DS-GA.3001 Embodied Learning and Vision

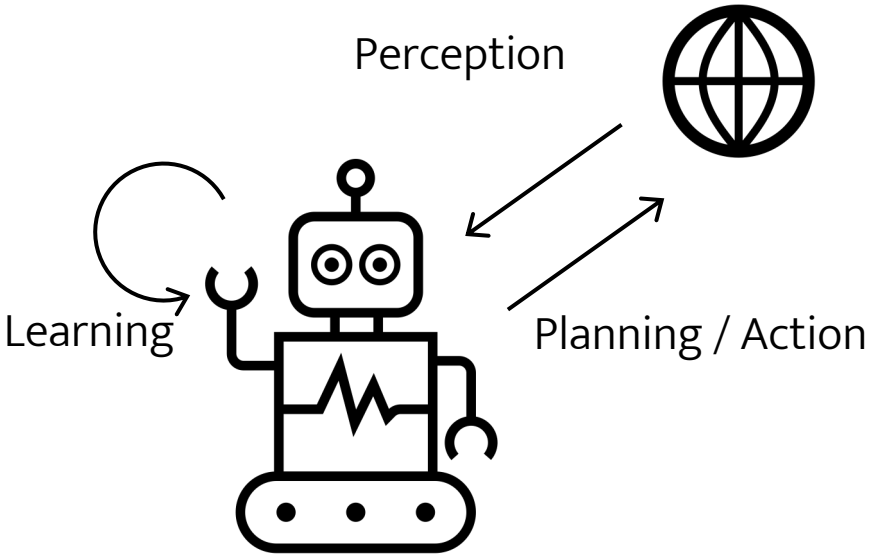
Mengye Ren

NYU

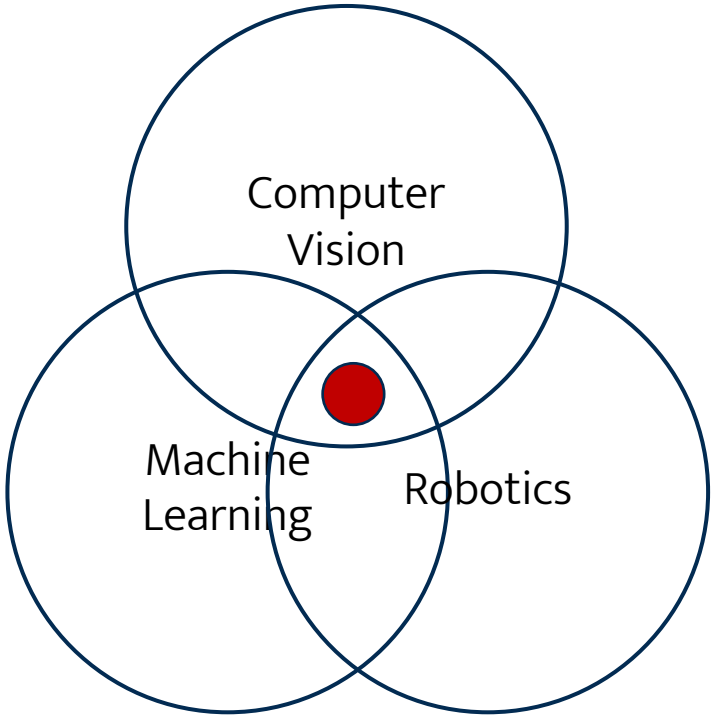
Spring 2025

embodied-learning-vision-course.github.io

What is this course about?

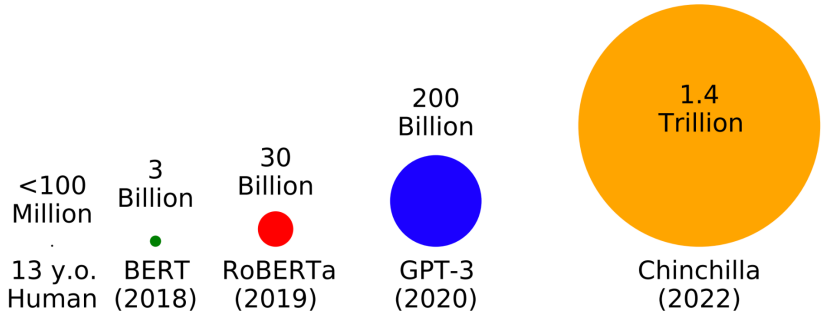


A General-Purpose Learning Agent



NYU

Human vs. Machine Learning



NYU

Language Vs. Embodied Video?

Yann LeCun
@ylecun

* Language is low bandwidth: less than 12 bytes/second. A person can read 270 words/minutes, or 4.5 words/second, which is 12 bytes/s (assuming 2 bytes per token and 0.75 words per token). A modern LLM is typically trained with 1x10^13 two-byte tokens, which is 2x10^13 bytes. This would take about 100,000 years for a person to read (at 12 hours a day).

* Vision is much higher bandwidth: about 20MB/s. Each of the two optical nerves has 1 million nerve fibers, each carrying about 10 bytes per second. A 4 year-old child has been awake a total 16,000 hours, which translates into 1x10^15 bytes.

NYU

- Multimodal
- Incremental
- Physical
- Explore
- Social
- Use language

The Development of Embodied Cognition: Six Lessons from Babies

Linda Smith
Psychology Department
Indiana University
Bloomington, IN 47405
smith4@Indiana.edu

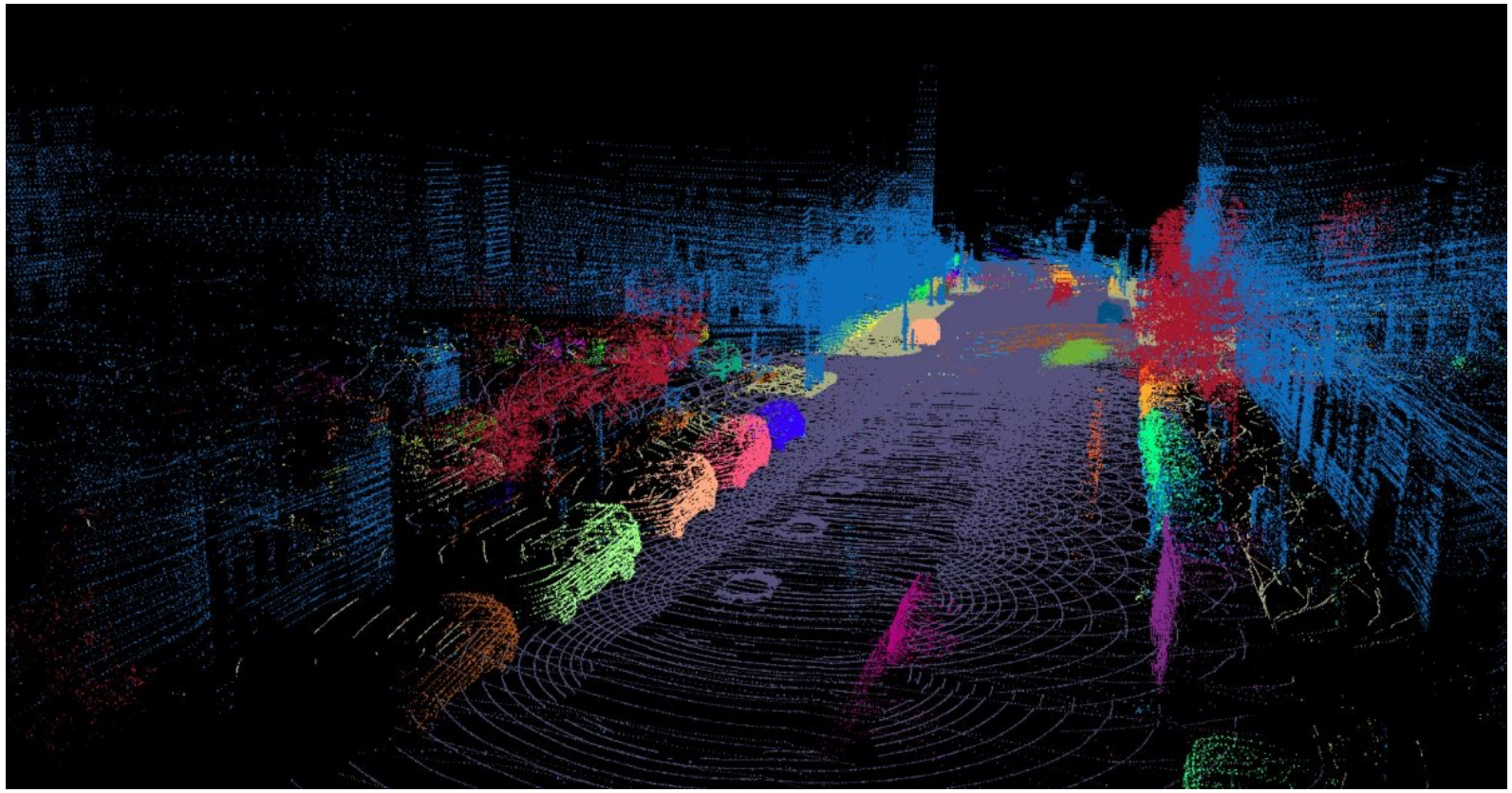
Michael Gasser
Computer Science Department
Indiana University
Bloomington, IN 47405
gasser@Indiana.edu

Abstract The embodiment hypothesis is the idea that intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity. We offer six lessons for *developing* embodied intelligent agents suggested by research in developmental psychology. We argue that starting as a baby grounded in a physical, social, and linguistic world is crucial to the development of the flexible and inventive intelligence that characterizes humankind.

Keywords
Development, cognition, language, embodiment, motor control

NYU

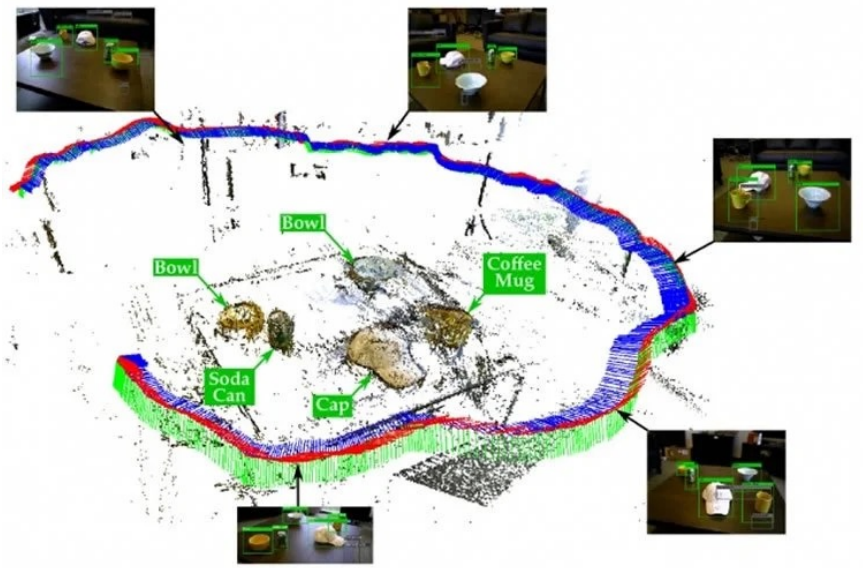
Physical: Geometric and Temporal Structure

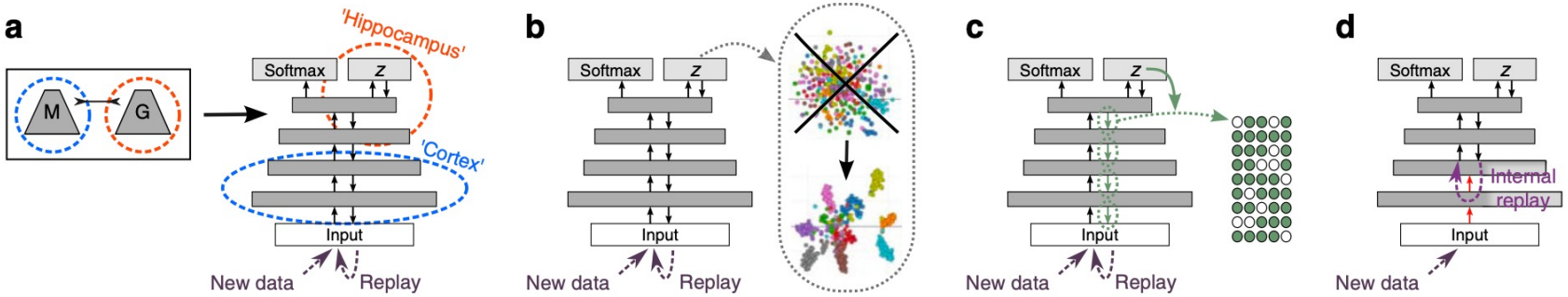


NYU

Incremental: Learning and Memory

- Spatial memory (mapping)
- Episodic memory (autobiography)
- Semantic memory (rule learning)
- Procedural memory (skill learning)
- Replay: Generative or storage





van de Ven, 2020

NYU

Explore: Learning Efficiency

NYU

Learning Objectives

- Solving embodied problems using deep learning tools
- Leverage geometric and temporal structure from real-world and simulated data

- Advanced graduate level course
- Develop research skills
- Conduct cutting edge research

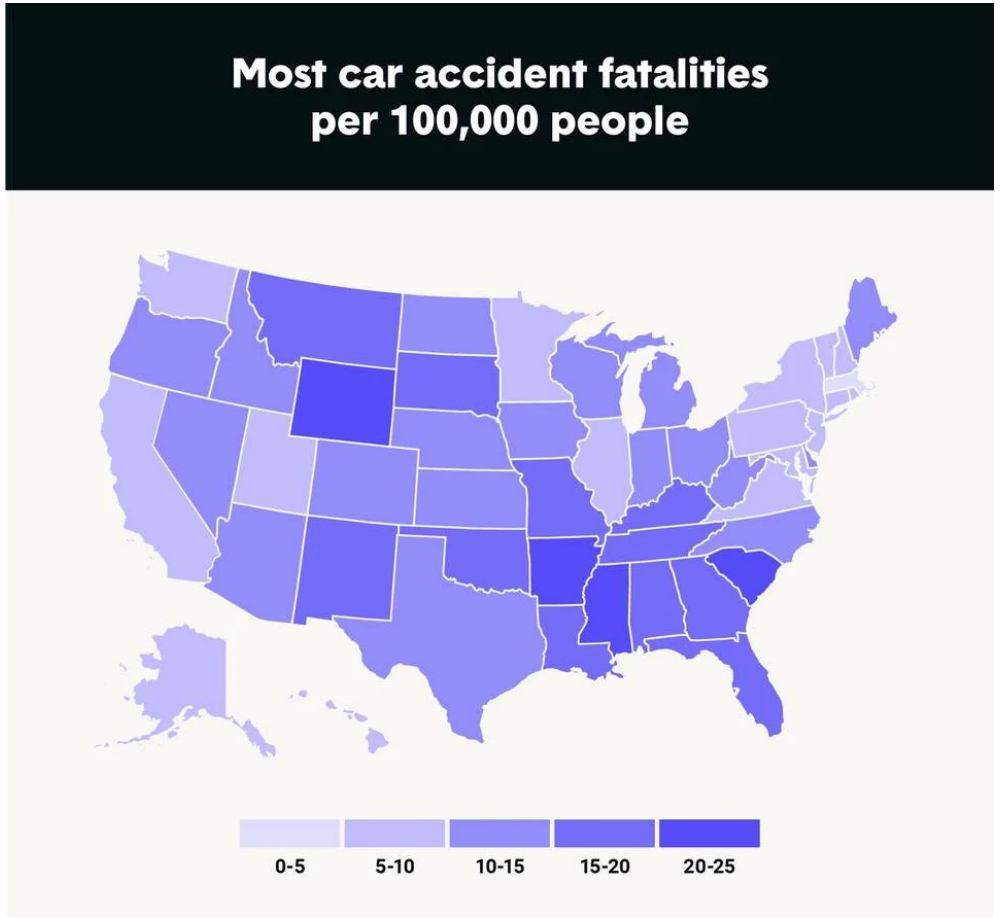
NYU

Applications

NYU

Why Self-Driving Cars

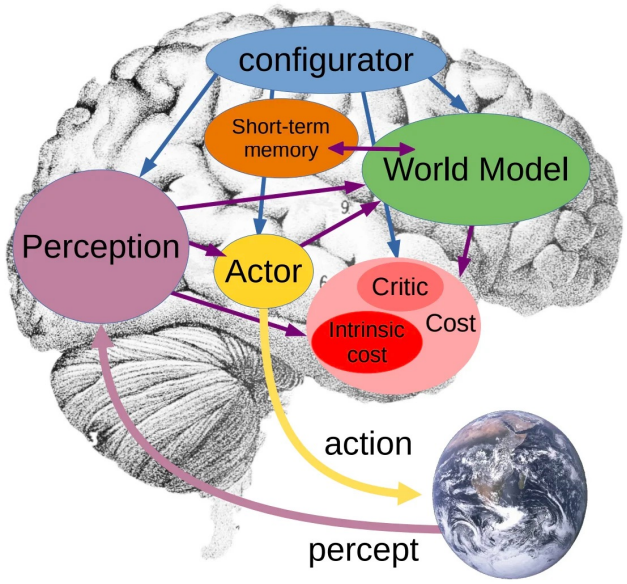
- >1 million people die every year from road
- One of the top 10 leading causes of death and injuries
- Environmental causes
  - Car/battery manufacturing

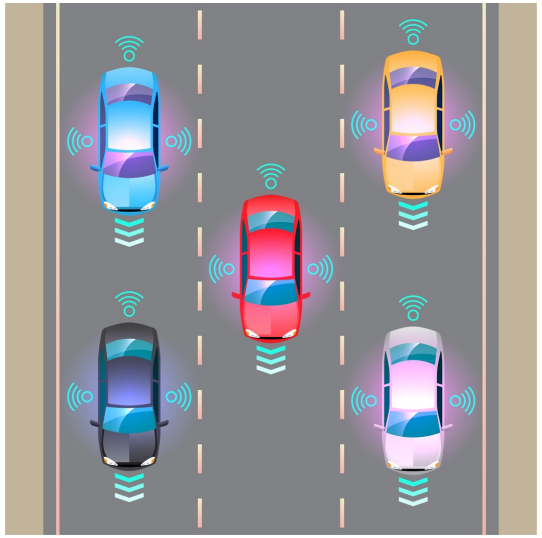


NYU

A Test-Bed for General Embodied Intelligence

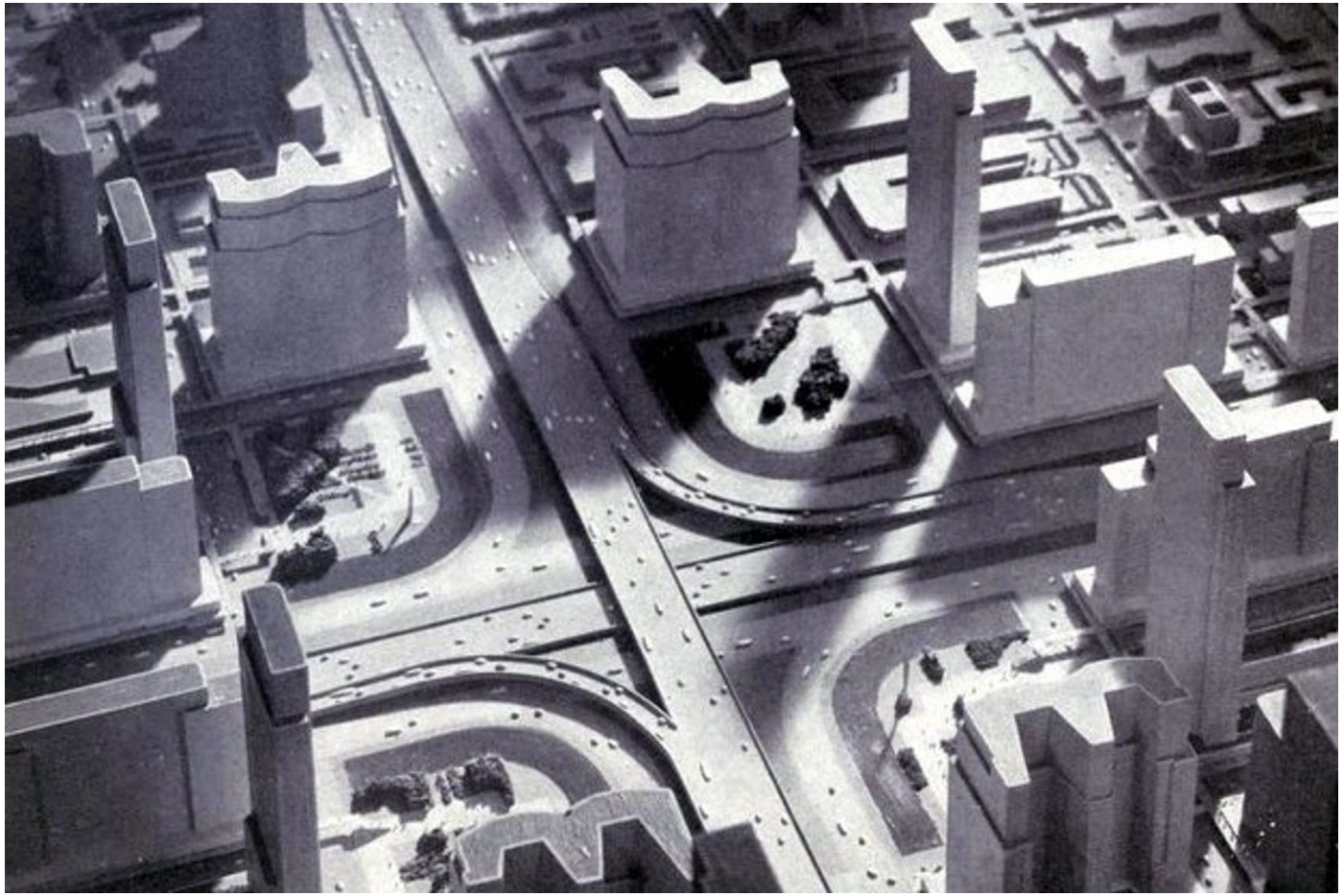
- Perception, world model, mapping, planning, and control
- Long-term hierarchical planning
- Closed loop learning
- Multi-agent social cognition, intent inference, communication
- Rule-based learning





NYU

A Brief History of Self-Driving Cars



Norman Bel Geddes' "Magic Motorways" 1939

NYU

1966 LUNAR Stanford University

NYU

Also in 1966

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group — July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

NYU

1988 ALVINN

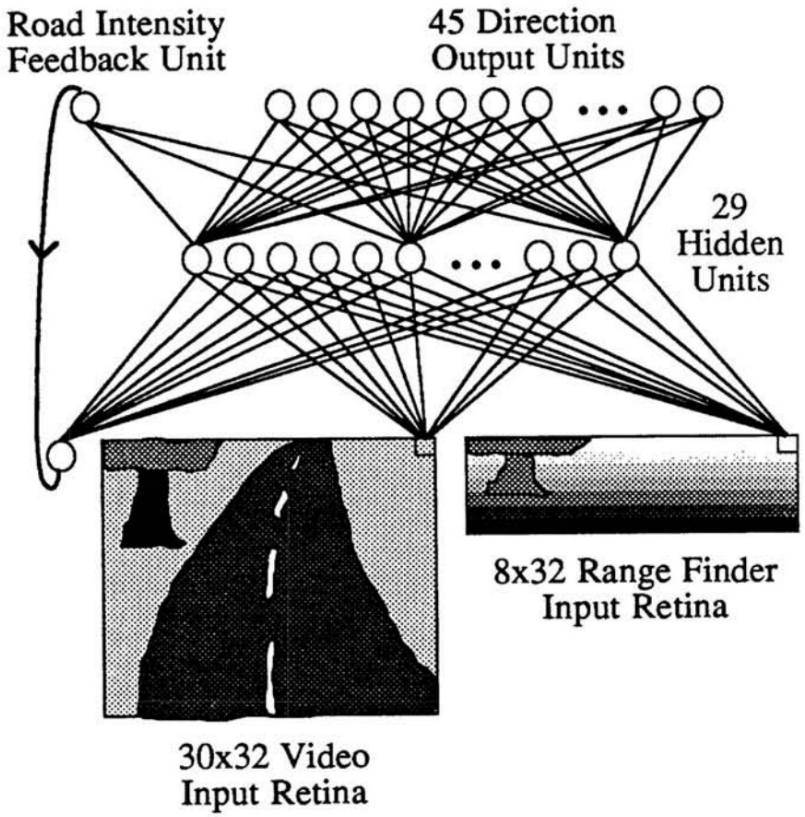


Figure 1: ALVINN Architecture

2004 DARPA Grand Challenge

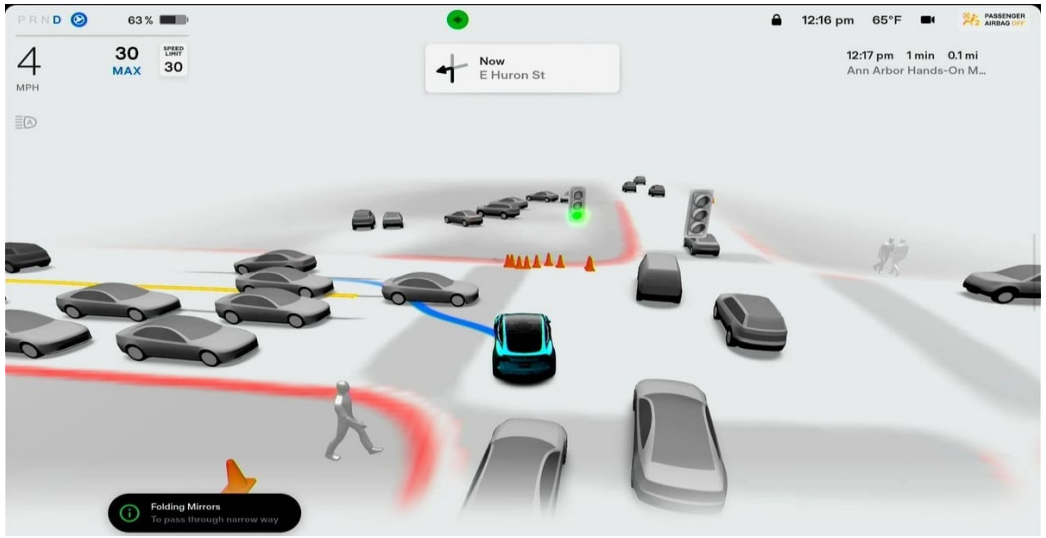
2005 DARPA Grand Challenge

2007 DARPA Urban Challenge

NYU

What Has Changed?

- Proof-of-Concept -> L4 autonomy
- Basic obstacle avoidance and planning -> Joint perception and planning, learning from massive data
- No deep learning involved -> Fully deep learning stack



NYU

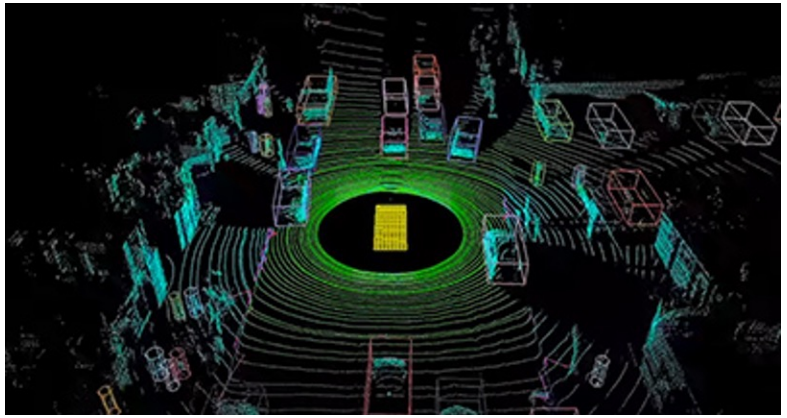
What is Missing General Embodied Intelligence?

- Self-driving as an example for a more general intelligence
- It has perception, mapping, planning, and multi-agent communication.
- Where are we in general embodied intelligence?
- Perception: Exploring new environments, recognizing new signs, objects, etc.
- Learning: Learning from world modeling, causal relations, from language instructions, etc.
- Memory and mapping: Efficient exploration of new environment without maps.
- Adaptation: Adapting to different hardware, and environments.

NYU

Better Perception

- Deep learning, semantic understanding
- Massively labeled data for training
- Sensor fusion: Camera, Multi-View, LiDAR, Radar, Motion

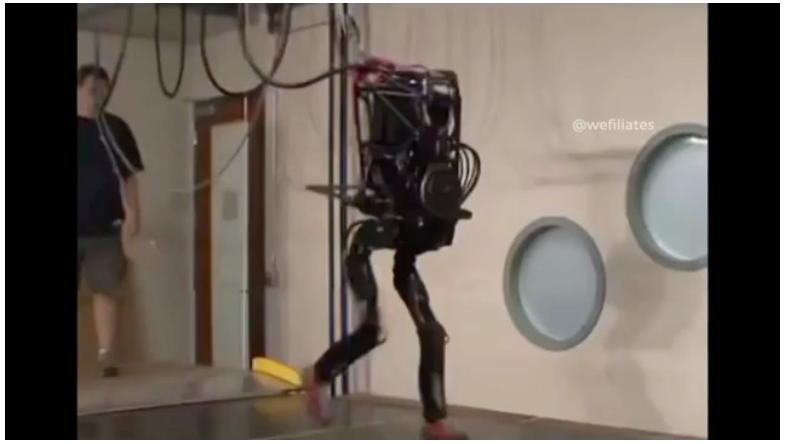


NYU

Better Control

Honda Asimo, 2011

Unitree B2W, 2024

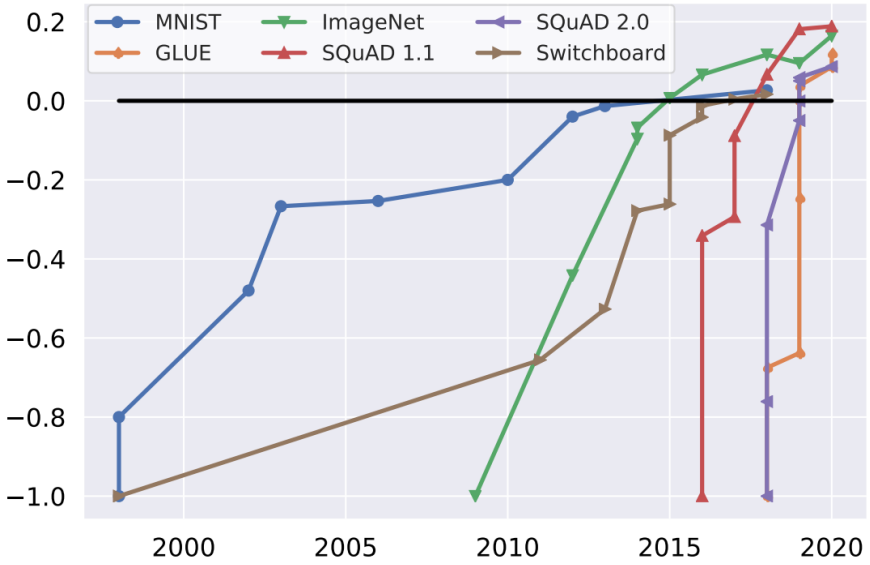


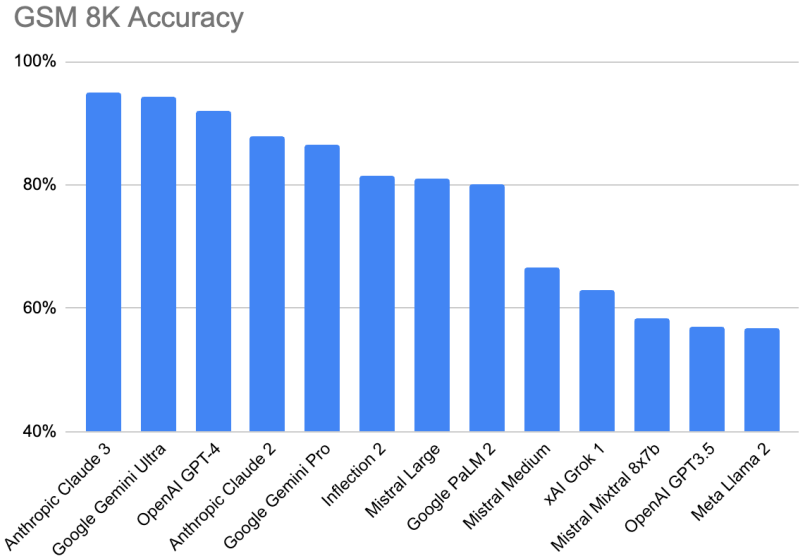
Boston Dynamics 2009-2022

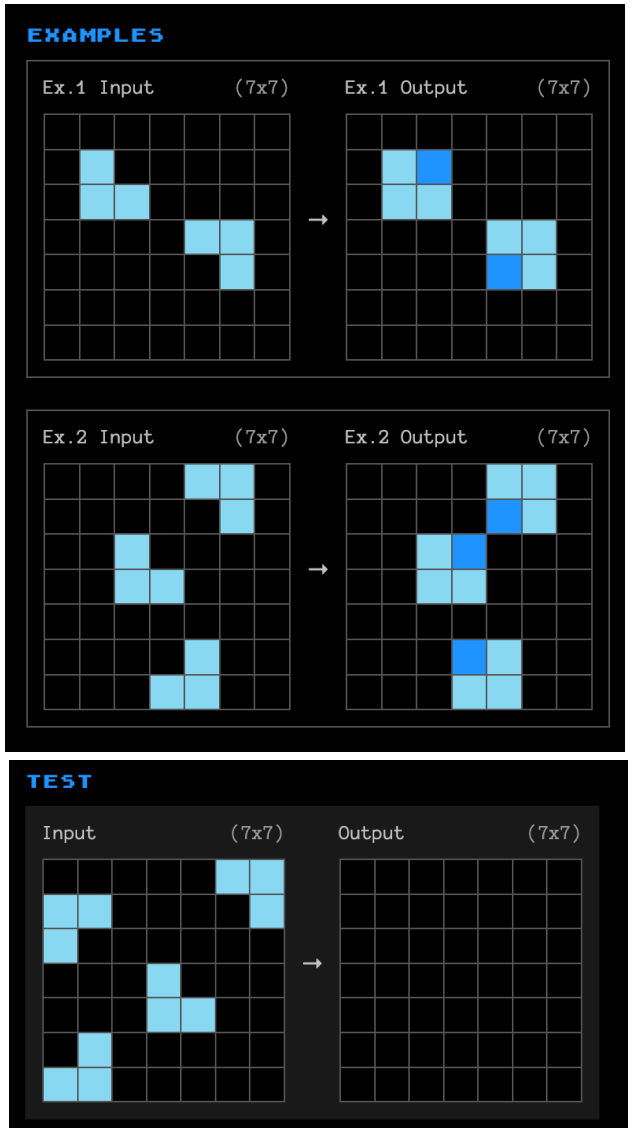
NYU

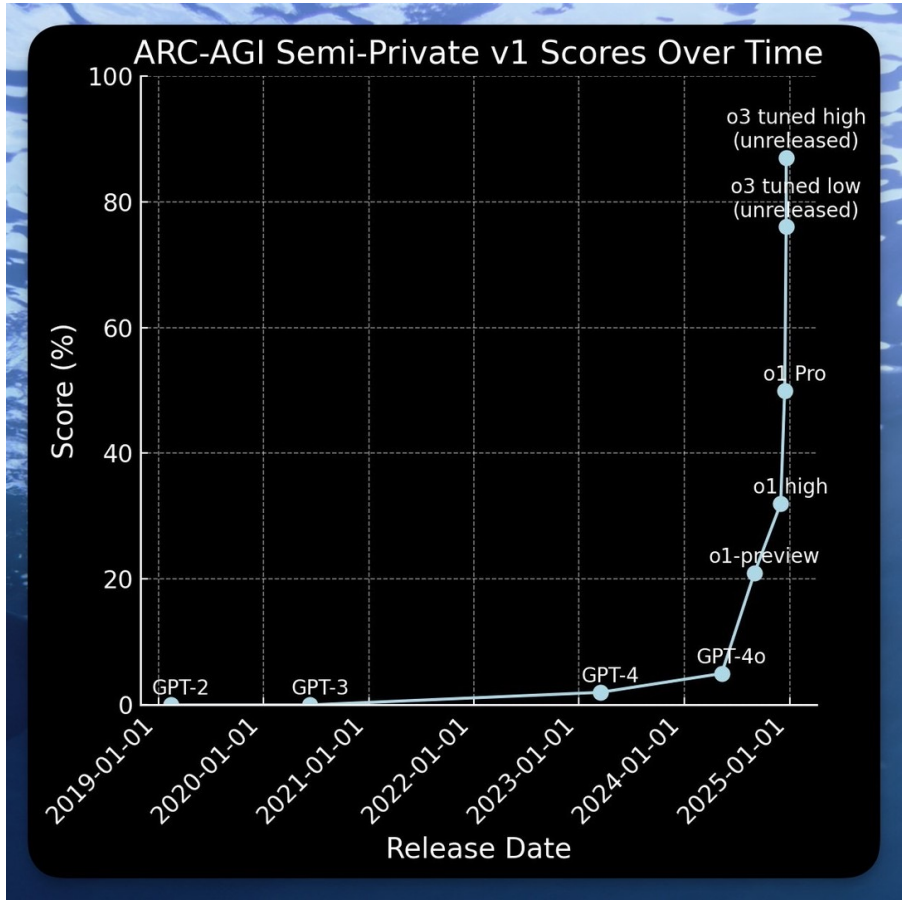
Better Reasoning and Abstraction

Kiela et al. 2021





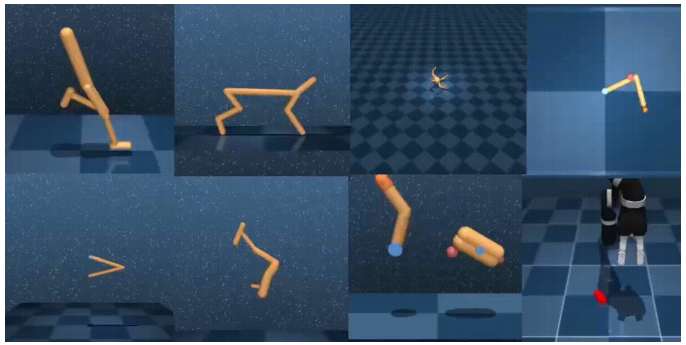




Riley Goodside, 2024

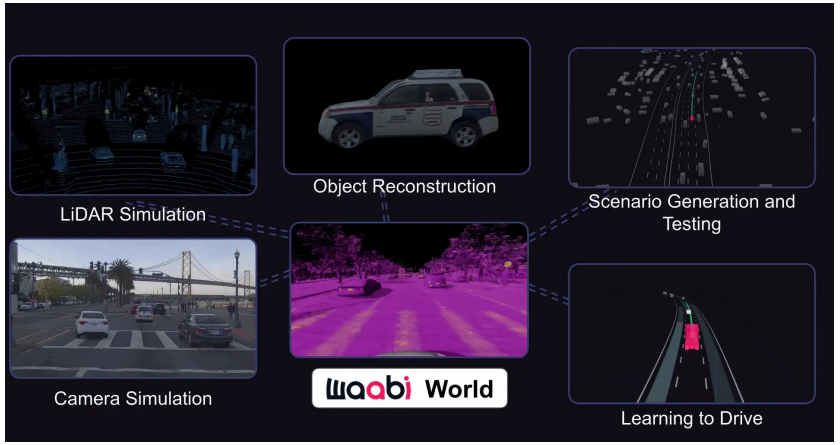
NYU

Better Simulation and Benchmarks

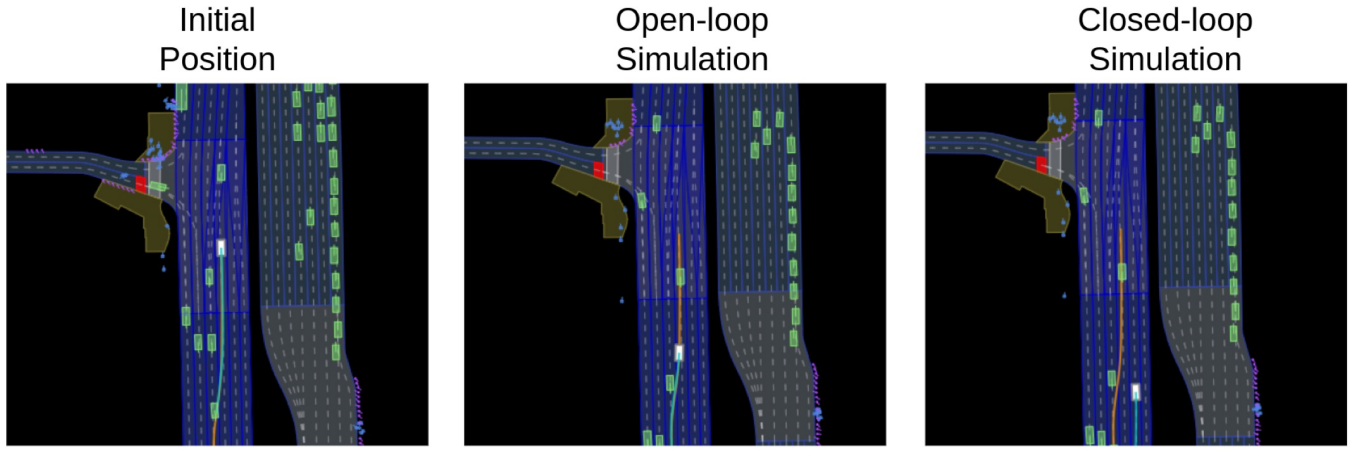


DM Control Suite

MineCraft & MineDojo



Waabi World



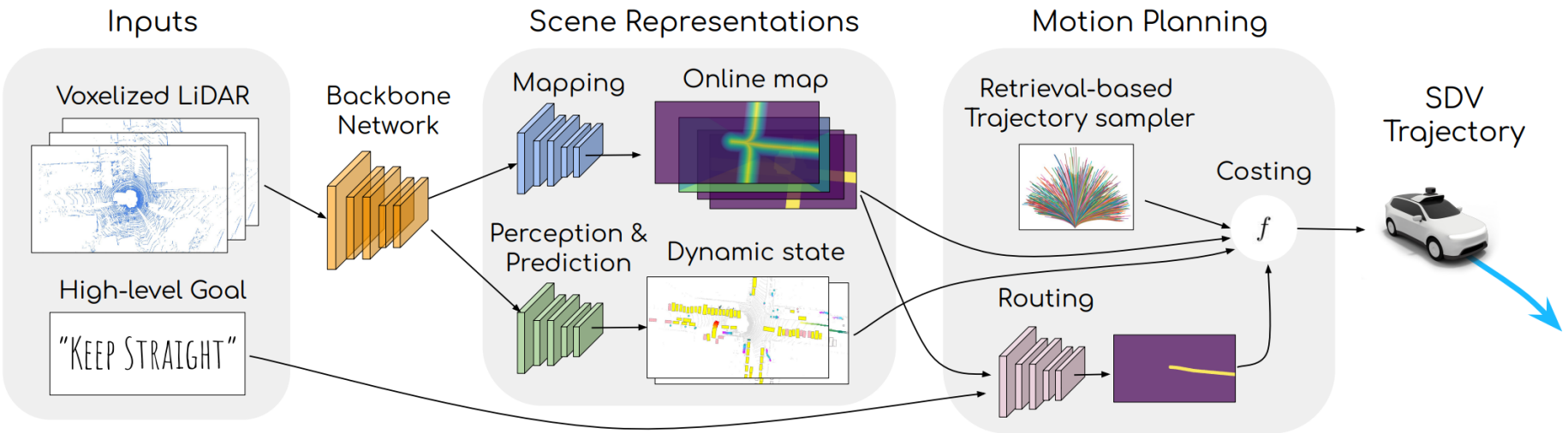
nuPlan

Nvidia Isaac Sim

NYU

End-to-End Learning

- Continuous and differentiable modules for end-to-end learning.
- We know how to optimize deep networks.
- Maintain rich information throughout decision making.
- Representations & output space modeling.



NYU

The Learning Question

- Humans can learn driving in 20 hours
- Current ML requires hundreds of millions of examples
- Ability to learn from noisy streaming data
- Ability to generalize and perform abstraction

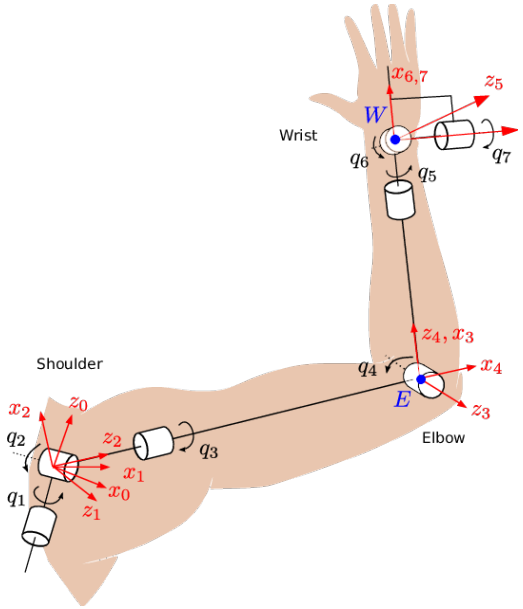
NYU

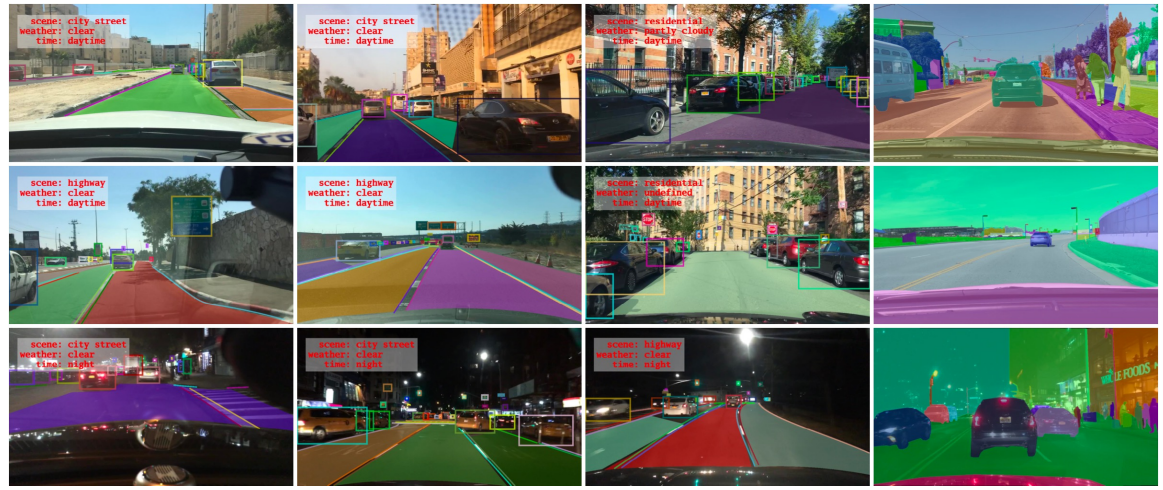
Learning Challenges?

- Learning from physical world
- Exploration, learning efficiency, causality
- Imbalance
- Flexibility
- Generalization
- Continuous adaptation
- Integration
  - Sensory modality, reasoning, and planning

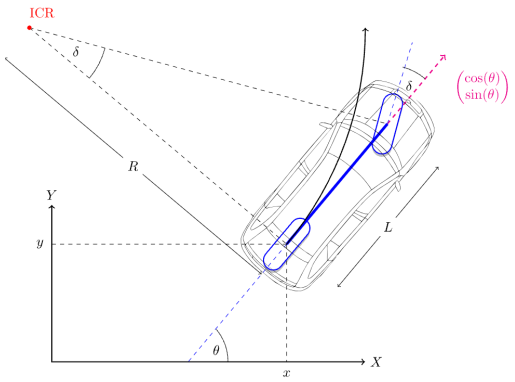
NYU

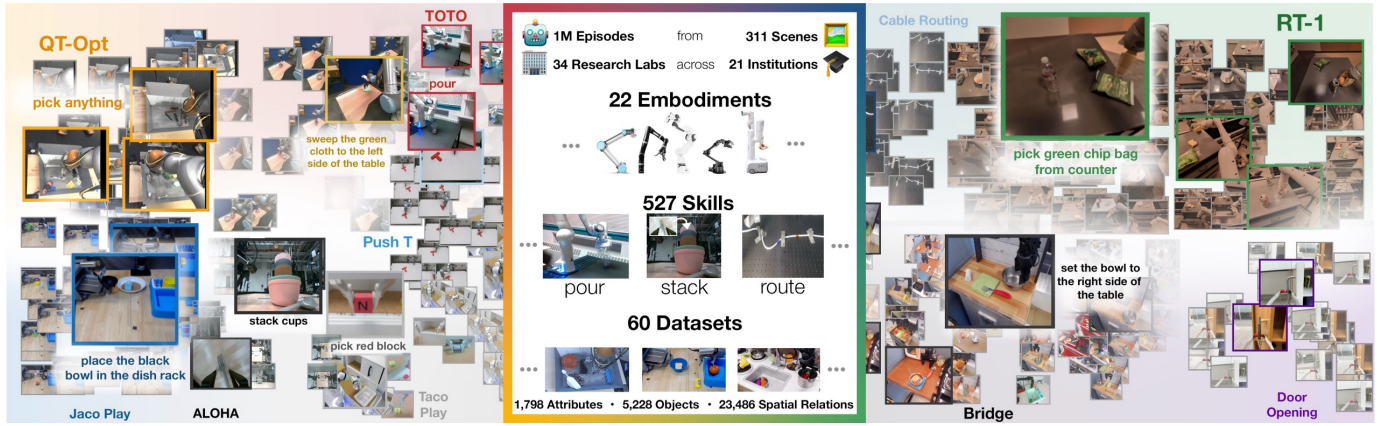
Model Driven Vs. Data Driven





Yu et al. 2018

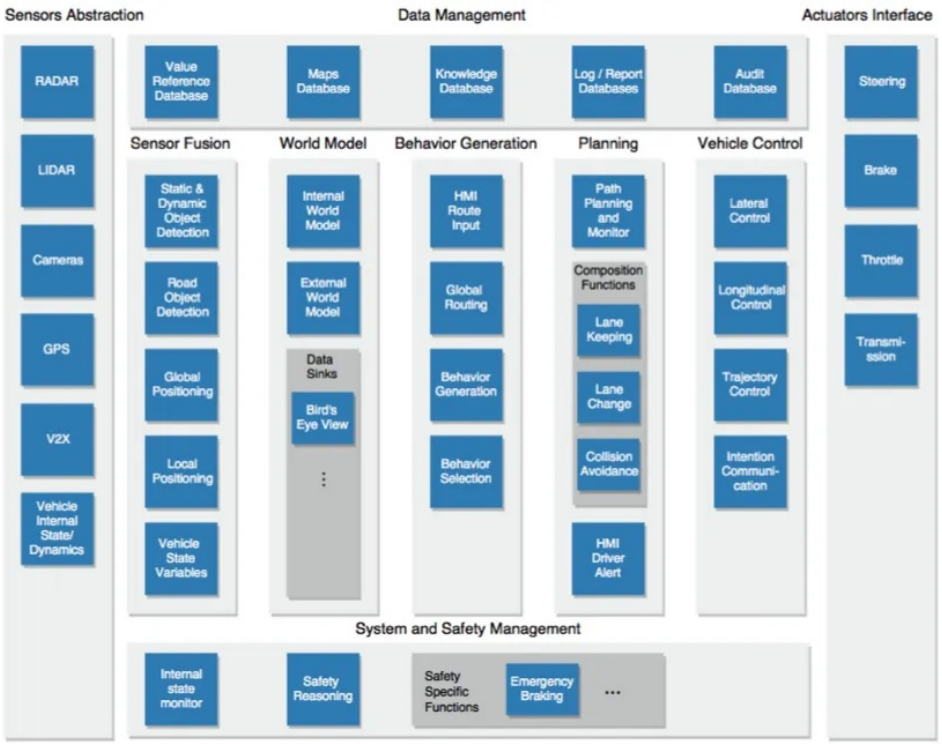


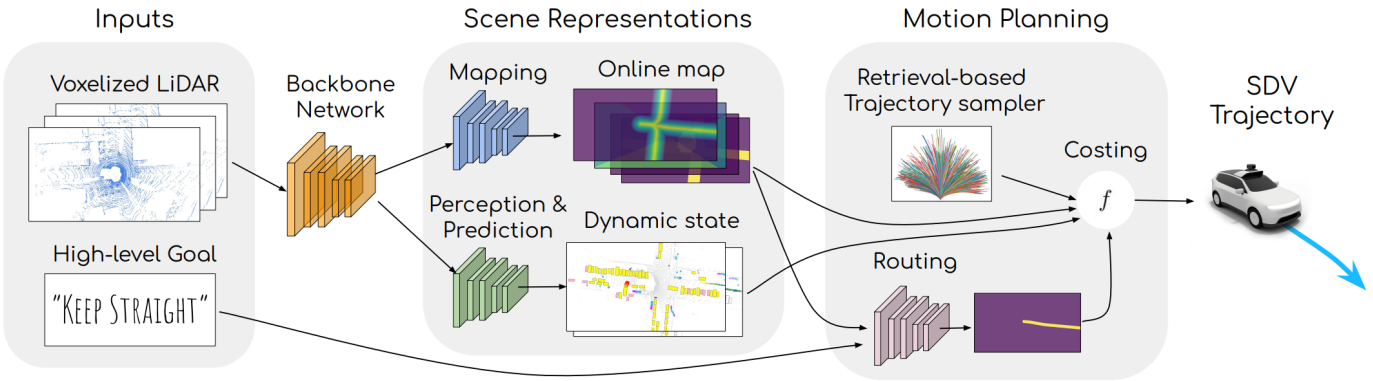


Open X-Embodiment, 2024

NYU

Modular/Symbolic Vs. End-to-End





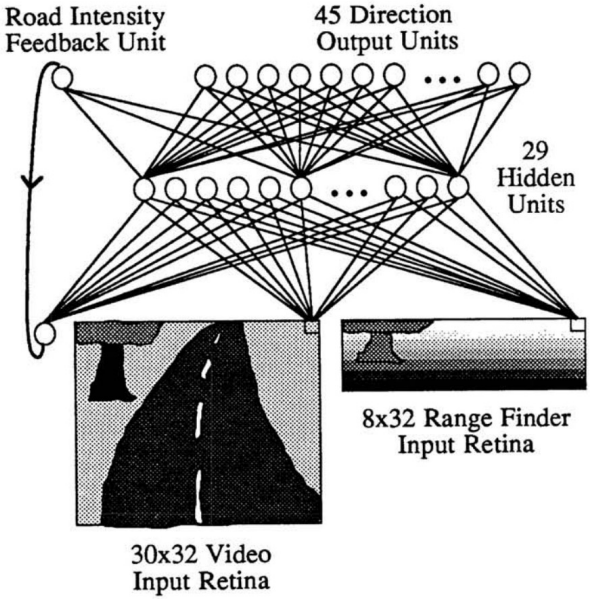
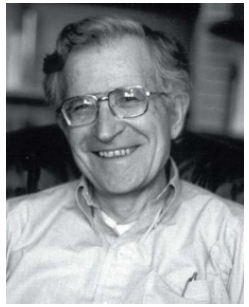


Figure 1: ALVINN Architecture

Nature Vs. Nurture

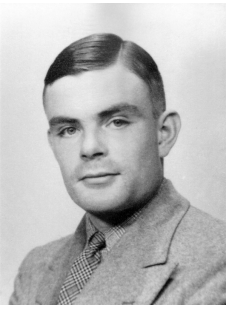


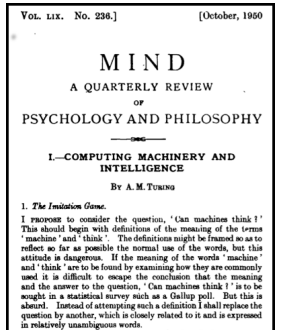
Noam Chomsky

There's an obvious answer to that: the knowledge is built in. You and I can learn English, as well as any other language, with all its richness because we are designed to learn languages based upon a common set of principles, which we may call universal grammar.

Yarek Waszul

In fact, if someone came along and said that a bird embryo is somehow "trained" to grow wings, people would just laugh.





Vol. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW OF PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. *The Imitation Game.*

I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

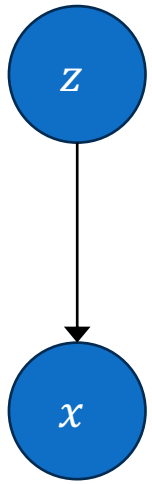
COMPUTING MACHINERY AND INTELLIGENCE (Turing, 1950)

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed.

NYU

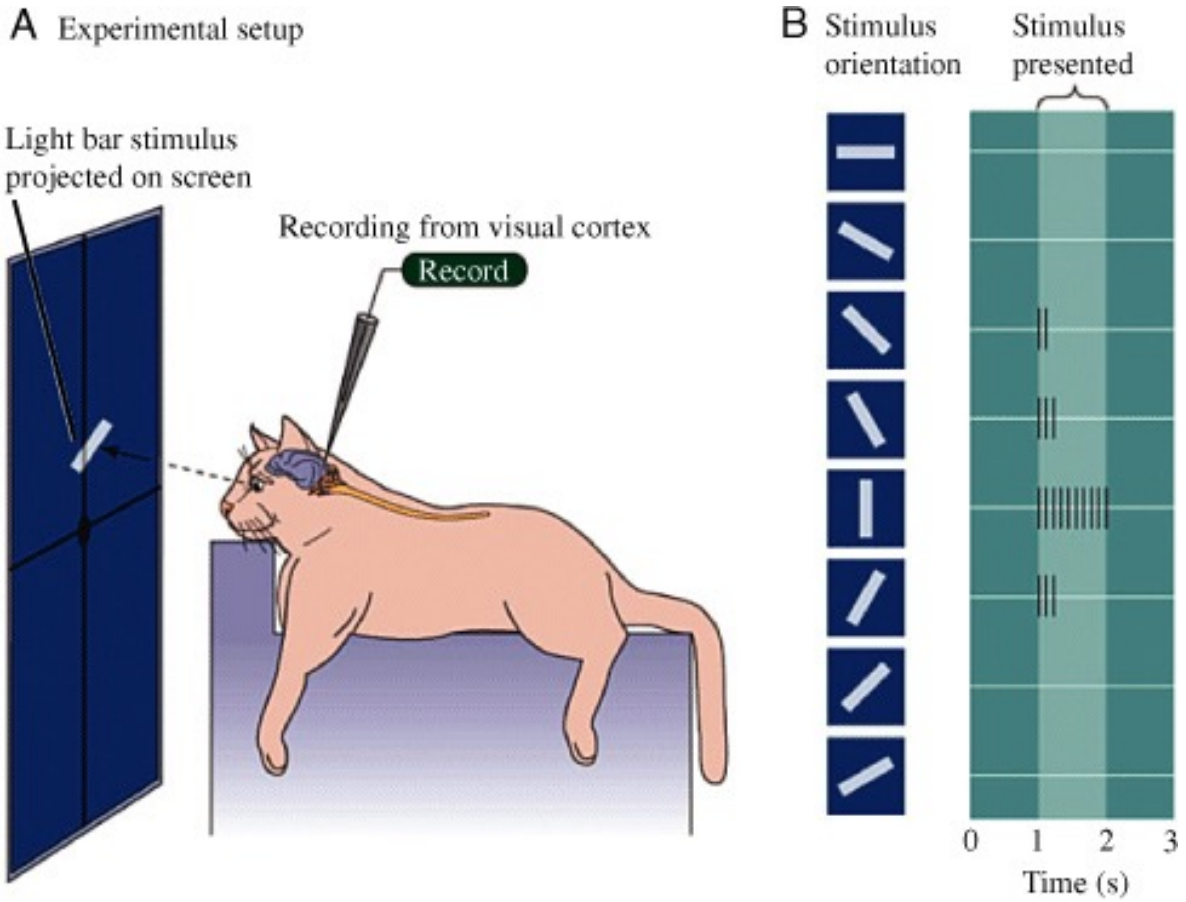
Why Do We Need Learning in Real-World Agents?

- Opinion 1: We always need learning in exploring new environments. There will always be something unknown. There will always be room for improvement. There won't be enough capacity to store all existing knowledge.
- Opinion 2: You can represent infinite variations with finite length description of an abstract symbolic system. We may not have seen all possible variations, but the underlying system remains the same.
- Opinion 3: While theoretically O2 might be true, empirically it is hard to realize. Given limited resource, you might be able to learn more abstract and invariant representations by compressing raw data. You can either be good at one thing without learning, or you need learning to be good at everything.

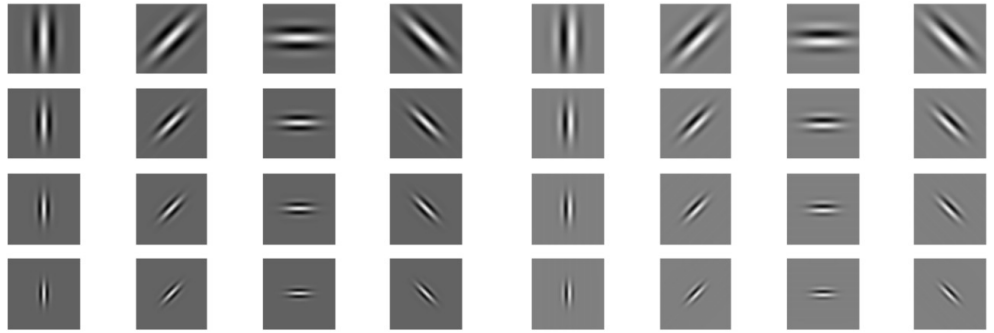


NYU

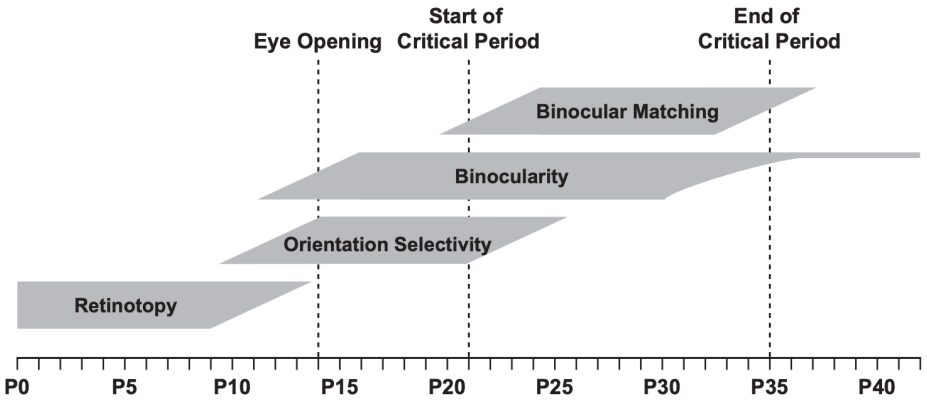
Hubel and Wiesel's Experiments



Wiesel, T.N., and Hubel, D.H. (1963)



Simple Cells, Gabor Filters



Espinosa and Stryker (2012)

NYU

Human Developmental Periods

Sensorimotor learning

Simple Reflexes (birth-1 month)

Infants use reflexes such as rooting, sucking, following moving objects with the eyes, and grasping objects. (For example: Infant closes their hand when a toys touches their palm.)

Primary Circular Reactions (1-4 months)

A primary circular reaction is when an infant tries to reproduce an event that happened by accident because they find it to be pleasurable. (For example: Intentionally mouthing a toy bunny.)

Secondary Circular Reactions (4-8 months)

Child becomes more focused on the world and begins to intentionally repeat an action in order to trigger an environmental response. (For example: purposefully picking up a pacifier to put it in their mouth.)

Coordination Of Secondary Circular Reactions (8-12 months)

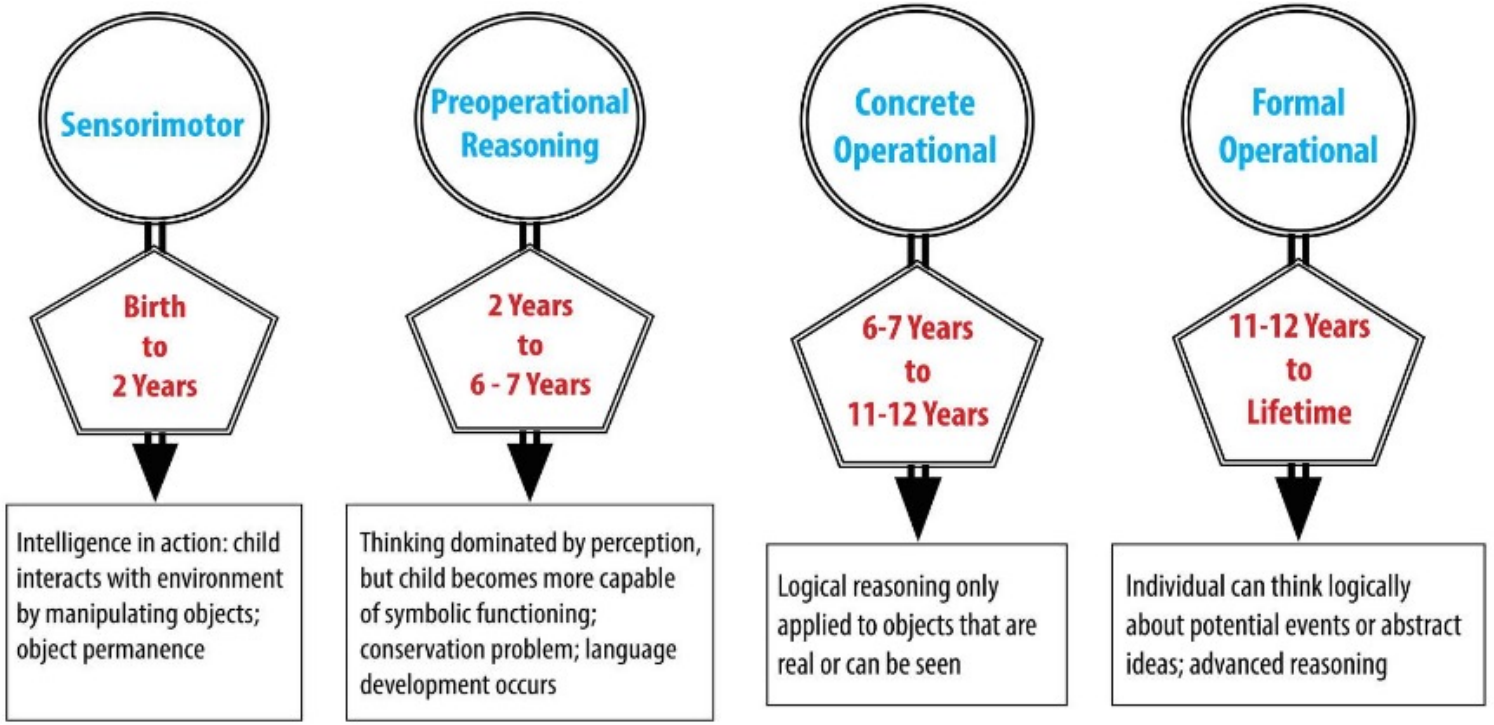
Child acts intentionally and follows steps to achieve goals. Child begin to do things intentionally and understands object permanence. (For example: Child will push one toy aside to get to a second toy partially concealed underneath.)

Tertiary Circular Reactions (12-18 months)

Child discovers new means to meet goals and begins to modify earlier behaviors to meet existing needs. Piaget described children in this stage as "young scientists". (For example: Child repeatedly drops/throws a set of plastic keys and observes how they move through space.)

Internalization of schemas (18-24 months)

Child begins to use symbols and form mental representations. The beginnings of insight and creativity are associated with this stage. (For example: Child pushes a chair across the kitchen and climbs up on it to reach a cookie on the counter.)

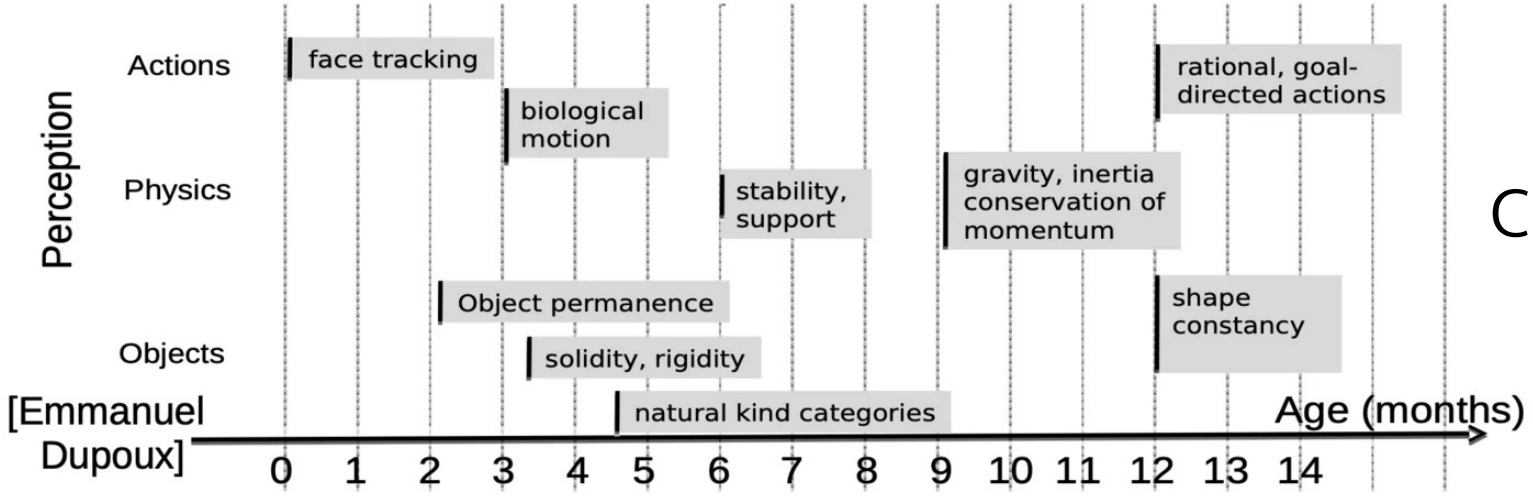


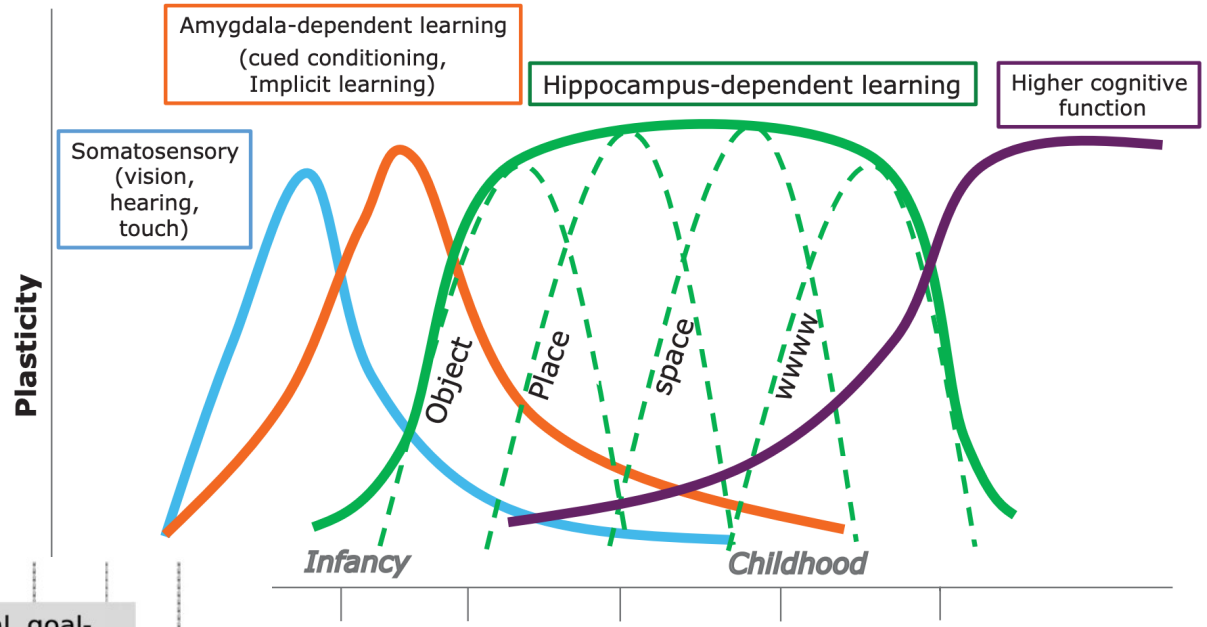
Piaget's Theory of Cognitive Development

NYU

Insights from the Brain

- Perception and motion
- Low-level to high-level representation
- Hippocampus and memory
- Abstraction





Critical Periods of Development

Logistics

NYU

Grading

- In-Class Participation (10%)
- Paper Review (15%)
- Paper Presentation (30%)
- Project (45%):
  - Project Proposal (10%).
  - Report (25%)
  - Presentation (10%)

Course Syllabus

NYU

CampusWire

0431

https://campuswire.com/p/G796EC30E

People who sign up using this link will get ‘student’ access to the class.

NYU

Introduce Yourself

- Send me an email about you.
  - Your knowledge background
  - Your research interest
  - What makes you excited to study this course
- (Optional) Share an introductory post on Campuswire.
  - Please update your profile picture.
  - This also helps you know each other and find a team partner.

NYU

In-Class Participation

- You get marks for asking good questions in Q&A periods of in-class presentations, guest lectures, etc.
- 10% worth of marks

Course Syllabus

NYU

Paper Reviews

- 15% of the total mark
- Select a paper from the suggested reading list (recent only), or find a recent paper of your interest (with approval)
- One topic each week
- W2 – W9

  - Week 2: Deep Learning for Structured Outputs
  - Week 3: 3D Vision and Mapping
  - Week 4: SSL and Object Discovery
  - Week 5: World Models and Forecasting
  - Week 6: End-to-End Planning
  - Week 7: Continual Learning
  - Week 8: Few-Shot Learning and Meta-Learning
  - Week 9: LLM Agents

NYU

Paper Reviews

Title: [Title of the paper]

Reviewer:
[Your Name]
[Net ID (e.g. ab1234)]

Summary:
What is the overall problem that the paper is solving? What approach does the paper take? How is the approach different from historical approaches? Why does it make sense for such an innovation? What does the result say?

Strengths:
Please describe the strengths of the paper, and explain why. Does the paper propose a new problem formulation? Does the paper take a new angle of looking at the problem, compared to the prior work? Are there ingenuity about the model and experiment design? Does the performance stand out? Make sure to include details.

Weaknesses:
Please describe the weaknesses of the paper, and explain why. Does the premise make sense? Is the methodology comprehensive? Are there any experiment comparisons that are missing to prove the point? What are the potential limitations of the approach? Be critical but back up your point.

Possible Future Extensions:
Please list 2-3 possible extensions based on this work. Explain why you think these extensions are viable and what are the potential risk factors.

Conclusion:
Give a concluding remark of the review. Summarize the contribution. If you encounter this paper in a paper review process, would you give a positive or negative score? Weigh in the strengths and weaknesses you wrote previously.

NYU

Topic Presentation

- Sign Up: Students have to sign up for a slot by Week 3 (Feb 6).
- Calendar: Week 7 – 13 (Tentatively).
- Approximately 3-4 students present on each topic.
- Each student will conduct a 30-minute presentation on 1-2 designated recent papers including necessary backgrounds.
- Panel Discussion: 30 minutes

NYU

Course Project

- 45% of the total mark
  - Project Proposal (10%)
  - Report (25%)
  - Presentation (10%)
- Project Consultation: 2 mandatory consultation by Feb 20 and Mar 20 with me and the TAs (one each).
- Week 14 + Week 15 Project Presentation.
- Week 14: 2% Bonus.

NYU

Project Key Dates

Key Dates (unless announced otherwise, every item is due before the lecture time at 4:55pm):

- Feb 6: Team registration of two students. Form link is here. If you cannot find a partner, you still need to submit the form.
- Feb 20: Complete one consultation meeting during the office hour with the instructor and/or the TAs.
- Feb 27: Course project proposal due. Submit the proposal on Gradescope. Each group only needs to submit one copy.
- Mar 20: Complete second consultation meeting during the office hour with the instructor and/or the TAs. You need to meet with the instructor and the TA at least once each. You are expected to bring preliminary results during the second meeting.
- Apr 10: Sign up for a presentation slot here.
- Apr 24/May 1: Presentation slides due the day before the presentation. Please submit your presentation slide deck here.
- May 2: Course final report due. Submit the report on Gradescope. Each group only needs to submit one copy.
- May 2: Finish mandatory peer evaluation survey. Form link is here.

NYU

Project Report Template

- Page limit: 8
- Including Table and Figures
- Does not include Appendix.
- What is enough?
- When you need to squeeze white space and cut down content.

DS-GA 3001 Course Project
Report Template and Instructions

Student Name 1
Affiliation
Address
email

Student Name 2
Affiliation
Address
email

Team [Team ID]

Abstract

This document outlines the instructions for the course project of DS-GA 3001. Your project proposal and final report should use this space for abstract.

1 Course Project

The course project constitutes 45% of your overall grade. The goal of the final project is to let students develop hands-on skills of implementing embodied learning systems for concrete real-world tasks such as toy games, long-form egocentric video understanding, self-driving simulation, indoor navigation, and robotic manipulation.

NYU

Direction 1: End-to-end self-supervised learning for perception and planning

- Existing end-to-end learning-based planning frameworks are mostly focused on supervised learning of labeled objects and human demonstrations.
- Demonstrations and rewards are forms of labels.
- How do we achieve label-efficient learning and exploration through self-supervision?
- Is planning and action necessary for a label-efficient algorithm for perception?
- Explore the full spectrum from end-to-end learning to modular designs.

NYU

Direction 2: Enhancing foundation models for spatial intelligence

- Foundation model are trained with discrete tokens and are less familiar with the 3D world to perform exact perception, inference and planning.
- Augment pretrained foundation models with the ability to perceive and plan under precision in embodied environments.
- How do we enhance robustness in real-world environments?
- Can be synthetic/realistic, 2D/3D environments.
- Can models with geometric designs beat generic foundation models in terms of learning efficiency?

NYU

Direction 3: Continual learning for embodied intelligence

- How do we apply continual learning algorithms to embodied tasks?
- Skill learning, open world learning
- Memory design, retrieval augmentation, continuous finetuning
- Incremental learning with experience/action abstraction
- Replay with physical constraints
- Actively choosing learning objectives

NYU

Other Directions?

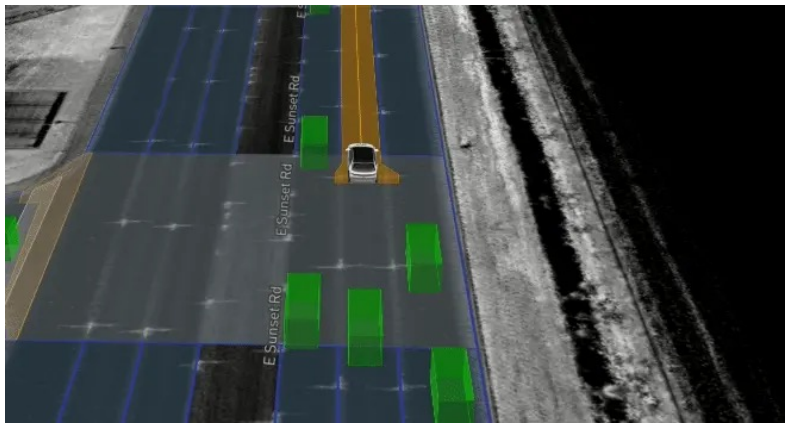
- You are allowed to form your own research ideas.
- Need to get my approval first. Talk to me early in the semester.

NYU

Embodied Environments

- You **must** demonstrate your project in an embodied environment.
- You can focus on one aspect of the algorithm. No need for a full stack.
- Your TAs will showcase demos on some exemplar environments.

Habitat indoor home



NuPlan self-driving

Ego-Exo4D Egocentric Videos

NYU

GenAI Policy

- AI may not be used in weekly paper reviews and paper presentations (except AI illustrations).
- AI may be used towards coding assistance and report writing assistance in the course project.
- The use of AI can still impact the grade if the report contains poor writings and non-factual statements.

NYU

Office Hours

- Myself: Thursday 1:00pm – 2:00pm Room 508, 60 5th Ave
- TAs:

Chris Hoang
Wed 2-3PM
Room 502

Ying Wang
Thu 2-3PM
Room 763

NYU

Content

- Introduction Brief History
- Deep Learning and Structured Outputs
- 3D Vision and Mapping
- Self-Supervised Representation Learning and Object Discovery
- World Models and Forecasting
- End-to-End Planning
- Continual Learning
- Few-Shot Learning
- LLM Agents

NYU

What's Next

- Today:
  - Introduction + Logistics
  - **Tutorial on HPC (Cloud Burst) by Ying**

- Next Week:
  - Deep Learning with Structured Outputs
  - **Tutorial on Learning with Simulators by Chris**

NYU