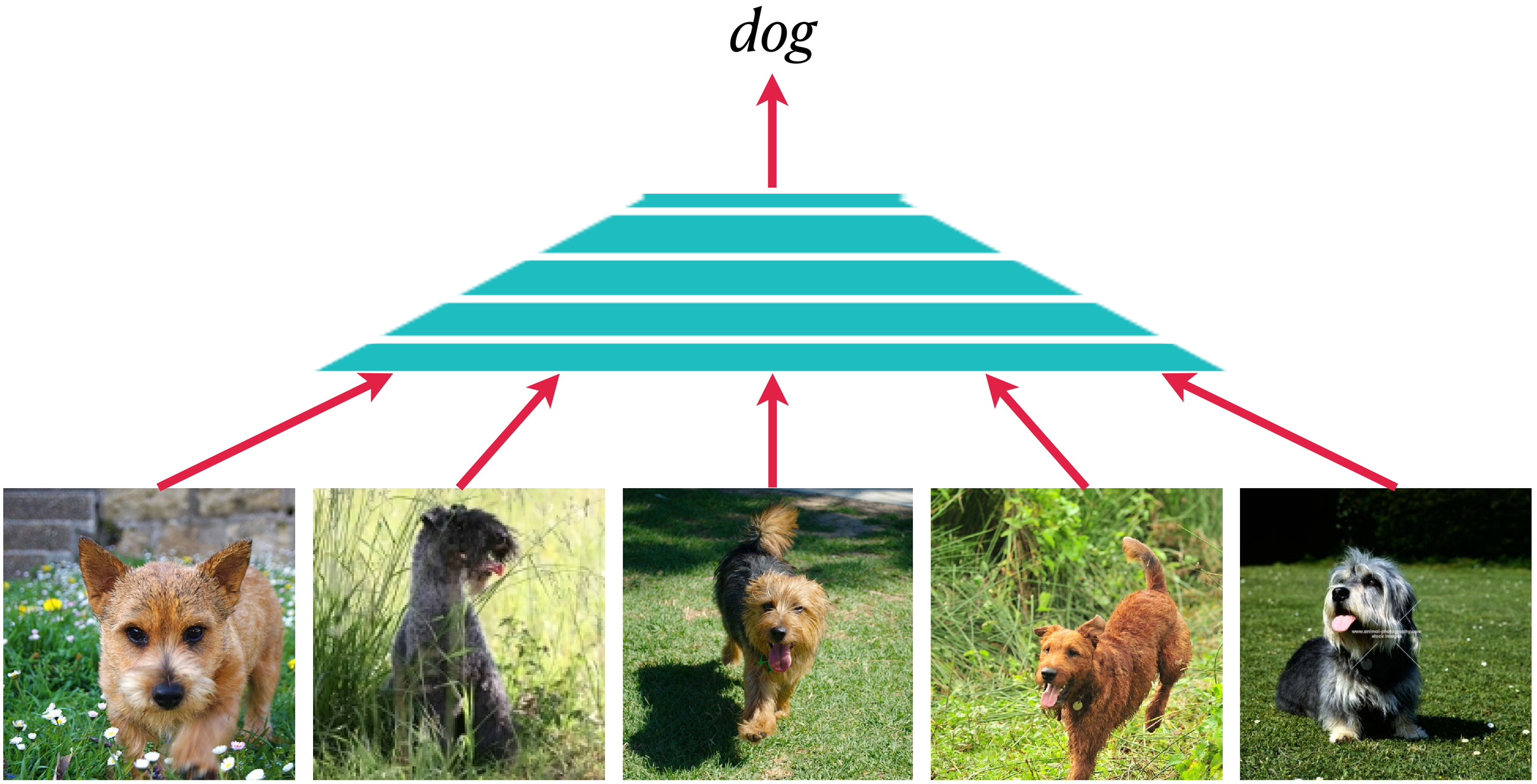
Introduction to EECS 598: Action and Perception

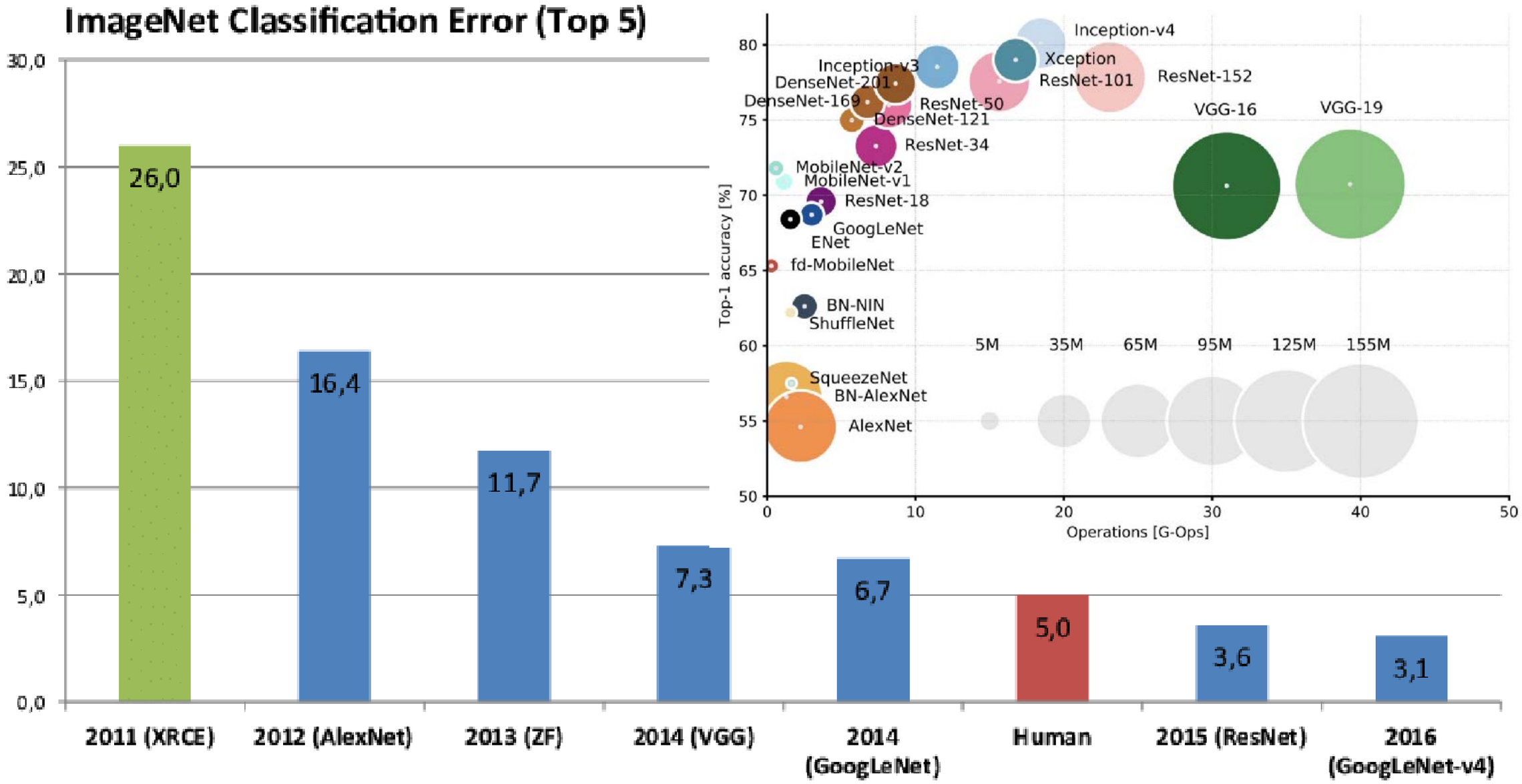
Stella Yu
University of Michigan

4 January 2023

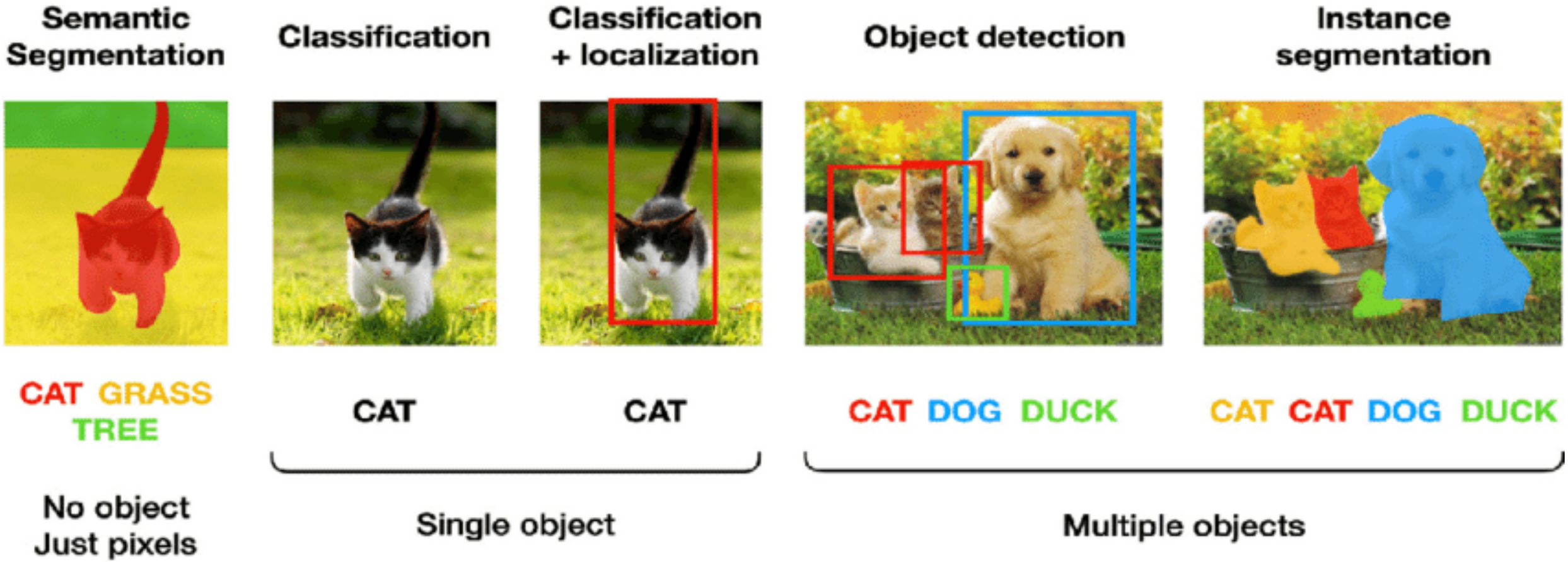
End-to-End Learning Approach to Visual Recognition



Supervised Learning Reaches Superhuman Performance



Deep Learning Is Successful But Too Specialized



High-level: Models trained for one task do not work for another.

Deep Learning Is Successful But Too Specialized

Low-level: Models trained on one data kind do not work for another.

Deep Learning Is Successful But Too Specialized

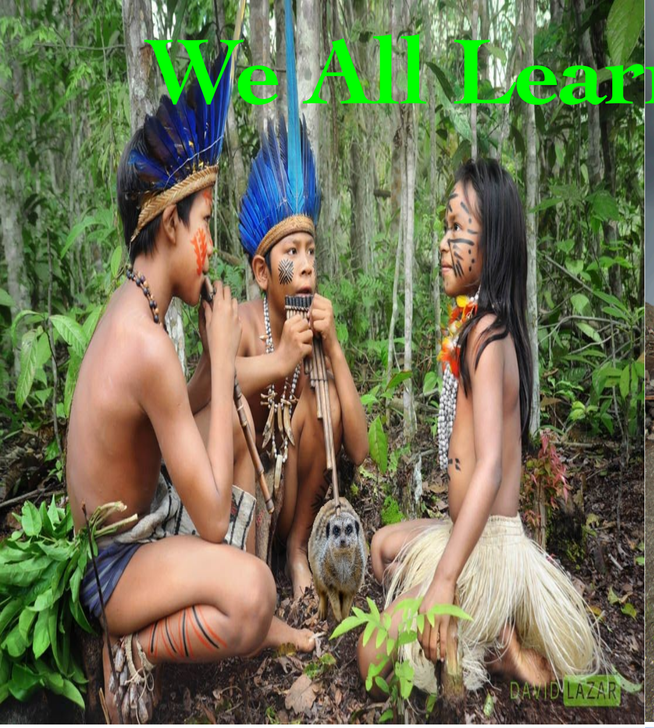
"A Car Parked On The Side of The Road"

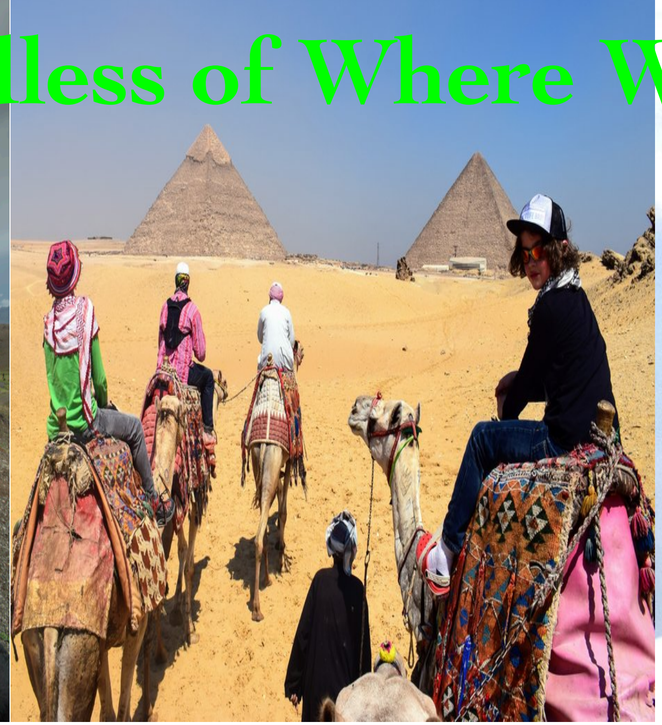
Low-level: Models trained on one data kind do not work for another.

Natural Learning of Vision: No Semantic Supervision

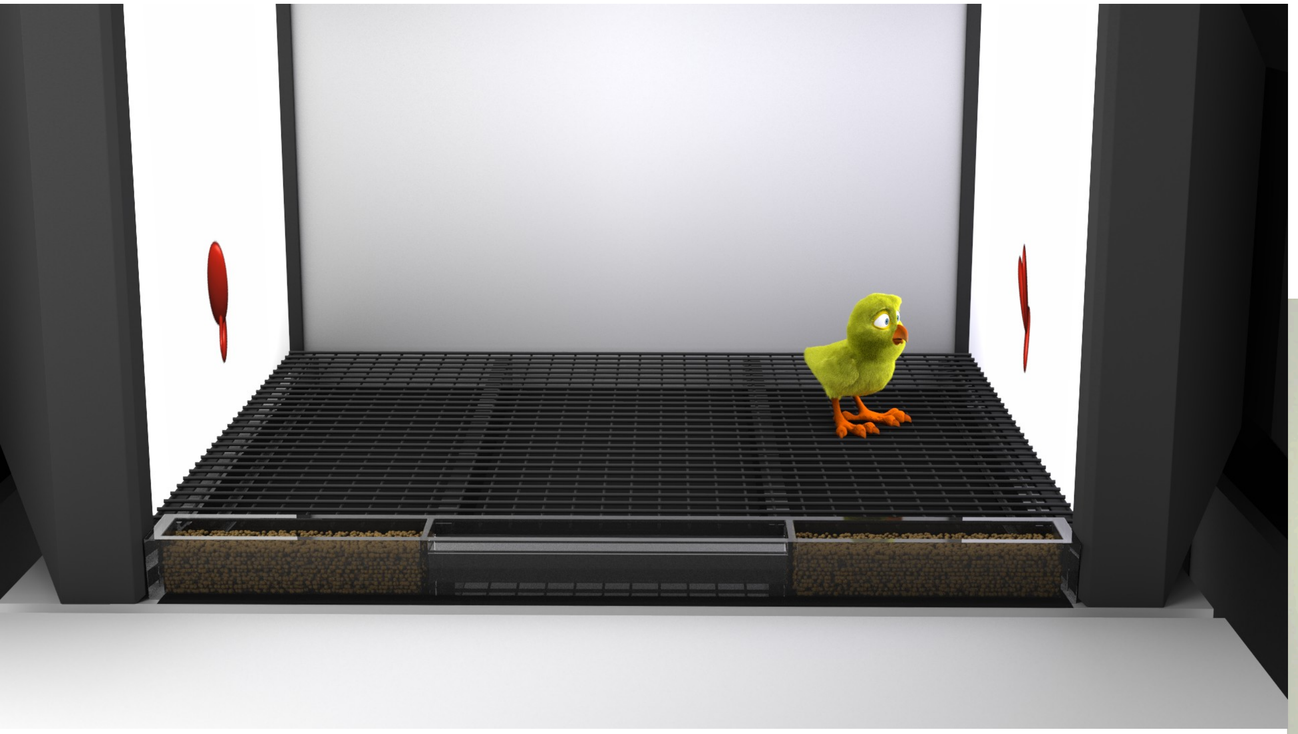
Linda Smith: [Jayaraman S. et al, PLoS ONE 2015; Clerkin et al, TRSB 2017; Slone et al, DS 2019]

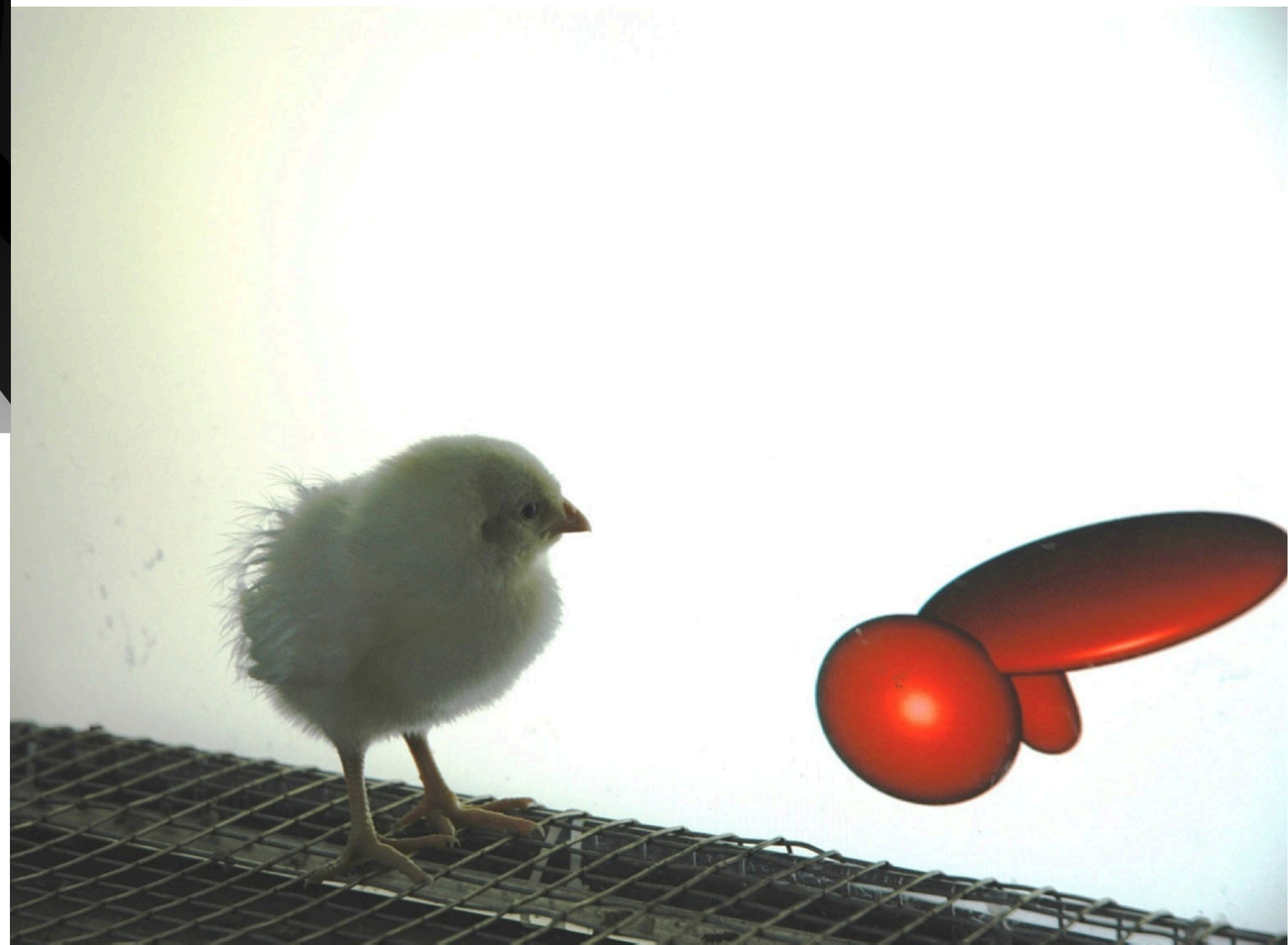
We All Learn to See Regardless of Where We Grow Up





Newborn Visual Recognition from Slow Smooth Videos

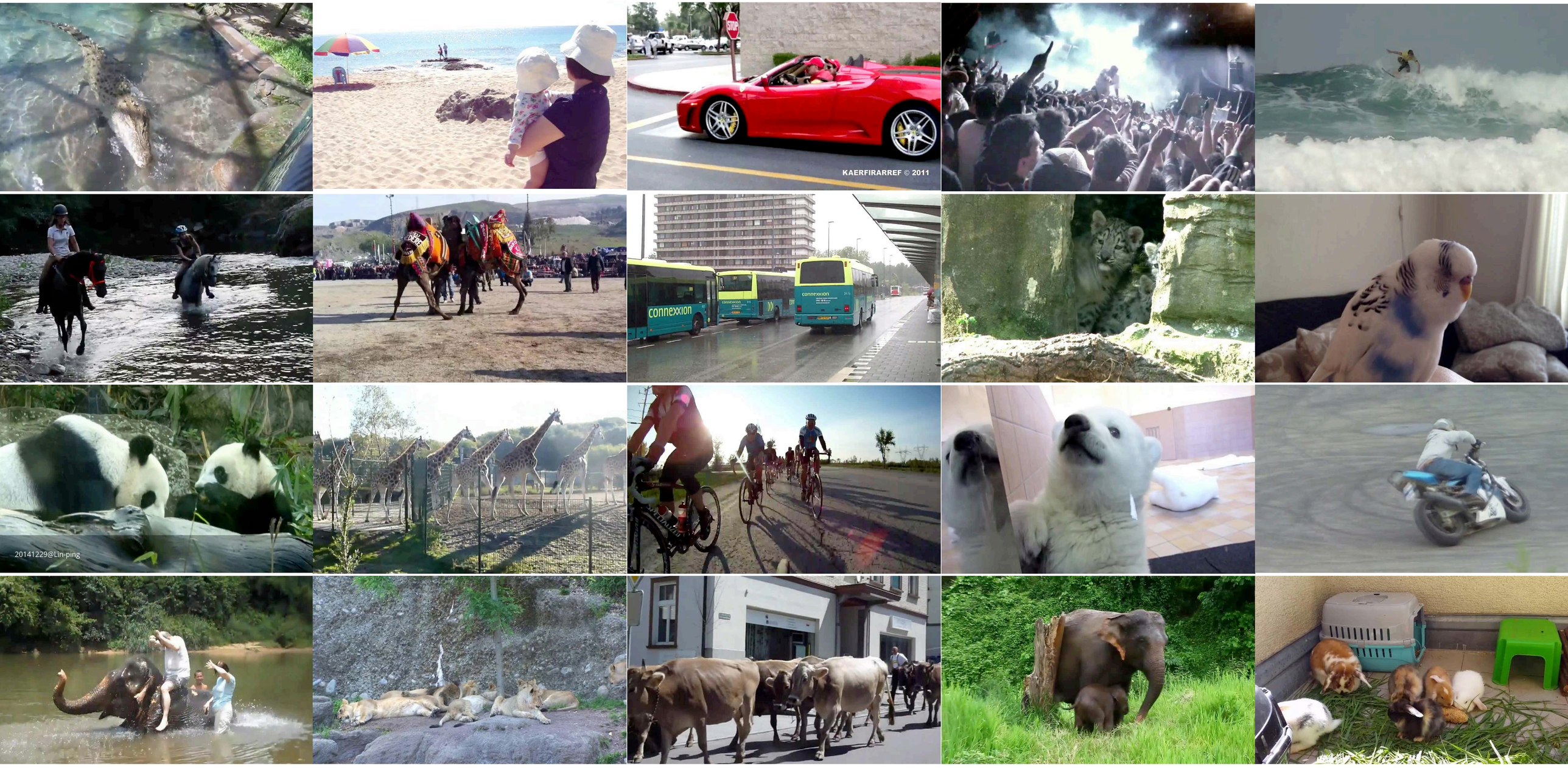




[Justin Wood et al, 2016] :
A smoothness constraint on the development of object recognition.

The development of newborn object recognition in fast and slow visual worlds.

What Can A Model Learn from Nothing but Data?



Baby Vision vs. Grown Vision

From Undivided Sensation to Bounded Rationality

blooming
buzzing
confusion

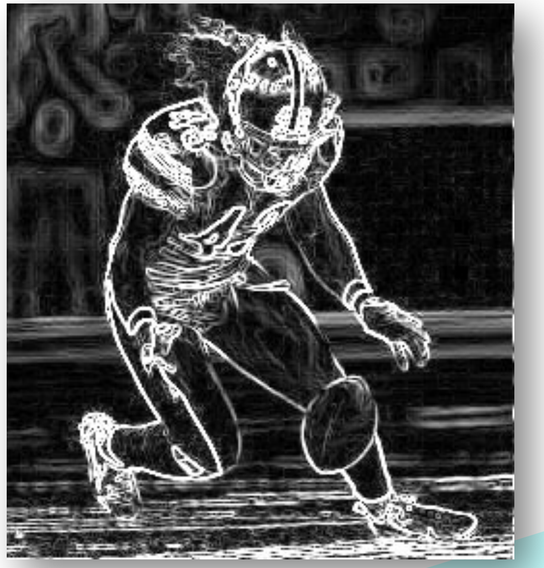
-William James

Bottom-Up Approach to Visual Recognition

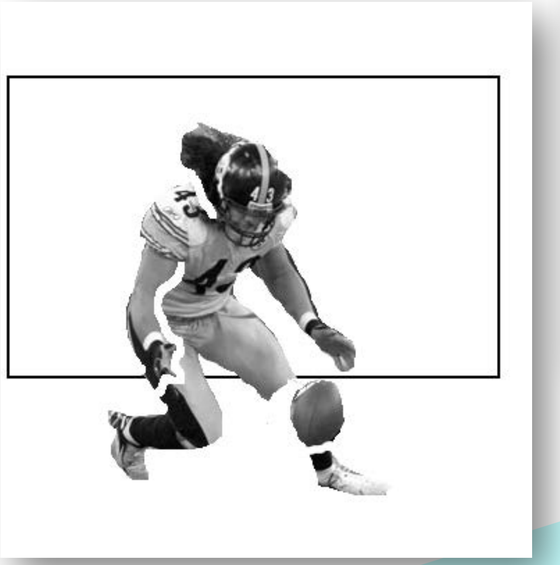
universal but brittle

image

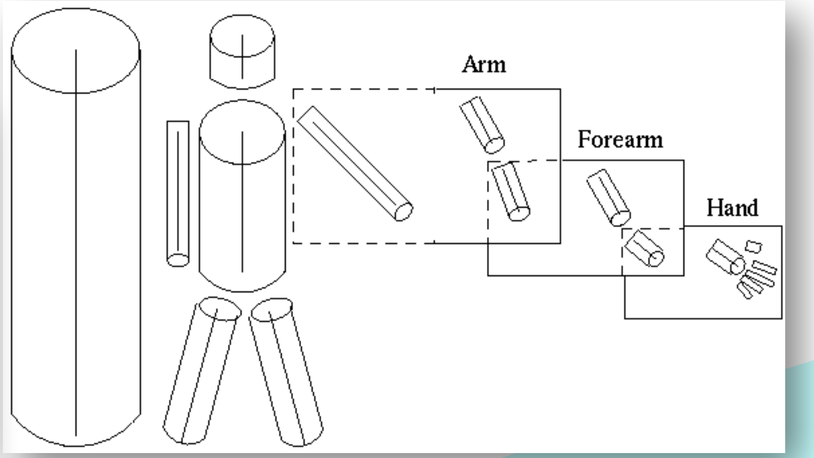
primal sketch



2.5D-sketch



3D models



low-level

mid-level

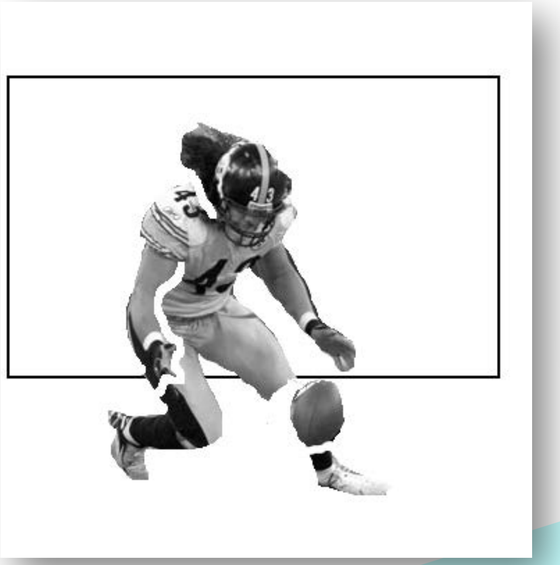
high-level

[Marr, 1982]

My Research: Unsupervised Learning of Mid-Level Vision

3D models

2.5D-sketch



primal sketch

image

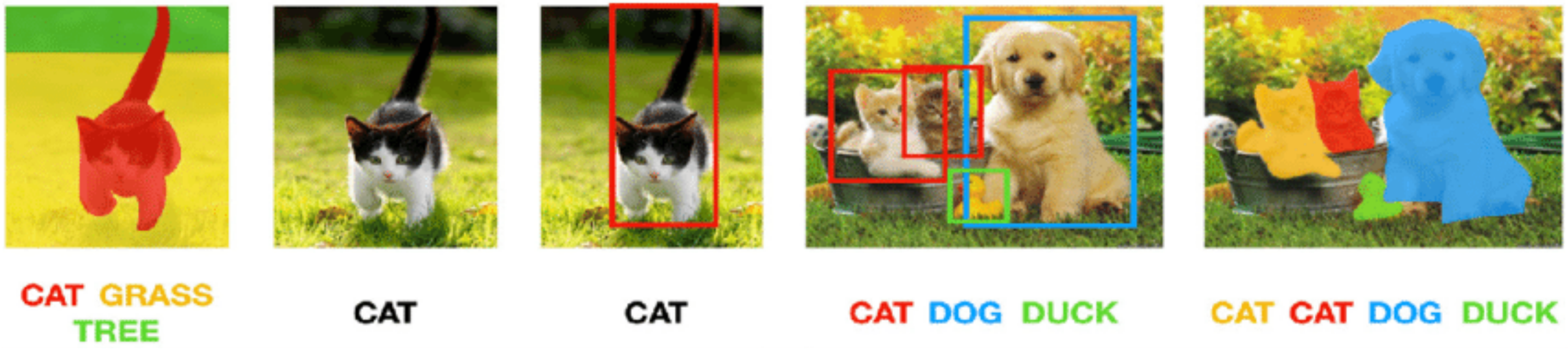
high-level

mid-level

low-level

My Research: Unsupervised Learning of Mid-Level Vision

high-level



mid-level

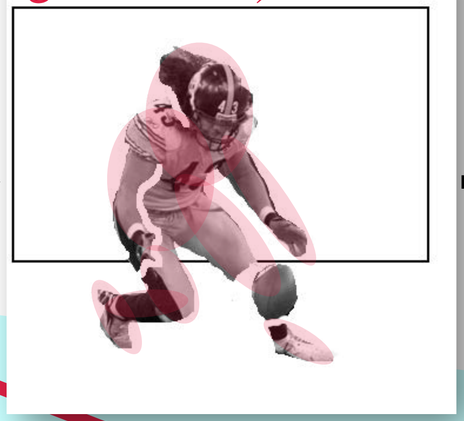
low-level

Unsupervised Learning: Segment Objects, Differentiate Them, Parse into Parts

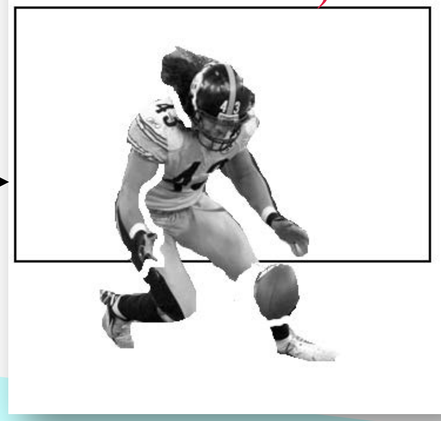
[AMD, NeurIPS 2021; NPID, CVPR 2018; CLD, CVPR 2021; HSG, CVPR 2022]

?

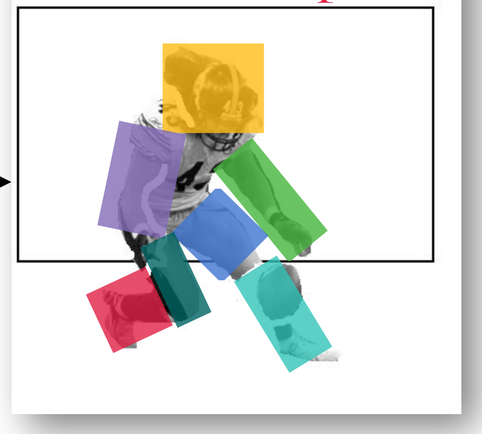
generic objectness

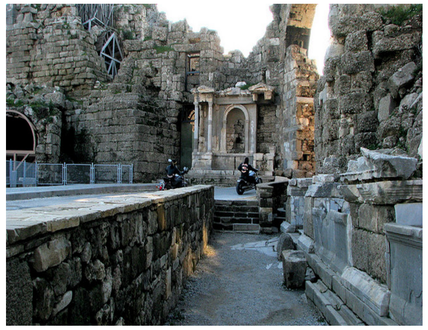


individual objects



hierarchical parts





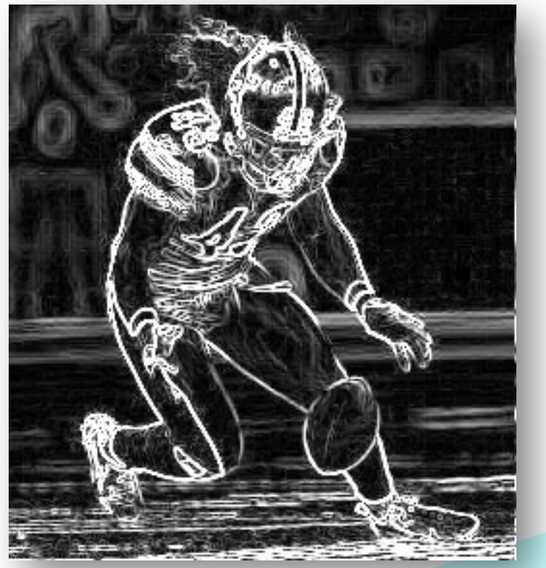
Mid-Level Vision Is the Key

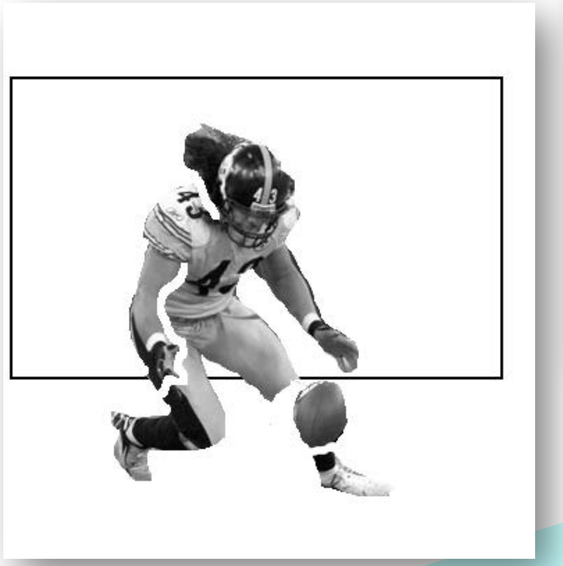
3D models

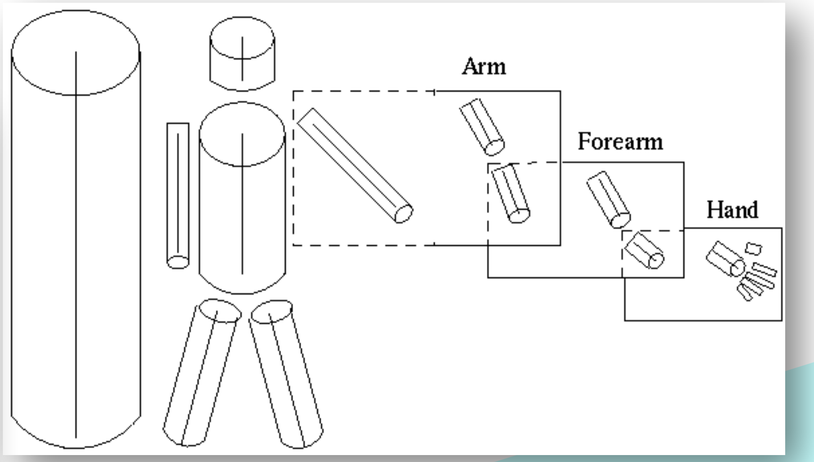
low-level

2.5D-sketch

primal sketch





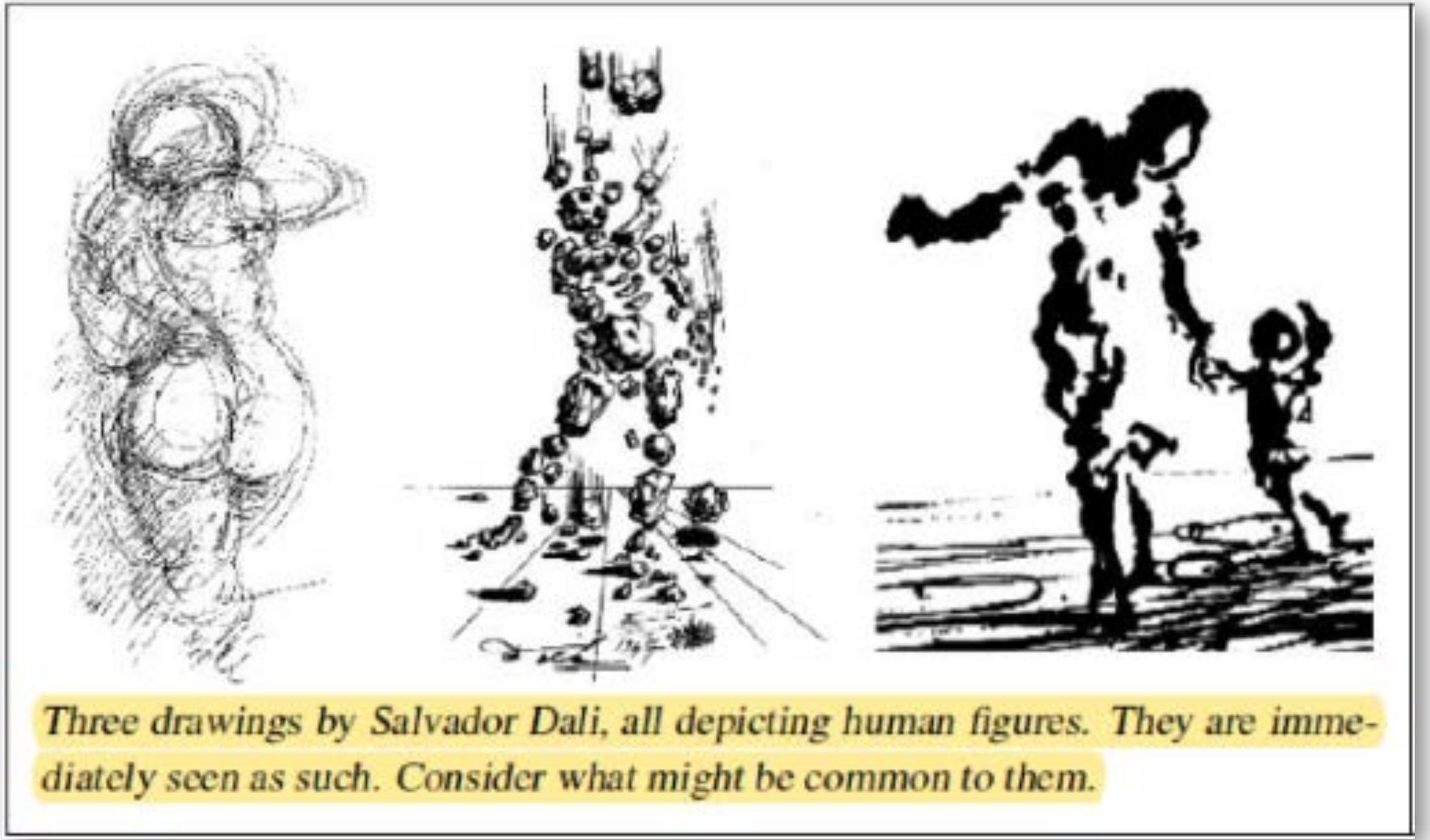


image

high-level

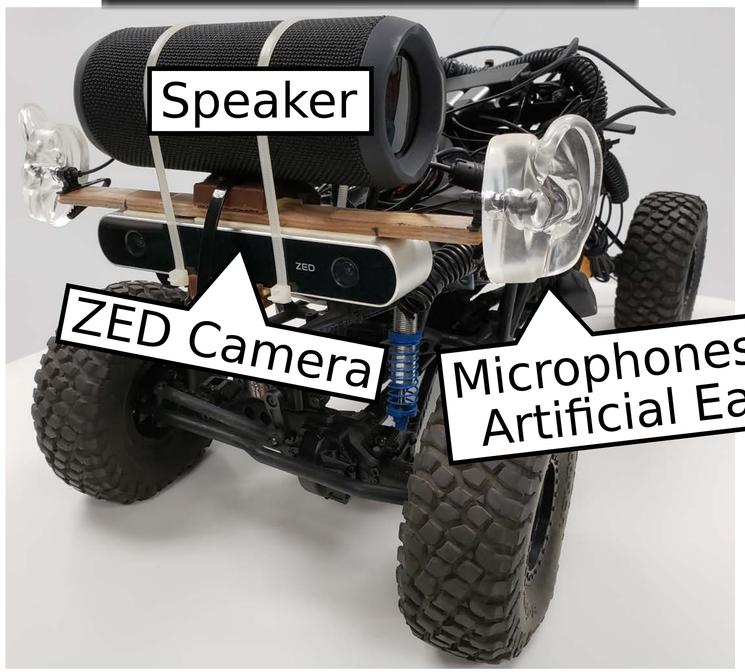
mid-level

J. J. Koenderink: Edges Are Imposed, Not Detected

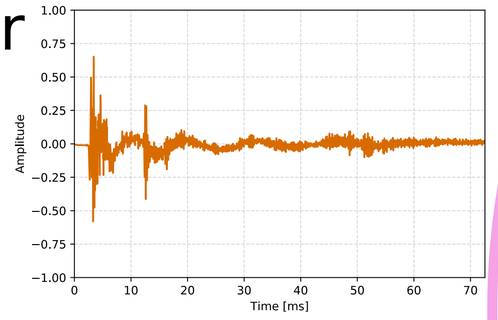


Three drawings by Salvador Dali, all depicting human figures. They are immediately seen as such. Consider what might be common to them.

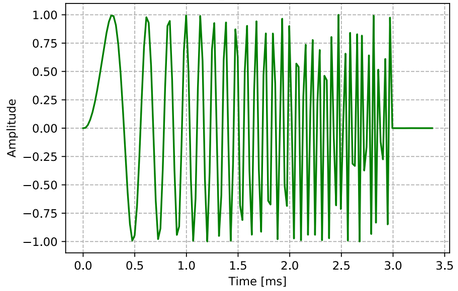
Unsupervised Ambient Sound Recognition for Localization / Navigation



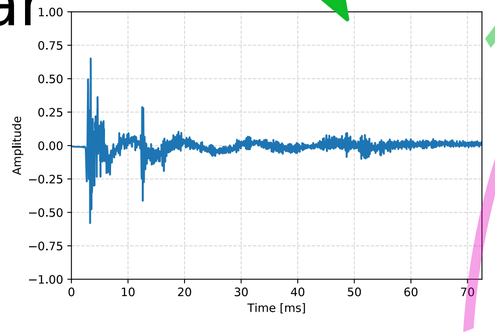
Left Ear



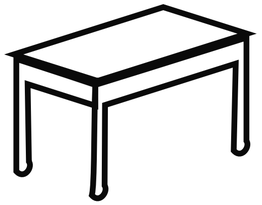
Sound Chirps

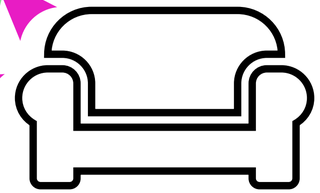


Right Ear



Echos of Environment

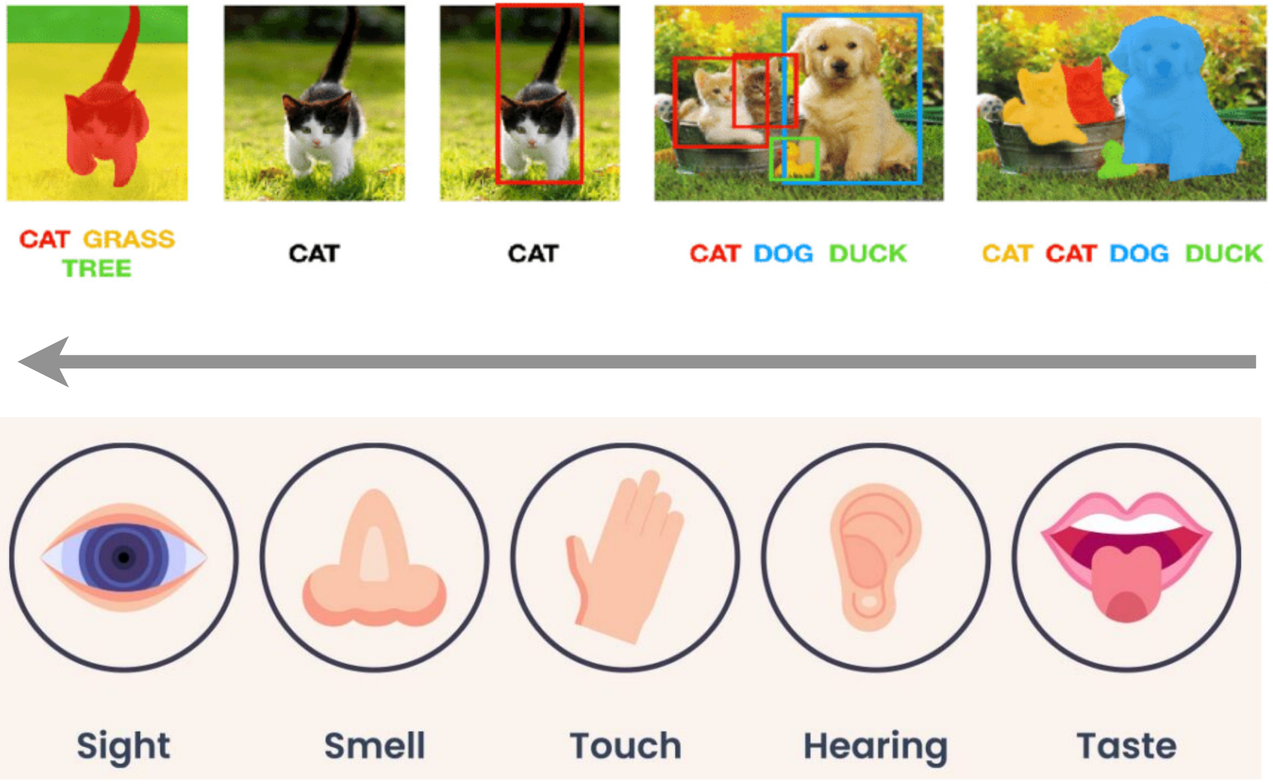




[BatVision, ICRA 2020, CVPRW 2020; Audio-NPID, ICASSP 2021]

Objective: What Is A Baby Supposed to Learn?

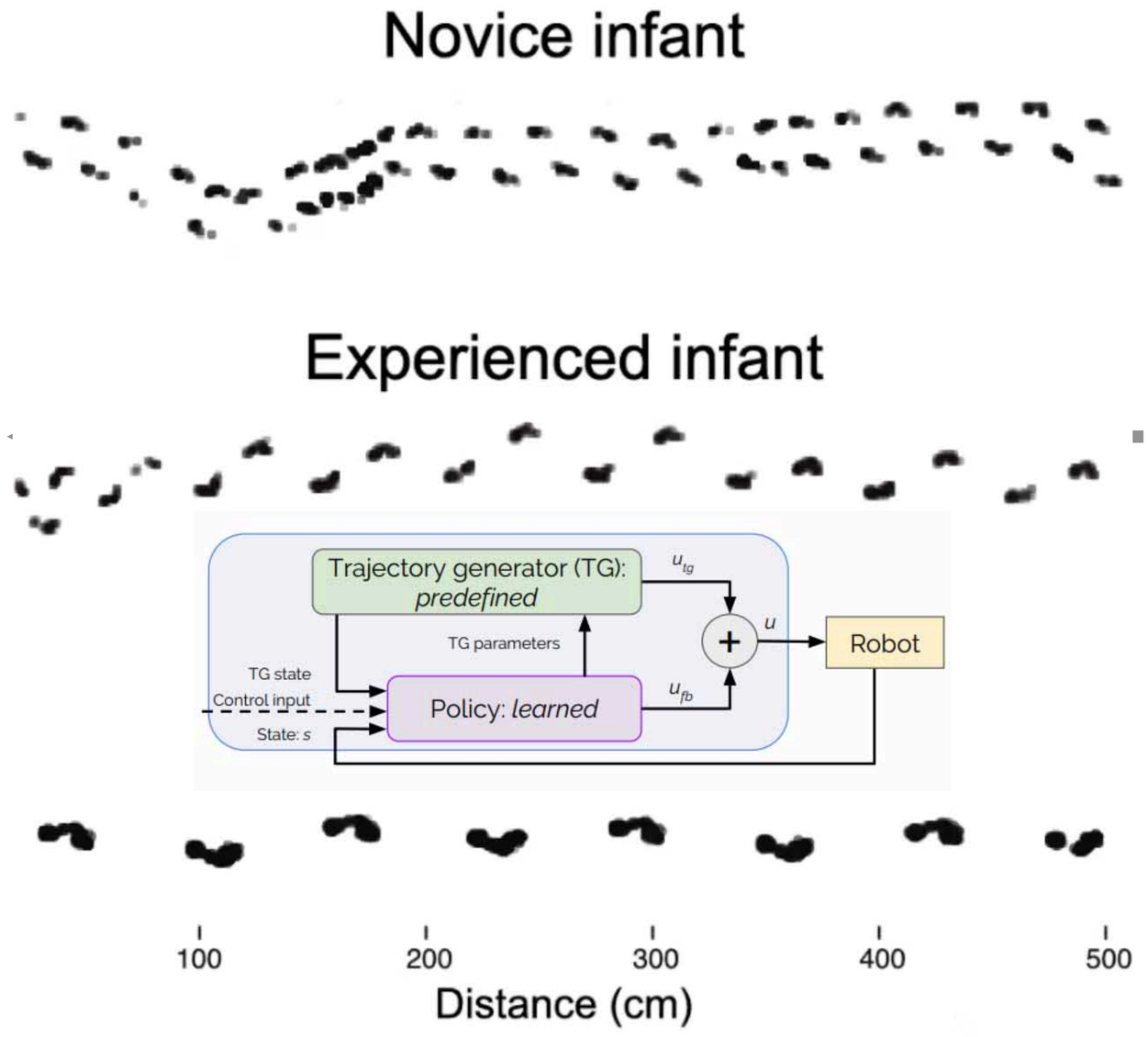
Model



Researcher

Objective: What Is A Baby Supposed to Learn?

Model



Researcher

Different Skills Learned at Different Times

Sit

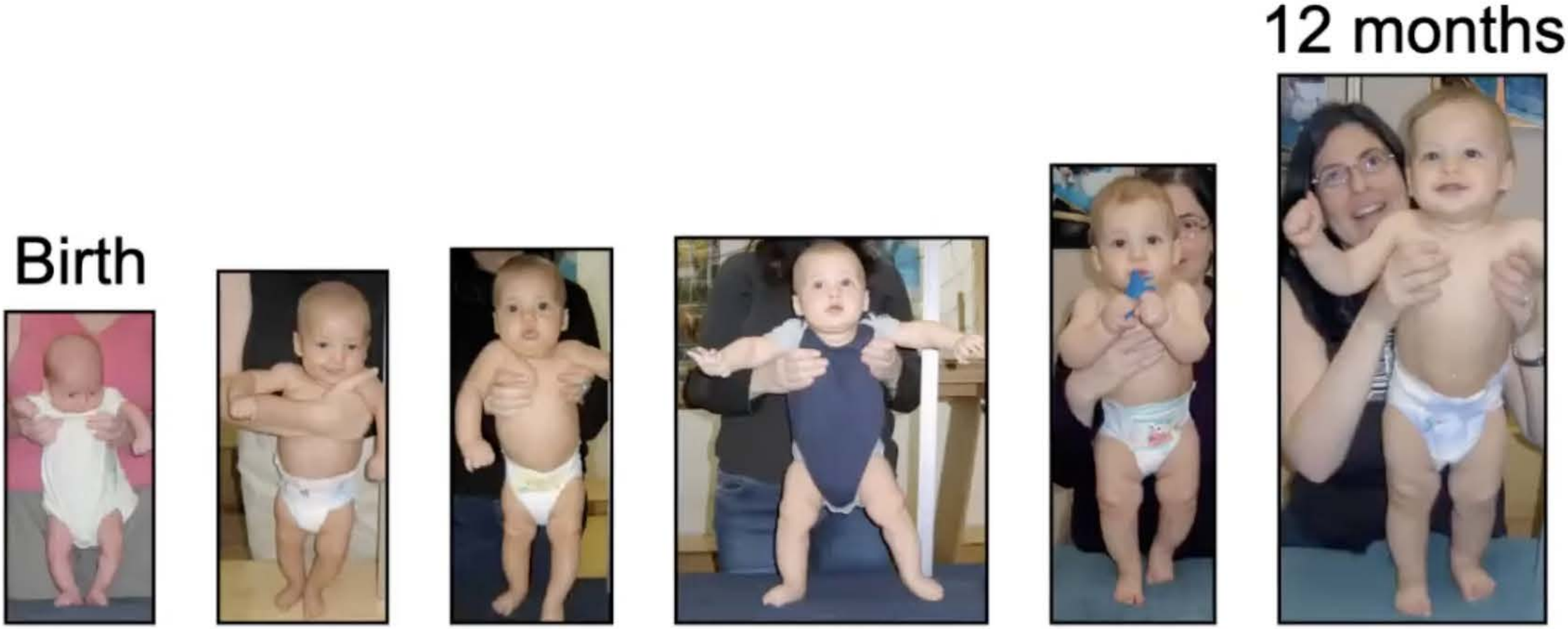
Crawl

Cruise

Walk

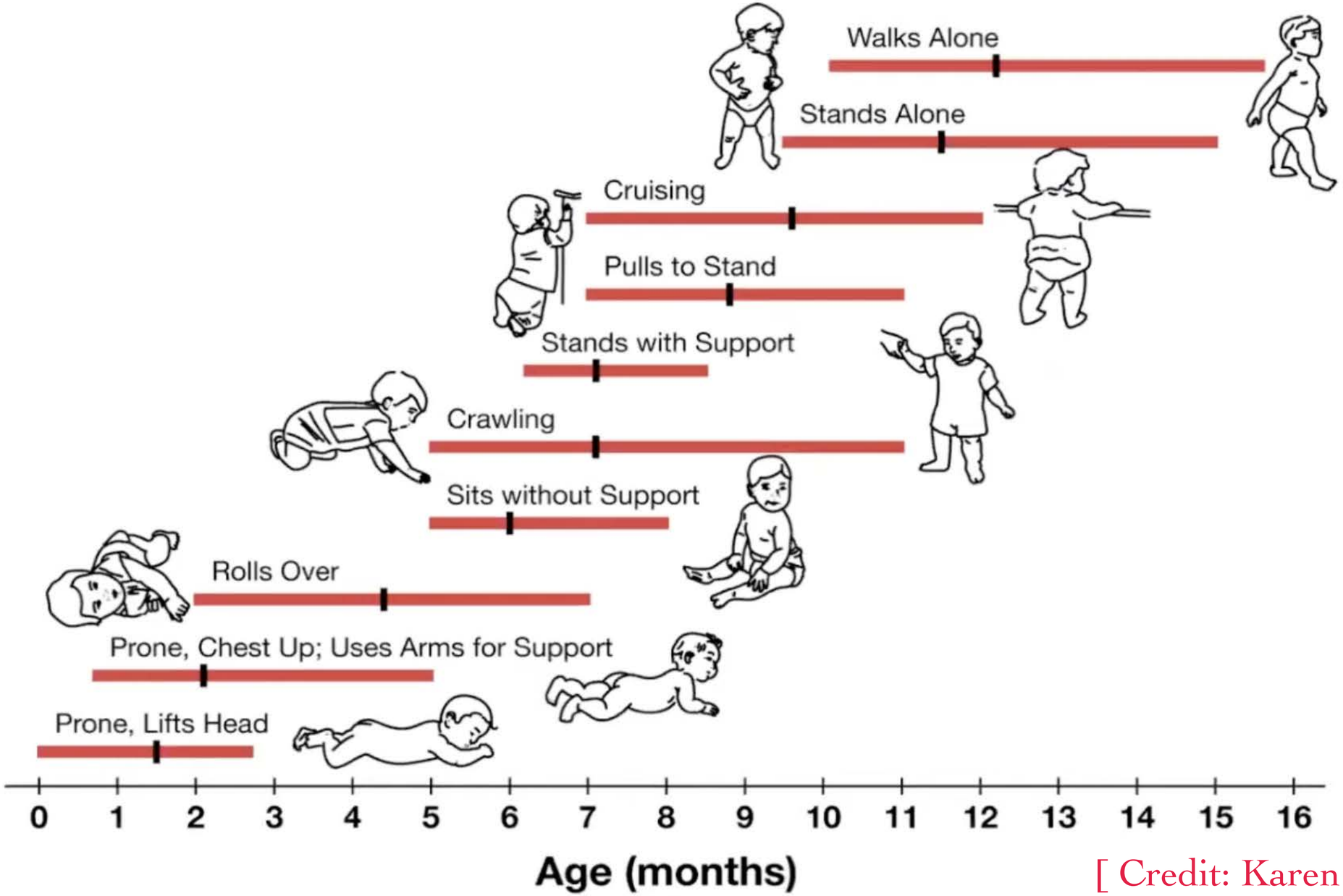
[Credit: Karen Adolph]

Different Skills Learned with Different Bodies



[Credit: Karen Adolph]

Different Skills Learned with Different Data



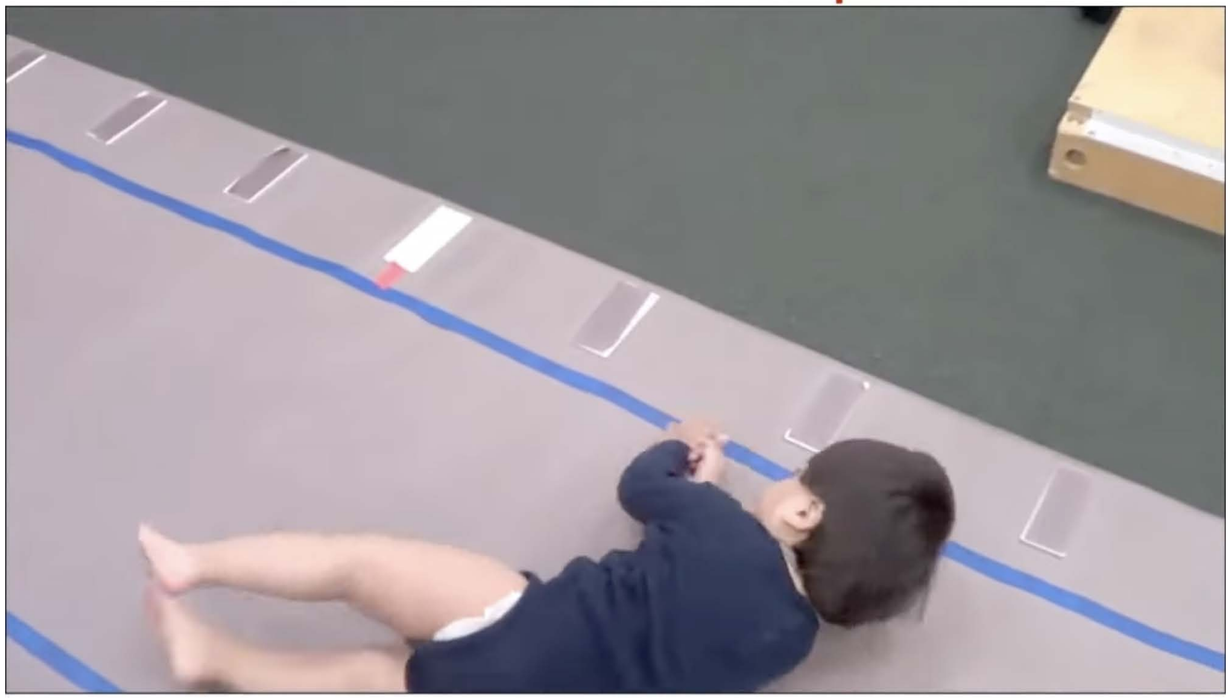
[Credit: Karen Adolph]

Skills Learned with Different Sensitivities to Errors

Example #2: High-impact errors (negative reward) shape learning

Robinovitch (2018). *Databrary.* https://nyu.databrary.org/volume/739

Infant walking is full of errors: But falls are low impact



Adolph (2012) *Psych Sci*; Han & Adolph (2021) *Dev Sci*

Learn the Simultaneous Development of Action and Perception

We see in order to move

We move in order to see

During the process we learn both how/what to see/move

TABLE 1. MARR'S THREE LEVELS OF EXPLANATION FOR COGNITIVE CAPACITIES (Marr 1982, 24).

Computational Theory	Representation and Algorithm	Hardware Implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Some Questions to Ponder

- How does our blurry visual system learn to acquire a clear image?
- How do we start to see depth from 2D images?
- How do we start to see colors? And colors of what?
- How do we learn ocular motor control?
- How do we learn reaching and grasping?
- How do we learn locomotion?
- How do we learn manipulation?
- …