CDS Seminar

SSL, JEPA, World Models and the Future of AI

Yann Le Cun
New York University
Meta – Fundamental AI Research

NYU CDS
2025-09-10

M-51, HSO

We Need Human-Level AI for Intelligent Assistant

- **In the near future, all of our interactions with the digital world will be mediated by AI assistants.**
- **Intelligent assistants that can helps us in our daily lives**
- **Smart glasses**
  - Communicates through voice, vision, display, EMG…
- **We need machines with human-level intelligence**
  - Machines that understand how the world works
  - Machines that can remember
  - Machines that can reason and plan.

"Her" (2013)

Meta Orion (2024)

The Ubiquitous AI Assistant is Becoming A Reality

- **Ray-Ban Meta (today)**
  - Cameras / microphone / speakers
  - no display
  - Voice interface to Meta AI assistant

- **Meta's Orion Demonstrator (future)**
  - Cameras / microphones
  - Augmented reality color display
  - Voice + EMG bracelet interface

But Machine Learning Sucks! (compared to humans and animals)

- **Supervised learning (SL) requires large numbers of labeled samples.**
- **Reinforcement learning (RL) requires insane amounts of trials.**
- **Self-Supervised Learning (SSL) works great but...**
  - Generative prediction only works for text and other discrete modalities

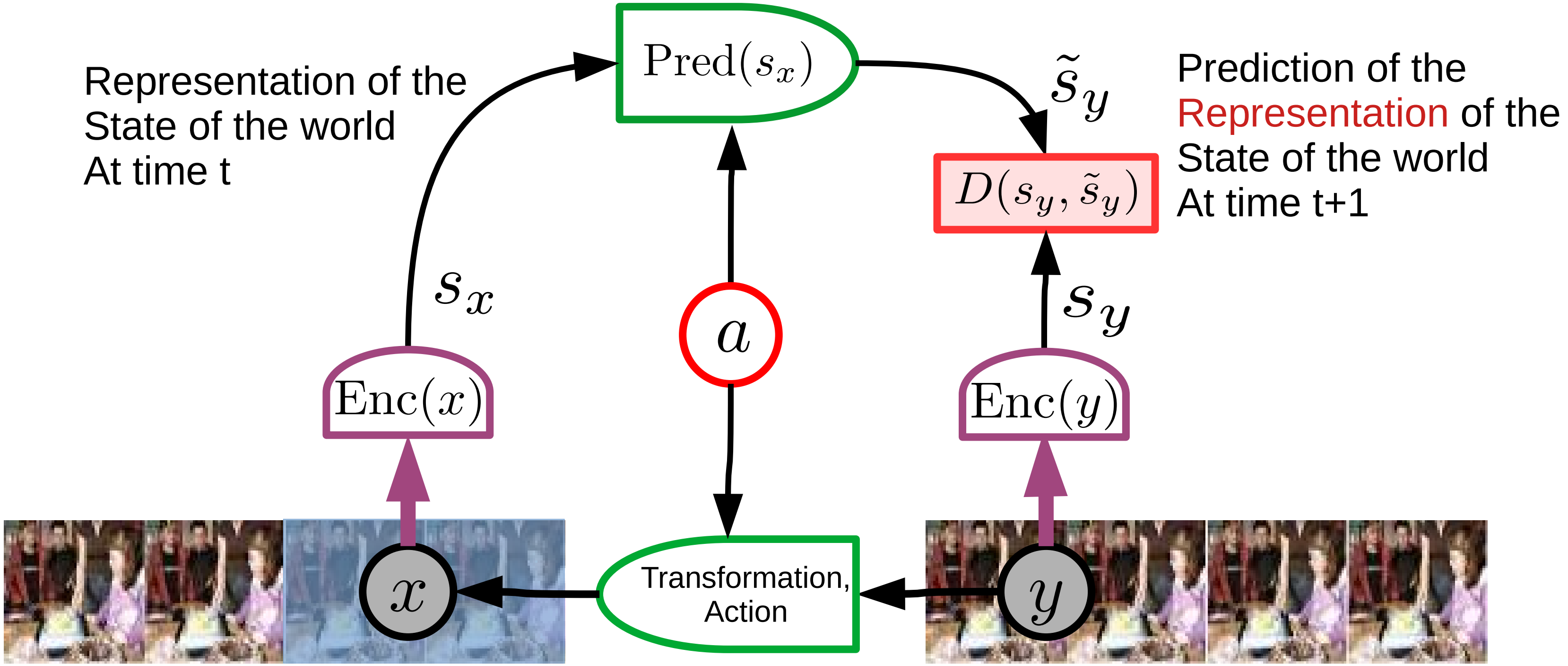
- **Animals and humans:**
  - Can learn new tasks **very** quickly.
  - Understand how the world works
  - Can reason an plan
- **Humans and animals have common sense**
- **Their behavior is driven by objectives (drives)**

What's a universal foundation model architecture

- **Captures structure in the data**
  - Discovers dependencies in a task-independent way
- **Trained with Self-Supervised Learning (SSL)**
  - No need for labels
- **Learns abstract representations in the data**
  - Representations that allow to make predictions
- **Learns a predictive model**
  - Observation x, transformed observation y=Trans(x,a)
  - Encoding : representations s_x = Enc(x), s_y = Enc(y)
  - Prediction of s_y : p_y = Pred(s_x, a)

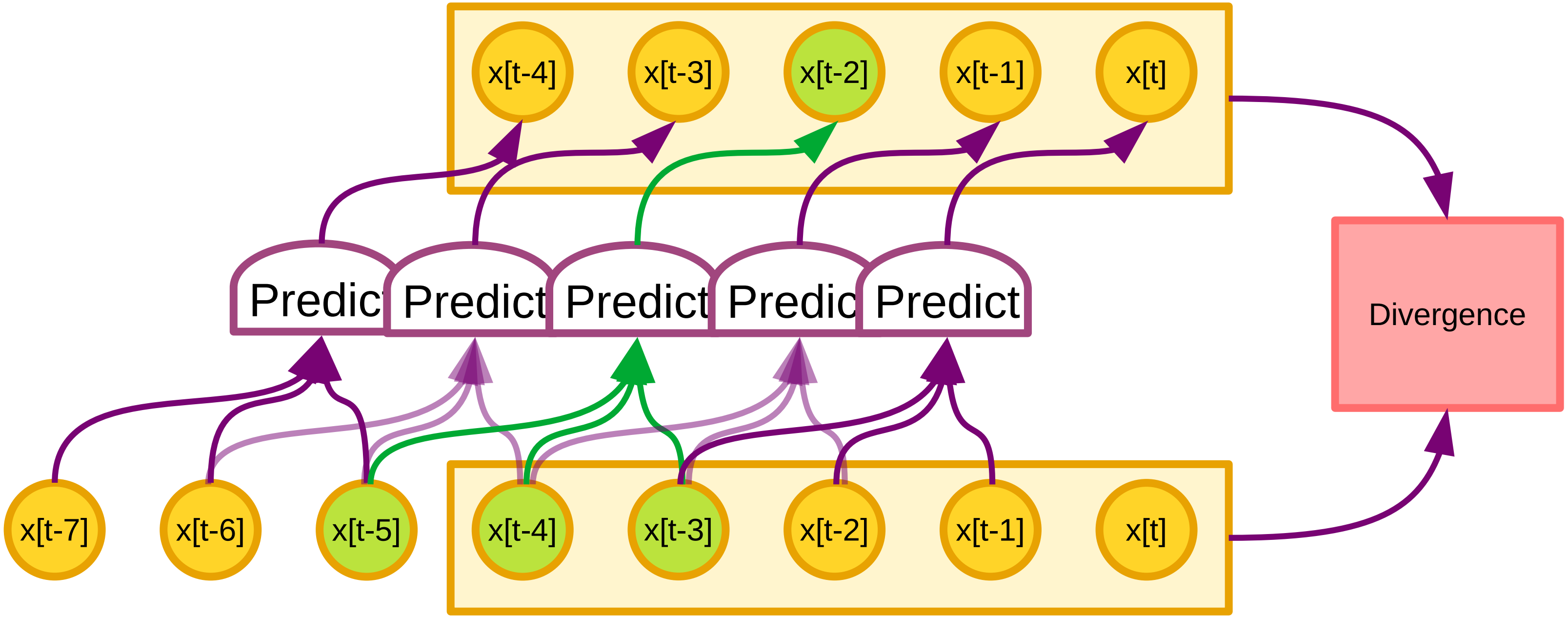
Predictive Model with JEPA

- **Joint Embedding Predictive Architecture (JEPA)**
  - [LeCun 2022], [Garrido 2023], [Bardes 2023], [Assran 2023], [Garrido 2024]



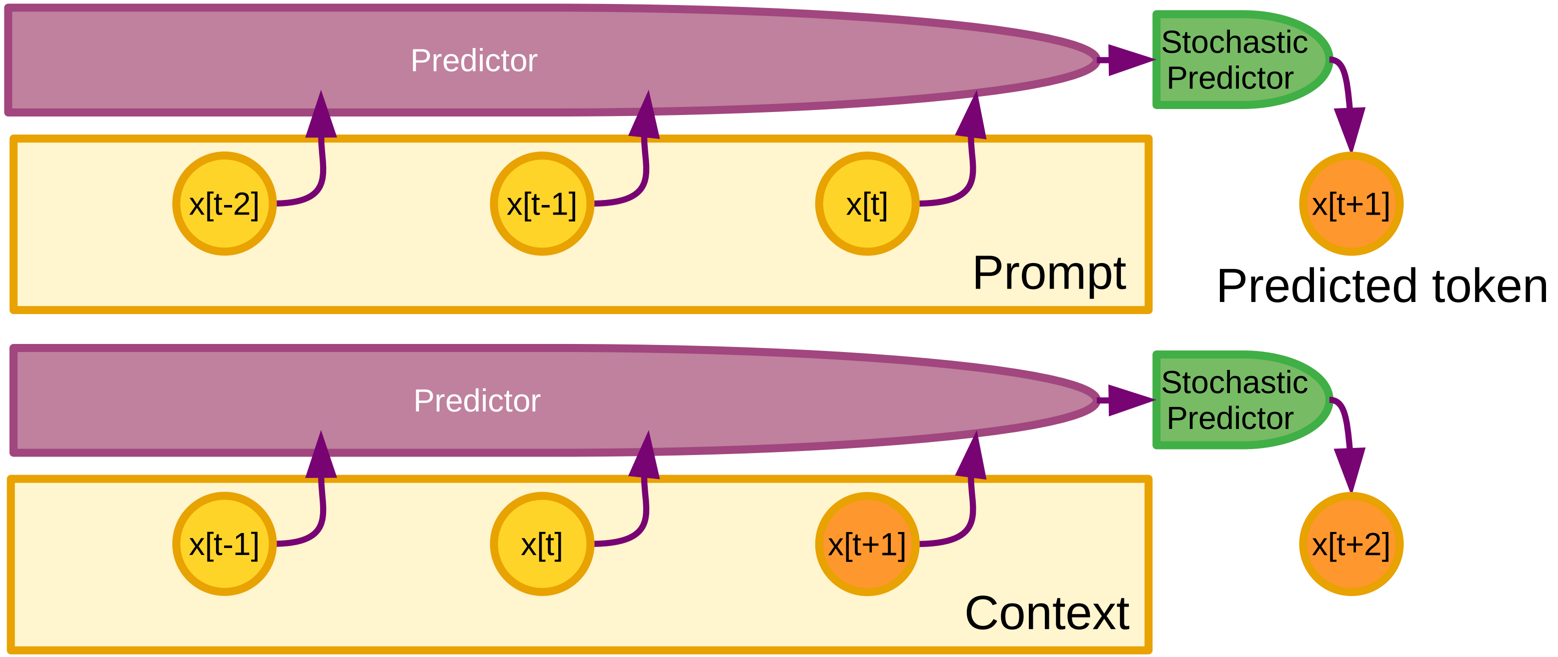
AE Collapse Prevention through Architectural Constraints

- **Train an auto-encoder with causal connections**
- **No connection between an input and its corresponding output**
- **LLMs / GPT architectures are the most popular example**
  - Trained to predict the next input.



Auto-Regressive LLM. Inject predicted token in the input

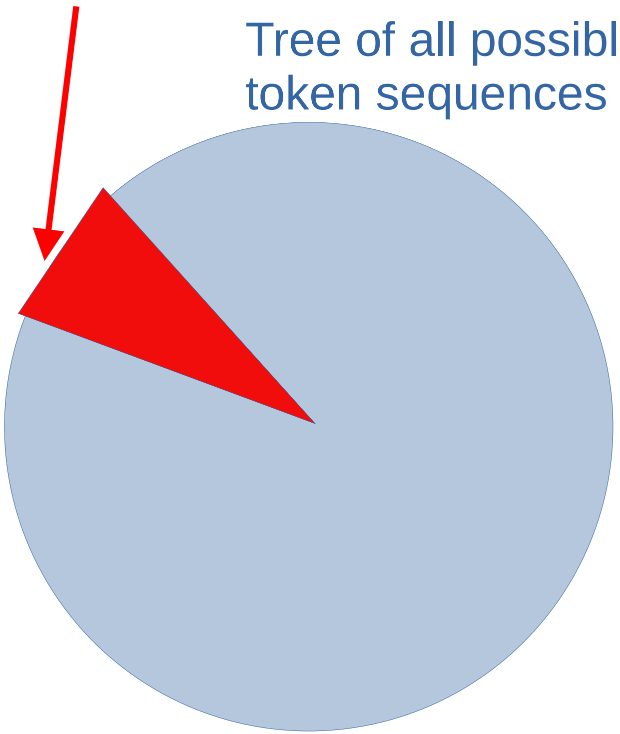
- Outputs one token after another through feed-forward prediction
- Tokens may represent words, image patches, speech segments…
- Predictor has a fixed number of layers
- Only works for discrete domains (text, DNA….)



Auto-Regressive Generative Models Suck!

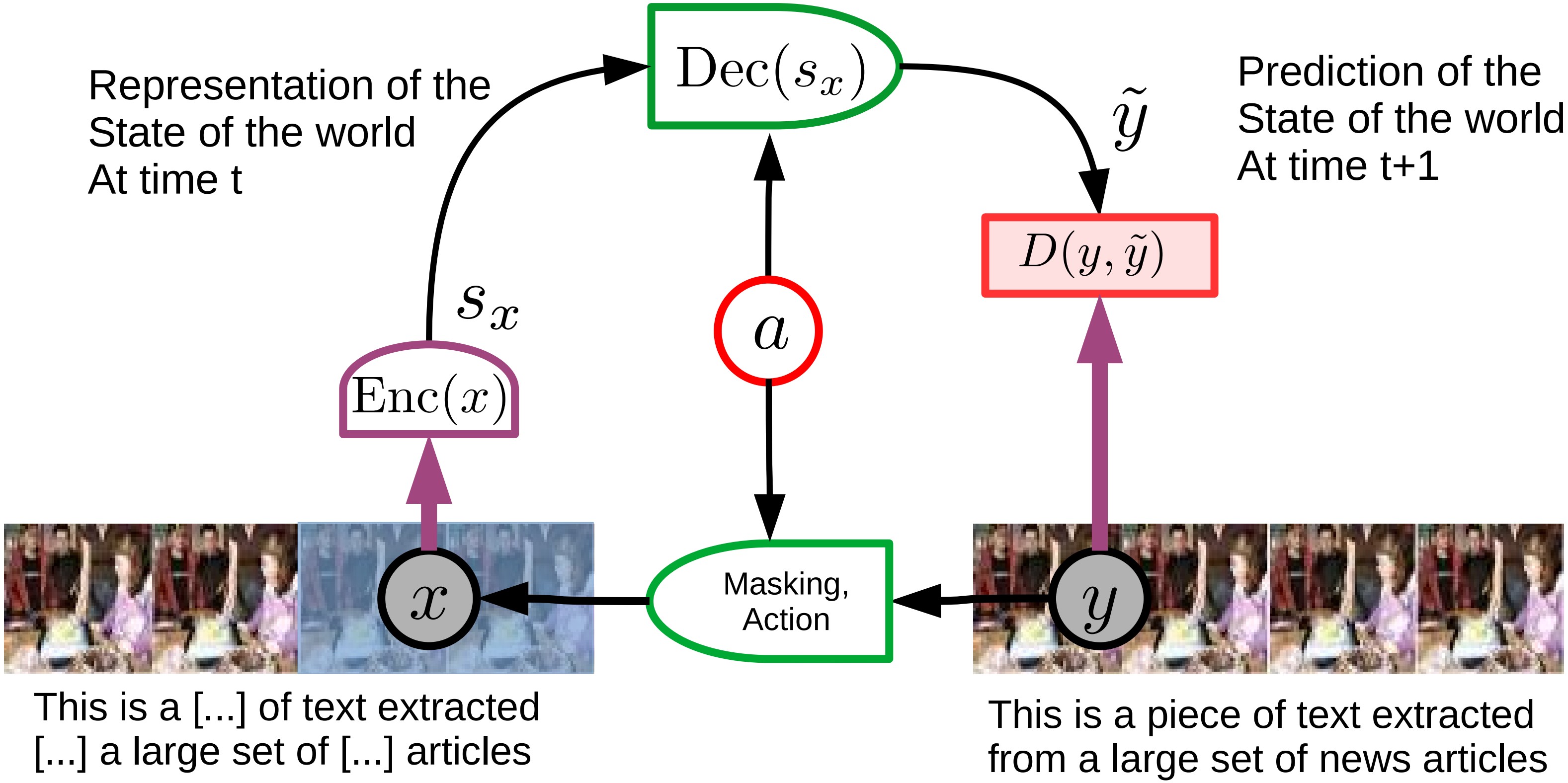
- **Auto-Regressive LLMs are doomed.**
- **They cannot be made factual, non-toxic, etc.**
- **They are not controllable**
- **Probability e that any produced token takes us outside of the set of correct answers**
- **Probability that answer of length n is correct (assuming independence of errors):**
  - $P(\text{correct}) = (1-e)^n$
- **This diverges exponentially.**
- **It's not fixable (without a major redesign).**

  - See also [Dziri...Choi, ArXiv:2305.18654]



Can we train Generative Architecture with Continuous Data?

- **Short answer: NO!!!**
- **It works for discrete domains, not high-dim domains**
- **Generative world model architecture**

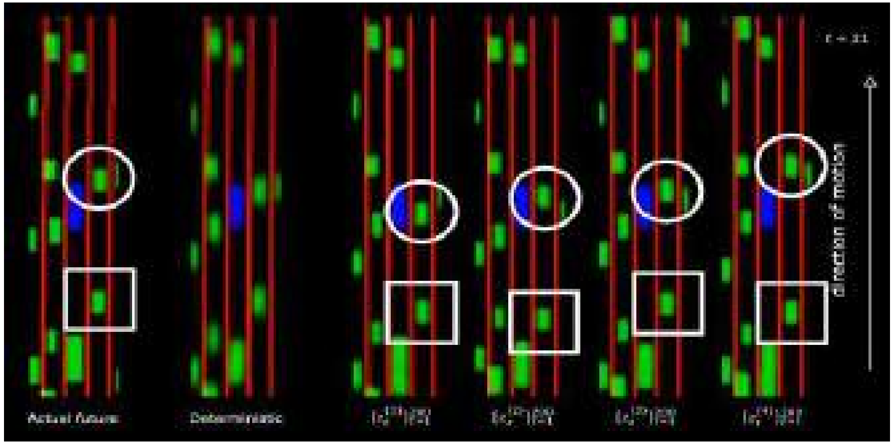


Generative Architectures DO NOT Work for Images and video

- **Because the world is only partially predictable**
- **A predictive model should represent multiple predictions**
- **Probabilistic models are intractable in high-dim continuous domains.**
- **Generative Models must predict every detail of the world**

- **My solution: Joint-Embedding Predictive Architecture**

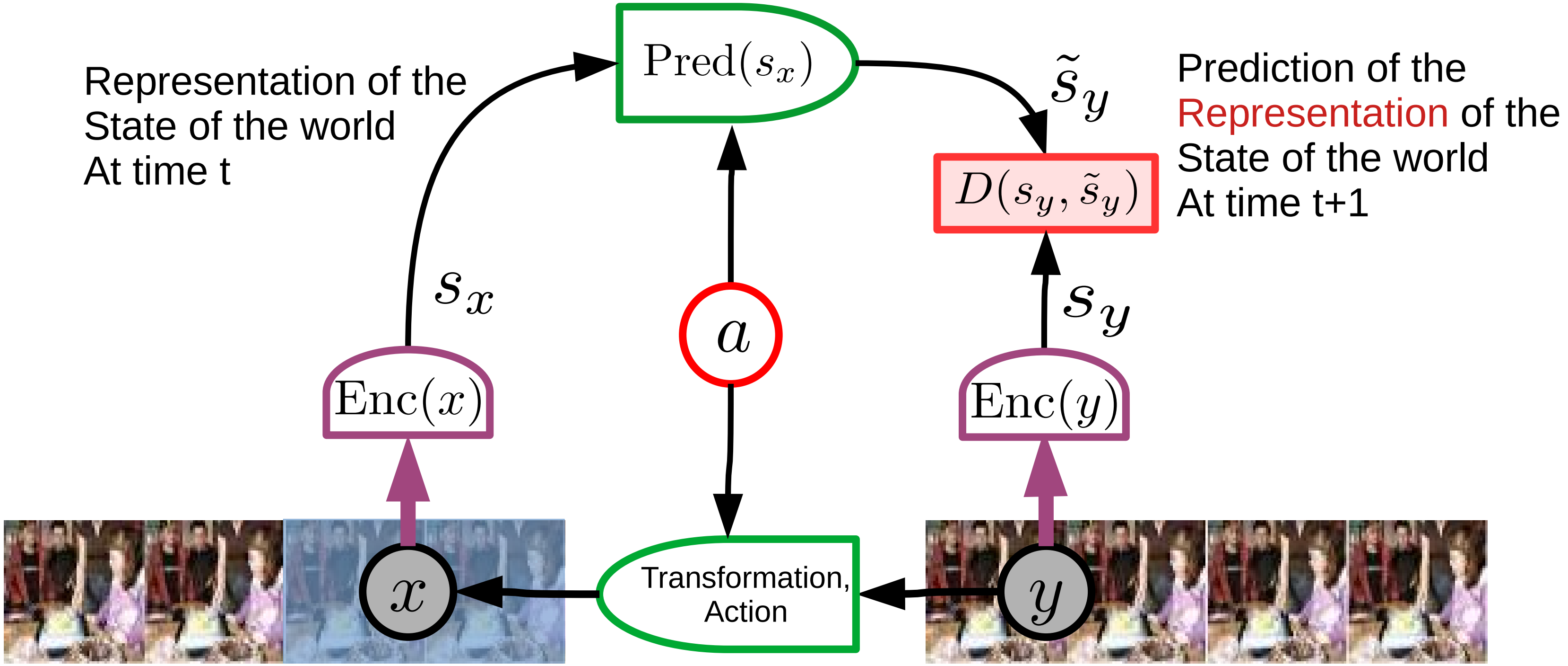
[Mathieu, Couprie, LeCun ICLR 2016]



[Henaff, Canziani, LeCun ICLR 2019]

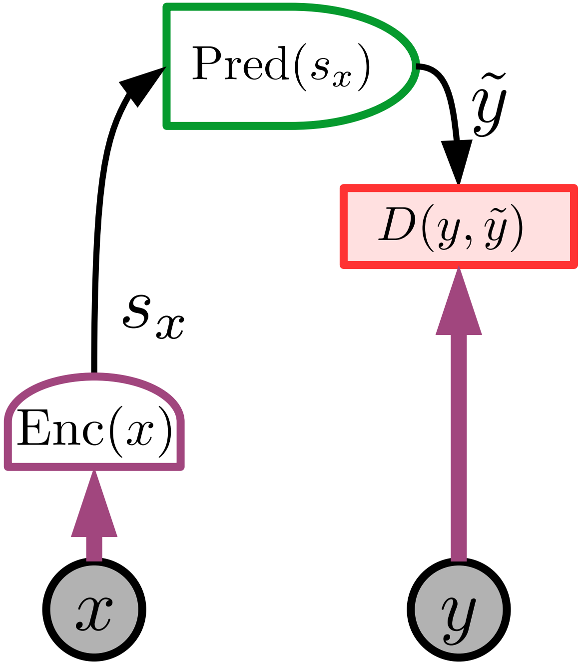
Joint Embedding World Model: Self-Supervised Training

- **Joint Embedding Predictive Architecture (JEPA)**
  - [LeCun 2022], [Garrido 2023], [Bardes 2023], [Assran 2023], [Garrido 2024]

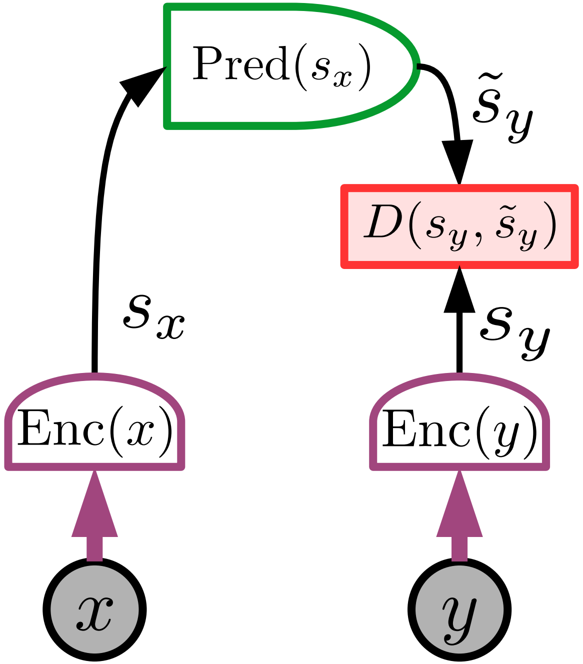


Architectures: Generative vs Joint Embedding

- **Generative: predicts y** (with all the details, including irrelevant ones)
- **Joint Embedding: predicts an abstract representation of y**
- **JEPA lifts the abstraction level, generative architectures do not.**



a) Generative Architecture
Examples: VAE, MAE...

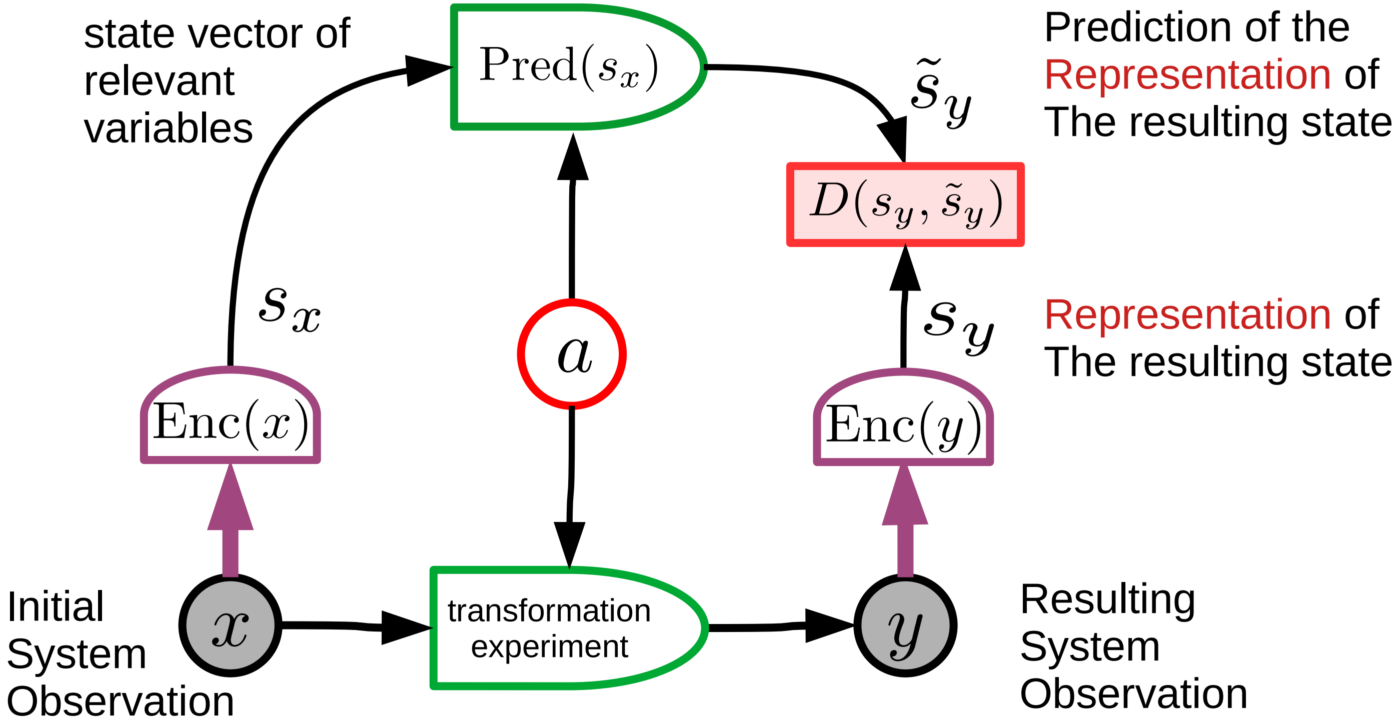


b) Joint Embedding Architecture

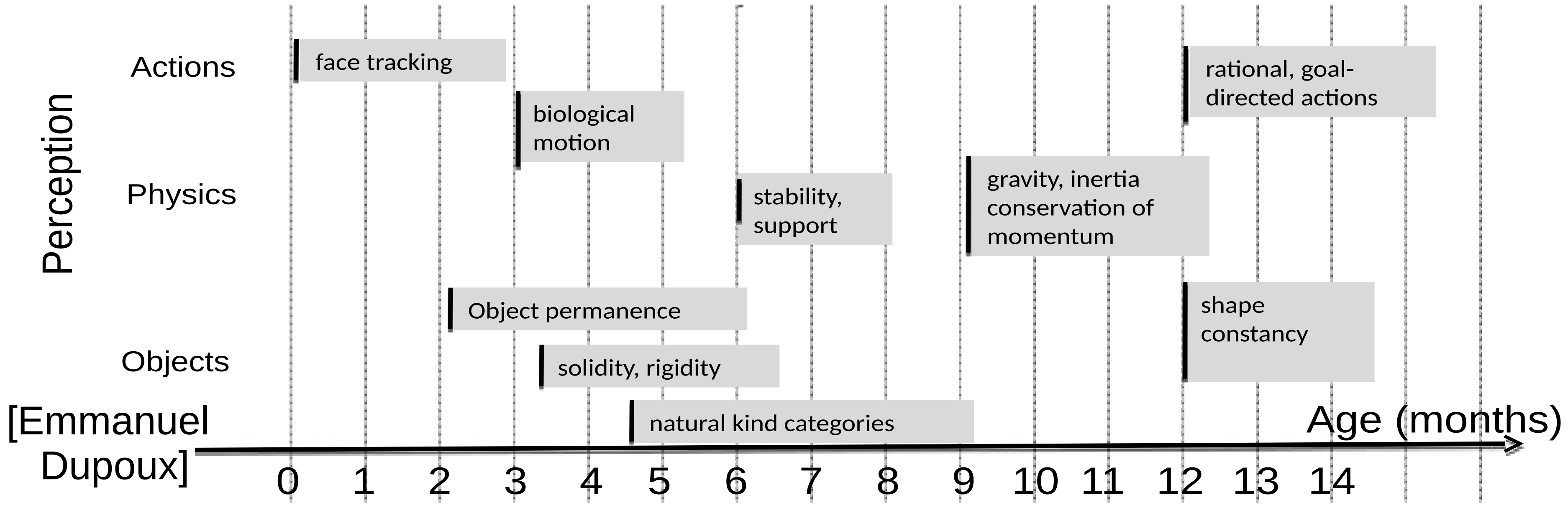
This is how models are built in traditional physics

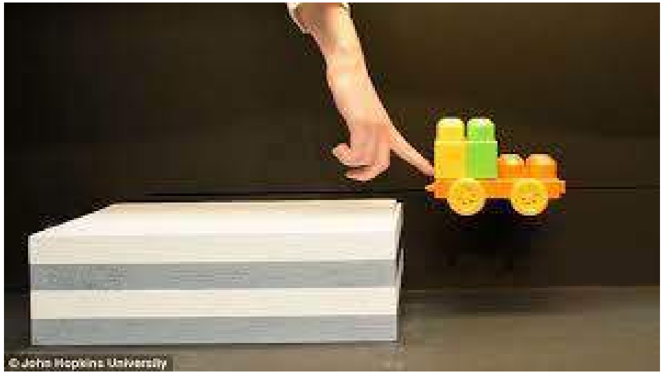
- **Find an abstract state representation that allows to make predictions**
- **Extract the state representation from observation/measurement**
- **Predict outcome resulting from an intervention/experiment**

- **Irrelevant and unpredictable information is eliminated from the representation**
- **The representation contains information that makes prediction possible**



How do babies learn how the world works?





- **How do we get machines to learn like babies?**

Current architectures are missing something really big!

- **Never mind humans, cats and dogs can do amazing feats**
  - Current robots intelligence doesn't come anywhere close
- **Any house cat can plan highly complex actions**
- **Any 10 year-old can clear up the dinner table and fill up the dishwasher without learning ("zero-shot")**
- **Any 17 year-old can learn to drive a car in 20 hours of practice**
- **AI systems that can pass the bar exam, do math problems, prove theorems….**
- **...but where are my Level-5 self-driving car and my domestic robot?**
- **We keep bumping into Moravec's paradox**
  - Things that are easy for humans are difficult for AI and vice versa.

Our world model needs to be trained from sensory inputs

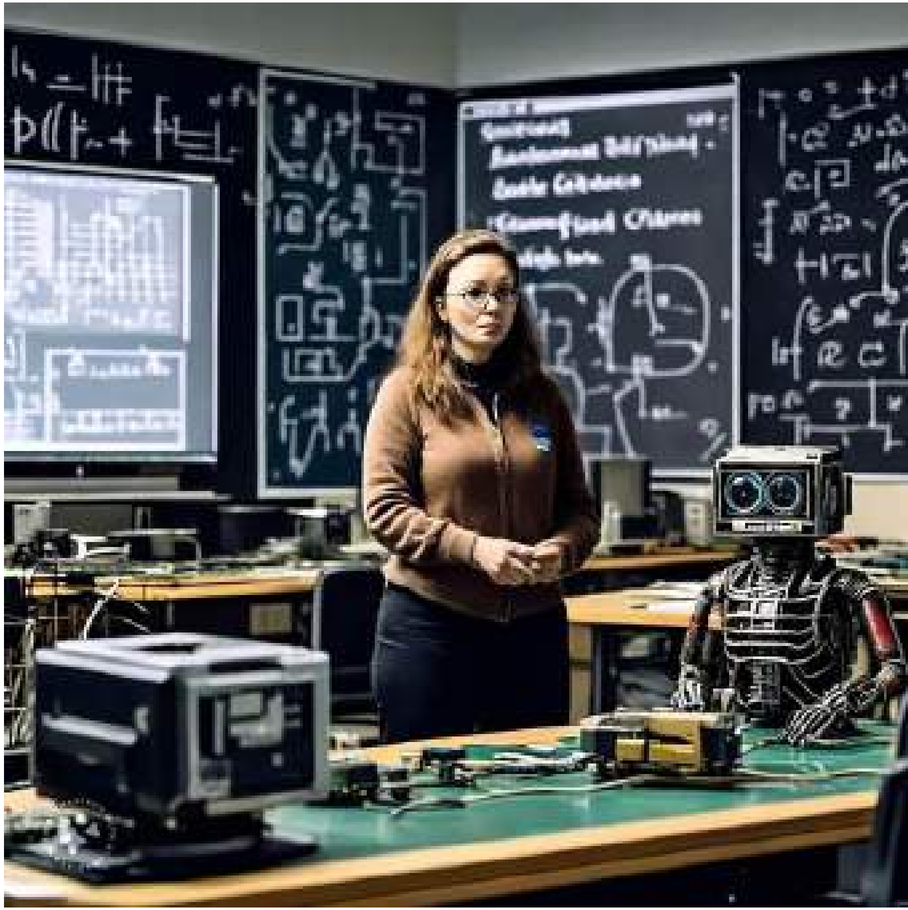
- **LLM**
  - Trained on 3.0E13 tokens (2E13 words). Each token is 3 bytes.
  - Data volume: 0.9E14 bytes.
  - Would take 450,000 years for a human to read (12h/day, 250 w/minute)

- **Human child**
  - 16,000 wake hours in the first 4 years (30 minutes of YouTube uploads)
  - 2 million optical nerve fibers, carrying about 1 byte/sec each.
  - Data volume: 1.1E14 bytes

- **A four year-old child has seen more data than an LLM !**

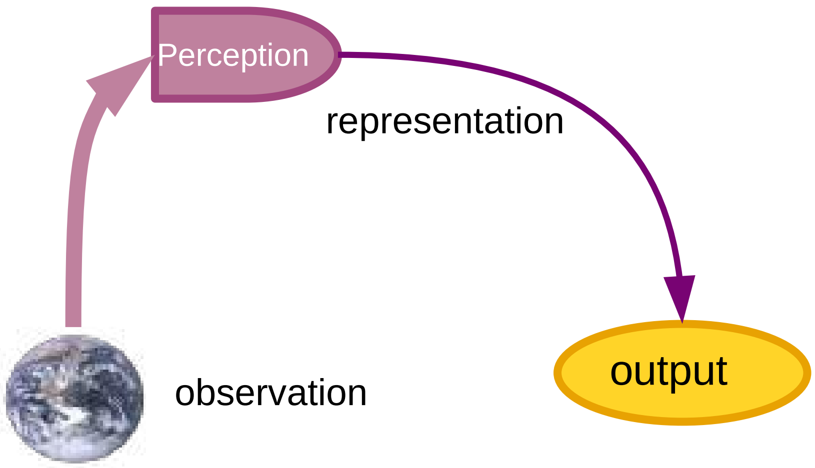
Desiderata for AMI (Advanced Machine Intelligence)

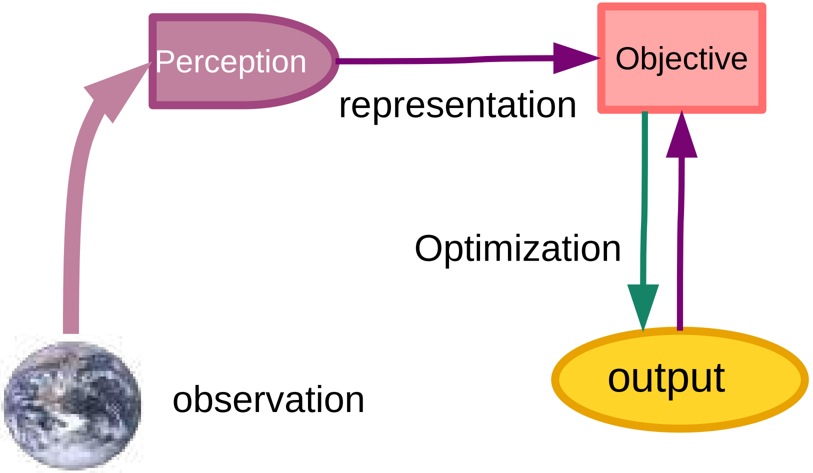
- **Systems that learn world models from sensory inputs**
  - E.g. learn intuitive physics from video
- **Systems that have persistent memory**
  - Large-scale associative memories
- **Systems that can plan actions**
  - So as to fulfill an objective
- **Systems that can reason**
  - Inventing new solutions to unseen problems
- **Systems that are controllable & safe**
  - By design, not by fine-tuning.



Inference: feed-forward propagation vs optimization

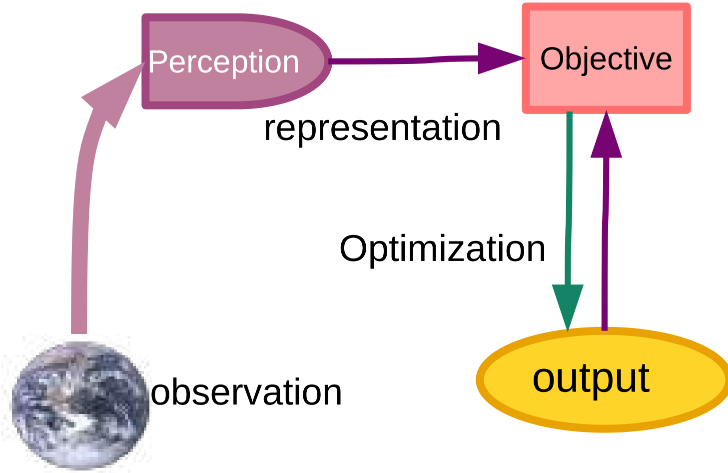
- **What is reasoning and planning?**
- **Feed-forward propagation is insufficient**
- **Complex inference requires the optimization of an objective**
- **Every computational problem can be reduced to optimization**
  - This includes every inference and planning problem.
- **Energy-Based Model**





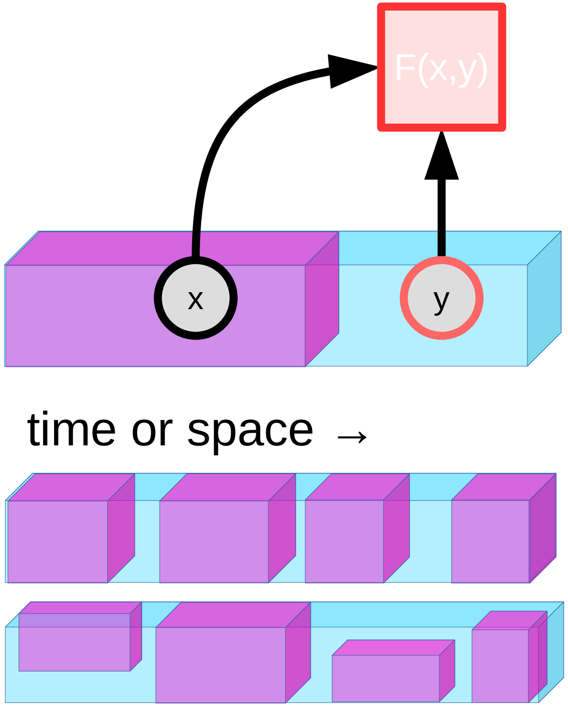
Inference through optimization: Objective-Driven AI.

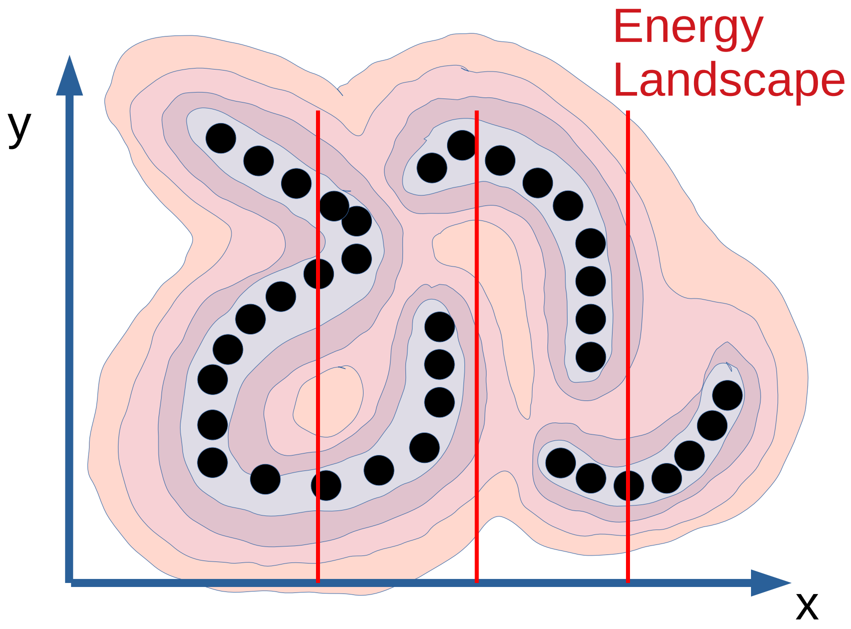
- **Inference through optimization is used in classical methods**
  - Probabilistic graphical models, Bayesian nets
  - Model-Predictive Control in robotics
  - Search & planning in “classical” AI
- **In the past, all of AI was viewed as a search or optimization problem**
  - Path planning, Block World, Towers of Hanoi, SAT, logical inference
- **Optimization-based inference enables zero-shot “learning”**
  - It can find innovative solutions to unseen problems.
  - All game-playing AI systems use search/planning
- **Optimization-based inference is “System 2”**



Capturing Dependencies with Energy-Based Models

- **The only way to formalize & understand all model types**
  - Gives low energy to compatible pairs of x and y
  - Gives higher energy to incompatible pairs

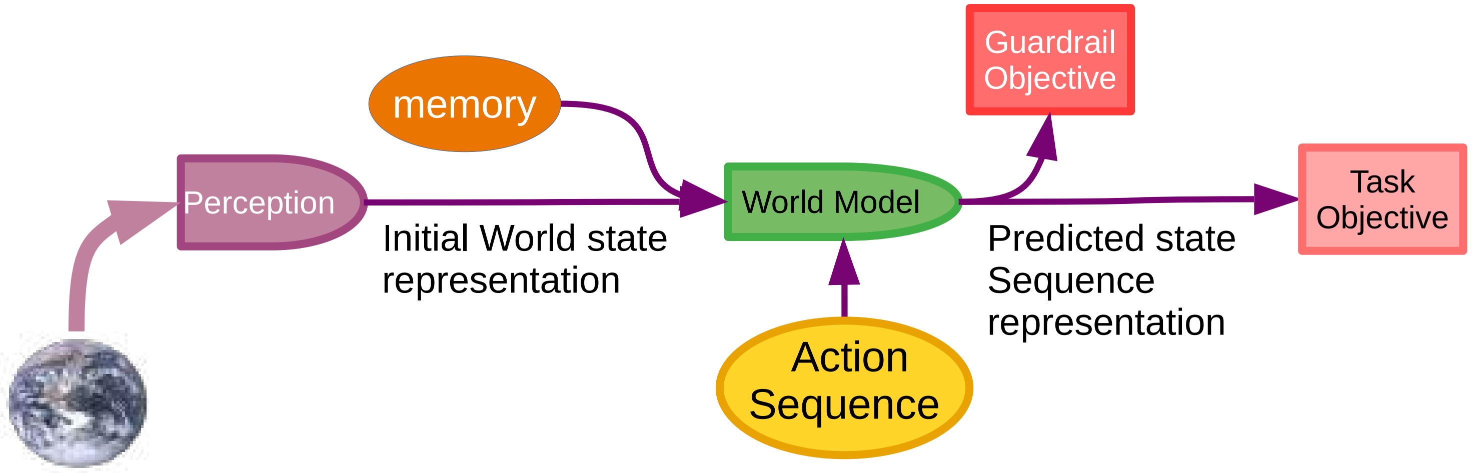




$$\check{y} = \mathrm{argmin}_y F(x, y)$$

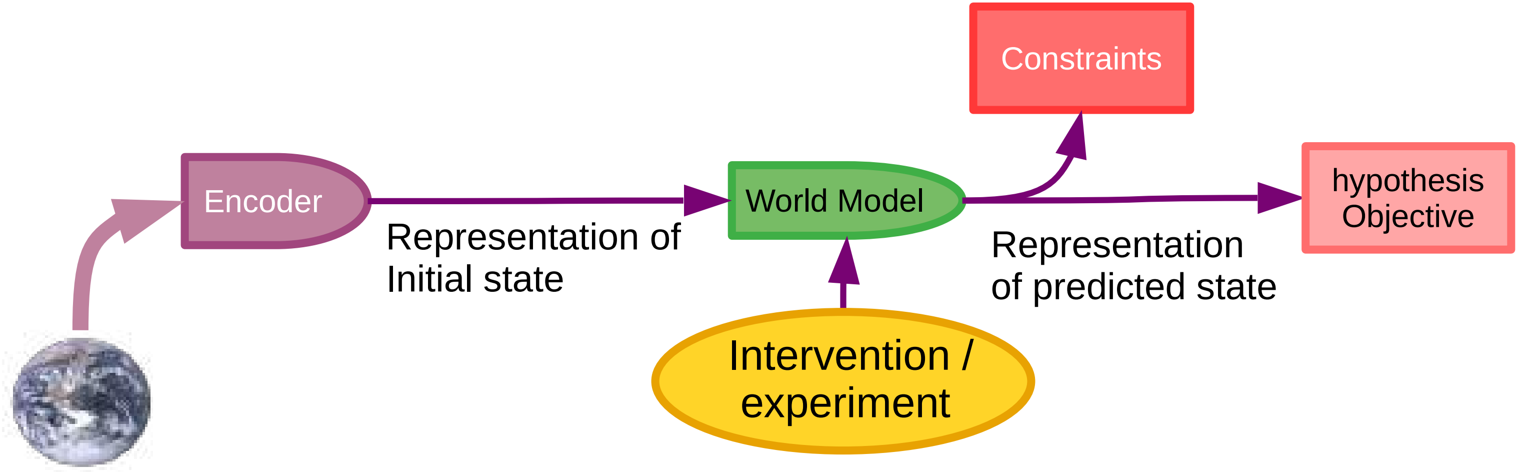
2. World Model for Planning/Reasoning

- **Perception:** Computes an abstract representation of the state of the world
  - Possibly combined with previously-acquired information in memory
- **World Model:** Predict the state resulting from an imagined action sequence
- **Task Objective:** Measures divergence to goal
- **Guardrail Objective:** Immutable objective terms that ensure safety
- **Operation: Finds an action sequence that minimizes the objectives**



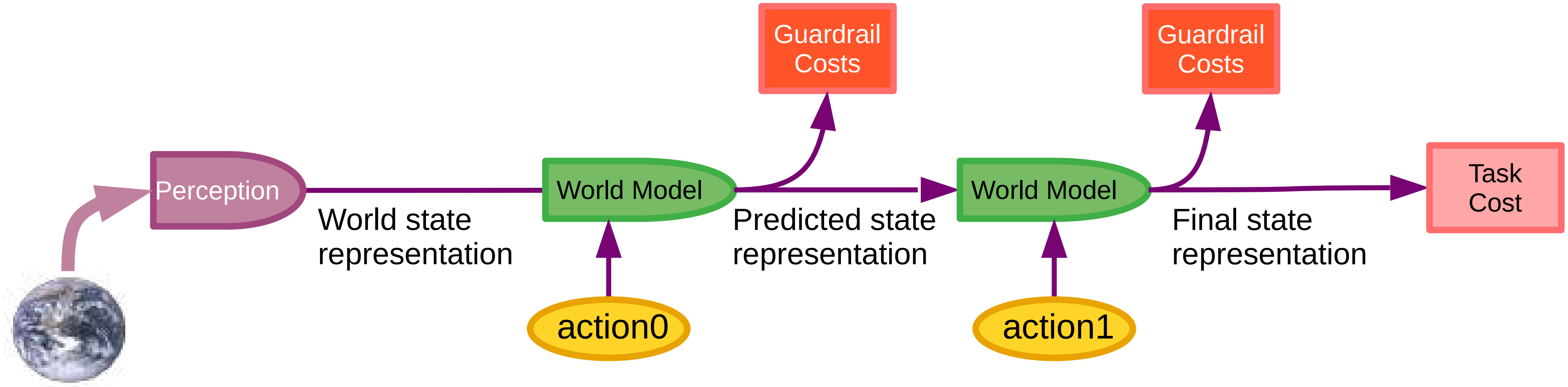
2. Models for Physics Experiments

- **Encoder:** Computes an abstract representation of the state of the system
- **World Model:** Predict the state resulting from an imagined experiment or intervention.
- **Hypothesis Objective:** Measures divergence to the result expected from the experiment
- **Constraints:** that the trajectory must satisfy.
- Find an action an experiment that validates or invalidates the hypothesis



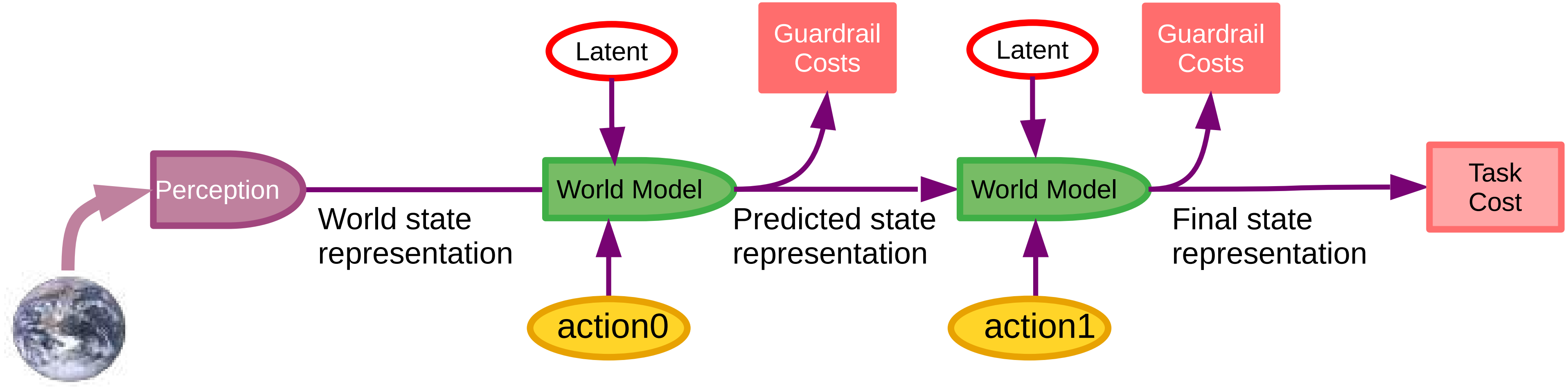
Objective-Driven AI: Multistep/Recurrent World Model

- **Same world model applied at multiple time steps**
- **Guardrail costs applied to entire state trajectory**
- **This is identical to Model Predictive Control (MPC)**
  - But with a trained world model
- **Action inference by minimization of the objectives**
  - Using gradient-based method, graph search, dynamic prog, A*, MCTS,….



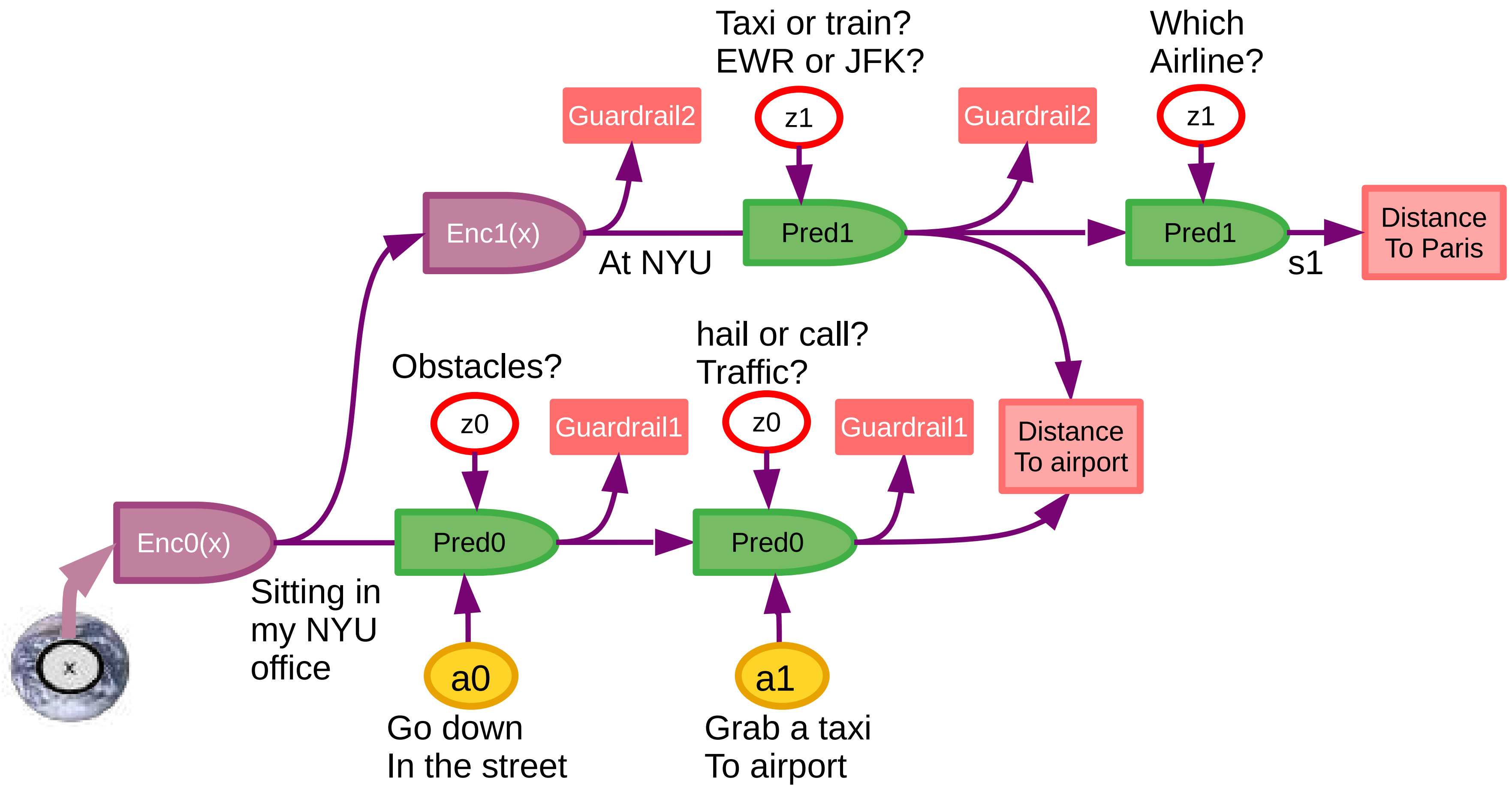
Objective-Driven AI: Non-Deterministic World Model

- **The world is not deterministic or fully predictable**
- **Latent variables parameterize the set of plausible predictions**
  - Can be sampled from a prior or swept through a set.
  - Planning can be done for worst case or average case
  - Uncertainty in outcome can be predicted and quantified



Objective-Driven AI: Hierarchical Planning

▶ Hierarchical Planning: going from NYU to Paris



Objective-Driven AI Systems

AI that can learn, understand the world,
reason, plan,
Yet is safe and controllable

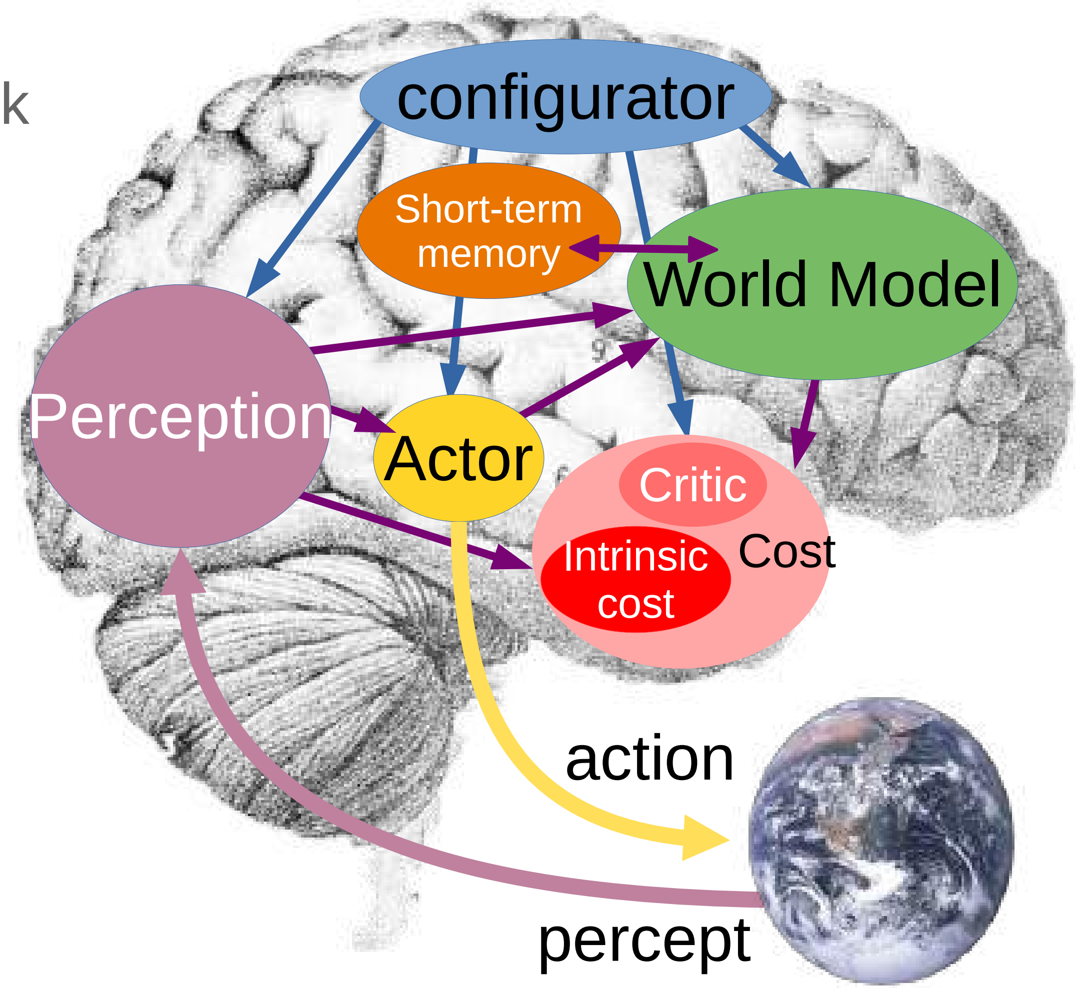
“A path towards autonomous machine intelligence”
https://openreview.net/forum?id=BZ5a1r-kVsf

[previous versions of this talk available on YouTube]

Y. LeCun

Modular Cognitive Architecture for AMI

- **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

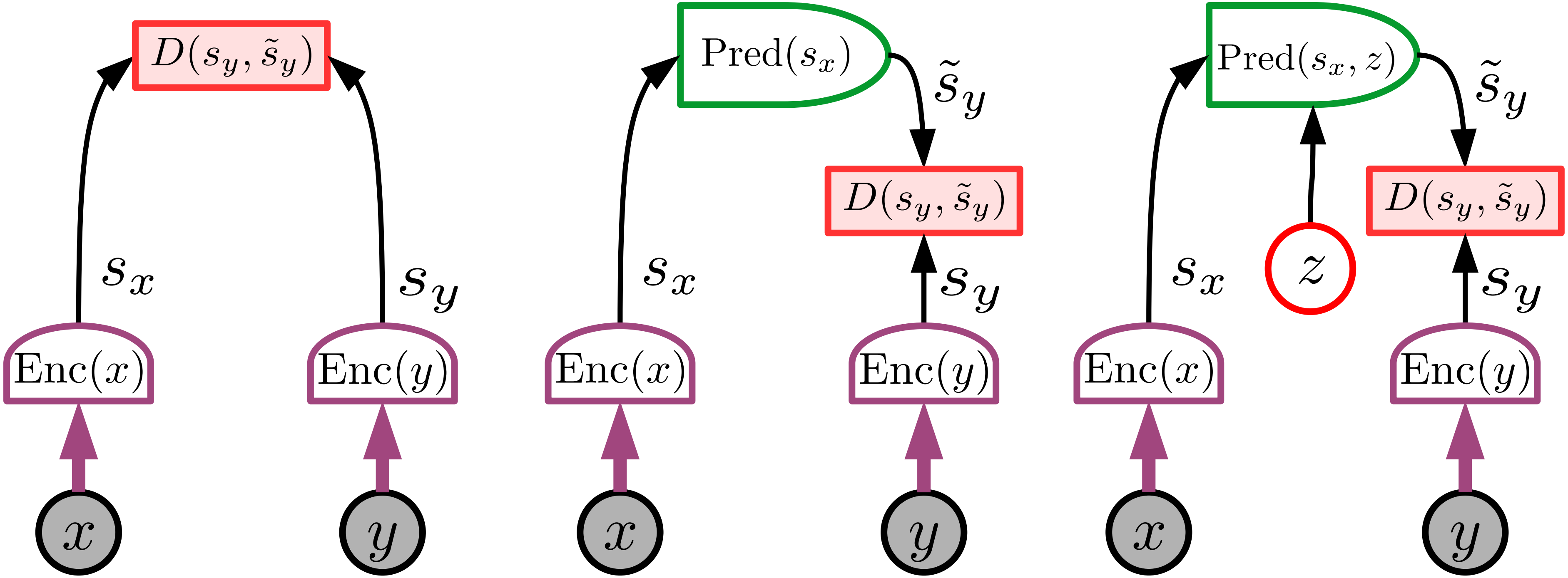


How could Machines Learn World Models from Observations?

Self-Supervised Learning

Joint Embedding Architectures

- Computes abstract representations for x and y
- Tries to make them equal or predictable from each other.



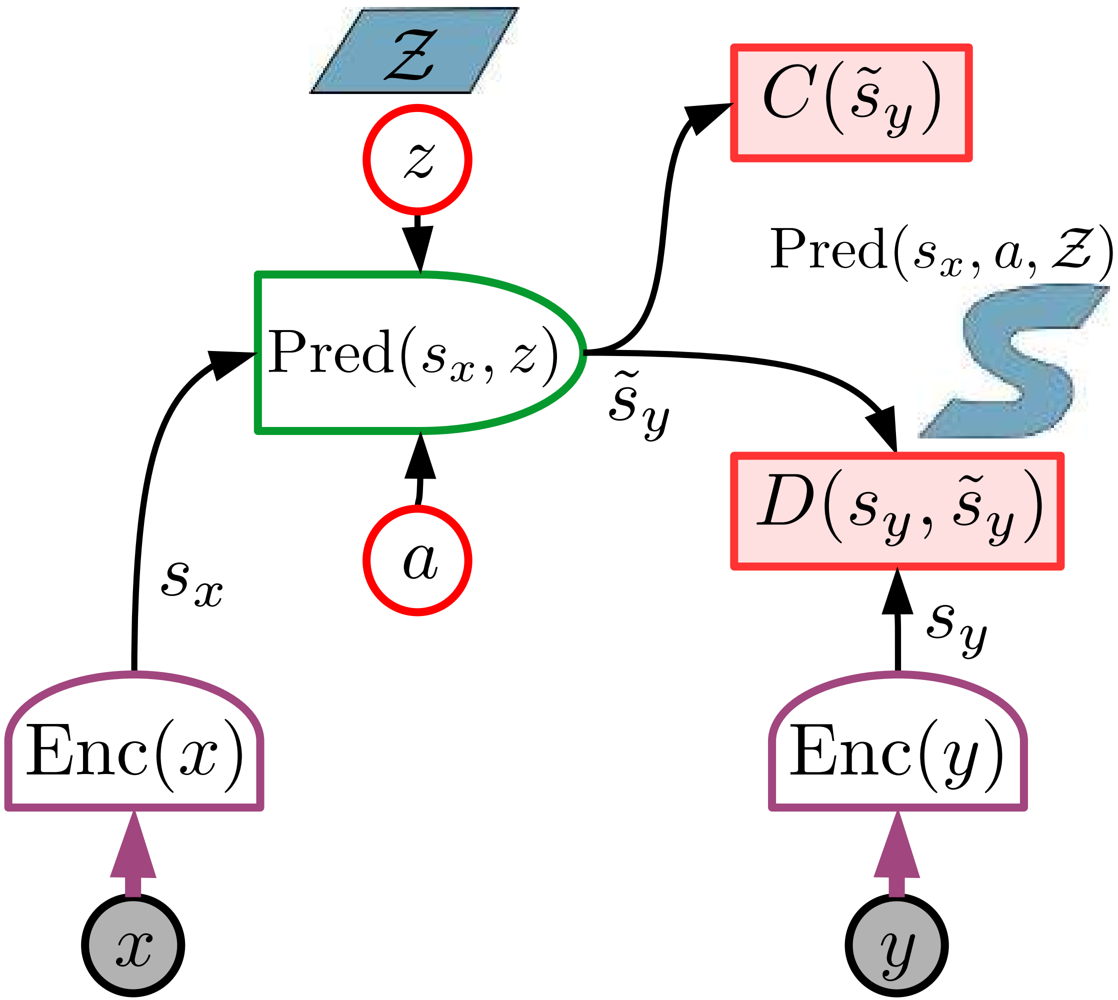
a) Joint Embedding Architecture (JEA)
Examples: Siamese Net, Pirl, MoCo, SimCLR, BarlowTwins, VICReg,

b) Deterministic Joint Embedding Predictive Architecture (DJEPA)
Examples: BYOL, VICRegL, I-JEPA

c) Joint Embedding Predictive Architecture (JEPA)
Examples: Equivariant VICReg
I-JEPA…..

Architecture for action-conditioned world models: 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
  - JEPA predicts a representation of the future Sy from a representation of the past and present Sx



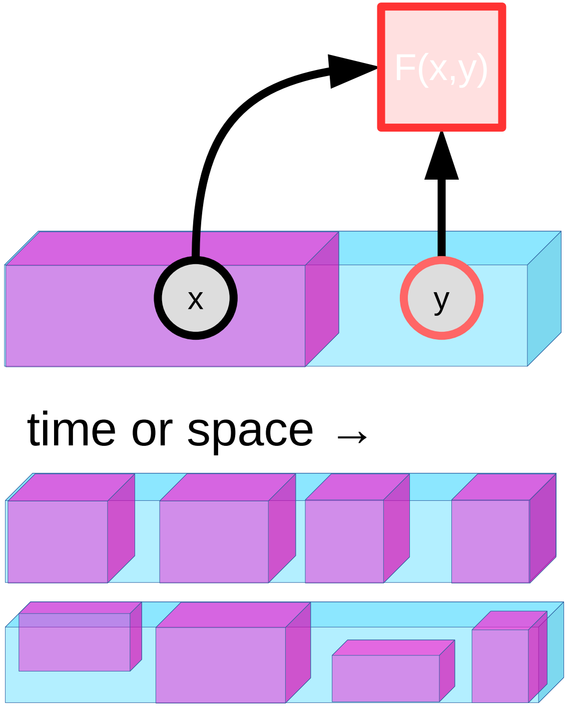
Energy-Based Models for Self-Supervised Learning

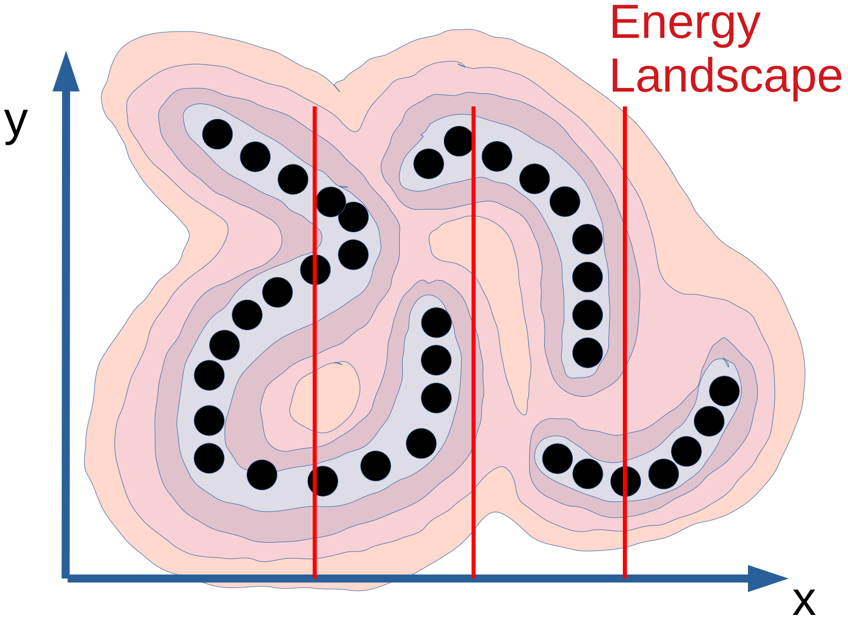
Capturing dependencies through an energy function

Probabilistic modeling is intractable in high-dimensional continuous domains.

Energy-Based Models: Implicit function

- **The only way to formalize & understand all model types**
  - Gives low energy to compatible pairs of x and y
  - Gives higher energy to incompatible pairs



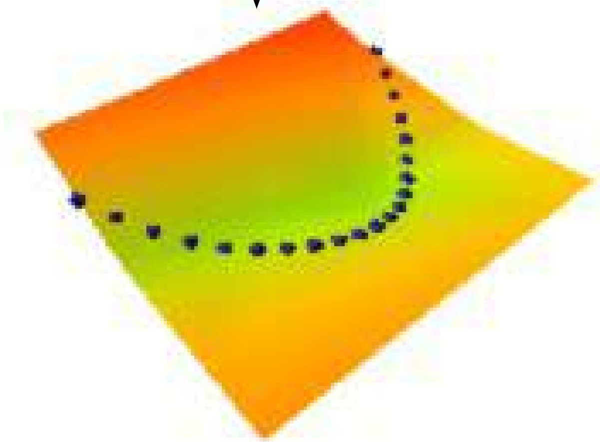


$$\check{y} = \mathrm{argmin}_y F(x, y)$$

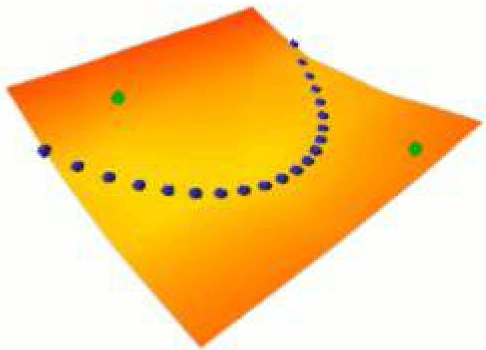
Training Energy-Based Models: Collapse Prevention

- **A flexible energy surface can take any shape.**
- **We need a loss function that shapes the energy surface so that:**
  - Data points have low energies
  - Points outside the regions of high data density have higher energies.

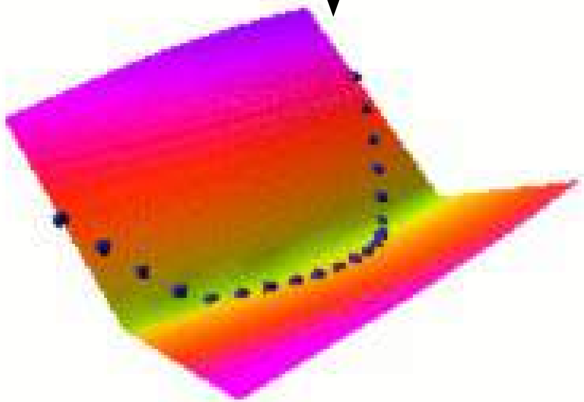
Collapse!



Contrastive Method

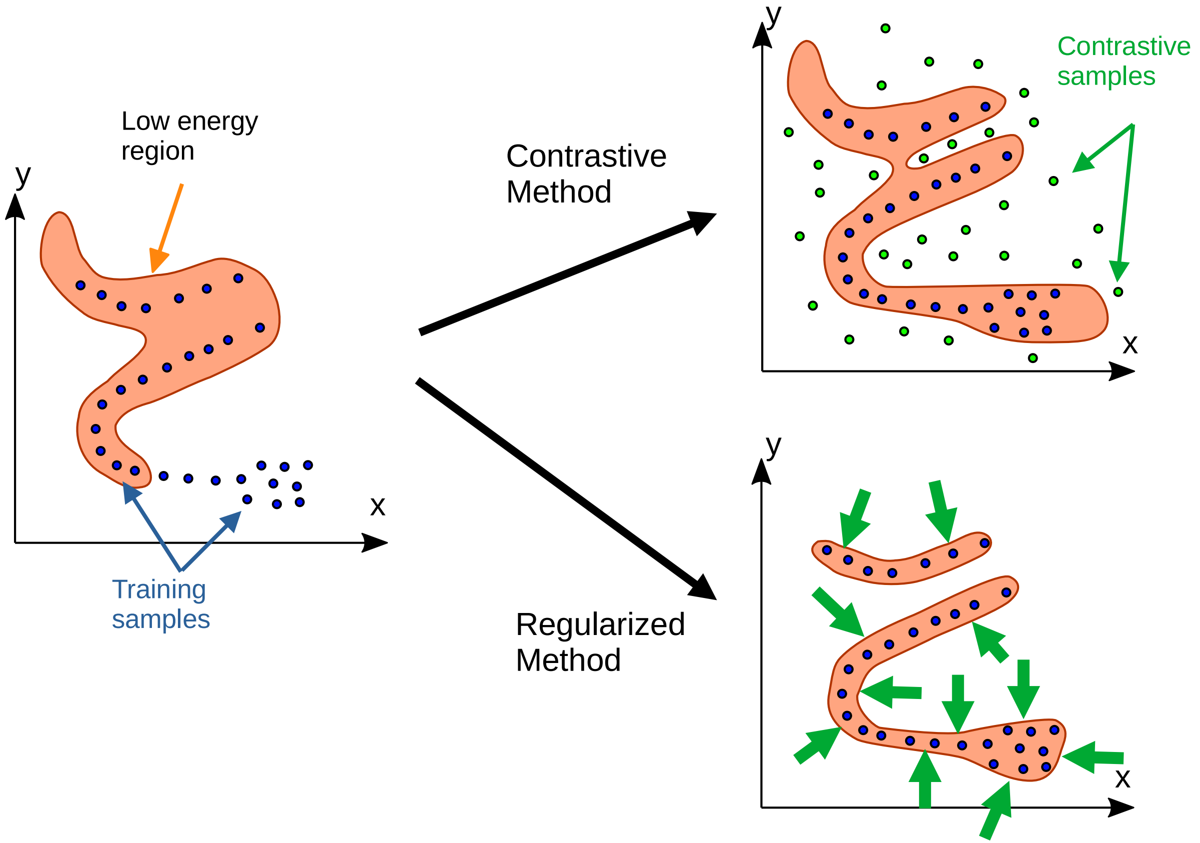


Regularized Methods



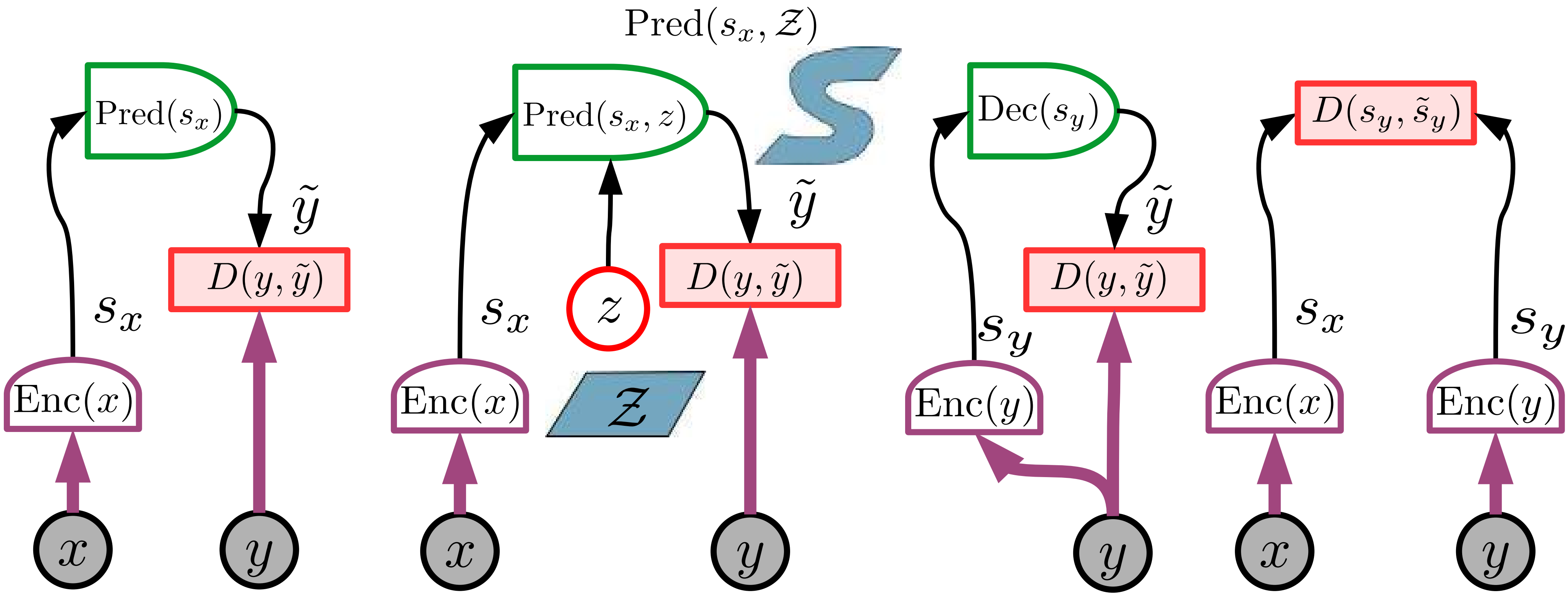
EBM Training: two categories of methods

- **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



EBM Architectures

- **Some architectures can lead to a collapse of the energy surface**



a) Prediction / regression
NO COLLAPSE

b) Generative latent-variable Architecture
CAN COLLAPSE

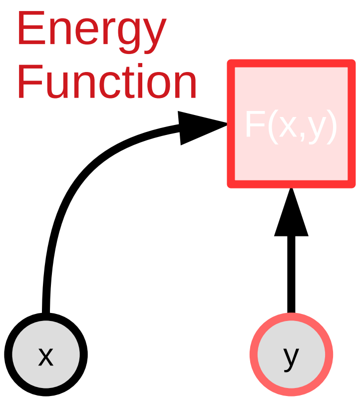
c) Auto-Encoder
CAN COLLAPSE

d) Joint Embedding Architecture
CAN COLLAPSE

Energy-Based Models vs Probabilistic Models

- **Probabilistic models are a special case of EBM**
  - Energies are like un-normalized negative log probabilities
- **Why use EBM instead of probabilistic models?**
  - EBM gives more flexibility in the choice of the scoring function.
  - More flexibility in the choice of objective function for learning
- **From energy to probability: Gibbs-Boltzmann distribution**
  - Beta is a positive constant

$$P(y|x) = \frac{e^{-\beta F(x,y)}}{\int_{y'} e^{-\beta F(x,y')}}$$

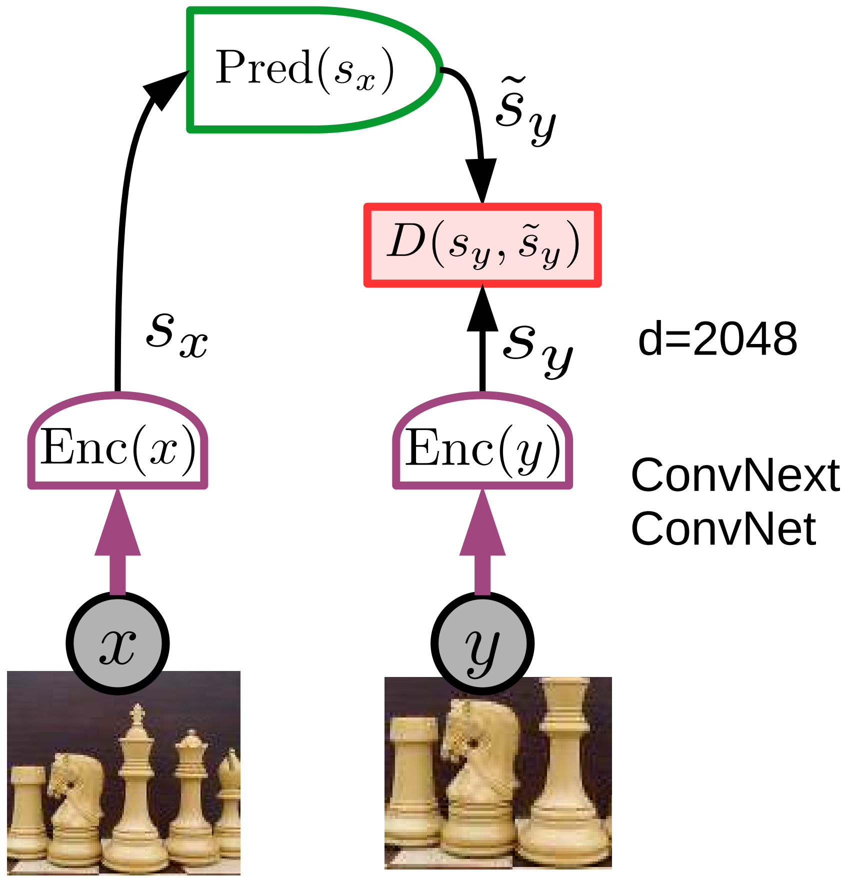


Contrastive Methods vs Regularized/Architectural Methods

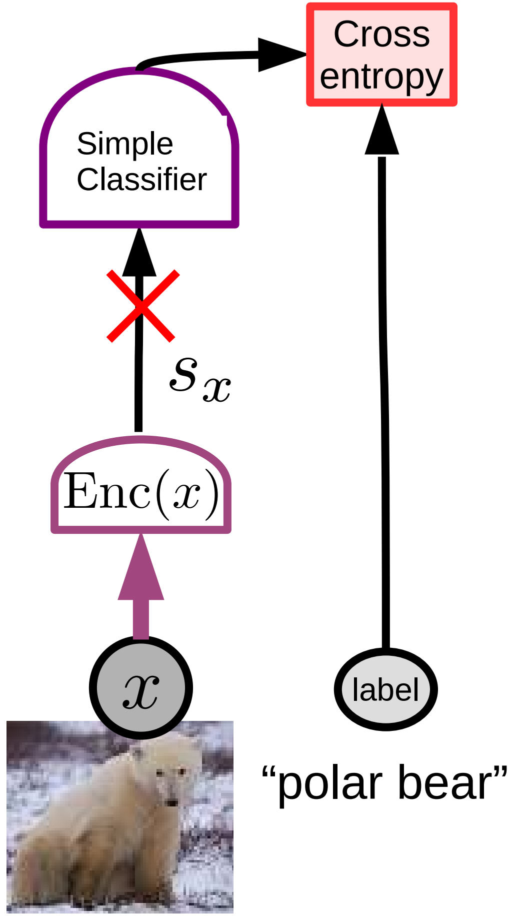
- **Contrastive: [they all are different ways to pick which points to push up]**
  - C1: push down of the energy of data points, push up everywhere else: Max likelihood (needs tractable partition function or variational approximation)
  - C2: push down of the energy of data points, push up on chosen locations: max likelihood with MC/MMC/HMC, Contrastive divergence, Metric learning/Siamese nets, Ratio Matching, Noise Contrastive Estimation, Min Probability Flow, adversarial generator/GANs
  - C3: train a function that maps points off the data manifold to points on the data manifold: denoising auto-encoder, masked auto-encoder (e.g. BERT)
- **Regularized/Architectural: [Different ways to limit the information capacity of the latent representation]**
  - A1: build the machine so that the volume of low energy space is bounded: PCA, K-means, Gaussian Mixture Model, Square ICA, normalizing flows…
  - A2: use a regularization term that measures the volume of space that has low energy: Sparse coding, sparse auto-encoder, LISTA, Variational Auto-Encoders, discretization/VQ/VQVAE.
  - A3: F(x,y) = C(y, G(x,y)), make G(x,y) as "constant" as possible with respect to y: Contracting auto-encoder, saturating auto-encoder
  - A4: minimize the gradient and maximize the curvature around data points: score matching

SSL-Pretrained Joint Embedding for Image Recognition

JEPA/JEA pretrained with SSL



Training a supervised classification head



(Sample) Contrastive Joint Embedding

- **Example:**
  - Siamese Networks
    [Bromley NIPS 1993]
    [Chopra CVPR 2005]
    [Hadsell CVPR 2006]
  - SimCLR
    [Chen 2020]
- **Can only produce low-dimensional image representations**
  - Around 200 D.