ACM TechTalk
2019-07-11

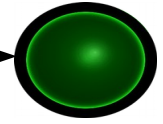
The Power and Limits Of Deep Learning

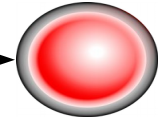
Yann LeCun
New York University
Facebook AI Research
http://yann.lecun.com

facebook
Artificial Intelligence Research

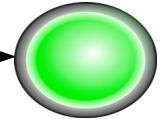
Supervised Learning works but requires too many samples

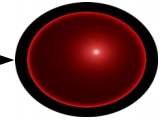
- **Training a machine by showing examples instead of programming it**
- **When the output is wrong, tweak the parameters of the machine**
- **Works well for:**
  - Speech→words
  - Image→categories
  - Portrait→ name
  - Photo→caption
  - Text→topic
  - ….





CAR



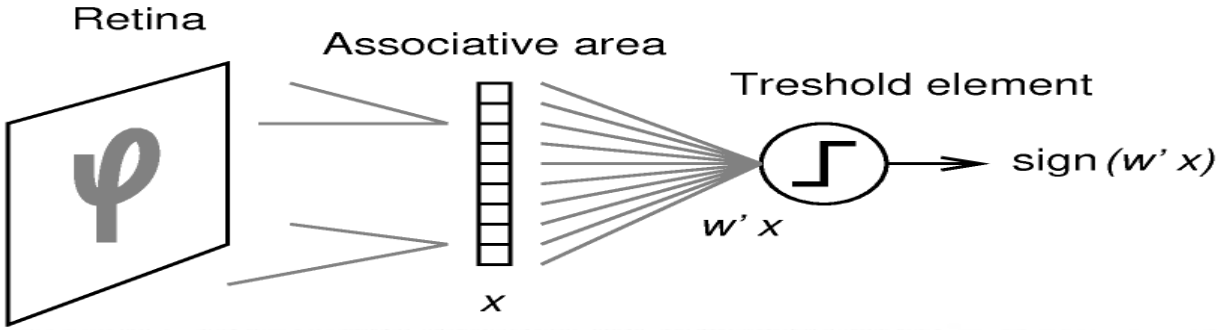


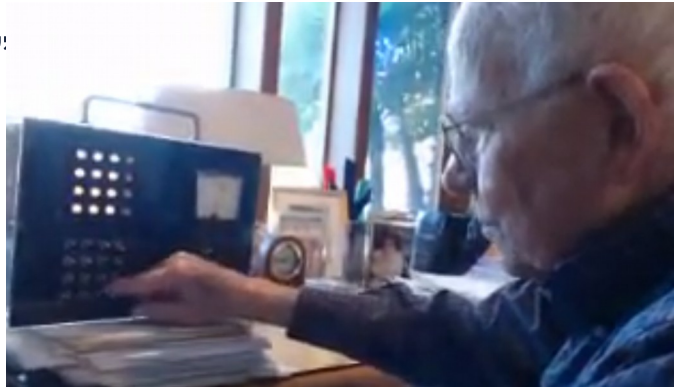
PLANE

Supervised Learning goes back to the Perceptron & Adaline

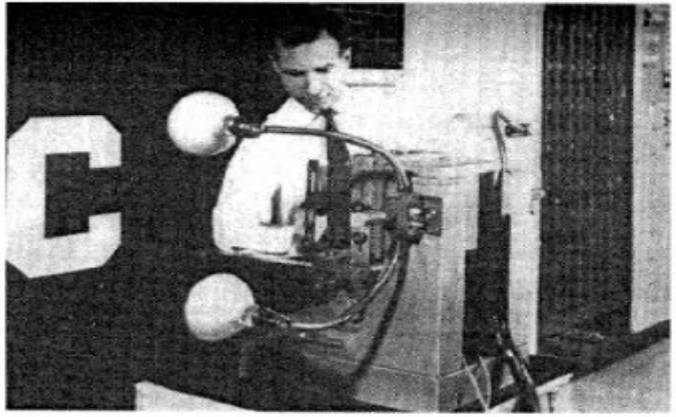
- **The McCulloch-Pitts Binar Neuron**
  - Perceptron: weights are motorized potentiometers
  - Adaline: Weights are electrochemical "memistors"

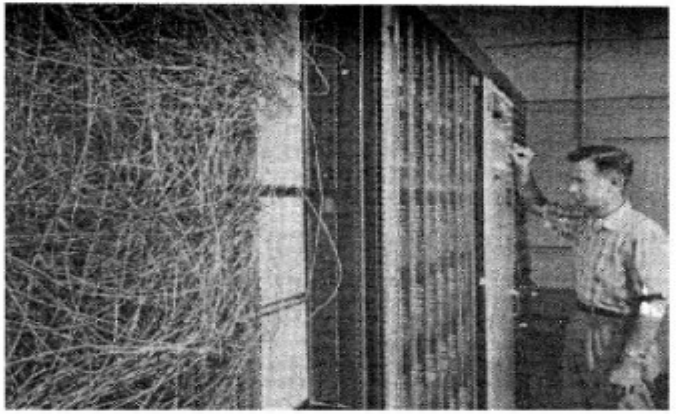
$$y = sign\left(\sum_{i=1}^{N} W_i X_i + b\right)$$

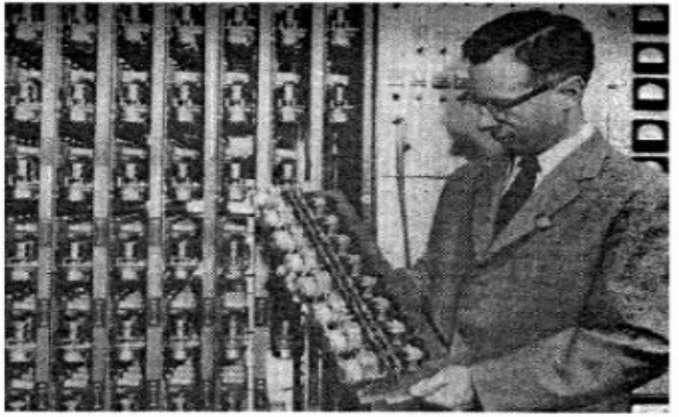




https://youtu.be/X1G2g3SiCwU







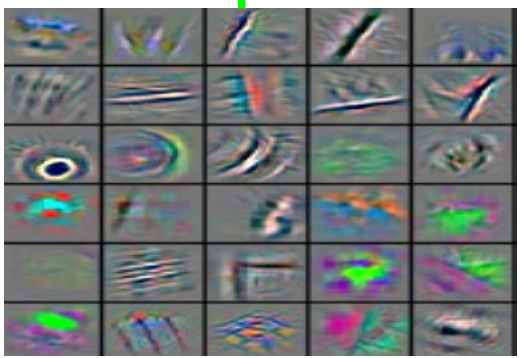
ConvNets: Multiple Trainable Layers, Hierarchical Representations

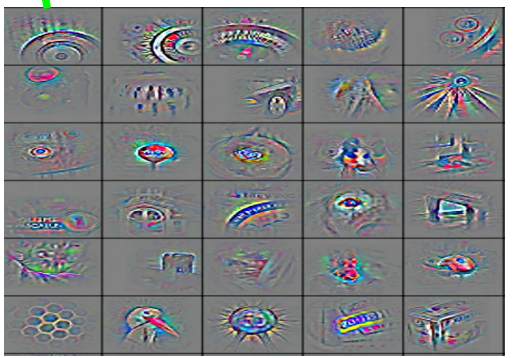
Traditional Pattern Recognition: Fixed/Handcrafted Feature Extractor

Feature Extractor → Trainable Classifier

Deep Learning: Representations are hierarchical and trained

Low-Level Features → Mid-Level Features → High-Level Features → Trainable Classifier

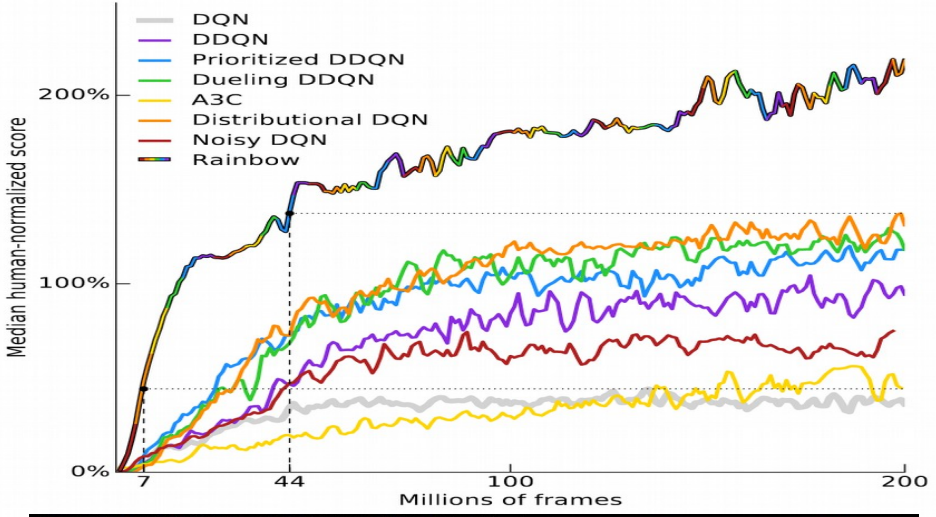


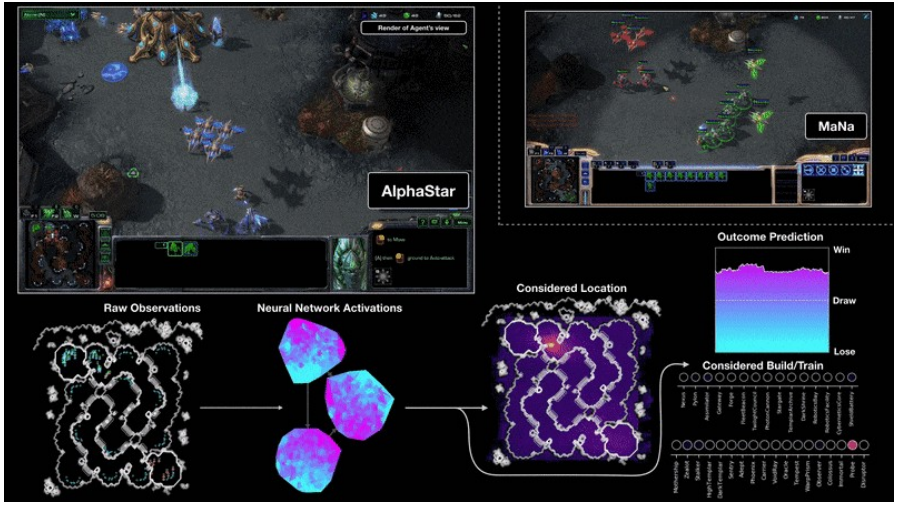


Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Reinforcement Learning: Model-Free RL works great for games.

- **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]
- **They all use ConvNets and a few other architectural concepts.**





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.**

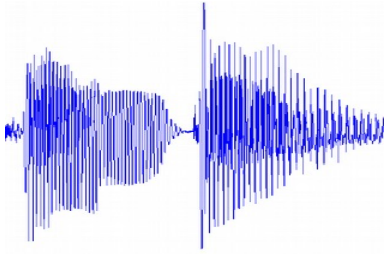
- **Anything you do in the real world can kill you**
- **You can't run the real world faster than real time**

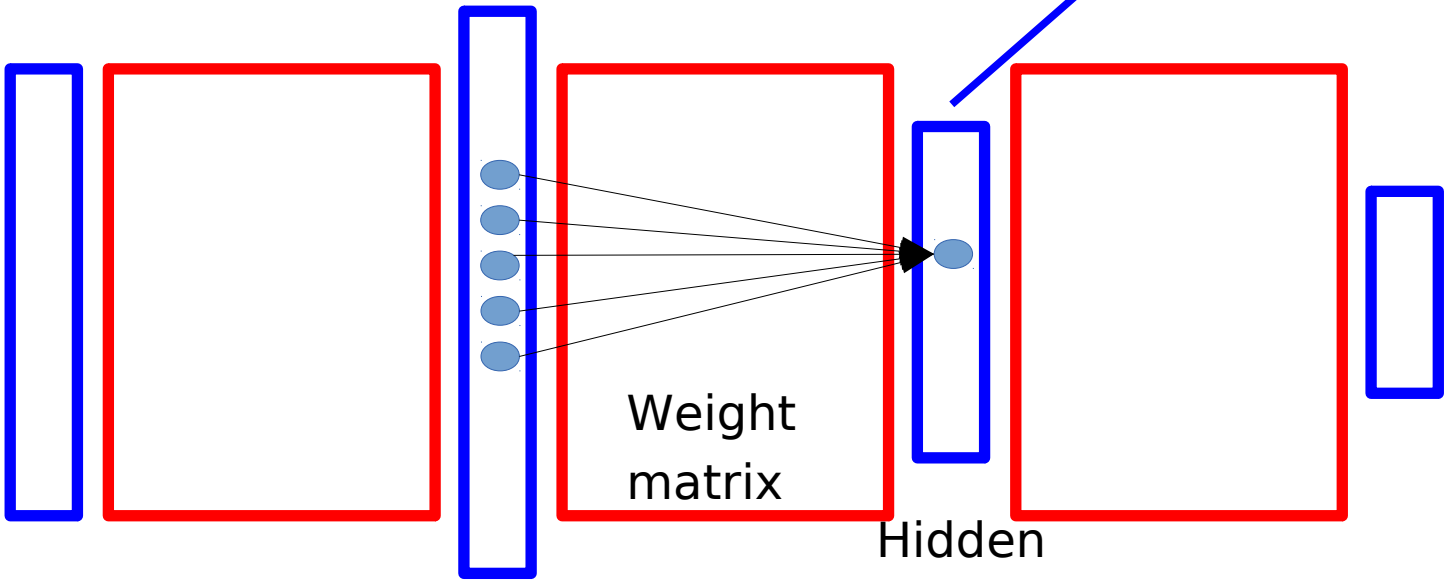
(Deep) Multi-Layer Neural Nets

- Multiple Layers of simple units
- Each units computes a weighted sum of its inputs
- Weighted sum is passed through a non-linear function
- The learning algorithm changes the weights

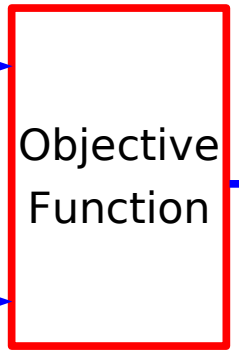
$$ReLU(x) = max(x, 0)$$

Ceci est une voiture

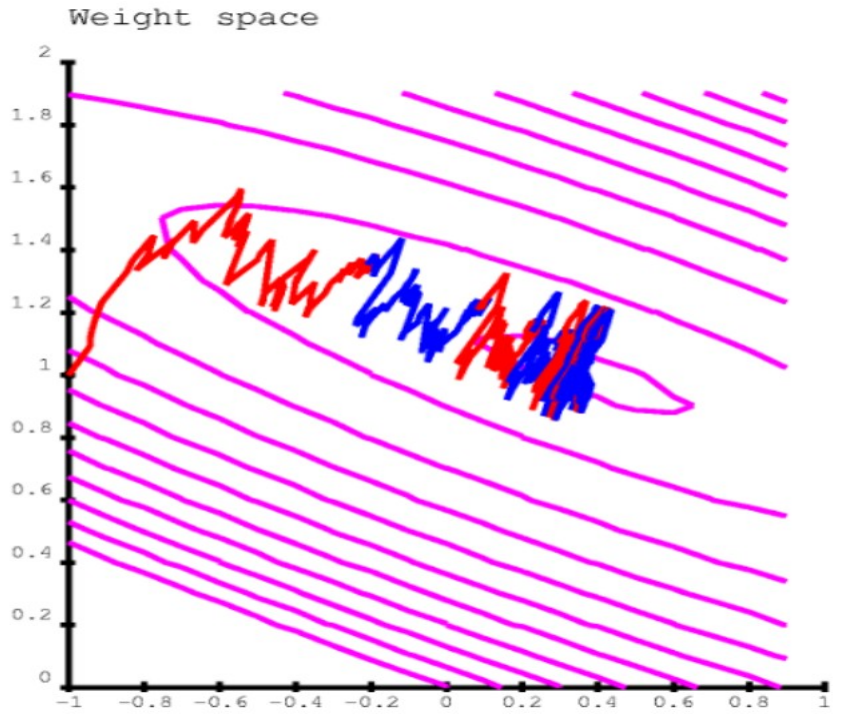




Supervised Machine Learning = Function Optimization



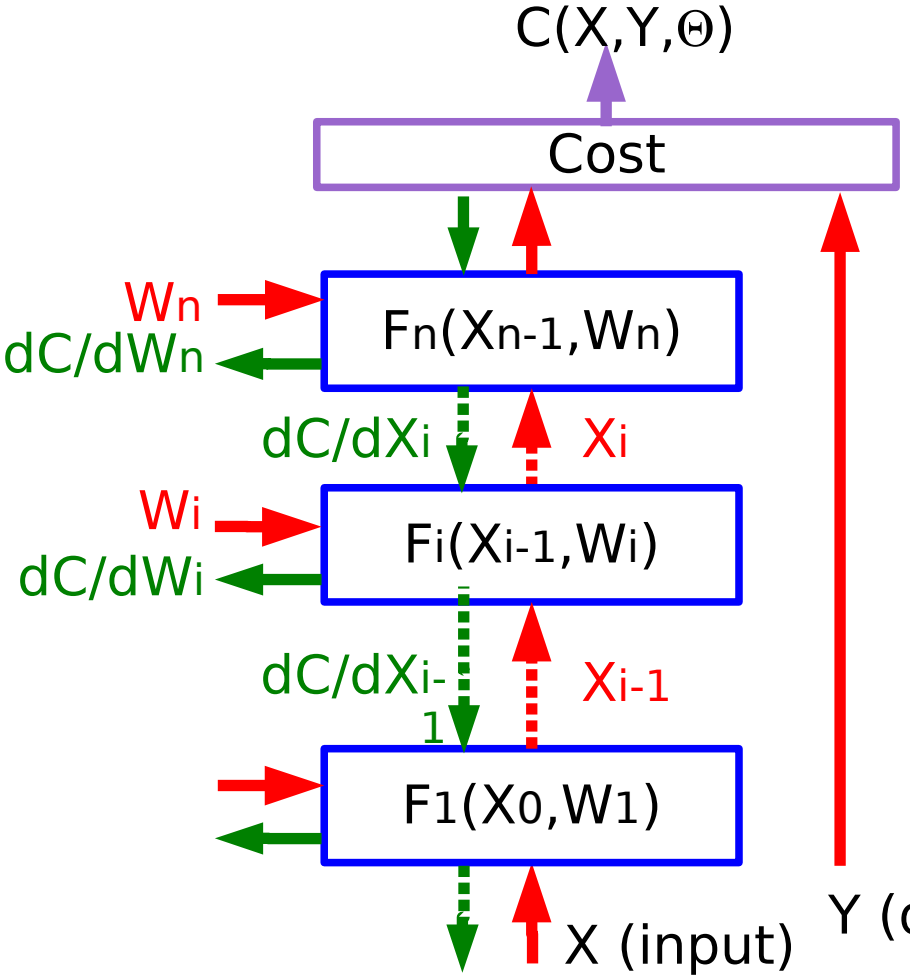
traffic light: -1



- **It's like walking in the mountains in a fog and following the direction of steepest descent to reach the village in the valley**
- **But each sample gives us a noisy estimate of the direction. So our path is a bit random.**
- **Stochastic Gradient Descent (SGD)**

$$W_i \leftarrow W_i - \eta \frac{\partial L(W, X)}{\partial W_i}$$

Computing Gradients by Back-Propagation



- **A practical Application of Chain Rule**

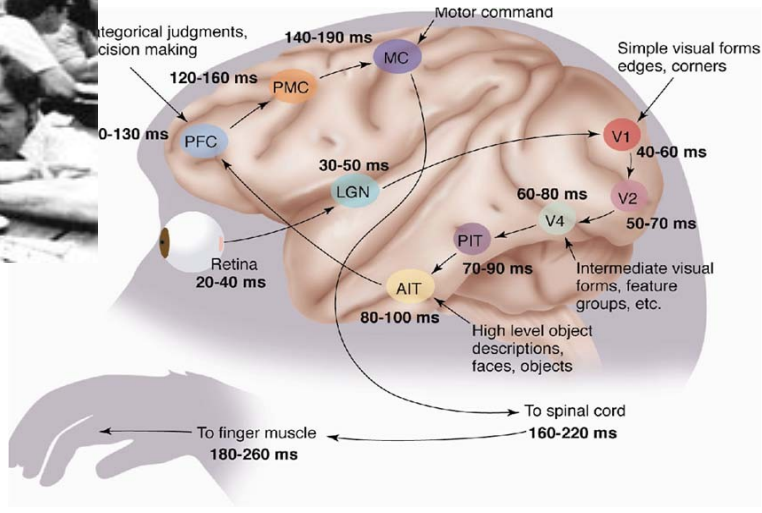
- **Backprop for the state gradients:**
- **$dC/dX_{i-1} = dC/dX_i \cdot dX_i/dX_{i-1}$**
- **$dC/dX_{i-1} = dC/dX_i \cdot dF_i(X_{i-1},W_i)/dX_{i-1}$**

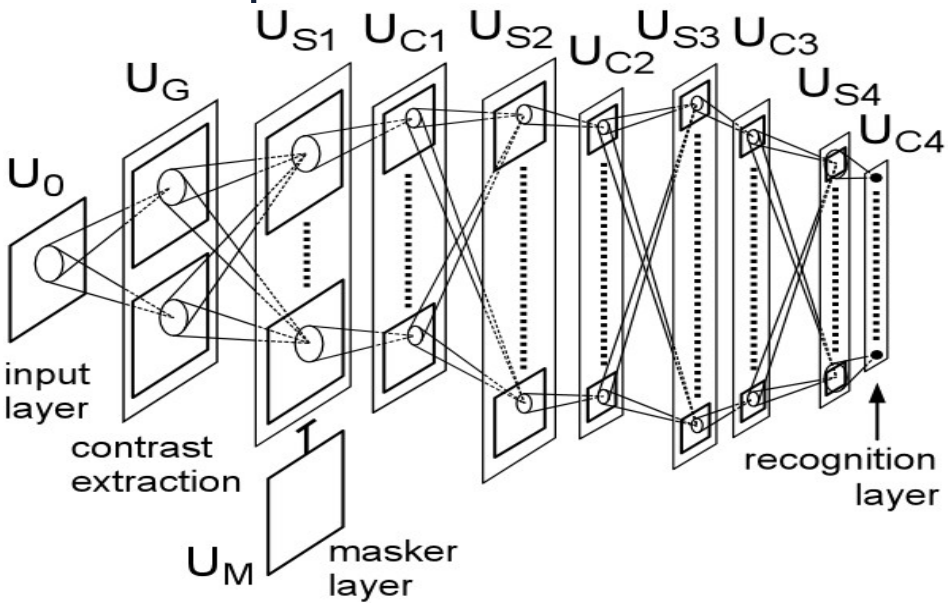
- **Backprop for the weight gradients:**
- **$dC/dW_i = dC/dX_i \cdot dX_i/dW_i$**
- **$dC/dW_i = dC/dX_i \cdot dF_i(X_{i-1},W_i)/dW_i$**

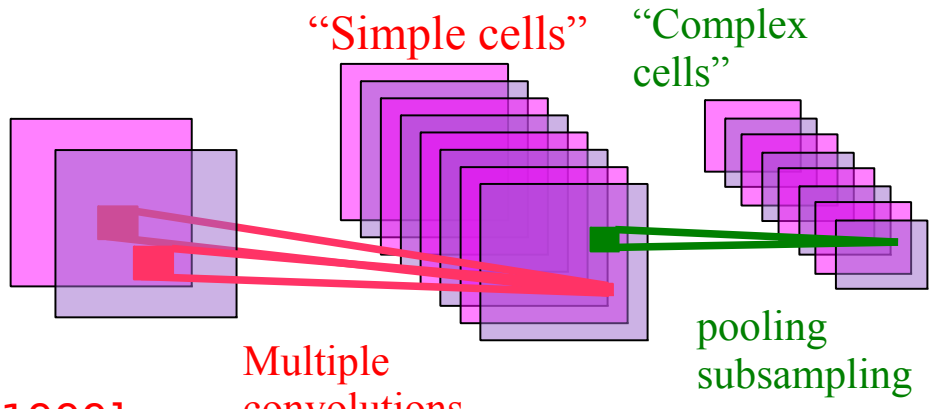
Hubel & Wiesel's Model of the Architecture of the Visual Cortex

- **[Hubel & Wiesel 1962]:**
  - simple cells detect local features
  - complex cells "pool" the outputs of simple cells within a

[Thorpe & Fabre-Thorpe 2001]

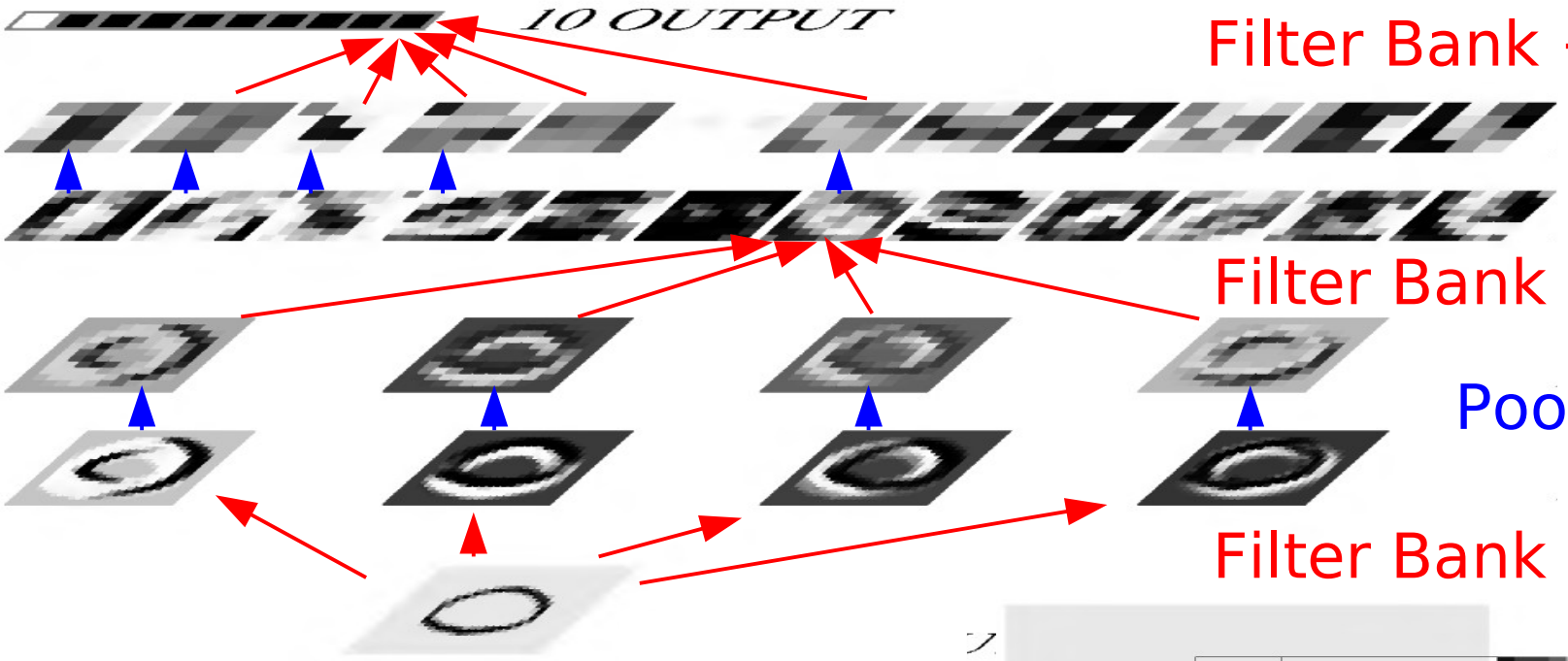






[Fukushima 1982][LeCun 1989, 1998],[Riesenhuber 1999].....

Convolutional Network Architecture [LeCun et al. NIPS 1989]



Filter Bank +non-linearity

Pooling

Filter Bank +non-linearity

Pooling

Filter Bank +non-linearity

- **Inspired by [Hubel & Wiesel 1962] & [Fukushima 1982] (Neocognitron):**
  - simple cells detect local features
  - complex cells "pool" the outputs of simple cells within a retinotopic neighborhood.

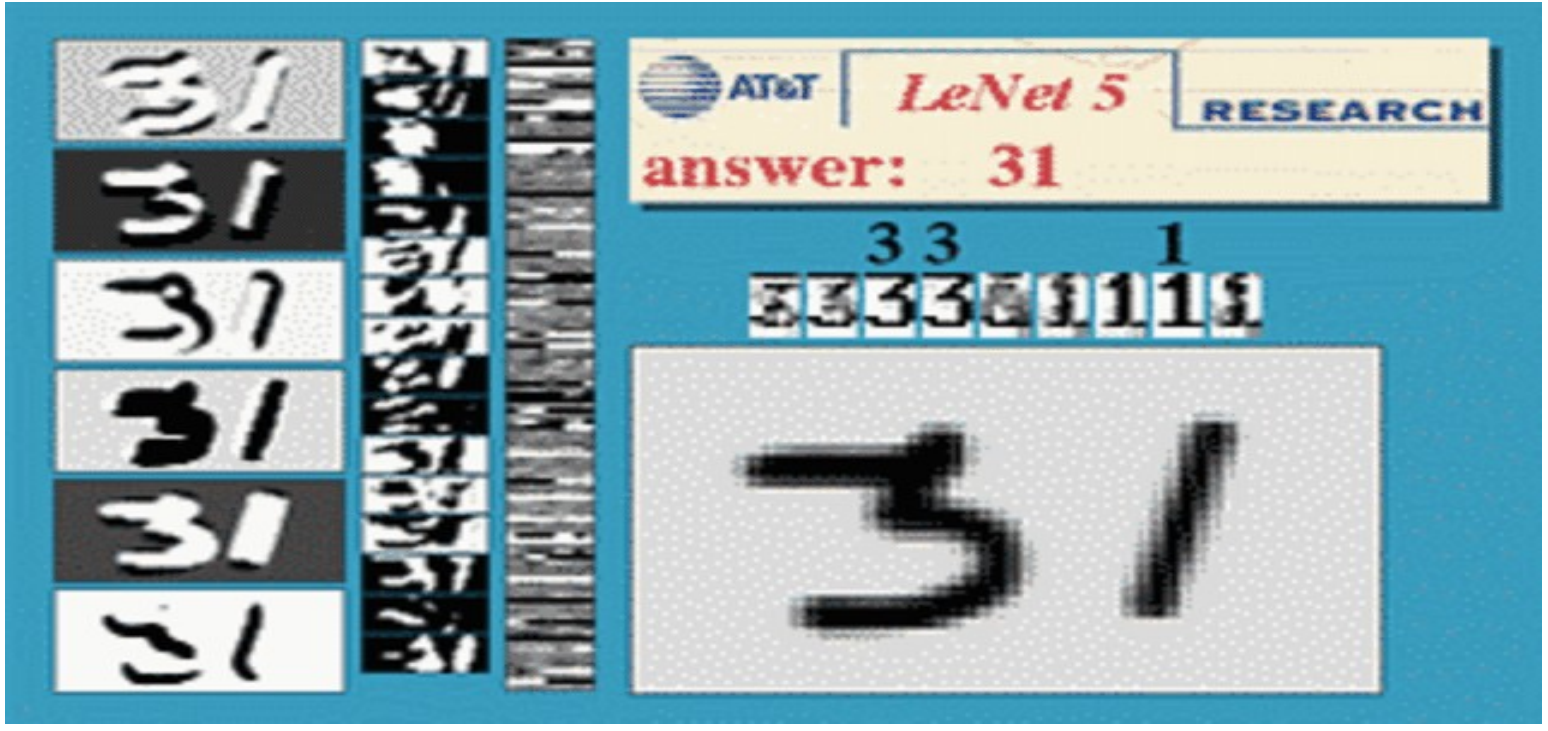
Convolutional Network (LeNet5, vintage 1990)

- **Filters-tanh → pooling → filters-tanh → pooling → filters-tanh**

LeNet character recognition demo 1992

ConvNets can recognize multiple objects

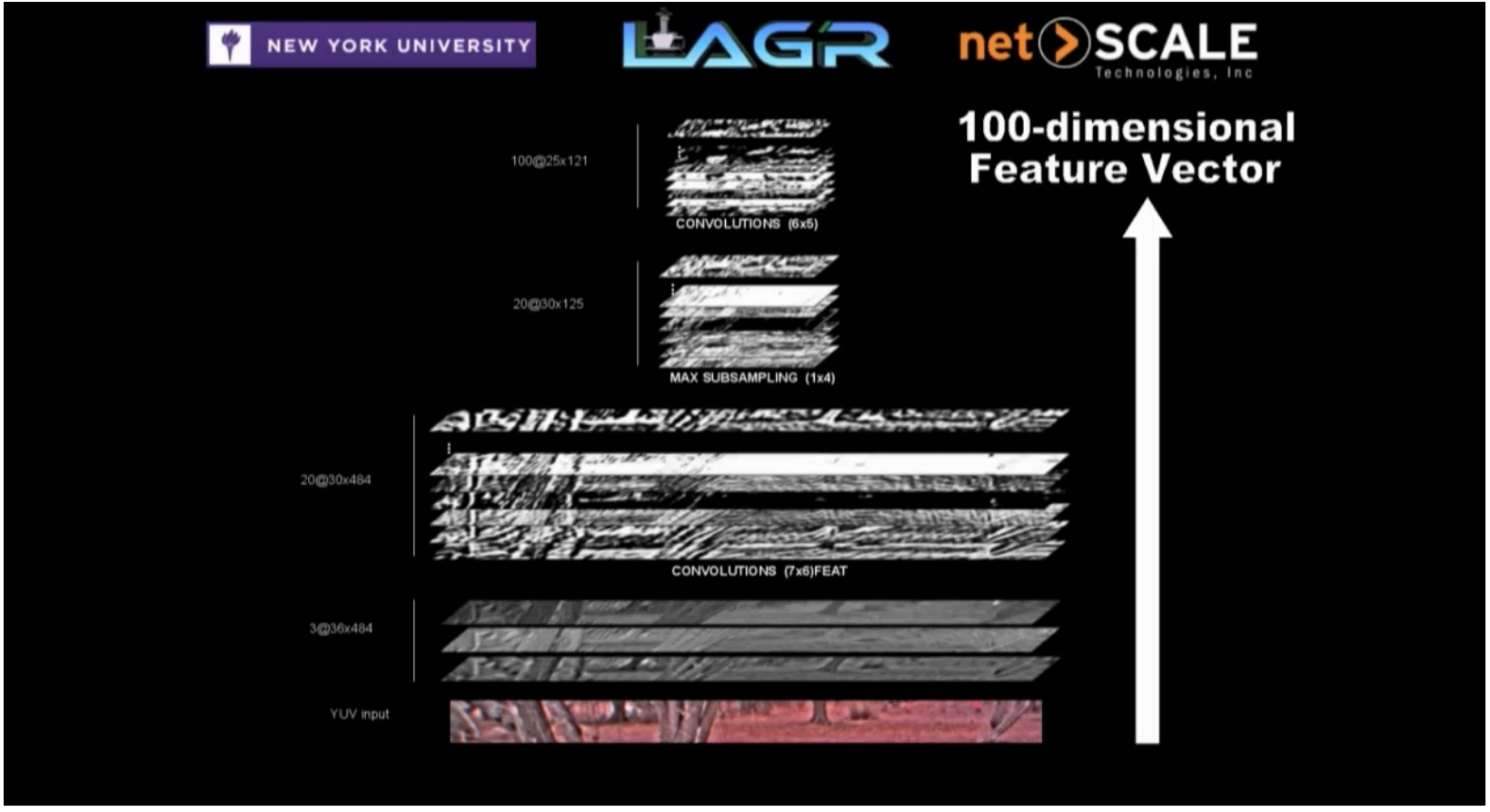
- All layers are convolutional
- Networks performs simultaneous segmentation and recognition



Face & Pedestrian Detection with ConvNets (1993-2005)

[Osadchy,Miller LeCun JMLR 2007],[Kavukcuoglu et al. NIPS 2010] [Sermanet et al. CVPR 2013]

Training a Robot to Drive Itself in Nature



Semantic Segmentation with ConvNets (33 categories)

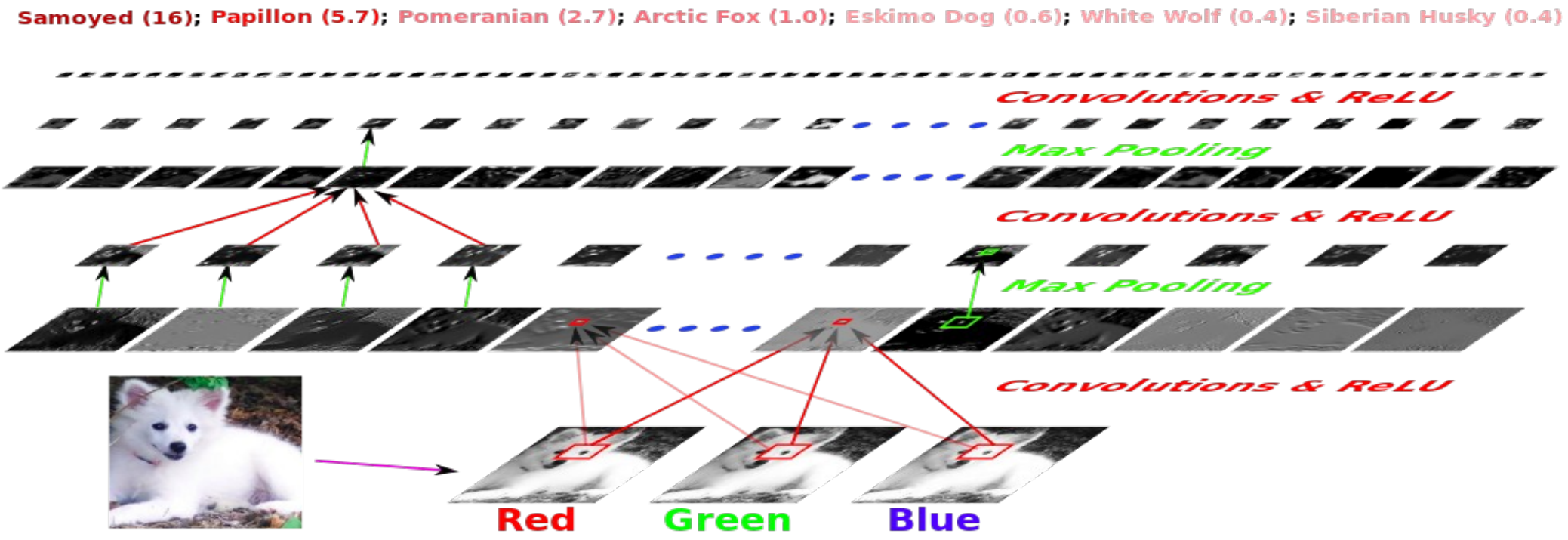
The Deep Learning Revolution

State of the Art

facebook
Artificial Intelligence Research

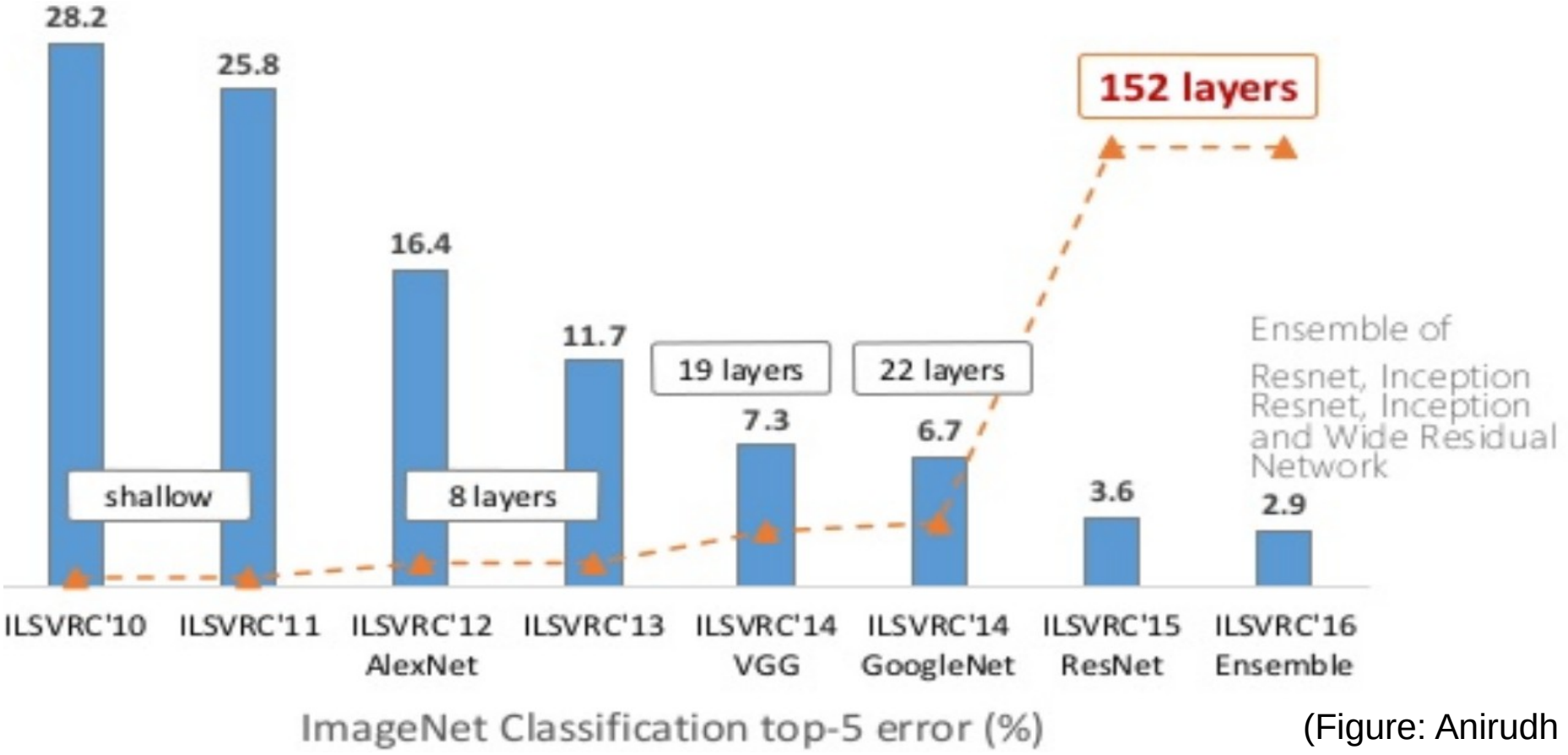
Deep ConvNets for Object Recognition (on GPU)

- AlexNet [Krizhevsky et al. NIPS 2012], OverFeat [Sermanet et al. 2013]
- 1 to 10 billion connections, 10 million to 1 billion parameters, 8 to 20 layers.



Error Rate on ImageNet

- **Depth inflation**

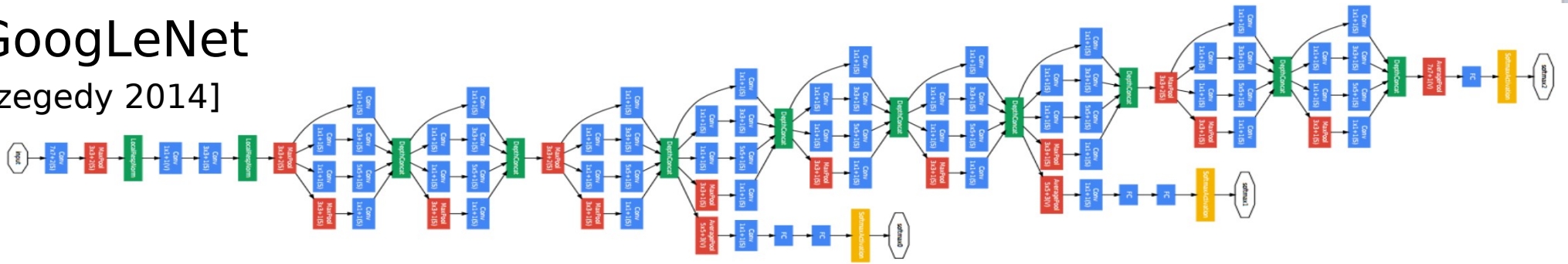


(Figure: Anirudh Koul)

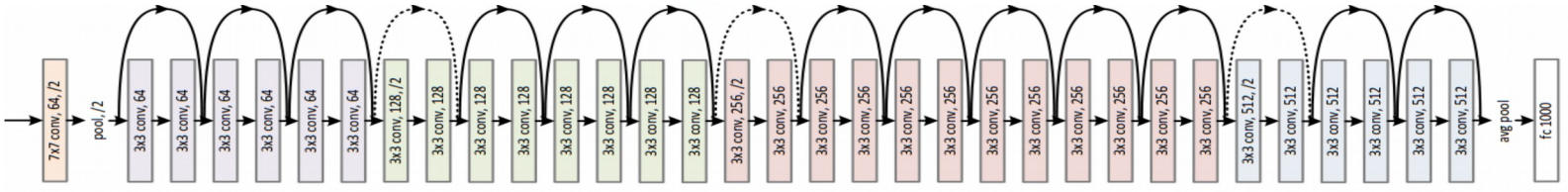
Deep ConvNets (depth inflation)

VGG
[Simonyan 2013]

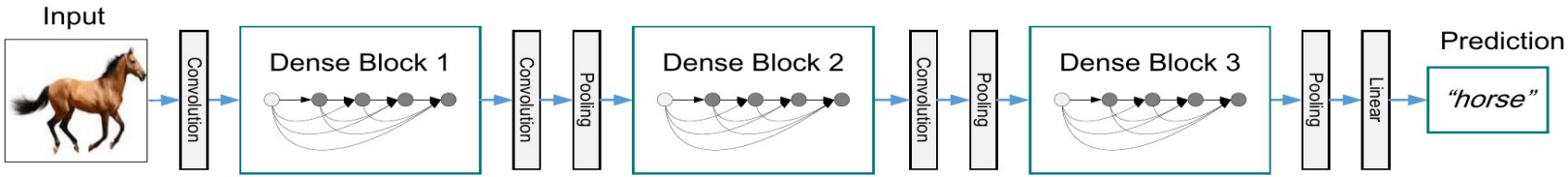
GoogLeNet
Szegedy 2014]



ResNet
[He et al. 2015]

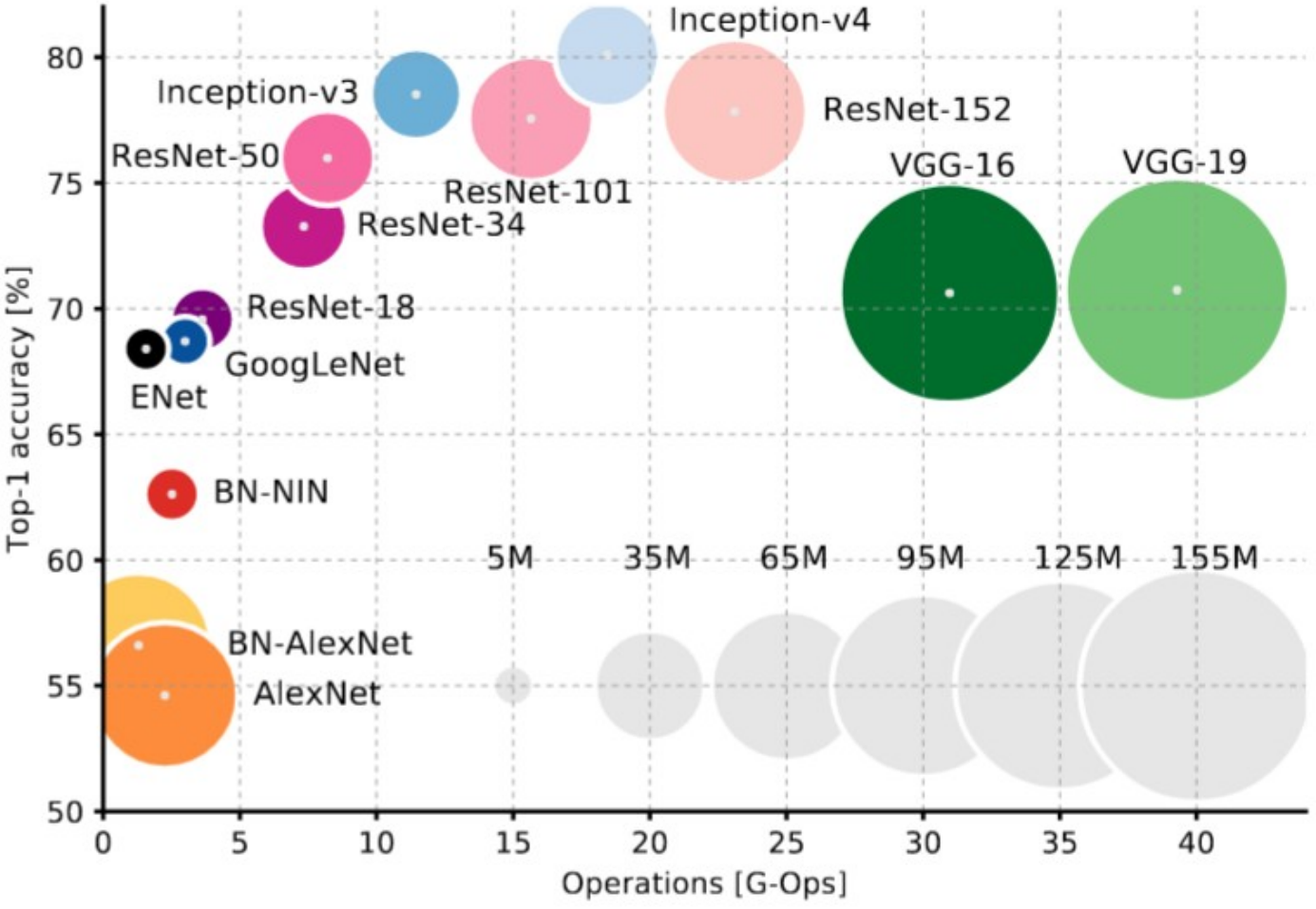


DenseNet
[Huang et al 2017]



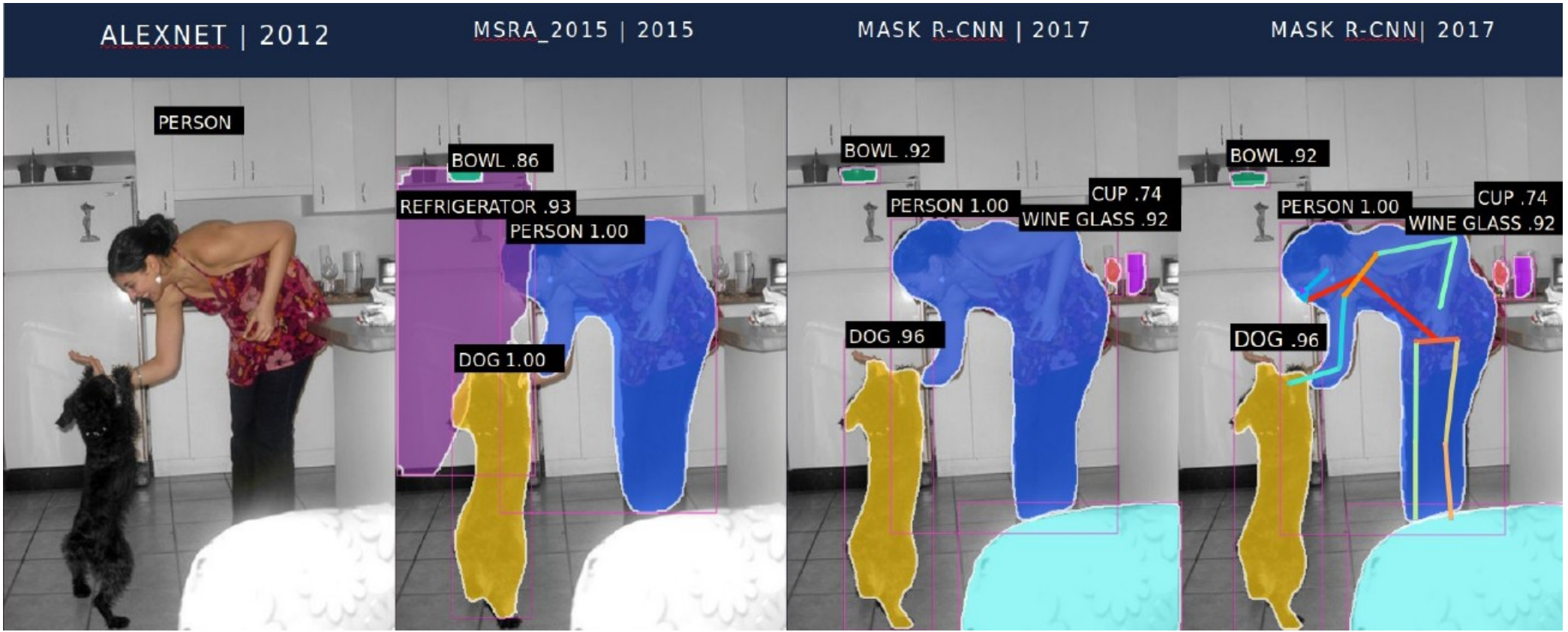
GOPS vs Accuracy on ImageNet vs #Parameters

- [Canziani 2016]
- ResNet50 and ResNet100 are used routinely in production.
- Each of the few billions photos uploaded on Facebook every day goes through a handful of ConvNets within 2 seconds.



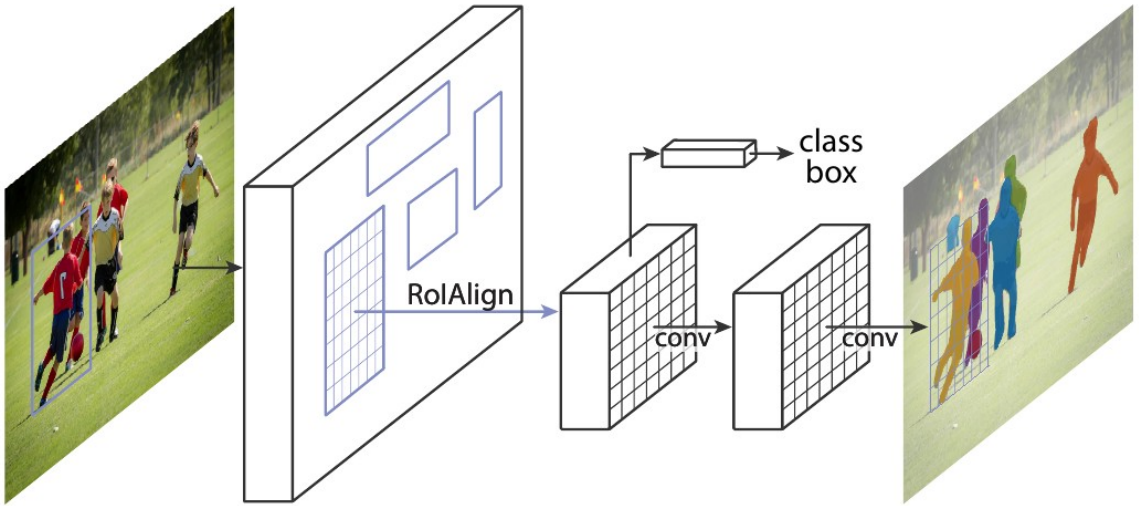
Progress in Computer Vision

- **[He 2017]**



Mask R-CNN: instance segmentation

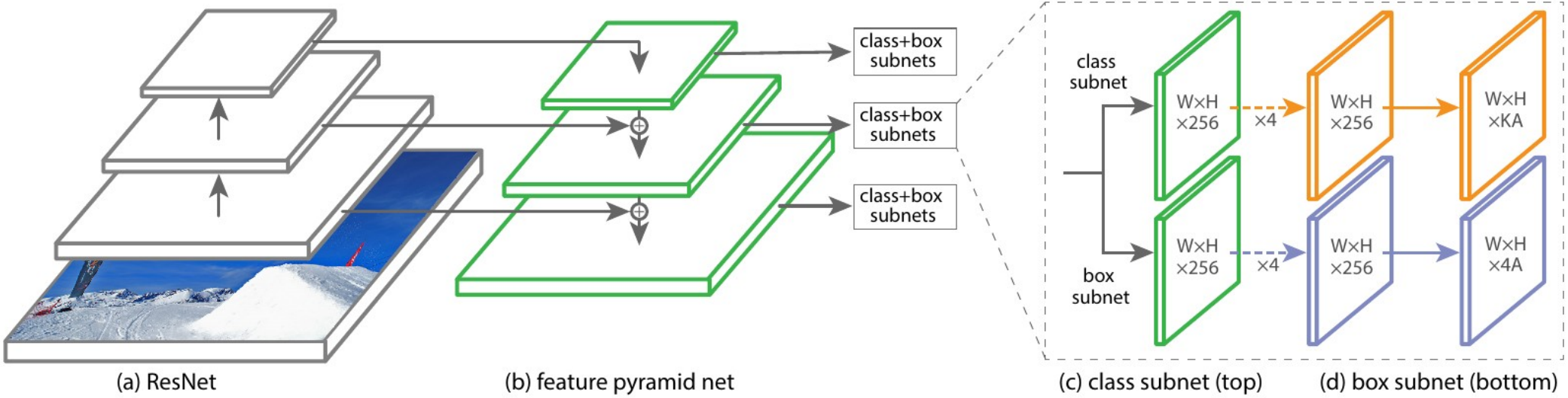
- [He, Gkioxari, Dollar, Girshick arXiv:1703.06870]
- ConvNet produces an object mask for each region of interest
- Combined ventral and dorsal pathways



	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [7]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [20] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [20] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

RetinaNet, feature pyramid network

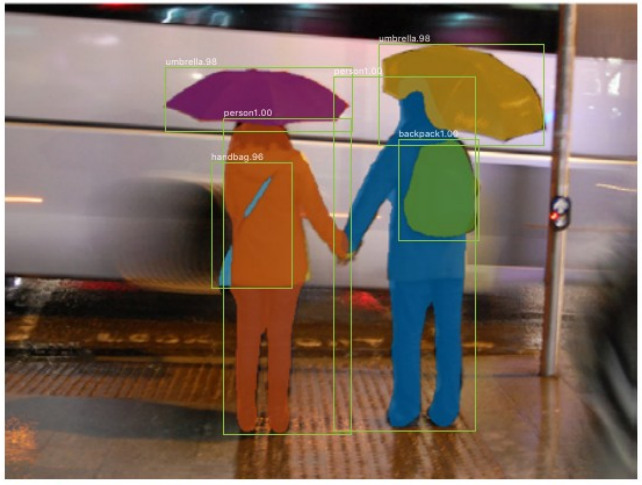
- **One-pass object detection**
- **[Lin et al. ArXiv:1708.02002]**

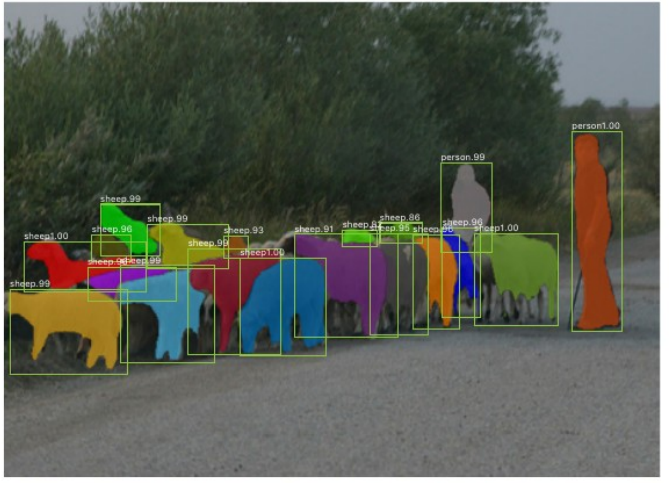


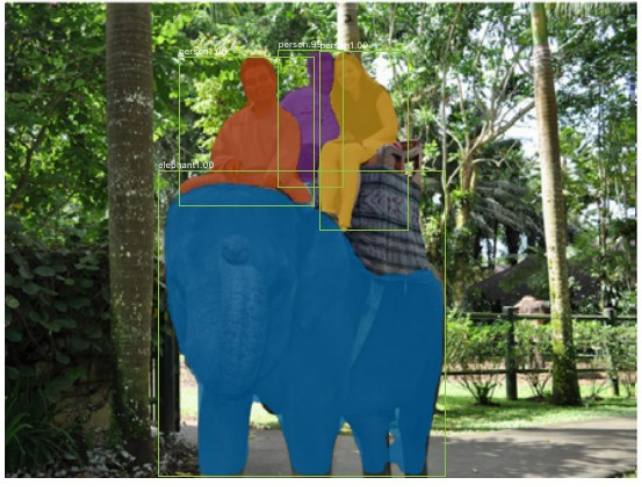
Mask-RCNN Results on COCO dataset

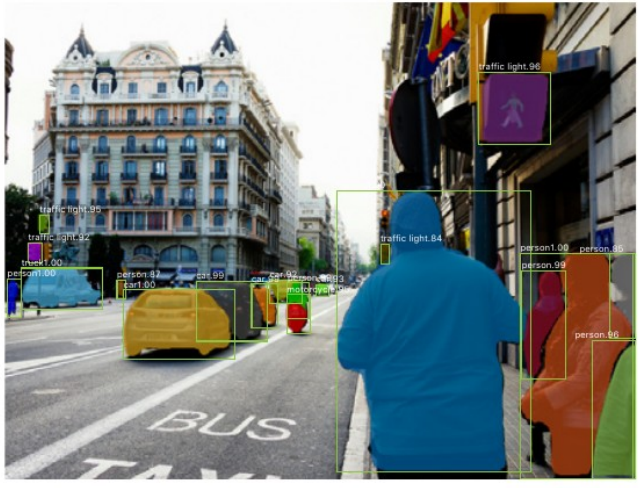
- **Individual objects are segmented.**

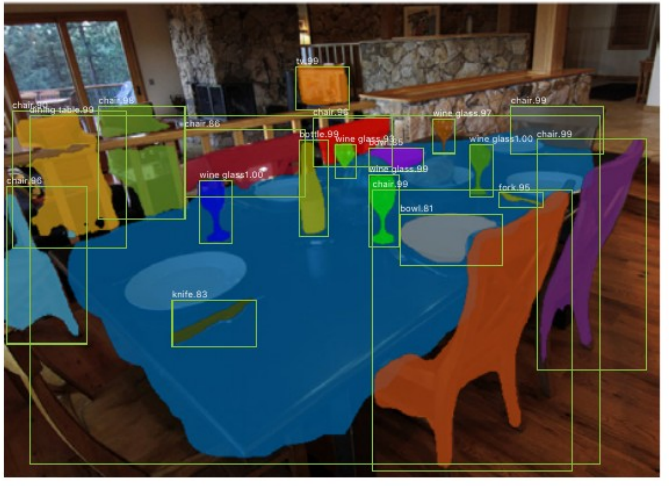
Mask R-CNN Results on COCO test set





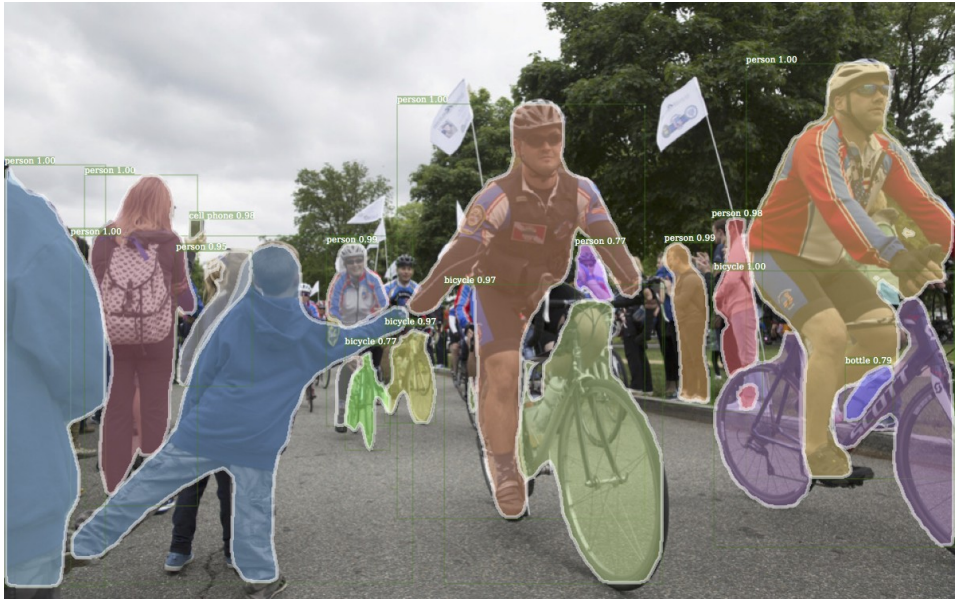


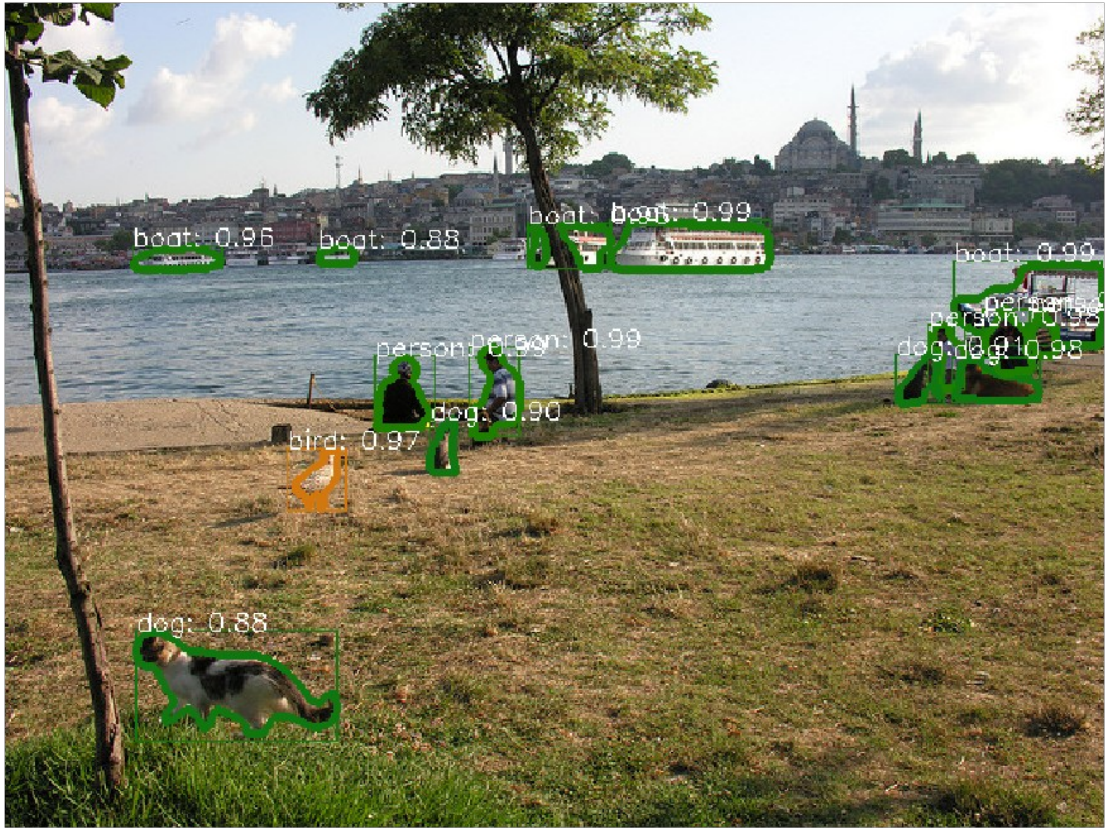




Detectron: open source vision in PyTorch

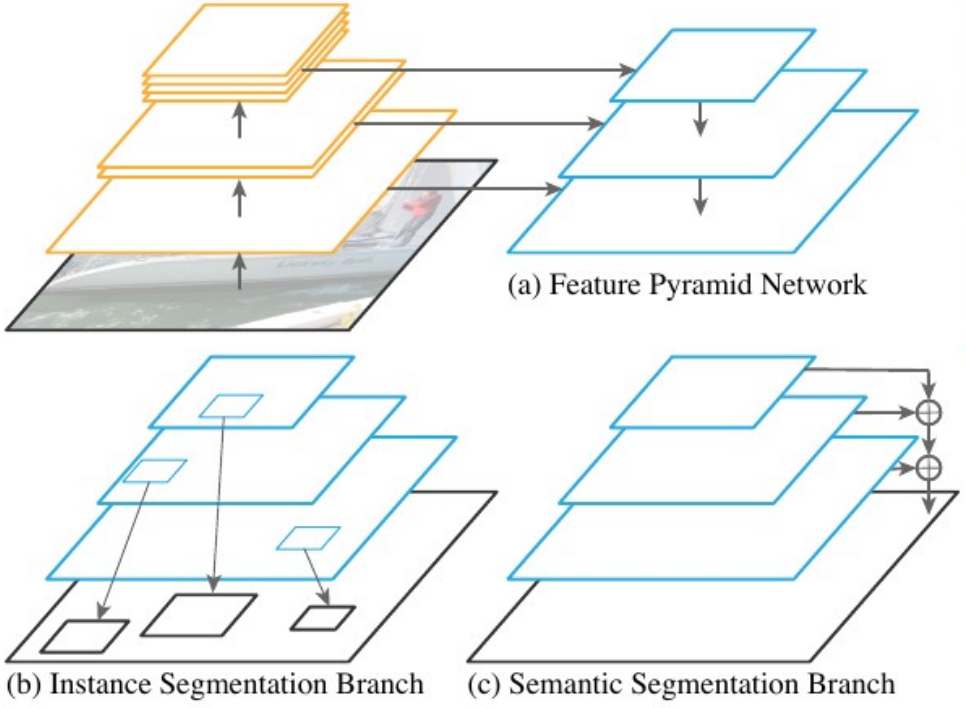
https://github.com/facebookresearch/maskrcnn-benchmark

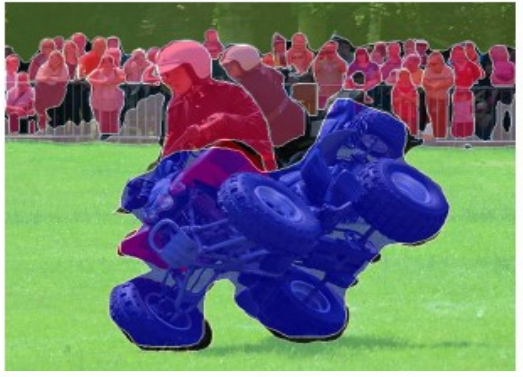




Panoptic Feature Pyramid Network

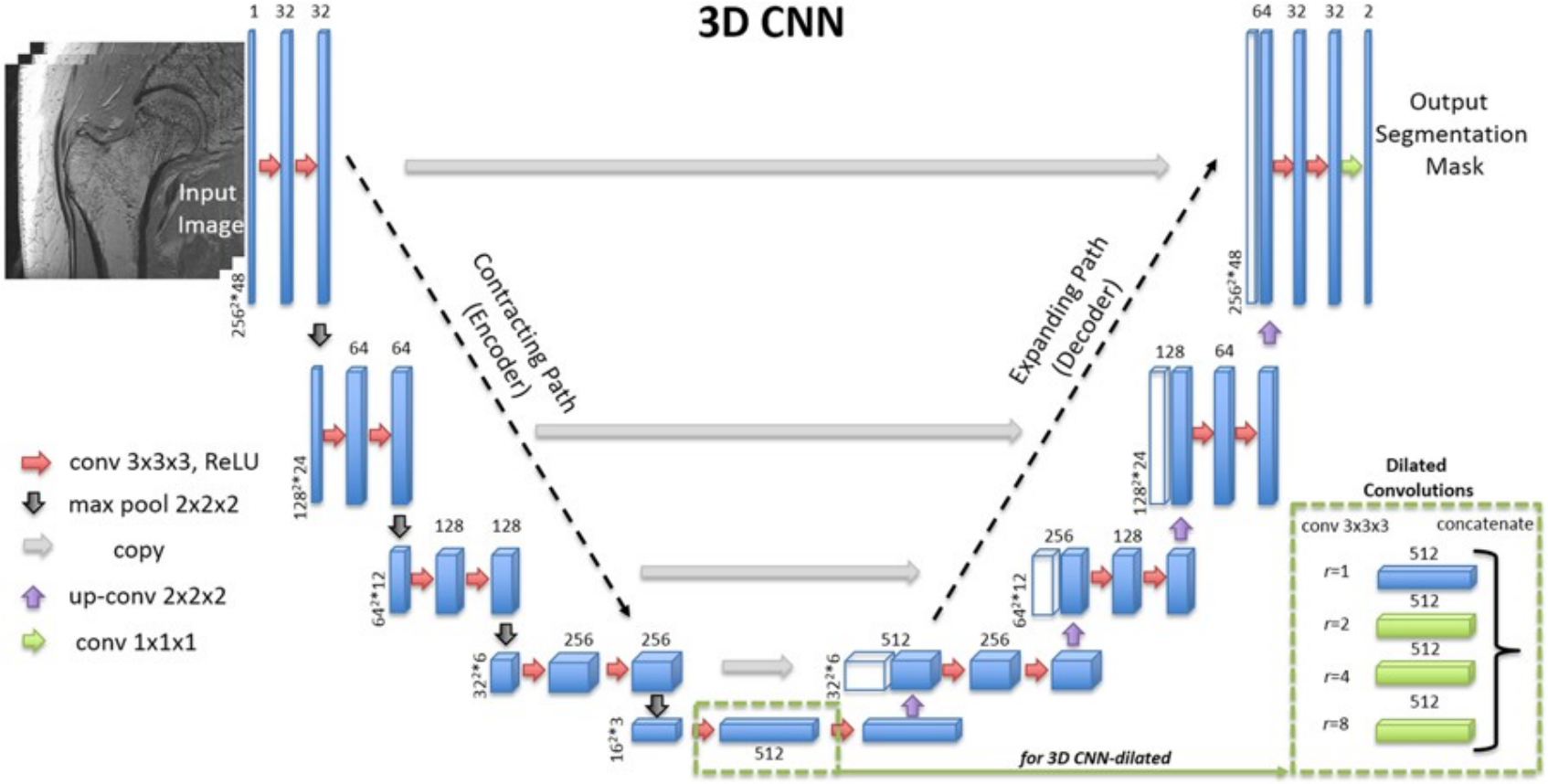
- **Segments and recognizes object instances and regions**
  - [Kirillov arXiv:1901.0244]



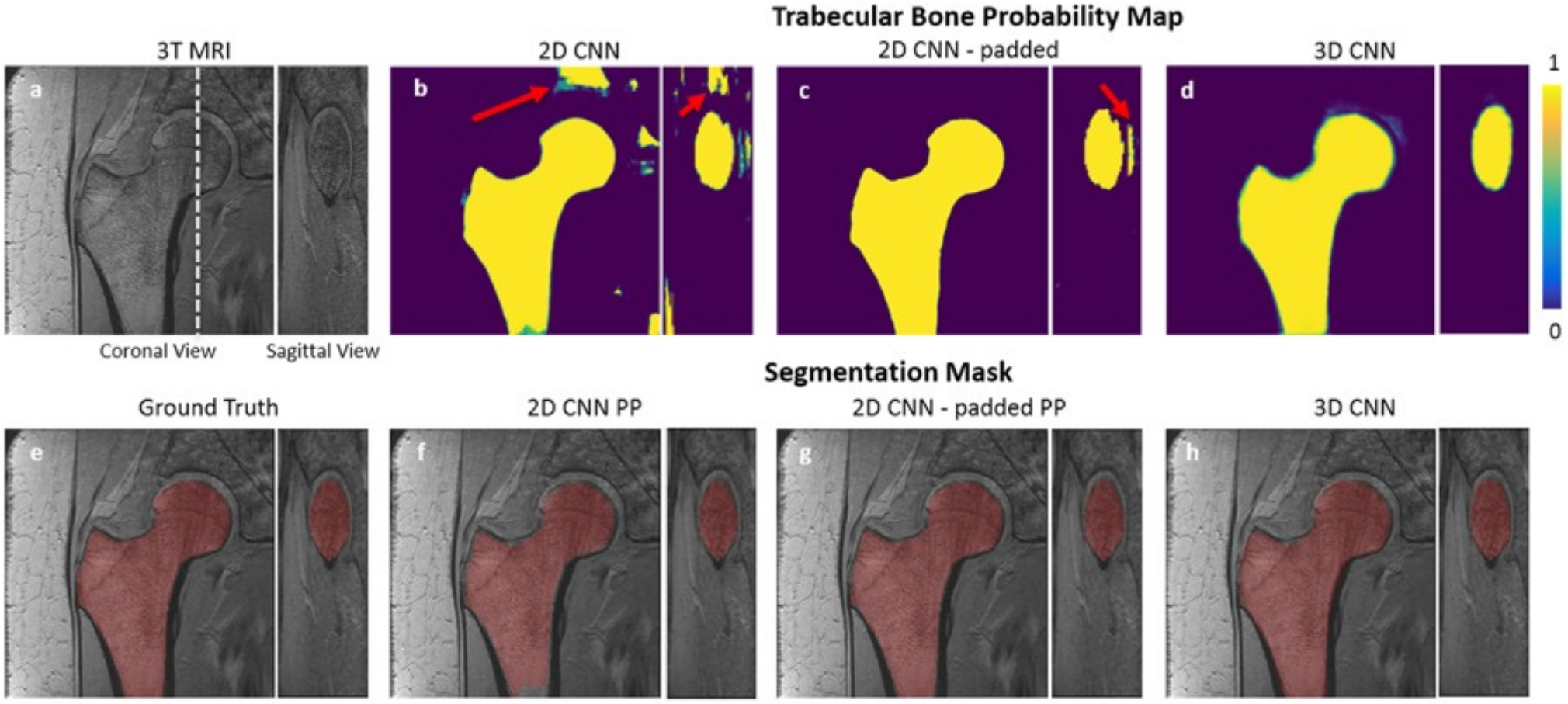


3D ConvNet for Medical Image Analysis (NYU)

- **Segmentation Femur from MR Images**
- **[Deniz et al. Nature 2018]**

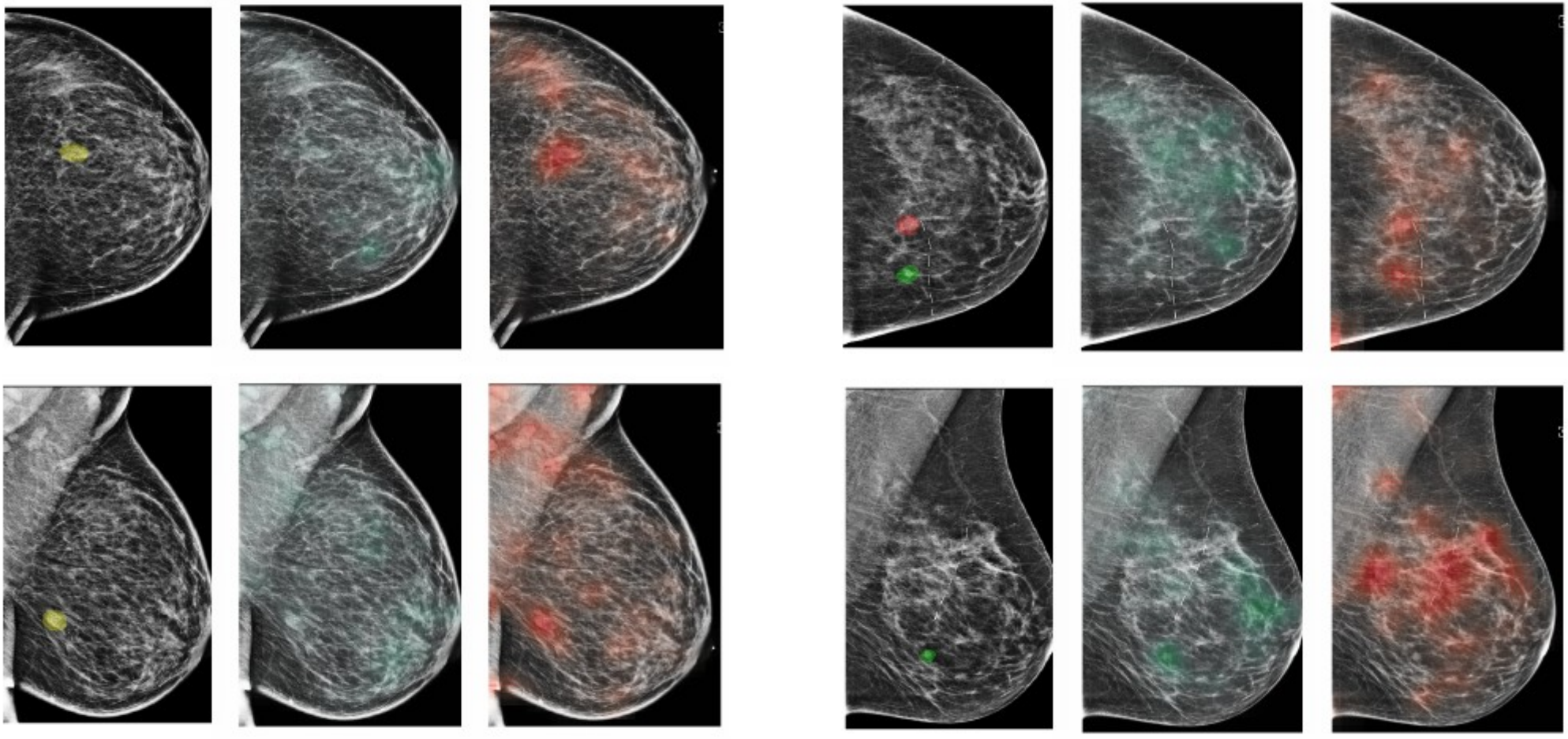


3D ConvNet for Medical Image Analysis



Breast Cancer Detection (NYU)

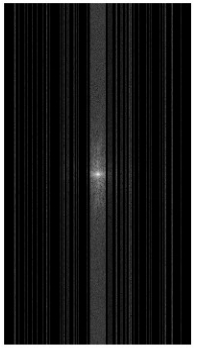
- **[Wu et al. ArXiv:1903.08297] https://github.com/nyukat/breast_cancer_classifier**



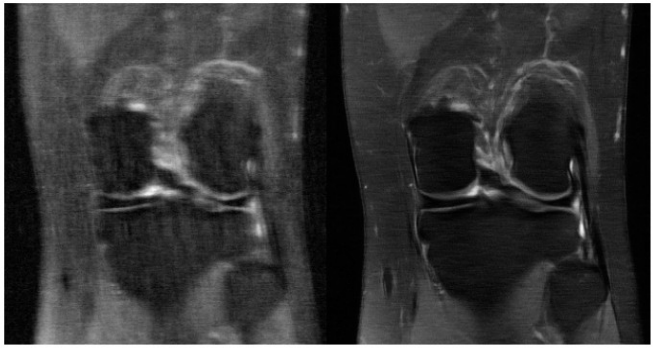
FastMRI (NYU+FAIR): 4x-8x speed up for MRI data acquisition

- MRI images subsampled (in k-space) by 4x and 8x
- [Zbontar et al. ArXiv:1811.08839]
- U-Net architecture
- 4-fold acceleration
- 8-fold acceleration
- K-space masks

(a) Cropped and vertically flipped reconstruction from fully sampled k-space data

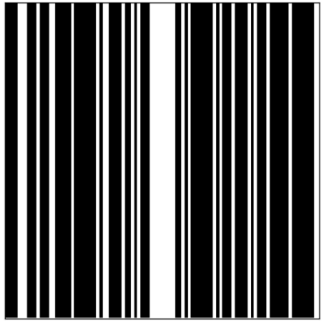


(b) Rectangular masked k-space

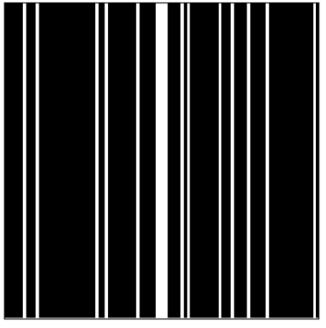


(c) Reconstruction via zero-filled IFFT

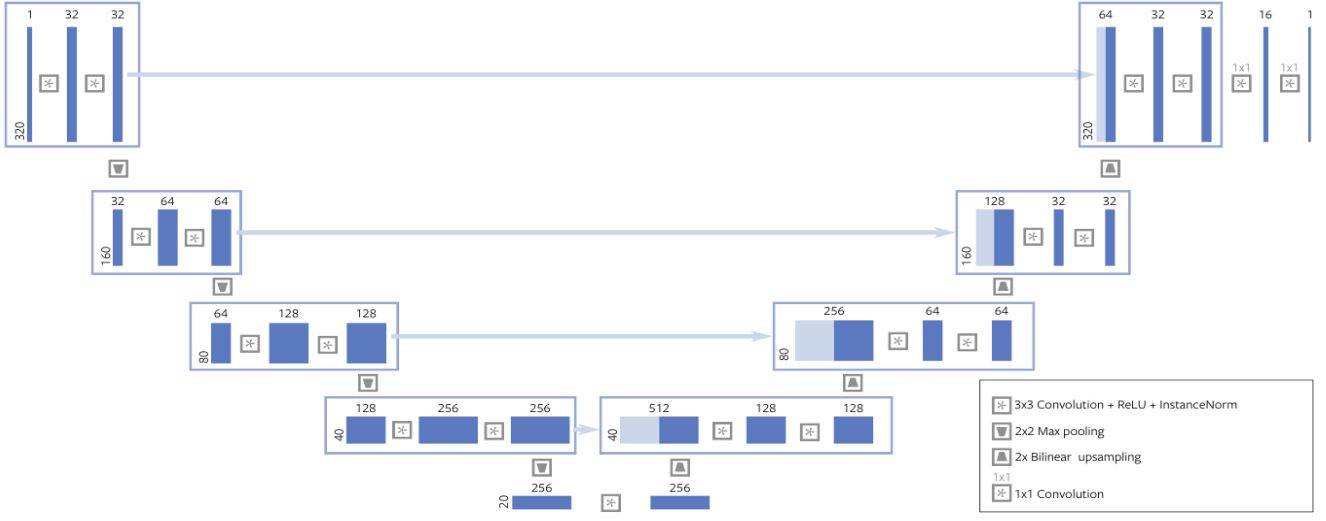
(d) Deep-learning baseline UNET reconstruction



(a) 4-fold acceleration

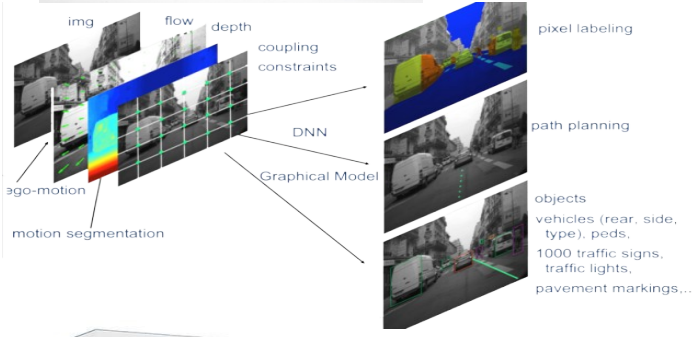


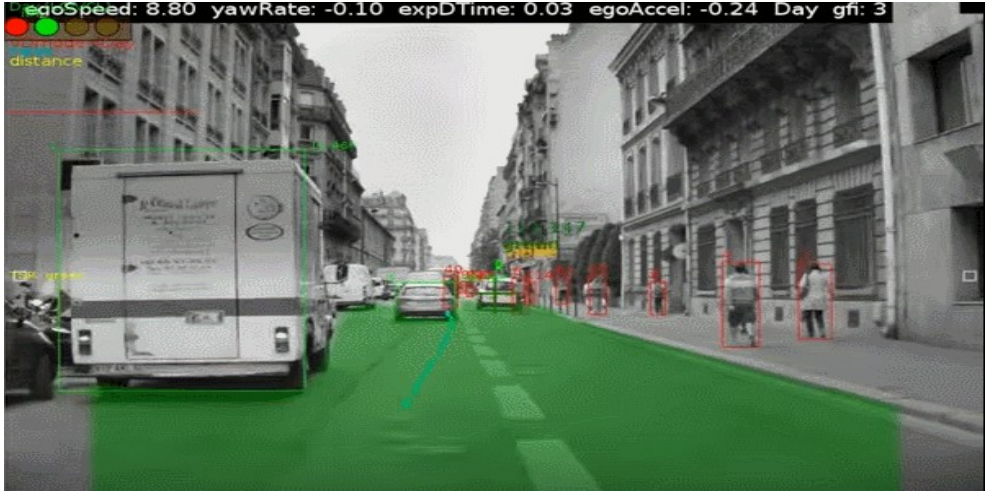
(b) 8-fold acceleration



Driving Cars with Convolutional Nets

- **MobilEye (2015)**





- **NVIDIA**

Learning to Reason?

Memory-Augmented networks, Dynamic networks

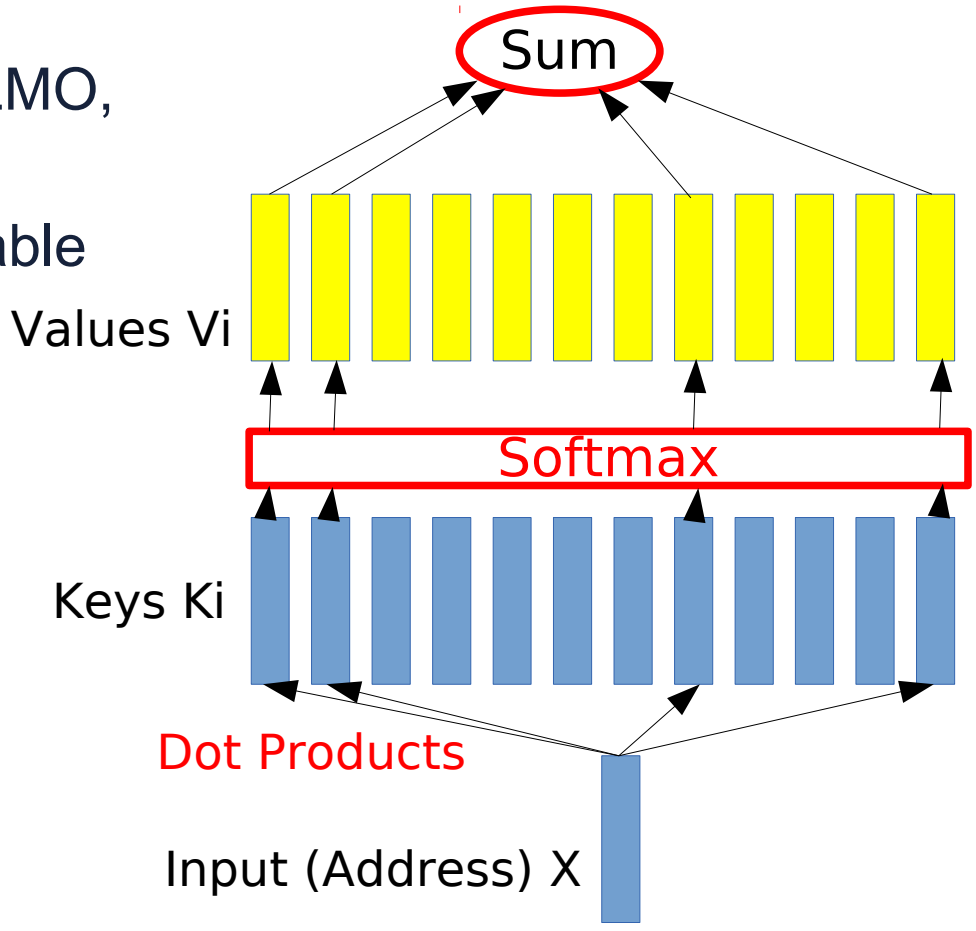
facebook
Artificial Intelligence Research

Differentiable Associative Memory

- Used very widely in NLP
- MemNN, Transformer Network, ELMO, GPT, BERT, GPT2, GLoMO
- Essentially a “soft” RAM or hash table

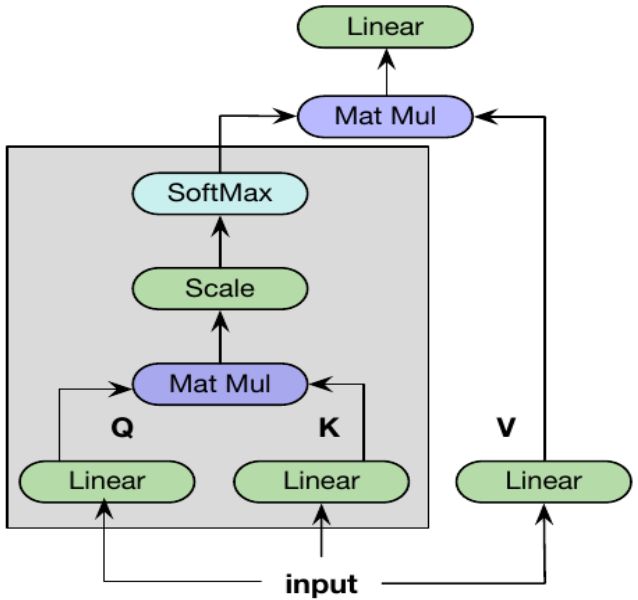
$$C_i = \frac{e^{K_i^T X}}{\sum_j e^{K_j^T X}}$$

$$Y = \sum_i C_i V_i$$

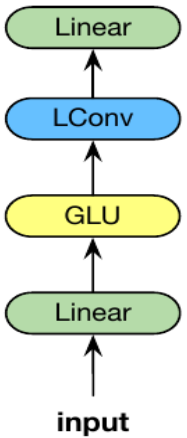


Language Translation with neural nets

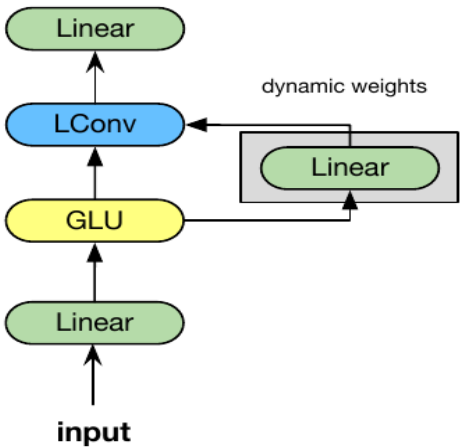
- **Dynamic convolutions with context-dependent kernel**
  - 200M to 300M parameters [Wu et al. ICLR 2019]
  - BLEU scores on WMT datasets: 29.7 En-De; 43.2 En-Fr; 24.4 Zh-En



(a) Self-attention



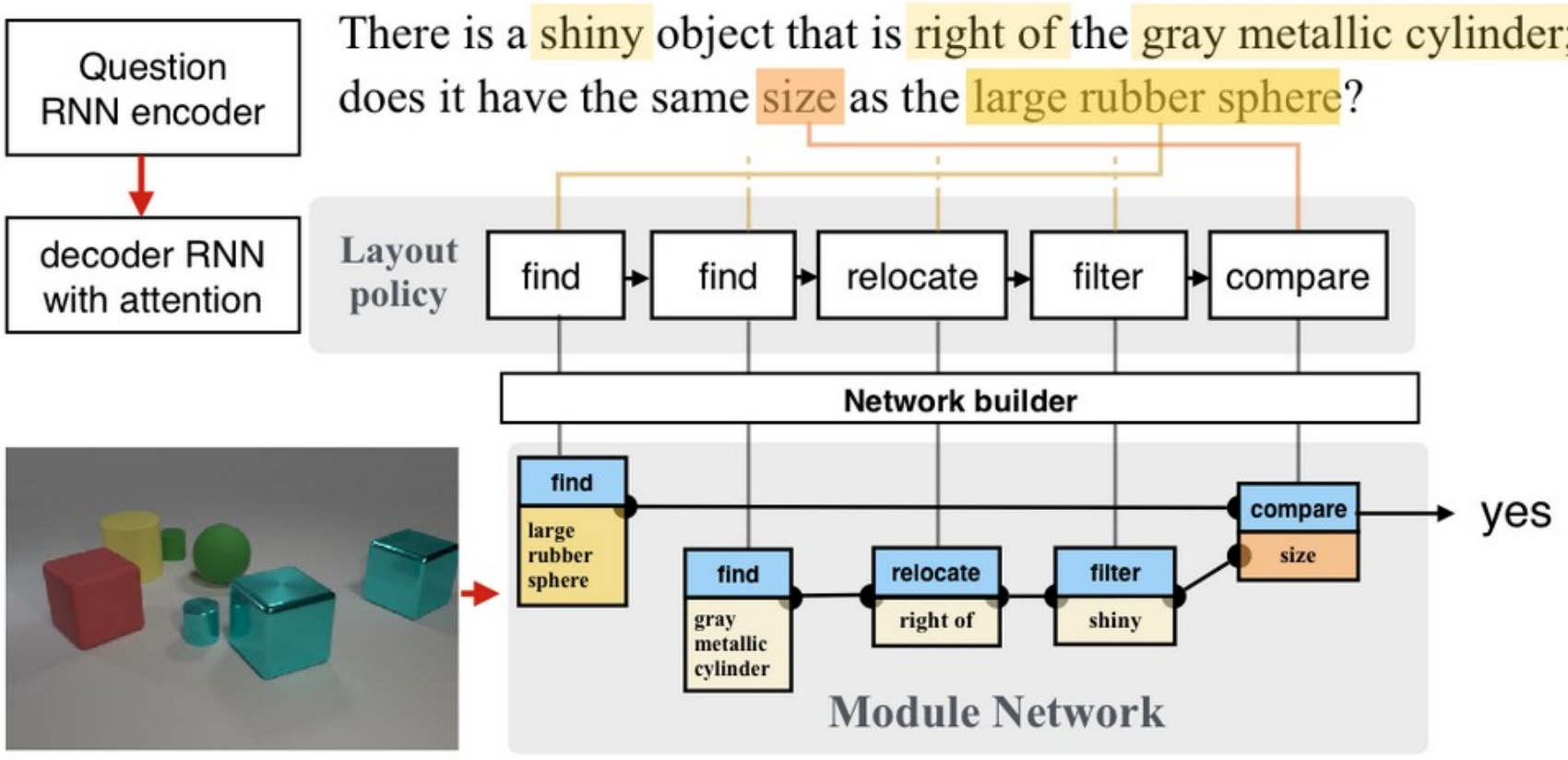
(b) Lightweight convolution



(c) Dynamic convolution

Learning to synthesize neural programs for visual reasoning

https://research.fb.com/visual-reasoning-and-dialog-towards-natural-language-conversations-about-visual-data/



PyTorch: differentiable programming

- **Software 2.0:**
  - The operations in a program are only partially specified
  - They are trainable parameterized modules.
  - The precise operations are learned from data, only the general structure of the program is designed.
- **Dynamic computational graph**
  - Automatic differentiation by recording a “tape” of operations and rolling it backwards with the Jacobian of each operator.
  - Implemented in PyTorch1.0, Chainer…
  - Easy if the front-end language is dynamic and interpreted (e.g Python)
  - Not so easy if we want to run without a Python runtime...

How do Humans and Animal Learn?

So quickly

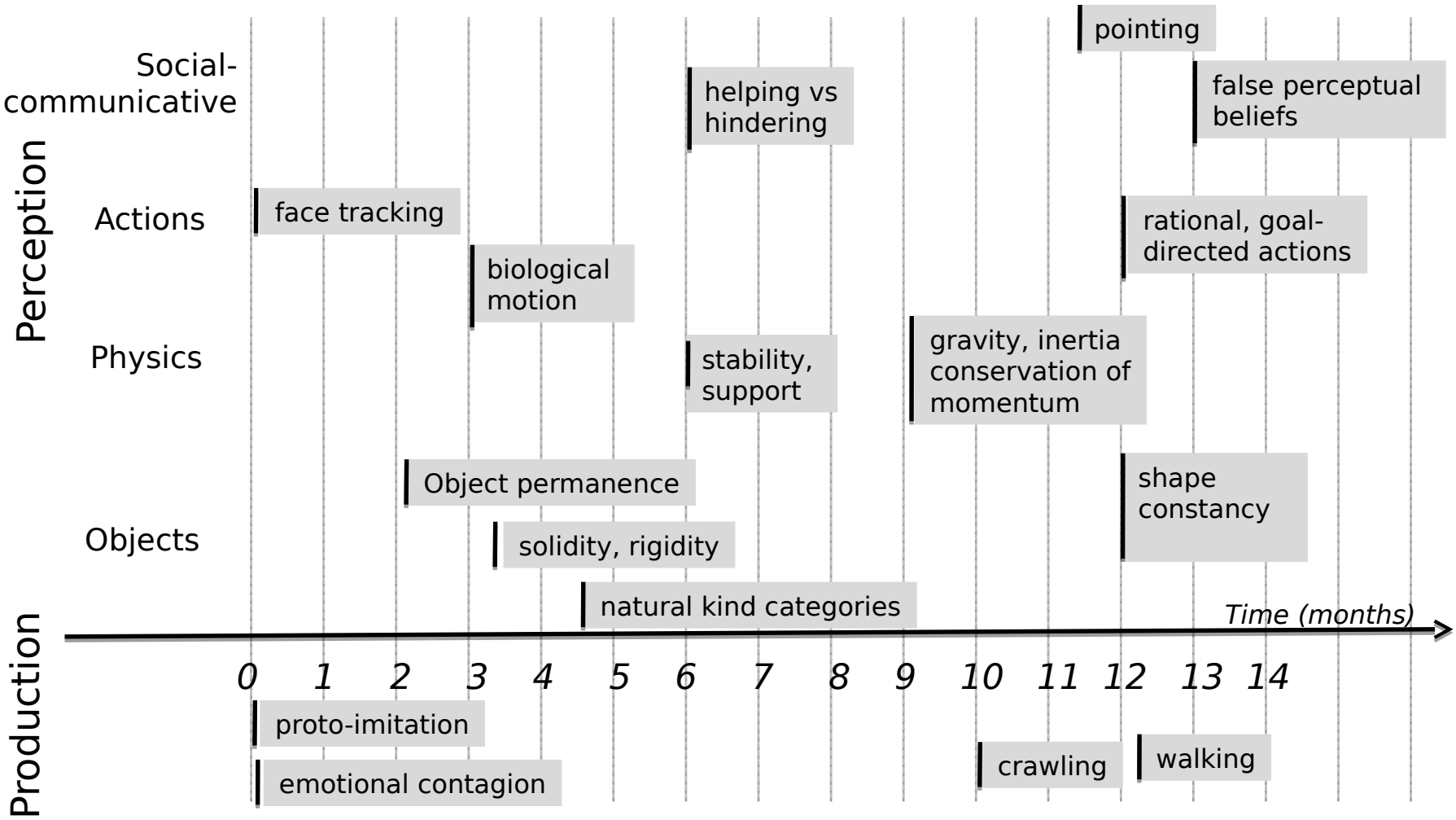
facebook
Artificial Intelligence Research

Babies learn how the world works by observation

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

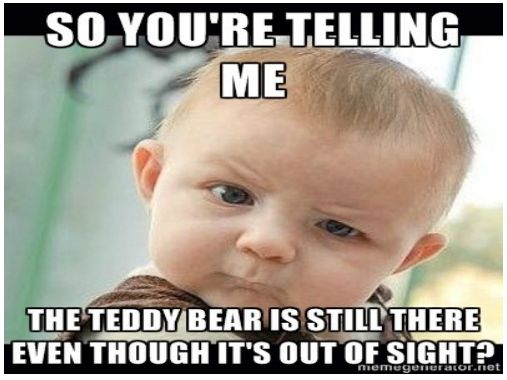
Photos courtesy of Emmanuel Dupoux

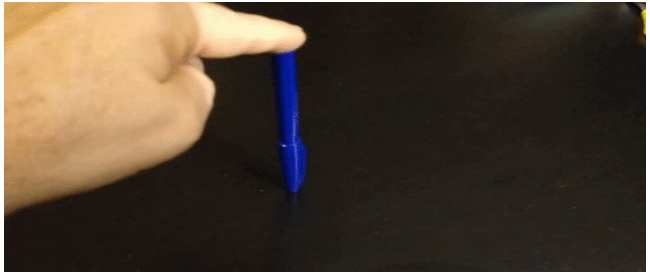
Early Conceptual Acquisition in Infants [from Emmanuel Dupoux]

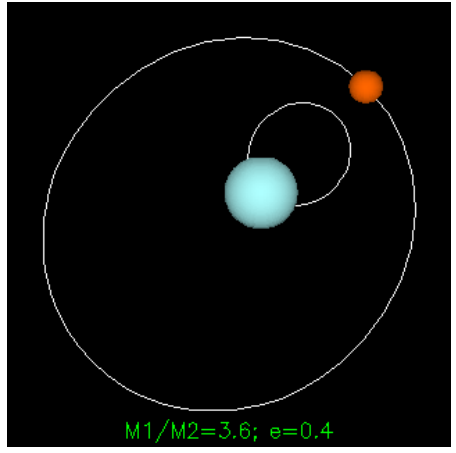


Prediction is the essence of Intelligence

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







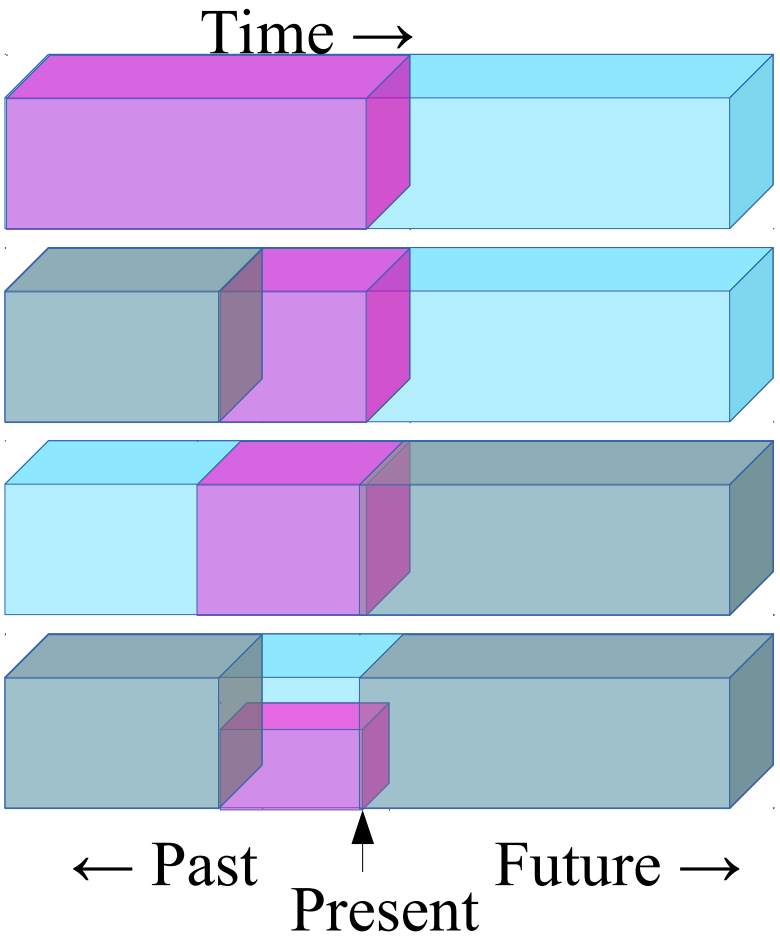
The Salvation?
Self-Supervised Learning

Training very large networks to Understand the world through prediction

facebook
Artificial Intelligence Research

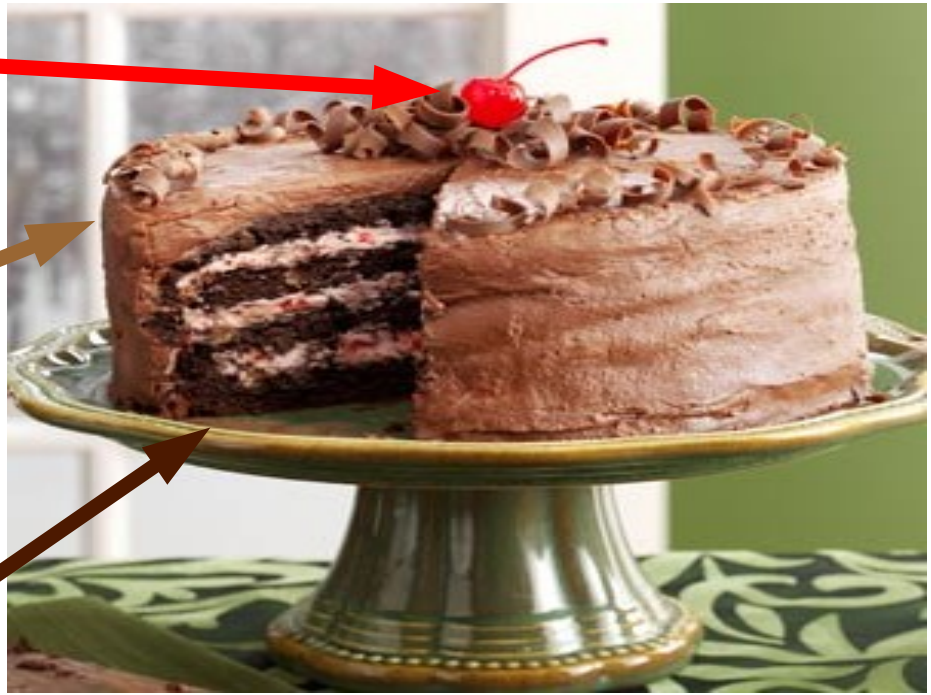
Self-Supervised Learning: Prediction & Reconstruction

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



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**



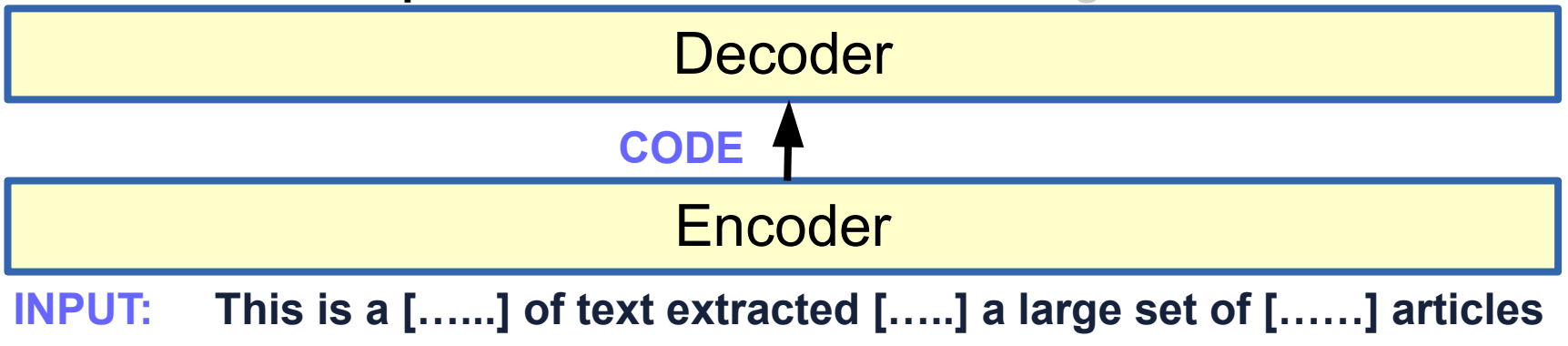
The Next AI Revolution

Get the T-shirt!

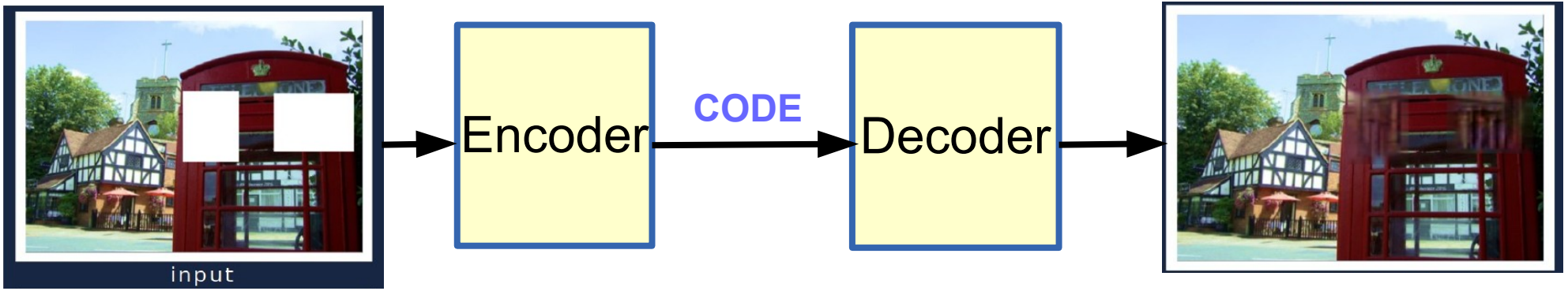
With thanks to Alyosha Efros and Gil Scott Heron

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 works well for text

- **Word2vec**
  - [Mikolov 2013]
- **FastText**
  - [Joulin 2016] (FAIR)
- **BERT**
  - Bidirectional Encoder Representations from Transformers
  - [Devlin 2018]
- **Cloze-Driven Auto-Encoder**
  - [Baevski 2019] (FAIR)

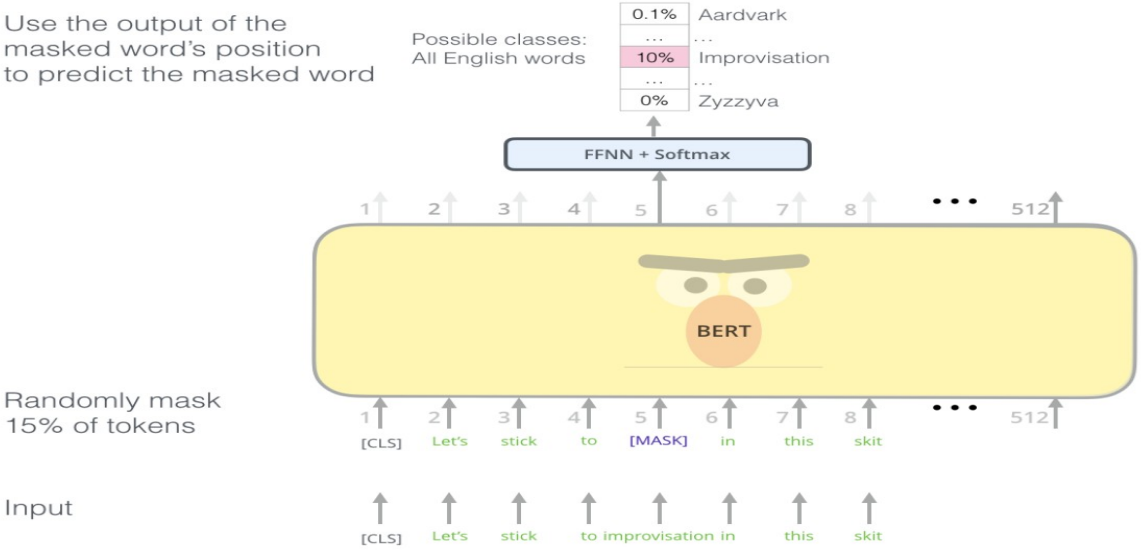
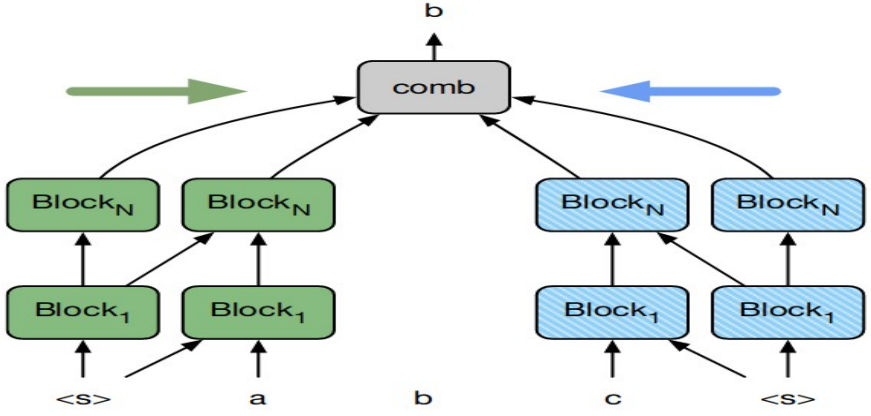
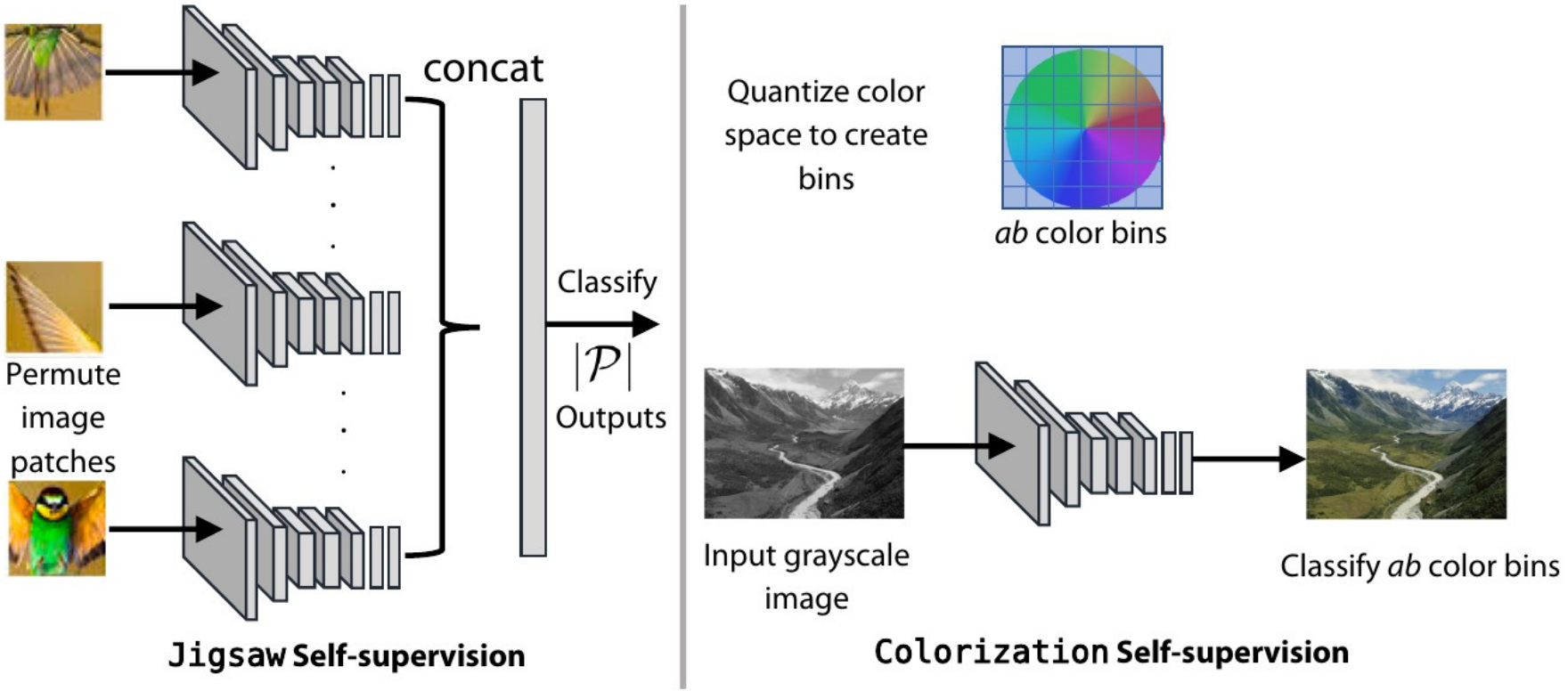


Figure credit: Jay Alammar http://jalammar.github.io/illustrated-bert/



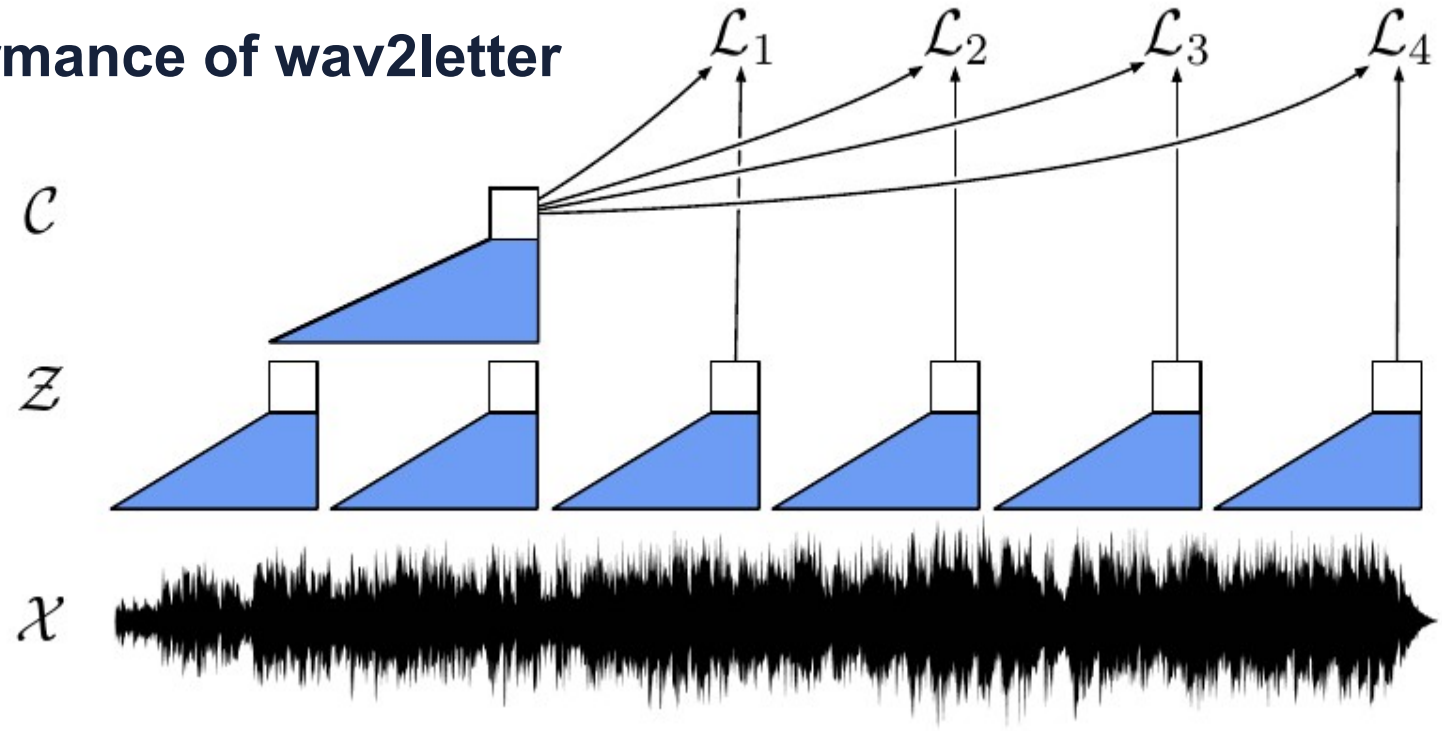
SSL for Vision [Goyal, Mahavan, Misra, Gupta 2019]

- **Self-supervised: jigsaw problem and colorization problem**



Wav2vec [Schneider et al. 2019]

- **Self-supervised speech features**
- **Classify a future audio segment as compatible or not**
- **Similar to jigsaw scenario**
- **Improves performance of wav2letter**



SSL works with discrete data

- **BERT / LM: discrete distribution on words**
- **Colorization: discrete distribution on quantized color bins**
- **Jigsaw: fundamentally a classification problem**
- **Wav2vec: fundamentally a 2-class classification problem**

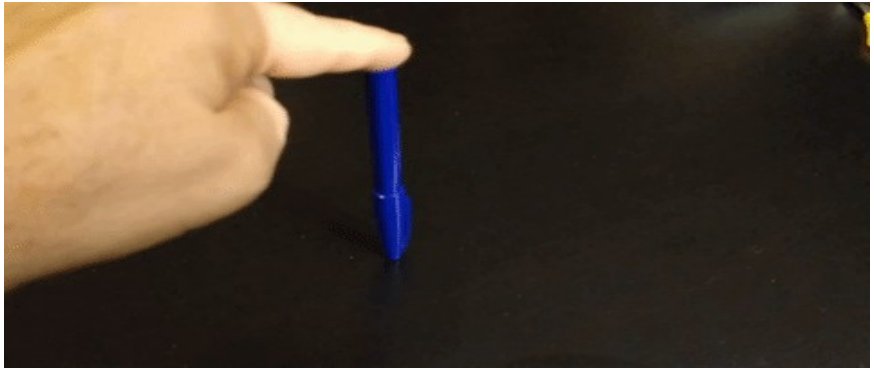
- **SSL works because we know how to represent uncertainty with discrete distributions (see also [Ranzato 2014] on video prediction)**

- **But how do we make it work with high-dimensional continuous data?**
  - Video prediction, audio prediction….

Self-Supervised Learning: Filling in the Blanks

The world is not entirely predictable / stochastic

- **Video prediction:**
  - Multiple futures are possible.
  - Training a system to make a single prediction results in “blurry” results
  - the average of all the possible futures



SSL through Video Prediction

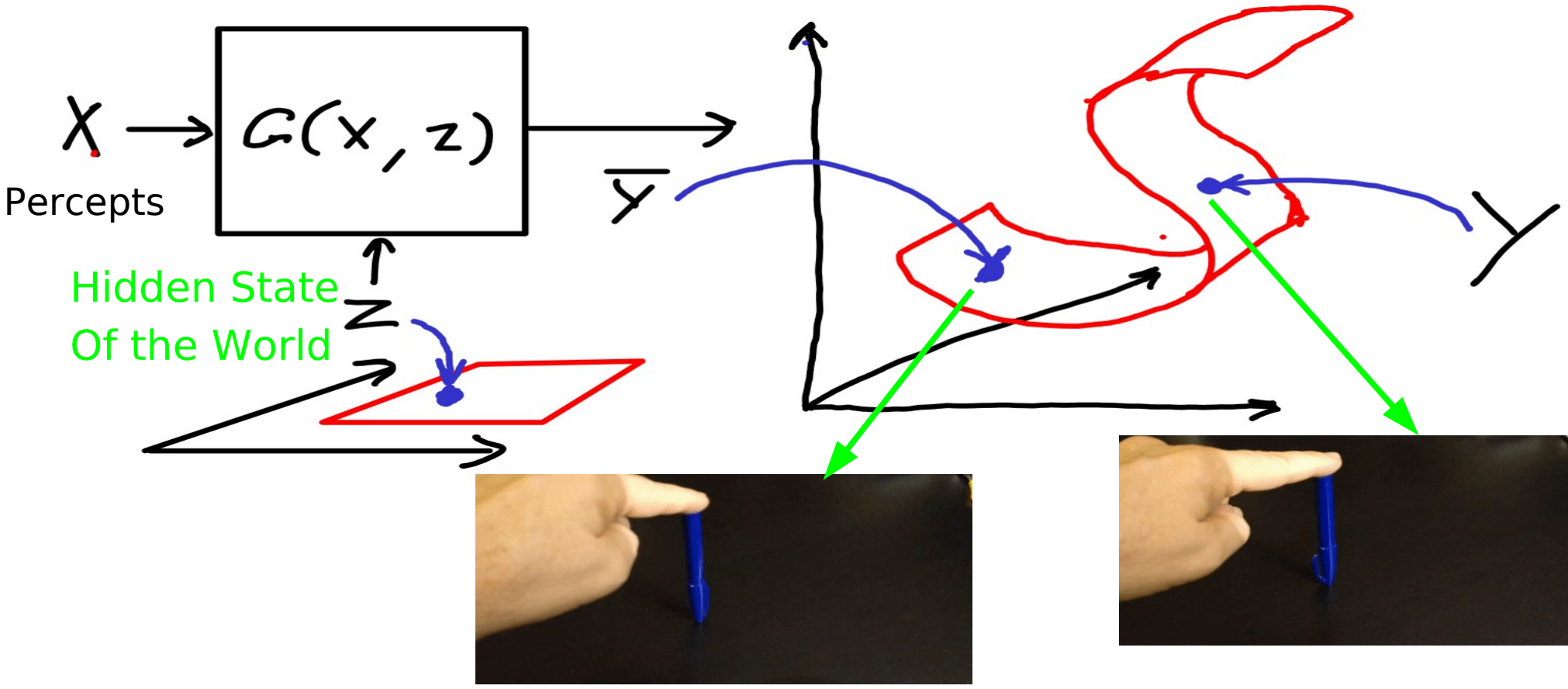
- **Some success**
  - [Mathieu, Couprie, YLC ICLR'16 arXiv:1511:05440]
  - [Luc, Couprie, LeCun, Verbeek ECCV 2018]
- **But we are far from a complete solution**

Adversarial Training & Video Prediction

facebook
Artificial Intelligence Research

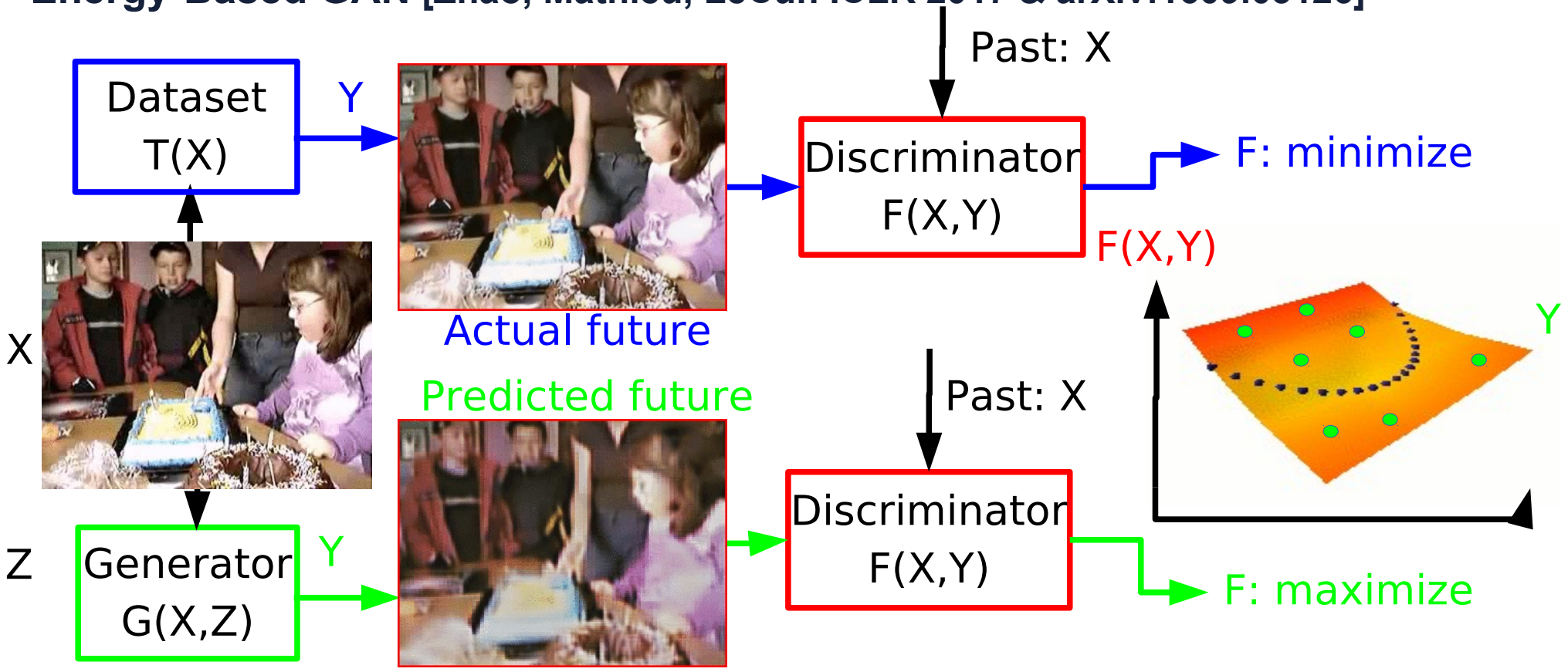
The Hard Part: Prediction Under Uncertainty

- **Invariant prediction: The training samples are merely representatives of a whole set of possible outputs (e.g. a manifold of outputs).**



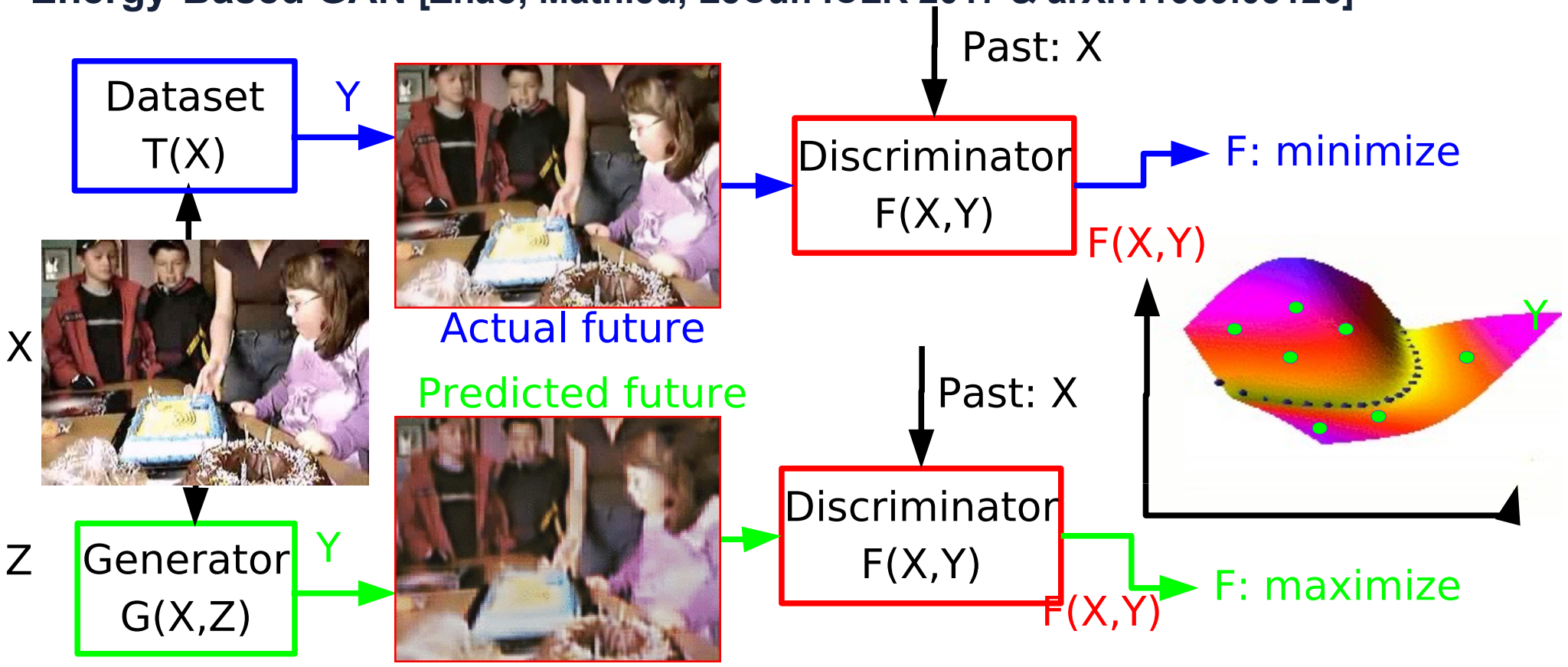
Adversarial Training: the key to prediction under uncertainty?

- **Generative Adversarial Networks (GAN) [Goodfellow et al. NIPS 2014],**
- **Energy-Based GAN [Zhao, Mathieu, LeCun ICLR 2017 & arXiv:1609.03126]**



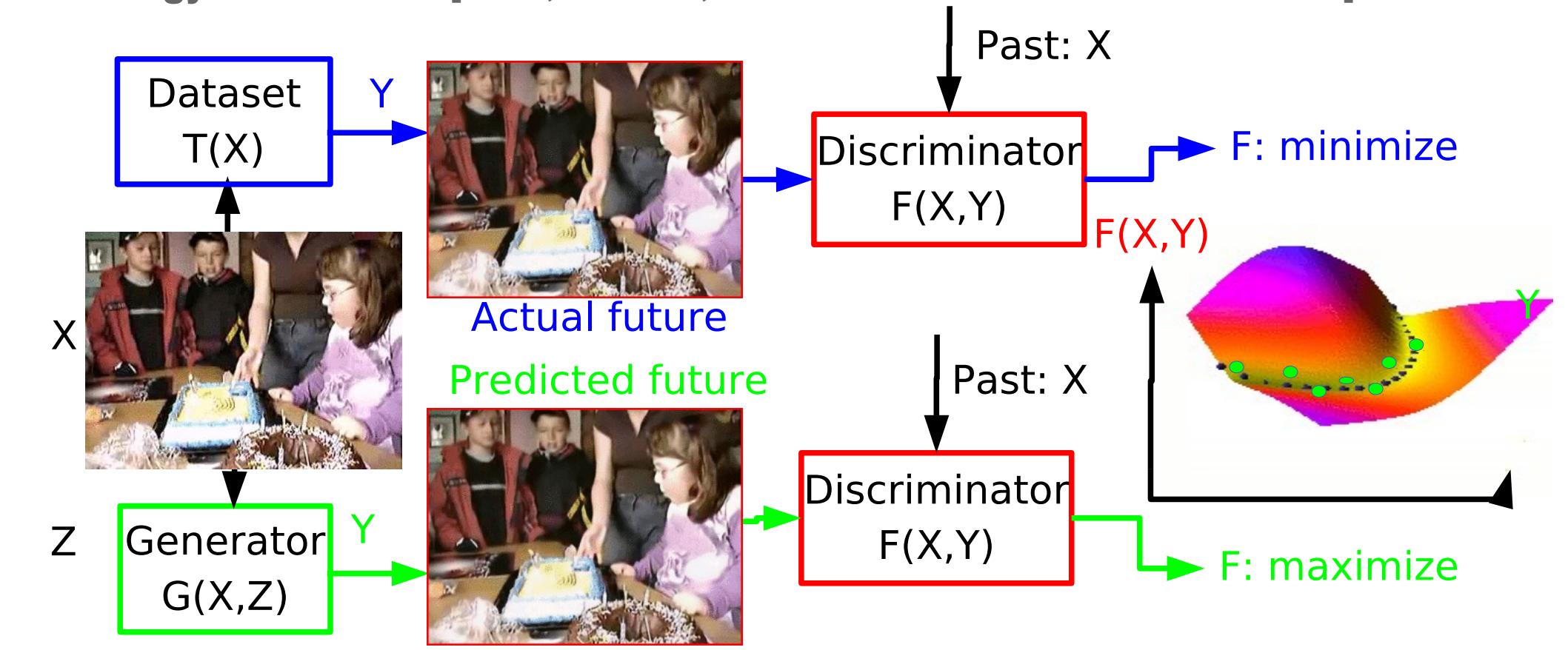
Adversarial Training: the key to prediction under uncertainty?

- **Generative Adversarial Networks (GAN) [Goodfellow et al. NIPS 2014],**
- **Energy-Based GAN [Zhao, Mathieu, LeCun ICLR 2017 & arXiv:1609.03126]**



Adversarial Training: the key to prediction under uncertainty?

- **Generative Adversarial Networks (GAN) [Goodfellow et al. NIPS 2014],**
- **Energy-Based GAN [Zhao, Mathieu, LeCun ICLR 2017 & arXiv:1609.03126]**



Faces “invented” by a GAN (Generative Adversarial Network)

- **Random vector → Generator Network → output image** [Goodfellow NIPS 2014]
  [Karras et al. ICLR 2018] (from NVIDIA)

Generative Adversarial Networks for Creation

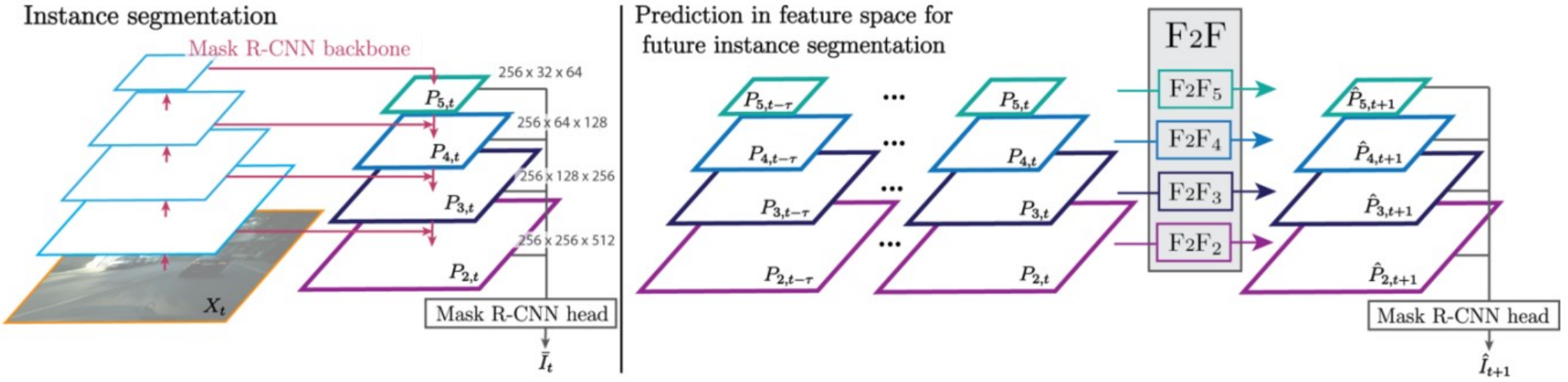
- **[Sbai 2017]**

Self-supervised Adversarial Learning for Video Prediction

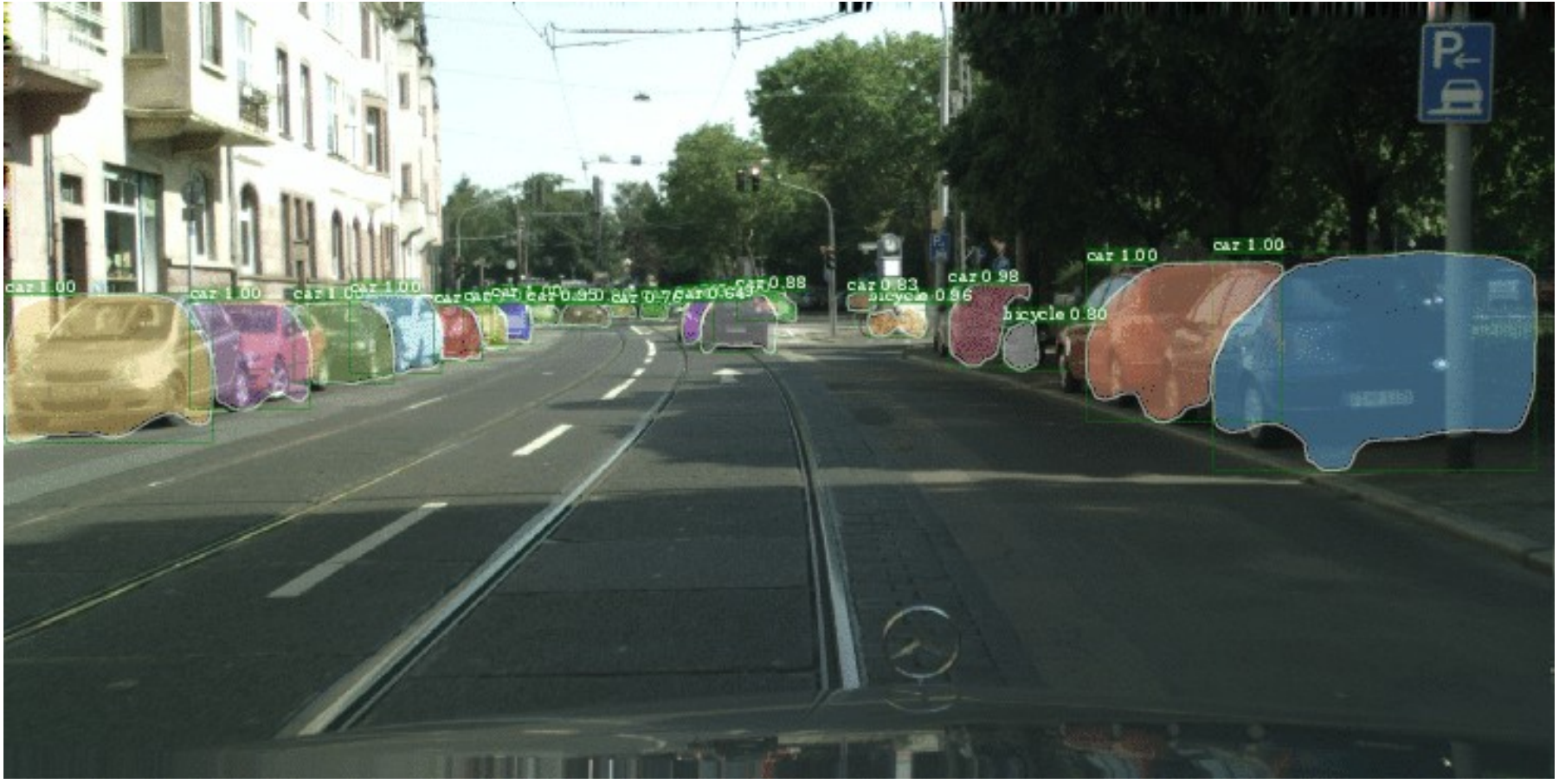
- **Our brains are "prediction machines"**
- **Can we train machines to predict the future?**
- **Some success with "adversarial training"**
  - [Mathieu, Couprie, LeCun arXiv:1511:05440]
- **But we are far from a complete solution.**

Predicting Instance Segmentation Maps

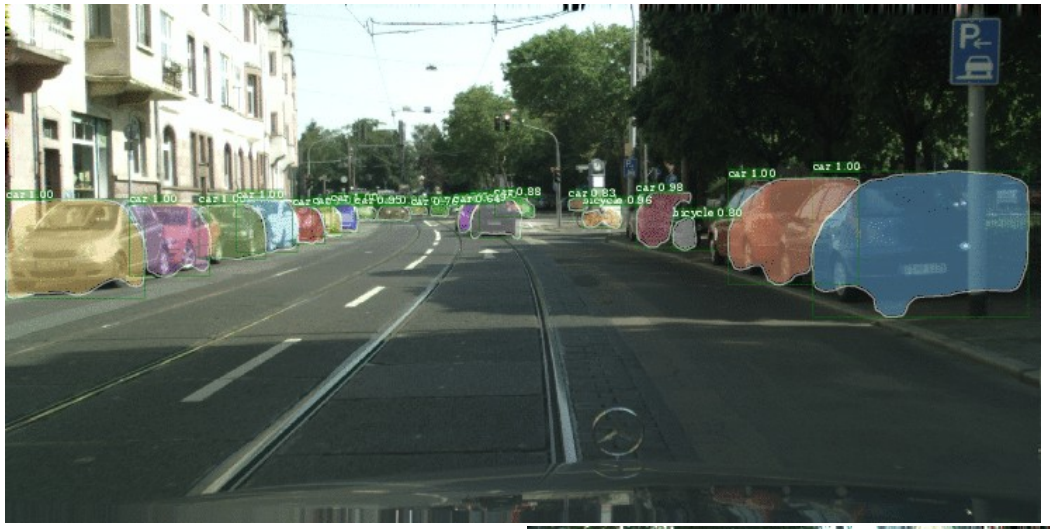
- [Luc, Couprie, LeCun, Verbeek ECCV 2018]
- Mask R-CNN Feature Pyramid Network backbone
- Trained for instance segmentation on COCO
- Separate predictors for each feature level

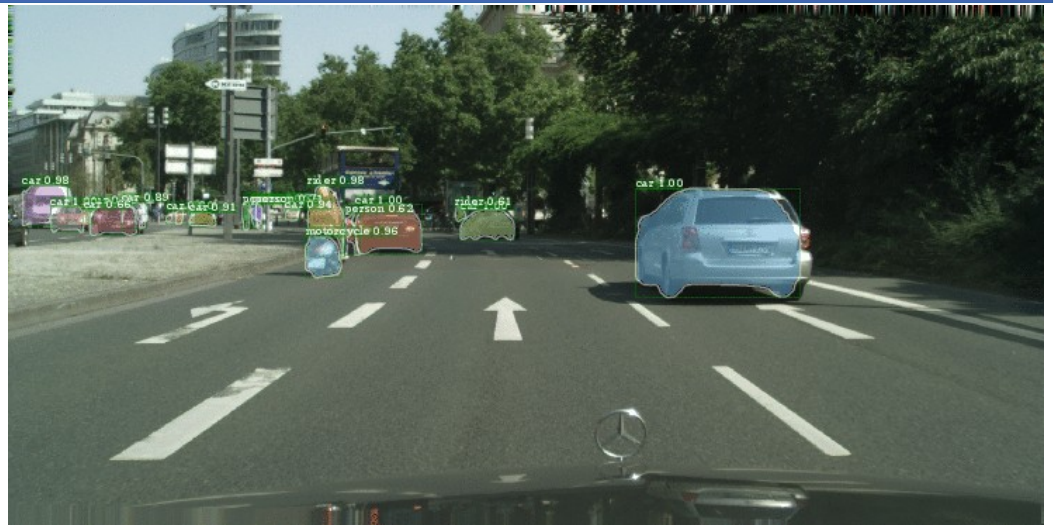


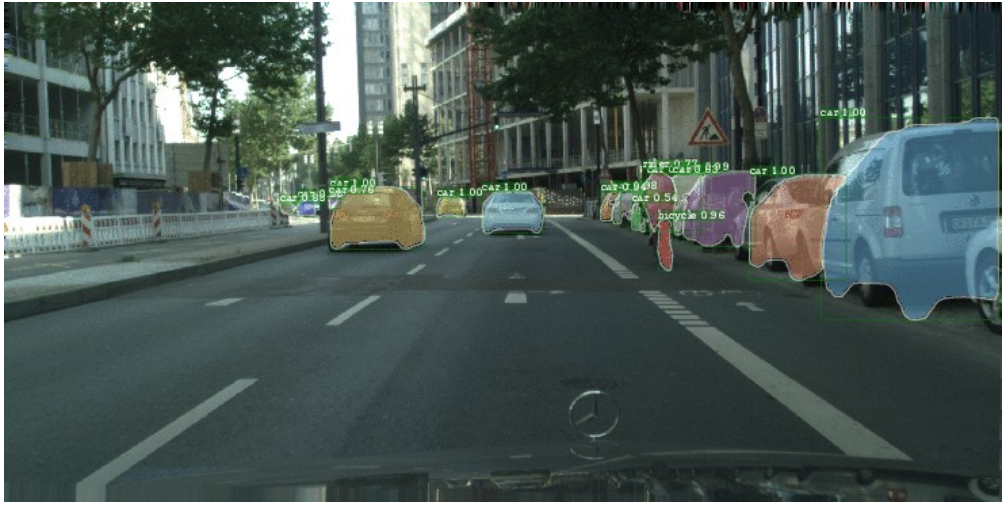
Predictions



Long-term predictions (10 frames, 1.8 seconds)







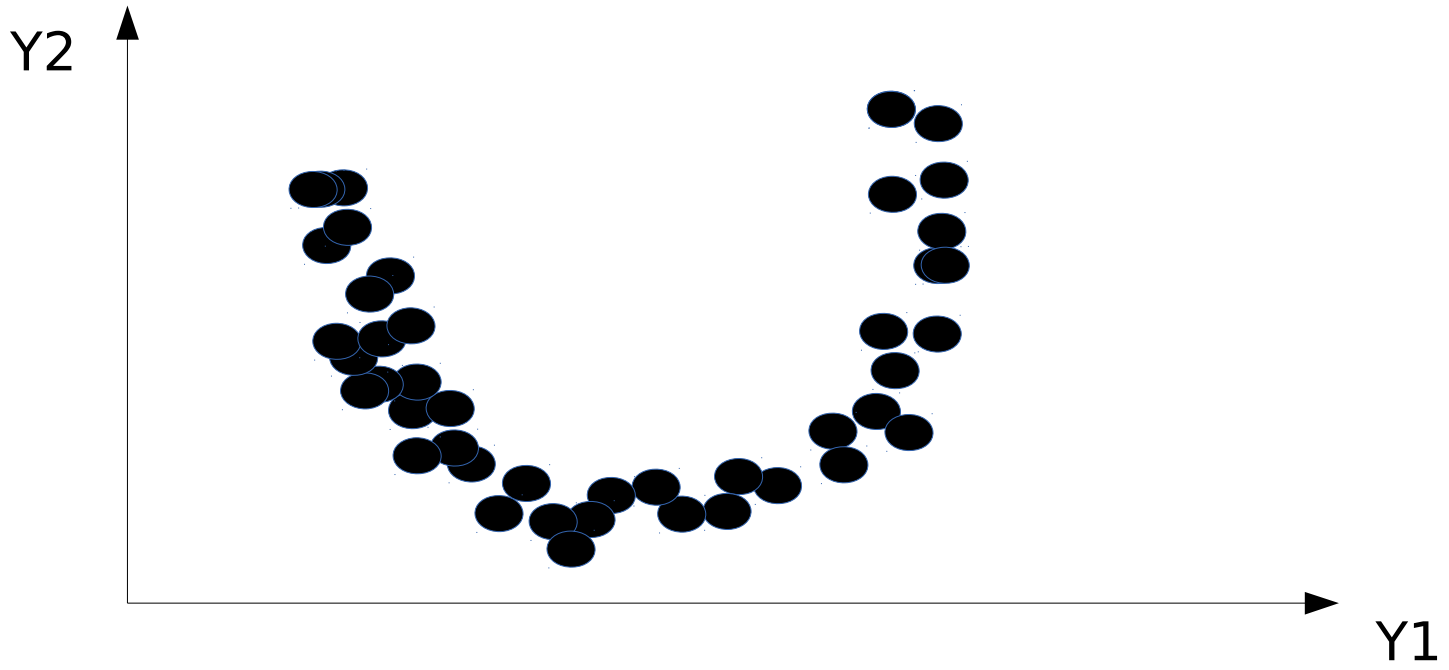
Energy-Based Self-Supervised Learning

Regularized Auto-Encoders

facebook
Artificial Intelligence Research

Energy-Based Unsupervised Learning

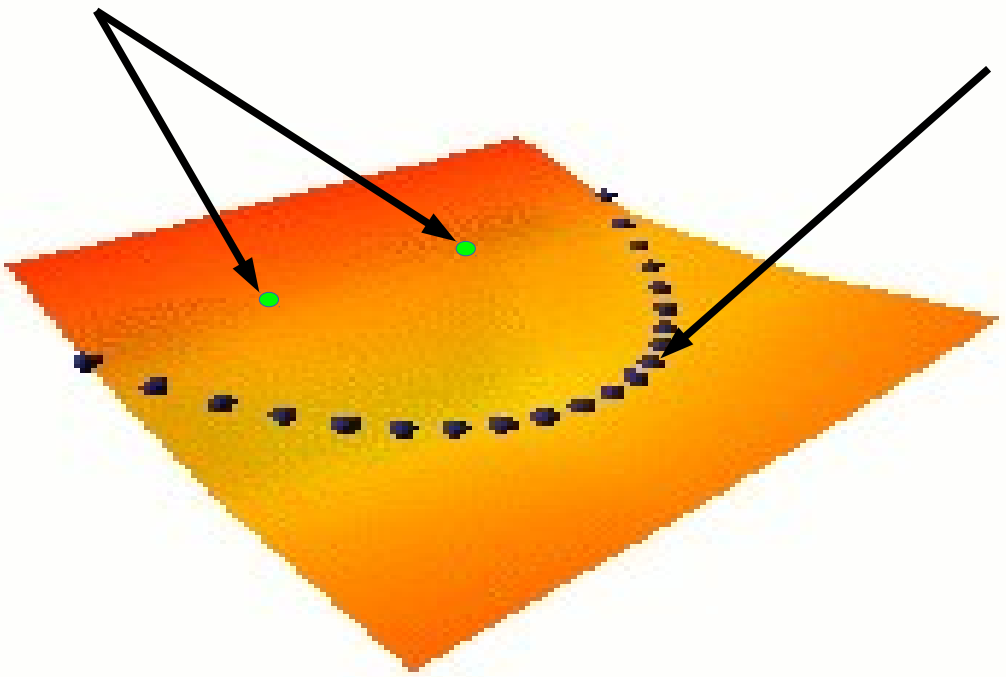
- **Learning an energy function (or contrast function) that takes**
  - Low values on the data manifold
  - Higher values everywhere else



Energy-Based Unsupervised Learning

- **Energy Function:** Takes low value on data manifold, higher values everywhere else
- Push down on the energy of desired outputs. Push up on everything else.
- But how do we choose where to push up?

Implausible futures (high energy)



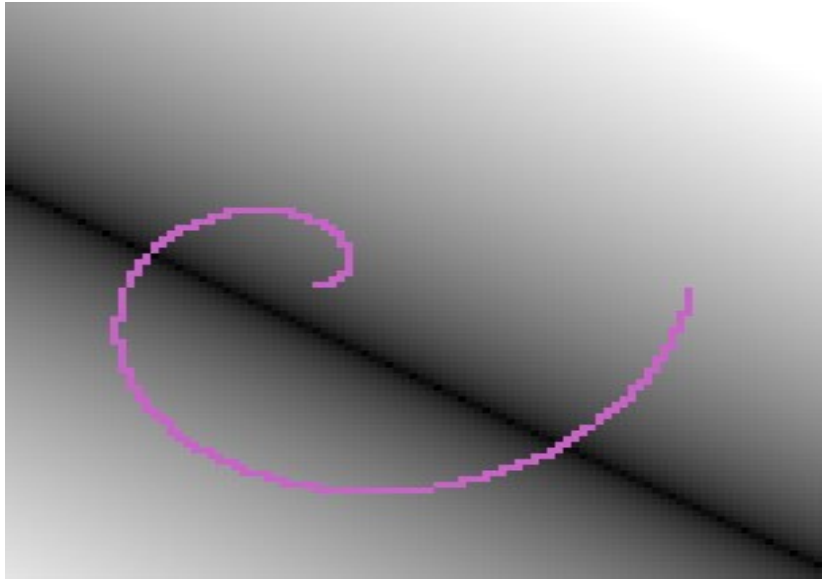
Plausible futures (low energy)

Energy surface for PCA and K-means

- **1. build the machine so that the volume of low energy stuff is constant**
  - PCA, K-means, GMM, square ICA...

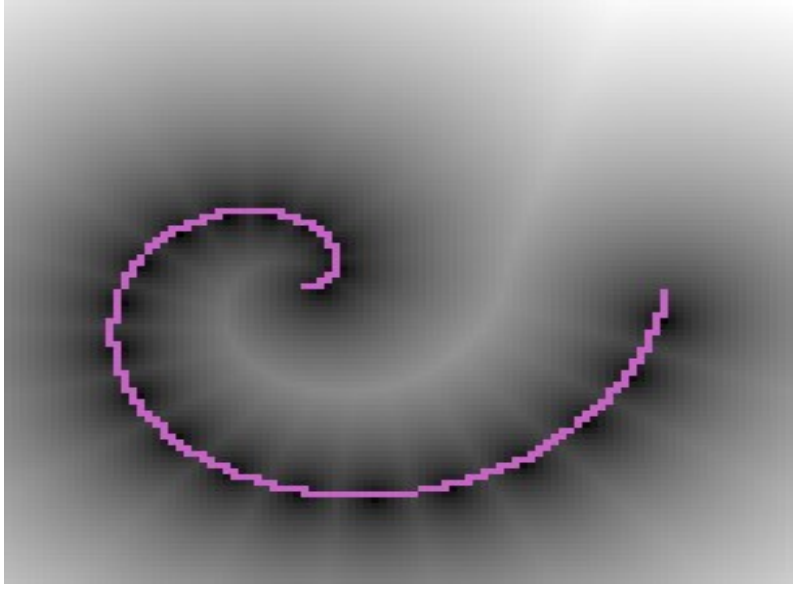
PCA

$$E(Y) = ||W^T WY - Y||^2$$



K-Means,
Z constrained to 1-of-K code

$$E(Y) = min_z \sum_i ||Y - W_i Z_i||^2$$



Seven Strategies to Shape the Energy Function

- **1. build the machine so that the volume of low energy stuff is constant**
  - PCA, K-means, GMM, square ICA
- **2. push down of the energy of data points, push up everywhere else**
  - Max likelihood (needs tractable partition function or variational approximation)
- **3. push down of the energy of data points, push up on chosen locations**
  - Contrastive divergence, Ratio Matching, Noise Contrastive Estimation, Min Probability Flow, adversarial generator/GANs
- **4. minimize the gradient and maximize the curvature around data points** score matching
- **5. if E(Y) = ||Y - G(Y)||^2, make G(Y) as "constant" as possible.**
  - Contracting auto-encoder, saturating auto-encoder
- **6. train a dynamical system so that the dynamics goes to the data manifold**
  - denoising auto-encoder, masked auto-encoder (e.g. BERT)
- **7. use a regularizer that limits the volume of space that has low energy**
  - Sparse coding, sparse auto-encoder, LISTA & PSD, Variational auto-encoders

Latent Variable Models

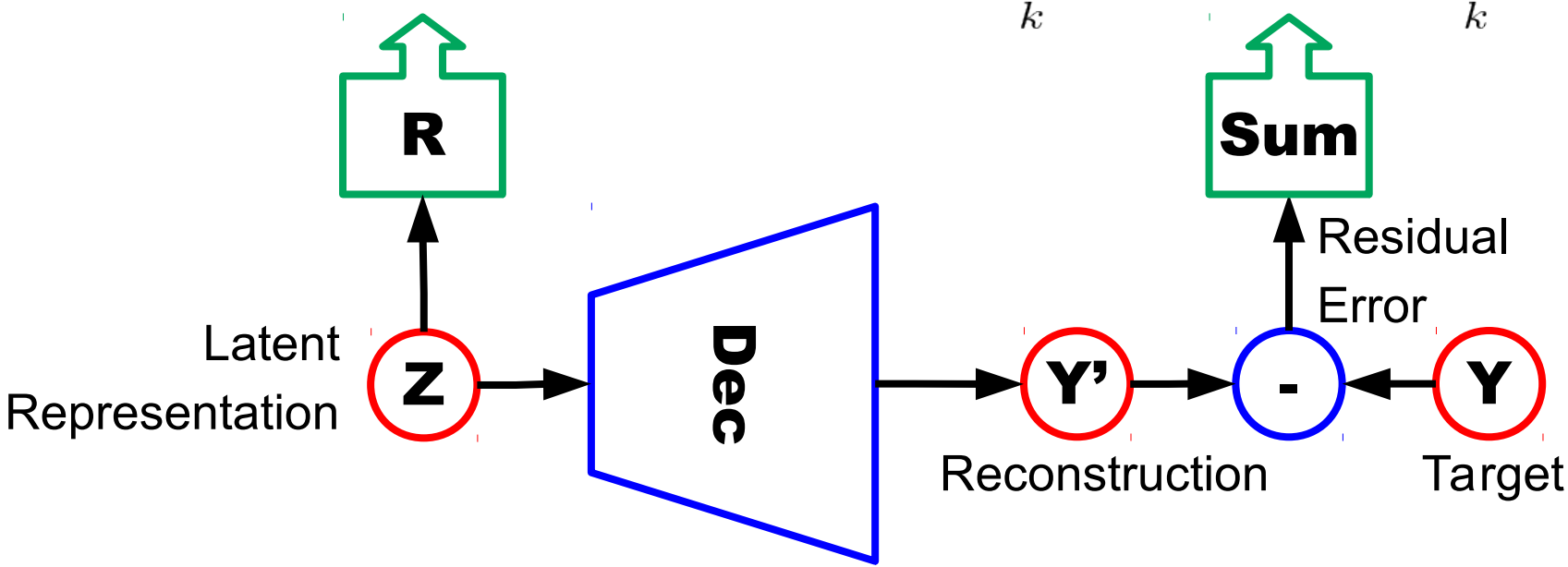
Regularized Auto-Encoders and such

facebook
Artificial Intelligence Research

The “Decoder with Regularized Latent Variable” Model

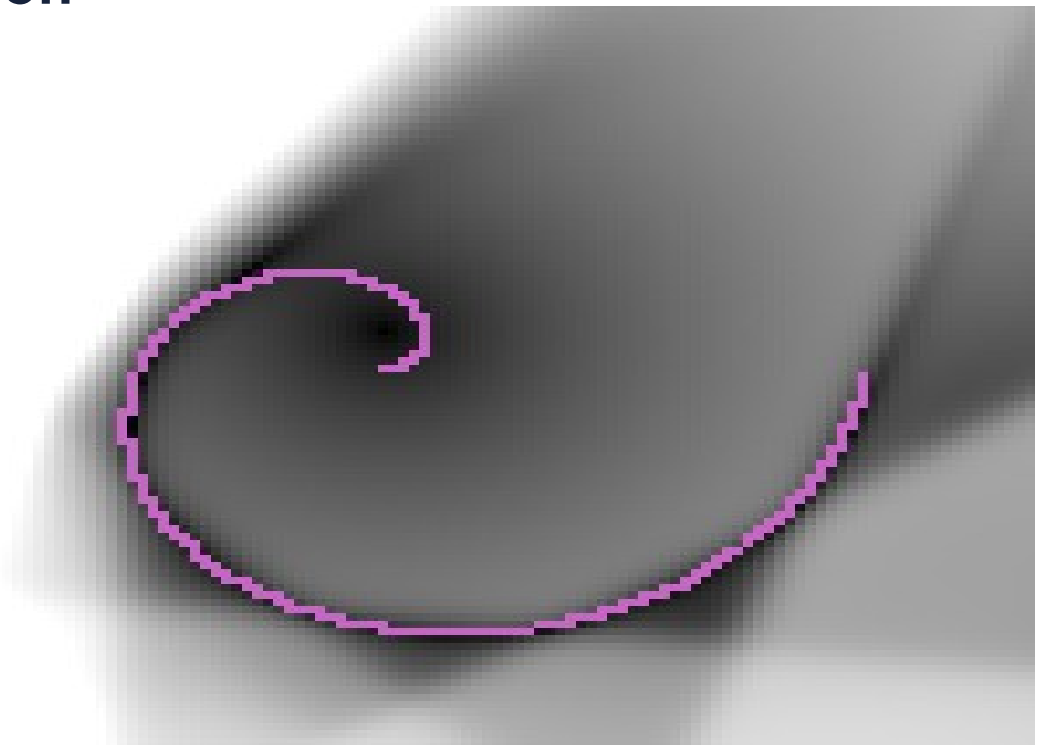
- **Y’ = Dec(Z) Z* = argmin || Y – Dec(Z) || + R(Z)**
- **Linear decoder: K-Means, basis pursuit, K-SVD, sparse coding,….**
- **Sparse modeling: [Olshausen Field 1997]**

$$E(Y,Z) = ||Y - \sum_k W_k Z_k||^2 + \alpha \sum_k |Z_k|$$



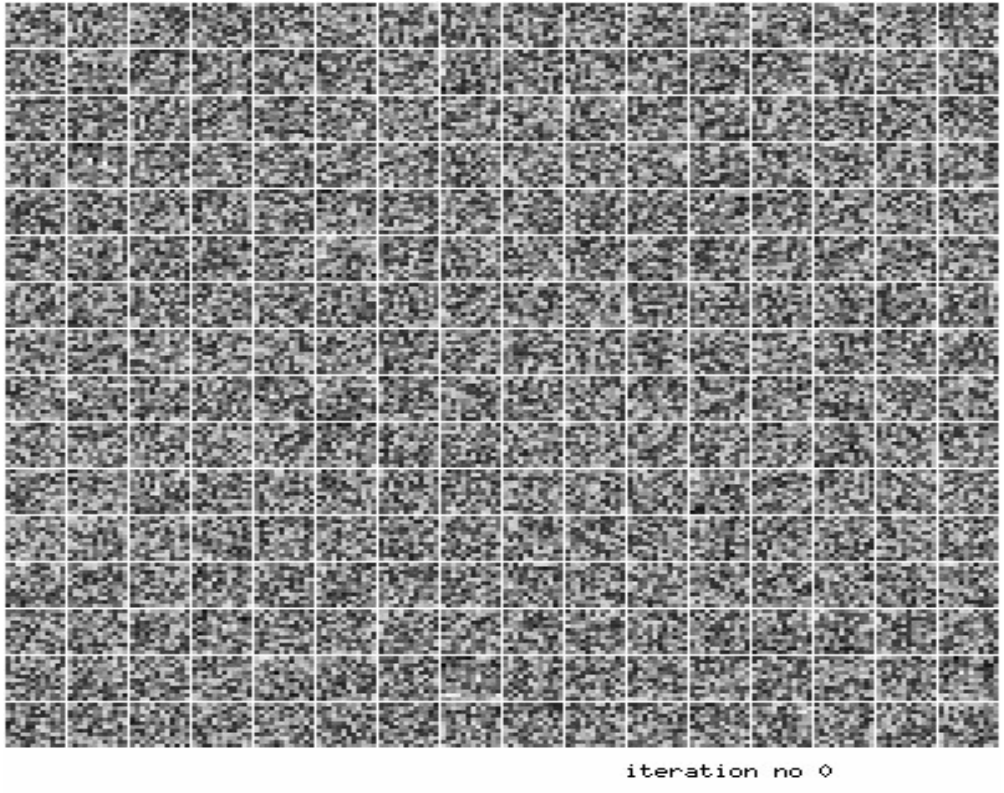
Energy Surface for Sparse Modeling

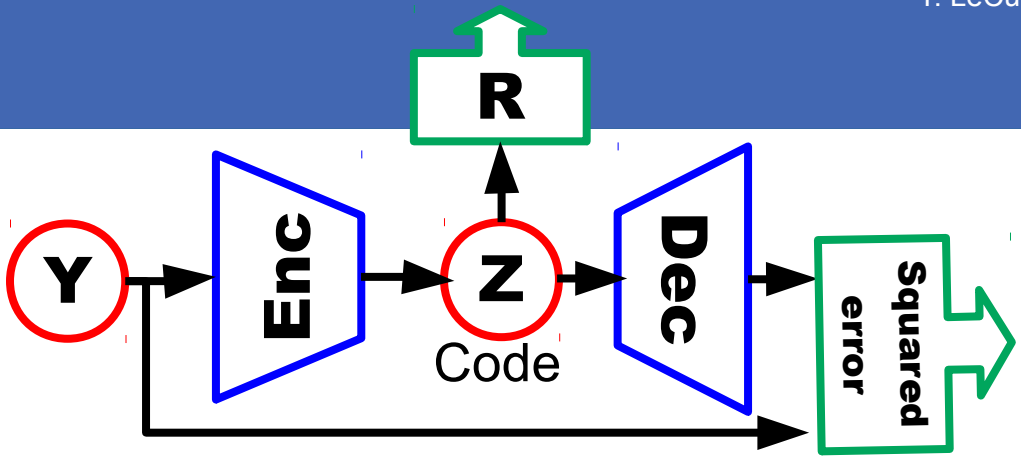
- **Sparse coding, sparse auto-encoder, Predictive Sparse Decomposition**

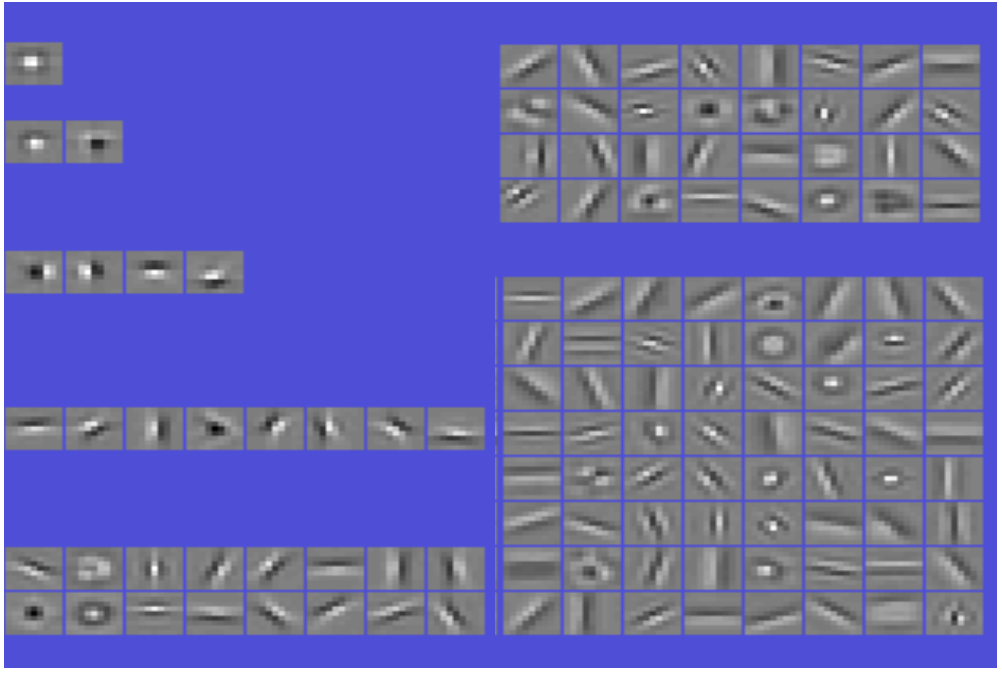


Sparse Auto-Encoder

- **Learns feature detectors unsupervised from images**
  - [Kavukcuoglu CVPR 2009]







Self-Supervised Forward Models: Learning a control task with few interactions

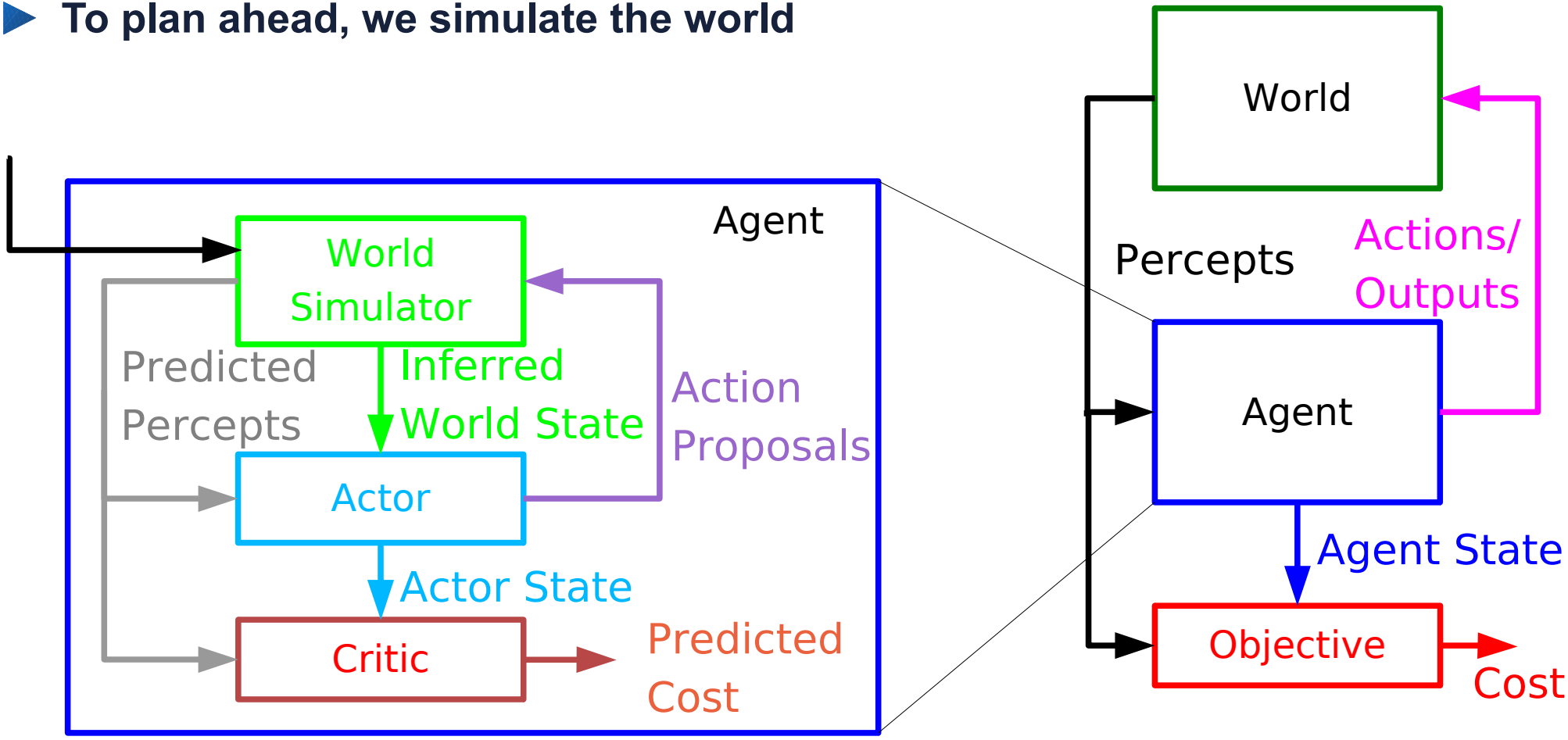
Learning motor skills with no interaction with the real world

[Henaff, Canziani, LeCun ICLR 2019]
[Henaff, Zhao, LeCun ArXiv:1711.04994]
[Henaff, Whitney, LeCun Arxiv:1705.07177]

facebook
Artificial Intelligence Research

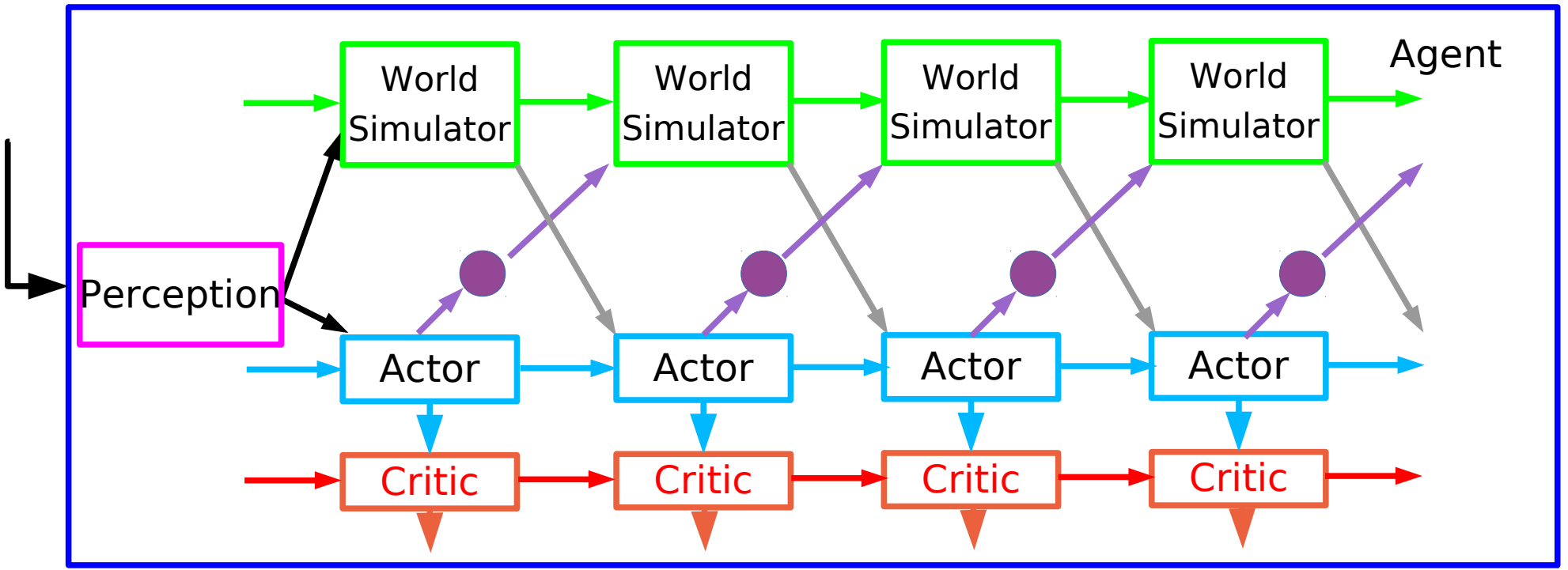
Planning Requires Prediction

- **To plan ahead, we simulate the world**



Training the Actor with Optimized Action Sequences

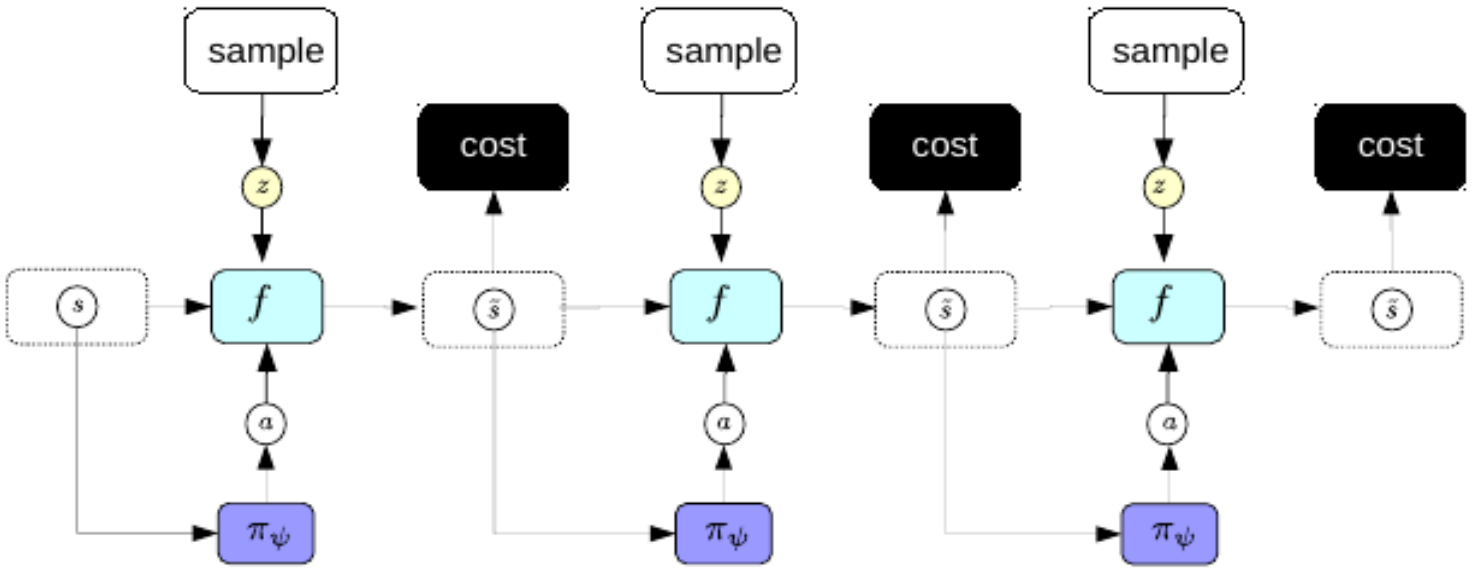
- **1. Find action sequence through optimization**
- **2. Use sequence as target to train the actor**
  - Over time we get a compact policy that requires no run-time optimization



Planning/learning using a self-supervised predictive world model

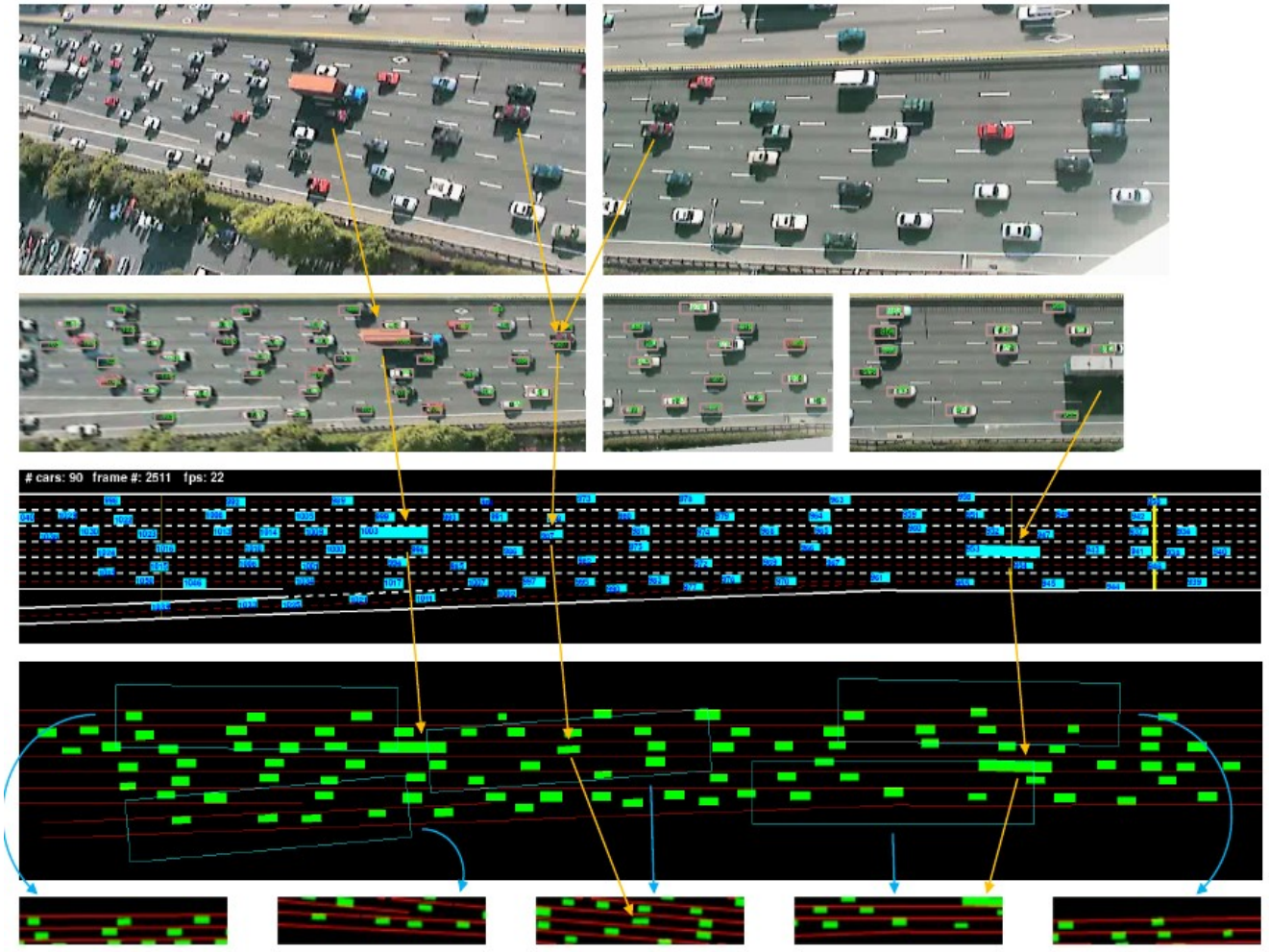
- **Feed initial state**
- **Run the forward model**
- **Backpropagate gradient of cost**
- **Act**
  - (model-predictive control)

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



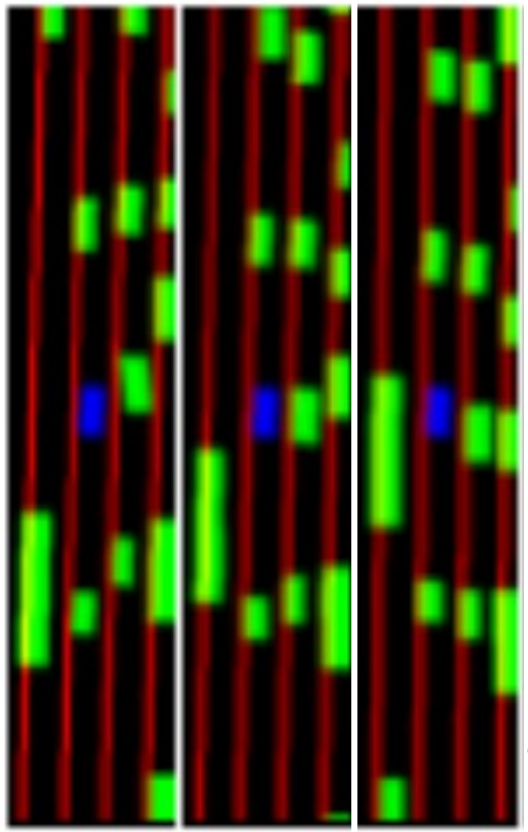
Using Forward Models to Plan (and to learn to drive)

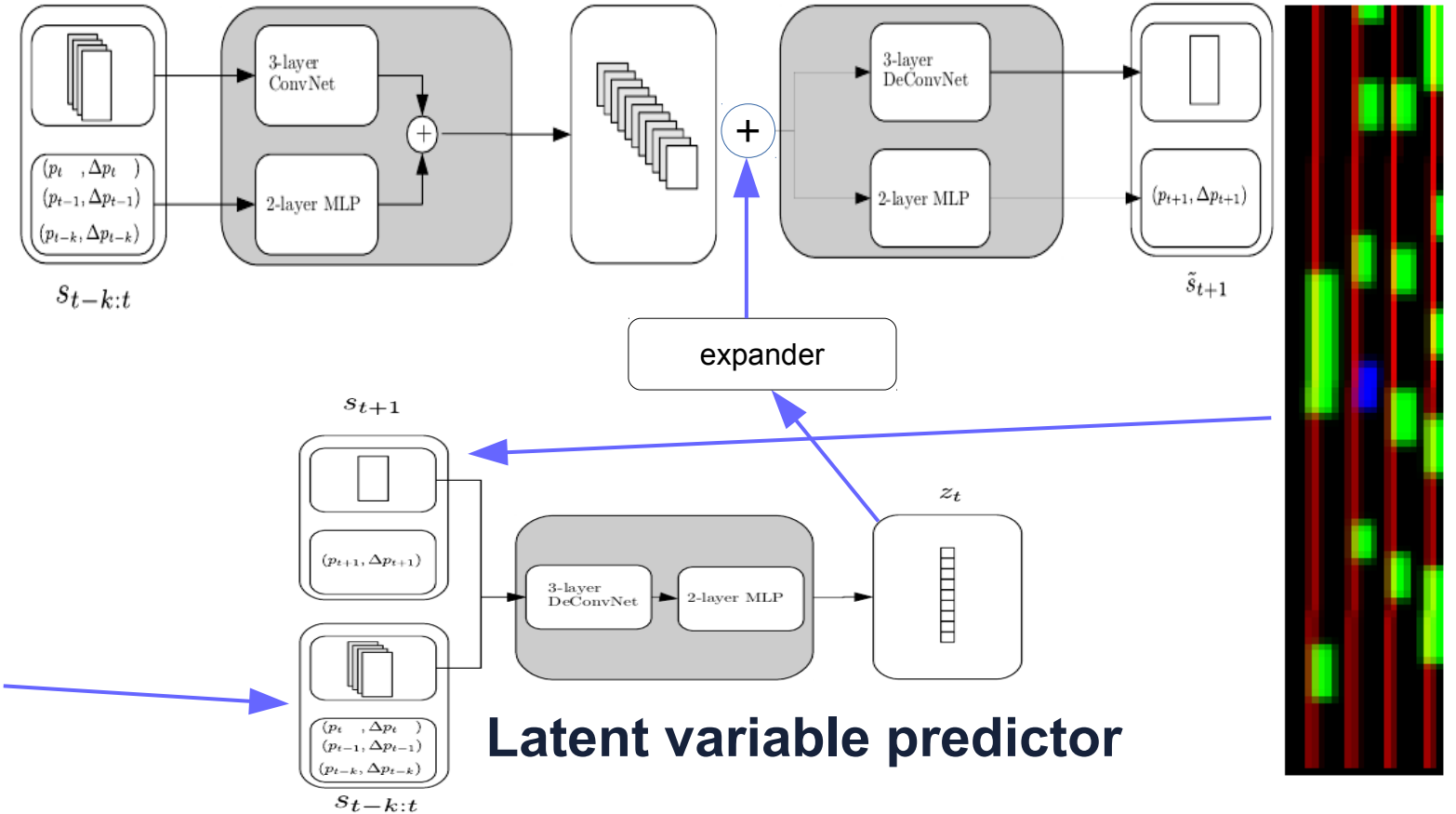
- **Overhead camera on highway.**
  - Vehicles are tracked
- **A "state" is a pixel representation of a rectangular window centered around each car.**
- **Forward model is trained to predict how every car moves relative to the central car.**
  - steering and acceleration are computed



Forward Model Architecture

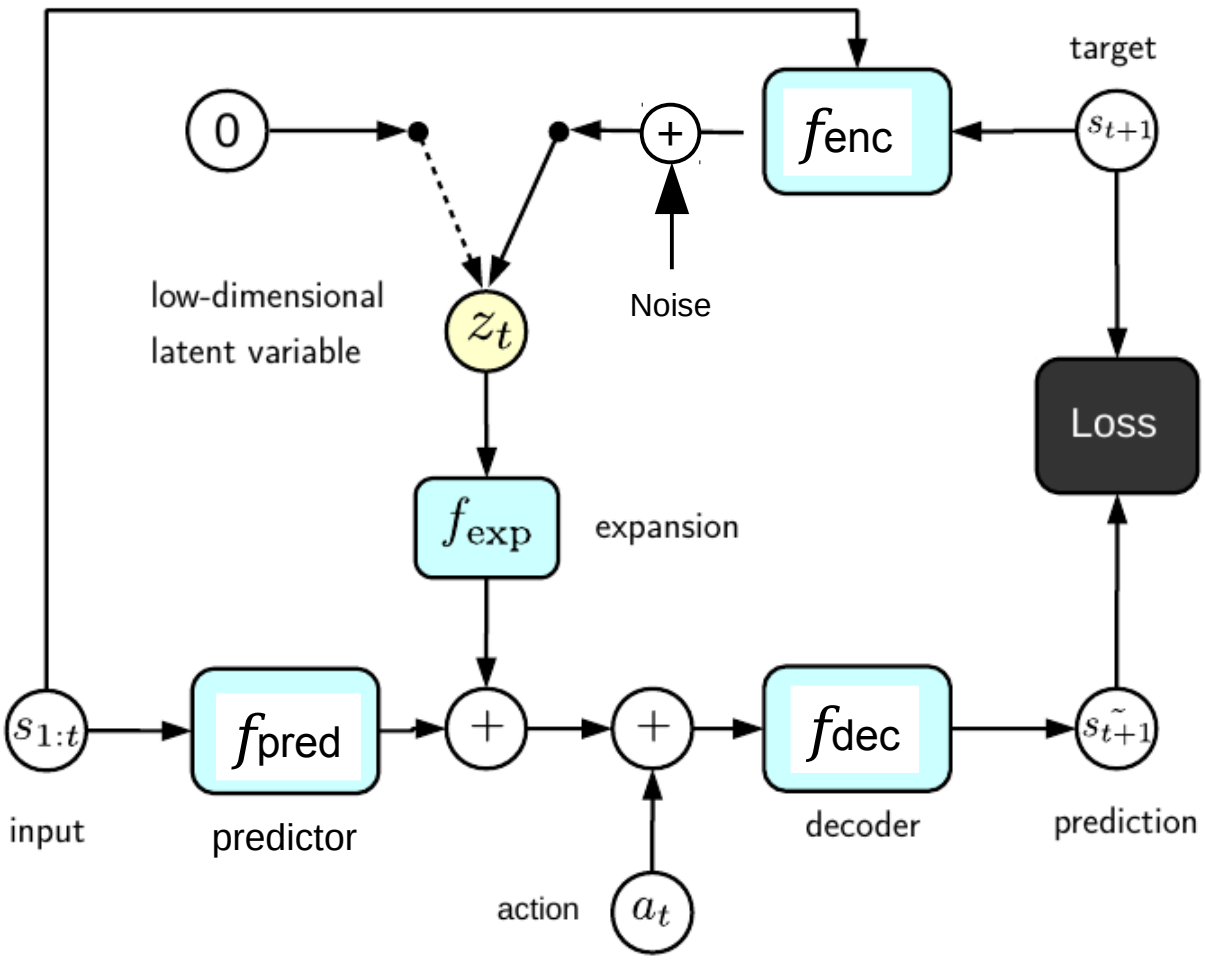
- **Architecture:**



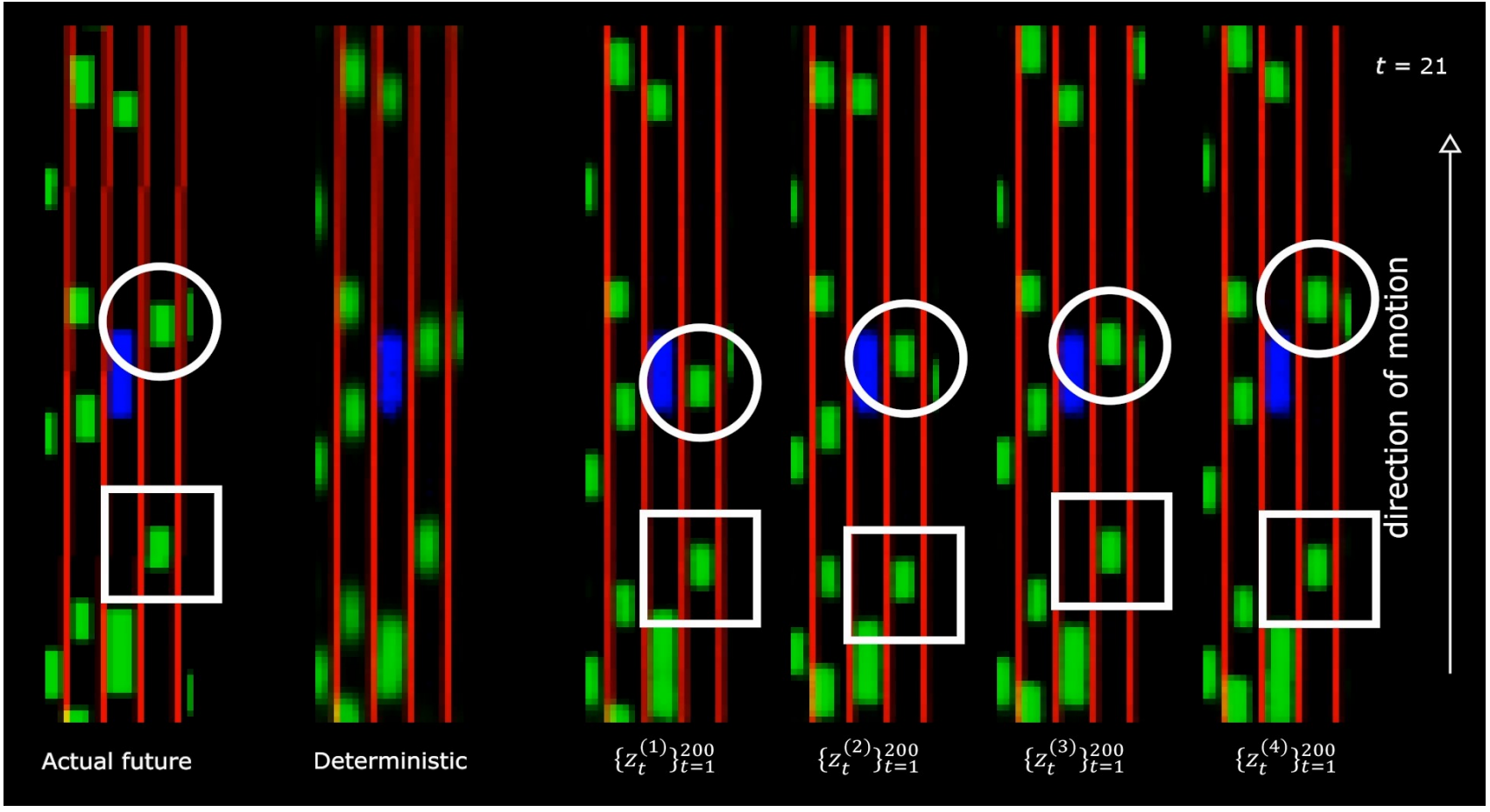


Stochastic Forward Modeling: regularized latent variable model

- **Latent variable is predicted from the target**
- **The latent variable is set to zero half the time during training (drop out) and corrupted with noise**
- **The model predicts as much as it can without the latent var.**
- **The latent var corrects the residual error.**

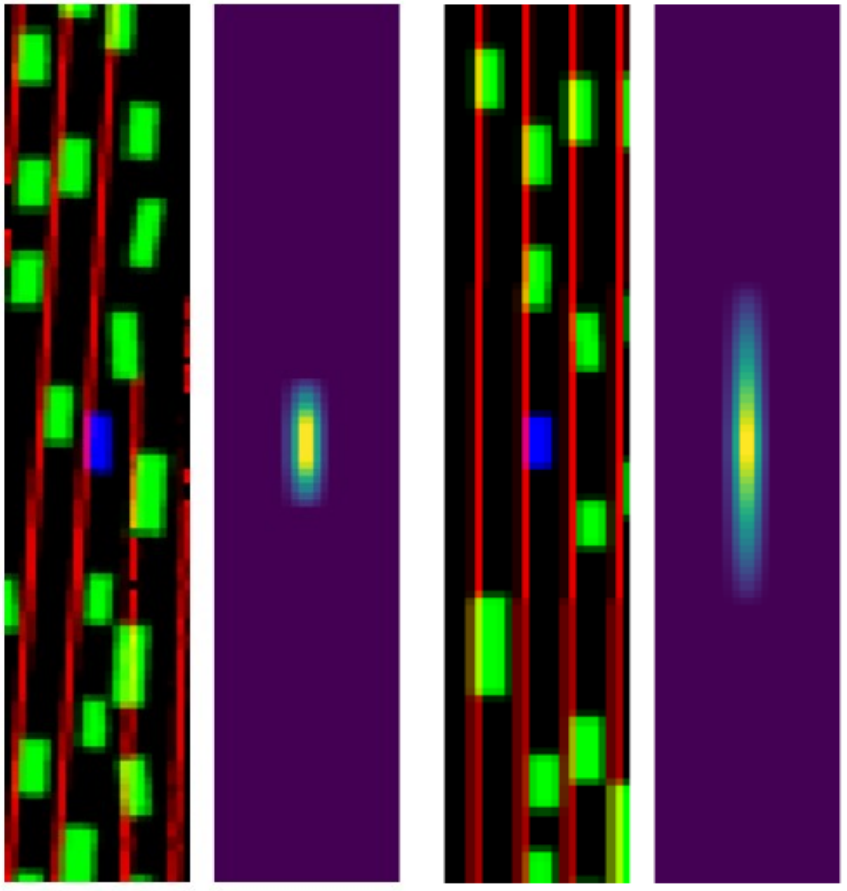


Actual, Deterministic, VAE+Dropout Predictor/encoder



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.

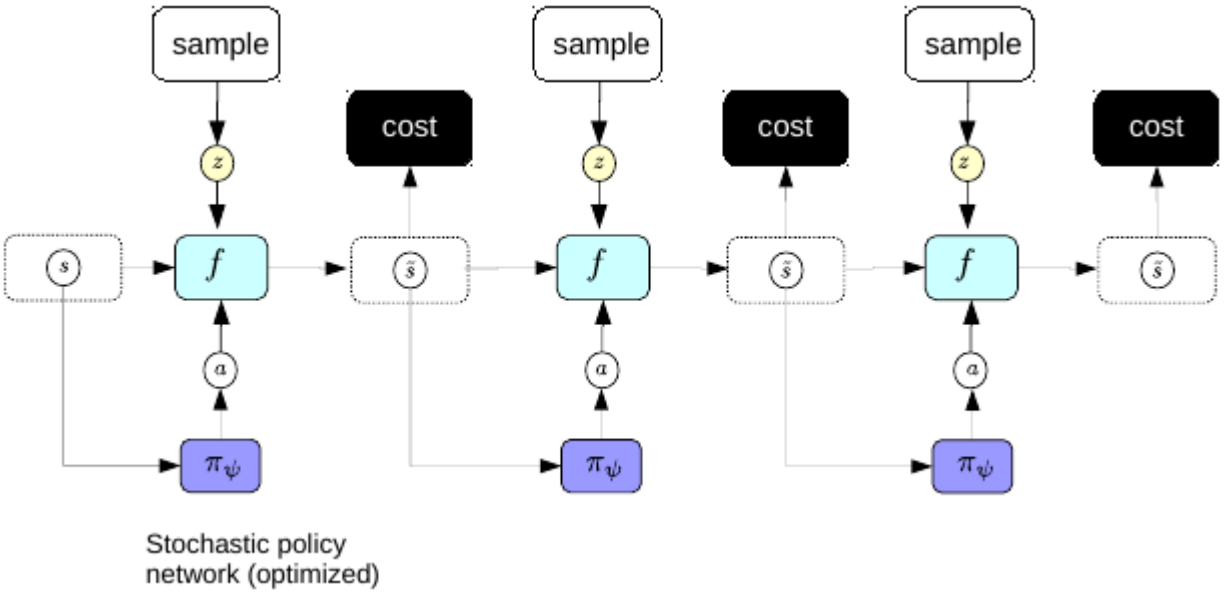


(a) 19.8 km/h (b) 50.3 km/h

Learning to Drive by Simulating it in your Head

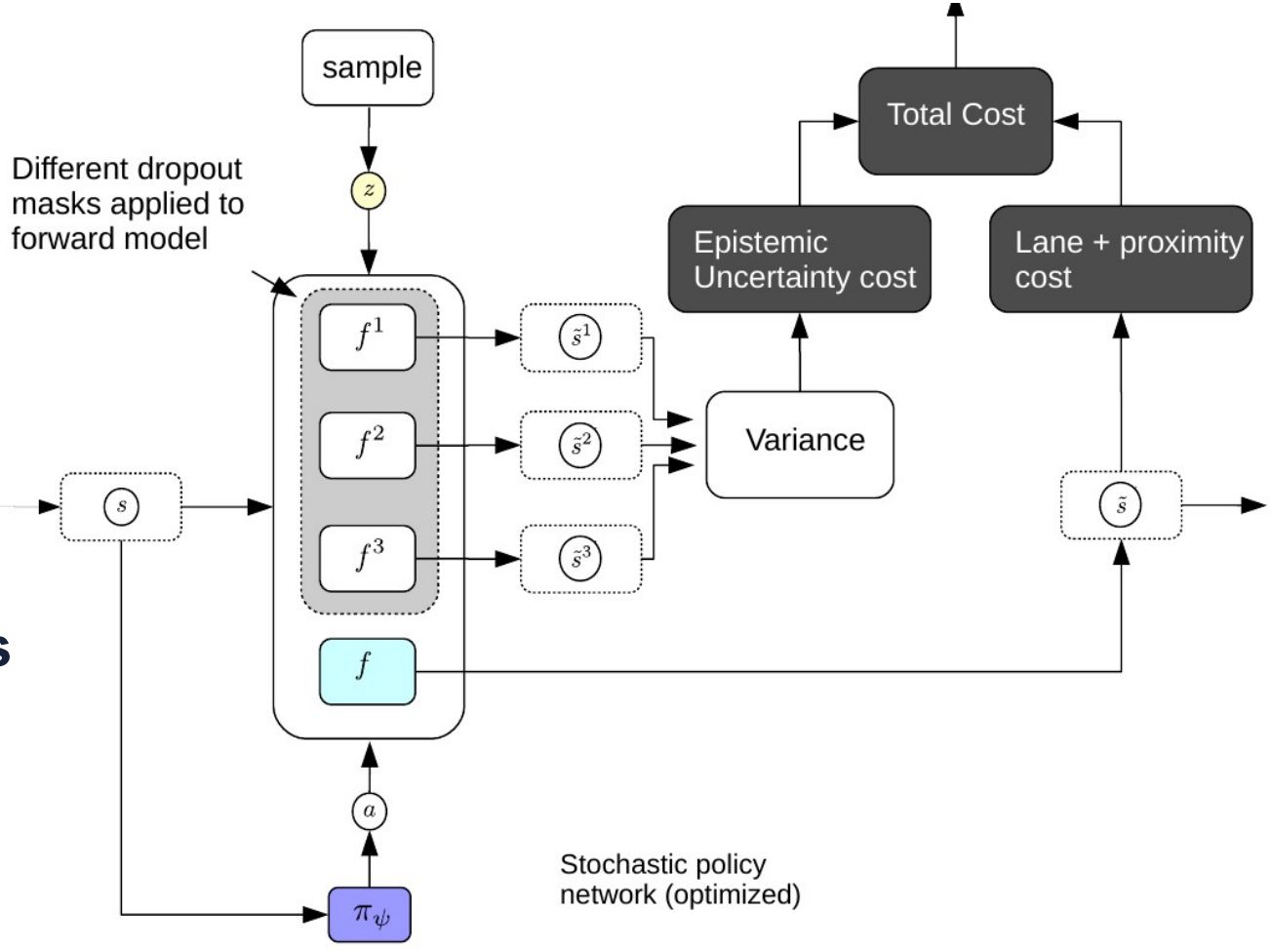
- Feed initial state
- Sample latent variable sequences of length 20
- Run the forward model with these sequences
- Backpropagate gradient of cost to train a policy network.
- Iterate

- No need for planning at run time.

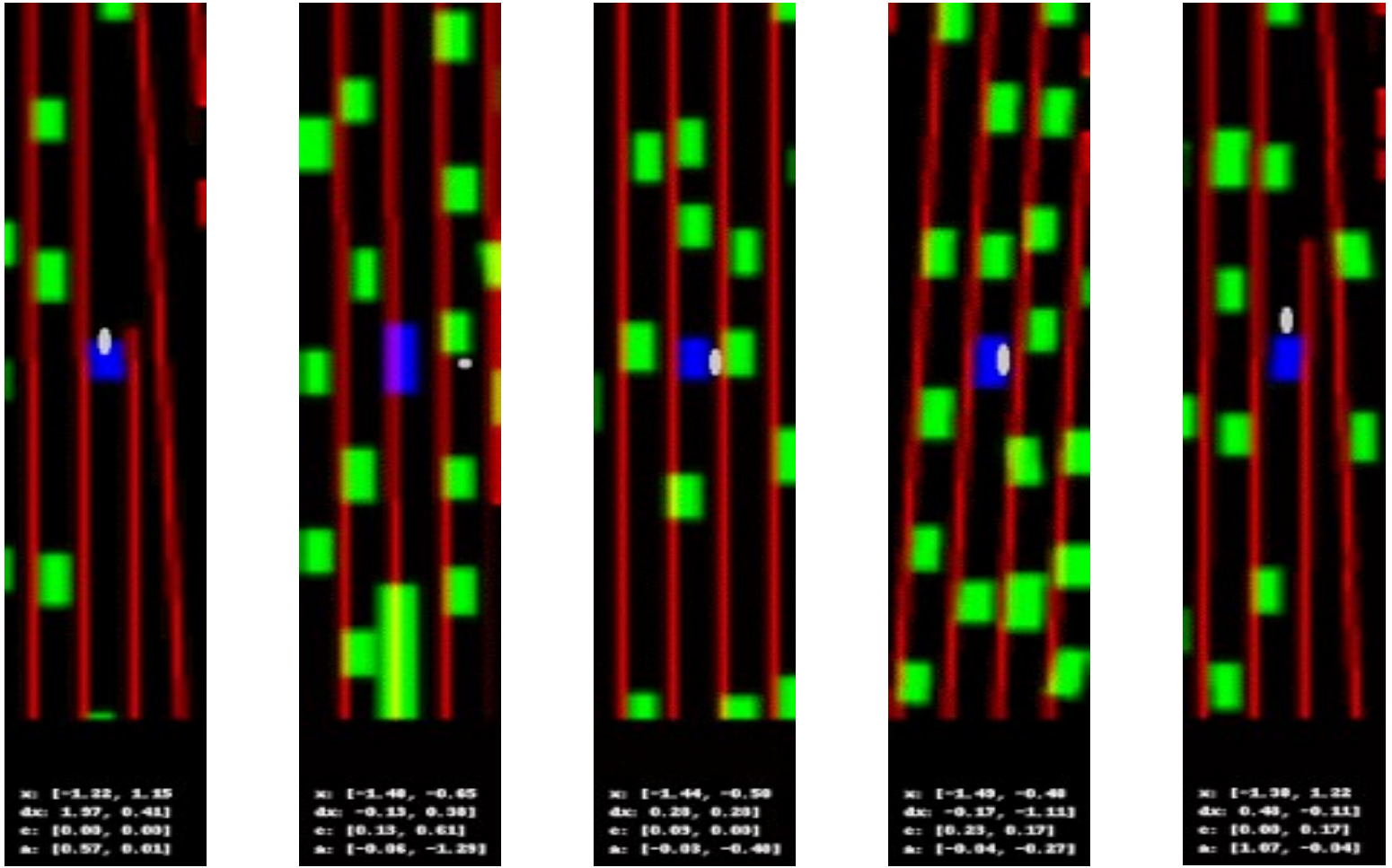


Adding an Uncertainty Cost (doesn't work without it)

- **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**

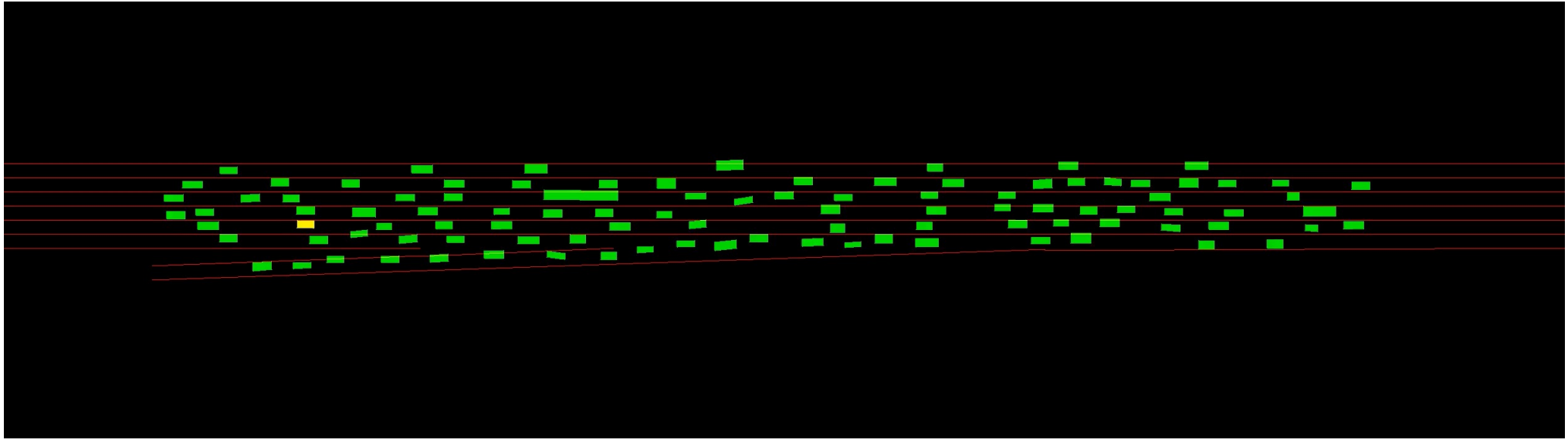


Driving an Invisible Car in “Real” Traffic



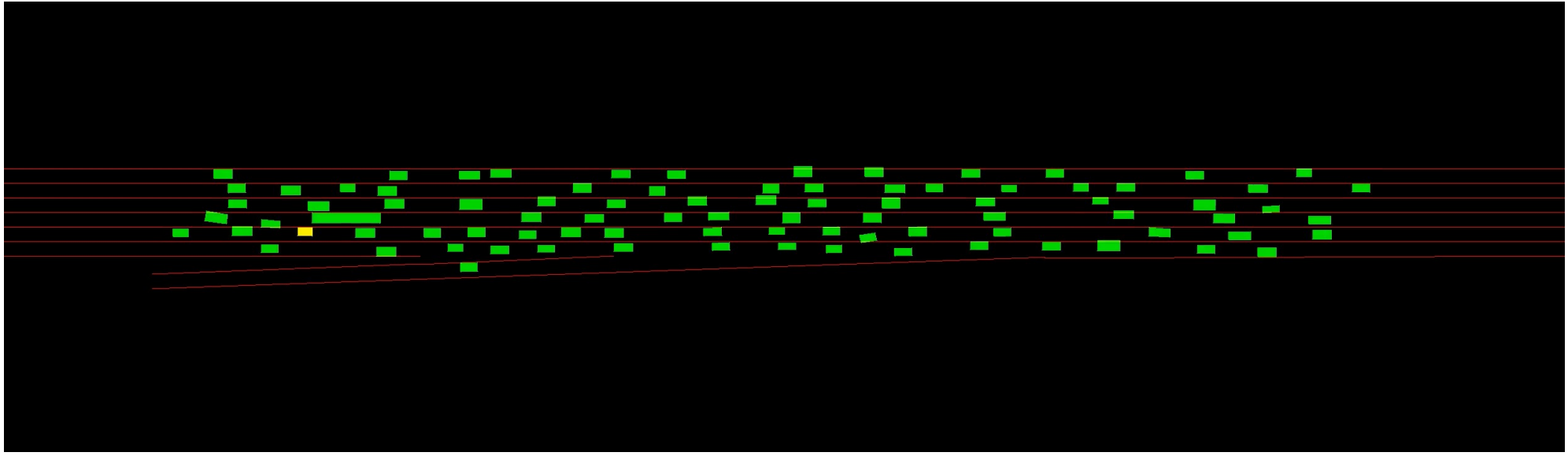
Driving!

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



Driving!

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



Theory often Follows Invention

- Telescope [1608]
- Steam engine [1695-1715]
- Electromagnetism [1820]
- Sailboat [???]
- Airplane [1885-1905]
- Compounds [???]
- Feedback amplifier [1927]
- Computer [1941-1945]
- Teletype [1906]

- Optics [1650-1700]
- Thermodynamics [1824-….]
- Electrodynamics [1821]
- Aerodynamics [1757]
- Wing theory [1907]
- Chemistry [1760s]
- Electronics [….]
- Computer Science [1950-1960]
- Information Theory [1948]

- Deep Learning

- Theory of Intelligence?

Biological Inspiration?

- **“Eole” took off from the ground on Oct 9, 1890**
  - (13 years before the Wright Brothers)
- **but you probably never heard of it**
  - (unless you are french).
- **L'Avion III (Musée du CNAM, Paris)**

1104. - L' “ Avion ” n° 3 d'Ader
Poids 258 kgr à vide ; approvisionnements charbon et eau pour 3 heures, 52 kgr ; force motrice : 2 moteurs à vapeur de 20 HP pesant chacun 21 kgr et actionnant séparément 2 hélices tractives avec armature bambou, pales en soie et papier.
Poids de chaque hélice 2 kgr 500
J. H.

Thank you

facebook
Artificial Intelligence Research

Lessons learned

- ***1 Model-free Reinforcement Learning is too slow in the real world***
  - *Requires too many "blind" interactions"*
- ***2: Regularized latent-variable energy-based models are a good way to learn features in an unsupervised fashion.***
- ***3: More generally, Self-Supervised learning is the future of DL***
  - *Networks will be much larger than today*
  - *We have unlimited amounts of data to train them*
  - *They will have sparse activation*
  - *Can electronic hardware take advantage of sparse activations?*
- ***4: Learning Models of the world accelerate learning of motor tasks***
- ***Prediction is the essence of intelligence***

When will the "True AI" revolution occur?

- **We won't have household robots and good digital friends (or assistants) until machines acquire common sense.**
- **This won't happen until we get machines to learn predictive world models**
- **Discovering the principles of it may take 2, 5, 10 or 20 years.**
- **Developing practical technology from it may take another 10 years**
  - The emergence of "true AI" will not be a singular event as in Hollywood movies.
- **We work on the assumption that there is "simple" principle (and a few algorithms) for AI, as there is for flight (aerodynamics) or engines (thermodynamics).**

What will super-intelligent AGI be like?

- **Will the "singularity" happen?**
  - No. Nothing is exponential forever
- **Future AI systems will have emotions and moral values**
  - How to align AI values with human values?

- **Will it take our jobs?**
  - No. But our jobs will change. Human experience will have high value.
  - it will empower humanity by amplifying our intelligence

- **Will it want to take over the world?**
  - No, the desire to dominate is not correlated with intelligence but with testosterone

Authentic human experience > material goods

- **Material goods:**
  - BlueRay player: $47
  - Handmade ceramic bowl: $750
- **Mozart's opera Die Zauberflöte**
  - Downloadable recording: $7
  - Ticket at the NYC Met: up to $807
- **Bright future for jazz musicians and artisans?**

Samsung

Samsung Smart Curved Design Blu-Ray Disc 1080p Player With Wired Ethernet Content Streaming Manufacturer Refurbished

19 customer reviews | 8 answered questions

Price: $46.88 & FREE Shipping

Mozart: Die Zauberflöte

Wiener Philharmoniker

January 1, 2012

19 customer reviews

Start your 30-day free trial of Unlimited to Prime pricing.

See all 50 formats and editions

Streaming Unlimited	MP3 $6.99	Audio CD $8.98

CENTER ORCHESTRA Row B-EE Email delivery by: 09/26/17	QTY 5	$751.00 ea.	BUY
ORCH Row B-O Email delivery by: 09/26/17	QTY 4	$772.00 ea.	BUY
CENTER ORCHESTRA Row A-DD Email delivery by: 09/26/17	QTY 5	$786.00 ea.	BUY
ORCH Row A-N Email delivery by: 09/26/17	QTY 4	$807.00 ea.	BUY

Scallop Bowl

$750.00 USD

Dimensions: 14"w x 7"h

H3-40

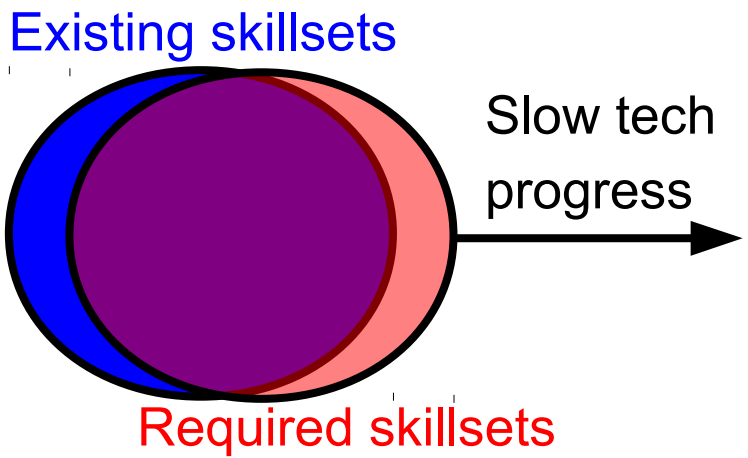
ADD TO CART

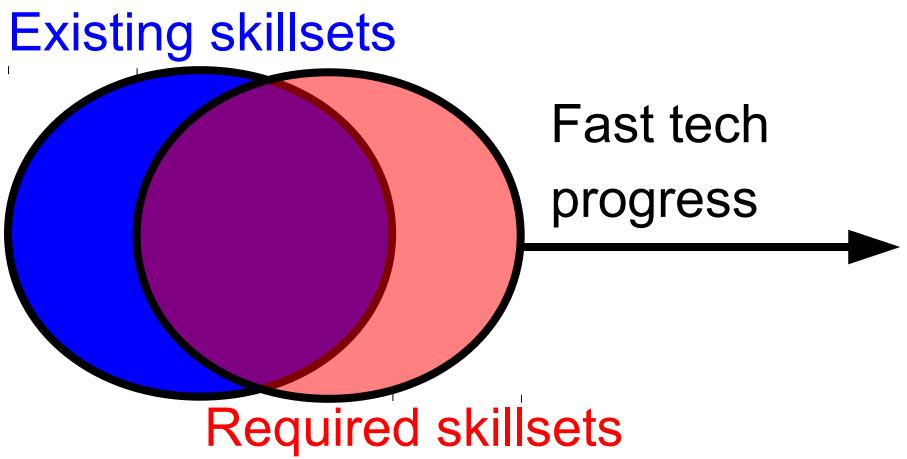
ADD TO REGISTRY

Tweet Like 0

AI is a "General Purpose Technology" (GPT)

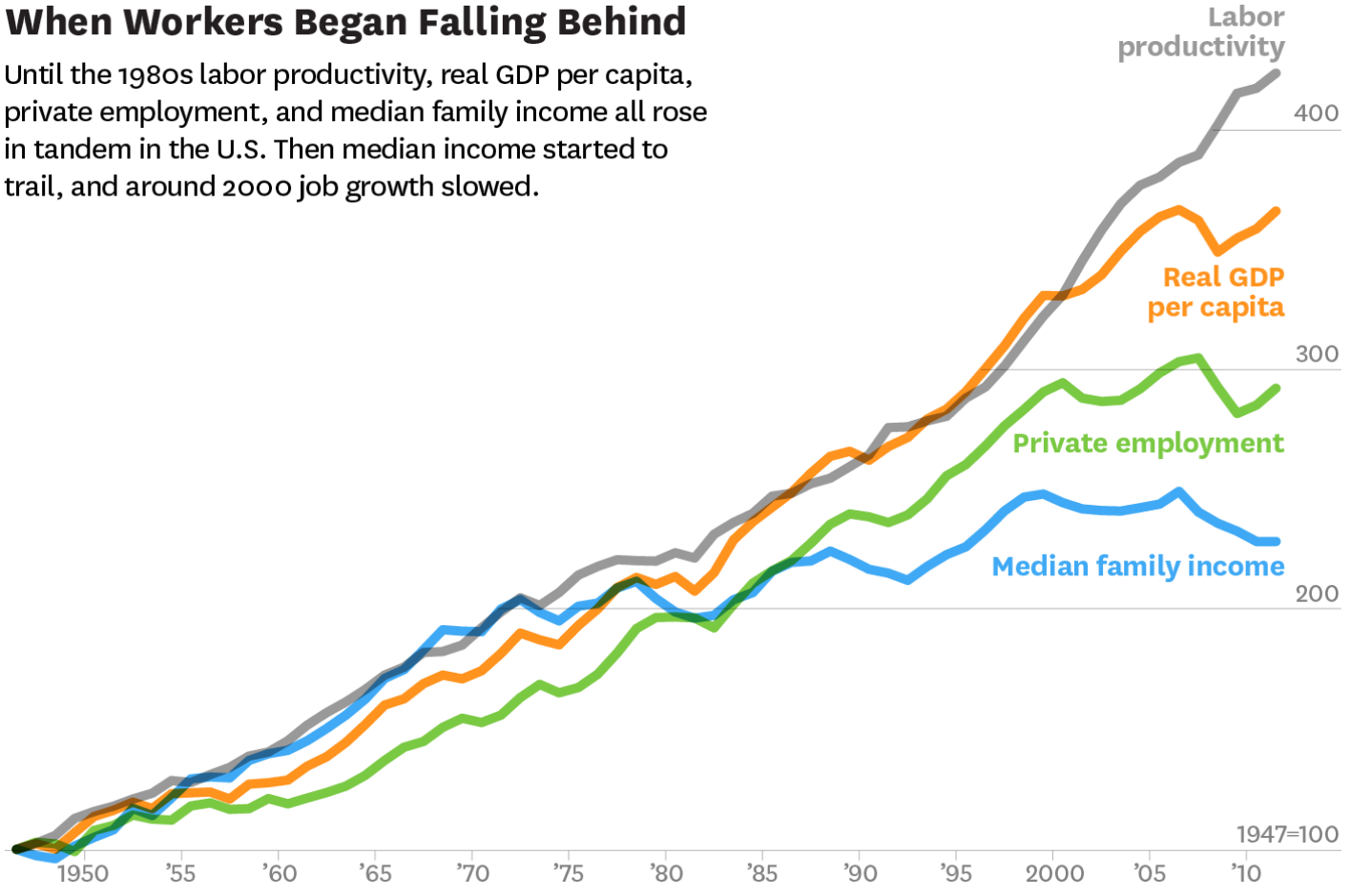
- **GPT: steam engine, electricity, computer...**
  - [Bresnahan & Trajtenberg 1995] "GPTs 'Engines of growth'?". J. Econometrics.
- **AI will affect many sector of the economy**
- **But it will take 10 or 20 years before we see the effect on productivity**
- **AI/automation → job displacement → technological unemployment**
- **Technology deployment is limited by how fast workers can train for it**





Does technological progress cause income inequality?

- **Erik Brynjolfsson & Andrew McAfee (MIT).**
- **Perhaps, but the fix is progressive taxation.**



Thank you

facebook
Artificial Intelligence Research