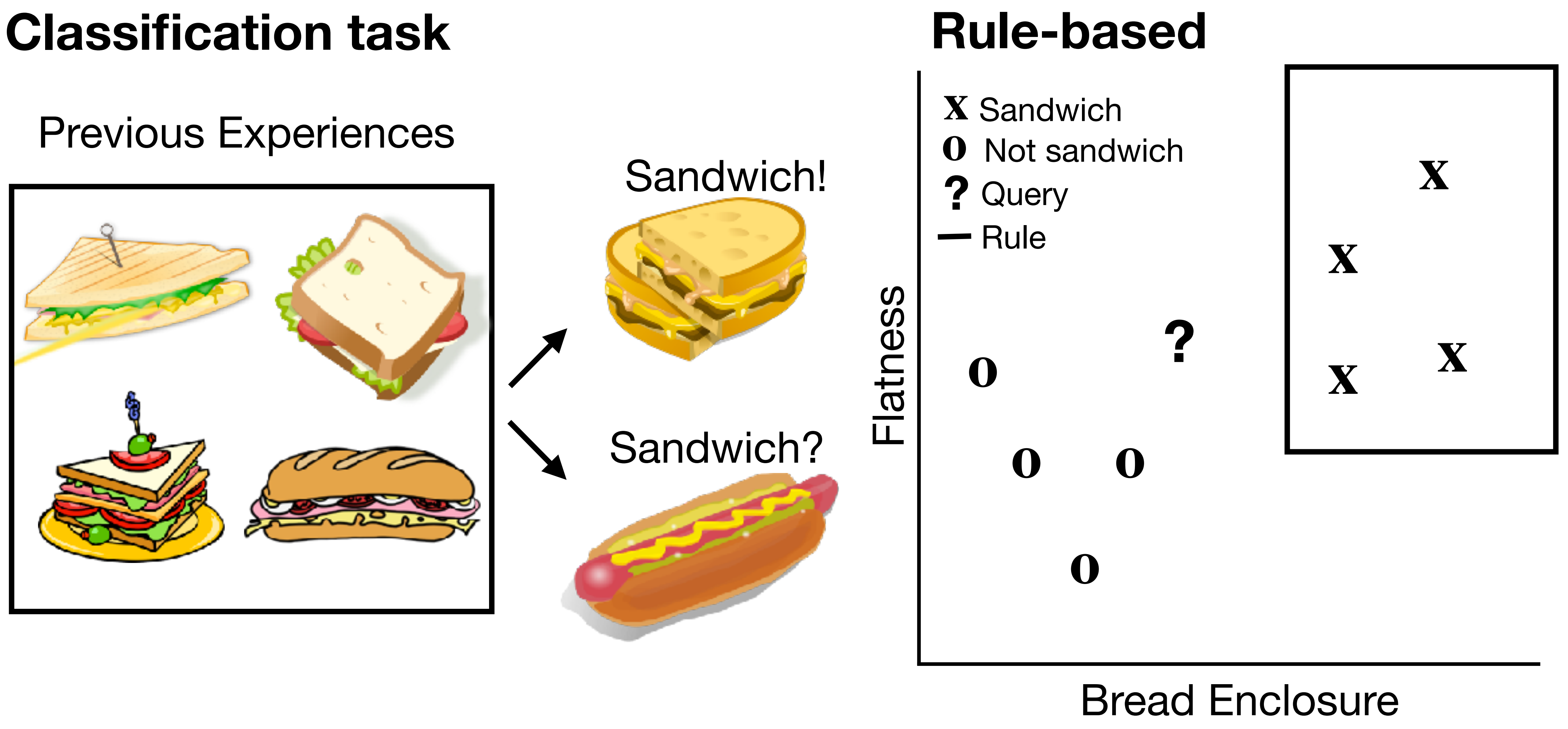


Rule-based



- *Rules* describe the explicit boundaries of category boundaries
  (Smith & Medin, 1981; Ashby & Gott, *JEP:LMC* 1988)

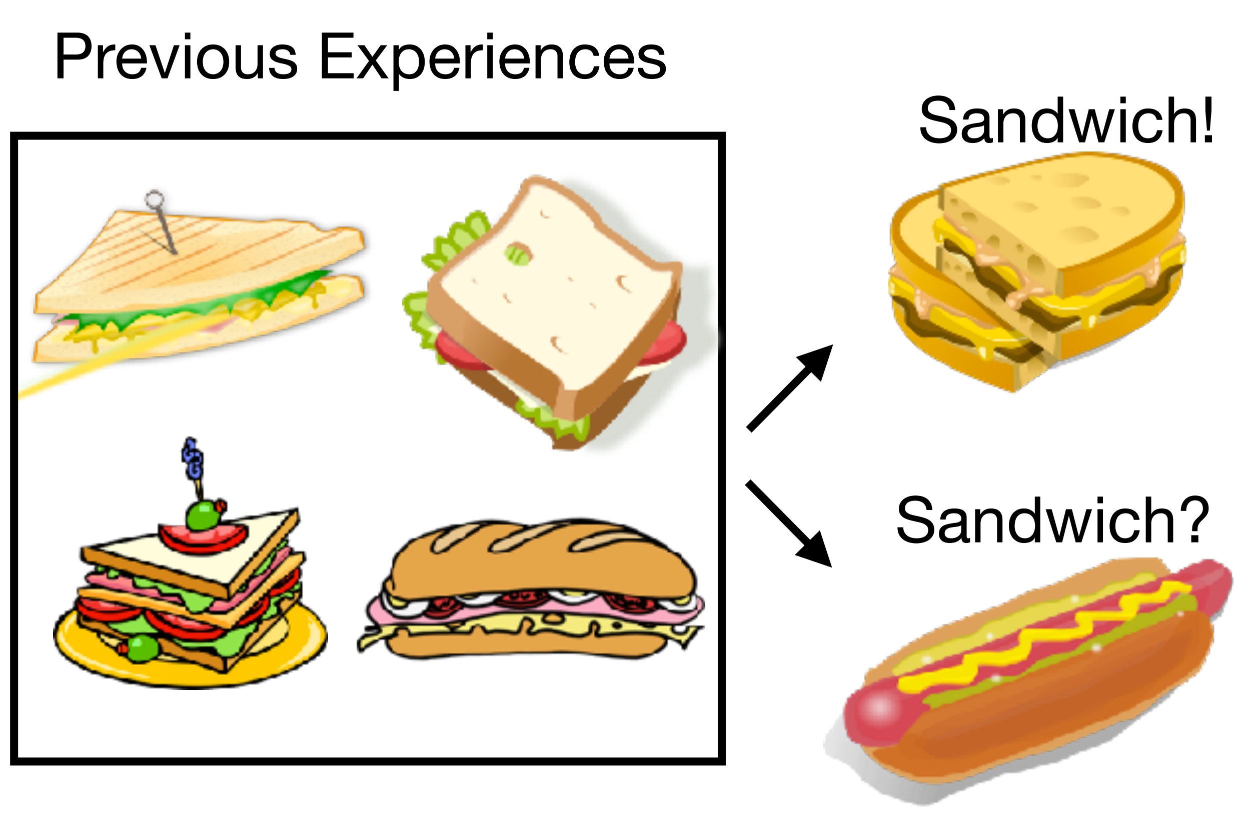
Theories of Concept Learning



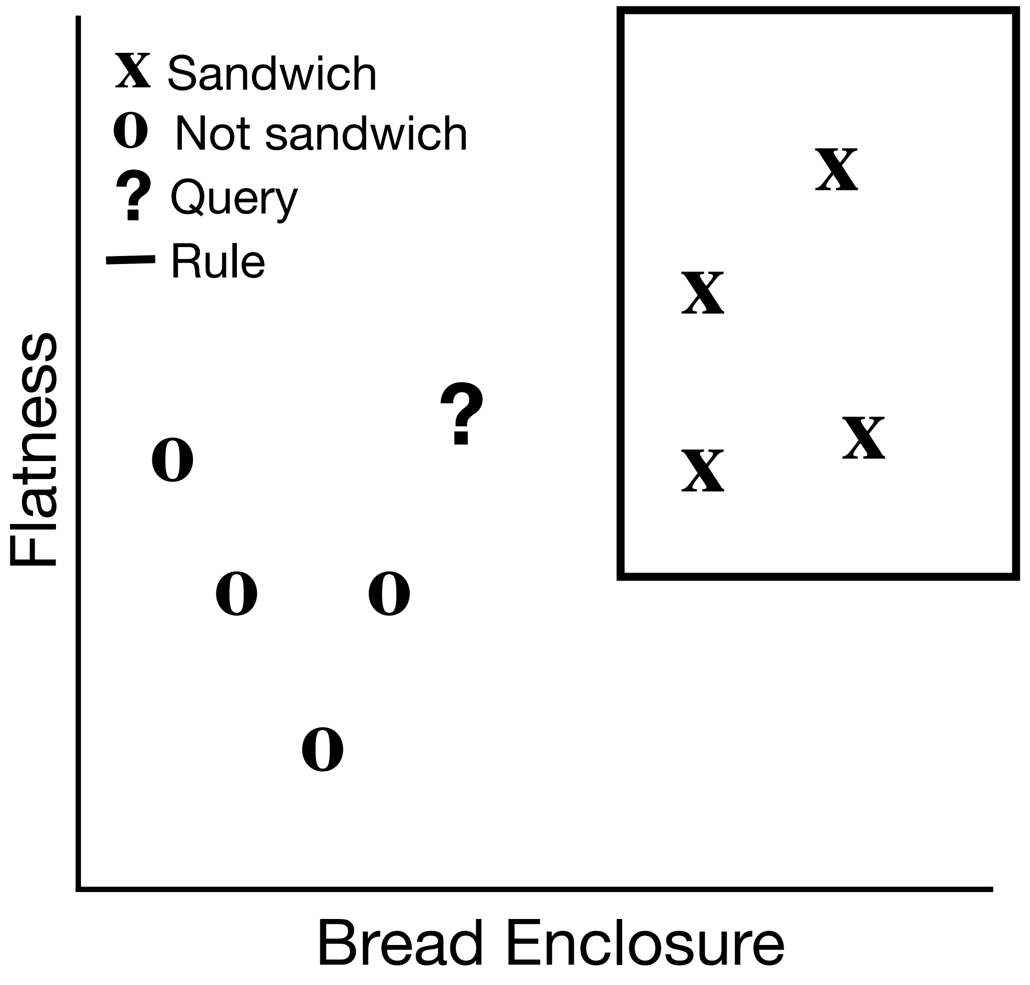
- *Rules* describe the explicit boundaries of category boundaries
(Smith & Medin, 1981; Ashby & Gott, *JEP:LMC* 1988)

Theories of Concept Learning

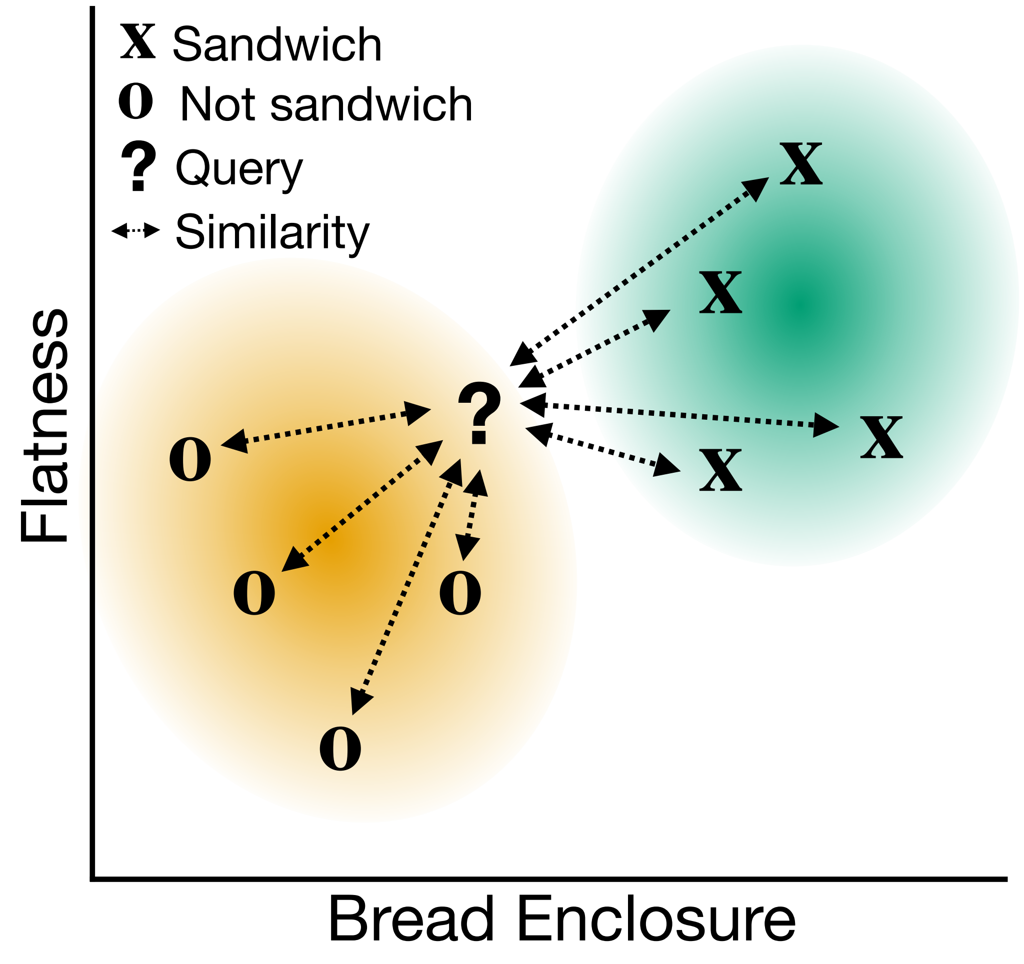
Classification task



Rule-based



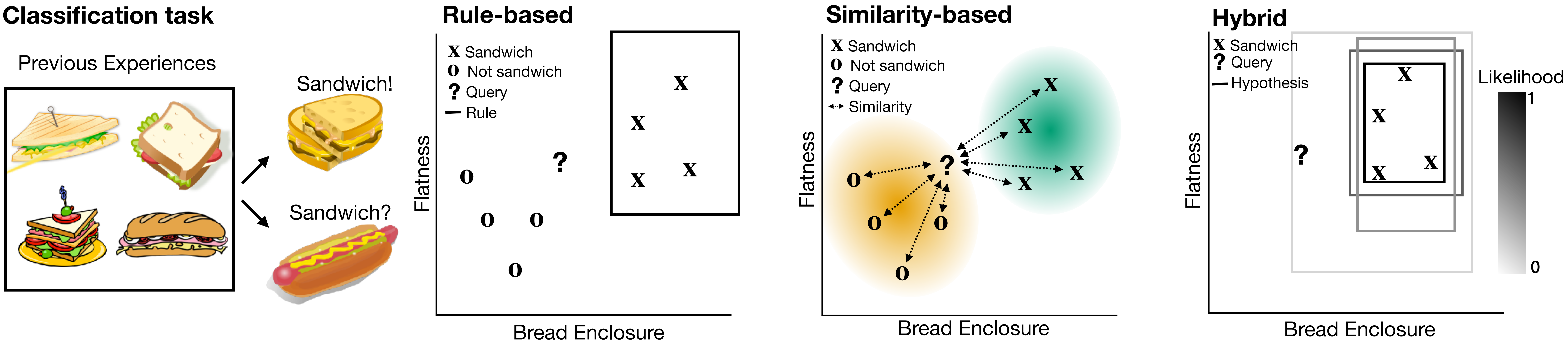
Similarity-based



- *Rules* describe the explicit boundaries of category boundaries
  (Smith & Medin, 1981; Ashby & Gott, *JEP:LMC* 1988)

- *Similarity* uses a comparison to previously encountered exemplars or a learned prototype (aggregated over multiple experiences) as the basis of generalization
  (Rosch, *CogPsy* 1973; Medin & Schaffer, *PsychRev* 1978 ; Nosofsky, *JEP:G* 1986; Smith & Minda *JEP:LMC* 1998)

Theories of Concept Learning



- *Rules* describe the explicit boundaries of category boundaries
(Smith & Medin, 1981; Ashby & Gott, *JEP:LMC* 1988)

- *Similarity* uses a comparison to previously encountered exemplars or a learned prototype (aggregated over multiple experiences) as the basis of generalization
(Rosch, *CogPsy* 1973; Medin & Schaffer, *PsychRev* 1978 ; Nosofsky, *JEP:G* 1986; Smith & Minda *JEP:LMC* 1998)

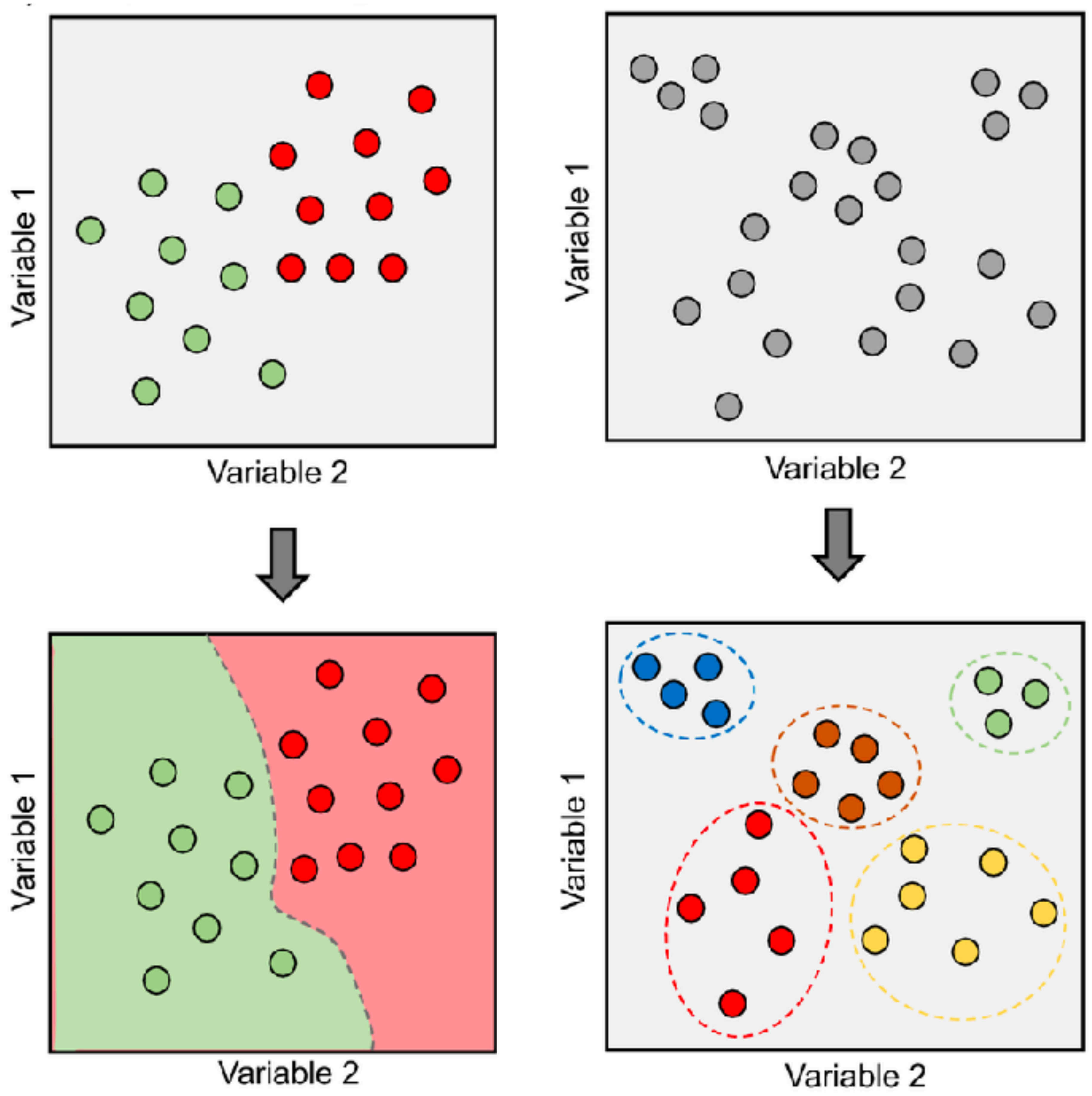
- *Hybrids* combine elements of both: Bayesian concept learning uses a distribution over rules, while reproducing predictions of two influential similarity-based approaches
(Tenenbaum & Griffiths, *BBS* 2001; Shepard, *Science* 1987; Tversky, *PsychRev* 1977)

General principles

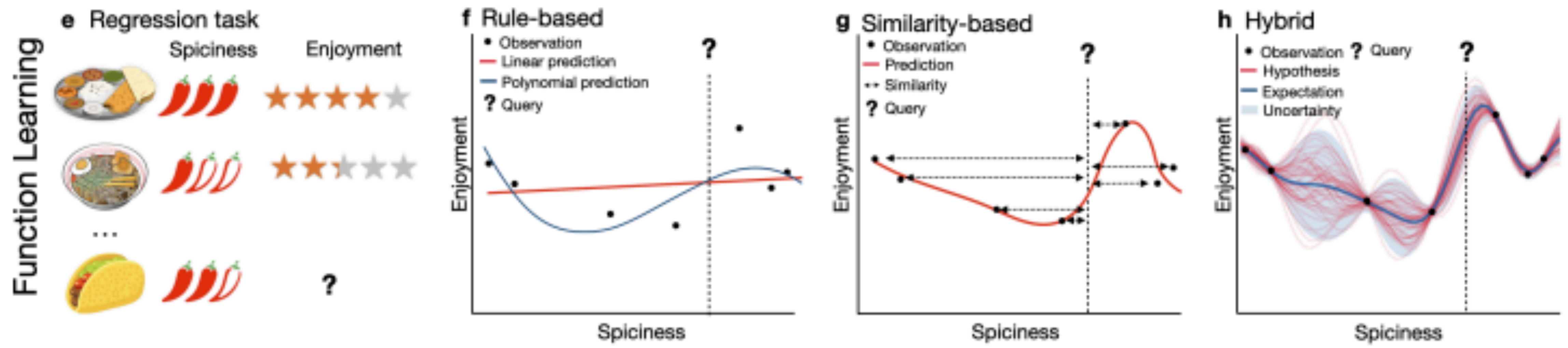
- Again, hybrid theories combining competiting mechanisms seem to provide the best answer
  - Rules have a symbolic flavor, offering rapid generalization and flexible composition
  - Similarity has a subsymbolic flavor, where previously encountered example exert influence on generalization based on similarity-weights
  - A hybrid using Bayesian inference combines the best of both worlds
- Concepts are not just passively learned associations (model-free RL), but seem to point towards generative representations about the structure of the world (Model-based RL)

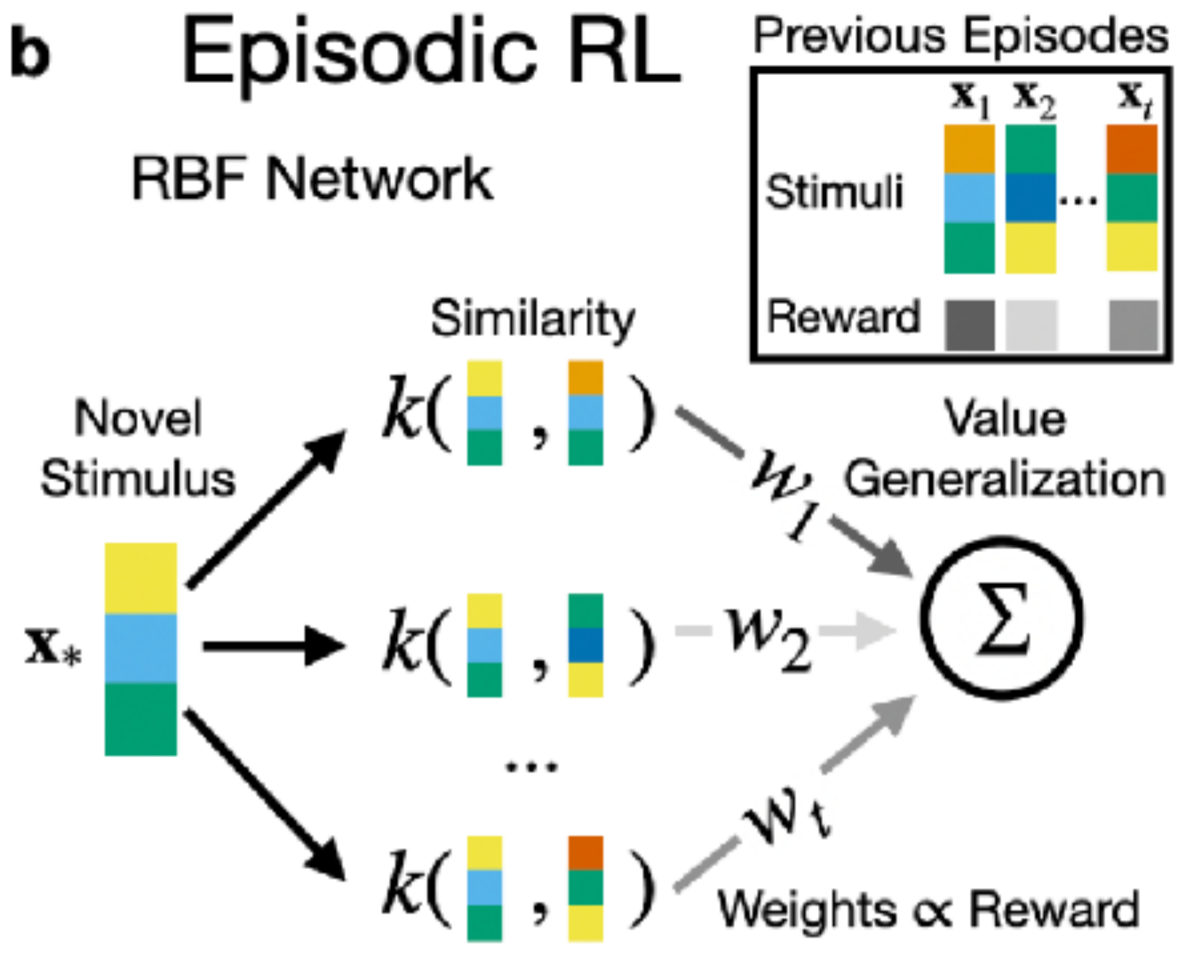
Next weeks

Supervised and unsupervised learning



*Function learning





*Note the change in topic and assigned reading. This is an in prep manuscript and I will send it via email