What is Cognitive Science?

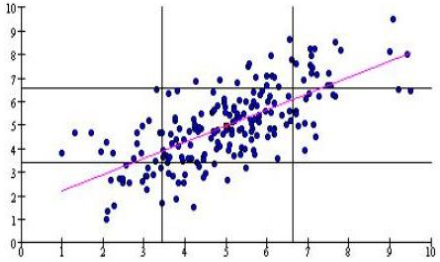
Josh Tenenbaum

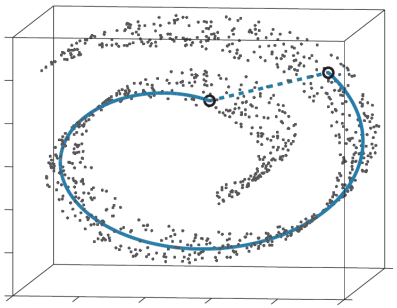
MLSS 2010

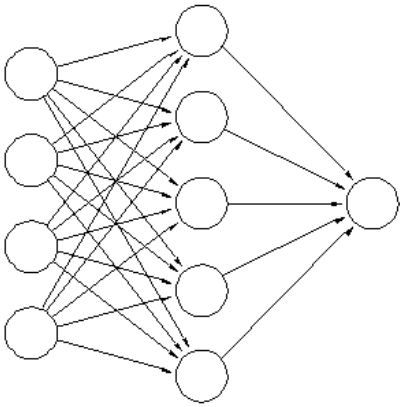
Pscyhology/CogSci and machine learning: a long-term relationship

- Unsupervised learning
  - Factor analysis
  - Multidimensional scaling
  - Mixture models (finite and infinite) for classification
  - Spectral clustering
  - Topic modeling by factorizing document-word count matrices
  - “Collaborative filtering” with low-rank factorizations
  - Nonlinear manifold learning with graph-based approximations
- Supervised learning
  - Perceptrons
  - Multi-layer perceptrons (“backpropagation”)
  - Kernel-based classification
  - Bayesian concept learning
- Reinforcement learning
  - Temporal difference learning

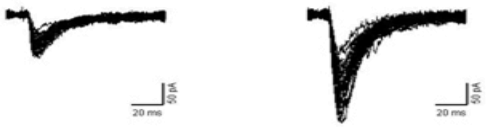
A success story in the 1980s-1990s: The "standard model of learning"







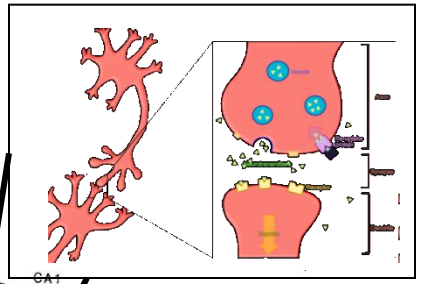
"Long term potentiation"

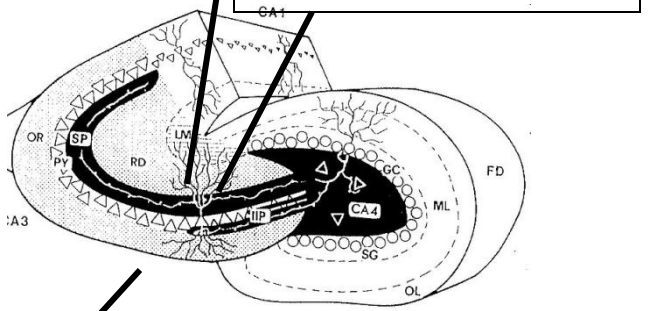


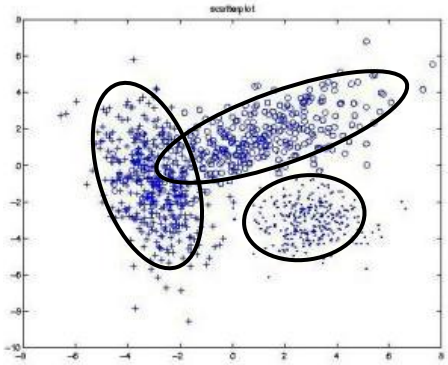
$$E = \frac{1}{2}\sum_{t=1}^{n} y_t^{\,2} = \frac{1}{2}\sum_{t=1}^{n} (w \cdot x_t)^2$$

$$\Delta w \propto -\frac{\partial E}{\partial w} = \sum_{t=1}^{n} y_t \cdot x_t$$

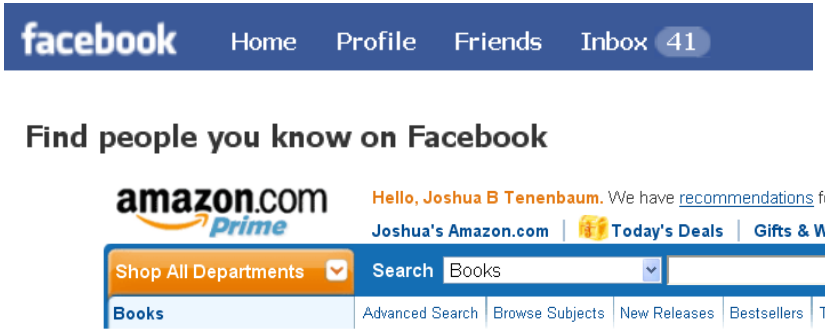
"Hebb rule"

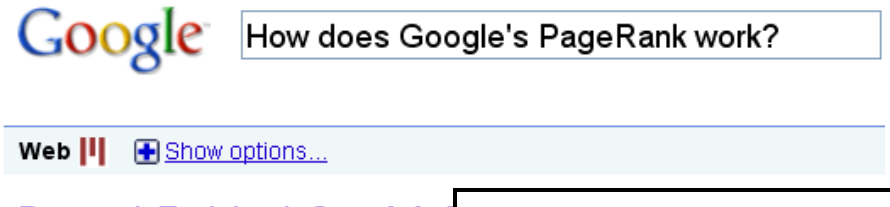


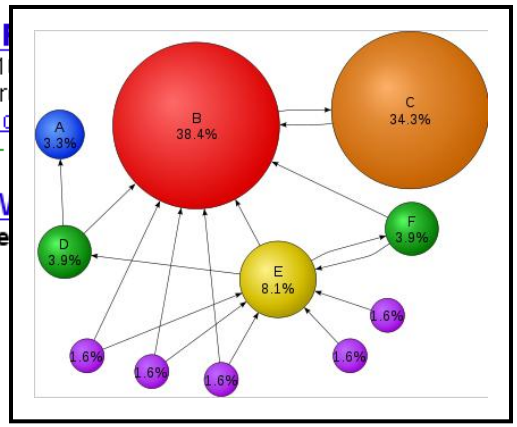


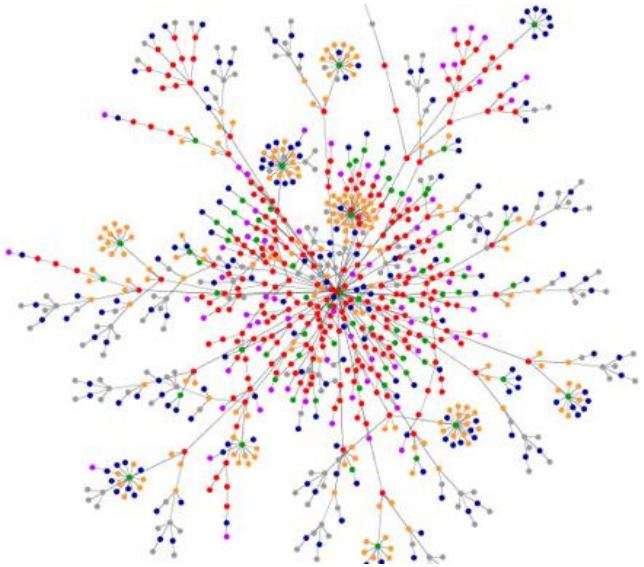


A success story in the 1980s-1990s: The "standard model of learning"









Outline

- The big problems of cognitive science.
- How machine learning can help.
- A brief introduction to cognition viewed through the lens of statistical inference and learning.

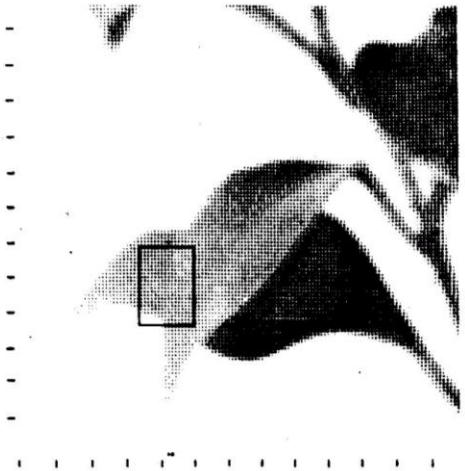
The big question

How does the mind get so much out of so little?

Our minds build rich models of the world and make strong generalizations from input data that is sparse, noisy, and ambiguous – in many ways far too limited to support the inferences we make.

How do we do it?

Visual perception

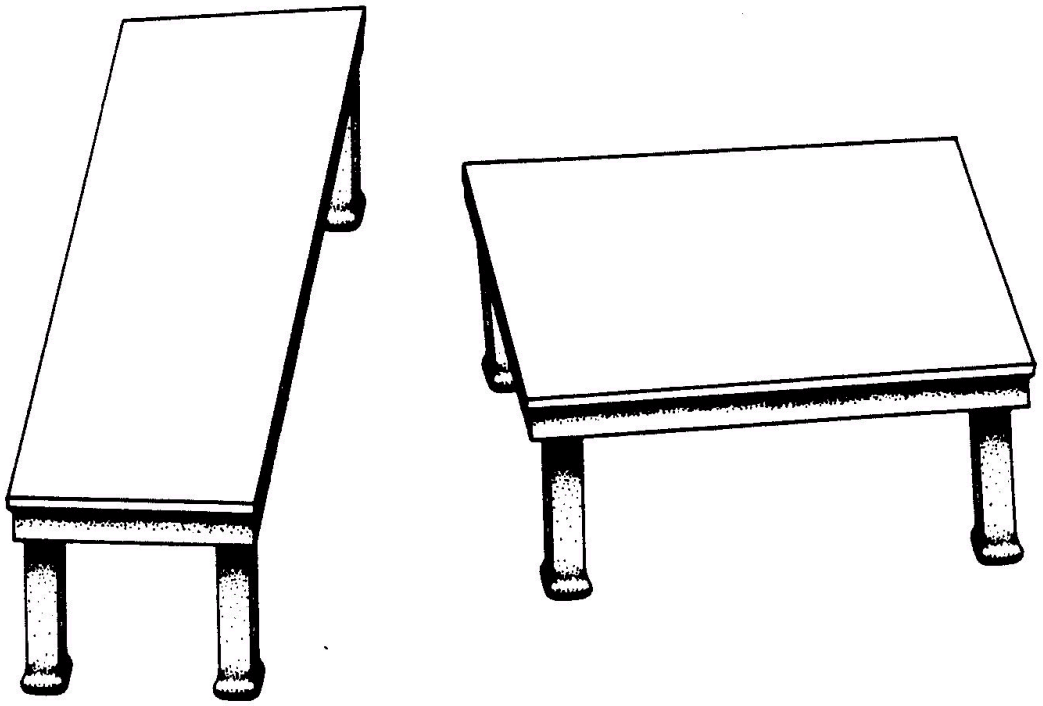


X =	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
Y																
58	171	169	167	167	166	165	166	164	167	171	171	174	174	175	173	171
57	168	168	168	167	166	167	167	165	169	168	174	176	175	175	175	172
56	168	167	167	165	166	166	167	167	168	170	178	177	176	174	174	173
55	168	168	165	169	167	168	167	165	168	175	177	177	175	175	172	171
54	169	170	167	169	169	168	163	166	172	169	174	173	175	178	173	173
53	171	169	170	168	169	168	169	168	168	170	175	173	175	177	178	176
52	172	171	170	168	169	169	167	168	173	172	173	177	174	175	178	176
51	172	174	171	170	166	168	167	168	172	172	172	177	179	172	175	175
50	171	167	176	169	170	169	168	169	171	172	174	174	173	173	174	178
49	174	172	173	173	173	174	171	171	172	174	172	172	172	169	173	173
48	173	173	173	176	178	172	171	174	174	173	175	175	175	173	173	171
47	173	175	178	173	173	171	171	175	175	177	178	175	174	173	175	178
46	178	175	174	169	173	175	177	175	177	177	174	175	176	177	177	174
45	173	175	173	174	172	173	174	175	174	171	173	174	175	174	172	171
44	177	174	175	175	172	171	172	176	172	173	172	172	173	170	170	175
43	173	171	174	168	176	172	173	173	173	174	171	174	175	173	174	174
42	175	173	171	172	170	171	176	175	178	172	174	175	175	175	175	172
41	181	179	177	172	170	170	169	179	175	174	175	174	172	175	174	175
40	188	184	179	178	176	176	176	174	172	178	172	174	173	172	174	173
39	195	191	188	186	185	183	180	177	178	175	174	176	175	174	176	176
38	200	199	197	193	190	187	185	180	176	175	180	177	175	175	176	177
37	202	202	199	202	199	194	187	180	175	179	177	176	174	175	176	173

(Marr)

Ambiguity in visual perception

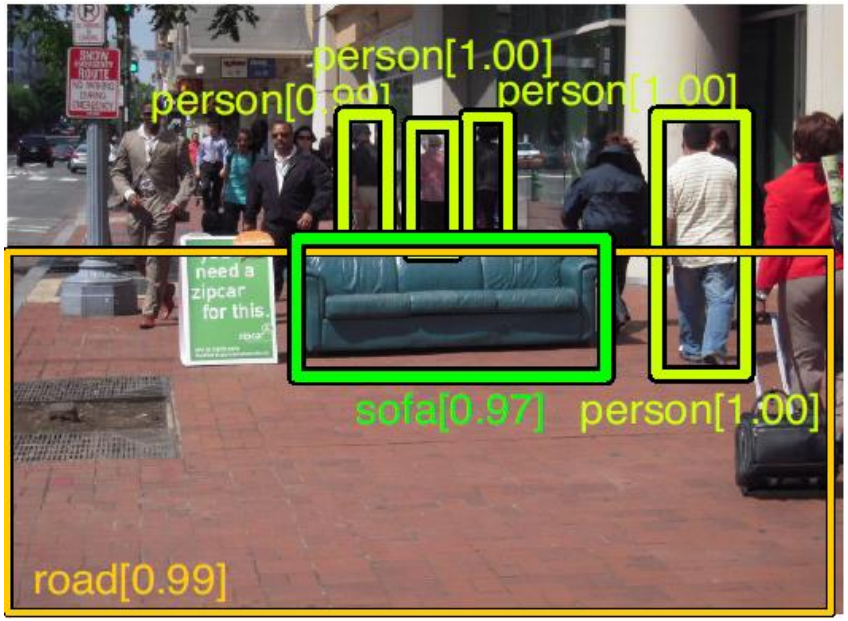
- Goal of visual perception is to recover world structure from visual images.
- Why the problem is hard: many world structures can produce the same visual input.
- Illusions reveal the visual system's implicit knowledge of the physical world and the processes of image formation.



(Shepard)

Learning-based machine vision: state of the art

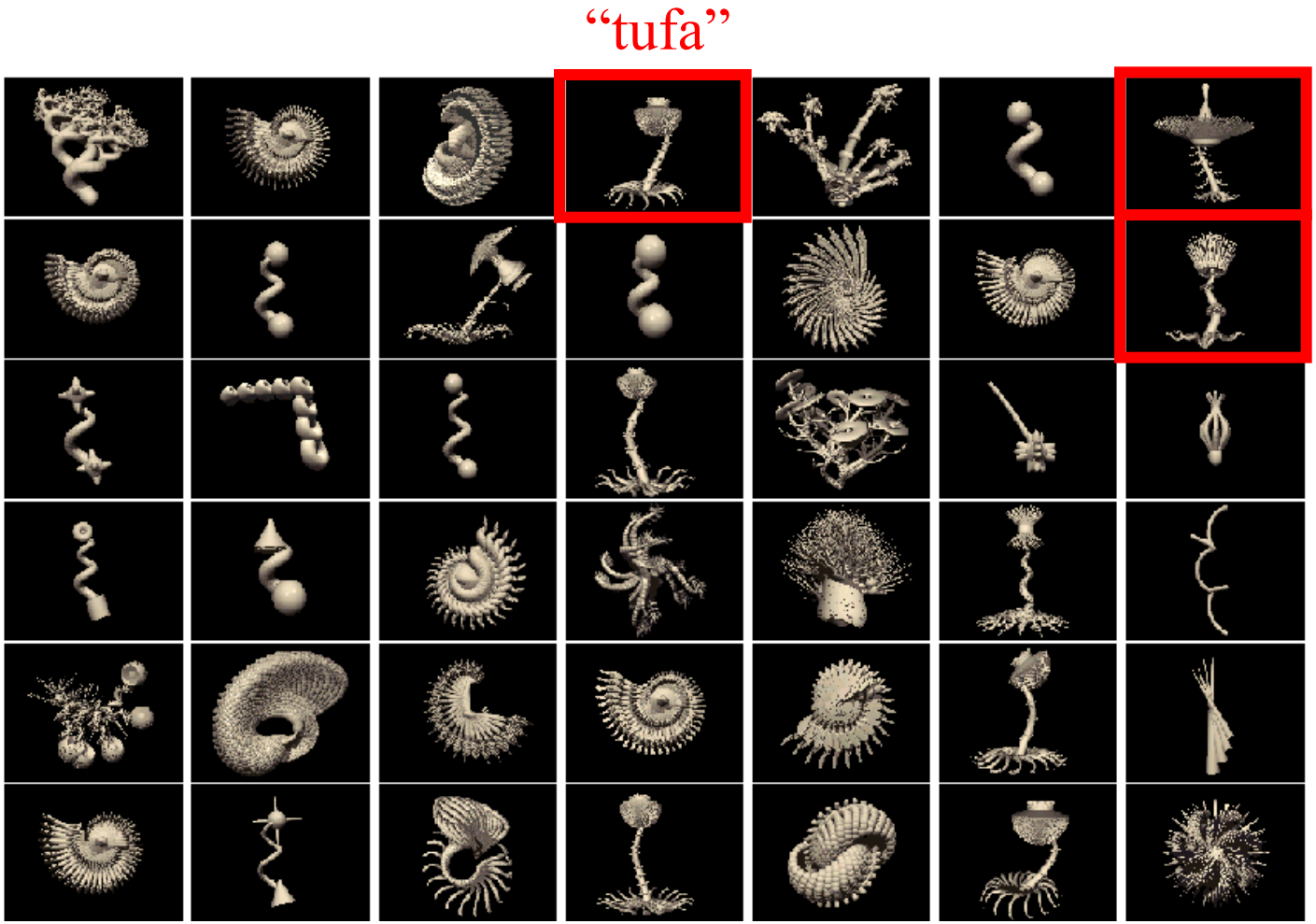
Input



Output

(Choi, Lim, Torralba, Willsky)

Learning concepts from examples



Humans and bumble bees

"According to the theory of aerodynamics, a bumble bee can't fly."

According to statistical learning theory, a person can't learn a concept from just one or a few positive examples…

Causal inference

	cold ≤ 1 week	cold > 1 week
Took drug	5	1
Didn't take drug	2	6

Does this drug help you get over a cold faster?

Causal inference

	Got burned	Didn't get burned
Touched stove	5	1
Didn't touch stove	2	6

How does a child learn not to touch a hot stove? (c.f. Hume)

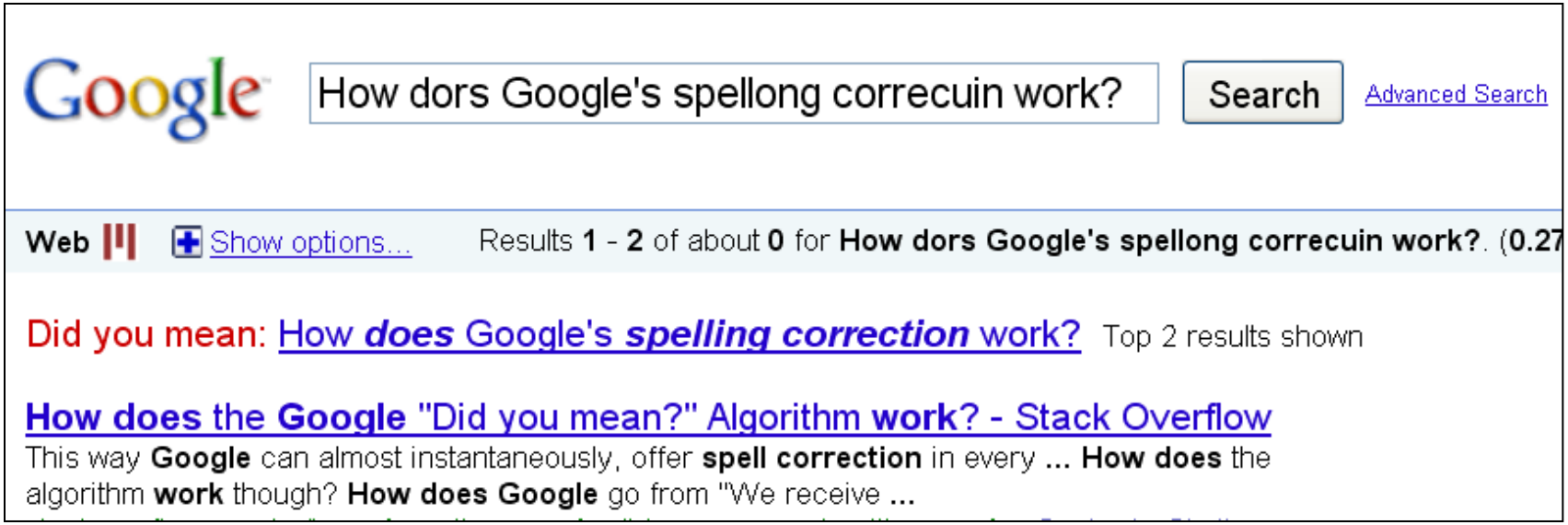
What happens if I press this button over here on the wall …?

Language

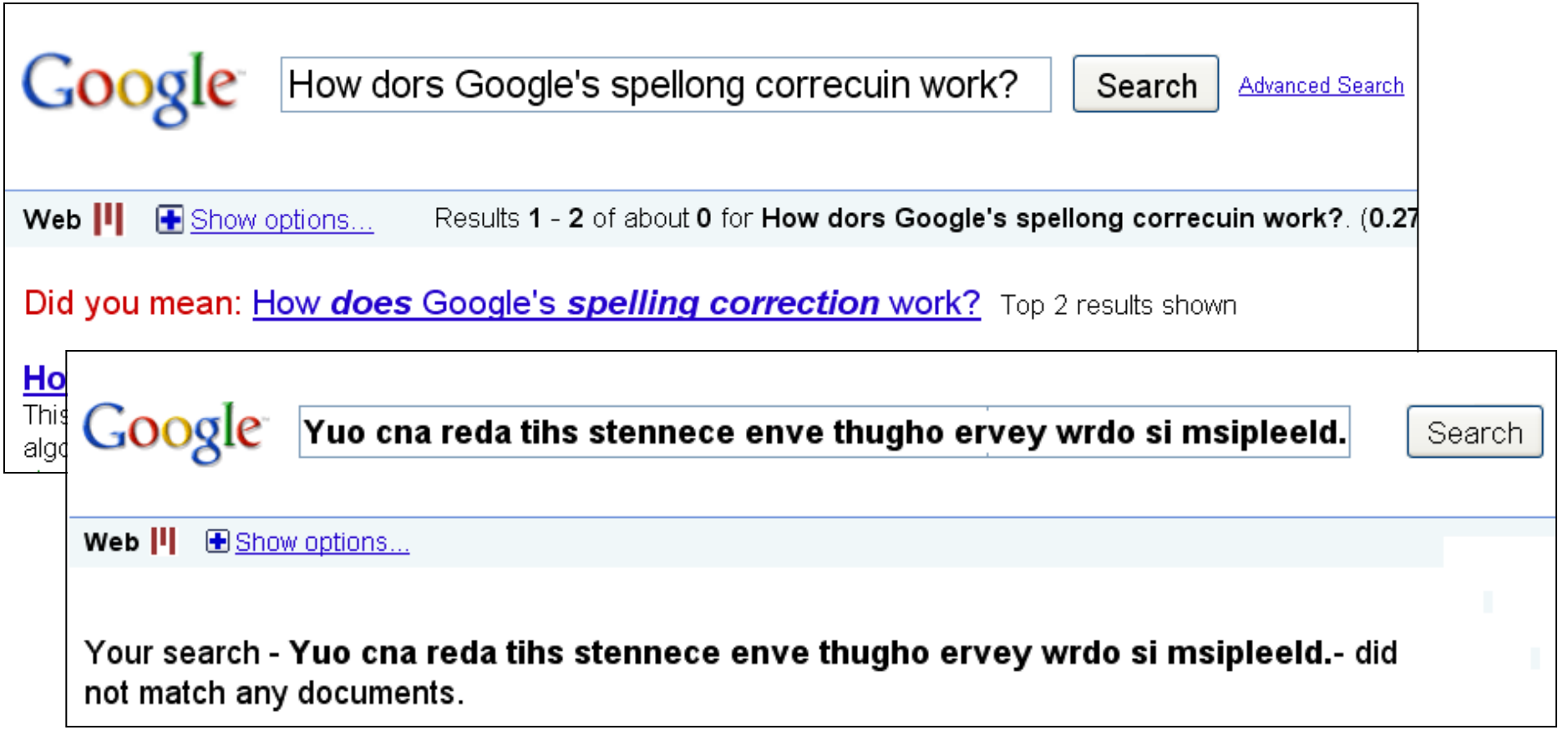
- Parsing:
  - The girl saw the boy with the telescope.
  - Two cars were reported stolen by the Groveton police yesterday.
  - The judge sentenced the killer to die in the electric chair for the second time.
  - No one was injured in the blast, which was attributed to a buildup of gas by one town official.
  - One witness told the commissioners that she had seen sexual intercourse taking place between two parked cars in front of her house.

(Pinker)

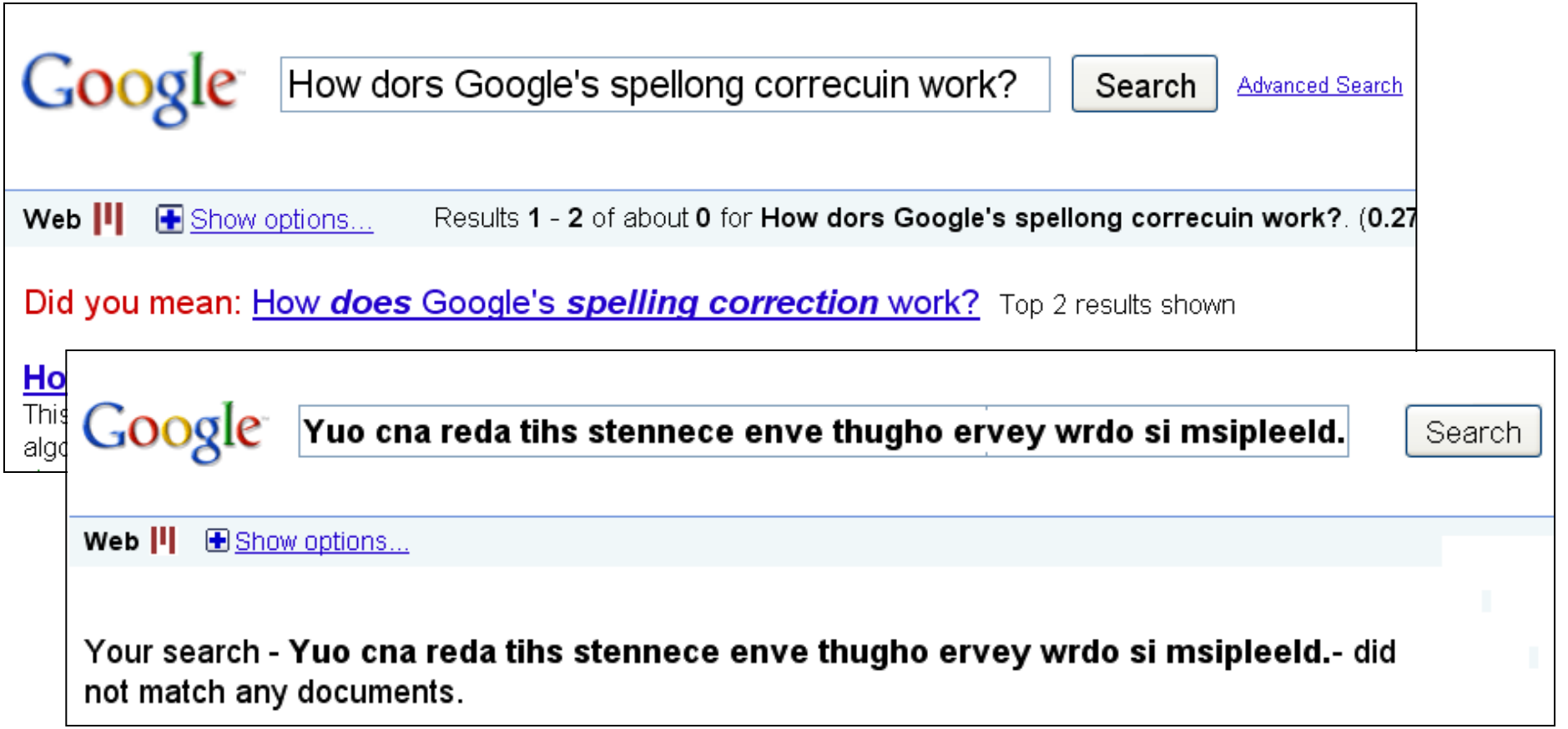
Language


Google
How dors Google's spellong correcuin work?
Search
Advanced Search
Web
Show options...
Results 1 - 2 of about 0 for How dors Google's spellong correcuin work?. (0.27
Did you mean: How does Google's spelling correction work?
Top 2 results shown
How does the Google "Did you mean?" Algorithm work? - Stack Overflow
This way Google can almost instantaneously, offer spell correction in every ... How does the algorithm work though? How does Google go from "We receive ...

Language


Google
How dors Google's spellong correcuin work?
Search
Advanced Search
Web
Show options...
Results 1 - 2 of about 0 for How dors Google's spellong correcuin work?. (0.27
Did you mean: How does Google's spelling correction work?
Top 2 results shown
Google
Yuo cna reda tihs stennece enve thugho ervey wrdo si msipleeld.
Search
Web
Show options...
Your search - Yuo cna reda tihs stennece enve thugho ervey wrdo si msipleeld.- did not match any documents.

Language

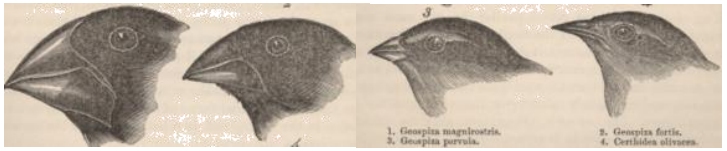


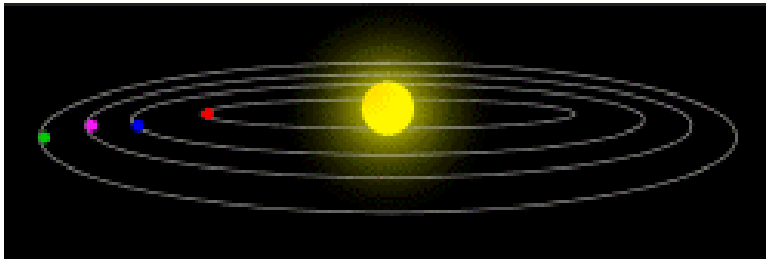
Ervey tihs si yuo enve msipleeld thugho wrdo cna stennece reda.

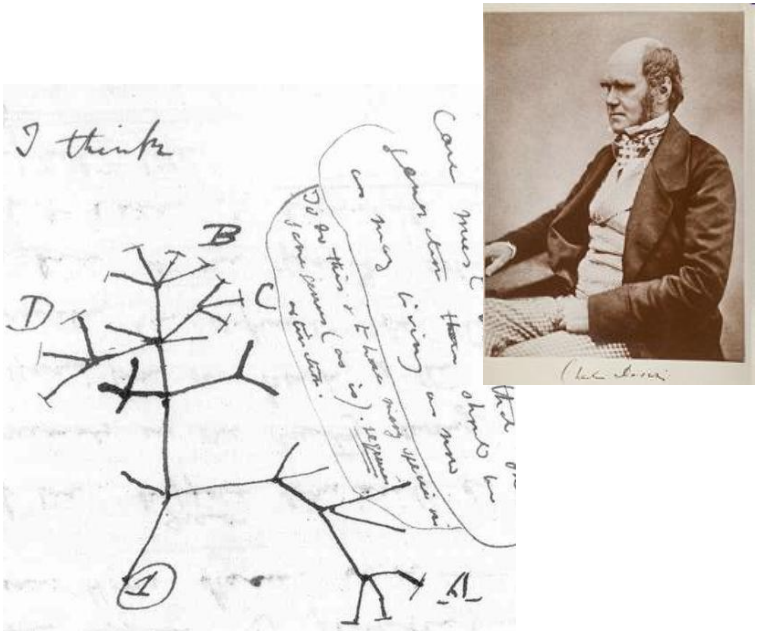
Language

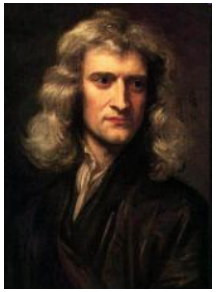
- Parsing
- Acquisition:
  - Learning verb forms
    - English past tense: rule vs. exceptions
    - Spanish or Arabic past tense: multiple rules plus exceptions
  - Learning verb argument structure
    - e.g., “give” vs. “donate”, “fill” vs. “load”
  - Learning to be bilingual

Theory construction in science

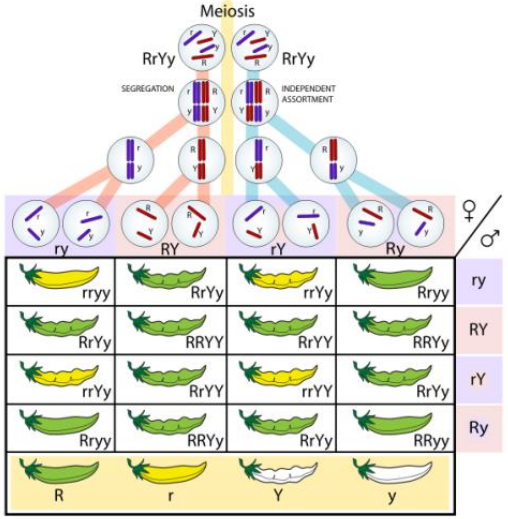






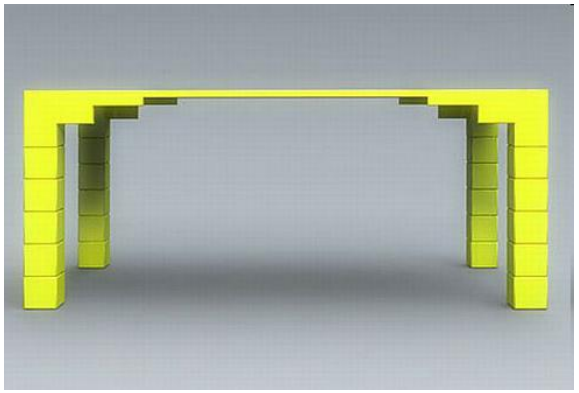


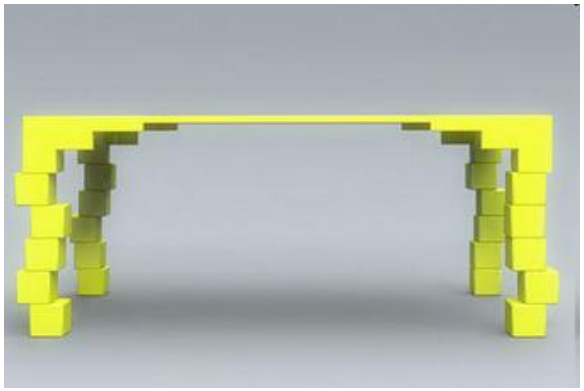
$$F = \frac{GMm}{r^2}$$



Intuitive theories

- Physics
  - Parsing: Inferring support relations, or the causal history and properties of an object.





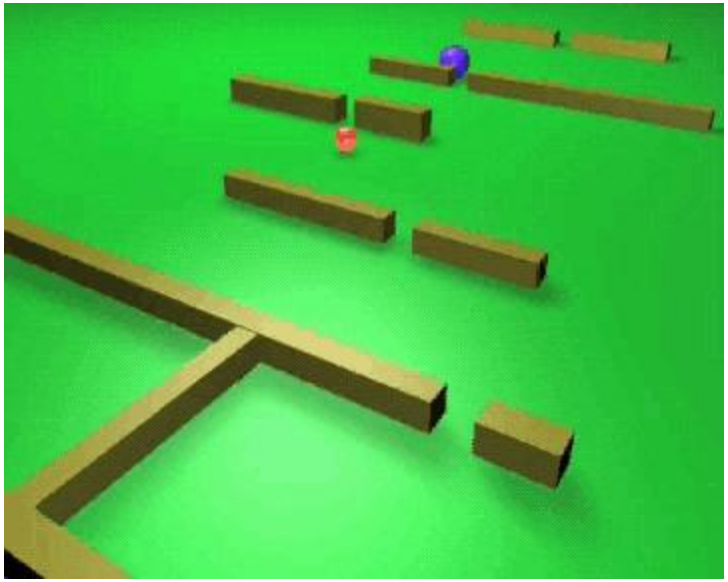
Intuitive theories

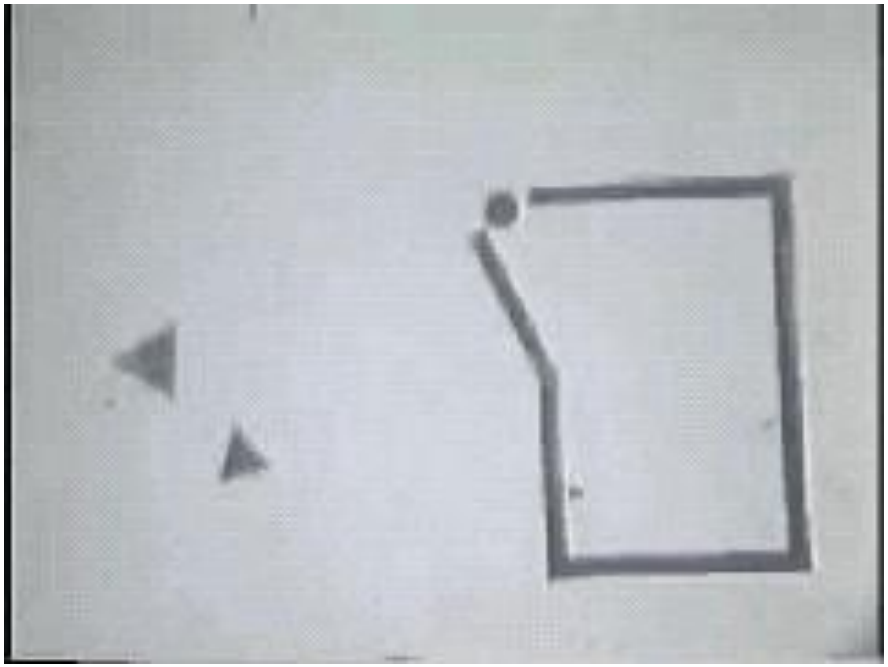
- Physics
  - Parsing: Inferring support relations, or the causal history and properties of an object.

Intuitive theories

- Physics
  - Parsing: Inferring support relations, or the causal history and properties of an object.
  - Acquisition: Learning about gravity and support.
    - Gravity -- what's that?
    - Contact is sufficient
    - Mass distribution and location is important
- A different intuitive theory…

  Two Demos.





"If you have a mate, and there is a rival, go and peck that rival…"

Intuitive theories

- Physics
  - Parsing: Inferring support relations, or the causal history and properties of an object.
  - Acquisition: Learning about gravity and support.
    - Gravity -- what's that?
    - Contact is sufficient
    - Mass distribution and location is important
- Psychology
  - Parsing: Inferring beliefs, desires, plans.
  - Acquisition: Learning about agents.
    - Recognizing intentionality, but without mental state reasoning
    - Reasoning about beliefs and desires
    - Reasoning about plans, rationality and "other minds".

Outline

- The big problems of cognitive science.
- How machine learning can help.
- A brief introduction to cognition viewed through the lens of statistical inference and learning.

The big questions

1. How does knowledge guide inductive learning, inference, and decision-making from sparse, noisy or ambiguous data?
2. What are the forms and contents of our knowledge of the world?
3. How is that knowledge itself learned from experience?
4. How do we balance constraint and flexibility, assimilating new data to our current model versus accommodate our model to the new data?
5. How can accurate inductive inferences be made efficiently, even in the presence of complex hypothesis spaces?

Machine learning provides a toolkit for answering these questions

1. Bayesian inference in probabilistic generative models
2. Probabilities defined over structured representations: graphs, grammars, predicate logic, programs
3. Hierarchical probabilistic models, with inference at all levels of abstraction
4. Adaptive nonparametric or “infinite” models, which can grow in complexity or change form in response to the observed data.
5. Approximate methods of learning and inference, e.g., Markov chain Monte Carlo (MCMC), importance sampling, and sequential importance sampling (particle filtering).

Basics of Bayesian inference

- Bayes' rule: $P(h \mid d) = \dfrac{P(d \mid h)P(h)}{\sum_{h_i \in H} P(d \mid h_i)P(h_i)}$
- An example
  - Data: John is coughing
  - Some hypotheses:
    1. John has a cold
    2. John has lung cancer
    3. John has a stomach flu
  - Likelihood $P(d|h)$ favors 1 and 2 over 3
  - Prior probability $P(h)$ favors 1 and 3 over 2
  - Posterior probability $P(h|d)$ favors 1 over 2 and 3

Grammar G

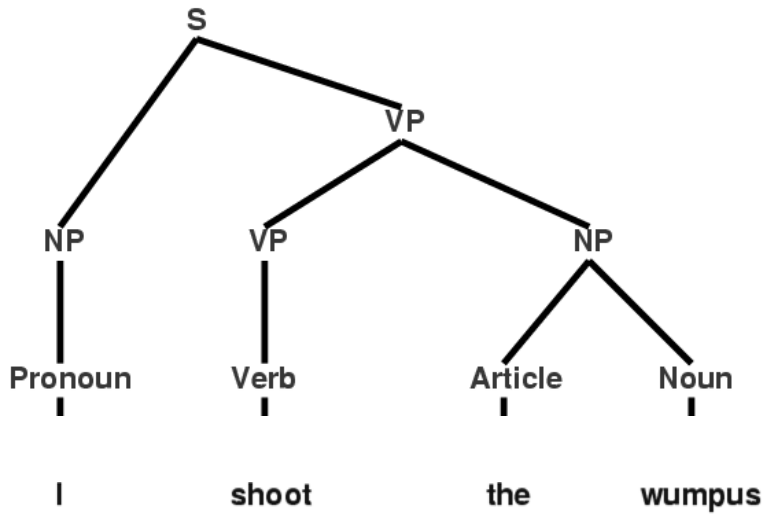
$P(S \mid G)$

Phrase structure S

$P(U \mid S)$

Utterance U

$$S \rightarrow NP\ VP$$
$$NP \rightarrow Det\,[Adj]\ Noun[RelClause]$$
$$RelClause \rightarrow [Rel]\ NP\ V$$
$$VP \rightarrow VP\ NP$$
$$VP \rightarrow Verb$$



$$P(S \mid U, G) \sim P(U \mid S) \times P(S \mid G)$$

Bottom-up Top-down

"Universal Grammar"

↓ *P*(grammar | UG)

Grammar

↓ *P*(phrase structure | grammar)

Phrase structure

↓ *P*(utterance | phrase structure)

Utterance

↓ *P*(speech | utterance)

Speech signal

Hierarchical phrase structure grammars (e.g., CFG, HPSG, TAG)

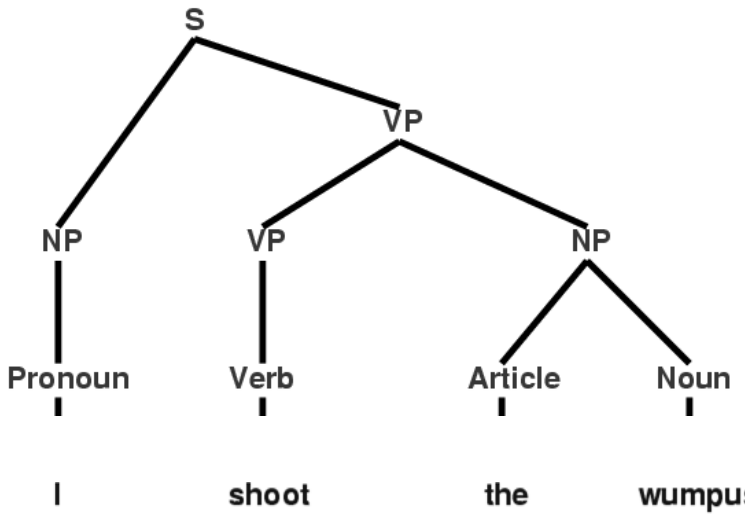
$$S \rightarrow NP\ VP$$

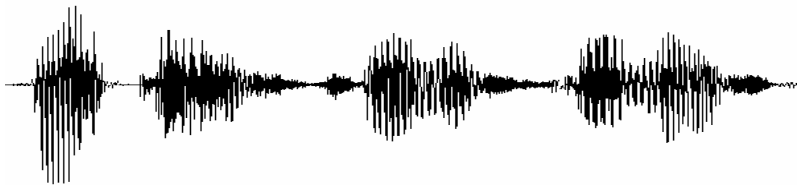
$$NP \rightarrow Det\,[Adj]\ Noun\,[RelClause]$$

$$RelClause \rightarrow [Rel]\ NP\ V$$

$$VP \rightarrow VP\ NP$$

$$VP \rightarrow Verb$$





“Universal Grammar”

↓ *P*(grammar | UG)

Grammar

↓ *P*(parsing graph | grammar)

Parsing graph

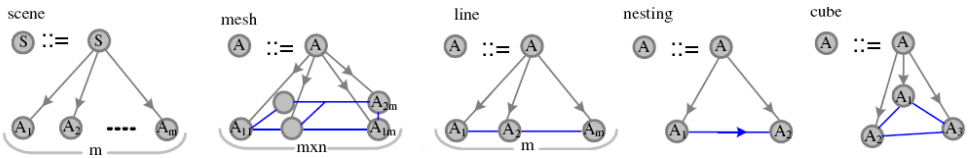
↓ *P*(surfaces | parsing graph)

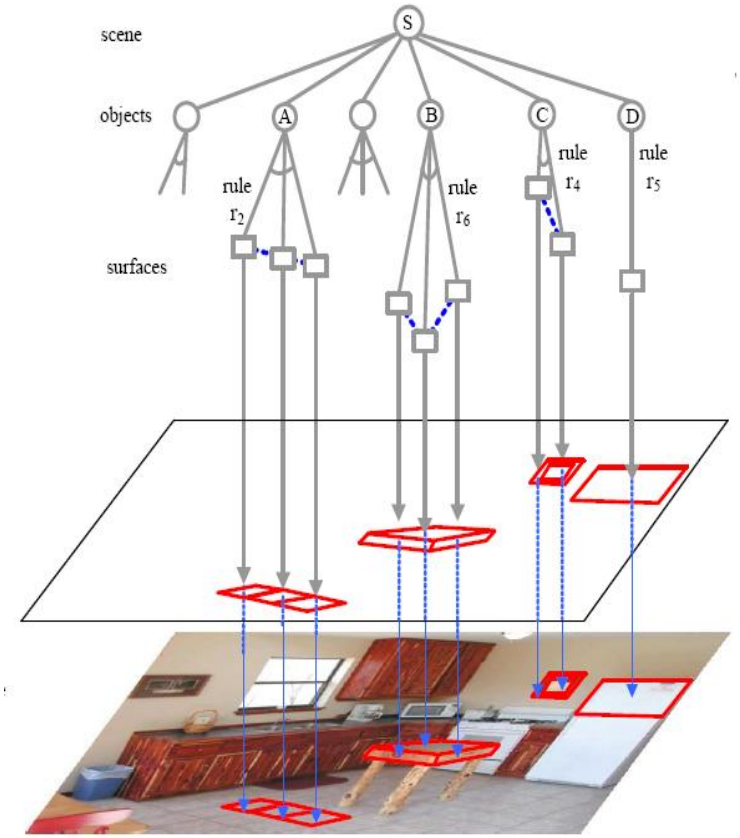
Surfaces

↓ *P*(image | surfaces)

Image

Compositional scene grammars
(e.g., attribute graph grammar, AND/OR grammar)



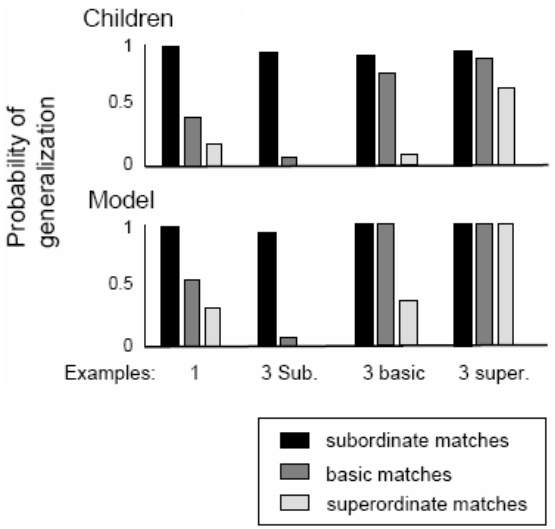


(Han & Zhu, 2006)

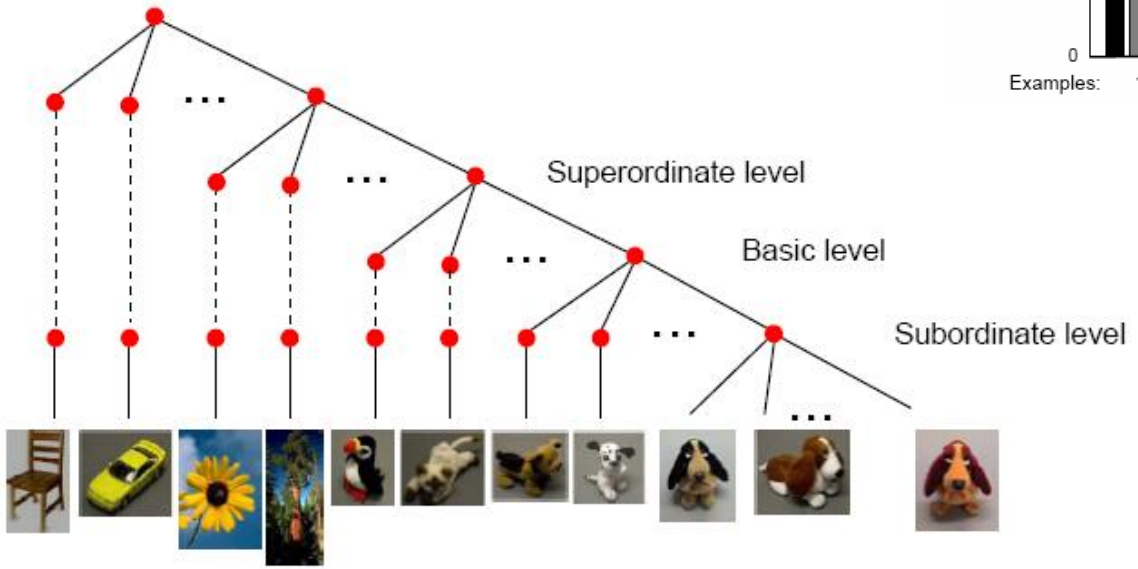
Learning word meanings

Principles

Whole-object principle
Shape bias
Taxonomic principle
Contrast principle
Basic-level bias



Structure



Data

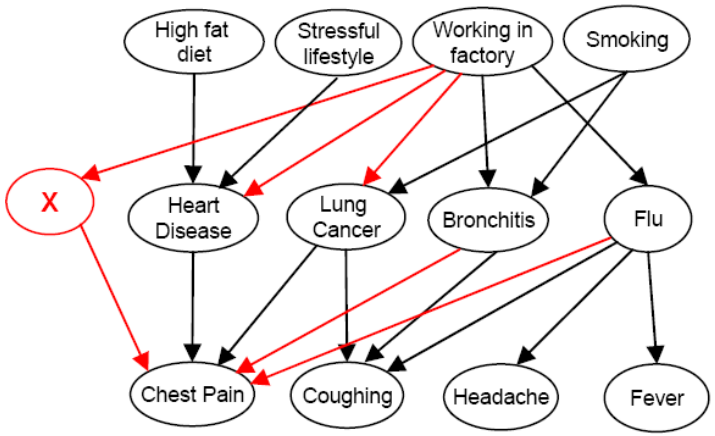
Causal learning and reasoning

Principles

Structure

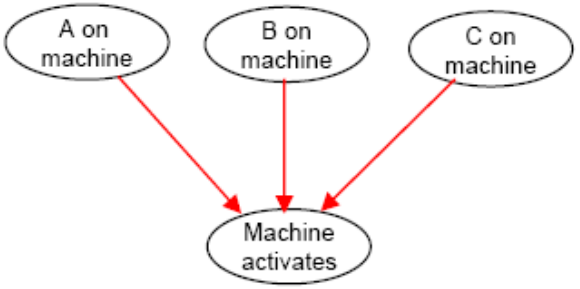
Data

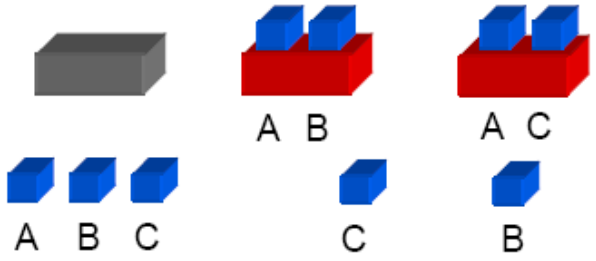
Classes: {R, D, S} (Risks, Diseases, Symptoms)
Causal laws: R → D, D → S



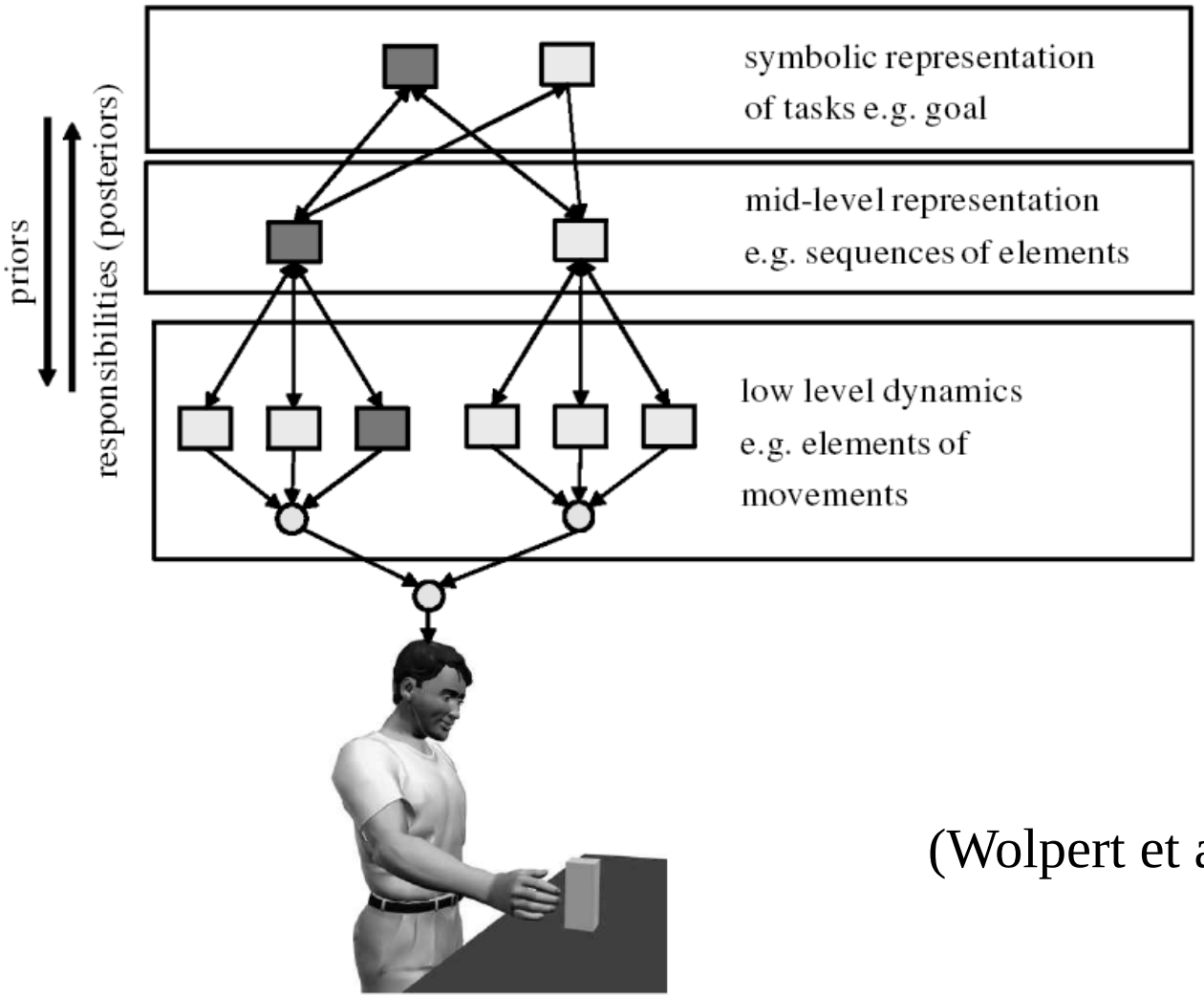
Patient 1: Stressful lifestyle
Chest Pain
Patient 2: Smoking
Coughing
Patient 3: Working in factory
Chest Pain
. . .

Objects can activate Machines
Activation requires contact
Machines are (near) deterministic





Goal-directed action (production and comprehension)



(Wolpert et al., 2003)

Bayes meets Marr: the Sampling Hypothesis

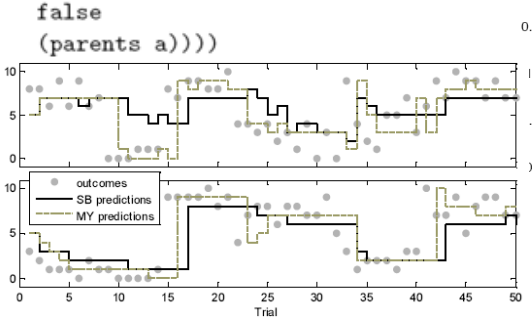
Marr's levels

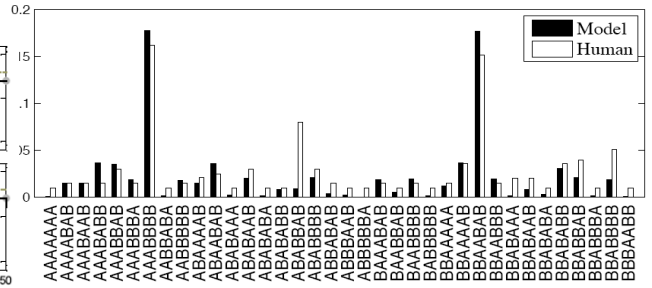
Computational

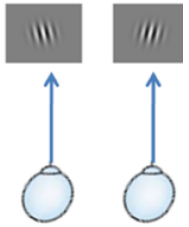
Algorithmic

Neural

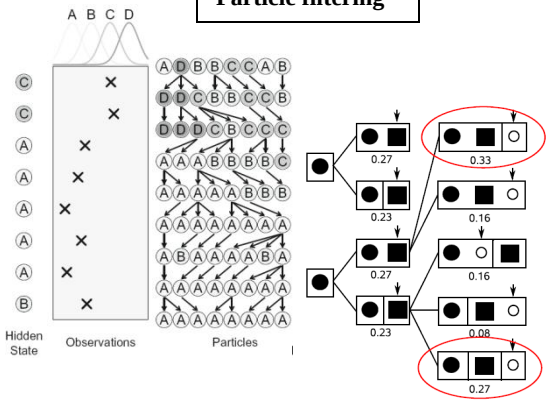
```
(define (occurs a t)
  (or (spontaneous a t)
      (do a t)
      (fold (lambda (x y) (noisy-or (occurs x t) (strength x a) y 1.0))
            false
            (parents a))))
```



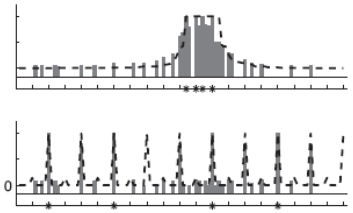


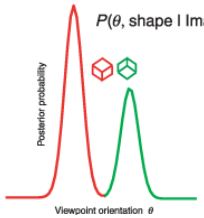


Particle filtering

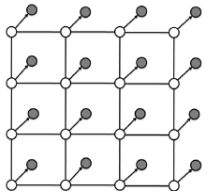


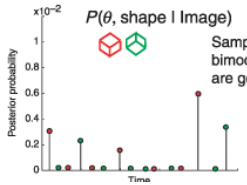
Importance sampling

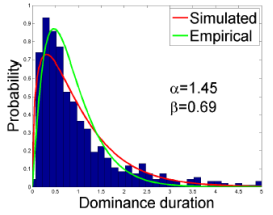


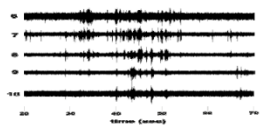


Markov Chain Monte Carlo (MCMC)



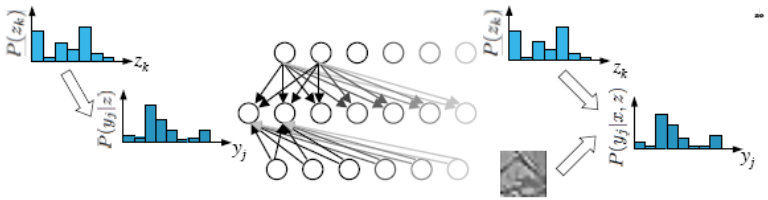


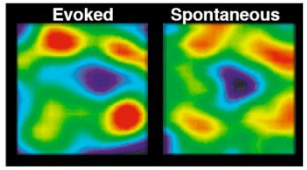


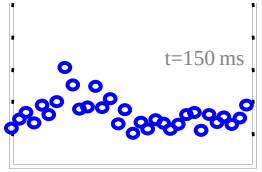


Spontaneous activity

Evoked activity







Outline

- The big problems of cognitive science.
- How machine learning can help.
- A *very* brief introduction to cognition viewed through the lens of statistical inference and learning.

Five big ideas

- Understanding human cognition as Bayesian inference over probabilistic generative models of the world.
- Building probabilistic models defined over structured knowledge representations, such as graphs, grammars, predicate logic, functional programs.
- Explaining the origins of knowledge by learning in hierarchical probabilistic models, with inference at multiple levels of abstraction.
- Balancing constraint with flexibility, via adaptive representations and nonparametric (“infinite”) models that grow in complexity or change form in response to the data.
- Tractable methods for approximate learning and inference that can react to new data in real time and scale up to large problems (e.g., Markov chain Monte Carlo, Sequential MC).

Cognition as probabilistic inference

Visual perception [Weiss, Simoncelli, Adelson, Richards, Freeman, Feldman, Kersten, Knill, Maloney, Olshausen, Jacobs, Pouget, ...]

Language acquisition and processing [Brent, de Marken, Niyogi, Klein, Manning, Jurafsky, Keller, Levy, Hale, Johnson, Griffiths, Perfors, Tenenbaum, …]

Motor learning and motor control [Ghahramani, Jordan, Wolpert, Kording, Kawato, Doya, Todorov, Shadmehr, …]

Associative learning [Dayan, Daw, Kakade, Courville, Touretzky, Kruschke, …]

Memory [Anderson, Schooler, Shiffrin, Steyvers, Griffiths, McClelland, …]

Attention [Mozer, Huber, Torralba, Oliva, Geisler, Yu, Itti, Baldi, …]

Categorization and concept learning [Anderson, Nosfosky, Rehder, Navarro, Griffiths, Feldman, Tenenbaum, Rosseel, Goodman, Kemp, Mansinghka, …]

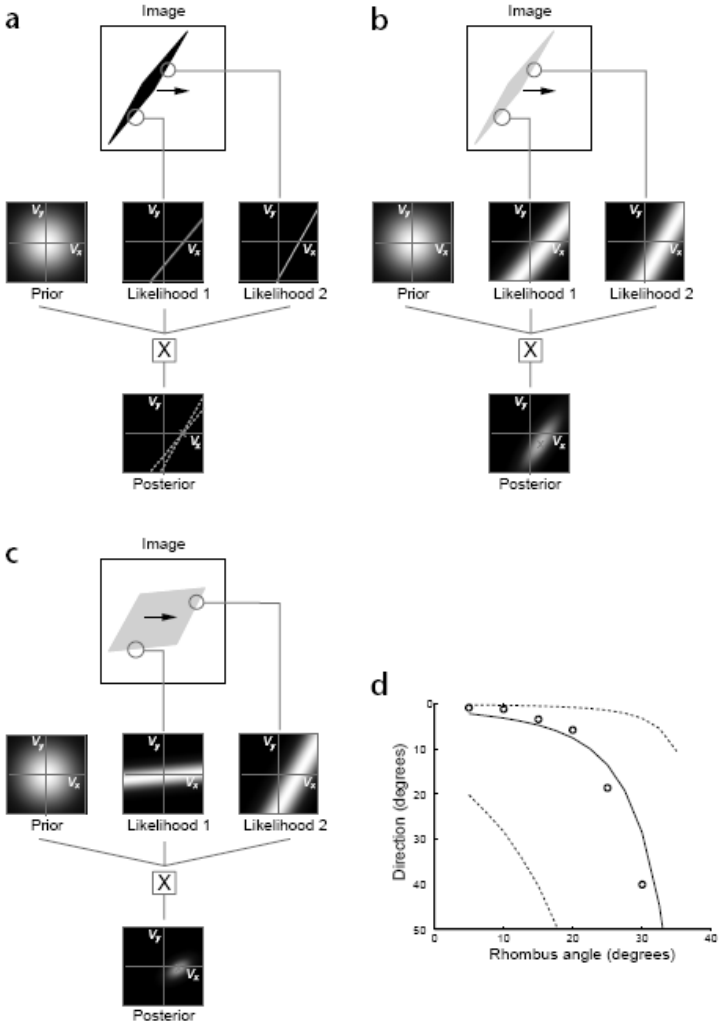
Reasoning [Chater, Oaksford, Sloman, McKenzie, Heit, Tenenbaum, Kemp, …]

Causal inference [Waldmann, Sloman, Steyvers, Griffiths, Tenenbaum, Yuille, …]

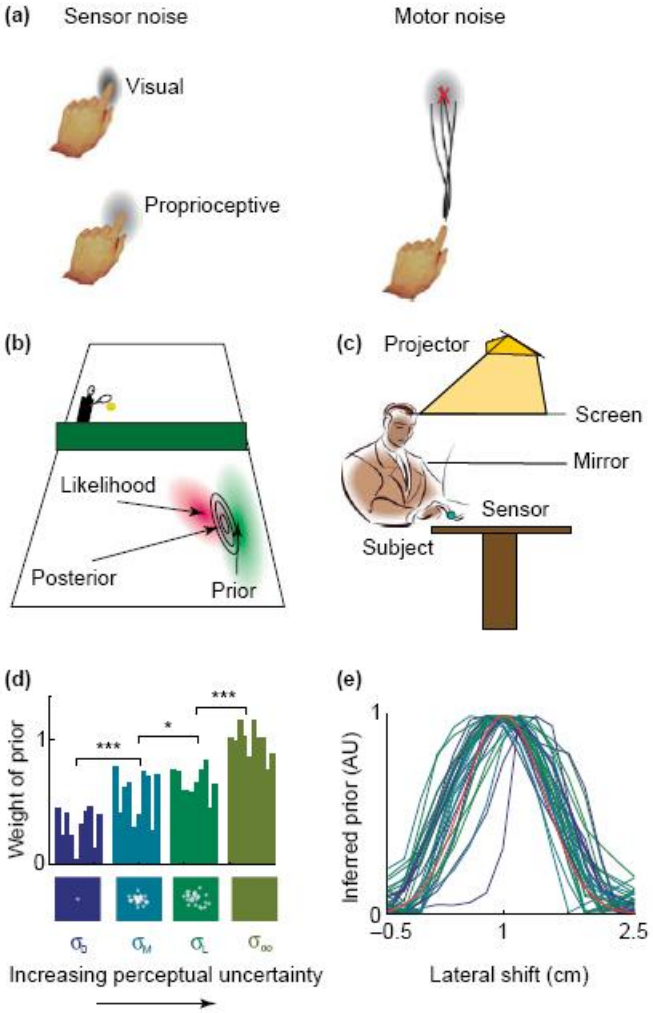
Decision making and theory of mind [Lee, Stankiewicz, Rao, Baker, Goodman, Tenenbaum, …]

Bayesian inference in perceptual and motor systems

Weiss, Simoncelli & Adelson (2002)

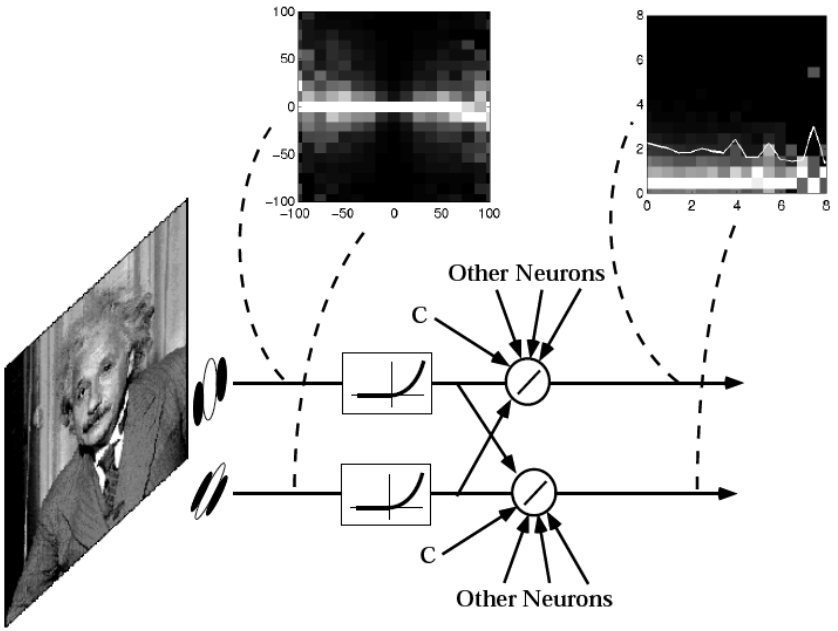


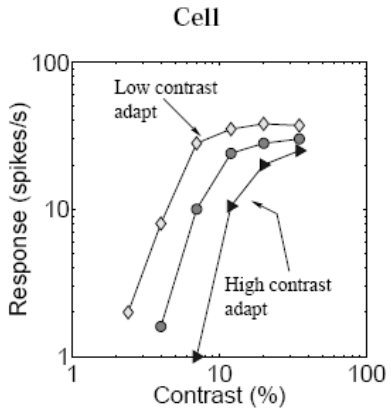
Kording & Wolpert (2004)

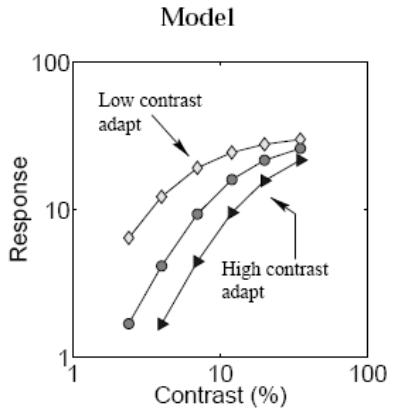


Bayesian ideal observers using natural scene statistics

Wainwright, Schwartz & Simoncelli (2002)







Does this approach extend to cognition?

Modeling basic cognitive capacities as intuitive Bayesian statistics

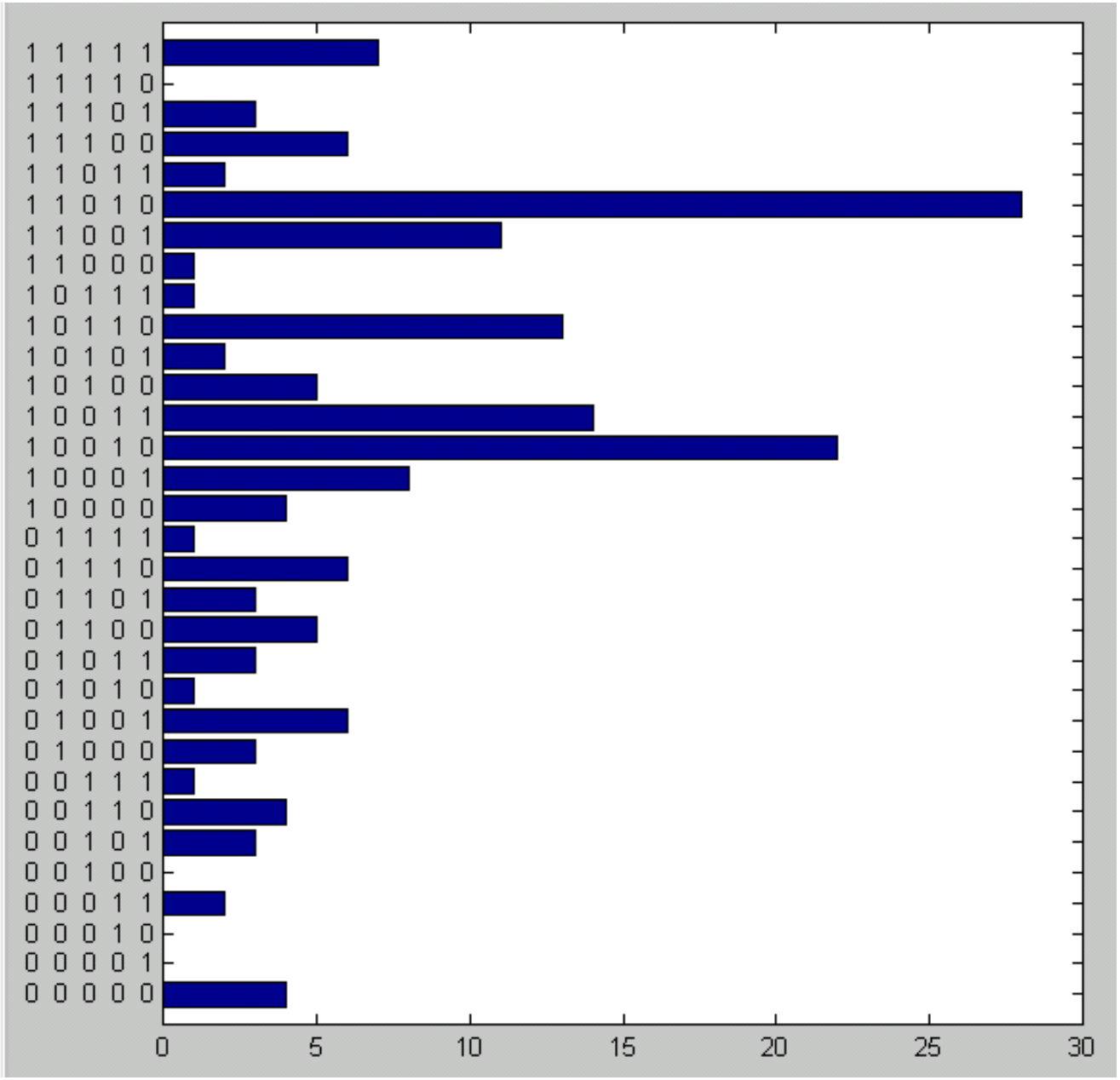
- Similarity (Tenenbaum & Griffiths, *BBS* 2001; Kemp & Tenenbaum, *Cog Sci* 2005)
- Representativeness and evidential support (Tenenbaum & Griffiths, *Cog Sci* 2001)
- Causal judgment (Steyvers et al., 2003; Griffiths & Tenenbaum, *Cog. Psych.* 2005)
- Coincidences and causal discovery (Griffiths & Tenenbaum, *Cog Sci* 2001; *Cognition* 2007; *Psych. Review*, in press)
- Diagnostic inference (Krynski & Tenenbaum, *JEP: General* 2007)
- Predicting the future (Griffiths & Tenenbaum, *Psych. Science* 2006)

Coin flipping

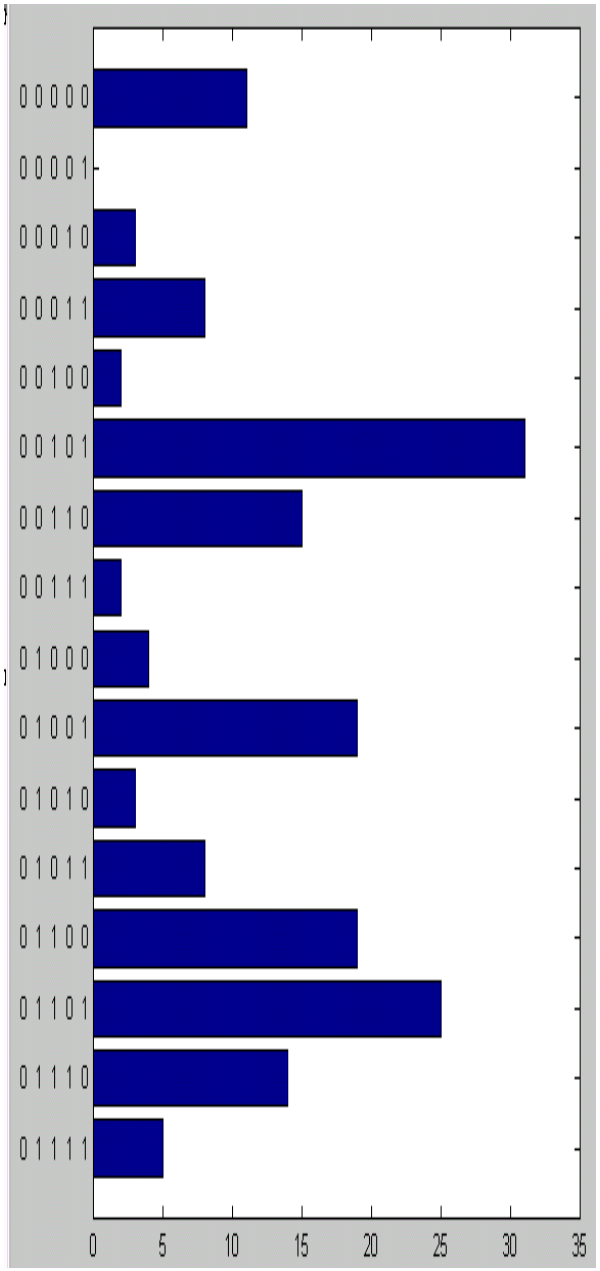
Which sequence is more likely to be produced by flipping a fair coin?

HHTHT $P(\text{HHTHT} \mid \text{fair coin}) = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$

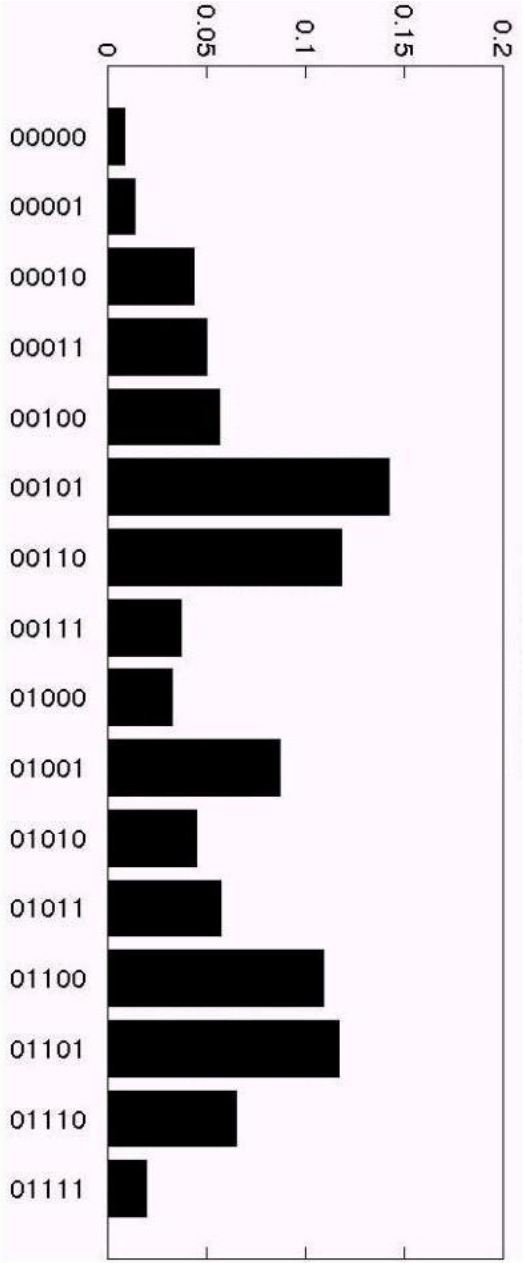
HHHHH $P(\text{HHHHH} \mid \text{fair coin}) = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$



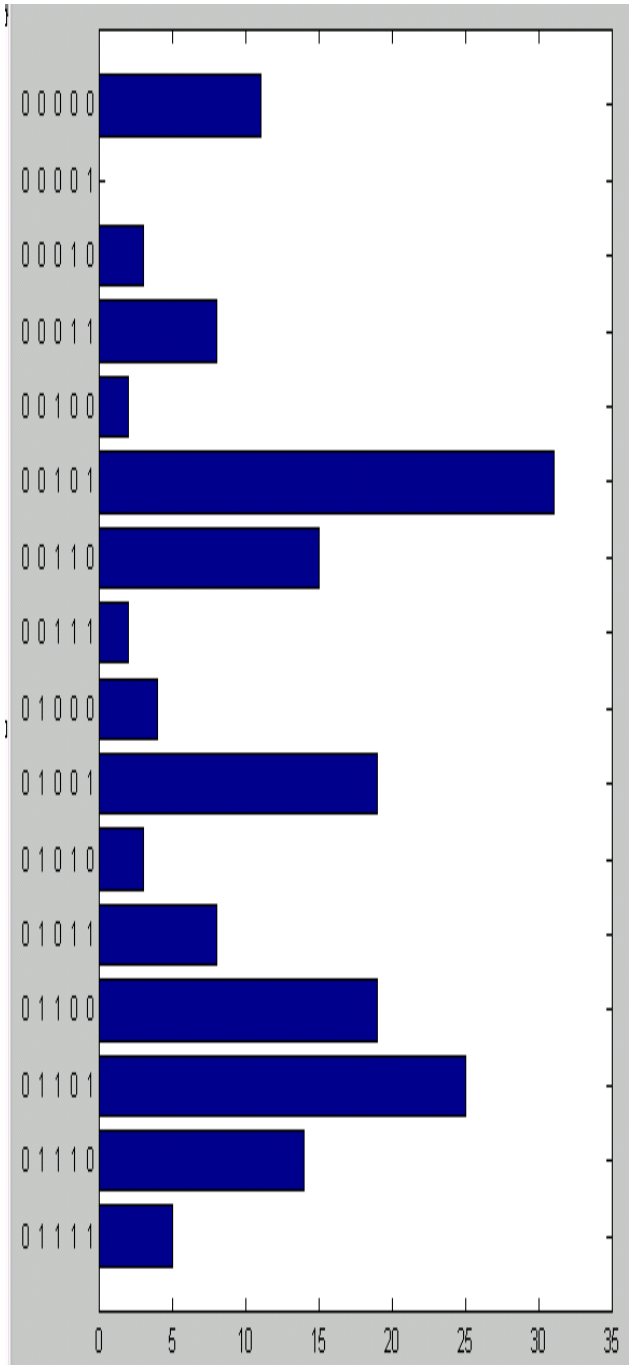
Predict a random sequence of coin flips: Mathcamp 2001, 2003

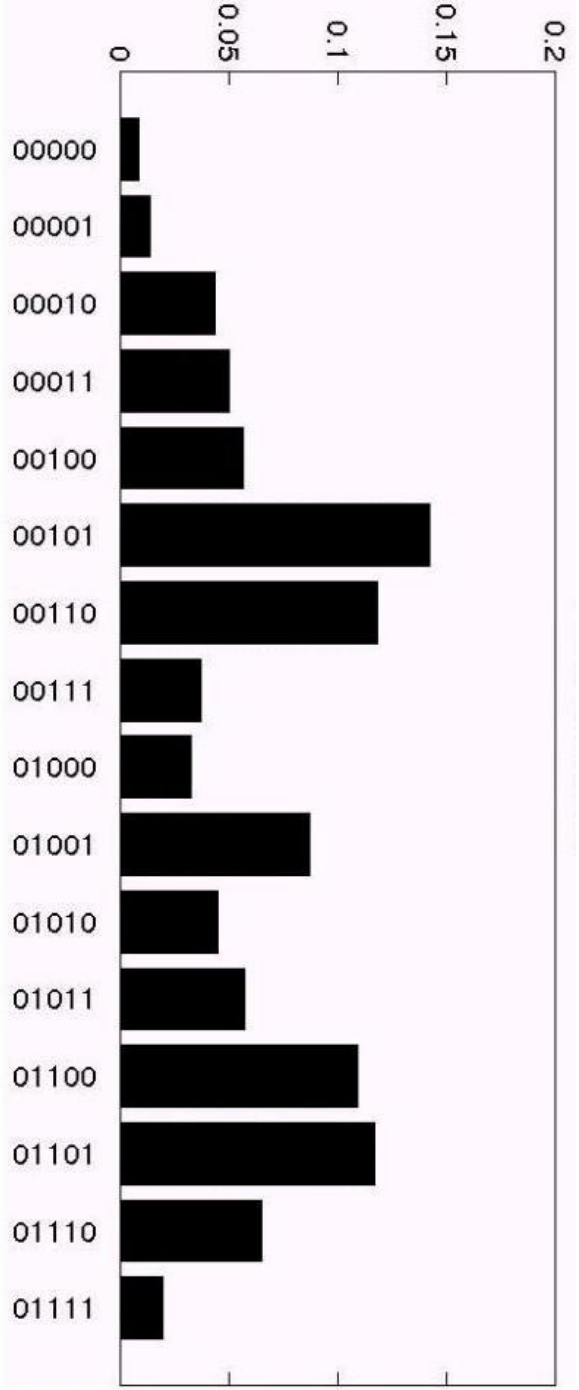


Mathcamp 2001, 2003 data: collapsed over parity



Zenith radio data (1930's): collapsed over parity


00000
00001
00010
00011
00100
00101
00110
00111
01000
01001
01010
01011
01100
01101
01110
01111
0
5
10
15
20
25
30
35


0
0.05
0.1
0.15
0.2
00000
00001
00010
00011
00100
00101
00110
00111
01000
01001
01010
01011
01100
01101
01110
01111

Coin flipping

Why do some sequences *appear* much more likely to be produced by flipping a fair coin?

HHTHT

HHHHH

"We can introspect about the outputs of cognition, not the processes or the intermediate representations of the computations."

Predictive versus inductive reasoning

Prediction

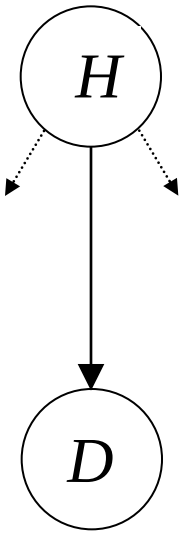
given

|

Likelihood: $P(D \mid H)$

↓

?



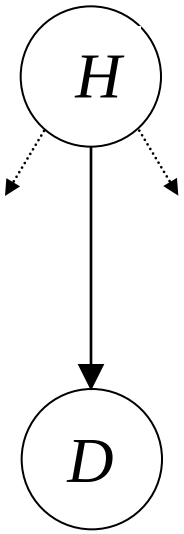
Predictive versus inductive reasoning

Prediction

given

Likelihood: $P(D \mid H)$

?



Induction

?

Likelihood ratio: $\frac{P(D \mid H_1)}{P(D \mid H_2)}$

given

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

Comparing two hypotheses

- Different patterns of observed data:
  - D = `HHTHT` or `HHHHH`
- Contrast simple hypotheses:
  - H_1: "fair coin", $P(\texttt{H}) = 0.5$
  - H_2:"always heads", $P(\texttt{H}) = 1.0$
- Bayes' rule in odds form:

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \times \frac{P(H_1)}{P(H_2)}$$

Comparing two hypotheses

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D: HHTHT

H_1, H_2: "fair coin", "always heads"

$P(D|H_1) = 1/2^5$ $\quad P(H_1) =$?

$P(D|H_2) = 0$ $\quad P(H_2) =$ 1-?

Comparing two hypotheses

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D: HHTHT

H_1, H_2: "fair coin", "always heads"

$P(D|H_1) = 1/2^5$ $\quad P(H_1) = \varepsilon$

$P(D|H_2) = 0$ $\quad P(H_2) = 1 - \varepsilon$

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{1/32}{0} \times \frac{\varepsilon}{1-\varepsilon} = \text{infinity}$$

Comparing two hypotheses

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D: HHHHH

H_1, H_2: "fair coin", "always heads"

$P(D|H_1) = 1/2^5$ $\quad P(H_1) = \varepsilon$

$P(D|H_2) = 1$ $\quad P(H_2) = 1 - \varepsilon$

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{1/32}{1} \times \frac{\varepsilon}{1-\varepsilon} = ?$$

Comparing two hypotheses

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D: HHHHH

H_1, H_2: "fair coin", "always heads"

$P(D|H_1) = 1/2^5$ $\quad P(H_1) = 999/1000$

$P(D|H_2) = 1$ $\quad P(H_2) = 1/1000$

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{1/32}{1} \times \frac{999}{1} \approx 30$$

Comparing two hypotheses

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D: HHHHHHHHHH

H_1, H_2: "fair coin", "always heads"

$P(D|H_1) = 1/2^{10}$ $\quad P(H_1) = 999/1000$

$P(D|H_2) = 1$ $\quad P(H_2) = 1/1000$

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{1/1024}{1} \times \frac{999}{1} \approx 1$$

Measuring prior knowledge

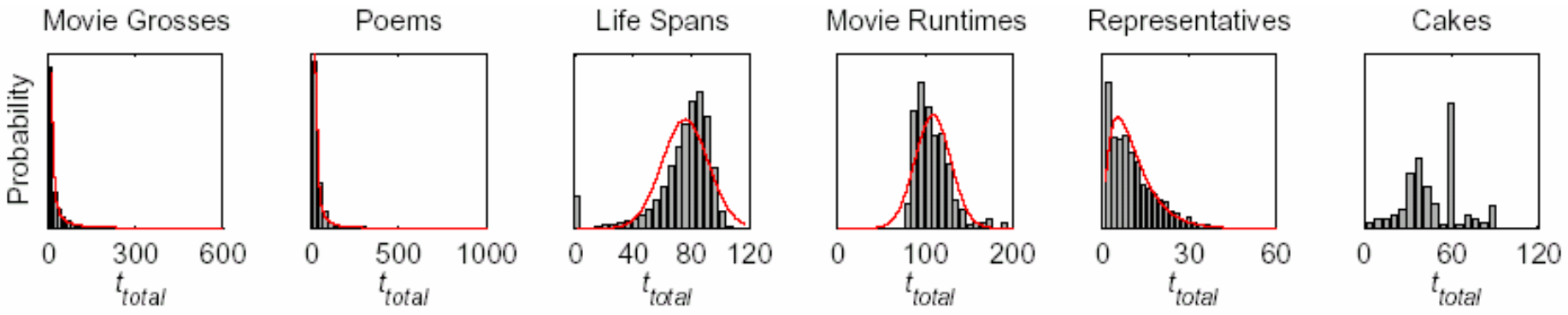
1. The fact that `HHHHH` looks like a "mere coincidence", without making us suspicious that the coin is unfair, while `HHHHHHHHHH` does begin to make us suspicious, measures the strength of our prior belief that the coin is fair.
   - If θ is the threshold for suspicion in the posterior odds, and D^* is the shortest suspicious sequence, the prior odds for a fair coin is roughly $\theta / P(D^* | \text{"fair coin"})$.
   - If $\theta \sim 1$ and D^* is between 10 and 20 heads, prior odds are roughly between 1/1,000 and 1/1,000,000.
2. The fact that `HHTHT` looks representative of a fair coin, and `HHHHH` does not, reflects our prior knowledge, intuitive theories about possible causal mechanisms in the world.
   - Easy to imagine how a trick all-heads coin could work: low (but not negligible) prior probability.
   - Hard to imagine how a trick "`HHTHT`" coin could work: extremely low (negligible) prior probability.

Everyday prediction problems

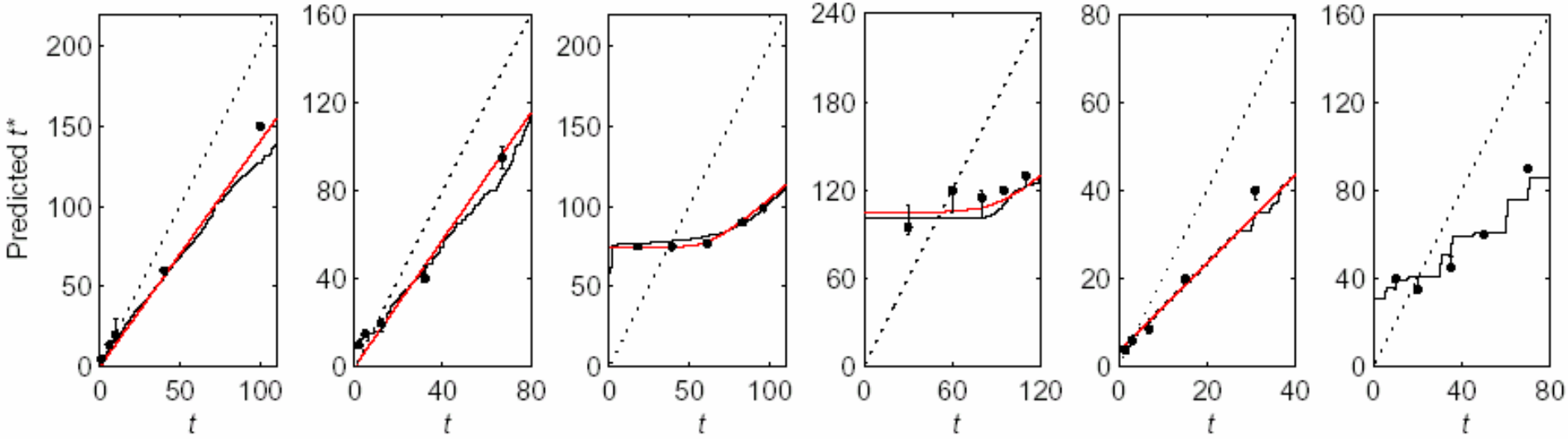
(Griffiths & Tenenbaum, *Psych. Science* 2006)

- You read about a movie that has made $60 million to date. How much money will it make in total?
- You see that something has been baking in the oven for 34 minutes. How long until it's ready?
- You meet someone who is 78 years old. How long will they live?
- Your friend quotes to you from line 17 of his favorite poem. How long is the poem?
- You meet a US congressman who has served for 11 years. How long will he serve in total?
- You encounter a phenomenon or event with an unknown extent or duration, t_{total}, at a random time or value of $t < t_{total}$. What is the total extent or duration t_{total}?

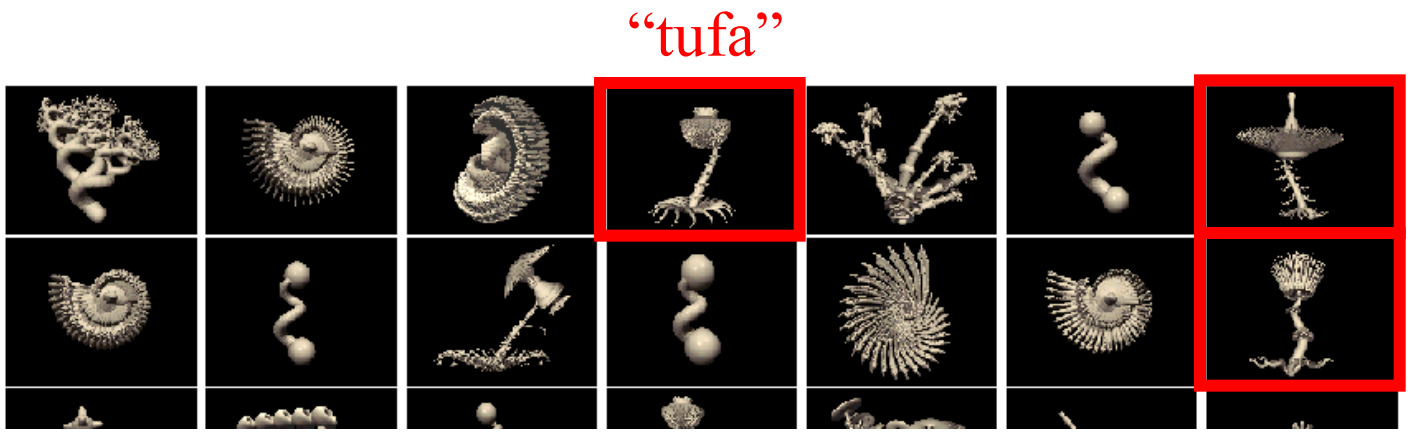
Priors $P(t_{total})$ based on empirically measured durations or magnitudes for many real-world events in each class:



Median human judgments of the total duration or magnitude t_{total} of events in each class, given one random observation at a duration or magnitude t, versus Bayesian predictions (median of $P(t_{total}|t)$).



Learning words for objects



What is the right prior?
What is the right hypothesis space?
How do learners acquire that background knowledge?