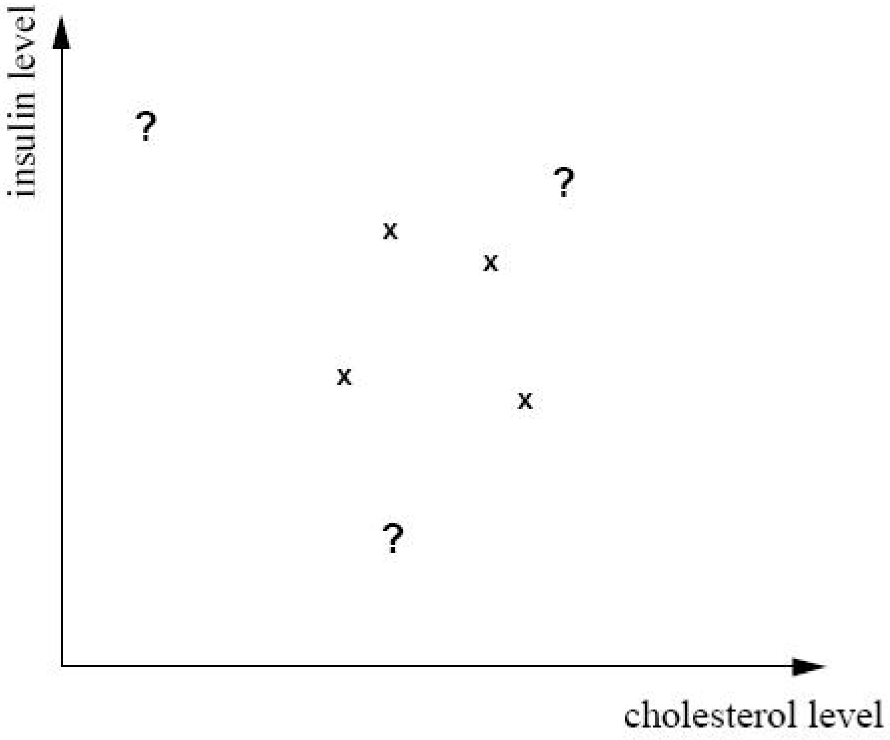
CS340

Bayesian concept learning cont'd

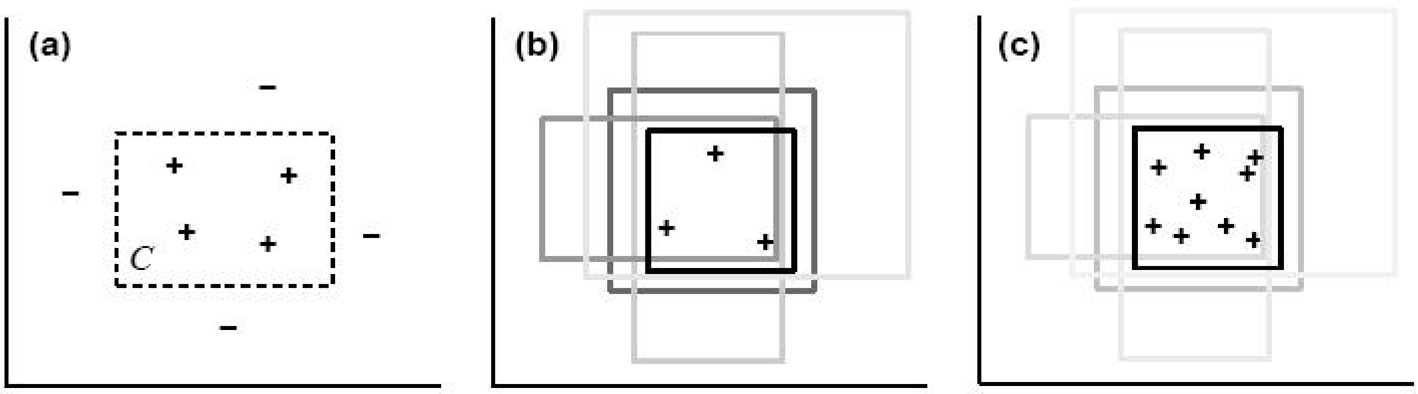
Kevin Murphy

Healthy levels game



"healthy levels"

Hypothesis space



$$h = (\ell_1, \ell_2, s_1, s_2)$$

Healthy levels of insulin/ cholestrol must lie between a minimum and maximum. Healthy levels of a chemical presumably lie between zero and a maximum.

Likelihood (strong sampling)

- $p(X|h) = 1/|h|^n$ if all $x_i \in h$,
  where $|h| = s_1 \times s_2$
- $p(X|h) = 0$ if any x_i outside h

Prior p(h)

- Use uninformative, but location and scale-invariant, prior (Jeffrey's principle)

$$p(h) \propto \frac{1}{s_1 s_2}$$

  This also happens to be conjugate to p(X|h).
- We will explain this later...

Posterior predictive

$$p(y \in C|X) \;\; = \;\; \int_{h \in H} p(y \in C|h) p(h|X) dh$$

Since the hypothesis space is continuous, we must use an integral instead of a sum...

Insert hairy math

$l - s \leq -r$, where s is size of the rectangle. Hence

$$
\begin{aligned}
p(X) &= \int_{h \in \mathcal{H}_X} \frac{p(h)}{|h|^n} dh && (1.34)\\
&= \int_{s=r}^{\infty} \int_{l=0}^{s-r} \frac{p(s)}{s^n} dl ds && (1.35)\\
&= \int_{s=r}^{\infty} \left[\int_{l=0}^{s-r} \frac{1}{s^{n+1}} dl \right] ds && (1.36)\\
&= \int_{s=r}^{\infty} \frac{1}{s^{n+1}} [l]_0^{s-r} ds && (1.37)\\
&= \int_{s=r}^{\infty} \frac{s-r}{s^{n+1}} ds && (1.38)
\end{aligned}
$$

Now, using **integration by parts**

$$
I = \int_a^b f(x) g'(x) dx = [f(x)g(x)]_a^b - \int_a^b f'(x) g(x) dx \qquad (1.39)
$$

with the substitutions

$$
\begin{aligned}
f(s) &= s - r && (1.40)\\
f'(s) &= 1 && (1.41)\\
f'(s) &= s^{-n-1} && (1.42)\\
g(s) &= \frac{s^{-n}}{-n} && (1.43)
\end{aligned}
$$

we have

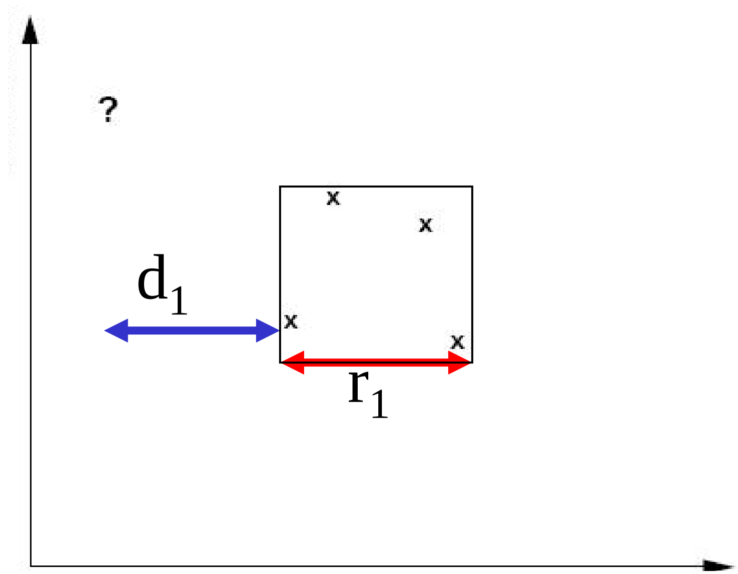
$$
\begin{aligned}
p(X) &= \left[\frac{(s-r)s^{-n}}{-n} \right]_r^{\infty} - \int_r^{\infty} \frac{s^{-n}}{-n} ds && (1.44)\\
&= \left[\frac{s^{-n+1}}{-n} + \frac{rs^{-n}}{n} - \frac{-1}{n} \frac{s^{-n+1}}{-n+1} \right]_r^{\infty} && (1.45)\\
&= \frac{r^{-n+1}}{n} - \frac{rr^{-n}}{n} + \frac{r^{-n+1}}{n(n-1)} && (1.46)\\
&= \frac{1}{nr^{n-1}} - \frac{r}{nr^{n-1}r} + \frac{1}{n(n-1)r^{n-1}} && (1.47)\\
&= \frac{1}{n(n-1)r^{n-1}} && (1.48)
\end{aligned}
$$

To compute the generalization function, let us suppose y is outside the range spanned by the examples (otherwise the probability of generalization is 1). Without loss of generality assume $y > 0$. Let d be the distance from y to the closest observed example. Then we can compute the numerator in Equation 1.33 by replacing r with $r + d$ in the limits of integration (since we have expanded the range of the data by adding y), yielding

$$
\begin{aligned}
p(y \in C, X) &= \int_{h \in \mathcal{H}_{X,y}} \frac{p(h)}{|h|^n} dh && (1.49)\\
&= \int_{r+d}^{\infty} \int_0^{s-(r+d)} \frac{p(s)}{s^n} dl ds && (1.50)\\
&= \frac{1}{n(n-1)(r+d)^{n-1}} && (1.51)
\end{aligned}
$$

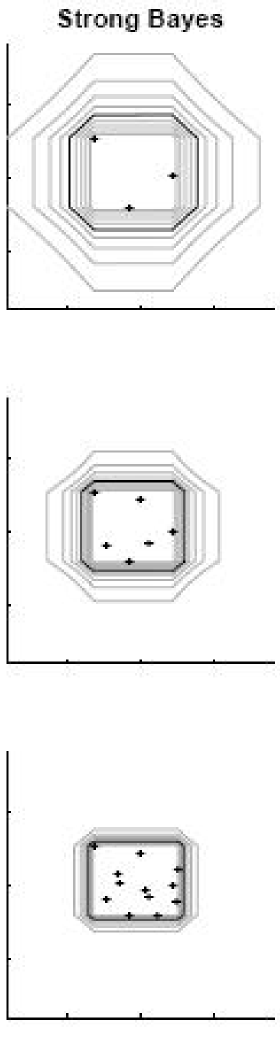
And the answer is…

$$p(y \in C|X) = \left[\frac{1}{(1+\tilde{d}_1/r_1)(1+\tilde{d}_2/r_2)}\right]^{n-1}$$



$$\begin{aligned}\tilde{d}_i &= 0 \text{ if } y \in \text{ range of } X_i \\ &= \text{ distance of y from closest } X_i\end{aligned}$$

Behavior for n=3, 6, 12

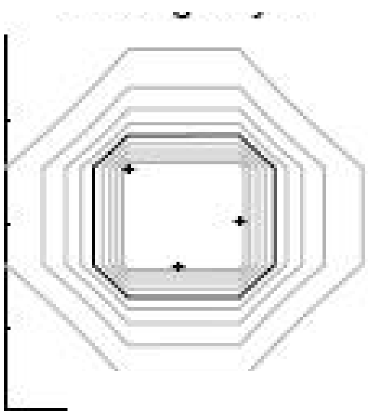


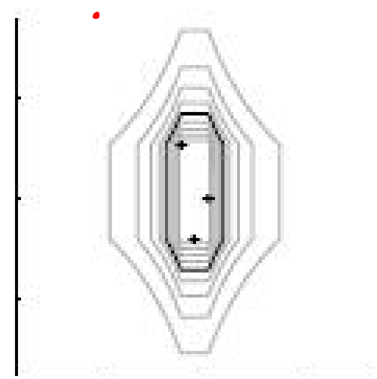
The size principle implies the smallest rectangle has highest likelihood, but there are many other consistent rectangles which are only slightly less likely. These get averaged to give a smooth generalization gradient.

As $N \to \infty$, the larger hypotheses become exponentially less likely, so we converge on the ML solution (the most specific/ MIN hypothesis)

Behavior for different shapes

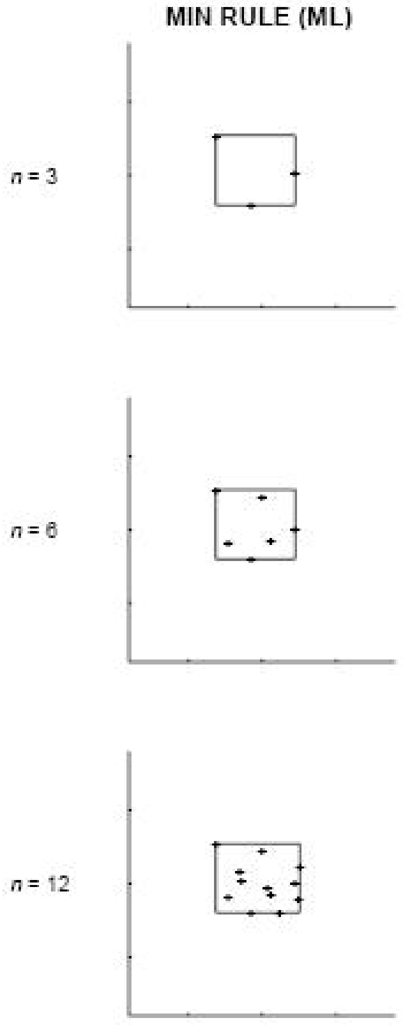
- n=3 in both cases, but on right, $r_1 << r_2$,
  so we generalize more along dimension 2
- Algebraically, d_1/r_1 is big, so $p(y \in C \mid X)$ is small unless y is inside X
- Intuitively, it would be a suspicious coincidence if the rectangle was wide but r_1





$$p(y \in C|X) = \left[\frac{1}{(1+\tilde{d}_1/r_1)(1+\tilde{d}_2/r_2)}\right]^{n-1}$$

Behavior of max likelihood/ MAP

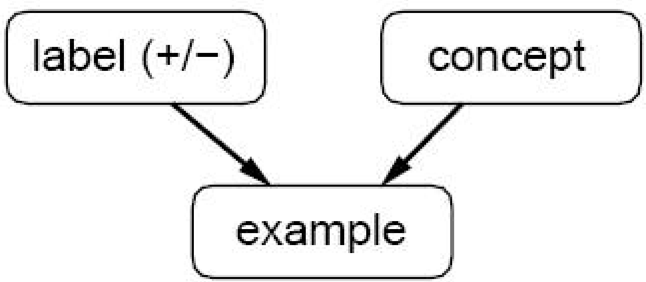


There is no generalization gradient (a point is either in or out of h).
The ML/MAP hyp. is the smallest enclosing rectangle.
This is a good approximation to Bayes when N is large.

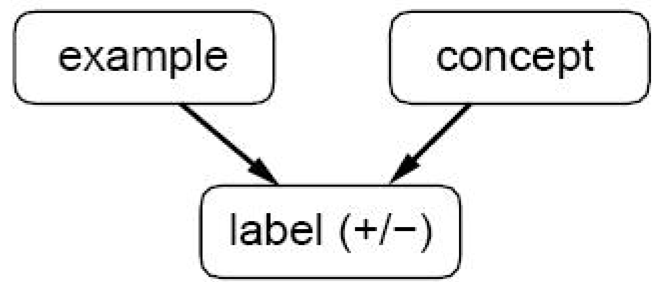
Weak sampling

- Examples are not sampled from the concept, they are just labeled as consistent or not.

$$p(X|h) = \begin{cases} 1 & \text{if } x_1, \ldots, x_n \in h \\ 0 & \text{if any } x_i \notin h \end{cases}$$

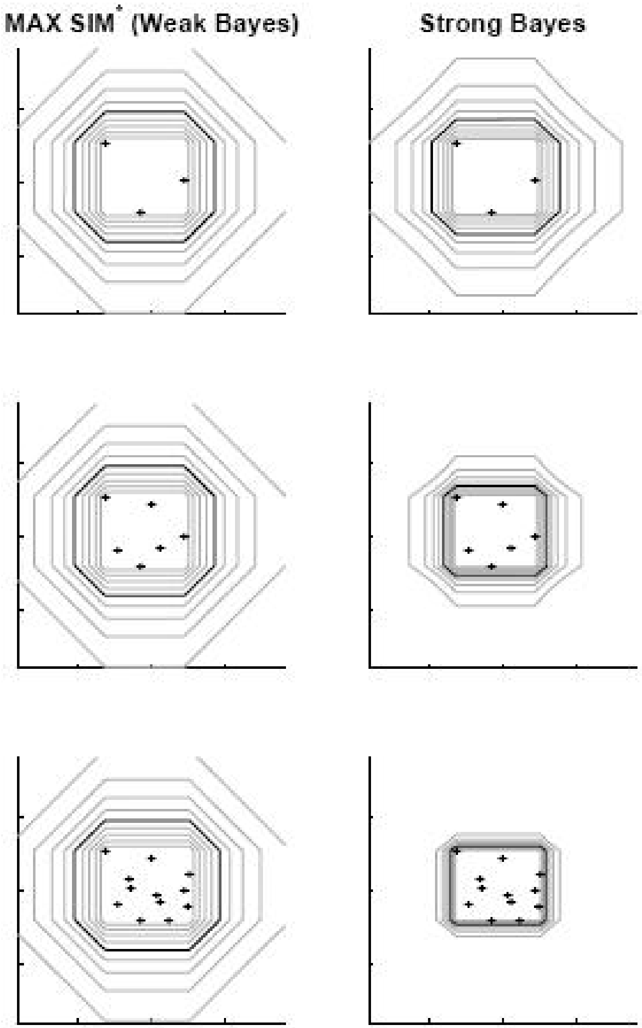


Strong sampling



Weak sampling

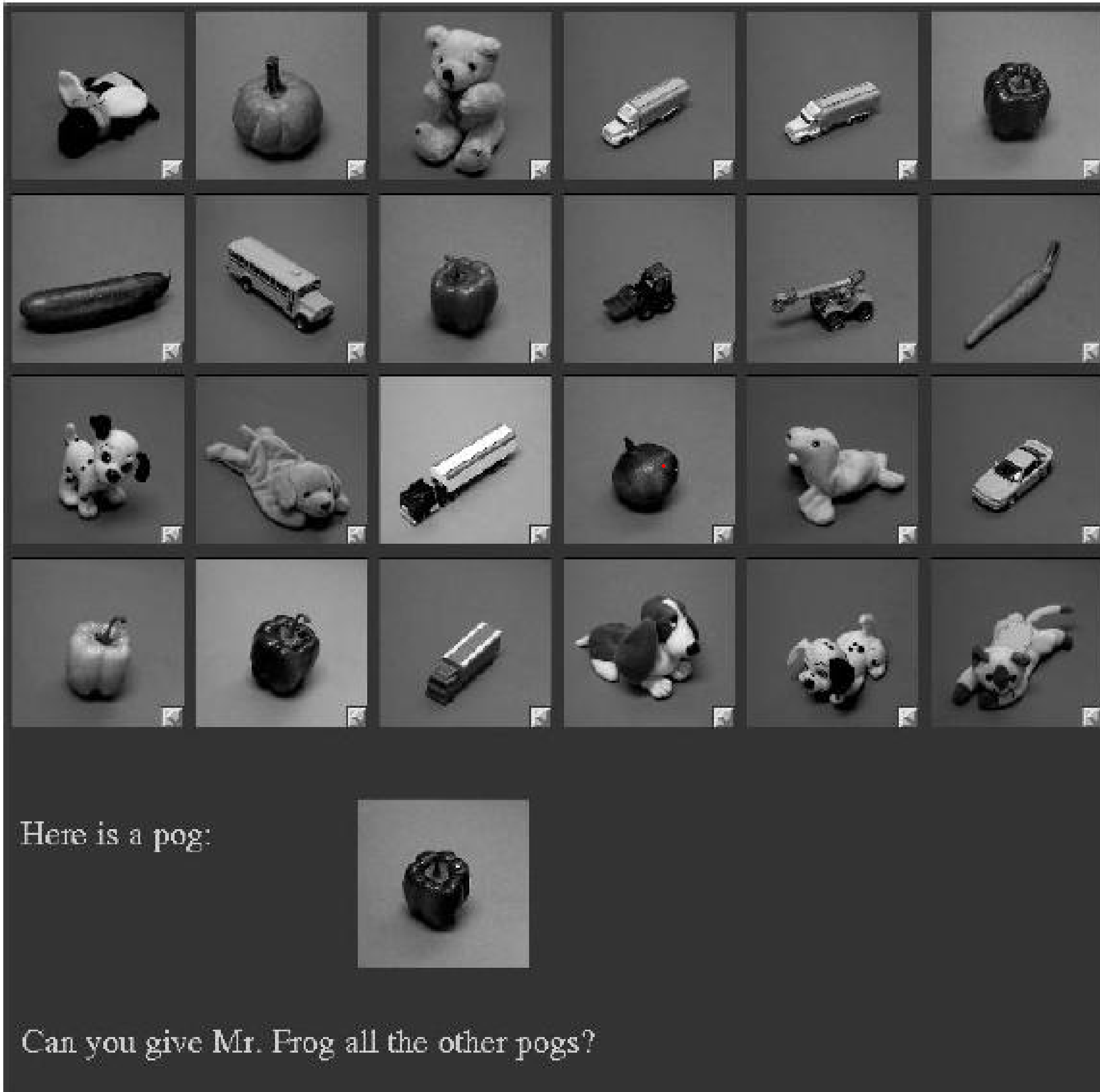
Behavior of weak Bayes



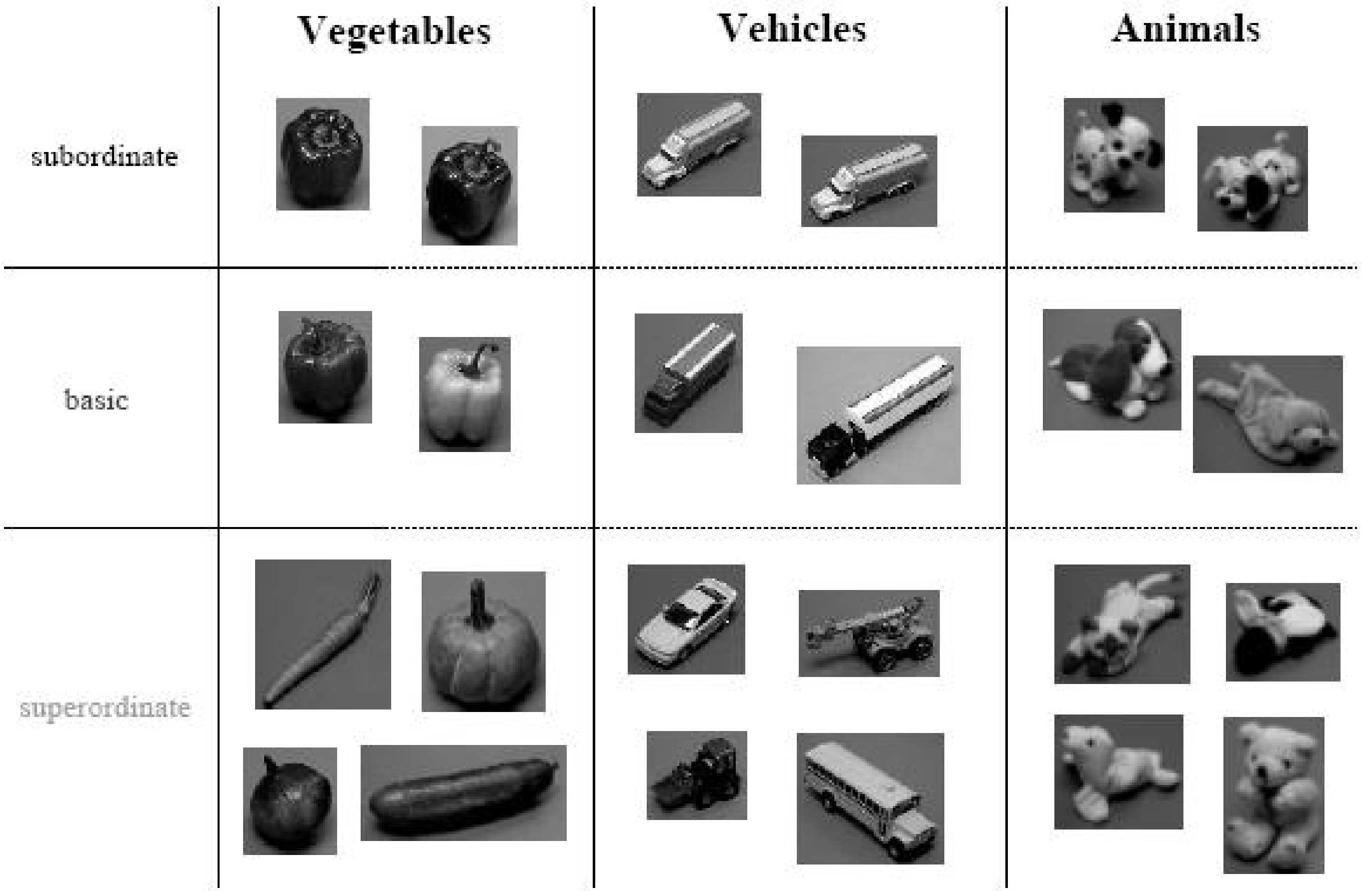
We do not get convergence to the ML hypothesis.
If truth is a rectangle, we do not converge to it
(not a consistent estimator).

A more realistic example

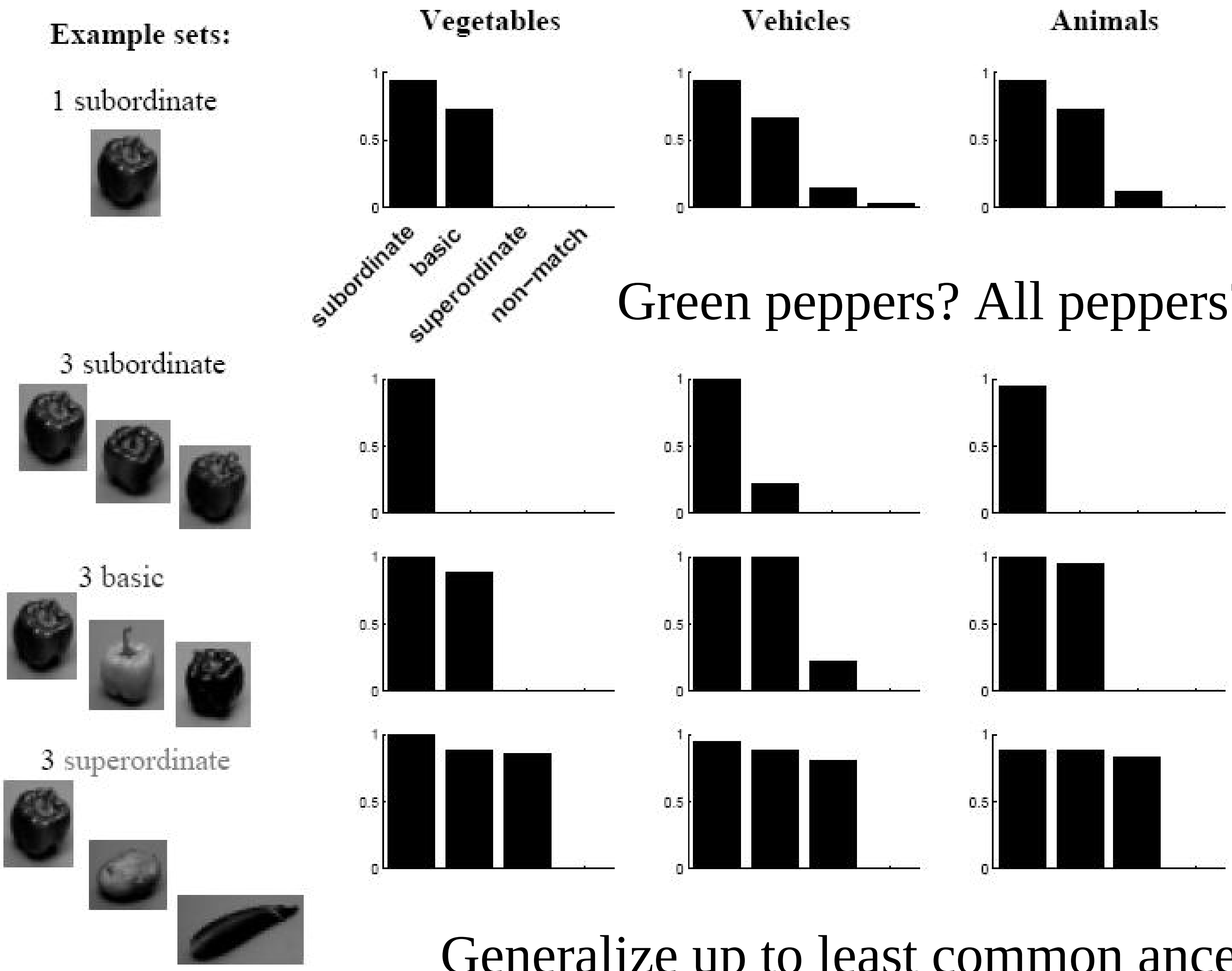
- A discrete hypothesis space (the number game)
- A continuous hypothesis space (the healthy levels concept)
- Word learning


Here is a pog:
Can you give Mr. Frog all the other pogs?

Hierarchical categories



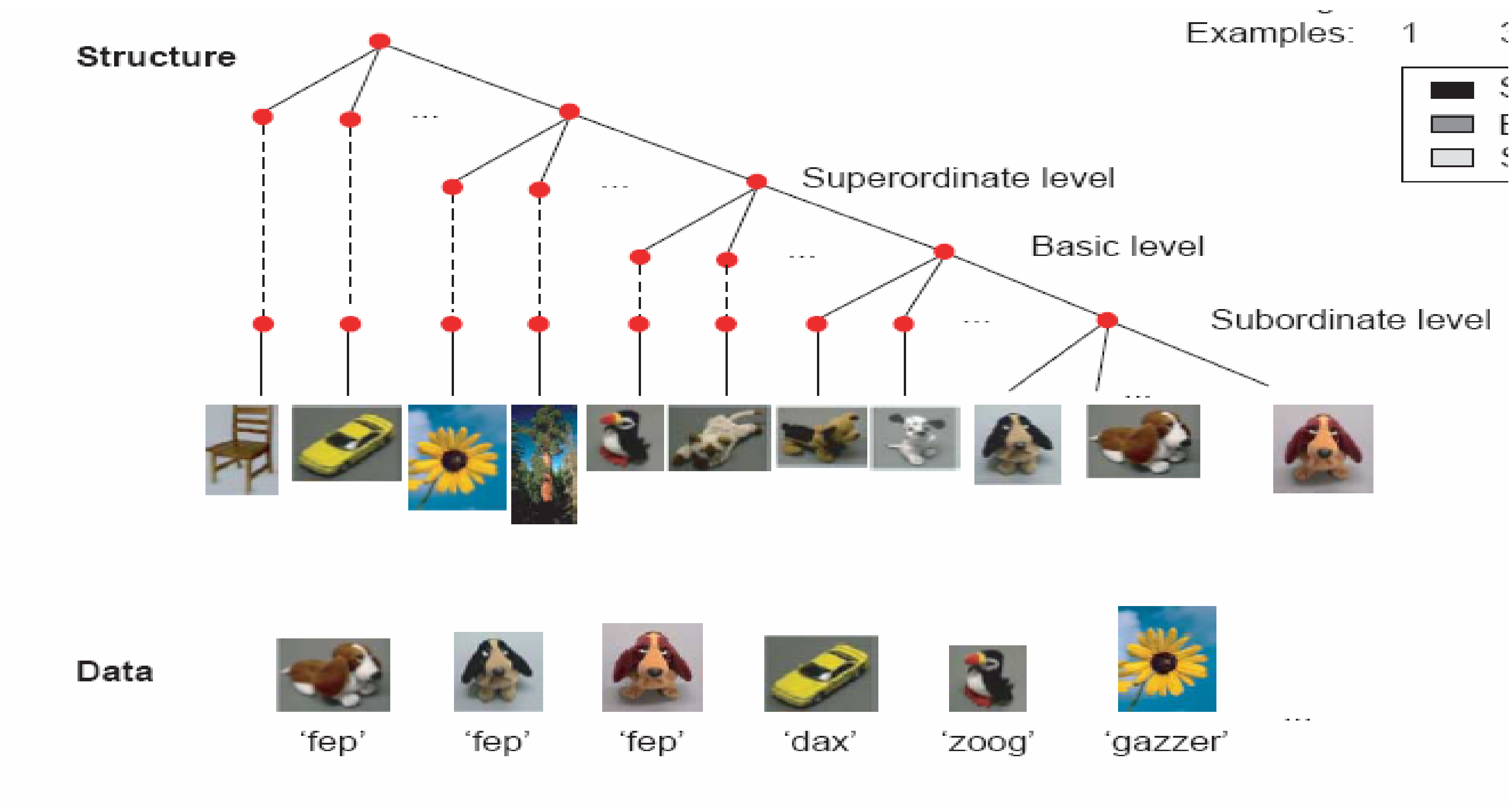
Human data



Green peppers? All peppers? All veg?

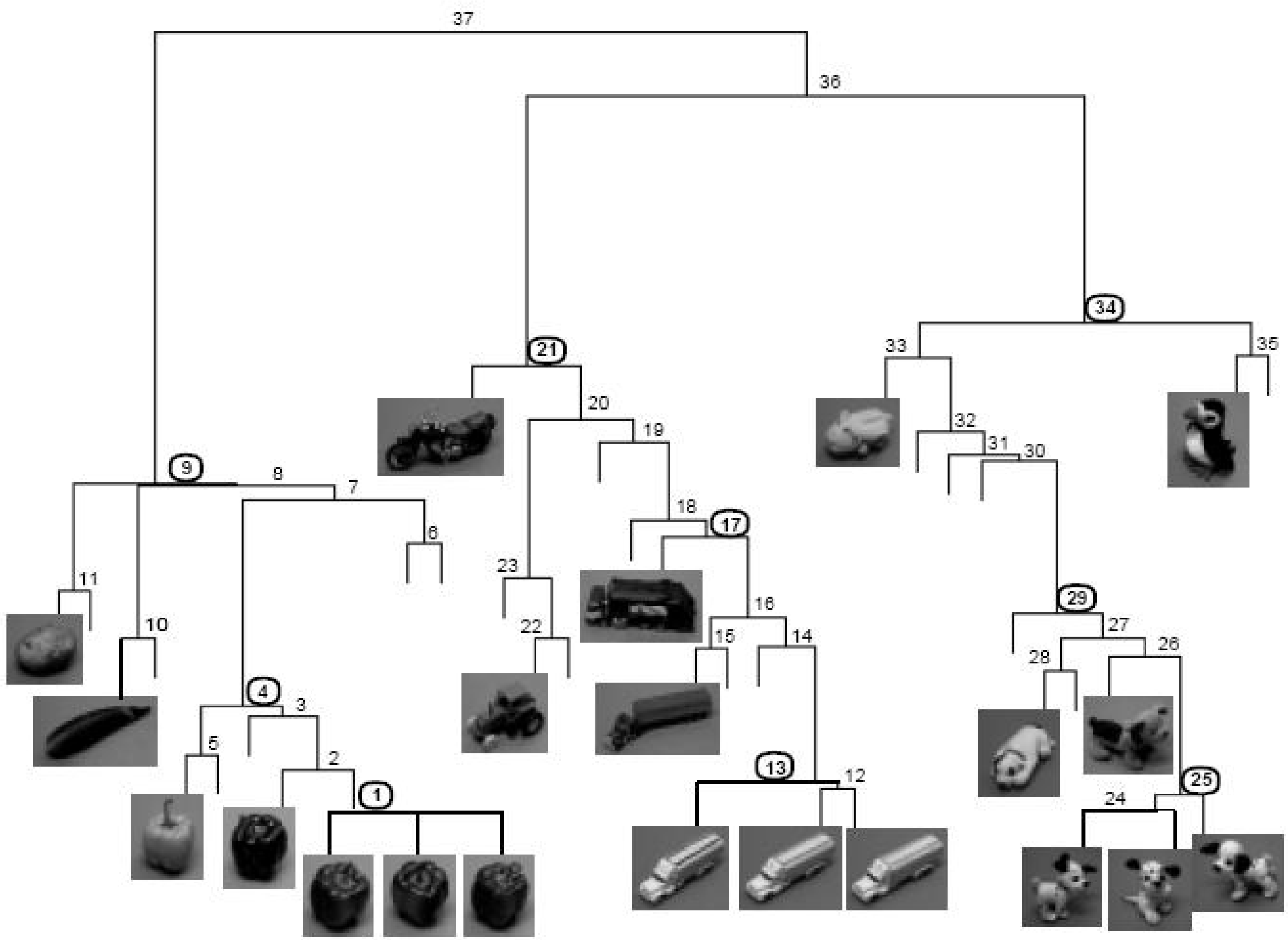
Generalize up to least common ancestor

Hypothesis space



Hypothesis space

Derived by applying agglomerative clustering to human similarity matrix



Hierarchical Clustering

- Cluster based on similarities/distances
- Distance measure between instances $\boldsymbol{x}^r$ and $\boldsymbol{x}^s$

  Minkowski (L_p) (Euclidean for $p = 2$)

$$d_m\left(\boldsymbol{x}^r, \boldsymbol{x}^s\right) = \left[\sum_{j=1}^{d}\left(x_j^r - x_j^s\right)^p\right]^{1/p}$$

City-block distance $d_{cb}\left(\boldsymbol{x}^r, \boldsymbol{x}^s\right) = \sum_{j=1}^{d}\left|x_j^r - x_j^s\right|$

Agglomerative Clustering

- Start with N groups each with one instance and merge two closest groups at each iteration
- Distance between two groups $\mathcal{G}_i$ and $\mathcal{G}_j$:
  - Single-link:

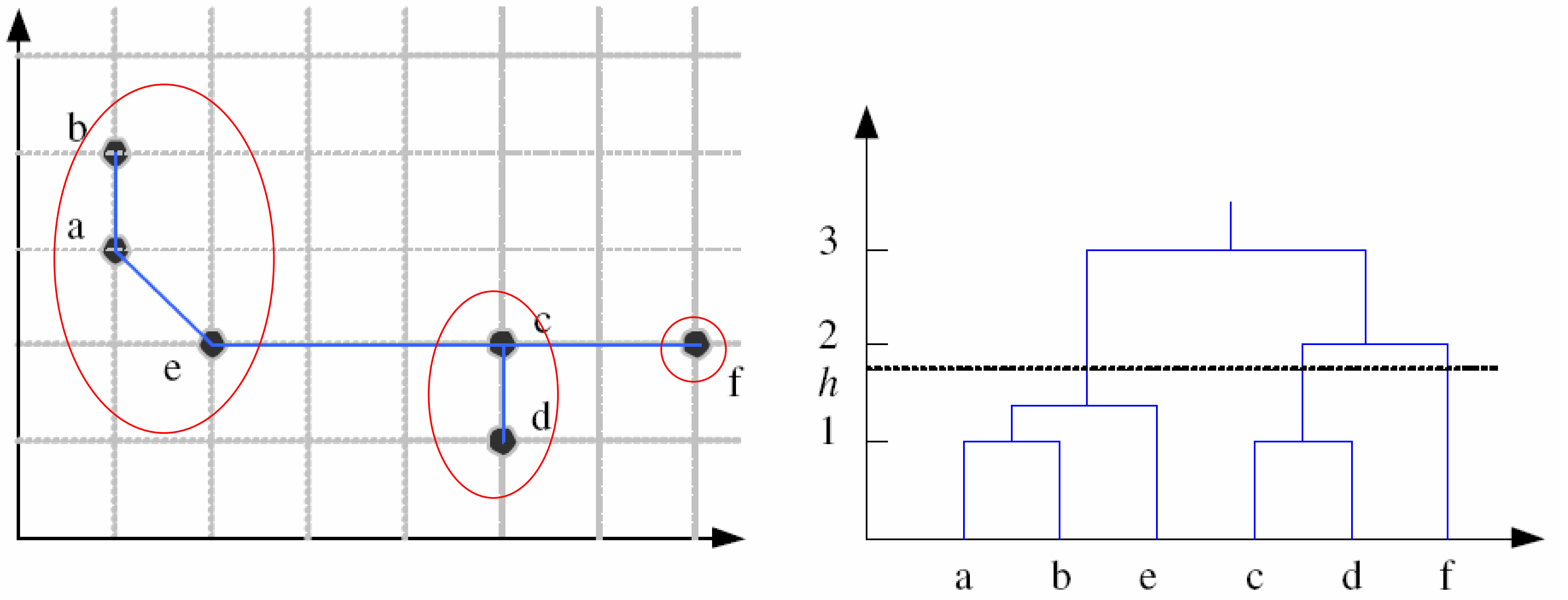
$$d(\mathcal{G}_i, \mathcal{G}_j) = \min_{\boldsymbol{x}^r \in \mathcal{G}_i, \boldsymbol{x}^s \in \mathcal{G}_j} d(\boldsymbol{x}^r, \boldsymbol{x}^s)$$

  - Complete-link:

$$d(\mathcal{G}_i, \mathcal{G}_j) = \max_{\boldsymbol{x}^r \in \mathcal{G}_i, \boldsymbol{x}^s \in \mathcal{G}_j} d(\boldsymbol{x}^r, \boldsymbol{x}^s)$$

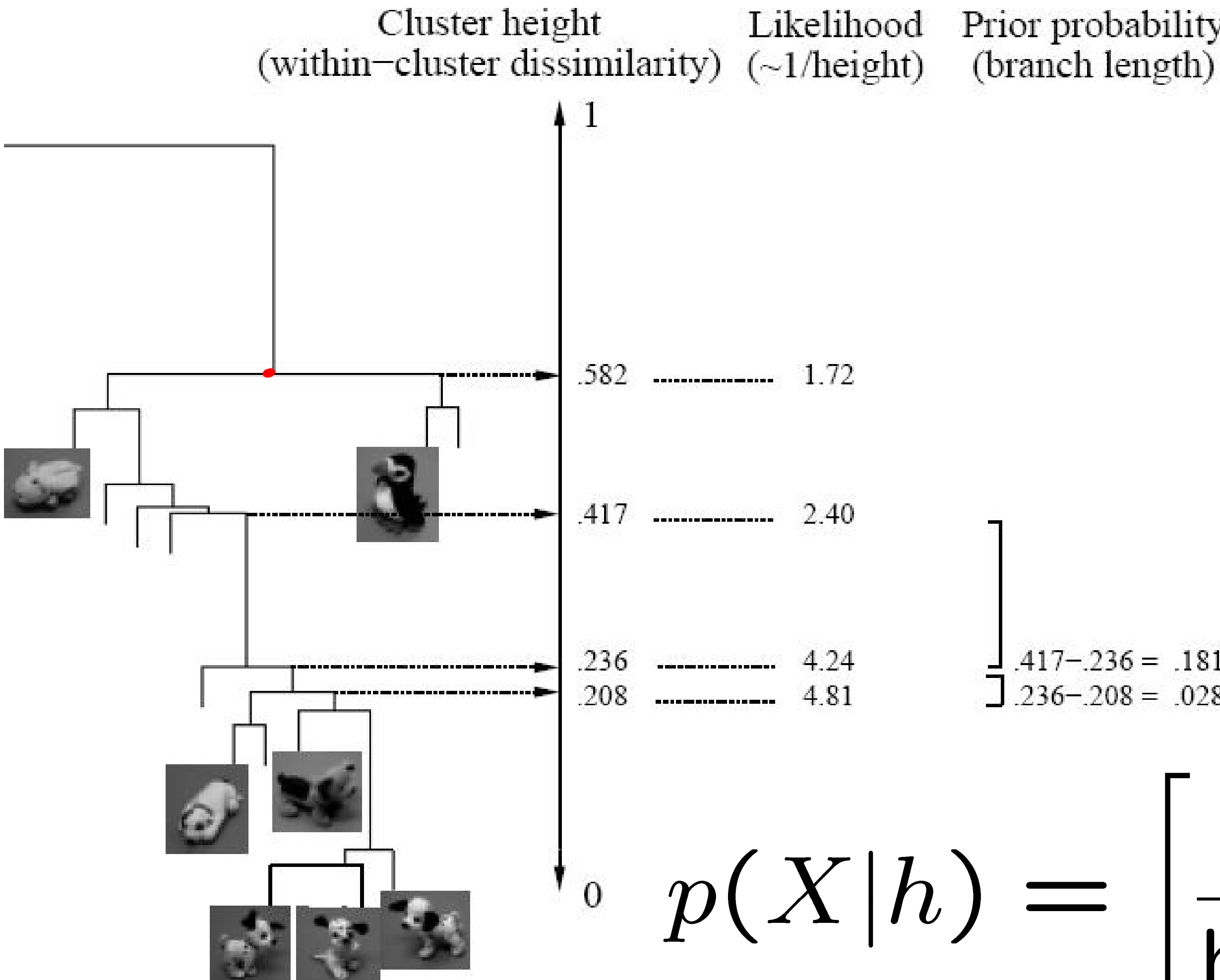
  - Average-link, centroid

Example: Single-Link Clustering



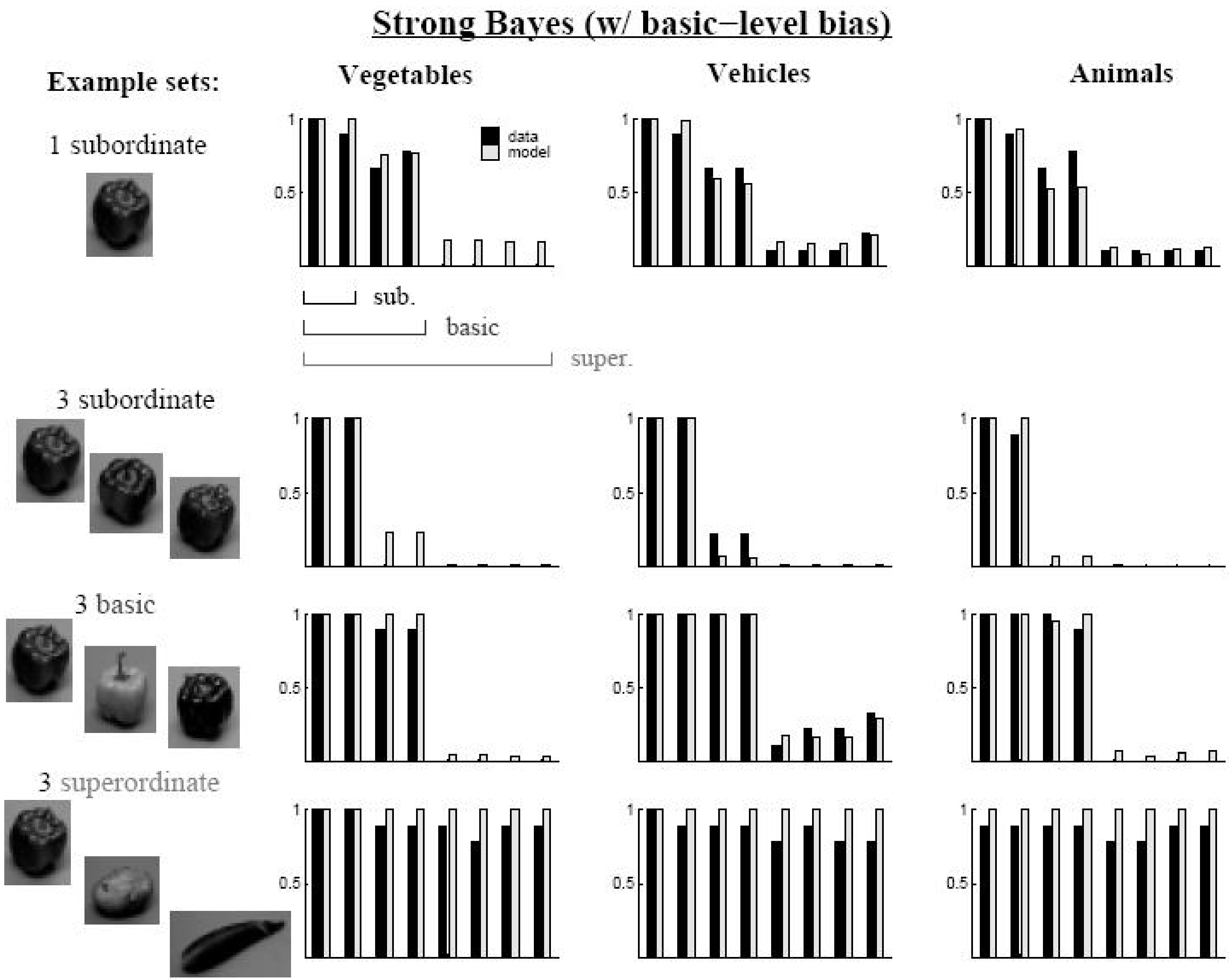
Dendrogram

Prior/ likelihood



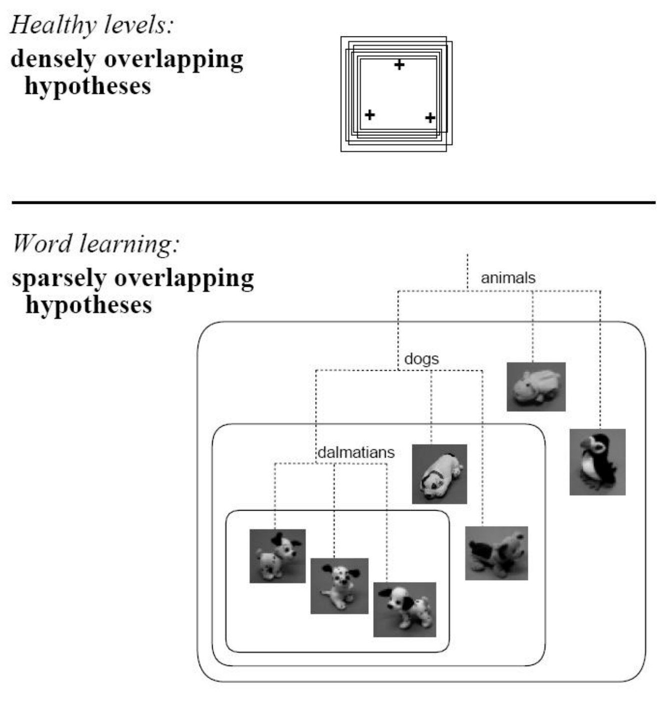
$$p(X|h) = \left[\frac{1}{\mathsf{height}(h)}\right]^n$$

$$p(h) = \mathsf{height}(\mathsf{parent}(h)) - \mathsf{height}(h)$$


Strong Bayes (w/ basic−level bias)
Example sets:
Vegetables
Vehicles
Animals
1 subordinate
3 subordinate
3 basic
3 superordinate
data
model
sub.
basic
super.
1
0.5

Word learning vs healthy levels

- In the word domain, after about N=3 we have an "aha" moment (rule-like learning), but for healthy levels, we need a large sample size, because in the former, hypotheses differ dramatically in size, so we rapidly prefer the smallest consistent, whereas latter averages many.



Rules and exemplars in the number game

- Hyp. space is a mixture of sparse (mathematical concepts) and dense (intervals) hypotheses.
- If data supports mathematical rule (eg X={16,8,2,64}), we rapidly learn a rule, otherwise (eg X={6,23,19,20}) we learn by similarity, and need many examples to get sharp boundary.