How Infants Learn Words by Interacting with the Visual World

Chen Yu
Indiana University

Children are prodigious word learners

- By the age of two, toddlers recognize instances of about 300 object categories. They also appropriately generalize newly learned label to instances that they've never seen before.

- By the age of six, children acquire up to 14,000 words (Carey, 1978)

- Children are *"lexical vacuum cleaners, inhaling a new word every two waking hours"* (Pinker, 1994)

How do children learn so many words so quickly?

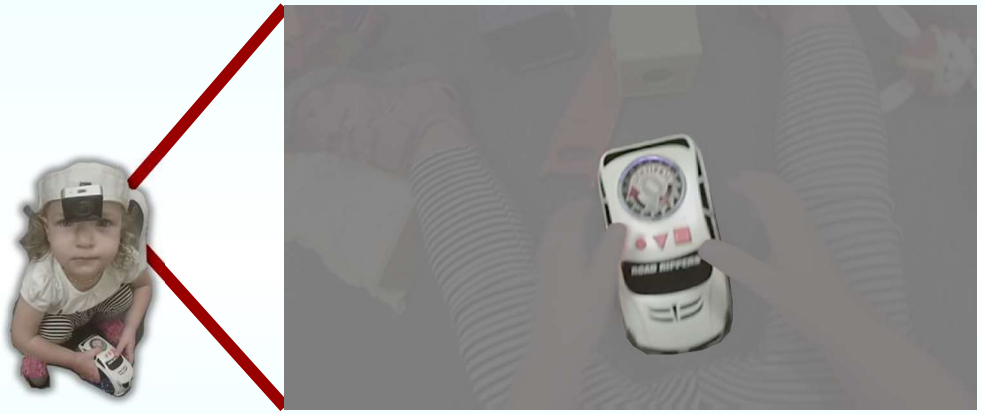
Word Learning from the Infant's perspective

Yurovsky, Smith & Yu (2013),
Developmental Science

Yurovsky, Smith & Yu (2013),
Developmental Science

Yurovsky, Smith & Yu (2013),
Developmental Science

Learning Problems



Segmentation Problem

Referential Uncertainty Problem

Recognition/Generalization

Head-Mounted Camera and Dual Eye Tracking

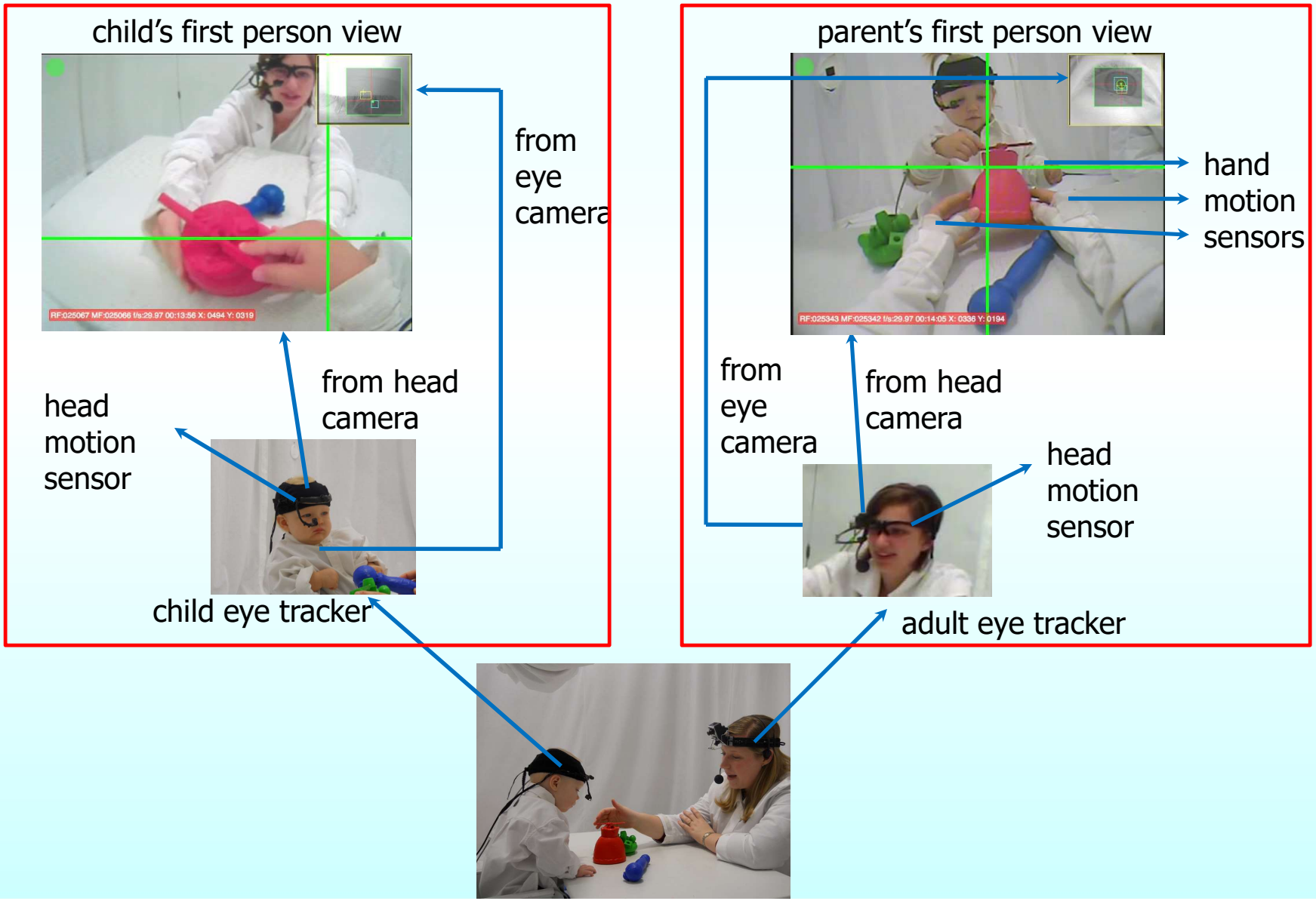


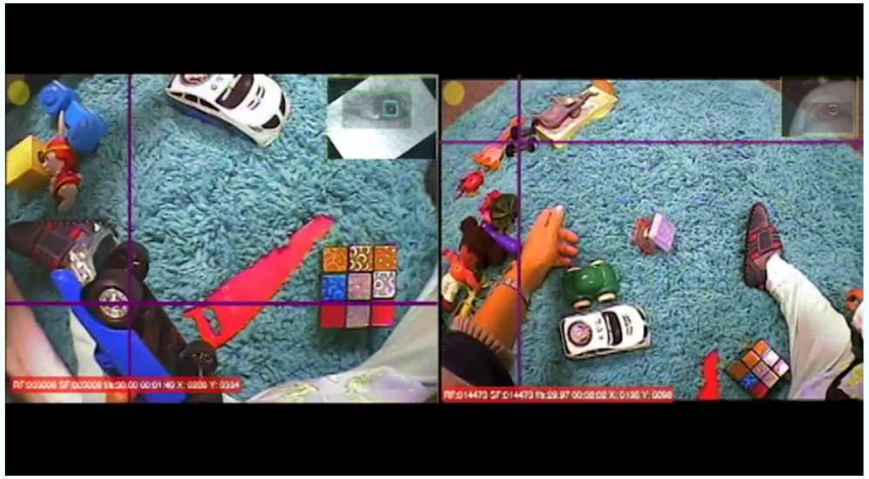






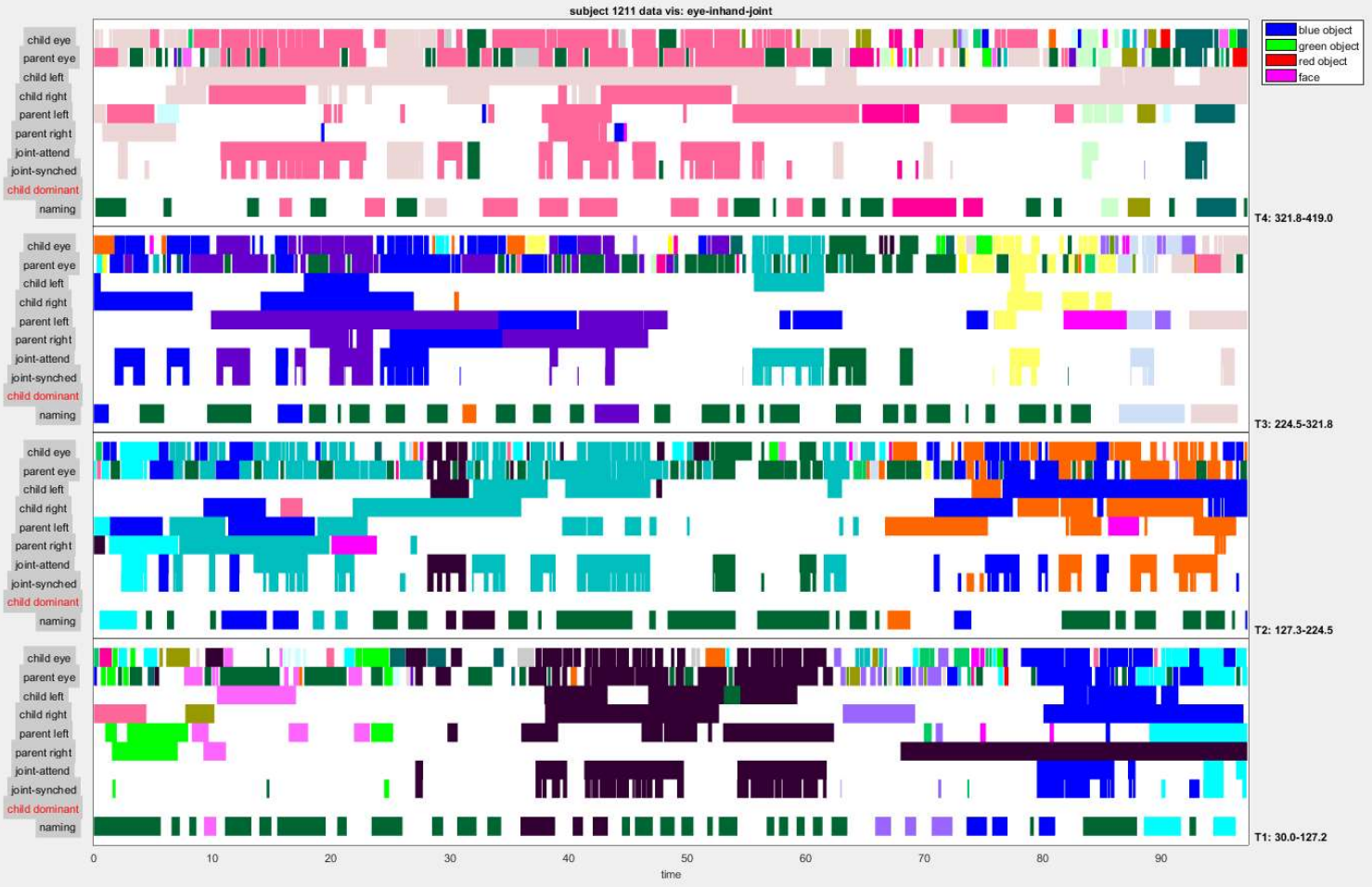
Dual Eye Tracking in Child-Parent Interaction






RF:003411 SF:003411 f/s:29.97 00:01:53 X: 0178 Y: 0238

Micro-Level Behavior



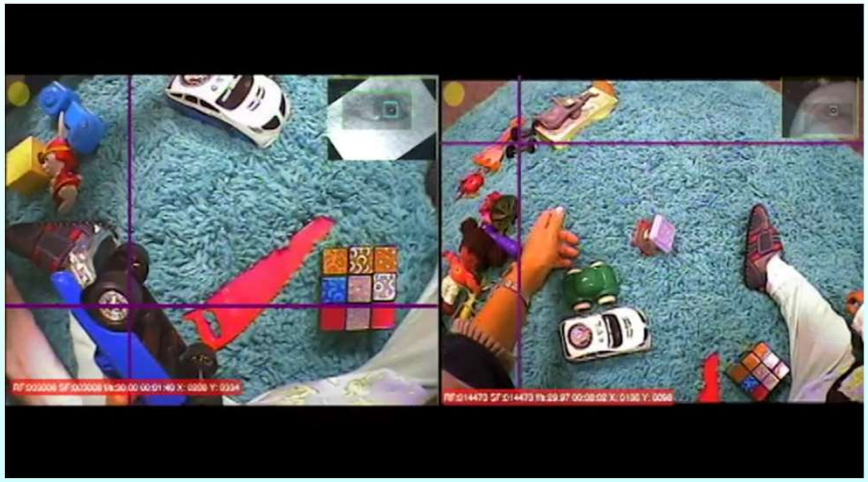
Study 1: Active Viewing in Toddlers Facilitates Visual Object Learning

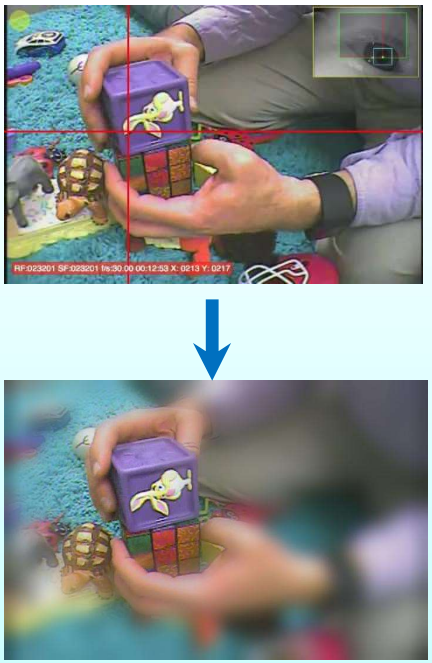
Sven Bambach

- In everyday contexts such as toy play, toddlers actively create many different views of the same object. Active viewing may create high-quality training data for visual object recognition.

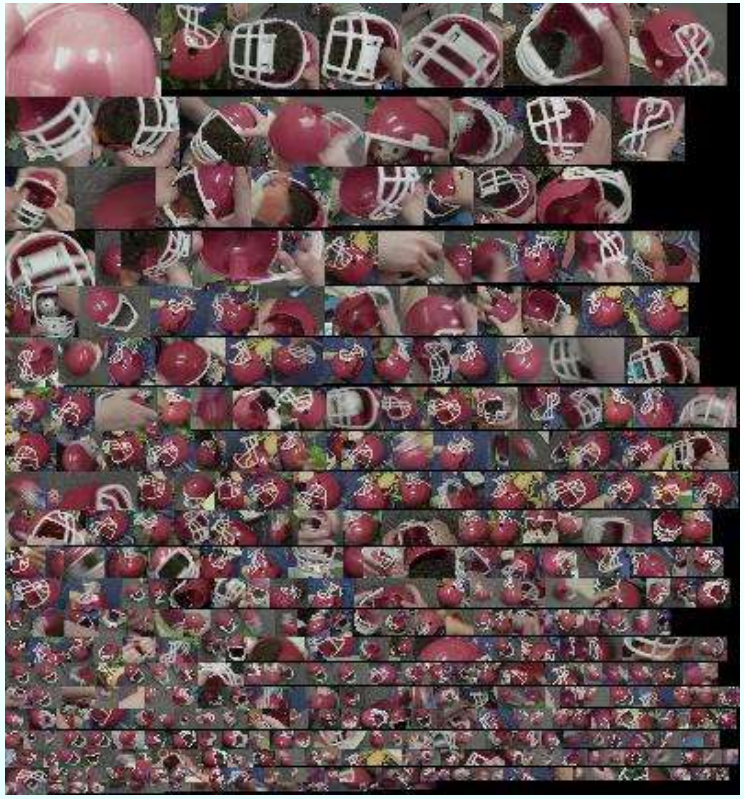
- We use these toddler-perspective images as inputs to machine learning models to test how the visual information supports the development of visual object recognition.

child's view parent's view

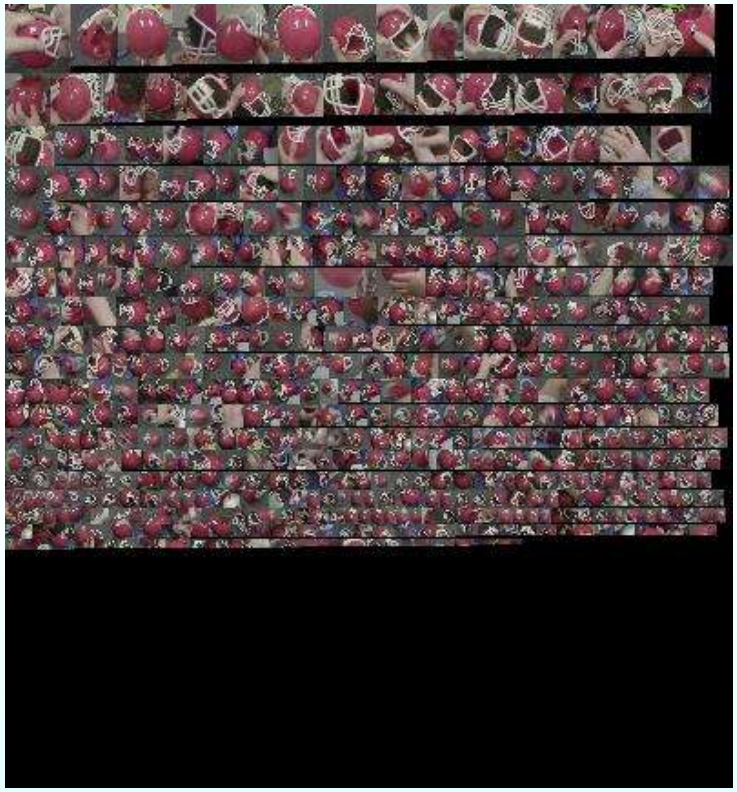




Different views of the world



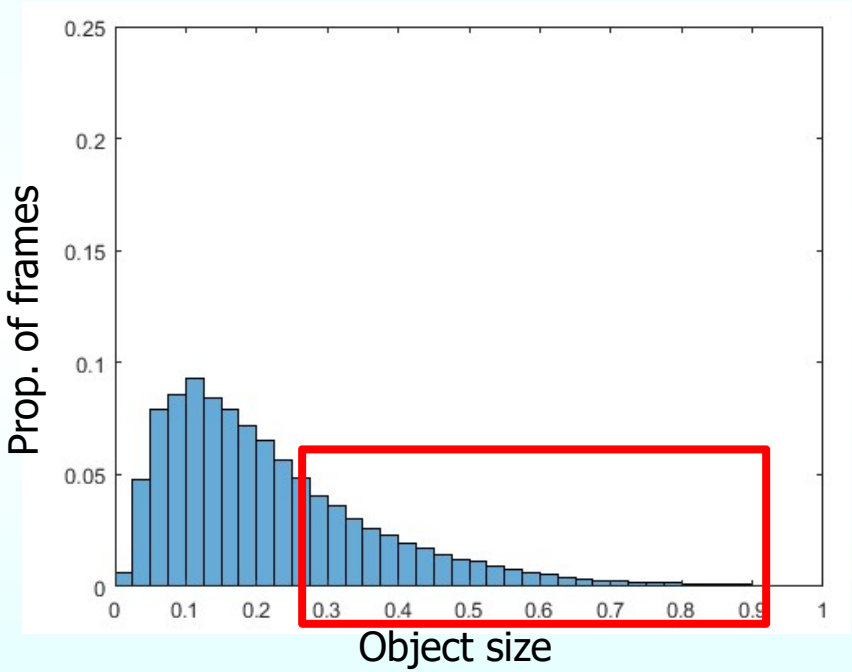
child’s view

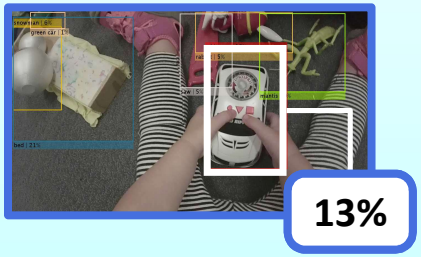


parent’s view

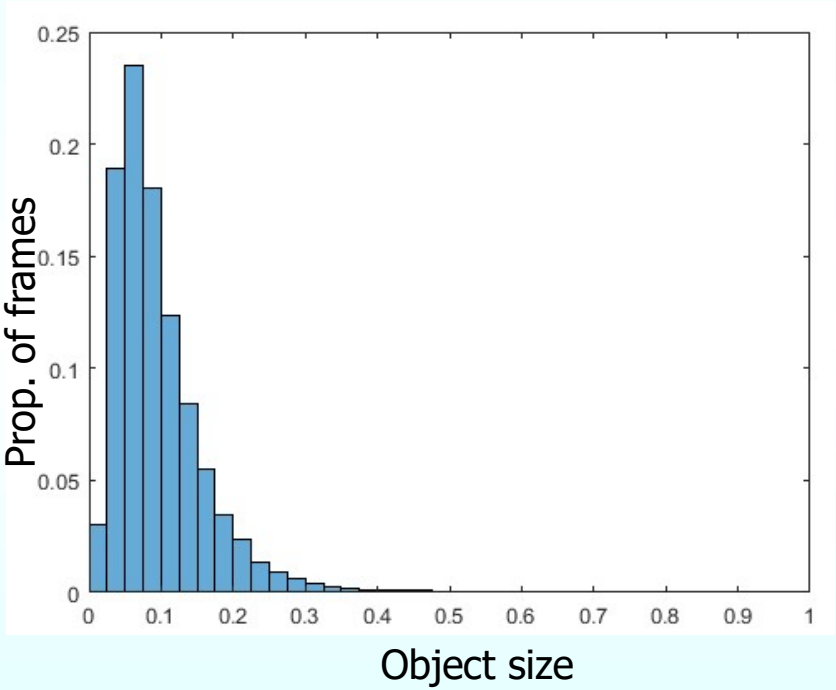
Object Size

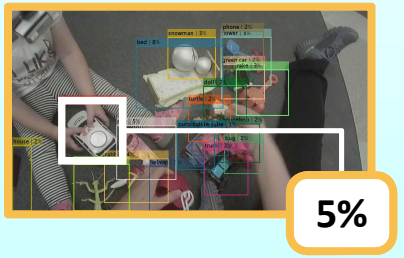
Child's view





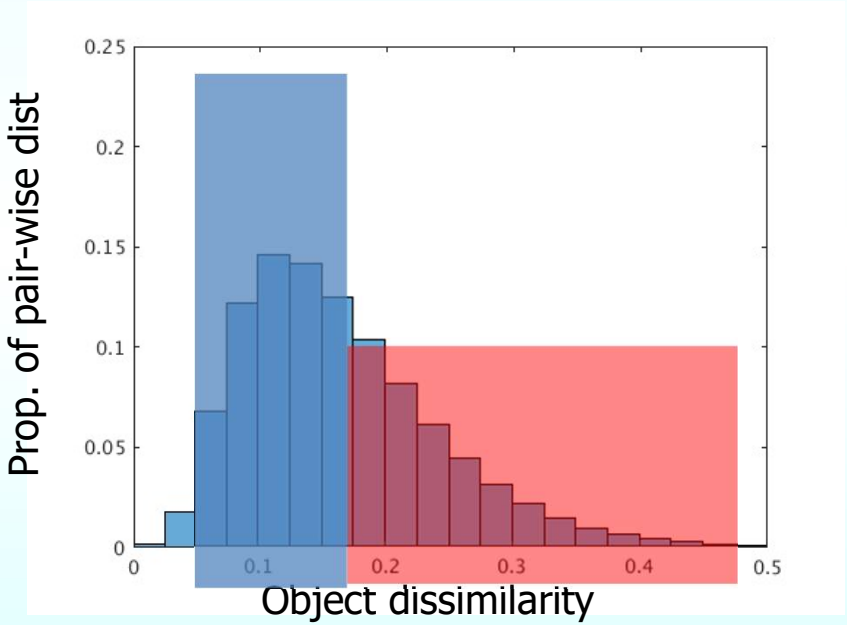
Parent's view



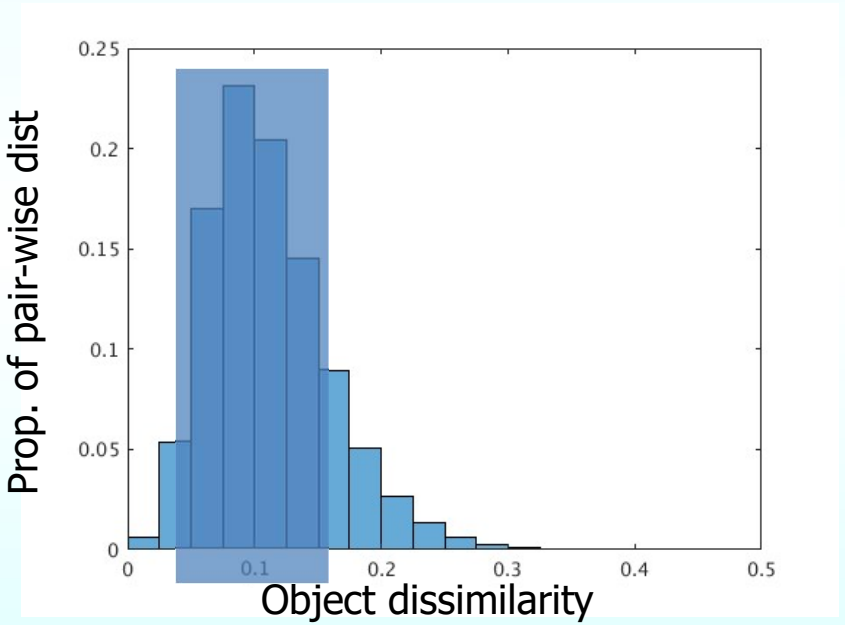


Visual properties of objects

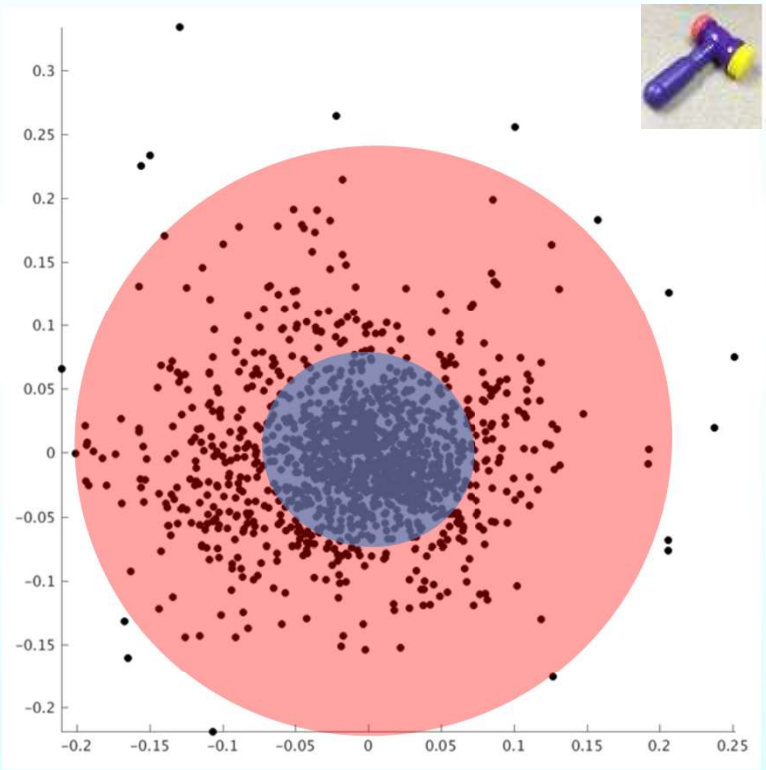
Child's view



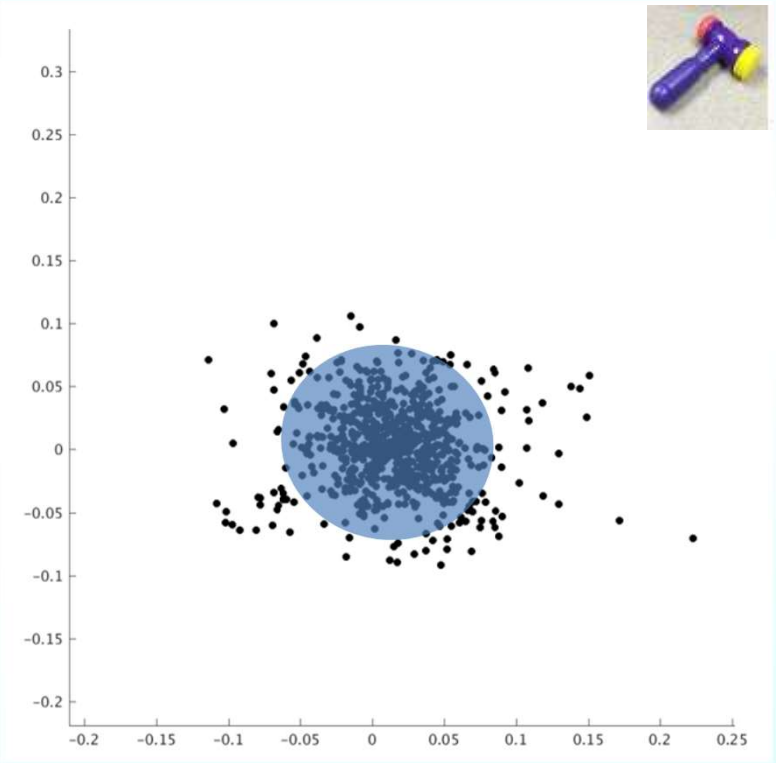
Parent's view



Consistency and Variability

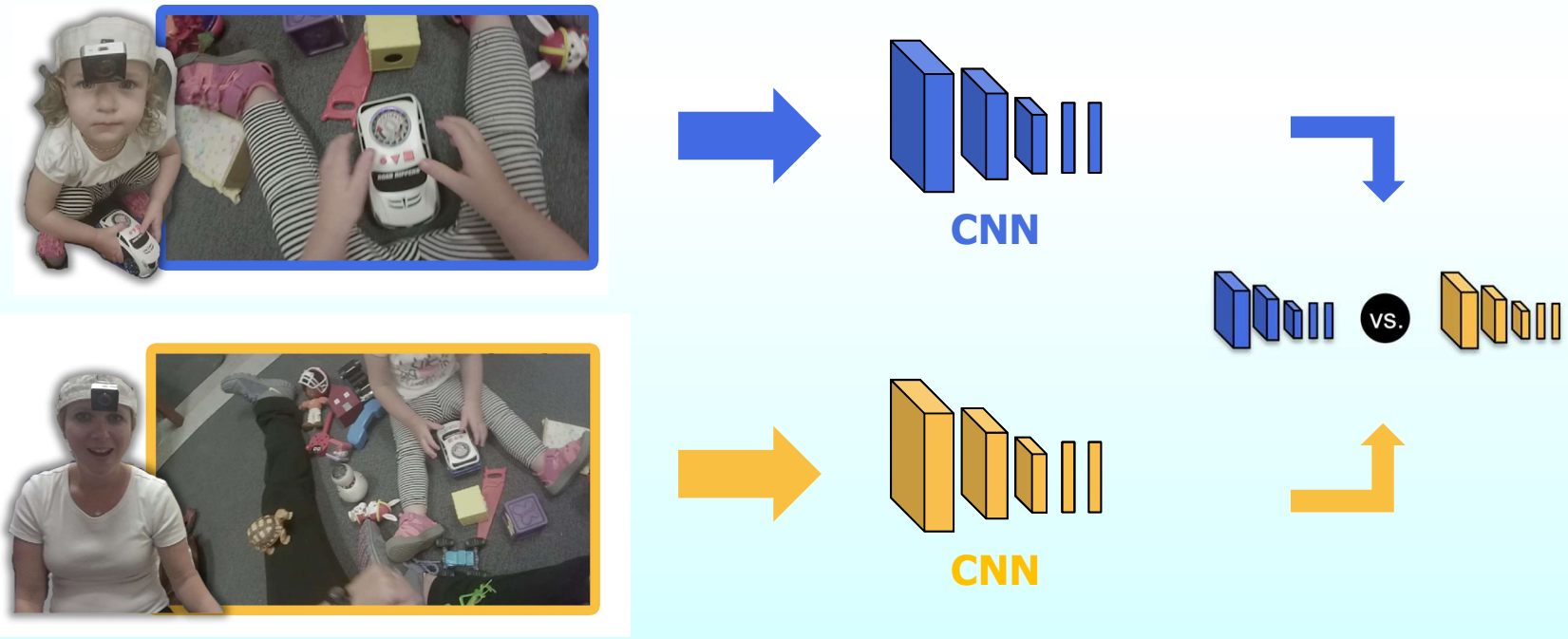


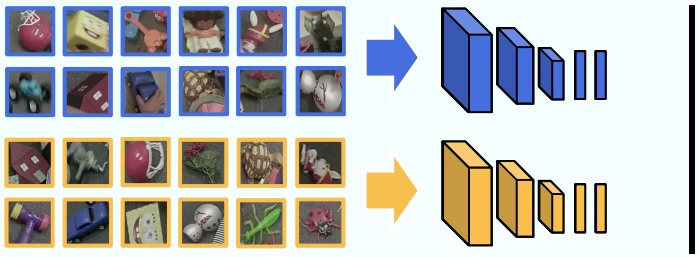
Child's view



Parent's view

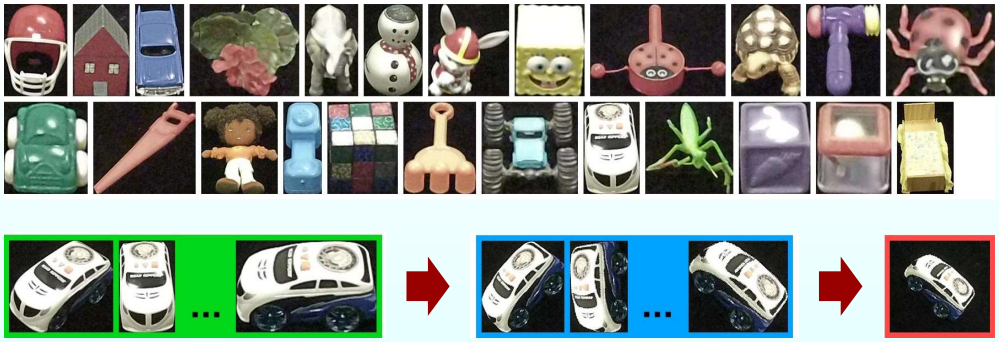
We trained machine learning models based on Convolutional Neural Networks (CNNs), and examine the extent to which these models take advantage of visual information created and perceived by toddlers.





Train

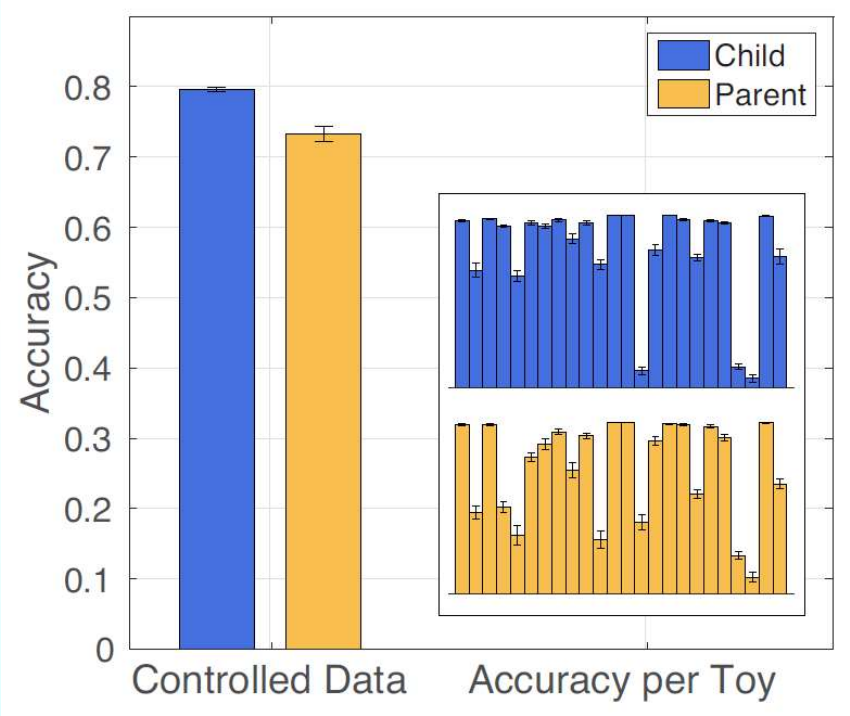
Test

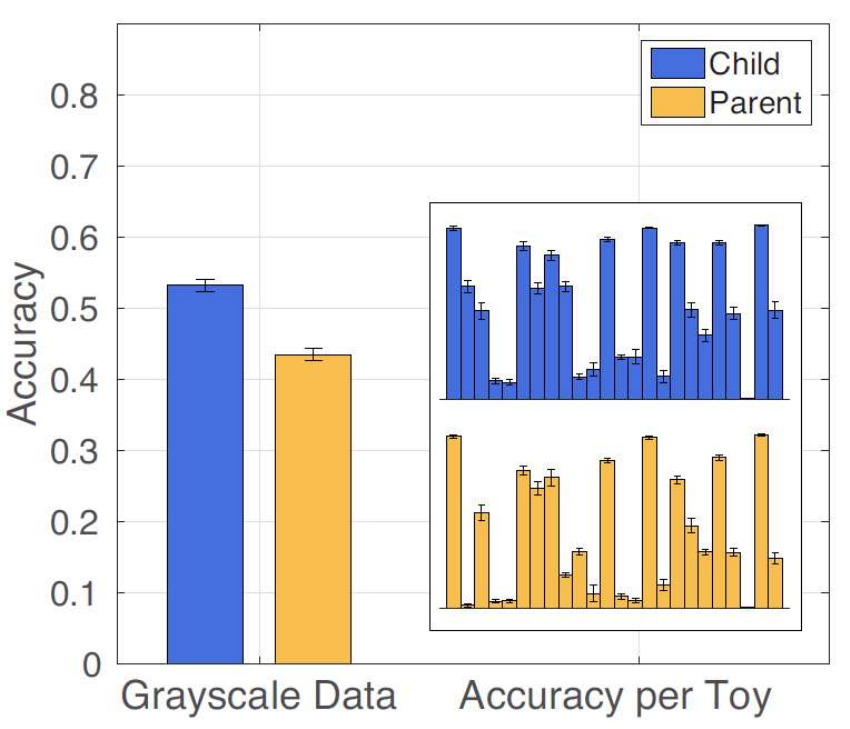


8x8x2 = 128 images per toy

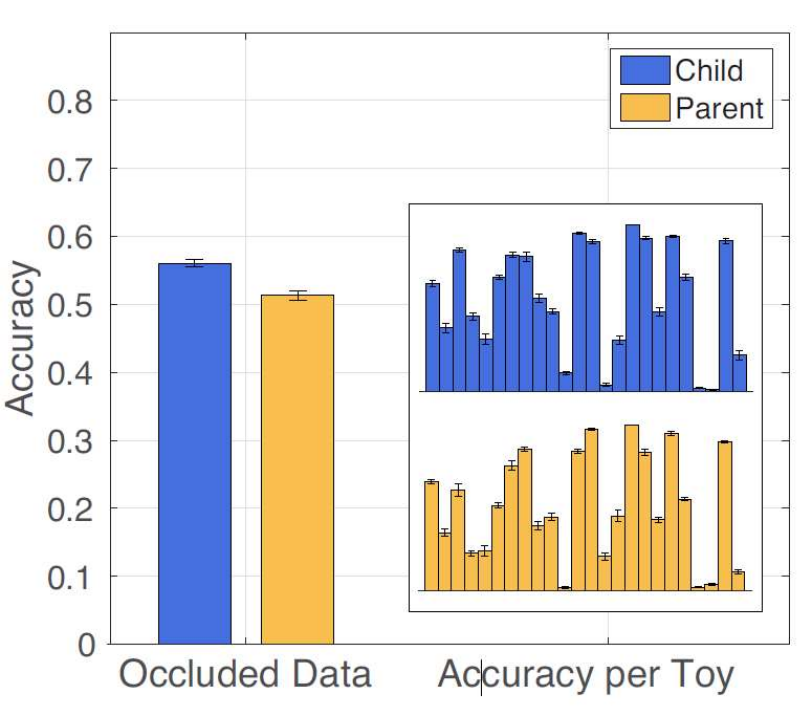
A large variety of clean, systematically-collected, unobstructed third-person views for each toy, to serve as a view-independent and therefore objective way to evaluate the performance of visual object recognition.

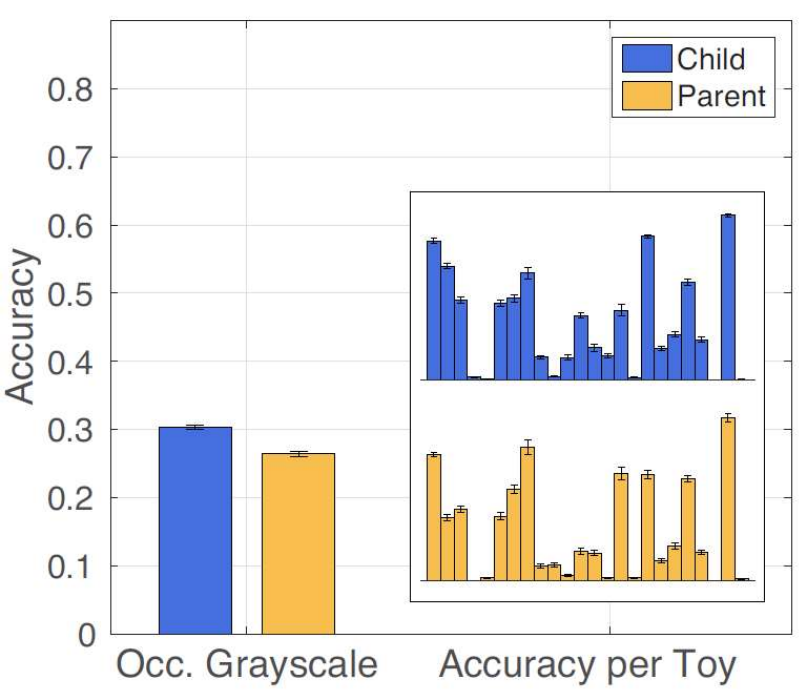
Performance





Performance





Our results showed that (1) CNNs were indeed able to learn object models of the toys in this first-person data and (2) that these models could generalize and recognize the same toys in a different context with different viewpoints. Finally, we showed that (3) the visual data collected by toddlers seemed to be of particularity high quality as models trained with toddler data consistently outperformed those trained with parent data in multiple simulation conditions.

Summary of Active Object Learning

(Bambach, Crandall, Smith, Yu, 2016,2017)

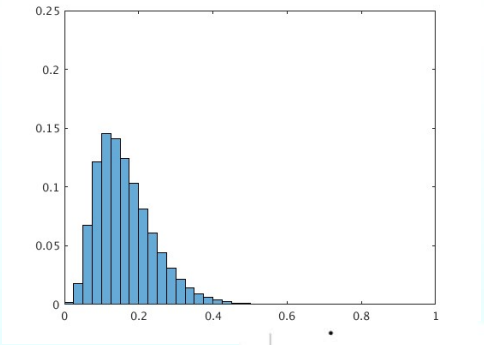
- In the real world, toddlers spend hours every day playing with toys. When they play, they actively manipulate toy objects and by doing so, they create learning experiences by self-selecting object views for visual learning. This object selection process seems to create more diverse viewpoints for the children than for the parents.

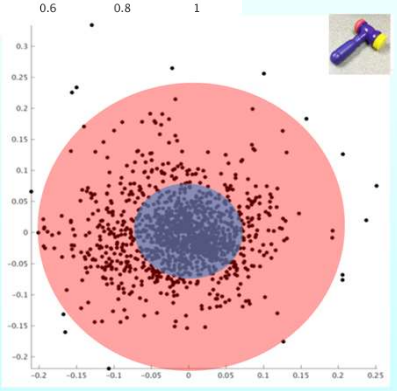
- Better data lead to better learning. If active viewing of toddlers creates high-quality training data for object recognition, then toddlers may rely on this active learning solution to become efficient learners of visual objects.

Creating own data

- ML approaches are based on different assumptions on data:
  - Data-hungry approaches assume that given enough data, the learner can discover the regularities that define different learning domains.
  - Powerful rational inference models, in contrast, achieve fast domain specific learning by seeding knowledge or biases into the internal inference machinery.

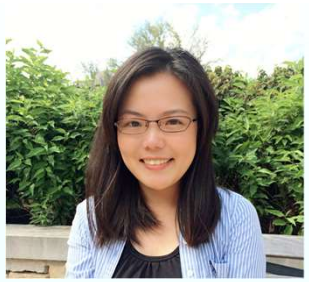
- The relevant data for learning are not the statistics of the physical and social world but only the samples that emerge within the learners' own experiences. Those experiences depend on the relation between the sensors and the events in the world and this relation changes with infants' actions which create their own data with unique properties and distributions.





Child's view

Study 2: Statistical Learning of Word-Object mapping (Zhang & Yu, 2016,2017)



Yayun Zhang

Child's first person view during toy play

Experimental Psychology Approach to the Interpretability Problem

Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study

Samuel Ritter [*1] David G.T. Barrett [*1] Adam Santoro [1] Matt M. Botvinick [1]

Abstract

Deep neural networks (DNNs) have achieved unprecedented performance on a wide range of complex tasks, rapidly outpacing our understand-

1. Introduction

During the last half-decade deep learning has significantly improved performance on a variety of tasks (for a review, see LeCun et al. (2015)). However, deep neural network (DNN) solutions remain poorly understood, leaving many to think of these models as black boxes, and to question

"Cognitive psychologists have long wrestled with the problem of understanding another opaque intelligent system: the human mind. We contend that the search for a better understanding of DNNs may profit from the rich heritage of problem descriptions, theories, and experimental tools developed in cognitive psychology."

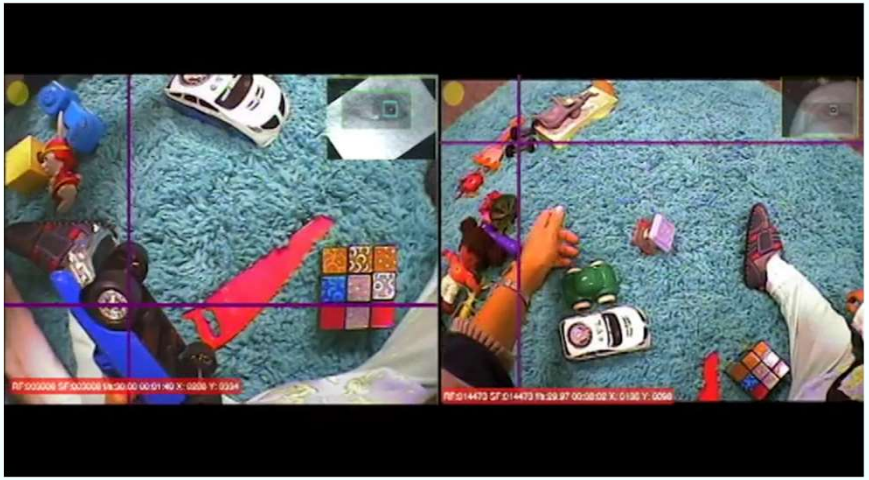
Experimental Psychology Approach

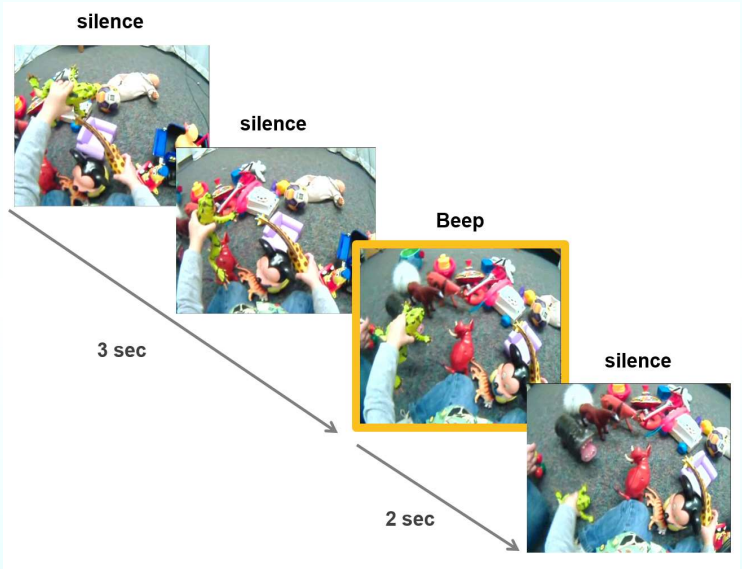
How toddlers learn object names through toy play?

How human learners discover correct word-object mappings from a sequence of naming instances with different degrees of uncertainty?

- Step 1: Extracting individual naming instances from naturalistic object play.
- Step 2: Quantifying how informative each individual naming instance is.
- Step 3: Examining how human learners aggregate statistical information across multiple learning instances by putting together a set of naming instances in a systematic way.

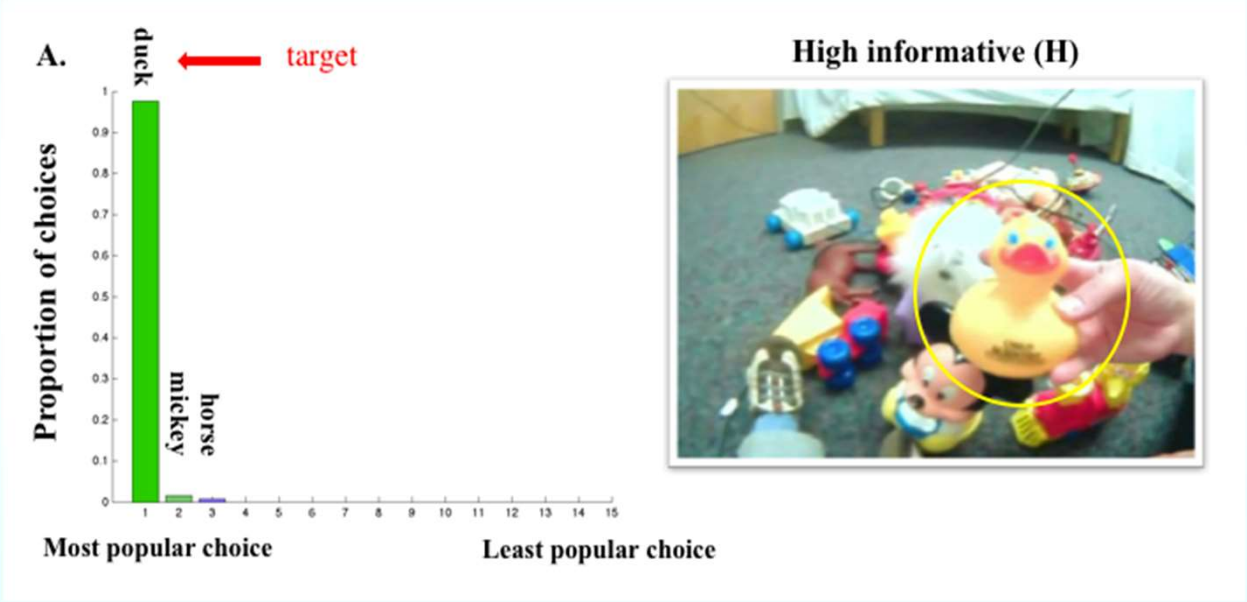
Experimental Approach



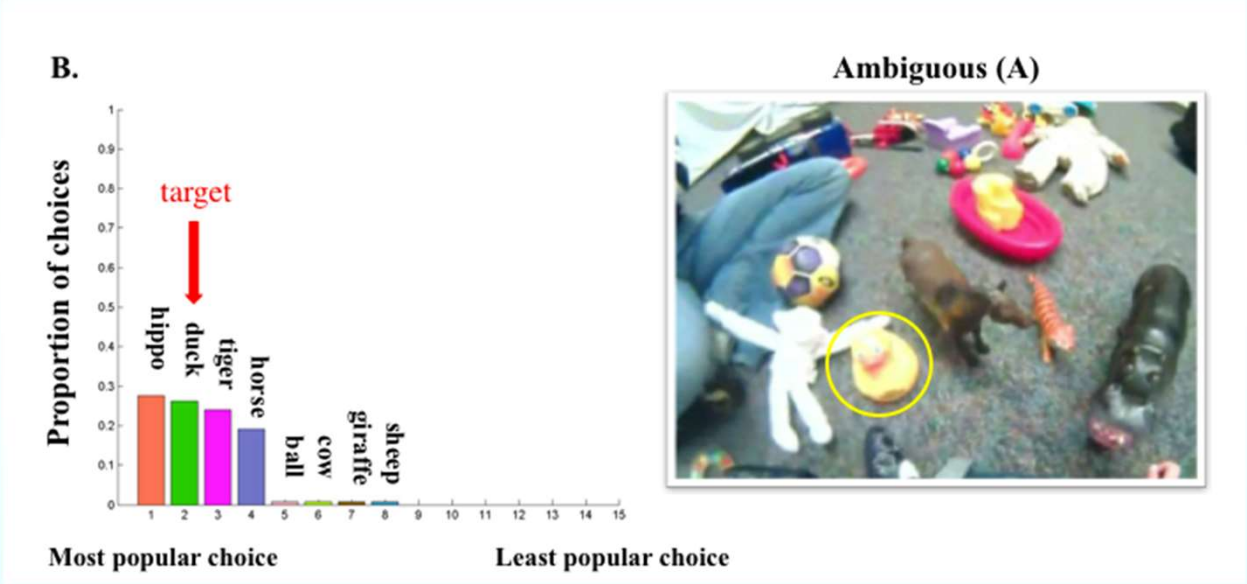


Gillette, Gleitman, Gleitman & Lederer (1999)

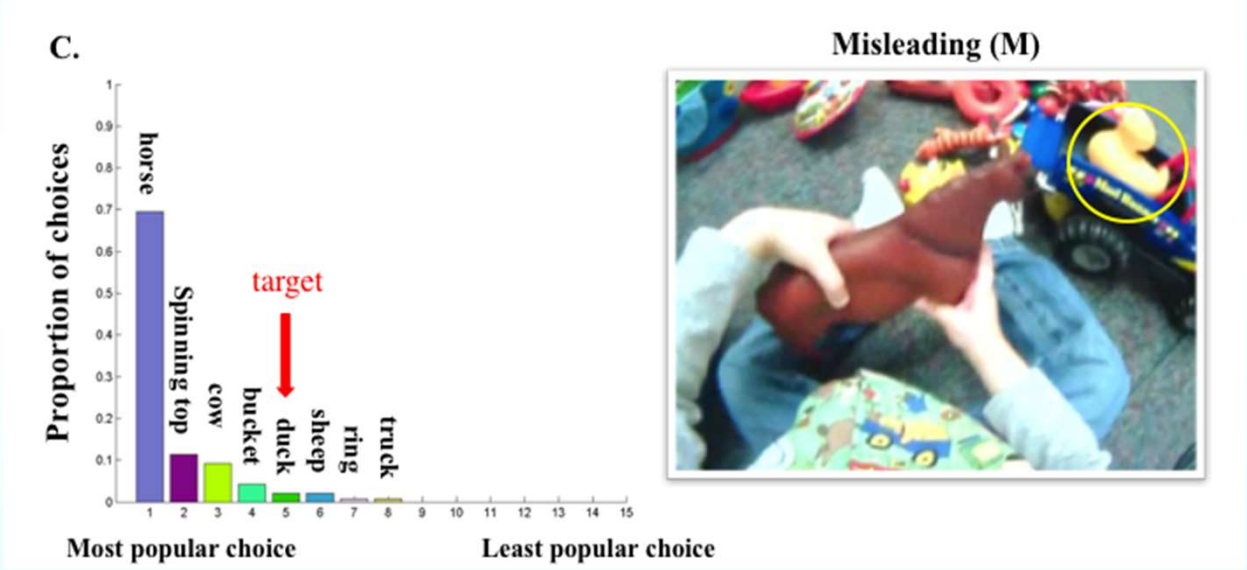
High informative case



Ambiguous case



Misleading case



Types of Individual Naming instances

- **High informative (H)** trials: above 90% accuracy. Only one object is dominant and it is the correct referent
- **Ambiguous (A):** the target toy is among the top 3 choices and accuracy of the top 1 choice does not exceed 50%. Multiple possible target options in view and participants are uncertain which one is correct
- **Misleading (M):** the target toy is not the top 1 choice and the possibility of choosing the wrong top 1 choice exceeds 50%. Only one object is dominant but it is not the correct referent.

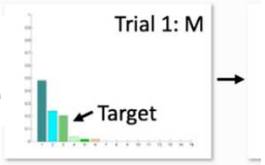
Learning from a sequence of naming instances

Trial 1: M

Statistical Learning

Trial 1: M test

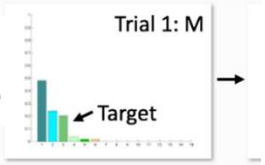
cumulative



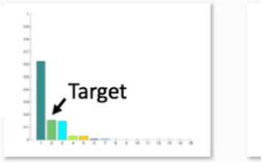
Statistical Learning

Trial 1: M test

cumulative



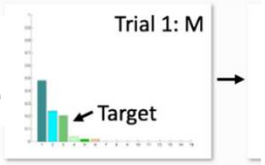
individual



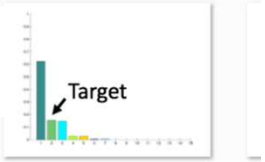
Statistical Learning

Trial 1: M test

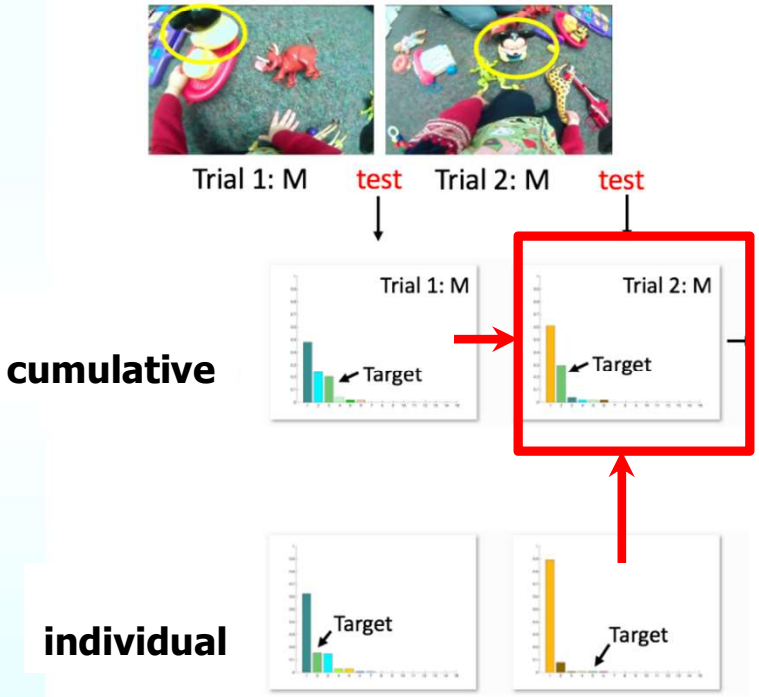
cumulative



individual



Statistical Learning

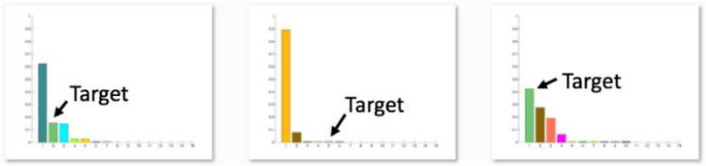


Statistical Learning

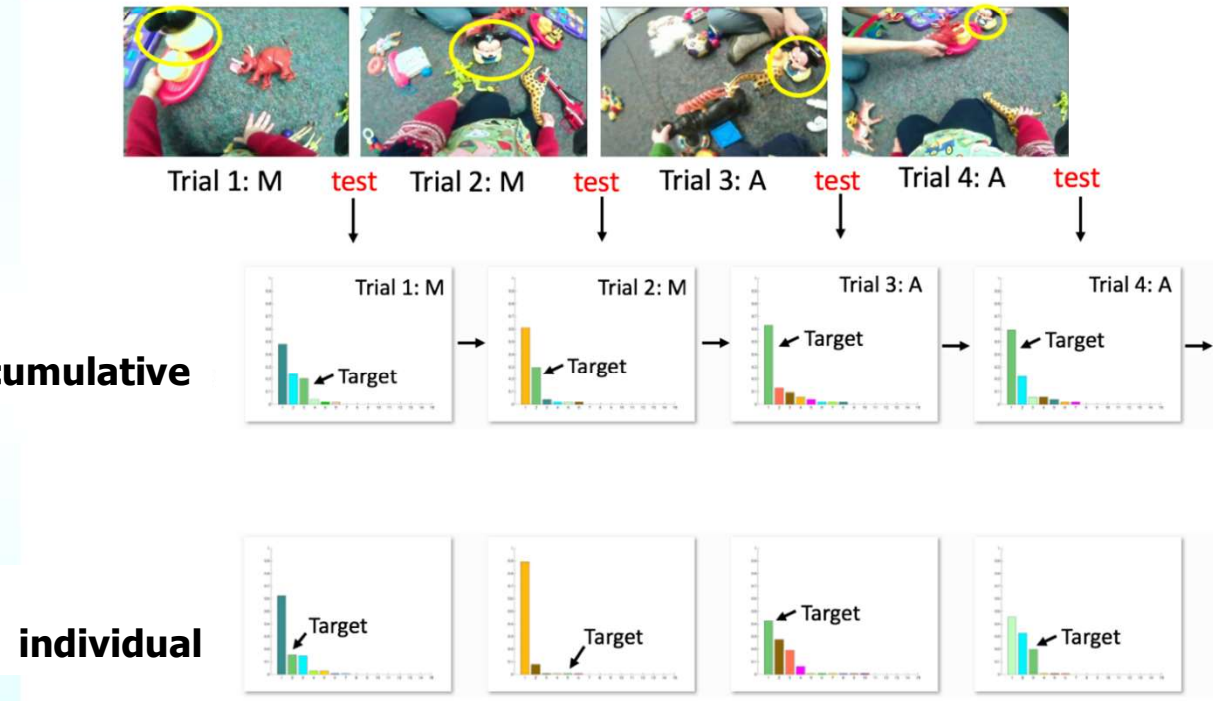
Trial 1: M test Trial 2: M test Trial 3: A test

cumulative

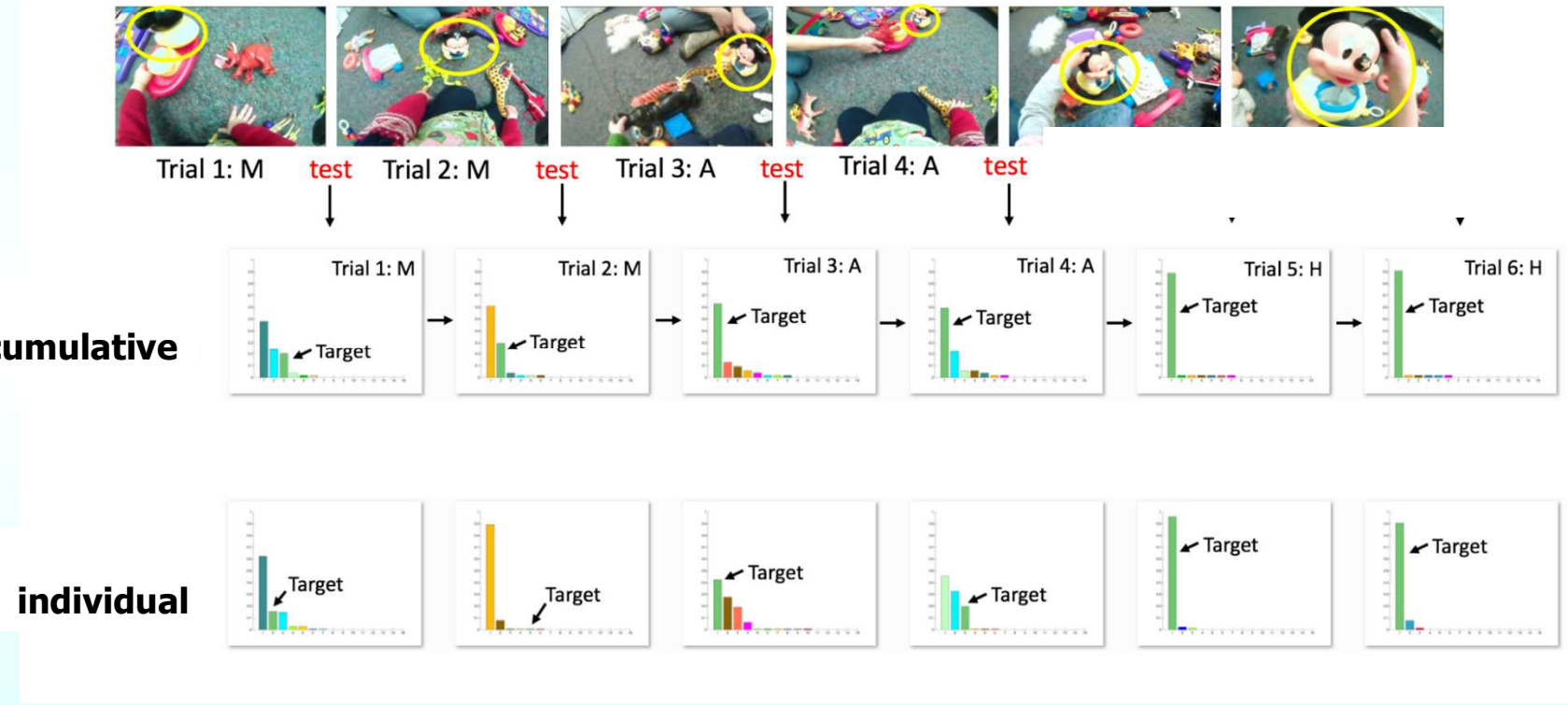
individual



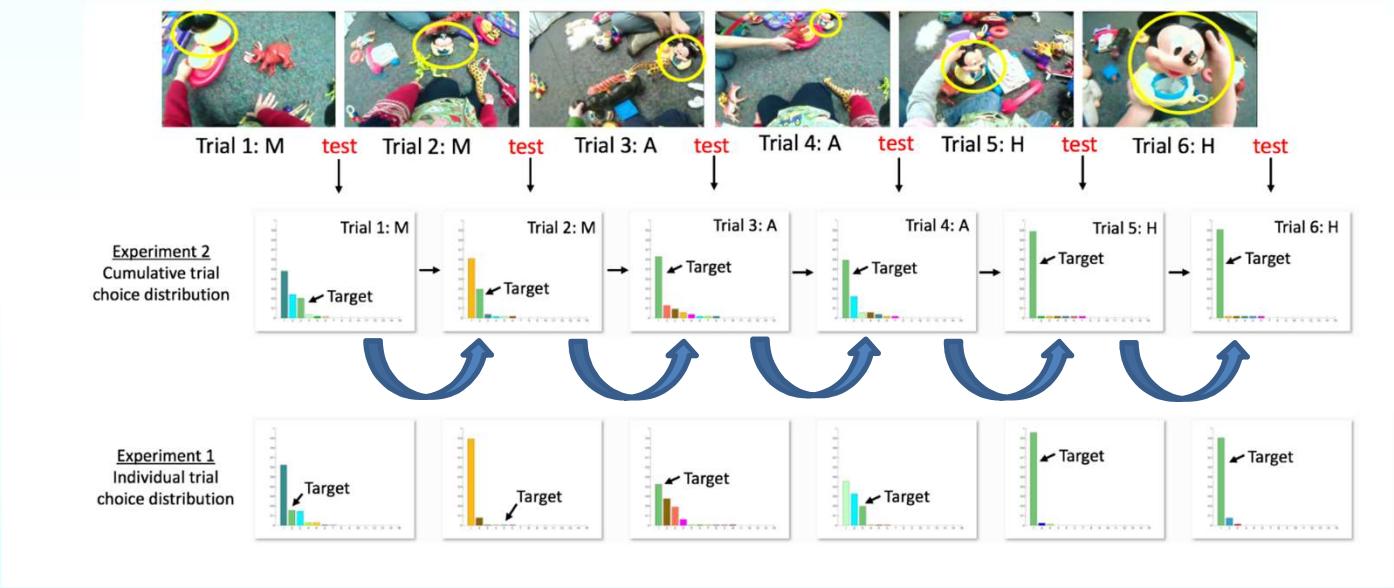
Statistical Learning



Statistical Learning



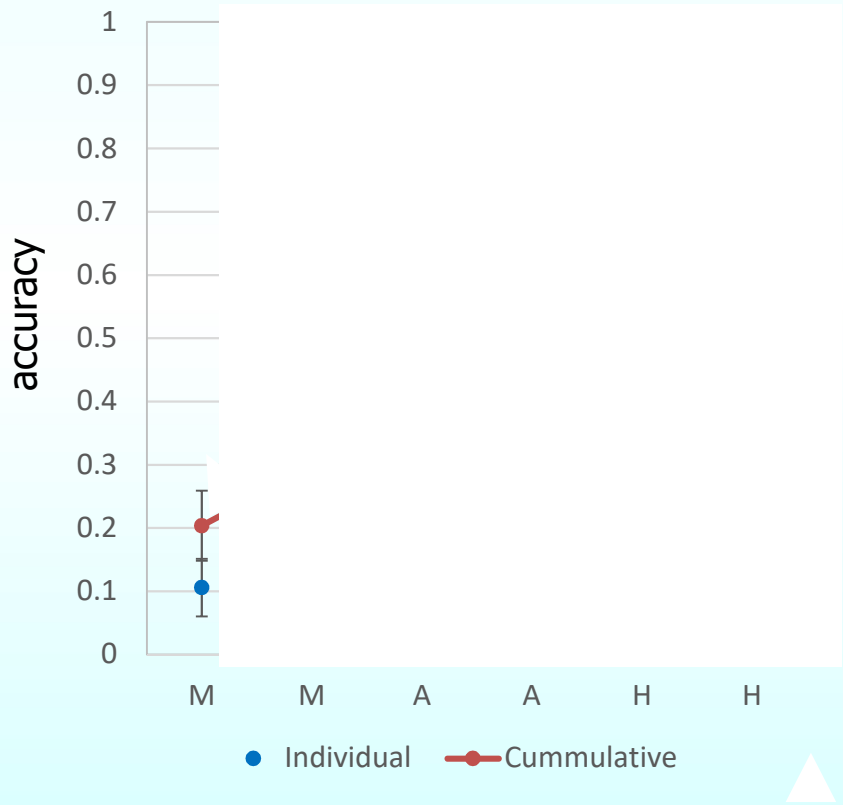
Analyses


Trial 1: M
test
Trial 2: M
test
Trial 3: A
test
Trial 4: A
test
Trial 5: H
test
Trial 6: H
test
Experiment 2
Cumulative trial
choice distribution
Trial 1: M
Target
Trial 2: M
Target
Trial 3: A
Target
Trial 4: A
Target
Trial 5: H
Target
Trial 6: H
Target
Experiment 1
Individual trial
choice distribution
Target
Target
Target
Target
Target
Target

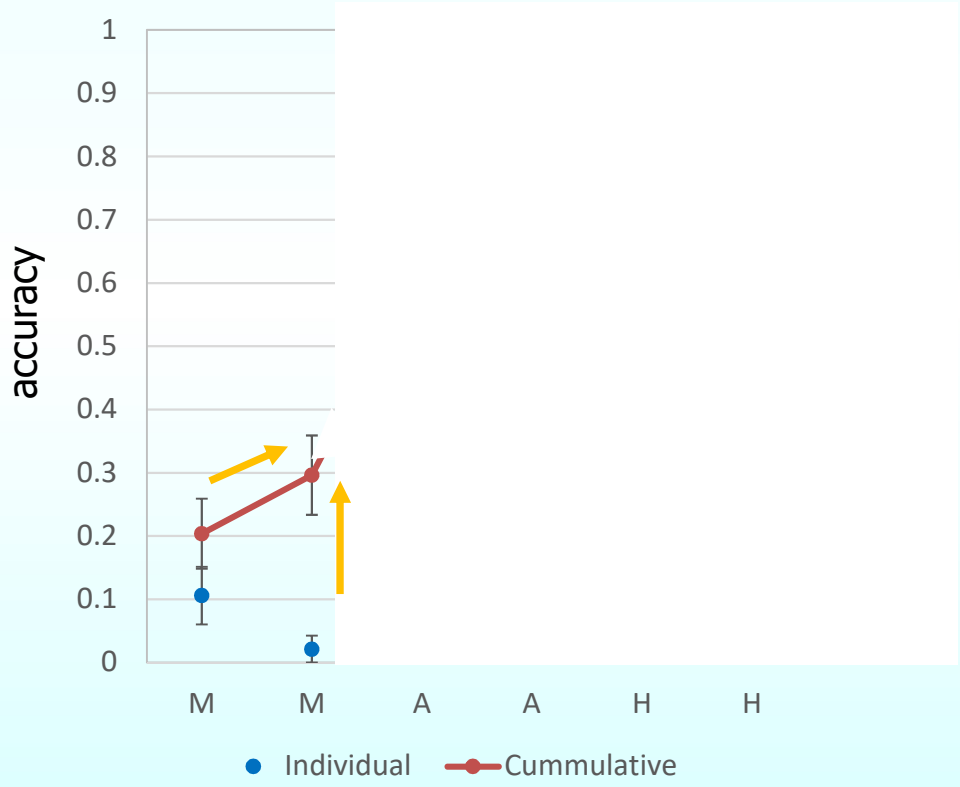
Statistical Learning Experiments

- Can learners accumulate information when they see misleading trials/ambiguous trials first?
  - MMAAHH
  - AAMMHH
- Will learners retain correct information when high informative trials are presented first, followed by misleading and ambiguous trials?
  - HHMMAA

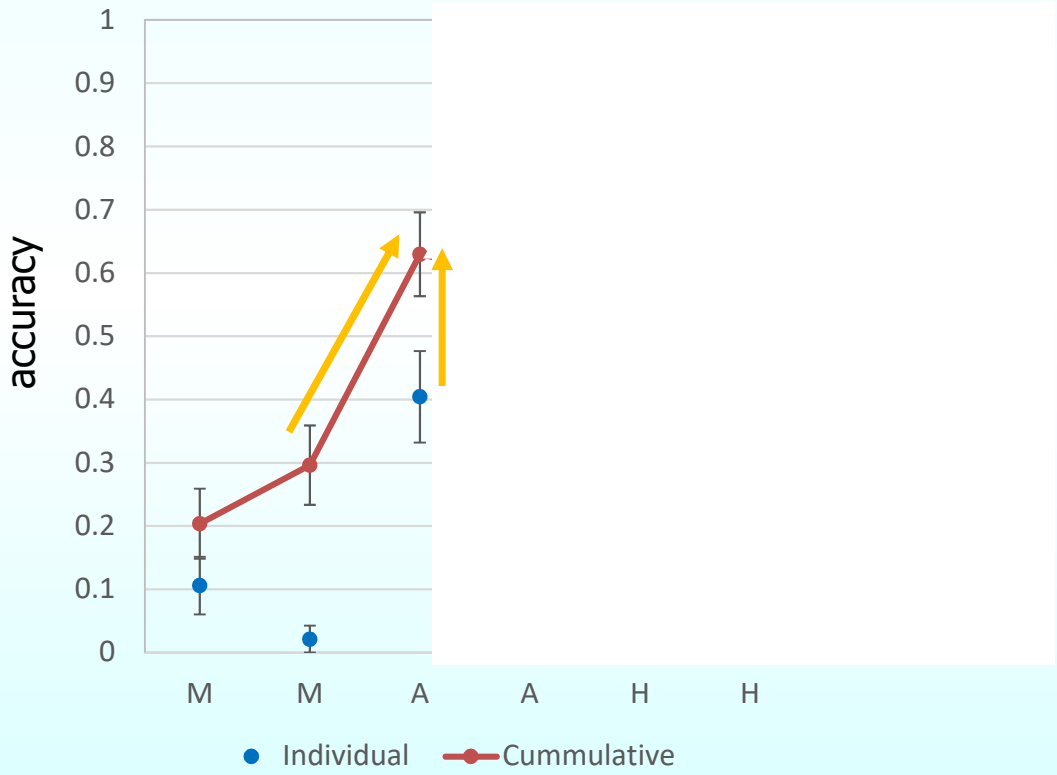
MMAAHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
M
M
A
A
H
H
Individual
Cummulative

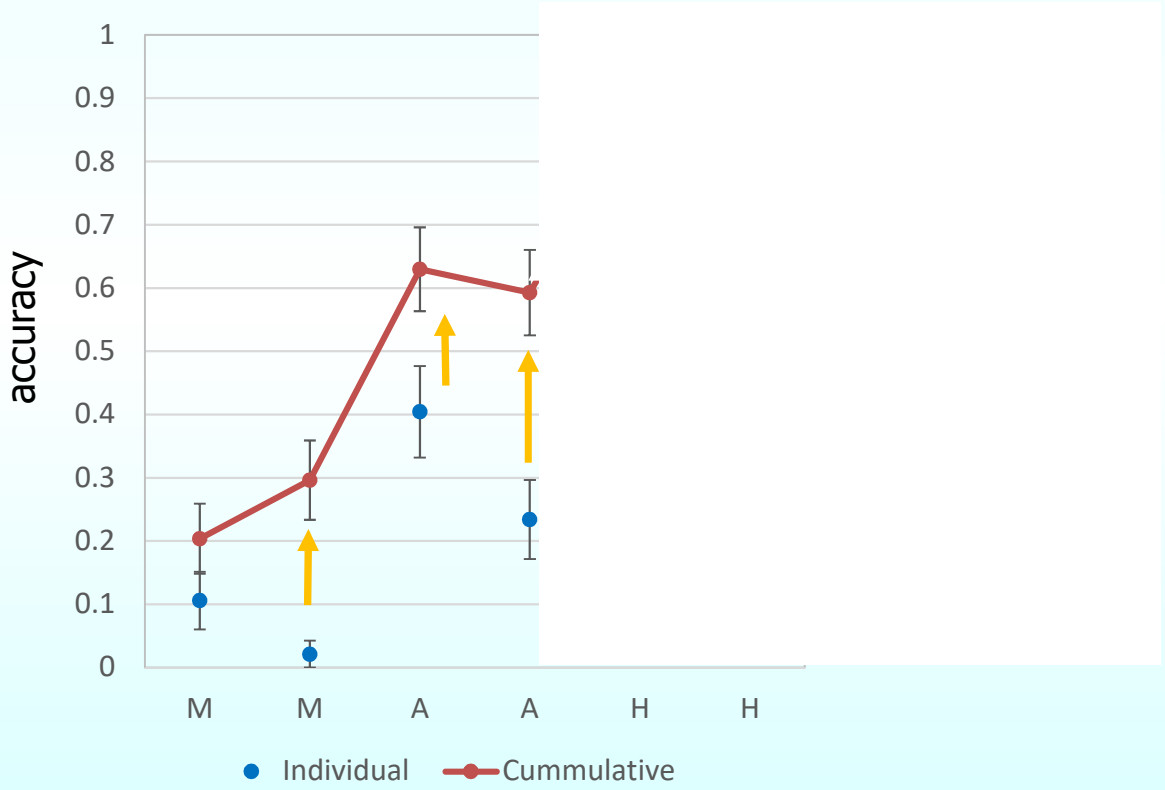
MMAAHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
M
M
A
A
H
H
Individual
Cummulative

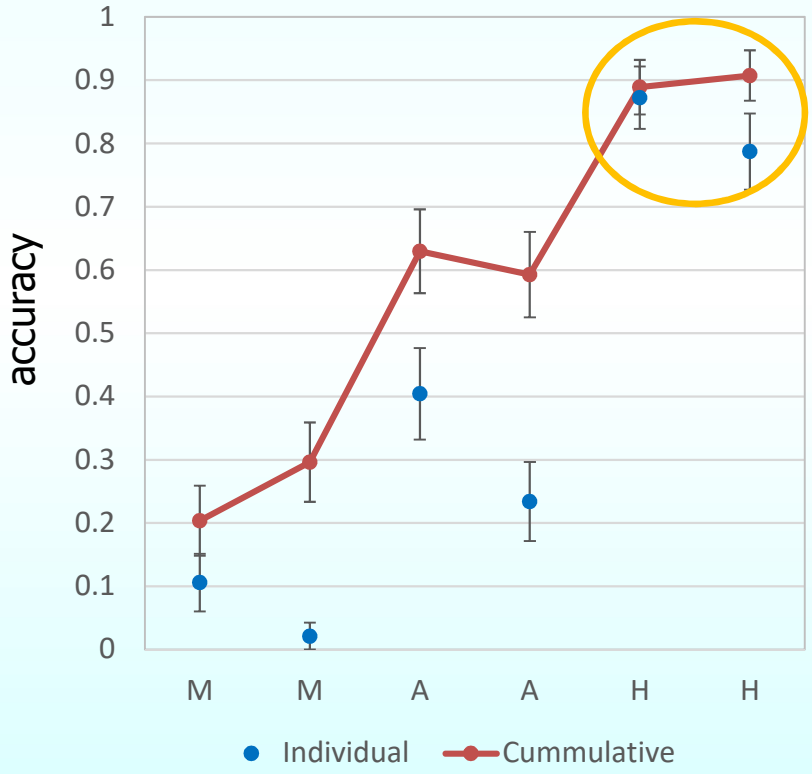
MMAAHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
M
M
A
A
H
H
Individual
Cummulative

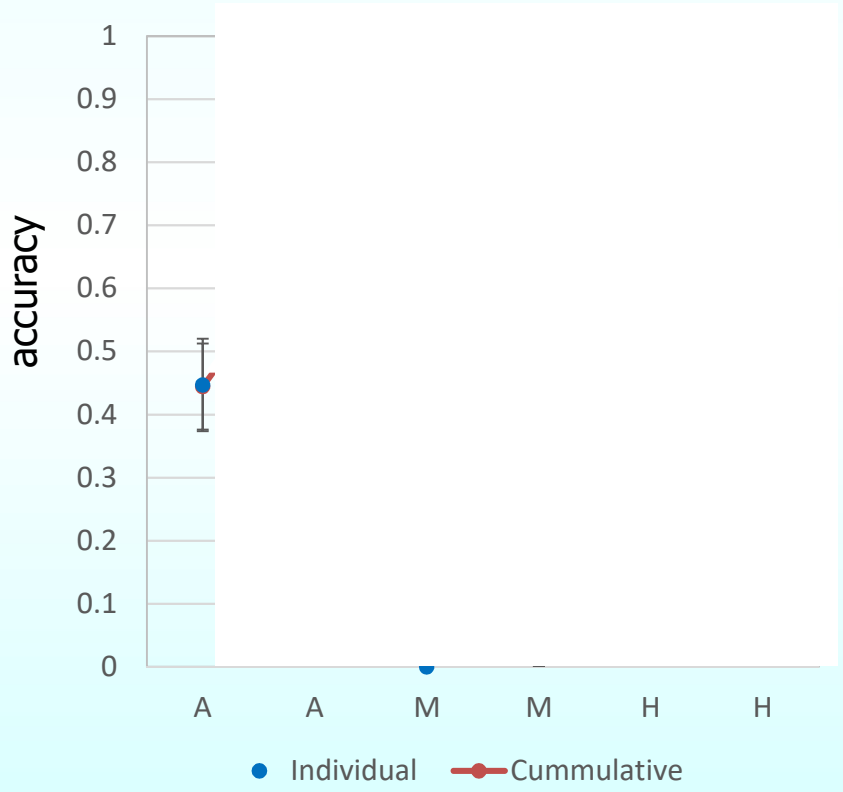
MMAAHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
M
M
A
A
H
H
Individual
Cummulative

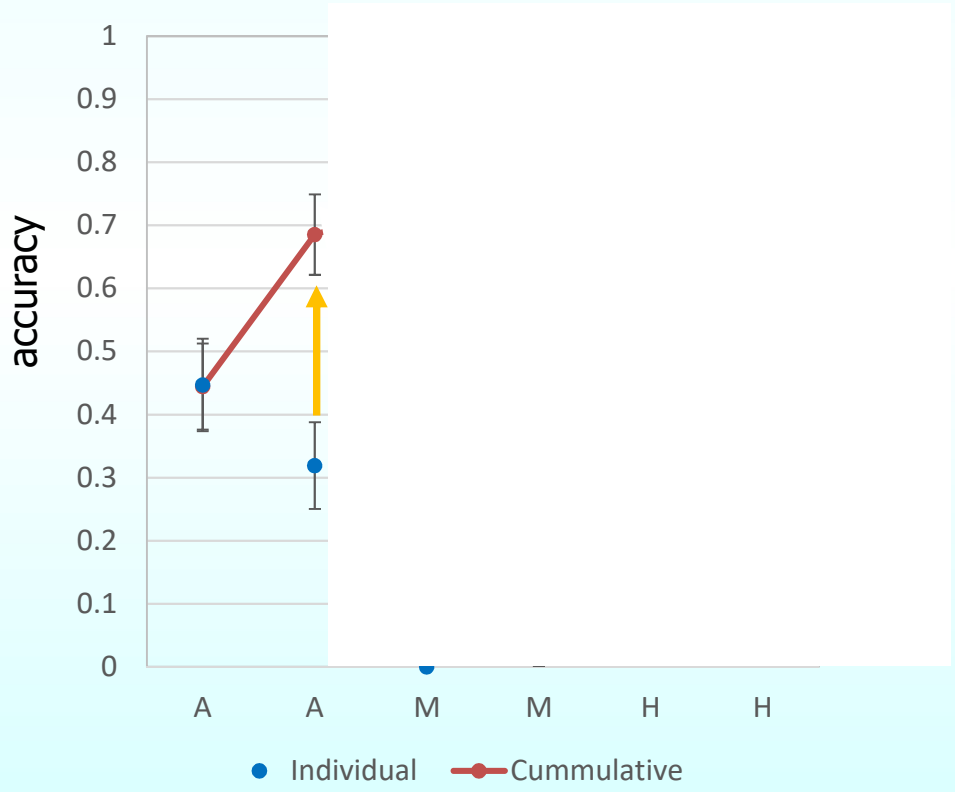
MMAAHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
M
M
A
A
H
H
Individual
Cummulative

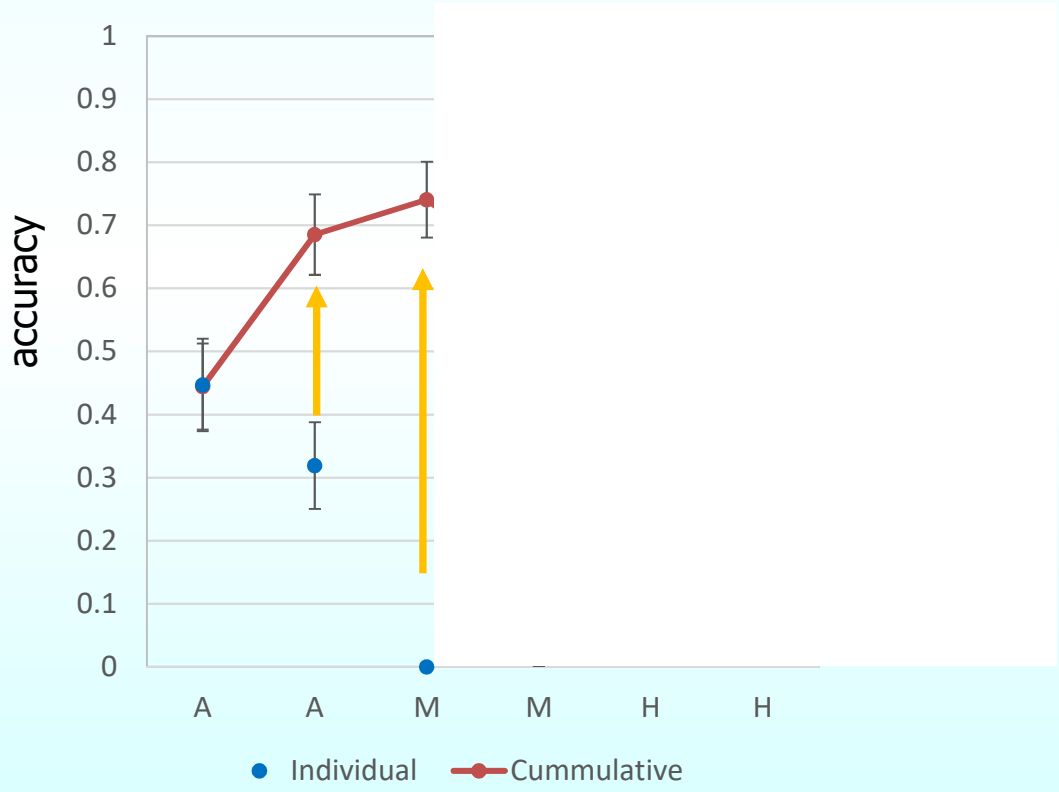
AAMMHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A
A
M
M
H
H
Individual
Cummulative

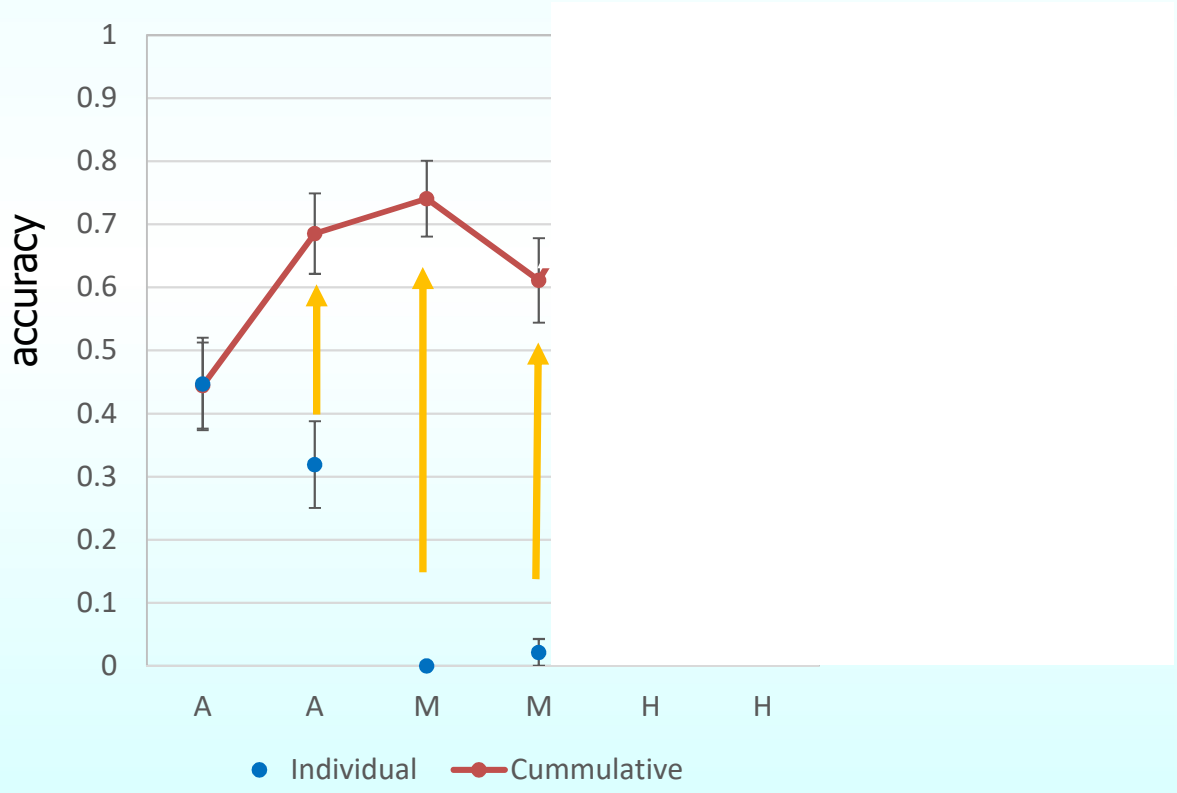
AAMMHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A
A
M
M
H
H
Individual
Cummulative

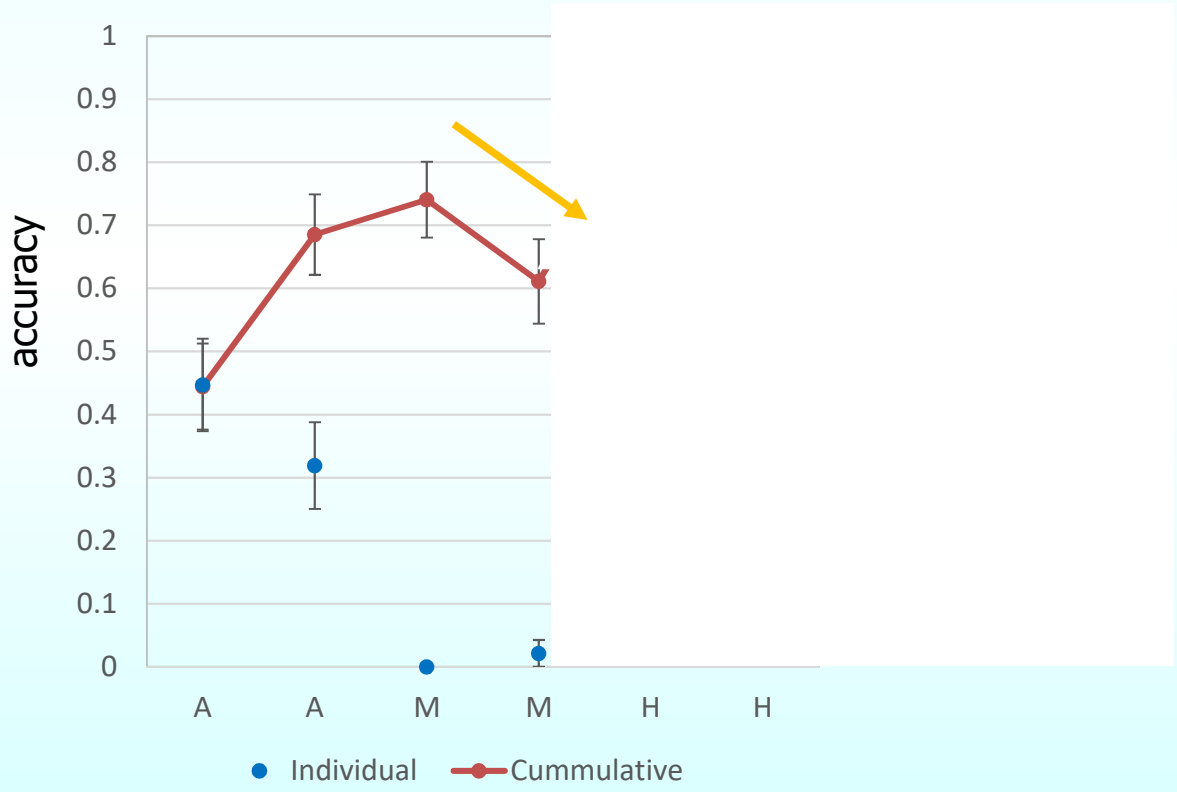
AAMMHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A
A
M
M
H
H
Individual
Cummulative

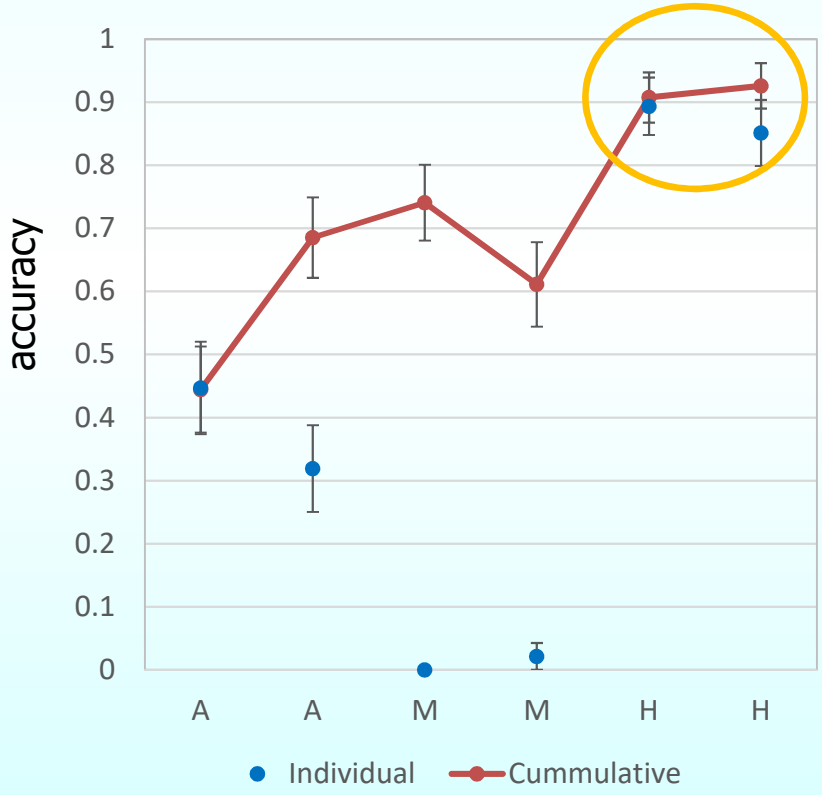
AAMMHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A
A
M
M
H
H
Individual
Cummulative

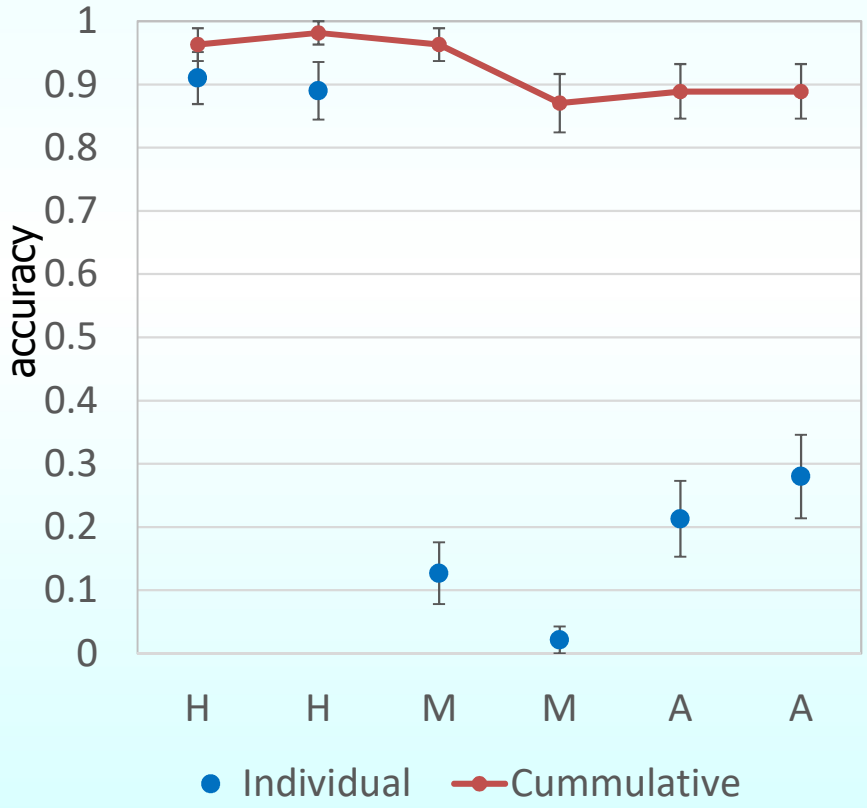
AAMMHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A
A
M
M
H
H
Individual
Cummulative

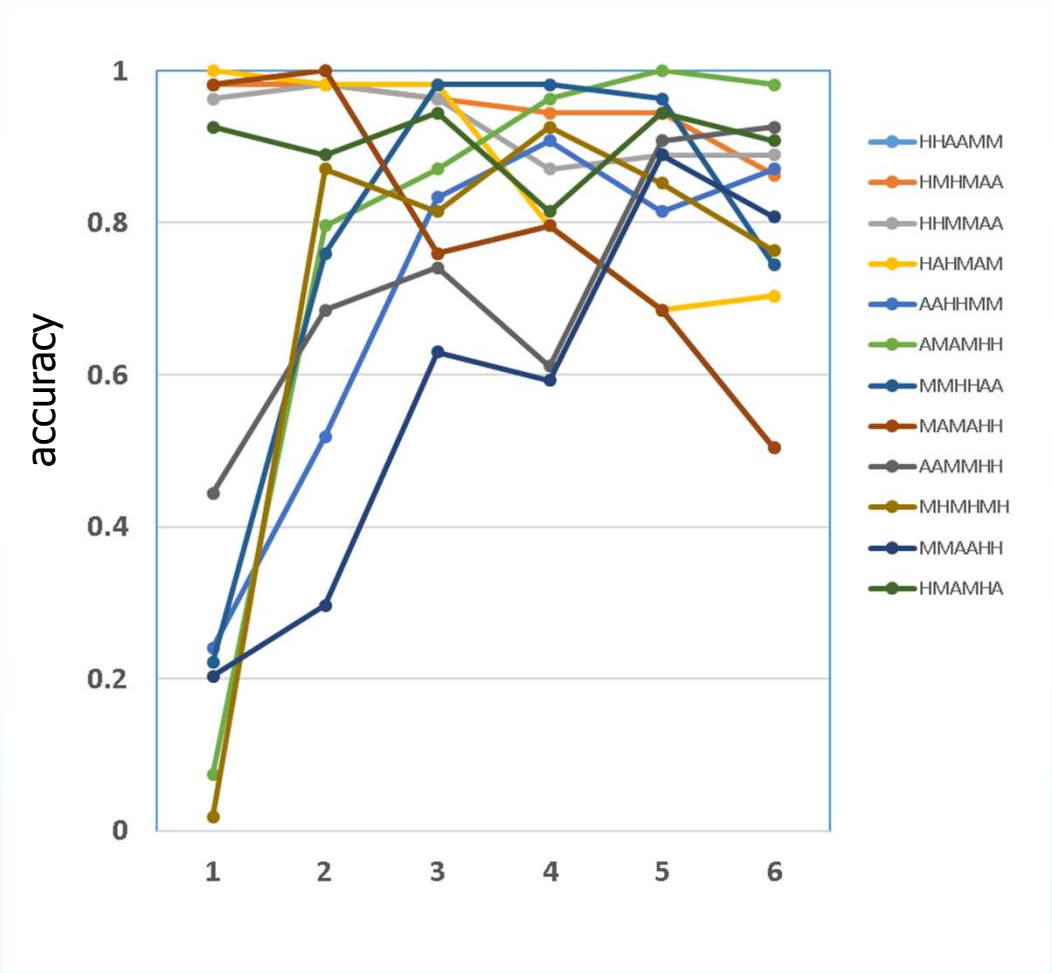
AAMMHH


accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A
A
M
M
H
H
Individual
Cummulative

HHMMAA

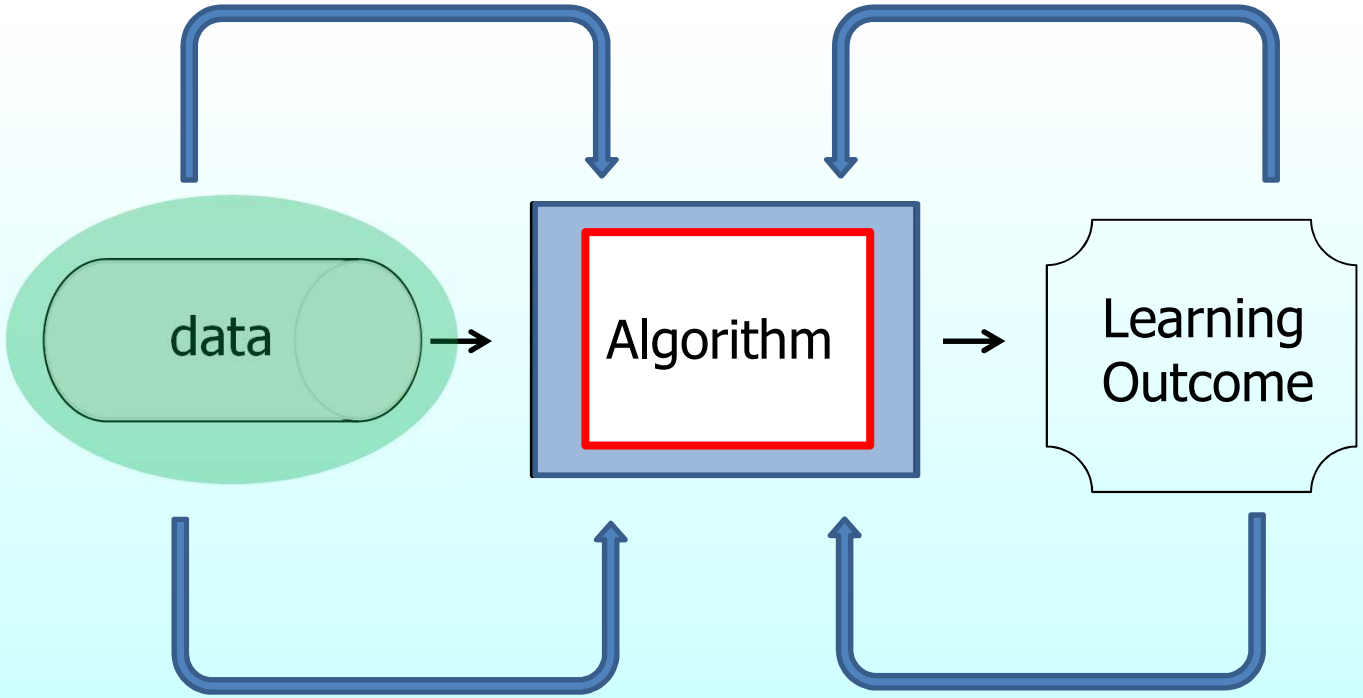

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
accuracy
H
H
M
M
A
A
Individual
Cummulative

Different orders lead to different learning outcomes



Summary

- In ML, most computational models work on batch statistics such that the order of individual learning experiences does not matter. Training materials are typically individual data points and do not present the temporal properties inherent in real-time real-world scene variation.
- Human learning systems are sensitive to timing and order of the data that matter to statistical learning (Zhang & Yu, 2016).



Study 3: Effects of Temporal Structure on Infant Word Learning

Lauren Slone

Drew Abney

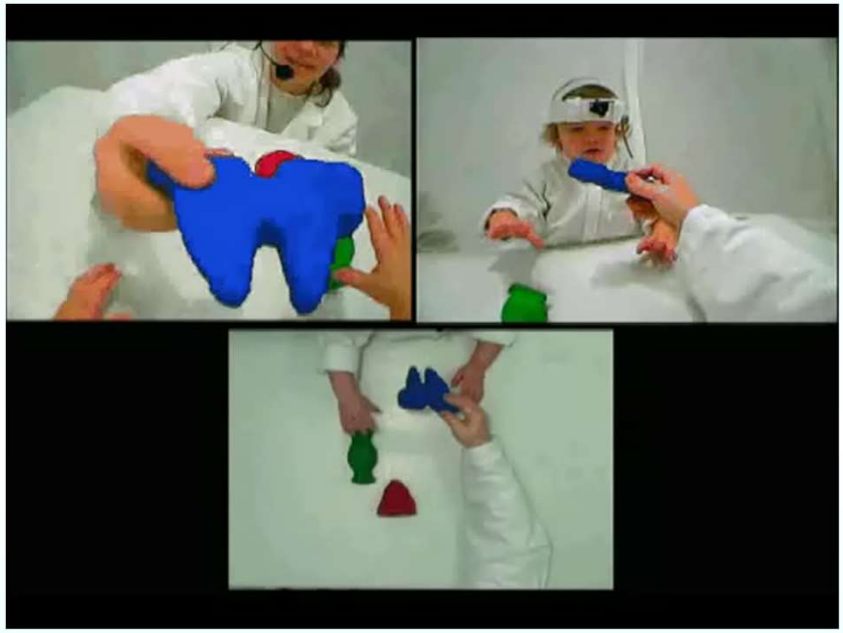
Words are not randomly sampled but rather are clustered in time (e.g., Altmann, Cristadoro, & Degli Esposti, 2012; Altmann, Pierrehumbert, & Motter, 2009; Church & Gale, 1995; Katz, 1996).

- i.e., the probability of hearing a word (e.g., "spoon") is higher if you just heard it compared to its probability in the distribution as a whole.

This is likely a property of parents' speech to their children (e.g., Brodsky, Waterfall, & Edelman, 2007).

→ Does this matter for children's word learning?

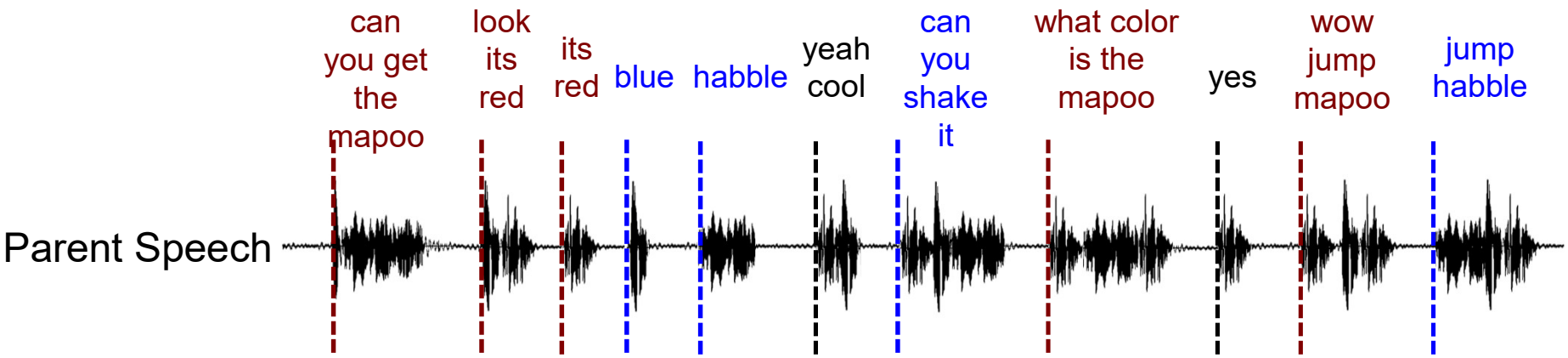
Effects of Temporal Structure on Infant Word Learning



- What is the temporal structure of parents' talk about objects to their infants during naturalistic play?
- Does the temporal structure of parents' talk about a novel object relate to infants' learning of that object's name?

Assessing the Temporal Structure of Parent Speech: IOI

IOI = Inter-Onset Interval of utterances



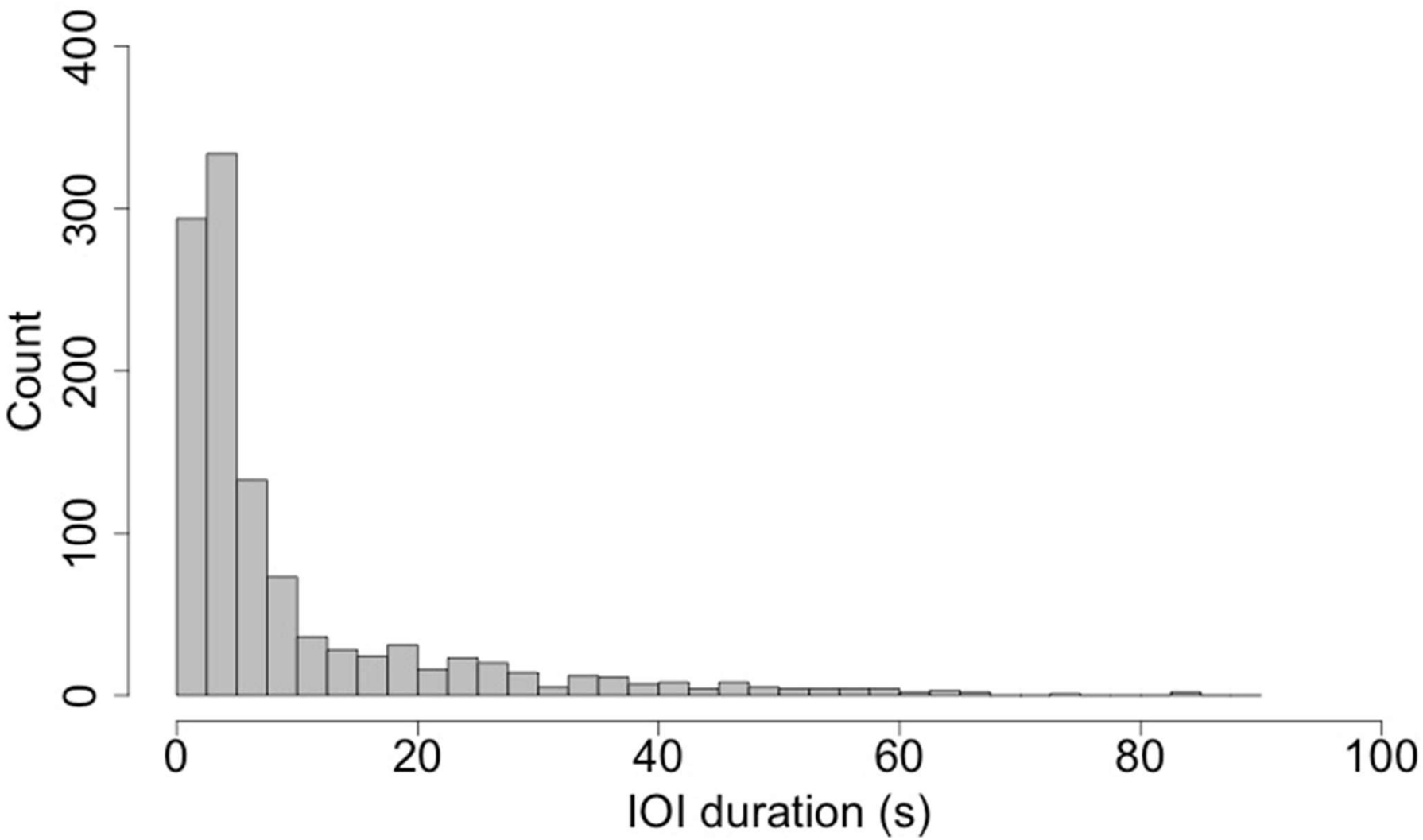
Object 1 (habble):

Object 2 (wawa):

Object 3 (mapoo):

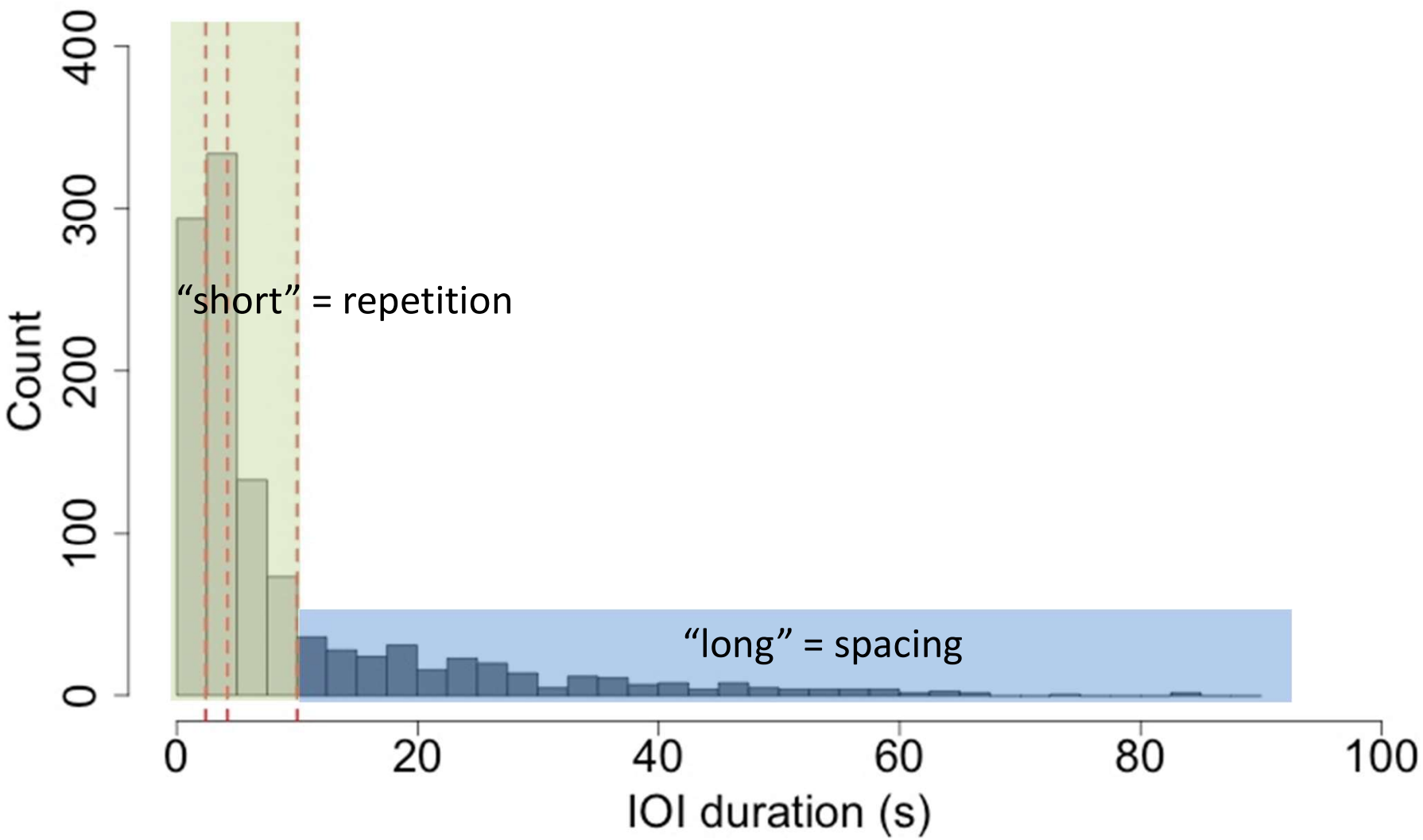
Assessing the Temporal Structure of Parent Speech

Dataset 1: n = 1112 IOIs



Assessing the Temporal Structure of Parent Speech

Dataset 1: n = 1112 IOIs

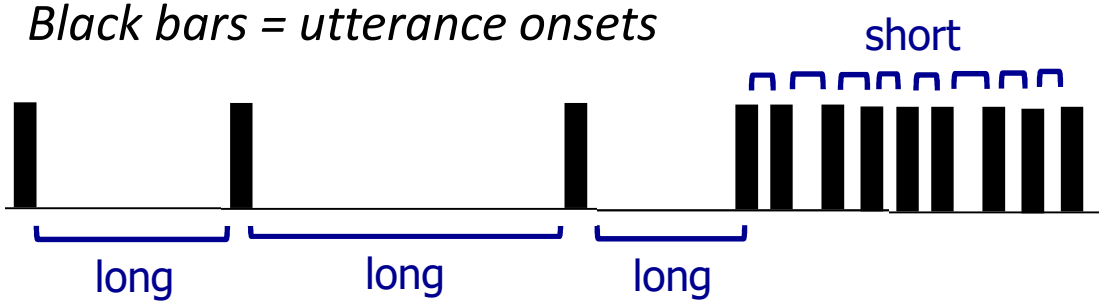


Assessing the Temporal Structure of Parent Speech

How do the IOIs go together for individual object talk distributions?

How do spacing (long IOIs) and repetition (short IOIs) play out *in time*?

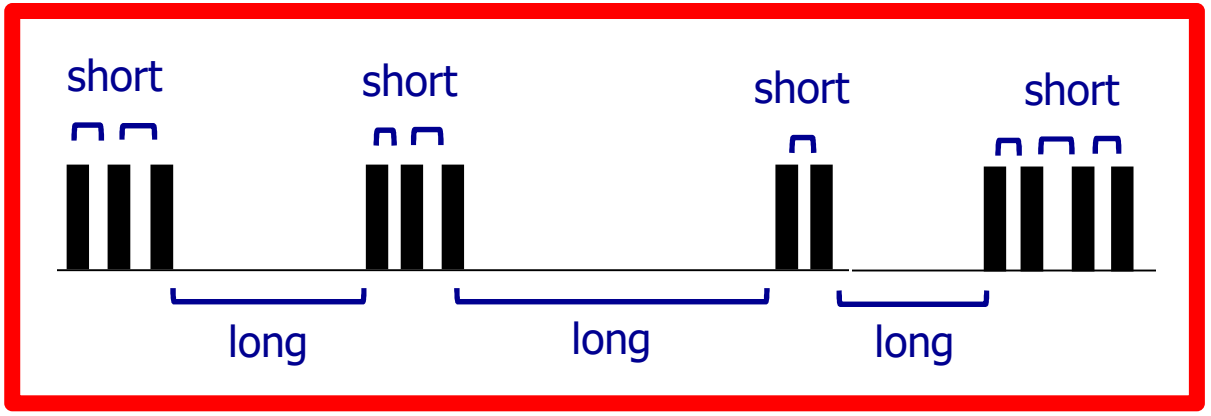
Black bars = utterance onsets



M = 10.7 IOIs per object

2.9 long IOIs of
$M_{duration}$= 30.7 s

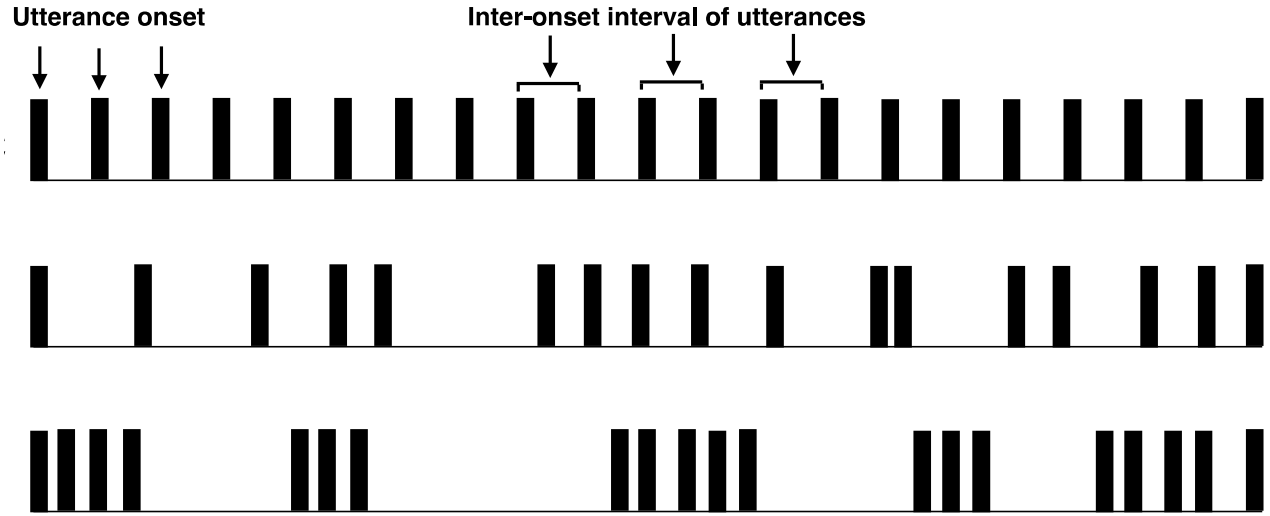
7.8 short IOIs of
$M_{duration}$= 4.2 s



How many short IOIs are between pairs of long IOIs?

M = 4.1
SD = 0.7

Quantify temporal structure in single measure: *Burstiness*



Periodic: $B = -1$

Poisson: $B = 0$

Bursty: $B = 1$

Infinite time series
(Goh & Barabasi, 2008)

$$B = \frac{\sigma - \mu}{\sigma + \mu} = \frac{r - 1}{r + 1}$$

Finite time series
(Kim & Jo, 2016)

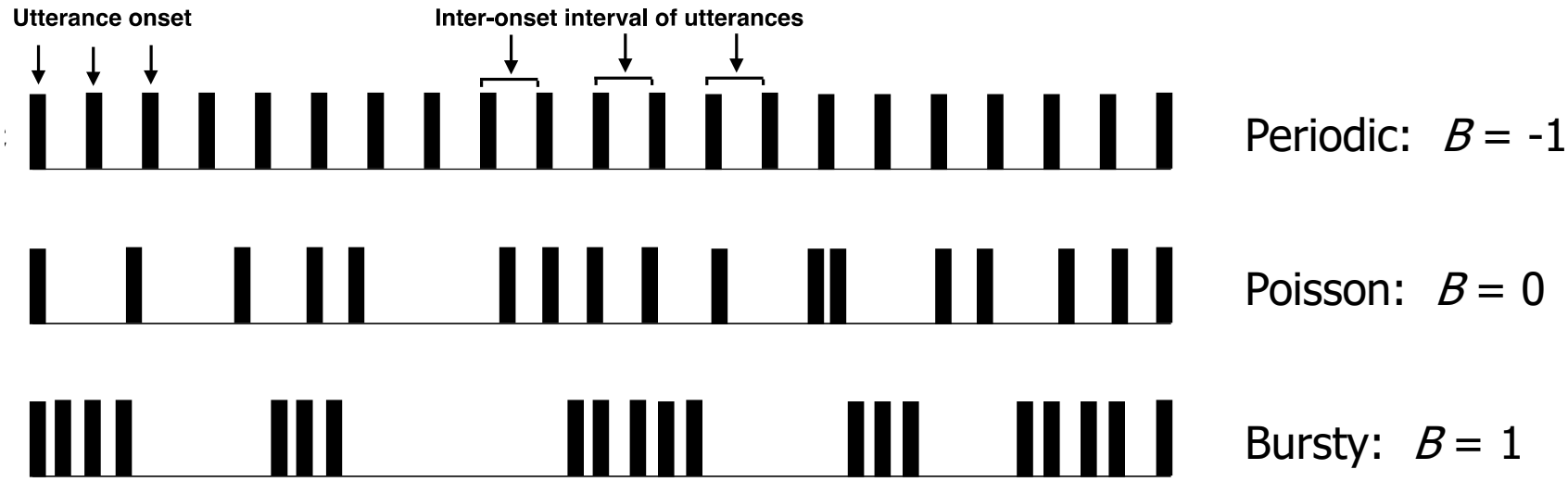
$$B = \frac{\sqrt{n+1}r - \sqrt{n-1}}{(\sqrt{n+1} - 2)r + \sqrt{n-1}}$$

Quantify temporal structure: Burstiness

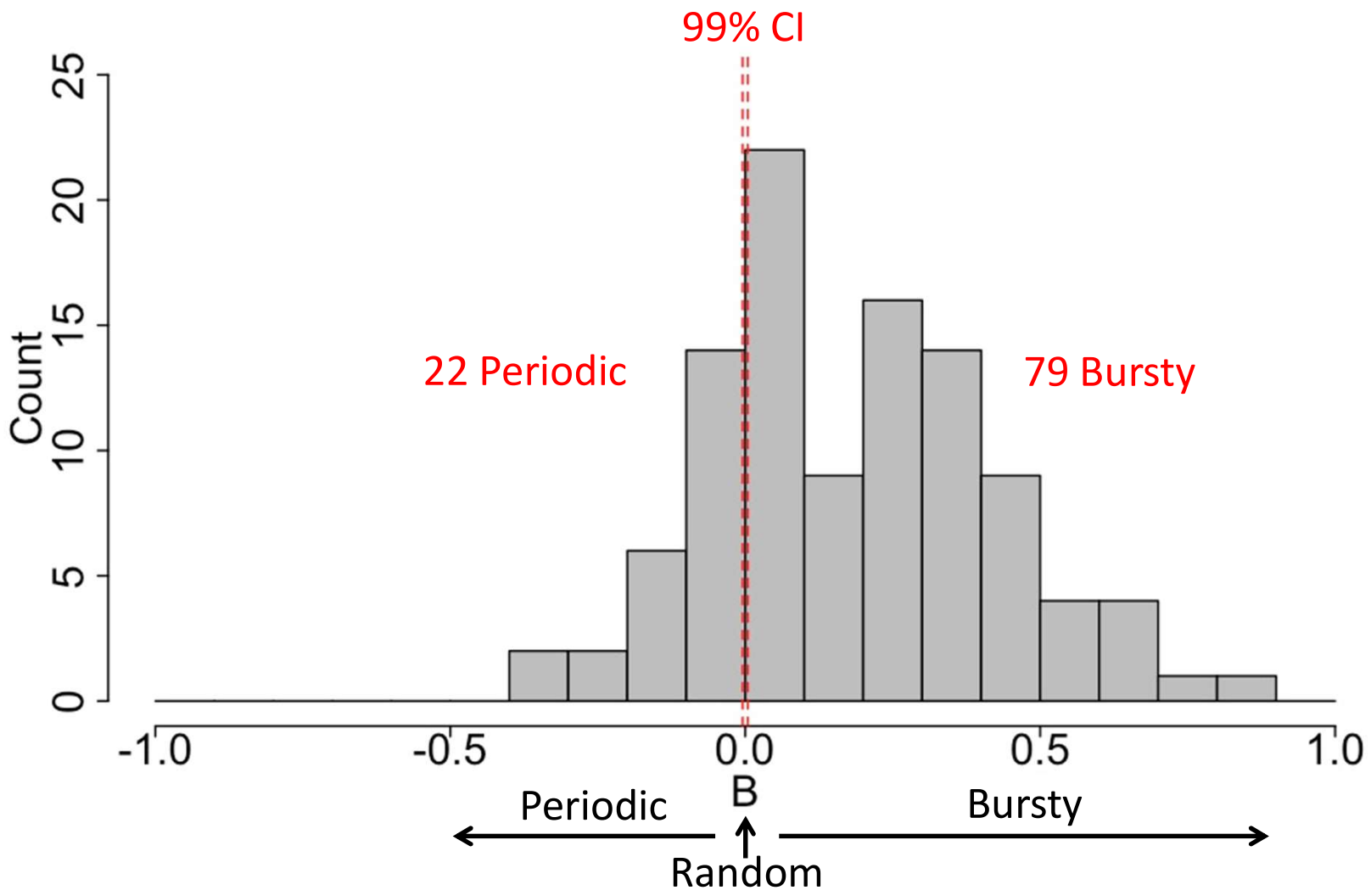
1. What is the temporal structure of parents' speech to their infants during naturalistic play?

It is predominantly *bursty:*

Mean B = .18 (95% CI = .13 - .23)



Classify temporal structure of talk about each object



Interim Discussion

Almost all parent talk to their infant deviated from a Poisson process to at least some extent, indicating that infants' language input was structured in time.

Parents' talk about individual objects was predominantly bursty.

- Bursty talk is characterized by clusters of utterances about the same object, spaced out in time.

Some objects were talked about in a periodic fashion.

- Periodic talk is characterized by more equal spacing of utterances about the same object.

Results: Relating Burstiness to Word Learning

2. Does the temporal structure of parents' talk about a novel object relate to infants' learning of that object's name?

i.e., Did an infant show differential learning of the objects talked about in a bursty vs. periodic way?

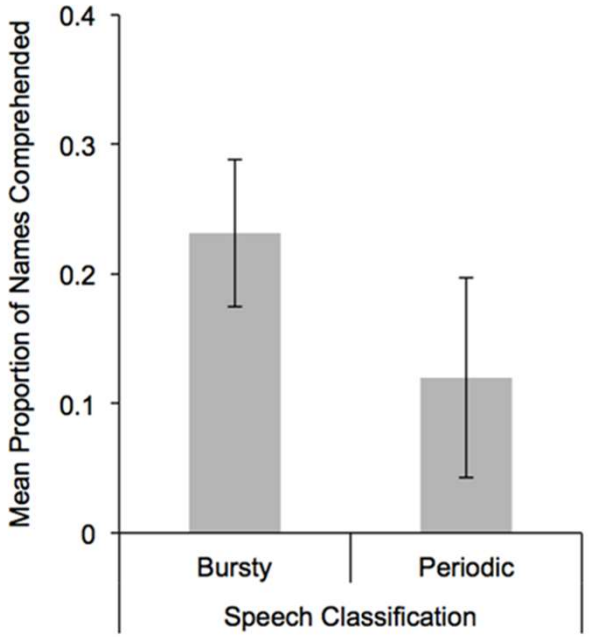


Figure 2. Mean comprehension scores for names of novel objects that parents talked about with Bursty and Periodic temporal structure. Error bars represent 95% confidence intervals.

Discussion

Bursty talk about an object promoted better learning of that object's name than did periodic talk.

Why?

Bursty distributions' combination of **repetition *and* spacing** may promote initial encoding and short-term retention of object-label pairs, as well as longer-term retention of those pairs.

Discussion

Repetition may help initial encoding and short-term retention of word-object pairs, by:

- enhancing processing of repeated object labels (Weisleder & Fernald, 2014)
- promoting visual attention to the referent (Suanda, Smith, & Yu, 2016)
- facilitating aggregation of cross-situational statistics, to resolve ambiguity & promote word-object mappings (Vlach & Johnson, 2013; Kachergis, Yu, & Shiffrin, 2009 (in adults))

Discussion

Spacing may promote long-term retention of word-object pairs, by:

- increasing encoding variability (Melton, 1970; Glenberg; 1979)
  - increases retrieval cues, which increase recall
- allowing time for forgetting, which promotes abstraction (Vlach, Sandhofer, & Kornell, 2008)
  - abstract memories are more durable than concrete memories (Brainerd & Reyna, 2002)
- fostering retrieval practice (Benjamin & Tullis, 2010; Thios & D'Agostino, 1976; Vlach et al., 2008, 2012).
- allowing time for consolidation (e.g., Atkinson & Shiffrin, 1968; Landauer, 1969; Wickelgren, 1970)
- facilitating discrimination learning (e.g., Birnbaum, Kornell, Bjork, & Bjork, 2013; Goldstone, 1996)
  - greater potential for juxtaposition of different objects, which highlights differences across objects

Human learning **Machine learning**

"We hope to be able to build a program that can learn, as a child does… instead of being spoon-fed the tremendous information necessary."

-- R.C. Schank (1972)

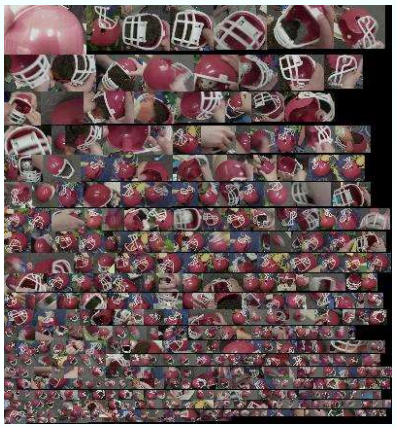
"…Only then may we be able to build intelligent machines that could learn to see—and think—without the need to be programmed to do it."

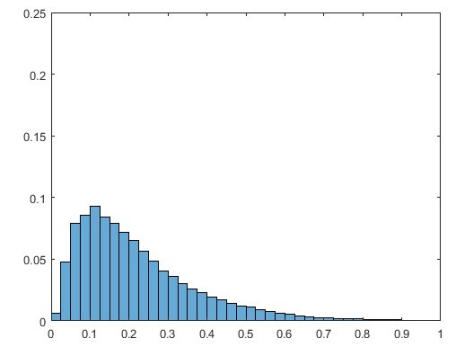
—T. Poggio (2010)

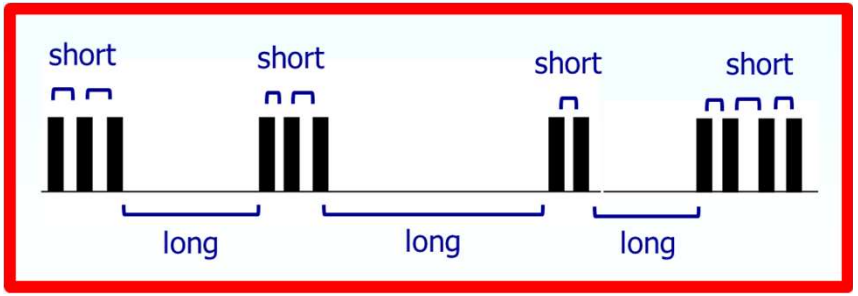
principles

Human learning

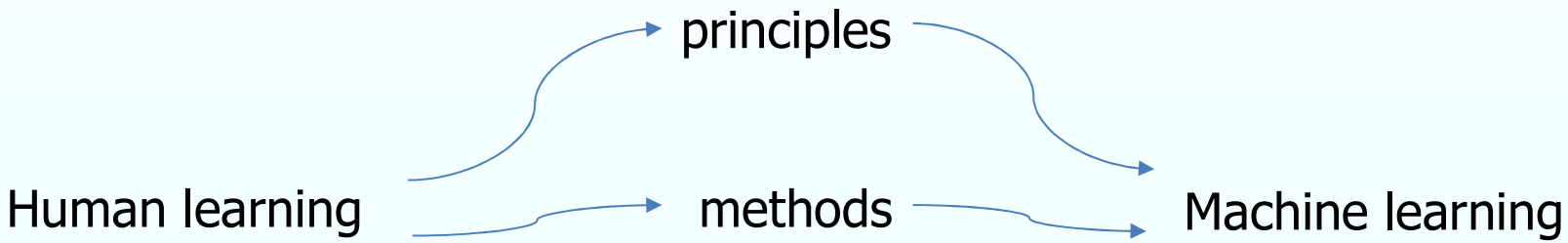
Machine learning

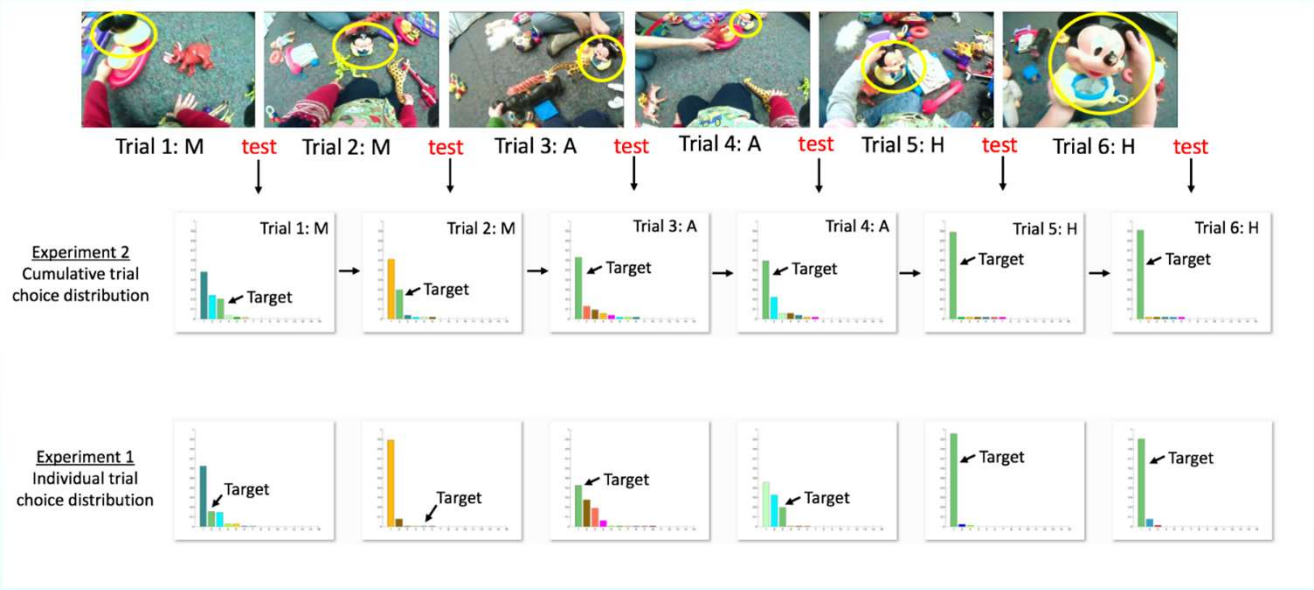




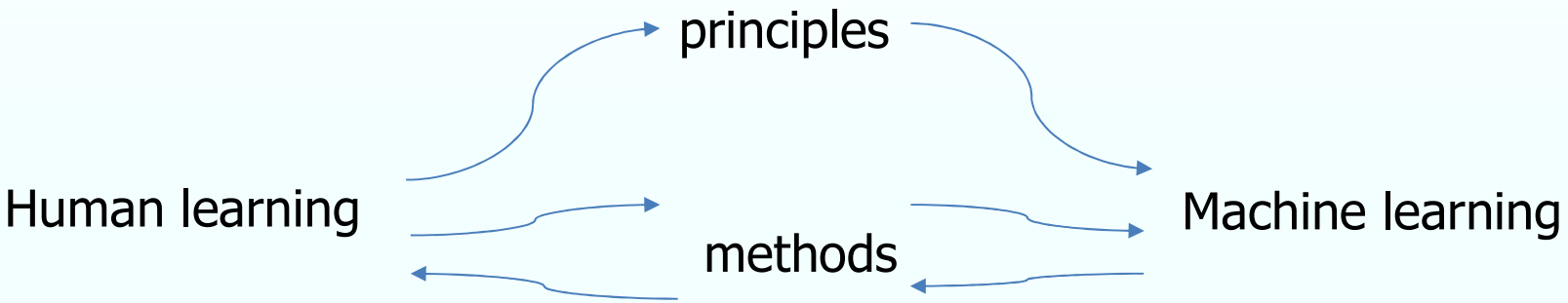


The samples of instances that young perceivers encounter are neither uniform, nor normal, and nor a random sample of the physical environment. Instead, it is constrained by the limits of time and place and by the young child's body, activities, and needs. The resulting distributional properties of experienced instances may not solve all the learning problems, but we propose that active learning makes those problems easier to solve.

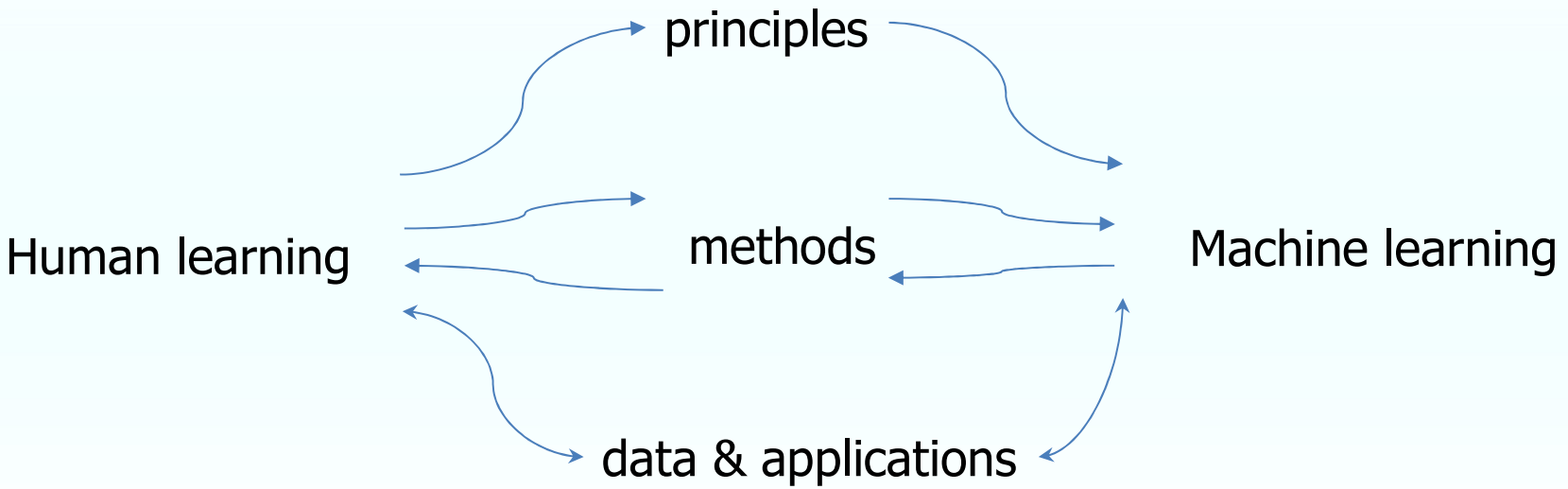




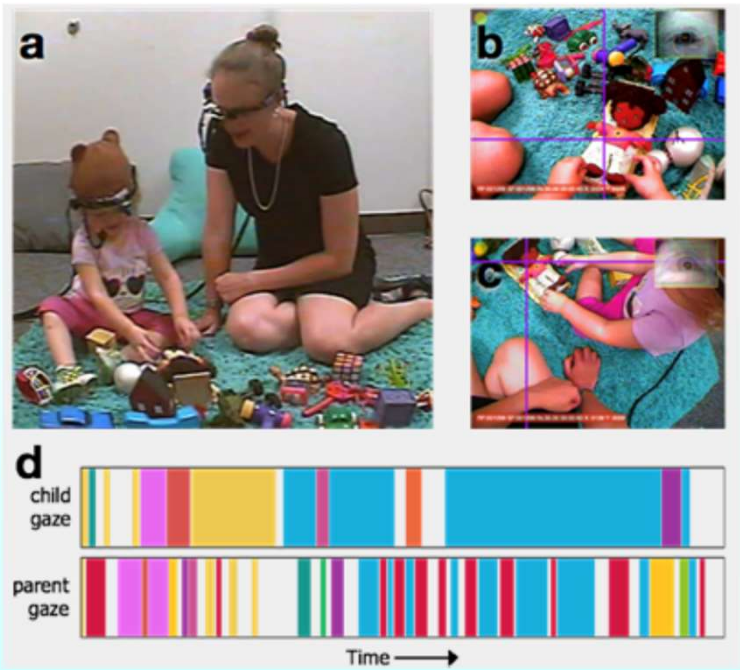
Experimental Psychology Approach



Train



Children with cochlear implants



Children with ASD

Human learning

Machine learning

Acknowledgement

Collaborators:


Linda B. Smith
David Crandall
Dan Kennedy
Derek Houston (OSU)
Irina Castellanosa (OSU)
Craig Ericsson (Univ. of Cincinatti)

Lab technicians:
Seth Foster (Duke Univ.)
Melisa Elston (Cornell Univ.)
Steven Elmlinger (Cornell Univ.)
Anting Chen (Univ. of Washington)
Melissa Hall
Daniel Pearcy

Postdoc Scientists & Graduate students
Umay Suanda (Univ. of Connecticut)
John Franchak (Univ. of California at Riverside)
Sven Bambach
Drew Abney
Lauren Slone
Lei Yuan
Tian (Linger) Xu
Charlene Ty
Yayun Zhang
Catalina Suarez

This research was supported by NIH R01HD074601H, R01 HD028675, and NSF BCS-15233982