Simulating Language
2: Word learning

Kenny Smith
kenny.smith@ed.ac.uk

THE UNIVERSITY OF EDINBURGH

Bayes' rule

- Bayes' rule provides a convenient way of expressing the quantity we want to know (probability of hypothesis given data) in terms of the quantities we already know (probability of data if that hypothesis is true; probability of that hypothesis before we have any data):

$$P(h \mid d) \propto P(d \mid h)P(h)$$

- Or, in full:

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)}$$

Breaking it down

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

$P(h|d)$

- The thing we want to know is called the **posterior**

$P(d|h)$

- The probability of a particular set of data given a particular hypothesis is true is called the **likelihood**

$P(h)$

- The probability that a particular hypothesis is the case, before I have any evidence from the data, is called the **prior**

$P(d)$

- The term on the bottom (the probability of the data independent of the hypothesis) is actually not very interesting to us, since it is the same for all hypotheses - it's a bit of book-keeping.

Bayesian language learning

- Evaluate hypotheses about language given some prior bias (perhaps provided by your biology?) and the data that you've heard

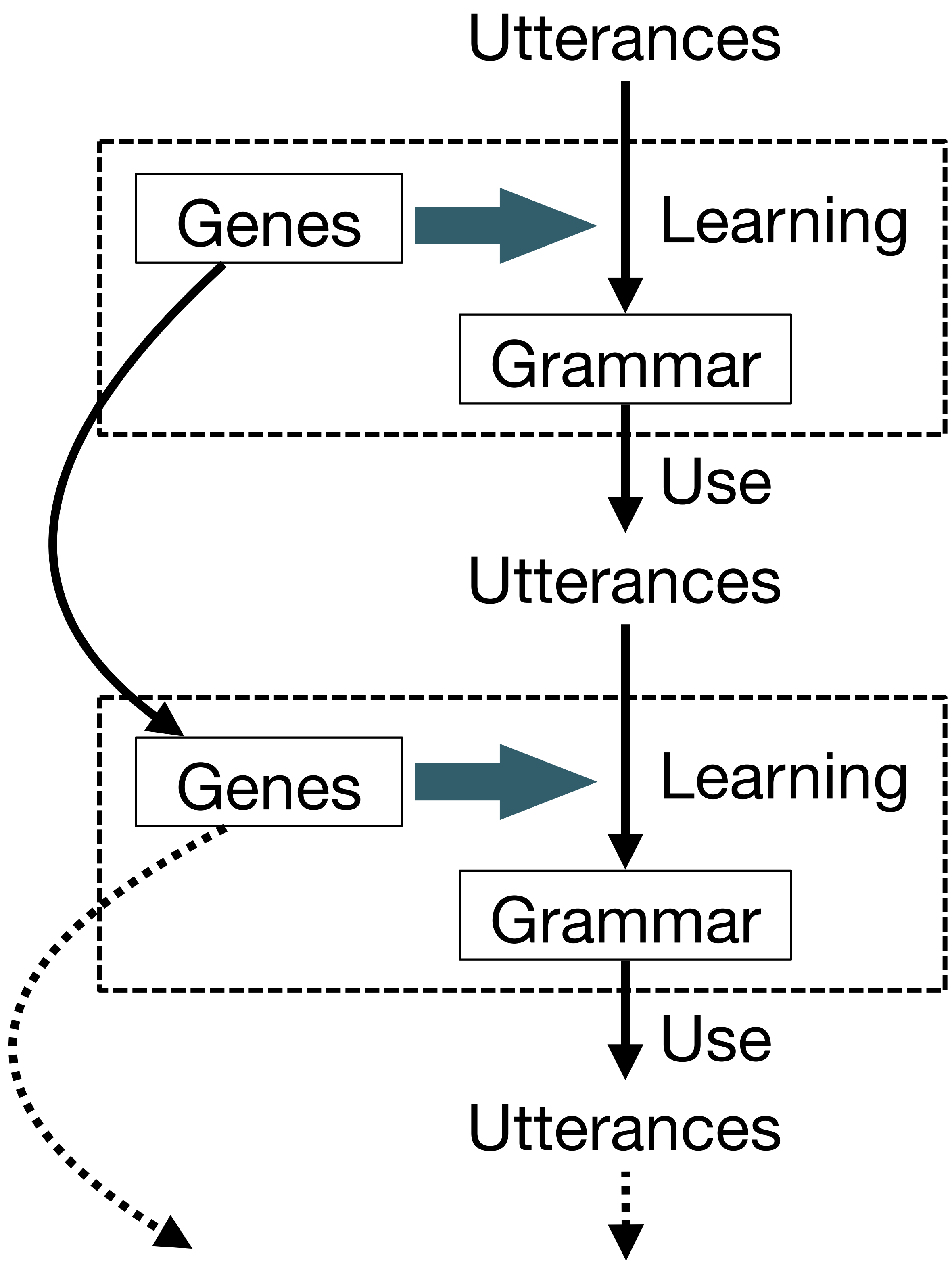
$$P(h|d) \propto P(d|h)P(h)$$

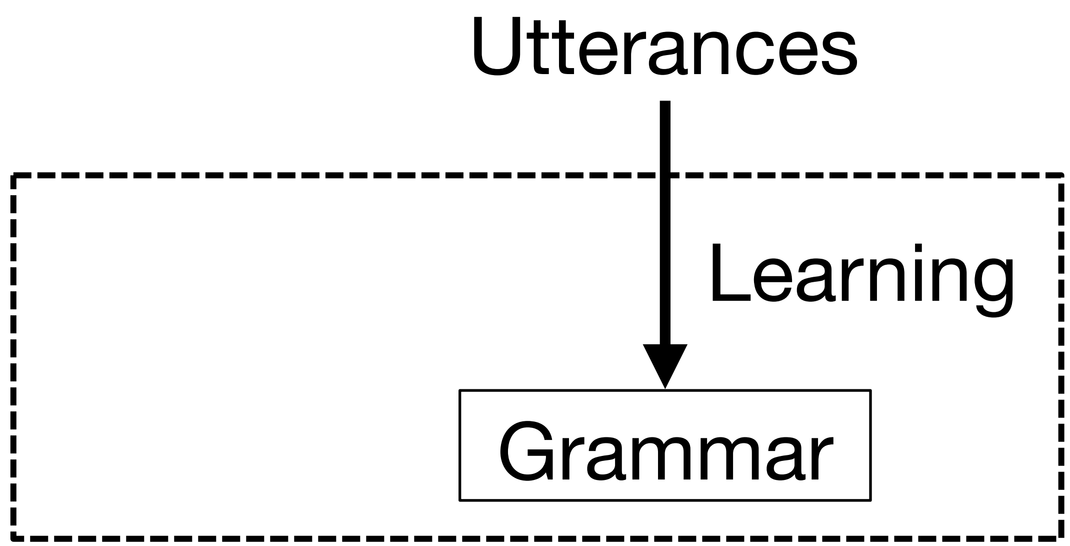
- h = hypothesis about the language
- d = linguistic data

It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness
- If the prior probability of a certain illness is high, this will increase the posterior probability of that illness
- If a particular illness has low prior probability, we need some really convincing evidence to make us believe it to be true


Utterances
Genes
Learning
Grammar
Use
Utterances
Genes
Learning
Grammar
Use
Utterances



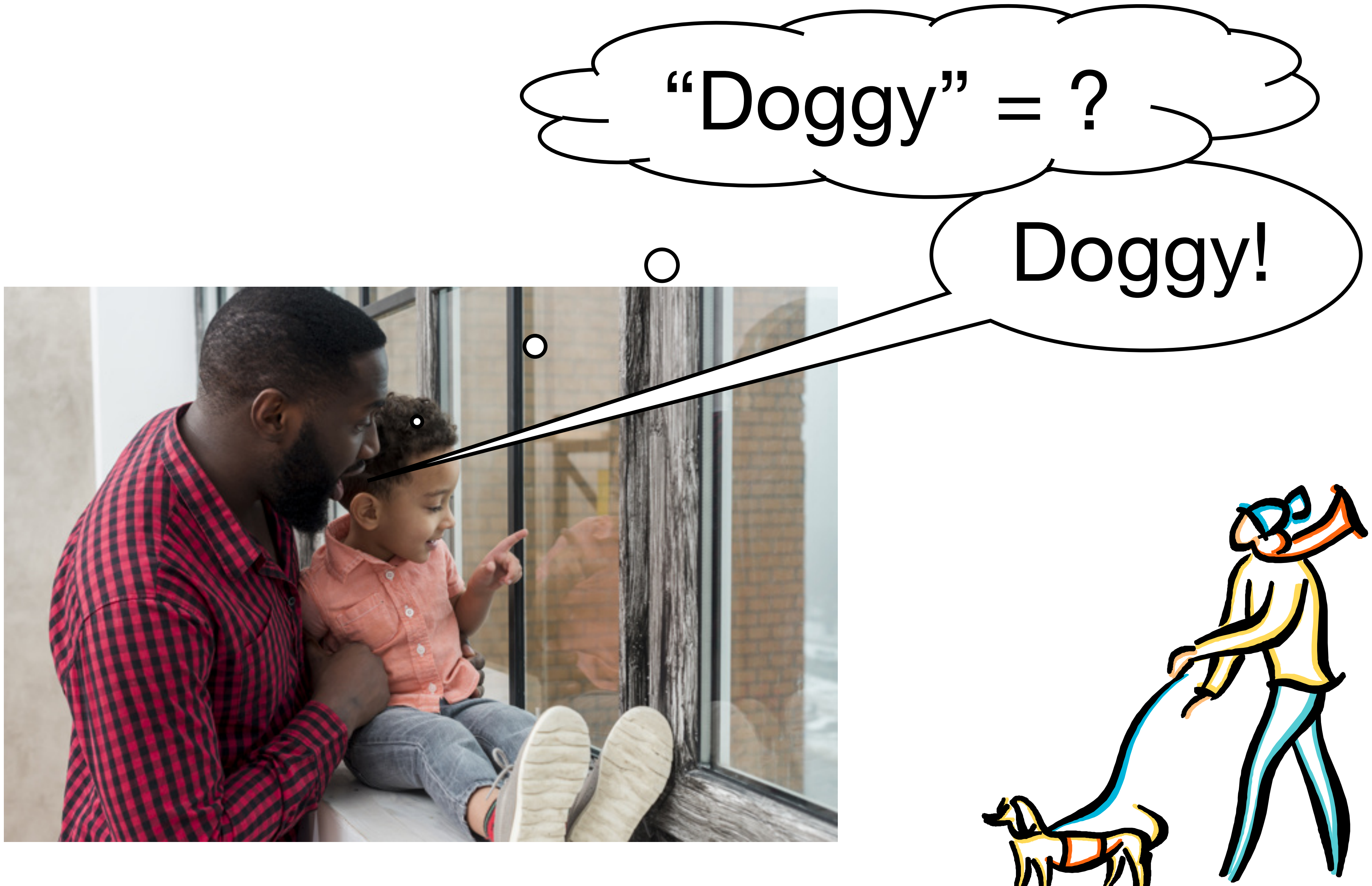
$$P(h|d) \propto P(d|h)P(h)$$

Bayesian language learning $P(h|d) \propto P(d|h)P(h)$

- This modelling approach provides several advantages for our purposes
  - Quantitative (we can put numbers on stuff)
  - Simple (just multiplying and dividing)
  - Transparent (nice clean representation of the role of **prior knowledge**)
  - Surprisingly powerful (as we'll see)

Word learning

Learning the meaning of words



Quine (1960): meaning underdetermined by data

- The four legged animal
- The two legged animal
- Some part of either (the leg, the hat, …)
- Some property of some part (the length of the leg, the material of the hat)
- Nothing to do with what you're seeing ("I'm hungry")
- Something weirder (a wet nose and a waggable tail, but only until Scotland win the World Cup)

There are in principle **infinitely many possible meanings** for "doggy" which would be consistent with this usage, and **any possible sequences of usages**

Learners must have **some** constraints on word meaning

Minimally: to rule out the extremely wacky word meanings

But maybe they are more detailed:

- Expectations about meanings (e.g. words refer to whole objects, words refer to basic-level categories, words generalise by shape of referent, …: Macnamara, 1972; Markman, 1989; Landau, Smith & Jones, 1988)

- Expectations about words (e.g. word meanings are mutually exclusive: Markman & Wachtel, 1988)

- …

If the constraints on learning are minimal, how is rapid word learning possible?

If the constraints on learning are strong, how do we learn words that don't fit the constraints?

Word learning as *Bayesian inference*

$$P(h|d) \propto P(d|h)P(h)$$

- Xu, F., & Tenenbaum, J. B. (2007) Word learning as Bayesian Inference. Psychological Review, 114, 245-272

- You are trying to use evidence provided by *instances of word use* to infer *unobservable word meaning*

hypotheses = word meanings

data = labelling events

likelihood = how word meanings lead to labelling events

prior = the kind of meanings I expect words to have

This is a *fep*

What does *fep* mean?

A. Dalmatian

B. Dog

C. Animal

These are also *feps*

What does *fep* mean?

A. Dalmatian

B. Dog

C. Animal

Here are 3 *daxes*



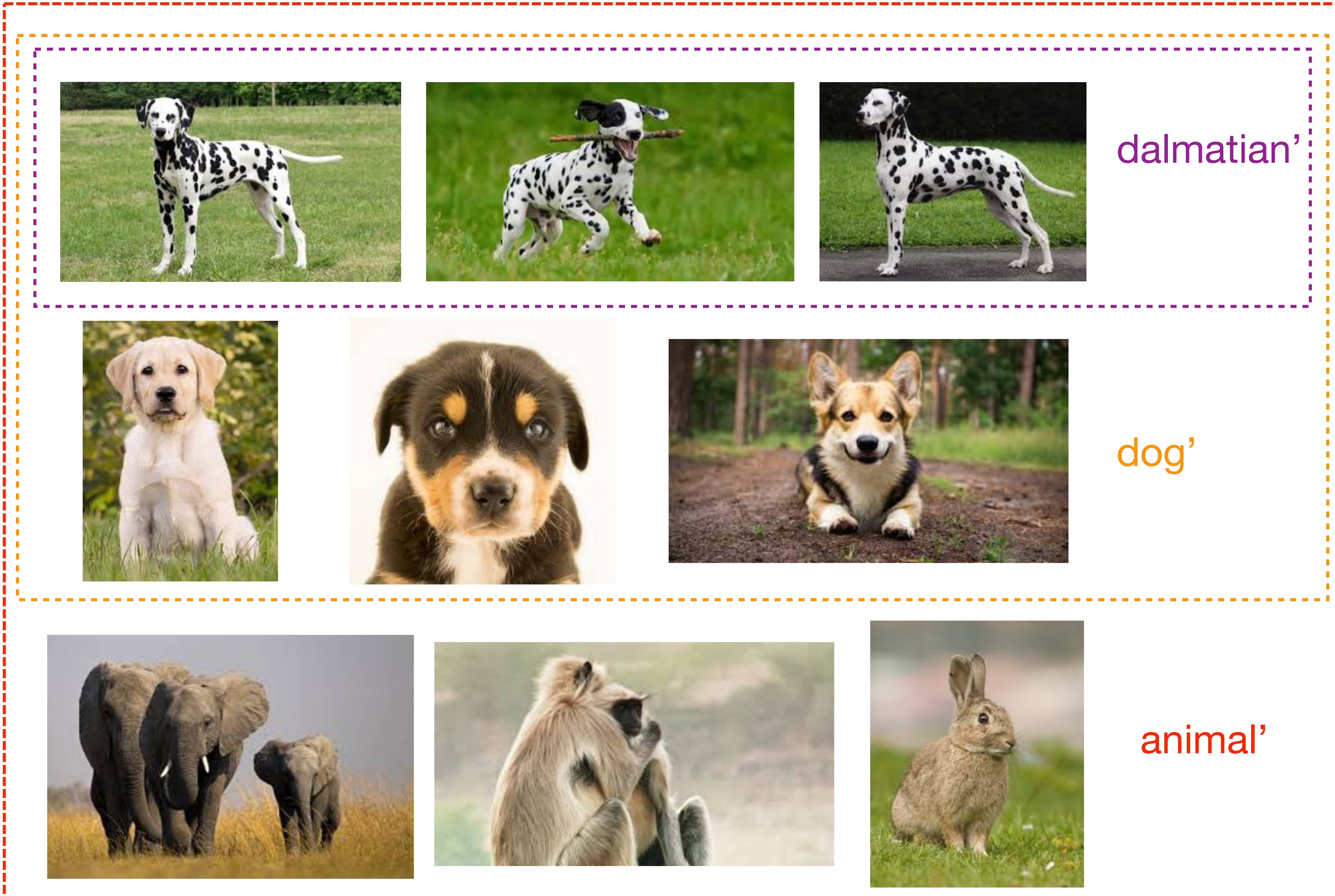
What does *dax* mean?

A. Dalmatian

B. Dog

C. Animal

Did you infer different meanings for *fep* and *dax*?
What factors influenced your decision?


dalmatian’
dog’
animal’

Quantifying a *suspicious coincidence*

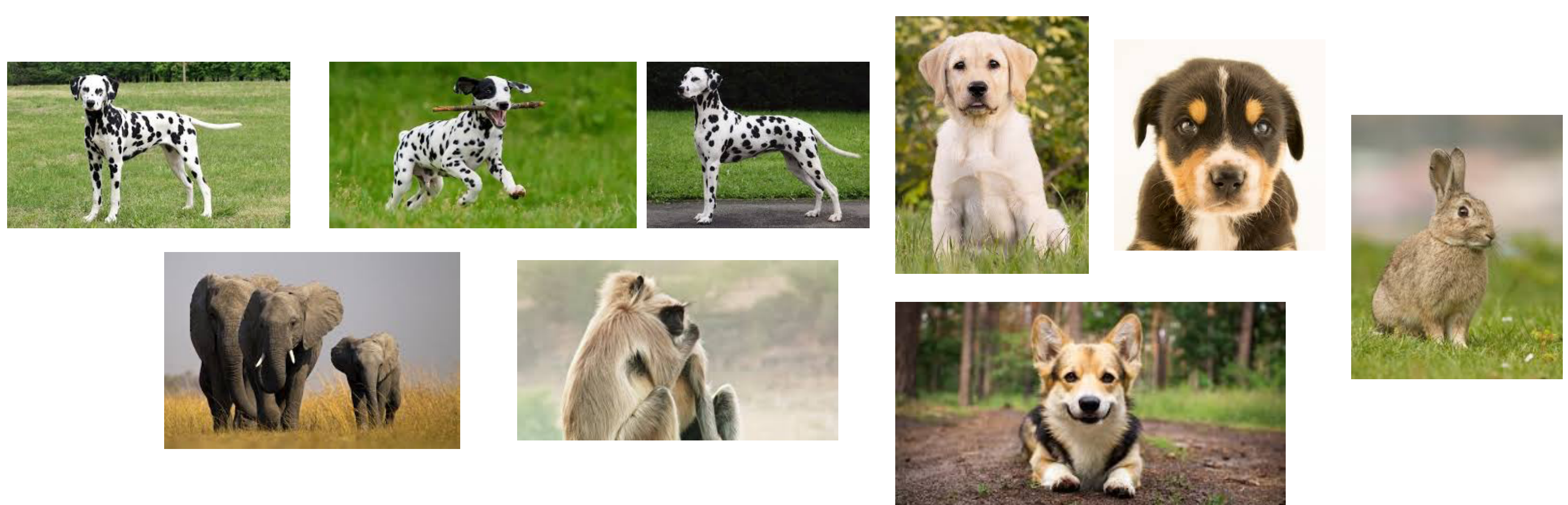
$$P(h|d) \propto P(d|h)P(h)$$

3 hypotheses under consideration

fep=dalmatian’

fep=dog’

fep=animal’



Quantifying a *suspicious coincidence*

$$P(h|d) \propto P(d|h)P(h)$$

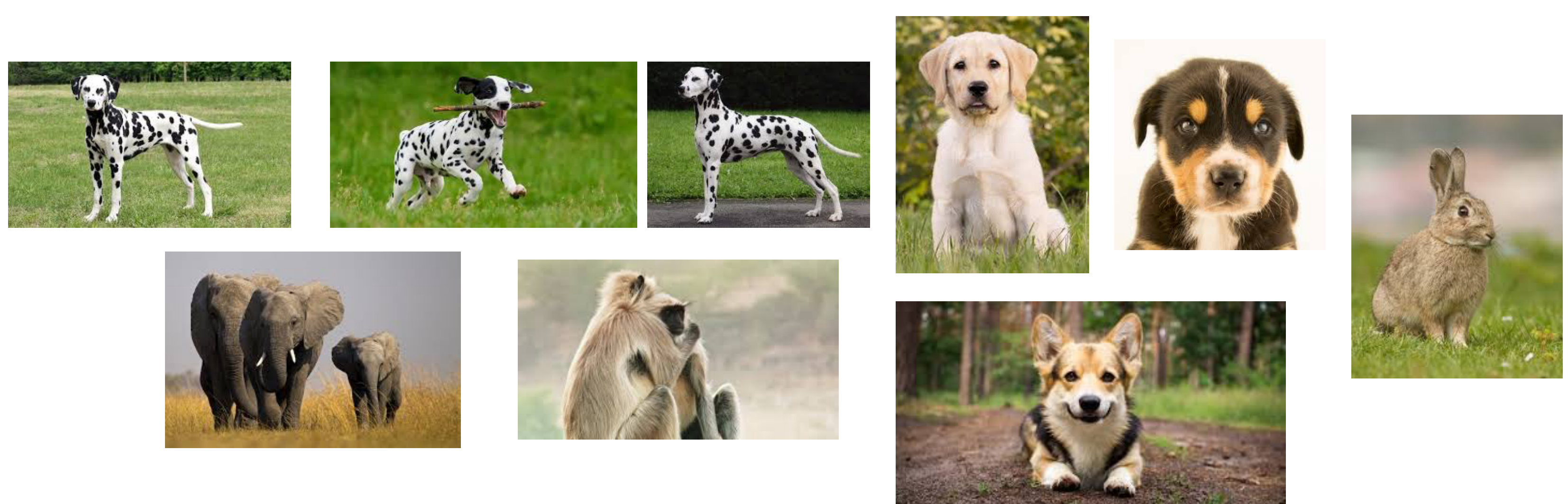
"fep"

Likelihood: P( | *fep*=dalmatian') = ???

fep=dalmatian'

fep=dog'

fep=animal'



Quantifying a *suspicious coincidence*

$$P(h|d) \propto P(d|h)P(h)$$

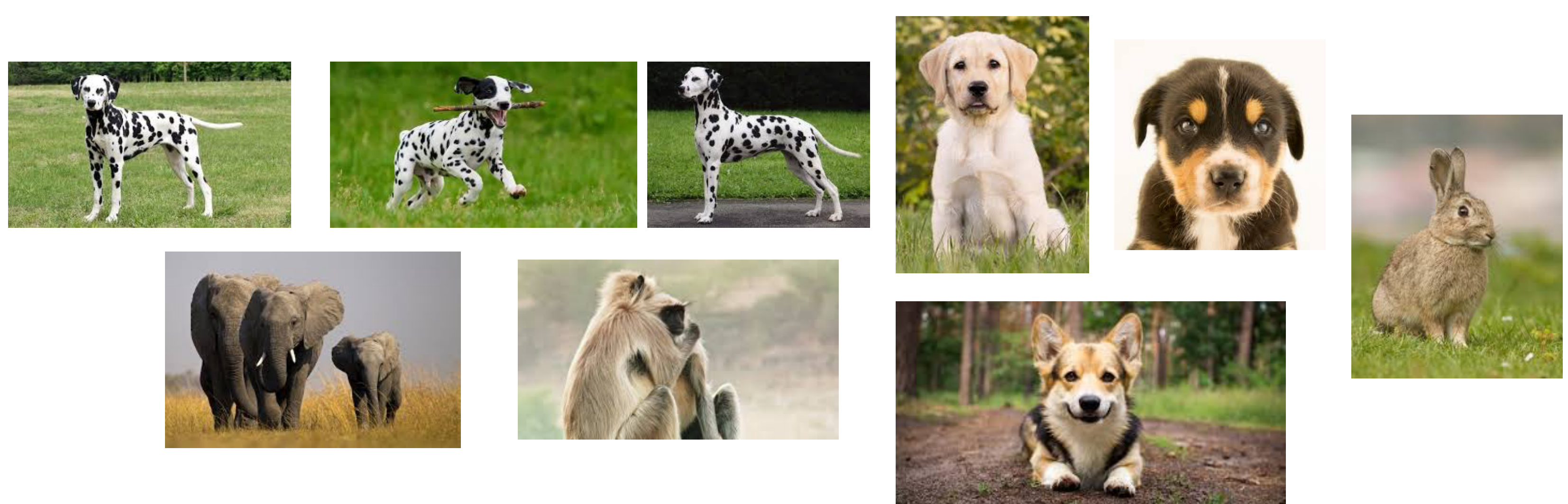
“*fep*”

Likelihood: P( | *fep*=dog’) = ???

fep=dalmatian’

fep=dog’

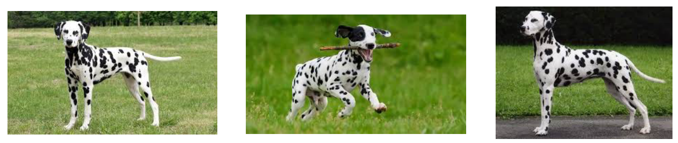
fep=animal’



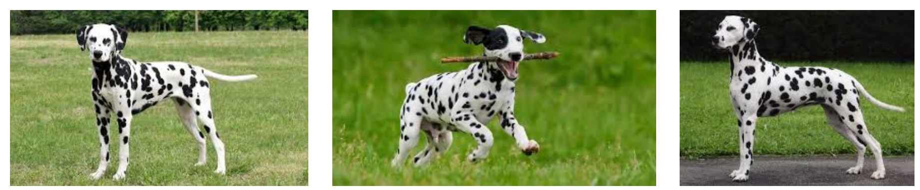
Quantifying a *suspicious coincidence*

$$P(h|d) \propto P(d|h)P(h)$$

“*fep*” “*fep*” “*fep*”

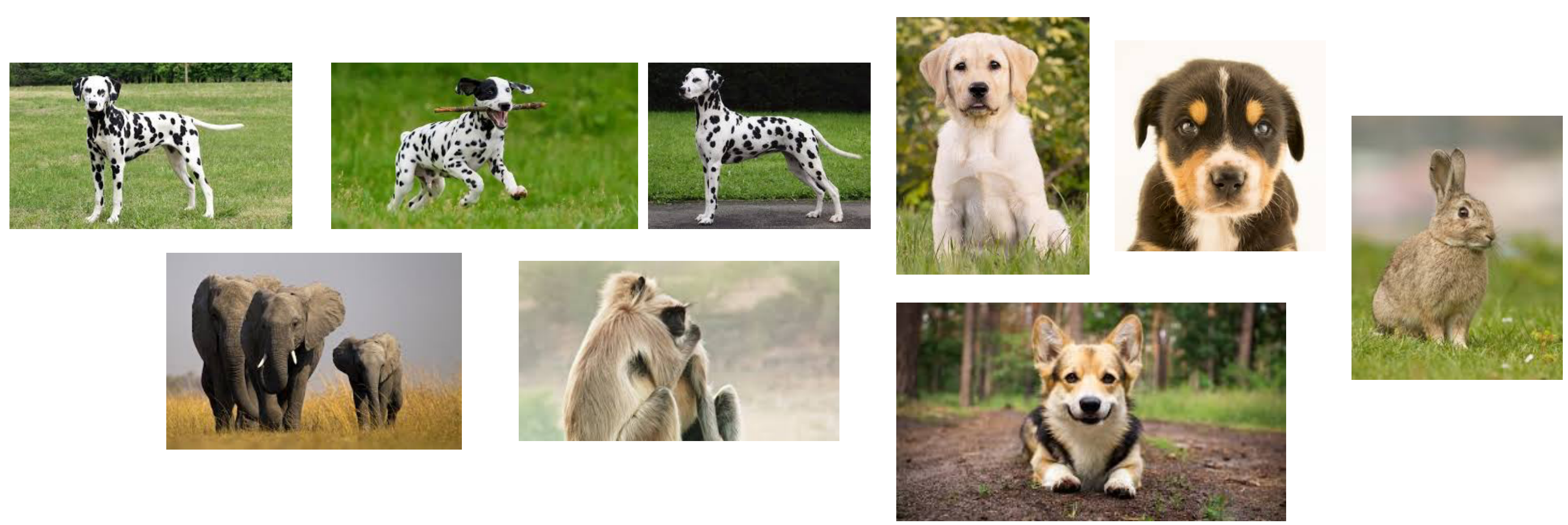
Likelihood: P( | *fep*=dalmatian’) = ???

fep=dalmatian’



fep=dog’

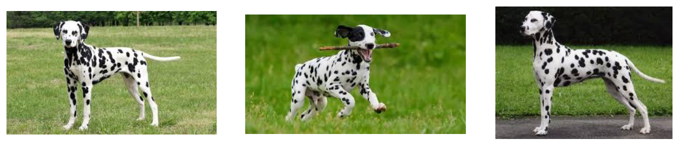
fep=animal’



Quantifying a *suspicious coincidence*

$$P(h|d) \propto P(d|h)P(h)$$

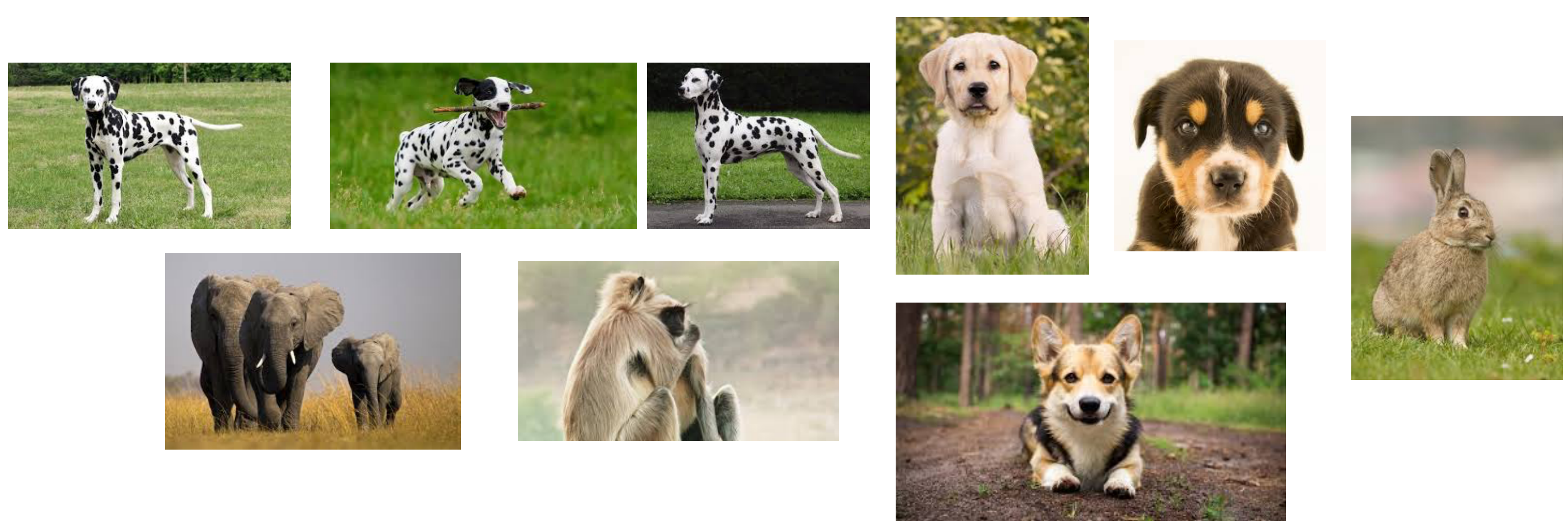
“fep” *“fep”* *“fep”*

Likelihood: P( | *fep*=dog’) = ???

fep=dalmatian’

fep=dog’

fep=animal’



Quantifying a *suspicious coincidence*

$$P(h|d) \propto P(d|h)P(h)$$

“*fep*”

P( | *fep*=dalmatian’) = 1/3

“*fep*”

P( | *fep*=dog’) = 1/6

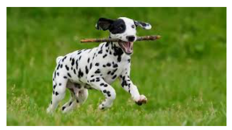
“*fep*”

P( | *fep*=animal’) = 1/9

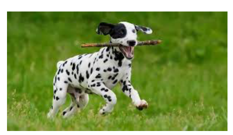
Quantifying a *suspicious coincidence*

$$P(h|d) \propto P(d|h)P(h)$$

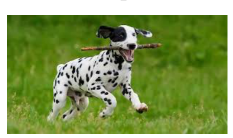
“*fep*” “*fep*” “*fep*”

P(   | *fep*=dalmatian’) = 1/3 x 1/3 x 1/3 = 1/27

“*fep*” “*fep*” “*fep*”

P(   | *fep*=dog’) = 1/6 x 1/6 x 1/6 = 1/216

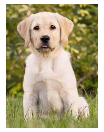
“*fep*” “*fep*” “*fep*”

P(   | *fep*=animal’) = 1/9 x 1/9 x 1/9 = 1/729

Quantifying a
suspicious coincidence

$$P(h|d) \propto P(d|h)P(h)$$

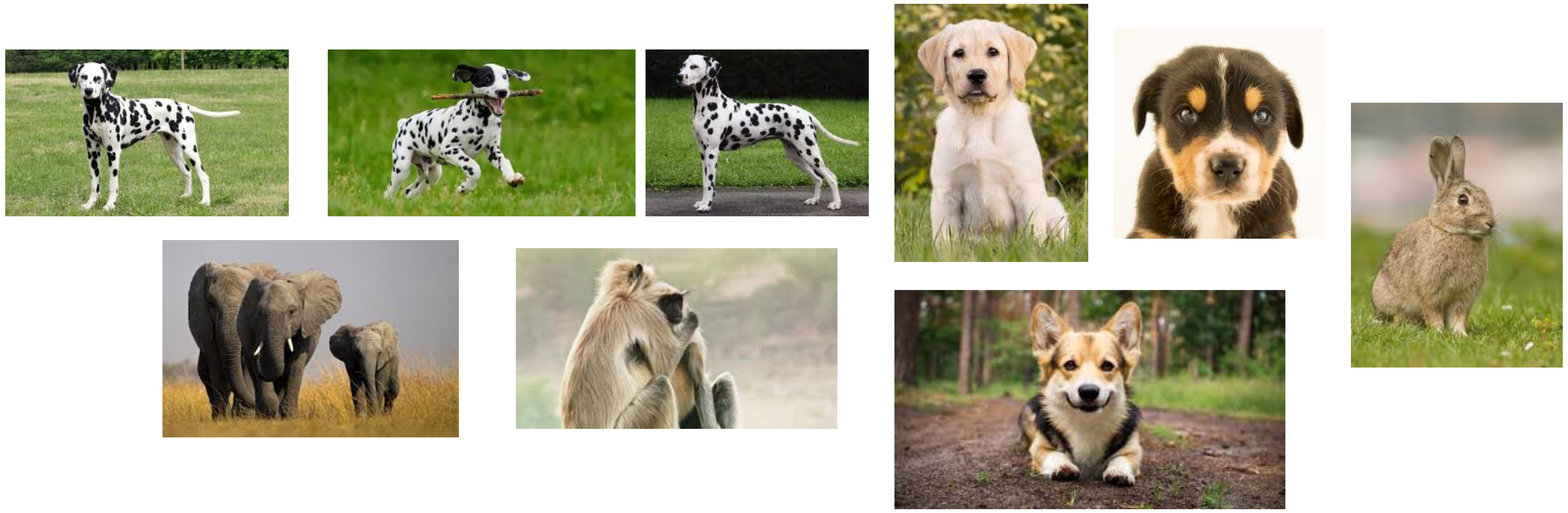
"*dax*" "*dax*" "*dax*"

Likelihood: P(   | *dax*=dalmatian') = ???

fep=dalmatian'

fep=dog'

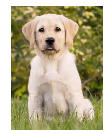
fep=animal'



Quantifying a *suspicious coincidence*

$$P(h|d) \propto P(d|h)P(h)$$

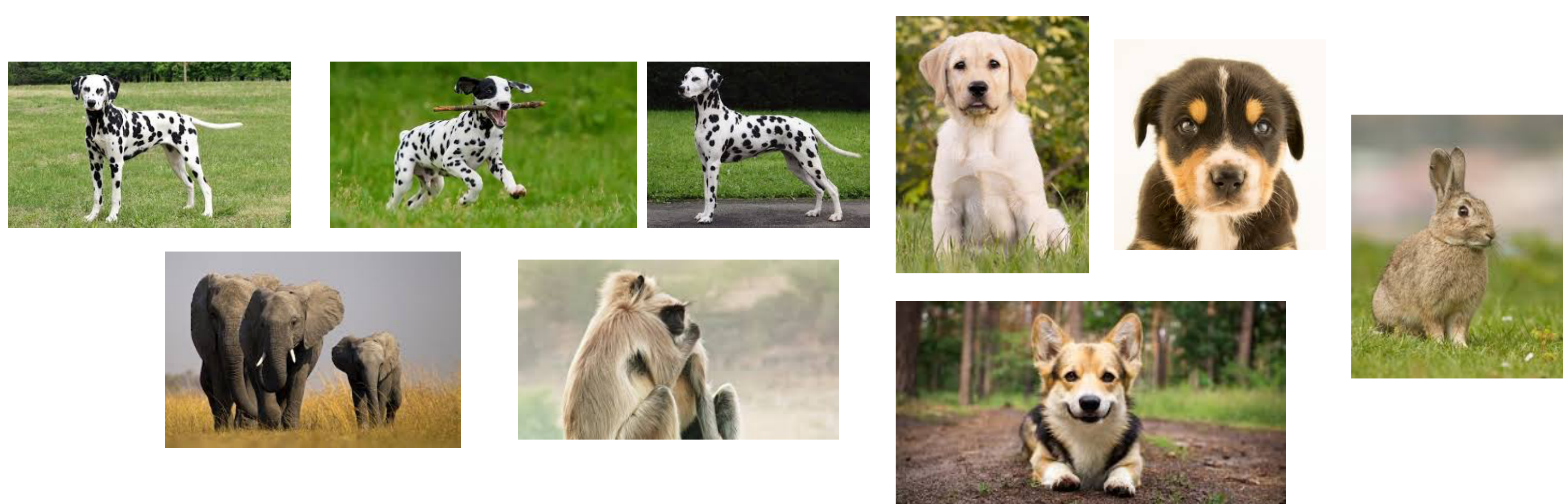
“*dax*” “*dax*” “*dax*”

Likelihood: P(   | *dax*=dog’) = ???

fep=dalmatian’

fep=dog’

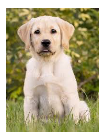
fep=animal’



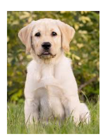
Quantifying a *suspicious coincidence*

$$P(h|d) \propto P(d|h)P(h)$$

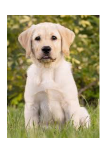
“*dax*” “*dax*” “*dax*”

P(   | *dax*=dalmatian’) = 1/3 x 0 x 0 = 0

“*dax*” “*dax*” “*dax*”

P(   | *dax*=dog’) = 1/6 x 1/6 x 1/6 = 1/216

“*dax*” “*dax*” “*dax*”

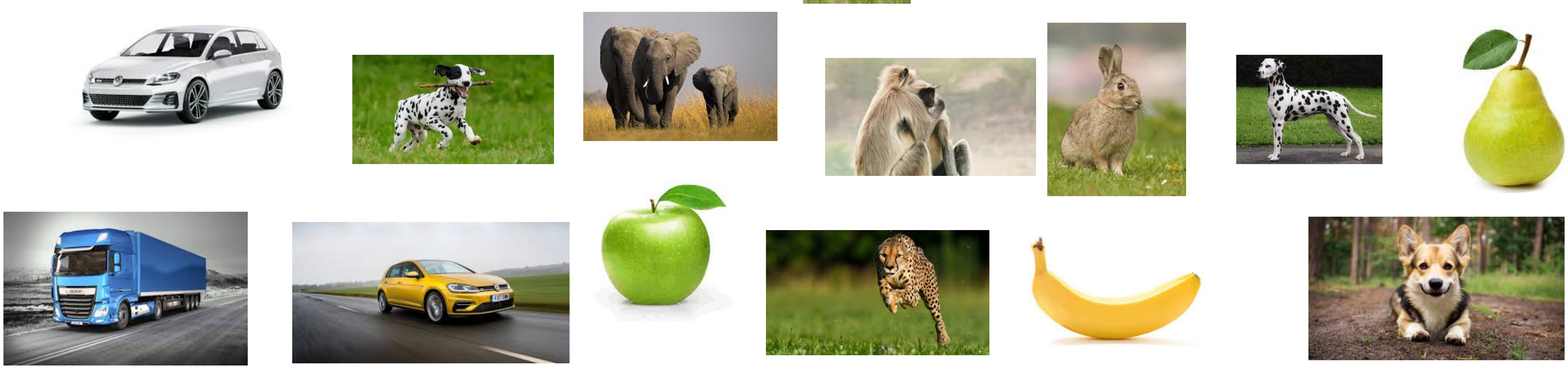
P(   | *dax*=animal’) = 1/9 x 1/9 x 1/9 = 1/729

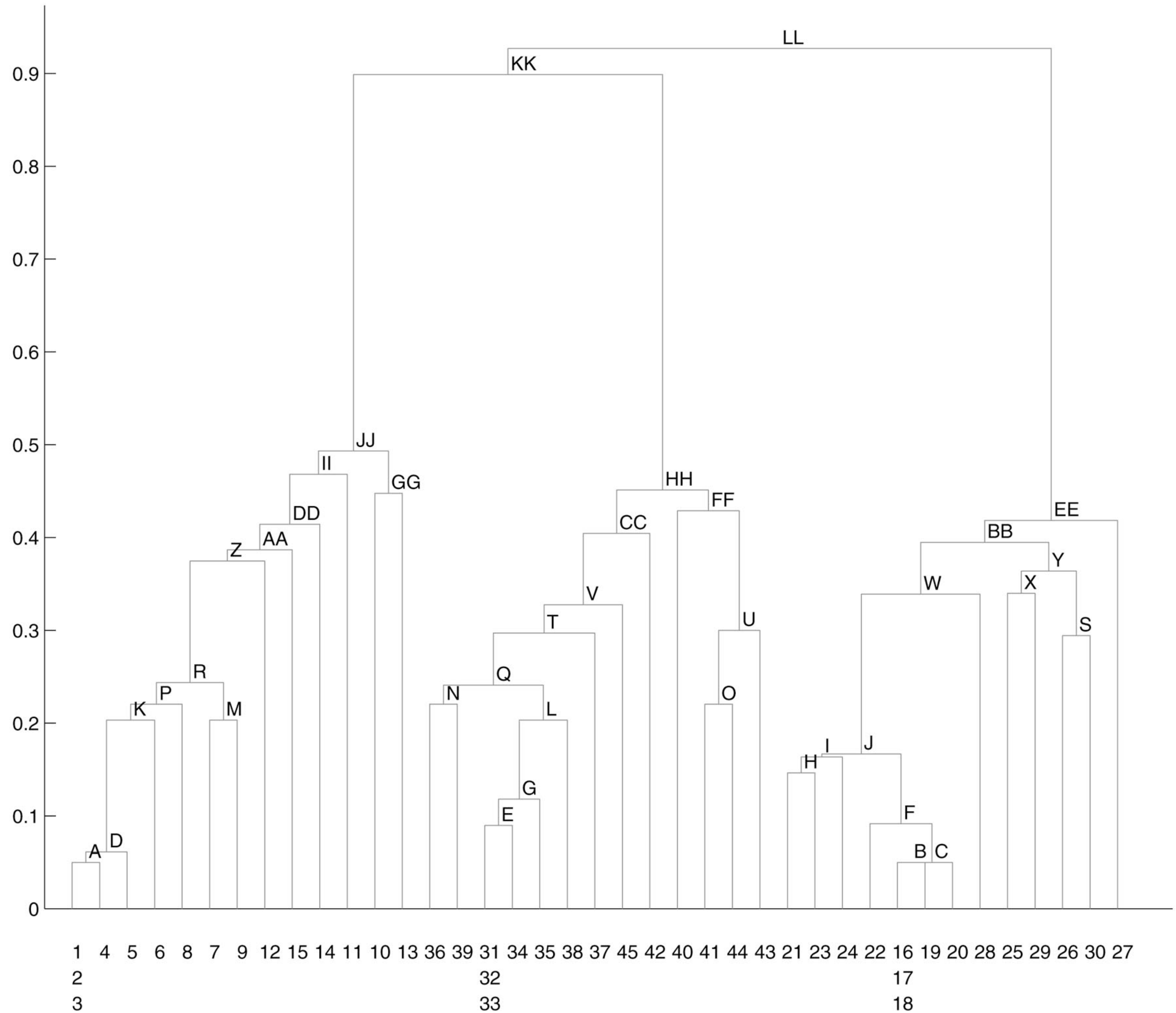
Xu, F., & Tenenbaum, J. B. (2007) Word learning as Bayesian Inference. Psychological Review, 114, 245-272

Their task

These are *feps*

Show me all the *feps*

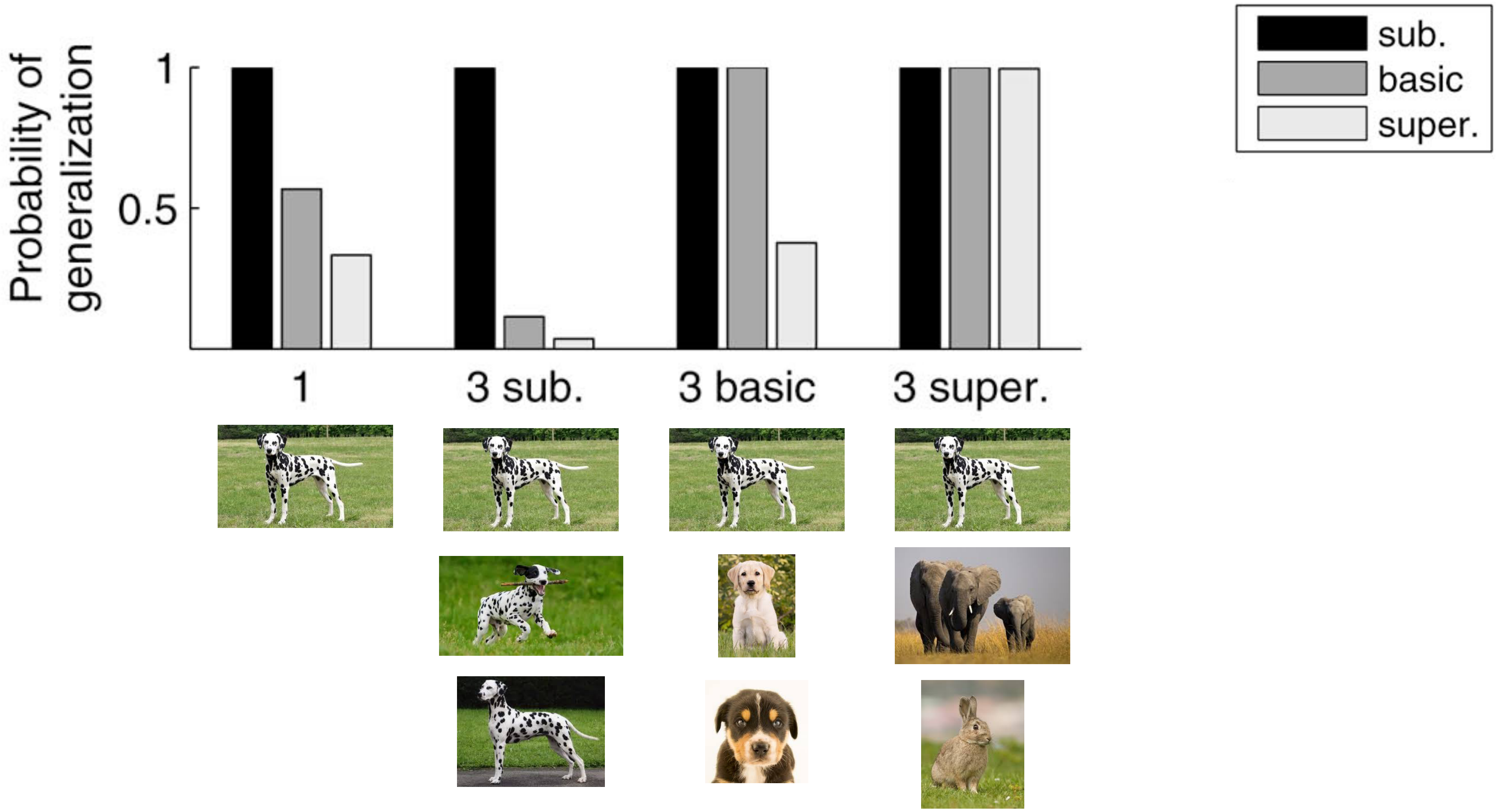




0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
LL
KK
JJ
II
GG
HH
FF
EE
DD
CC
BB
AA
Z
Y
X
W
V
U
T
S
R
Q
P
O
N
M
L
K
J
I
H
G
F
E
D
C
B
A
1 4 5 6 8 7 9 12 15 14 11 10 13 36 39 31 34 35 38 37 45 42 40 41 44 43 21 23 24 22 16 19 20 28 25 29 26 30 27
2
3
32
33
17
18

Model predictions

$$P(h|d) \propto P(d|h)P(h)$$

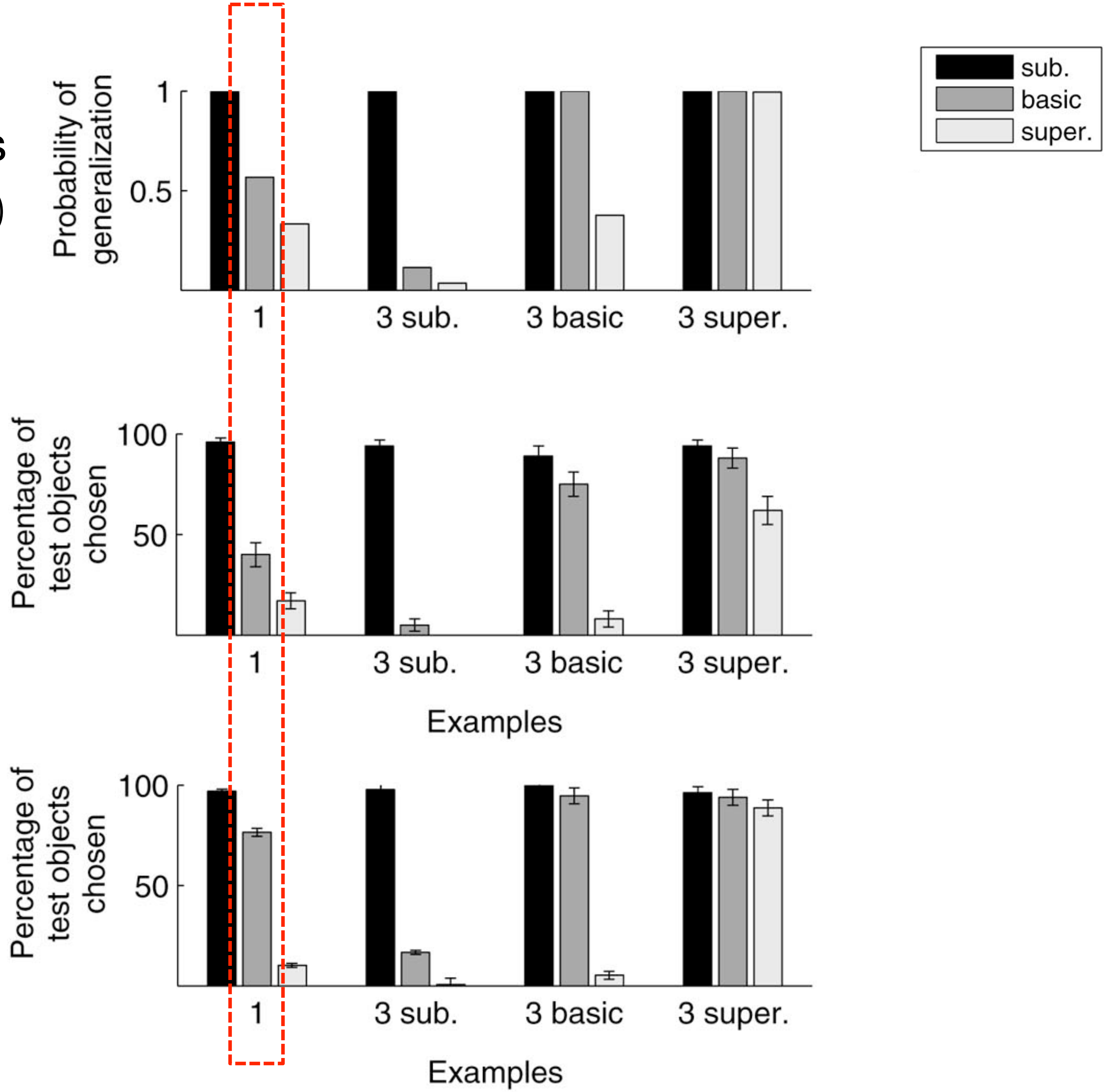


Model predictions

$$P(h|d) \propto P(d|h)P(h)$$

Children

Adults



Add a basic-level bias

$$P(h|d) \propto P(d|h)P(h)$$

Uniform prior

P(*fep*=dalmatian') = P(*fep*=dog') = P(*fep*=animal')

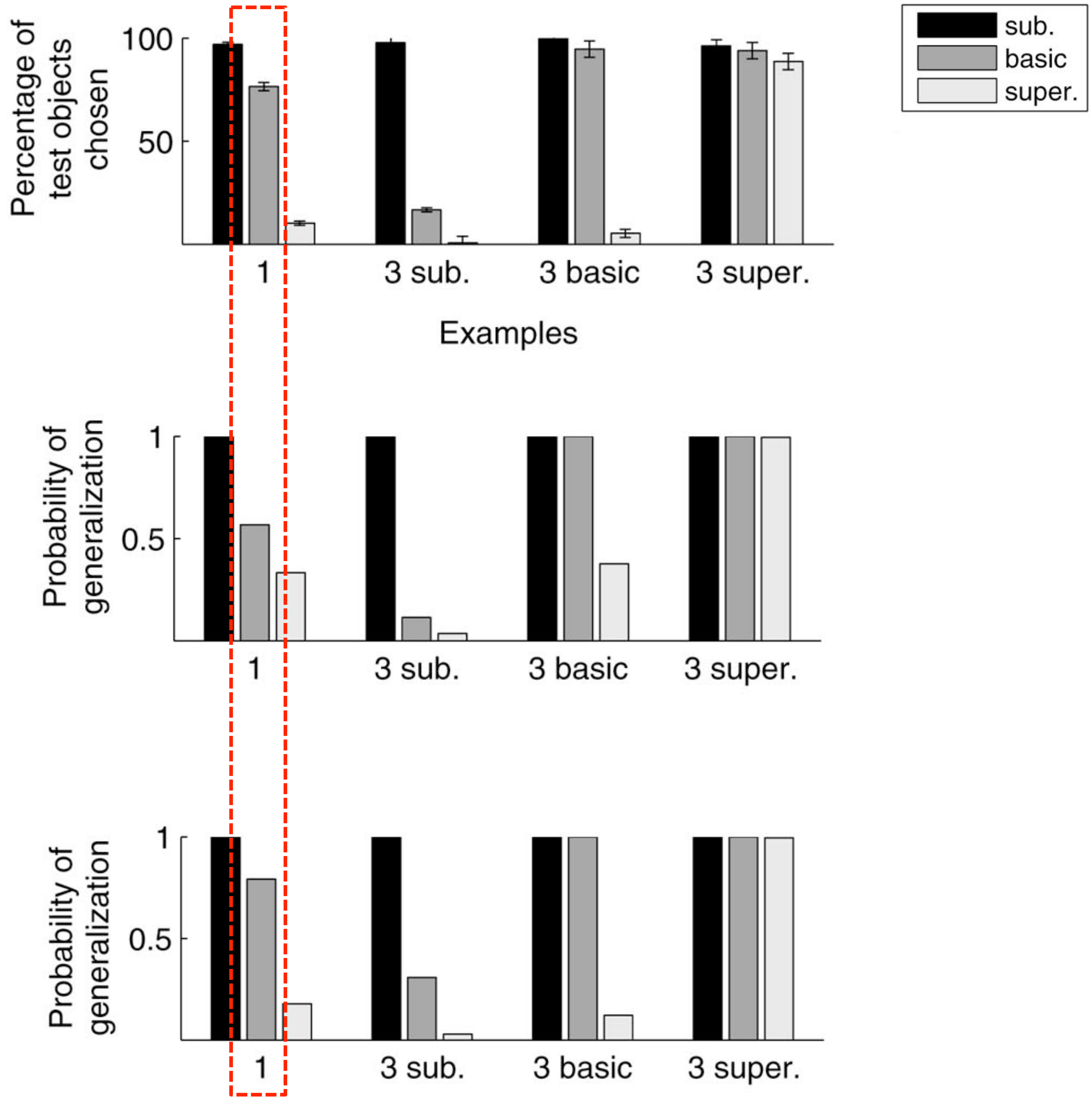
Prior with a basic-level bias

P(*fep*=dog') > P(*fep*=dalmatian') = P(*fep*=animal')

Adults

Model without basic-level bias
$P(h|d) \propto P(d|h)P(h)$

Model **with** basic-level bias
$P(h|d) \propto P(d|h)P(h)$



Why might adults and children come to this word learning task with different priors?

Coming up next!

- This week's lab: a simple Bayesian model of word learning
  - Basic framework for Bayesian models
  - Play around with suspicious coincidences, the prior
- Next week: a Bayesian model of frequency learning
  - No pre-reading for lecture 3: catch up on the intro to probabilities and Bayes set for today…
  - …or read Xu & Tenenbaum (2007), it's very rich

References

Landau, B., Smith, L. B., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 5,* 287–312.

Macnamara, J. (1972). The cognitive basis of language learning in infants. *Psychological Review, 79,* 1–13.

Markman, E. M. (1989). *Categorization and naming in children.* Cambridge, MA: MIT Press.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20,* 121–157.

Quine, W. V. O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Xu, F., & Tenenbaum, J. B. (2007) Word learning as Bayesian Inference. *Psychological Review, 114,* 245-272.