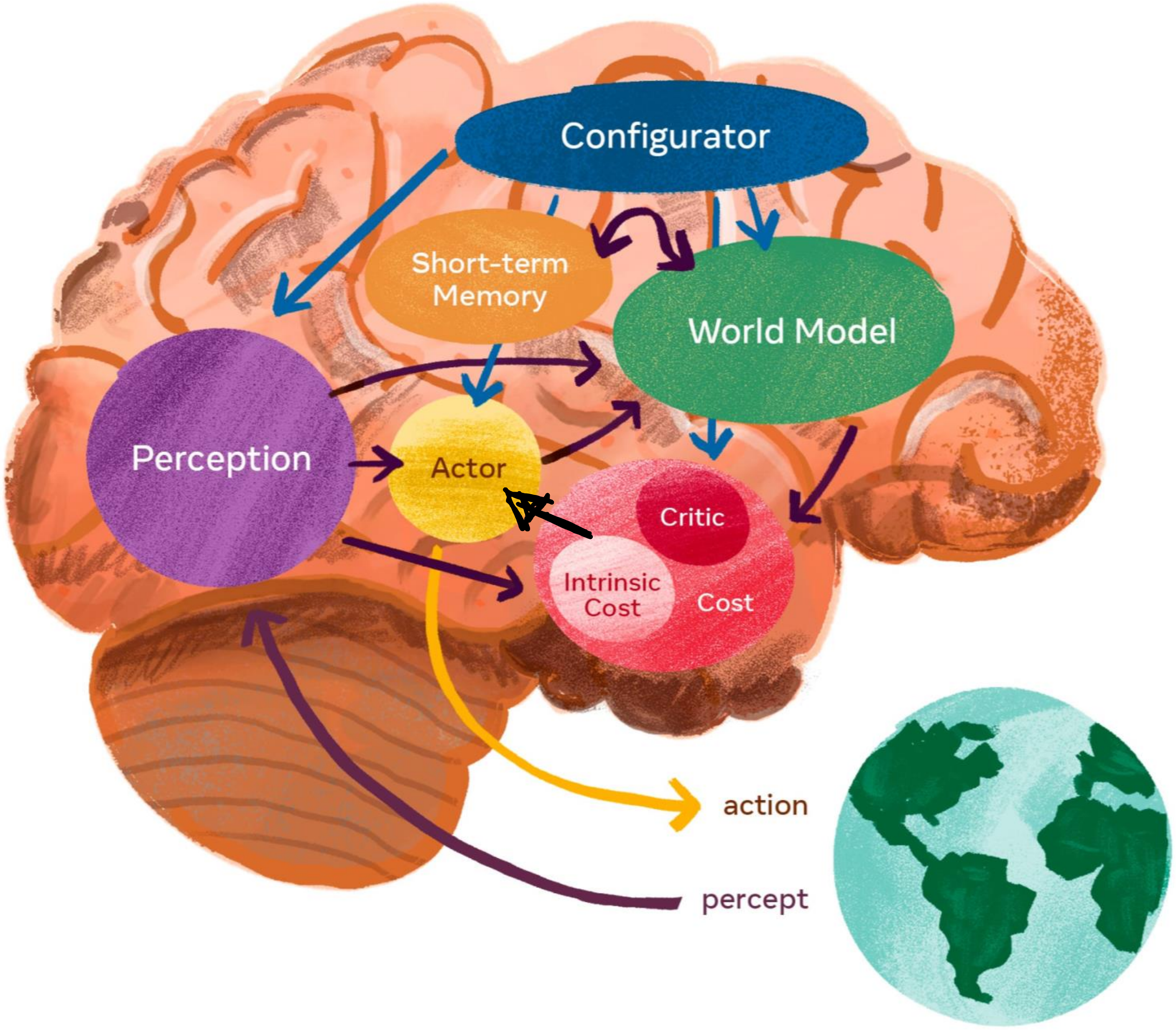
A Path Towards Autonomous Machine Intelligence

Yann LeCun



Motivation

- Deduction on how human brain operates influence research into artificial intelligence

- There are several unaddressed problems in the field of AI

Three major questions:

How can machines learn as efficiently as humans and animals?

How can machines learn to reason and plan?

How can machines learn multiple levels of abstraction at multiple timescales

1. How can machines learn as efficiently as humans and animals?

- Learning “world models”
    - + observation; - interaction
    - *common sense* as a world model of what is impossible
    - Single world model at the prefrontal cortex
        - Dynamically configurable

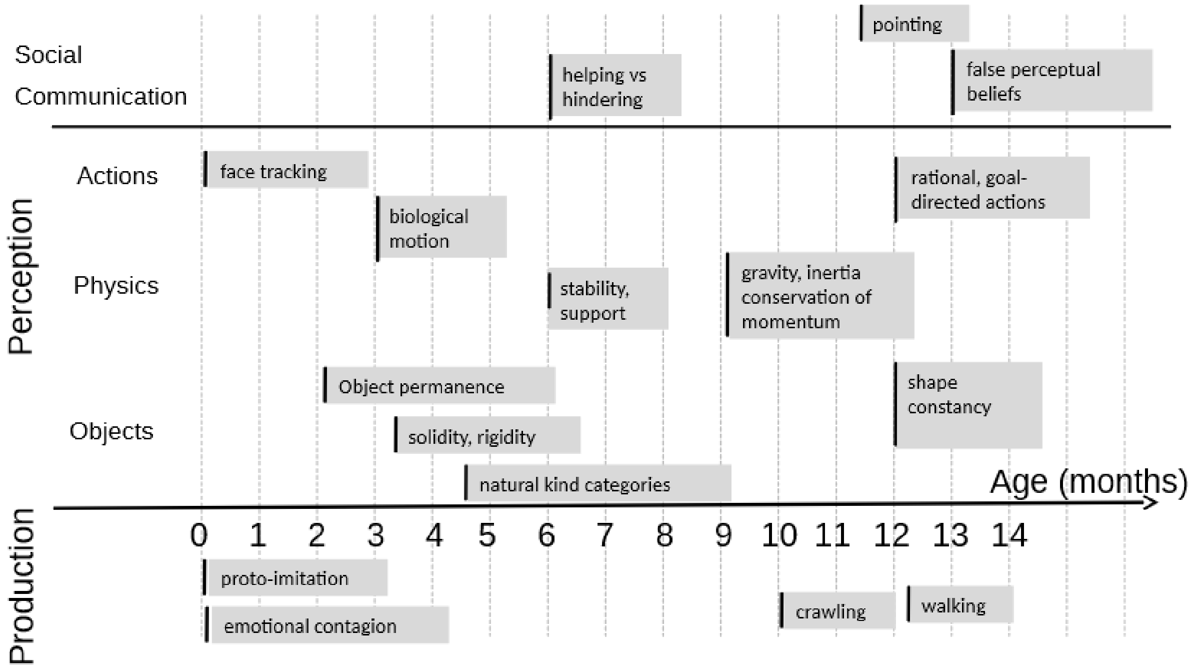
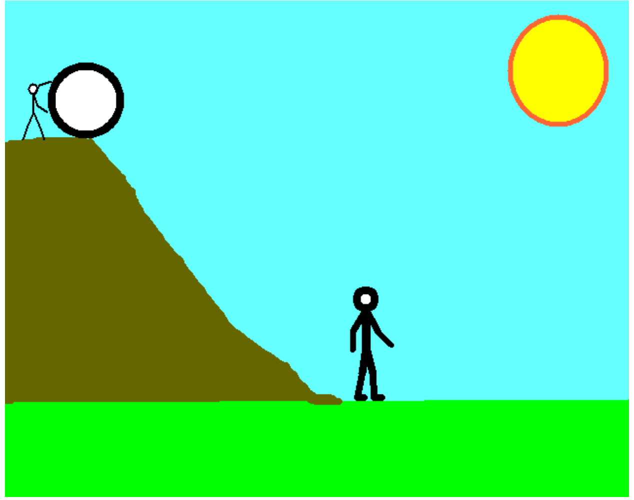


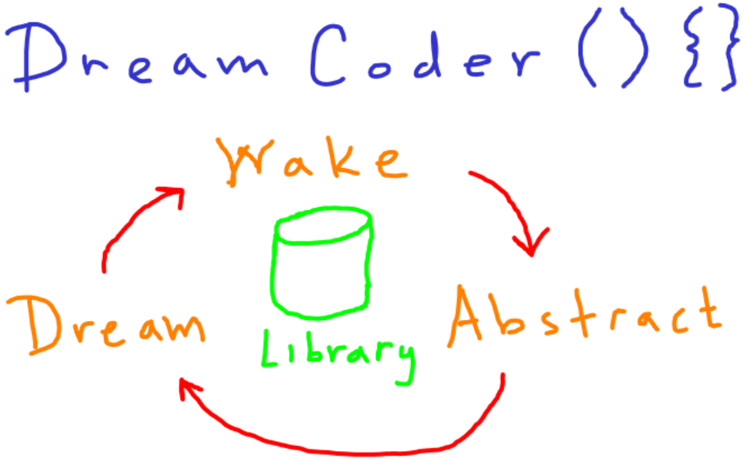
Figure 1: *This chart, (courtesy of Emmanuel Dupoux), indicates at what age infants generally acquire various concepts about how the world works. It is consistent with the idea that abstract concepts, such as the fact that objects are subject to gravity and inertia, are acquired on top of less abstract concepts, like object permanence and the assignment of objects to broad categories. Much of this knowledge is acquired mostly by observation, with very little direct intervention, particularly in the first few weeks and months.*

https://www.nature.com/articles/d41586-022-01921-7

World model?

$$W = \int_{k<\Lambda} \overbrace{[Dg][DA][D\psi][D\Phi]\exp}^{\text{quantum mechanics}} \left\{ i \int \overbrace{d^4x\sqrt{-g}}^{\text{spacetime}} \left[\overbrace{\frac{m_p^2}{2}R}^{\text{gravity}} \underbrace{-\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu}}_{\text{other forces}} + \underbrace{i\bar{\psi}^i\gamma^\mu D_\mu\psi^i + \left(\bar{\psi}^i_L V_{ij}\Phi\psi^j_R + \text{h.c.}\right)}_{\text{matter}} \underbrace{- |D_\mu\Phi|^2 - V(\Phi)}_{\text{Higgs}} \right] \right\}$$





2. How can machines learn to reason and plan?

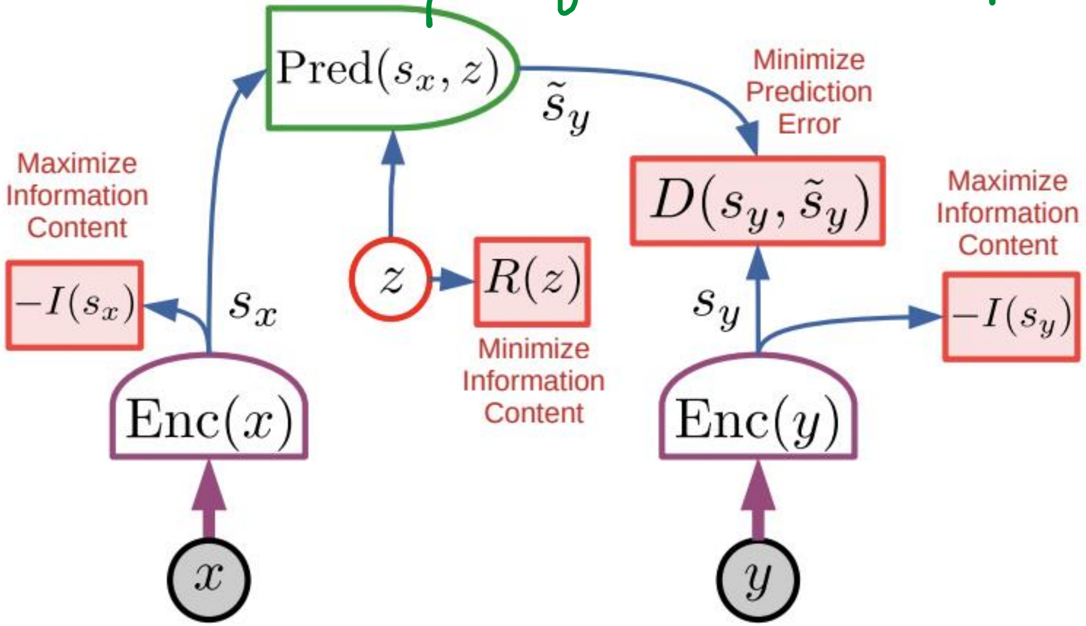
- Everything is differentiable
- Reasoning and planning as energy minimization
- Not feedforward computation

3. How can machines learn multiple levels of abstraction at multiple timescales?

- EBM world model: The Joint Embedding Predictive Architecture (JEPA)

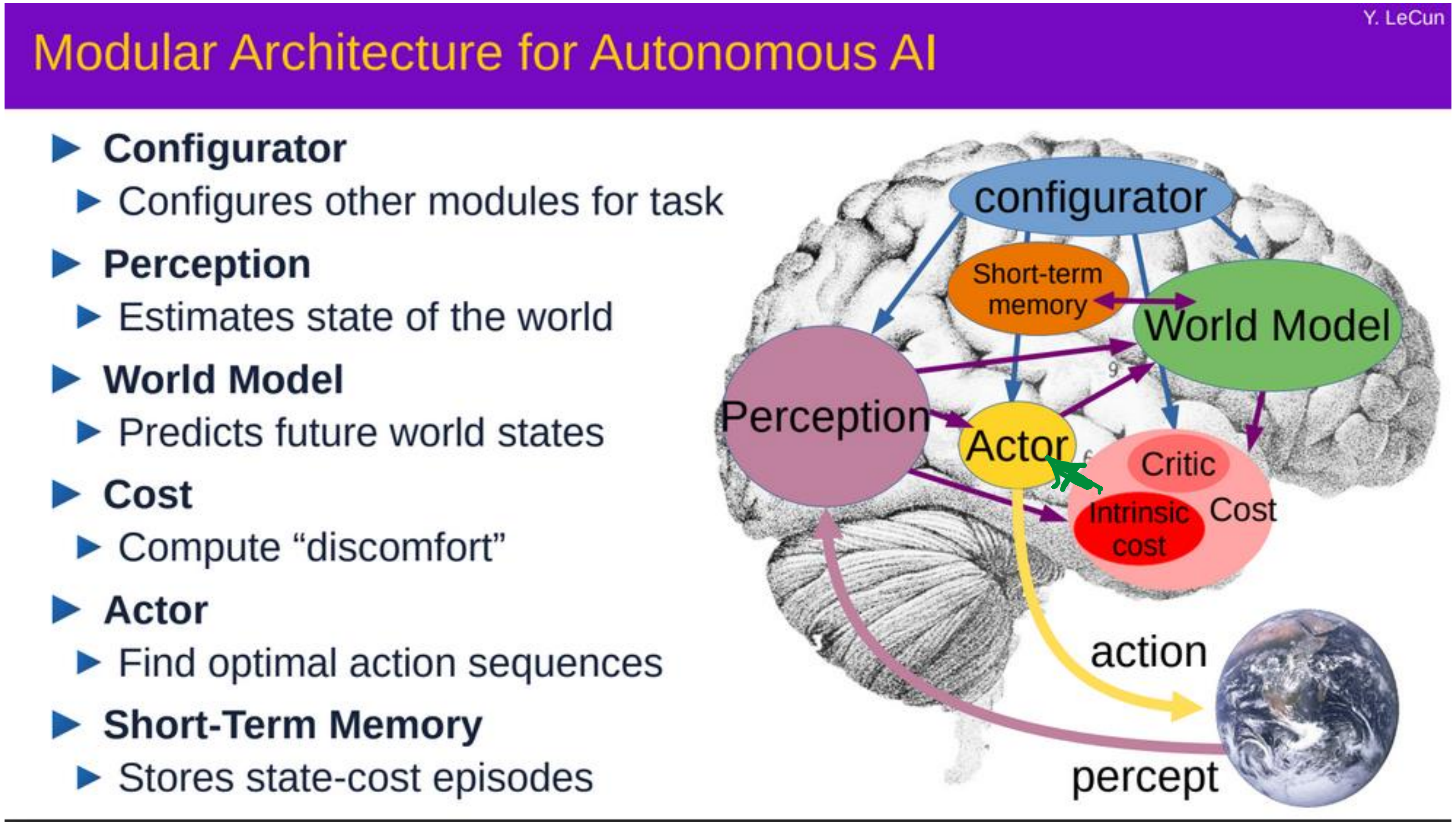
- Addresses:
  - Intractability of probabilistic models
  - Ignore details as opposed to generative models

global latent space ; "common sense"



Pred(s_x, z)
$\tilde{s}_y$
Minimize Prediction Error
$D(s_y, \tilde{s}_y)$
Maximize Information Content
$-I(s_x)$
s_x
z
$R(z)$
Minimize Information Content
s_y
Maximize Information Content
$-I(s_y)$
Enc(x)
Enc(y)
x
y

Architecture breakdown

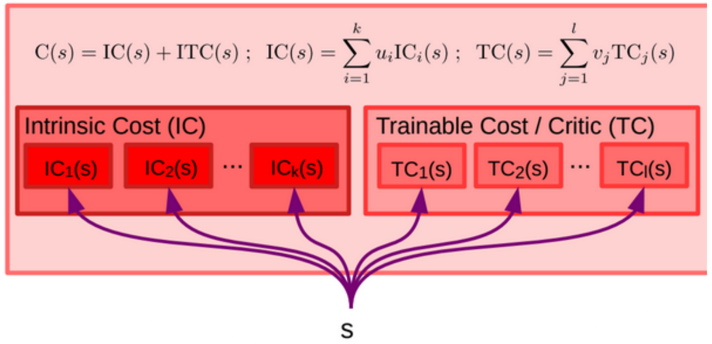


Y. LeCun

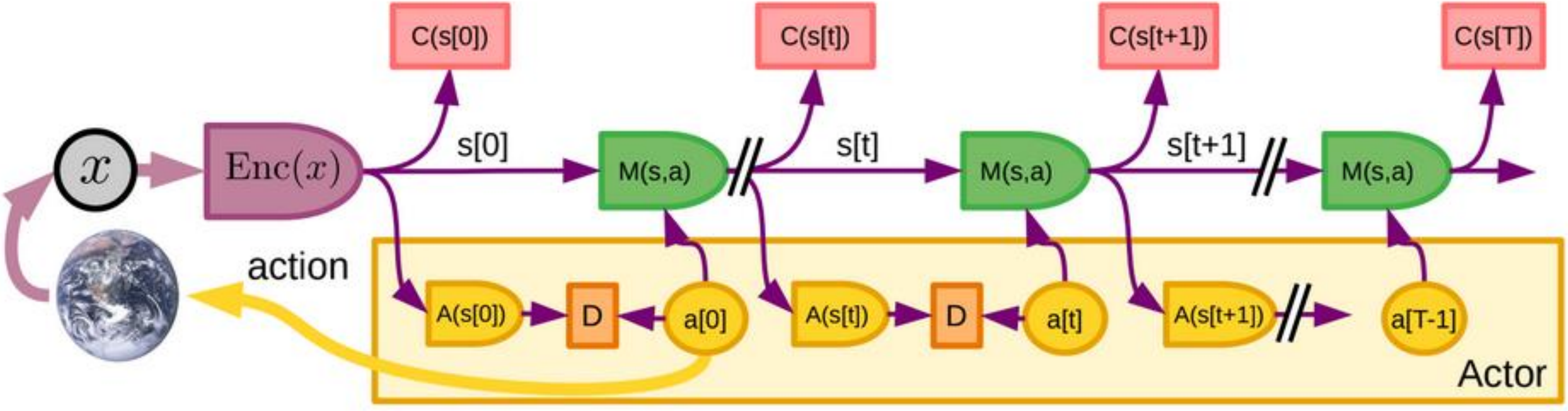
Modular Architecture for Autonomous AI

- **Configurator**
  - Configures other modules for task
- **Perception**
  - Estimates state of the world
- **World Model**
  - Predicts future world states
- **Cost**
  - Compute “discomfort”
- **Actor**
  - Find optimal action sequences
- **Short-Term Memory**
  - Stores state-cost episodes

1. Perception, planning, action

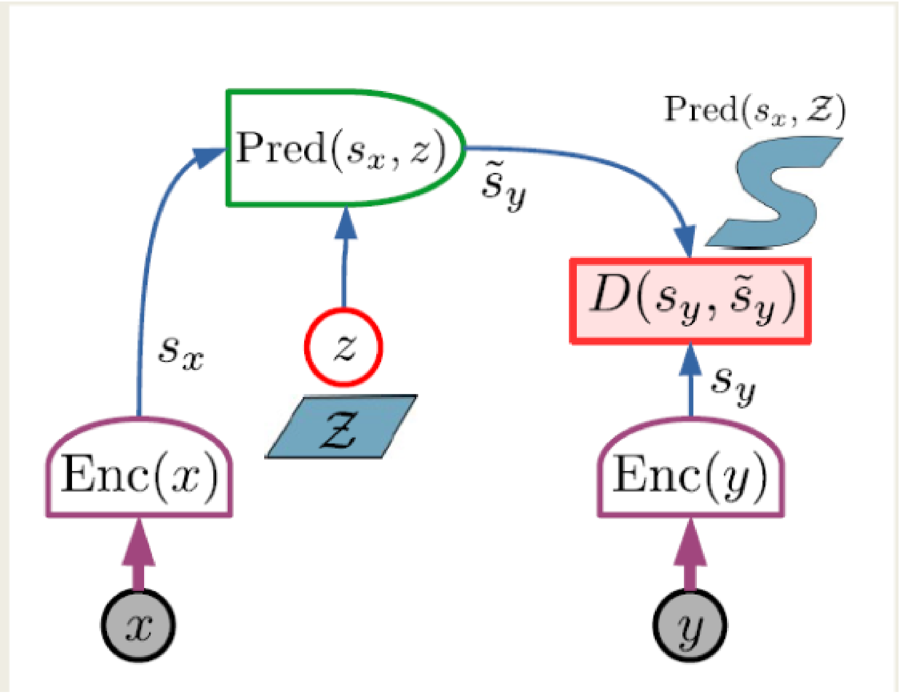


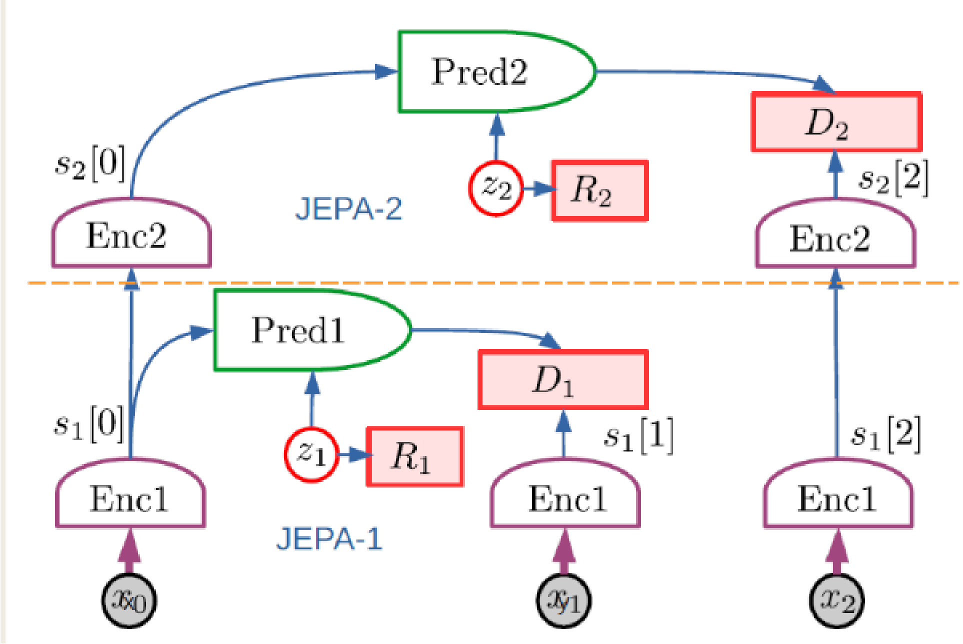
$$C(s) = IC(s) + ITC(s)\ ;\ \ IC(s) = \sum_{i=1}^{k} u_i IC_i(s)\ ;\ \ TC(s) = \sum_{j=1}^{l} v_j TC_j(s)$$



World model (zoomed-in)

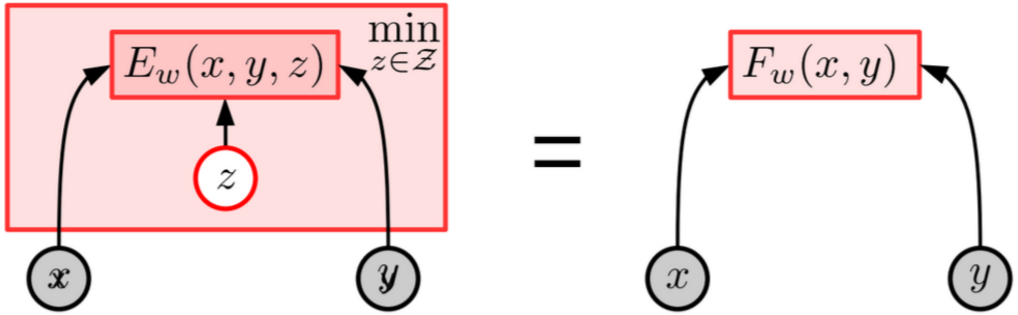
- Trained to predict future states of the world

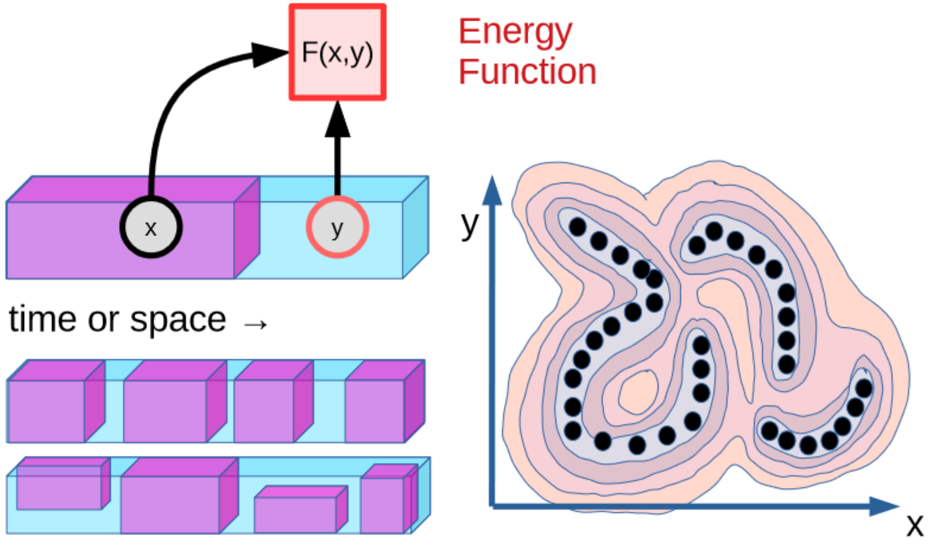




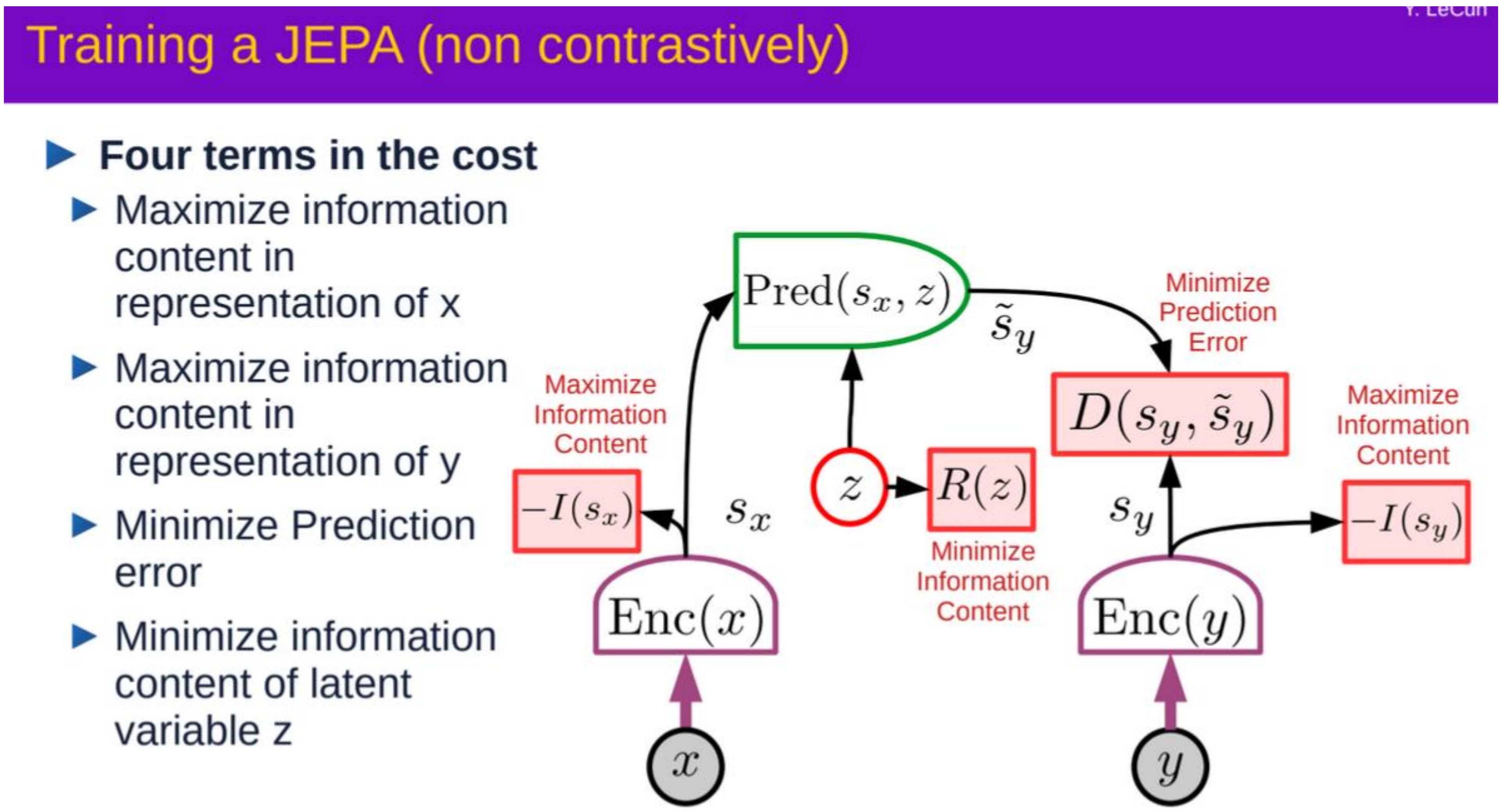
Training (energy-based)

$\check{z} = \mathrm{argmin}_{z \in \mathcal{Z}} E_w(x, y, z)$ $\qquad$ $F_w(x, y) = E_w(x, y, \check{z})$





Training



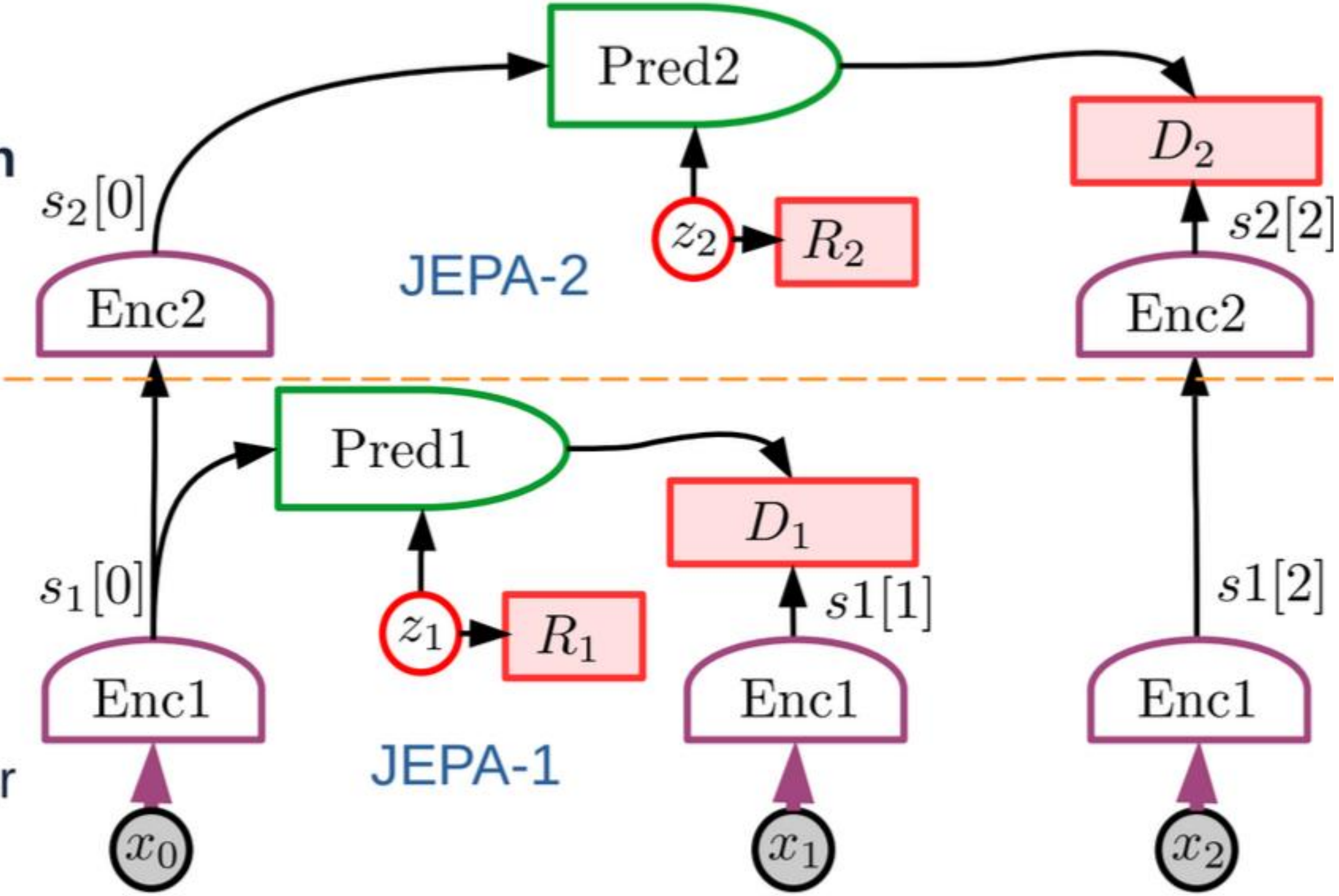
Y. LeCun

Training a JEPA (non contrastively)

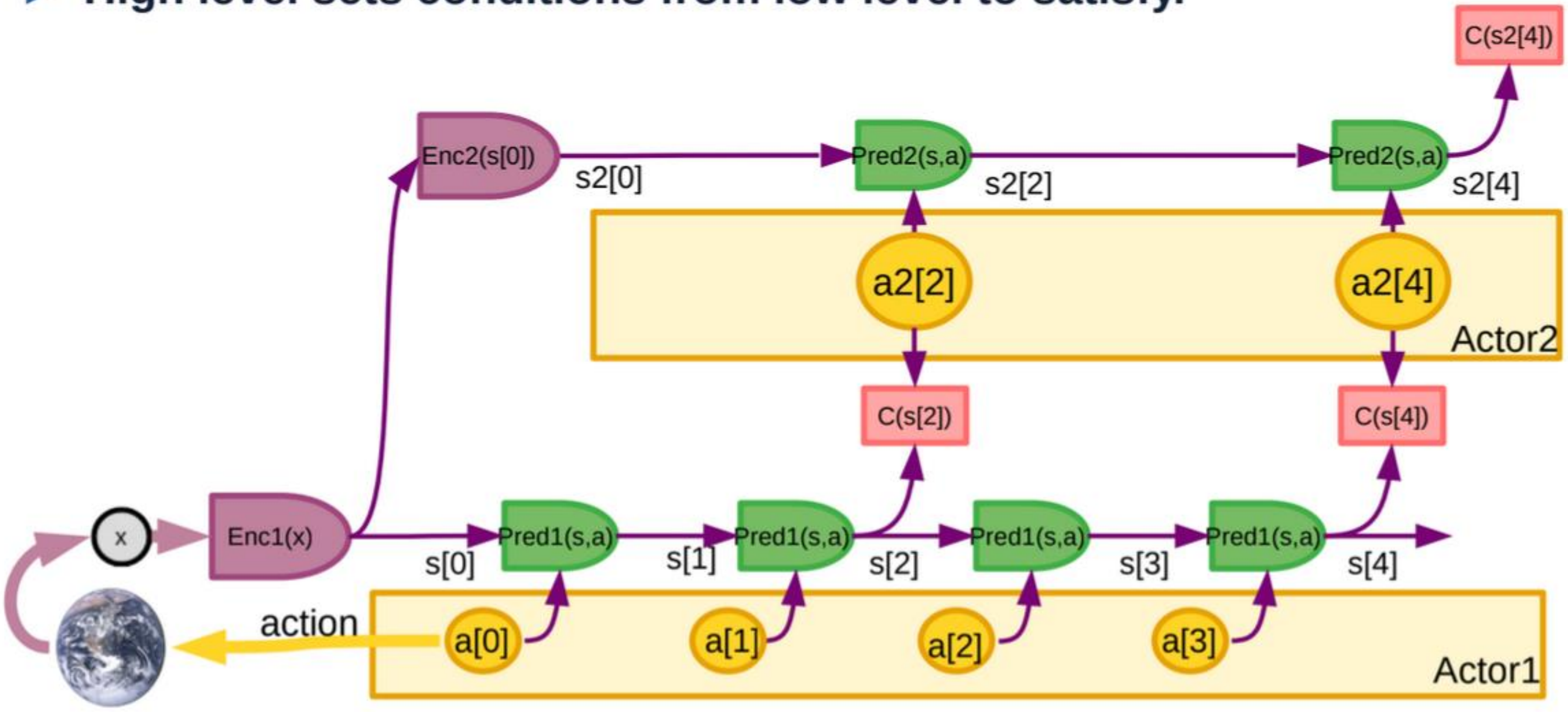
- **Four terms in the cost**
  - Maximize information content in representation of x
  - Maximize information content in representation of y
  - Minimize Prediction error
  - Minimize information content of latent variable z

Multi time-scale Predictions

- **Low-level representations can only predict in the short term.**
  - Too much details
  - Prediction is hard
- **Higher-level representations can predict in the longer term.**
  - Less details.
  - Prediction is easier

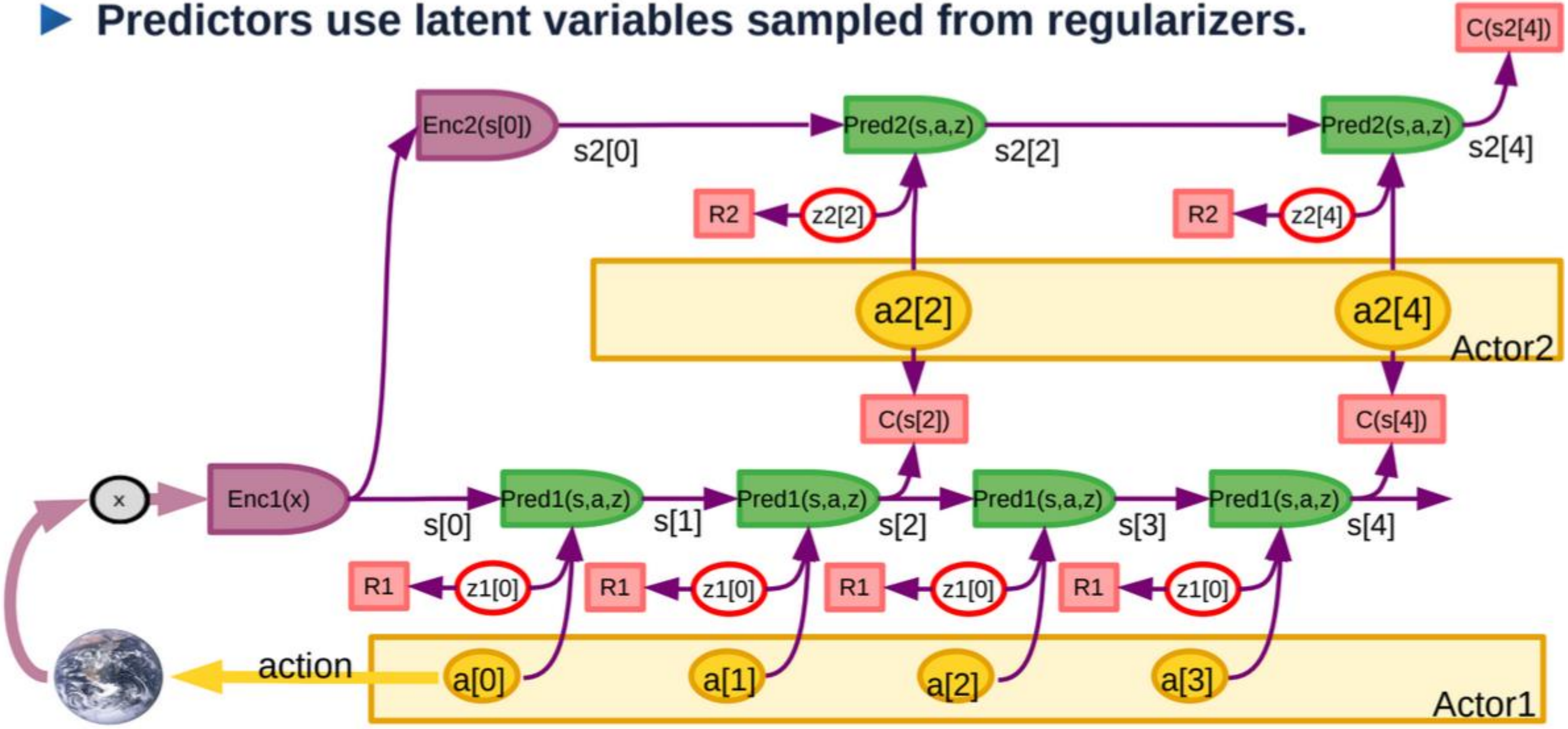


- High level sets conditions from low level to satisfy.



Hierarchical Planning with Uncertainty

- **Predictors use latent variables sampled from regularizers.**



Personal opinions

- Language as a means of thought hypothesis
- Embodied cognition
- Causal prediction ?= capture input dependencies through latent representation
- Self-reference
- One hierarchical model vs multiple models