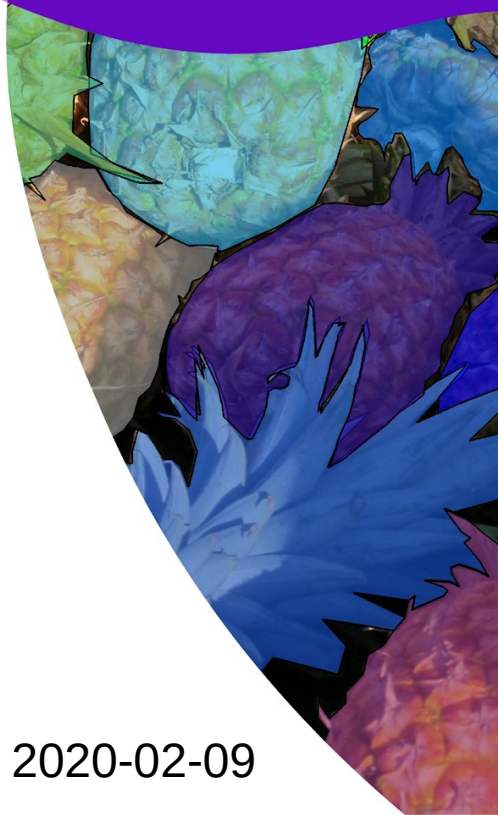
FACEBOOK AI NEW YORK UNIVERSITY

Self-Supervised Learning

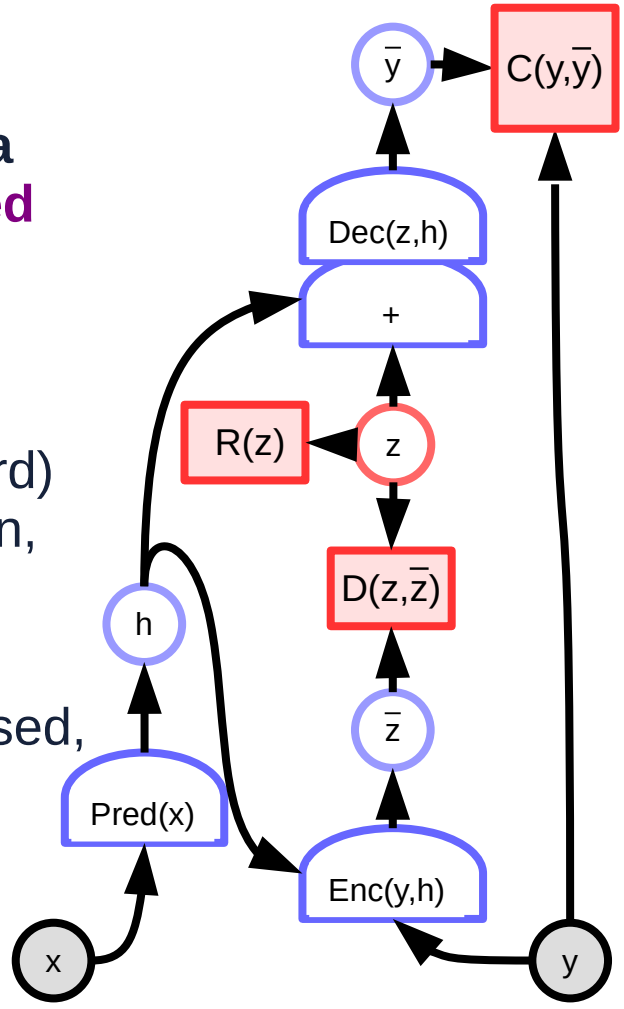
Yann LeCun
NYU - Courant Institute & Center for Data Science
Facebook AI Research
http://yann.lecun.com



AAAI - 2020-02-09

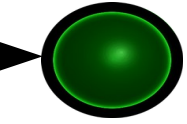
What is Deep Learning?

- **Definition: Deep Learning is building a system by assembling parameterized modules into a (possibly dynamic) computation graph, and training it to perform a task by optimizing the parameters using a gradient-based method.**
  - Graph can be defined dynamically by input-dependent programs: differentiable programming
  - Output may be computed through complex (non feed-forward) process, e.g. by minimizing some energy function: relaxation, constraint satisfaction, structured prediction,….
  - Learning paradigms and objective functions are up to the designer: supervised, reinforced, self-supervised/unsupervised, classification, prediction, reconstruction,….
  - Note: the limitations of Supervised Learning are sometimes mistakenly seen as intrinsic limitations of DL

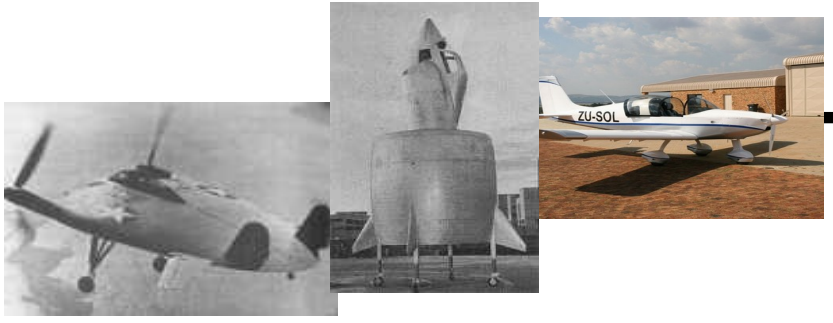


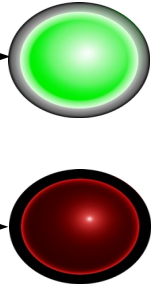
Supervised Learning works but requires many labeled samples

- **Training a machine by showing examples instead of programming it**
- **When the output is wrong, tweak the parameters of the machine**
- **Works well for:**
  - Speech → words
  - Image → categories
  - Portrait → name
  - Photo → caption
  - Text → topic
  - ….



CAR





PLANE

Supervised DL works amazingly well, when you have data

- **And services like Facebook, Instagram, Google, Youtube,… are built around it.**
  - Content understanding, filtering, ranking, translation, accessibility….
  - Content understanding, filtering, ranking, translation, accessibility….

Supervised Symbol Manipulation

- **Solving integrals and differential equations symbolically with a transformer architecture**
  - [Lample & Charton arXiv:1912.01412]
  - Accuracy on various problems →

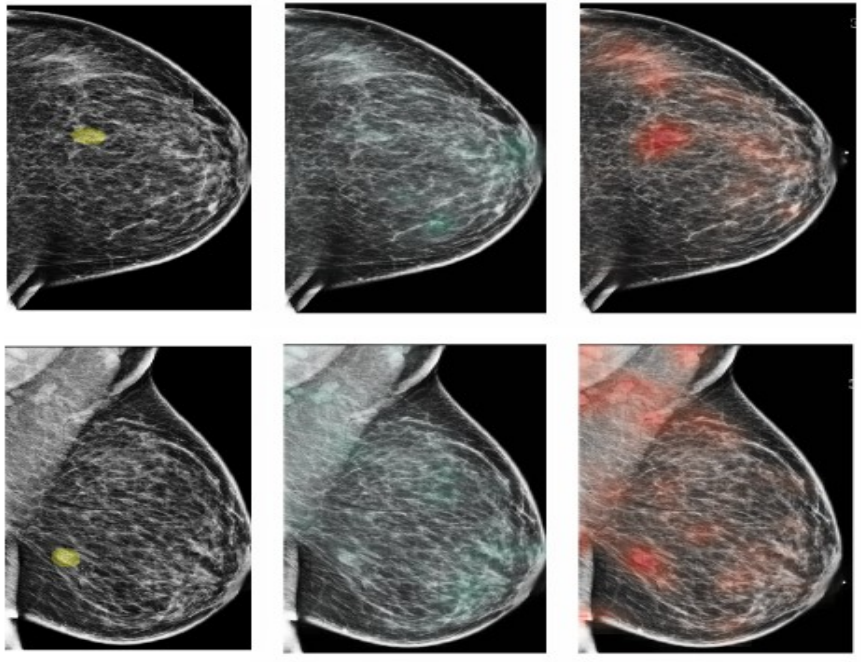
	Integration (BWD)	ODE (order 1)	ODE (order 2)
Mathematica (30s)	84.0	77.2	61.6
Matlab	65.2	-	-
Maple	67.4	-	-
Beam size 1	98.4	81.2	40.8
Beam size 10	99.6	94.0	73.2
Beam size 50	99.6	97.0	81.0

Equation	Solution
$y' = \frac{16x^3 - 42x^2 + 2x}{(-16x^8 + 112x^7 - 204x^6 + 28x^5 - x^4 + 1)^{1/2}}$	$y = \sin^{-1}(4x^4 - 14x^3 + x^2)$
$3xy\cos(x) - \sqrt{9x^2\sin(x)^2 + 1}y' + 3y\sin(x) = 0$	$y = c\exp\left(\sinh^{-1}(3x\sin(x))\right)$
$4x^4yy'' - 8x^4y'^2 - 8x^3yy' - 3x^3y'' - 8x^2y^2 - 6x^2y' - 3x^2y'' - 9xy' - 3y = 0$	$y = \frac{c_1 + 3x + 3\log(x)}{x(c_2 + 4x)}$

Deep Learning Saves Lives

- **Automated emergency Braking Systems**
  - Reduce collisions by 40%
  - Use Convolutional nets.

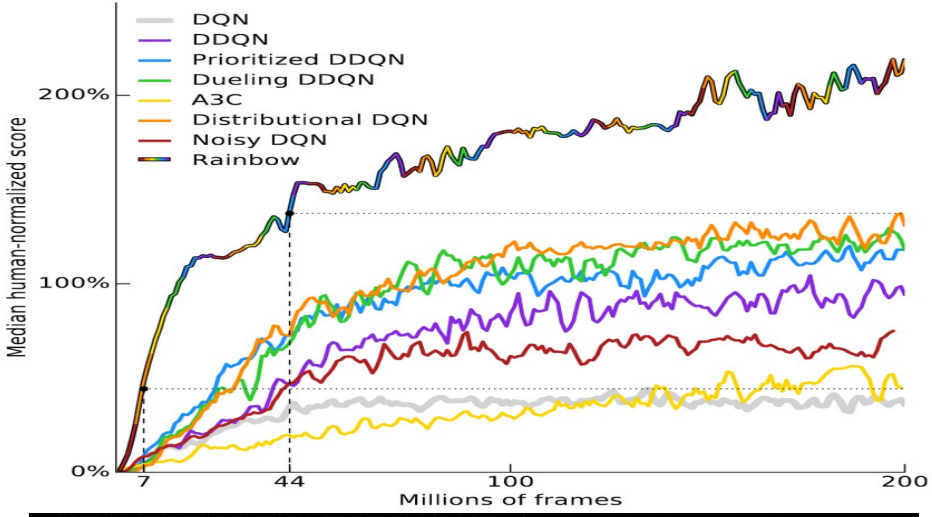
- **Tumor detection in mammograms**
  - [Wu et al. ArXiv:1903.08297] https://github.com/nyukat/breast_cancer_classifier

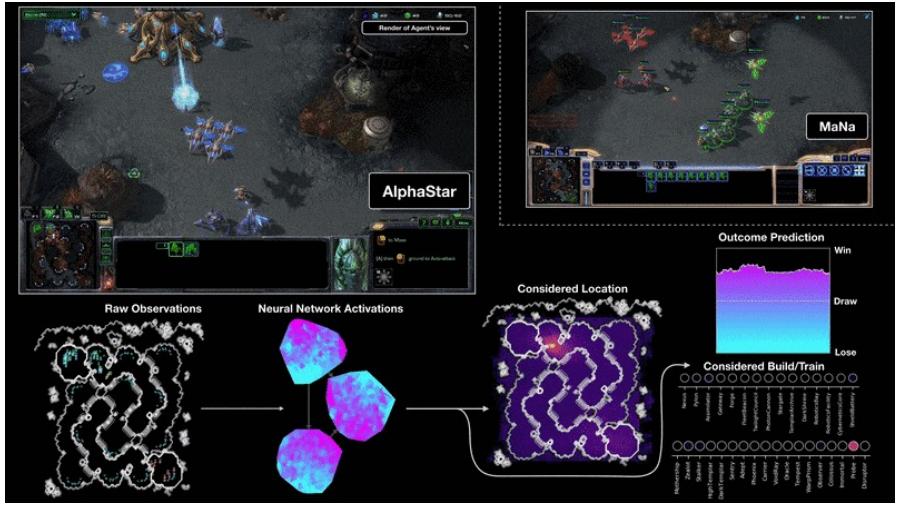


- **Content filtering.**
  - Hate speech, calls to violence, weapon sales, terrorist propaganda….

Reinforcement Learning: works great for games and simulations.

- **57 Atari games: takes 83 hours equivalent real-time (18 million frames) to reach a performance that humans reach in 15 minutes of play.**
  - [Hessel ArXiv:1710.02298]
- **Elf OpenGo v2: 20 million self-play games. (2000 GPU for 14 days)**
  - [Tian arXiv:1902.04522]
- **StarCraft: AlphaStar 200 years of equivalent real-time play**
  - [Vinyals blog post 2019]
- **OpenAI single-handed Rubik's cube**
  - 10,000 years of simulation





But RL Requires too many trials in the real world

- **Pure RL requires too many trials to learn anything**
  - it's OK in a game
  - it's not OK in the real world
- **RL works in simple virtual world that you can run faster than real-time on many machines in parallel.**

- **Anything you do in the real world can kill you**
- **You can't run the real world faster than real time**

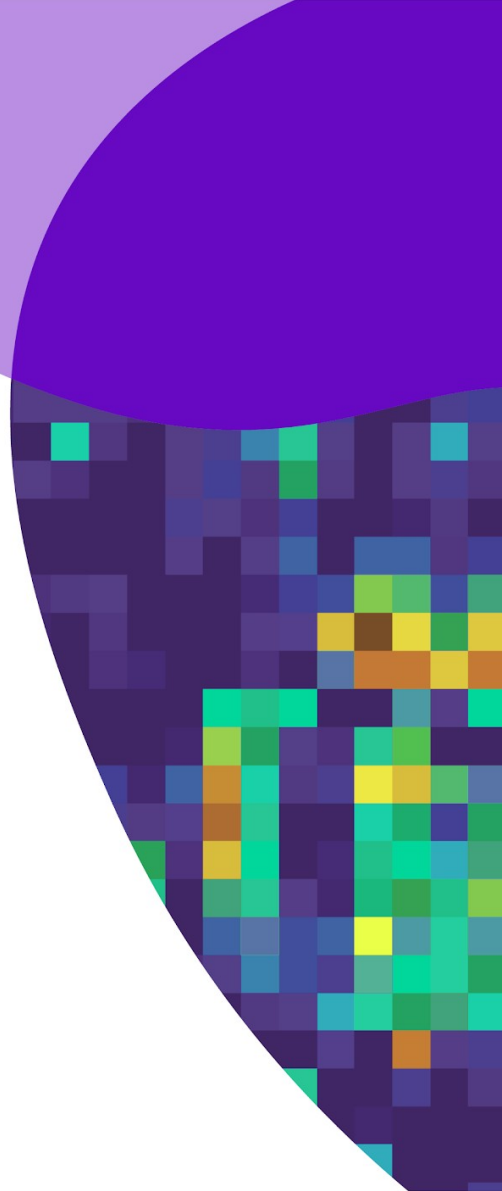
Three challenges for Deep Learning

- **Deep Supervised Learning works well for perception**
  - When labeled data is abundant.
- **Deep Reinforcement Learning works well for action generation**
  - When trials are cheap, e.g. in simulation.
- **Three problems the community is working on:**
- **1. Learning with fewer labeled samples and/or fewer trials**
  - Self-supervised learning / unsup learning / learning to fill in the blanks
    - learning to represent the world before learning tasks
- **2. Learning to reason,** beyond “system 1” feed-forward computation.
  - Making reasoning compatible with gradient-based learning.
- **3. Learning to plan complex action sequences**
  - Learning hierarchical representations of action plans

FACEBOOK AI

How do humans and animals learn so quickly?

Not supervised.
Not Reinforced.

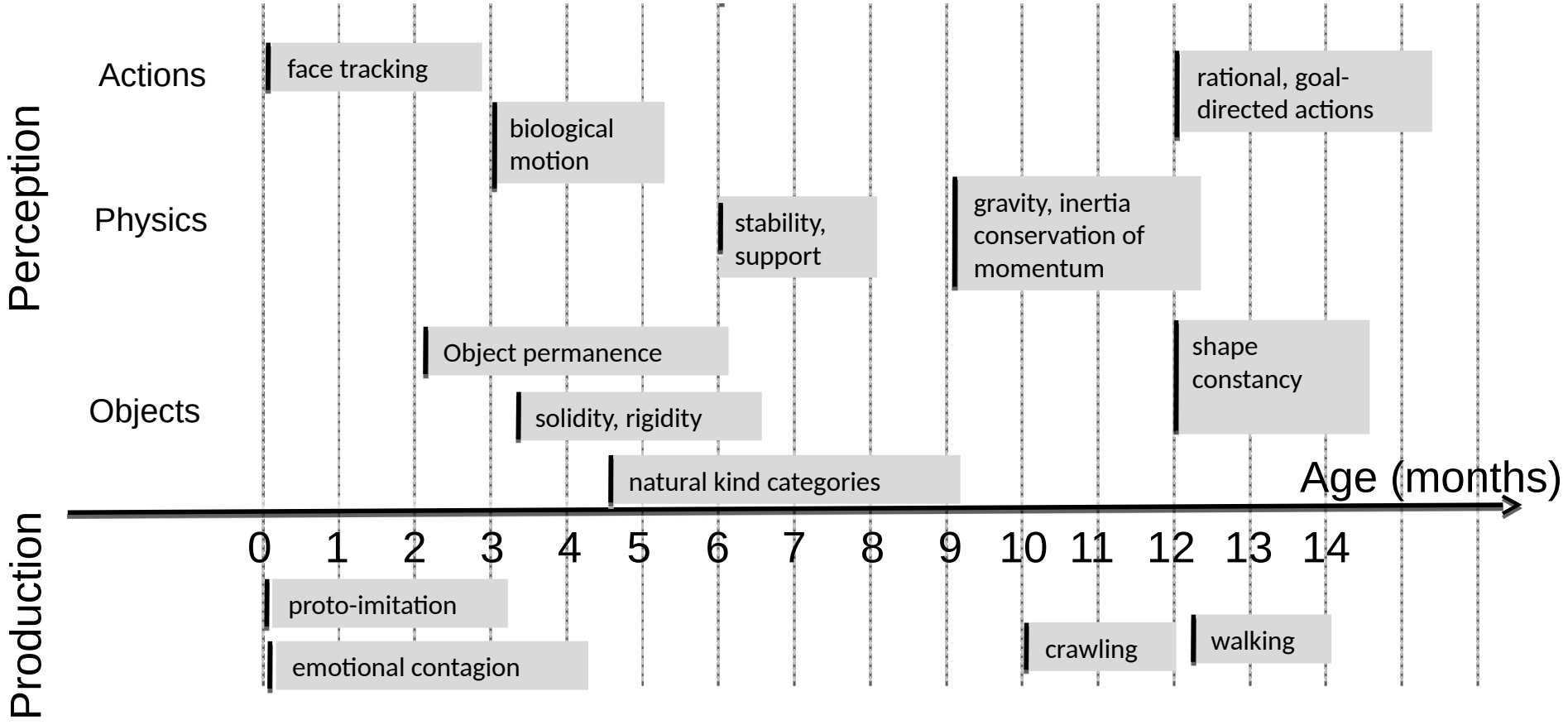


Babies learn how the world works by observation

- **Largely by observation, with remarkably little interaction.**

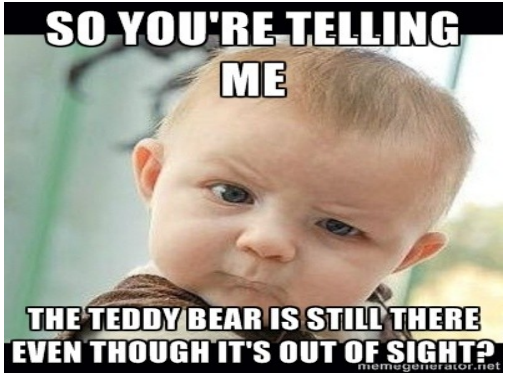
Photos courtesy of Emmanuel Dupoux

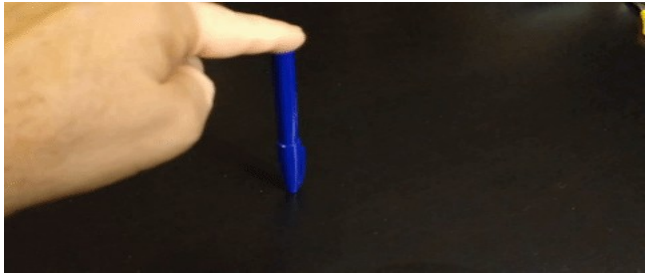
Early Conceptual Acquisition in Infants [from Emmanuel Dupoux]

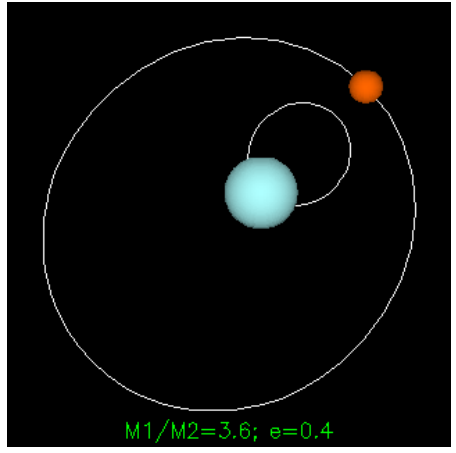


Prediction is the essence of Intelligence

- **We learn models of the world by predicting**



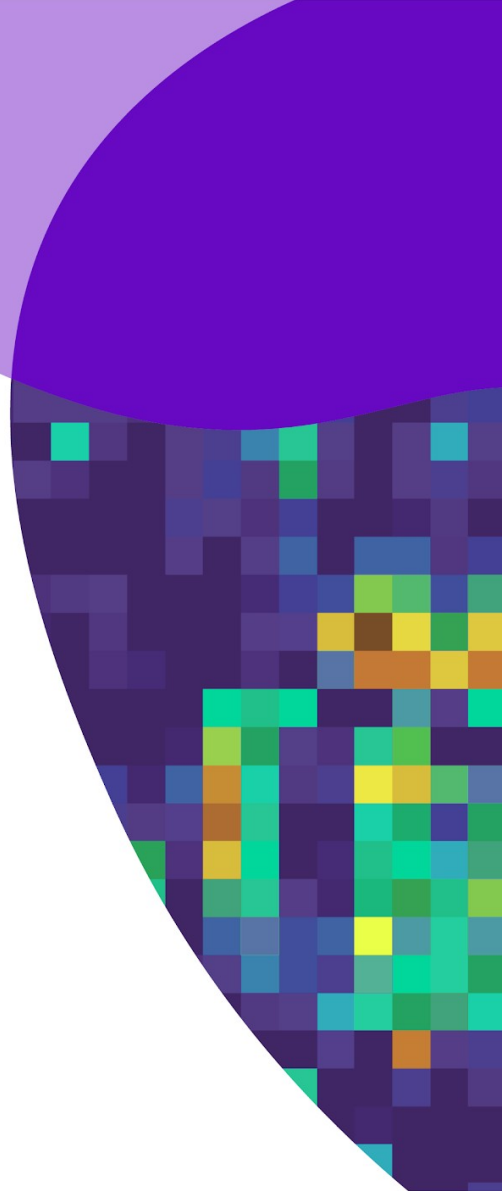




FACEBOOK AI

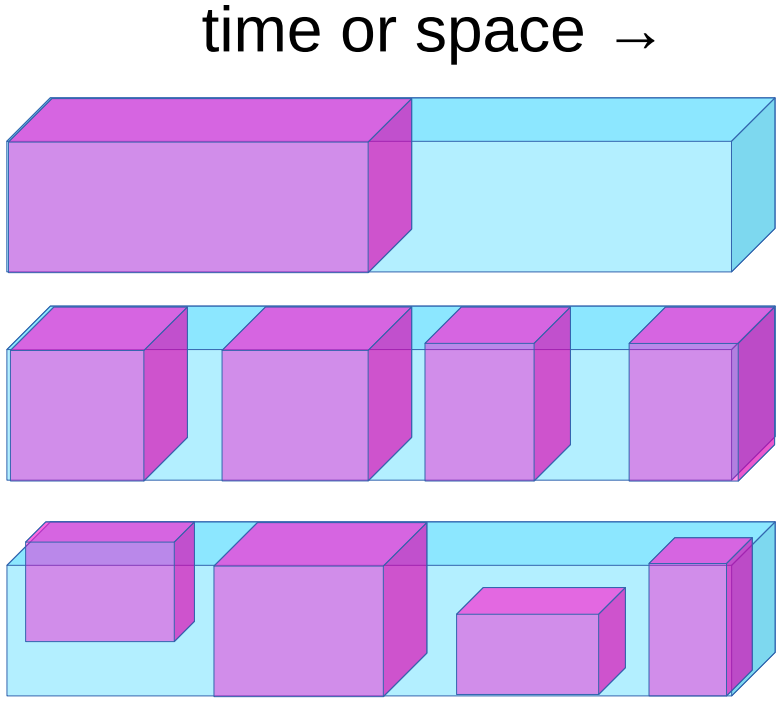
Self-Supervised Learning

Predict everything
from everything else



Self-Supervised Learning = Filling in the Blanks

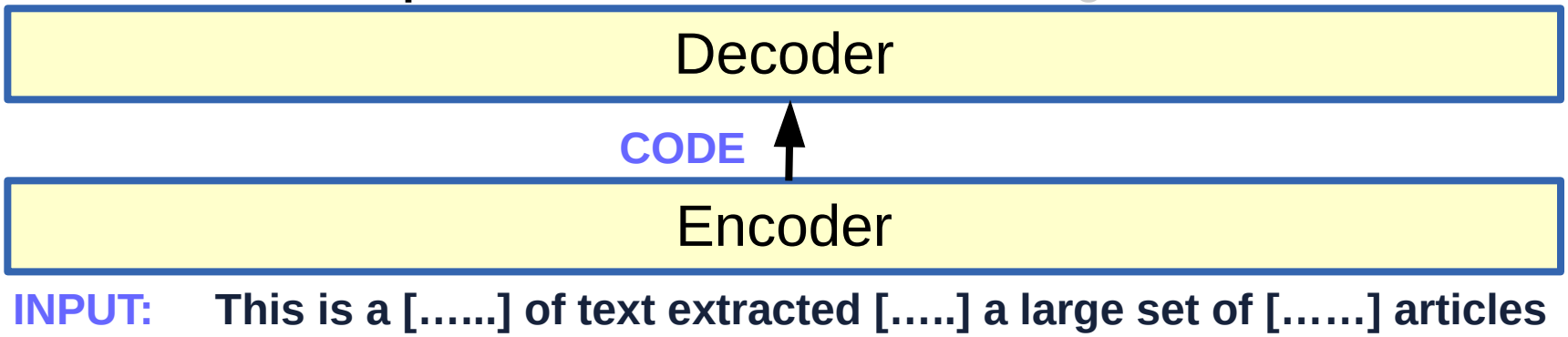
- **Predict any part of the input from any other part.**
- **Predict the future from the past.**
- **Predict the masked from the visible.**
- **Predict the any occluded part from all available parts.**



- **Pretend there is a part of the input you don't know and predict that.**
- **Reconstruction = SSL when any part could be known or unknown**

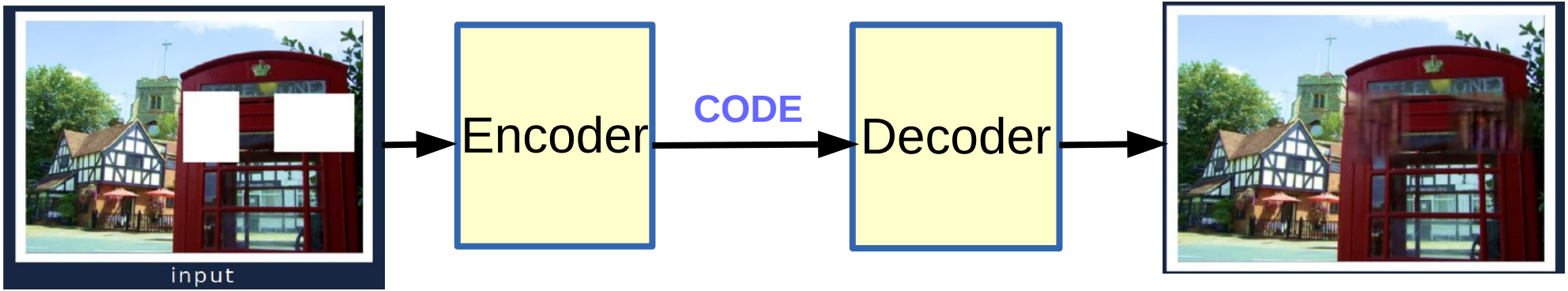
Self-Supervised Learning: filling in the bl_nks

- **Natural Language Processing: works great!**



- **Image Recognition / Understanding: works so-so**

[Pathak et al 2014]

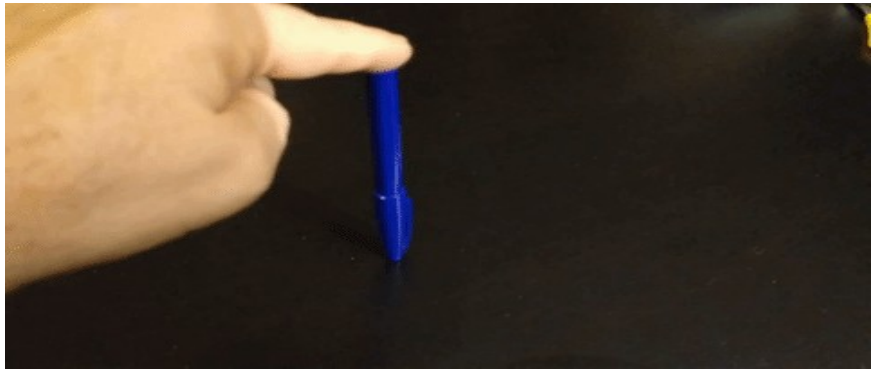


Self-Supervised Learning for Video Prediction

- The world is not entirely predictable
- There are many plausible continuations to a video segment

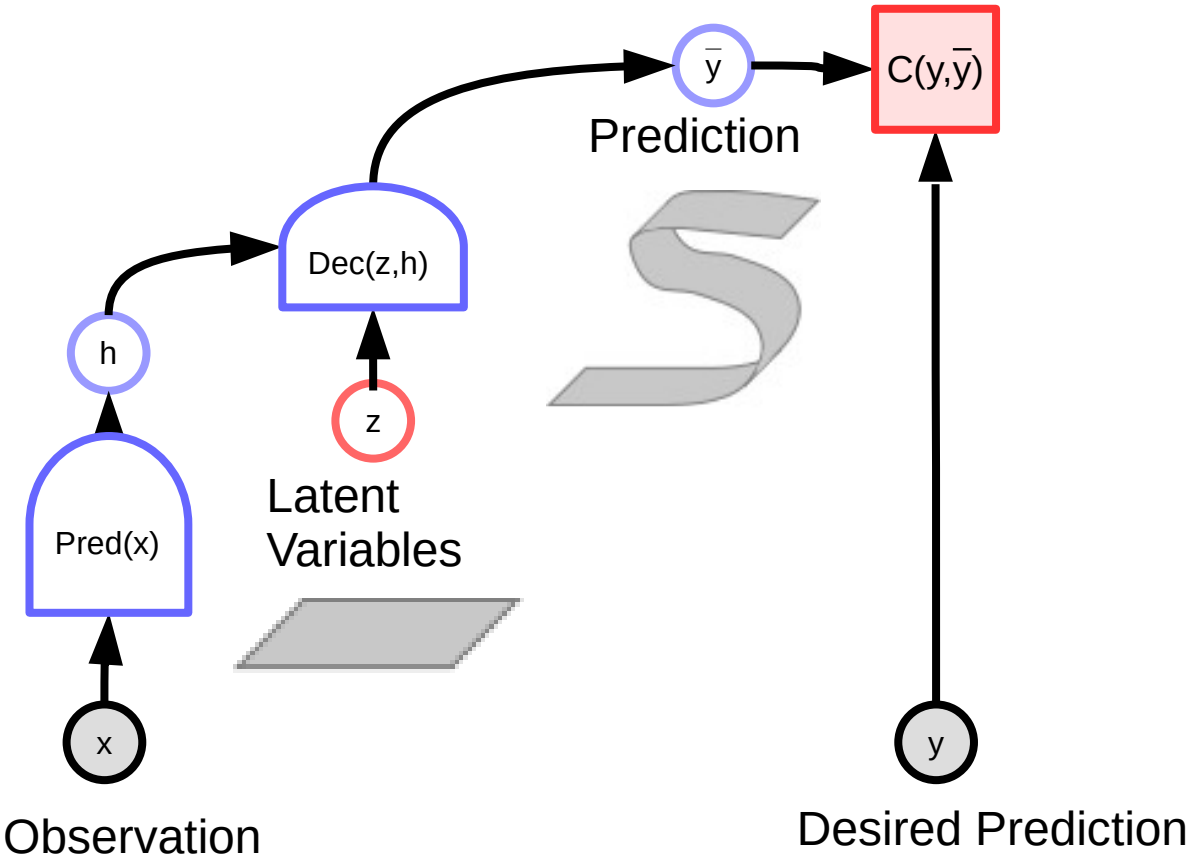
The world is stochastic

- Training a system to make a single prediction makes it predict the average of all plausible predictions
- Blurry predictions!



Solution: latent variable energy-based models

- **Latent variables allows system to make multiple predictions**



Self-supervised Adversarial Learning for Video Prediction

- **Our brains are “prediction machines”**
- **Can we train machines to predict the future?**
- **Some success with “adversarial training”**
  - [Mathieu, Couprie, LeCun arXiv:1511:05440]
- **But we are far from a complete solution.**

Three Types of Learning

- **Reinforcement Learning**
  - The machine predicts a scalar reward given once in a while.
  - **weak feedback**

- **Supervised Learning**
  - The machine predicts a category or a few numbers for each input
  - **medium feedback**

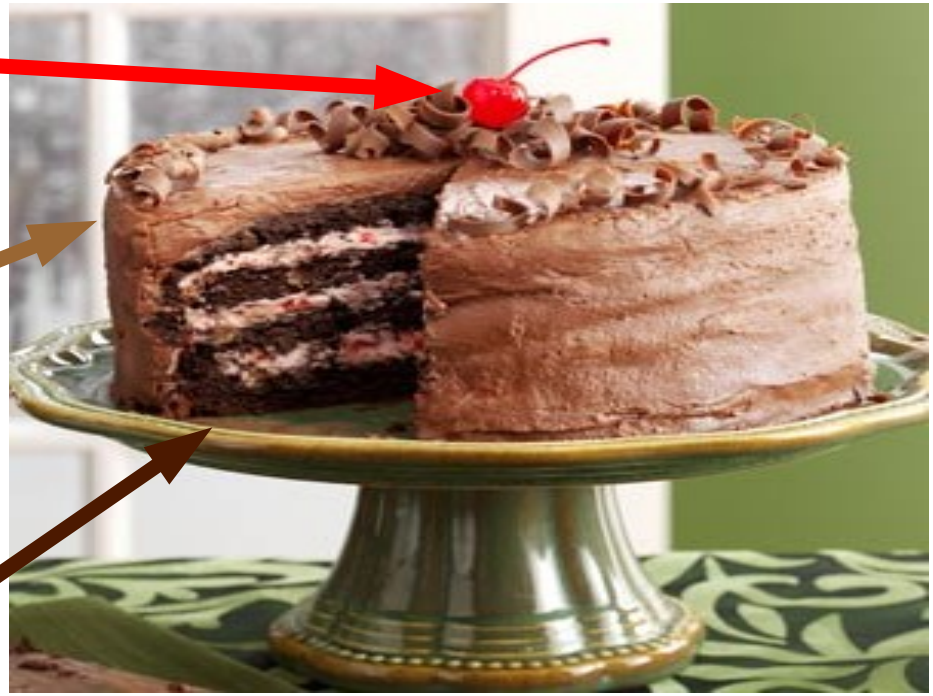
PLANE

CAR

- **Self-supervised Learning**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **A lot of feedback**

How Much Information is the Machine Given during Learning?

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**



The Next AI Revolution

With thanks to Alyosha Efros and Gil Scott Heron

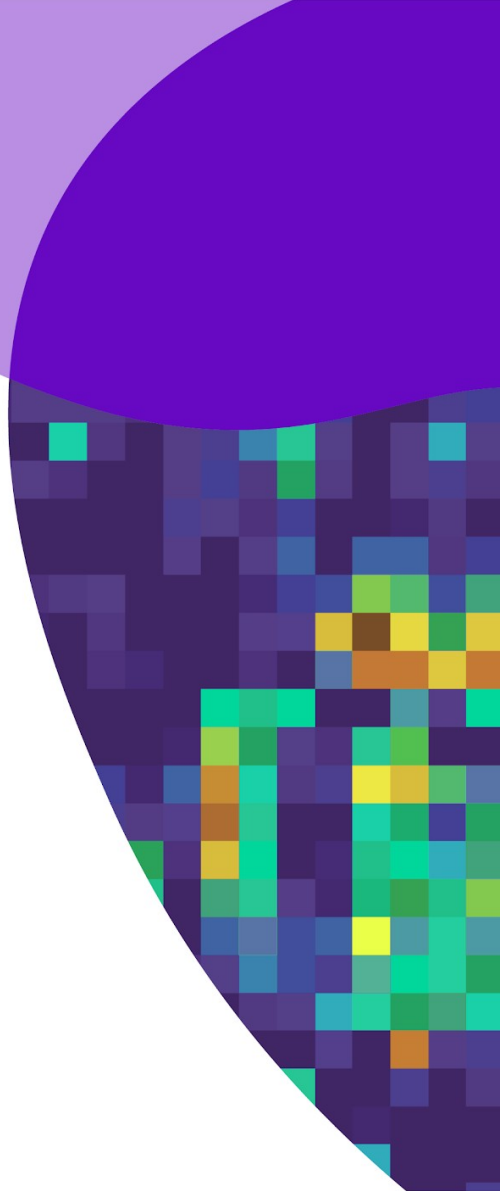
Get the T-shirt!

Jitendra Malik: “Labels are the opium of the machine learning researcher”

FACEBOOK AI

Energy-Based Models

Learning to deal with uncertainty while eschewing probabilities

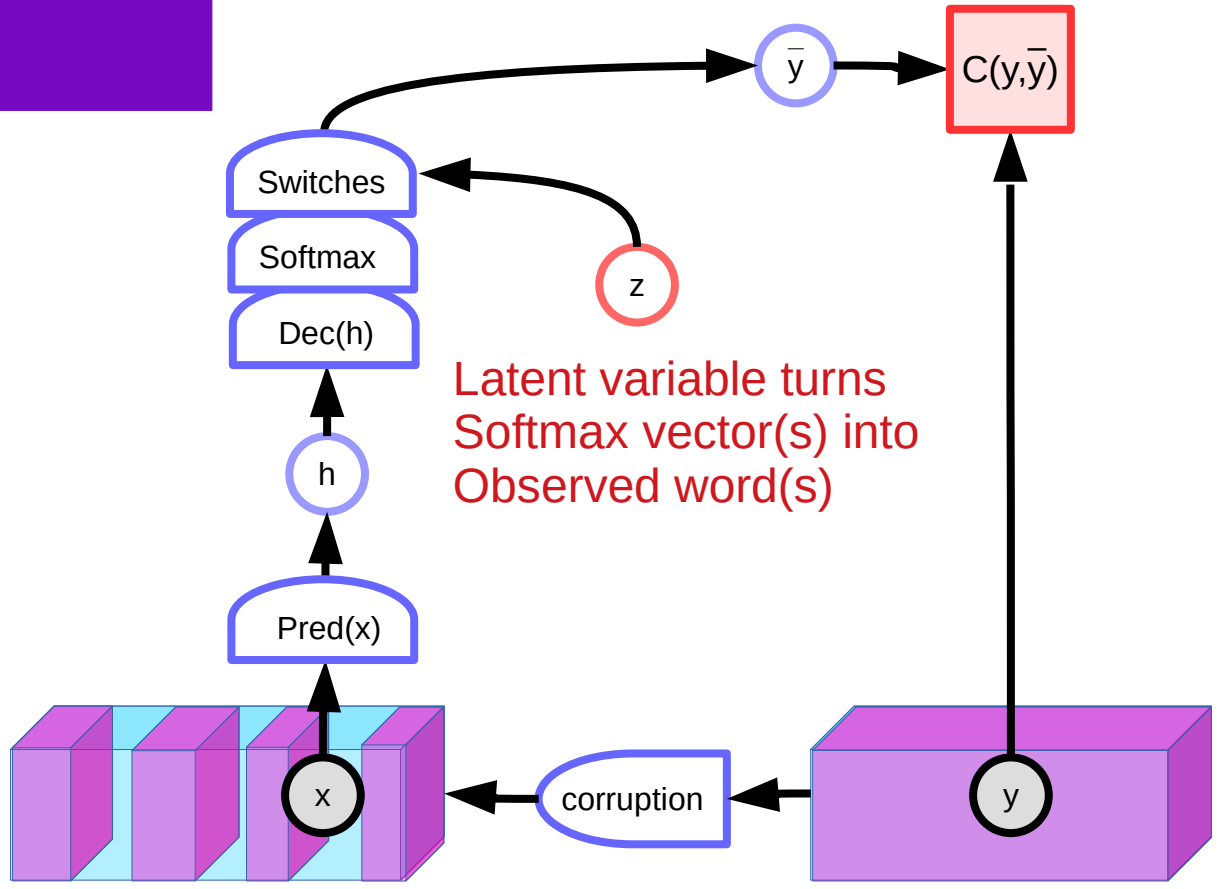


Seven Strategies to Shape the Energy Function

- **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, 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)
- **Architectural: [they all are different ways to limit the information capacity of the code]**
  - A1: build the machine so that the volume of low energy stuff is bounded: PCA, K-means, Gaussian Mixture Model, Square ICA…
  - A2: use a regularization term that measures the volume of space that has low energy: Sparse coding, sparse auto-encoder, LISTA, Variational auto-encoders
  - 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

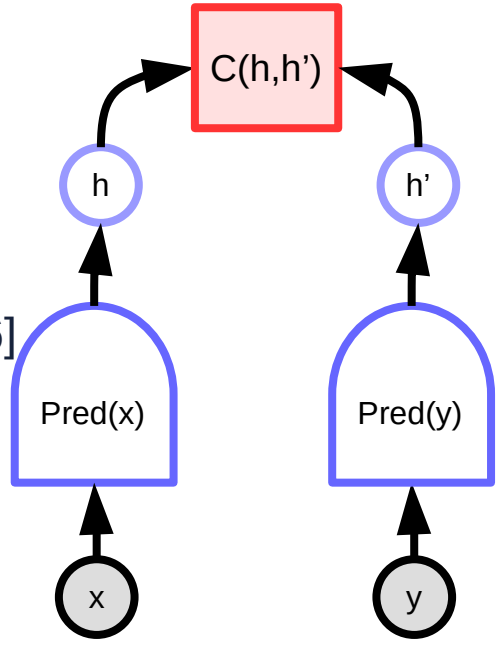
Denoising AE: discrete

- **[Vincent et al. JMLR 2008]**
- **Masked Auto-Encoder**
  - [BERT et al.]
- **Issues:**
  - latent variables are in output space
  - No abstract LV to control the output
  - How to cover the space of corruptions?



Contrastive Embedding

- **Distance measured in feature space**
- **Multiple “predictions” through feature invariance**
- **Siamese nets, metric learning** [YLC NIPS’93,CVPR’05,CVPR’06]
- **Advantage: no pixel-level reconstruction**
- **Difficulty: hard negative mining**
- **Successful examples for images:**
  - DeepFace [Taigman et al. CVPR’14]
  - PIRL [Misra et al. Arxiv:1912.01991]
  - MoCo [He et al. Arxiv:1911.05722]
- **Video / Audio**
  - Temporal proximity [Taylor CVPR’11]
  - Slow feature [Goroshin NIPS’15]

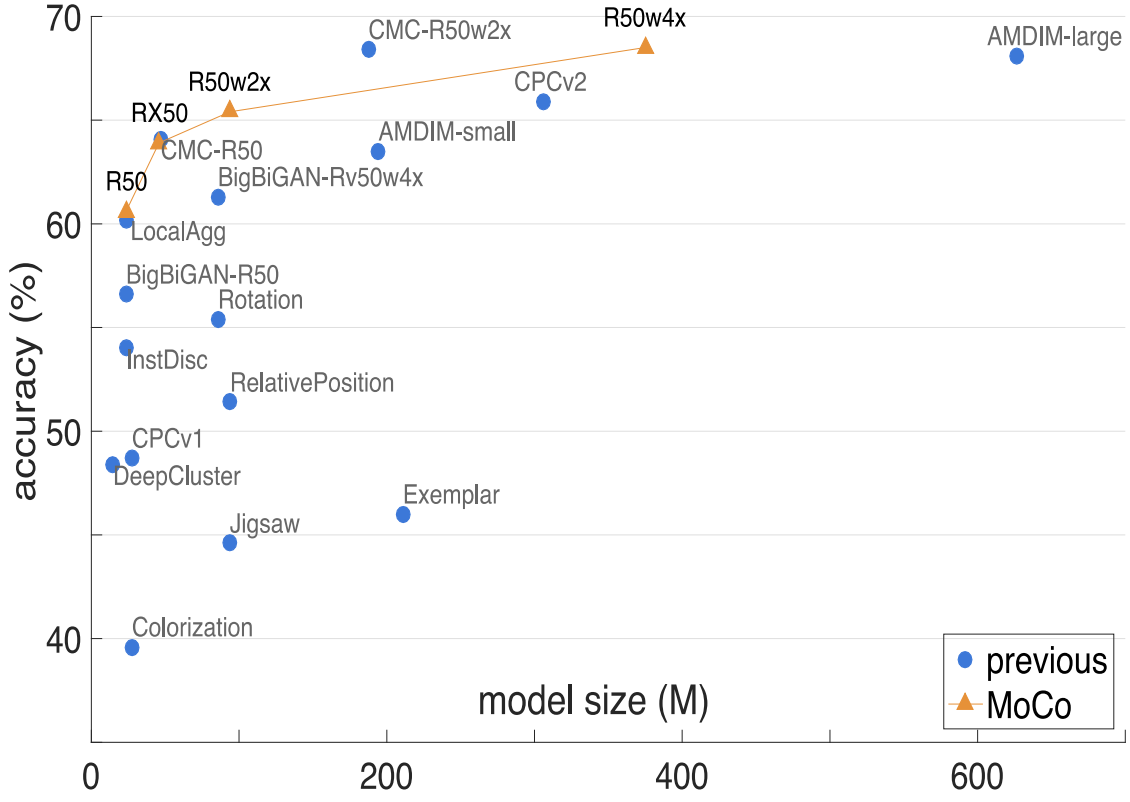


Positive pair:
Make F small

Negative pair:
Make F large

MoCo on ImageNet

MoCo



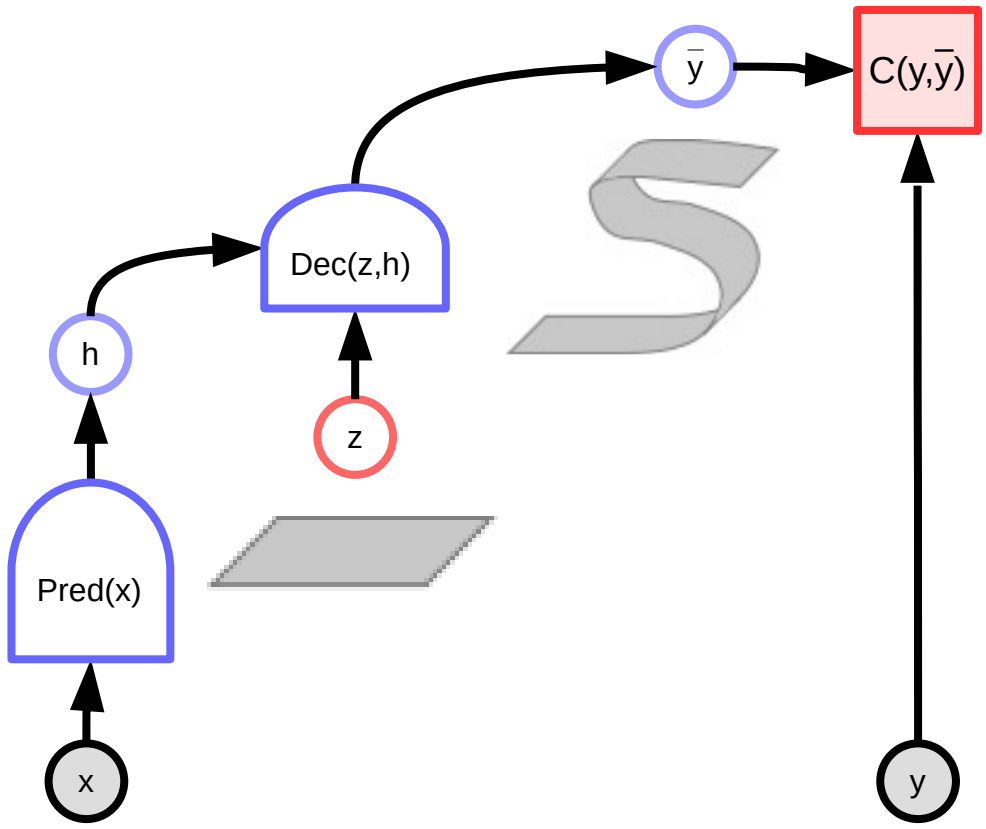
PIRL

Method	Data fraction → Backbone	1% Top-5 Accuracy	10% Top-5 Accuracy
Random initialization [66]	R-50	22.0	59.0
NPID [66]	R-50	39.2	77.4
Jigsaw [18]	R-50	45.3	79.3
NPID++ [66]	R-50	52.6	81.5
S^4L Exemplar [68]	R-50v2	47.0	83.7
S^4L Rotation [68]	R-50v2	53.4	**83.8**
PIRL (**ours**)	R-50	**57.2**	**83.8**
Colorization [31]	R-152	29.8	62.0
CPC-Largest [22]	R-170 and R-11	64.0	84.9

Latent-Variable EBM for inference & multimodal prediction

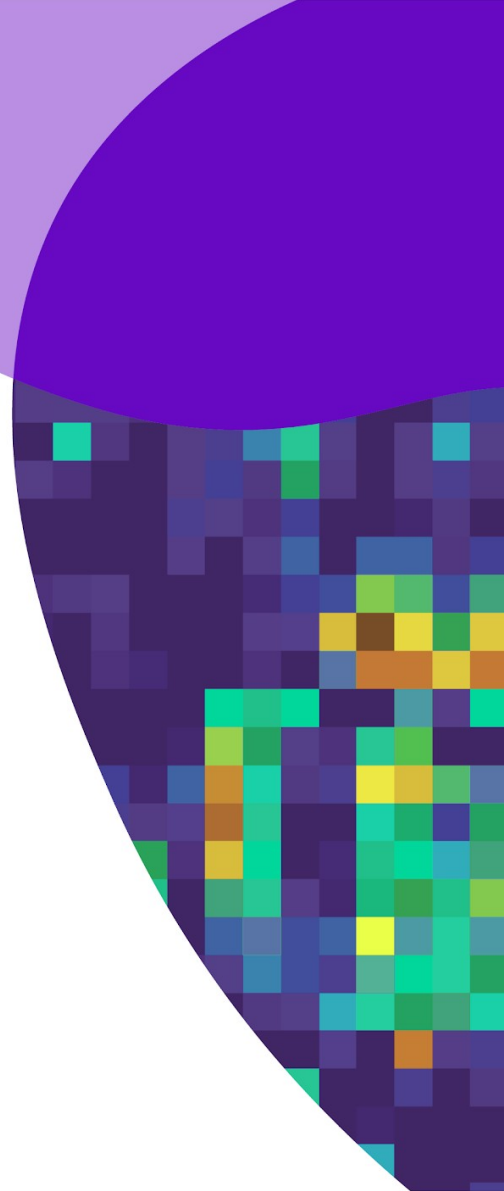
- **Allowing multiple predictions through a latent variable**
- **Allowing inference through minimization**
  - Structured prediction
- **As z varies over a set, y varies over the manifold of possible predictions**

$$F(x,y) = min_z E(x,y,z)$$



FACEBOOK AI

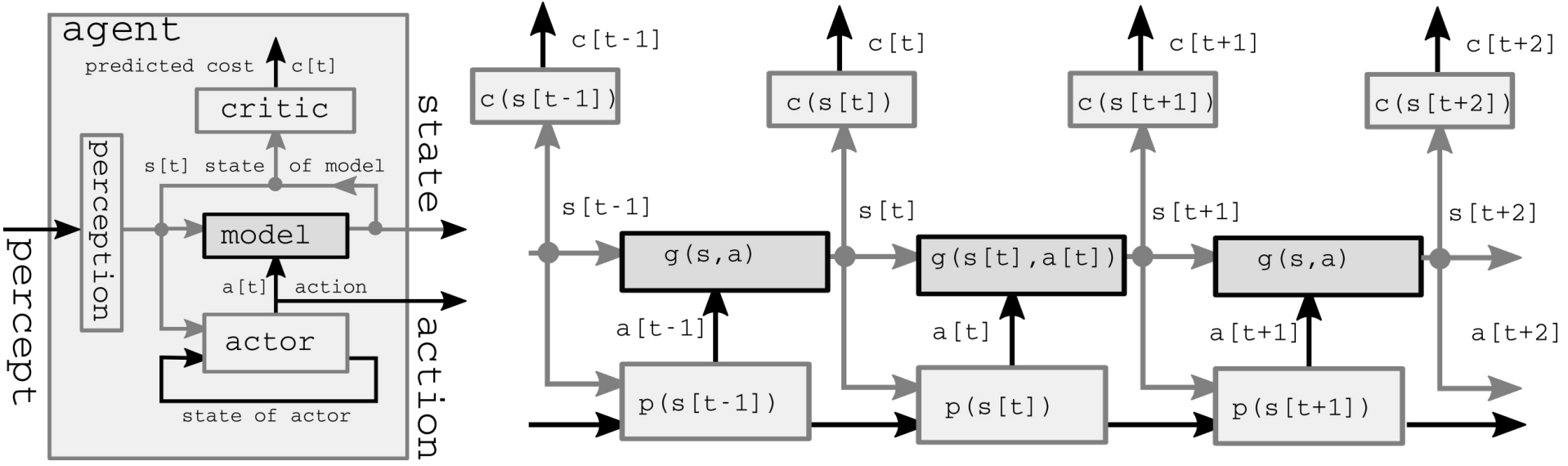
Learning a (stochastic) Forward Model for Autonomous Driving



Learning to predict what others around you will do

A Forward Model of the World

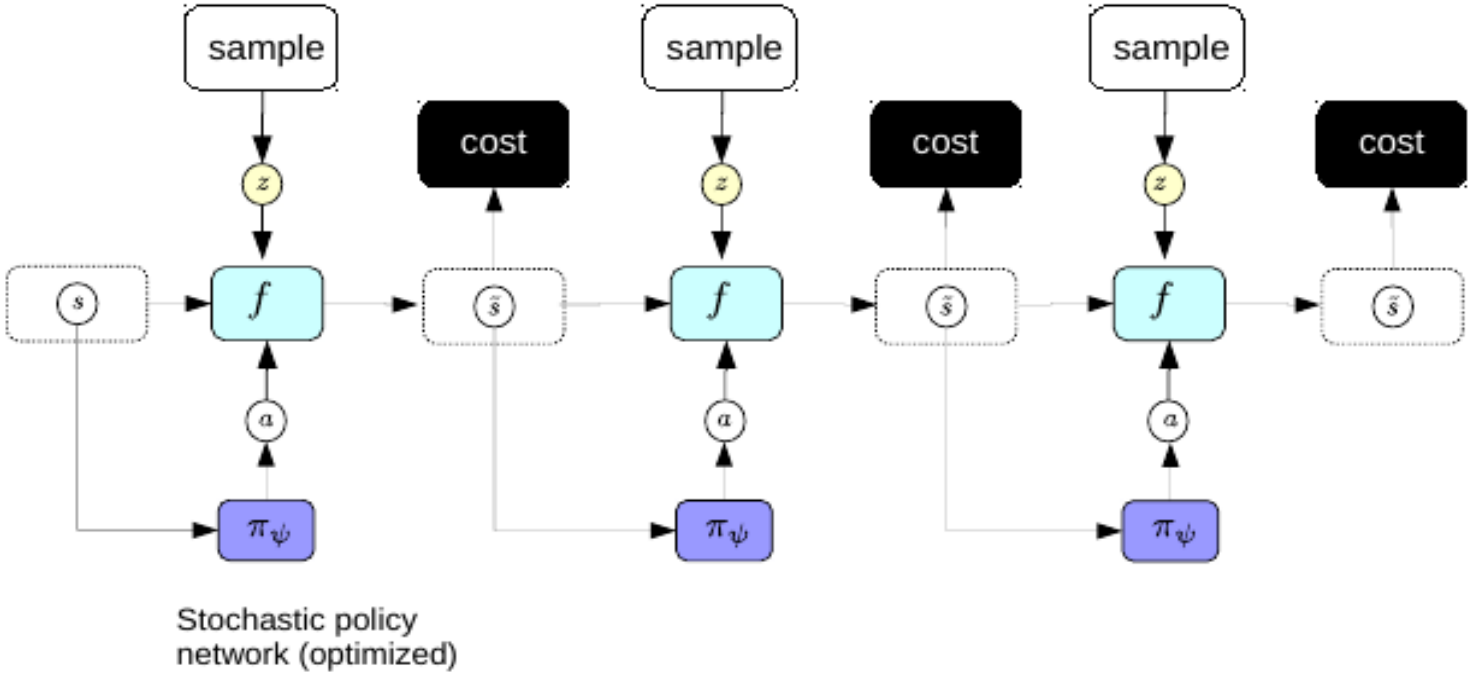
- **Learning forward models for control**
  - s[t+1] = g(s[t], a[t], z[t])
  - Classical optimal control: find a sequence of action that minimize the cost, according to the predictions of the forward model



Planning/learning using a self-supervised predictive world model

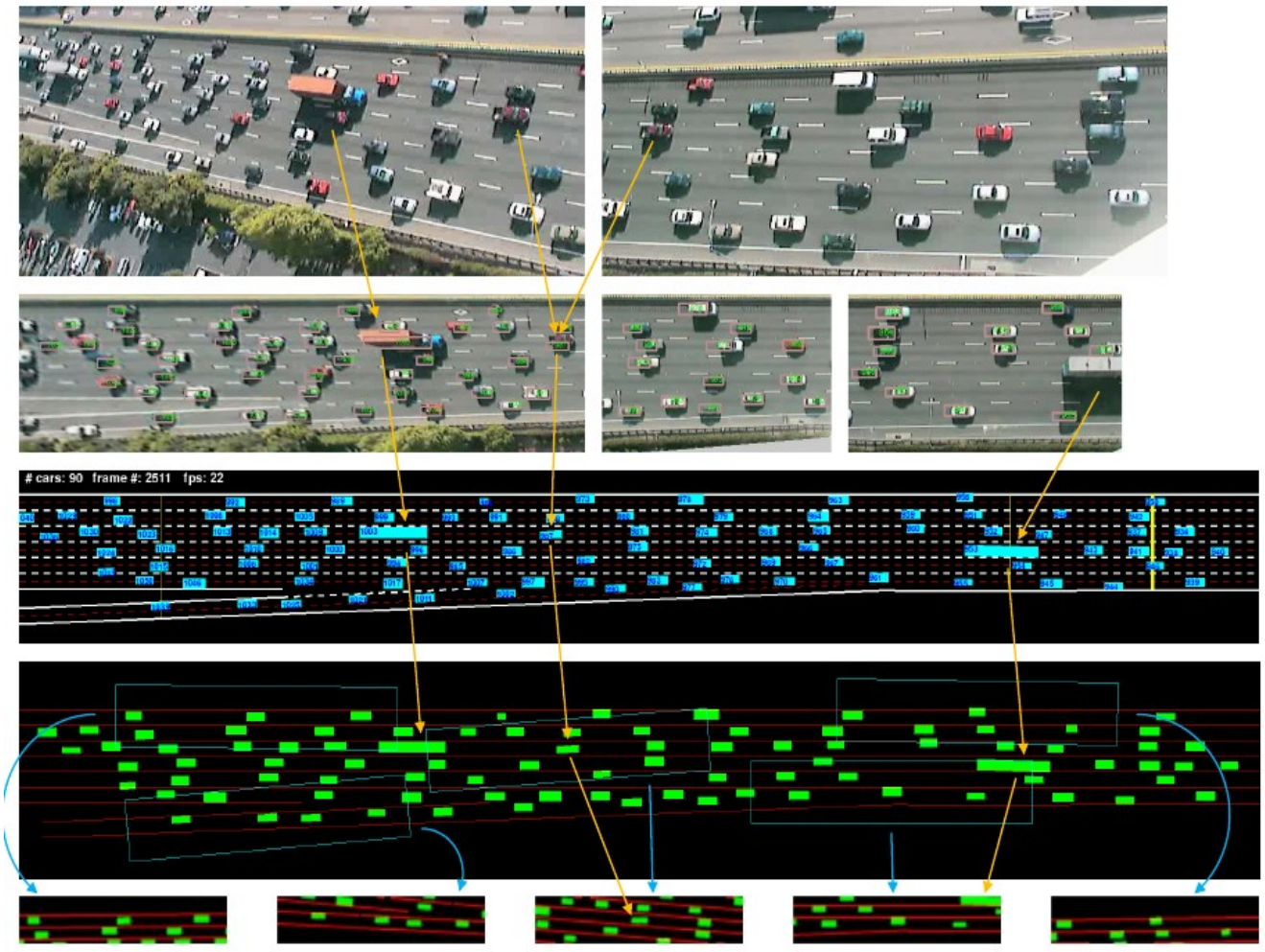
- **Feed initial state**
- **Run the forward model**
- **Backpropagate gradient of cost**
- **Act**
  - (model-predictive control)

  or
- **Use the gradient to train a policy network.**
- **Iterate**



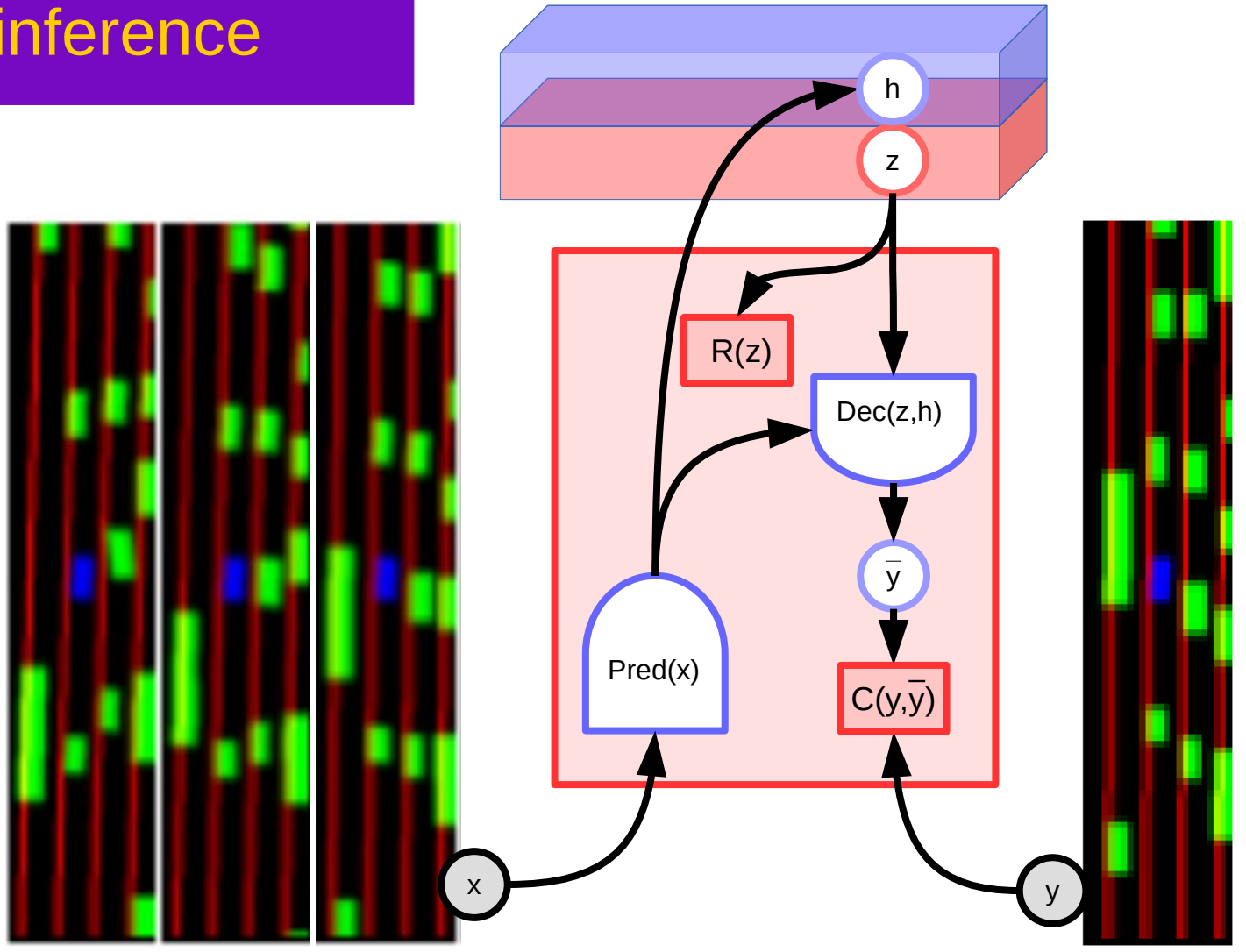
Using Forward Models to Plan (and to learn to drive)

- **Overhead camera on highway.**
  - Vehicles are tracked
- **A "state" is a pixel representation of a rectangular window centered around each car.**
- **Forward model is trained to predict how every car moves relative to the central car.**
  - steering and acceleration are computed



Video Prediction: inference

- **After training:**
  - Observe frames
  - Compute h
  - Sample z
  - Predict next frame

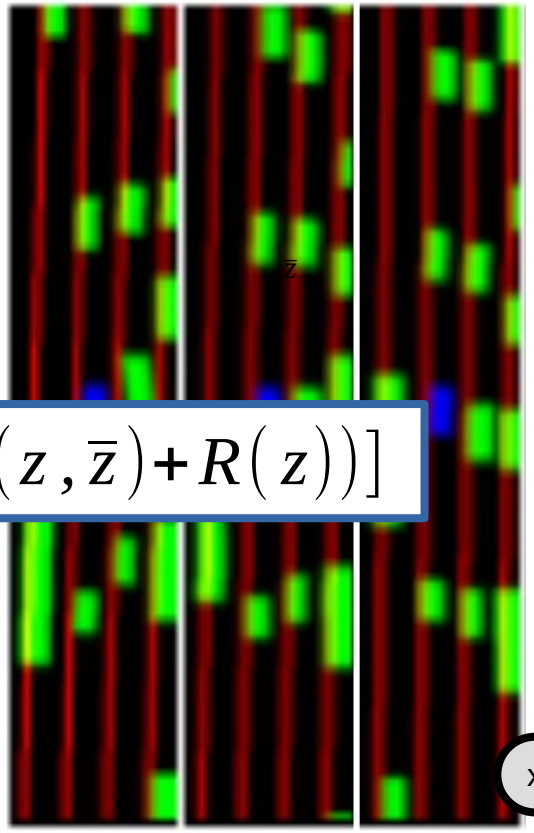


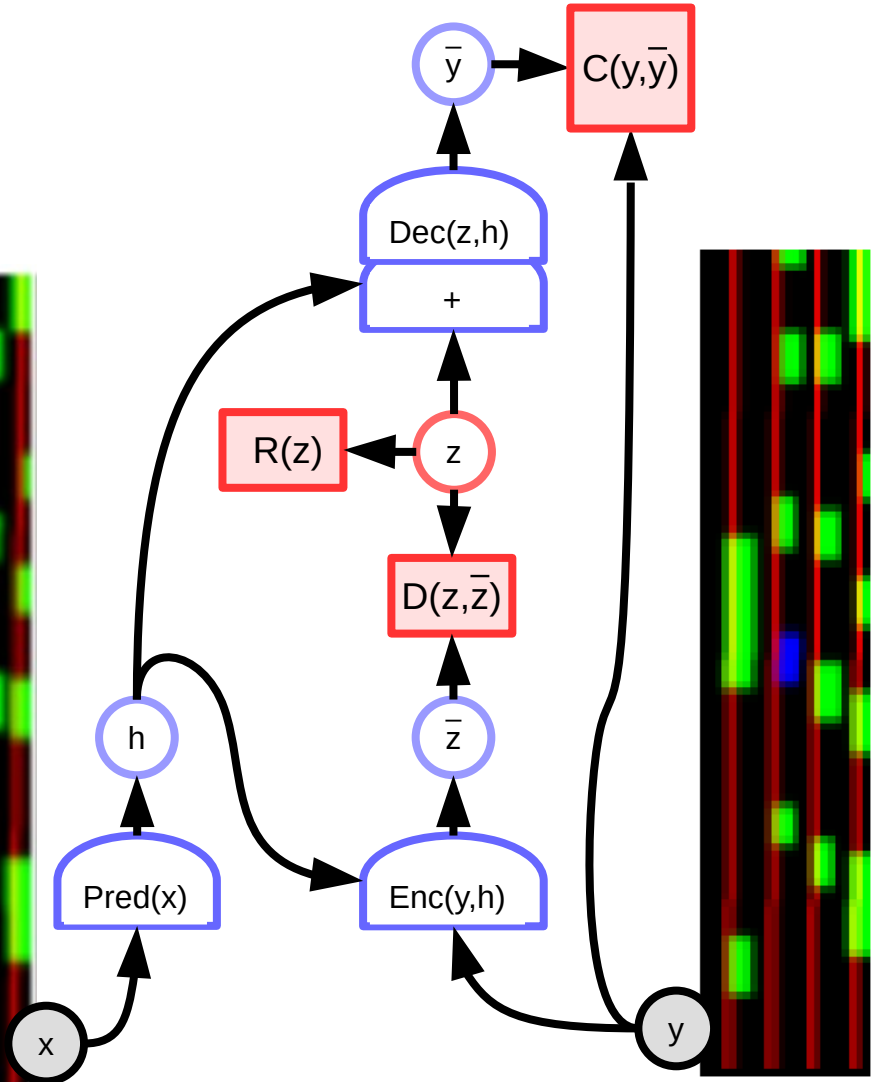
Video Prediction: training

- **Training:**
  - Observe frames
  - Compute h
  - Predict $\bar{z}$ from encoder
  - Sample z, with:

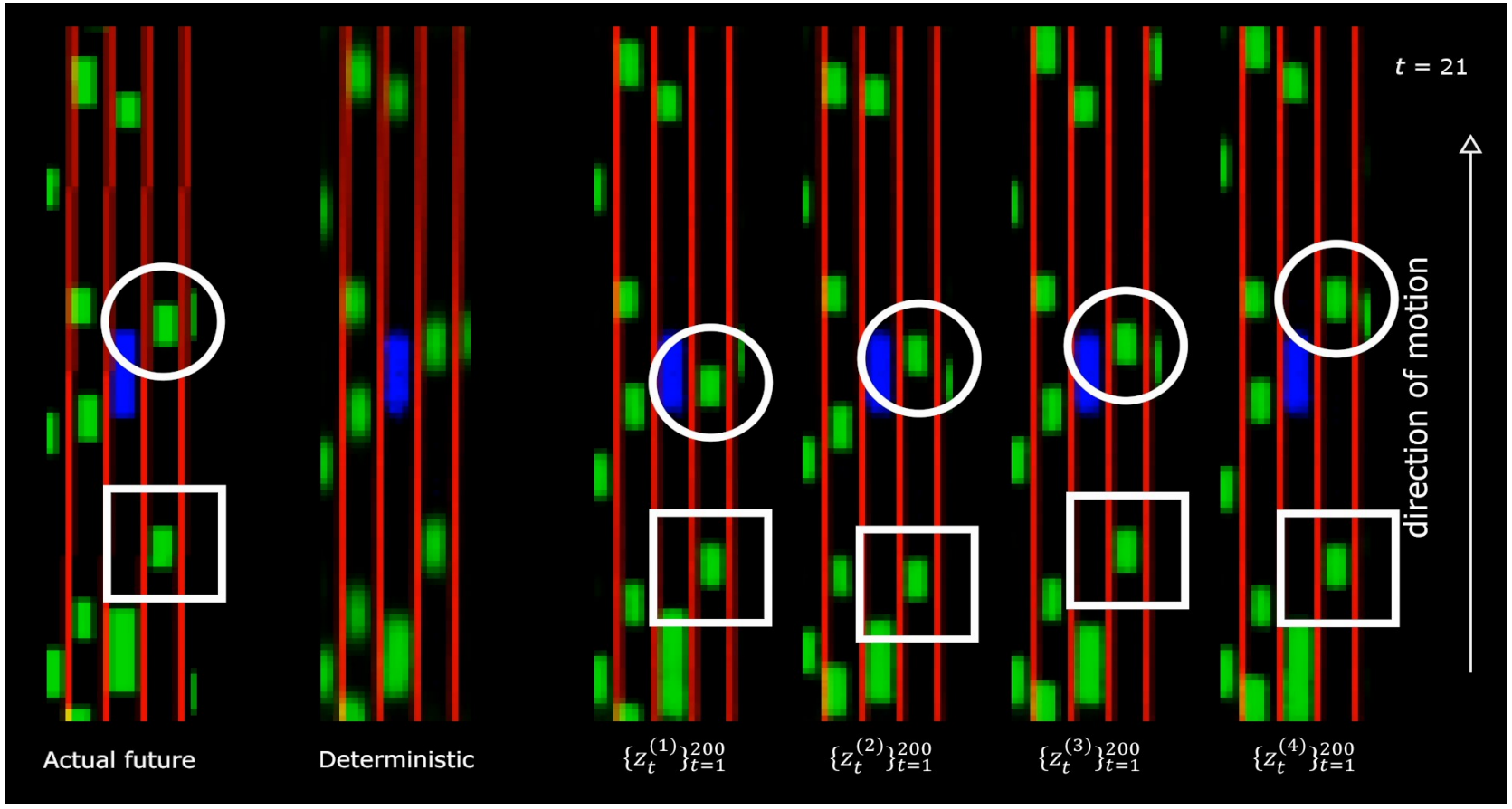
$$P(z/\bar{z}) \propto \exp[-\beta(D(z,\bar{z})+R(z))]$$

  - Predict next frame
  - backprop



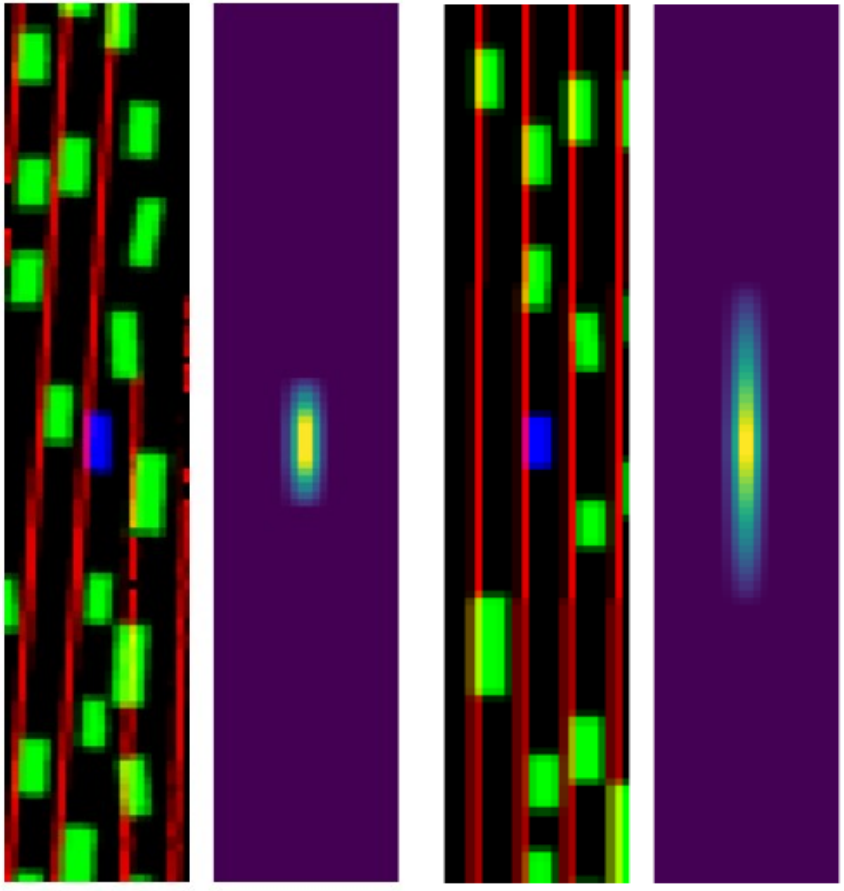


Actual, Deterministic, VAE+Dropout Predictor/encoder



Cost optimized for Planning & Policy Learning

- **Differentiable cost function**
  - Increases as car deviates from lane
  - Increases as car gets too close to other cars nearby in a speed-dependent way
- **Uncertainty cost:**
  - Increases when the costs from multiple predictions (obtained through sampling of drop-out) have high variance.
  - Prevents the system from exploring unknown/unpredictable configurations that may have low cost.

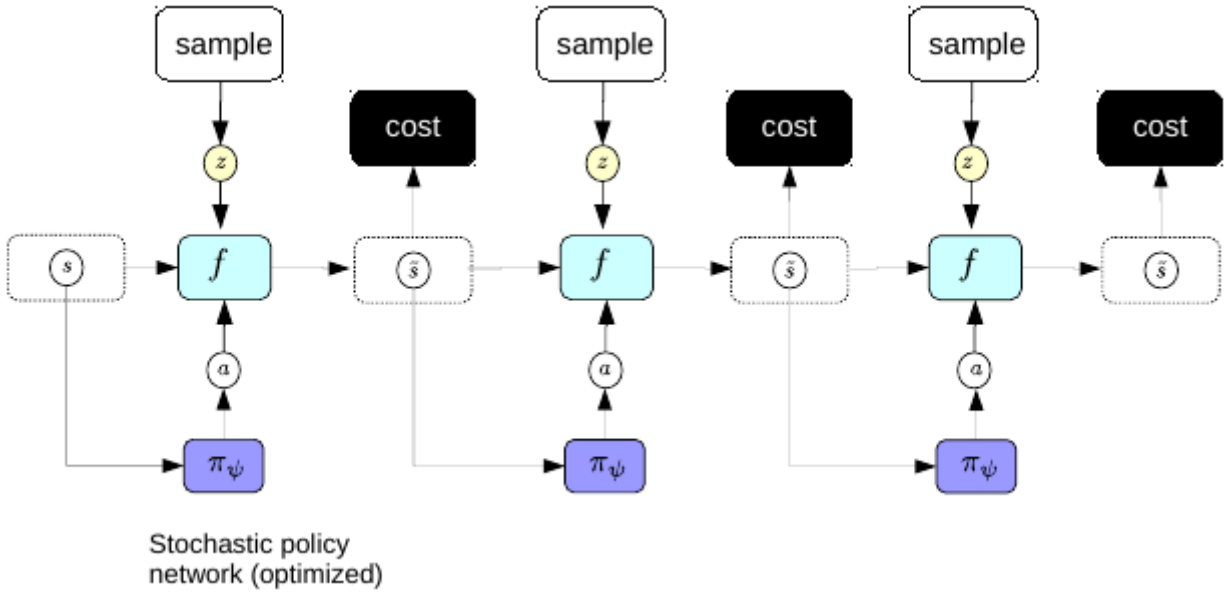


(a) 19.8 km/h (b) 50.3 km/h

Learning to Drive by Simulating it in your Head

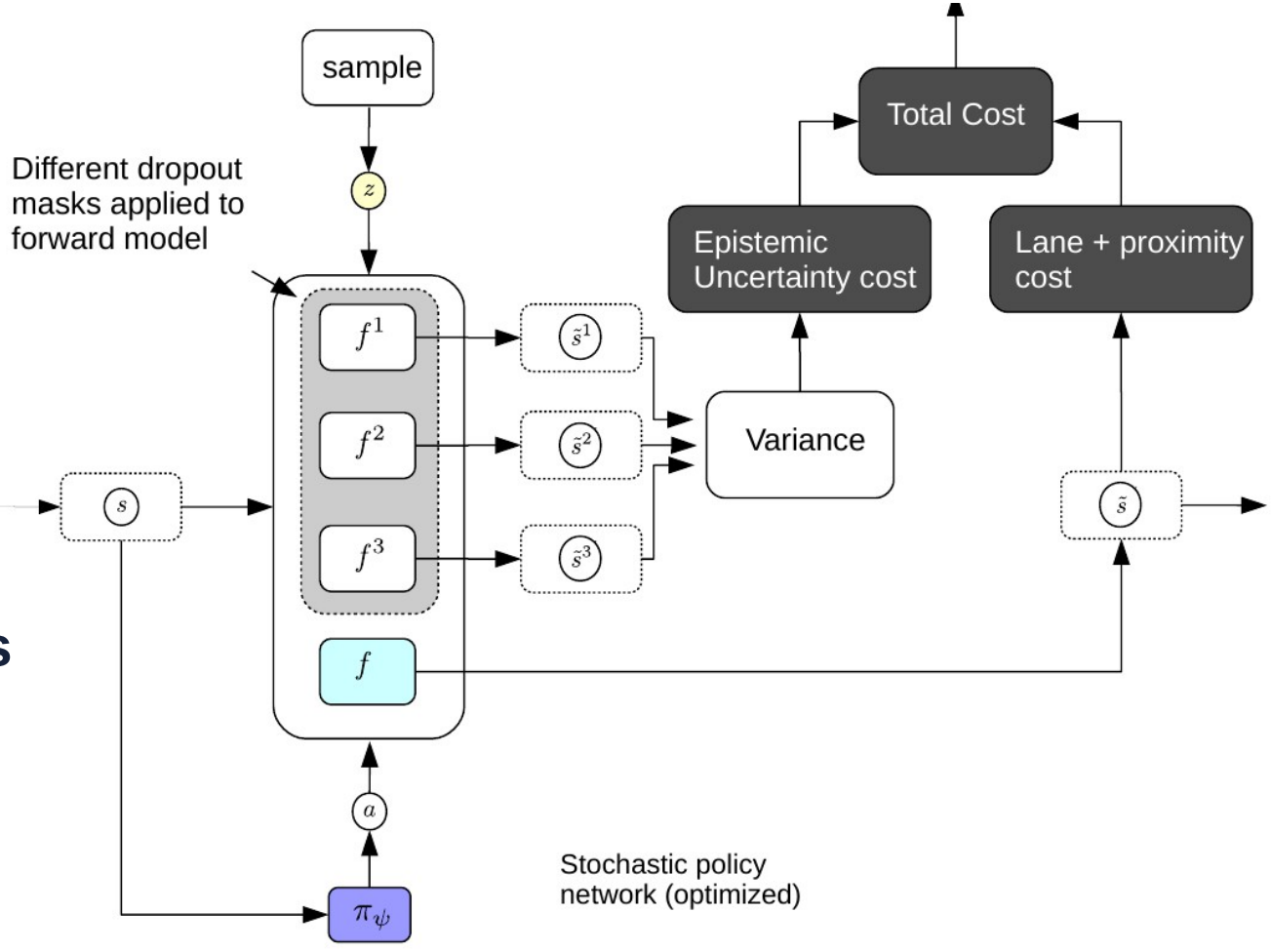
- Feed initial state
- Sample latent variable sequences of length 20
- Run the forward model with these sequences
- Backpropagate gradient of cost to train a policy network.
- Iterate

- No need for planning at run time.

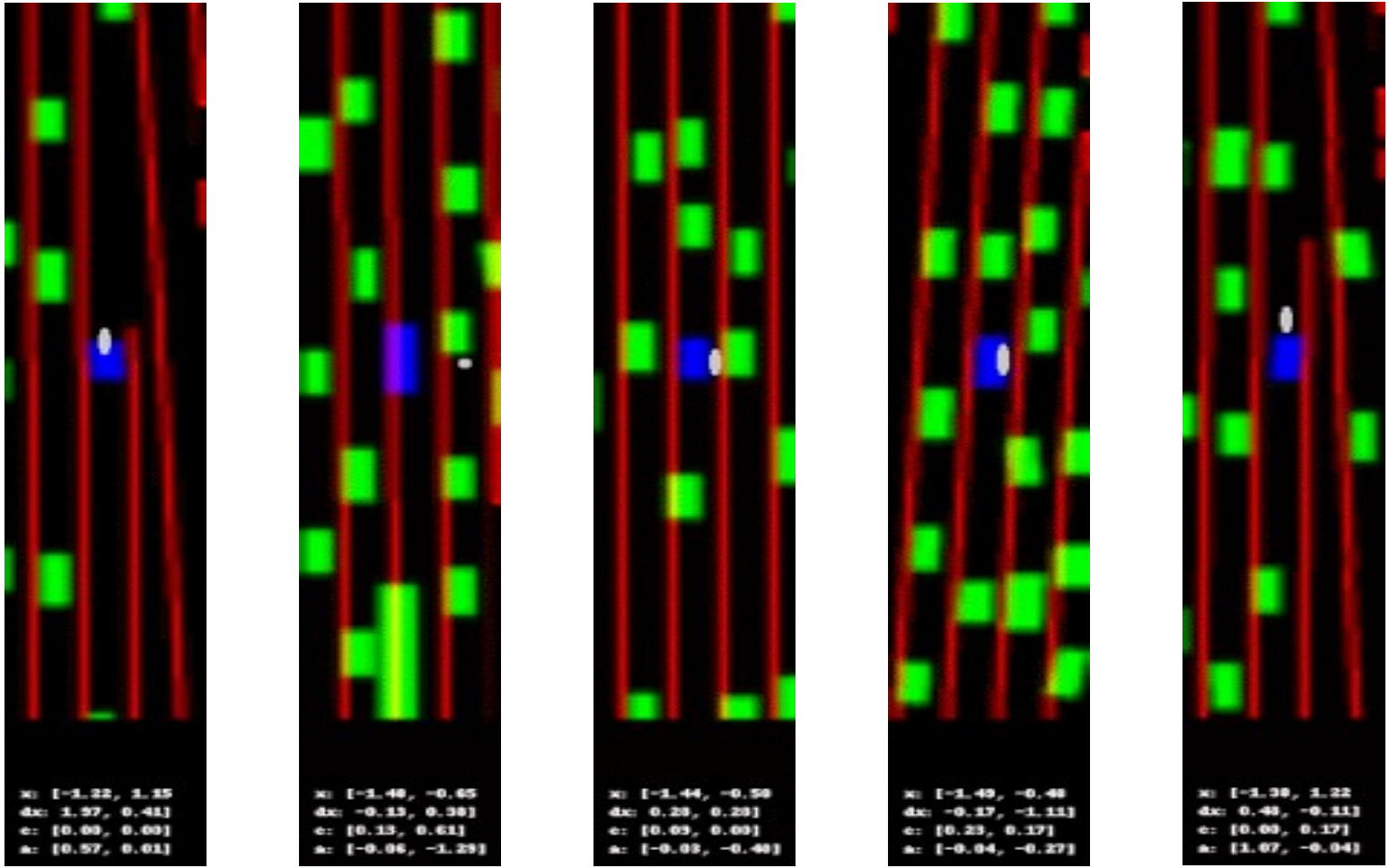


Adding an Uncertainty Cost (doesn't work without it)

- **Estimates epistemic uncertainty**
- **Samples multiple drop-puts in forward model**
- **Computes variance of predictions (differentiably)**
- **Train the policy network to minimize the lane&proximity cost plus the uncertainty cost.**
- **Avoids unpredictable outcomes**

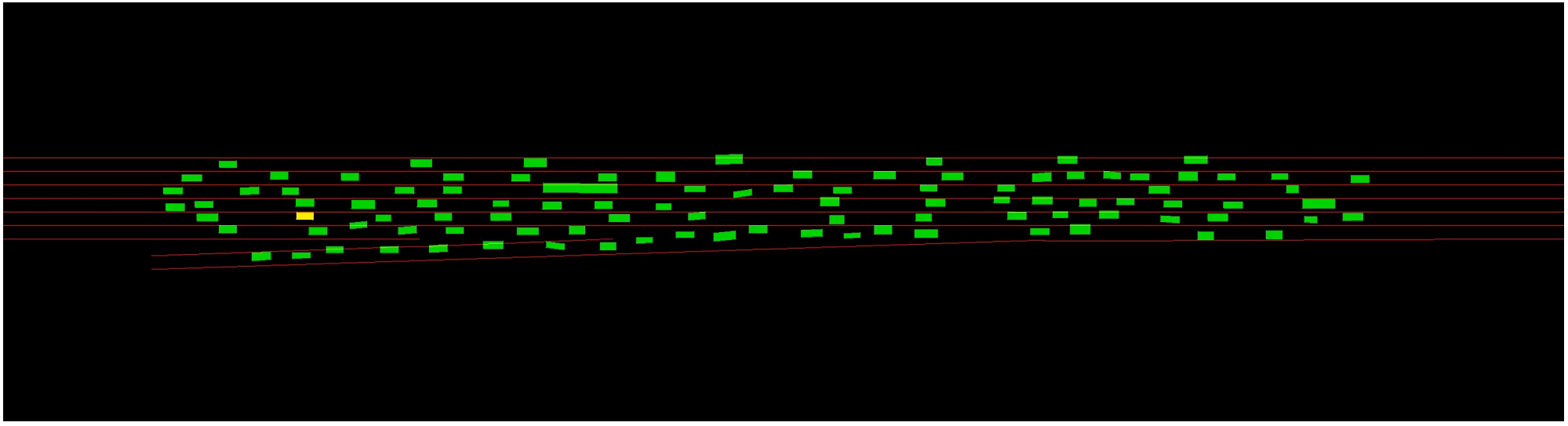


Driving an Invisible Car in “Real” Traffic



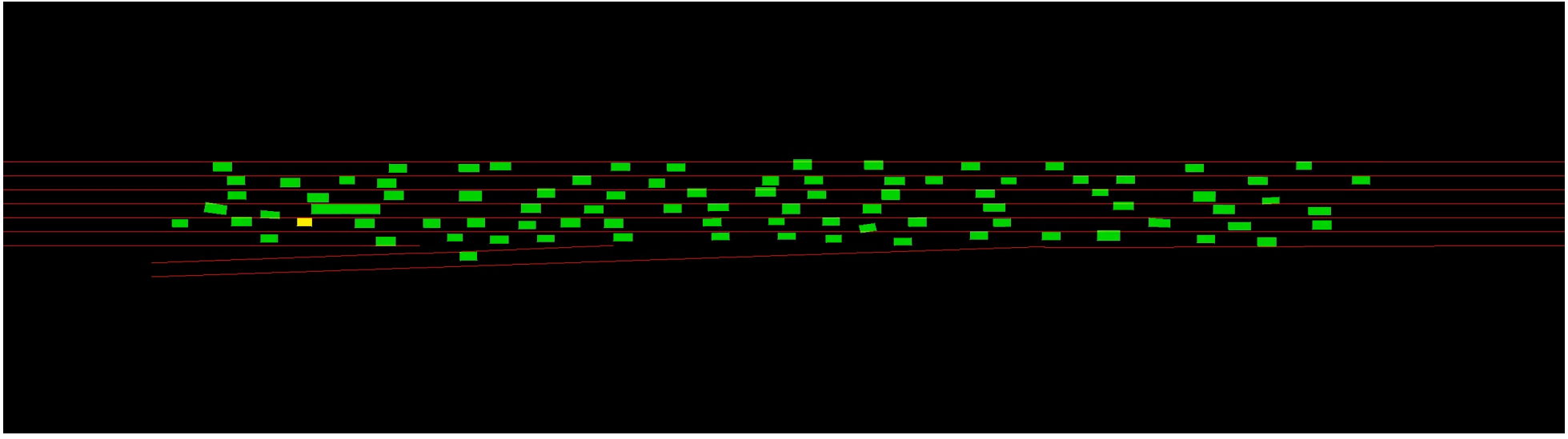
Driving!

- **Yellow: real car**
- **Blue: bot-driven car**



Driving!

- Yellow: real car
- Blue: bot-driven car



conclusions

- **SSL is the future**
  - Learning hierarchical features in a task-invariant way
  - Plenty of data, **massive** networks
  - Learning Forward Models for Model-Based Control
  - Challenge: handling uncertainty in the prediction: energy-based models
- **Reasoning through vector representations and energy minimization**
  - Energy-Based Models with latent variables
  - Replace symbols by vectors and logic by continuous functions.

- **Learning hierarchical representations of action plans**
  - No idea how to do that!

FACEBOOK AI

Thank You!