Yann LeCun - A Path Towards Autonomous Machine Intelligence

Sabína Ságová

The text is written with as little jargon as possible, and using as little mathematical prior knowledge as possible, so as to appeal to readers with a wide variety of backgrounds including neuroscience, cognitive science, and philosophy, in addition to machine learning, robotics, and other fields of engineering. I hope that this piece will help contextualize some of the research in AI whose relevance is sometimes difficult to see.

ML Sucks! (Compared to Humans and Animals)

Supervised Learning (SL)

- Requires a large number of labeled samples

Reinforcement Learning (RL)

- Requires an **insane** amount of trials

SL/RL-Trained ML Systems

- **Specialized** and **brittle**
- Make **“stupid” mistakes**
- Do not **reason** or **plan**

Animals and Humans

- Can learn new tasks **very quickly**
- **Understand** how the world works
- Can **reason** and **plan**
- Have **common sense** (which machines do not)

ML Sucks! (plain ML/DL, at least)

ML Systems (Most of Them)

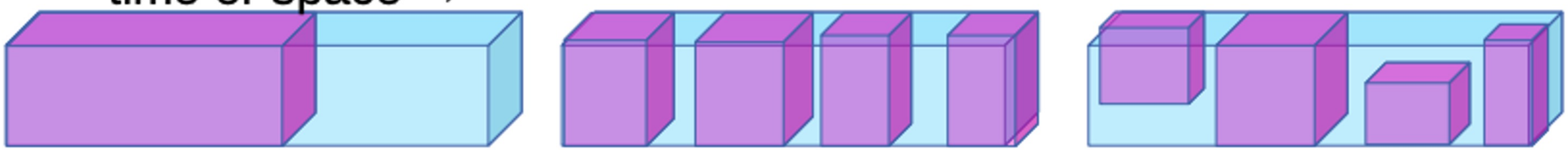
- Have a **constant** number of computational steps between input and output (Auto-regressive LLMs - fixed amount of computation to compute every token → limits the reasoning ability of these systems)
- **Do not reason**
- **Cannot plan** (Auto-regressive = produce things one after another)

Humans and Some Animals

- **Understand** how the world works
- Can **predict** the consequences of their actions
- Can perform **chains of reasoning** with an unlimited number of steps
- Can **plan complex tasks** by decomposing them into sequences of subtasks

SSL = Learning to Fill in the Blanks

- **SSL** has taken over the world for understanding and generation of: **Images, Audio, Text**
  - **Reconstruct the input or Predict missing parts of the input.**



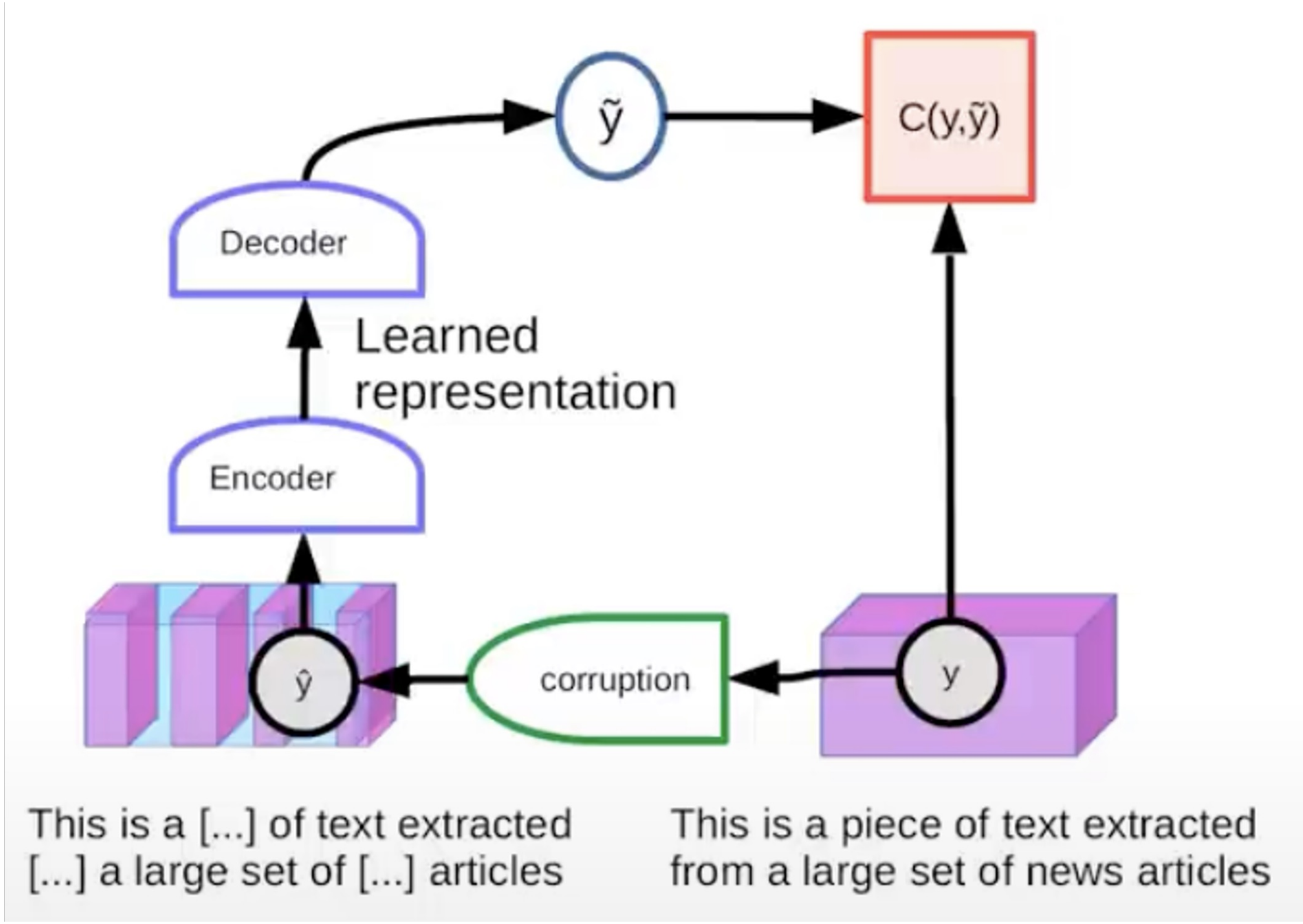
This is a [...] of text extracted [...] a large set of [...] articles

Denoising Auto-Encoders

Multilingual

those systems find some sort of
internal representation
that is **language independent**
→ content moderation

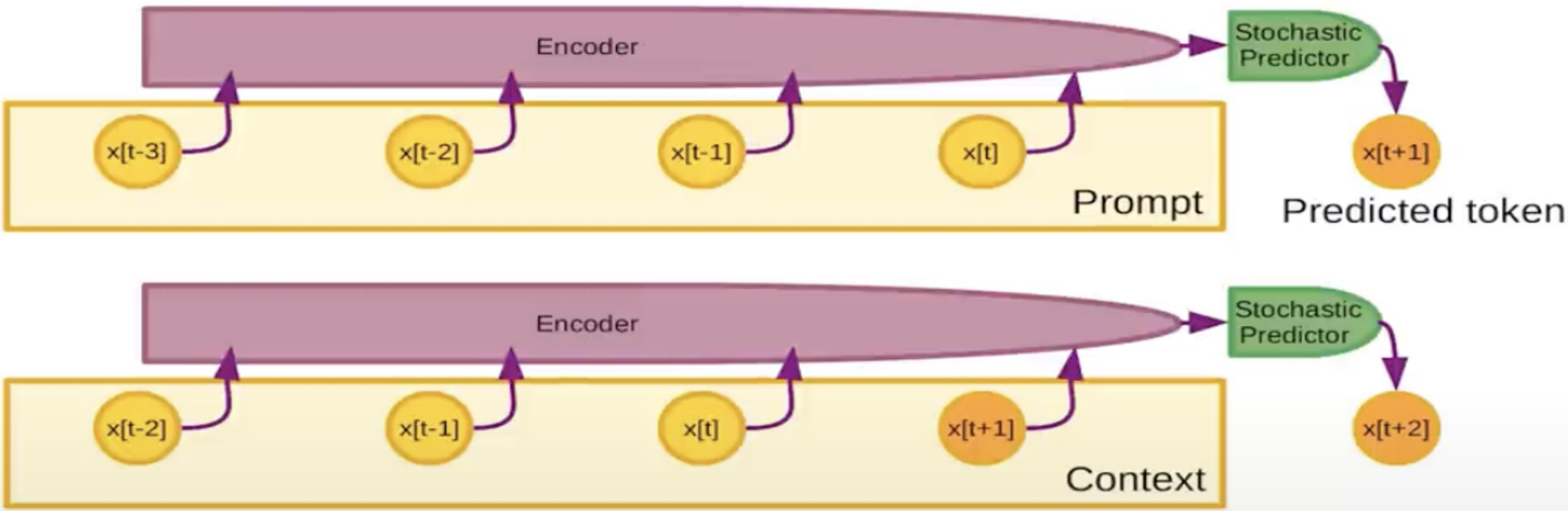
(hate speech)



Auto-Regressive Generative Architectures

Outputs one **“token”** at a time. Tokens can represent: **Words, image patches, Speech segments**

just predict the last word in a long sequence of a few thousand words taken from a corpus



Auto-Regressive LLMs

- Outputs one **text token** after another
    - Tokens may represent **words** or **subwords**
- **Encoder/predictor** is a **transformer architecture**
- **Billions of parameters**: typically from **1B to 500B**
- **Training data**: **1 to 2 trillion tokens**
- **can produce texts that kind of make sense**

LLMs for Dialog/Text Generation → scaling them up, having access to more data (ethics)

- **BlenderBot**, **Galactica**, **LLaMA (FAIR)**, **Alpaca (Stanford), LaMDA/Bard (Google)**, **Chinchilla (DeepMind)**, **ChatGPT (OpenAI)**

Performance

- **Amazing (code generation)**, but… They make **stupid mistakes (no mental model)**:
        - **Factual errors**, **logical errors, inconsistencies**, **limited reasoning**, **toxicity, hallucinate**

Limitations

- **No knowledge** of the underlying reality
- No **common sense**, cannot **plan** answers

What are Auto-Regressive LLMs Good For?

Auto-Regressive LLMs: Good For

- Writing assistance, first draft generation, stylistic polishing
- Code writing assistance

Auto-Regressive LLMs: Not Good For

- Producing factual and consistent answers (hallucinations!)
- Taking into account recent information (anterior to the last training)
- Behaving properly (they mimic behaviors from the training set)
- Reasoning, planning, math
- Using “tools”, such as search engines, calculators, database queries…

Important Note

- We are easily fooled by their fluency.
- But they do not know how the world works.

Unpopular Opinion about Auto-Regressive LLMs

Auto-Regressive LLMs Are Doomed

- They **cannot be made factual**, **non-toxic**, etc.
- They are **not controllable**

Key Problem

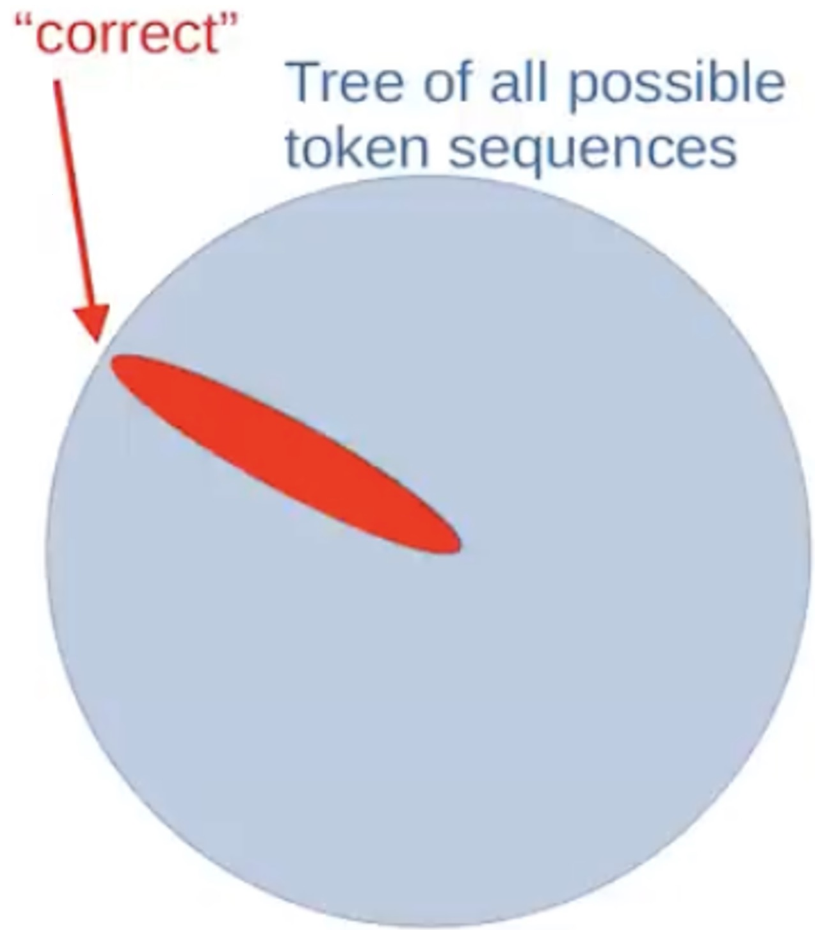
- Probability **e** that any produced token takes us

  outside of the set of correct answers

- Probability that an answer of length **n** is correct:
  - **P(correct) = (1 - e)^n**
  - This diverges **exponentially**

Conclusion

- **It is not fixable** (without a major redesign)



Auto-Regressive Generative Models Suck!

AR-LLMs

- Have a **constant number** of computational steps between input and output for each token
- **Weak representational power**
- **Do not really reason**
- **Do not really plan**

Humans and Many Animals

- **Understand** how the world works
- Can **predict** the consequences of their actions
- Can perform **chains of reasoning** with an **unlimited number of steps**
- Can **plan complex tasks** by decomposing them into sequences of subtasks

How Do Humans and Animals Learn So Quickly?

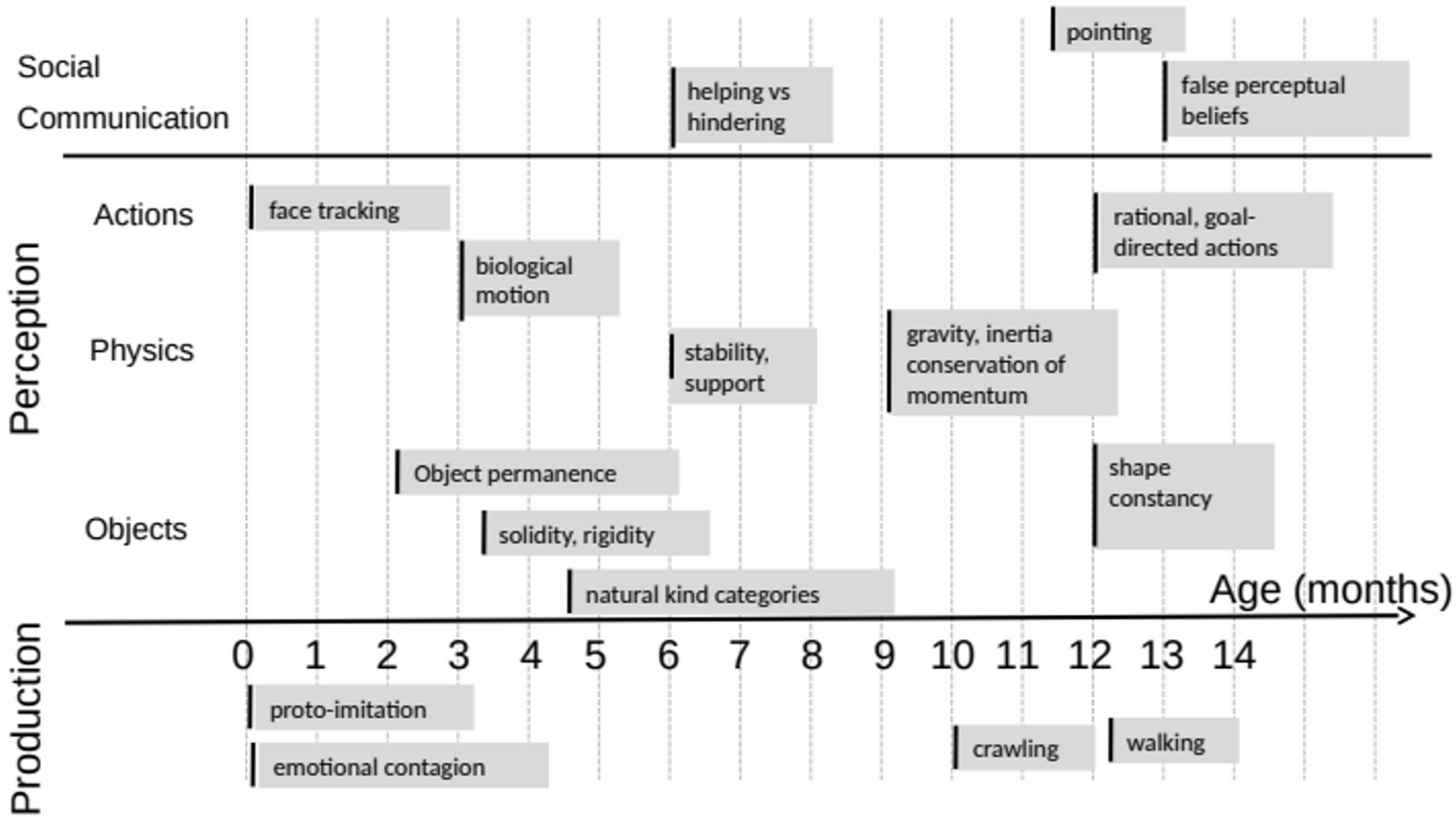
- **Not supervised, Not reinforced, At least, not much**
- **observation, interaction**

How Could Machines Learn Like Humans and Animals?

How Can Babies Learn How the World Works?

How Can Teenagers Learn to Drive with Just 20 Hours of Practice?

paradox



How Do Human and Animal Babies Learn?

How Do They Learn How the World Works?

- **Largely by observation**, with remarkably little interaction (initially)
- They accumulate **enormous amounts of background knowledge**
    - About the **structure of the world**, like **intuitive physics**
- Perhaps **common sense** emerges from this knowledge?

Three challenges for AI & ML

Learning Representations and Predictive Models of the World

- **Supervised and RL**: Require too many samples/trials
- **SSL / Learning Dependencies**:
  - Learning to **fill in the blanks**
  - Learning to represent the world in a **non-task-specific** way
  - Learning **predictive models** for planning and control

Learning to Reason

- **Beyond feed-forward**
- Making reasoning **compatible with learning**
  - Reasoning and planning as **energy minimization**

Learning to Plan Complex Action Sequences

- Learning **hierarchical representations** of action plans

Modular Architecture for Autonomous AI

Configurator

- Configures other modules for the task

Perception

- Estimates the state of the world

World Model

- Predicts future world states

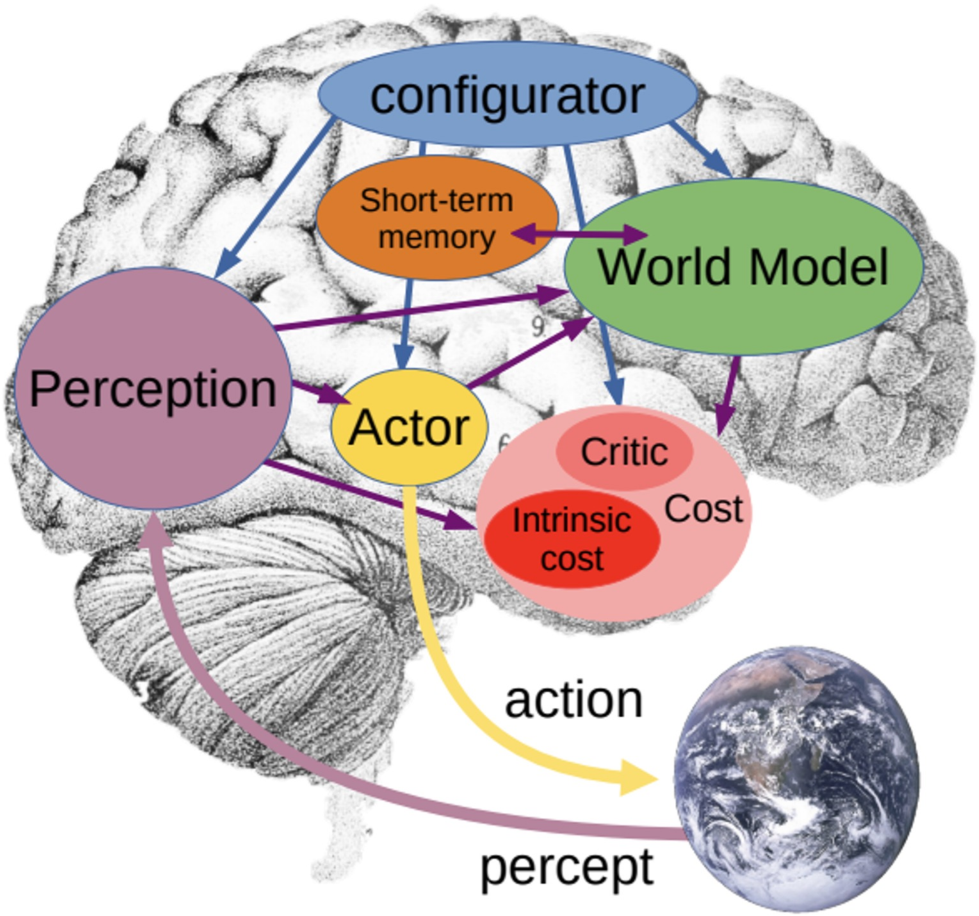
Cost

- Computes "discomfort"

Actor

- Finds optimal action sequences

Short-Term Memory



Mode-1 Perception Action Cycle

Perception Module

- **s[0] = Enc(x)**
  - Extracts representation of the world

Policy Module

- **A(s[0])**
  - Computes an action **reactively**

Cost Module

- **C(s[0])**
  - Computes the **cost** of the state

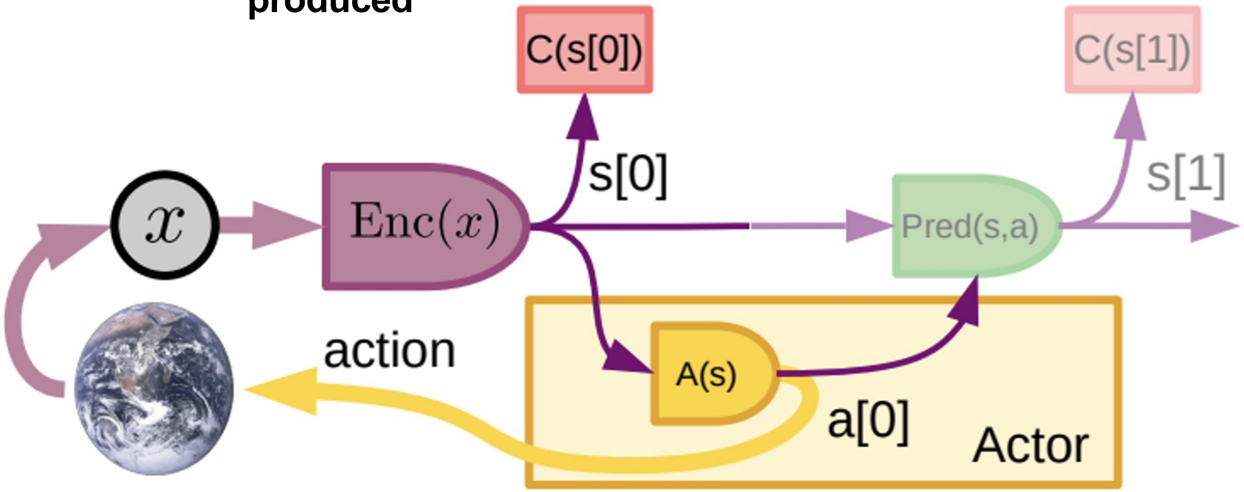
Optionally:

- **World Model**
  - **Pred(s, a)**: Predicts **future state**
  - Stores states and costs in **short-term memory**

Execution

- **perceive the world → extract internal representation of state → run through NN to produce and action**

World = windows of previous worlds that have been produced



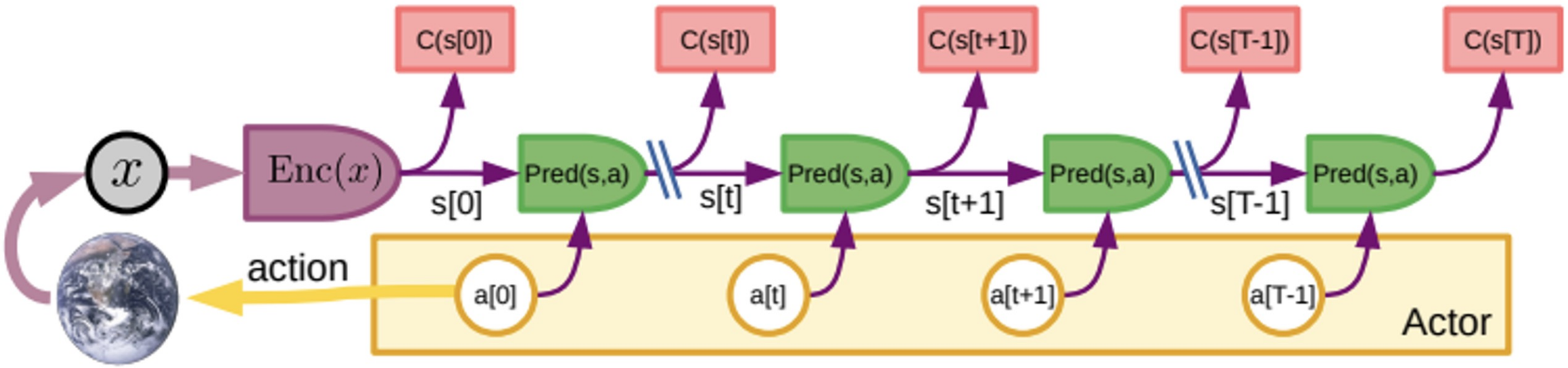
Mode-2 Perception-Planning-Action Cycle

Akin to Classical Model-Predictive Control (MPC)

- **Actor** proposes an **action sequence, world model** predicts the **outcome, actor** optimizes the action sequence to minimize **cost** e.g., using **gradient descent**, **dynamic programming**, **MC tree search**, etc.
- this is not auto-regressive, can correct hallucinations, toxicity by designing cost functions in appropriate ways

Execution

- **perceive the world → run encoder (estimate state) → run world model (predictor = from state t the action you mi**

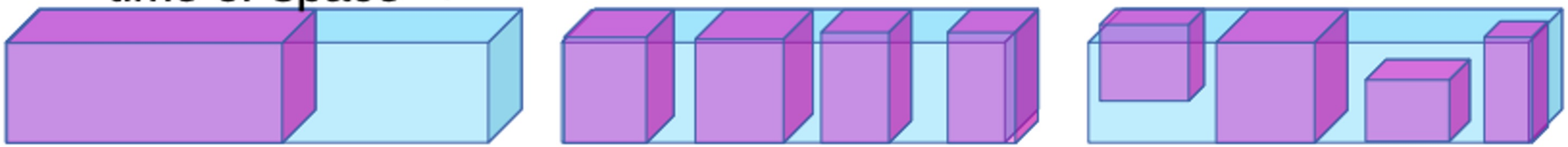


the cat catches the ______

Building & Training the World Model

- **Reconstruct the input or Predict missing parts of the input.**

time or space →



This is a [...] of text extracted [...] a large set of [...] articles

SSL works really well for text (a probability distribution), for video we do not have a proper way to represent distribution over all video clips

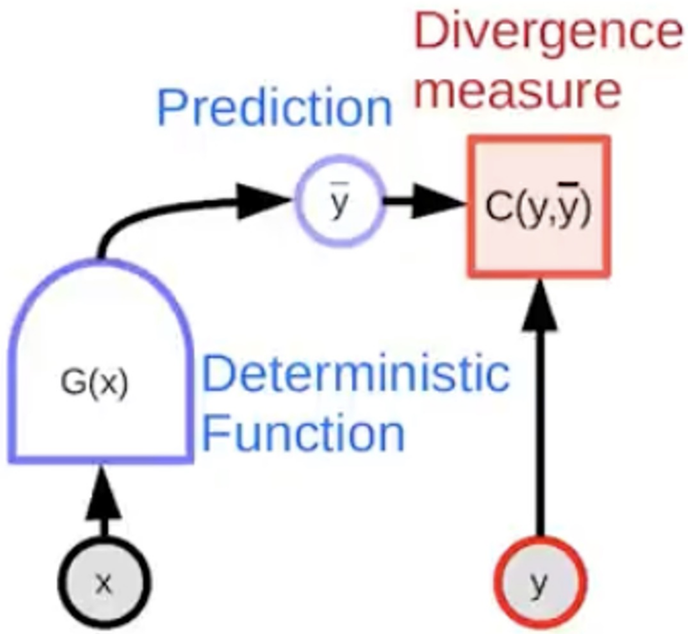
The World is stochastic

Training a System to Make a Single Prediction

- It tends to predict the **average** of all **plausible predictions**

Result

- **Blurry predictions! → no SSL trained from video, we do not know how to deal with that problem**



How Do We Represent Uncertainty in the Predictions?

The World is Only Partially Observable

- How can a predictive model represent **multiple predictions**?
- **Probabilistic models** are intractable in **continuous domains**
- **Generative models** must predict **every detail** of the world

Solution

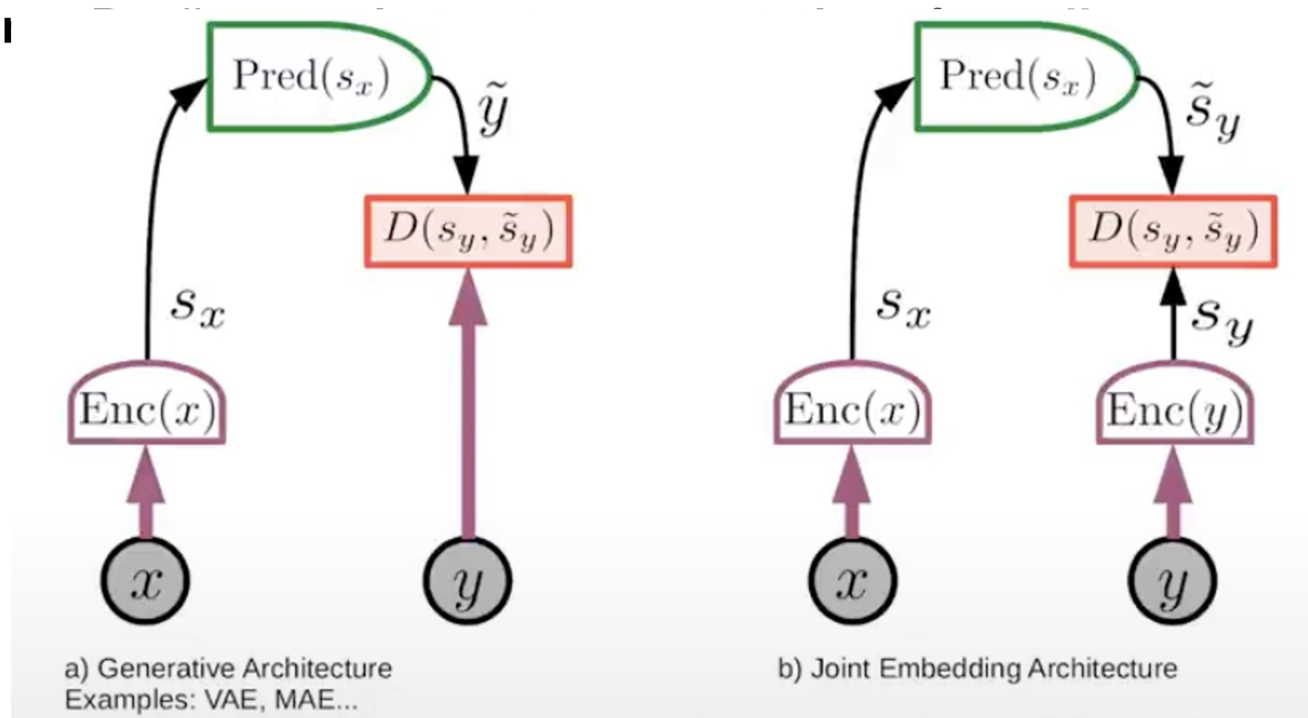
- **Joint Embedding Predictive Architecture**

Architectures: Generative vs Joint Embedding

Generative: Predicts **y** with all the details, includes even **irrelevant information.**

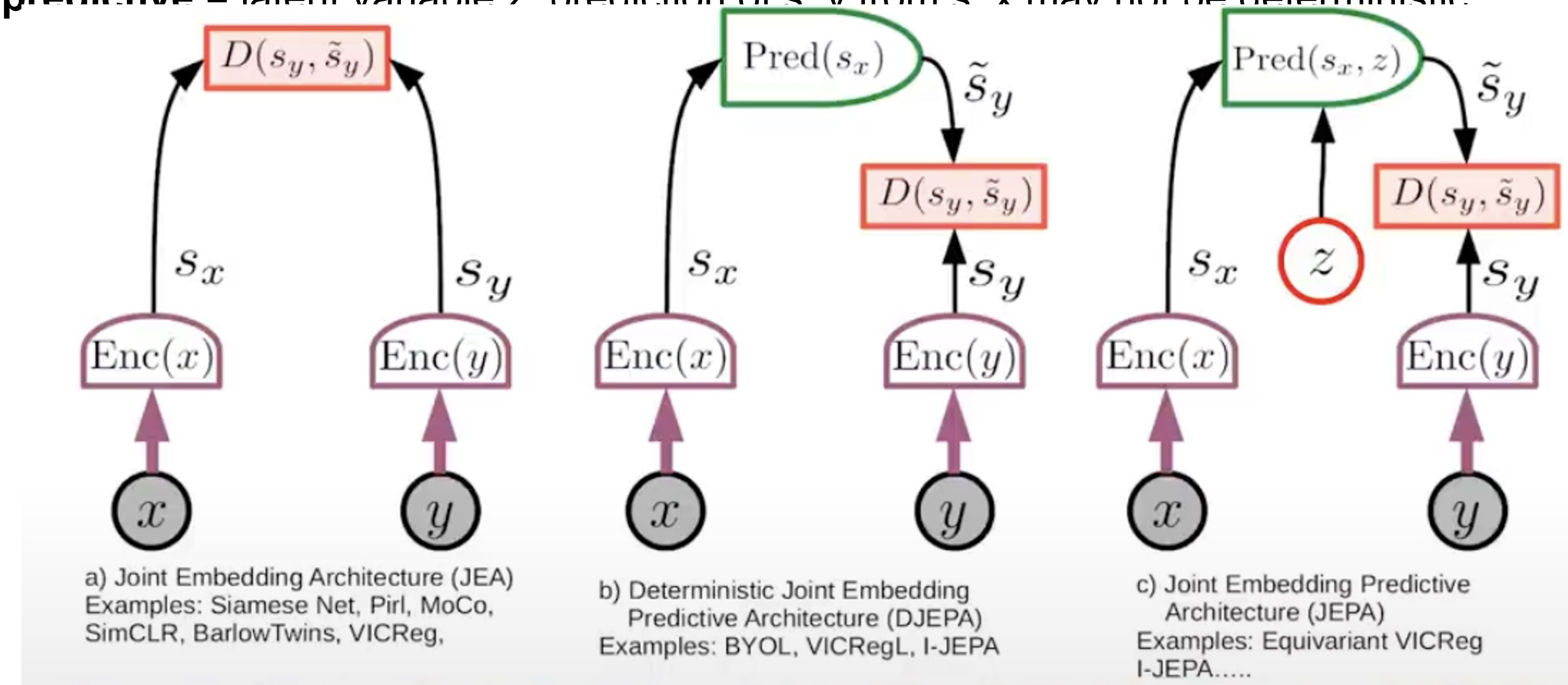
Run x through encoder → run representation to predictor → measure reconstruction error

Joint Embeddi



Joint Embedding Architectures

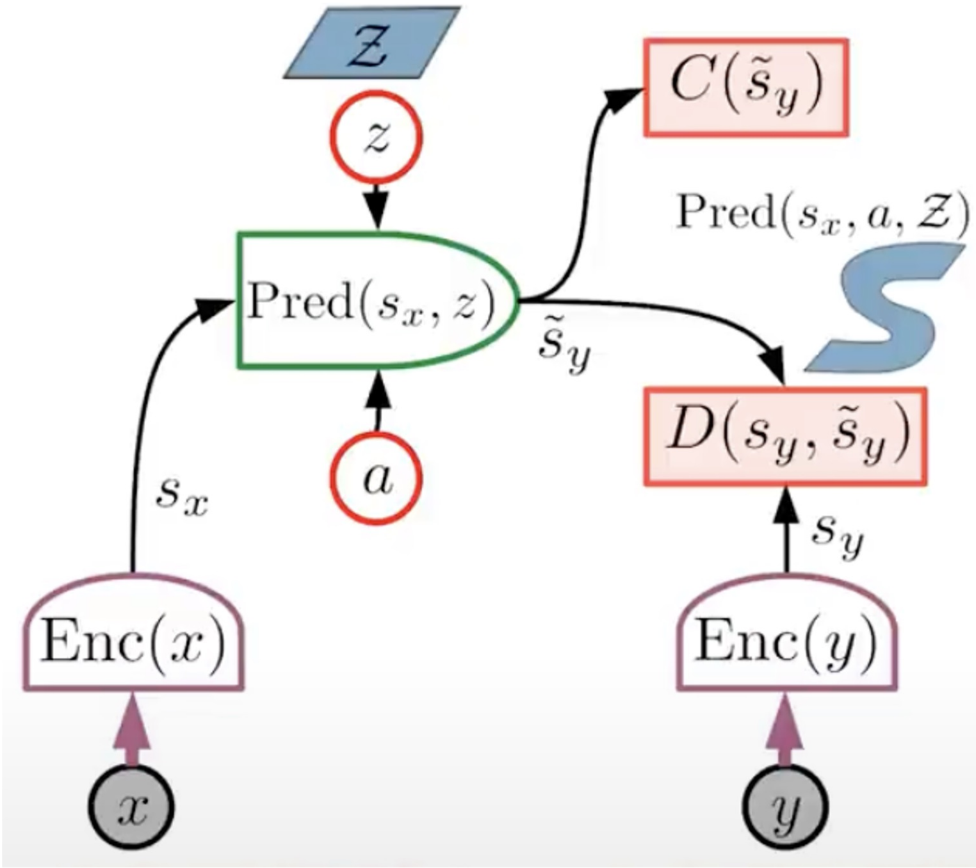
- Computes **abstract representations** for **X** and **Y**
- Tries to make them **equal** or **predictable** from each other
- **predictive** = latent variable z, prediction of s_y from s_x may not be deterministic



Architecture for the World Model: JEPA

JEPA: Joint Embedding Predictive Architecture

- **x**: observed past and present
- **y**: future
- **a**: action
- **z**: latent variable (unknown)
- **D()**: prediction cost
- **C()**: surrogate cost



Core Idea

- JEPA predicts a representation of the future (**S_y**)

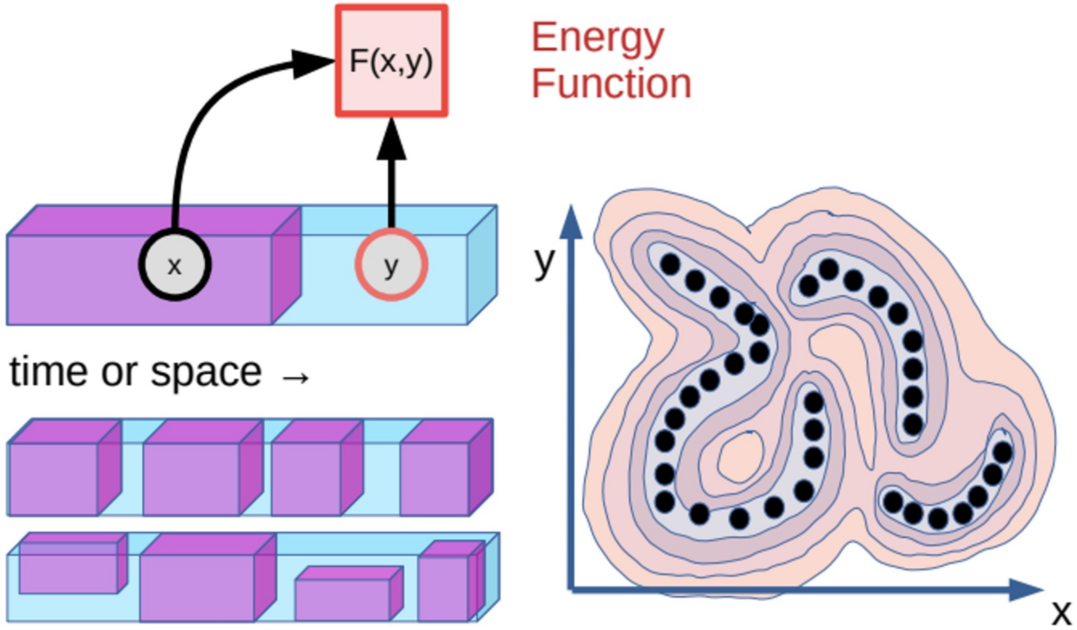
  From a representation of the past and present (**S_x**)

Energy-Based Models: Implicit function

The only way to **formalize and understand** all model types (abandon probability theory)

Assign **low energy** to compatible pairs of **X** and **Y**

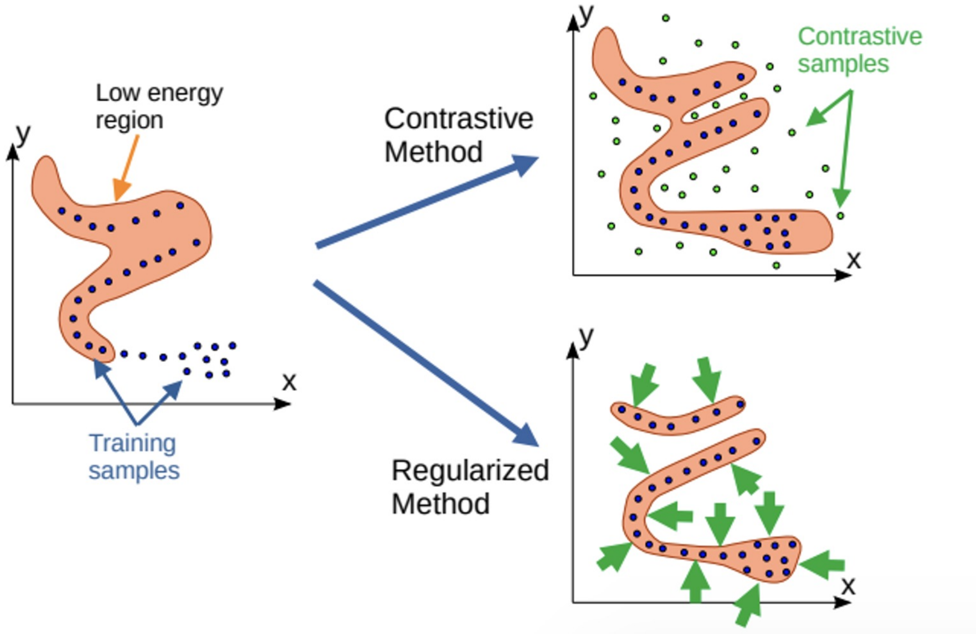
Assign **higher energy** to **incompatible** pairs



EBM Training: Two Categories of Methods

Contrastive Methods

- Push down on energy of **training samples**
- Pull up on energy of **suitably-generated contrastive samples**
- **Scales very badly** with dimension



Regularized Methods

- **Regularizer** minimizes the **volume** of space that can take **low energy**

Recommendations:

Abandon generative models
→ in favor of **joint-embedding architectures**

Abandon probabilistic models
→ in favor of **energy-based models**

Abandon contrastive methods
→ in favor of **regularized methods**

Abandon reinforcement learning (RL)
→ in favor of **model-predictive control**

Training a JEPA Non-Contrastively

This is the cool stuff!

- **Push down** on the energy of **compatible sample pairs**
- **Maximize** the information capacity of **representations**

Four Terms in the Cost

1. **Maximize** information content in the representation of **x**
2. **Minimize** information content in the representation of **y**
3. **Minimize** prediction error
4. **Minimize** information content of latent variable **z**

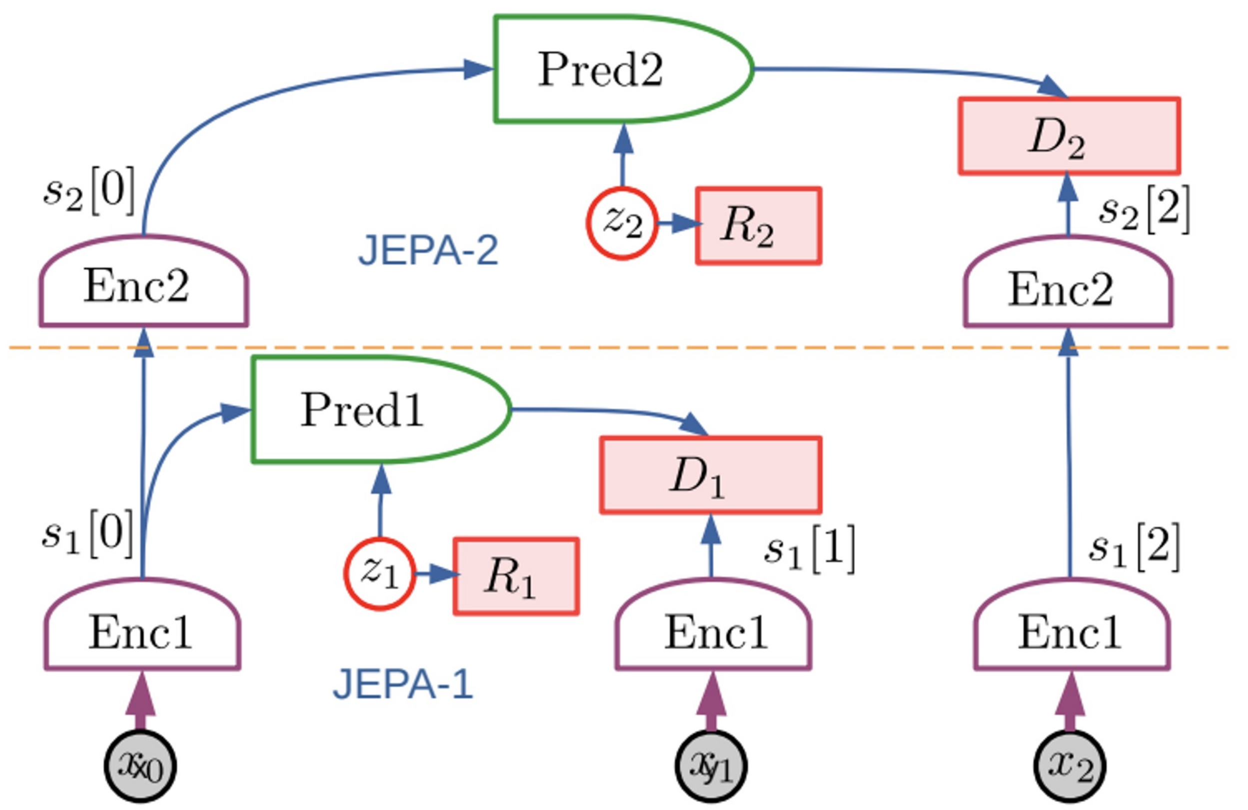
Multi-Time Scale Predictions

Higher-Level Representations

- Can predict in the longer term (we can fine tune if we observe the next side of the world)
- Contain fewer details
- Prediction is easier

People plan hierarchically. We want abstraction representation of the world to make longer term prediction.

Hierarchical JEPA = makes a prediction at multiple levels



Hierarchical Planning with Uncertainty

Hierarchical World Model

- **Hierarchical Planning**:
  - An action at level **k** specifies an objective for level **k-1**
- **Prediction**:
  - Predictions at higher levels are more **abstract** and **longer-range**

Missing from Current Architectures

- Planning/reasoning by minimizing cost with respect to **“action” variables**
  - This is lacking in current architectures, including:
    - **LLMs**
    - **Multimodal systems**
    - **Learning robots**, etc.

Steps Towards Autonomous AI Systems

1. SSL

- to learn representations of the world
- to learn predictive models of the world

2. Handling uncertainty in predictions

- Joint-embedding predictive architectures
- energy-based model framework

3. Learning world models from observation

- like animals and human babies?

4. Reasoning and planning

- that is compatible with gradient-based learning
- no symbols, no logic, vectors & continuous functions

Towards Human-Level Machine Intelligence

SSL
learning models of the world from observation

Learning to reason and plan:

- by learning to predict consequences of action
- by being driven by objectives / costs

Will machines become more intelligent than humans?
Yes, but not tomorrow.

Will machines have emotions, consciousness, moral sense?
Almost certainly yes.

Will they want to take over the world?
No!

Conclusions

Can we get machines to learn like humans and animals?
SSL, H-JEPA, Energy-Based Models, new mathematics

Will machines eventually reach human-level intelligence (HLAI)?
YES!

We hear a lot about **artificial general intelligence**,
but there is **no such thing** as general intelligence.
Intelligence is always specialized, including human intelligence.

We should talk about:
rat-level, cat-level, or human-level AI (HLAI)

Thank you!