Deep Learning Limitations and New Frontiers

Ava Amini

MIT Introduction to Deep Learning

January 10, 2024

T-shirts! Tomorrow!

Class Schedule

Intro to Deep Learning

Lecture 1
Jan. 8, 2024
[Slides] [Video] *coming soon!*

Deep Sequence Modeling

Lecture 2
Jan. 8, 2024
[Slides] [Video] *coming soon!*

Intro to TensorFlow; Music Generation

Software Lab 1
[Code]

Deep Computer Vision

Lecture 3
Jan. 9, 2024
[Slides] [Video] *coming soon!*

Deep Generative Modeling

Lecture 4
Jan. 9, 2024
[Slides] [Video] *coming soon!*

Facial Detection Systems

Software Lab 2
[Paper] [Code] *coming soon!*

Deep Reinforcement Learning

Lecture 5
Jan. 10, 2024
[Slides] [Video] *coming soon!*

New Frontiers

Lecture 6
Jan. 10, 2024
[Slides] [Video] *coming soon!*

Large Language Models

Software Lab 3
[Code] *coming soon!*

Generative AI for Media

Lecture 7
Jan. 11, 2024
[Info] [Slides] [Video] *coming soon!*

Stories from Models in the Wild

Lecture 8
Jan 11, 2024
[Info] [Slides] [Video] *coming soon!*

Final Project

Work on final projects

Final Project

Work on final projects
Jan 12, 2024

Project Presentations

Pitch your ideas!
Jan 12, 2024

Awards Ceremony

Final awards and celebration!

- Lab competition: 1/11/24
- Proposal slides: 1/12/24
- Proposal pitch: 1/12/24

Labs and Prizes

Lab 1: Music Generation

Lab 2: Computer Vision

Lab submission: 1/11/24 at 11:00pm ET

Instructions: bit.ly/6s191-syllabus

github.com/aamini/introtodeeplearning/

Final Class Project

Option 1: Proposal Presentation

- At least 1 registered student to be prize eligible
- Present a novel deep learning research idea or application
- 5 minutes (strict)
- Presentations on **Friday, Jan 12**
- Submit groups by **Wed 1/10 11:00pm ET** to be eligible
- Final slides by **Fri 1/12 3:00pm ET**
- Instructions: bit.ly/6s191-syllabus

- Judged by a panel of judges
- Top winners are awarded:

NVIDIA 3080 GPU

Smartwatches

Display Monitors

Final Class Project

Option 1: Proposal Presentation

- At least 1 registered student to be prize eligible
- Present a novel deep learning research idea or application
- 3 minutes (strict)
- Presentations on Friday, Jan 29
- Submit groups by Wednesday 11:59pm ET to be eligible
- Submit slide by Thursday 11:59pm ET to be eligible
- Instructions:

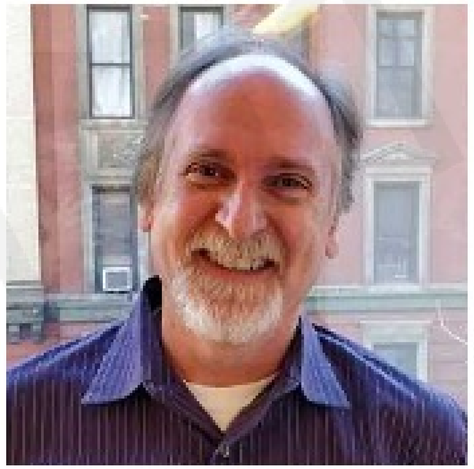
Option 2: Write a 1-page review of a deep learning/AI paper

- Grade is based on clarity of writing and technical communication of main ideas
- Due **Fri Jan 12 3:00pm ET**
- **Instructions:** bit.ly/6s191-syllabus

Program Guest Lectures

Douglas Eck
Google DeepMind

Niko Laskaris
Comet ML



Douglas Blank
Comet ML

So far in Introduction to Deep Learning…

The Rise of Deep Learning

So far in Introduction to Deep Learning…

Data

- Signals
- Images
- Sensors

…

Decision

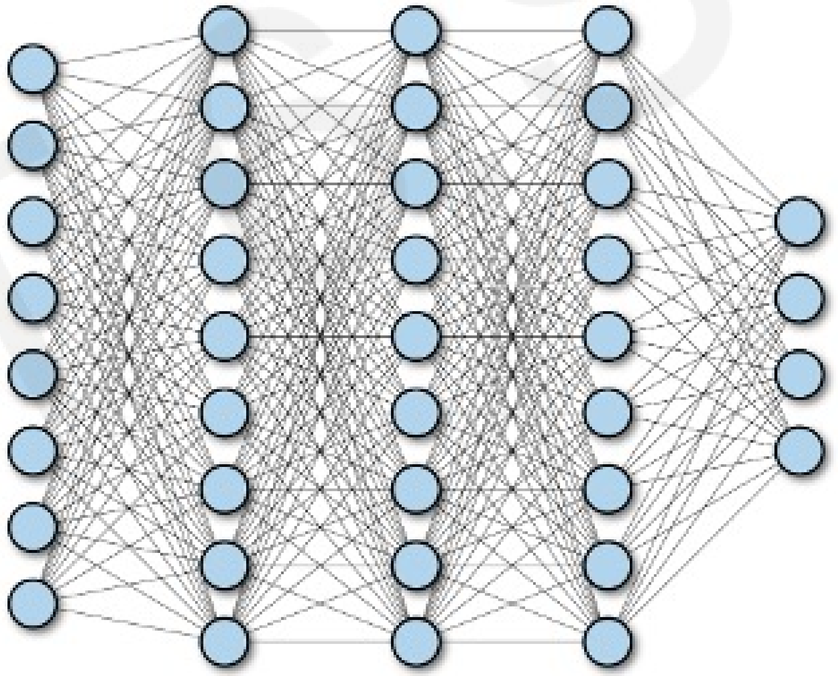
- Prediction
- Detection
- Action

…

Power of Neural Nets

Universal Approximation Theorem

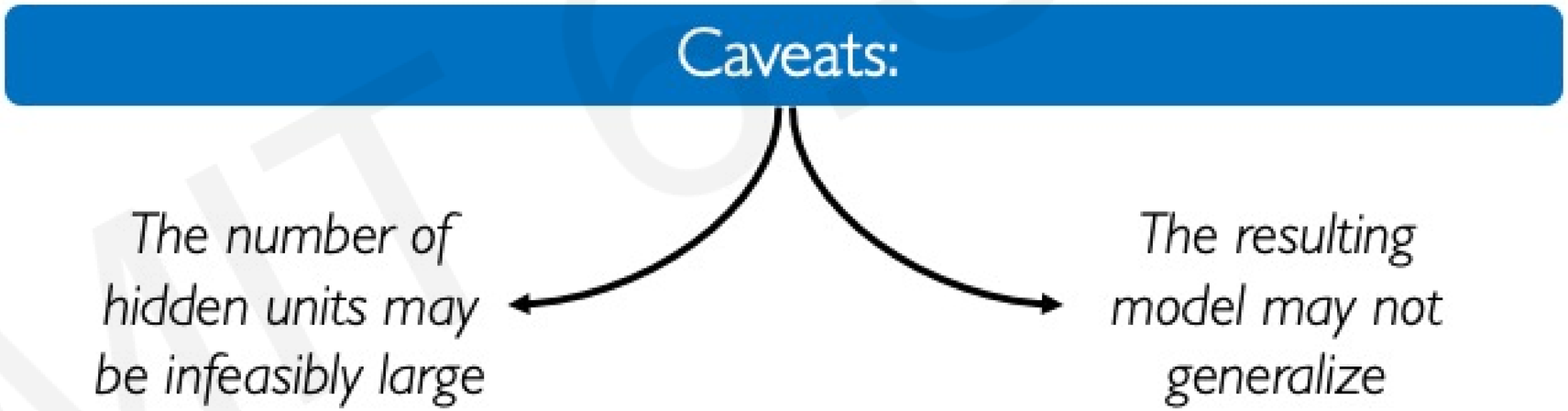
A feedforward network with a single layer is sufficient to approximate, to an arbitrary precision, any continuous function.



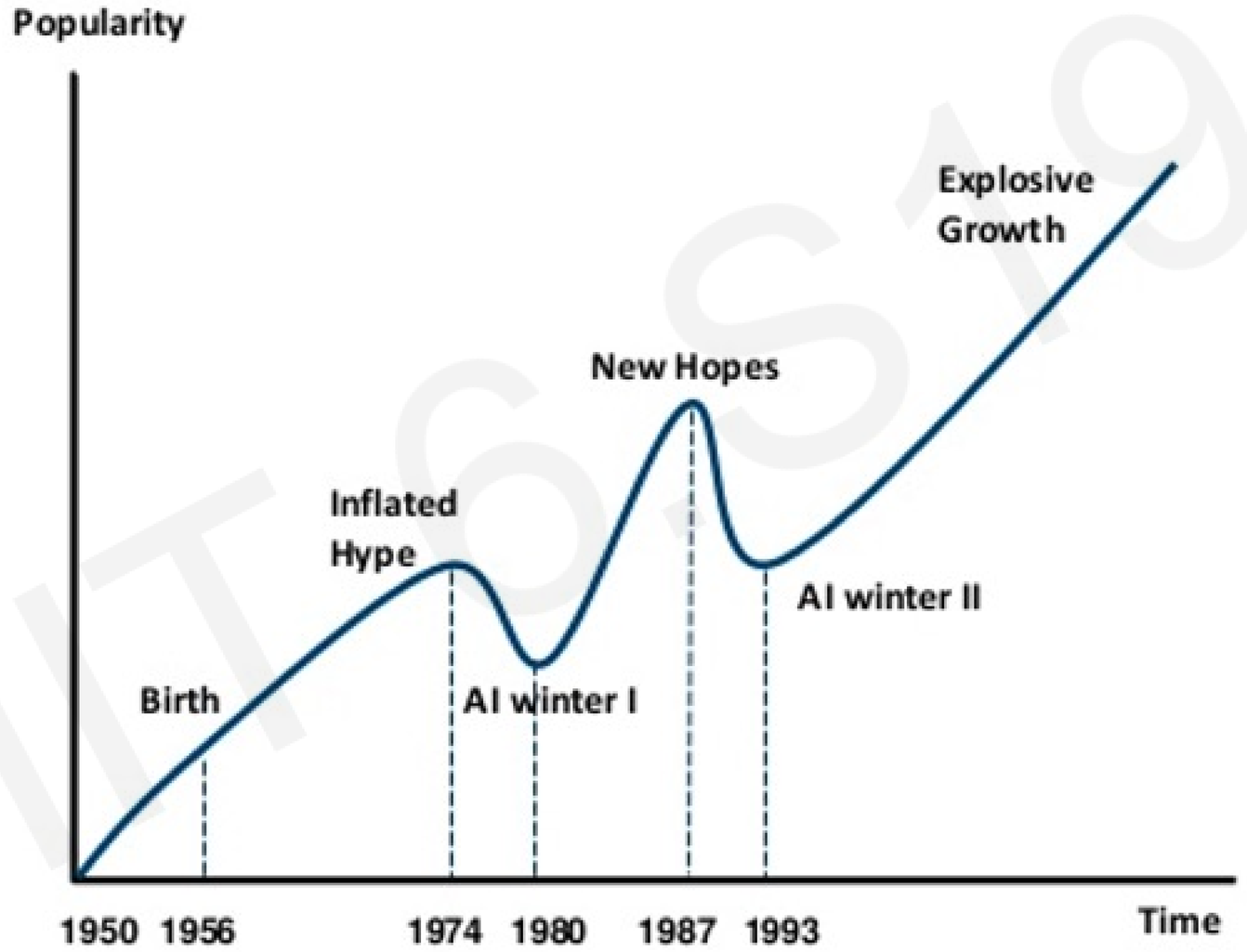
Power of Neural Nets

Universal Approximation Theorem

A feedforward network with a single layer is sufficient to approximate, to an arbitrary precision, any continuous function.



Artificial Intelligence "Hype": Historical Perspective



Limitations

Rethinking Generalization

"Understanding Deep Neural Networks Requires Rethinking Generalization"

dog

banana

dog

tree

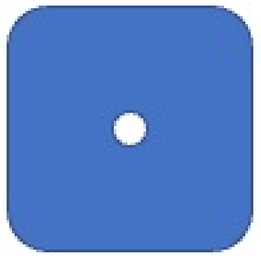
Rethinking Generalization

"Understanding Deep Neural Networks Requires Rethinking Generalization"

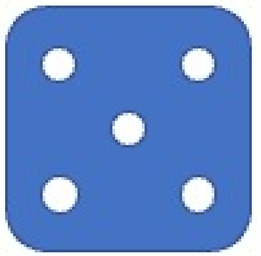
dog

banana

dog



tree



Rethinking Generalization

"Understanding Deep Neural Networks Requires Rethinking Generalization"

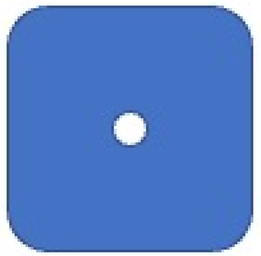
dog

banana

banana

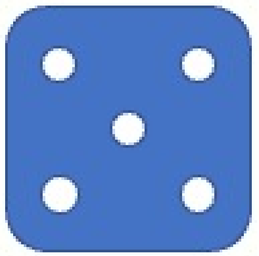
dog

dog



tree

tree



dog

Rethinking Generalization

"Understanding Deep Neural Networks Requires Rethinking Generalization"

~~dog~~

banana

~~banana~~

dog

~~dog~~

tree

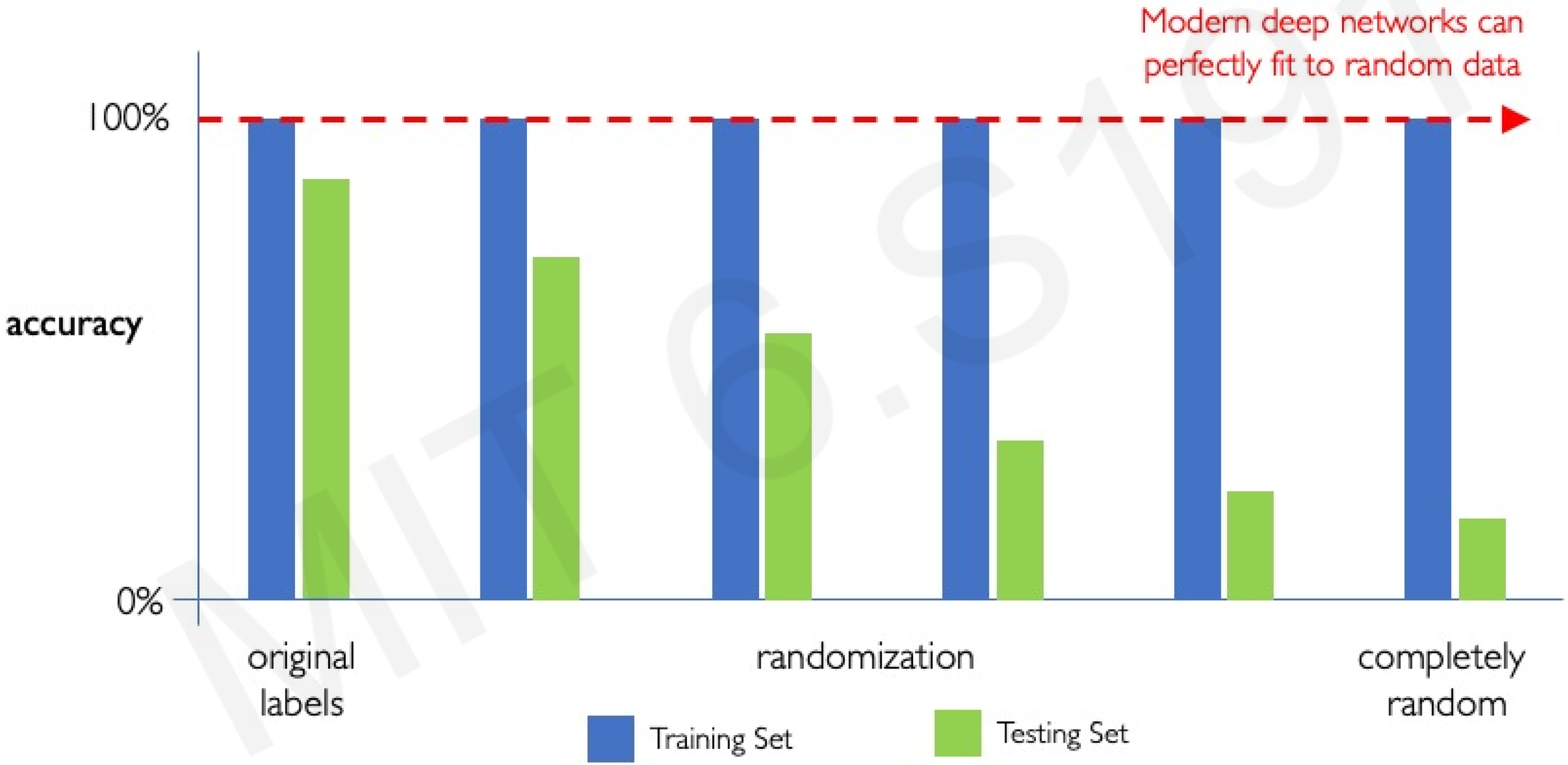
~~tree~~

dog

Capacity of Deep Neural Networks

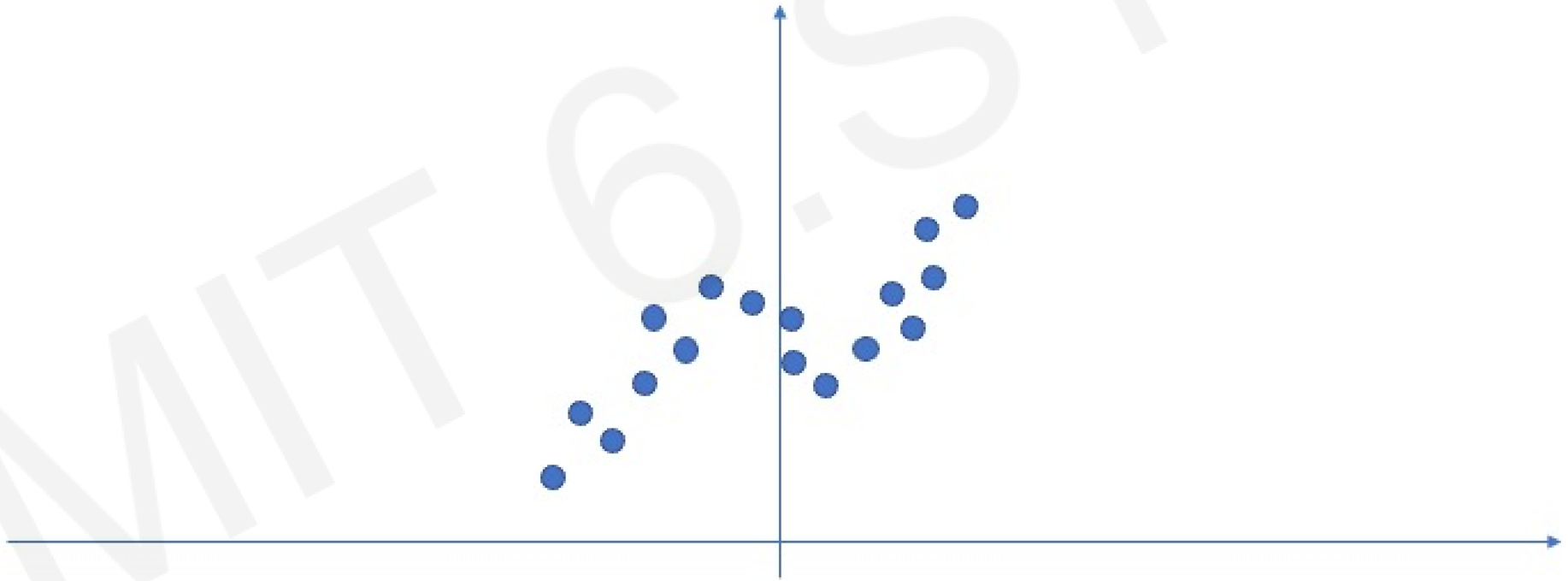
Capacity of Deep Neural Networks

Capacity of Deep Neural Networks



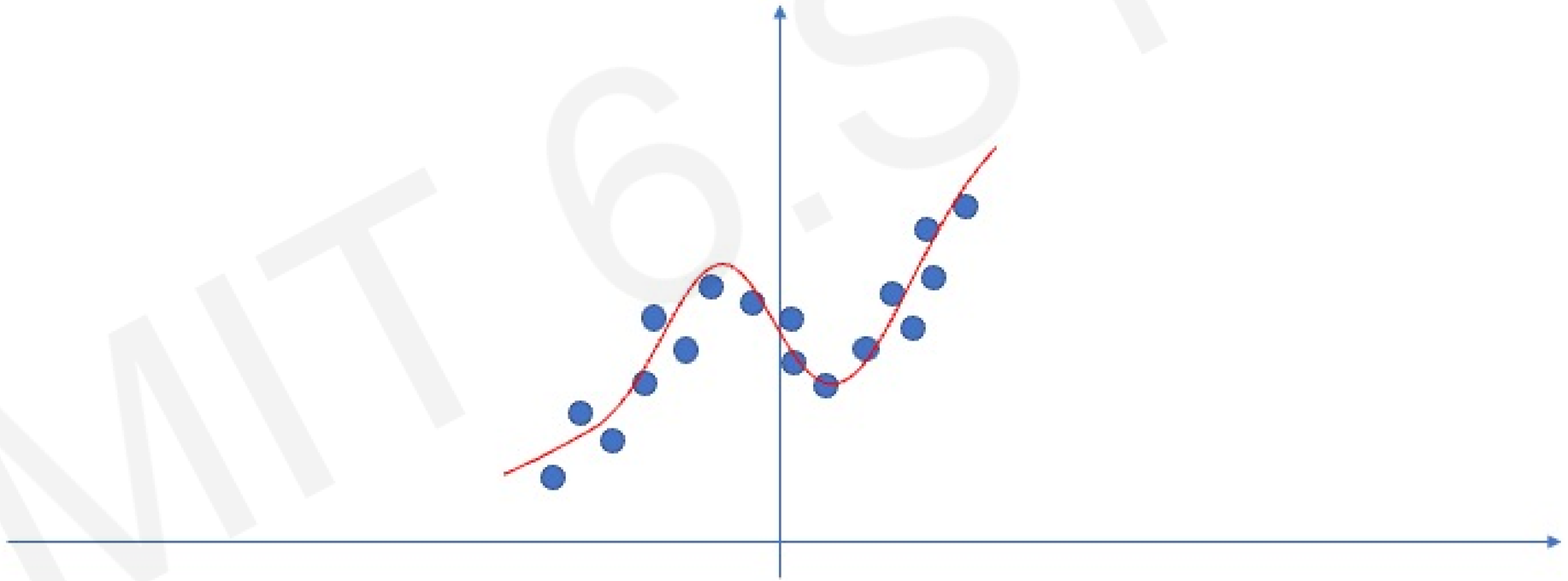
Neural Networks as Function Approximators

Neural networks are **excellent** function approximators



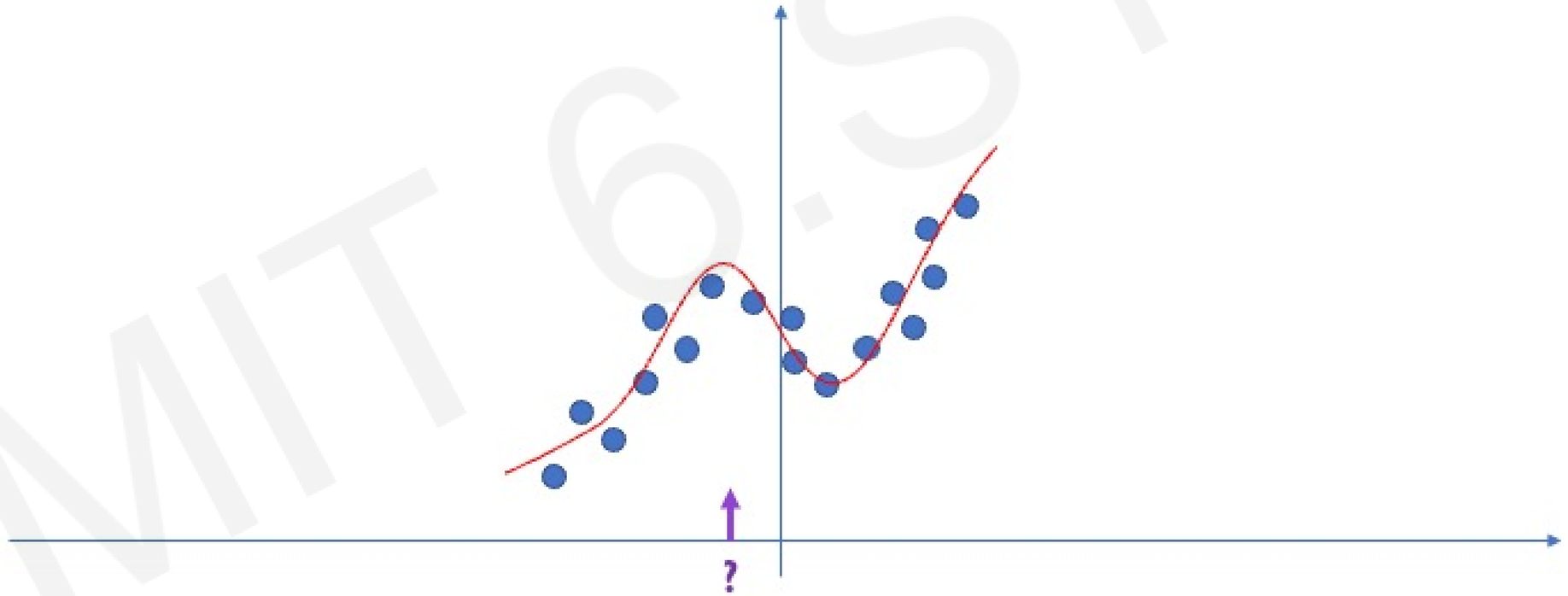
Neural Networks as Function Approximators

Neural networks are **excellent** function approximators



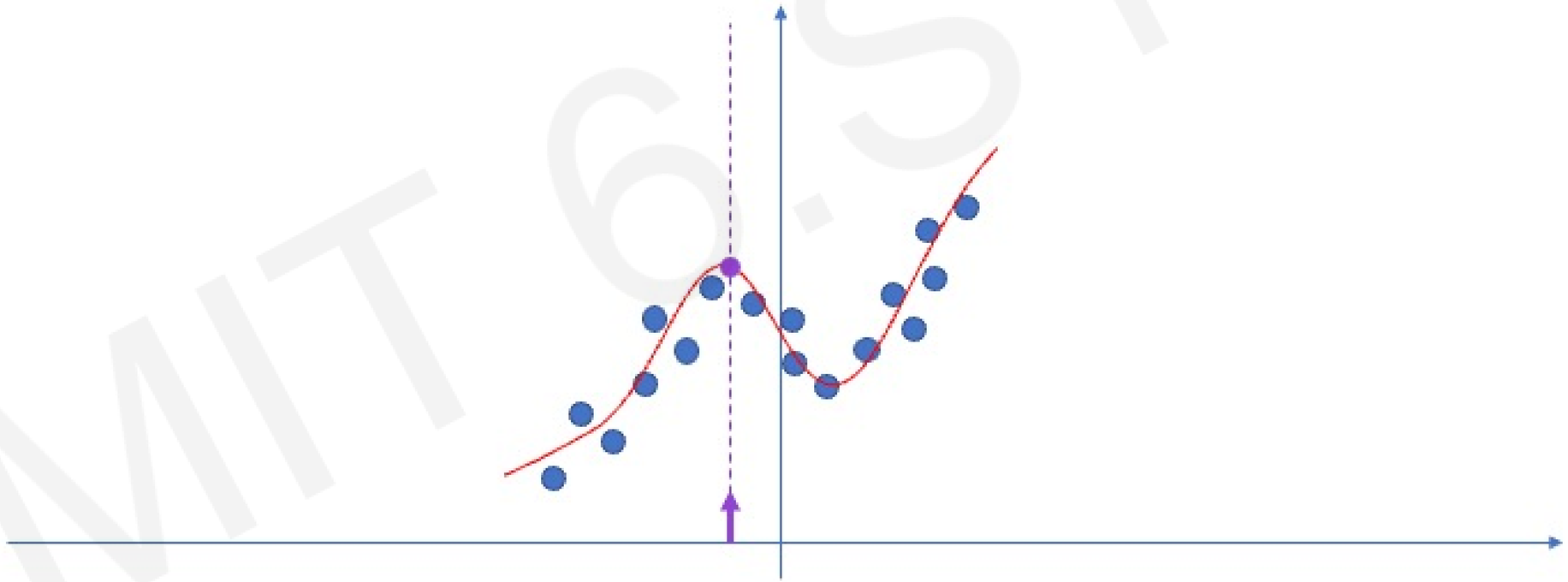
Neural Networks as Function Approximators

Neural networks are **excellent** function approximators



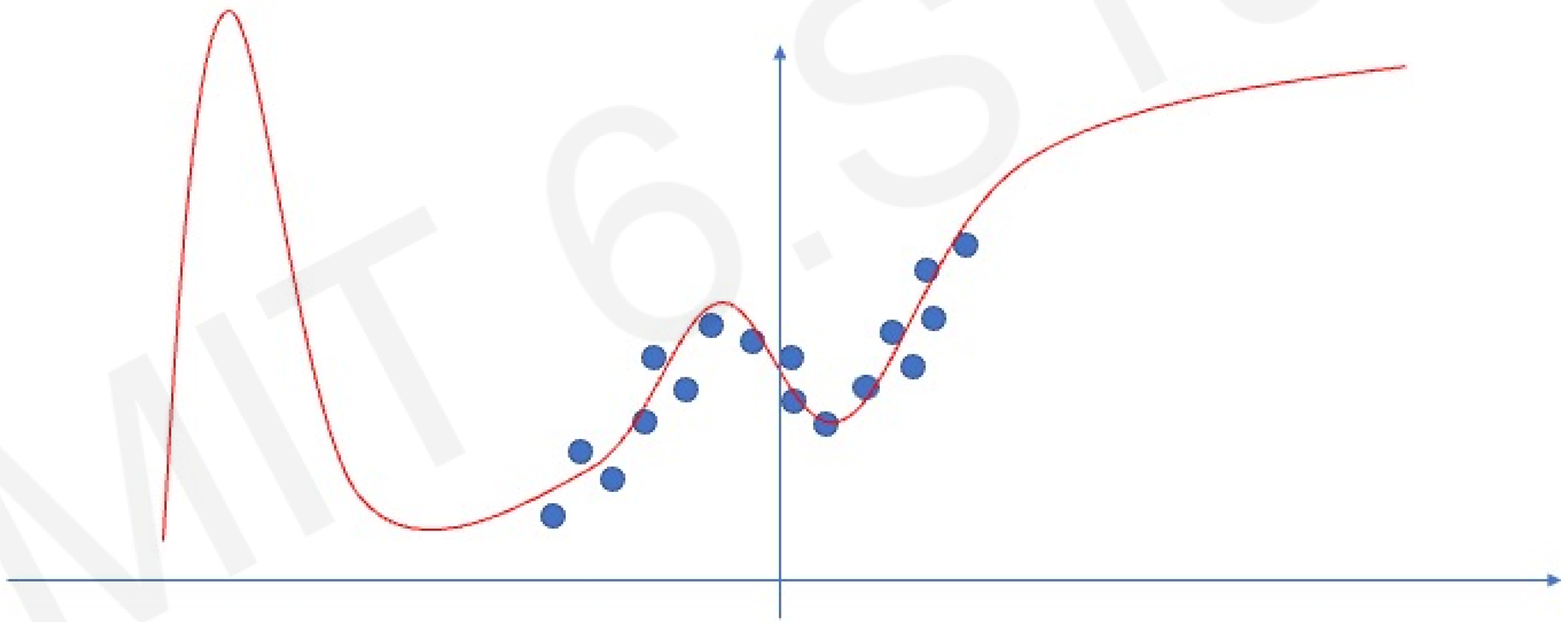
Neural Networks as Function Approximators

Neural networks are **excellent** function approximators



Neural Networks as Function Approximators

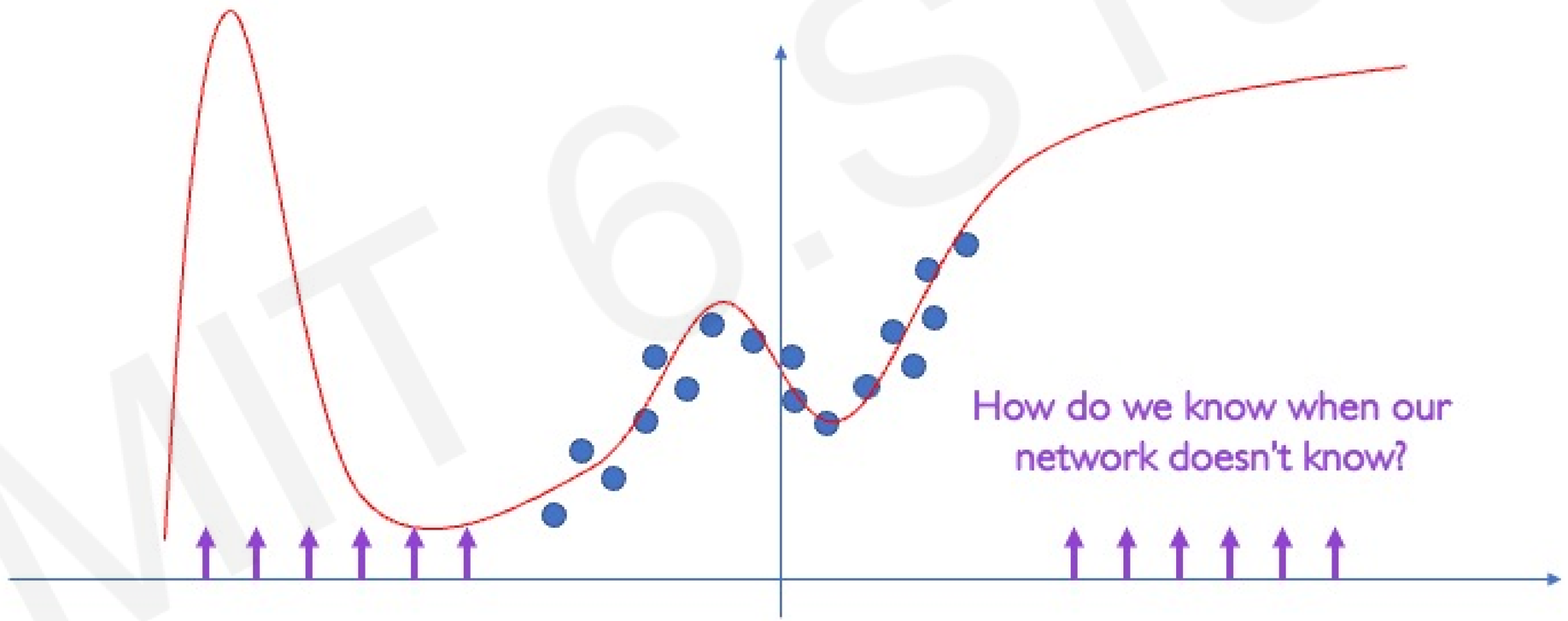
Neural networks are **excellent** function approximators



Neural Networks as Function Approximators

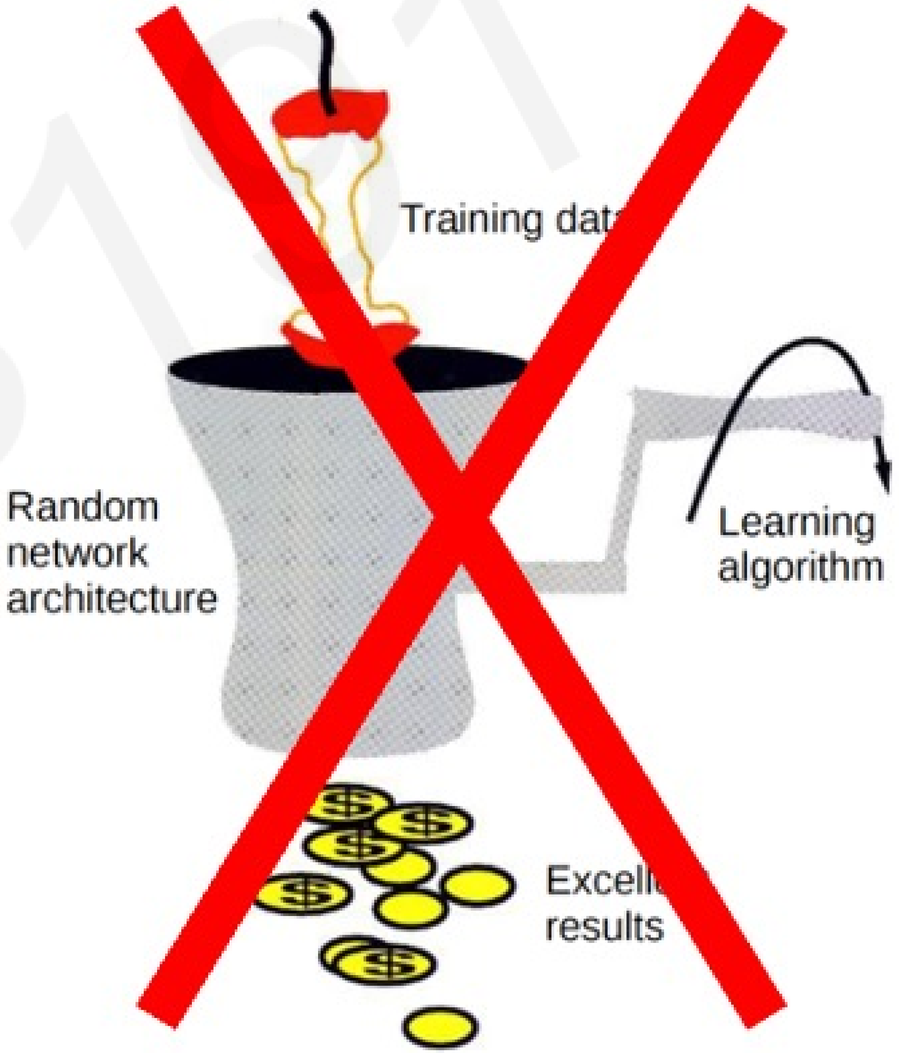
Neural networks are **excellent** function approximators

…when they have training data

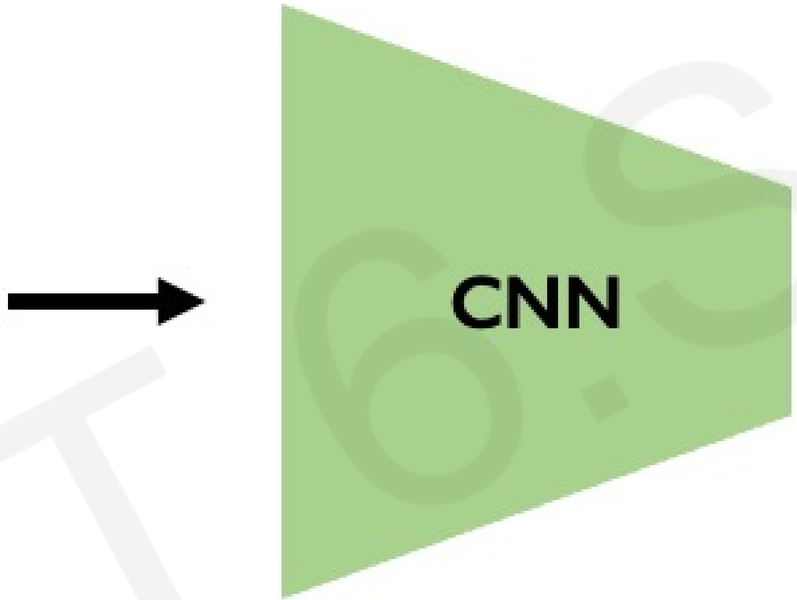


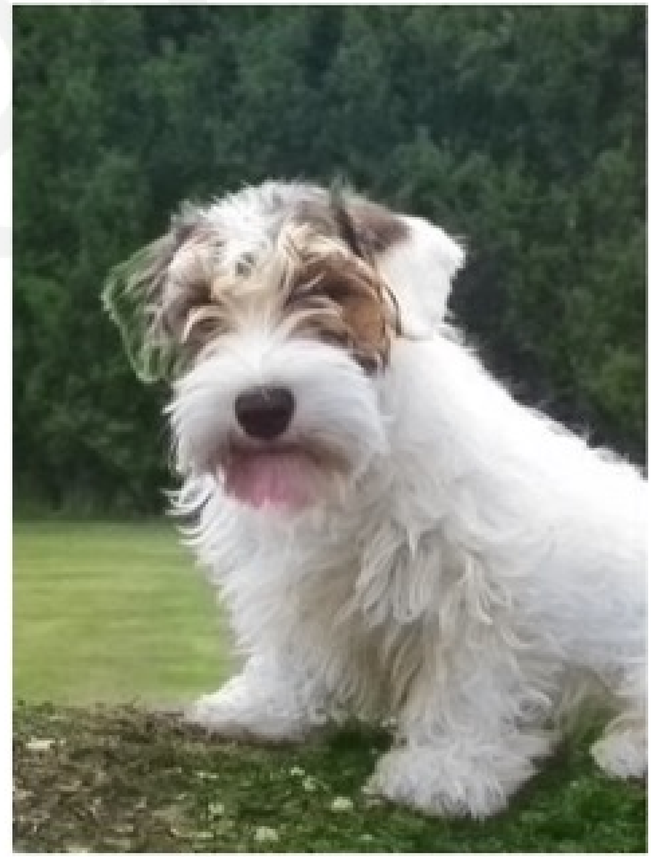
Deep Learning = Alchemy?

NO!



Neural Network Failure Modes, Part 1

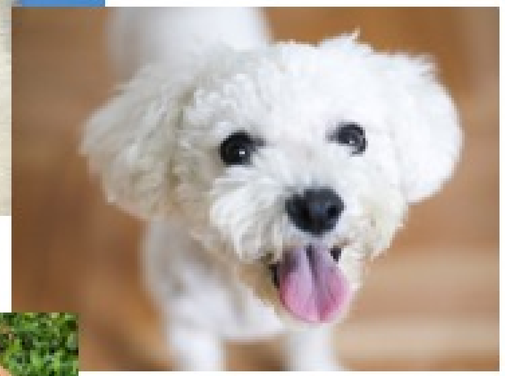
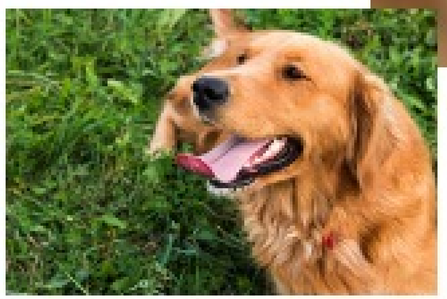
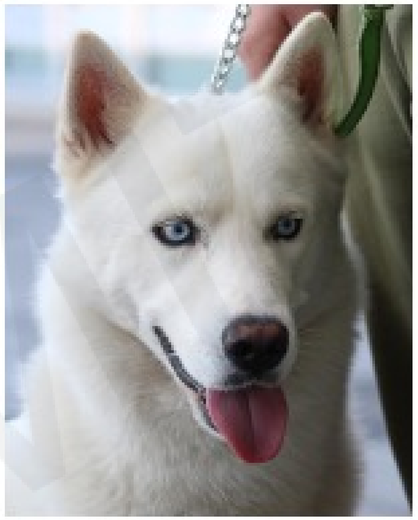




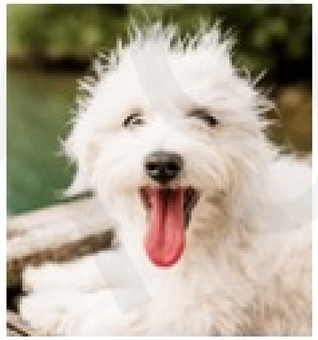
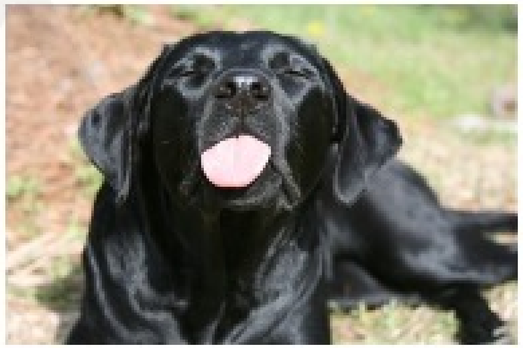
Train network to
colorize BW images.

Why could this be the case?

What Happens During Training…


CNN

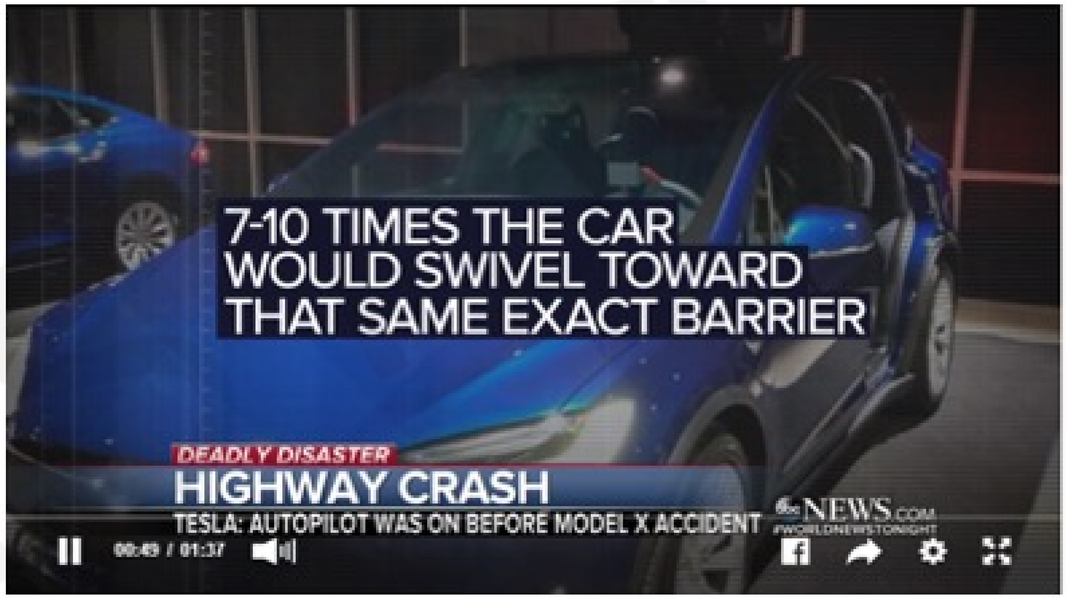

Neural Network Failure Modes, Part II

Tesla car was on autopilot prior to fatal crash in California, company says

The crash near Mountain View, California, last week killed the driver.

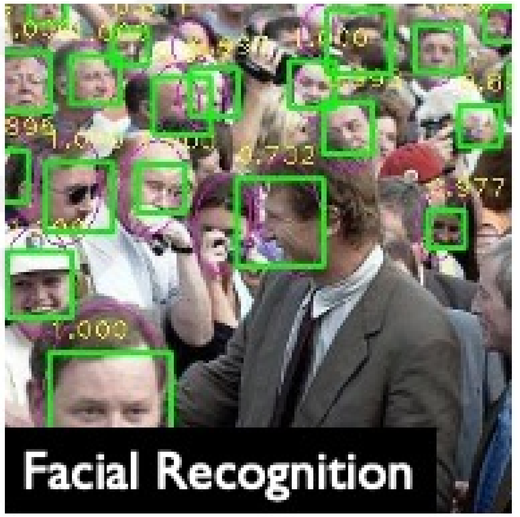
By **Mark Osborne**
March 31, 2018, 1:37 AM • 5 min read

abc NEWS

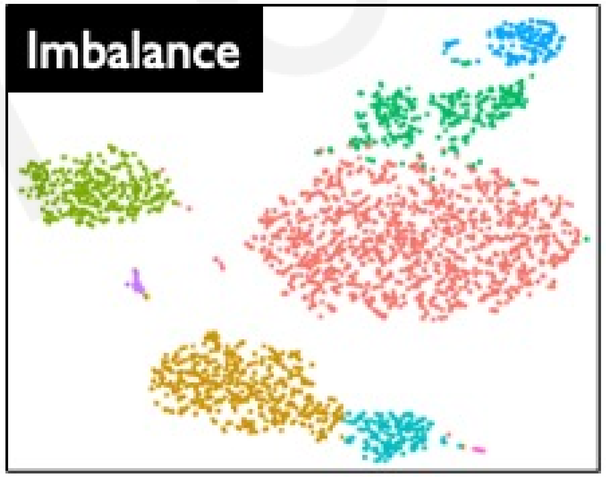


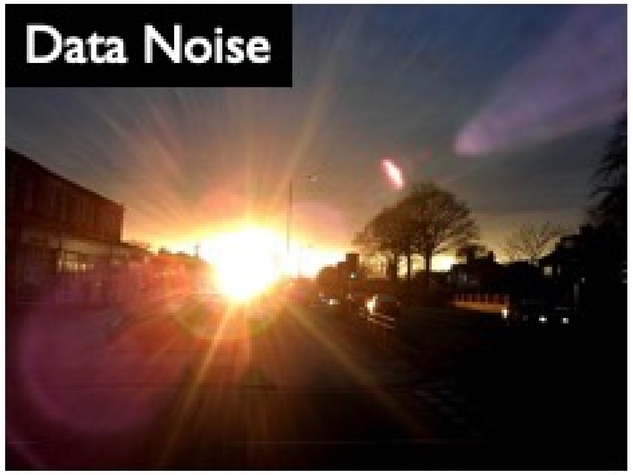
Uncertainty in Deep Learning

Safety-critical applications



Sparse and/or noisy datasets

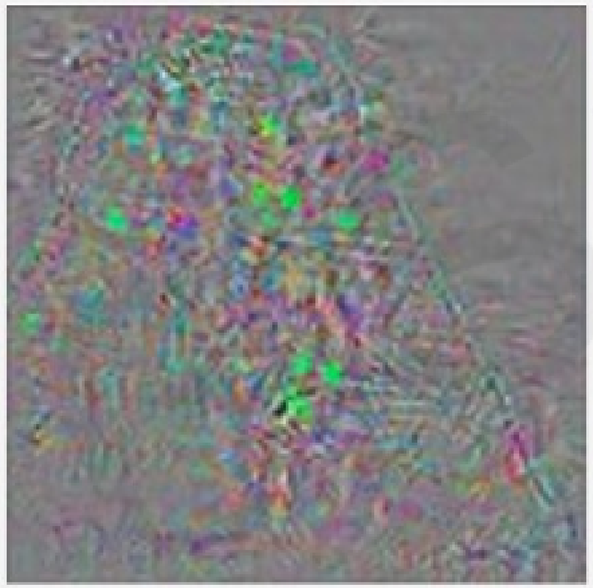




Neural Network Failure Modes, Part III

Original image

Temple (97%)

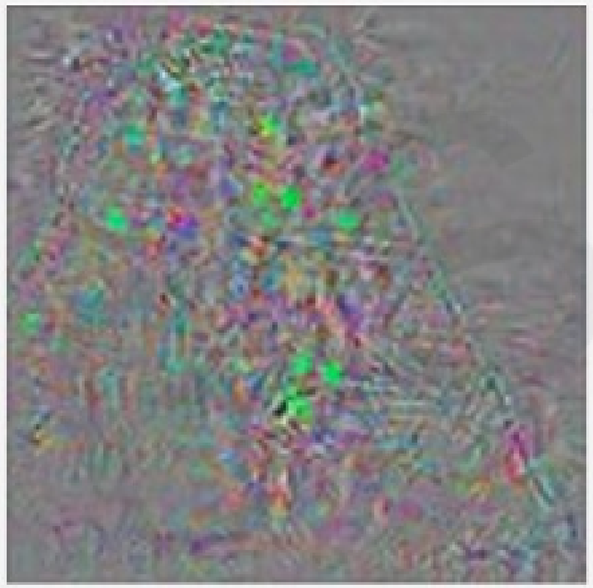


Perturbations

Adversarial example

Ostrich (98%)

Adversarial Attacks on Neural Networks



Perturbations

Adversarial Attacks on Neural Networks

Remember:

We train our networks with gradient descent

$$W \leftarrow W - \eta \frac{\partial J(W, x, y)}{\partial W}$$

"How does a small change in weights decrease our loss"

Adversarial Attacks on Neural Networks

Remember:

We train our networks with gradient descent

$$W \leftarrow W - \eta \frac{\partial J(W, x, y)}{\partial W}$$

"How does a small change in weights decrease our loss"

Adversarial Attacks on Neural Networks

Remember:

We train our networks with gradient descent

$$W \leftarrow W - \eta \frac{\partial J(W, x, y)}{\partial W}$$

Fix your image x, and true label y

"How does a small change in weights decrease our loss"

Adversarial Attacks on Neural Networks

Adversarial Image:

Modify image to increase error

$$x \leftarrow x + \eta \frac{\partial J(W, x, y)}{\partial x}$$

"How does a small change in the input increase our loss"

Adversarial Attacks on Neural Networks

Adversarial Image:

Modify image to increase error

$$x \leftarrow x + \eta \frac{\partial J(W, x, y)}{\partial x}$$

"How does a small change in the input increase our loss"

Adversarial Attacks on Neural Networks

Adversarial Image:

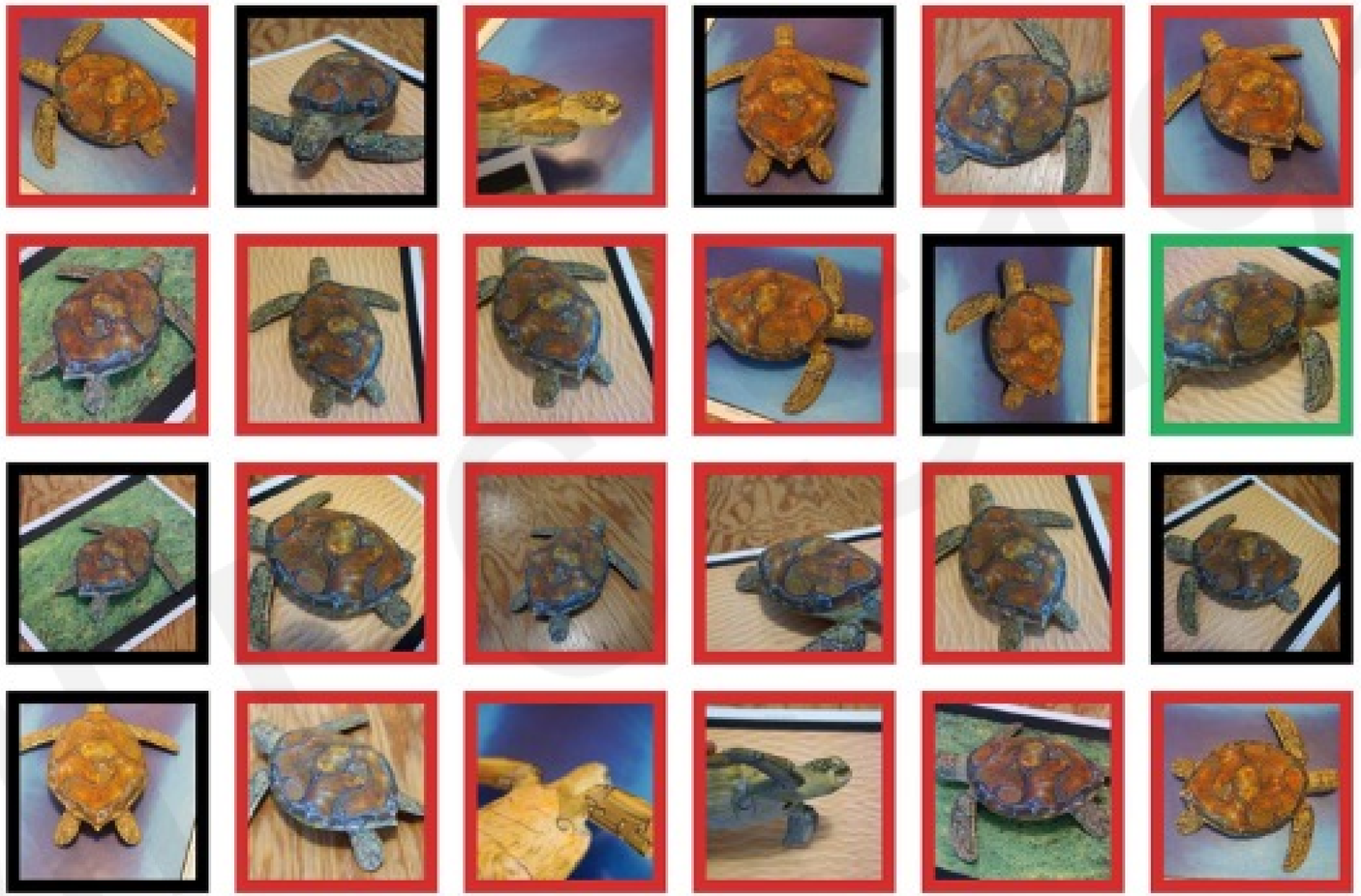
Modify image to increase error

$$x \leftarrow x + \eta \frac{\partial J(W, x, y)}{\partial x}$$

Fix your weights θ, and true label y

"How does a small change in the input increase our loss"

Synthesizing Robust Adversarial Examples



classified as turtle classified as rifle

classified as other

Algorithmic Bias

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk – A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

Artificial Intelligence has a gender bias problem – just ask Siri

6.S191 Lab

Neural Network Limitations…

- Very **data hungry** (eg. often millions of examples)
- **Computationally intensive** to train and deploy (tractably requires GPUs)
- Easily fooled by **adversarial examples**
- Can be subject to **algorithmic bias**
- Poor at **representing uncertainty** (how do you know what the model knows?)
- Uninterpretable **black boxes**, difficult to trust
- Often require **expert knowledge** to design, fine tune architectures
- Difficult to **encode structure** and prior knowledge during learning
- **Extrapolation**: struggle to go beyond the data

Neural Network Limitations…

- Very **data hungry** (eg. often millions of examples)
- **Computationally intensive** to train and deploy (tractably requires GPUs)
- Easily fooled by **adversarial examples**
- Can be subject to **algorithmic bias**
- Poor at **representing uncertainty** (how do you know what the model knows?)
- Uninterpretable **black boxes**, difficult to trust
- Often require **expert knowledge** to design, fine tune architectures
- Difficult to **encode structure** and prior knowledge during learning
- **Extrapolation**: struggle to go beyond the data

6.S191 Lab

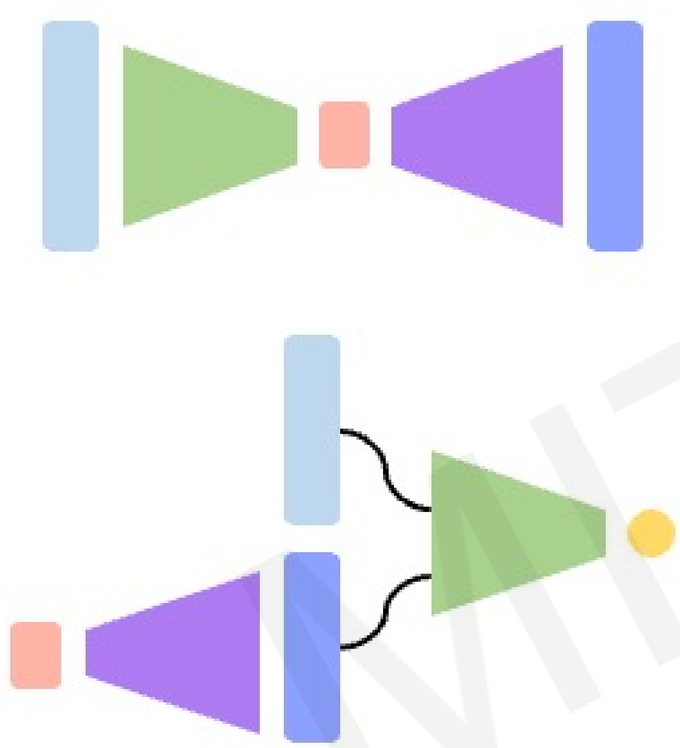
Neural Network Limitations…

- Very **data hungry** (eg. often millions of examples)
- **Computationally intensive** to train and deploy (tractably requires GPUs)
- Easily fooled by **adversarial examples**
- Can be subject to **algorithmic bias**
- Poor at **representing uncertainty** (how do you know what the model knows?)
- Uninterpretable **black boxes**, difficult to trust
- Often require **expert knowledge** to design, fine tune architectures
- Difficult to **encode structure** and prior knowledge during learning
- **Extrapolation**: struggle to go beyond the data

New Frontiers I:
Generative AI & Diffusion Models

The Landscape of Generative Modeling

Lecture 4:
VAEs and GANs



Limitations

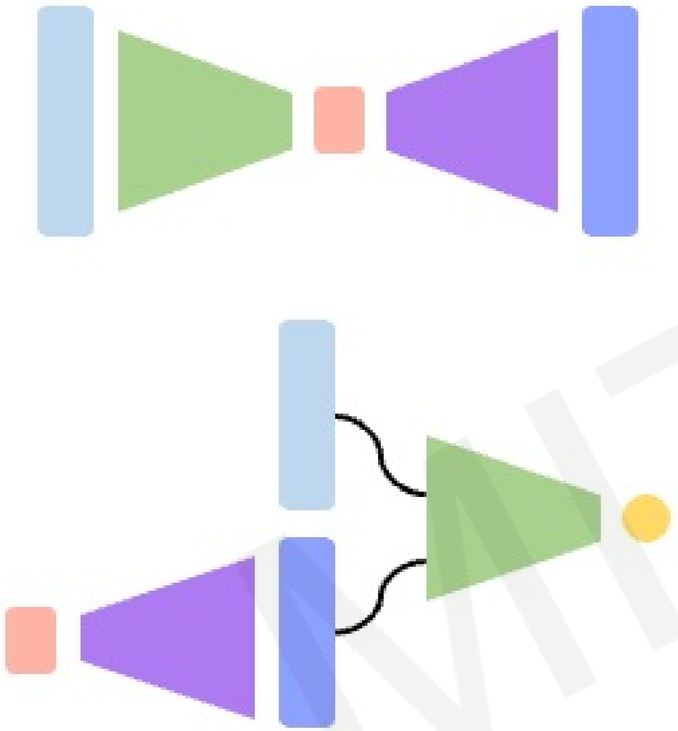
- Mode collapse
- Generating OOD
- Hard to train

Challenges

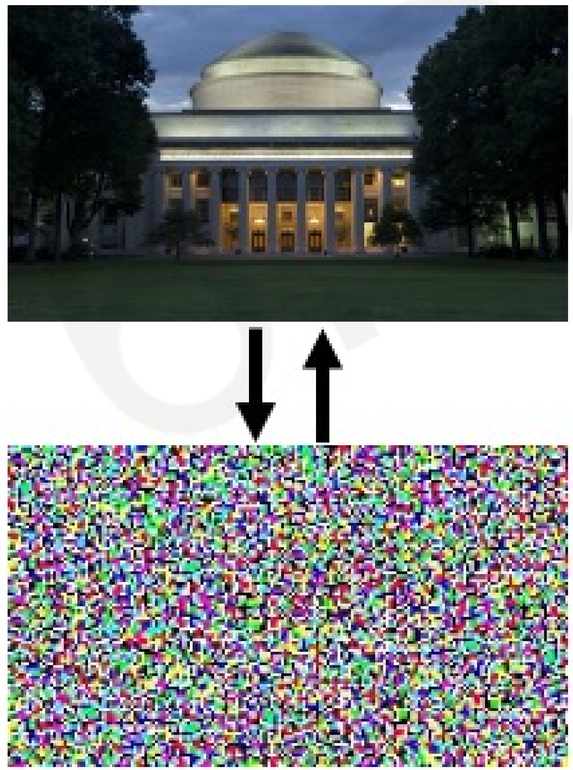
- Stability
- Efficiency
- Quality
- Novelty

The Landscape of Generative Modeling

Lecture 4:
VAEs and GANs



Diffusion Models

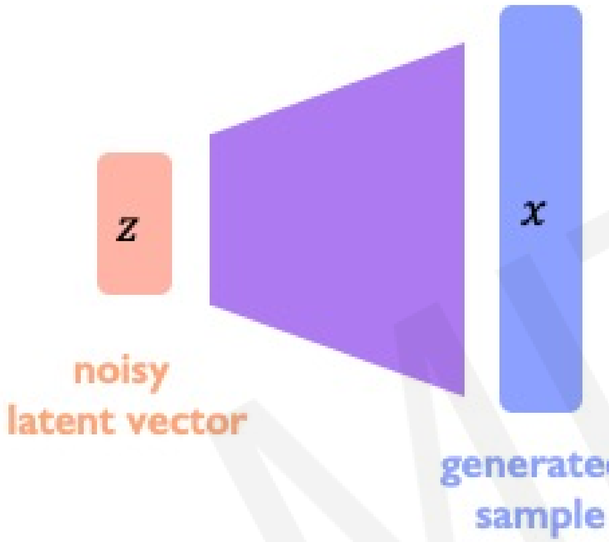


Text-to-Image

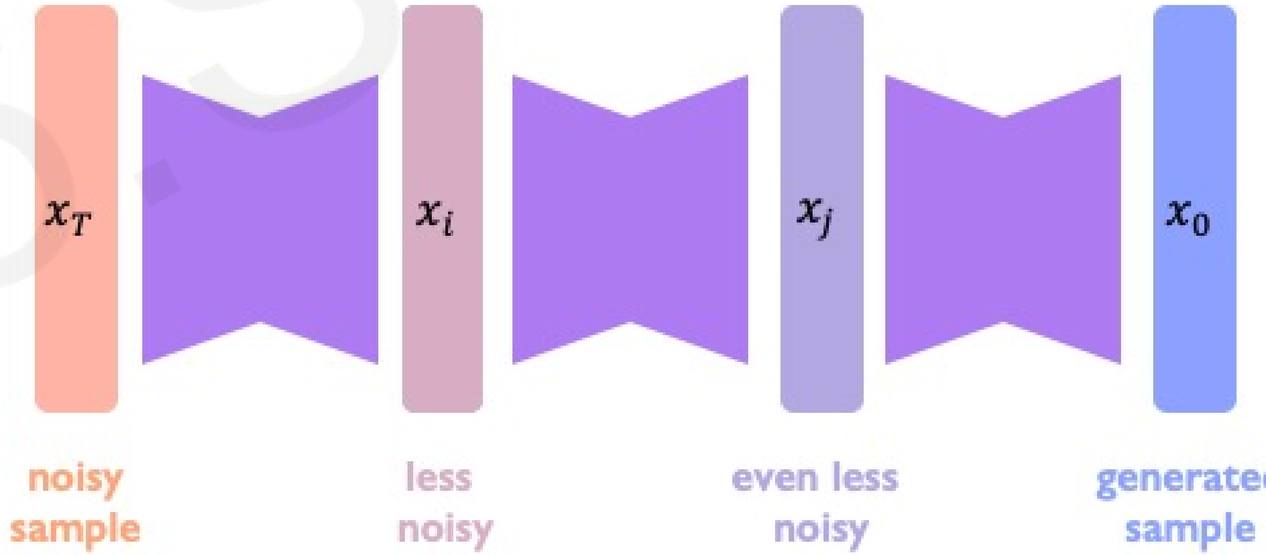
"Two cats doing research"

Diffusion Models

VAEs/GANs: Generating samples <u>in one-shot</u> directly from low-dimensional latent variables



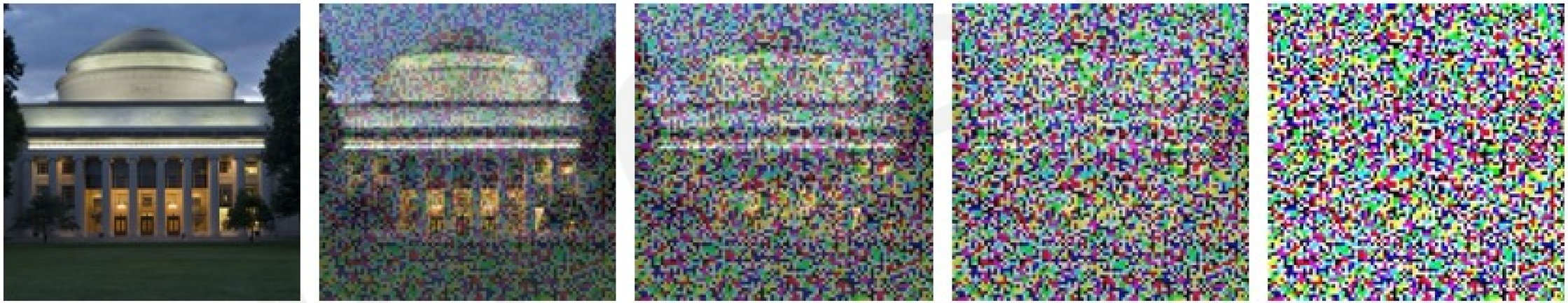
Diffusion: Generating samples <u>iteratively</u> by repeatedly refining and removing noise



The Diffusion Process

Forward noising

(data-to-noise)

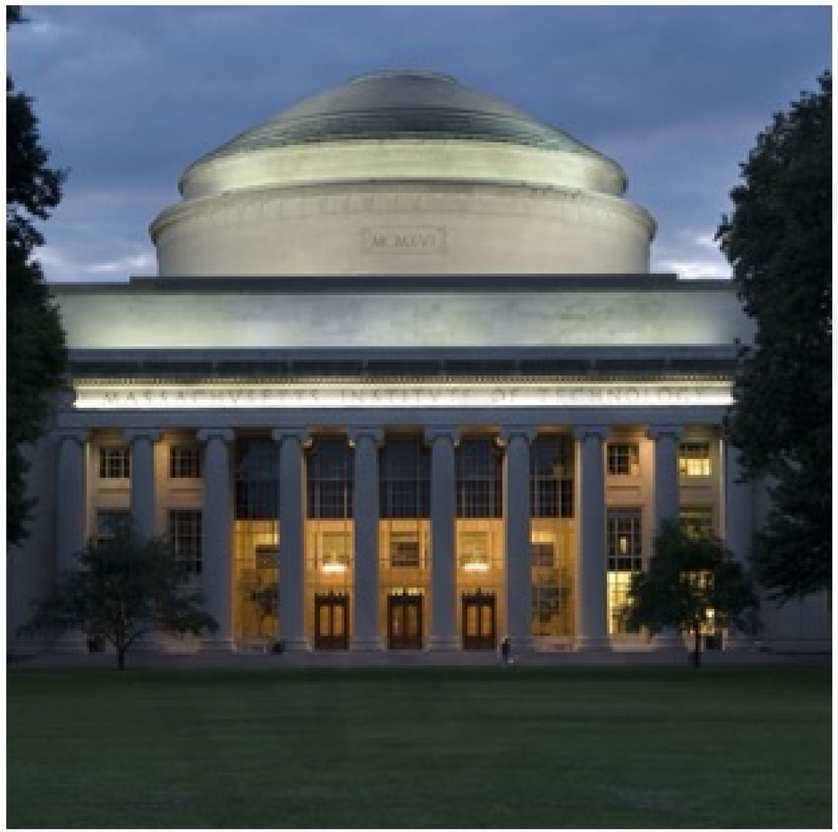


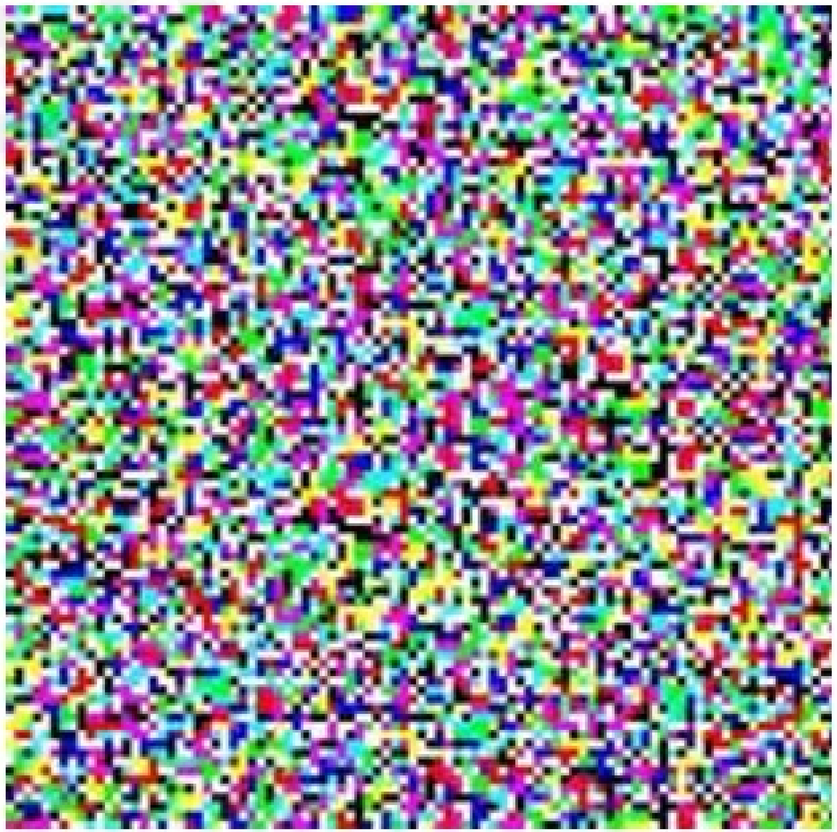
Reverse denoising

(noise-to-data)

Forward Noising

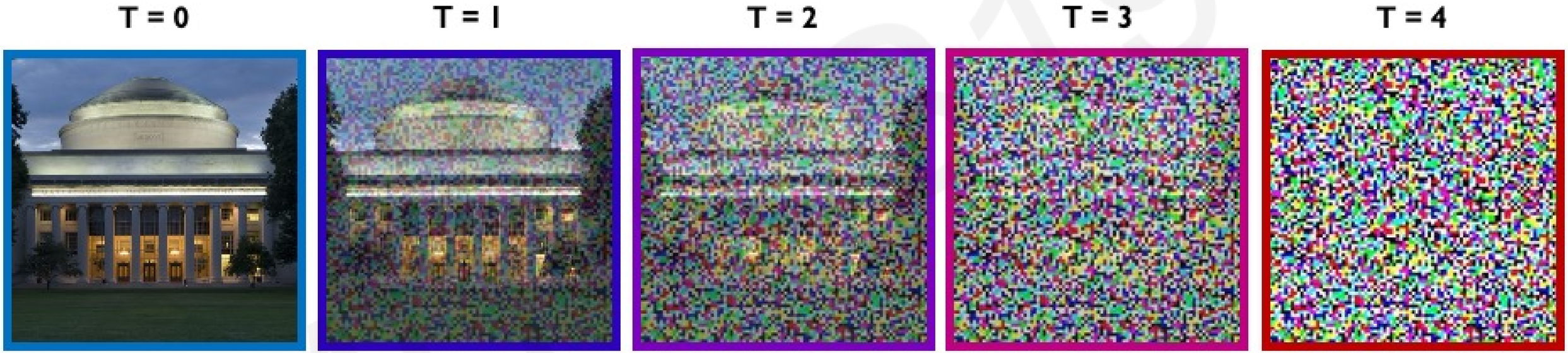
Step 1: Given an image (left), randomly sample a random noise pattern (right)



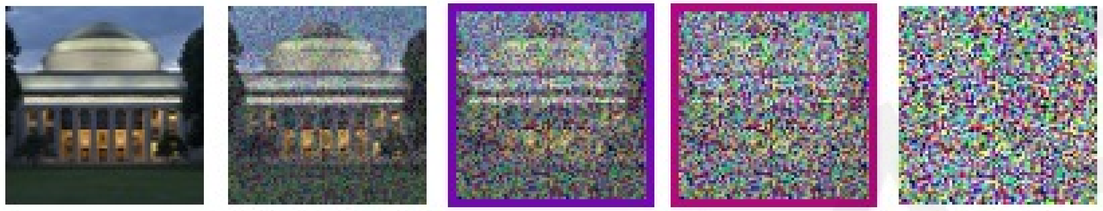


Forward Noising

Step 2: Progressively add more and more of the noise to your image



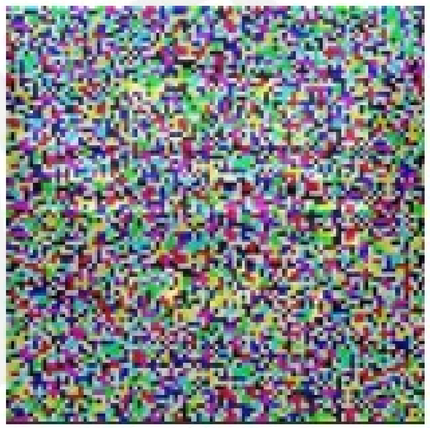
Reverse Denoising



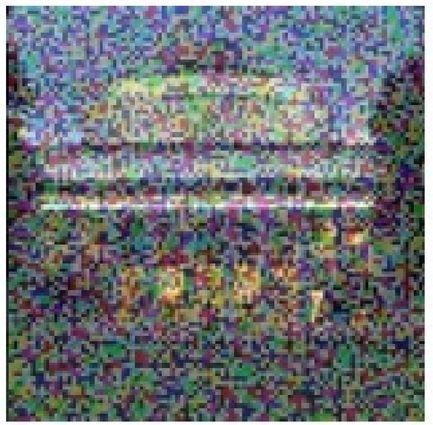
T = 0 ... T = 4

Goal: Given image at **T**, can we **learn** to estimate image at **T-1**?

T = 3



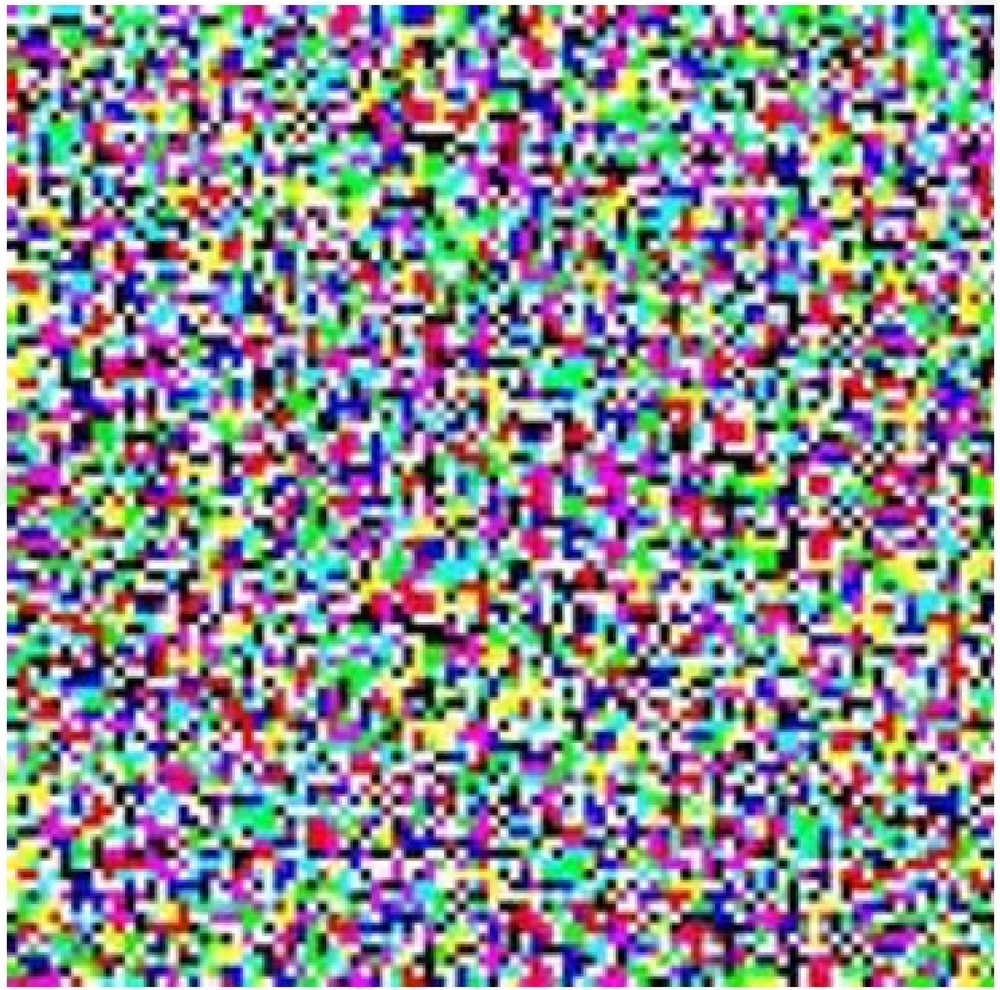
T = 2



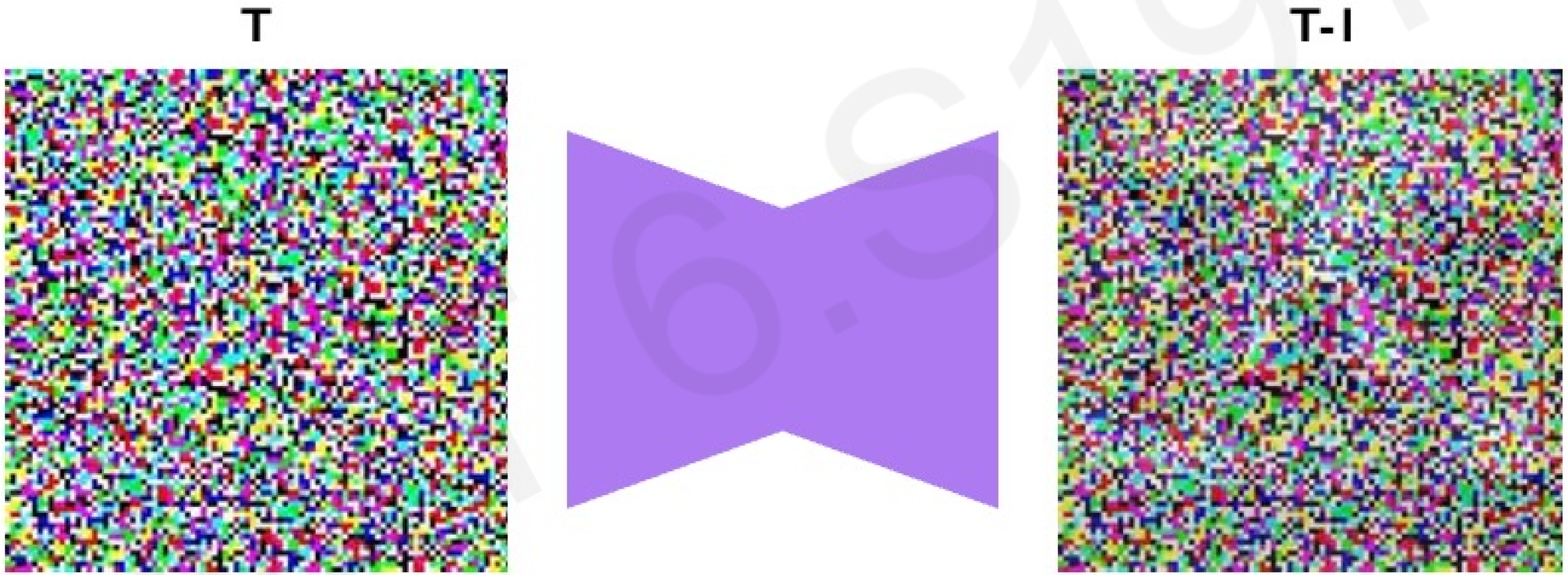
How can we train this network?

Sampling Brand New Generations

T



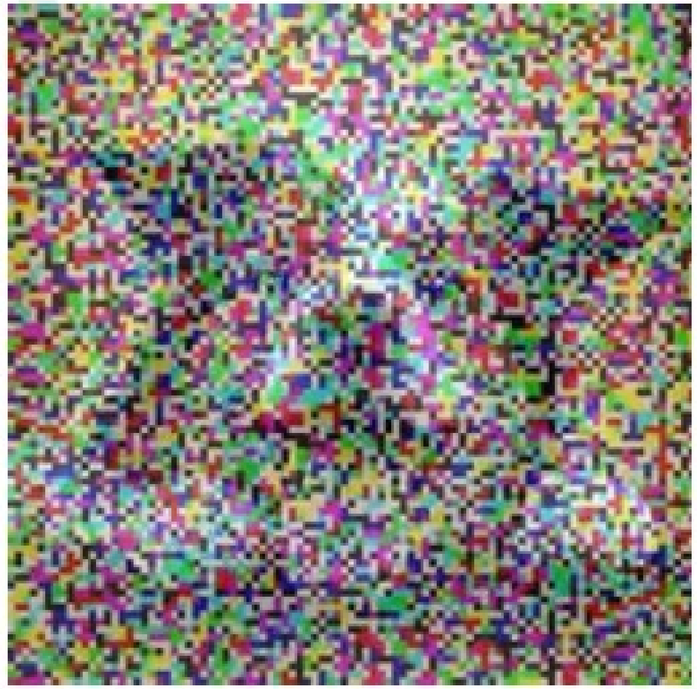
Sampling Brand New Generations



Sampling Brand New Generations

T-1

T-2



Sampling Brand New Generations

T-2

T-3

Sampling Brand New Generations

T-3

T-4

Sampling Brand New Generations

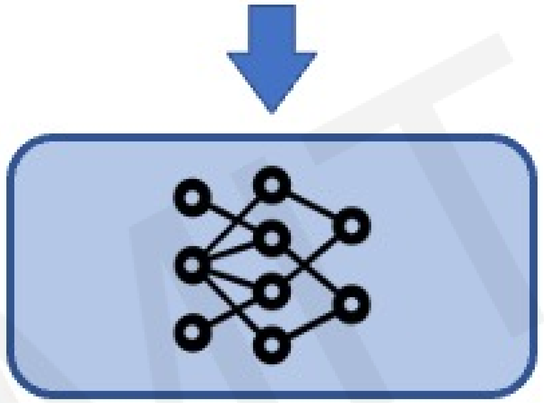
T-4

T0 (end)

Sampling Brand New Generations

Generating Images from Natural Language

"A photo of an astronaut riding a horse."



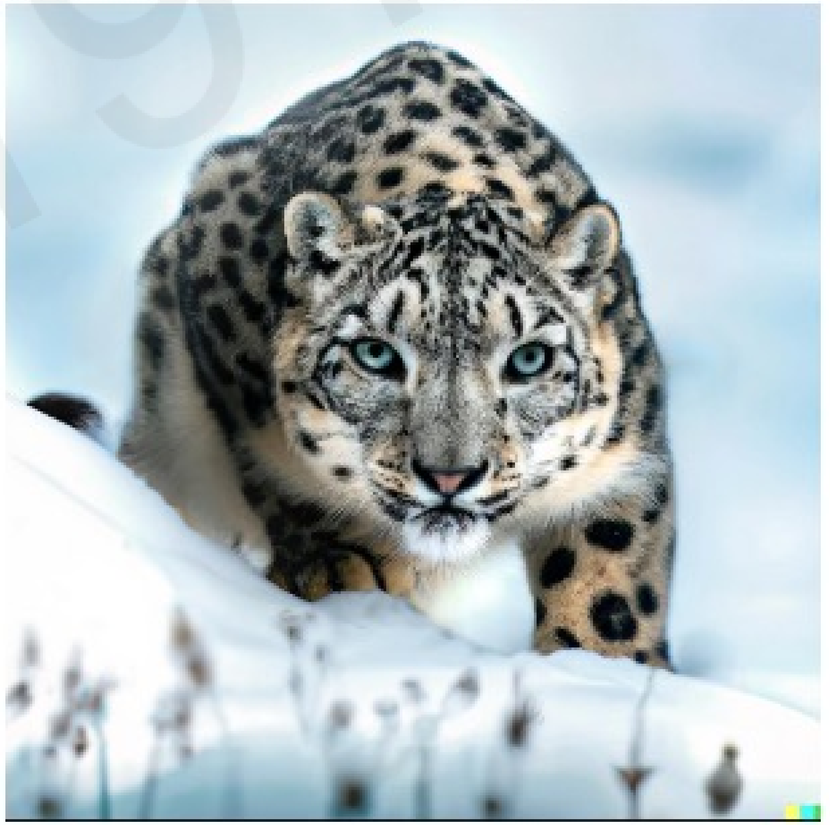
Ramesh+ *arXiv* 2022

Text-to-Image Generation

"a painting of a fox sitting in a field at sunrise in the style of Claude Monet"

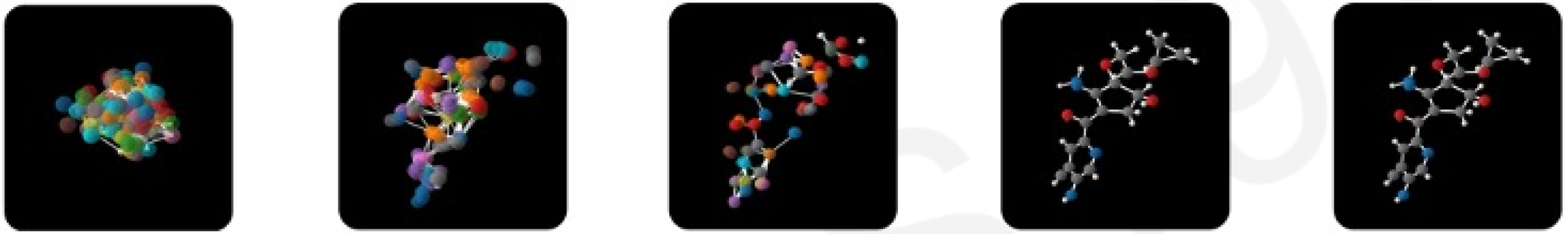
"an ibis in the wild, painted in the style of John Audubon"

"close-up of a snow leopard in the snow hunting, rack focus, nature photography"



Beyond Images: Molecular Design

Chemistry: Generating Molecules in 3D



Noise ---------- Molecule

Hoogeboom+ *ICML* 2022, Jing+ *NeurIPS* 2022.

Biology: Generating Novel Proteins

Wu+ arXiv 2022, Anand+ *arXiv* 2022, Trippe+ *arXiv* 2022, Jumper+ *arXiv* 2022, and more …

Generative Models for Protein Design

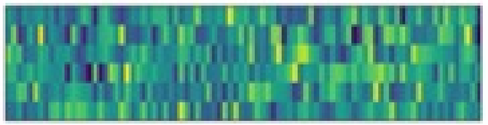
Can we design **new proteins** with new biological or therapeutic functions?

Protein function is encoded in sequence and structure.

Protein Structure Generation via Folding Diffusion

Structure, unfolded

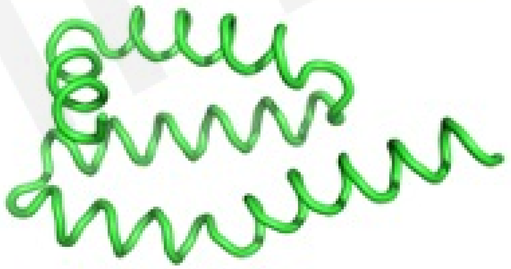
Angles, unfolded



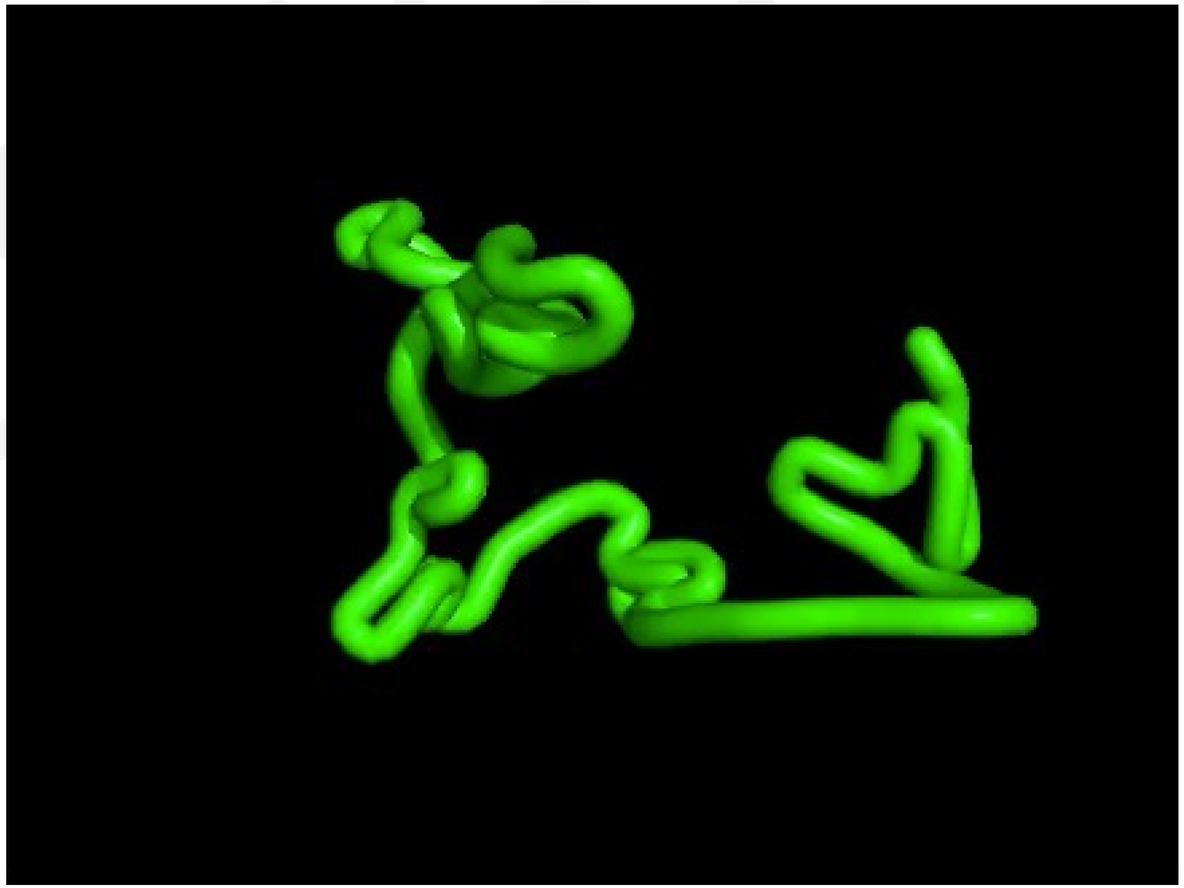
Diffusion Model

Angles, generated

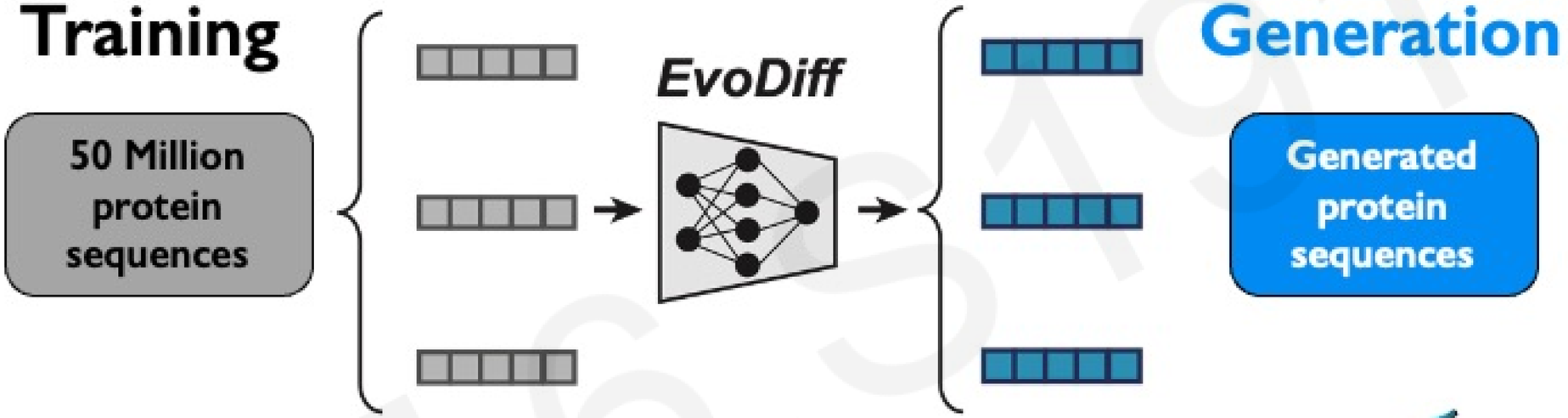
Structure, generated

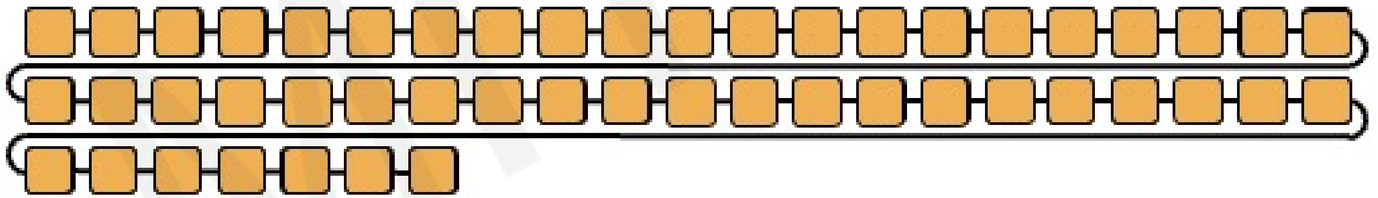


Denoising to generate a structure!



Evolutionary-scale Diffusion for Protein Design

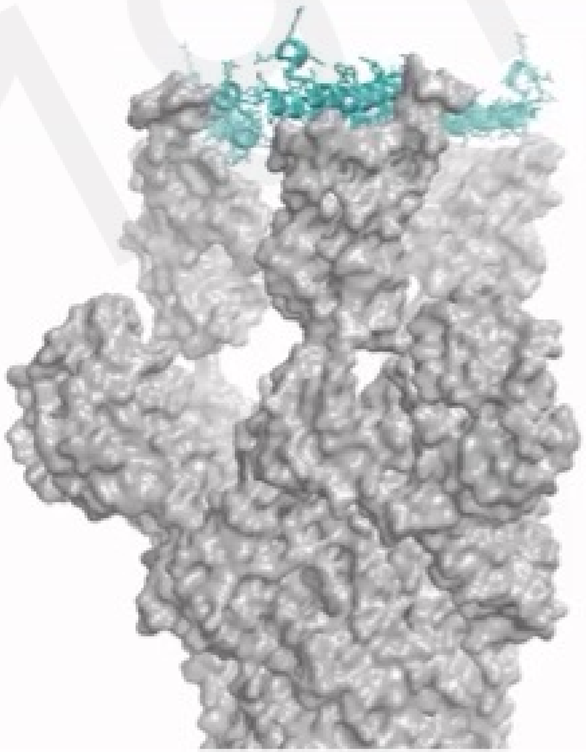




Diffusion Models: Foundation for Programmable Protein Design

AI-generated protein design

Real-world structure



Generating a novel binder to COVID spike

New Frontiers II:
Large Language Models

Large Language Models (LLMs) and the World

ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

GPT-4

OpenAI

What are LLMs?

Artificial Intelligence

Any technique that enables computers to mimic human behavior

Deep Learning

Extract patterns from data using neural networks

Large Language Models

Very, very large neural networks trained on very, very large sets of text

A B C D E F
G H I J K L
M N O P Q R

What is GPT?

GPT is a large language model (LLM). LLMs learn a **probability distribution** over natural language.

Generative: generates text-based outputs.

Pre-trained: previously trained on a large, unannotated body of text.

Transformer: the architecture used.

How do LLMs like GPT work?

Training:

Dataset
Common Crawl, WebText, etc
Split into chunks – "tokens"

Model
GPT
175B parameters (GPT3)

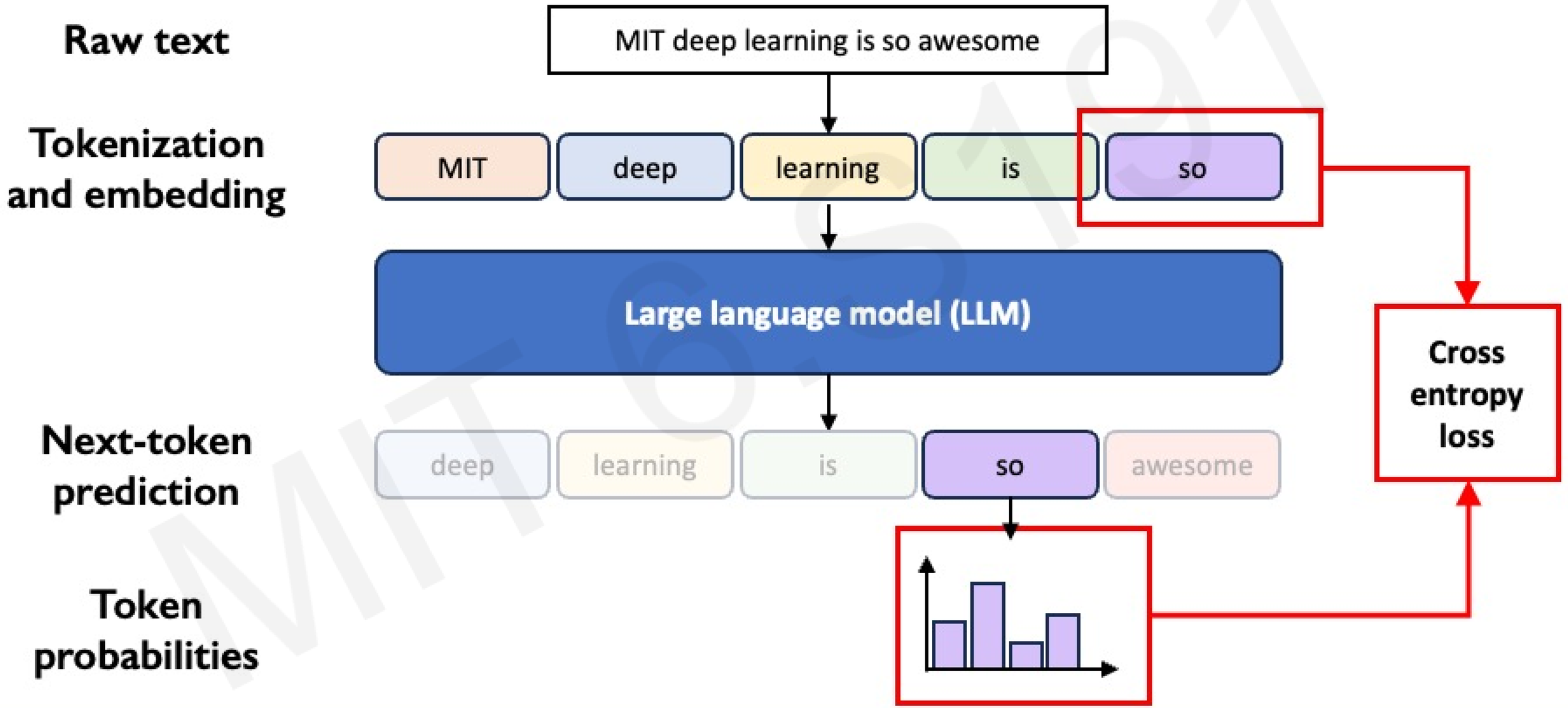
Task and Objective:

Given a sequence of tokens, predict the next token.

Update model parameters given how good next-token prediction is.

How does next token prediction work? 🤔

Next Token Prediction



Using LLMs to Generate Text

Training:

Dataset

Common Crawl, WebText, etc
Split into chunks – "tokens"

Model

GPT
175B parameters (GPT3)

Task and Objective:

Given a sequence of tokens, predict the next token.

Update model parameters given how good next-token prediction is.

Deployment:

I'm giving a talk on AI at MIT.
Can you outline it?

Introduction
What is AI?
How does AI work?
How can we use AI?

What capabilities do LLMs have?

Capabilities that are feasible and reliable now:

Knowledge Retrieval

Writing Co-Pilot

Planning Co-Pilot

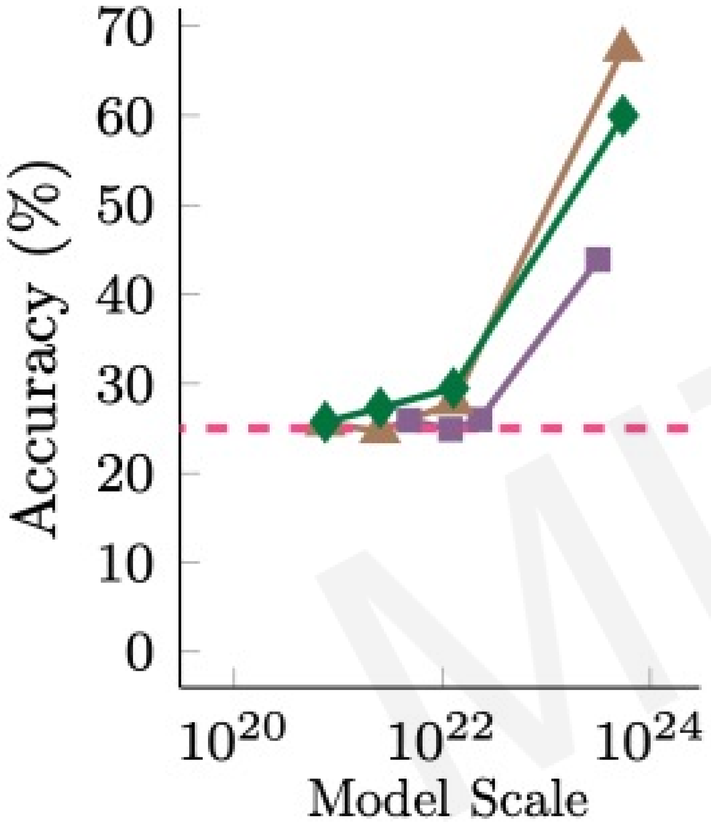
LLMs like GPT have shown mastery over natural language.

Can these models go beyond knowledge retrieval?

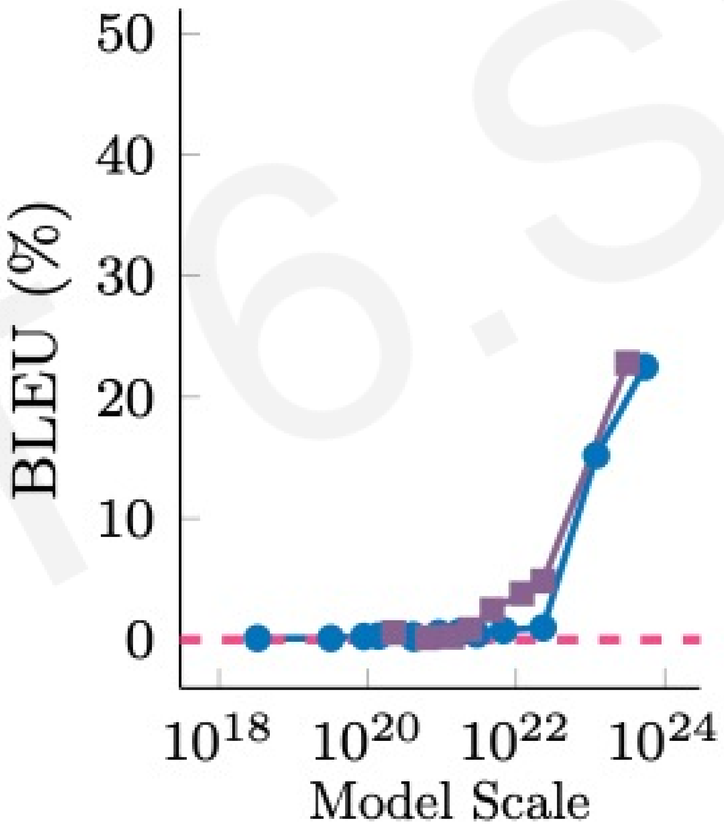
What can LLMs do?
Emergent Abilities with Scale.

An ability is **emergent** if it is not present in smaller models but is present in larger models.

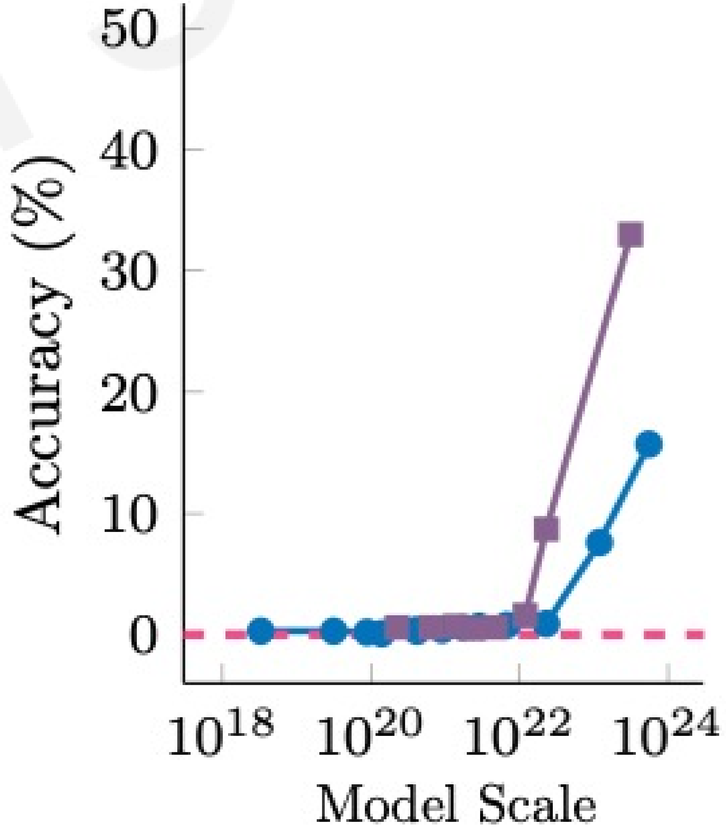
Structuring Language



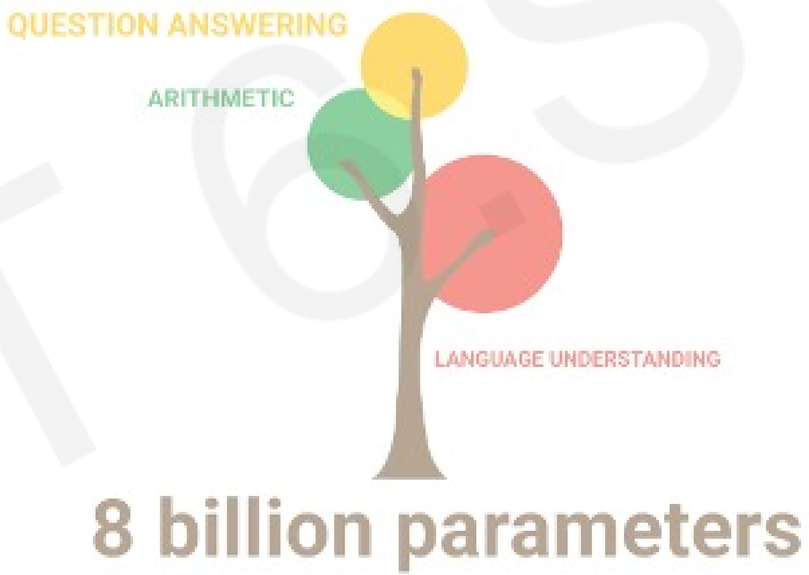
Understanding Phonetics



Performing Arithmetic



Emergent Abilities: Towards Intelligence



Limitations of LLMs

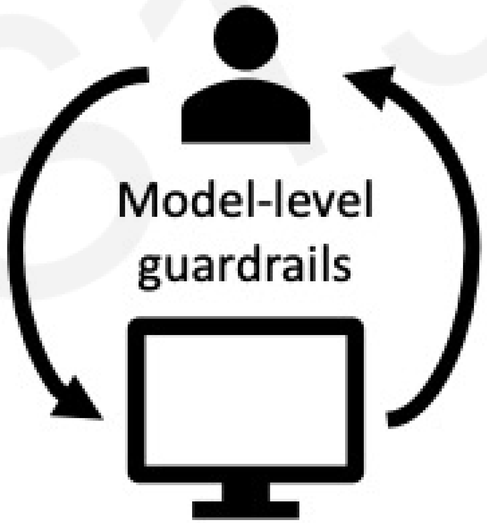
Robustness: How confident?

Cn @uN66rN you translate ths from Spanish to English?

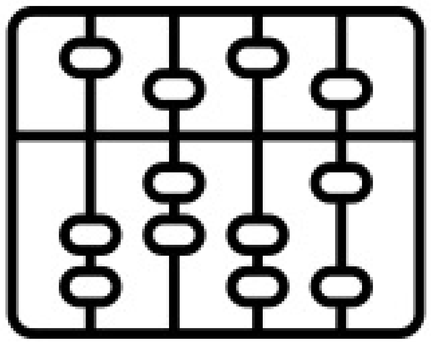
Wang+ *arXiv* 2023.

"Hallucinations": Confidently wrong

Guardrails and Jailbreaks



Logic and Numerics



Key challenges motivated by the high-level thinking process: robustness + confidence; long-term planning; logic and discovery

Generative AI Spawns a Powerful Idea

"What I cannot create, I cannot understand."
Richard Feynman

- Images, language, biology, and more
- Can Generative LLMs provide a central reasoning system?
- Design AI to improve and evolve AI itself
- Power and Caution of Generative AI

Connections and distinctions between artificial and human intelligence