# Cognition: Methods and Models

PSYC 2040

L10: Language

Part 2

# recap: Apr 18, 2023


- what we covered:
  - debates in language (innate/learned)
  - statistical learning and prediction
- your to-dos were:
  - *finish:* L10 readings
  - *post (by 10 am)*: conceptual reflection
  - *complete (by 10 am)* : language experiment

# conceptual reflection #saffran

- experiment 1: testing familiarity
  - familiar: words (bulado)
  - novel: non-words (ladobi)
- experiment 2: pretty#baby
  - familiar: words (pretty)
  - novel: part-words (ty#ba)
- experiment 2 was a stronger test of the statistical learning hypothesis as infants were truly learning word boundaries

**Table 1.** Mean time spent listening to the familiar and novel stimuli for experiment 1 (words versus nonwords) and experiment 2 (words versus part-words) and significance tests comparing the listening times.

| Experiment | Mean listening times (s) | | Matched-pairs *t* test |
|---|---|---|---|
| | Familiar items | Novel items | |
| 1 | 7.97 (SE = 0.41) | 8.85 (SE = 0.45) | $t(23) = 2.3, P < 0.04$ |
| 2 | 6.77 (SE = 0.44) | 7.60 (SE = 0.42) | $t(23) = 2.4, P < 0.03$ |

# conceptual reflection #animallearning

## Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins

Marc D Hauser [a], Elissa L Newport [b], Richard N Aslin [b]

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/S0010-0277(00)00132-3 Get rights and content

### Abstract

Previous work has shown that human adults, children, and infants can rapidly compute sequential statistics from a stream of speech and then use these statistics to determine which syllable sequences form potential words. In the present paper we ask whether this ability reflects a mechanism unique to humans, or might be used by other species as well, to acquire serially organized patterns. In a series of four experimental conditions, we exposed a New World monkey, the cotton-top tamarin (*Saguinus oedipus*), to the same speech streams used by Saffran, Aslin, and Newport (Science 274 (1996) 1926) with human infants, and then tested their learning using similar methods to those used with infants. Like humans, tamarins showed clear evidence of discriminating between sequences of syllables that differed only in the frequency or probability with which they occurred in the input streams. These results suggest that both humans and non-human primates possess mechanisms capable of computing these particular aspects of serial order. Future work must now show where humans' (adults and infants) and non-human primates' abilities in these tasks diverge.

## Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate

Elissa L. Newport [a], Marc D. Hauser [b], Geertrui Spaepen [b], Richard N. Aslin [a]

Trends in Cognitive Sciences

Review

## Constraints on Statistical Learning Across Species

Chiara Santolin[1,*] and Jenny R. Saffran[2]

Both human and nonhuman organisms are sensitive to statistical regularities in sensory inputs that support functions including communication, visual processing, and sequence learning. One of the issues faced by comparative research in this field is the lack of a comprehensive theory to explain the relevance of statistical learning across distinct ecological niches. In the current review we interpret cross-species research on statistical learning based on the perceptual and cognitive mechanisms that characterize the human and non-human models under investigation. Considering statistical learning as an essential part of the cognitive architecture of an animal will help to uncover the potential ecological functions of this powerful learning process.

# today’s agenda

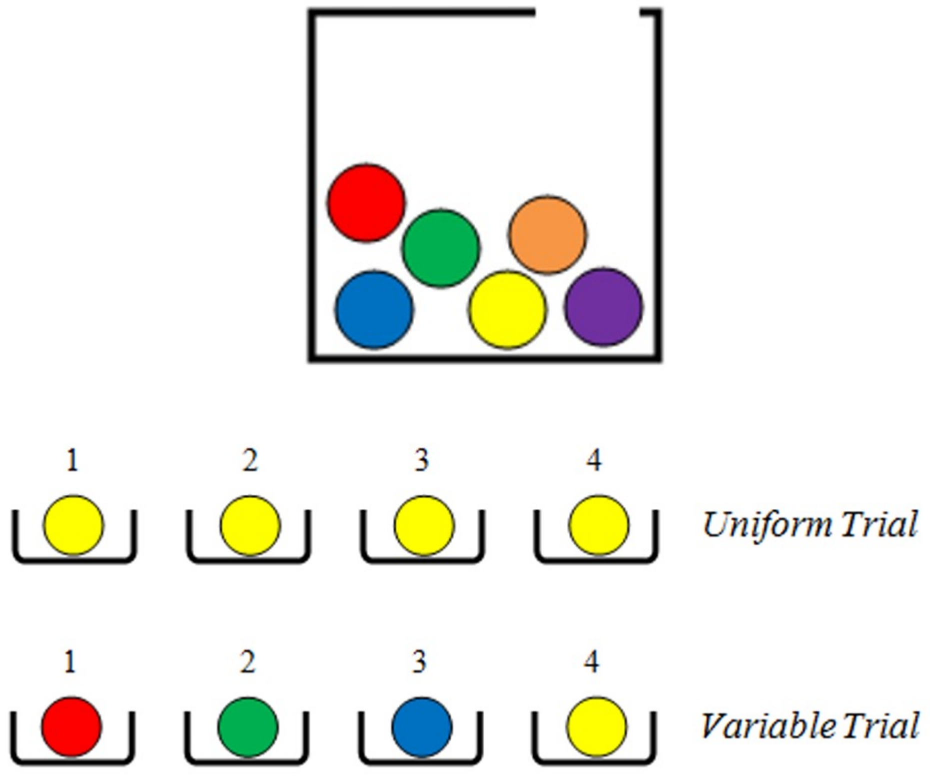
- statistical learning and curiosity
- co-occurrence
- error-free vs. error-driven learning
    - language models

# statistical learning and curiosity

- while there is evidence that statistical learning can inform predictions, it may also inform what to learn about in the first place
- curiosity may be particularly important in creating learning opportunities and minimizing uncertainty in the environment

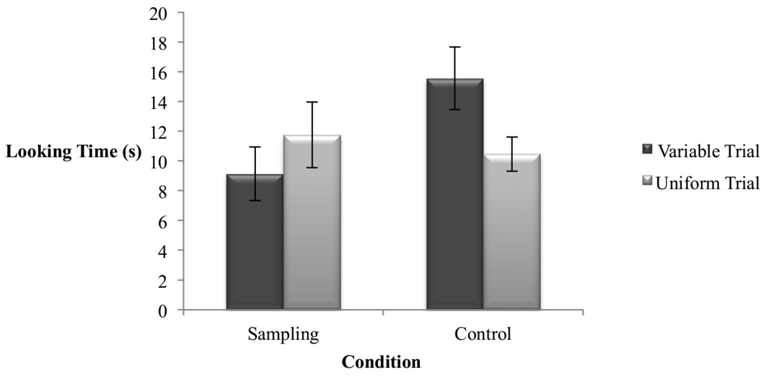
# statistical learning and curiosity

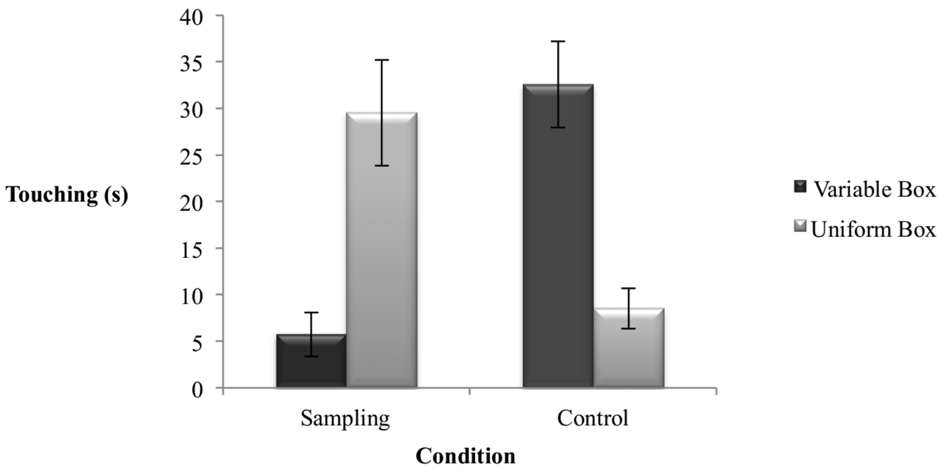
- Sim & Xu (2017) tested 13-month-old infants in a violation of expectation (VOE) and crawling paradigm
    - draw: could be “uniform” or “variable”
    - condition: control condition (experimenter looked into the box before drawing out the balls) or sampling (no looking)
- two experiments: looking time (VOE) vs. touching/reaching time (crawling)

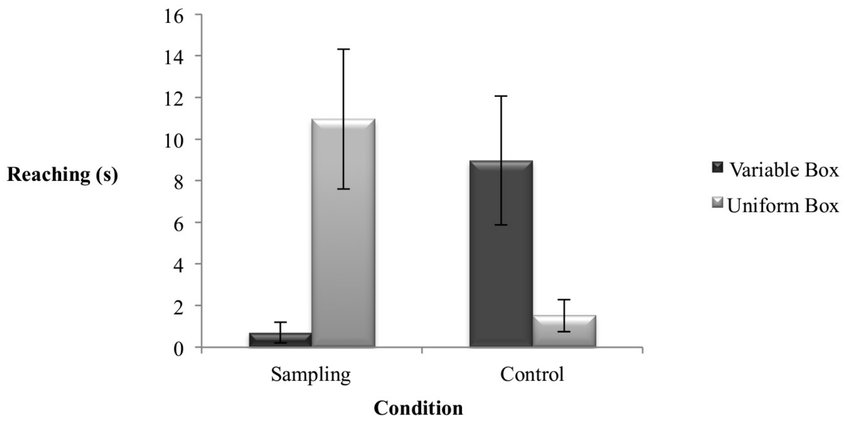


# statistical learning and curiosity

- Sim & Xu (2017) showed that 13-month-old infants preferentially explore sources of unexpected events



# conceptual reflection #curious

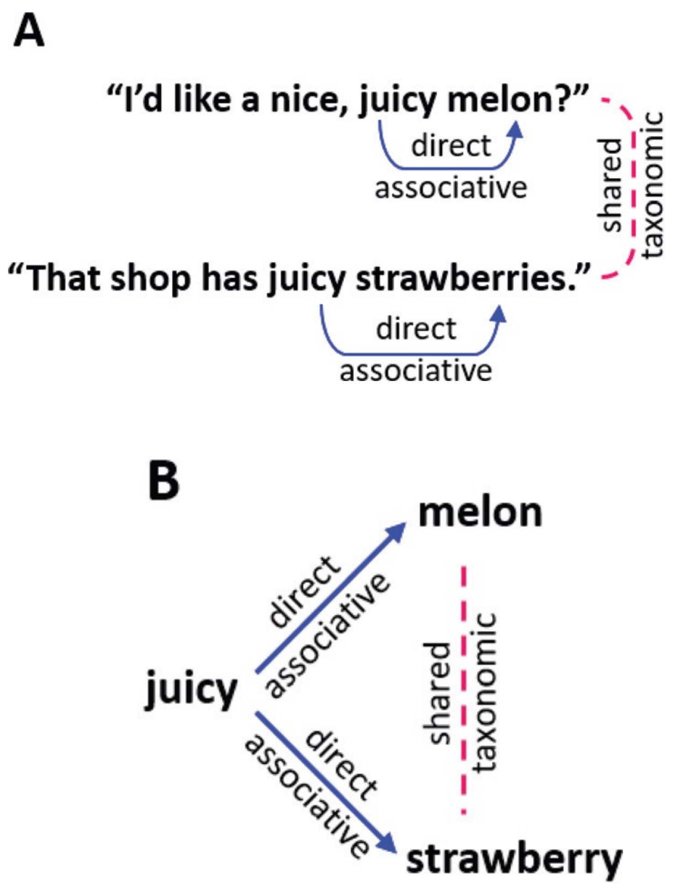
- I am quite interested in the claim that learning and reducing uncertainty can be a motivator in and of itself. This is applicable to many topics we have talked about so far. If rats are motivated by learning and reducing uncertainty sufficiently, why do we not see a steady increase in their mistakes when leaving a maze until they are presented with a reward? It seems like an underlying theme that, if true, should make us question much of the learning research we have discussed so far.

# review of findings/inferences

- infants track statistical regularities
- children learn from prediction error
- children are inherently curious and want to reduce uncertainty
- but…..
- what are we learning from, and are there alternatives to prediction?

# learning from co-occurrence

- a prominent view in language research is that the meaning of words is learned based on which words it co-occurs with in natural language
  - "you shall know a word by the company it keeps" (Firth, 1957)
- co-occurrence can be defined in two ways:
  - direct: if words occur together in the same context (e.g., eat-food, sit-chair, etc.)
  - indirect/shared: if words occur in similar contexts (e.g., strawberries are red, apples are red)
- co-occurrences ***are*** statistical regularities and can extend to any type of input (tones, figures, words, etc.)

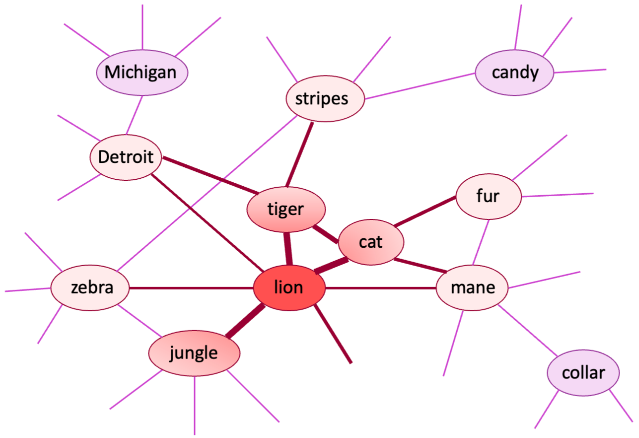


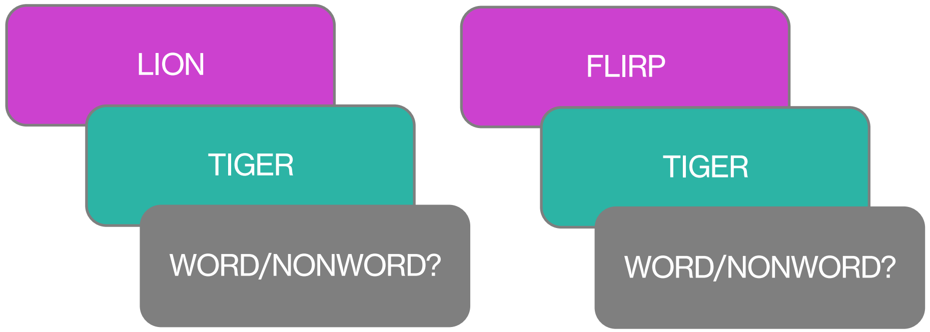
# experiment review

- think back to the language experiment you did
- what kinds of tasks did you perform?
- what do you think the experiment was about?

# semantic priming

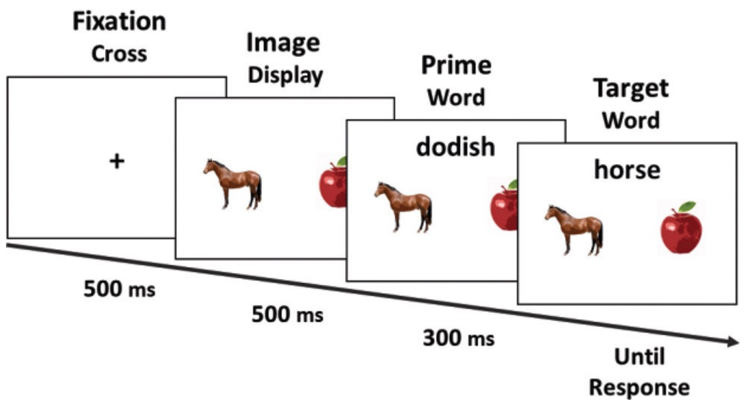
- semantic priming tasks are widely used to study how concepts influence the processing of other concepts (spreading activation theory)
  - what's semantic about this?
- a key finding from priming tasks is that related words are responded to faster than unrelated words



# learning new words

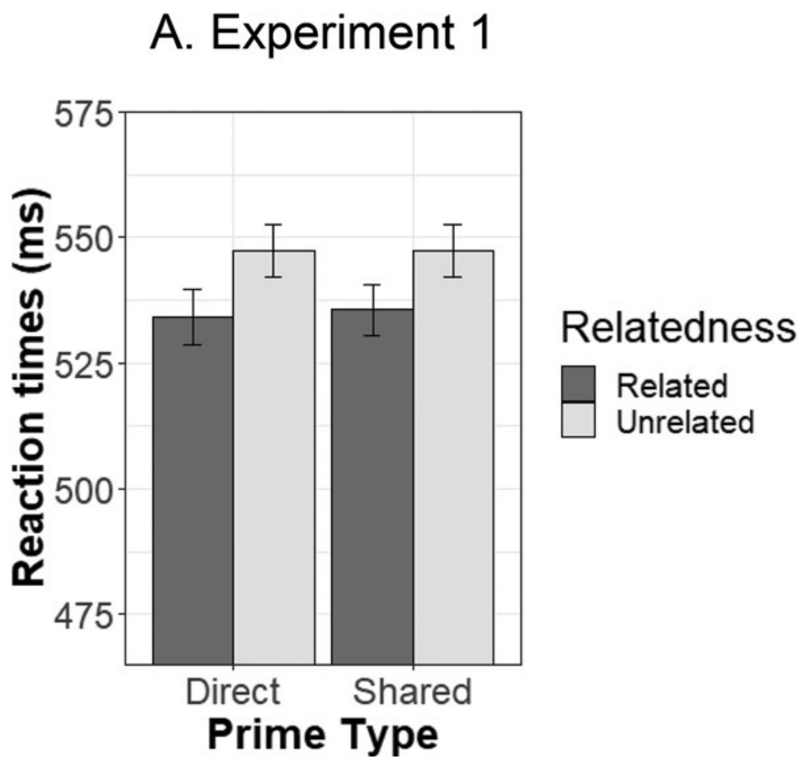
- Savic et al. (2022) had participants read sentences with novel and familiar words
  - novel words co-occurred with familiar words (direct or indirect)
- participants tested in a semantic priming experiment
- novel – familiar words were paired based on whether the pairs were related or unrelated and whether there was direct/indirect co-occurrence




| | related | unrelated |
|---|---|---|
| direct | dodish-horse | foobly-horse |
| indirect/shared | geck-horse | mipp-horse |

# semantic priming and co-occurrences

- reaction time to identify targets was faster when they were preceded by novel pseudowords/primes with which they directly co-occurred or shared co-occurrence in training
- pattern did not differ for direct and indirect co-occurrences
- inference: co-occurrences in natural language can drive semantic integration of new words

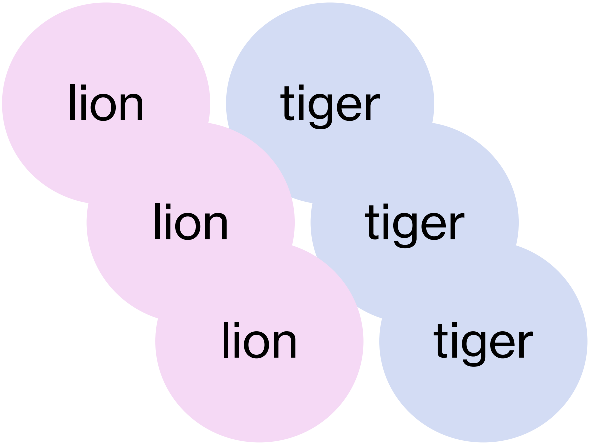


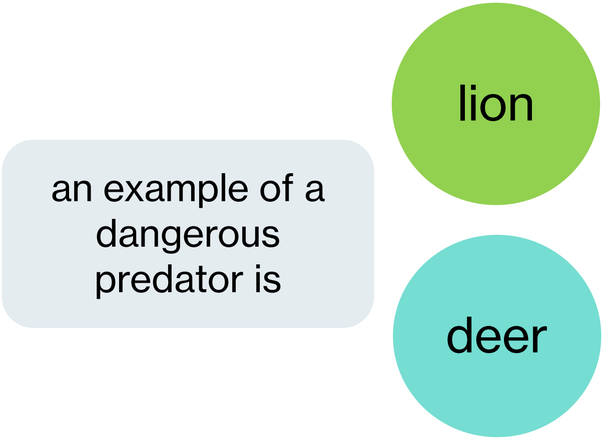
# connecting co-occurrence to learning

- assuming that we have a way of tracking co-occurrences or statistical regularities, ***how*** do they contribute to learning?
- two proposals for learning from co-occurrences:
  - error-driven (prediction)
  - error-free (no prediction)

# error-free vs. error-driven learning

- associative (error-free) learning
  - simply attending to statistical regularities/co-occurrence in the environment is sufficient to develop a conception of meaning
  - inspired by Hebbian learning within neurons
- predictive (error-driven) learning
  - use co-occurrence as a signal for word prediction at the sentence level
  - inspired by behaviorism/reinforcement learning



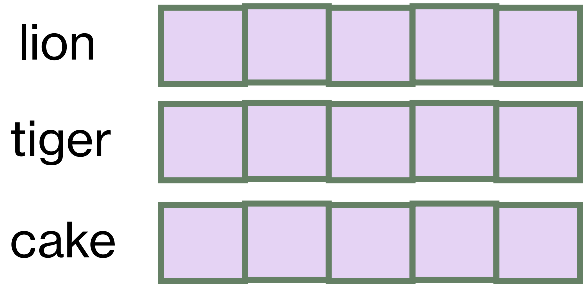



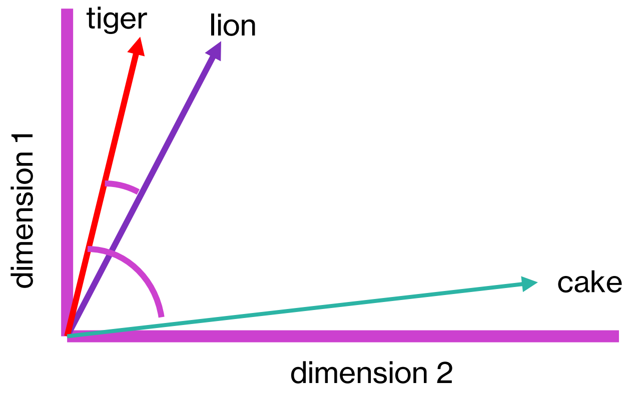
# testing the claims

- how would you test whether learning is error-free or error-driven?
- one possible solution: model learning in both ways and compare!
  - compare how they represent and relate concepts
  - compare language abilities on specific tasks
  - compare to human performance

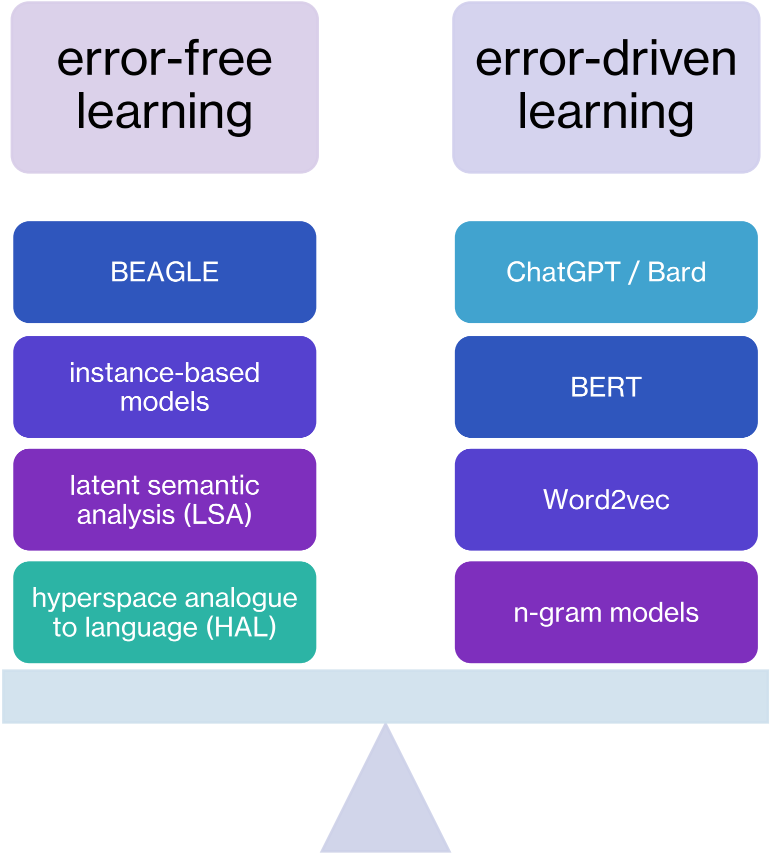
# what does it mean?

- language models are typically "trained" on large databases of text (e.g., Wikipedia, Google News, etc.)
- algorithm: specific models prioritize creating a co-occurrence matrix or predicting the next word(s) in the sentence
- after training, we can look under the hood at what 'representations' the models have acquired
- these representations are usually a collection of numbers but they are meaningfully related to each other in a high-dimensional space



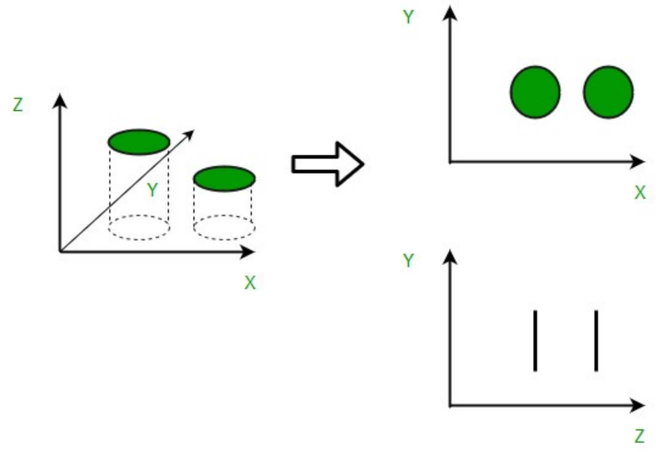



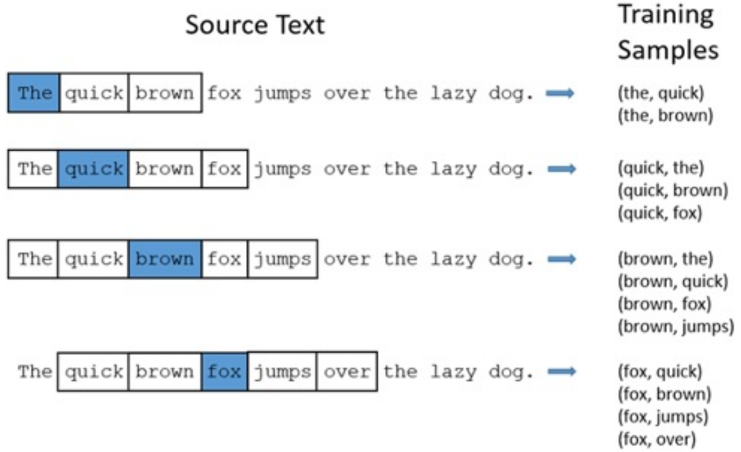
# language models



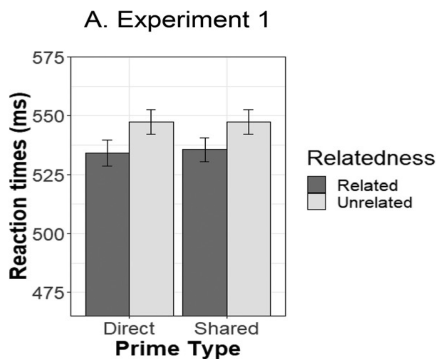
# error-free vs. error-driven models

- error-free models focus on the co-occurrence matrix
    - some models only emphasize direct co-occurrence (HAL), whereas others emphasize direct and indirect co-occurrence through some type of higher-level abstraction process (LSA, BEAGLE)
- error-driven models focus on prediction
    - prediction can occur at multiple levels and different models emphasize different aspects of the prediction process (n-word windows, attention-based, etc.)
    - neural networks are a family of models that can learn from prediction error

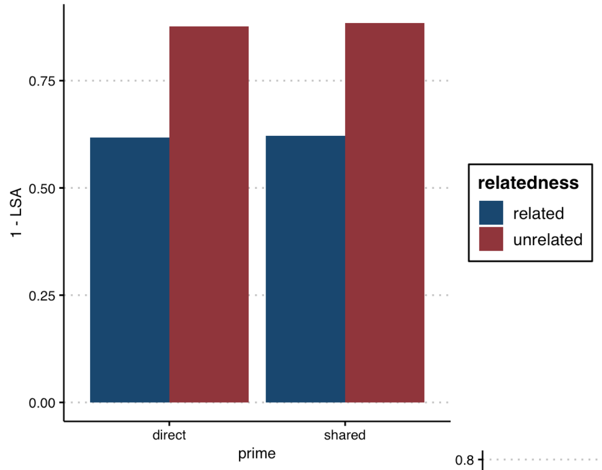


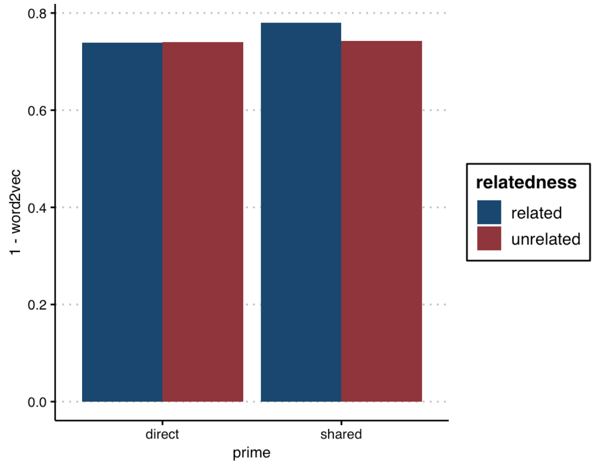



# error-free (LSA) vs. error-driven (word2vec))




- I submitted the 40 sentences you read to Latent Semantic Analysis to construct the co-occurrence matrix and perform abstraction and word2vec to perform prediction
- LSA was able to determine dodish-horse and geck-horse are more related than horse and mipp-horse
- word2vec struggled with comparing the similarities between dodish-horse/geck-horse and foobly-horse/mipp-horse
- this is very impressive: LSA does not have the input you have as adult learners!



# real empirical demonstrations

- word2vec and other prediction-based models, such as ChatGPT are extremely powerful when trained on large datasets with billions of parameters
- some work has shown that on smaller datasets (e.g., child-directed speech), error-free approaches may be more useful in predicting behavior (Asr, Willits, & Jones, 2016)

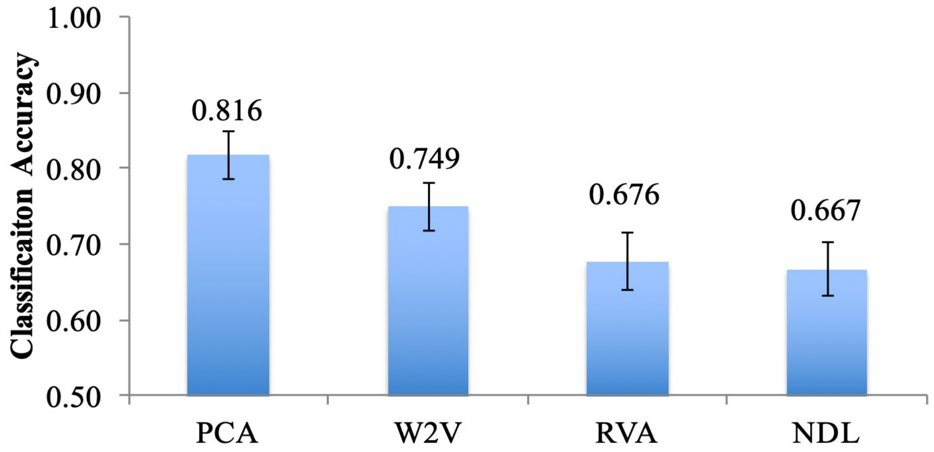



**Figure 3.** Mean classification accuracy (averaging across the 30 different categories and 95% CI) for the four DSMs.

# the wins of language models

- truly start from "scratch" and display remarkable language abilities
- dispel the need for hard-wiring many abilities by showing "emergent" behavior
  - hierarchical structure
  - separating syntax from semantics
- have widespread applications
  - translation
  - content writing/marketing
  - assistance

**Modern language models refute Chomsky's approach to language**

Steven T. Piantadosi[a,b]
[a]UC Berkeley, Psychology [b]Helen Wills Neuroscience Institute

The rise and success of large language models undermines virtually every strong claim for the innateness of language that has been proposed by generative linguistics. Modern machine learning has subverted and bypassed the entire theoretical framework of Chomsky's approach, including its core claims to particular insights, principles, structures, and processes. I describe the sense in which modern language models implement genuine *theories* of language, including representations of syntactic and semantic structure. I highlight the relationship between contemporary models and prior approaches in linguistics, namely those based on gradient computations and memorized constructions. I also respond to several critiques of large language models, including claims that they can't answer "why" questions, and skepticism that they are informative about real life acquisition. Most notably, large language models have attained remarkable success at discovering grammar without using any of the methods that some in linguistics insisted were necessary for a science of language to progress.

# activity: test ChatGPT!

- in groups of 3, go to activity spreadsheet
- you will test ChatGPT on how well it can understand internal beliefs
- you will:
  - read a prompt
  - record ChatGPT's response
  - evaluate its response
  - debrief

# ChatGPT

## Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks

**Tomer D. Ullman**
Department of Psychology
Harvard University
Cambridge, MA, 02138
tullman@fas.harvard.edu

# potential concerns: thinking & reasoning

- despite the successes, the models often fail on logical reasoning and thinking tasks

Language and thought are not the same thing: evidence from neuroimaging and neurological patients

Evelina Fedorenko[1,2,3] and Rosemary Varley[4]

Author information Copyright and License information Disclaimer

The publisher's final edited version of this article is available at Ann N Y Acad Sci

Abstract Go to:

Is thought possible without language? Individuals with global aphasia, who have almost no ability to understand or produce language, provide a powerful opportunity to find out. Astonishingly, despite their near-total loss of language, these individuals are nonetheless able to add and subtract, solve logic problems, think about another person's thoughts, appreciate music, and successfully navigate their environments. Further, neuroimaging studies show that healthy adults strongly engage the brain's language areas when they understand a sentence, but not when they perform other nonlinguistic tasks like arithmetic, storing information in working memory, inhibiting prepotent responses, or listening to music. Taken together, these two complementary lines of evidence provide a clear answer to the classic question: many aspects of thought engage distinct brain regions from, and do not depend on, language.

**Prompt:** `Get your sofa onto the roof of your house.`

**GPT-3 response:** `I would start by getting a very strong ladder and a very strong friend...` (goodness: 3.6 out of 7)

**Human response:** `You may need to rent a Genie lift large enough to carry the sofa. You will need at least one other person...` (goodness: 4.8 out of 7)

To manipulate how "out-of-distribution" the query is, the prompts get progressively more constrained:

**Prompt:** `Get your sofa onto the roof of your house, without using a pulley.`

**GPT-3 response:** `Use a rope to tie around the sofa and connect it to a car.` (goodness: 3.0 out of 7)

**Human response:** `I would get a giant crane... and use the crane to lift it to the roof of my house.` (goodness: 5.1 out of 7)

With more and more constraints, human responses remain at approximately the same level of goodness, whereas model performance breaks down completely:

**Prompt:** `Get your sofa onto the roof of your house, without using a pulley, a ladder, a crane...`

**GPT-3 response:** `Cut the bottom of the sofa so that it would fit through the window...break the windows to make room for the sofa.` (goodness: 2.7 out of 7)

**Human response:** `I will build a large wooden ramp...on the side of my house with platforms every 5 feet...` (goodness: 5.0 out of 7)

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

# potential concerns: data

- the size of the corpora that models are trained on is 1000 times more than the input available to children
  - “poverty of stimulus” argument!
- most models are based on the English language (Bender rule)

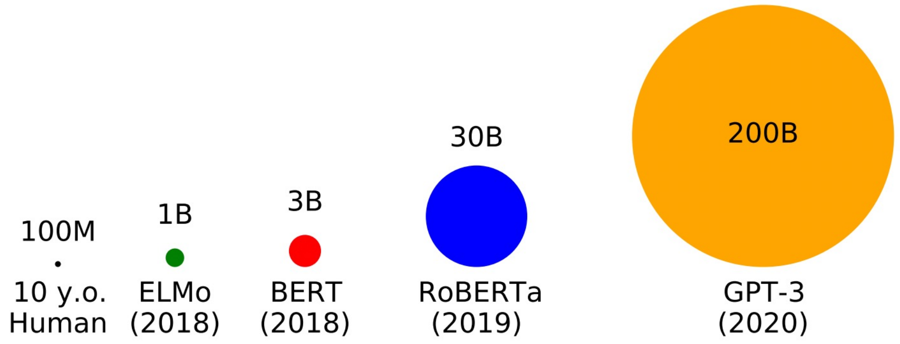
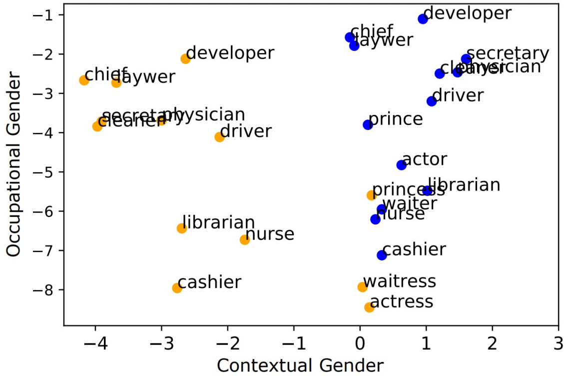



**Figure 1:** Comparison of human and model linguistic input (# of word tokens).

# potential concerns: biases and costs

- due to being trained on natural language text (often from the internet), they learn stereotypes and biases
- there are sizeable costs to the environment and climate of training these models




| Consumption | $CO_2e$ (lbs) |
|---|---|
| Air travel, 1 passenger, NY↔SF | 1984 |
| Human life, avg, 1 year | 11,023 |
| American life, avg, 1 year | 36,156 |
| Car, avg incl. fuel, 1 lifetime | 126,000 |
| **Training one model (GPU)** | |
| NLP pipeline (parsing, SRL) | 39 |
| w/ tuning & experimentation | 78,468 |
| Transformer (big) | 192 |
| w/ neural architecture search | 626,155 |

Table 1: Estimated $CO_2$ emissions from training common NLP models, compared to familiar consumption.[1]

# the path forward

- situating language within the broader conversation about human intelligence
- linguistic: sign language, prosody
- non-linguistic:
  - multimodal input
    - "intuitive physical reasoning"
  - interactive/social learning
    - "intuitive psychology"

Building machines that learn and think like people

**Brenden M. Lake**
*Department of Psychology and Center for Data Science, New York University, New York, NY 10011*
brenden@nyu.edu
http://cims.nyu.edu/~brenden/

**Tomer D. Ullman**
*Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139*
tomeru@mit.edu
http://www.mit.edu/~tomeru/

**Joshua B. Tenenbaum**
*Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139*
jbt@mit.edu
http://web.mit.edu/cocosci/josh.html

**Samuel J. Gershman**
*Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139*
gershman@fas.harvard.edu
http://gershmanlab.webfactional.com/index.html

# big takeaways

- get in groups of 3 and report key takeaways from today
- [takeaways document]

# next class


- **before** class:
  - *finish:* L10 quiz/assignment
  - *work on:* project milestone #5
  - *read*: L11 reading
- **during** class:
  - social cognition