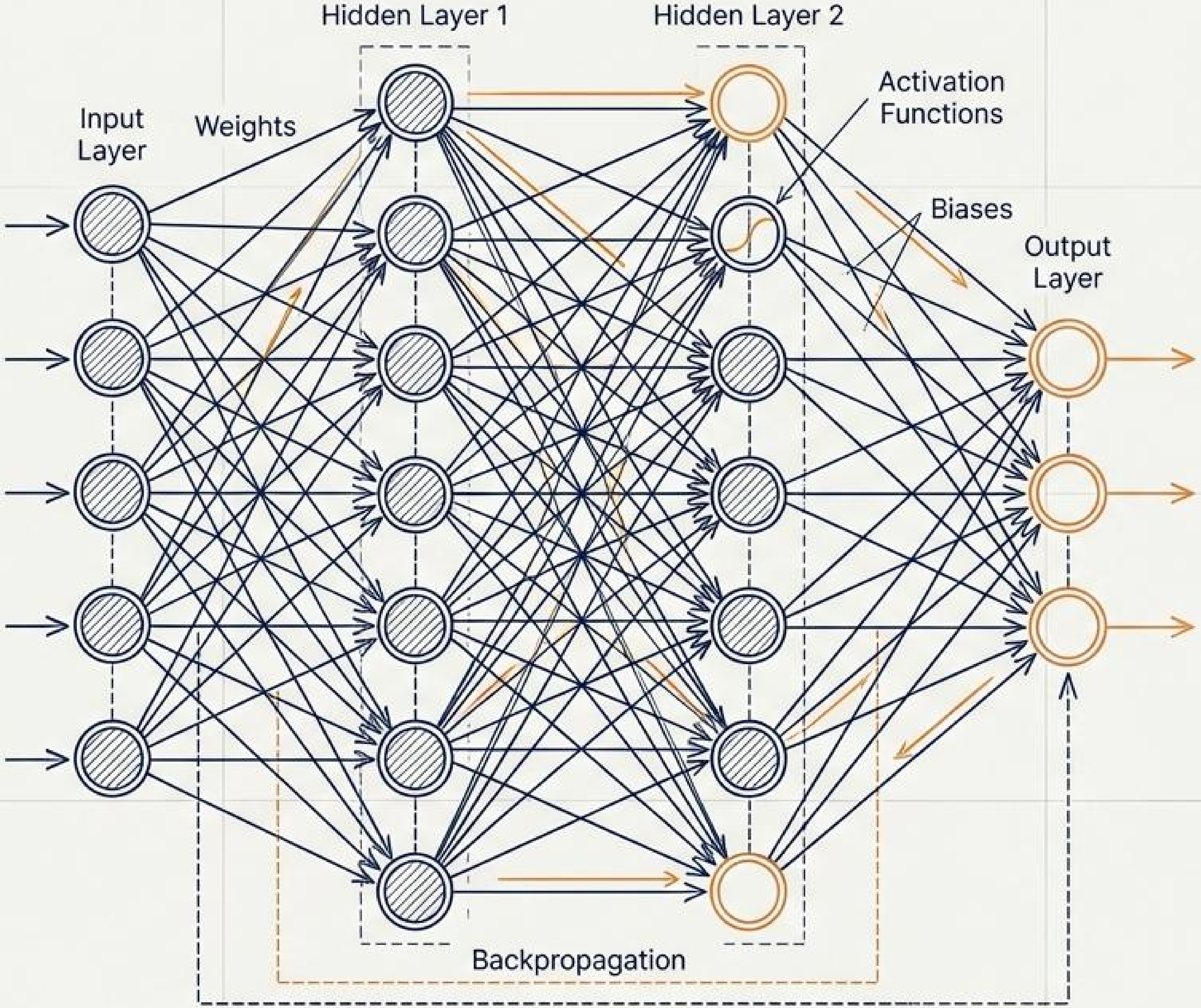


CS-878 DEEP LEARNING

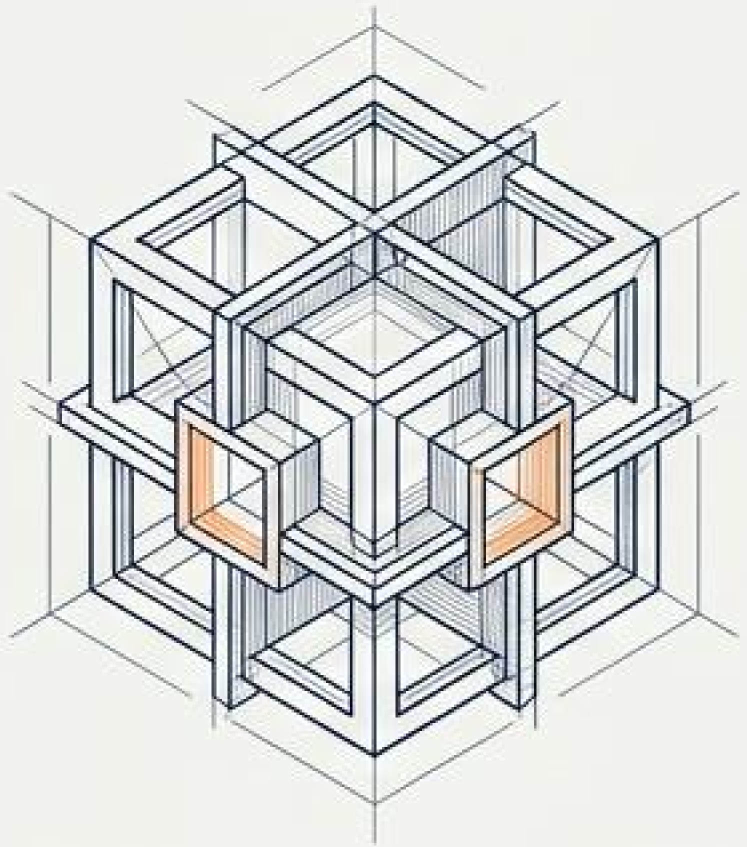
Week 01 Introduction

BIOMISA

- Introduction & History
- Core Concepts
- Why Now?

Course Learning Objectives

01. BUILD



Design, implement, and optimize architectures. Master Convolutional, Recurrent, and Transformer networks to solve complex problems in Computer Vision and NLP.

02. INNOVATE

Gain proficiency in advanced techniques. Apply Generative Models, Attention Mechanisms, and Self-Supervised learning to create novel solutions.

03. RESPONSIBILITY

Develop critical understanding. Navigate ethical implications, interpretability, and fairness in AI deployment.

The Roadmap | Phase I

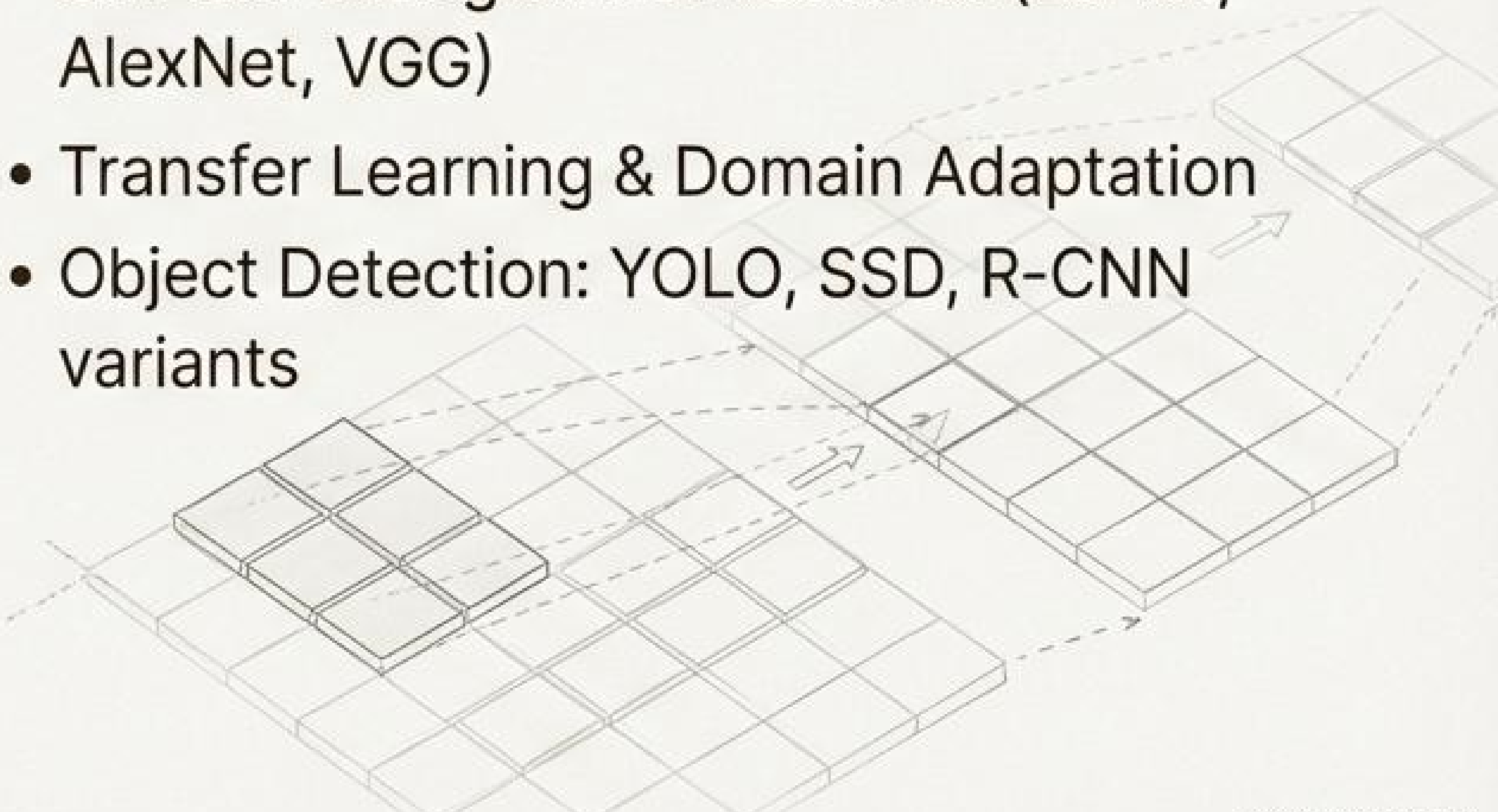
Foundations & Vision

Topic Cluster: The Basics

- Introduction & Historical Context
- Perceptron & Feedforward Networks
- Training: Backpropagation, Gradient Descent, Optimization
- Regularization & Handling Overfitting

Topic Cluster: Computer Vision

- CNNs: Pooling & Architectures (LeNet, AlexNet, VGG)
- Transfer Learning & Domain Adaptation
- Object Detection: YOLO, SSD, R-CNN variants



The Roadmap | Phase II

Sequence, Context & Segmentation

Topic Cluster: Sequence Modeling

- Recurrent Neural Networks (RNNs)
- Long Short-Term Memory (LSTM) & GRU
- Advanced Sequence Applications

Topic Cluster: Space & Language

- Semantic Segmentation: U-Net & Fully Convolutional Networks
- NLP: Word Embeddings & Sequence-to-Sequence
- Modern NLP: BERT & GPT Architectures

The Roadmap | Phase III

The Generative Frontier

Topic Cluster: Latent Spaces

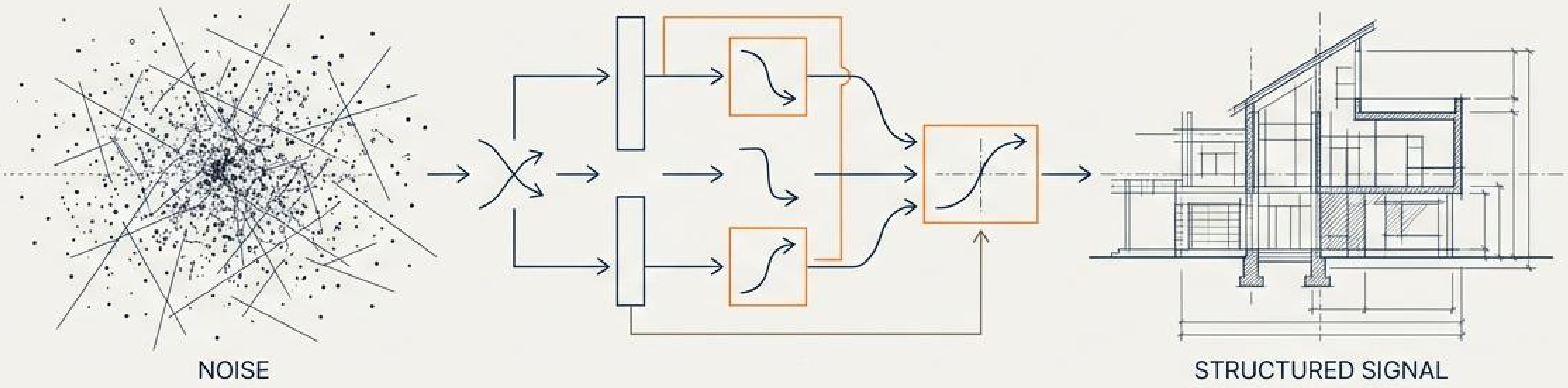
- Semi-Supervised Learning
- Autoencoders
- Variational Autoencoders (VAEs)

Topic Cluster: Adversarial

- Generative Adversarial Networks (GANs)
- Principles & Applications

Topic Cluster: Modern Synthesis

- Diffusion Models
- Comparison: Diffusion vs. GANs



Phase IV

The Roadmap | Phase IV

Advanced Architectures & Deployment

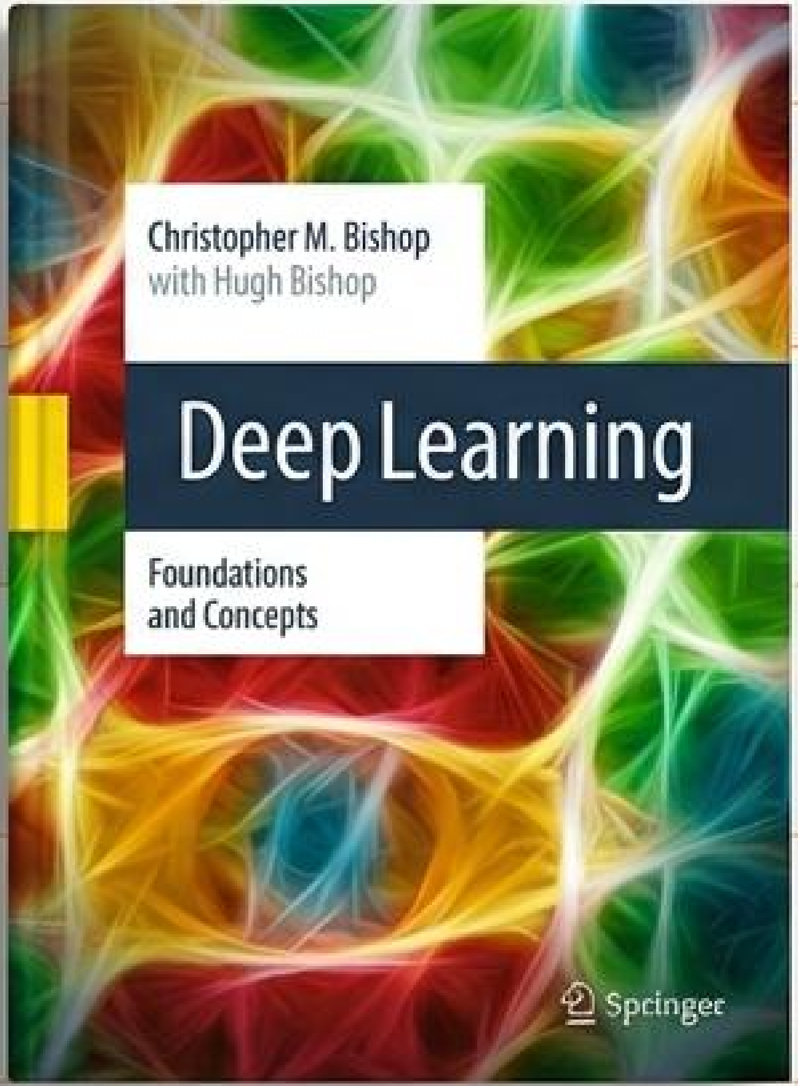
Topic Cluster: Attention & Integration

- Attention Mechanisms & Self-Attention
- Transformers Architecture
- Vision Transformers (ViTs)
- CLIP & Multi-modal Learning

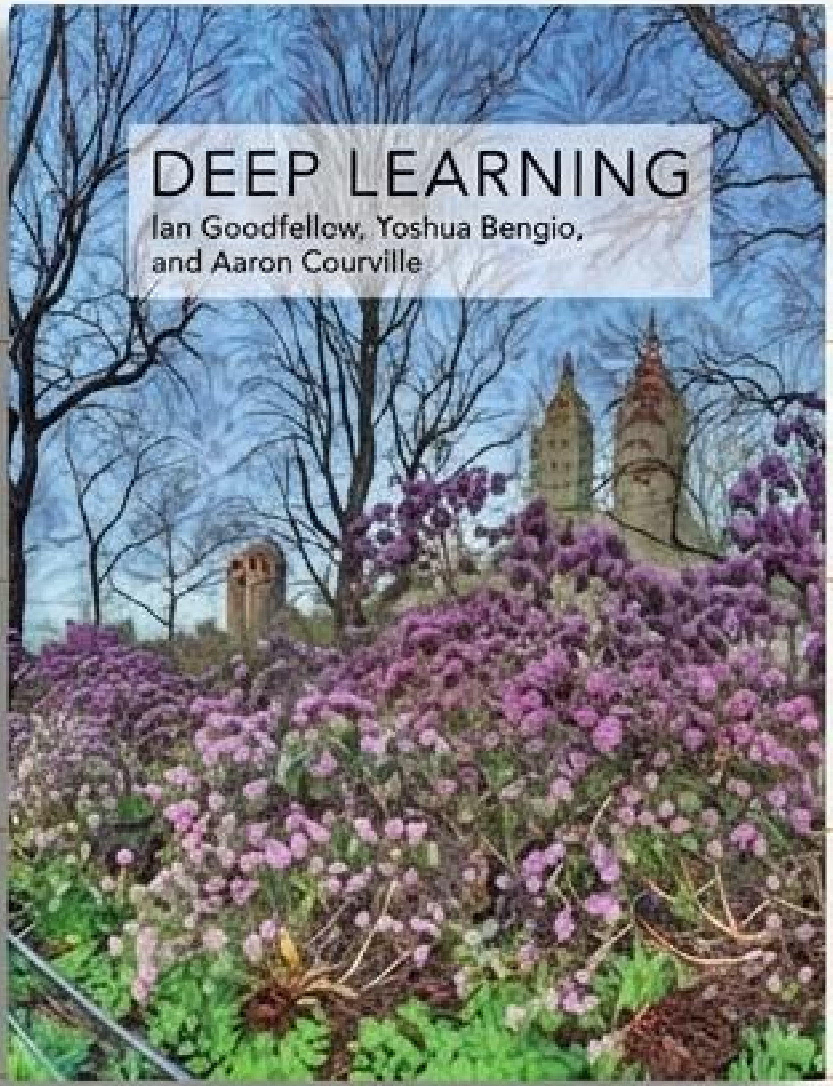
Topic Cluster: Responsibility & Scale

- Explainability (XAI) & Interpretability
- Model Optimization
- Edge Deployment Strategies
- Scalability Issues

Core Textbooks



Primary Text: *Deep Learning: Foundations and Concepts*
Christopher M. Bishop & Hugh Bishop (2024)
bishopbook.com



Classic Reference: *Deep Learning*
Ian Goodfellow, Yoshua Bengio, & Aaron Courville (2016)
deeplearningbook.org

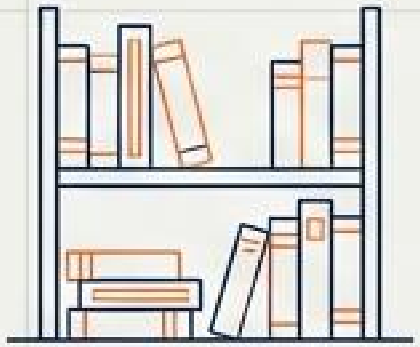
Extended Knowledge Base

Additional Reference Books

- Pattern Recognition and Machine Learning (Bishop)
- Dive into Deep Learning (Zhang et al.)
- Mathematics for Machine Learning (Deisenroth et al.)
- The Matrix Cookbook

Top University Courses

- Stanford CS231n (CNNs) & CS230
- CMU (Deep Learning Systems)
- Berkeley (Deep Unsupervised Learning)
- UoT (Intro to Neural Networks)



Prerequisites

Required Baseline Knowledge

- **Mathematics & Theory**
  Linear Algebra, Probability, Information Theory. (Remediation: Goodfellow Ch 1-4)
- **Computer Science Fundamentals**
  Variables, functions, loops, classes, and algorithms.
- **Programming Proficiency**
  Python (Refer to official docs.python.org tutorials).

Development Environment

Python
NumPy Tutorial provided.

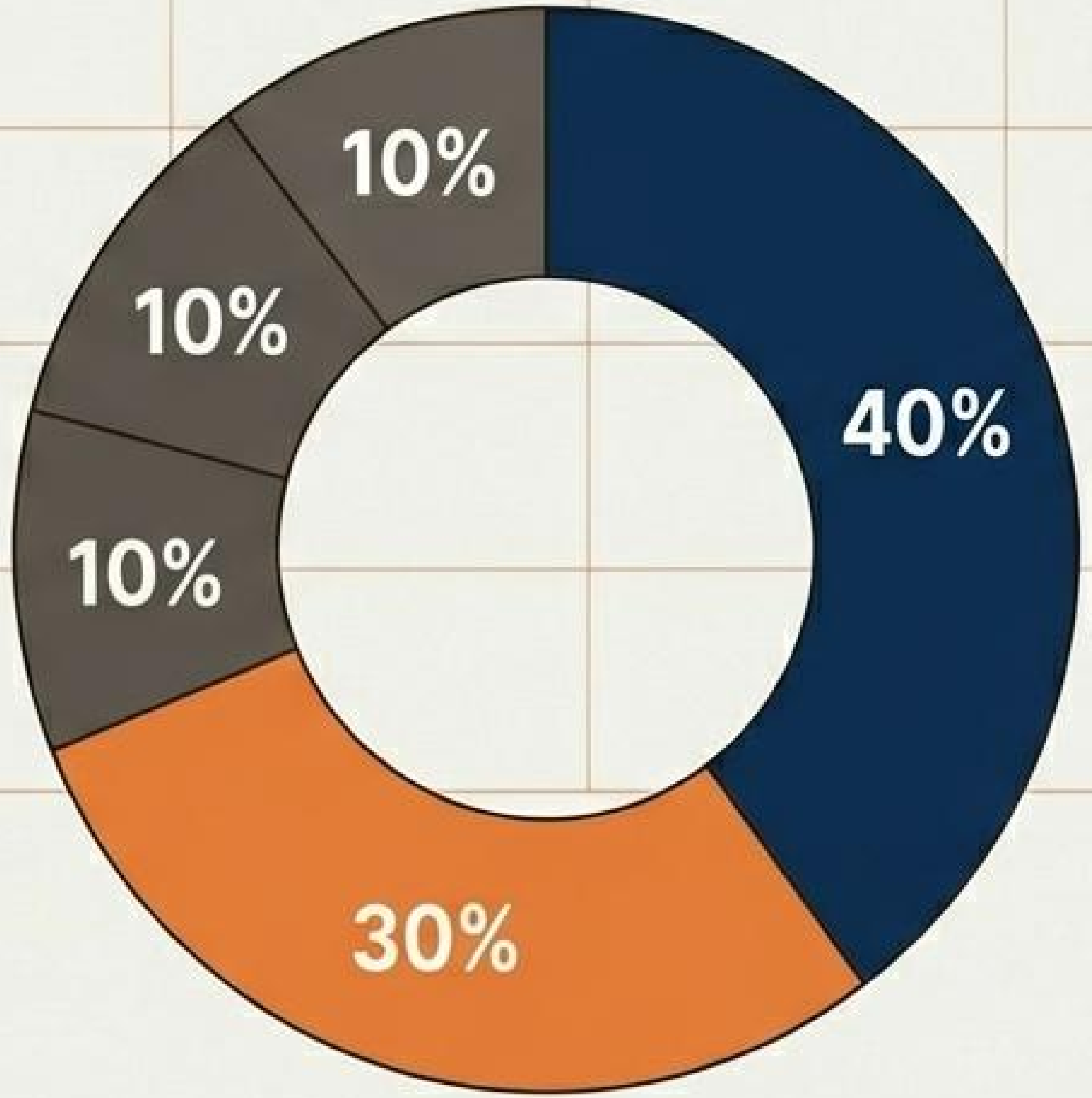
PyTorch
Official Tutorials.

Visual Studio Code
Recommended IDE.

Google Colab
Cloud Resource.

Grading Scheme

100% Theory Component



- **40%:** Final Exam
- **30%:** Mid-term Exam
- **10%:** Semester Project (Design Report + Presentation)
- **10%:** Quizzes
- **10%:** Assignments (3 Graded)

Note: Lab Reports: N/A

Operational Policies

EXAMS

Open Book Policy.

DEADLINES

No extensions on assignments.

QUIZZES

Unannounced.

ATTENDANCE

Mandatory. Arrive on time.

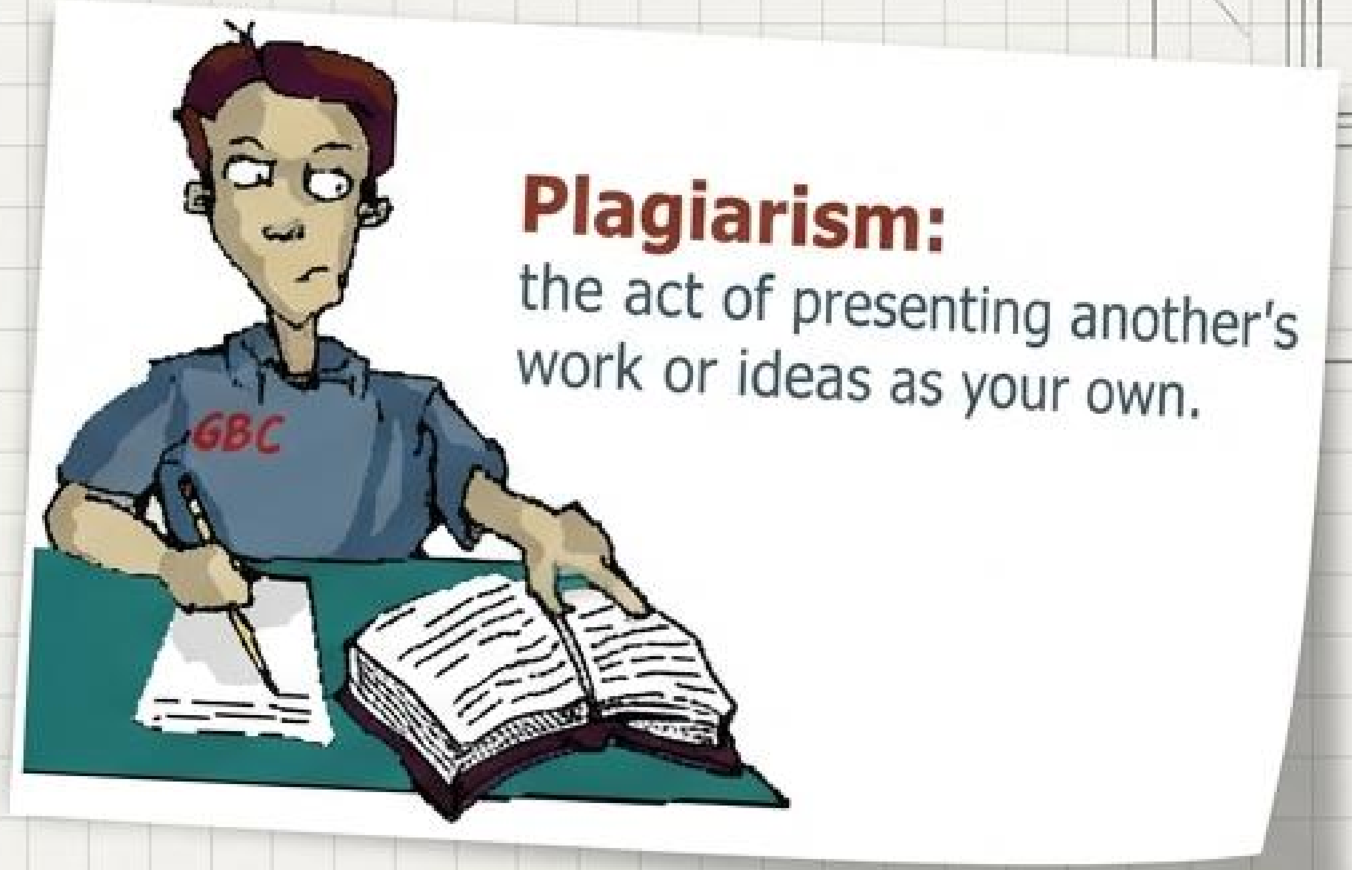
CONDUCT

No gadgets during lecture. Respect peers and faculty.

Code of Ethics & Integrity

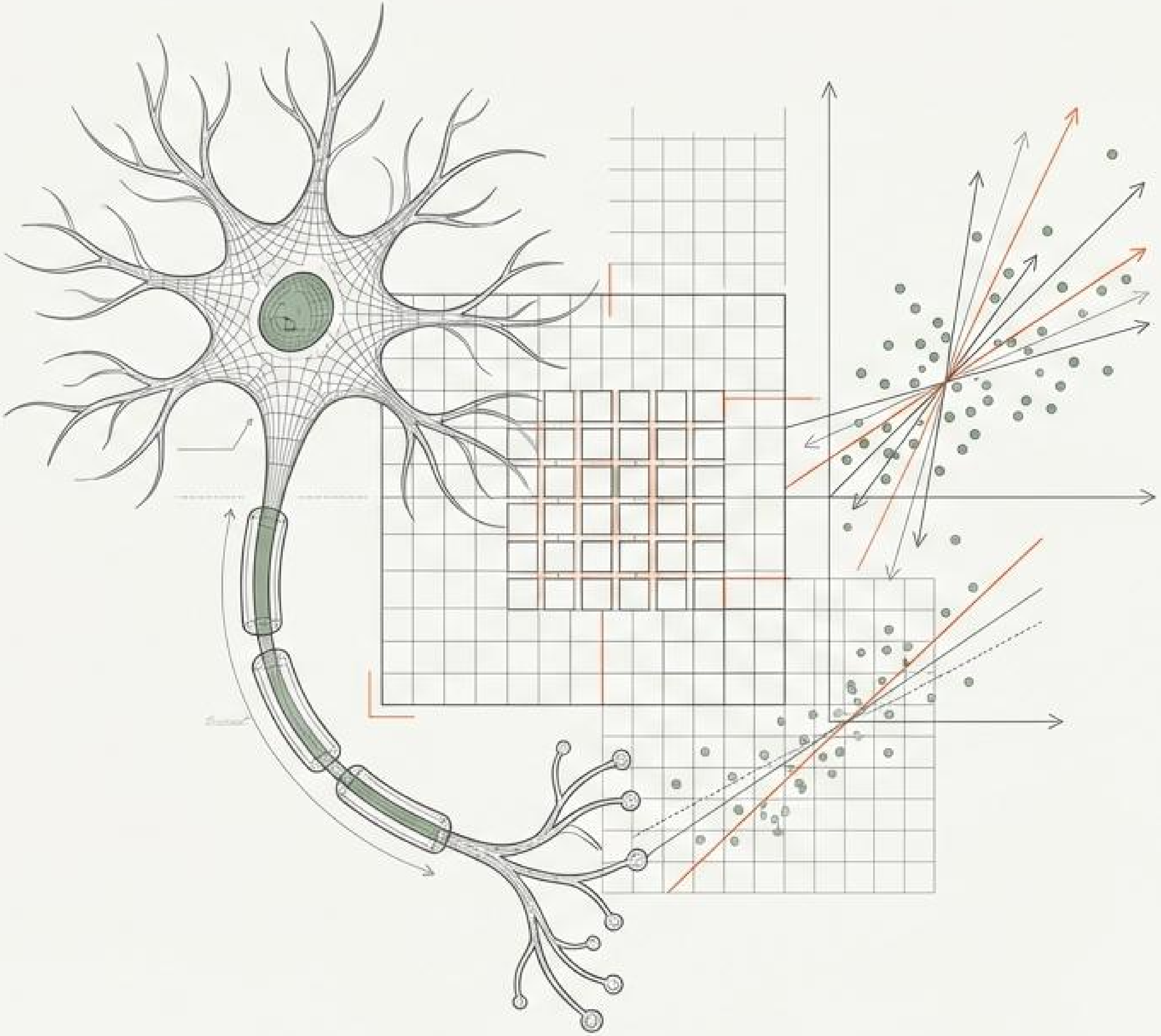
"Better fail NOW or else will fail somewhere LATER in life."

- **Plagiarism:** Presenting another's work or ideas as your own will face strict penalties.
- Obedience to laws, discipline code, and community norms.



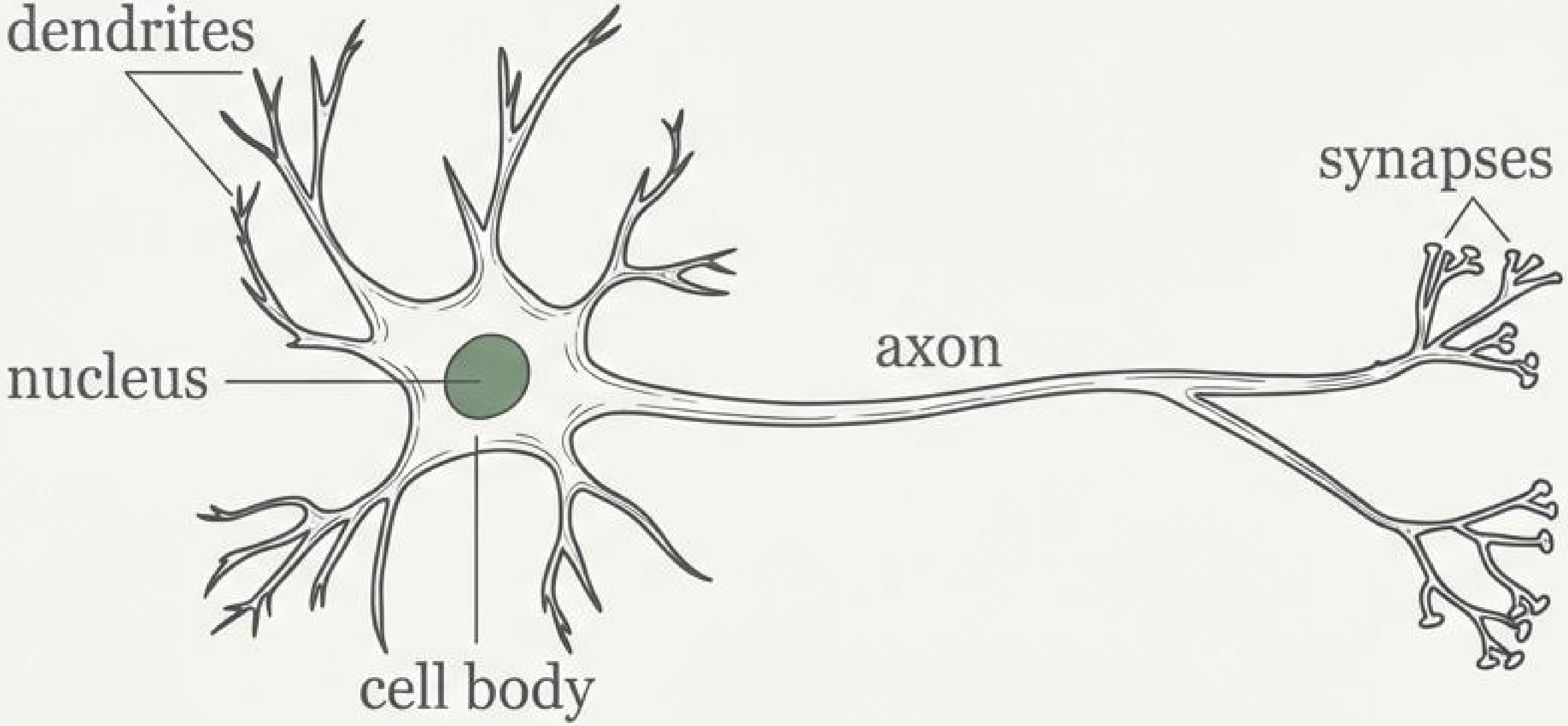
The Deep Learning Landscape

Architectures, Applications, and the Functional Evolution of Artificial Intelligence



Biological Inspiration

Computational models inspired by biological machinery.



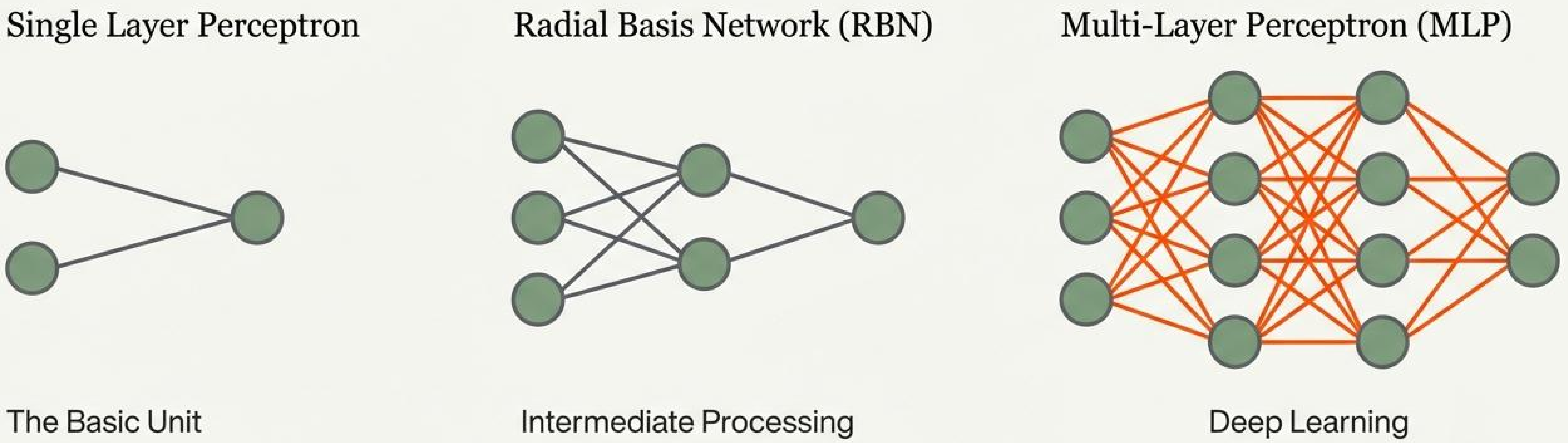
Artificial Neural Networks (ANNs) are not magic; they are mathematical abstractions of this biology.

1. Dendrites = Input Receivers.
2. Cell Body = Processing Unit (Summation).
3. Axon/Synapses = Signal Transmission to the next unit.

Key Takeaway: Just as biological neurons connect to process signals, artificial neurons are mathematical functions connected to transmit information.

Network Architectures

Complexity arises from connection.



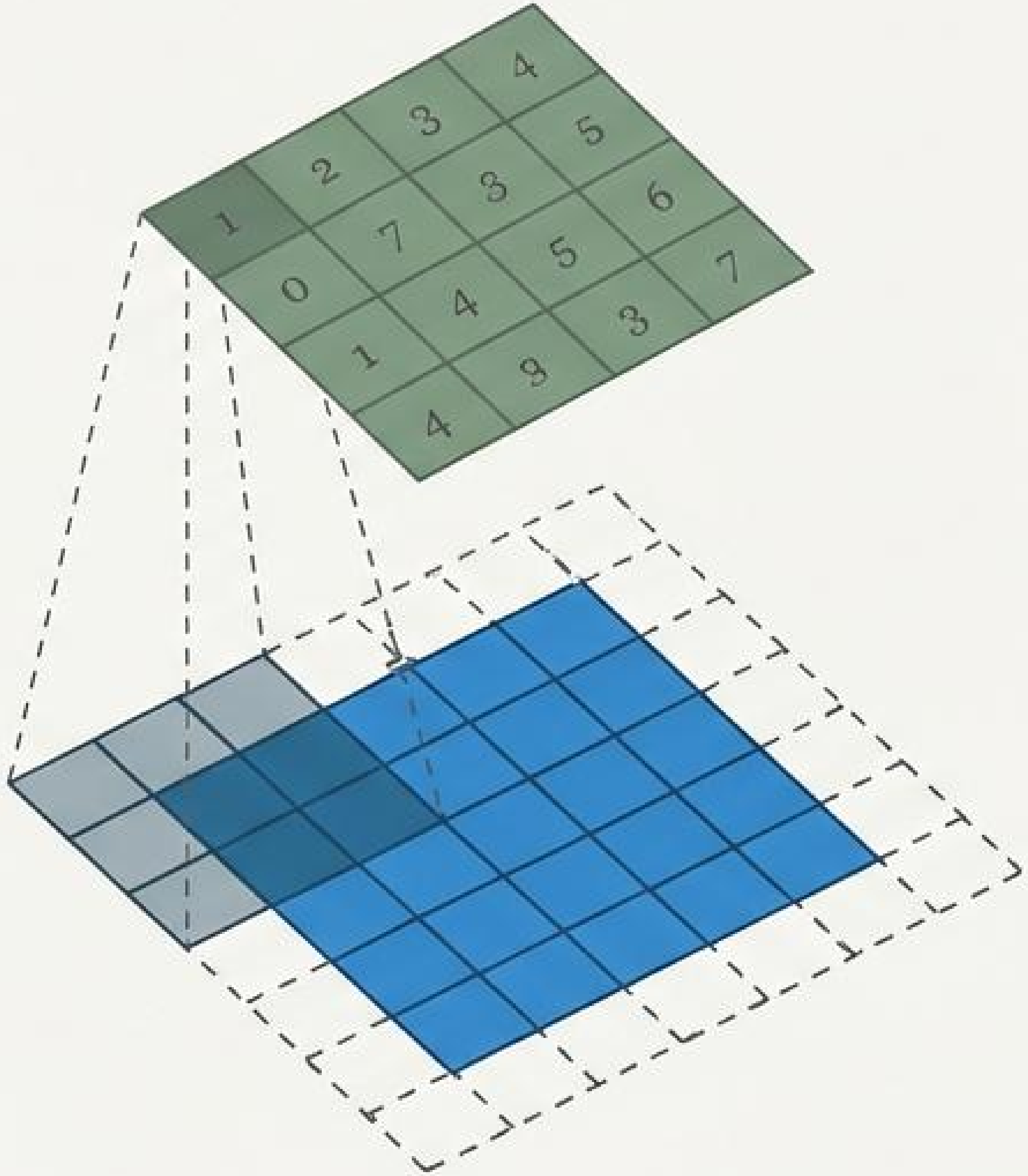
The evolution from single computational units to Deep Learning involves scaling the number of layers (Depth). This allows the machine to learn hierarchical representations of data, solving non-linear and complex problems that a single layer cannot.

Computer Vision: Teaching Machines to See

Convolutional Neural Networks (CNNs) & Spatial Invariance.

55	120	240
90	110	235
85	115	242
70	124	257
85	127	246
66	120	240

Matrix representation



The Filter Scans for Patterns

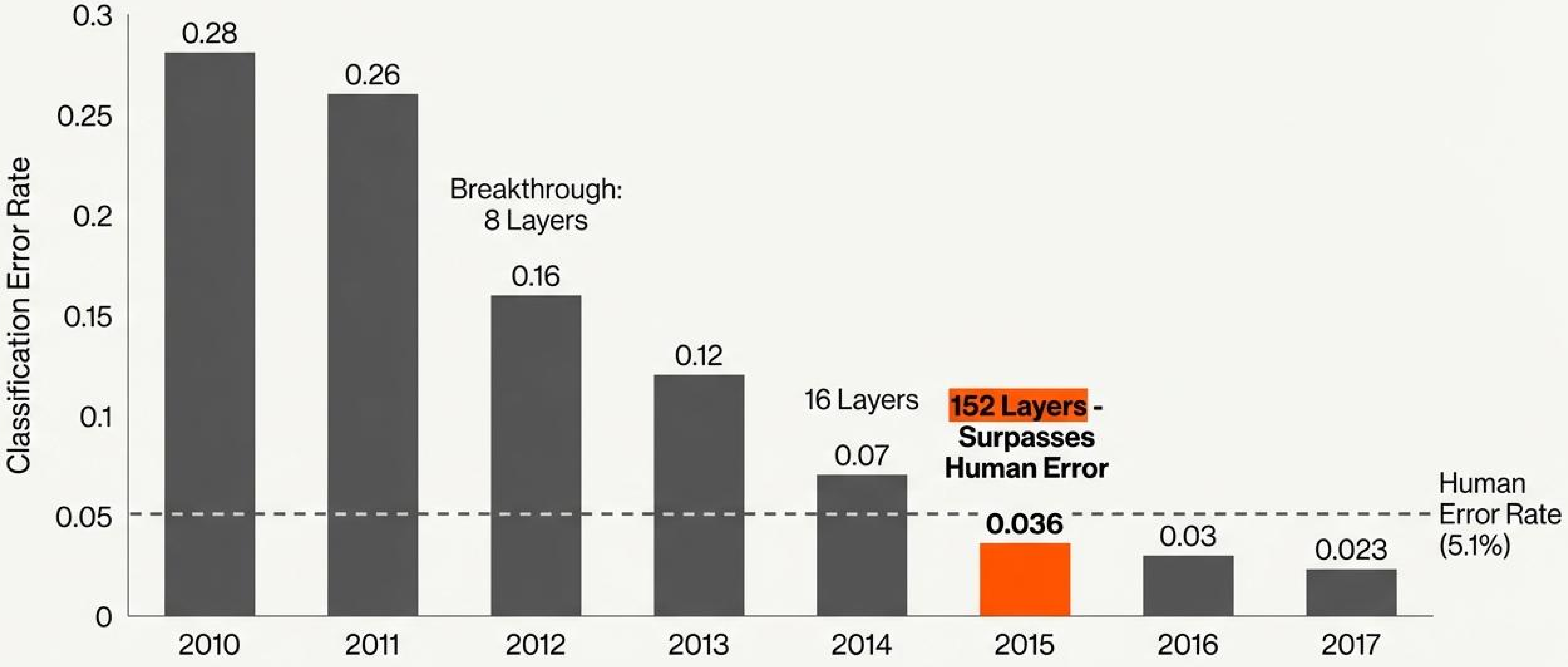
How it works:

1. **Input**: The computer sees a matrix of numbers, not a picture.
2. **Convolution**: Filters slide over the image to detect features (edges, textures) regardless of location.
3. **Output**: A probability score (e.g., 82% Cat, 15% Dog).

→ 82% cat
15% dog
2% hat
1% mug

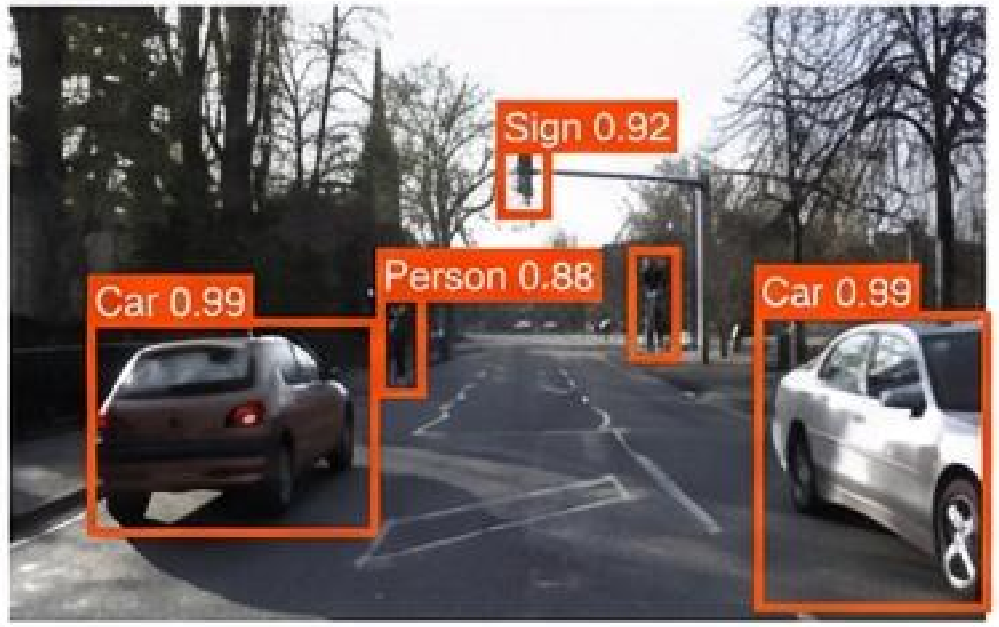
The Evolution of Accuracy (2010–2017)

Deep architectures drove error rates below the human threshold.



Applications: Beyond Simple Classification

Object Detection



Semantic Segmentation

Image Captioning

A man riding a large elephant.

Identifying WHAT and WHERE (Bounding Boxes).

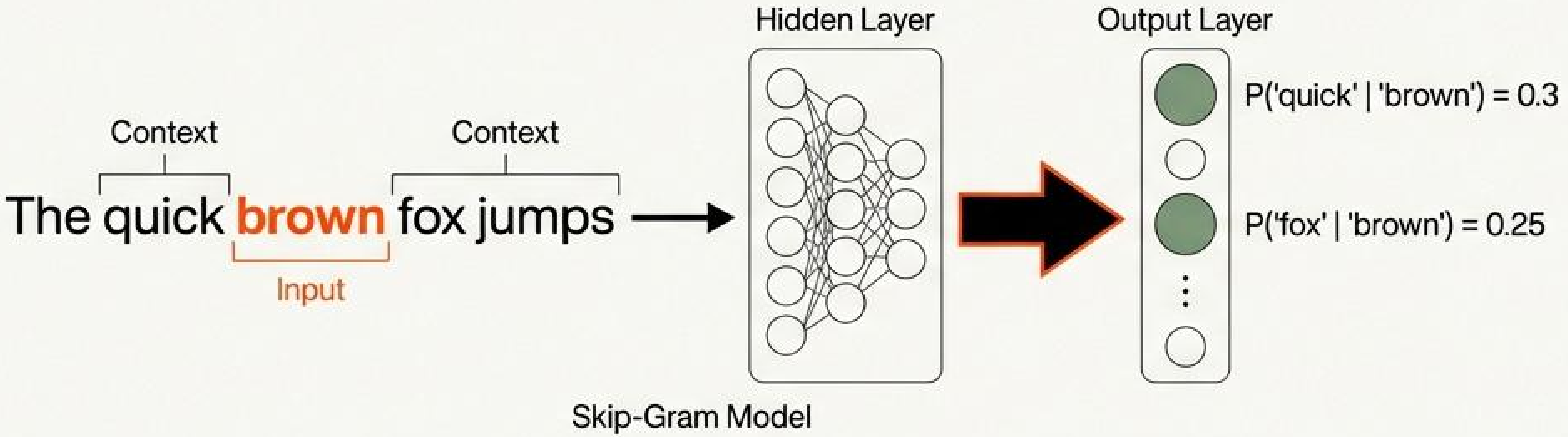
Classifying every single pixel (Context).

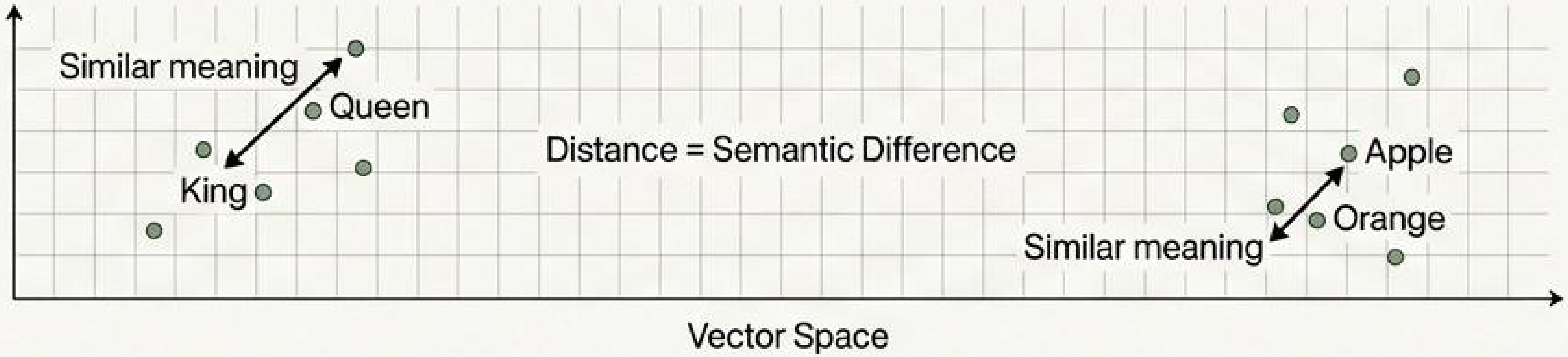
Generating textual descriptions from visual inputs.

NLP: Word Embeddings

Translating language into mathematics (Word2Vec).

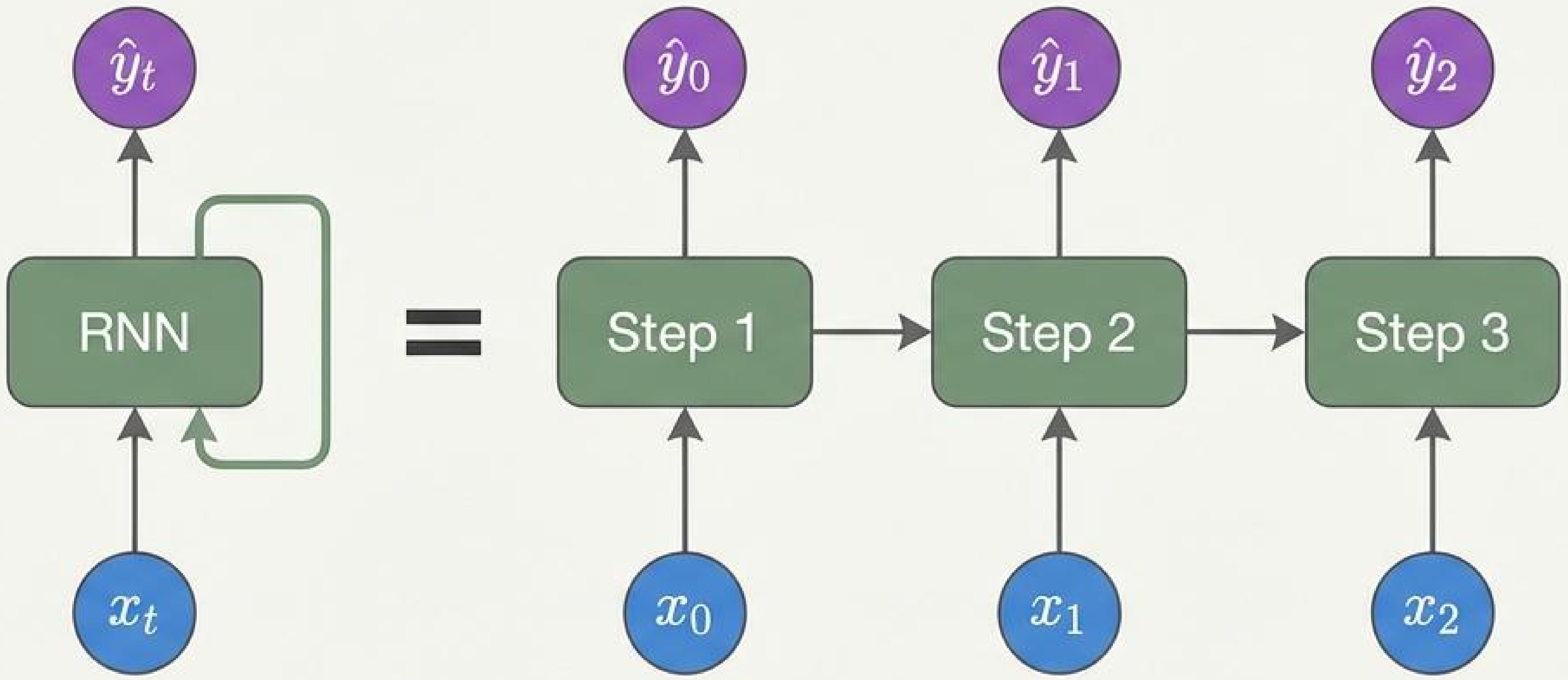
Words are converted into vectors. In this mathematical space, words with similar meanings share similar coordinates. The model learns these relationships by predicting context.





Recurrent Neural Networks (RNNs)

Processing sequences where order matters.



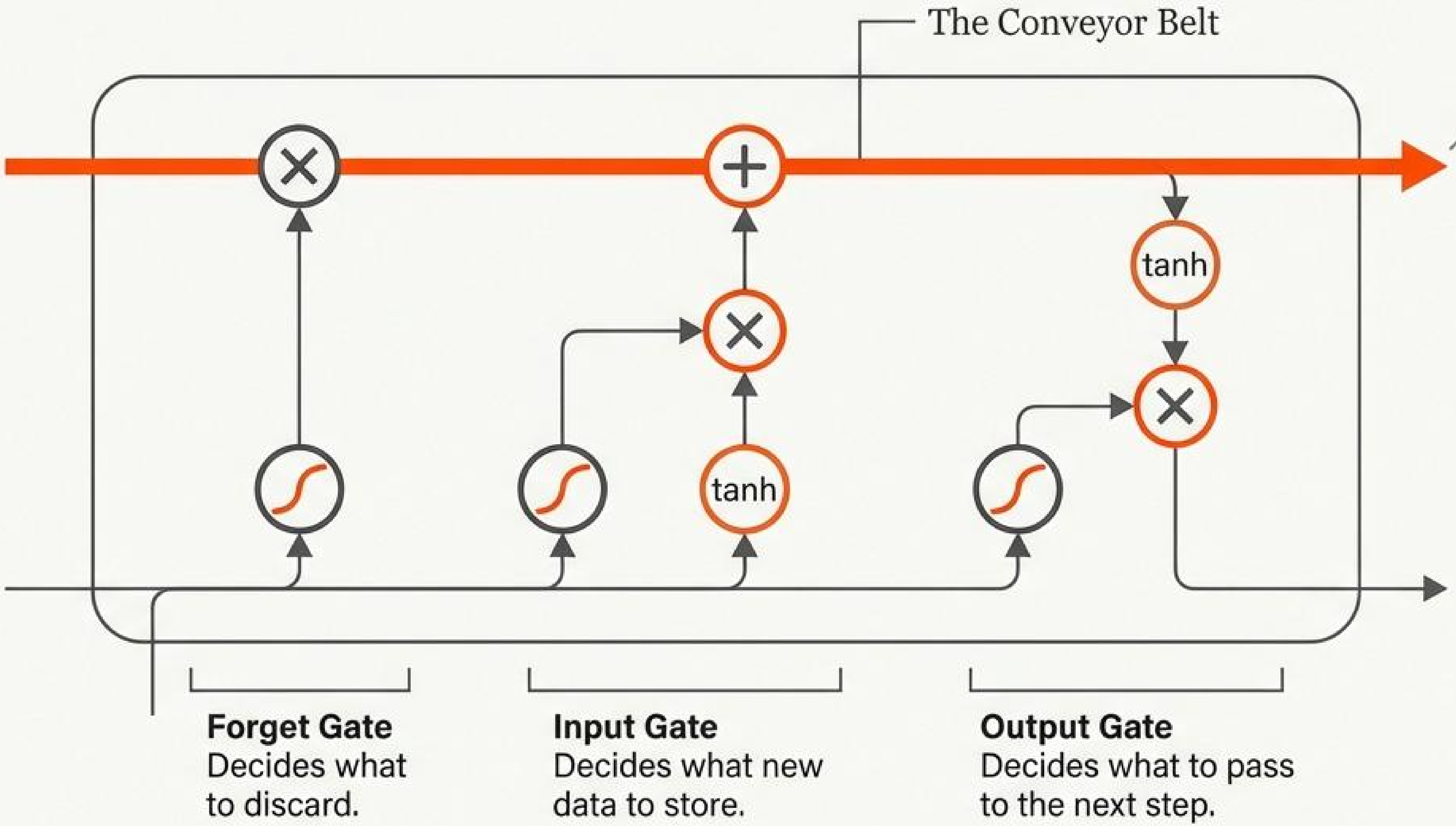
Standard networks have no memory. RNNs introduce a loop, allowing information to persist.

Ideally suited for:

- Time-series data (Stock markets)
- Speech Recognition
- Text prediction (where the next word depends on the previous).

Long Short-Term Memory (LSTM)

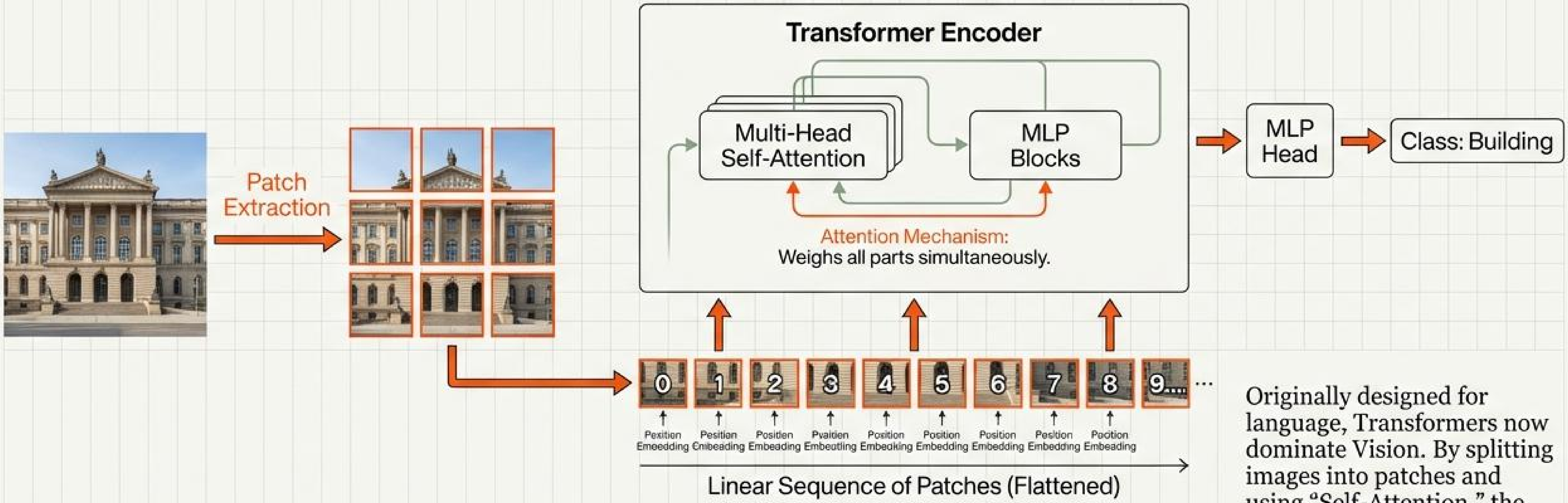
Solving the memory loss problem.



The LSTM acts like a manager for the data conveyor belt. It selectively forgets irrelevant history and adds important new context, allowing it to remember dependencies over long sequences.

The Transformer Revolution

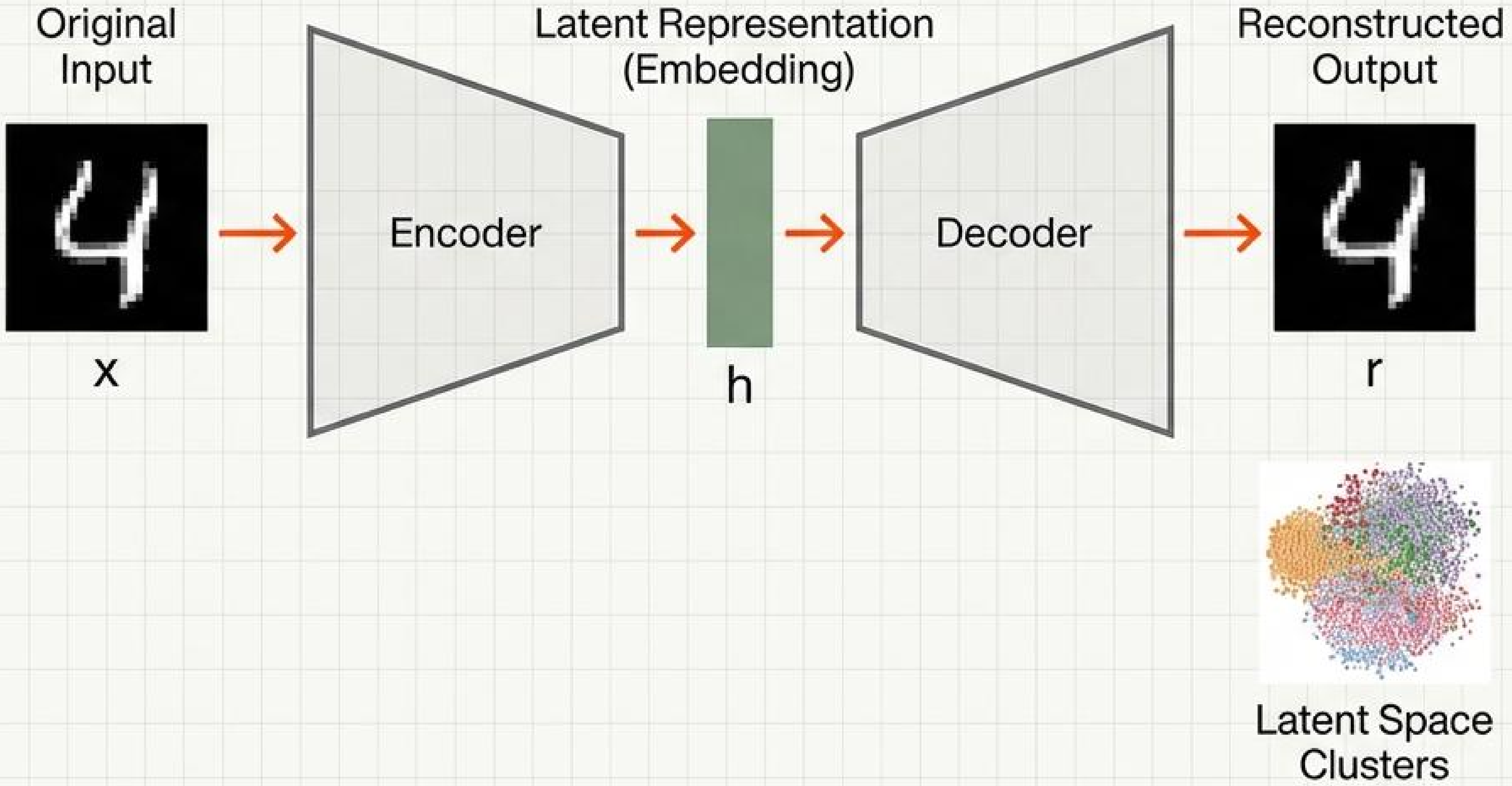
Vision Transformers (ViT) and Self-Attention.



Originally designed for language, Transformers now dominate Vision. By splitting images into patches and using "Self-Attention," the model weighs the importance of every part of the image simultaneously, rather than scanning it pixel by pixel.

Generative AI: Autoencoders

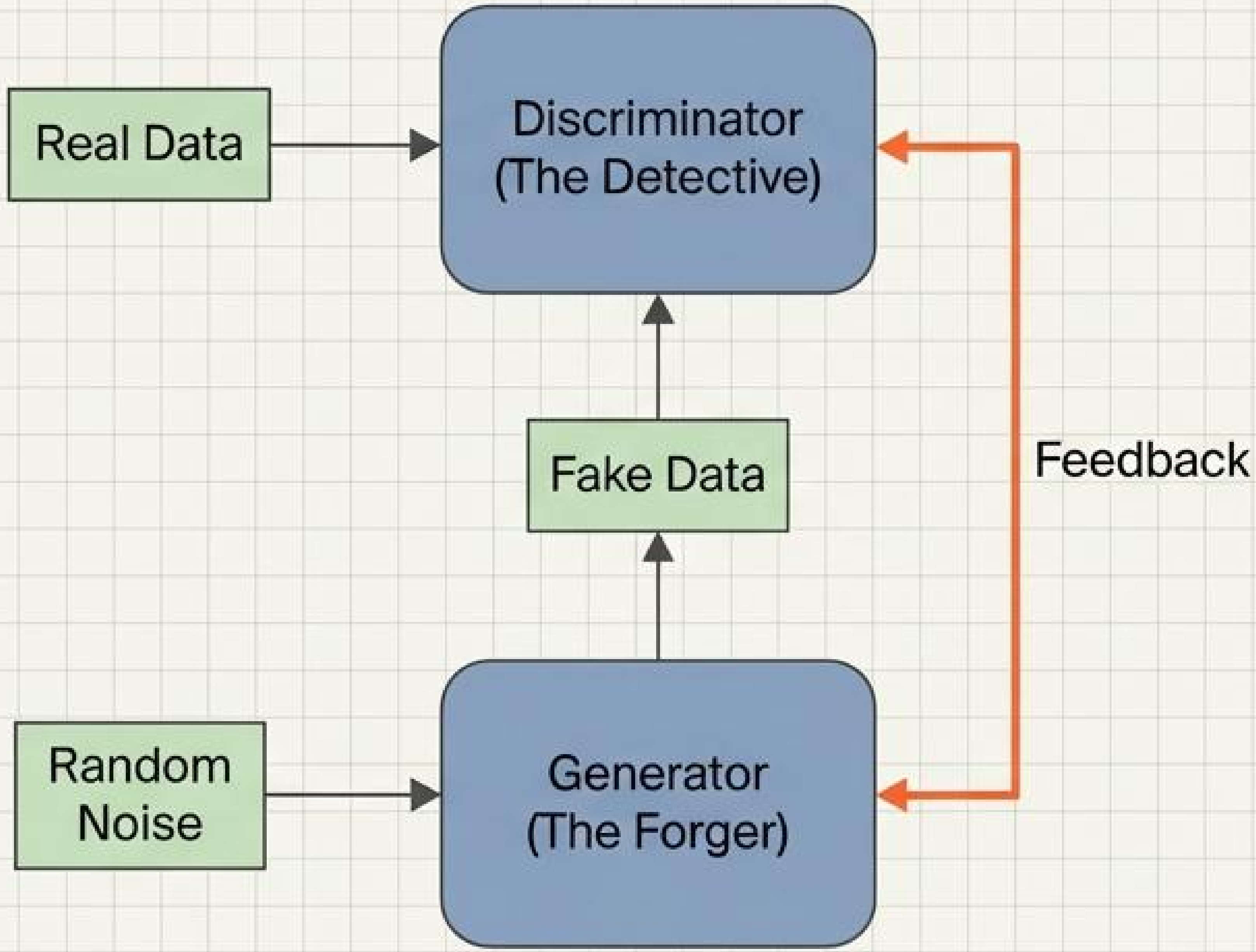
Learning through compression.



Autoencoders are unsupervised. They learn to compress data into efficient embeddings and then reconstruct it. This is fundamental for *denoising* and *dimensionality reduction*.

Generative Adversarial Networks (GANs)

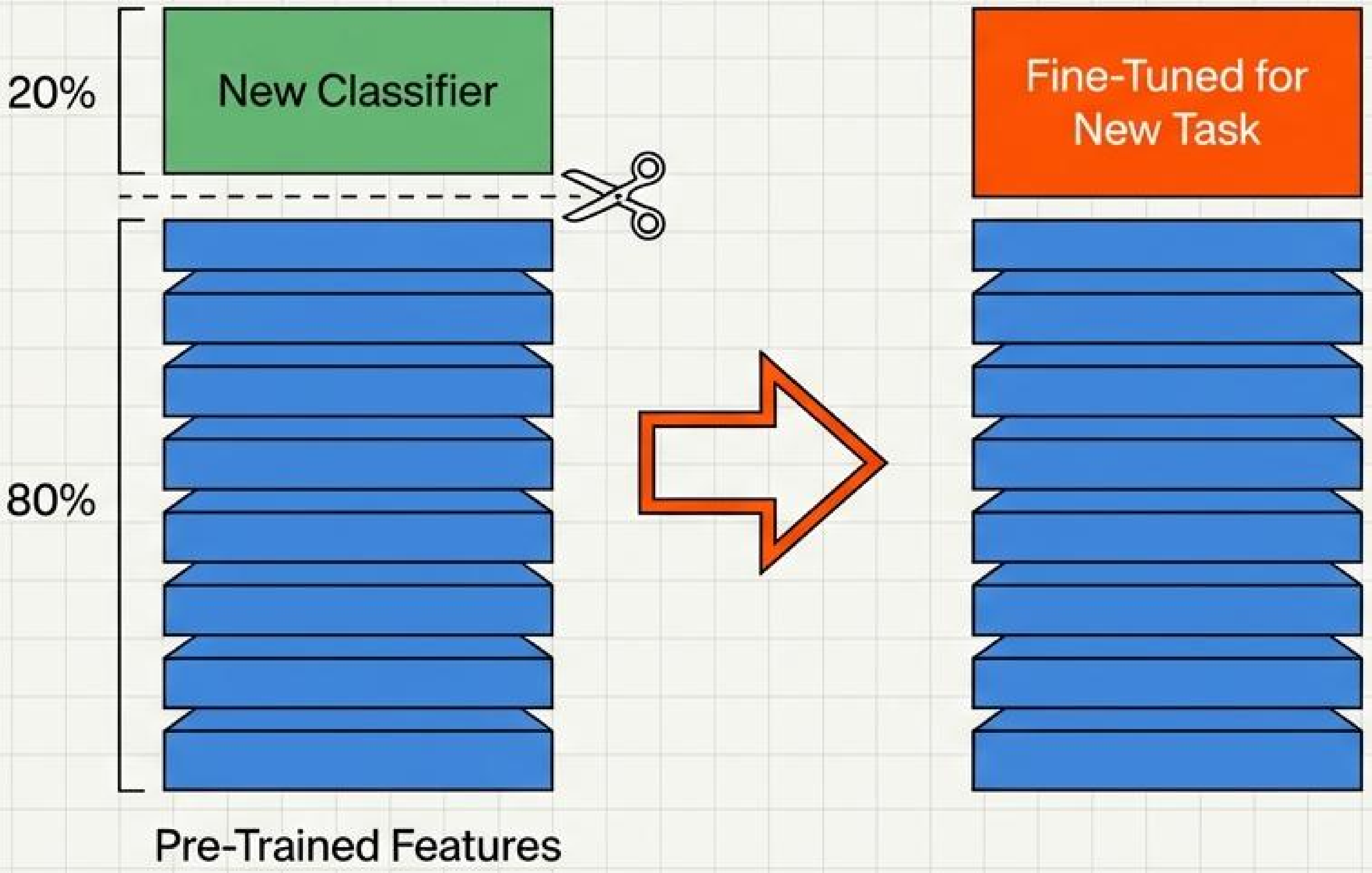
Creativity through competition.



Two networks compete. The Generator creates fakes to fool the Discriminator. The Discriminator learns to spot the fakes. This arms race results in hyper-realistic synthetic data.

Efficiency: Transfer Learning

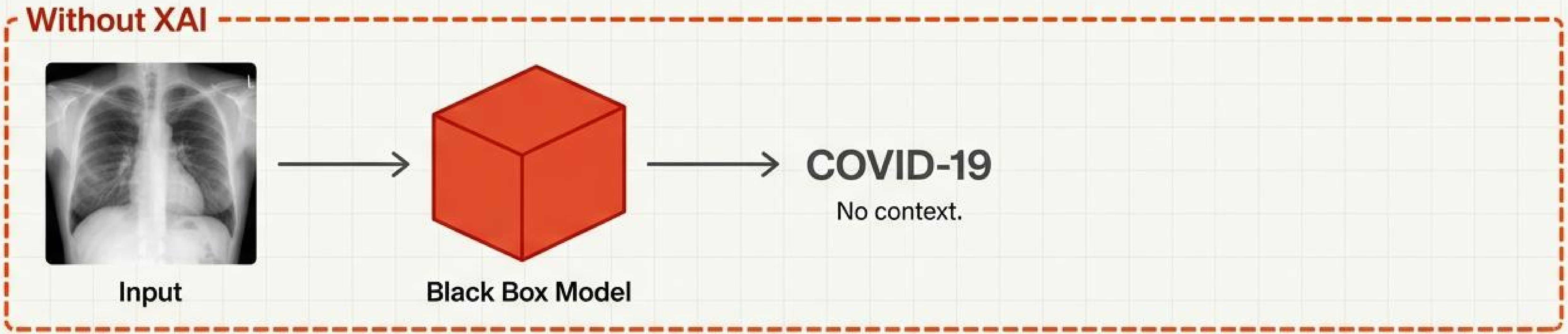
Standing on the shoulders of giants.

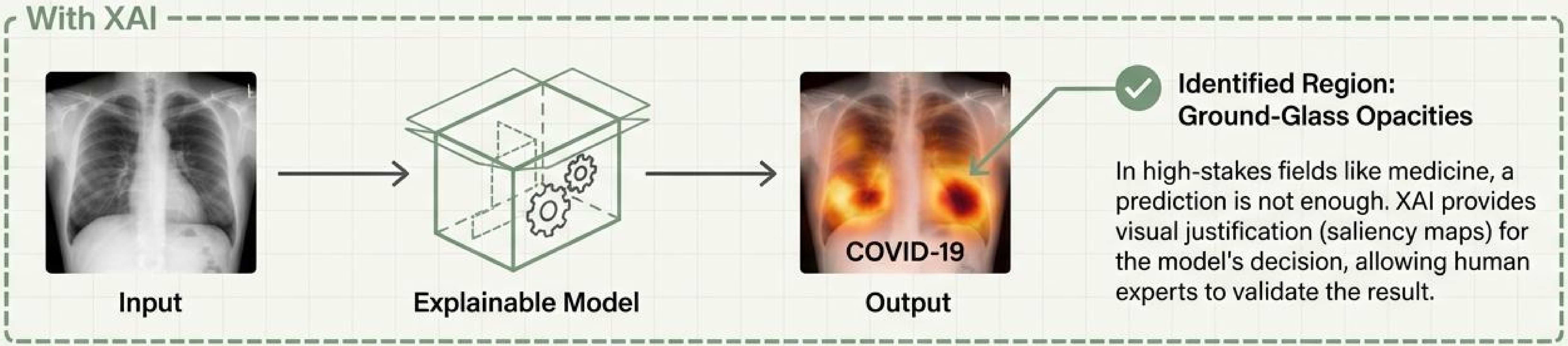


Training a model from scratch is expensive. Transfer Learning takes a model pre-trained on massive datasets (like ImageNet), “freezes” the feature-detection layers, and only retrains the final classification layers for a specific new task.

Trust: Explainable AI (XAI)

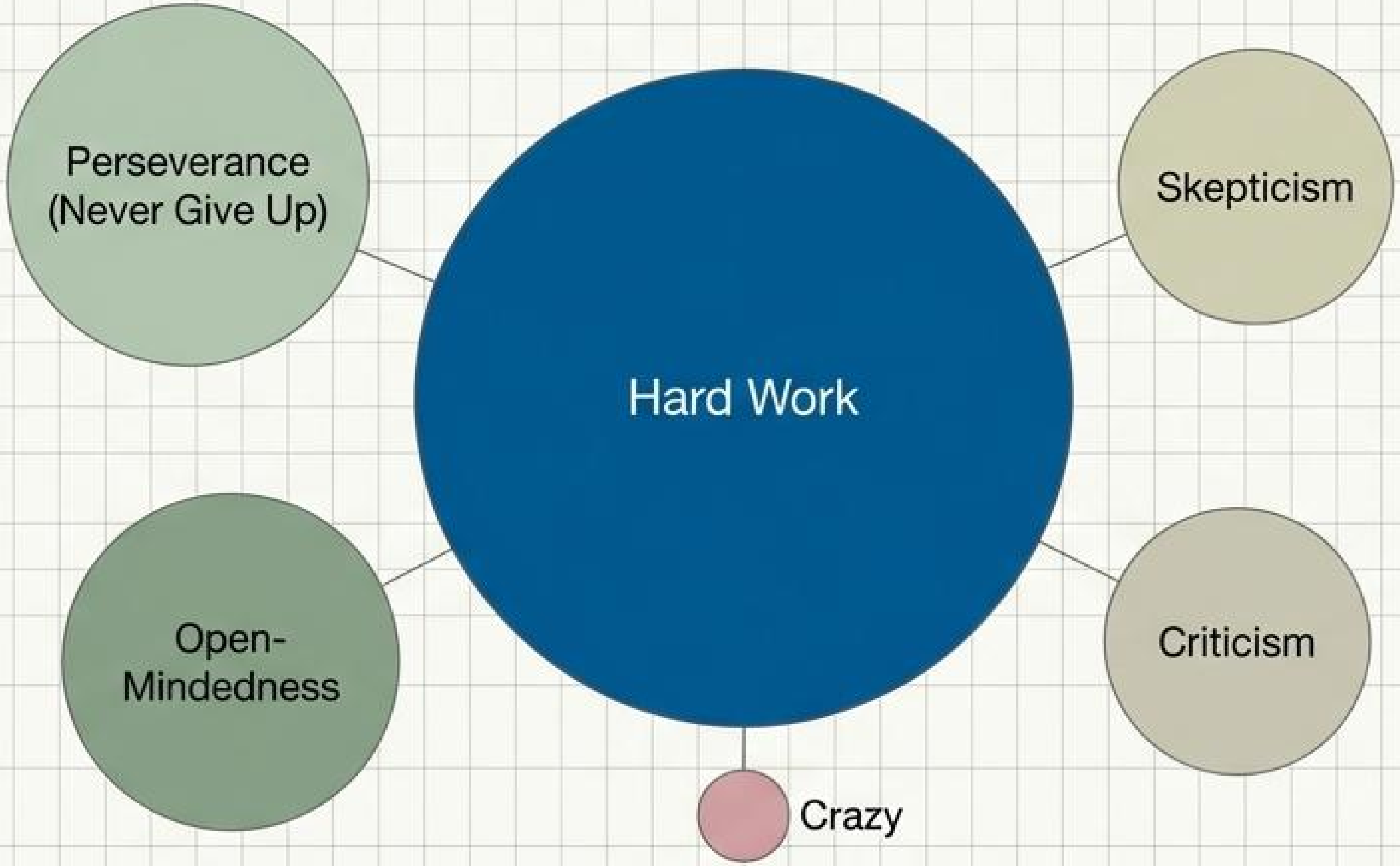
Opening the Black Box.





In high-stakes fields like medicine, a prediction is not enough. XAI provides visual justification (saliency maps) for the model's decision, allowing human experts to validate the result.

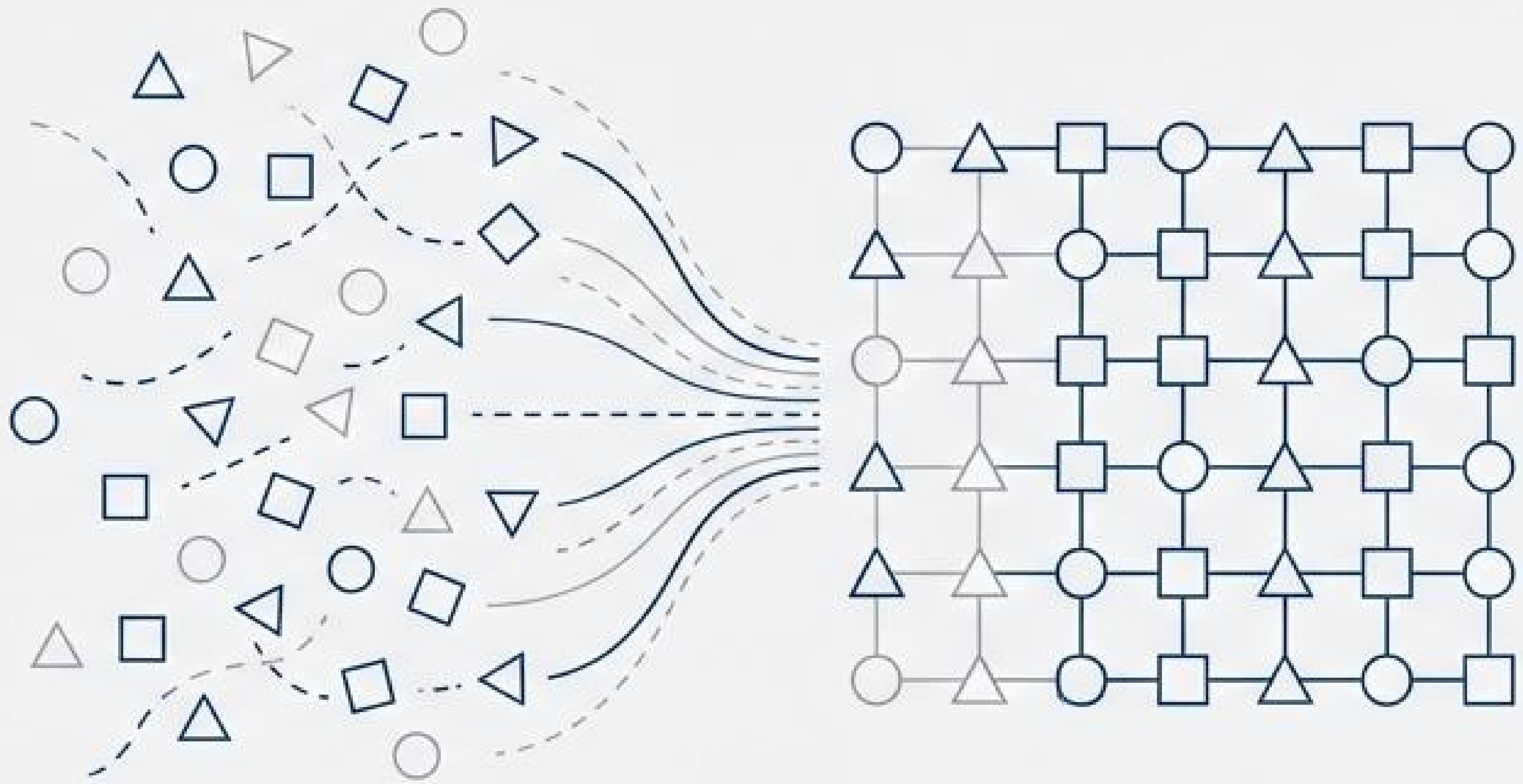
The Recipe for Progress



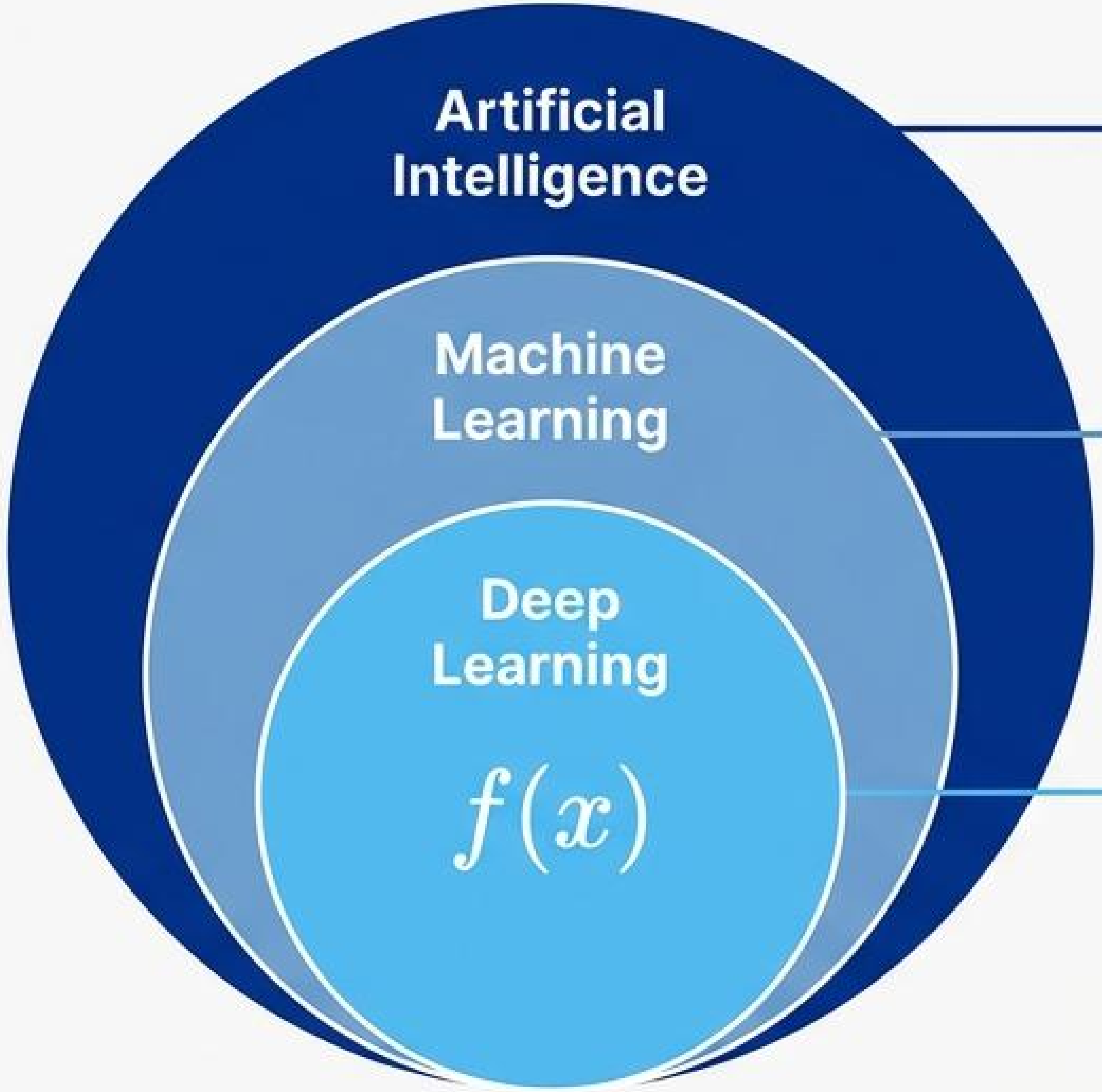
"The future depends on some graduate student who is deeply suspicious of everything I have said."
— Geoffrey Hinton

Introduction to Machine Learning

From Traditional Computing
to Intelligent Systems



The Hierarchy of Intelligence



ARTIFICIAL INTELLIGENCE

A technique which enables machines to mimic human behaviour.

MACHINE LEARNING

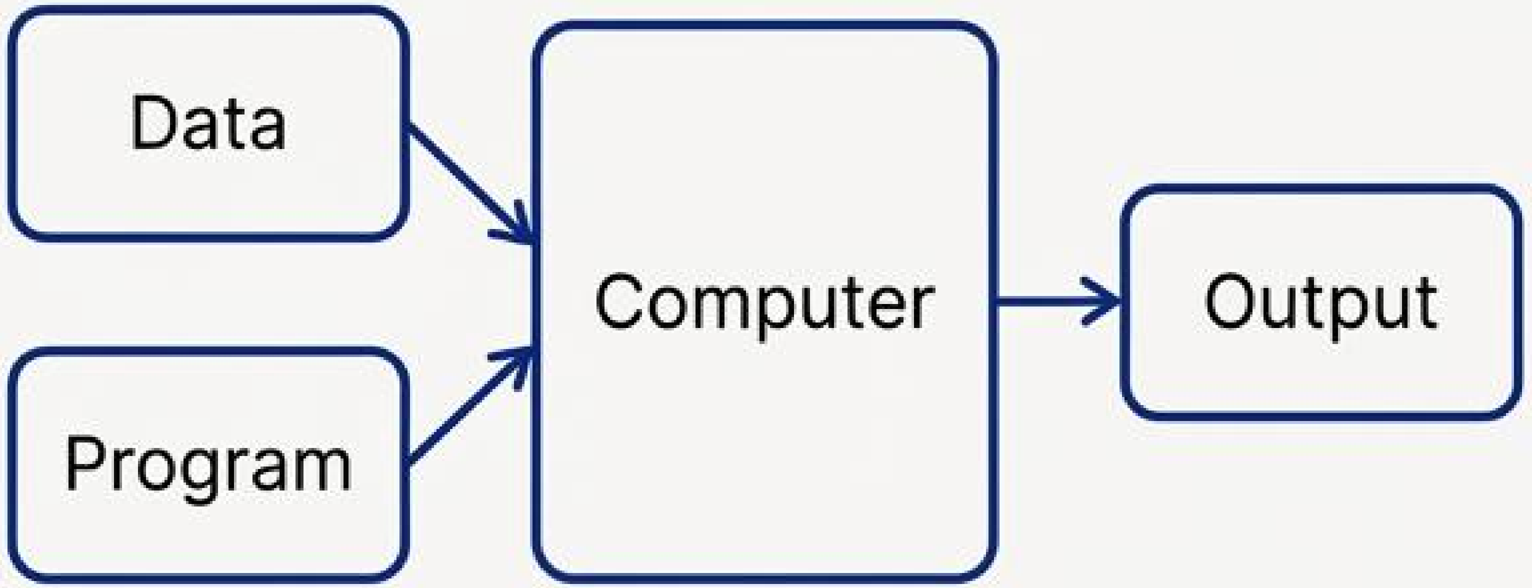
A subset of AI techniques which use statistical methods to enable machines to improve with experience.

DEEP LEARNING

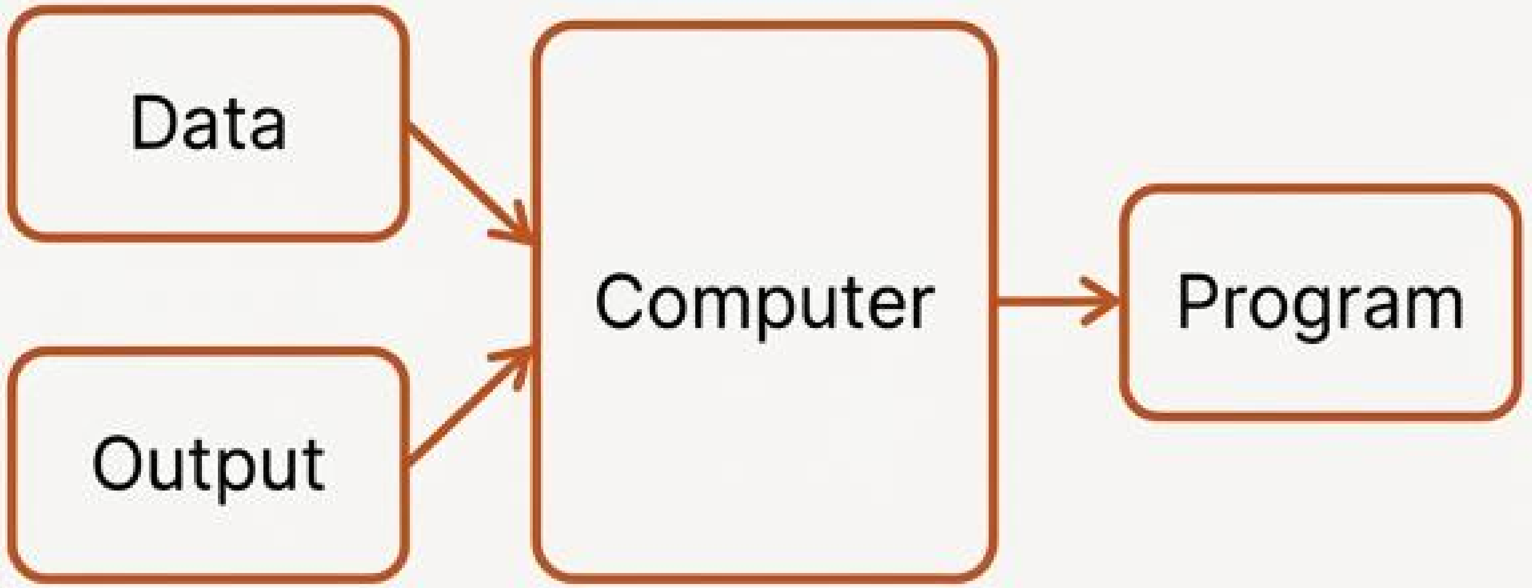
A subset of ML which makes the computation of multi-layer neural networks feasible.

The Paradigm Shift

Traditional Programming



Machine Learning



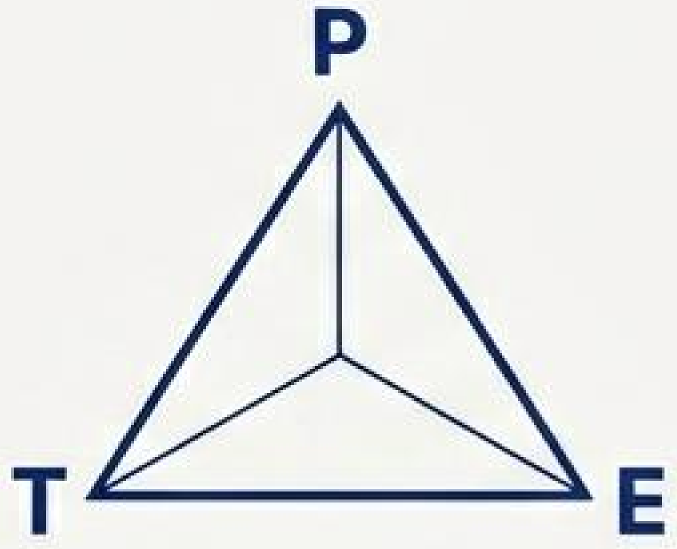
Insight: In traditional programming, we code the rules.
In Machine Learning, the system learns the rules.

Defining the Learning Process

“Learning is any process by which a system improves performance from experience.” — Herbert Simon

---

“Machine Learning is the study of algorithms that improve their performance **P** at some tasks **T** with experience **E**.” — Tom Mitchell (1998)



The Learning Triad

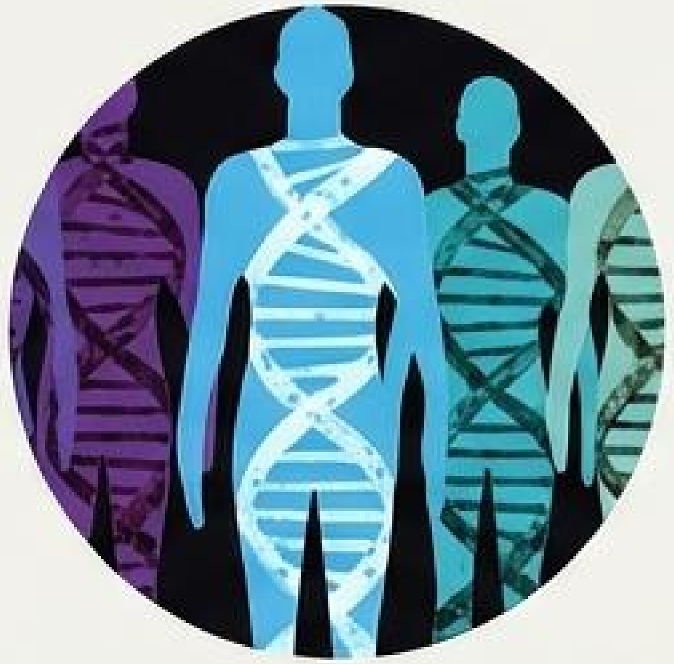
The Framework in Practice

	Task (T)	Performance (P)	Experience (E)
	Playing checkers.	Percentage of games won against an arbitrary opponent.	Playing practice games against itself.
	Recognizing hand-written words.	Percentage of words correctly classified.	Database of human-labeled images.
	Driving on four-lane highways using vision sensors.	Average distance traveled before a human-judged error.	Sequence of images and steering commands recorded from human drivers.
	Categorise email as spam or legitimate.	Percentage classified correctly.	Database of emails with human-given labels.

When Do We Use Machine Learning?

Human expertise does not exist. (e.g. Navigating on Mars).

Humans cannot explain their expertise. (e.g. Speech recognition).



Need for customization. (e.g. Personalized medicine).

Scale of data. (e.g. Models based on massive datasets like astronomical analysis).

The Taxonomy of Learning

- **Supervised (Inductive) Learning**
  Given training data + desired outputs (labels).
- **Unsupervised Learning**
  Given training data (without desired outputs).
- **Semi-supervised Learning**
  Given training data + a few desired outputs.
- **Reinforcement Learning**
  Learning via rewards from a sequence of actions.
- **Association Analysis**
  Based on pattern mining.

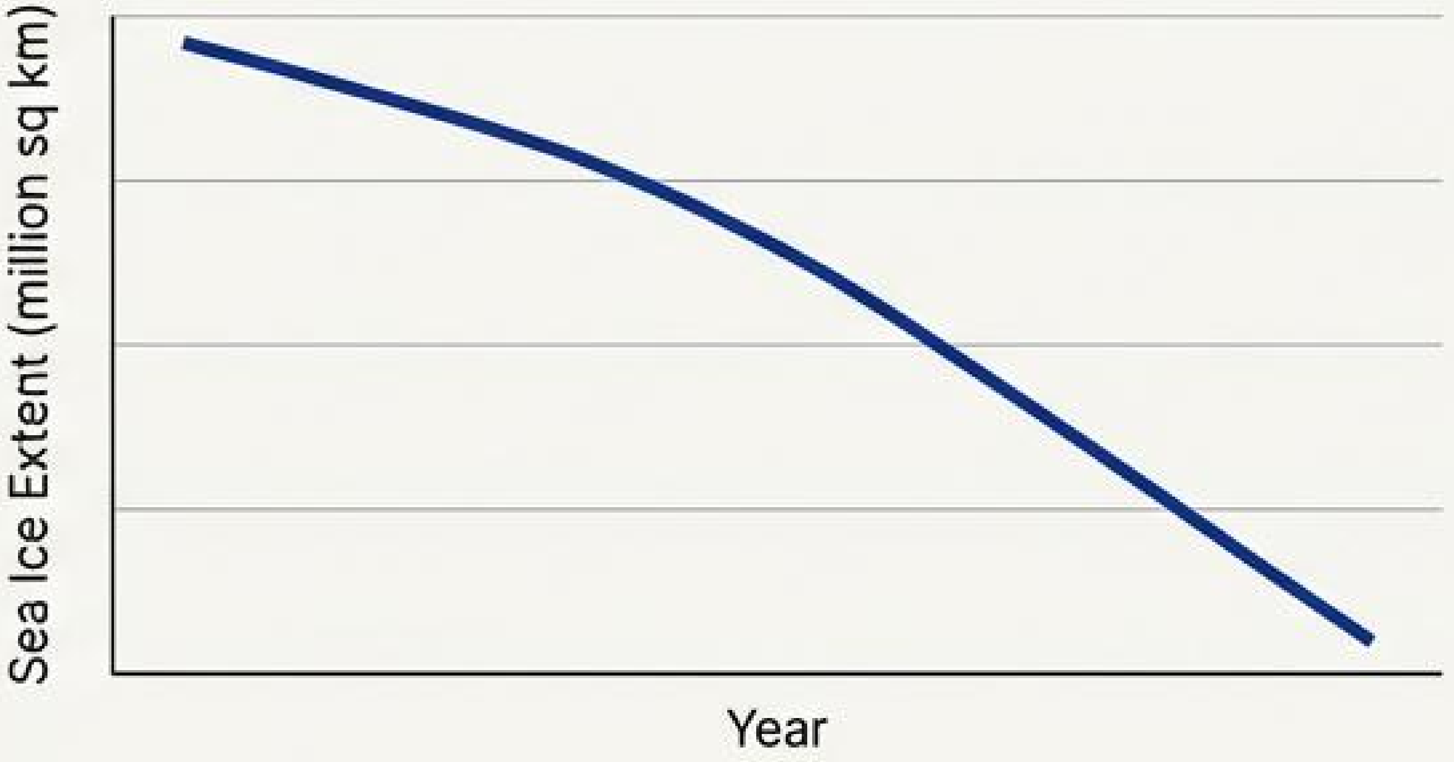
Supervised Learning: Learning with Labels

Core Concept: Learn a function f(x) to predict y given x.

Type 1: Regression

y is real-valued.

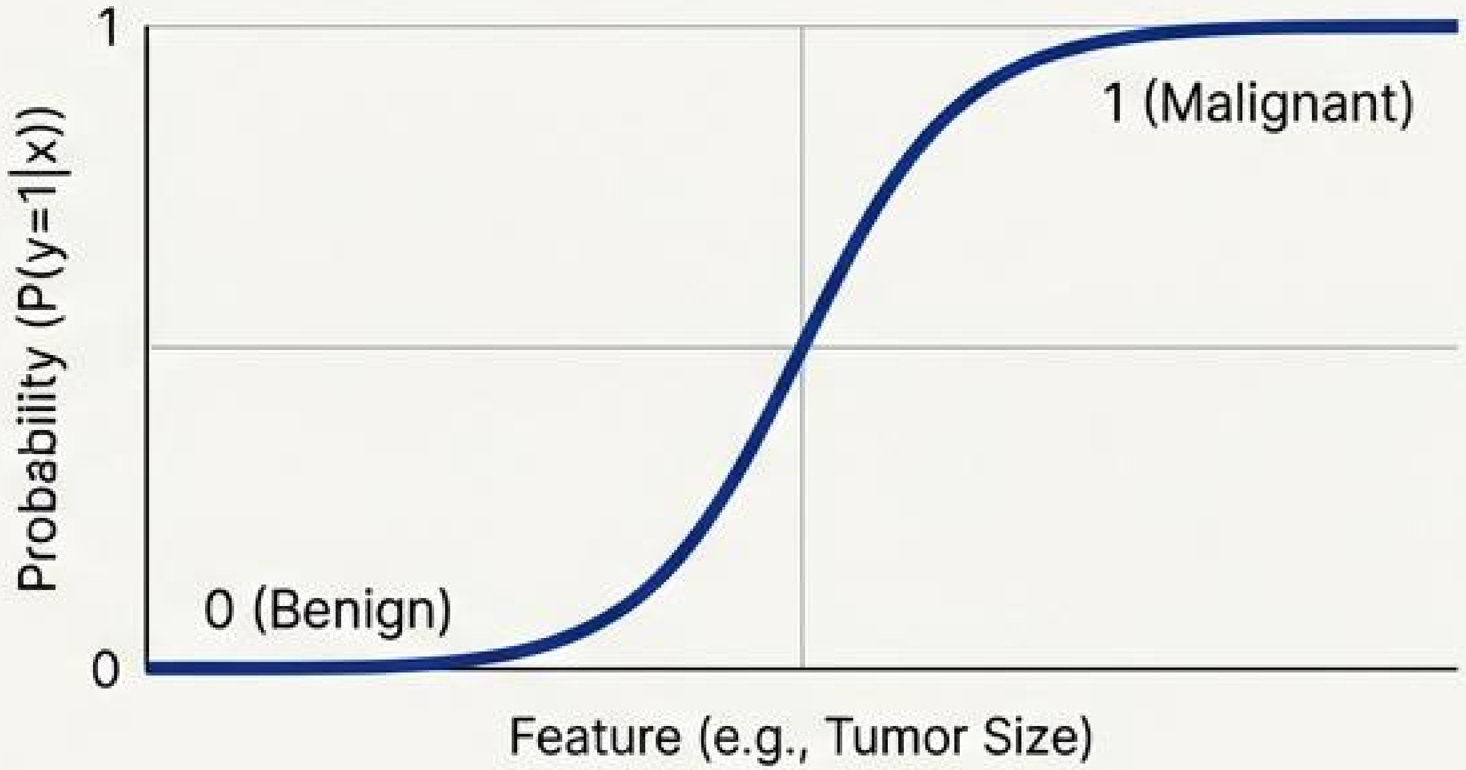
Predicting Arctic Sea Ice Extent over time.



Type 2: Classification

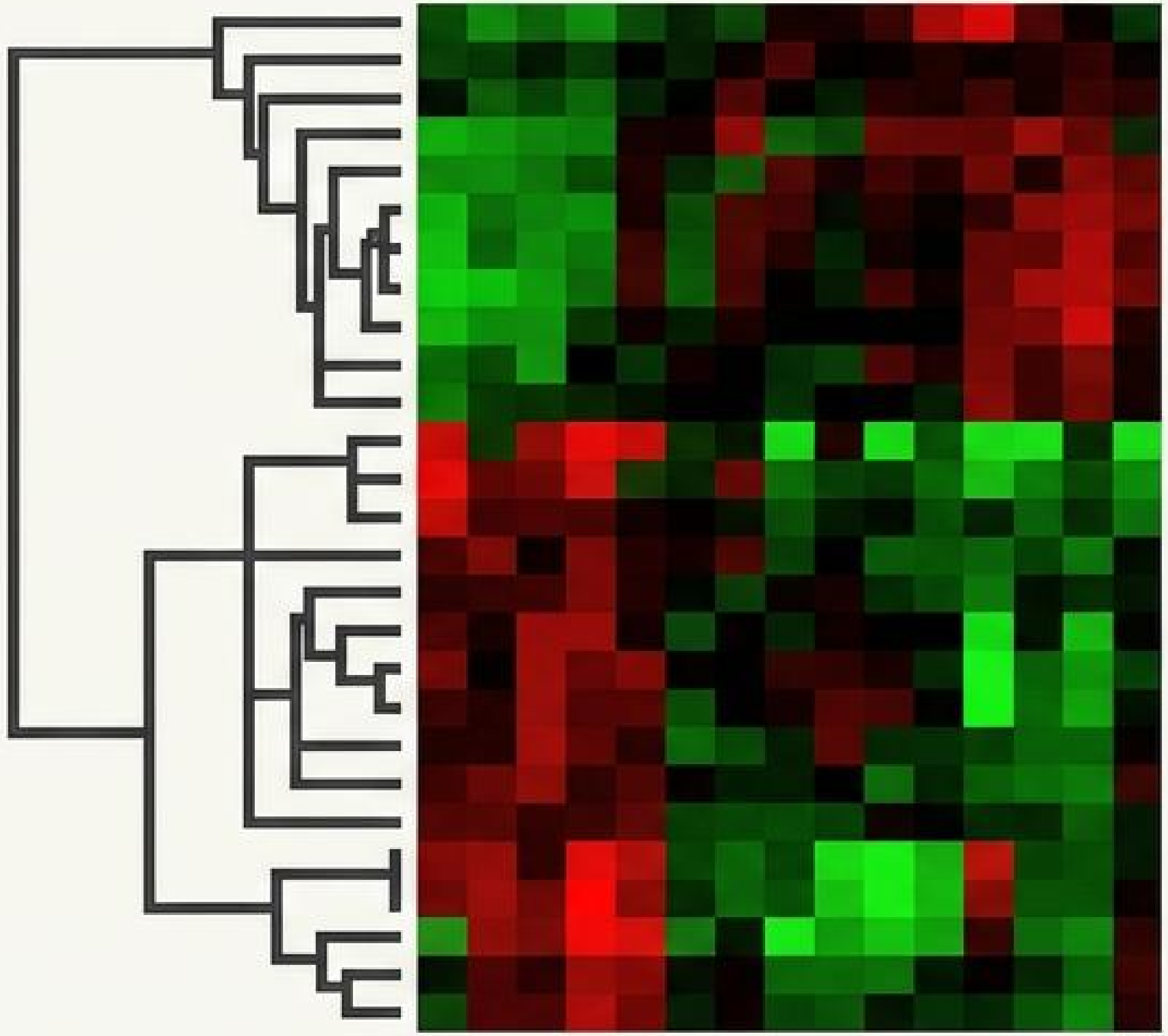
y is categorical.

Breast Cancer diagnosis (Malignant/Benign).



Unsupervised Learning: Finding Hidden Structure

Core Concept: Given inputs **x** (without labels), output the hidden structure.



Document Retrieval: Grouping similar news stories.

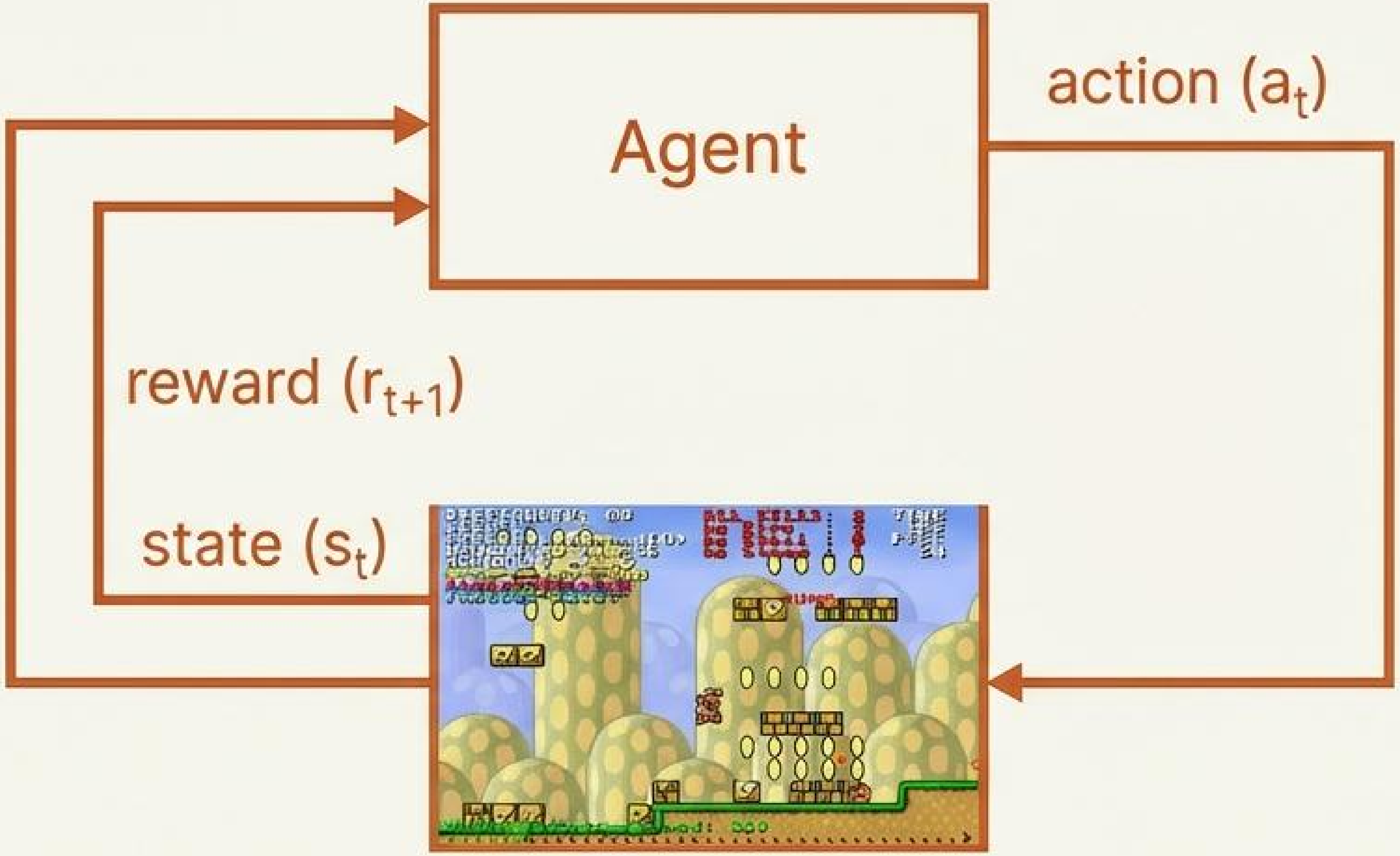
Genomics: Analyzing gene expression data.

Market Segmentation: Grouping customers by behaviour.

Astronomy: Analyzing star/galaxy data.

Reinforcement Learning

Learning through interaction.



Goal: Output a **Policy**—a mapping from states to actions that determines what to do.

Examples: Robot in a maze, Game playing, Balancing a pole.

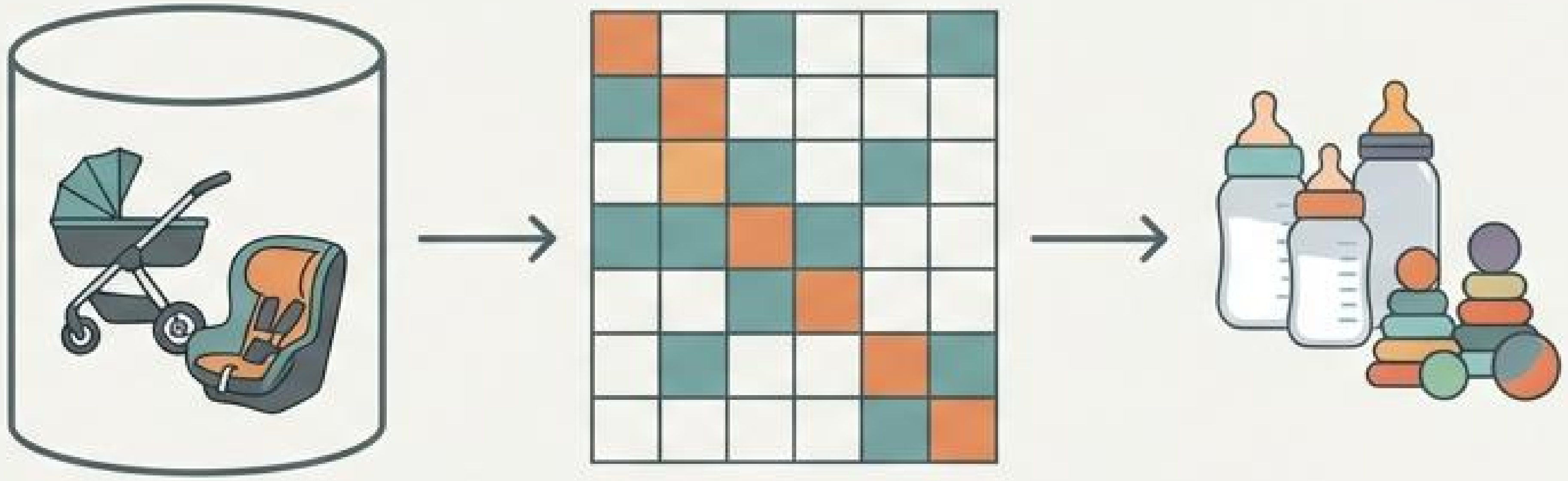
Association Analysis & Recommender Systems

Discovering relationships in transaction data.

$$P(Y \mid X)$$

The probability that someone who buys X also buys Y.

If $X = \text{Cold Drink}$, then $P(\text{Bread} \mid \text{Cold Drink}) = 0.7$.



Data
Your past purchases

ML Method

Intelligence
Recommended items

Market Basket Analysis: “People who bought diapers also bought beer.”

Classi Classification Complexity

Multi-Class

Mutually Exclusive Categories

Example: An image is either a Sun, a Moon, OR a Cloud.

Label: [1 0 0] (One-hot encoding)

Multi-Label

Non-Exclusive Categories

Example: An image can contain a Sun AND a Cloud.

Label: [1 0 1]

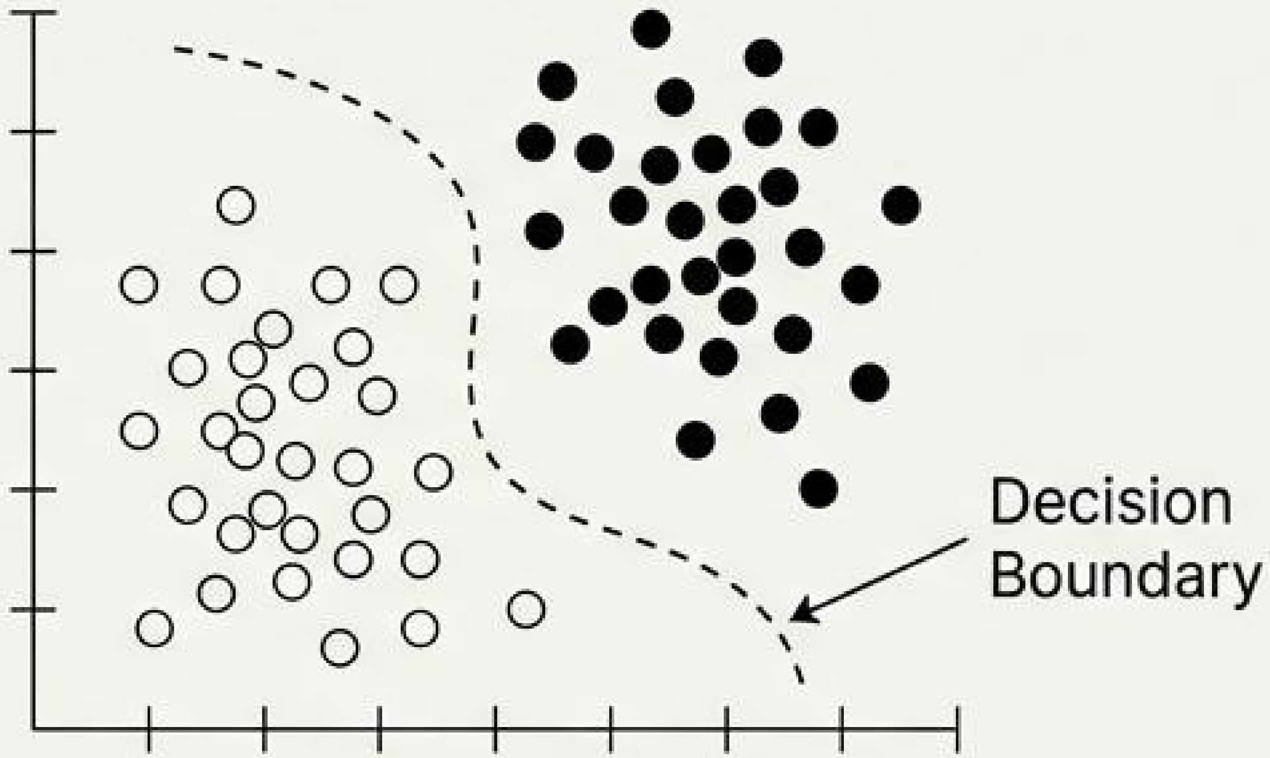
A History of Algorithms: Statistical Roots

Probabilistic Modelling

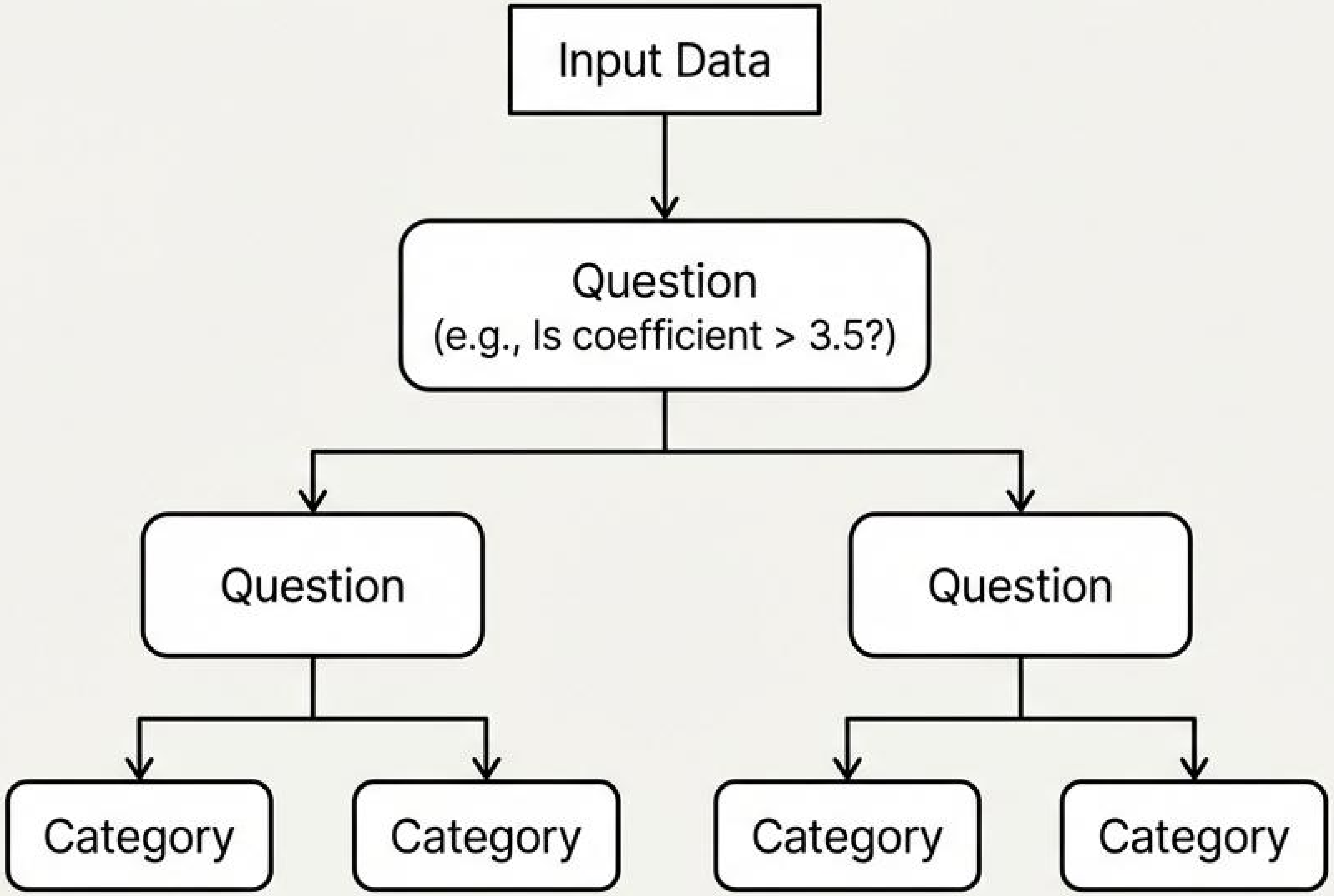
- **Naive Bayes:** Assumes features are independent.
- **Logistic Regression:** The ‘Hello World’ of modern machine learning.

Kernel Methods (SVM)

- **Support Vector Machines (SVM):** Finds the best decision boundary separating two categories.



A History of Algorithms: Decision Trees

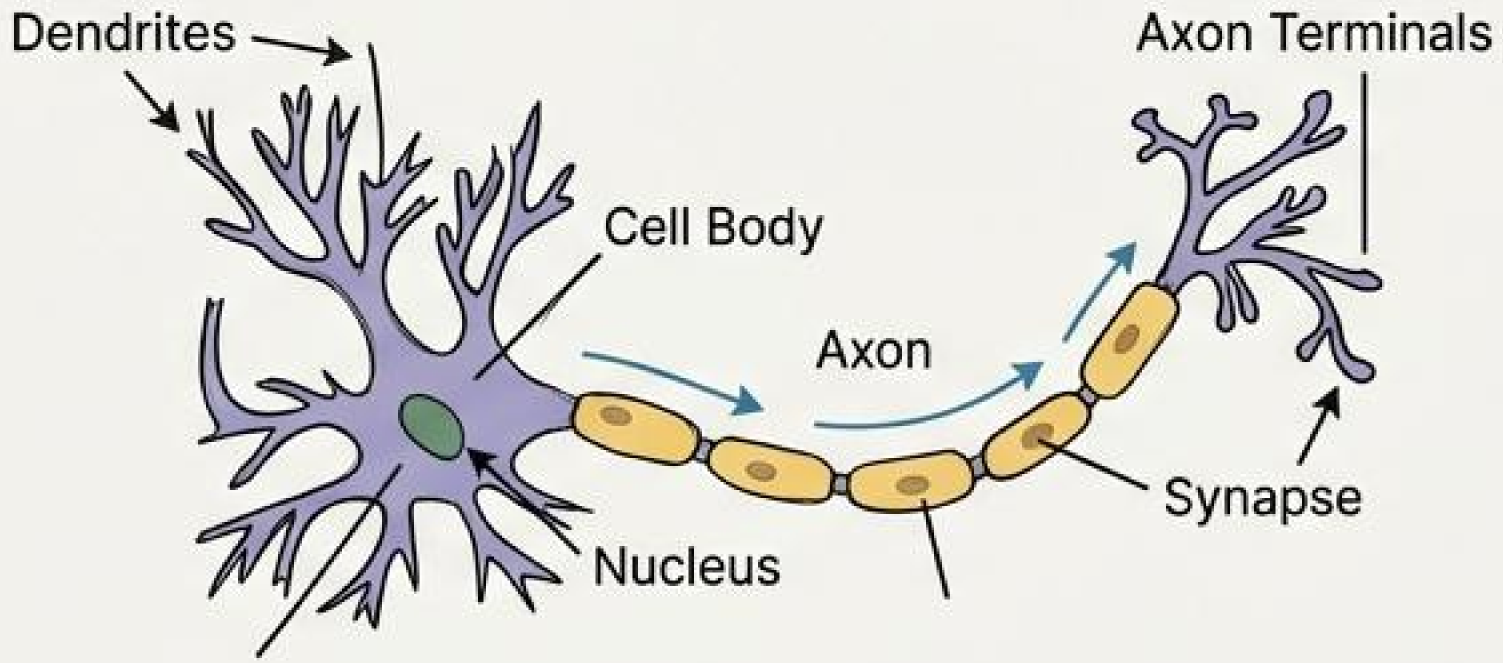


Decision Trees: Flowchart-based logic that learns questions to split data.

Evolution: Single trees led to powerful ensemble methods like **Random Forests** and **Gradient-Boosting Machines**.

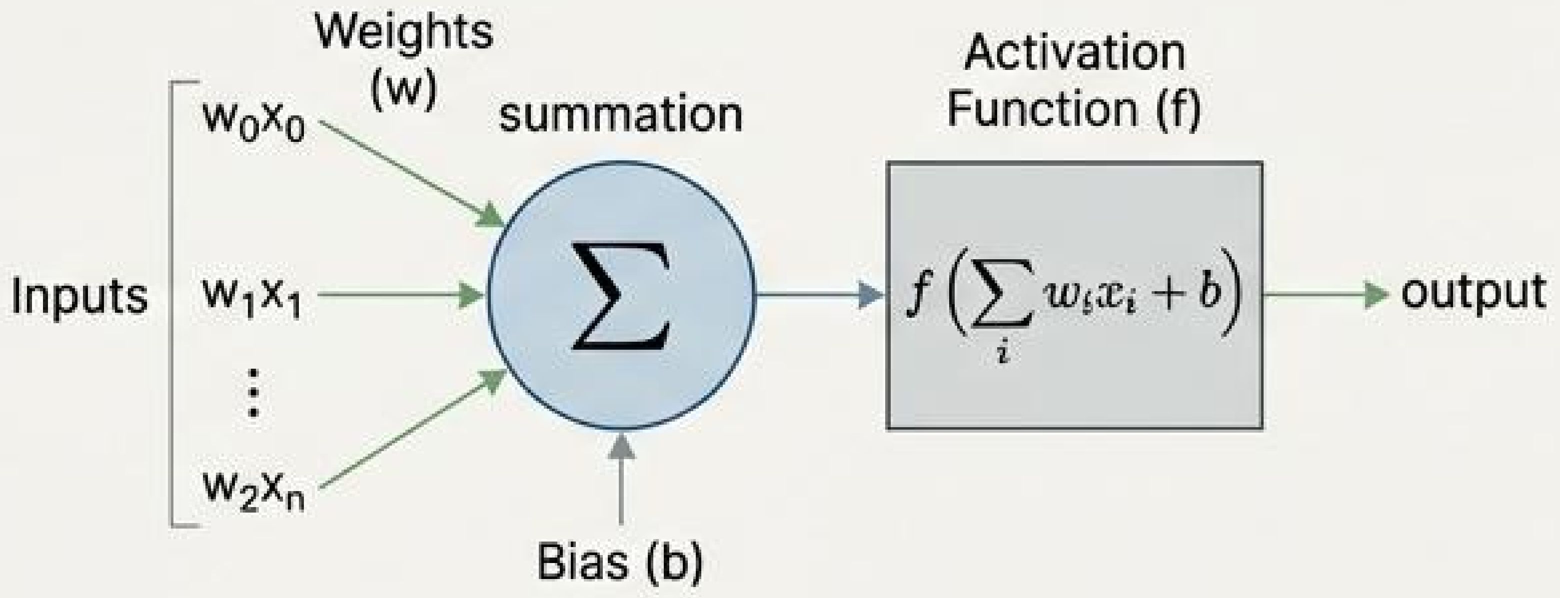
Biological Inspiration: Neural Networks

Biological Neuron



- Async topology
- Low power consumption
- Massive connections (~10,000,000x)
- Continuous learning

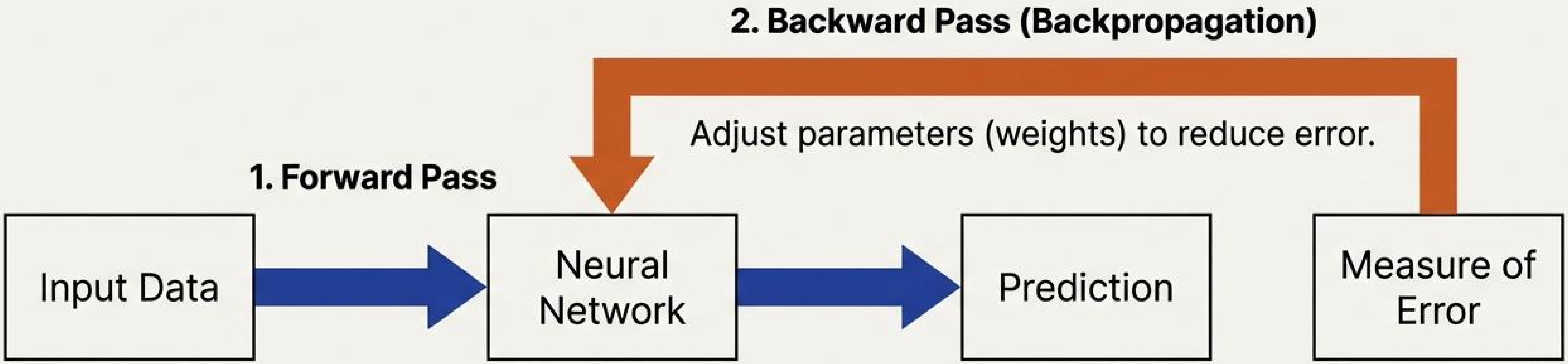
Artificial Neuron



- Sync topology (Layers)
- High power consumption
- Simplified connectivity
- Train then test

The Modern Era: Backpropagation

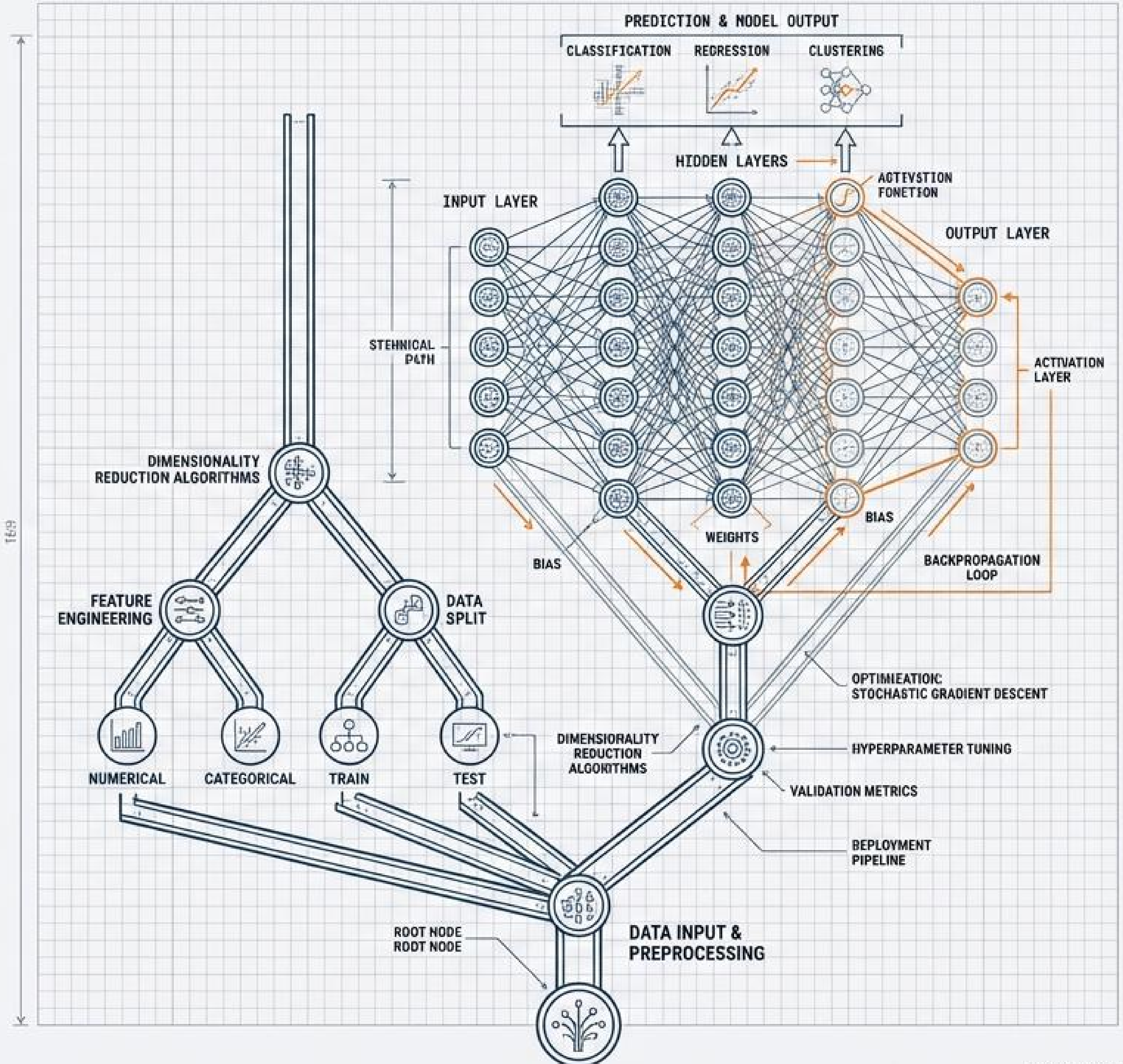
The mechanism that powers the Deep Learning revolution.



Mid-80s Breakthrough: This algorithm allowed us to train chains of parametric operations efficiently.

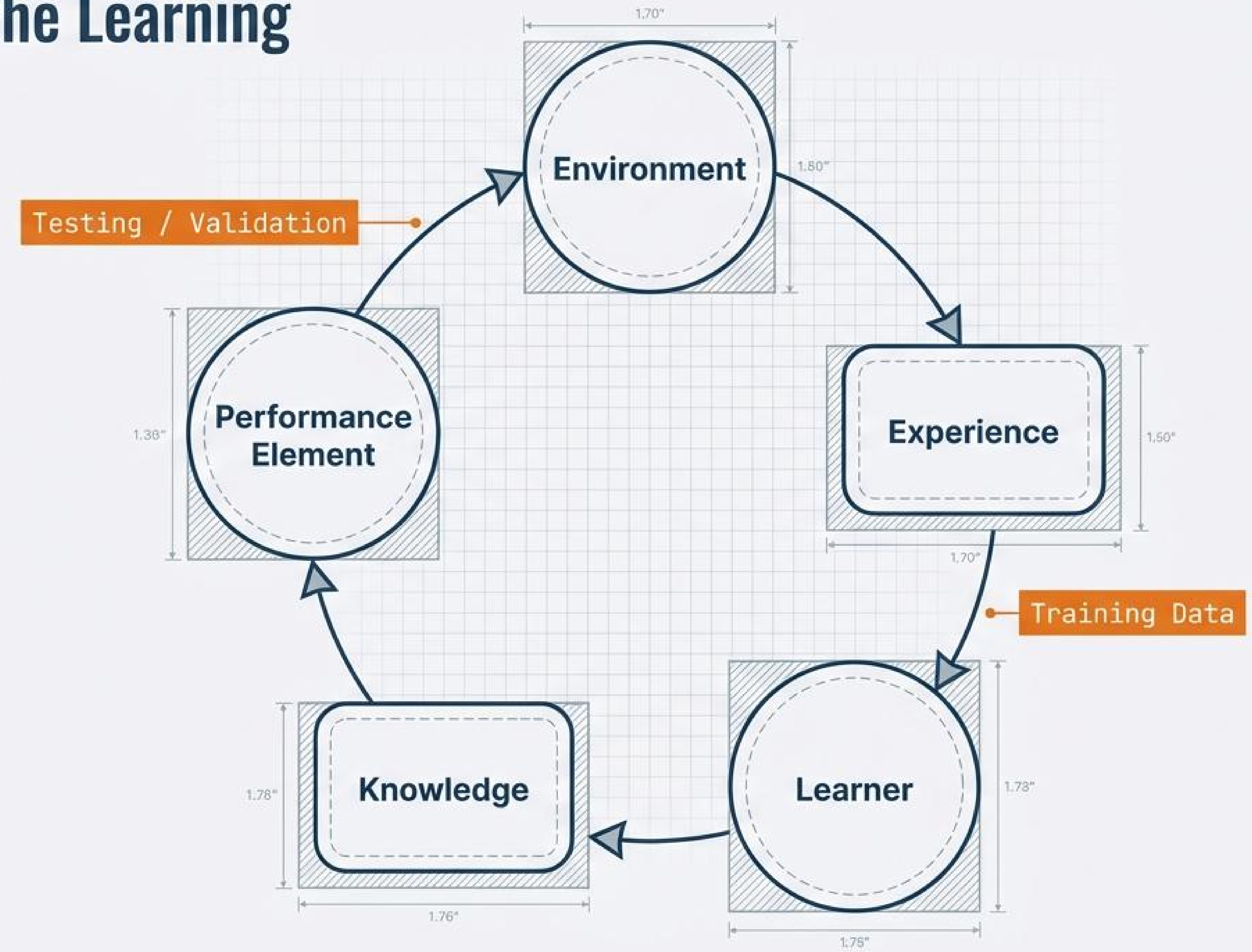
THE BLUEPRINT OF MACHINE LEARNING

Designing Systems, Core Components, and Practical Application



A structural guide to building functional learning engines: From conceptual design to operational execution.

Framing the Learning Problem



The Four Design Pillars

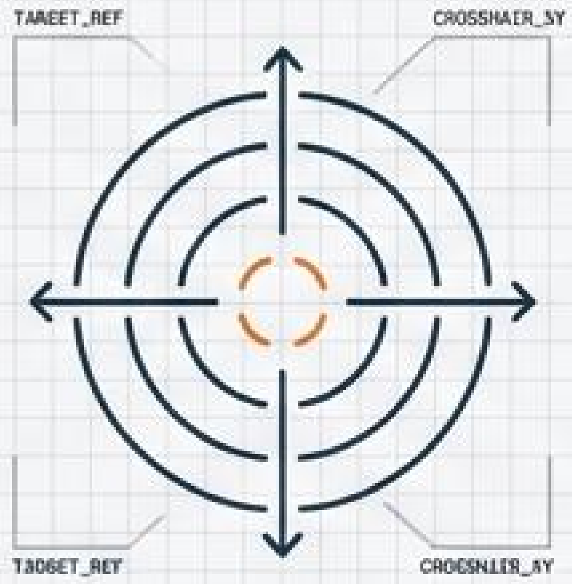
Every learning system requires four specific design choices to define its scope:

01. Training Experience

What data defines the environment?

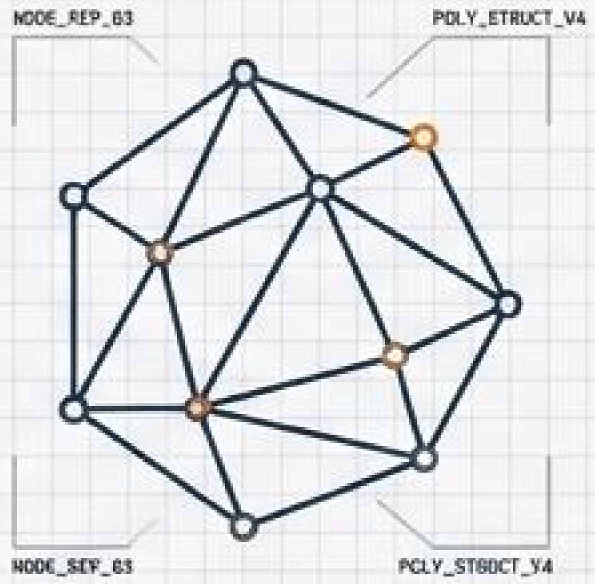
02. Target Function

Exactly what is to be learned?



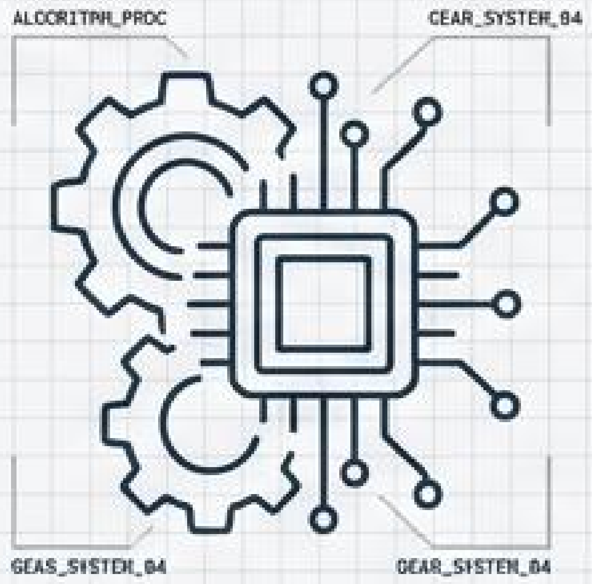
03. Representation

How do we mathematically describe this function?



04. Learning Algorithm

How do we infer the target function from the experience?



The Anatomy of an Algorithm

There are tens of thousands of machine learning algorithms, with hundreds of new ones created every year. Yet, they all share a common structure.

Component 1: Representation

How the system structures its view of the world.

Numerical Functions

- Linear regression
- Neural networks
- Support vector machines

Symbolic Functions

- Decision trees
- Rules in propositional logic
- Rules in first-order predicate logic

Instance-based Functions

- Nearest-neighbour
- Case-based

Probabilistic Graphical Models

- Naïve Bayes
- Bayesian networks
- Hidden-Markov Models (HMMs)
- Probabilistic Context-Free Grammar (PCFGs)
- Markov networks

Component 2: Optimisation

Inter The search methods used to infer the target function.

Gradient Descent → Perceptron, Backpropagation

Dynamic Programming → HMM Learning, PCFG Learning

Divide and Conquer → Decision tree induction, Rule learning

Evolutionary Computation → Genetic Algorithms (GAs), Genetic Programming (GP), Neuro-evolution

Component 3: Evaluation

Defining what “good” looks like.

EVALUATION METRICS DASHBOARD

- Accuracy
- Precision and Recall
- Squared Error
- Likelihood
- Posterior Probability

DATA_POINTS

SYS_EVAL_CONF24

- Cost / Utility
- Margin
- Entropy
- K-L Divergence

PERFORMANCE_INDICATORS

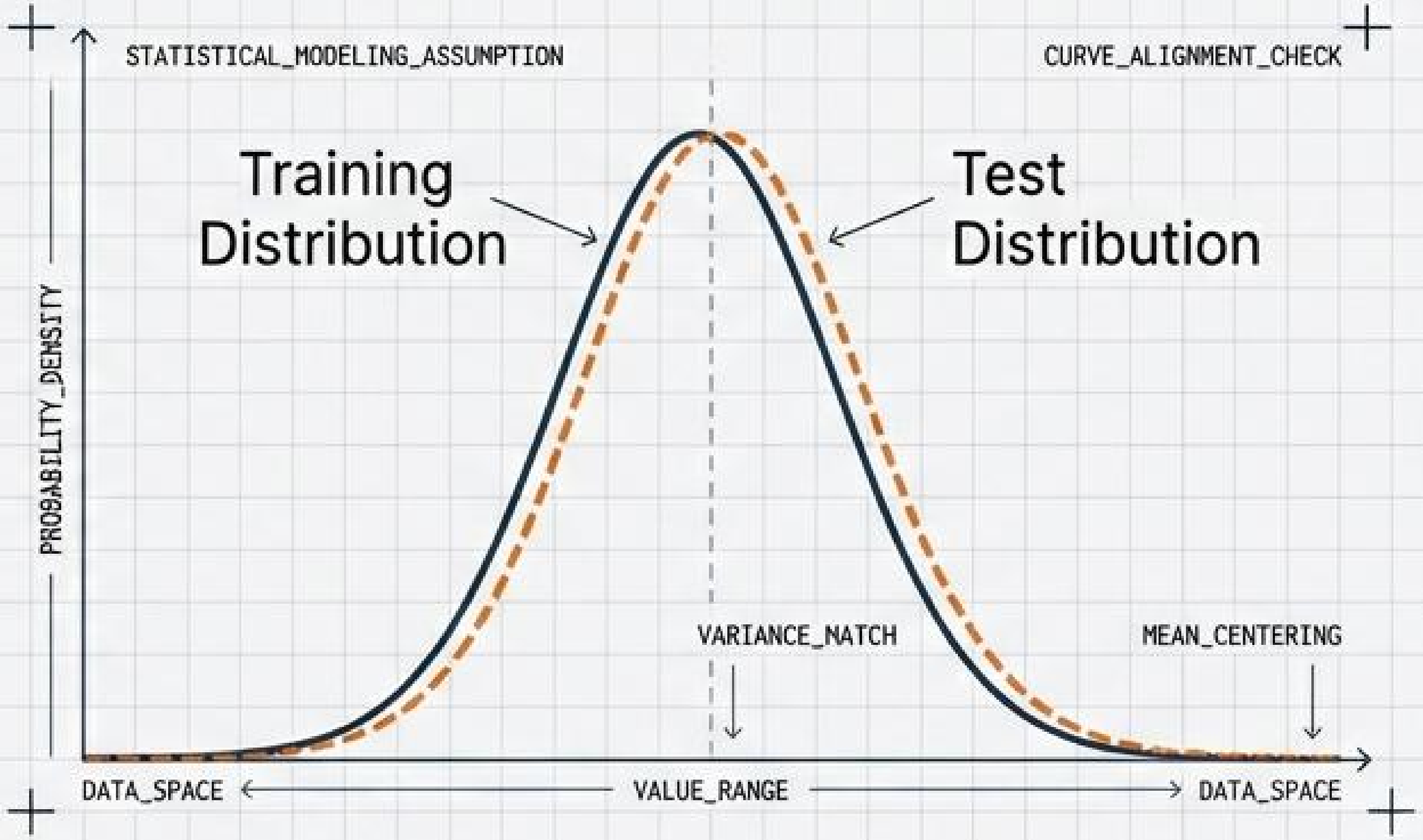
SYS_OPT_PARAMS

Insight: The choice of metric defines the success of the system.

The Distribution Assumption

The Golden Rule: i.i.d.

Independent and Identically Distributed

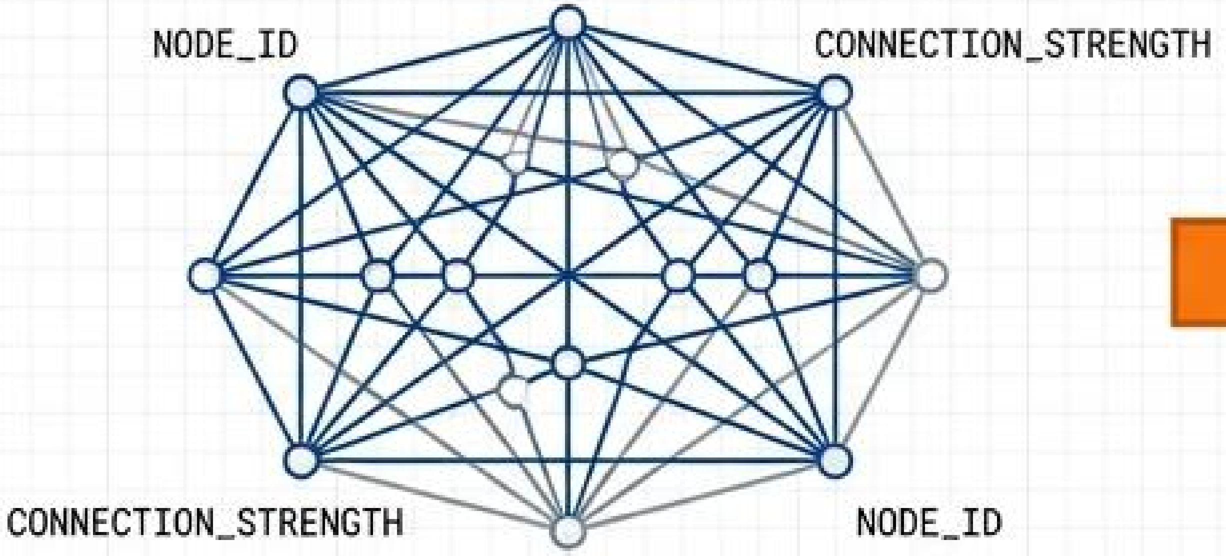


We assume that training examples and test examples are independently drawn from the same overall distribution of data. The past must statistically resemble the future.

When Assumptions Fail

Scenario A: Dependency

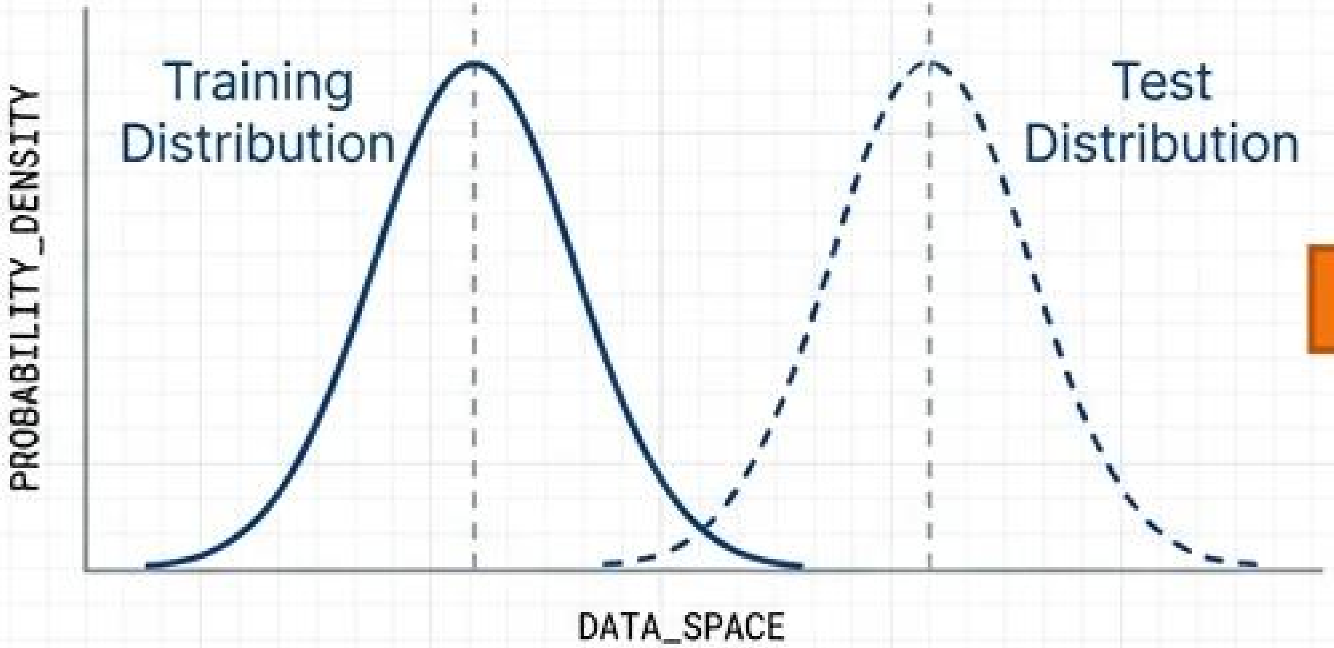
If examples are not independent...



Requires **Collective Classification**

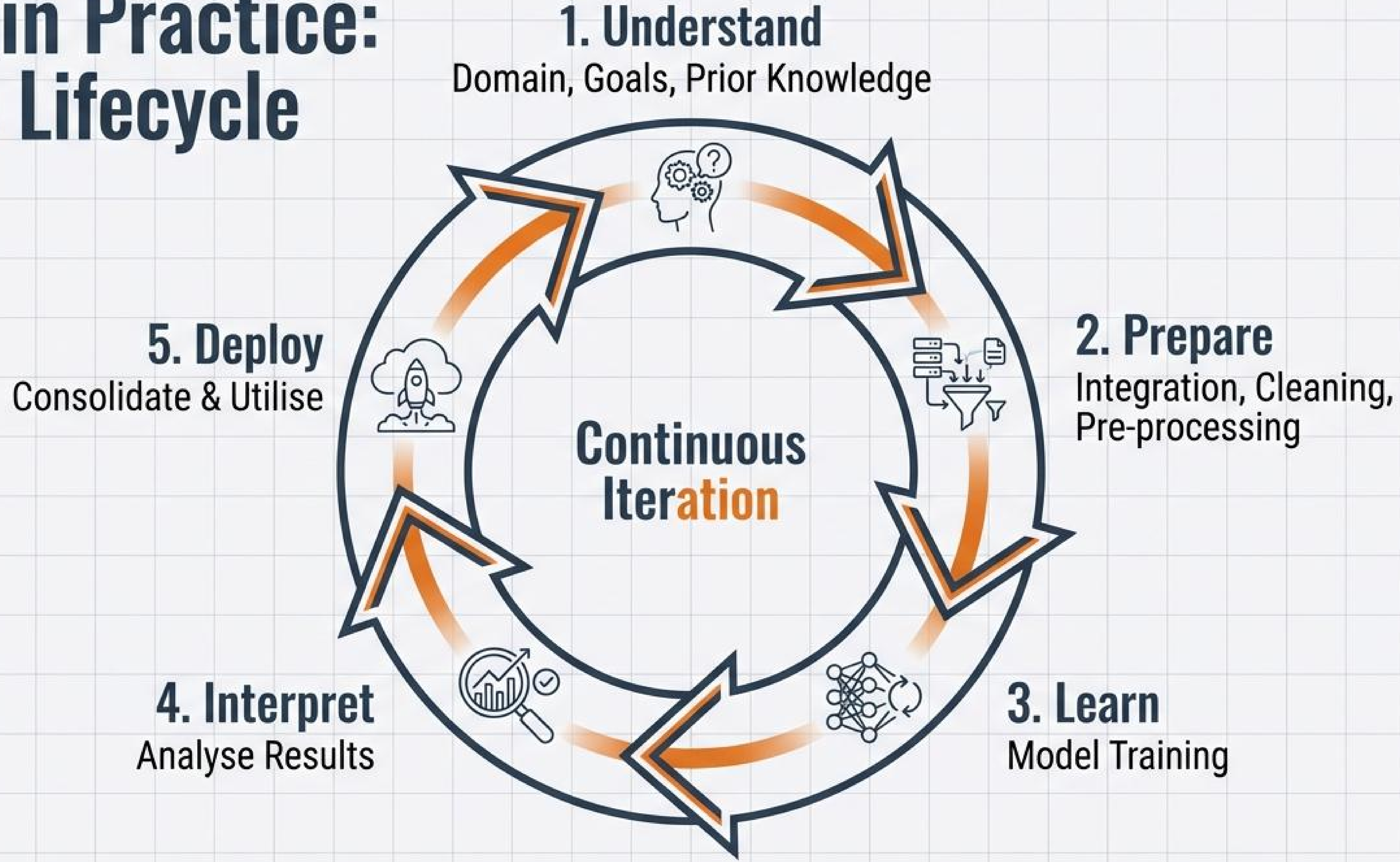
Scenario B: Drift

If the test distribution is different from the training distribution...



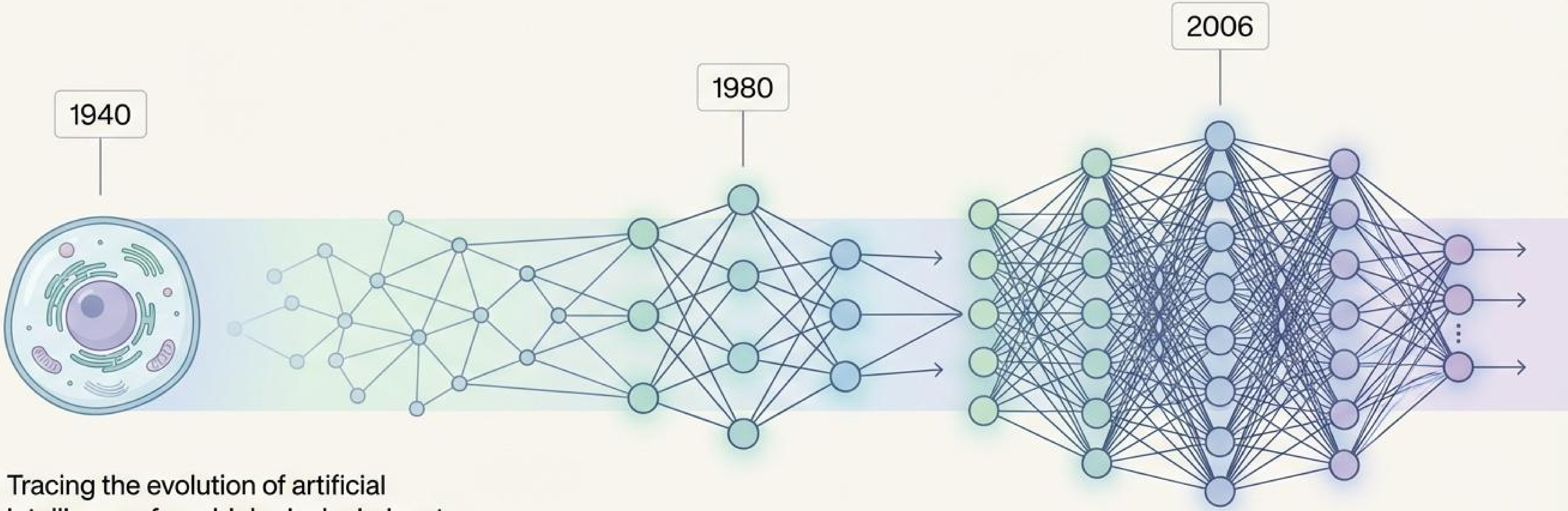
Requires **Transfer Learning**

ML in Practice: The Lifecycle



A Brief History of Deep Learning

From Cybernetics to the Revolution Age (1940–Present)



Tracing the evolution of artificial intelligence from biological mimicry to massive engineering scale.

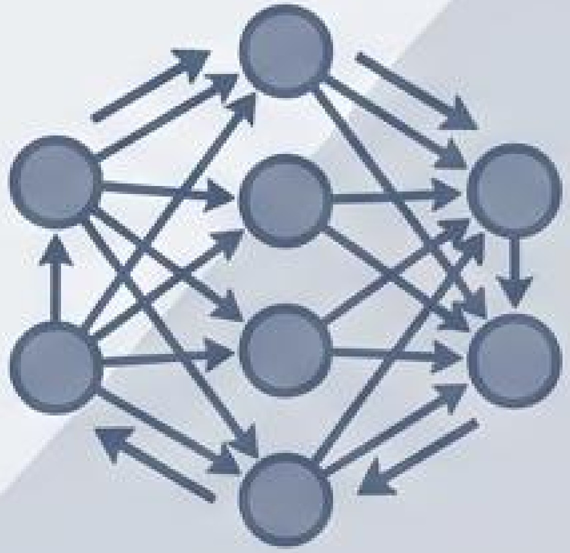
Three Waves of Development

1940–1970: Cybernetics (The Golden Age)

Simple computational models of biological learning and simple learning rules.

1980–2000: Connectionism (The Dark Age)

Intelligent behaviour achieved through large numbers of simple units using Backpropagation.



2006–Now: Deep Learning (The Revolution Age)

Deeper networks, larger datasets, and massive computation leading to **state-of-the-art** results.

The Linear Threshold Neuron

McCulloch & Pitts (1943)

- Proposed an early model for neural activation: the linear threshold neuron (binary).
- Mathematically demonstrated to be more powerful than standard logical AND/OR gates.
- The Limitation: While the structure structure existed, there was no procedure to learn weights yet.

$$f_{\mathbf{w}}(\mathbf{x}) = \begin{cases} +1 & \text{if } \mathbf{w}^T\mathbf{x} \geq 0, \\ -1 & \text{otherwise} \end{cases}$$

1958–1962

Act I

The Perceptron and The Hype

Frank Rosenblatt

- The first algorithm and implementation to train a single linear threshold neuron.
- Novikoff proved the convergence of the optimization of the perceptron criterion.

$$\mathcal{L}(\mathbf{w}) = -\sum_{n \in \mathcal{M}} \mathbf{w}^T \mathbf{x}_n y_n$$

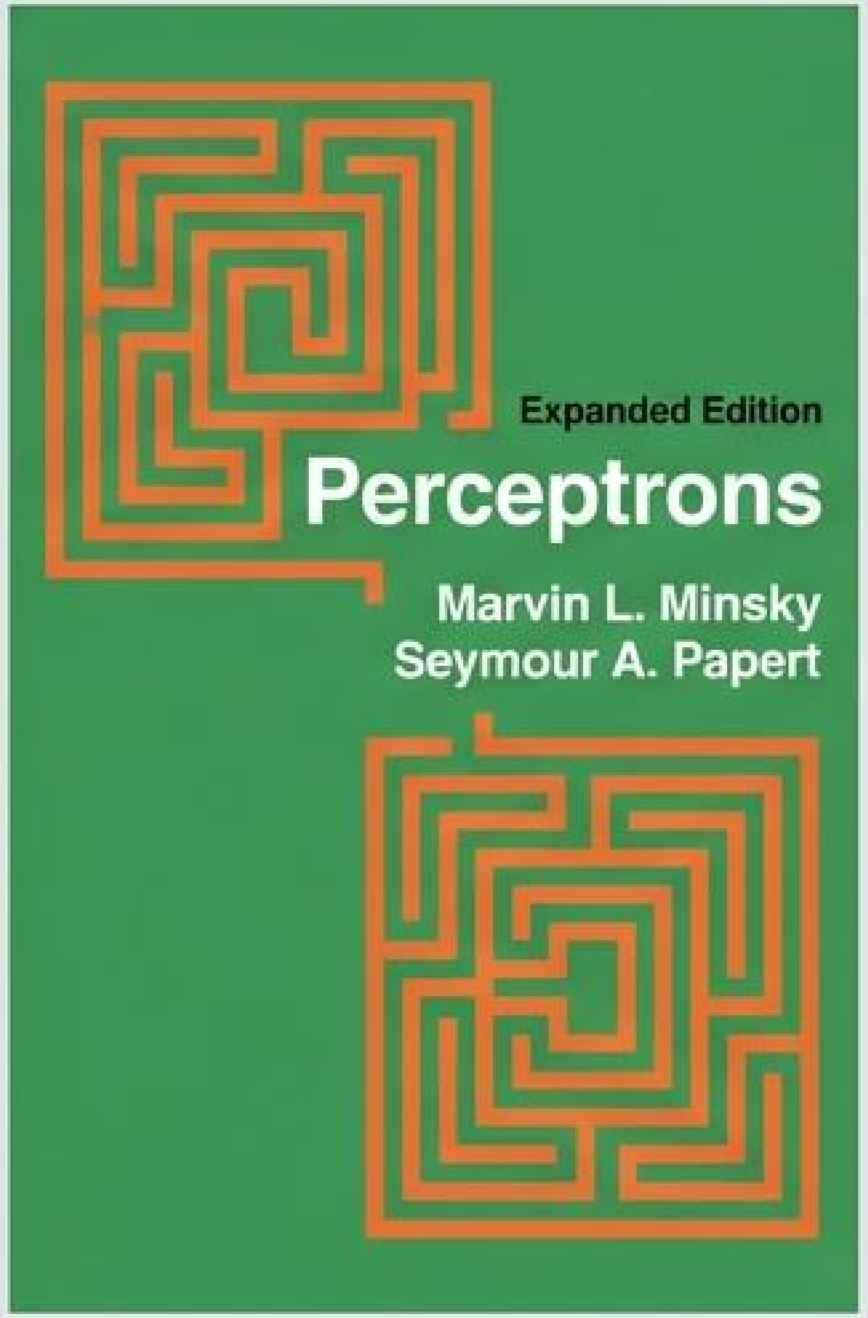
> *The perceptron will lead to computers that walk, talk, see, write, reproduce, and are conscious of of their existence.”*
>
> — Frank Rosenblatt, 1958

The Collapse

1969

Minsky & Papert

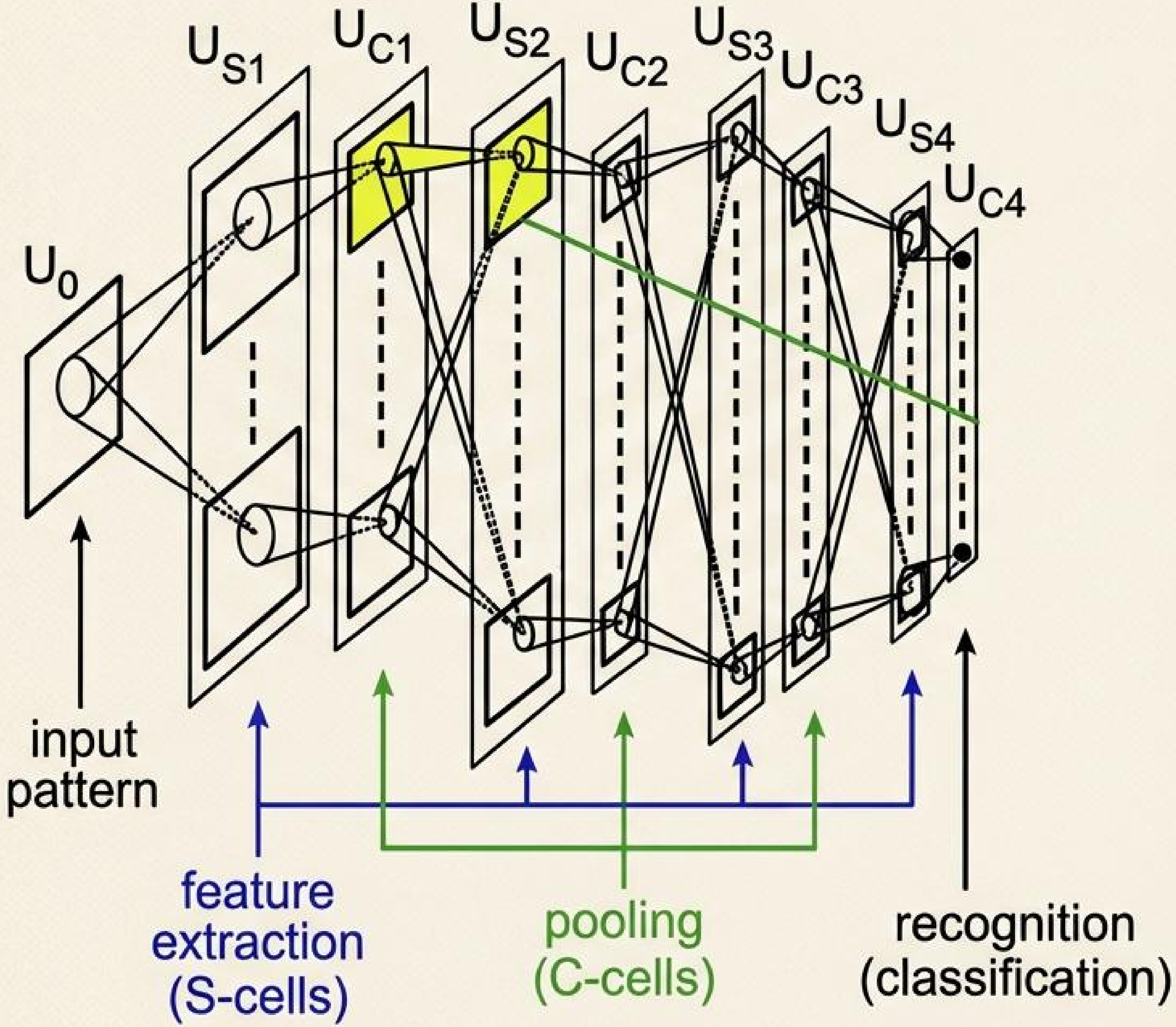
- **Event:** Publication of the book "Perceptrons".
- **The Problem:** Proved that single-layer perceptrons cannot solve simple non-linear problems like the XOR problem or counting.
- **The Aftermath:** Symbolic AI research dominates the 70s.



1979

The Neocognitron

Fukushima

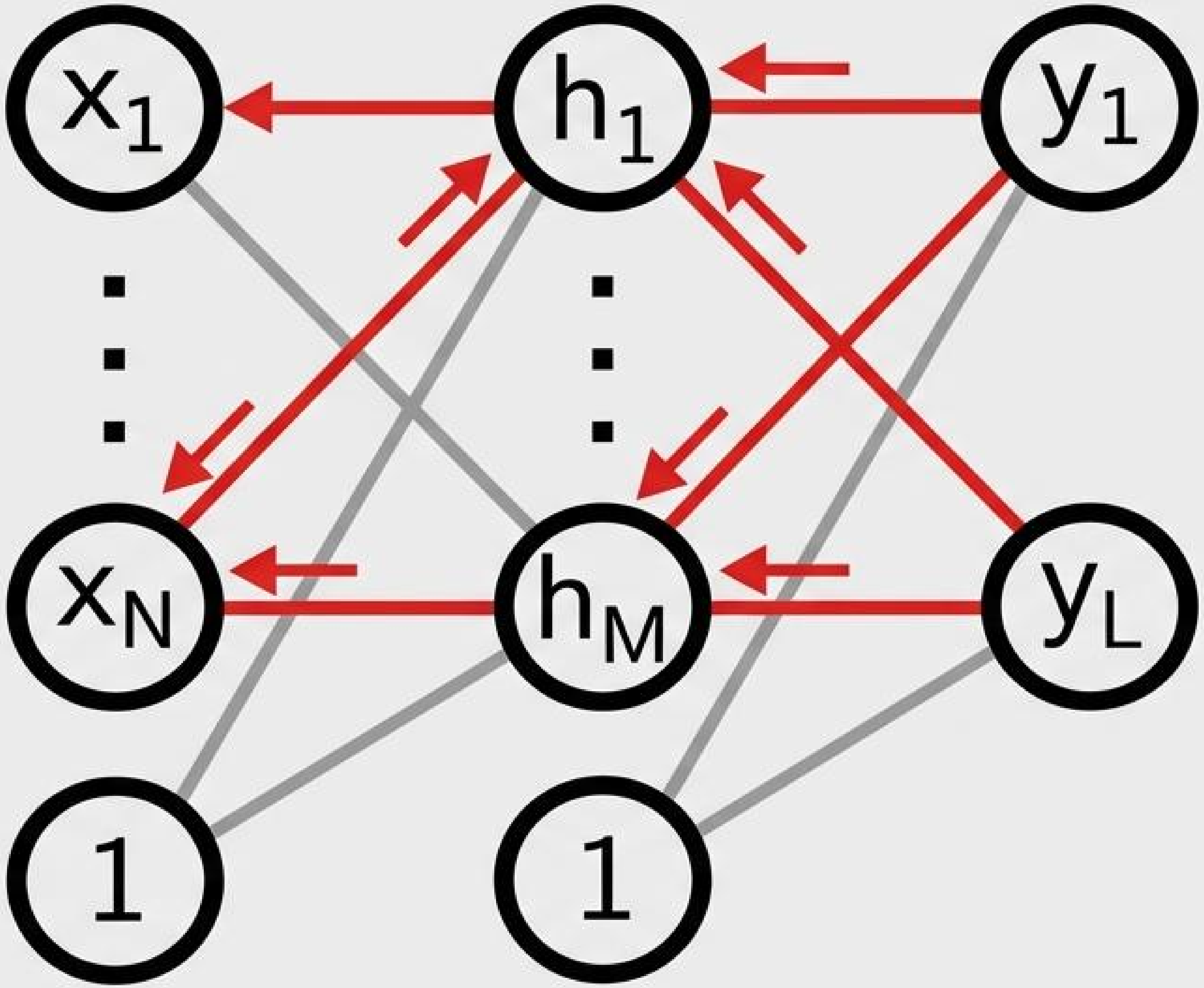


- **Inspired** by Hubel and Wiesel's Nobel Prize-winning experiments on the visual cortex of cats.
- **Biological Insight:** Visual cells are sensitive to edge orientation but insensitive to position (Simple vs. Complex cells).
- **The Innovation:** Fukushima used this to create a multi-layer processing architecture using 'Simple (S)' and 'Complex (C)' cells to implement convolution and pooling.
- **Legacy:** The direct inspiration for modern ConvNets.

1986

The Engine of Learning: Backpropagation

Rumelhart, Hinton, and Williams



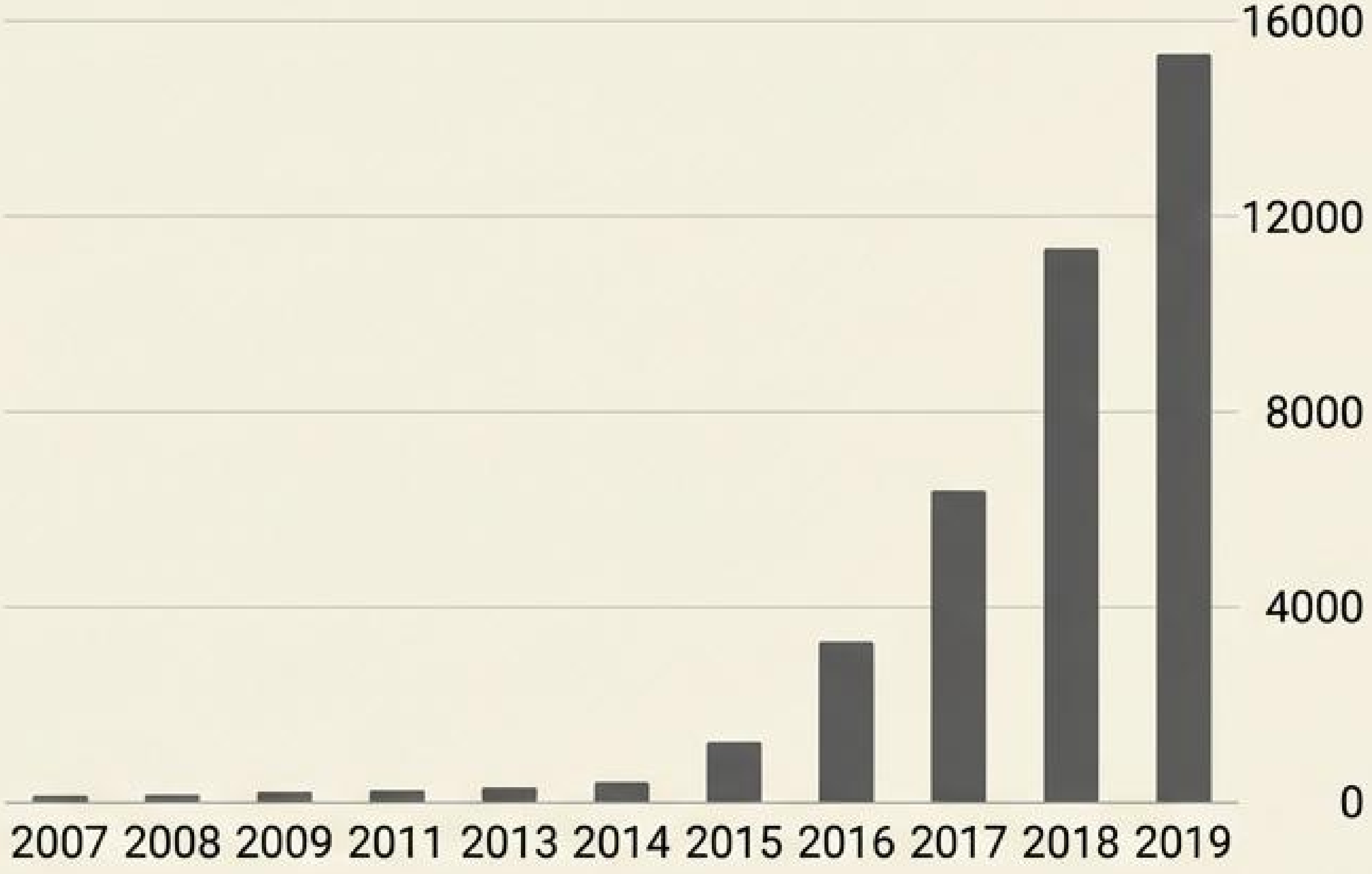
- **Definition:** An efficient method for calculating gradients in a deep network with respect to network weights.
- **Significance:** It enabled gradient-based learning to finally be applied effectively to deep networks.
- **Context:** While known since 1961, 1986 marked its first empirical success. It remains the main workhorse of AI today.

Solving Memory: LSTM

Hochreiter & Schmidhuber

1997

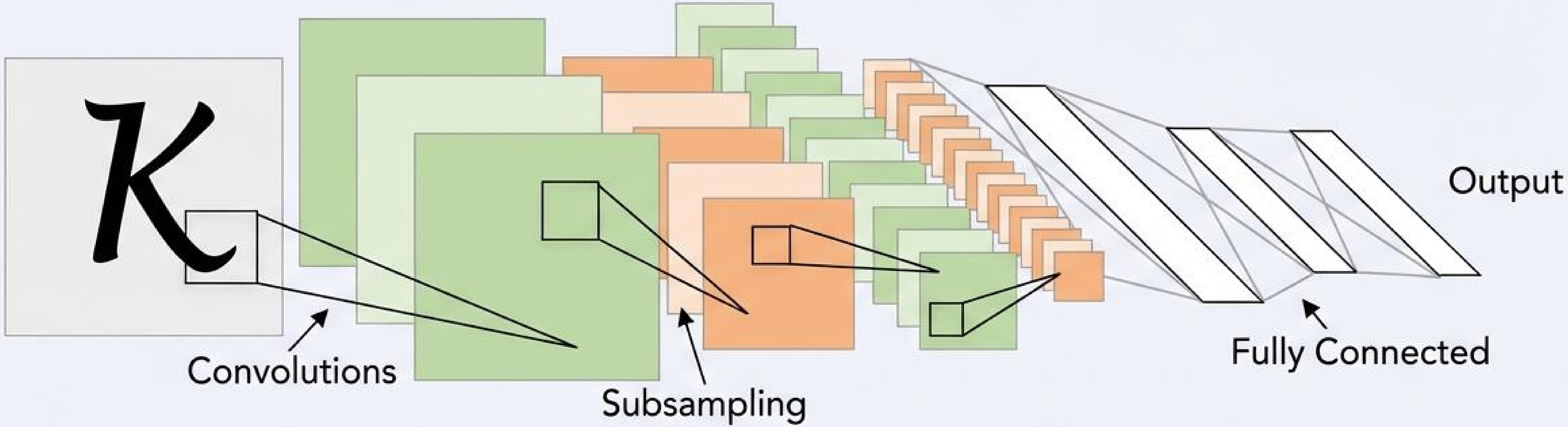
- **The Problem:** Standard networks struggled with vanishing/exploding gradients (Hochreiter, 1991).
- **The Solution:** Long Short-Term Memory (LSTM) introduced feedback loops and "forget/keep gates" to handle sequence modelling.
- **The Payoff:** Revolutionized Natural Language Processing (e.g., Google Translate) nearly 20 years later.



Convolutional Neural Networks

1998

LeCun, Bottou, Bengio, Haffner

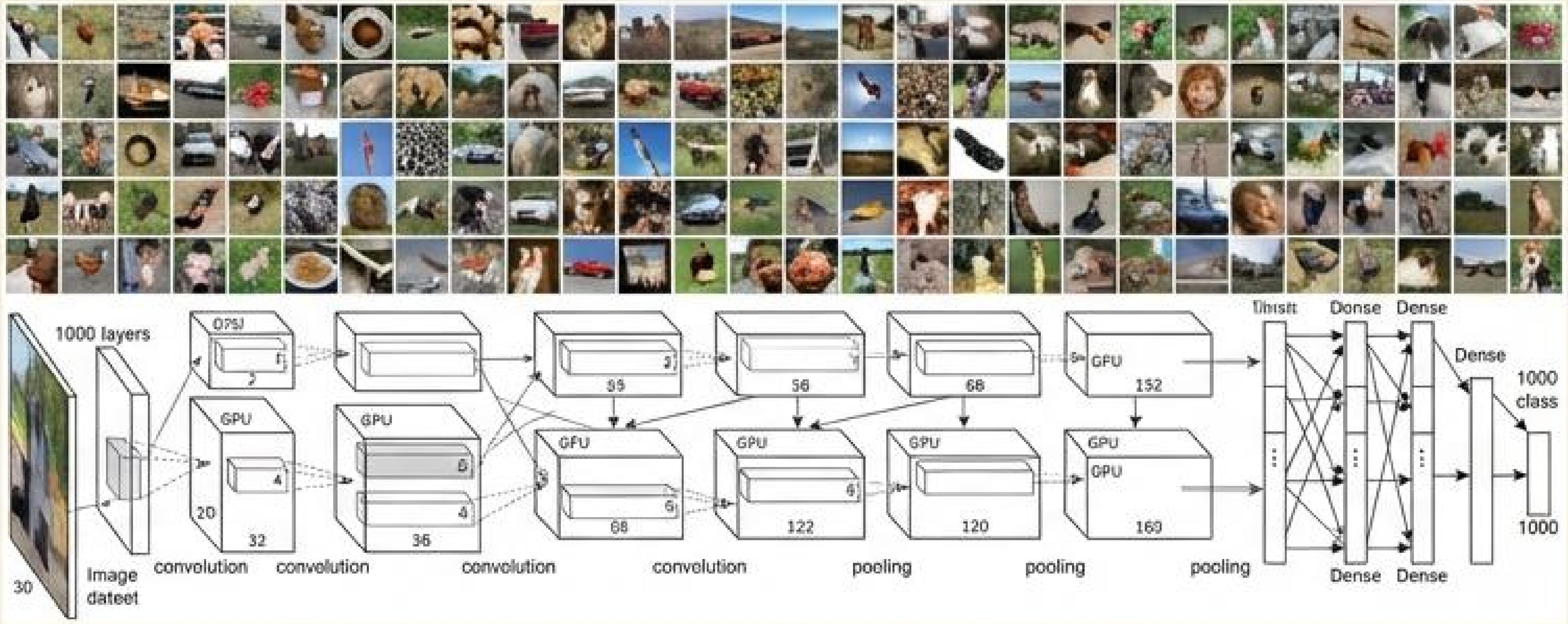


- **Architecture:** Similar to the Neocognitron but trained end-to-end using backpropagation.
- **Key Features:** Implements spatial invariance via convolutions and max-pooling; uses weight sharing to reduce parameters.
- **Result:** Achieved good results on MNIST (document recognition).
- **The Ceiling:** Did not scale up to complex images (yet).

2012

The Big Bang: AlexNet

Krizhevsky, Sutskever, Hinton

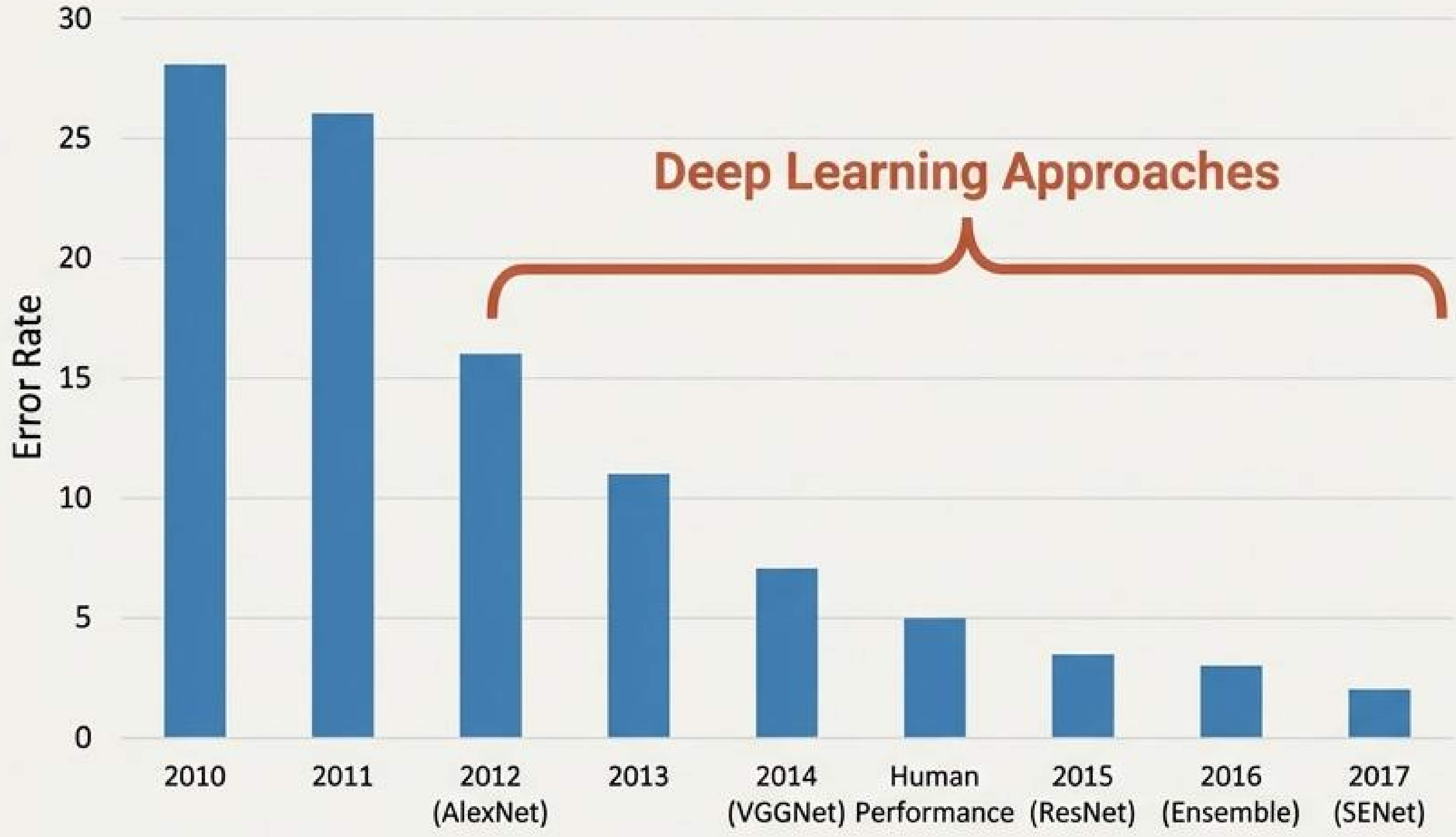


- **Context:** The ImageNet competition (ILSVRC) provided 10 million annotated images.
- **The Breakthrough:** AlexNet was the first neural network to win ILSVRC by combining Deep Models + GPU Training + Massive Data.
- **Impact:** Sparked the modern deep learning revolution.

The Impact of Deep Learning

2017

ILSVRC Top-5 Error on ImageNet



- **ILSVRC:** The ImageNet Large Scale Visual Recognition Challenge, a benchmark for image classification.
- **Dramatic Error Reduction:** The error rate dropped significantly after 2012, directly attributed to the adoption of deep learning.
- **Key Models:** AlexNet (2012), VGGNet (2014), ResNet (2015), and SENet (2017) represent major architectural advancements.
- **Surpassing Human Performance:** Deep learning approaches achieved and then surpassed human-level performance (~5% error) around 2015.

2012–Now

The Golden Age of Datasets

- **Self-driving:** KITTI, Cityscapes
- **Recognition:** PASCAL, MS COCO
- **3D Deep Learning:** ShapeNet, ScanNet
- **Language Understanding:** GLUE
- **Visual QA:** Visual Genome
- **Medical:** MITOS (Breast cancer)

Algorithms need data. Progress is now driven by massive, specialized benchmarks.

The Future of Fuel: Synthetic Data

- **The Problem:** Annotating real data is expensive and slow.
- **The Solution:** A surge in synthetic datasets generated via game engines and simulation.
- Bypassing the bottleneck of human annotation.

2012–2015

2012–2015: Learning from the Fake World

The Constraint:
Annotating real-world data was prohibitively expensive.

The Solution: Synthetic Data.
Researchers used simple 3D assets to train complex models.

The Insight: Pre-training on very simple, synthetic 3D datasets (like flying chairs) proved surprisingly effective for real-world tasks like optical flow.

Synthetic 'Flying Chairs' dataset used to train FlowNet.

Dosovitskiy et al.: FlowNet: Learning Optical Flow with Convolutional Networks. ICCV, 2015.

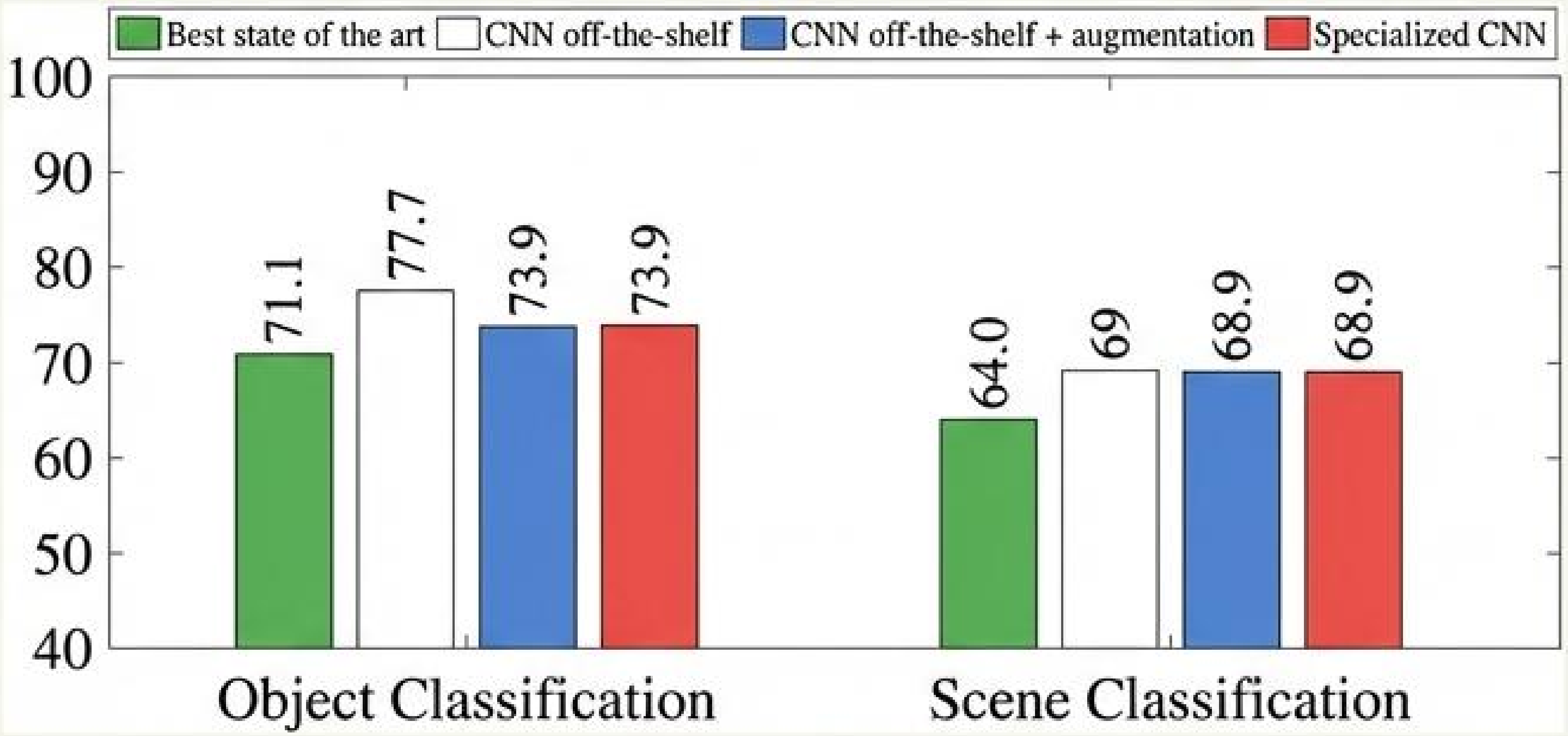
2014: The Power of Generalisation

Core Argument: Deep representations are not rigid memorisation; they are transferable skills.

The Methodology:

1. Pre-train a CNN on large generic datasets (ImageNet).
2. Fine-tune only the last layers for a specific new task.

The Result: Established the 'Off-the-Shelf' paradigm, yielding state-of-the-art performance with minimal new data.



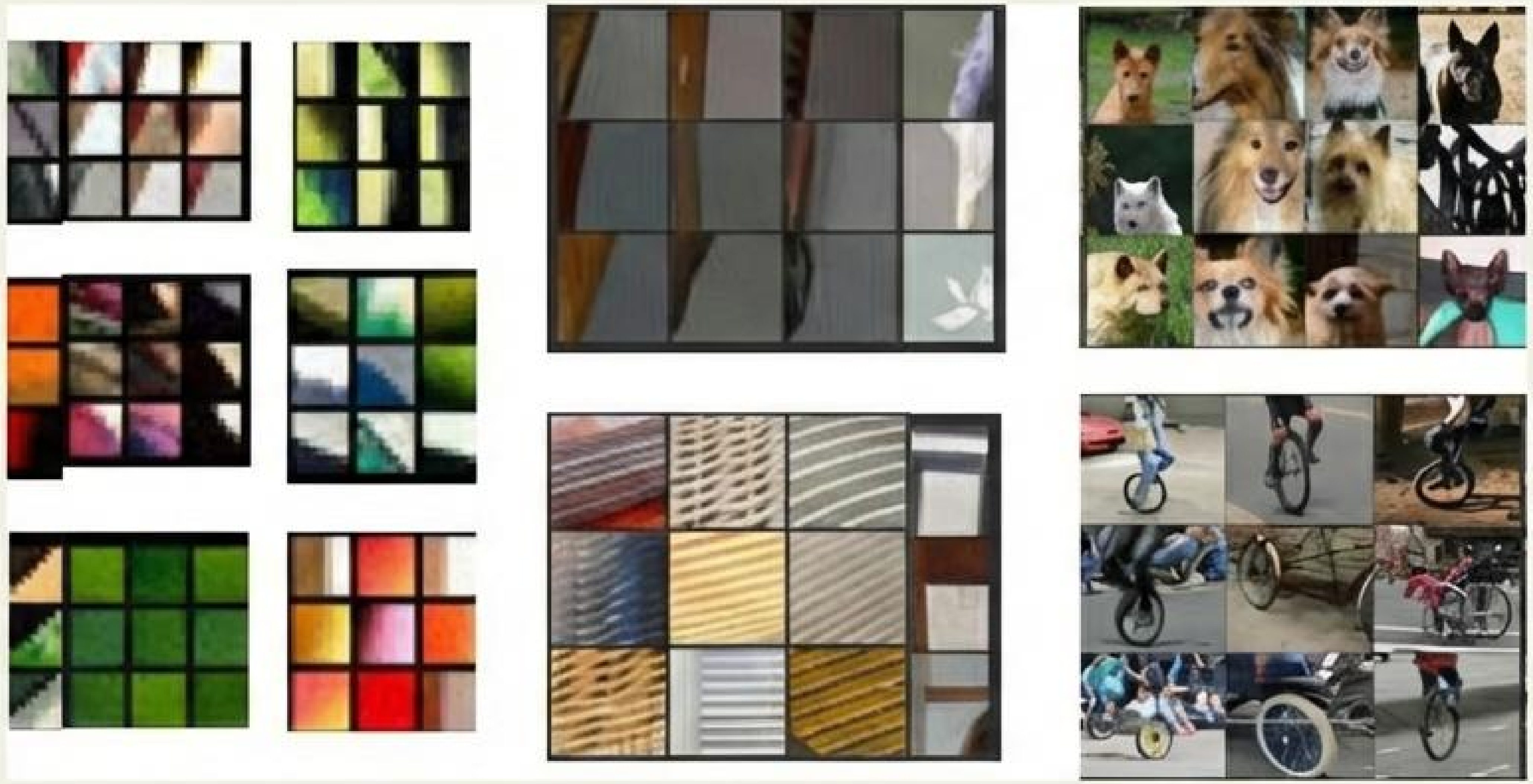
Razavian et al.: CNN Features Off-the-Shelf. CVPR Workshops, 2014.

2014: Opening the Black Box

The Goal: To understand *how* neural networks see.

The Discovery: Visualisation revealed a clear hierarchy. Lower layers activate on basic edges and textures. Higher layers capture abstract semantic information like faces and wheels.

Impact: Validated the hierarchical theory of Deep Learning.



Layer 1 Layer 2 Layer 5

Zeiler and Fergus: Visualizing and Understanding Convolutional Networks. ECCV, 2014.

2014: The Illusion of Perfection

The Vulnerability: High accuracy does not imply robust understanding.

Adversarial Examples: Imperceptible noise added to an image can catastrophically fool a network.

The Example: A school bus + invisible noise = “Ostrich”.

$$x + \underset{\Delta x}{\operatorname{argmin}} \{\|\Delta x\|_2 : f(x + \Delta x) \neq f(x)\}$$

Classified as: OSTRICH

Szegedy et al.: Intriguing properties of neural networks. ICLR, 2014.

2014

2014: The Expansion of Capability

Deep Learning moved beyond simple classification to conquer three distinct frontiers simultaneously:

Language

Seq2Seq models revolutionised Machine Translation.

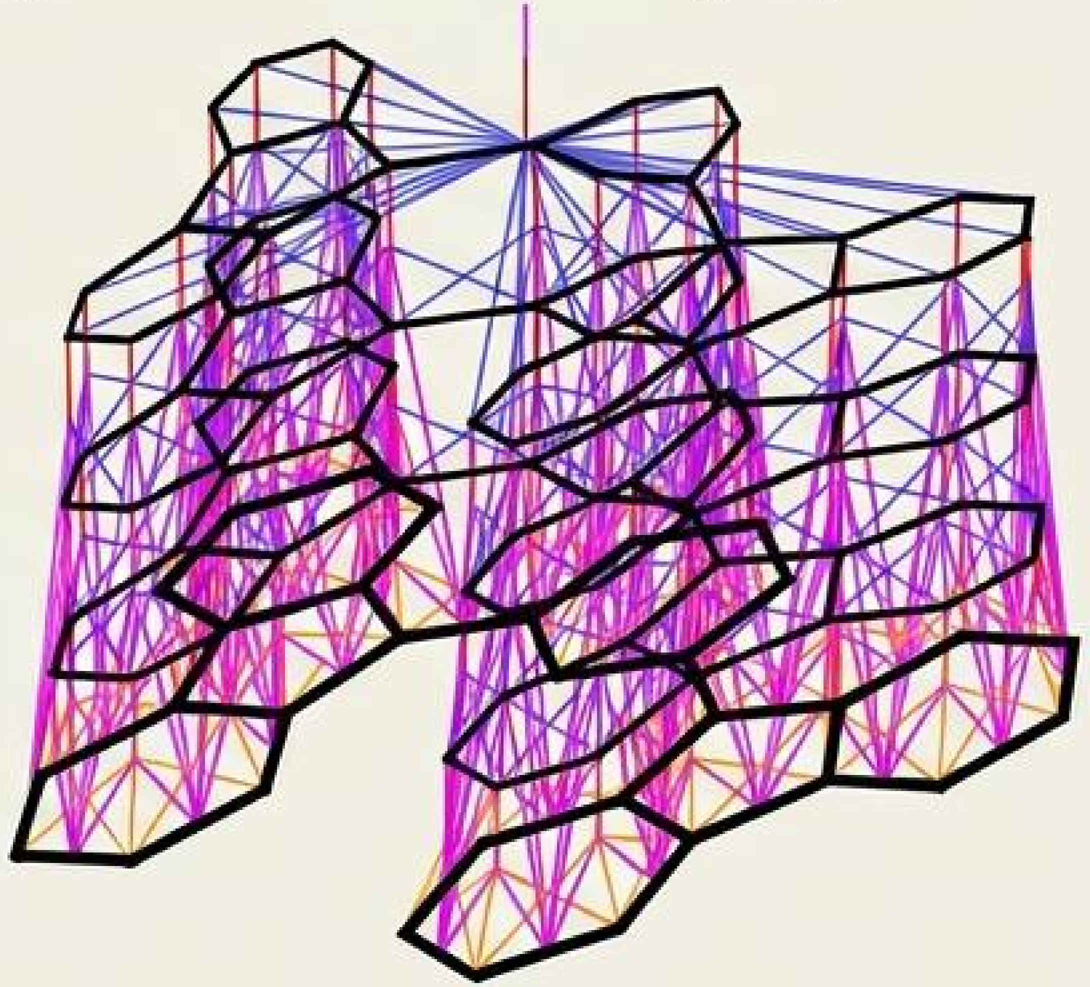
Type	Sentence
Our model	Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .
Truth	Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .
Our model	" Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air " , dit UNK .
Truth	" Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord " , a déclaré Rosenker .
Our model	Avec la crémation , il y a un " sentiment de violence contre le corps d' un être cher " , qui sera " réduit à une pile de cendres " en très peu de temps au lieu d' un processus de décomposition " qui accompagnera les étapes du deuil " .
Truth	Il y a , avec la crémation , " une violence faite au corps aimé " , qui va être " réduit à un tas de cendres " en très peu de temps , et non après un processus de décomposition , qui " accompagnerait les phases du deuil " .

Creation

Generative models (GANs) began 'dreaming' up images.

Science

Graph Neural Networks (GNNs) predicted molecular properties.



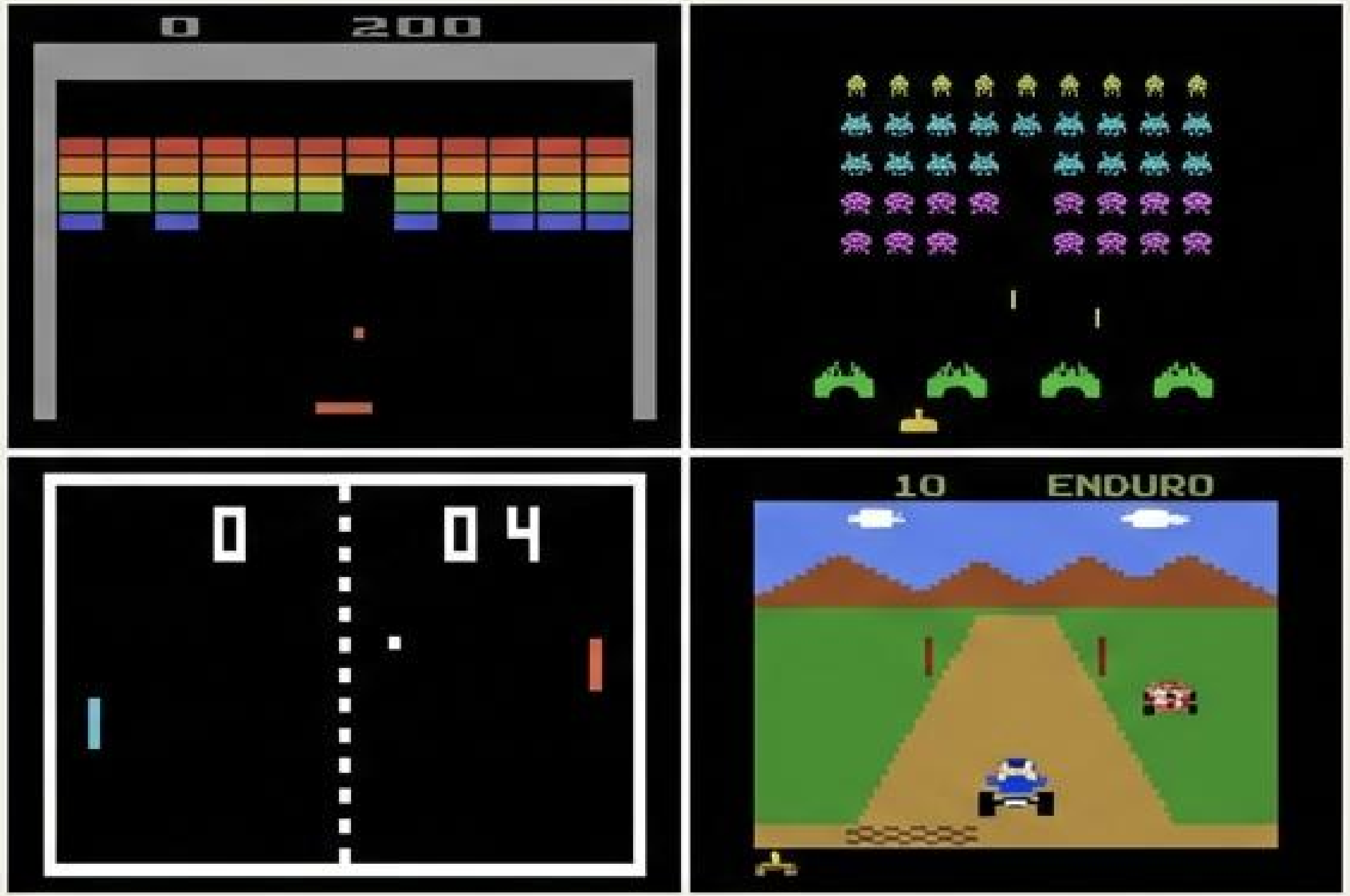
Sutskever et al. (2014); Goodfellow et al. (2014); Duvenaud et al. (2015)

2015 2015: From Passive to Active (Deep RL)

The Shift: AI moved from processing static images to making dynamic decisions.

Deep Reinforcement Learning: Learning a policy (*state* → *action*$) through random exploration and reward signals.

Constraint: Zero supervision. The agent learned solely by playing and observing the score.



DeepMind's agent achieving superhuman performance on Atari games directly from pixels.

Mnih et al.: Human-level control through deep reinforcement learning. *Nature*, 2015.

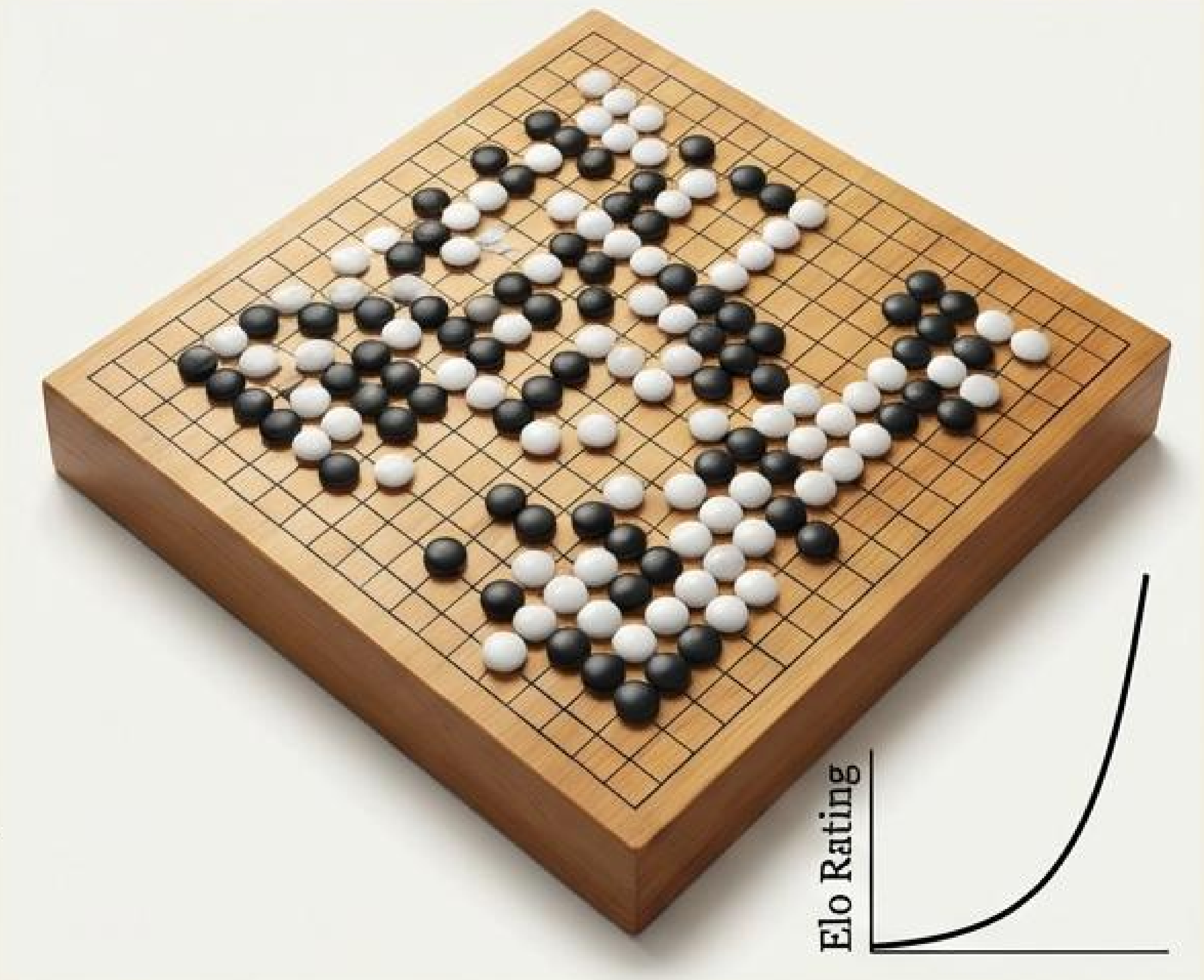
2016: Solving the Impossible Game

The Milestone: AlphaGo defeated world champion Lee Sedol in a game previously thought too complex for computers.

Methodology:

Methodology: A hybrid of Deep Learning (to evaluate board states) and Monte Carlo tree search (for long-term planning).

Evolution: AlphaZero (2017) later discarded human data entirely, learning via self-play.



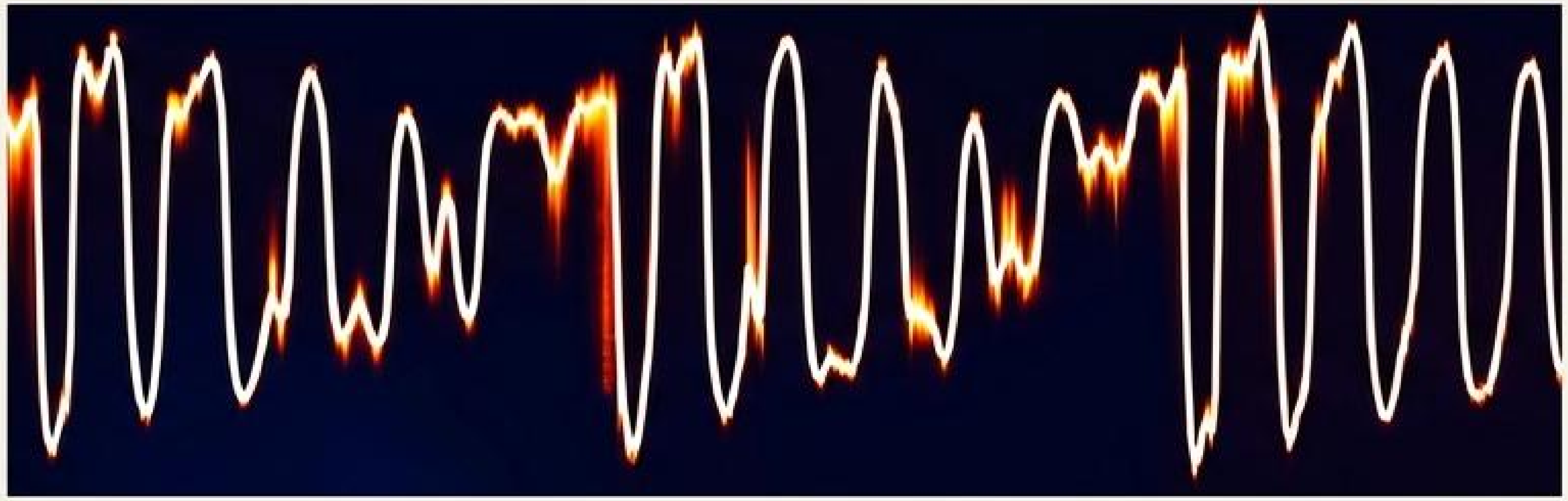
Silver et al.: Mastering the game of Go without human knowledge. Nature, 2017.

2016: Computational Creativity

2016

WaveNet (Audio)

Generative models for raw audio waveforms, mimicking human speech and music.



Style Transfer (Vision)

Disentangling ‘content’ from ‘style’ to render photos as paintings.

Input Content Style Reference Output

Oord et al. (2016); Gatys et al. (2016).

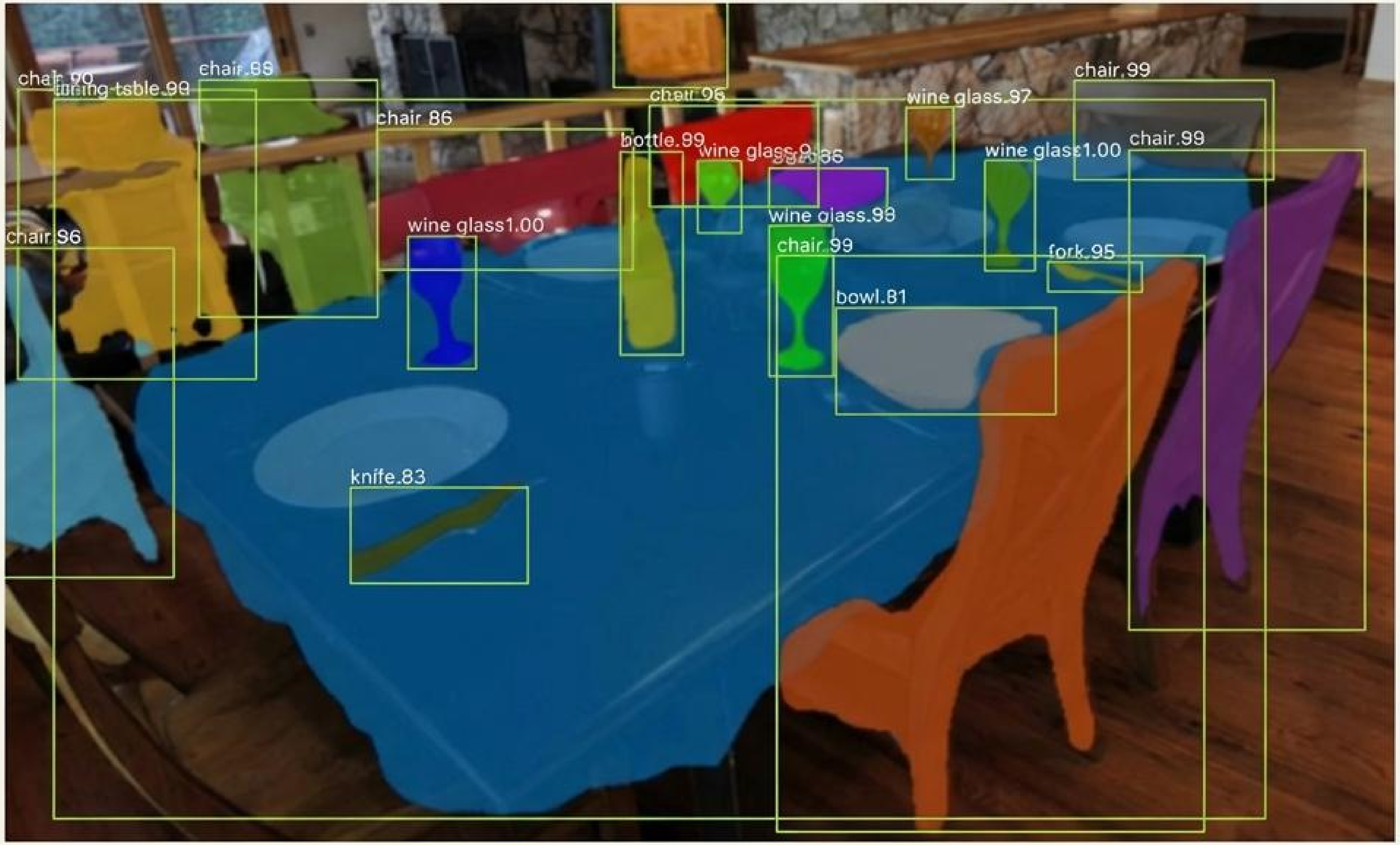
2017

2017: Precision Vision (Mask R-CNN)

The Problem:
Bounding boxes are too coarse. We need to know exactly which pixels belong to an object.

The Solution:
Mask R-CNN. Joint object detection and instance segmentation.

The Output:
A “structured object” with precise boundaries, not just a label.



He et al.: Mask R-CNN, ICCV, 2017.

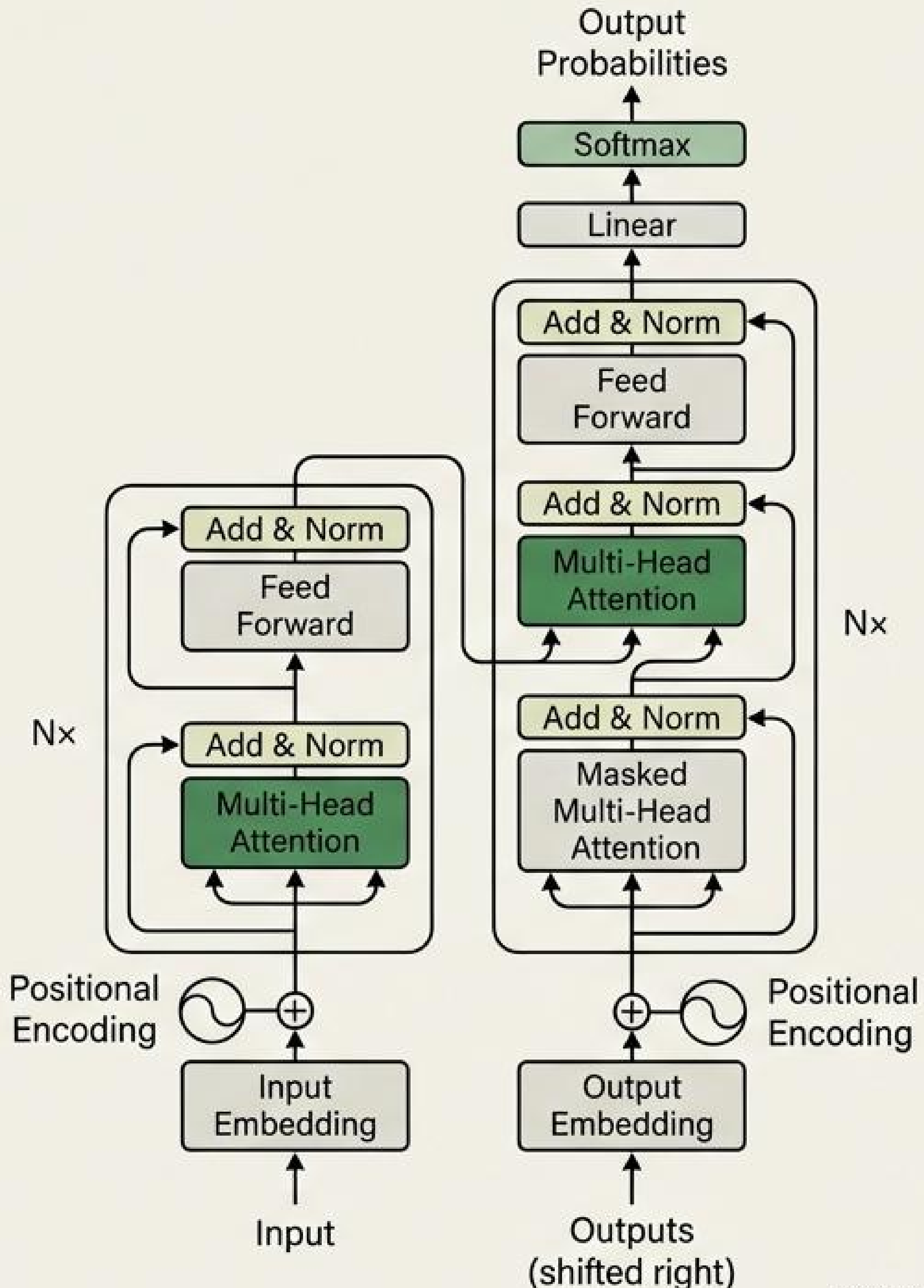
2017

2017: The Attention Revolution

The Architecture: The Transformer.

Key Mechanism: "Attention is All You Need."

The Breakthrough: Replaced recurrence (RNNs) with attention mechanisms, allowing the model to weigh the importance of all words in a sentence simultaneously. This enabled massive parallelisation.

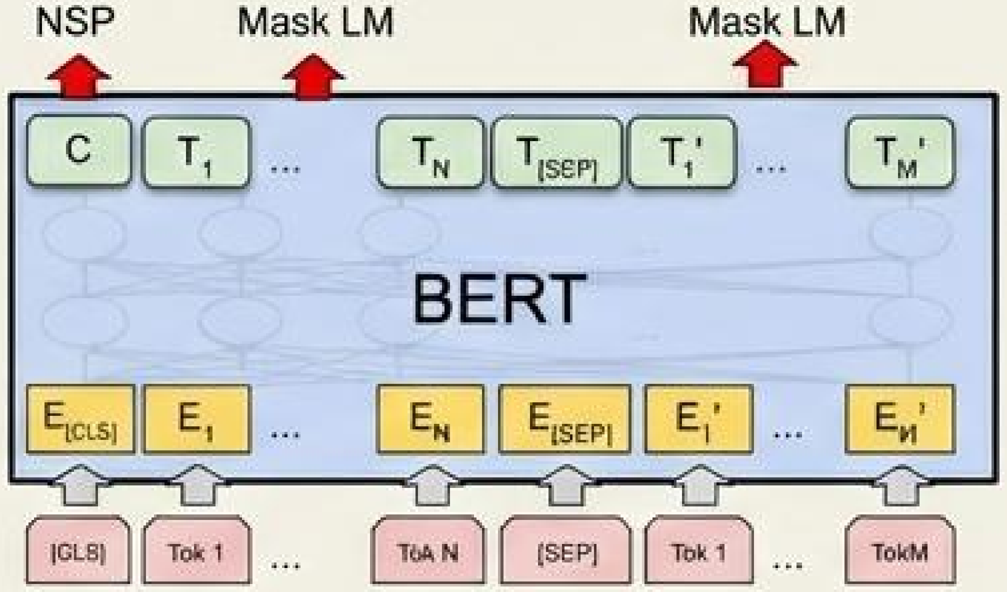


Vaswani et al.: Attention is All you Need. NIPS 2017.

2018

2018: Mastering Language

BERT: Pre-training bidirectional transformers on massive unlabeled text.



	Raok	Neme	Model	URL	Ssore
	1	HFL iFLYTEK	MscALSERT + DKM		90.7
+	2	Allbabs DAMO NLP	StrucISERT + TAPT		90.6
+	3	PING-AN Gmni-Sinitic	ALSERT + DAAF + NAS		90.6
	4	ERNIE Team - Saidu	ERNIE		90.4
	5	T5 Team - Google	T5		90.3
	6	Microsefl D365 AI & MSR AI & GATECH	MTDNN SMART		89.9
+	7	Zihang Dei	Funnel-Tranformer (Encembie 819-10-10H1024)		89.7
+	8	ELECTRA Team	ELECTRA Large + Standard Tricks		89.4
+	9	Huawei Noali's Ark Lab	NEZHA-Large		89.1
+	10	Microsoft D365 AI & UMD	FreeLB RoBERTa (ensemble)		88.4
	11	Junjie Yang	HIRE RoBERTa		88.3
	12	Facebock AI	RoBERTa		88.1
+	13	Microsoft D365 AI & MS9 AI	MT DNN-ensemble		87.6
	14	GLUE Human Baselines	GLUE Human Baselines		87.1

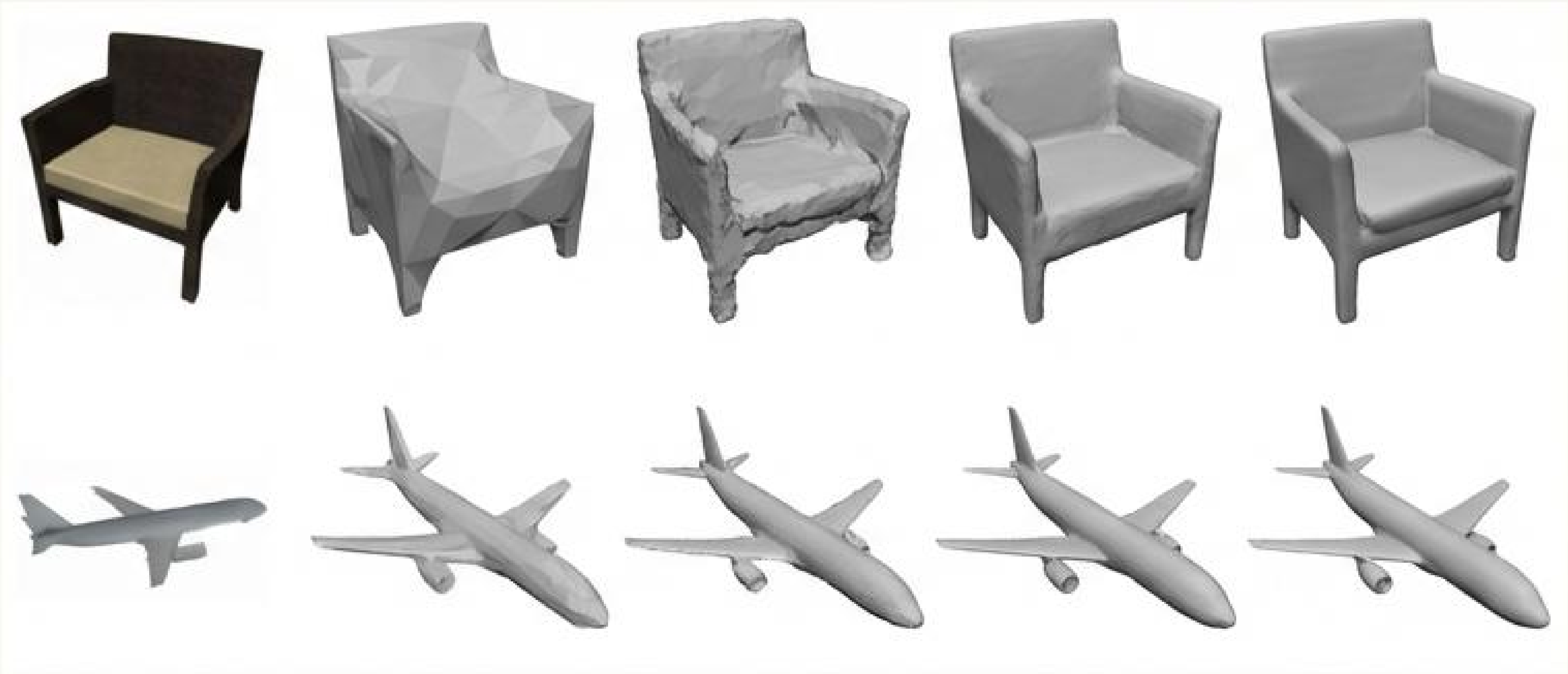
The Impact: Achieved superhuman performance on the GLUE benchmark (General Language Understanding Evaluationg Evaluation).

Recognition: The 2018 Turing Award was granted to Bengio, Hinton, and LeCun.

Devlin et al. (2018); Wang et al. (2019).

2016–2020: Entering the Third Dimension

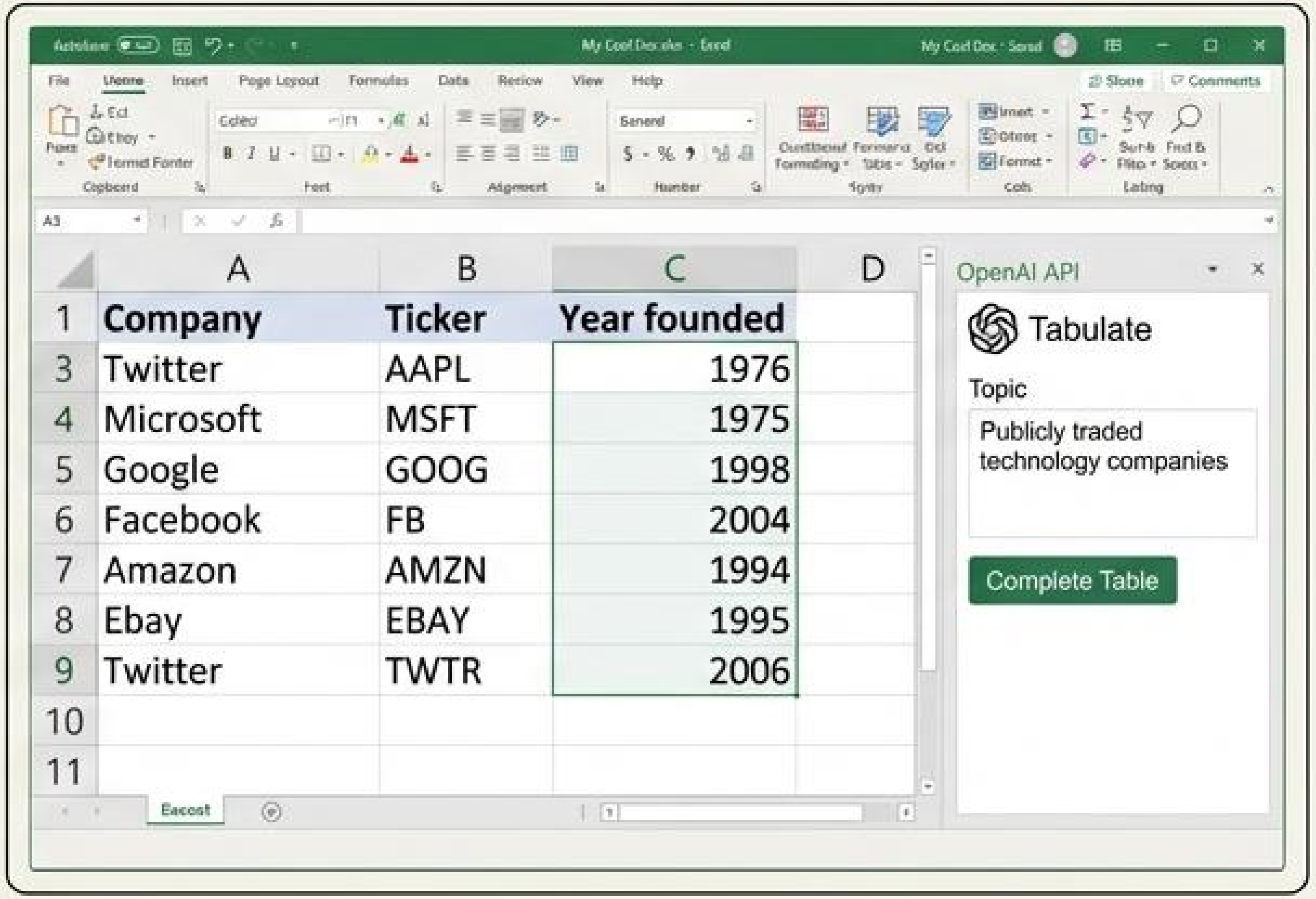
- **The Frontier:** Inferring 3D geometry from 2D inputs.
- **Capabilities:** Models learned to output voxels, point clouds, and implicit representations (NeRFs).
- **Inference:** Predicting full 3D shape, material, and lighting from a single flat image.



Niemeyer et al. : Differentiable Volumetric Rendering. CVPR, 2020.

2020: The Era of Scale (GPT-3)

- **The Model:** 175 Billion parameters.
- **The Interface:** A general-purpose “Text-in / Text-out” engine.
- **Capabilities:** Few-shot learning enabled it to write code, compose poetry, and generate news articles with minimal instruction.
- **Business:** Licensed exclusively to Microsoft in Sept 2020.

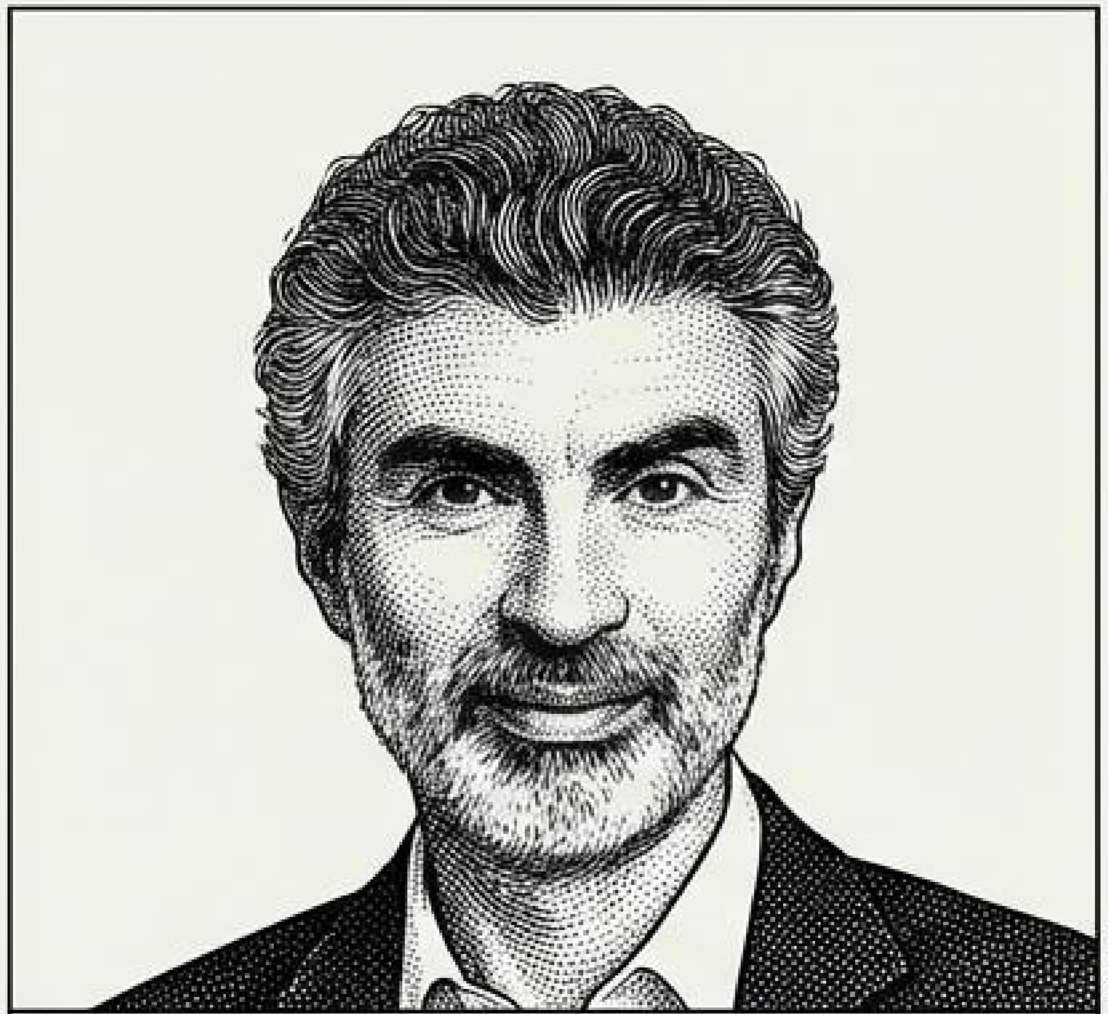


The Scale Hypothesis Verified (GPT-3)

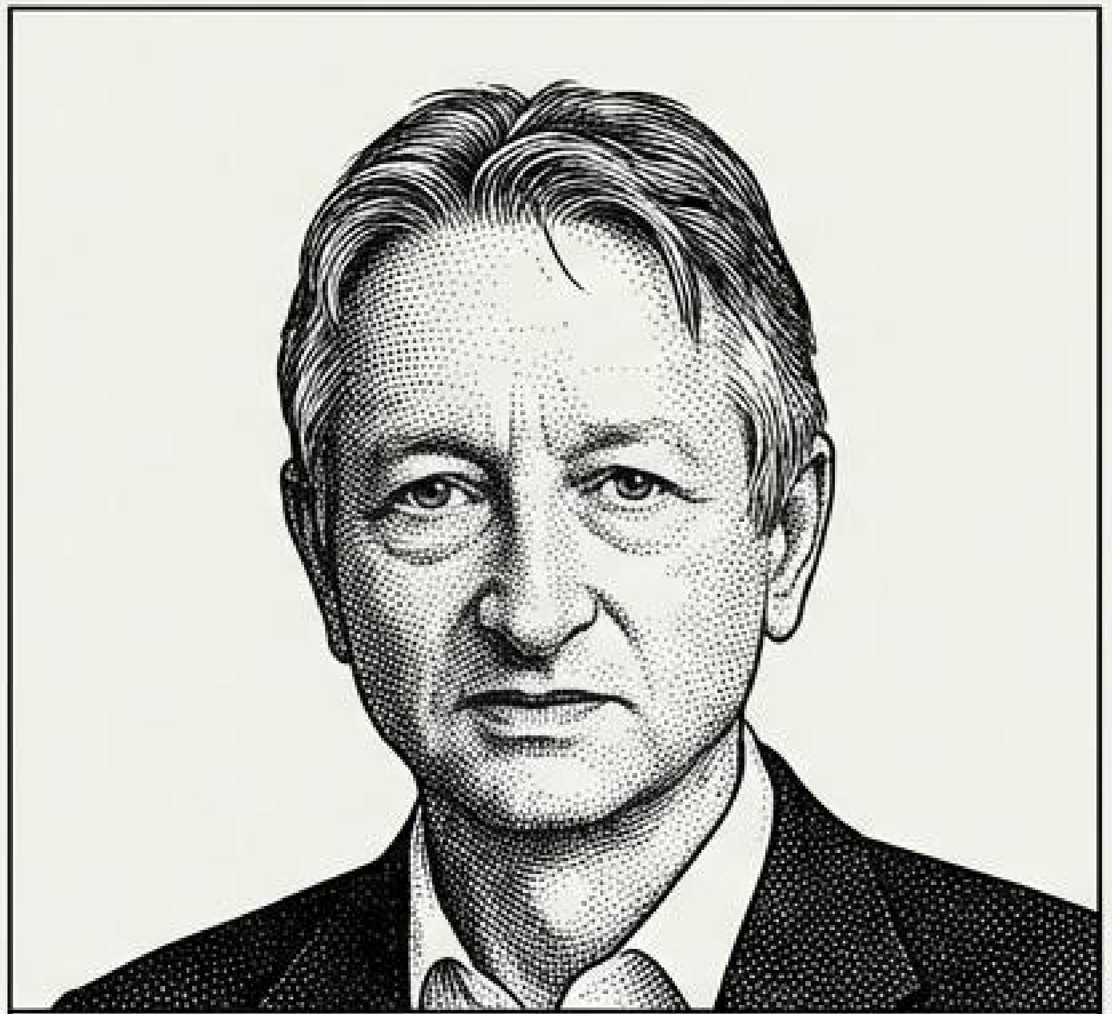
Licensed exclusively to Microsoft, Sept 2020.

Honoring the Godfathers

2018 Turing Award Recipients



Yoshua Bengio



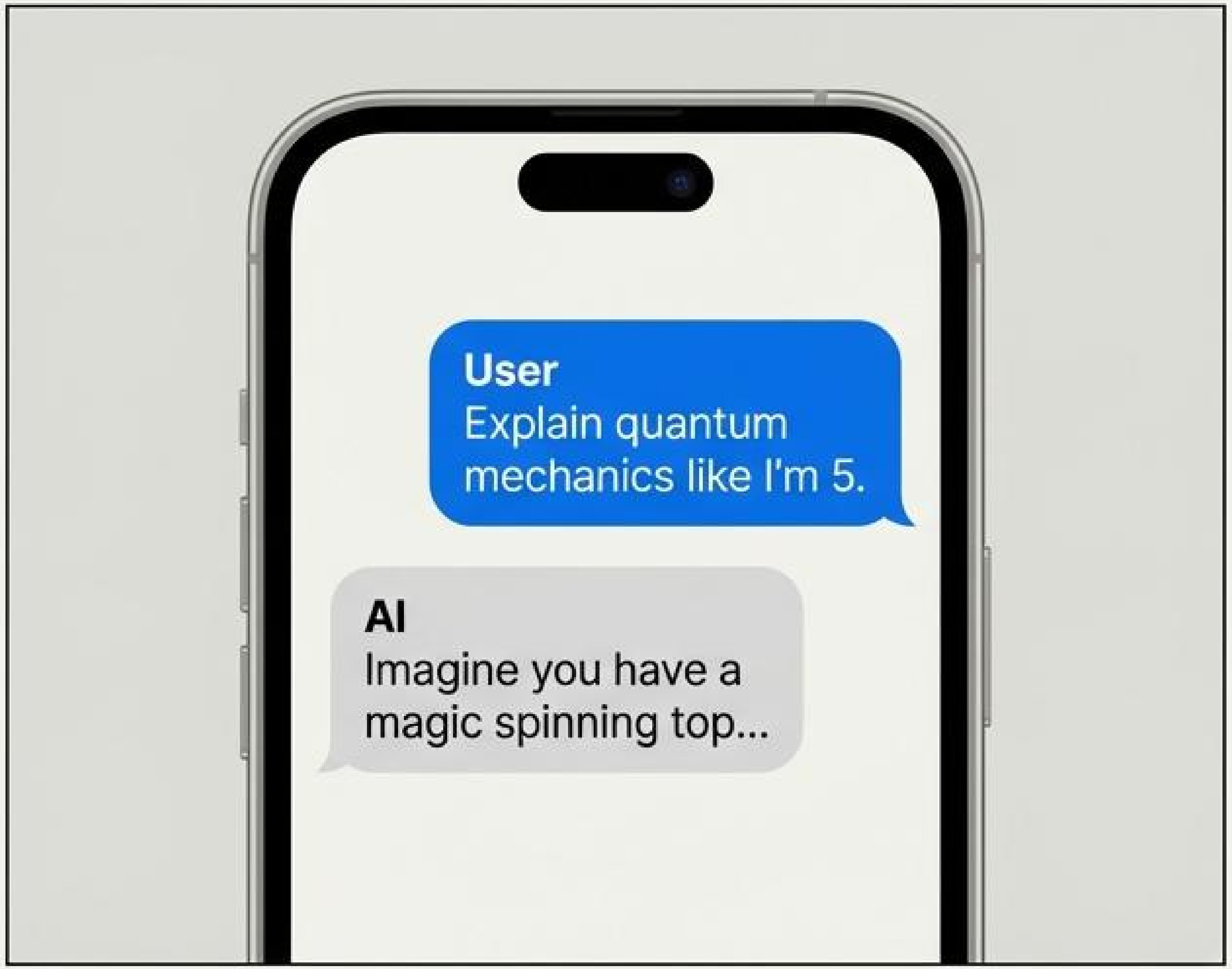
Geoffrey Hinton

Yann LeCun

Recognized with the "Nobel Prize of Computing" for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing.

AI Enters the Group Chat (2021–2022)

Neue Haas Grotesk Display Pro



ChatGPT: LLMs as Utility

Diffusion Models: Text-to-Image

Breaking the Walls Between Senses

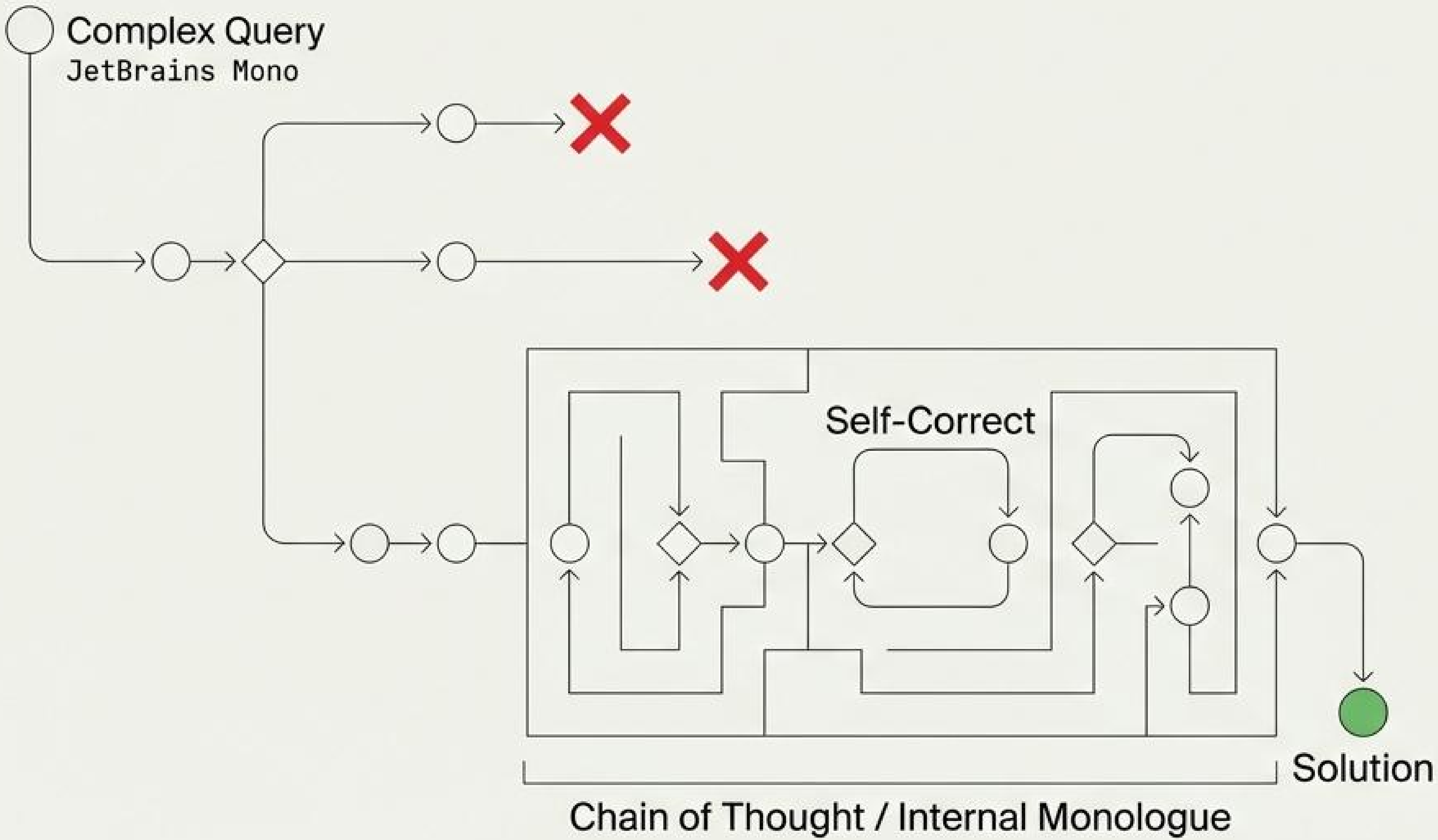
2023–2024: Multimodality & Video

Multimodality: Models like GPT-4 and Gemini now process text, audio, and vision simultaneously.

World Simulation: Video generation models (Sora) begin to simulate physical laws (gravity, collision) purely from observing video data.

From Pattern Matching to Reasoning

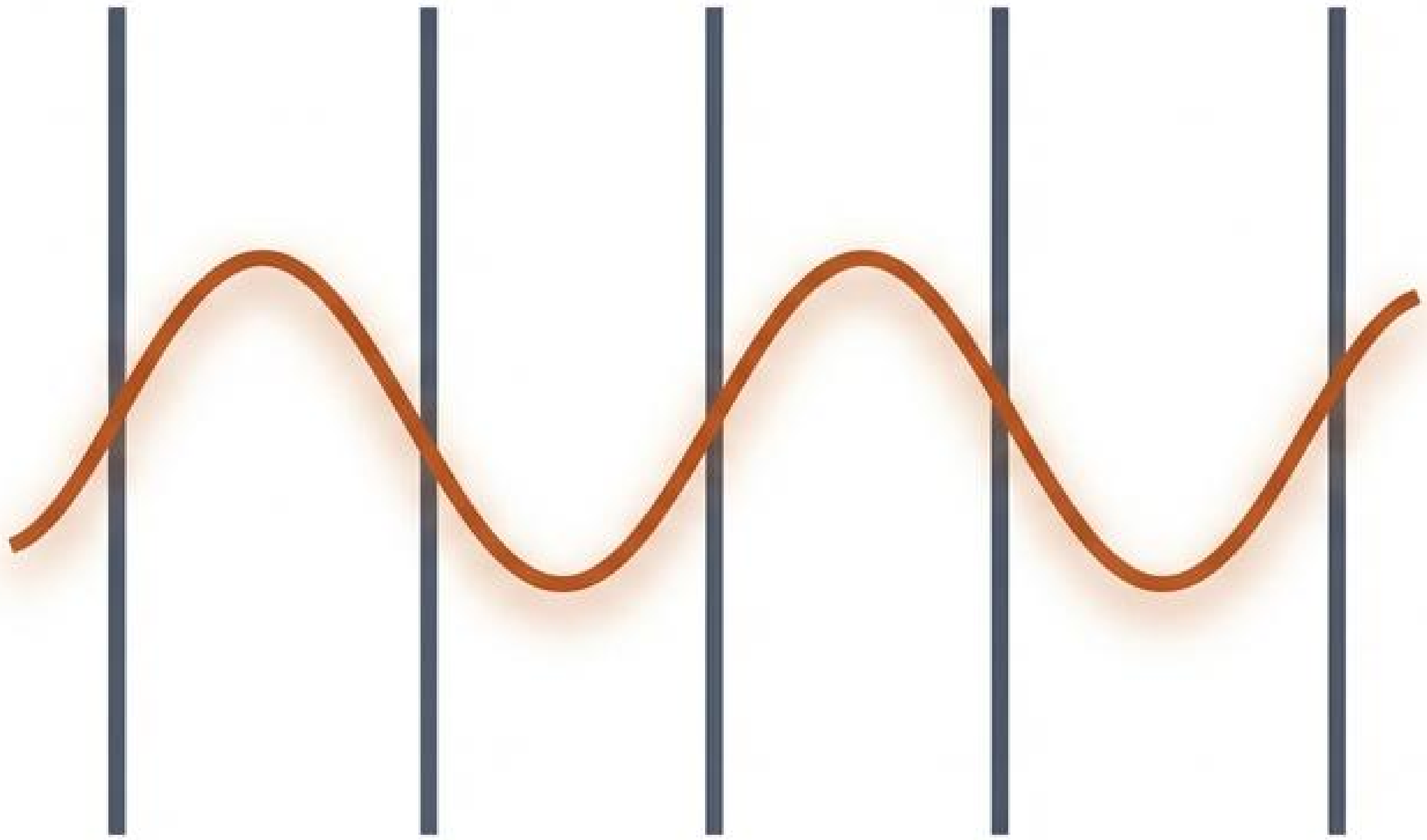
2024–2025: System 2 Thinking & Agents



System 2 Thinking: Moving from fast, intuitive responses to slow, deliberate reasoning.

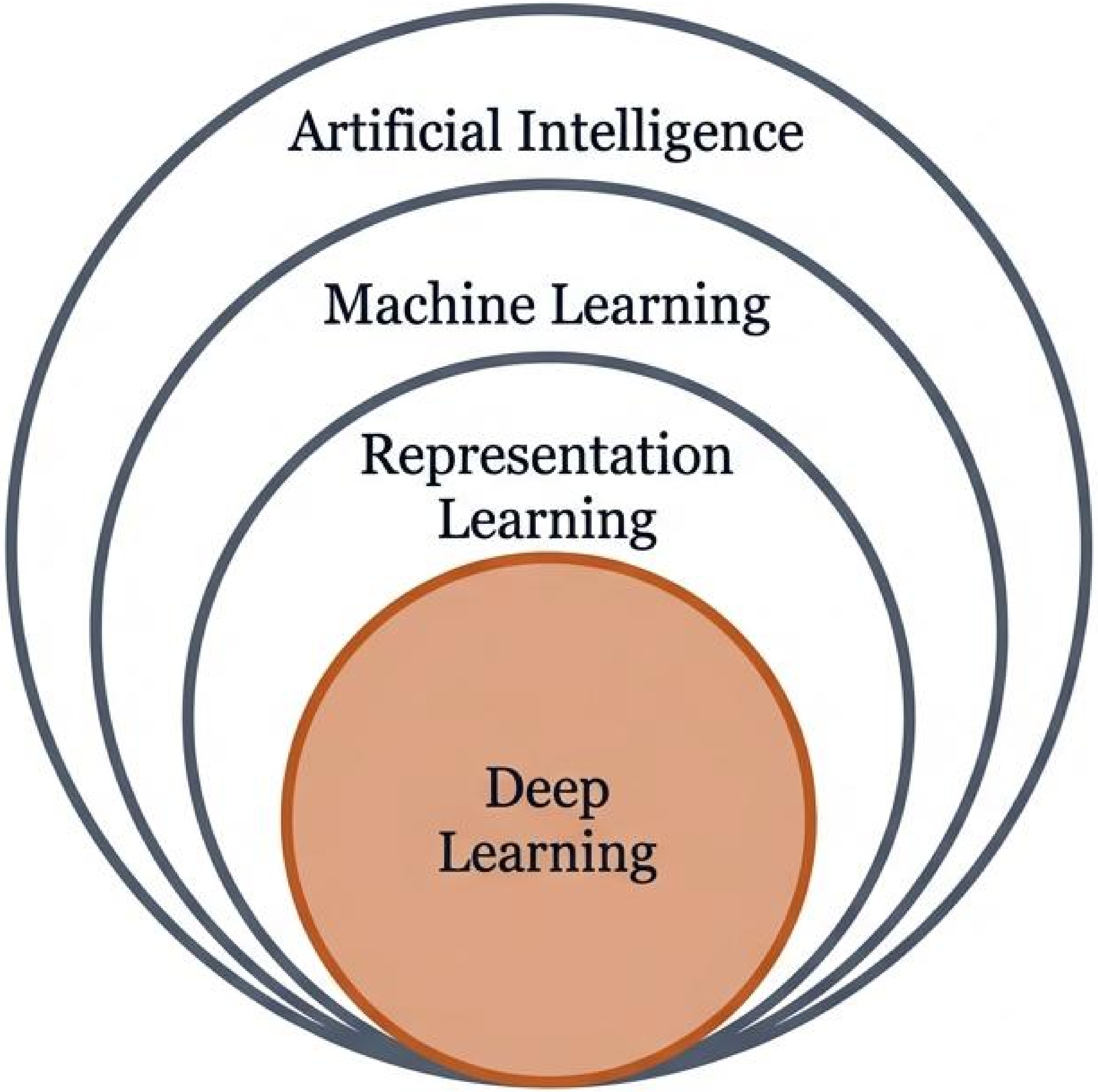
Agents: AI systems that plan, backtrack, self-correct, and execute multi-step tasks autonomously.

Key Concepts in Deep Learning



From Representation Learning to Hierarchical Feature Extraction

The Deep Learning Hierarchy



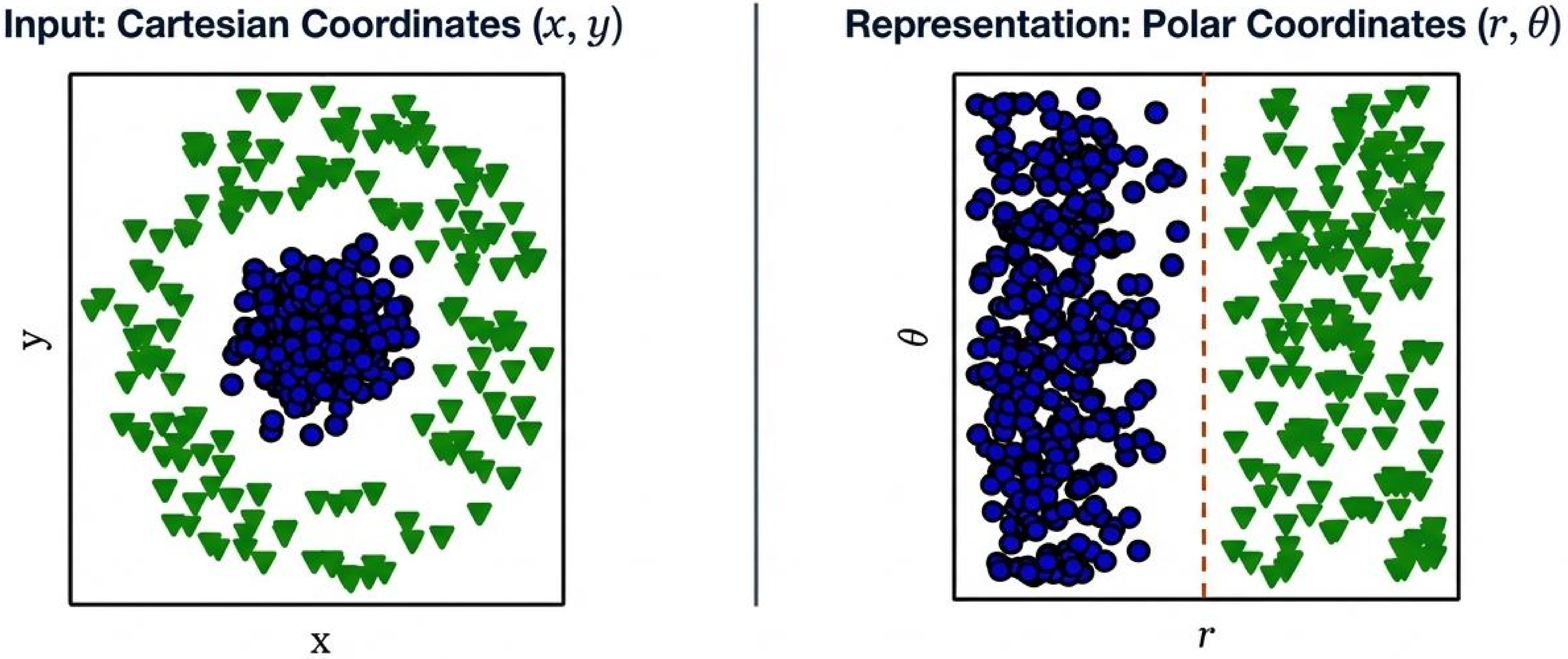
Representation Learning

The technique of transforming raw data into a format that is easier for a machine to understand. It replaces manual feature engineering with automated discovery.

The 'Deep' Factor

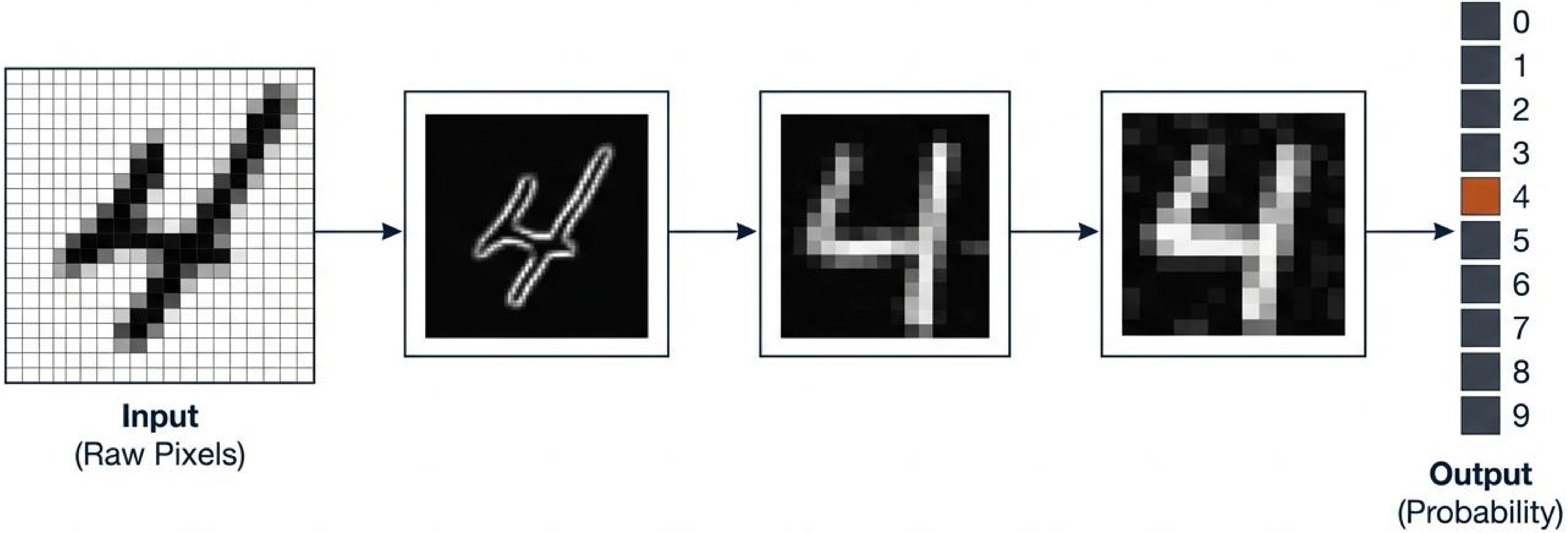
Depth refers to the number of successive layers of representation. While shallow learning uses one or two layers, modern deep learning involves tens or hundreds, all learned automatically from training data.

Why Representation Matters: Disentangling Data



Deep learning acts as a multi-stage information distillation process. It transforms messy, tangled inputs into ordered representations that clarify the solution.

The Filter Analogy



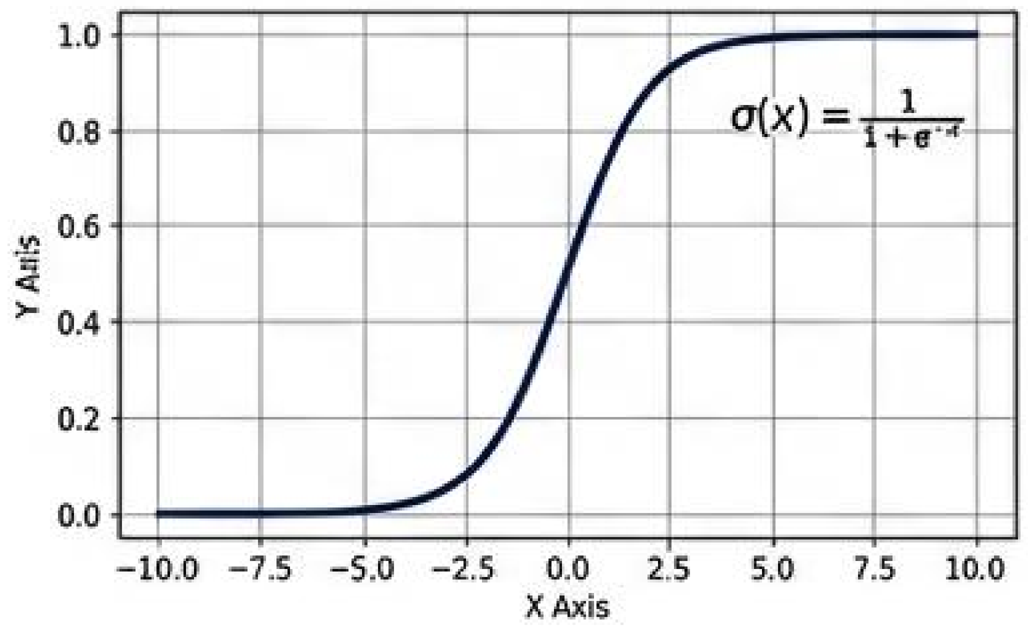
The network functions as successive filters, purifying information layer by layer.

The input is transformed into representations that are increasingly different from the original image and increasingly informative about the final result.

The Mechanics: Activation Functions

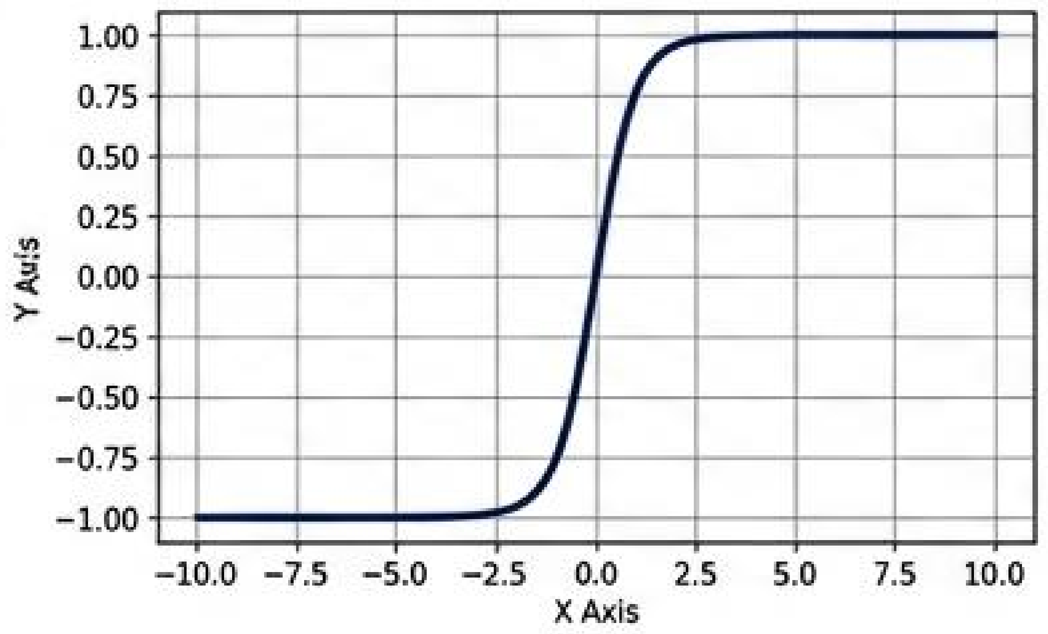
Without non-linearity, a neural network is just linear regression.

Sigmoid



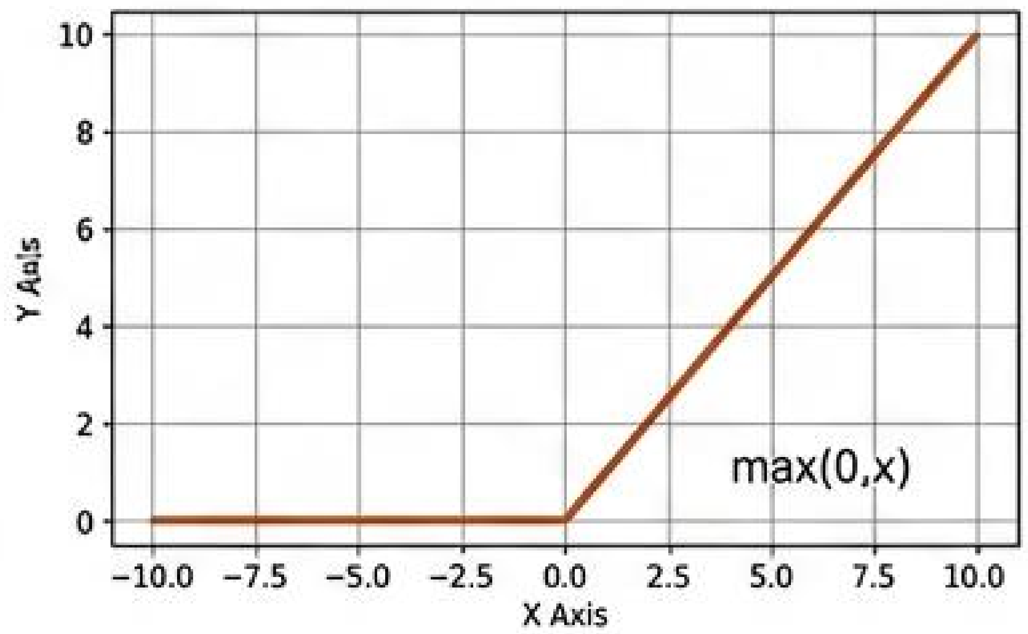
- Historical standard
- S-curve shape (0 to 1)
- Cons: Suffers from vanishing gradients; not zero-centred

Tanh



- Zero-centred
- Cons: Still suffers from vanishing gradients

ReLU (Rectified Linear Unit)



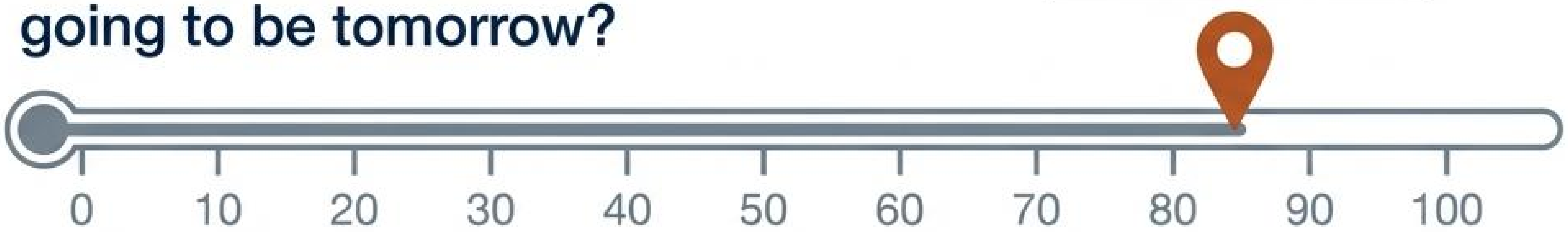
- Modern standard
- Computationally efficient
- Solves vanishing gradients
- Cons: Not zero-centred

Architecture Defines the Task

Regression

What is the temperature going to be tomorrow?

PREDICTION 84°

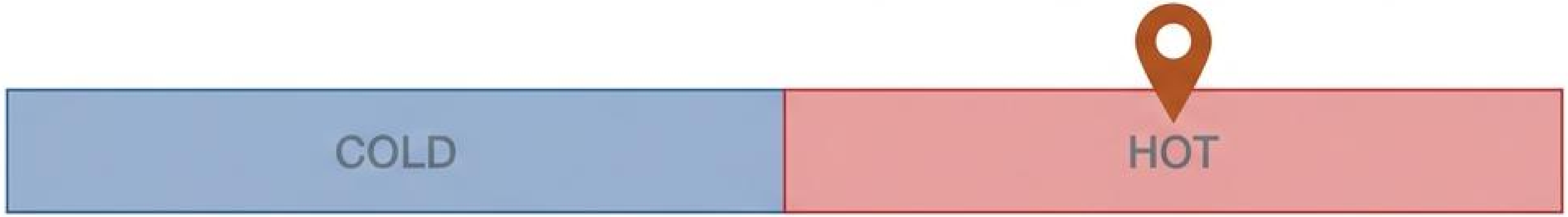


Predicting a continuous numerical value.

Classification

Will it be Cold or Hot tomorrow?

PREDICTION HOT



Assigning an input to a specific discrete category.

The Scorecard: Loss Functions

Quantifying the gap between Prediction and Ground Truth.

Regression Task

Penalizes large errors

$$\mathrm{MSE} = \frac{1}{N}\sum (t_i - s_i)^2$$

Ground Truth

Prediction

Mean Squared Error measures the average squared difference.

Classification Task

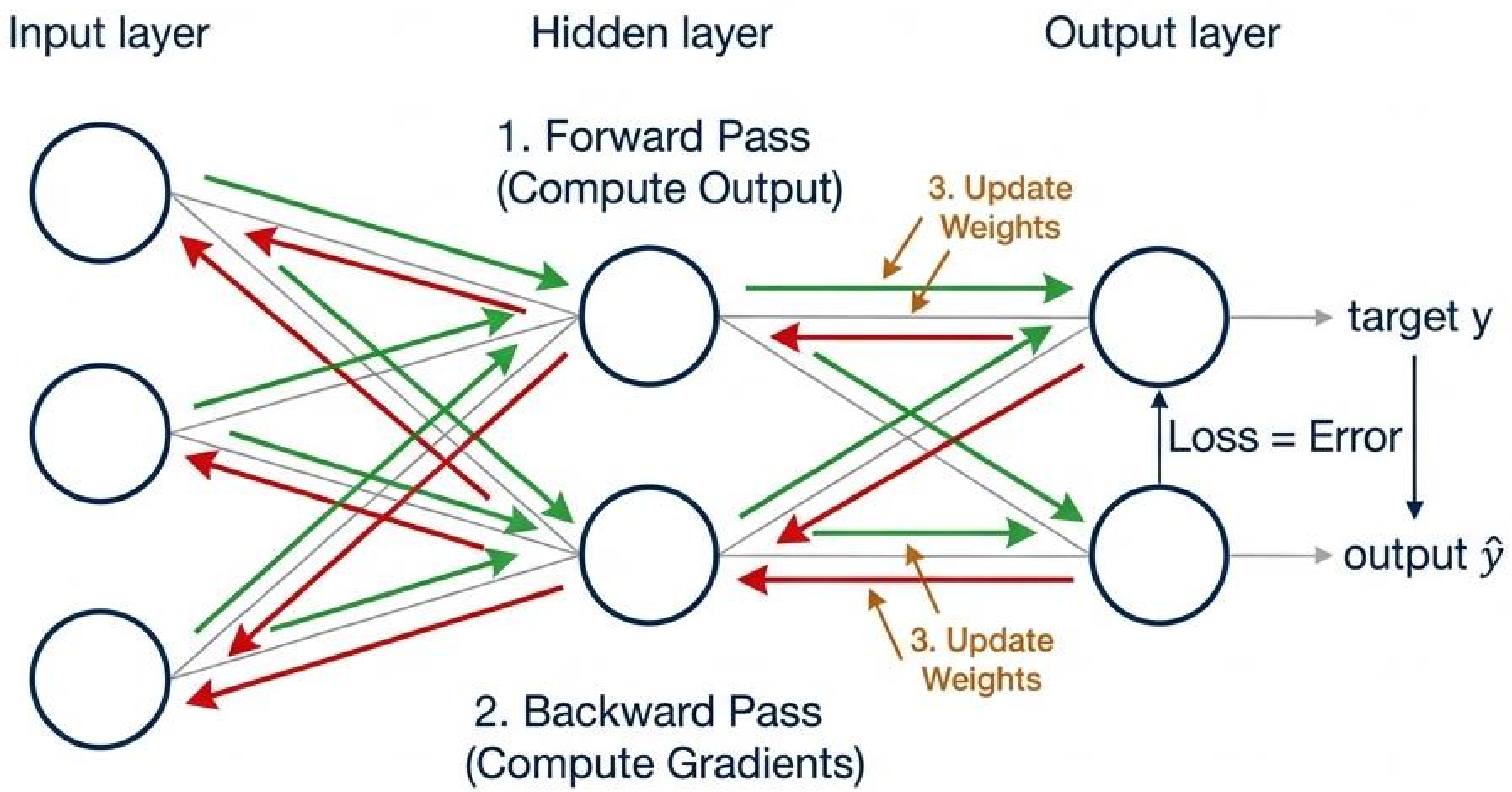
$$\mathrm{CE} = -\sum t_i \log(s_i)$$

Actual Class Label

Predicted Probability

Cross Entropy Loss measures the divergence between two probability distributions.

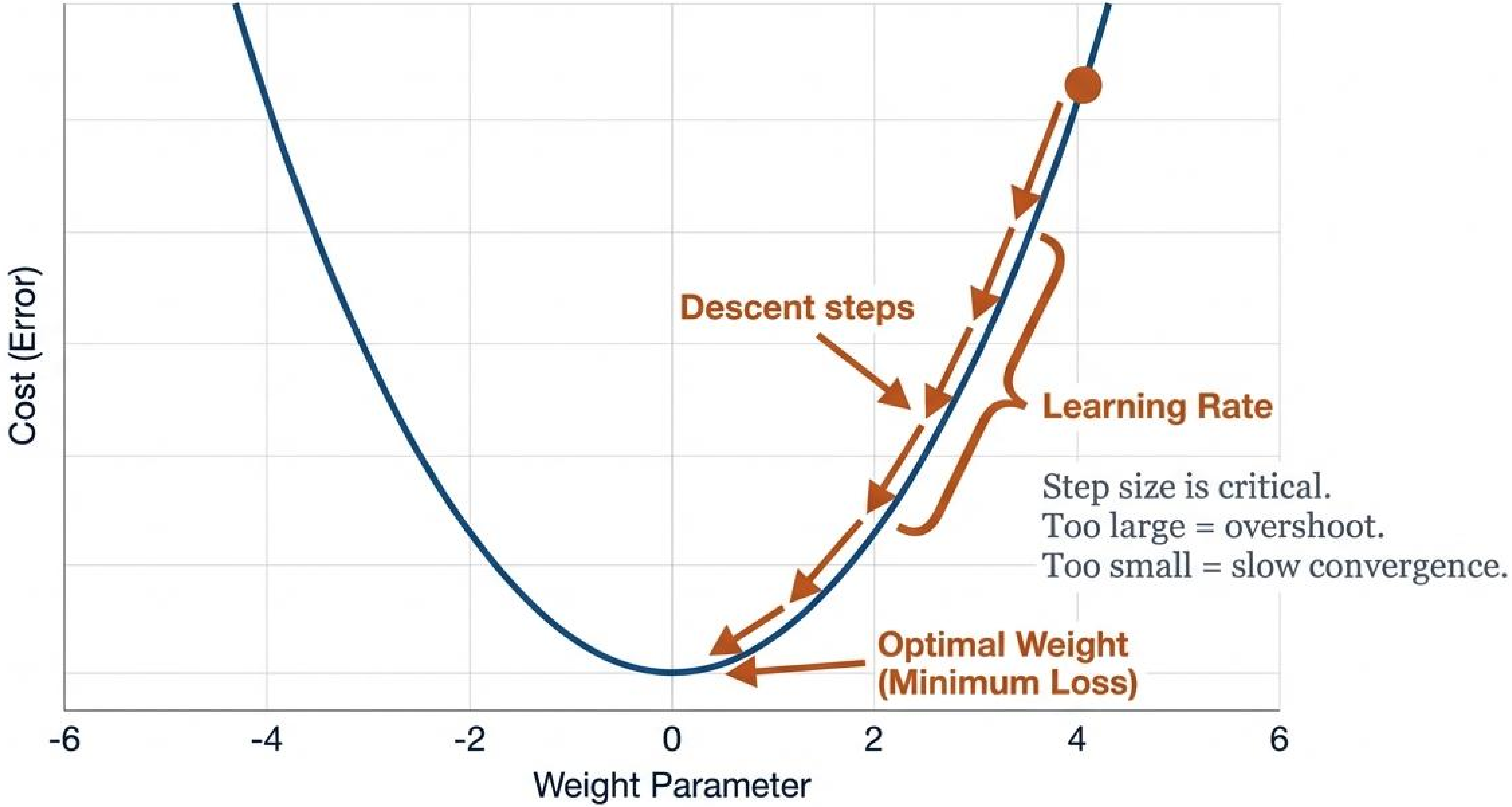
The Learning Cycle: Backpropagation



Workflow:

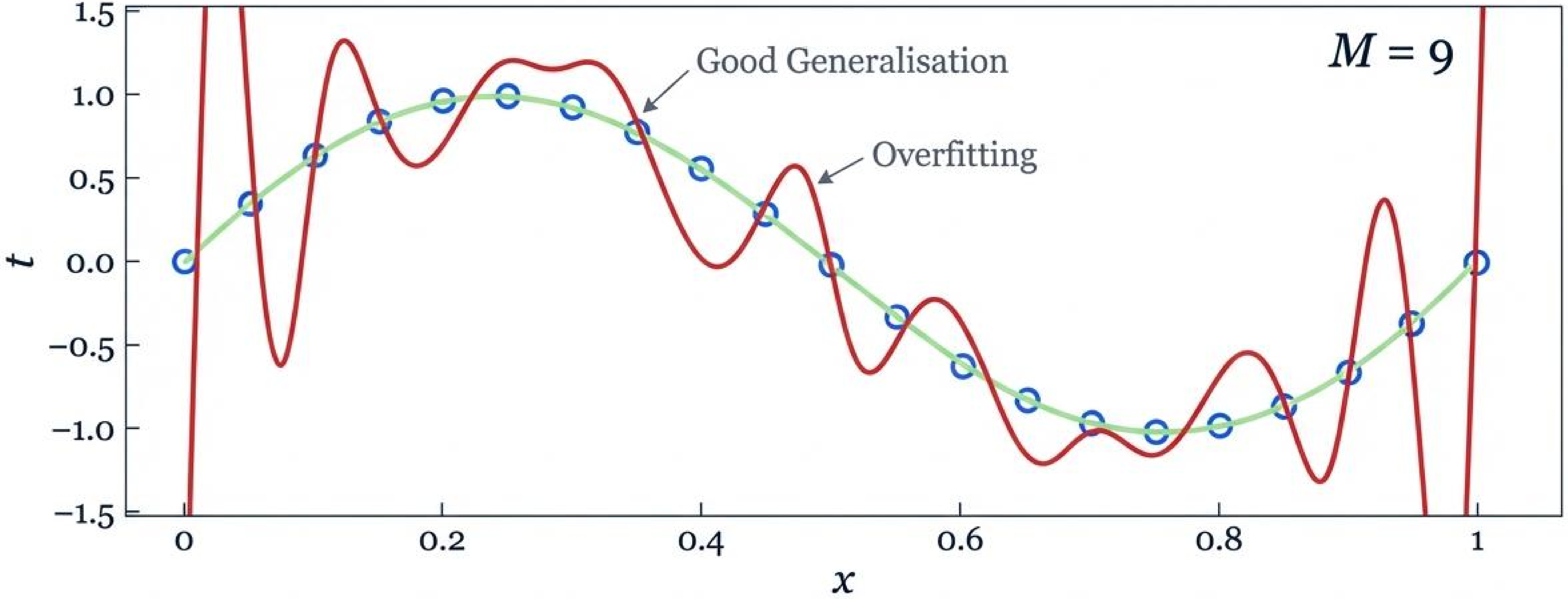
1. Input travels forward to generate a prediction.
2. Error is calculated against the target.
3. Gradients (responsibility for error) are propagated backward.
4. Weights are adjusted to reduce future error.

Optimisation: Stochastic Gradient Descent (SGD)



The Challenge: Overfitting

Memorisation vs. Understanding

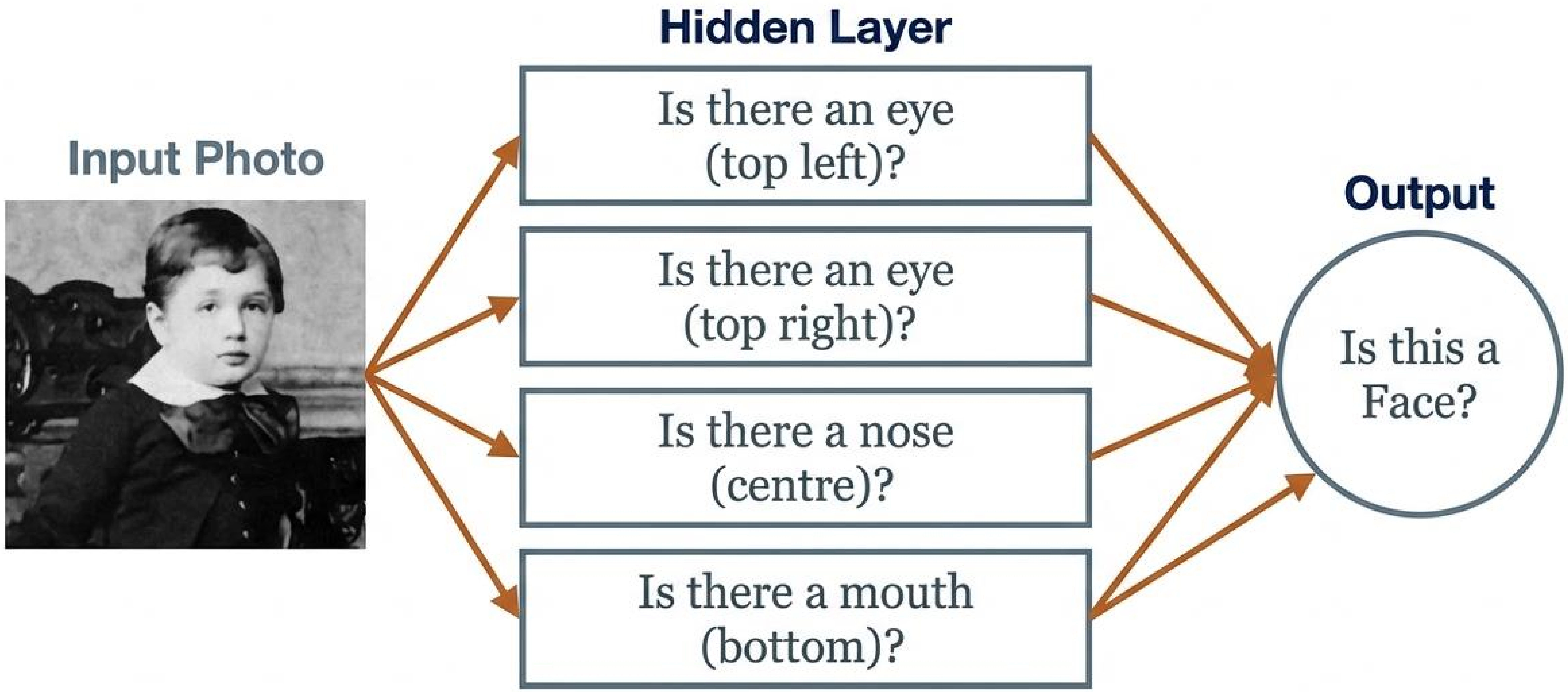


The overfitting model has zero error on the training data but fails to capture the true underlying pattern. It has memorised the noise rather than the signal.

The Solution: Regularisation

Regularisation techniques penalise complexity, forcing the model to stay in the sweet spot where it generalises to unseen test data.

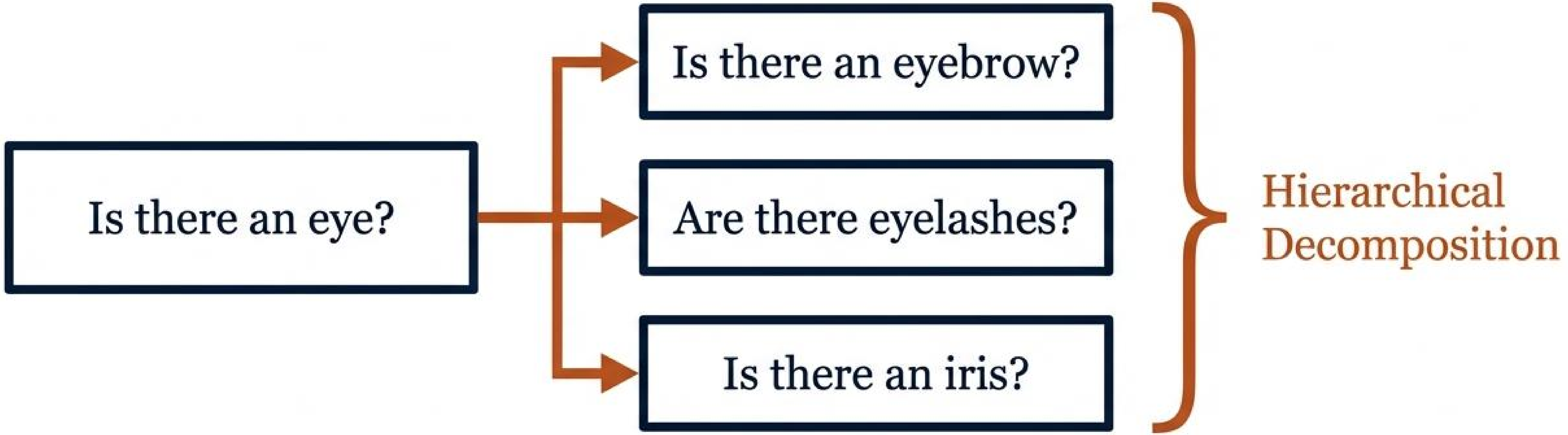
Intuition: The Hierarchy of Features



Complex recognition tasks are decomposed into simpler, local sub-problems.

Turtles All The Way Down

Deep Sub-Networks

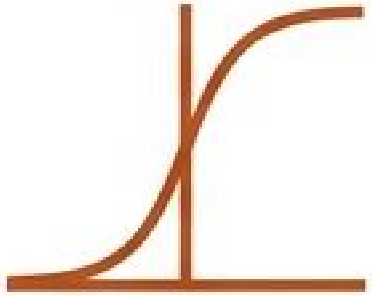


Deep Learning automates this **hierarchy**. We start with **pixels**, form **edges**, form **shapes (eyes)**, and finally form **concepts (faces)**. The depth allows the network to answer complex questions by solving thousands of tiny, simple ones.

Summary: The Automated Discovery of Structure

Representation

Data is transformed to make it linearly separable and useful.



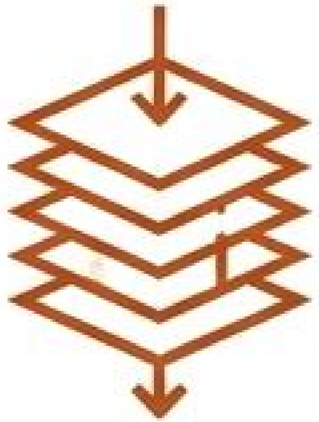
Mechanics

Activation, Loss, and Backpropagation work together to tune the engine.

&

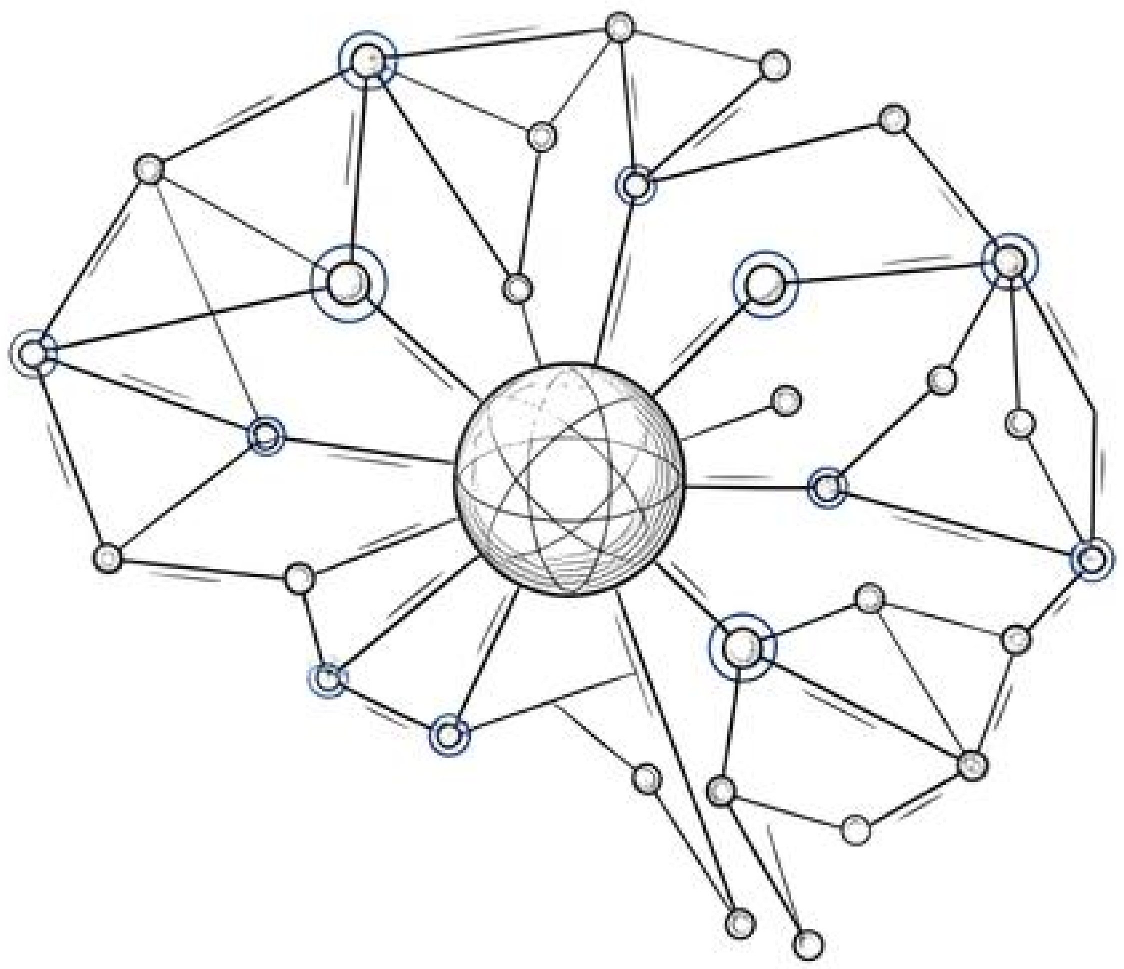
Generalisation

Regularisation prevents memorisation, ensuring real-world performance.



Hierarchy

Deep networks build complex concepts from simple foundations.



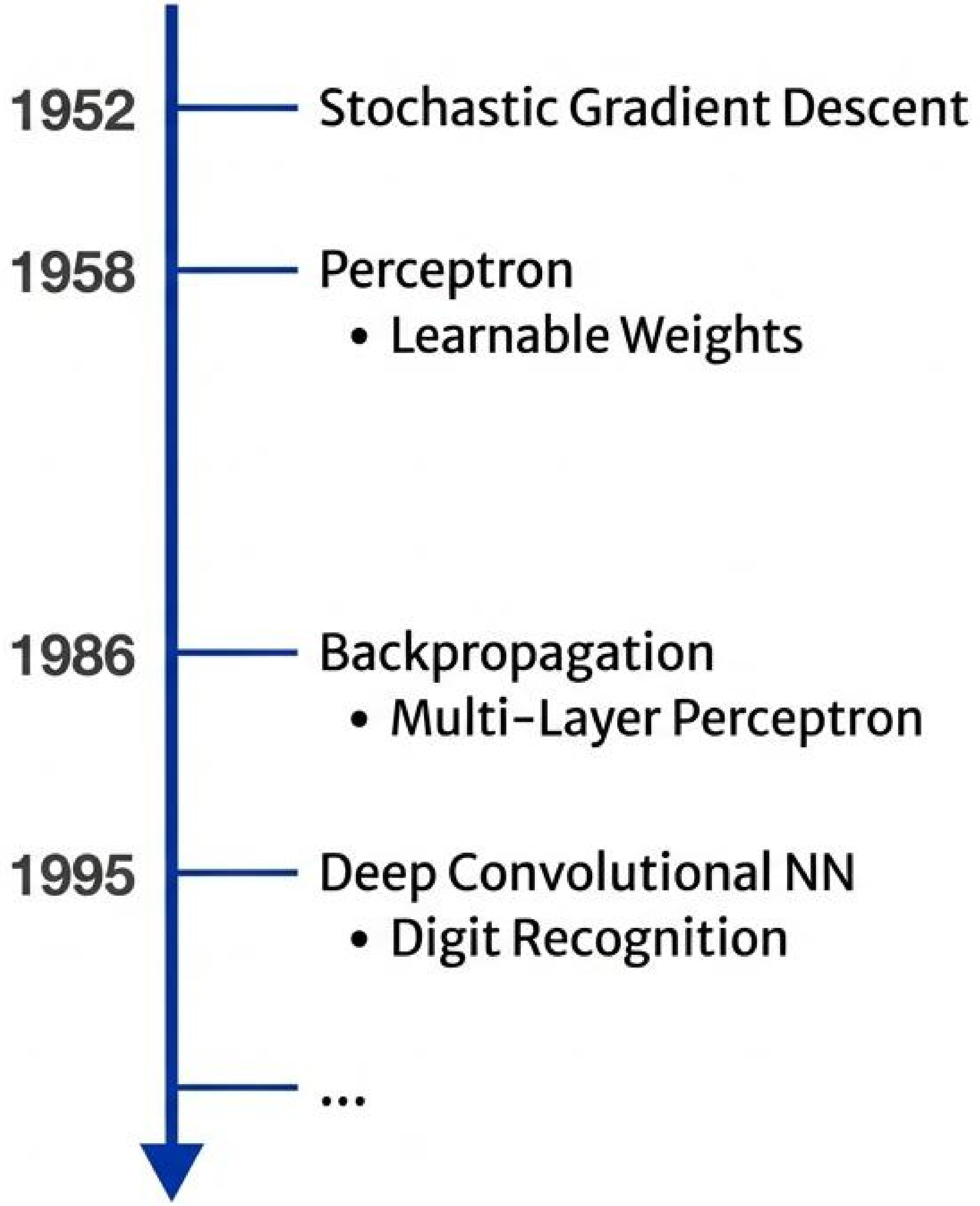
Deep Learning: The Renaissance of Intelligence

From Hand-Engineered Features to End-to-End Representation

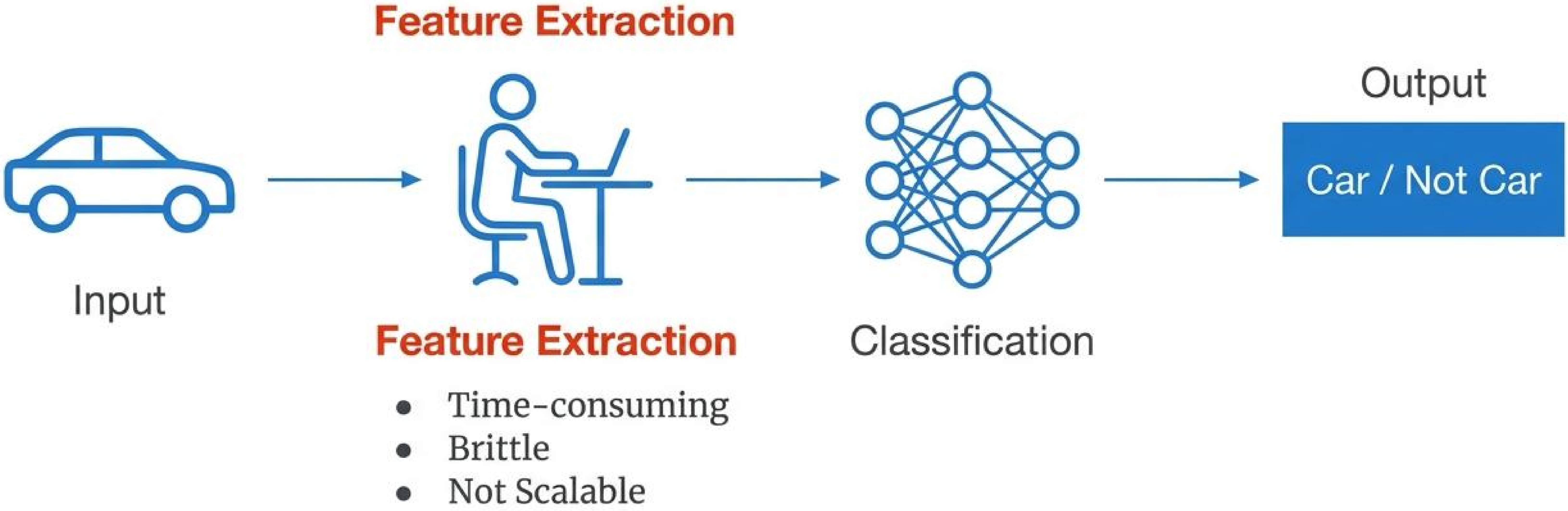
A Self-Guided Reading Deck

Decades of Mathematical Theory Waiting for Capability

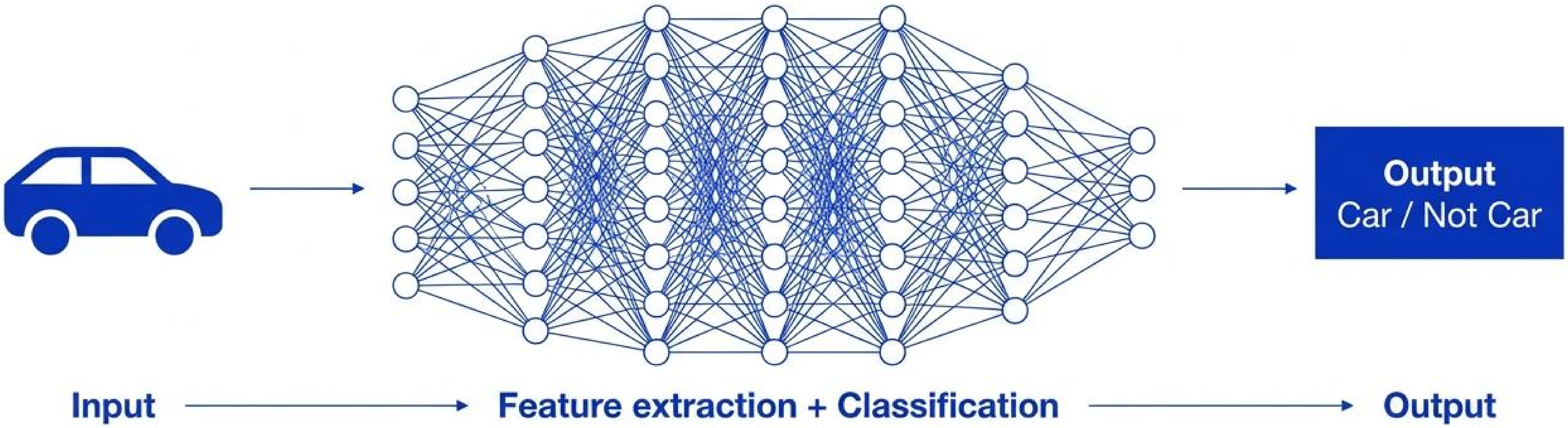
Deep Learning is not a sudden invention, but a resurgence of established ideas. The mathematical foundations were laid over half a century ago but remained dormant during the ‘AI Winters’ until until recent breakthroughs.



The Bottleneck of Traditional Machine Learning



The Paradigm Shift: End-to-End Representation



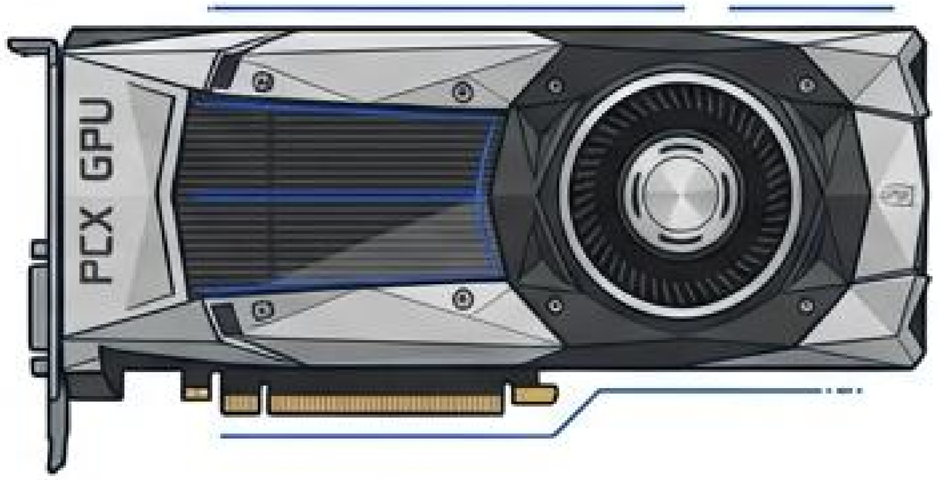
Deep Learning changes the fundamental workflow. We no longer tell the computer *what* to look for. The model learns the features itself directly from the data.

The Perfect Storm: Why the Resurgence is Happening Now

Big Data

- Large Datasets (ImageNet, Wikipedia)
- Easier Collection & Storage

Hardware



- Graphics Processing Units (GPUs)
- Massively Parallelisable

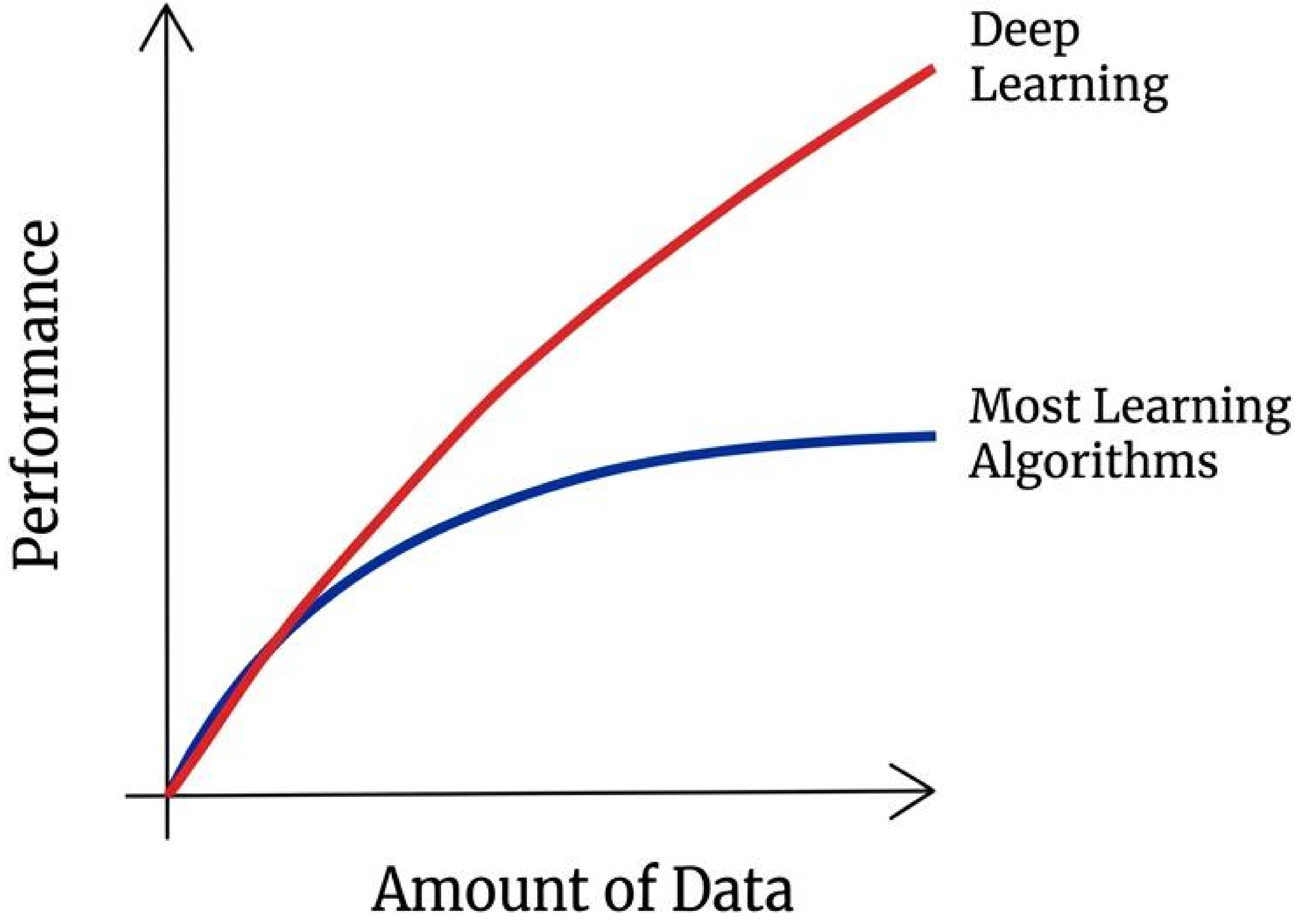
Software

- Improved Techniques
- New Models
- Accessible Toolboxes

Hardware: The Engine of Parallelisation

CPU (Central Processing Unit)
Serial, general-purpose. Everyone has one. Great for sequential logic, slow for training.
GPU (Graphics Processing Unit)
Parallelisable. Originally for graphics, now the workhorse of Deep Learning. High throughput.
TPU (Tensor Processing Unit)
Custom ASIC by Google. Specialised specifically for machine learning with low precision.

The Unreasonable Effectiveness of Data

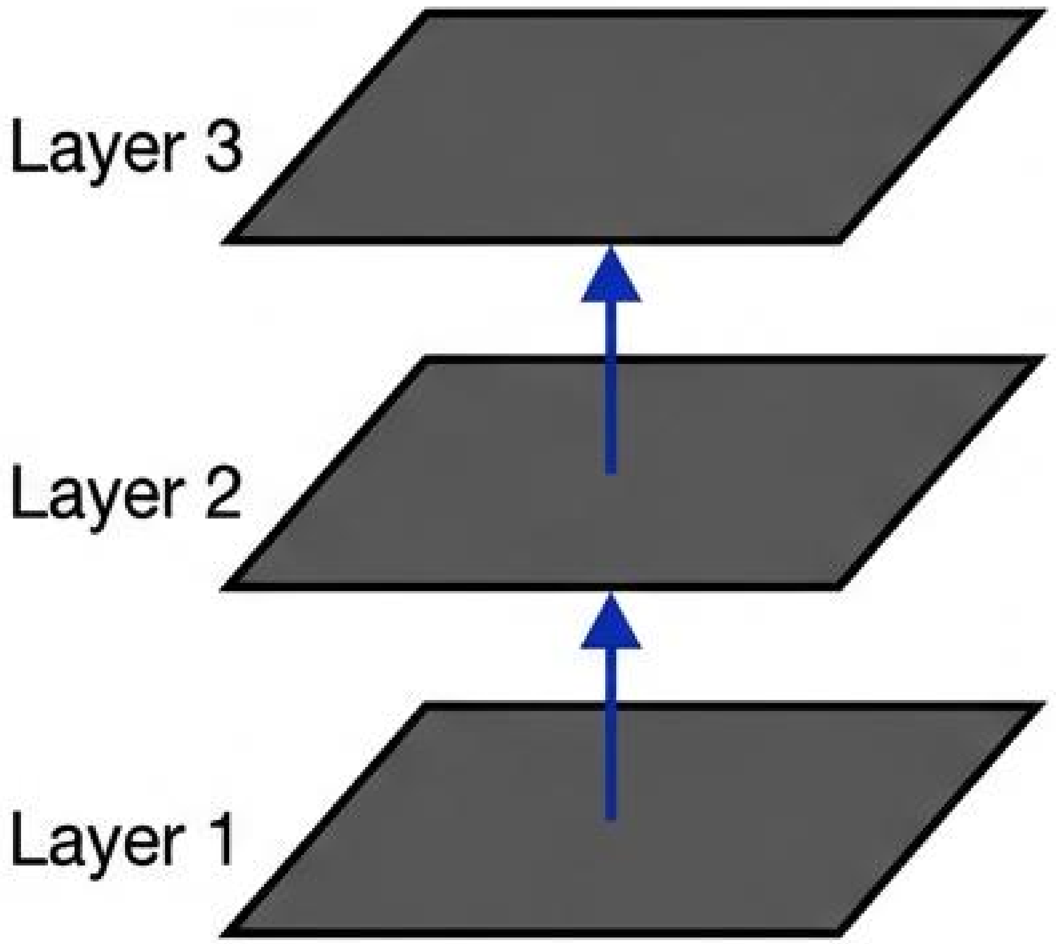


Traditional algorithms have a performance ceiling; adding more data eventually yields diminishing returns.

Deep Learning models are “data hungry”—their performance continues to improve as the dataset size increases.

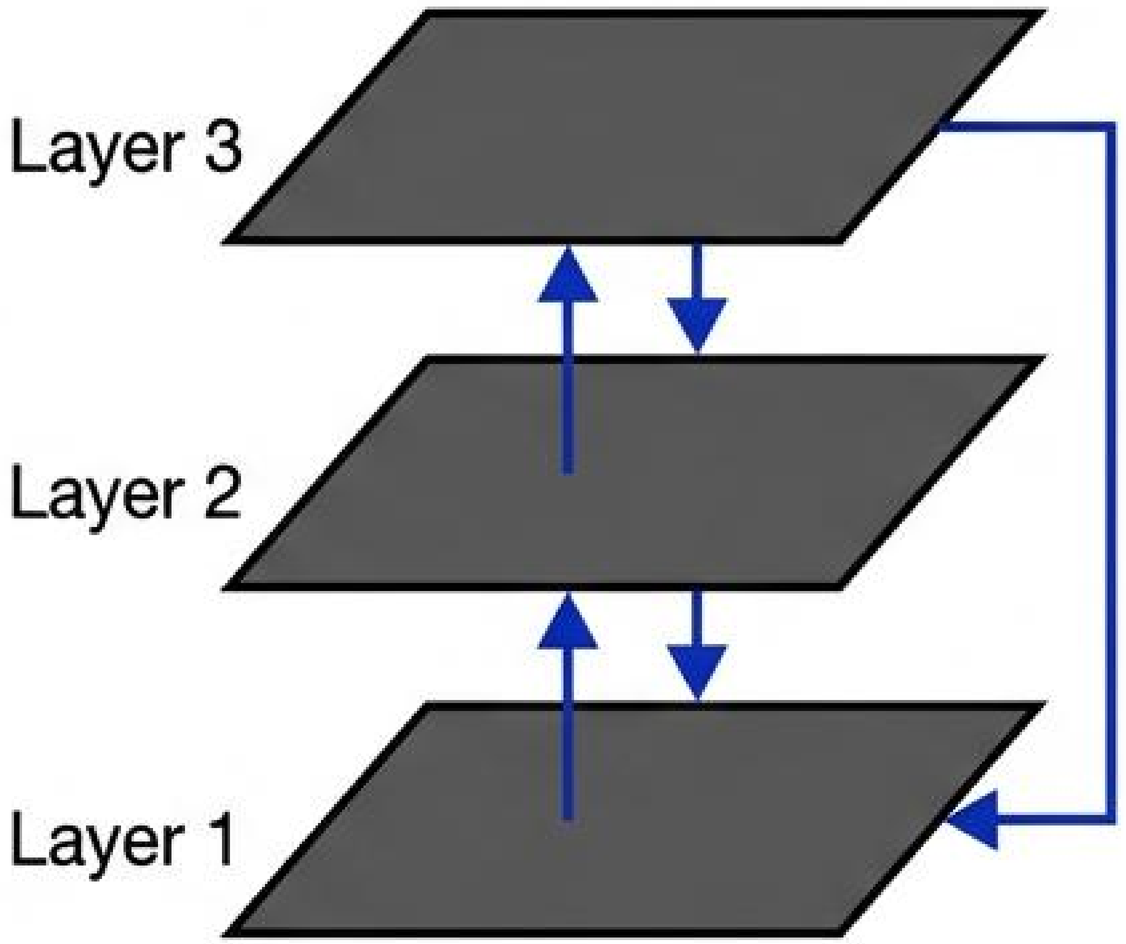
The Mechanics of Depth: Joint vs. Greedy Learning

Greedy Learning (Shallow)



Layers learned one by one in succession.

Joint Learning (Deep)

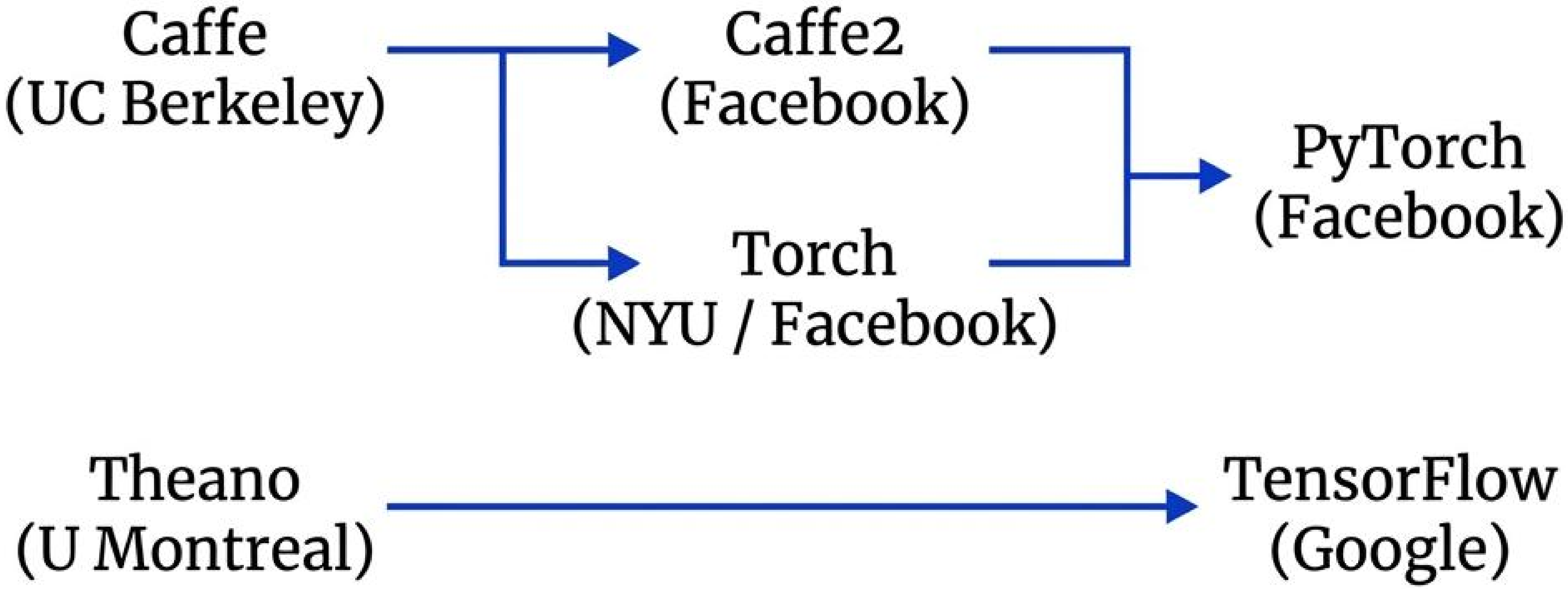


All layers learned simultaneously.
Each layer is updated to satisfy the needs of the layer above AND the layer below.

The optimal first layer in a deep model is NOT the same as the optimal first layer in a shallow model.

Modern Academic Industrial

From Academic Fragmentation to Industrial Consolidation



Other Players
PaddlePaddle (Baidu)
MXNet (Amazon)
CNTK (Microsoft)
JAX (Google)
And others...

The Great Convergence: TensorFlow and PyTorch

TensorFlow

PYTORCH

TensorFlow 2.0

- Keras integration + promotion
- TensorFlow Lite (Mobile)
- TensorFlow Serving (Production)
- TensorFlow.js

PyTorch 1.3

- Dynamic graphs (Pythonic)
- TorchScript
- PyTorch Mobile
- TPU Support

```
>>> print 'Goodbye World'
```

Python 2 support ended Jan 1, 2020. The ecosystem has modernized.

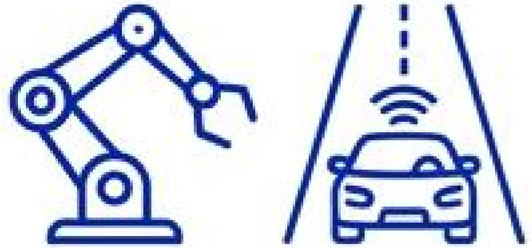
Moving from Theory to Ubiquity

Perception

- Face recognition
- Image classification
- Handwriting transcription

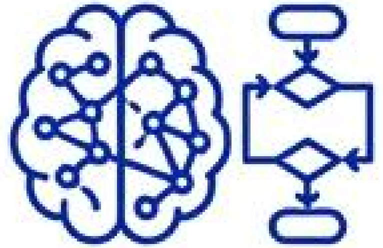
Communication

- Speech recognition
- Text-to-speech
- Machine translation



Action

- Autonomous driving (lane keeping)
- Robotics



Decision Making

- Medical diagnosis
- Digital assistants
- Game playing (Deep RL)

Deep Learning: The Essential Summary

What is it?	Extracting useful patterns from data.
How does it work?	Neural Networks + Optimisation.
Why now?	Convergence of Data, Hardware, and Investment.
The Tooling:	Python + TensorFlow / PyTorch.
The Hard Part:	Asking Good Questions + Curating Good Data.

Courses:

- ►Stanford CS231n
  https://www.youtube.com/playlist?list=PLoROMvodv4rOmsNzYBMe0gJY2XS8AQg16
- ►Stanford CS230
  https://www.youtube.com/playlist?list=PLoROMvodv4rNRRGdS0rBbXOUGA0wjdh1X
- ►McAllester (TTI-C): Fundamentals of Deep Learning
  http://mcallester.github.io/ttic-31230/Fall2020/
- ►Kolterl, Chen (CMU): Deep Learning Systems
  https://dlsyscourse.org/lectures/
- ►Leal-Taixe, Niessner (TUM): Introduction to Deep Learning
  http://niessner.github.io/I2DL/
- ►Grosse (UoT): Intro to Neural Networks and Machine Learning
  http://www.cs.toronto.edu/~rgrosse/courses/csc321_2018/
- ►Abbeel, Chen, Ho, Srinivas (Berkeley): Deep Unsupervised Learning
  https://sites.google.com/view/berkeley-cs294-158-sp20/home

Tutorials:

- ► Python NumPy Tutorial
  https://cs231n.github.io/python-numpy-tutorial/
- ► The Python Tutorial
  https://docs.python.org/3/tutorial/
- ► NumPy Quickstart
  https://numpy.org/devdocs/user/quickstart.html
- ► PyTorch Tutorial
  https://pytorch.org/tutorials/

Frameworks / IDEs:

- ► Visual Studio Code
  https://code.visualstudio.com/
- ► Google Colab
  https://colab.research.google.com

Acknowledgements

Various contents in this presentation have been taken from different books, lecture notes, and the web. These solely belong to their owners and are here used only for clarifying various educational concepts. Formatted has been done using NotebookLM. Any copyright infringement is *not* intended.