UVA CS 4774: Machine Learning

S3: Survey (Extra)

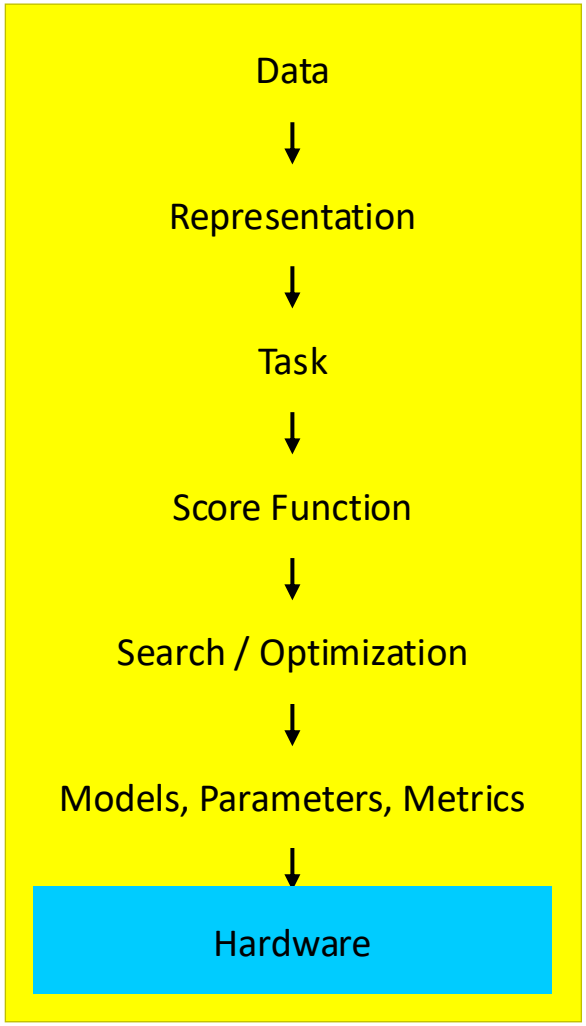
Deep Learning, 2020 to 2026: What Changed at Each Stage

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Survey

How to read this deck: follow the nutshell



One question at every stage:

What changed from 2020 to 2026, and what lasted?

- Each story runs: 2020 idea → what replaced it → what lasted
- Earlier topics are pointers, not repeats
- Today's focus is what past lectures left out

The whole map: 2020 vs 2026

Stage	The 2020 story	The 2026 story
Data	Grids, sequences, some graphs	Molecules, weather, video, 3D
Representation	CNN vs Transformer	Hybrids, sparse experts
Task	Classify, generate faces, play games	Video with sound, world models, proofs
Score function	Adversarial and reconstruction	Flow matching, joint embedding
Search	Search a defined space (NAS)	Agents edit code; scale drives progress
Metrics	Accuracy, saliency, robustness	Circuits, verification, time horizons
Hardware	GPUs, TPUs, edge pruning	Inference chips, heavy quantization

UVA CS 4774: Machine Learning

S3: DNN Survey (20-26)

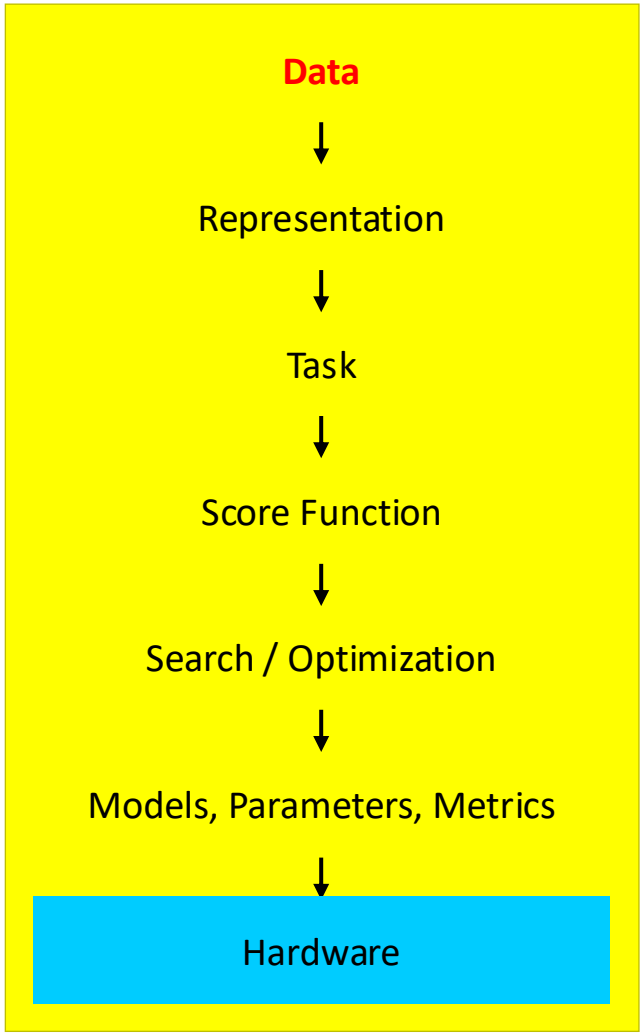
Data and Representation

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Module I

Nutshell: I.1 Data



In 2020

- Graphs, trees and sets were a new trend
- Examples from chemistry tutorials

By 2026

- Graphs and 3D geometry in major science results
- Protein structure, weather forecasting
- Prizes, and operational systems

2024: Nobel Prizes reach neural networks

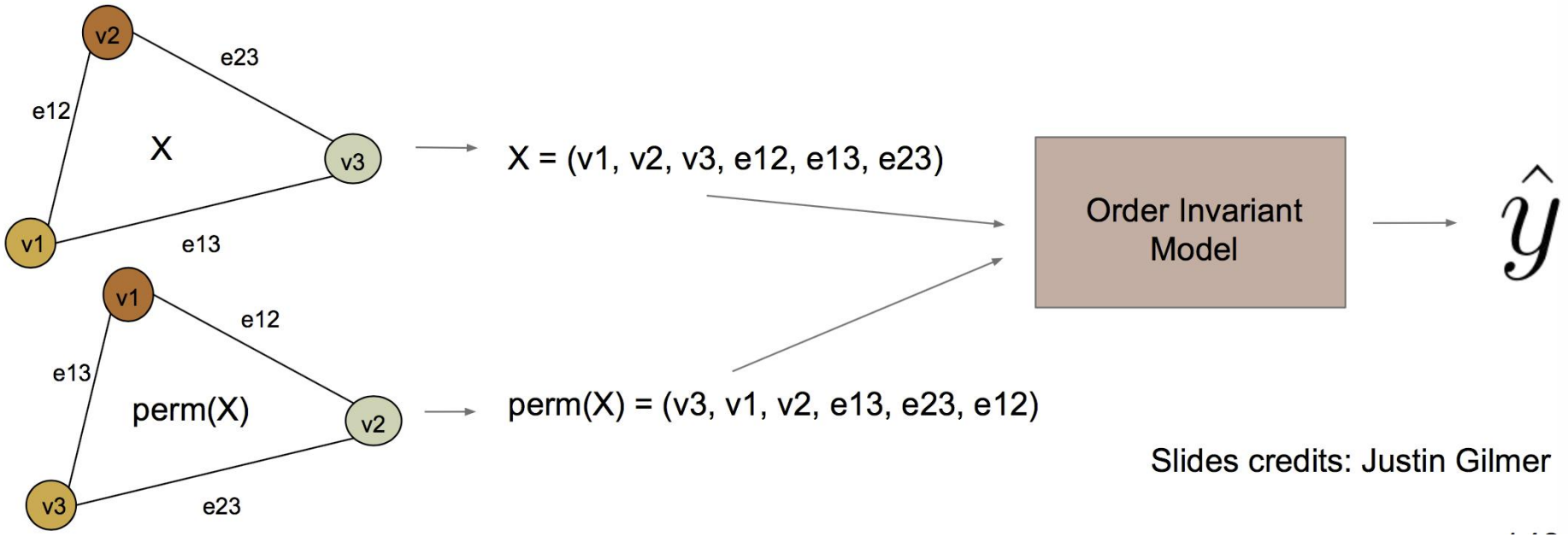
- Physics: John Hopfield and Geoffrey Hinton
  - “Foundational discoveries and inventions that enable machine learning with artificial neural networks”
- Chemistry: David Baker, Demis Hassabis, John Jumper
  - Baker for computational protein design
  - Hassabis and Jumper for protein structure prediction: AlphaFold

source: nobelprize.org, 2024 Physics and Chemistry

Inductive Bias for Graphs



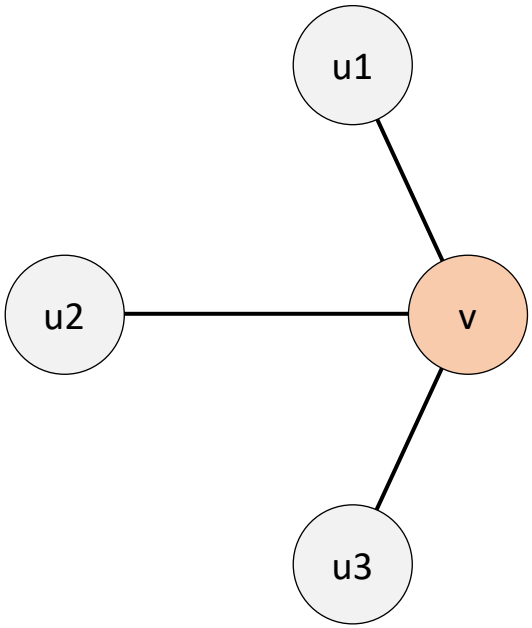
- If we have a graph on N nodes, there are N! possible orderings of the nodes.
- Ideally want a model invariant to the order of nodes.



Slides credits: Justin Gilmer

credit: Y. Qi, CS 4774 survey deck (2020); J. Gilmer, NIPS 2017 tutorial

Message passing: the core GNN idea

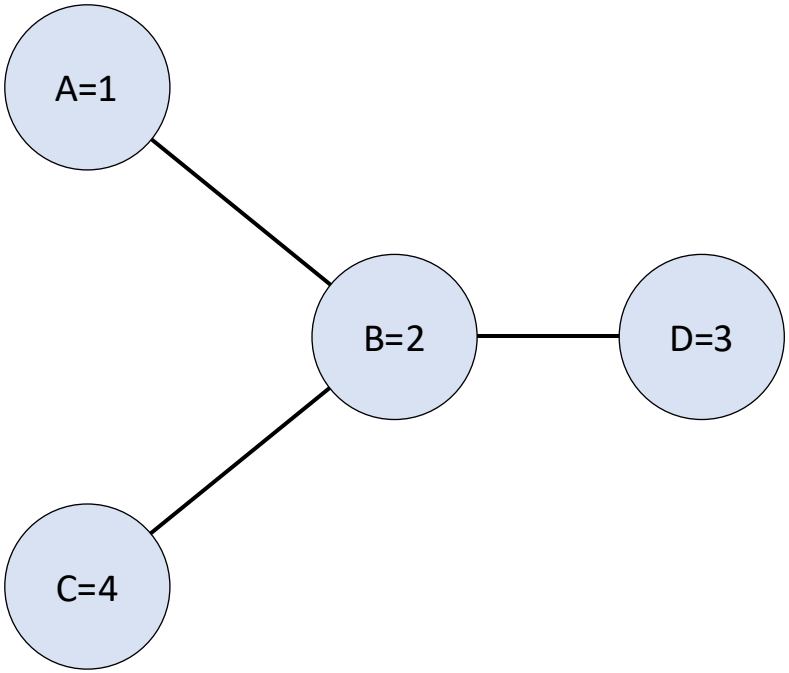


- Each neighbor sends a message
- Aggregate: sum or mean, so order does not matter
- Update the node's vector
- Stack K layers: each node sees K hops
- Same weights at every node

$$h_v^{(k+1)} = \phi\left(h_v^{(k)}, \sum_{u \in N(v)} \psi\left(h_u^{(k)}\right)\right)$$

source: Gilmer et al. 2017; Stanford CS224W

Worked example: one message-passing step



Node	Neighbors	New value
A	B	1.50
B	A, C, D	2.33
C	B	3.00
D	B	2.50

$$h'_v = \frac{1}{2}h_v + \frac{1}{2}\,\mathrm{mean}_{u \in N(v)}(h_u)$$

Symmetry as a design rule: 3D data

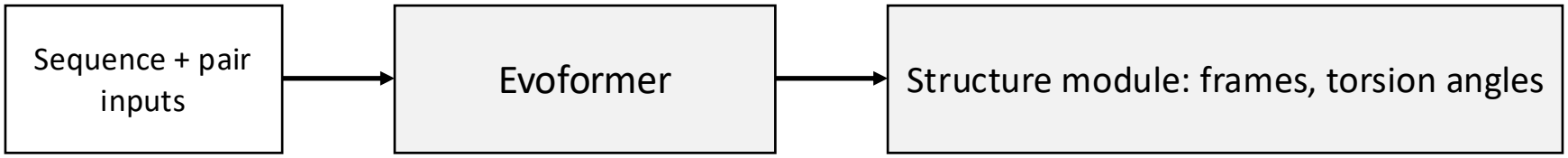
- Rotate a molecule: its energy stays the same
- Rotate a molecule: its forces rotate with it
- Invariant output: unchanged by the transformation
- Equivariant output: changes in step with the transformation
- Build the symmetry in, so the network need not learn it

Geometric deep learning: pick the data's symmetry, then design the layers

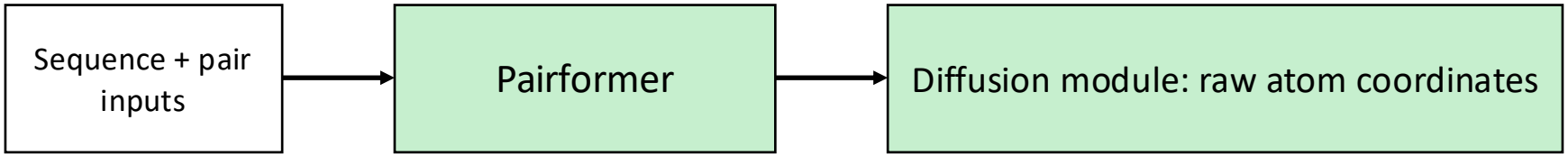
source: Bronstein et al. 2021, Geometric Deep Learning

AlphaFold 2 → 3: what changed

AlphaFold 2



AlphaFold 3

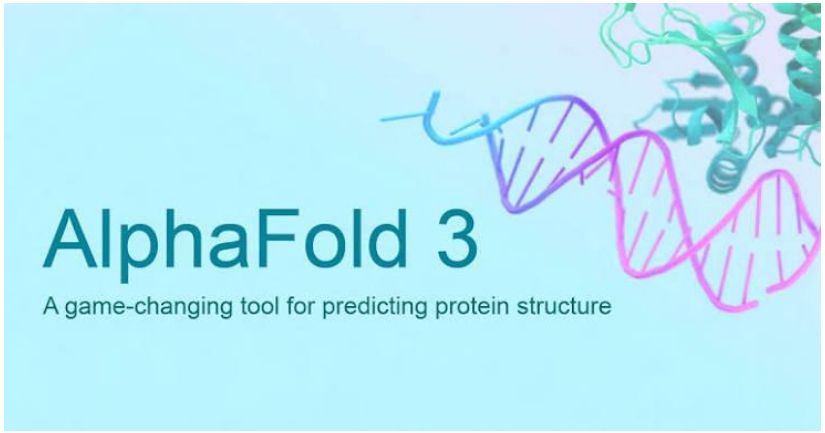


Same overall plan: a pairwise trunk, then a module that makes 3D structure.
New scope: proteins, DNA, RNA, ligands, ions

source: Abramson et al., Nature 2024

AlphaFold 3's diffusion module

- Input: noisy atom positions, plus the trunk's representations
- Output: denoised atom coordinates
- Repeat from pure noise to get a structure
- Multiscale noise: local chemistry and global shape
- The paper drops stereochemical losses and most special bond handling

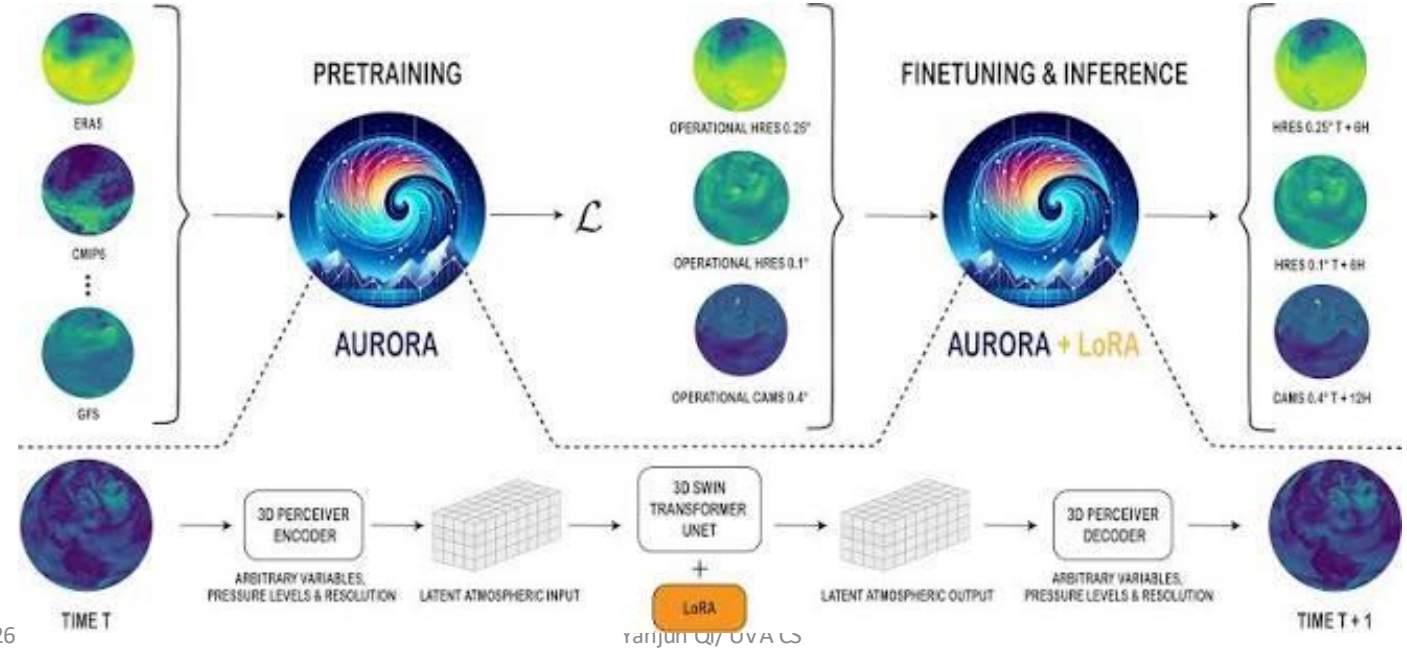


A generic generative tool replaced hand-built geometry

source: Abramson et al., Nature 2024

AI weather models: GenCast and Aurora

- Train on decades of reanalysis data, not physics equations
- GenCast (DeepMind): diffusion ensemble
  - 0.25°, 15 days, beat ECMWF ENS on 97.2% of 1,320 targets
  - An ensemble in about 8 minutes
- Aurora (Microsoft): a foundation model, then fine-tuned
  - Trained on over a million hours of weather and climate data



source: Price et al., Nature 2025 (GenCast); Bodnar et al. 2025 (Aurora)

From paper to operations: ECMWF AIFS

- AIFS Single: operational since 25 Feb 2025
- AIFS ENS (ensemble): operational since 1 Jul 2025
- Version 2 went live on 12 May 2026
- Runs four times a day, 6-hour steps, out to 15 days
- Open data under a CC-BY licence
- Still being tested: one preprint finds Aurora's surface extremes weaken after about 7–10 days

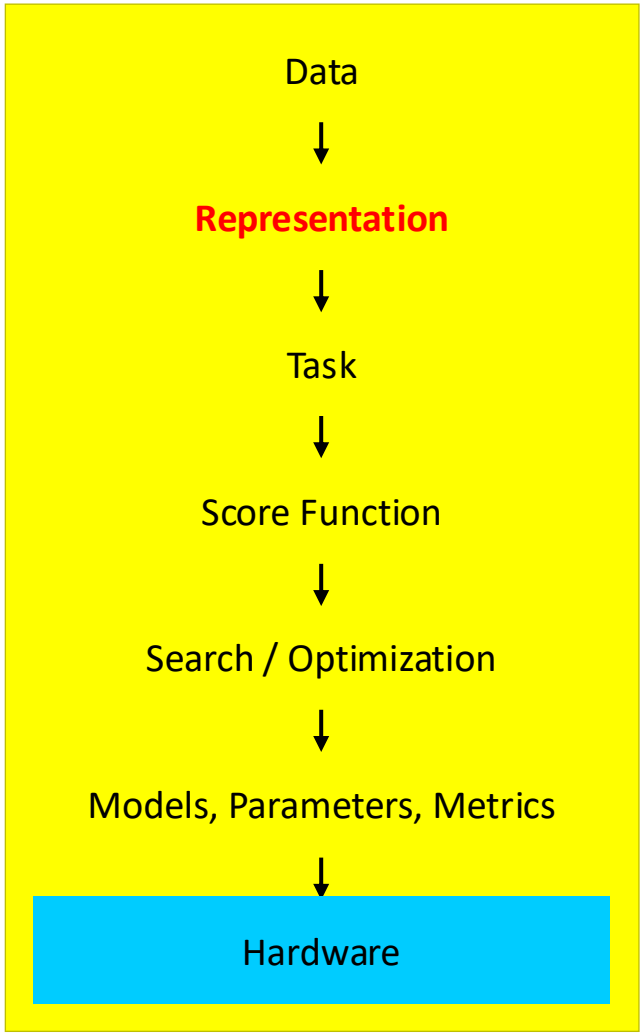
AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the Continuous Ranked Probability Score

Simon Lang, Mihai Alexe, Mariana C. A. Clare
Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue
Matthew Chantry, Jesper Dramsch, Peter D. Dueben
Sara Hahner, Pedro Maciel, Ana Prieto-Nemesio
Cathal O'Brien, Florian Pinault, Jan Polster
Baudouin Raoult, Steffen Tietsche, Martin Leutbecher

European Centre for Medium-Range Weather Forecasts (ECMWF)

source: ECMWF AIFS dataset page; Preprints.org 202606.1070 (preprint)

Nutshell: I.2 Representation



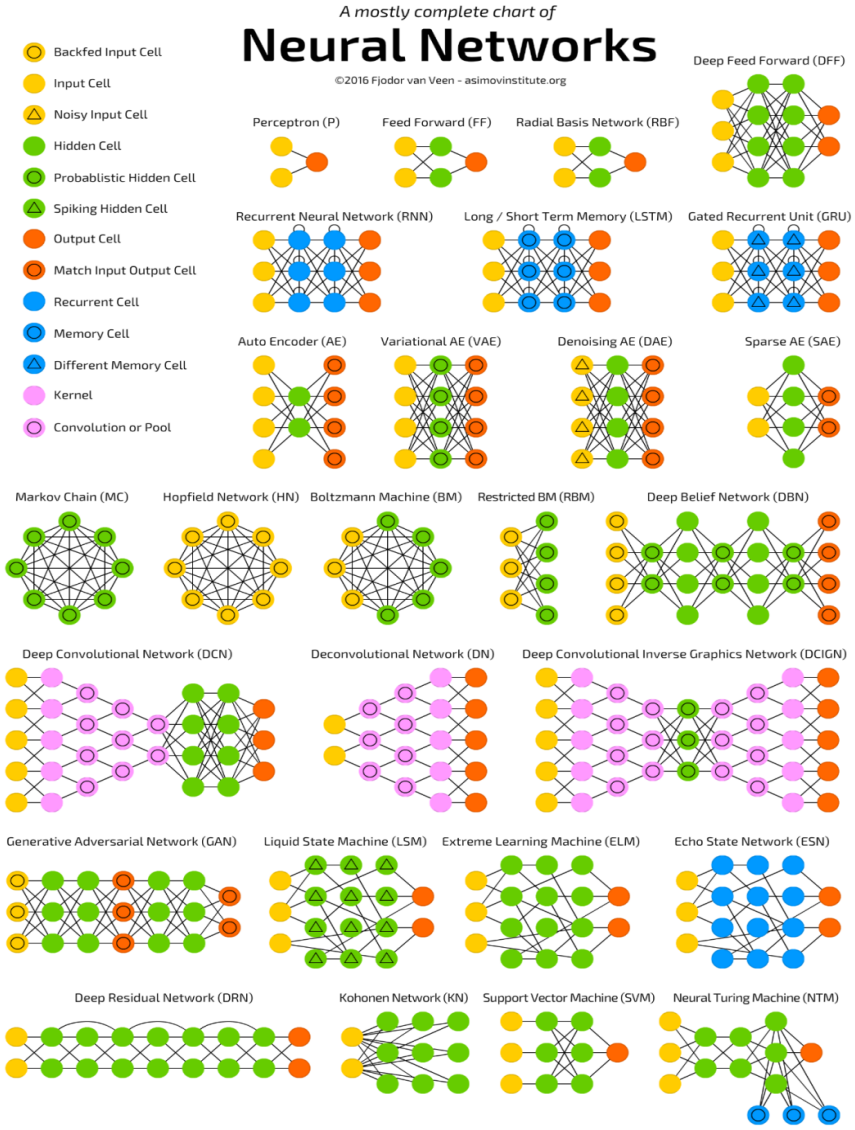
In 2020

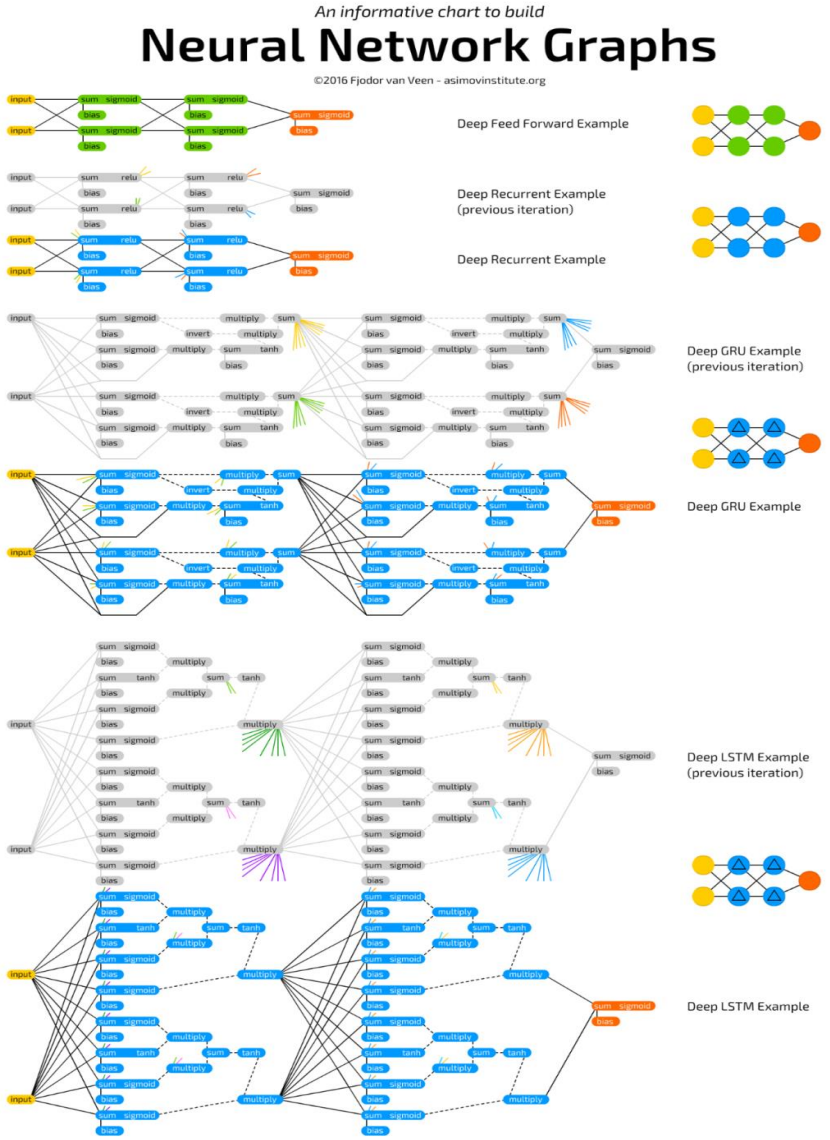
- CNNs for grids, Transformers for language
- “Which architecture wins?” was the question

By 2026

- ViT and CNN backbones are mixed
- Sparse experts make giant models usable
- Transformers still dominate language

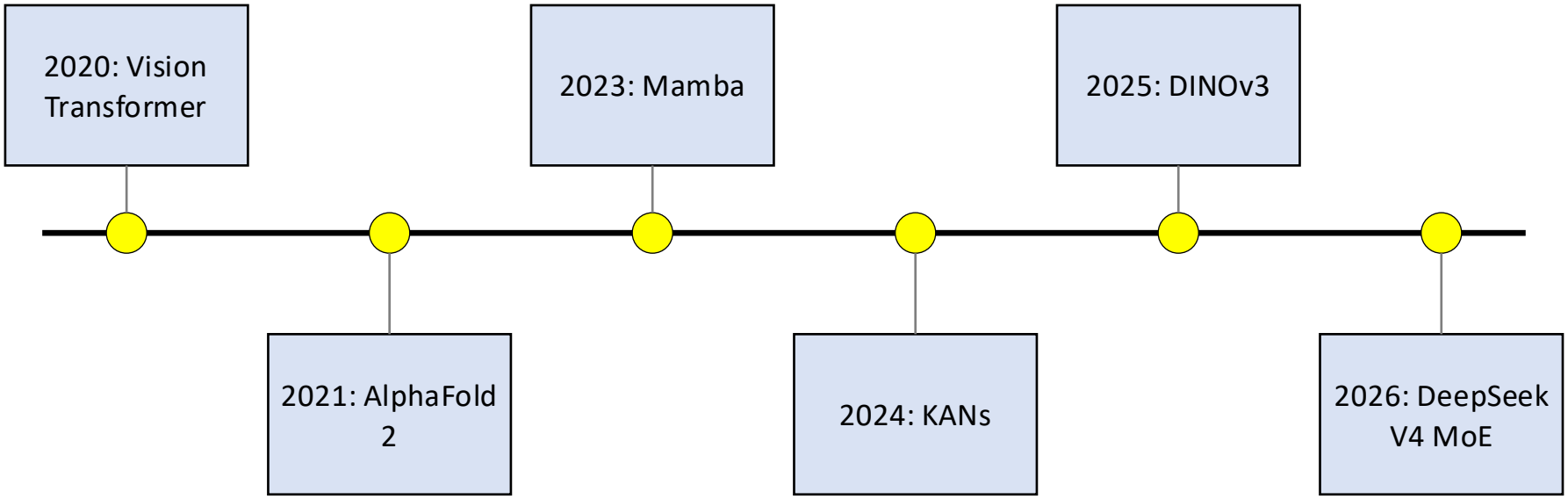
https://www.asimovinstitute.org/neural-network-zoo/





credit: Asimov Institute, Neural Network Zoo; Y. Qi, CS 4774 survey deck (2020)

2020–2026: a new architecture timeline



source: original papers, see References

CNN vs ViT stopped being a contest

- ViT itself: covered earlier
- DINOv3 (2025): one large ViT teacher, no labels
- The teacher is distilled into smaller students
  - Students include both ViT and ConvNeXt backbones
- A convolutional network can learn from a Transformer
- Practical pattern: a frozen, general-purpose backbone

Train once, reuse many times: the pretraining pattern came to vision

source: Siméoni et al. 2025; Meta model cards

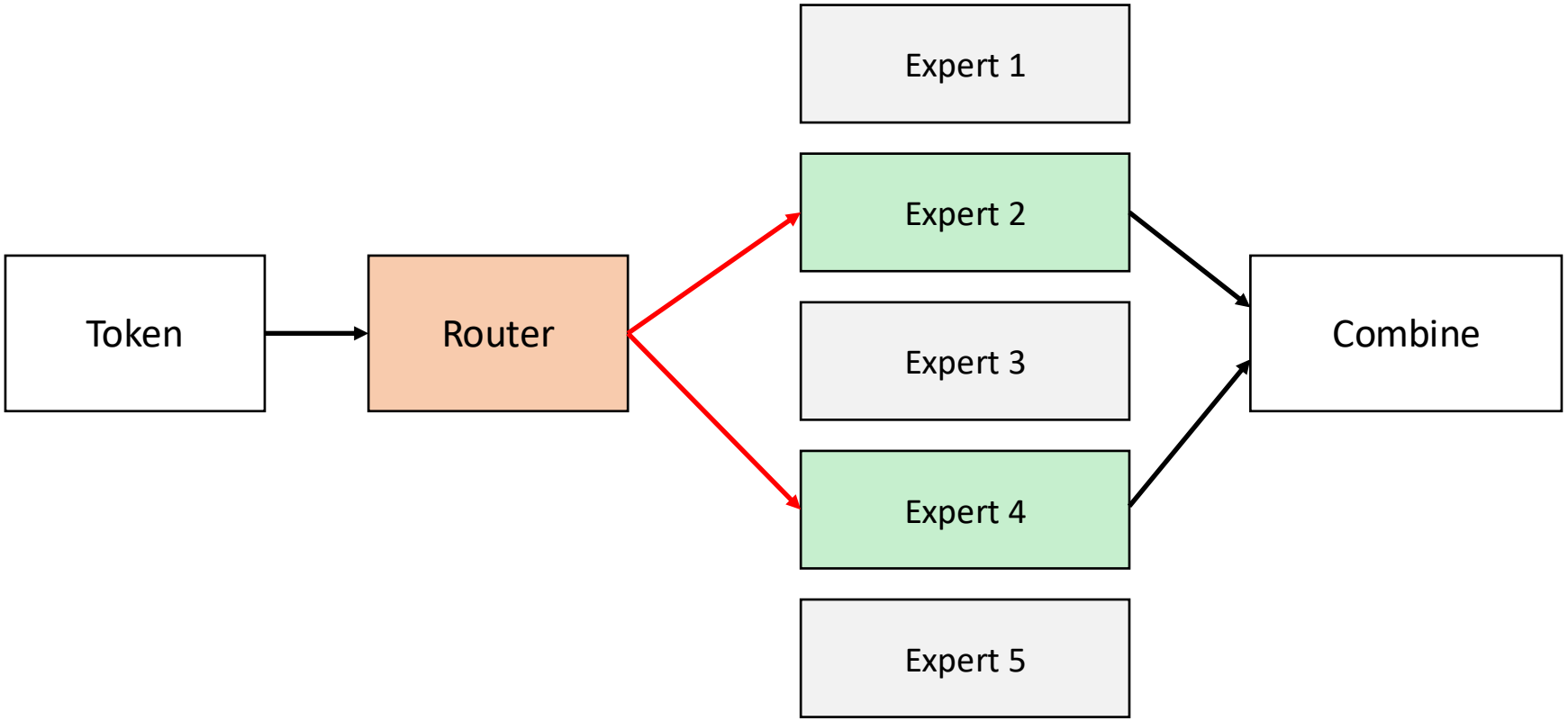
SAM 3: segmentation by text prompt

- SAM and SAM 2: one object per click or box
- SAM 3 (Nov 2025): a short noun phrase or example image
  - “yellow school bus” → find, segment, track every instance
- A detector and a tracker share a vision encoder
- Limit: no complex queries or referring expressions

Vision models are starting to take prompts

source: Meta, SAM 3 (arXiv 2511.16719)

Mixture-of-experts: the idea



Memory follows total parameters. Compute per token follows active parameters

source: Shazeer et al. 2017

Mixture-of-experts: one real example

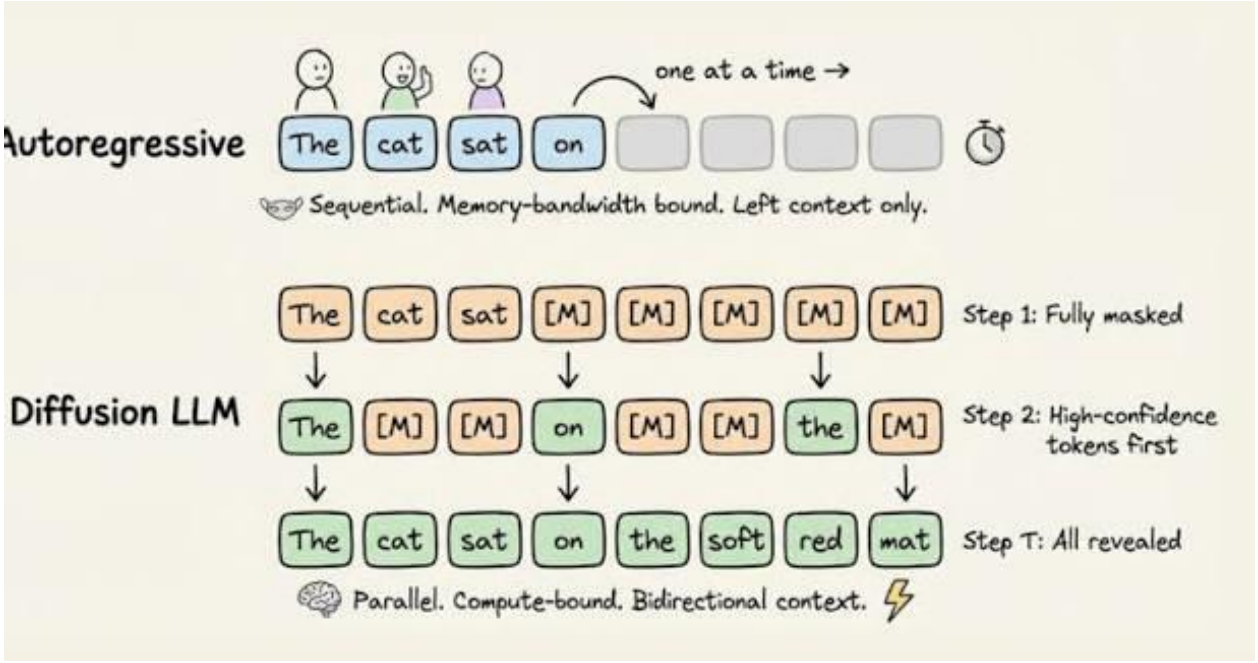
Model	Total	Active per token	Active share
DeepSeek-V4-Pro	1.6 T	49 B	3.1%
DeepSeek-V4-Flash	284 B	13 B	4.6%

- Benefit: a large store of knowledge at a smaller compute cost
- Cost: all parameters still need memory
- Model sizes change fast: read this as a snapshot

source: DeepSeek-V4 model cards on Hugging Face, as of Oct 2026

Challengers: Mamba and diffusion LLMs

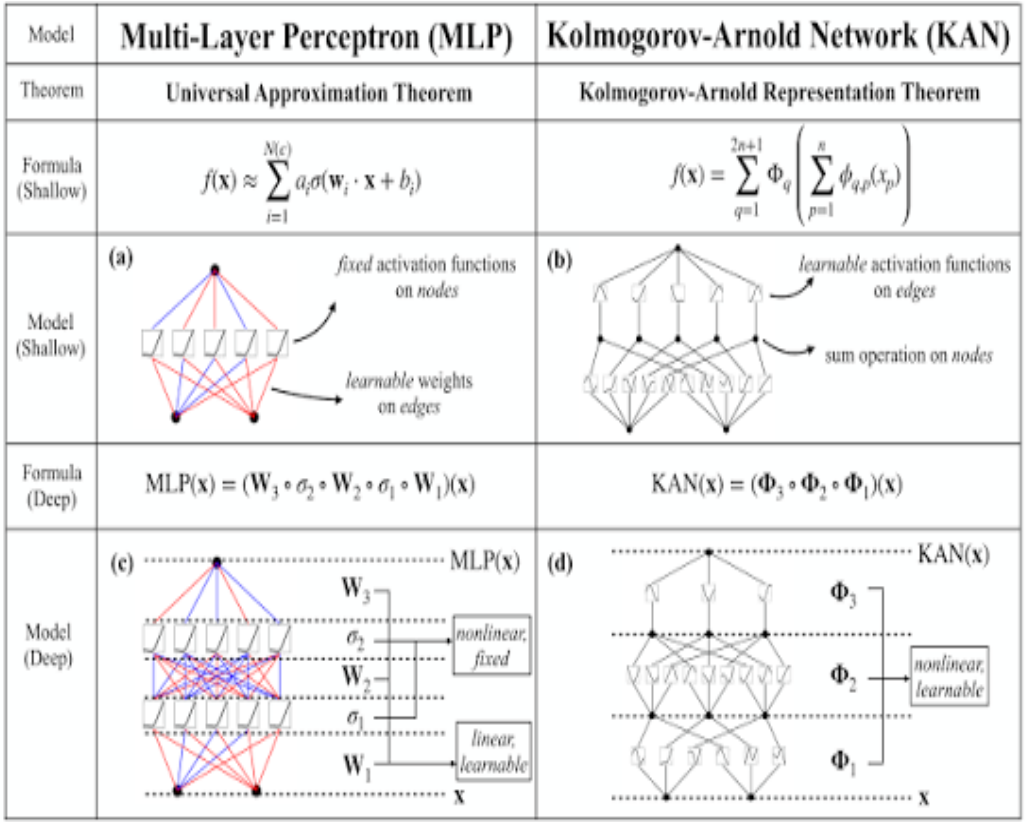
- Attention costs grow quadratically with sequence length
- Mamba (2023): selective state spaces, linear in length
  - Paper: 3B Mamba beats same-size Transformers
- Diffusion language models: generate tokens in parallel, not one by one
  - ICML 2026 award paper: free token order can hurt reasoning



source: Gu and Dao 2023; Ni et al., ICML 2026

KANs, and why the Transformer lasted

- KANs (2024): -- KAN: Kolmogorov-Arnold Networks
  - learnable functions on edges, not fixed activations
- Fair test: match parameters and FLOPs
  - MLPs generally did better, outside symbolic formulas
  - A 2026 review: results conflict; no simple verdict
- So why did Transformers last? Discuss: 2 minutes



Reading habit: when a paper says “better”, ask what was held equal

source: Yu et al. 2024 (arXiv 2407.16674); arXiv 2510.25781

Quiz examples: Module I

1. In the worked example, swap the labels C and D. What happens to the new values of A and B, and why?

2. A model has 1.6 T total parameters and 49 B active per token. What share runs per token?

3. Why did the KAN vs MLP result change once the comparison was made fair?

UVA CS 4774: Machine Learning

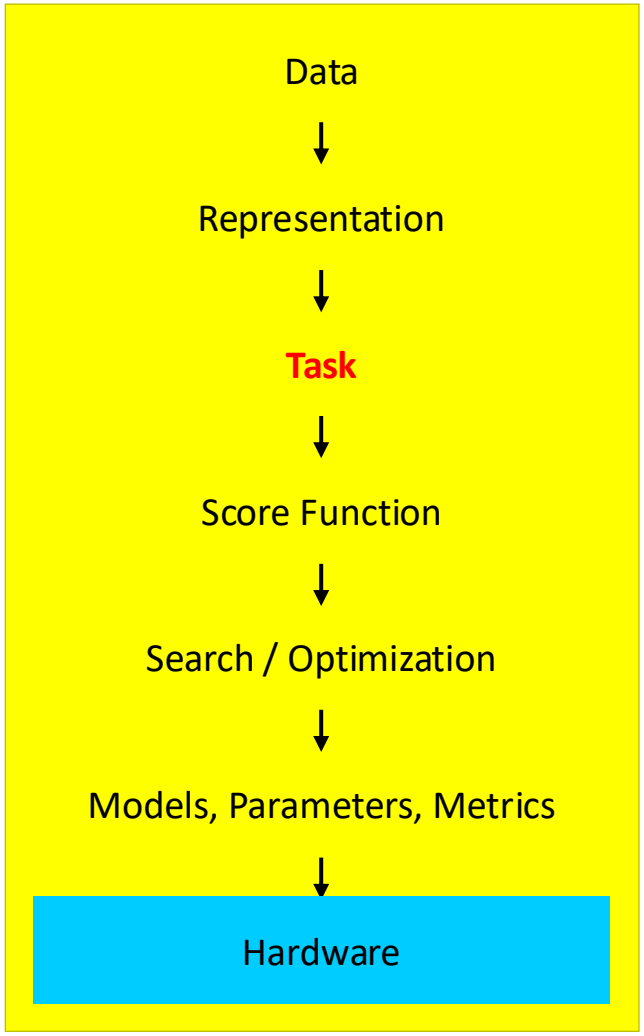
S3: DNN Survey (20-26)
Tasks

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Module II

Nutshell: II.1 Task: Generation



In 2020

- Super-resolution, label-to-image, faces
- Deepfakes meant face swaps

By 2026

- Text to video with sound, text to 3D
- Fakes need marking, detection and law
- Rules are now in force

Why generate? The 2017 task gallery

- Why generative models: planning, missing data, multi-modal outputs
- Image super-resolution (Ledig et al. 2017)
- Label to image and edges to image (Isola et al. 2017)
- Text to image, at low resolution (Reed et al. 2016)
- Faces: Progressive GAN, StyleGAN

source: Goodfellow 2016; papers named on slide; Y. Qi, CS 4774 survey deck (2020)

The 2026 task gallery

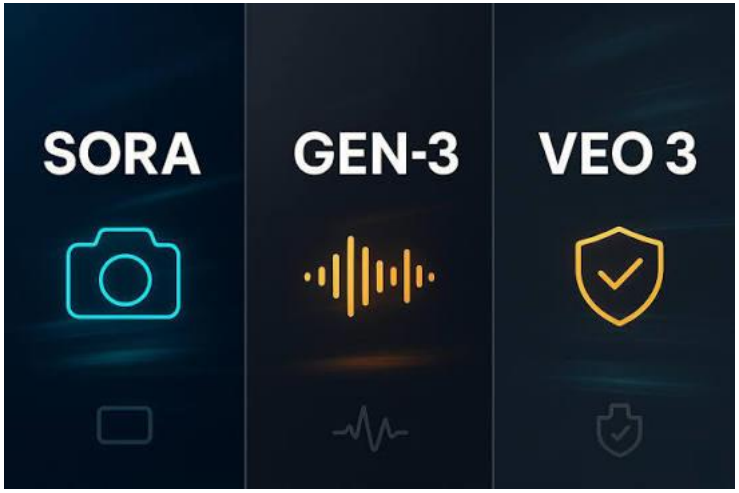
Task	2017	2026
Image	One model per task, paired data	One model, text prompt, edits
Video	Short clips, research demos	Text to video with sound
3D	Hard, slow per scene	Multi-view diffusion + splats
Science	Rare	Molecules, weather ensembles

The general condition, usually text, replaced one model per task

source: Veo, Sora 2, SAM 3 pages; Chen and Wang, 3DGS survey

Video with sound: Veo and Sora 2

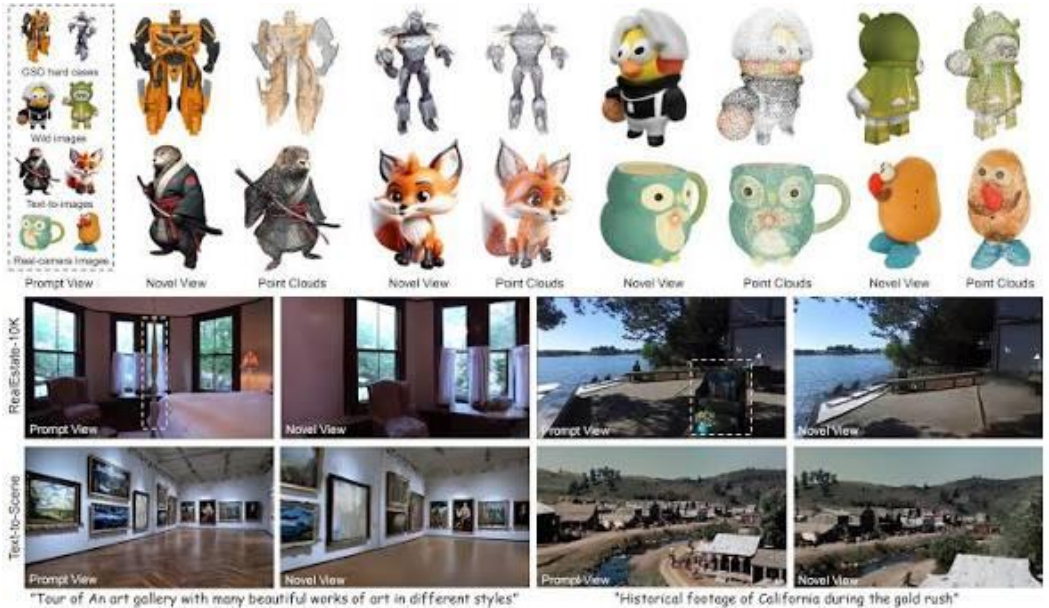
- Veo 3 (Google): sound effects, ambient noise, dialogue
  - Audio generated natively, with the video
  - Google still lists natural speech as active development
- Sora 2 (OpenAI, Sep 2025): video and audio model
- Both ship with provenance marks
- Comparing models is hard: few public benchmarks



source: deepmind.google/models/veo; OpenAI Sora 2 system card

3D: diffusion plus Gaussian splatting

- NeRF (2020): a network stores the scene; slow to render
- 3D Gaussian Splatting: millions of explicit learnable Gaussians
  - Real-time rendering and direct editing
- A common recipe for generation
  - Step 1: diffusion makes several views
  - Step 2: fit a 3D representation to those views



(ICCV), 2025

nd Wang, A Survey on 3D Gaussian Splatting (ACM CSUR)

Deepfakes: then and now

- 2020: face swap, puppet-master, lip-sync; tools such as DeepFaceLab
- 2020 defense: train a detector to spot artifacts
- 2026: whole video, voice and text from a prompt
- Detectors chase each new generator
- So the focus moved: mark content when it is made

source: DeepFaceLab; Black Hat USA 2019 (Price); Y. Qi, CS 4774 survey deck (2020)

Marking generated content

	Watermark (SynthID)	**Signed metadata (C2PA)**
Where	Inside pixels or audio	Attached to the file
Strength	Survives more edits	Cryptographic signature
Weakness	Only if the maker added it	Lost on screenshot or re-save

"No mark found" does not mean "real".
A mark detector only finds content that was marked

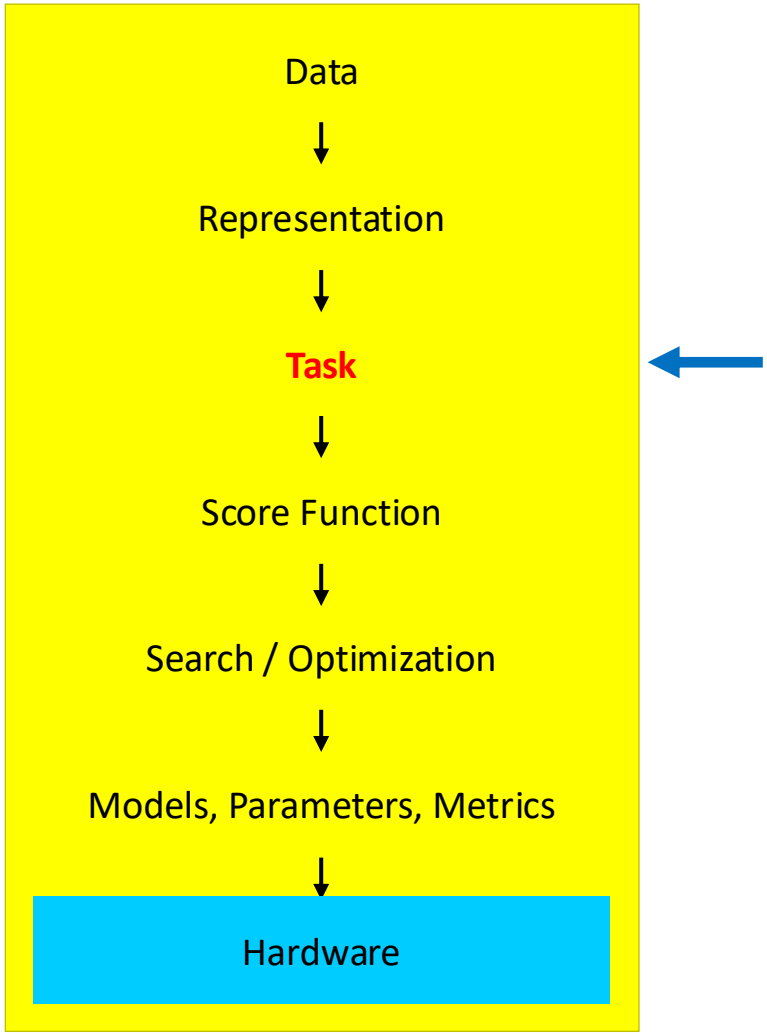
source: Google DeepMind SynthID; C2PA specification

Law catches up

- US TAKE IT DOWN Act (signed May 2025)
  - Platforms must remove reported intimate deepfakes within 48 hours
  - FTC handles reports
- EU AI Act Article 50 (applies from 2 Aug 2026)
  - Providers: mark outputs in a machine-readable way
  - Deployers: disclose deepfakes
  - Systems on the market before 2 Aug 2026: marking due 2 Dec 2026

source: takeitdown.ftc.gov; EU AI Act Art. 50; AI Omnibus, Reg. (EU) 2026/1744; as of 31 Oct 2026

Nutshell: II.2 Task: Decisions



In 2020

- Games: AlphaGo and AlphaZero
- Deep RL needed huge simulation

By 2026

- Learn a model, then plan inside it
- Generated worlds you can act in
- Real robots remain the hard test

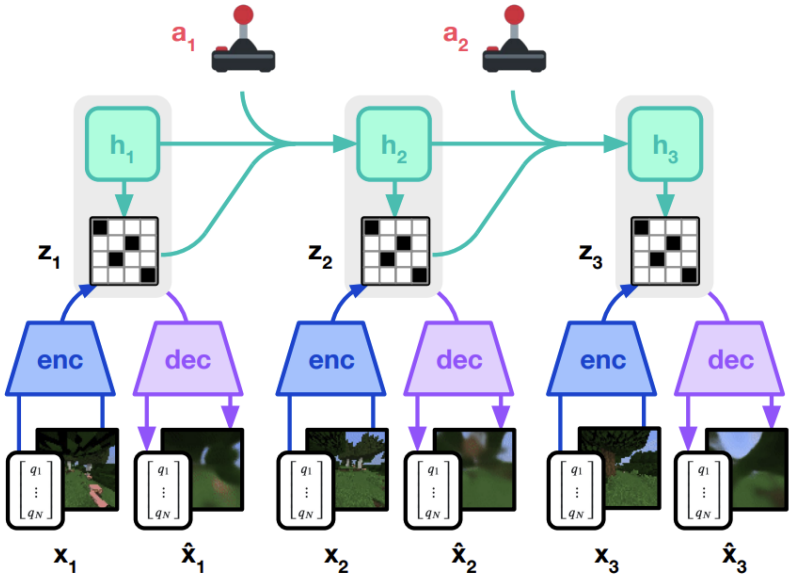
Where our RL lecture left off

- AlphaGo (2016): supervised and RL plus tree search
- AlphaGo Zero (2017): self-play only
- A game gives perfect rules and unlimited practice
- The real world gives neither
- ICML 2026 Test of Time: A3C (Mnih et al. 2016)
- Next step: learn the rules

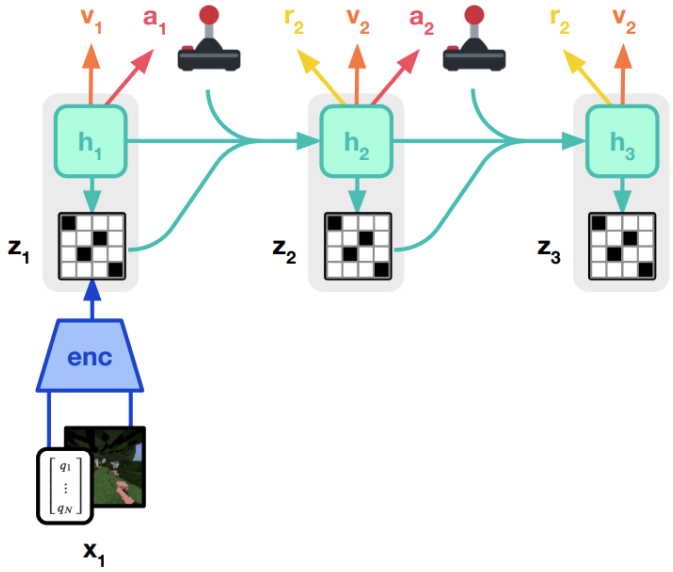
source: Silver et al. 2016, 2017; ICML 2026 awards

Learning inside a model: DreamerV3

- Learn a world model from experience
- Train the policy inside imagined rollouts: seen earlier
- New result: one configuration across 150+ tasks
- First to collect Minecraft diamonds from scratch, with no human data
- Published in Nature, April 2025



(a) World Model Learning



(b) Actor Critic Learning

, Nature 2025

Interactive generated worlds: Genie 3

- From a text prompt to a world you can move through
- 720p at 24 fps, consistent for a few minutes
- Limits stated by DeepMind:
  - Few agent actions; many-agent interaction is still hard
  - Minutes of interaction, not hours
  - Limited research preview
- Use: a training ground for agents

source: Google DeepMind, Genie 3 (Aug 2025)

Robot learning and the sim-to-real gap

- Train in simulation: cheap, safe, fast
- Deploy on a robot: friction, lighting, sensors all differ
- Tools: domain randomization, fine-tuning on real data
- V-JEPA 2-AC: under 62 hours of robot video
  - Zero-shot pick-and-place on Franka arms in two labs
- This is distribution shift, in physical form

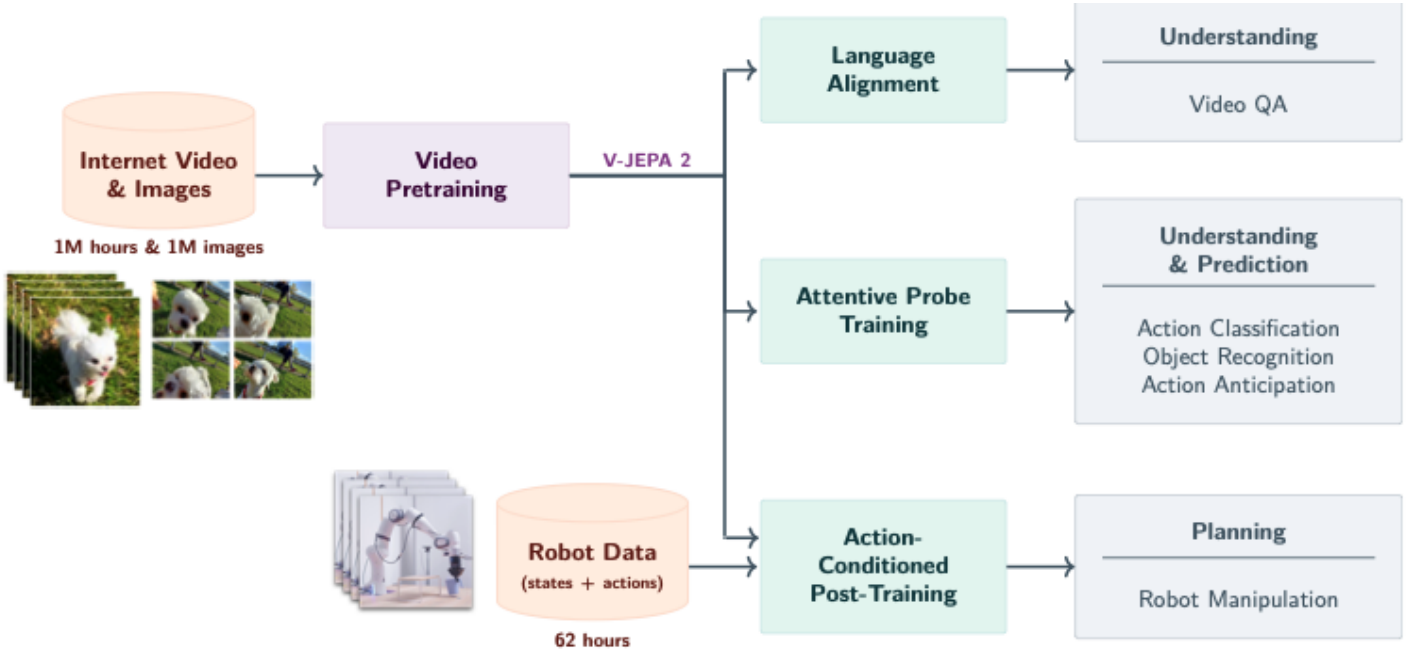
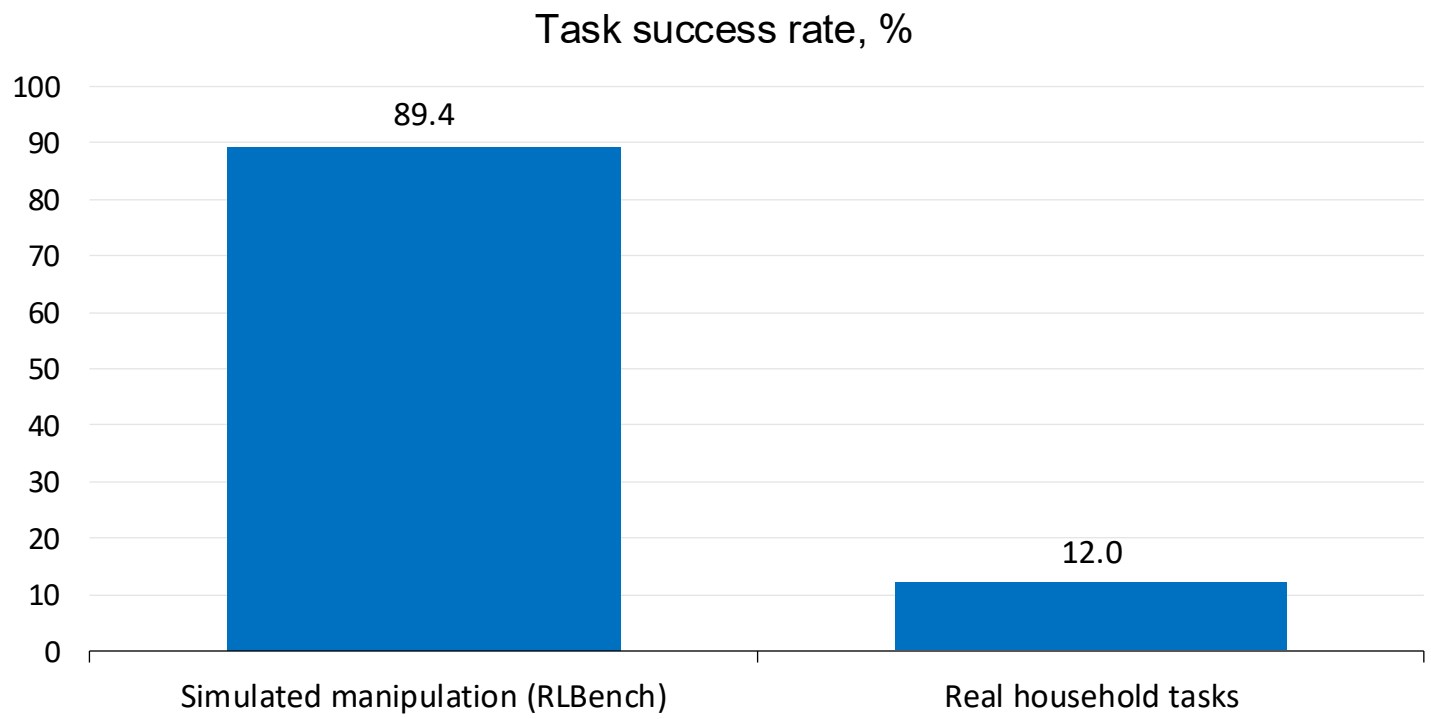


Figure 1 V-JEPA 2 Overview. Leveraging 1M hours of internet-scale video and 1M images, we pretrain the V-JEPA 2 …tics (arXiv 2510.20808)

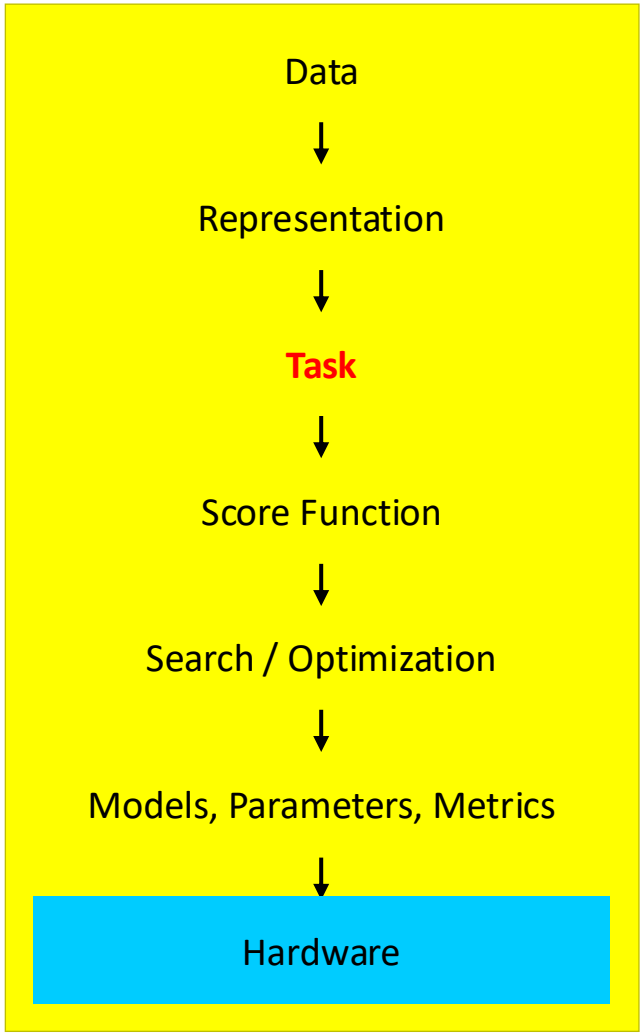
Reality check: simulated vs real robots



Different benchmarks, not a controlled comparison. The size of the gap is the message

source: Stanford AI Index 2026, Technical Performance

Nutshell: II.3 Task: Reasoning



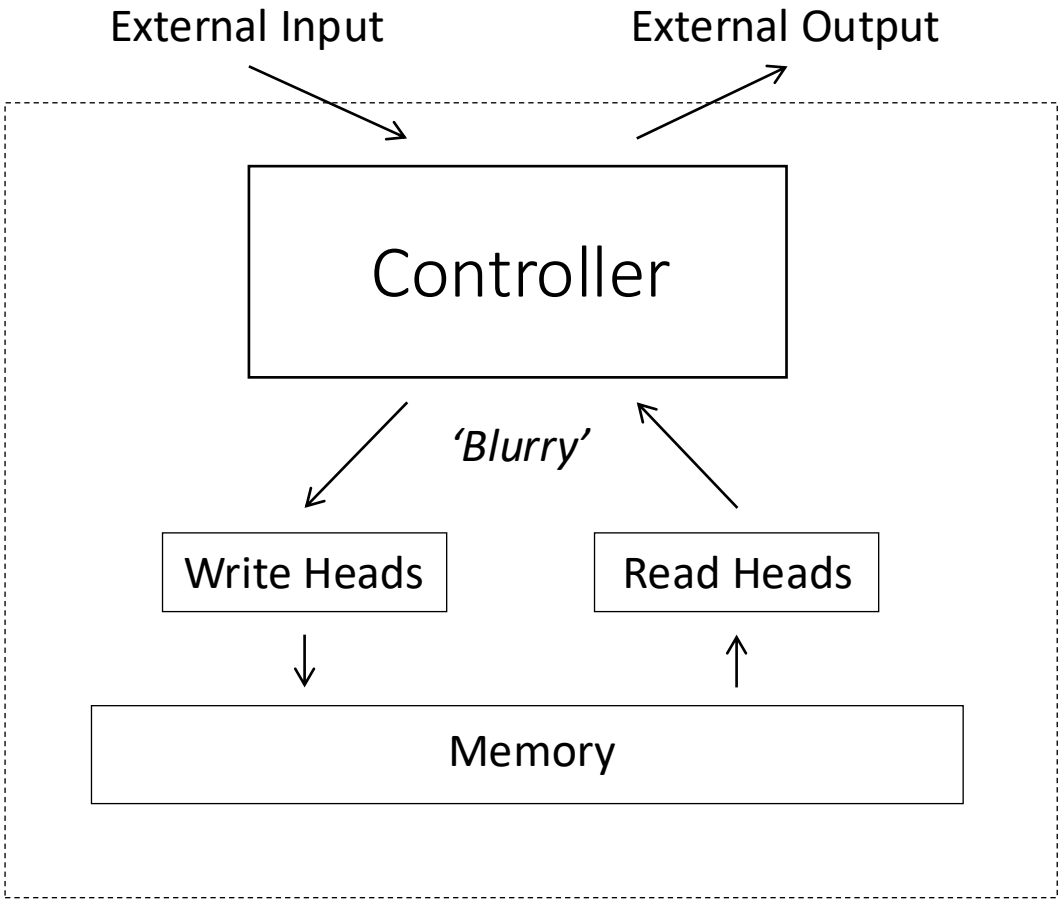
In 2020

- Neural Turing Machines
- Program induction on toy tasks

By 2026

- Formal proofs with RL
- Gold-medal level at the IMO
- Natural language, end to end

Neural Turing Machines (2014)



credit: Graves et al. 2014; Y. Qi, CS 4774 survey deck (2020)

Symbolic tasks: 2020 vs 2026

	2020	**2026**
Architecture	Special: differentiable memory	General: large Transformer
Signal	Input and output pairs	Verifiable reward
Symbols	Inside the network	Outside: Lean, tests, tools
Scale	Toy tasks	Olympiad problems

Symbols came back as a checker outside the network

source: Y. Qi, CS 4774 survey deck (2020); Hubert et al. 2025; DeepSeek-R1 (Nature 2025)

AlphaProof: RL writing Lean proofs

- AlphaZero-style RL, in the Lean proof language
- Trained on about 80 million auto-formalized statements
- A proof either checks or fails: a perfect reward
- IMO 2024: 3 of 5 non-geometry problems
- With AlphaGeometry 2: 28 of 42 points, silver level
- Published in Nature, November 2025

source: Hubert et al., Nature 2025; Nature Asia press release

IMO 2025: formal and natural-language

	Natural language	**Formal (Lean)**
Example	Gemini Deep Think: 35 of 42	Aristotle: 5 of 6 problems
Output	A written proof	A machine-checked proof
Who checks	Human graders	The proof checker
Trade-off	Flexible, harder to verify	Verifiable, harder to produce

Discuss: which output would you trust more, and for what?

source: Google DeepMind IMO 2025 post; Aristotle (arXiv 2510.01346); arXiv 2609.10712

Learning from few examples

- 2020: meta-learning, or learning to learn
  - MAML: find an initialization that adapts in a few steps
- 2026: a pretrained LLM adapts from a few examples in the prompt
  - few-shot prompting
  - much of the goal moved here
- MAML-en-LLM (2024) applies meta-training to LLMs

source: Finn et al. 2017; Sinha et al., KDD 2024

Quiz examples: Module II

1. A detector finds no watermark in a viral video. Can you conclude that the video is real? Why?

2. What did games offer RL that the real world does not? Why does Lean offer the same thing?

Roadmap:

So far: Data, Representation, Task

Next: Score Function, Search, Metrics, Hardware

UVA CS 4774: Machine Learning

S3: DNN Survey (20-26)

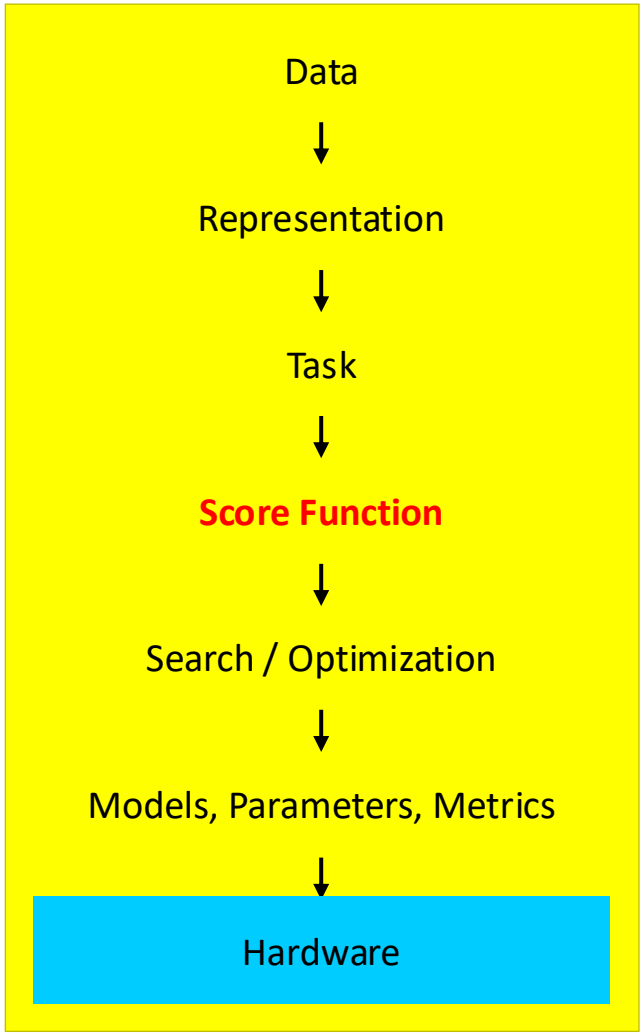
Training: Losses and Search

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Module III

Nutshell: III.1 New Losses



In 2020

- Adversarial loss: GANs
- Reconstruction: autoencoders
- Contrastive learning just arriving

By 2026

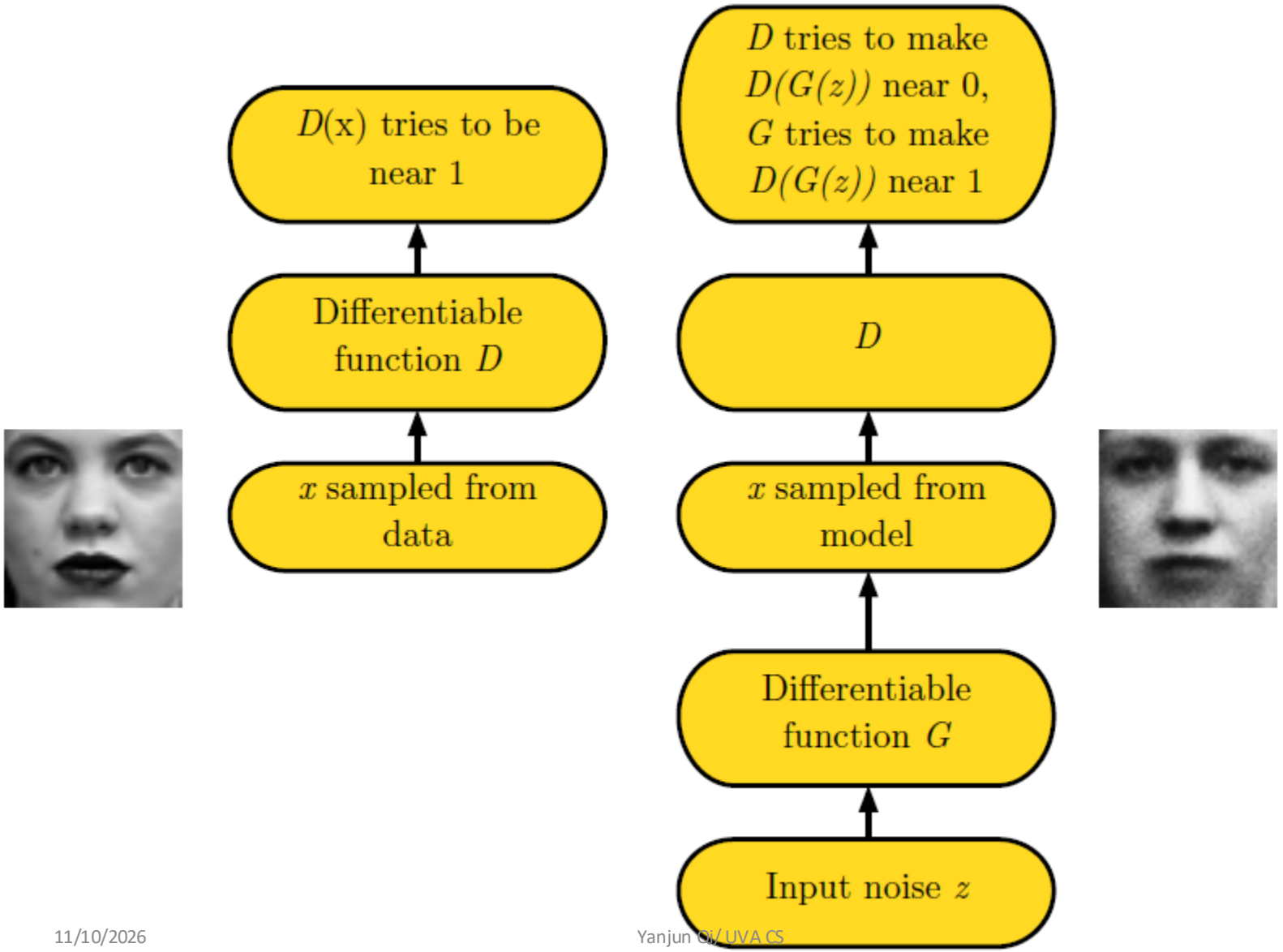
- Denoising, then flow matching
- One-step generators
- Joint-embedding prediction

One table: what each loss asks

Family	The network is asked to…	Where
GAN	Fool a critic	This part
Diffusion	Predict the added noise	Multimodal
Flow matching	Predict a velocity, noise to data	This part
Contrastive	Match two views of one image	This part
Masked AE	Fill in hidden patches	This part
JEPA	Predict hidden content as an embedding	This part

Each row makes a supervised problem out of unlabeled data

Adversarial Nets Framework (2016)

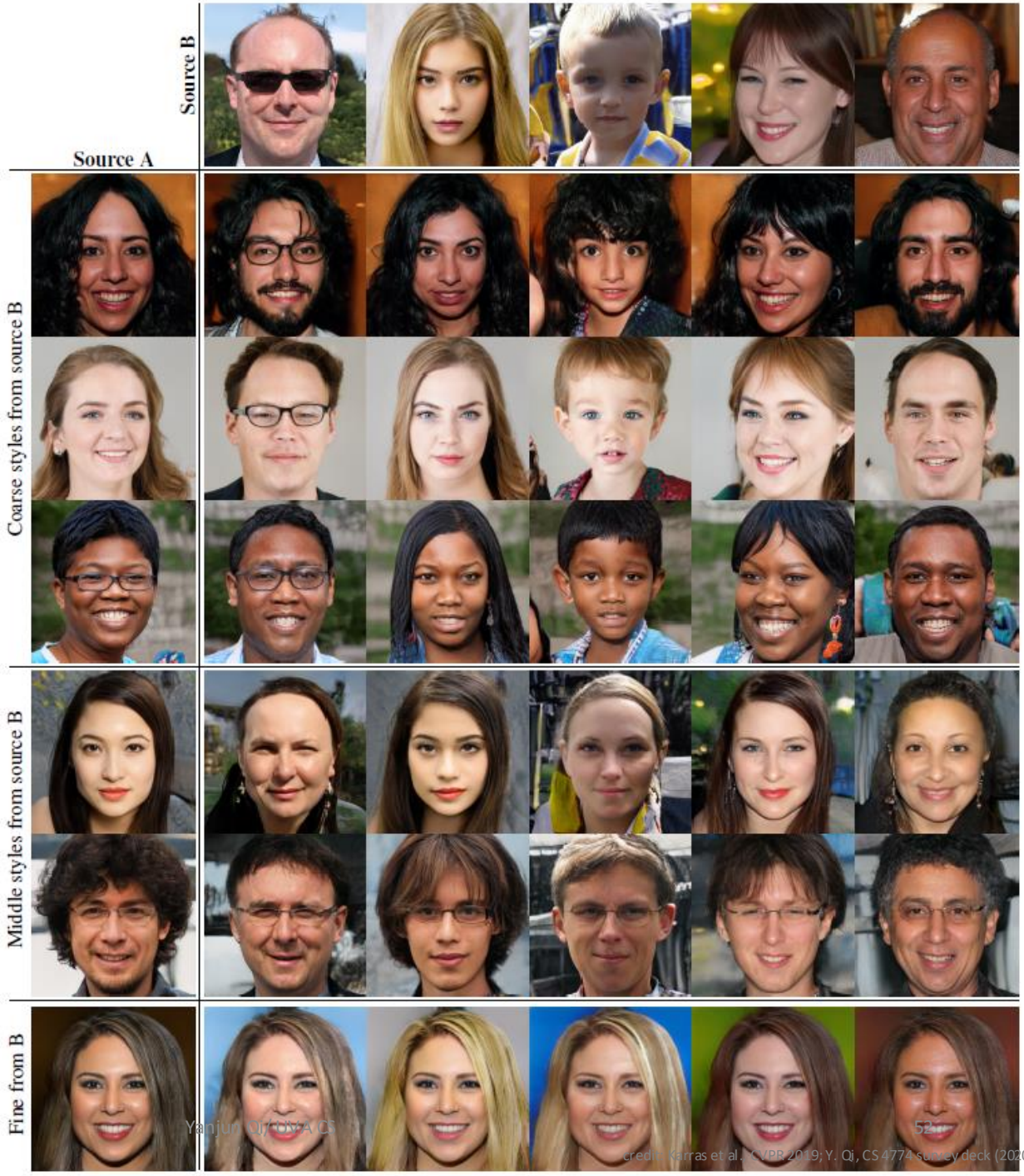


(Goodfellow 2016)

credit: Goodfellow, NIPS 2016 tutorial; Y. Qi, CS 4774 survey deck (2020)

StyleGAN

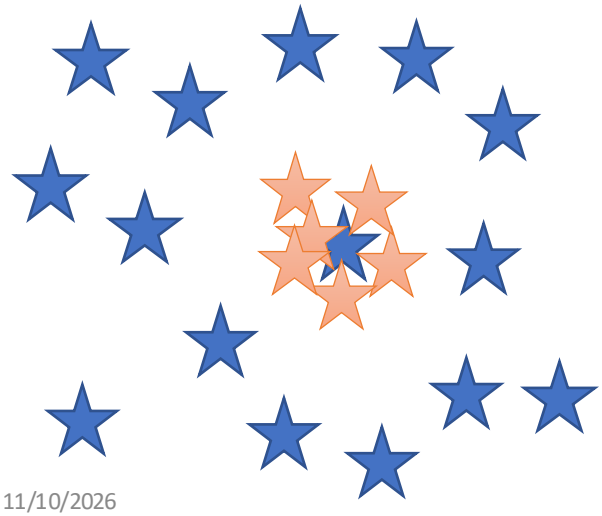
- propose a new generator architecture for GAN which can generate high quality images (called StyleGAN)
- leads to an automatically learned, unsupervised separation of high-level attributes (latent space disentangled)
- propose two new automated metrics for quantifying disentanglement
- propose a high quality dataset of human faces (FFHQ)



Credit: Karras et al., CVPR 2019; Y. Qi, CS 4774 survey deck (2020)

R3GAN: is the GAN dead?

- "The GAN is dead; long live the GAN!" (NeurIPS 2024)
- Claim: the tricks were not fundamental
- A better-behaved regularized relativistic loss removes them
- Then swap in modern ConvNet backbones
- Reports results beyond StyleGAN2 on FFHQ, ImageNet, CIFAR
- Lesson: a field's reputation is not a theorem



: real data

: generated data

Diversity - Mode Collapse

source: Huang et al., NeurIPS 2024

credit: Prof. Hung-yi Lee, NTU ML 2021

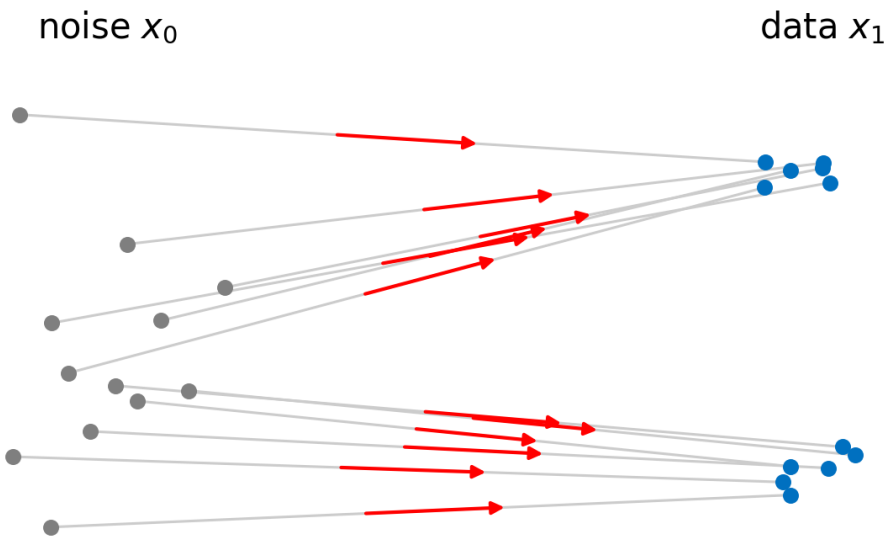
Diffusion recap, then the next question

$$\mathcal{L}_{\text{diff}} = \mathbb{E} \left|\left| \epsilon - \epsilon_\theta(x_t, t) \right|\right|^2$$

- Covered earlier: DDPM and latent diffusion
- Win over GANs: plain regression, one fixed loss
- Cost: sampling needs many network calls

Can we keep stable training and shorten the path from noise to data?

Flow matching: learn a velocity field



- Sample noise
- Follow the velocity
- Arrive at data
- Diffusion is one special path
- Straight paths need fewer steps

Flow matching is a training objective, not a network architecture

source: Lipman et al. 2022; Holderrieth and Erives 2026 (figure redrawn)

Flow matching: the loss, simply

$$x_t = (1-t)\,x_0 + t\,x_1$$

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,x_0,x_1}\left\| v_\theta(x_t,t) - (x_1 - x_0)\right\|^2$$

x_0 (noise)	x_1 (data)	t	x_t	Target	Prediction	Loss
2	6	0.25	3	4	3	1

Same recipe as linear regression: a target, a prediction, mean squared error

source: Lipman et al. 2022; Liu et al. 2022

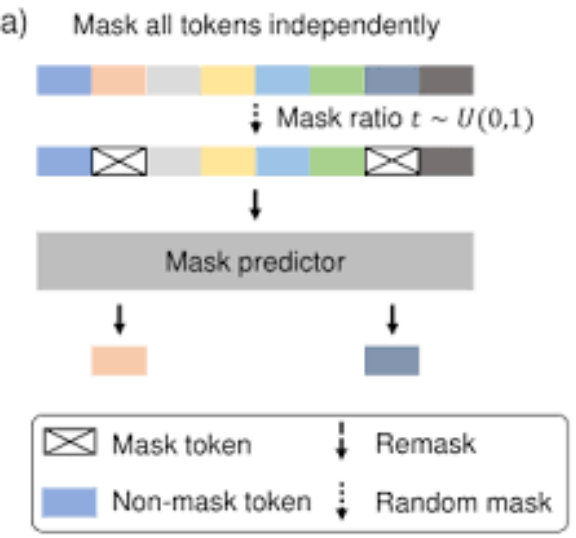
Fewer steps: from many to one

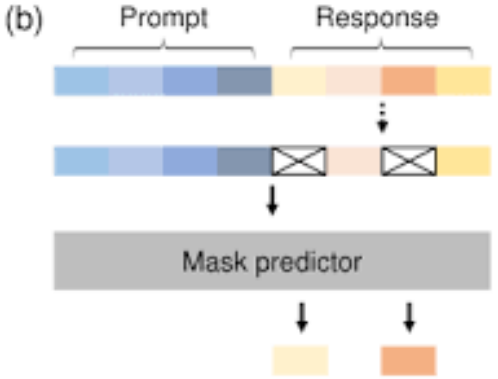
- Rectified flow (2022): recursively straighten the paths
  - Straight paths can be simulated with coarse steps
- Consistency models (2023): map noise directly to data
  - One-step CIFAR-10 FID of 3.55; more steps trade compute for quality
- Distillation: teach a fast student from a slow diffusion teacher
- Same pressure as Module I: pay once, then reuse

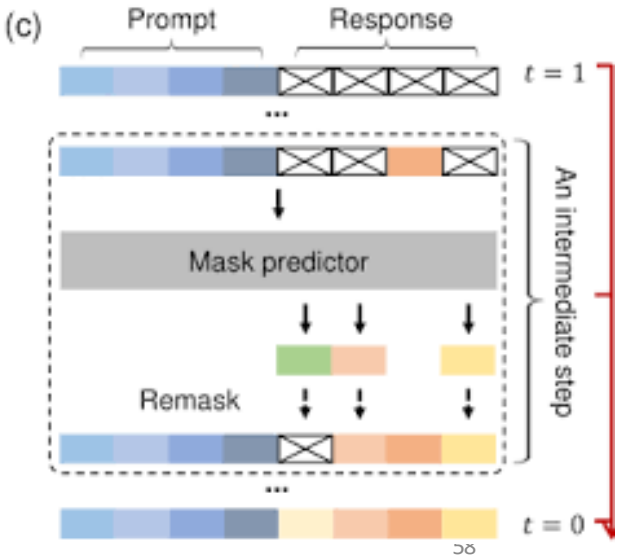
source: Liu et al. 2022; Song et al. 2023; arXiv 2502.08364

ICML 2026: diffusion took both top papers

- Two Outstanding Papers, both about diffusion
- Chen, Chewi, Daskalakis, Rakhlin
  - Sampling to a given error in polylogarithmic steps
- Ni et al., “The Flexibility Trap”
  - Diffusion for language: arbitrary token order can hurt reasoning
- Theory and application are still moving fast







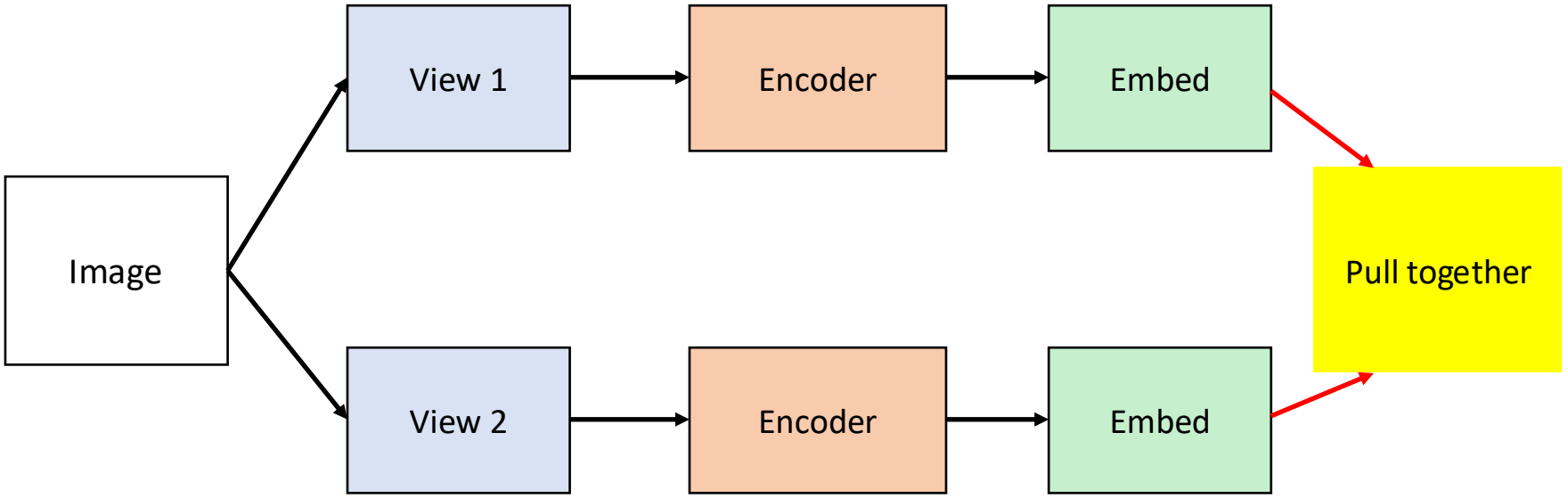
source: ICML 2026 awards, via AIhub

History: layer-wise pretraining

- 2006–2012: train one layer at a time, using reconstruction
- Then fine-tune the whole network with labels
- Why: deep networks were hard to optimize
- Replaced by better activations, residuals, normalization
- What lasted: pretrain without labels, then fine-tune

source: Hinton et al. 2006; Y. Qi, CS 4774 survey deck (2020)

Contrastive learning: SimCLR (2020)



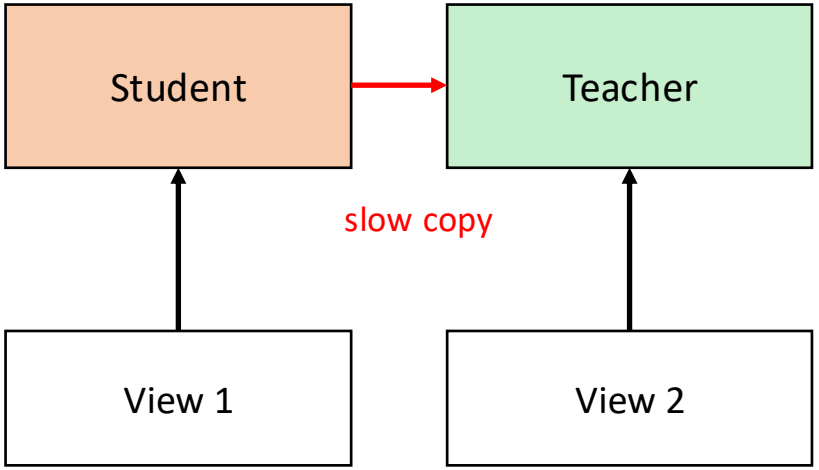
- Other images in the batch are pushed away
- Linear classifier: 76.5% top-1 on ImageNet

source: Chen et al., ICML 2020

Masked autoencoders and DINO (2022)

MAE: hide 75%, rebuild pixels

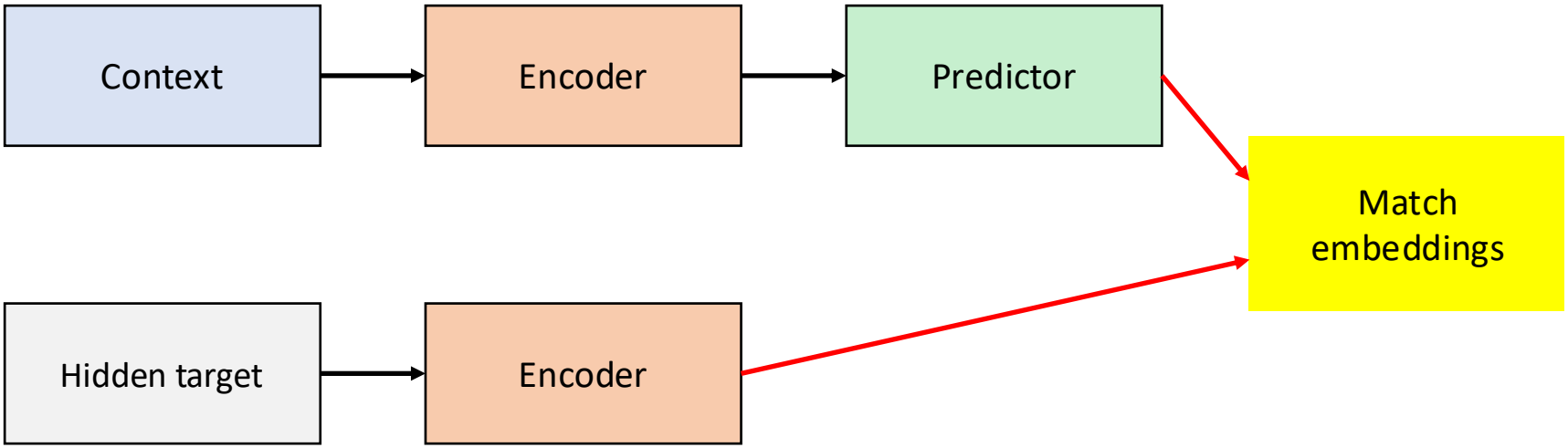
DINO: match a momentum teacher



MAE needs a decoder and pixels. DINO needs neither. Both train with no labels

source: He et al., CVPR 2022; Caron et al., ICCV 2021

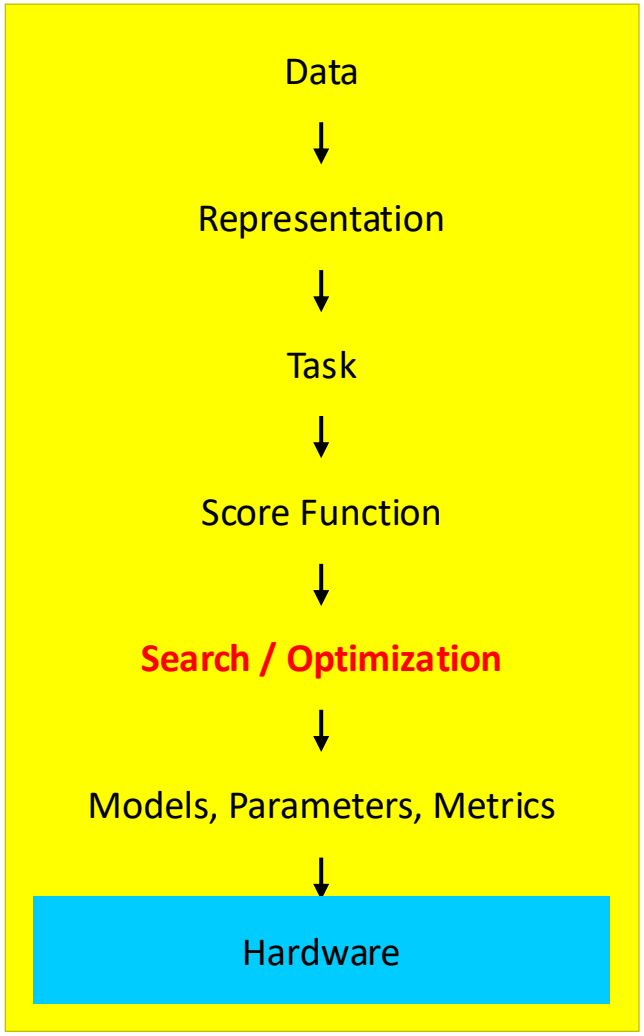
JEPA: predict in latent space (2025)



- No pixel decoder: unpredictable detail can be ignored
- Risk: collapse to a constant; regularizers prevent it
- V-JEPA 2 trained on over 1 million hours of video

source: Assran et al. 2023, 2025; Balestriero and LeCun 2025

Nutshell: III.2 Search



In 2020

- NAS: search a defined space
- RL or evolution over architectures
- Optimizers searched too

By 2026

- Compute and data drive much progress
- Agents edit code, not cells in a grid
- Peer review is a new test

Neural Architecture Search (ICLR 2017)

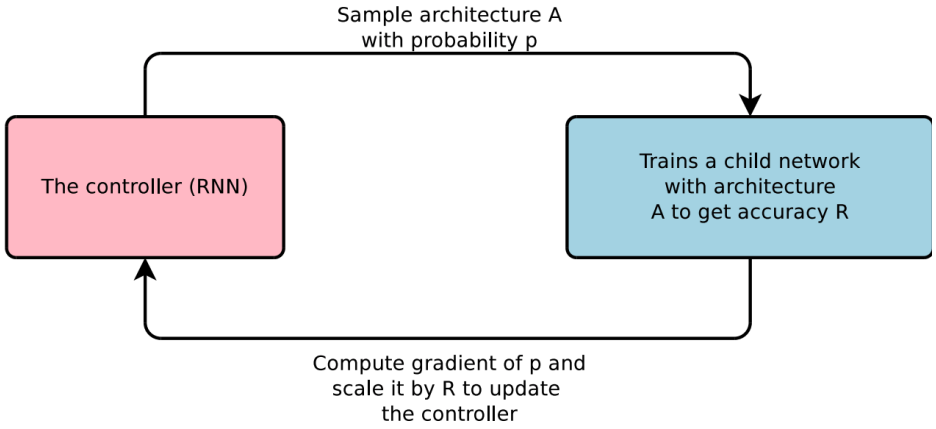
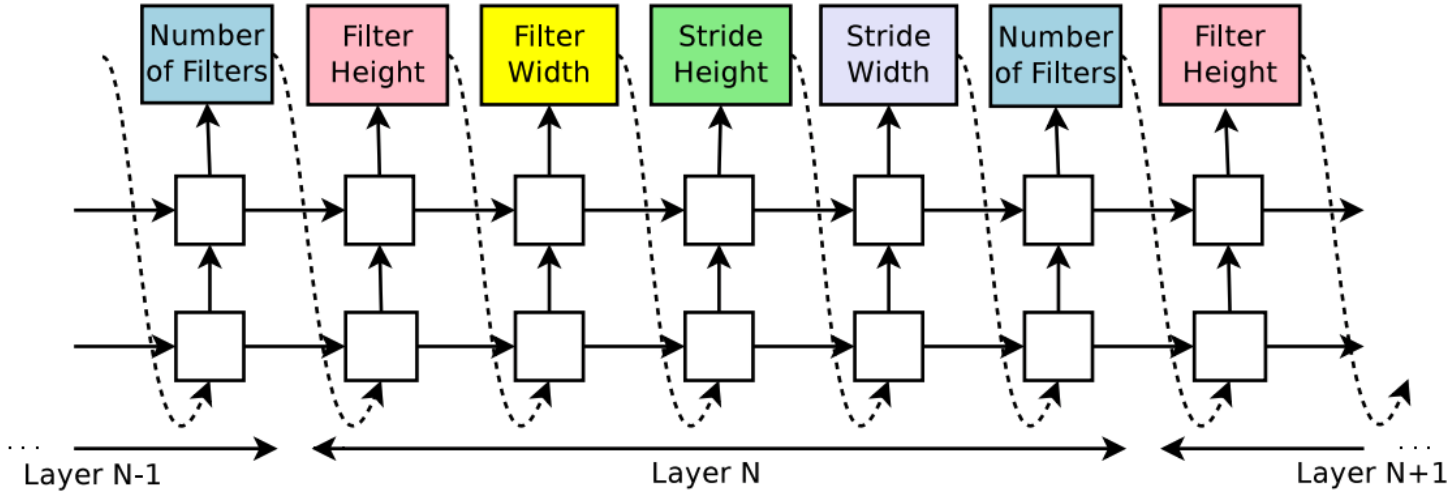


Figure 1: An overview of Neural Architecture Search.



credit: Zoph and Le, ICLR 2017; Y. Qi, CS 4774 survey deck (2020)

Neural Optimizer Search with Reinforcement Learning, ICML17

- e.g. hyperpara search

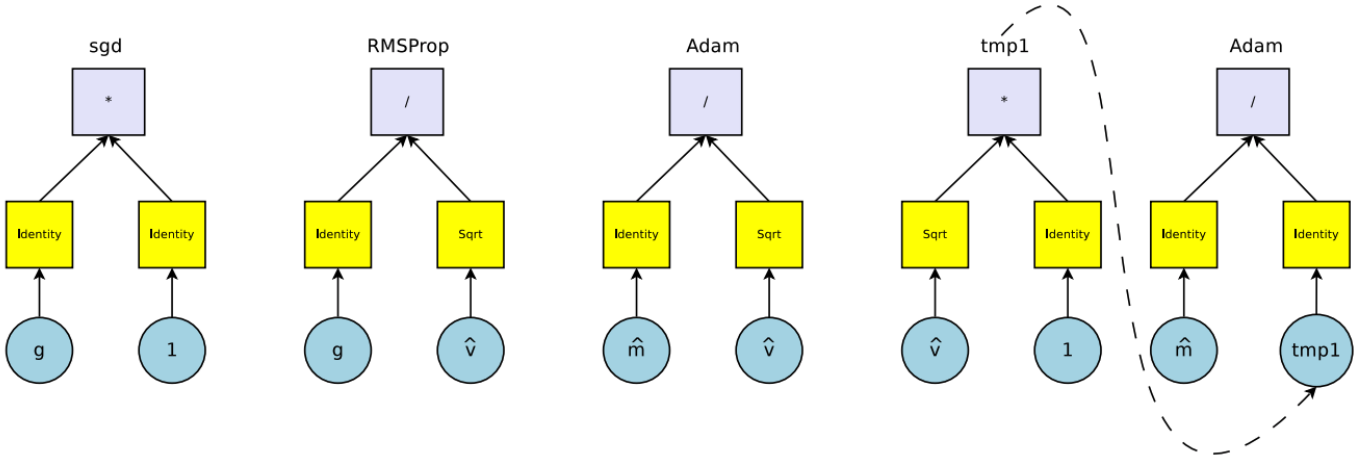


Figure 2. Computation graph of some commonly used optimizers: SGD, RMSProp, Adam. Here, we show the computation of Adam in 1 step and 2 steps. Blue boxes correspond to input primitives or temporary outputs, yellow boxes are unary functions and gray boxes represent binary functions. g is the gradient, $\hat{m}$ is the bias-corrected running estimate of the gradient, and $\hat{v}$ is the bias-corrected running estimate of the squared gradient.

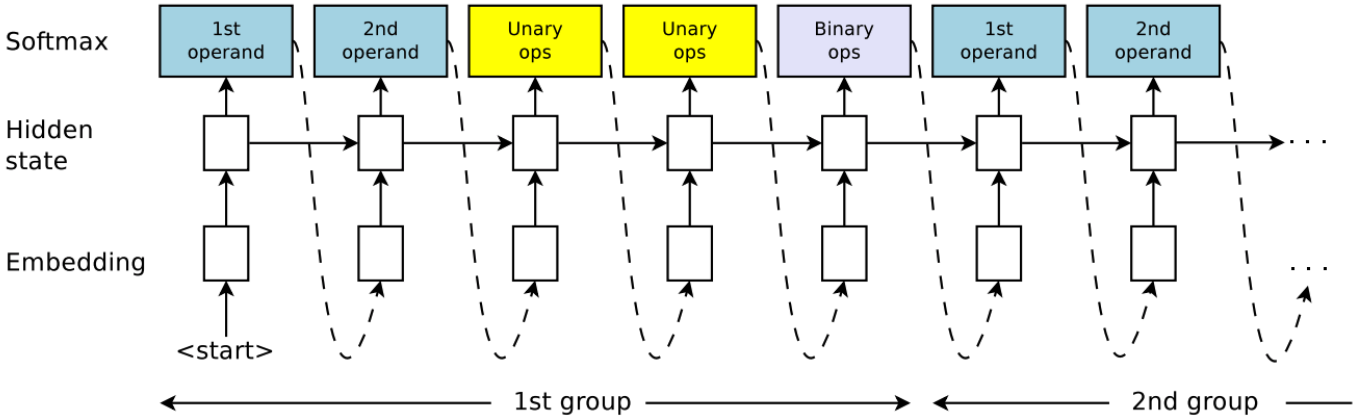
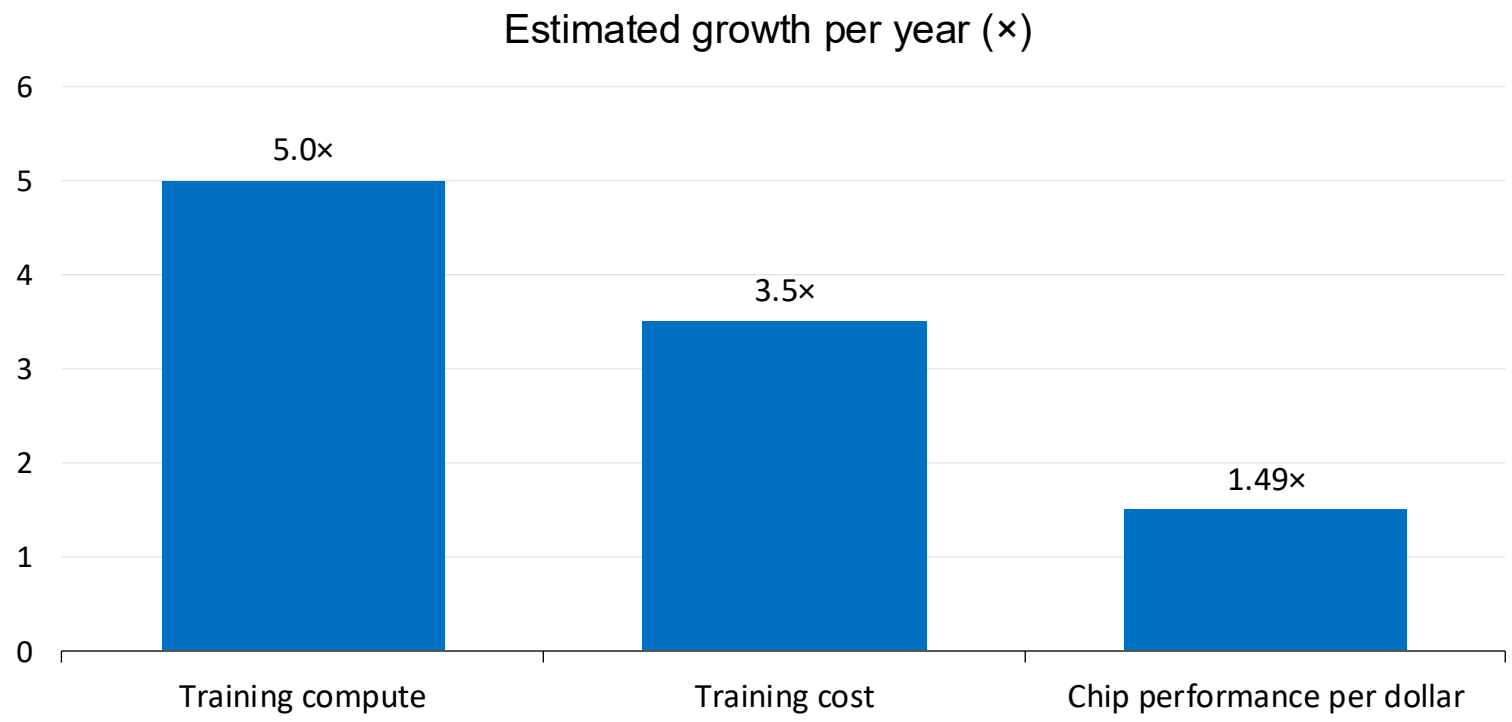


Figure 3. Overview of the controller RNN. The controller iteratively selects subsequences of length 5. It first selects the 1st and 2nd operands op_1 and op_2, then 2 unary functions u_1 and u_2 to apply to the operands and finally a binary function b that combines the outputs of the unary functions. The resulting $b(u_1(op_1), u_2(op_2))$ then becomes an operand that can be selected in the subsequent group of predictions or becomes the update rule. Every prediction is carried out by a softmax classifier and then fed into the next time step as input.

credit: Bello et al., ICML 2017; Y. Qi, CS 4774 survey deck (2020)

Scale: compute and cost trends (2026)



Compute grew much faster than efficiency. Spending filled the gap

source: Epoch AI Trends, as of Oct 2026

Agents that edit their code (ICLR 2026)

- Darwin Gödel Machine (ICLR 2026)
- A coding agent modifies its own code, then tests the change
- Keeps an archive of diverse agents
- SWE-bench: 20.0% → 50.0%
- Polyglot: 14.2% → 30.7%
- The search space is code, not a human-defined grid

source: Zhang et al., ICLR 2026 (arXiv 2505.22954)

The AI Scientist, and its limits (Nature 2026)

- End to end: ideas, code, experiments, a written paper (Nature 2026)
- One paper at an ICLR 2025 workshop scored 6.33, above the human threshold
- It was withdrawn before publication, as planned
- Stated limits: naive ideas, weak rigor, hallucinations
  - Inaccurate citations, duplicated figures
- Computational experiments only

source: Sakana AI, Nature 2026 announcement; Beel et al. (arXiv 2502.14297)

UVA CS 4774: Machine Learning

S3: DNN Survey (20-26)

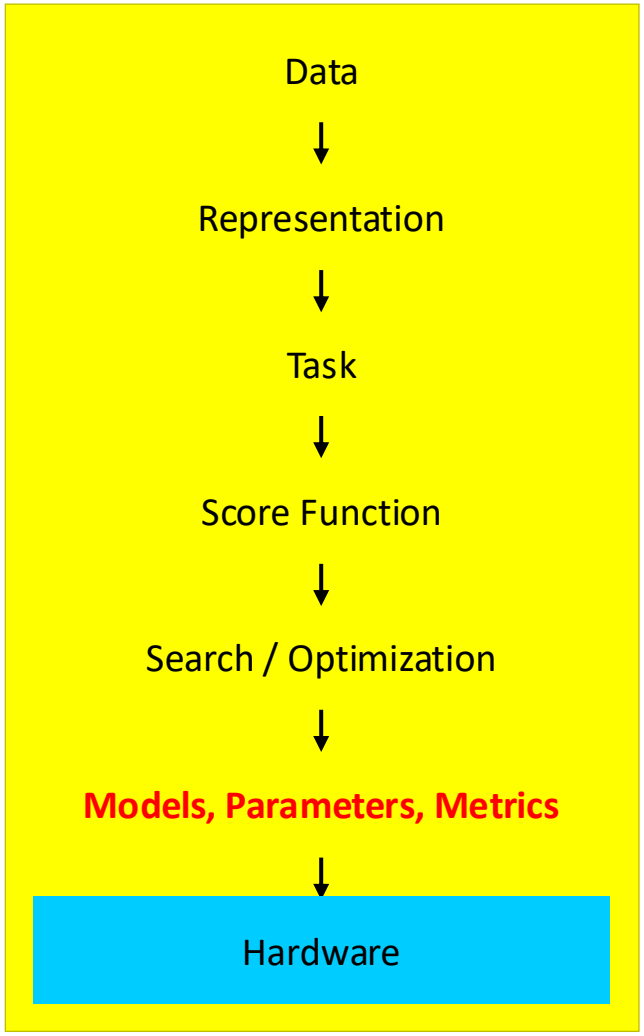
Deployment: Trust and Hardware

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Module IV

Nutshell: IV.1 Trust and Measurement



In 2020

- Saliency and feature visualization
- Adversarial images
- Small-network verification

By 2026

- Tracing circuits inside LLMs
- Verifiers that scale further
- Benchmarks that saturate in months

Interpretation Methods (2020) Overview

- Post-hoc explanations
  - Feature visualization
  - Feature attribution
    - Instance-wise vs. model-wide
  - Feature visualization + attribution
  - Training example attribution
- Inherently interpretable models



credit: Olah et al., Distill; Y. Qi, CS 4774 survey deck (2020)

Mechanistic interpretability: inside an LLM (2025)

- Goal: find which internal parts cause a behavior
- Sparse autoencoders split activations into readable features
- Circuit tracing (Anthropic, 2025): map paths from input to output
  - Studied on Claude 3.5 Haiku
- Test by intervention: suppress or inject a feature
  - “Dallas → Texas → Austin”; swap Texas for California, get Sacramento

source: Anthropic, Circuit Tracing and On the Biology of a Large Language Model, 2025

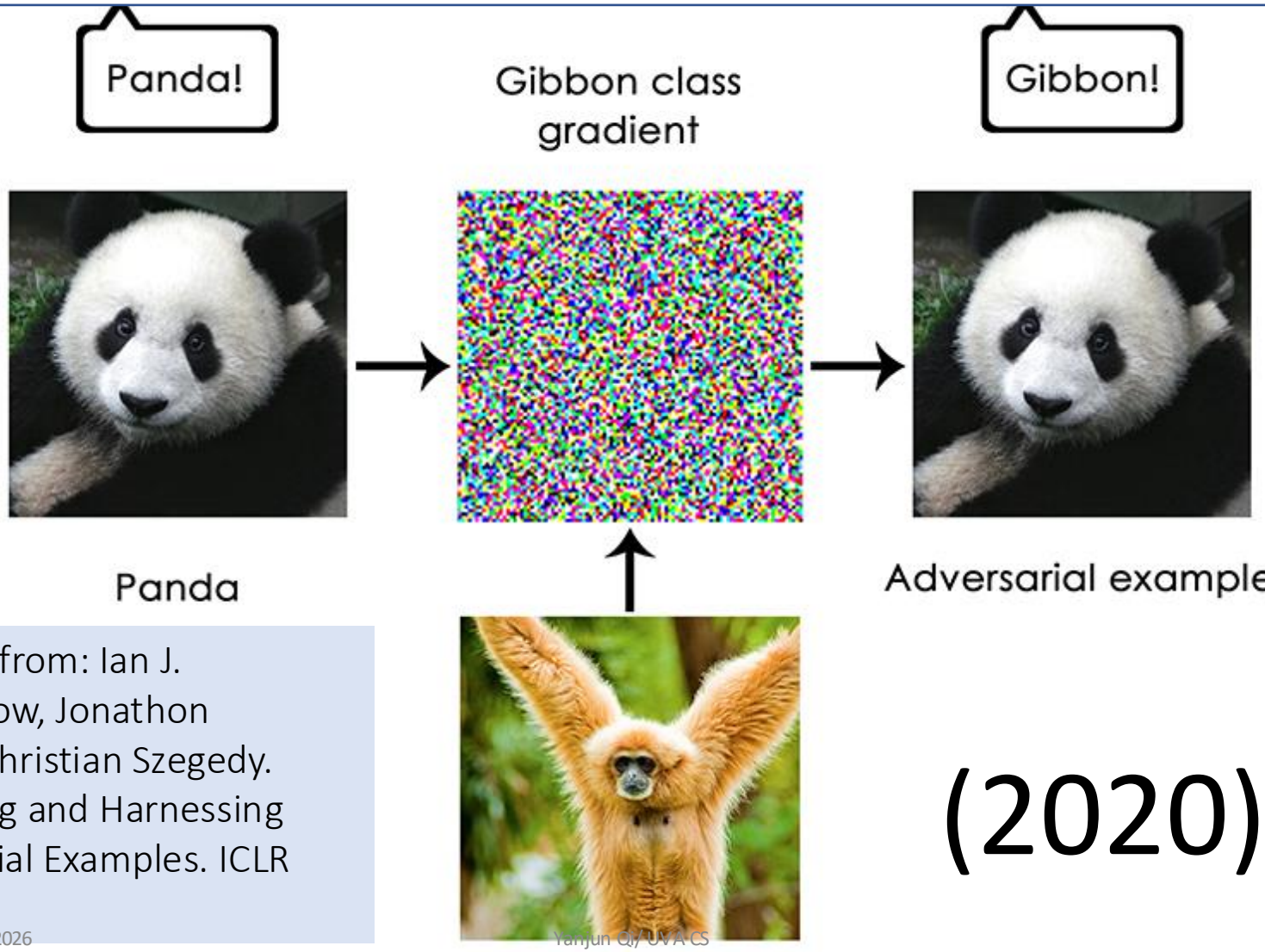
An open debate: features or raw neurons?

- Common view: raw neurons are mixed, so train a sparse autoencoder
- Arora et al. (Jan 2026): circuits are sparse in the neuron basis
  - About 100 neurons control subject-verb agreement
  - No extra sparse autoencoder to train
- Practical survey: locate, steer, improve
- Open: how much of a model can we really explain?

source: Arora et al. (arXiv 2601.22594); Zhang et al. (arXiv 2601.14004)

To Fool DNN, e.g. Adversarial Examples (AE)

"panda" + $0.007 \times [noise]$ "gibbon"

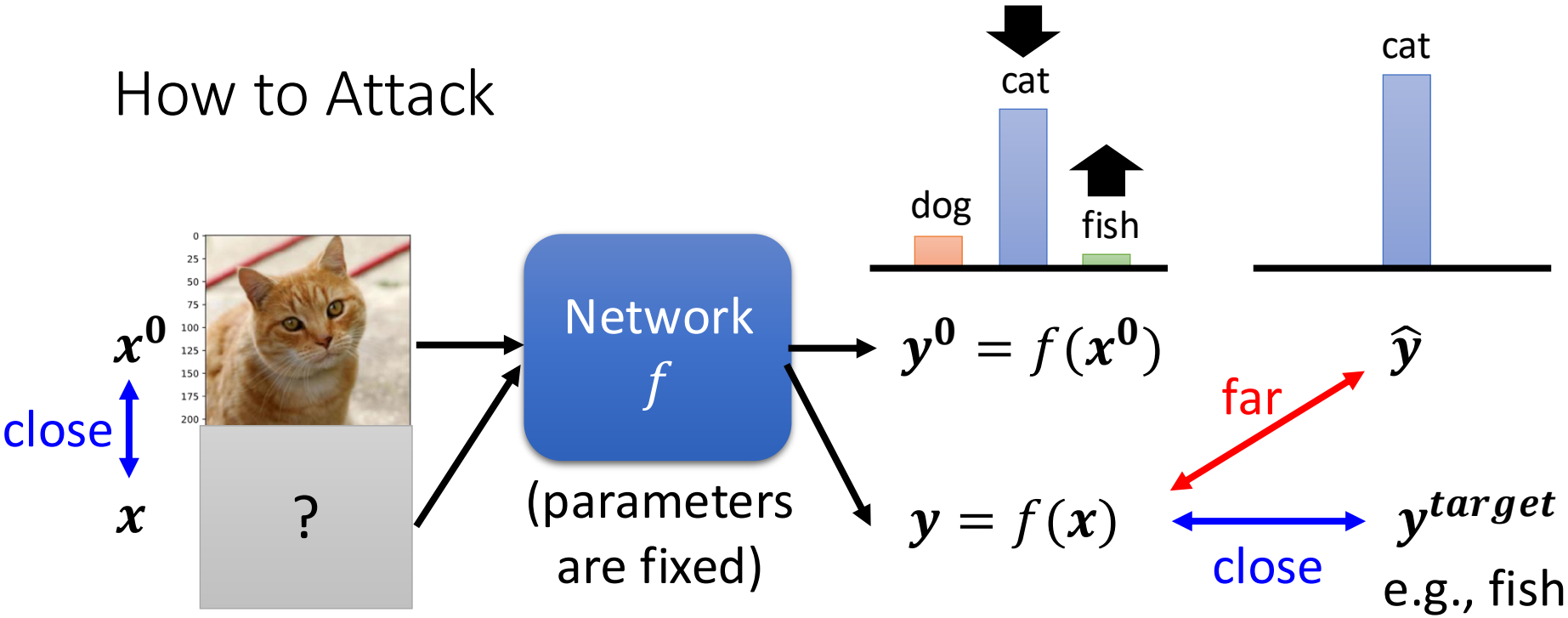


Example from: Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. Explaining and Harnessing Adversarial Examples. ICLR 2015.

(2020)

credit: Goodfellow et al., ICLR 2015; F. Chollet; Y. Qi, CS 4774 survey deck (2020)

How to Attack



Non-targeted $x^* = arg \min L(x)$

not perceived by humans

$L(x) = -e(y, \hat{y})$

Targeted $L(x) = -e(y, \hat{y}) + e(y, y^{target})$

credit: Prof. Hung-yi Lee, NTU ML 2021

Verifying a network: Reluplex (CAV 2017)

Provably robust?

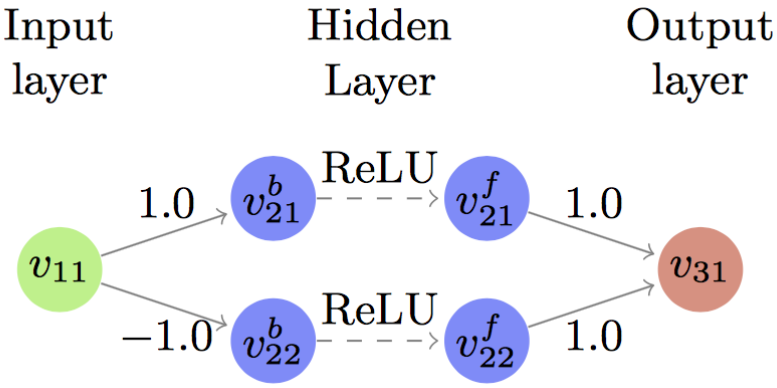


Table 3: Local adversarial robustness tests. All times are in seconds.

	$\delta = 0.1$		$\delta = 0.075$		$\delta = 0.05$		$\delta = 0.025$		$\delta = 0.01$		Total
	Result	Time	Result	Time	Result	Time	Result	Time	Result	Time	Time
Point 1	SAT	135	SAT	239	SAT	24	UNSAT	609	UNSAT	57	1064
Point 2	UNSAT	5880	UNSAT	1167	UNSAT	285	UNSAT	57	UNSAT	5	7394
Point 3	UNSAT	863	UNSAT	436	UNSAT	99	UNSAT	53	UNSAT	1	1452
Point 4	SAT	2	SAT	977	SAT	1168	UNSAT	656	UNSAT	7	2810
Point 5	UNSAT	14560	UNSAT	4344	UNSAT	1331	UNSAT	221	UNSAT	6	20462

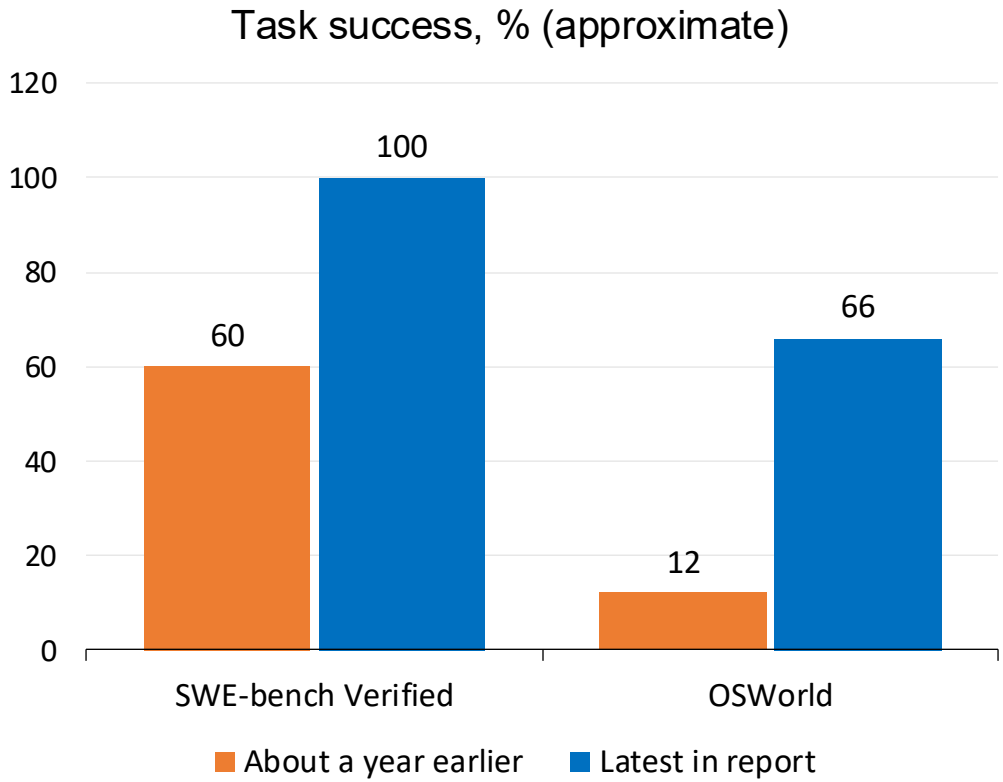
credit: Katz et al., CAV 2017; Y. Qi, CS 4774 survey deck (2020)

Verification at scale: α,β-CROWN

- Exact verification is NP-complete
- Bound propagation: compute cheap outer bounds, on a GPU
- Branch and bound: split unstable ReLUs where bounds are loose
- α,β-CROWN won VNN-COMP 2021 through 2025
- Open: networks the size of an LLM are still far out of reach
- A different use of formal methods: check outputs (Module II)

source: Verified-Intelligence/alpha-beta-CROWN; VNN-COMP 2025 report (arXiv 2512.19007)

The jagged frontier



- Benchmarks saturate in months
- IMO gold, yet analog clocks at 50.6%
- Agents still fail about 1 in 3 structured attempts
- Reported incidents: 233 → 362

Capability is not one number. See the evaluation lecture

source: Stanford AI Index 2026

METR time horizons (2026)

- Task length, in expert human time, that an agent completes with 50% reliability
- Example: GPT-5 at about 2 hours 17 minutes
- Fits an exponential trend better than linear or logistic
- METR's own limits:
  - Above 16 hours is unreliable with the current tasks
  - Mostly software, ML and cybersecurity tasks

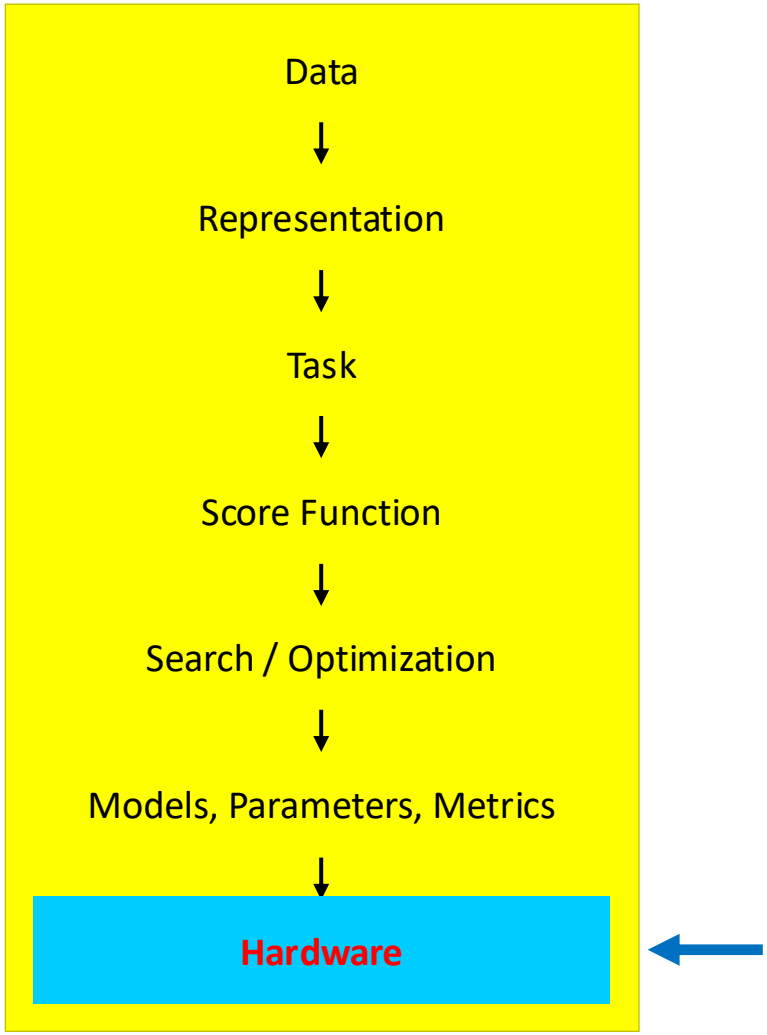
source: metr.org/time-horizons, as of Oct 2026

ARC-AGI-3: replacing the AGI slide

- 2020 deck: a definition of AGI vs narrow AI
- 2026: test novel skill acquisition directly
- Interactive, turn-based environments, with no instructions
- Scored on efficiency against human action baselines
- March 2026: humans solve 100%, frontier AI below 1%
- ARC Prize 2025 top Kaggle score on ARC-AGI-2: 24.0%

source: Chollet et al. (arXiv 2603.24621); ARC Prize 2025 results, as of Oct 2026

Nutshell: IV.2 Hardware



In 2020

- GPUs, TPUs, and mobile devices
- Compression for the edge

By 2026

- Compute is a constraint on research
- Separate chips for inference
- Heavy quantization of large models

Compute trends: what the money buys

Measure	Estimate
Chip performance per dollar	About +49% per year since 2023
Doubling time for that	About 1.7 years
Frontier training compute	About 5× per year since 2020
Largest known AI data center	About 1.1M H100 equivalents

Frontier research now needs capital on an industrial scale

source: Epoch AI Trends, as of Oct 2026

Inference gets its own silicon

- Training: large batches and huge communication
- Inference: low latency, many small requests
- NVIDIA Vera Rubin platform: seven chips, adding the Groq 3 LPU for inference
- Google Ironwood TPU: built for inference
- Wafer-scale chips: Cerebras
- Why now: an agent calls the model many times

source: HPCwire, Mar 2026; Google Cloud Ironwood announcement

Supply concentration

- One company, TSMC, fabricates almost every leading AI chip
- The US hosts 5,427 data centers, over 10× any other country
- Top US model leads top Chinese model by 2.7% (Mar 2026)
- Model production concentrated in the US and China

Hardware now includes supply chains, energy and policy

source: Stanford AI Index 2026

DNN Model Compression Methods: Overview (2020)

(a) Model pruning (neurons, filters, kernels, layer)

- Magnitude-based method
  - Iterative pruning + Retraining
  - Pruning with rehabilitation
- Hessian-based method
  - Diagonal Hessian-based method
  - Full Hessian-based method

(b) Construct simpler filters / e.g.

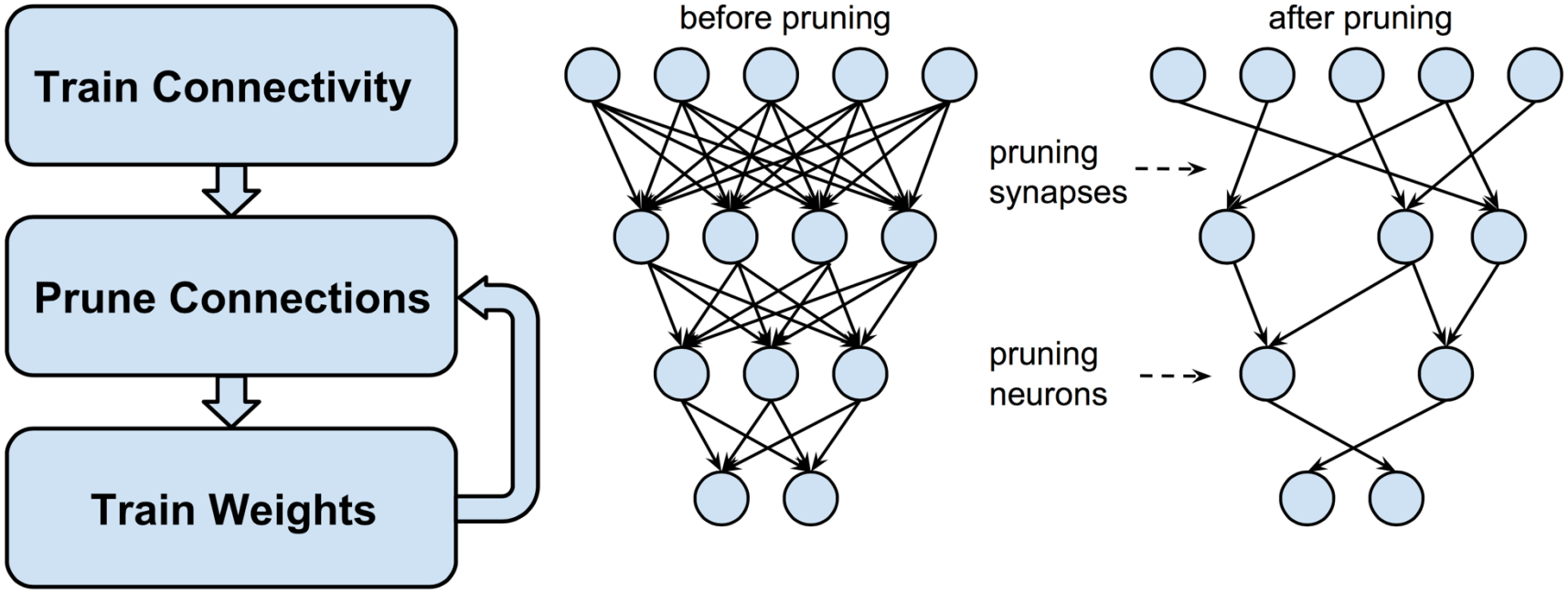
- Decomposition, Matrix Factorization , Singular Value Decomposition (SVD)
- Flattened Convolutions

(c) Quantization of activations, weights, and even gradients.

- Full Quantization
  - Fixed-point format
  - Code book
- Quantization with full-precision copy

credit: Y. Qi, CS 4774 survey deck (2020)

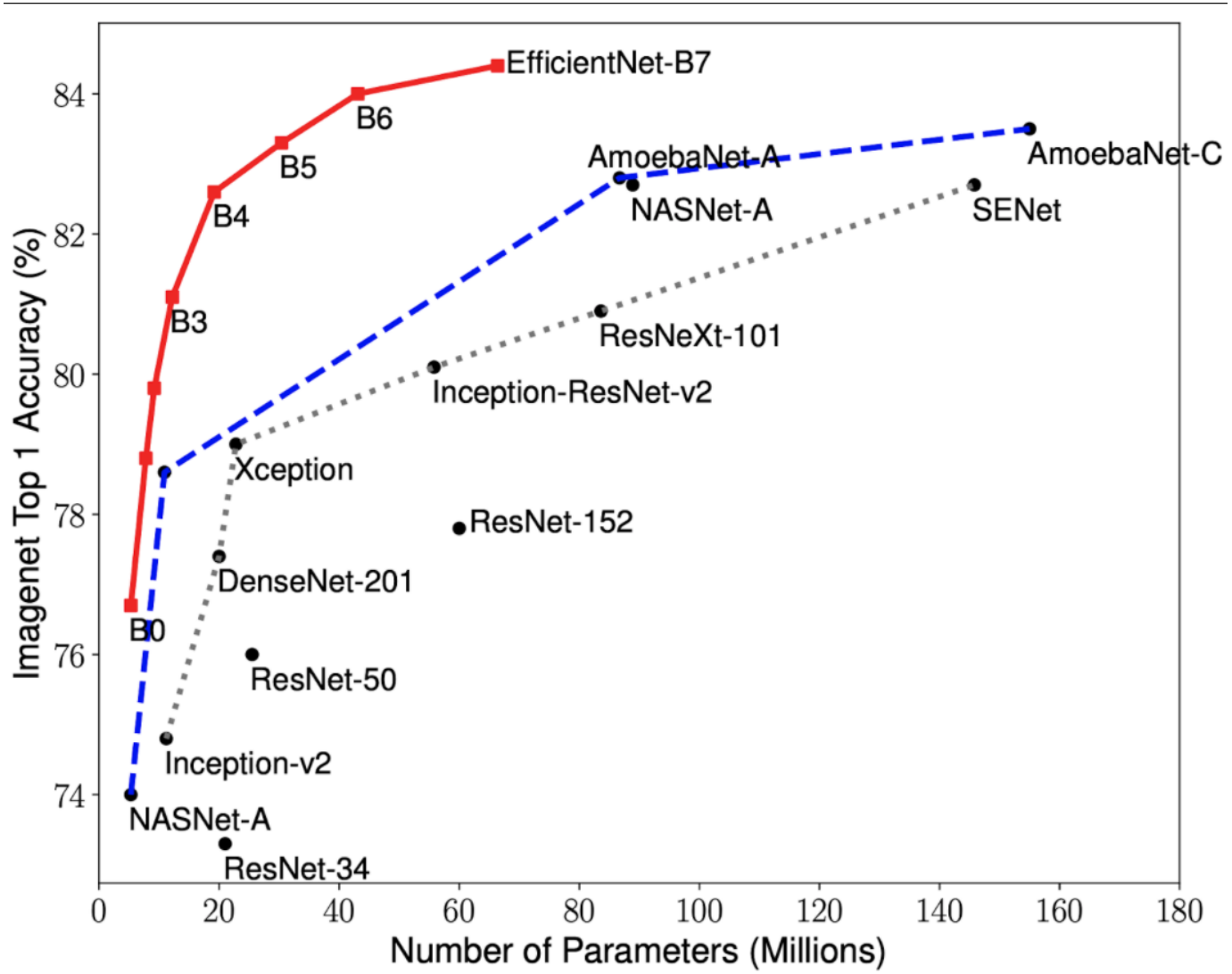
Magnitude pruning: prune, then retrain (2020)



Han, Song, et al. "Learning both weights and connections for efficient neural network." NIPS. 2015.

credit: Han et al., NIPS 2015; Y. Qi, CS 4774 survey deck (2020)

EfficientNet: compound scaling (2019)



credit: Tan and Le, ICML 2019; Google AI Blog; Y. Qi, CS 4774 survey deck (2020)

On-device: big and quantized beats small

Model	Bits per weight	Memory
2 B parameters	16	4 GB
8 B parameters	4	4 GB

- Finding: heavily quantized large models consistently beat smaller high-precision ones
- Reported threshold: about 3.5 effective bits per weight
- Table: illustrative arithmetic, not results from the paper

source: Song et al. (arXiv 2505.15030)

Quiz examples: Module IV

1. How does circuit tracing go further than a saliency map?

2. A 1 B parameter model at 16 bits takes 2 GB. How many parameters fit in 2 GB at 4 bits?

3. About 66% on OSWorld, under 1% on ARC-AGI-3 (Mar 2026). How can both be true?

Recap: with earlier lecture links

Stage	Today's message	Earlier lectures
Data	Build the symmetry in: graphs, 3D	CNNs, multimodal
Representation	Families mix; experts add scale	CNNs, Transformers, VLMs, MoE
Task	One general model, with a condition	RL, prompting, VLMs
Score function	Same trick: make a supervised problem	deep NNs, autoencoders, GPT, VLMs
Search	From a defined space to code	Agents
Metrics	Capability is not one number	Evaluation, safety
Hardware	Inference chips, heavy quantization	LoRA, agents, serving

What lasted from 2020

- Inductive bias: put the data's structure in the model
- Pretrain without labels, then fine-tune
- Residual connections, attention, gradient descent
- Regression on a good target: simple training wins
- Check the result: a verifier, a benchmark, a held-out set
- The pipeline itself still organizes the field

References: field reports and courses

Stanford HAI, AI Index Report 2026. hai.stanford.edu/ai-index/2026-ai-index-report

AI Index 2026, Technical Performance chapter. hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance

Epoch AI, Trends in AI. epoch.ai/trends

Benaich et al., State of AI Report 2025. www.stateof.ai/2025-report-launch

Stanford CS231n, Spring 2026 schedule. cs231n.stanford.edu/schedule.html

Stanford CS231n 2026, Lecture 13 (self-supervised). cs231n.stanford.edu/slides/2026/lecture_13.pdf

Stanford CS224W, GNN lecture slides. web.stanford.edu/class/cs224w/slides/06-theory.pdf

MIT 6.S191, Introduction to Deep Learning 2026. introtodeeplearning.com/

MIT 6.S184, Flow Matching and Diffusion Models 2026. diffusion.csail.mit.edu/2026/

Holderrieth and Erives, An Introduction to Flow Matching and Diffusion Models. arxiv.org/abs/2506.02070

Cornell CS 5788, Flow Matching lecture, Spring 2026. www.cs.cornell.edu/courses/cs5788/2026sp/slides/lec16-flowmatching.pdf

Hung-yi Lee, NTU Machine Learning 2021 (slides credited in this deck). speech.ee.ntu.edu.tw/~hylee/ml/2021-spring.php

Y. Qi, Reading Deep Learning (2020 topic map). qdata.github.io/deep2Read/

References: Module I, Data and Representation

2024 Nobel Prize in Physics. www.nobelprize.org/prizes/physics/2024/summary/

2024 Nobel Prize in Chemistry. www.nobelprize.org/prizes/chemistry/2024/popular-information/

Gilmer et al. 2017, Neural Message Passing for Quantum Chemistry. arxiv.org/abs/1704.01212

Bronstein et al. 2021, Geometric Deep Learning. arxiv.org/abs/2104.13478

Abramson et al. 2024, AlphaFold 3, Nature. pmc.ncbi.nlm.nih.gov/articles/PMC11168924

Price et al., GenCast: Diffusion-based ensemble forecasting. arxiv.org/abs/2312.15796

Bodnar et al., Aurora: A Foundation Model of the Atmosphere. arxiv.org/abs/2405.13063

ECMWF, AIFS machine-learning data. www.ecmwf.int/en/forecasts/dataset/aifs-machine-learning-data

Siméoni et al. 2025, DINOv3. arxiv.org/abs/2508.10104

Meta, SAM 3: Segment Anything with Concepts. arxiv.org/abs/2511.16719

Shazeer et al. 2017, Outrageously Large Neural Networks. arxiv.org/abs/1701.06538

DeepSeek-V4-Pro model card. huggingface.co/deepseek-ai/DeepSeek-V4-Pro

Gu and Dao 2023, Mamba. arxiv.org/abs/2312.00752

Yu et al., KAN or MLP: A Fairer Comparison. arxiv.org/abs/2407.16674

Review of KANs, 2026. arxiv.org/abs/2510.25781

References: Module II, Tasks

Google DeepMind, Veo. deepmind.google/models/veo/

OpenAI, Sora 2 System Card. openai.com/index/sora-2-system-card/

Chen and Wang, A Survey on 3D Gaussian Splatting. arxiv.org/abs/2401.03890

Google DeepMind, SynthID. deepmind.google/models/synthid/

C2PA, Coalition for Content Provenance and Authenticity. c2pa.org/

FTC, TAKE IT DOWN Act reporting. takeitdown.ftc.gov/

EU AI Act, Article 50. artificialintelligenceact.eu/article/50/

Hafner et al. 2025, DreamerV3. danijar.com/project/dreamerv3/

Google DeepMind 2025, Genie 3. deepmind.google/blog/genie-3-a-new-frontier-for-world-models/

The Reality Gap in Robotics, 2025. arxiv.org/abs/2510.20808

Hubert et al. 2025, AlphaProof. research.google/pubs/olympiad-level-formal-mathematical-reasoning-with-reinforcement-learning/

Google DeepMind 2025, Gemini Deep Think at IMO. deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/

Harmonic, Aristotle. arxiv.org/abs/2510.01346

DeepSeek-R1, reported in Scientific American. www.scientificamerican.com/article/secrets-of-chinese-ai-model-deepseek-revealed-in-landmark-paper

Sinha et al. 2024, MAML-en-LLM. arxiv.org/abs/2405.11446

Finn et al. 2017, MAML. arxiv.org/abs/1703.03400

Graves et al. 2014, Neural Turing Machines. arxiv.org/abs/1410.5401

References: Module III, Losses and Search

Goodfellow et al. 2014, Generative Adversarial Nets. arxiv.org/abs/1406.2661

Karras et al. 2019, StyleGAN. arxiv.org/abs/1812.04948

Huang et al. 2024, R3GAN. arxiv.org/abs/2501.05441

Lipman et al. 2022, Flow Matching for Generative Modeling. arxiv.org/abs/2210.02747

Liu et al. 2022, Rectified Flow. arxiv.org/abs/2209.03003

Song et al. 2023, Consistency Models. arxiv.org/abs/2303.01469

Survey on Pre-Trained Diffusion Model Distillations, 2025. arxiv.org/abs/2502.08364

ICML 2026 awards, via AIhub. aihub.org/2026/07/20/congratulations-to-the-icml2026-award-winners/

Chen et al. 2020, SimCLR. arxiv.org/abs/2002.05709

He et al. 2021, Masked Autoencoders. arxiv.org/abs/2111.06377

Caron et al. 2021, DINO. arxiv.org/abs/2104.14294

Assran et al. 2023, I-JEPA. arxiv.org/abs/2301.08243

Assran et al. 2025, V-JEPA 2. arxiv.org/abs/2506.09985

Balestriero and LeCun 2025, LeJEPA. arxiv.org/abs/2511.08544

Zoph and Le 2017, Neural Architecture Search. arxiv.org/abs/1611.01578

Zhang et al., Darwin Gödel Machine. arxiv.org/abs/2505.22954

Sakana AI, The AI Scientist in Nature. sakana.ai/ai-scientist-nature/

Beel et al. 2025, Evaluating Sakana's AI Scientist. arxiv.org/abs/2502.14297

References: Module IV, Trust and Hardware

Anthropic 2025, Tracing the thoughts of a large language model. www.anthropic.com/research/tracing-thoughts-language-model

Zhang et al. 2026, Locate, Steer, and Improve. arxiv.org/abs/2601.14004

Arora et al. 2026, Language Model Circuits Are Sparse in the Neuron Basis. arxiv.org/abs/2601.22594

Goodfellow et al. 2015, Explaining and Harnessing Adversarial Examples. arxiv.org/abs/1412.6572

Katz et al. 2017, Reluplex. arxiv.org/abs/1702.01135

α,β-CROWN verifier. github.com/Verified-Intelligence/alpha-beta-CROWN

VNN-COMP 2025 report. arxiv.org/abs/2512.19007

METR, Task-Completion Time Horizons. metr.org/time-horizons/

Chollet et al. 2026, ARC-AGI-3. arxiv.org/abs/2603.24621

ARC Prize 2025 Results and Analysis. arcprize.org/blog/arc-prize-2025-results-analysis

HPCwire 2026, NVIDIA Vera Rubin chips. www.hpcwire.com/2026/03/16/nvidia-boasts-7-chips-in-production-for-vera-rubin-platform-including-groq-3-lpu/

Google 2025, Ironwood TPU. blog.google/products/google-cloud/ironwood-tpu-age-of-inference/

Han et al. 2015, Deep Compression. arxiv.org/abs/1510.00149

Tan and Le 2019, EfficientNet. arxiv.org/abs/1905.11946

Song et al. 2025, A Systematic Evaluation of On-Device LLMs. arxiv.org/abs/2505.15030