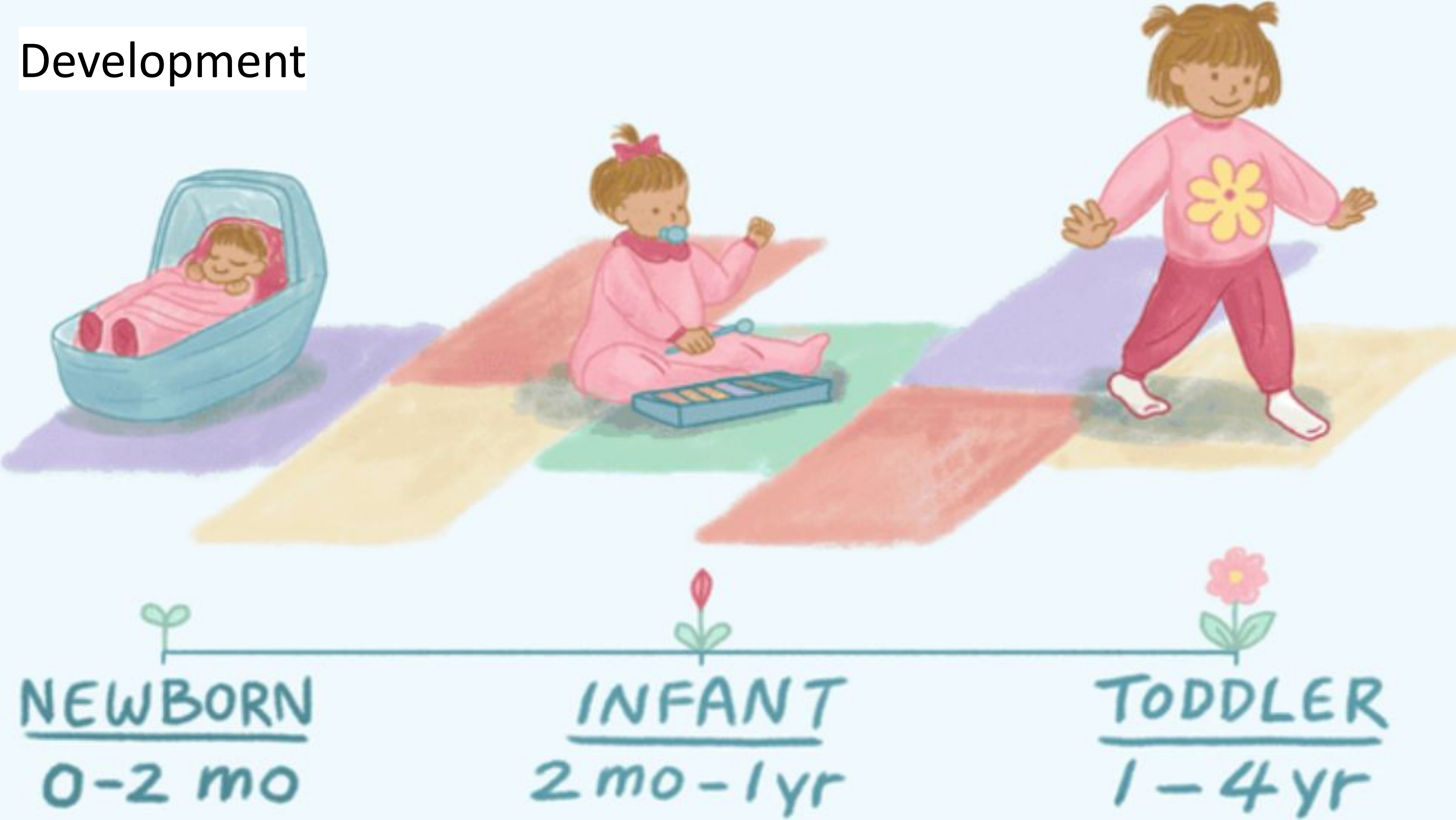
The developing infant creates a curriculum for statistical learning

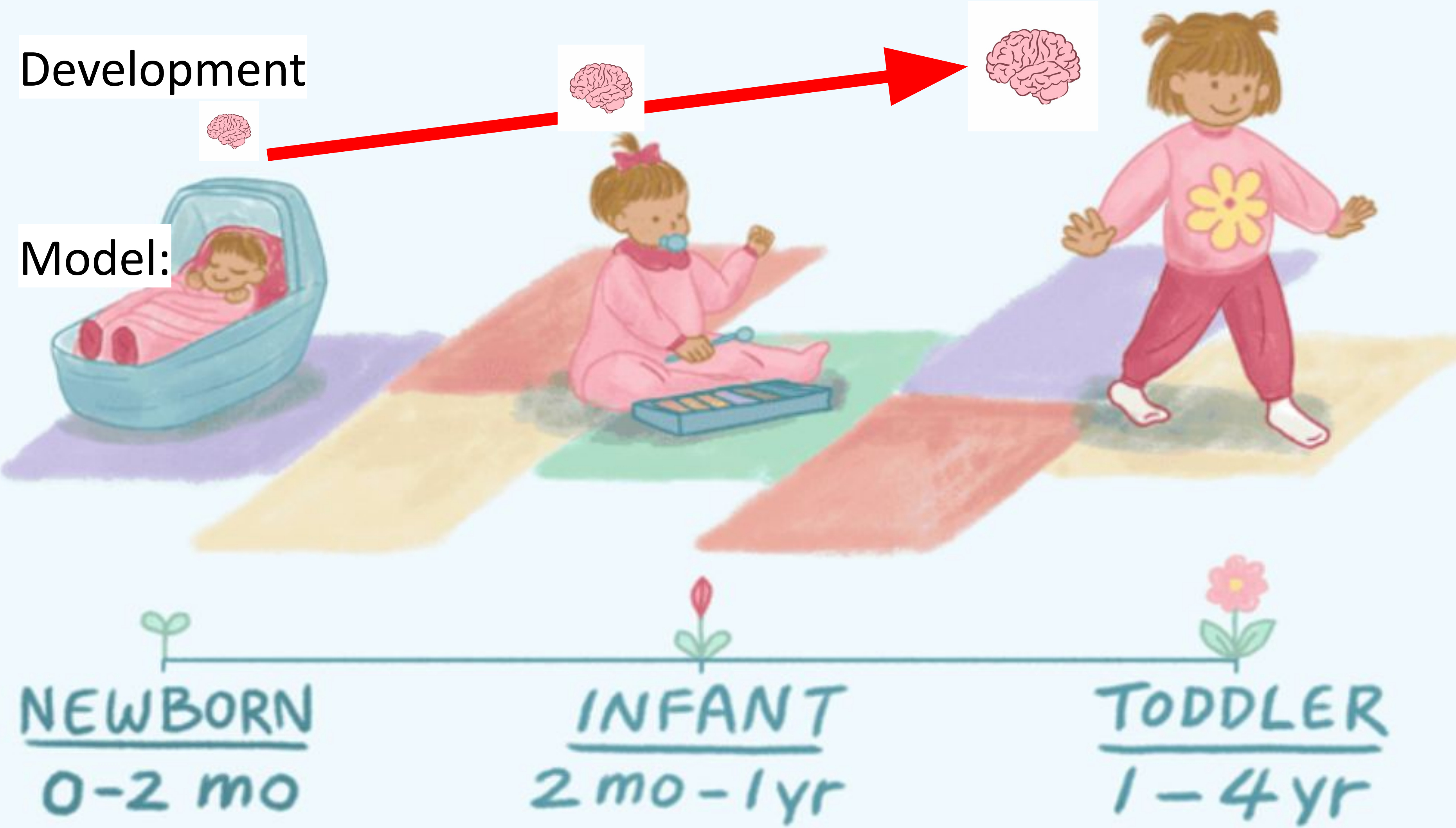
Linda B. Smith, Swapnaa Jayaraman, Elizabeth Clerkin, and Chen Yu

Presented by Mohamed El Banani and Richard E.L. Higgins

Development

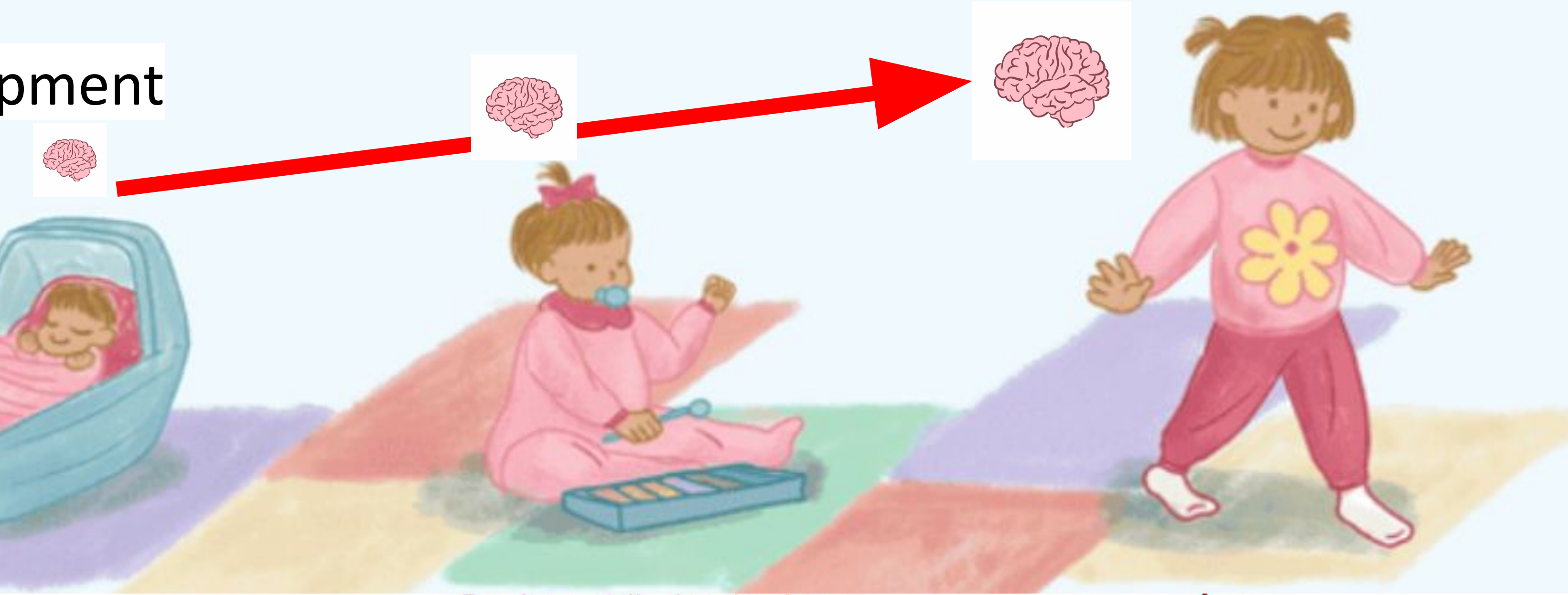


Development Model:



Development

Model:

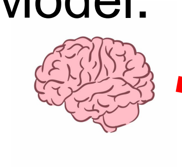


Data:

Typical View of Development

Rich but noisy data for many tasks mixed together:

Model:

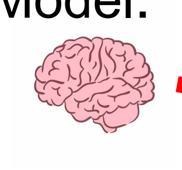


Sample things to learn

Rich but noisy data for many tasks mixed together:

<u>Typical View of Development</u>

Model:



Sample things to learn

Researchers thus:

- test learning mechanisms
- develop machinery that uses statistical regularities
- expect this to work

Visual Processing vs. Visual Development

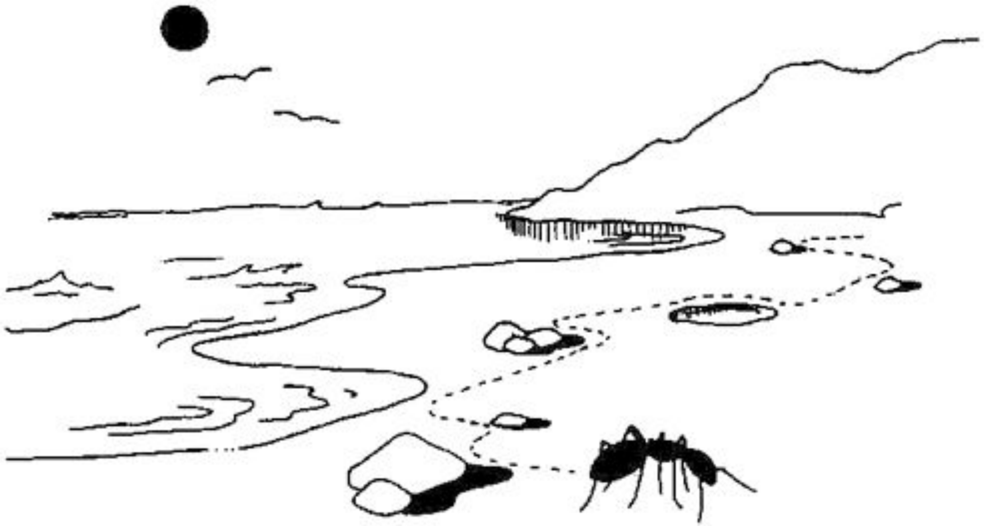
- “The study of vision must therefore include not only the study of **how to extract from images the various aspects of the world that are useful to us**, but also an inquiry into the **nature of the internal representations by which we capture this information** [...]”

  – David Marr (1982)

- “All statistical learning depends on both the **internal machinery that does the learning** and the **regularities in the data** on which that machinery operates.”

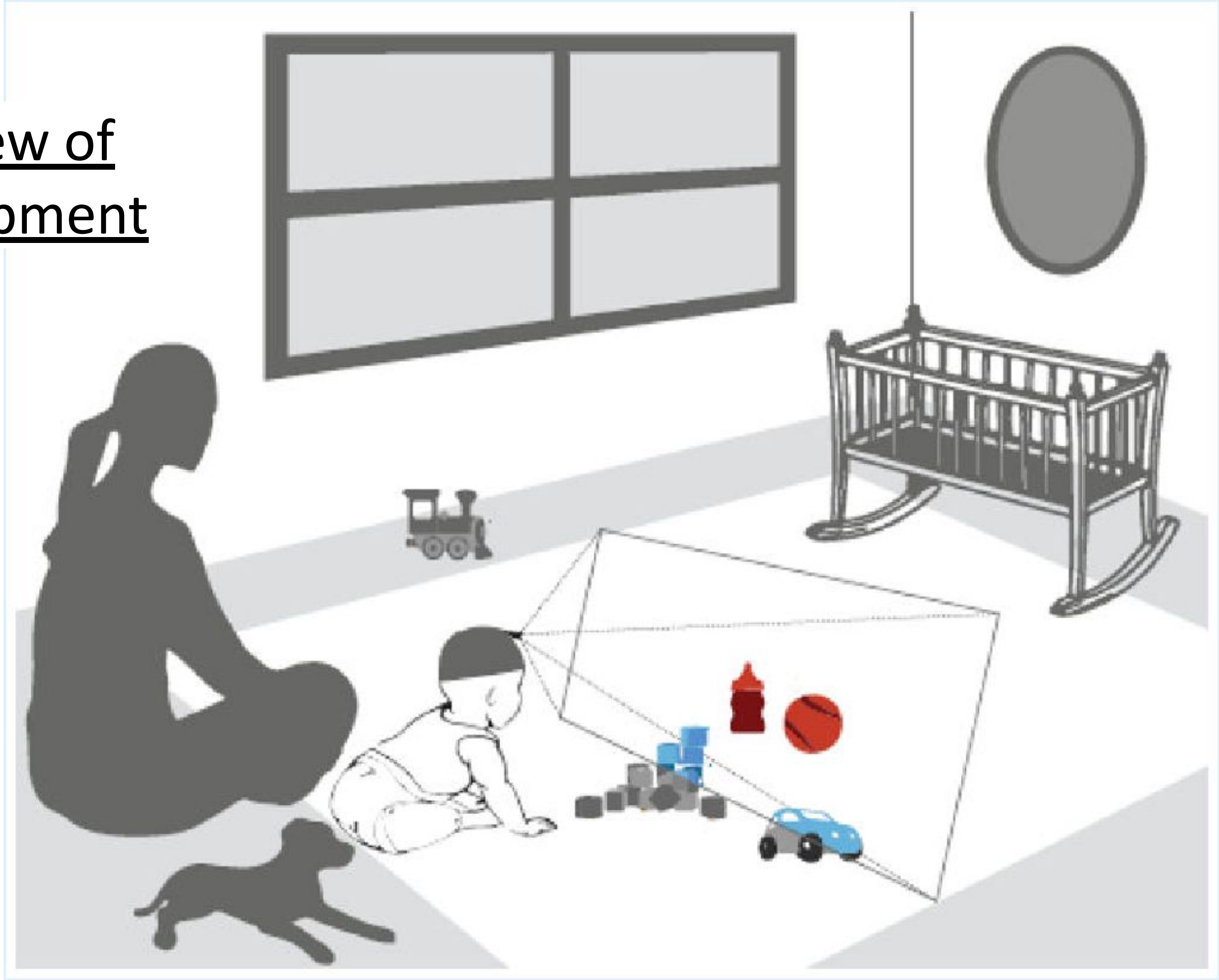
  – Smith et al. (2018)

The analogy of Simon’s Ant



- The complexity of an ant’s path is a combination of its **simplistic navigation** and the **complexity of its environment**.

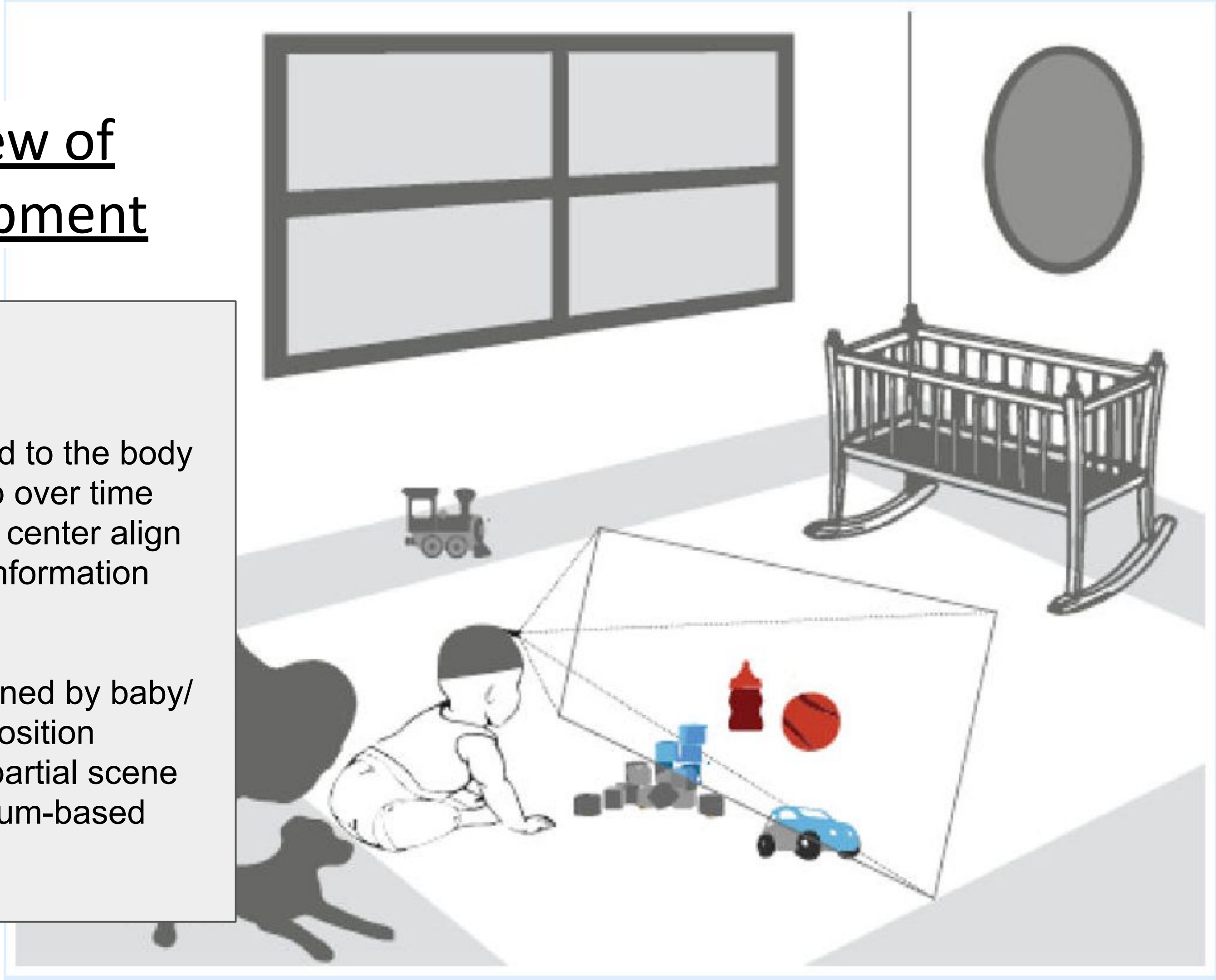
This View of Development



This View of Development

Sensors are:
- attached to the body
- develop over time
- used to center align visual information

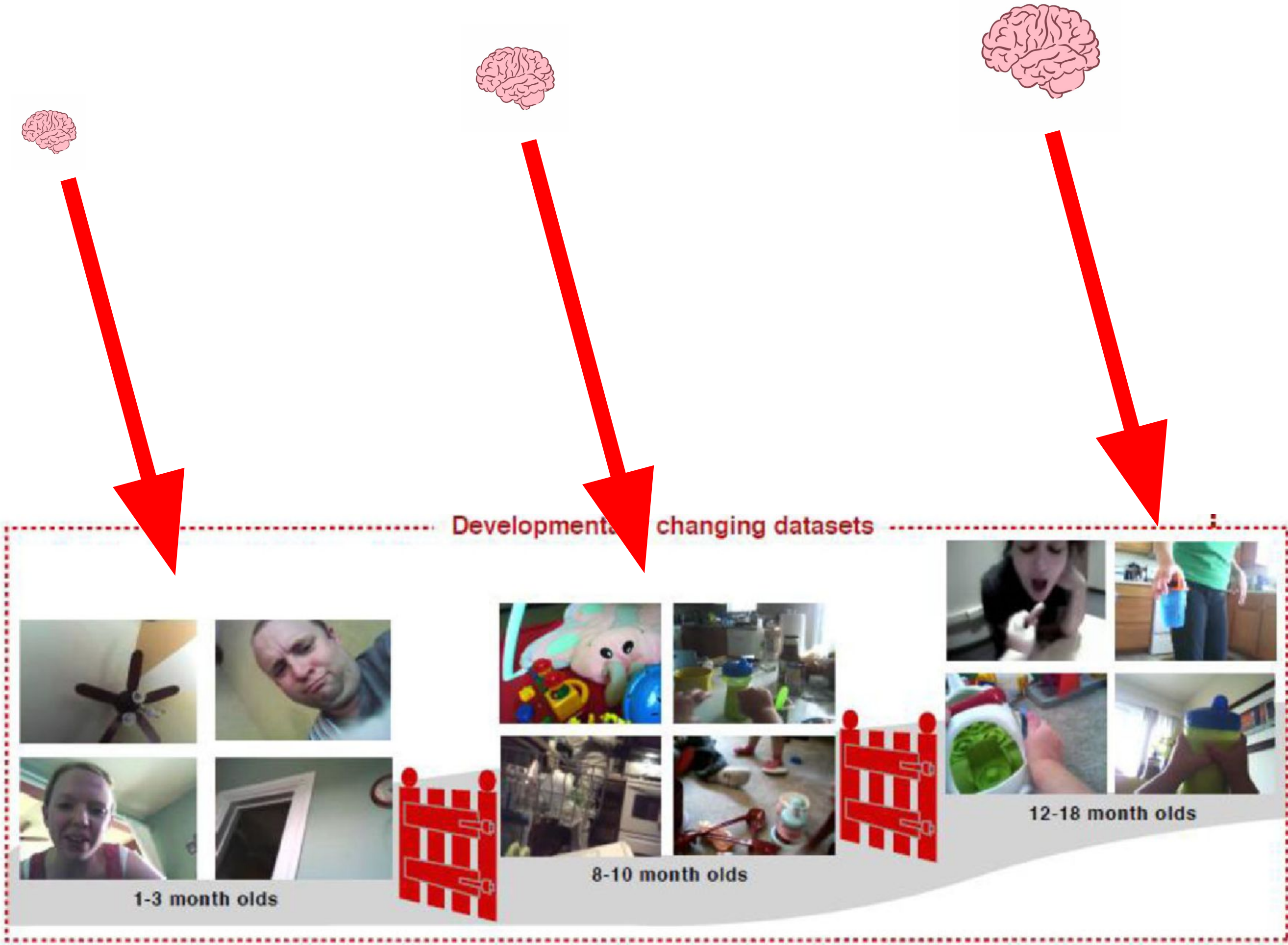
Data is:
- determined by baby/ infant position
- only a partial scene
- curriculum-based



This View of Development

Sensors are:
- attached to the body
- develop over time
- used to center align visual information

Data is:
- determined by baby/ infant position
- only a partial scene
- curriculum-based



What can infants see?

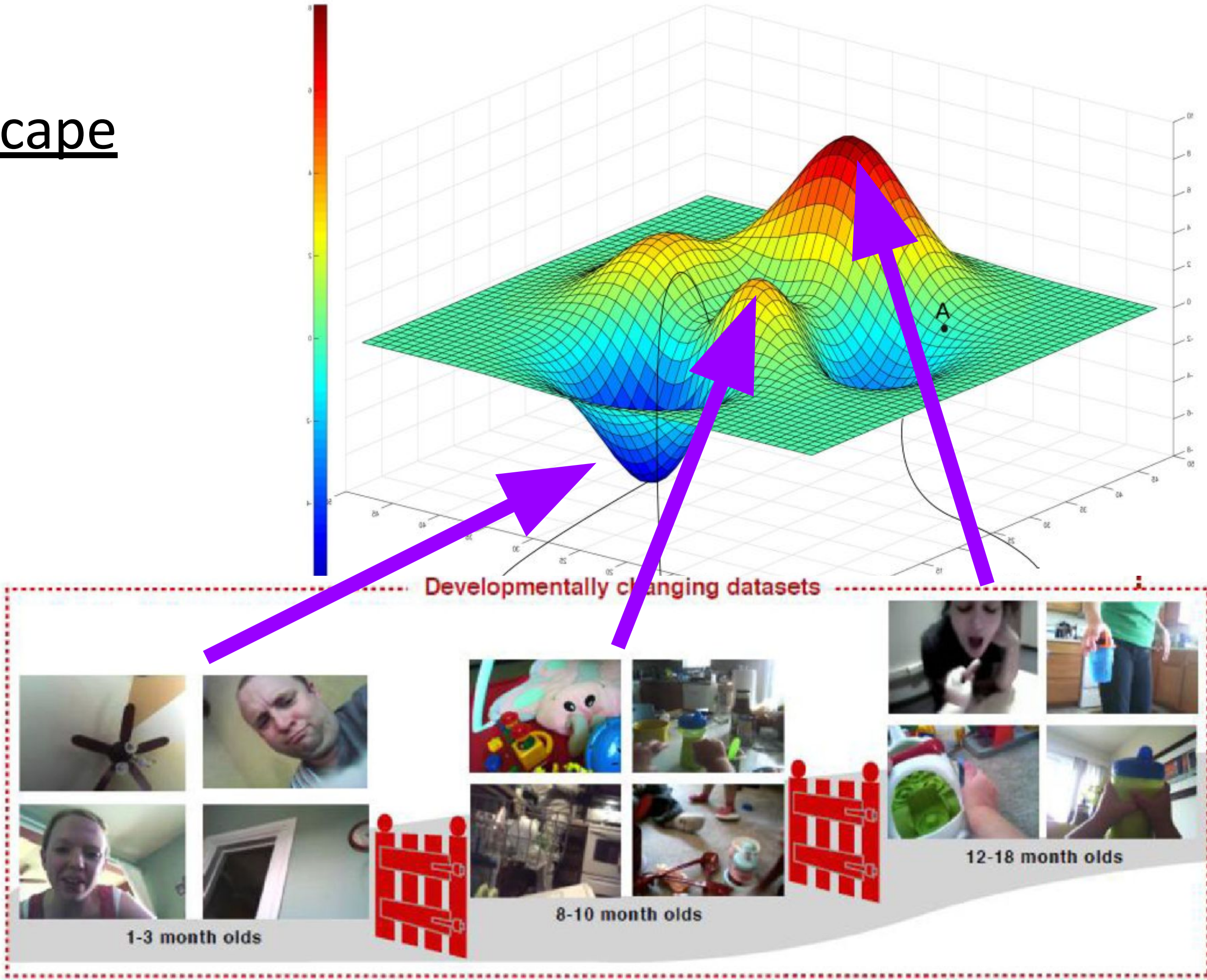
Source: Clinic Compare UK

Environment Landscape

1-3 months:
- limited mobility
- faces
- ceiling
- rolling over
- reaching

8-10 months
- crawling
- touching

12-18 months old
- walking
- manipulating
- using

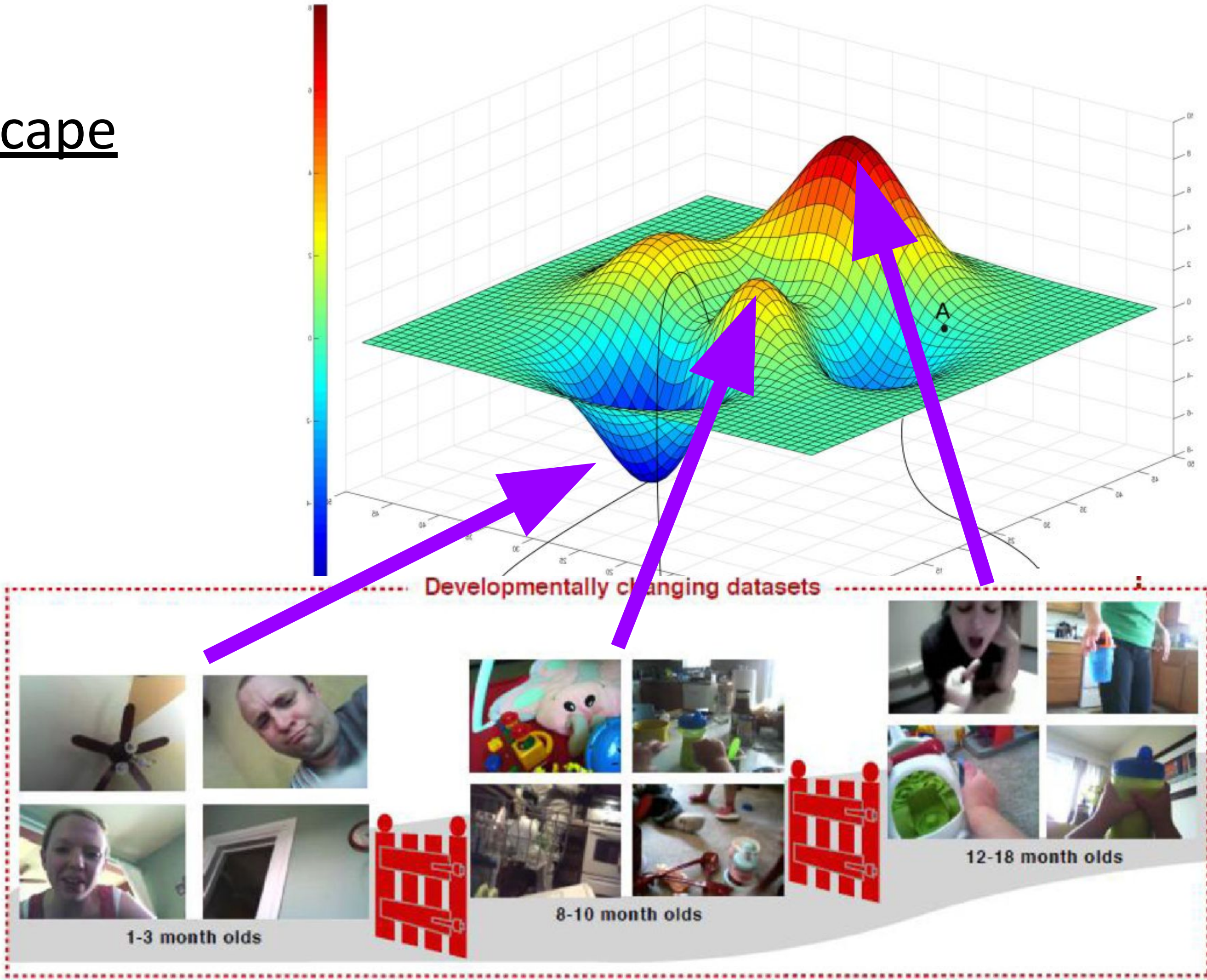


Environment Landscape

1-3 months:
- limited mobility
- faces
- ceiling
- rolling over
- reaching

8-10 months
- crawling
- touching

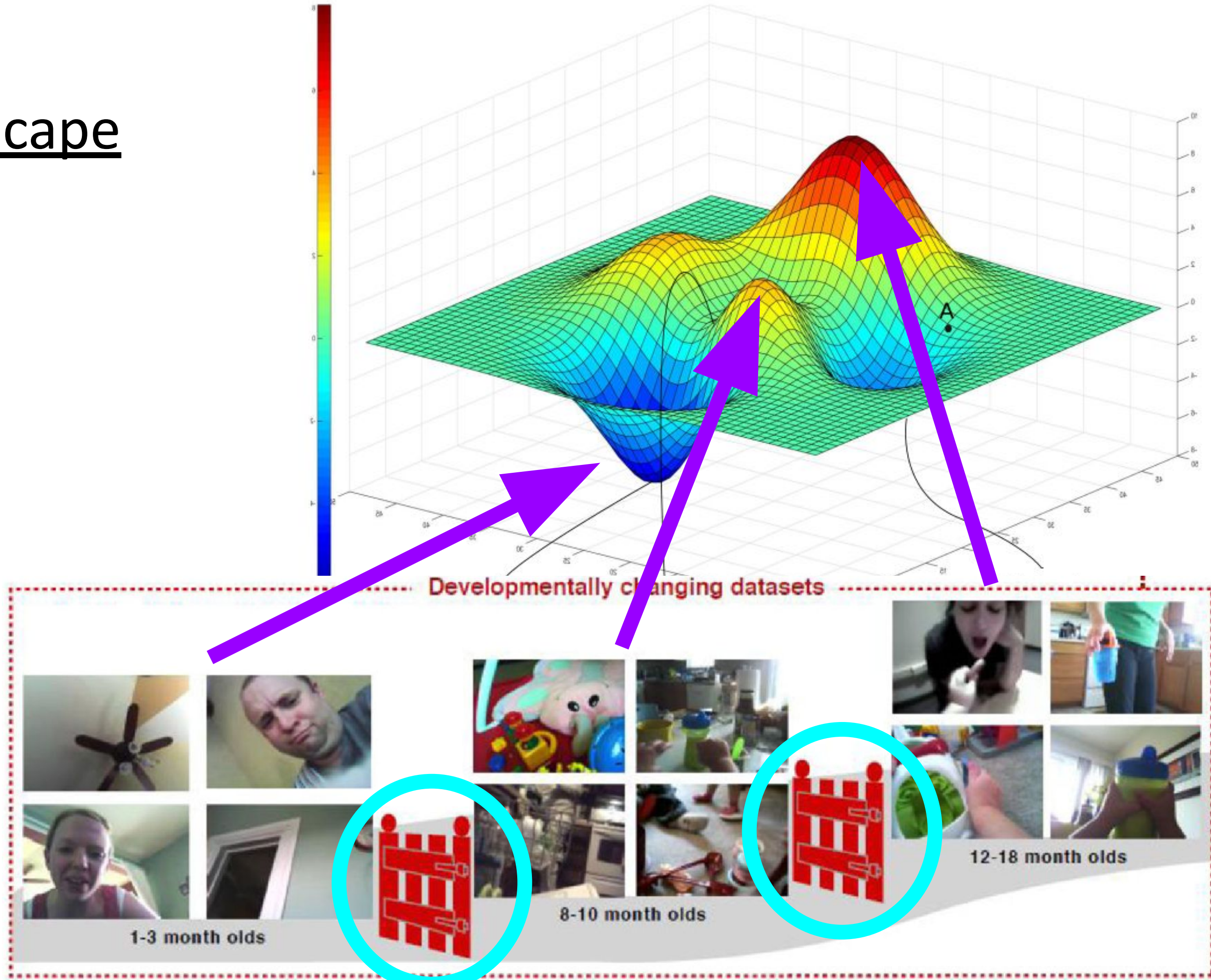
12-18 months old
- walking
- manipulating
- using



Environment Landscape

Gate 1:
- Visual ability
- Physical ability

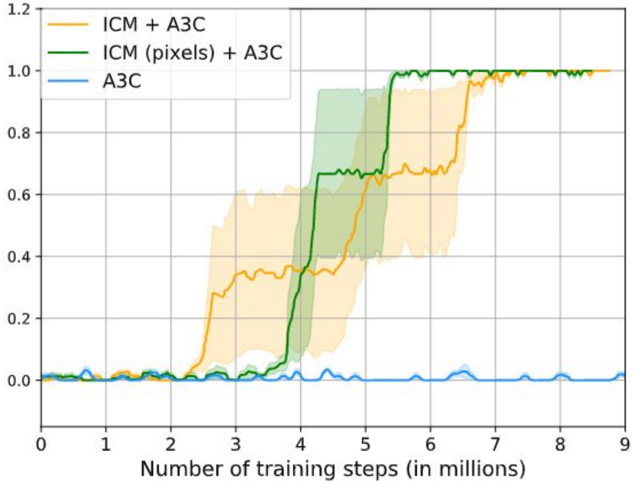
Gate 2:
- Walking
- Focusing beyond some number of ft



Environment Landscape

Gate 1:
- Visual ability
- Physical ability

Gate 2:
- Walking
- Focusing beyond some number of ft

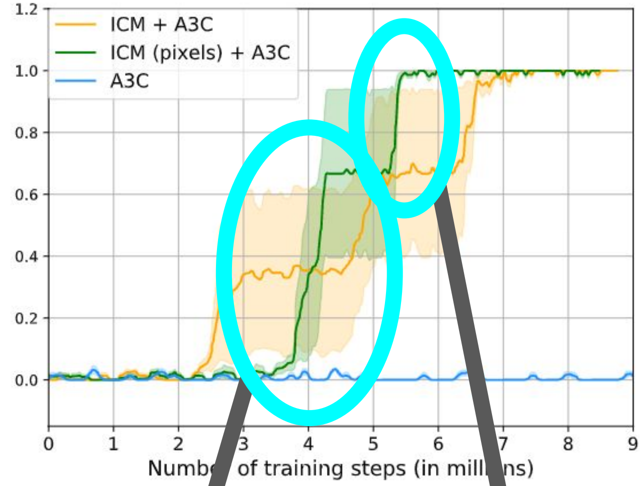


(b) “sparse reward” setting

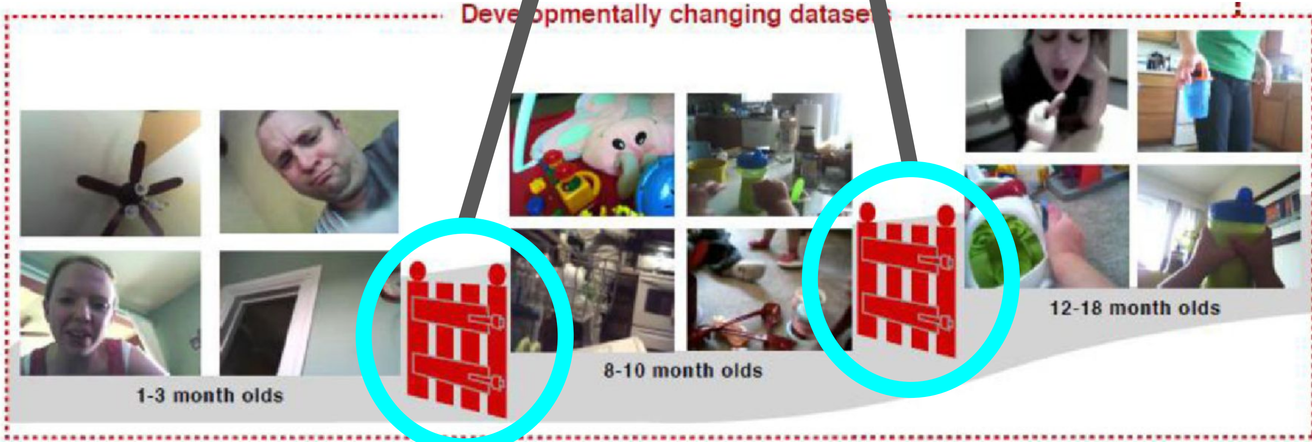
Environment Landscape

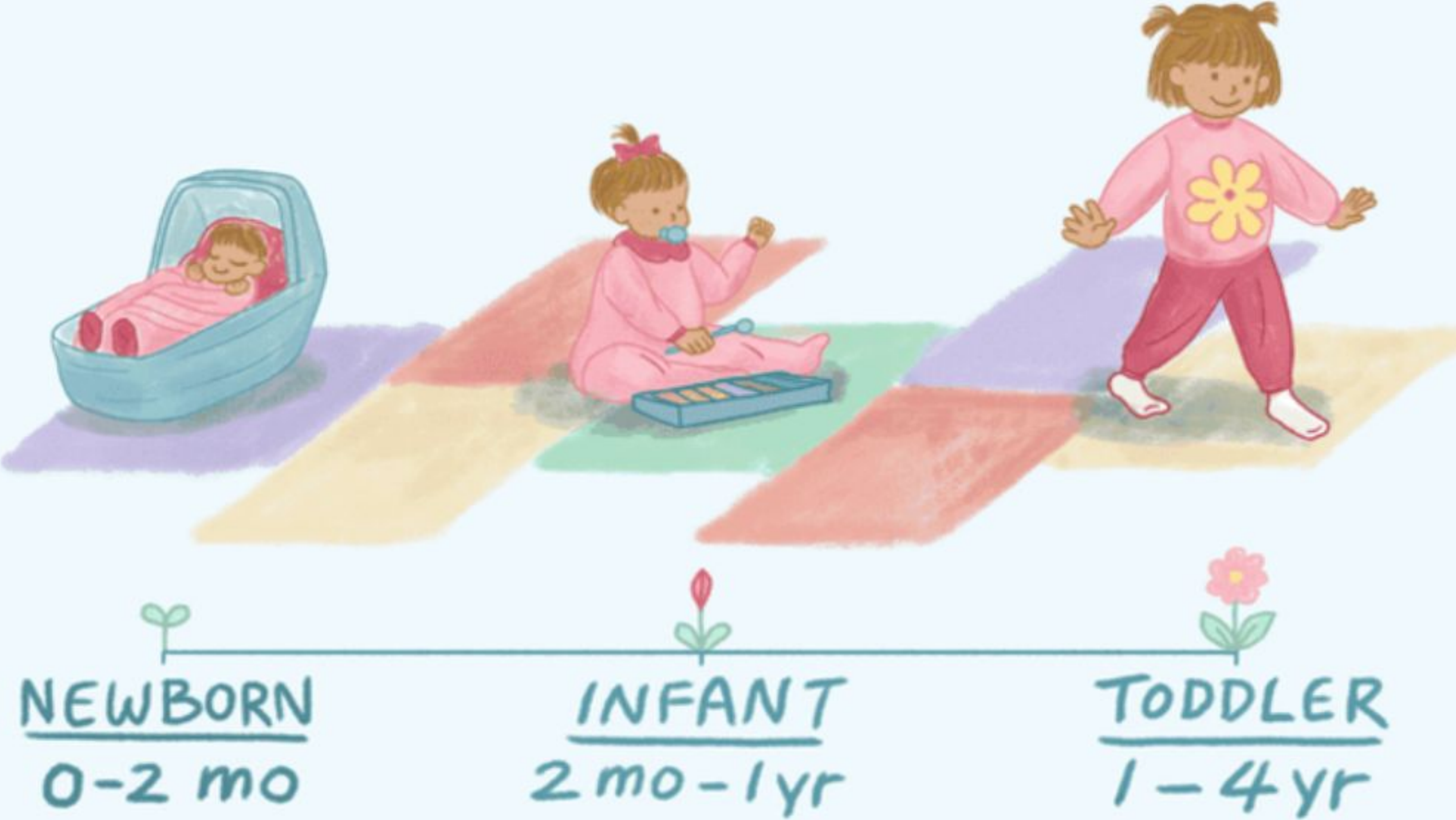
Gate 1:
- Visual ability
- Physical ability

Gate 2:
- Walking
- Focusing beyond some number of ft



(b) “sparse reward” setting




NEWBORN
0-2 mo
INFANT
2 mo-1 yr
TODDLER
1-4 yr

NEWBORN
0-2 mo

- limited visual acuity
- can do little
- what's in front of their face
(normally caregiver face)
- close and frontal views
- 15 min of face / every hour

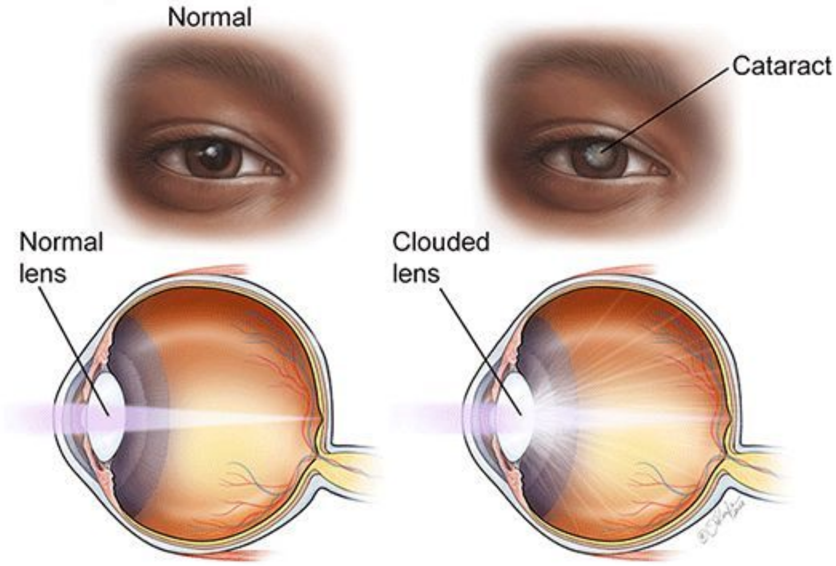
INFANT
2 mo-1yr

- see further, and move to see far object up close
- crawling creates dynamic visual input and optic flow
- manipulating for more views of the same object
- needs to sit up to see social partners or objects

TODDLER
1-4yr

- walking
- rarely sees other's faces
- other's hands provide manual examples
- 6 min of face / every hour

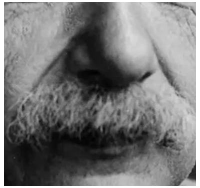
Timing Matters | Congenital Cataracts and Facial Processing



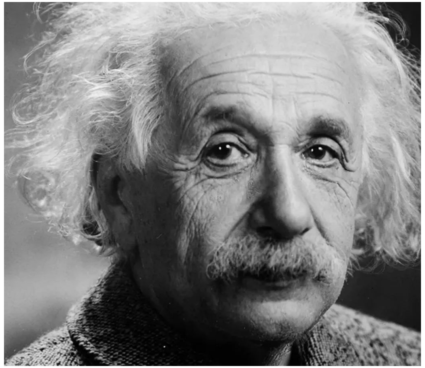
Timing Matters | Configural Facial Processing

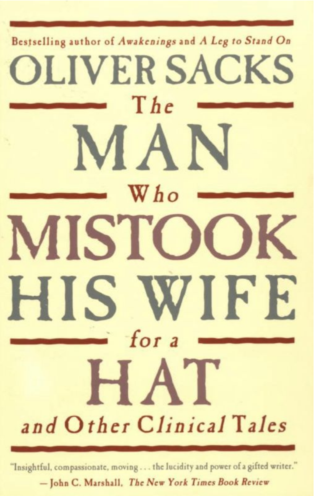
Timing Matters | Configural Facial Processing

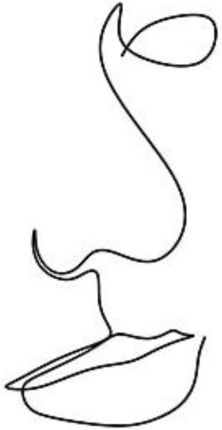
Who is this?



Who is this?







<u>Feature Based Processing</u>

"He recognized a portrait of Einstein because he picked up the characteristic hair and moustache"

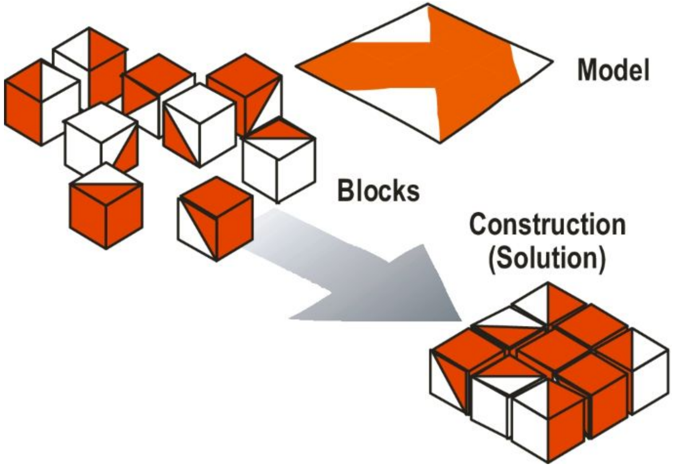
– Oliver Sacks

<u>Configural Processing</u>

Recognizing faces based on relations between features; location and spacing.

Source: Sacks, FreePNG, Maurer et al. 2002

Configural Face Processing and Coherence



Journal of Autism and Developmental Disorders, Vol. 36, No. 1, January 2006 (© 2006)
DOI 10.1007/s10803-005-0039-0
Published Online: February 1, 2006

The Weak Coherence Account: Detail-focused Cognitive Style in Autism Spectrum Disorders

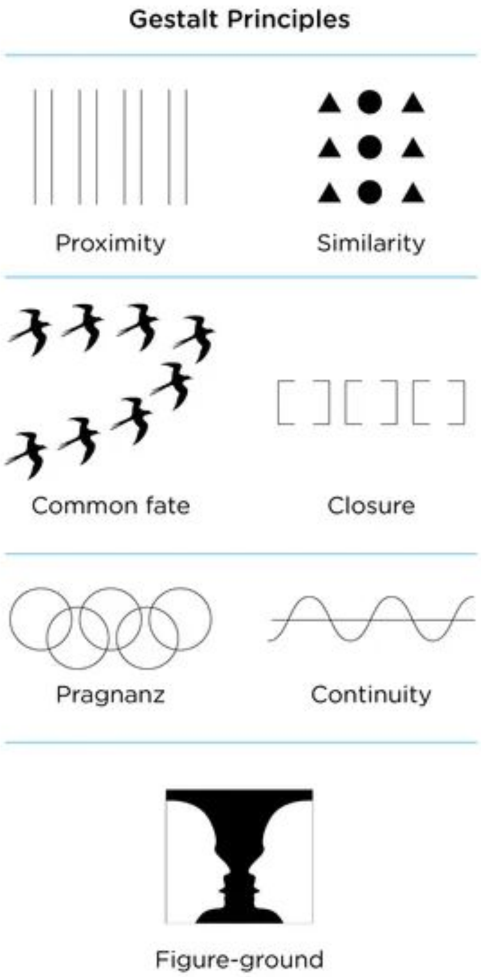
Francesca Happé[1,3] and Uta Frith[2]

"Weak central coherence" refers to the detail-focused processing style proposed to characterise autism spectrum disorders (ASD). The original suggestion of a core deficit in central processing resulting in failure to extract global form/meaning, has been challenged in three ways. First, it may represent an outcome of superiority in local processing. Second, it may be a processing *bias*, rather than deficit. Third, weak coherence may occur alongside, rather than explain, deficits in social cognition. A review of over 50 empirical studies of coherence suggests robust findings of local bias in ASD, with mixed findings regarding weak global processing. Local bias appears not to be a mere side-effect of executive dysfunction, and may be independent of theory of mind deficits. Possible computational and neural models are discussed.

Atypical development of configural face recognition in children with autism, Down syndrome and Williams syndrome

D. Dimitriou,[1] H. C. Leonard,[2] A. Karmiloff-Smith,[2] M. H. Johnson[2] & M. S. C. Thomas[2]

1 *Institute of Education, Department of Psychology and Human Development, University of London, London, UK*
2 *Centre for Brain and Cognitive Development, Birkbeck, University of London, London, UK*



Counter | The Molyneaux Problem

- "Would a blind person, on regaining sight, be able to immediately visually recognize an object previously known only by touch?"
- Test this ability in 8-17 year children with congenital cataracts.
- Results:
    - No immediate transfer;
    - Transfer developed in 5 days!!

nature neuroscience

The newly sighted fail to match seen with felt

Richard Held[1], Yuri Ostrovsky[1], Beatrice de Gelder[2], Tapan Gandhi[3], Suma Ganesh[4], Umang Mathur[4] & Pawan Sinha[1]

Would a blind subject, on regaining sight, be able to immediately visually recognize an object previously known only by touch? We addressed this question, first formulated by Molyneux three centuries ago, by working with treatable, congenitally blind individuals. We tested their ability to visually match an object to a haptically sensed sample after sight restoration. We found a lack of immediate transfer, but such cross-modal mappings developed rapidly.

Could we restore the facial abilities through late retraining or do we lose it after the critical period?

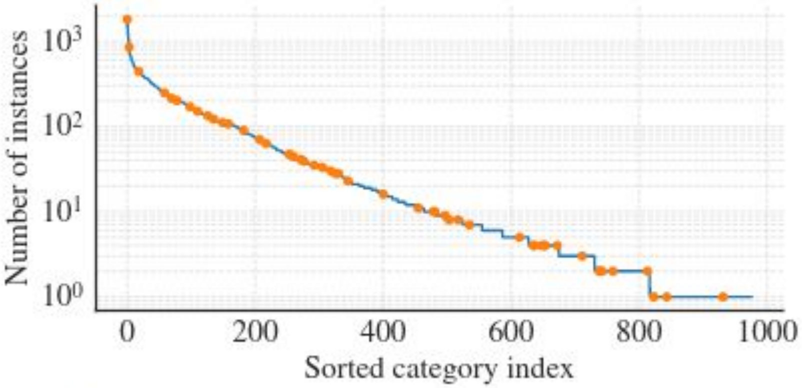
Limited data and heavy tails!

- Early experience is highly limited!
- A child typically only a few environments and mostly 2-3 people.
- Very skewed data!

How do kids learn from this kind of data?

Source: Pintrest

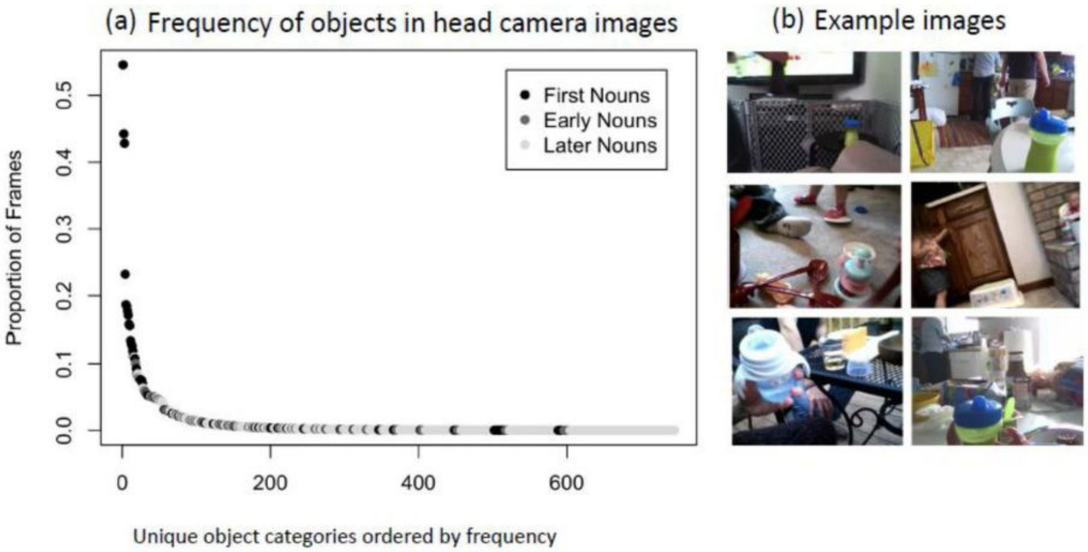
Long-tail distributions are **real** and **difficult**!



(b) The number of instances per category (on 5k images) reveals the long tail with few examples. Orange dots: categories in common with COCO.

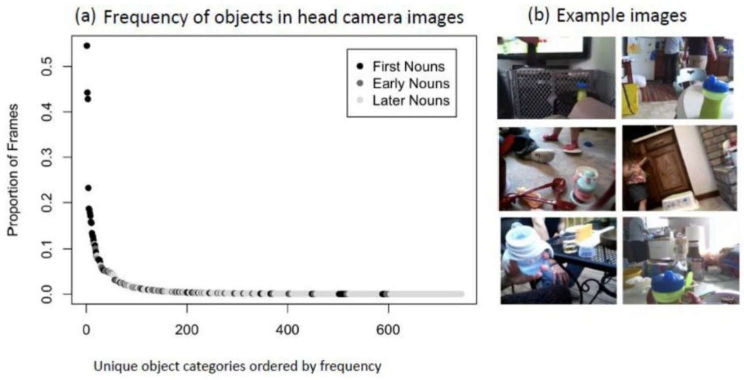
Source: LVIS dataset

Infants focus on important objects!



Limited data and heavy tails!

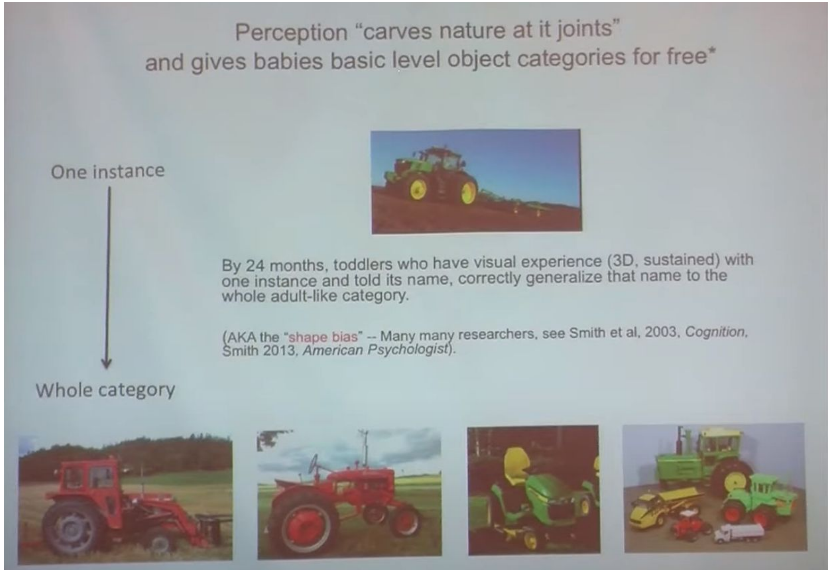
- 3 months of development may not seem long, but this is a critical period
  - 15 min/hr x 12 hr/day: 270 hrs of faces.
- Sampling is selective: few objects appear but they appear frequently.
- Are kids learning common objects or sampling learnable objects?



Skewed data and amazing generalization .. how?

- 2yo can generalize a single instance to a full category … 1-shot learning!

- Three interrelated hypotheses:
    - Consistency;
    - Bootstrapping;
    - Desirable difficulty.



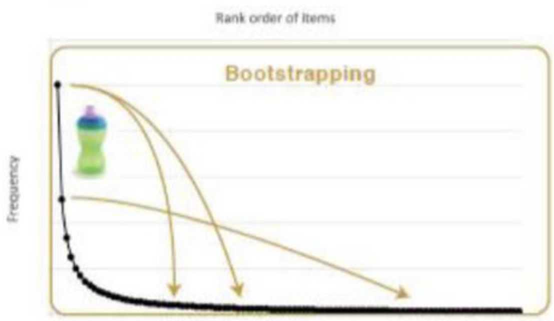
single instance of that category. For example, if a 2-year-old child encounters their very first tractor – say, a green John Deere working in a field – while hearing its name, the child is likely from that point forward to recognize all variety of tractors as tractors – red Massey-Fergusons, rusty antique tractors, ride-on mowers – but not backhoes or trucks [48 – 50].

Three Hypotheses | Consistency

- *Rare things don't matter*
- The world will provide a set of repeating and consistent items, the infant learns to differentiate between them and slowly grows their classifier.

Three Hypotheses | Bootstrapping

- *Common and understood items help us understand the world better*
- Rare items will occur with common items. Knowing the common items helps scaffold and structure the learning of rare items.



Three Hypotheses | Desirable Difficulty

- *Difficult environments help you learn more robustly*
- While the infant might only encounter one instance, it can learn a more robust model of it by differentiating it from all the clutter.

Where do we go now?

- Current approaches are too static:
    - Same task,
    - Same dataset,
    - Same underlying algorithm!
- How can we take a developmental approach to statistical learning?

Where do we go now? Starting small

- Early work has shown that:
    - Models can benefit from learning from a growing corpus;
    - Learning can improve when networks reconfigure or grow!

Cognition, 48 (1993) 71–99 71
0010-0277/93/$06.00

Learning and development in neural networks: the importance of starting small

Jeffrey L. Elman
Departments of Cognitive Science and Linguistics, University of California, San Diego, La Jolla, CA 92093-0526, USA

Received August 27, 1992, final version accepted April 26, 1993

Abstract

It is a striking fact that in humans the greatest learning occurs precisely at that point in time – childhood – when the most dramatic maturational changes also occur. This report describes possible synergistic interactions between maturational change and the ability to learn a complex domain (language), as investigated in connectionist networks. The networks are trained to process complex sentences involving relative clauses, number agreement, and several types of verb argument structure. Training fails in the case of networks which are fully formed and 'adultlike' in their capacity. Training succeeds only when networks begin with limited working memory and gradually 'mature' to the adult state. This result suggests that rather than being a limitation, developmental restrictions on resources may constitute a necessary prerequisite for mastering certain complex domains. Specifically, successful learning may depend on starting small.

constant during learning. Plunkett and Marchman (1990) have shown that while the basic influences of type/token frequency and phonological predictability are similar to the condition of non-incremental learning, better overall learning is achieved when the training corpus for a connectionist model is allowed slowly to grow in size. We might also ask what the consequences are when the learning mechanism itself is changing. Allowing networks to reconfigure dynamically or acquire additional nodes has been shown to facilitate learning (Ash, 1989; Fahlman & Lebiere, 1990; Shultz & Schmidt, 1991).

Where do we go now? Curriculum Learning

- Maybe we can change the data distribution as we learn?
- Captures the idea of changing the data distribution.

Curriculum Learning

Yoshua Bengio[1] YOSHUA.BENGIO@UMONTREAL.CA
Jérôme Louradour[1,2] JEROMELOURADOUR@GMAIL.COM
Ronan Collobert[3] RONAN@COLLOBERT.COM
Jason Weston[3] JASONW@NEC-LABS.COM
(1) U. MONTREAL, P.O. BOX 6128, MONTREAL, CANADA (2) A2IA SA, 40BIS FABERT, PARIS, FRANCE
(3) NEC LABORATORIES AMERICA, 4 INDEPENDENCE WAY, PRINCETON, NJ, USA

Abstract

Humans and animals learn much better when the examples are not randomly presented but organized in a meaningful order which illustrates gradually more concepts, and gradually more complex ones. Here, we formalize such training strategies in the context of machine learning, and call them "curriculum learning". In the context of recent research studying the difficulty of training in the presence of non-convex training criteria (for deep deterministic and stochastic neural networks), we explore curriculum learning in various set-ups. The experiments show that significant improvements in generalization can be achieved. We hypothesize that curriculum learning has both an effect on the speed of convergence of the training process to a minimum and, in the case of non-convex criteria, on the quality of the local minima obtained: curriculum learning can be seen as a particular form of continuation method (a general strategy for global optimization of non-convex functions).

1. Introduction

Humans need about two decades to be trained as fully functional adults of our society. That training is highly organized, based on an education system and a curriculum which introduces different concepts at different times, exploiting previously learned concepts to ease the learning of new abstractions. By choosing which examples to present and in which order to present them to the learning system, one can *guide* training and remarkably increase the speed at which learning can occur. This idea is routinely exploited in *animal training* where it is called **shaping** (Skinner, 1958; Peterson, 2004; Krueger & Dayan, 2009).

Previous research (Elman, 1993; Rohde & Plaut, 1999; Krueger & Dayan, 2009) at the intersection of cognitive science and machine learning has raised the following question: can machine learning algorithms benefit from a similar training strategy? The idea of training a learning machine with a curriculum can be traced back at least to Elman (1993). The basic idea is to *start small*, learn easier aspects of the task or easier subtasks, and then gradually increase the difficulty level. The experimental results, based on learning a simple grammar with a recurrent network (Elman, 1993), suggested that successful learning of grammatical structure depends, not on innate knowledge of grammar, but on starting with a limited architecture that is at first quite restricted in complexity, but then expands its resources gradually as it learns. Such conclusions are important for developmental psychology, because they illustrate the adaptive value of starting, as human infants do, with a simpler initial state, and then building on that to develop more and more sophisticated representations of structure. Elman (1993) makes the statement that this strategy could make it possible for humans to learn what might otherwise prove to be unlearnable. However, these conclusions have been seriously questioned in Rohde and Plaut (1999). The question of guiding learning of a recurrent neural network for learning a simple language and increasing its capacity along the way was recently revisited from the cognitive perspective (Krueger & Dayan, 2009), providing evidence for faster convergence using a shaping procedure. Similar ideas were also explored in robotics (Sanger, 1994), by gradually making the learning task more difficult.

We want to clarify when and why a curriculum or

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

Where do we go now? Active Learning

- How can we choose the samples we learn from?

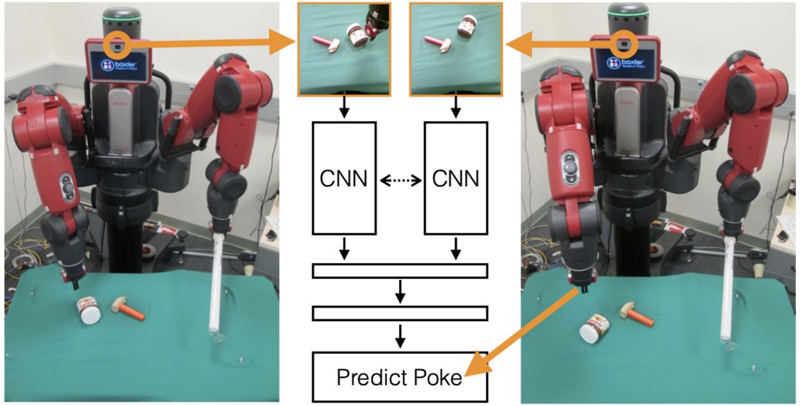


Figure 1: Infants spend years worth of time playing with objects in a seemingly random manner. They might use this experience to learn a model of physics relating their actions with the resulting motion of objects. Inspired by this hypothesis, we let a robot interact with objects by randomly poking them. The robot pokes objects and records the visual state before (left) and after (right) the poke. The triplet of before image, after image and the applied poke is used to train a neural network (center) for learning the mapping between actions and the accompanying change in visual state. We show that this learn model can be used to push objects into a desired configuration.

Queries and Concept Learning

DANA ANGLUIN (ANGLUIN@YALE.EDU)
Department of Computer Science, Yale University, P.O. Box 2158, New Haven, CT 06520, U.S.A.

(Received: July 20, 1987)
(Revised: January 26, 1988)

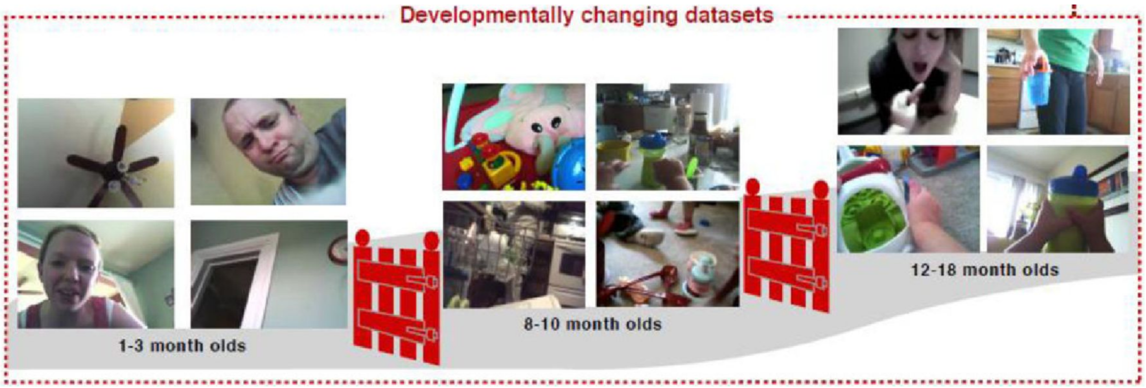
Keywords: Concept learning, supervised learning, queries

Abstract. We consider the problem of using queries to learn an unknown concept. Several types of queries are described and studied: membership, equivalence, subset, superset, disjointness, and exhaustiveness queries. Examples are given of efficient learning methods using various subsets of these queries for formal domains, including the regular languages, restricted classes of context-free languages, the pattern languages, and restricted types of propositional formulas. Some general lower bound techniques are given. Equivalence queries are compared with Valiant's criterion of probably approximately correct identification under random sampling.

1. Introduction

A successful learning component in an expert system will probably rely heavily on queries to its instructors. For example, Sammut and Banerji's (1986) system uses queries about specific examples as part of its strategy for efficiently learning a target concept. Shapiro's (1981, 1982, 1983) Algorithmic Debugging System uses a variety of types of queries to the user to pinpoint errors in Prolog programs. In this paper we use a formal framework to study the power of several types of queries for concept-learning tasks.

Infant’s visual world is very different from a CNN



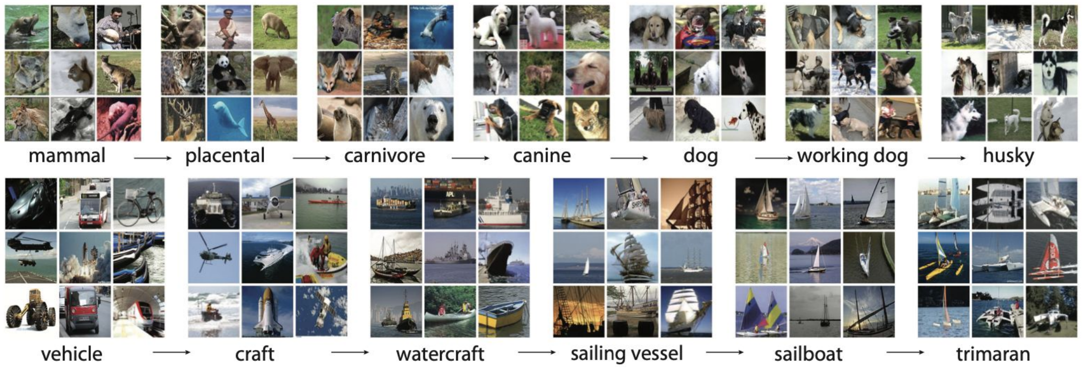


Image Source: Deng et al. (CVPR 2009)

Some approaches build on this observation!

Figure 1. **Internet vision versus robotic vision.** Pictures taken by humans (top row) (and uploaded on the web) are the *output* of visual perception of a well-trained agent, the human photographer.

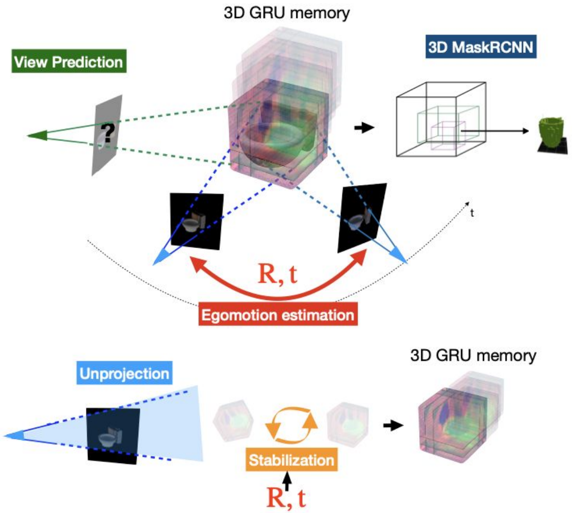
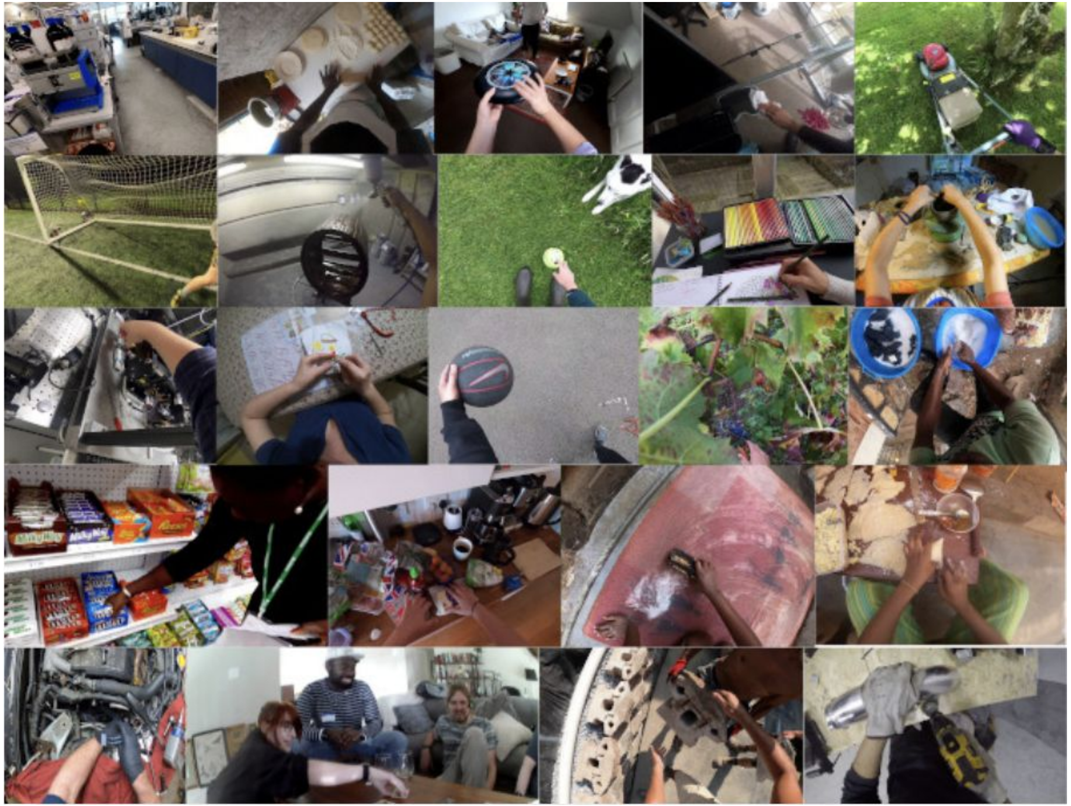


Figure 2. **Geometry-aware Recurrent Neural Networks (GRNNs)** integrate visual information over time in a 3D geometrically-consistent deep feature memory of the visual scene. At each frame, RGB images are *unprojected* into corresponding 3D feature tensors, which are oriented to the coordinate frame of the memory map built thus far (2nd row). A 3D convolutional GRU memory is then updated using the egomotion-stabilized features as input.

Source: Fish Tung et al. (CVPR 2019)

Datasets that focus on Ego-centric Motion



Source: Ego4D

What are the relevant questions to us?

- Do we need to rethink our architectures?
- Do machines need developmental learning?
- What does development look like for a machine?
- Are critical periods a function of **learning** or **human learning**?
  - Does omitting a critical period risk catastrophic forgetting in a critical domain?

Outstanding questions

- How can we experimentally test whether and how the structures found in first-person recordings of infant visual environments provide the required curriculum for infant learning?
- What are the real-time properties of data used for learning that change both over developmental time and over the learners' real time activities? How do they interact with potentially changing or different learning mechanisms?
- Does the order of developmentally segregated data sets – such as first faces and then objects matter to developmental outcomes? Do early face experiences support later visual development in other domains, in object perception, in letter recognition?
- If sensitive periods are formed in part by the closing of sensory-motor gates on critical experiences, can a sensitive period for learning be re-opened by re-opening those sensory-motor gates?
- What role do disruptions in the real-world data for learning play in the cognitive developmental trajectories of children with developmental disorders? This will shed light on the cognitive developmental disorders that are characterized by atypical patterns of sensory-motor development.