Maths of AI
Introduction

Instructor - Simon Lucey

THE UNIVERSITY of ADELAIDE

AUSTRALIAN INSTITUTE FOR MACHINE LEARNING

Today

- **About the Course**
- What is AI?
- The Rise of Deep Learning
- Why Math in AI Matters

Instructors

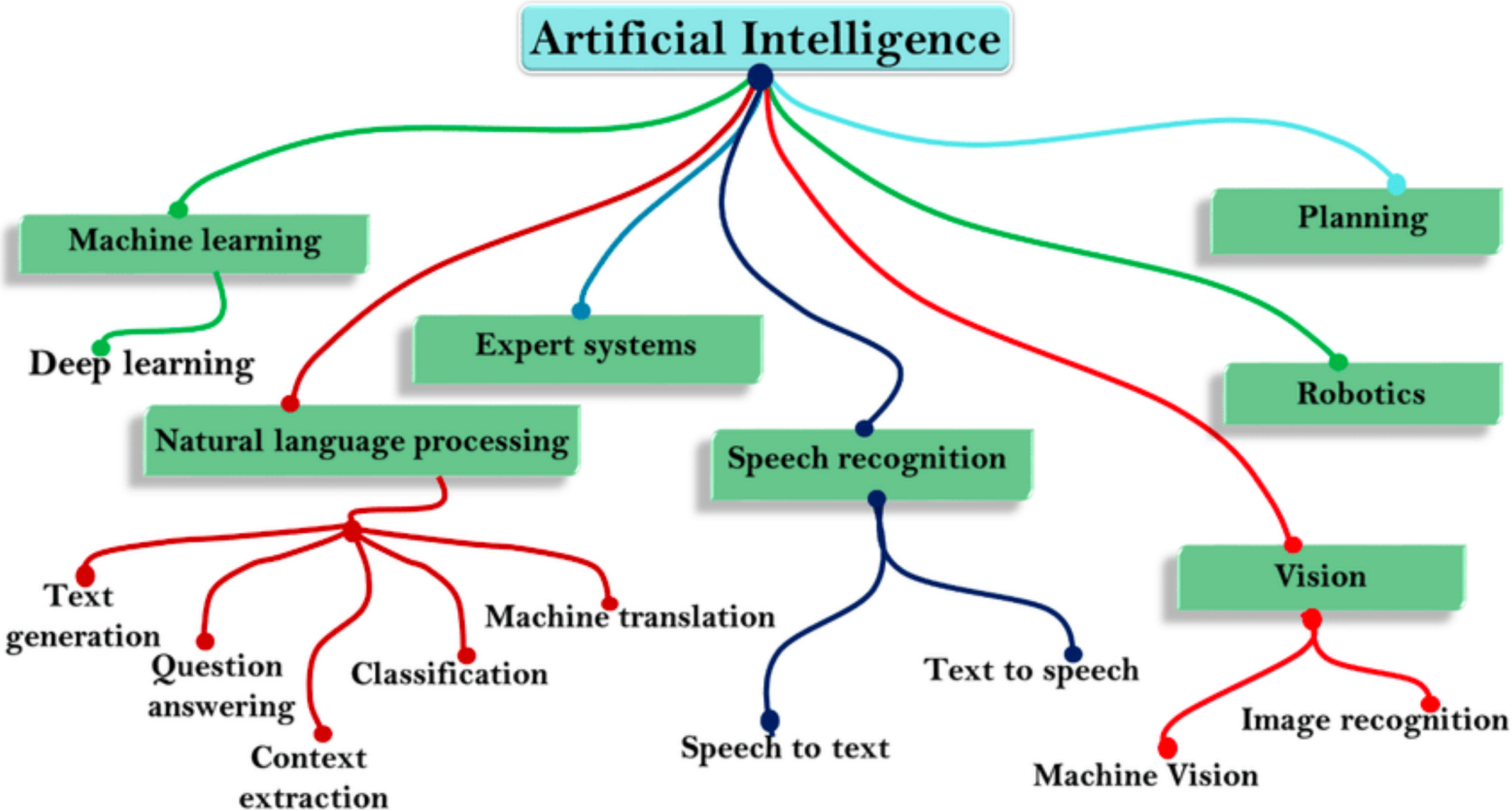
Prof. Simon Lucey

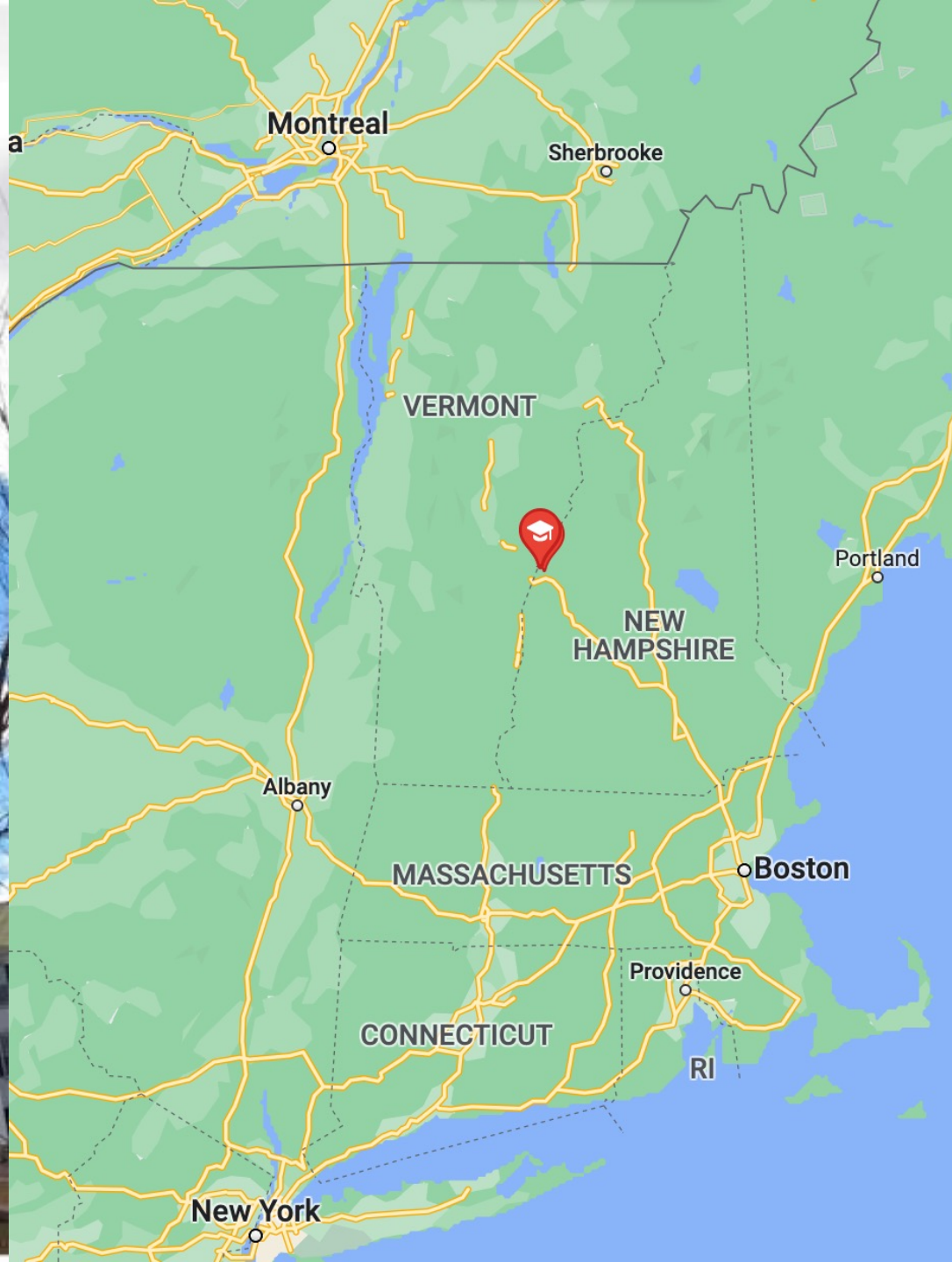
Prof. Matthew Roughan

Prof. Lewis Mitchell

Today

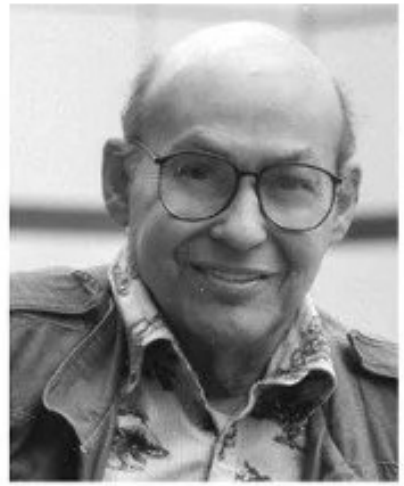
- About the Course
- **What is AI?**
- The Rise of Deep Learning
- Why Math in AI Matters


Artificial Intelligence
Machine learning
Deep learning
Expert systems
Natural language processing
Text generation
Question answering
Context extraction
Classification
Machine translation
Speech recognition
Speech to text
Text to speech
Planning
Robotics
Vision
Image recognition
Machine Vision

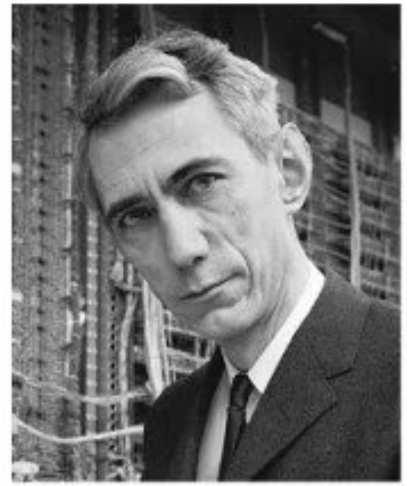

Montreal
Sherbrooke
VERMONT
Portland
NEW
HAMPSHIRE
Albany
MASSACHUSETTS
Boston
Providence
CONNECTICUT
RI
New York

1956 Dartmouth Conference: The Founding Fathers of AI

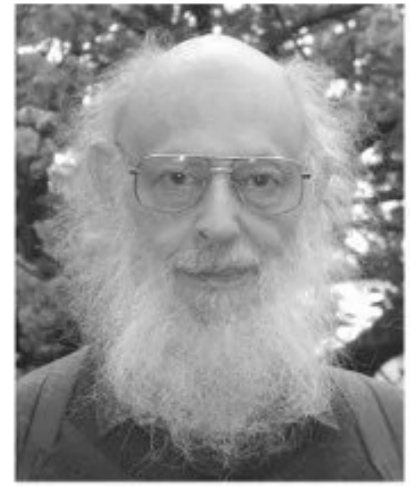
John MacCarthy



Marvin Minsky



Claude Shannon

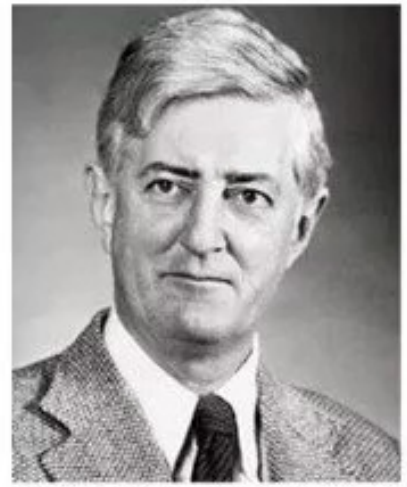


Ray Solomonoff

Alan Newell

Herbert Simon

Arthur Samuel

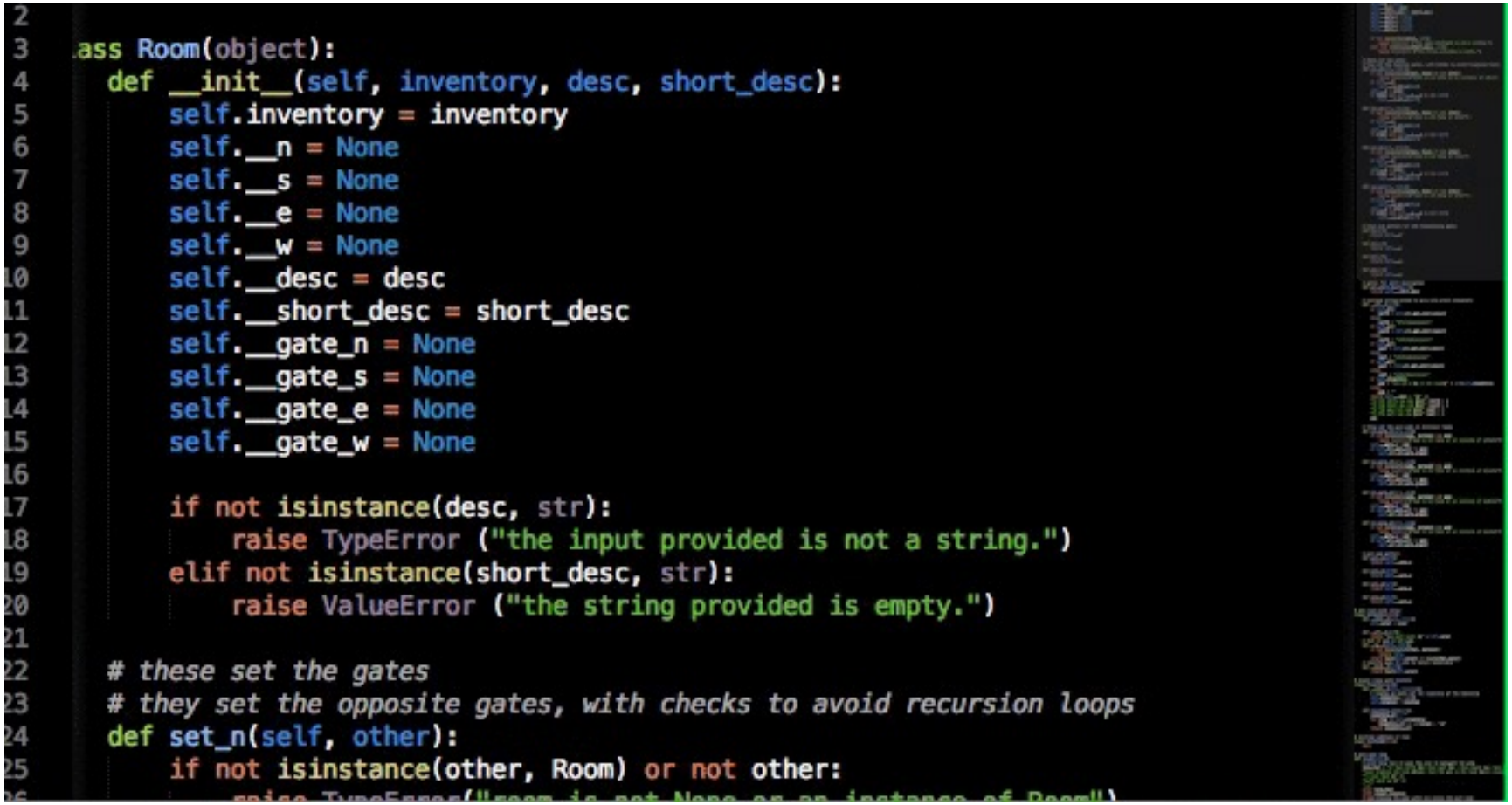


Oliver Selfridge

Nathaniel Rochester

Trenchard More

Conventional programming

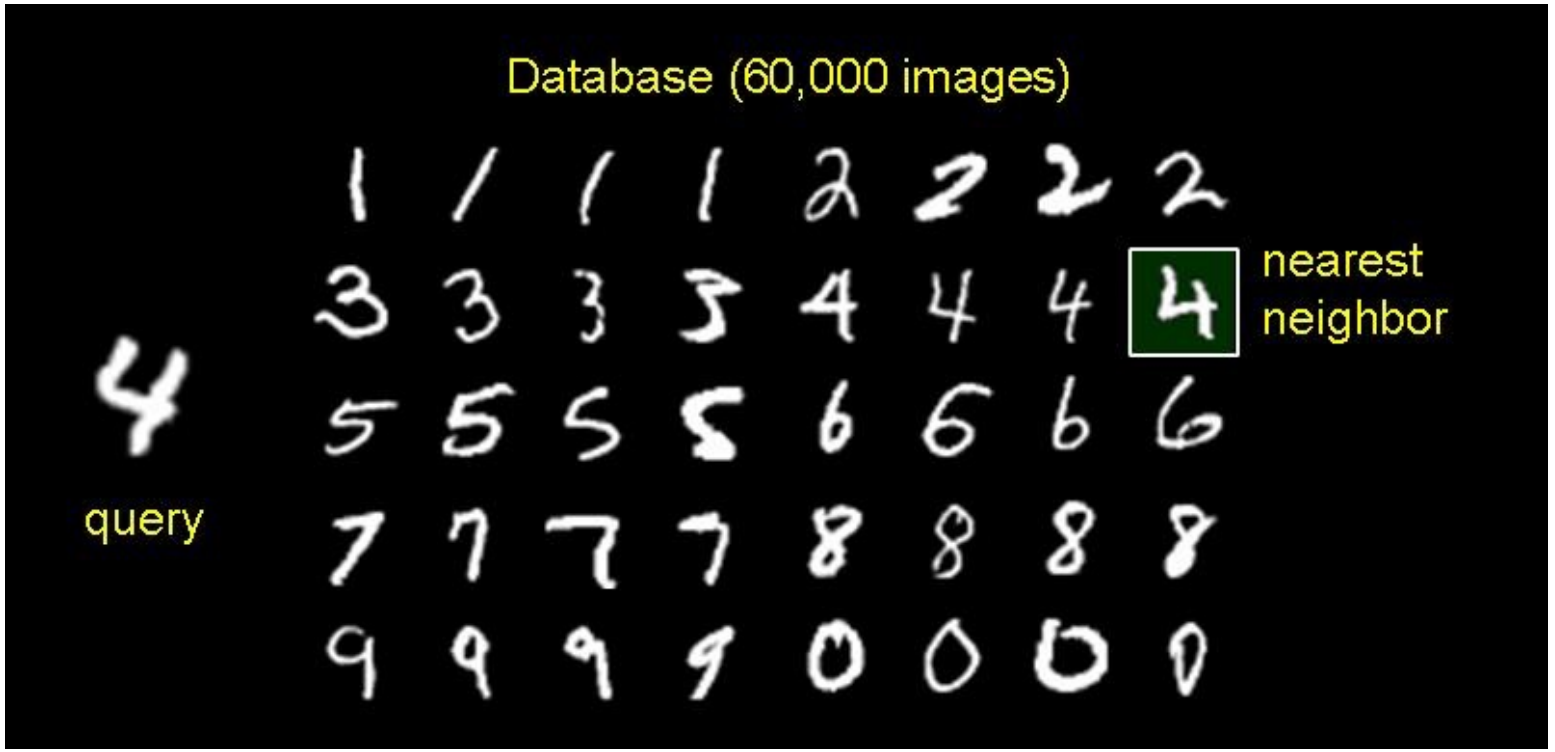


```
ass Room(object):
    def __init__(self, inventory, desc, short_desc):
        self.inventory = inventory
        self.__n = None
        self.__s = None
        self.__e = None
        self.__w = None
        self.__desc = desc
        self.__short_desc = short_desc
        self.__gate_n = None
        self.__gate_s = None
        self.__gate_e = None
        self.__gate_w = None

        if not isinstance(desc, str):
            raise TypeError ("the input provided is not a string.")
        elif not isinstance(short_desc, str):
            raise ValueError ("the string provided is empty.")

    # these set the gates
    # they set the opposite gates, with checks to avoid recursion loops
    def set_n(self, other):
        if not isinstance(other, Room) or not other:
```

Machine Learning



Perceptron

- Rosenblatt simulated the perceptron on a IBM 704 computer at Cornell in 1957.
- Input scene (i.e. printed character) was illuminated by powerful lights and captured on a 20x20 cadmium sulphide photo cells.
- Weights of perceptron were applied using variable rotary resistors.
- Often times referred to as the very first neural network.

“Frank Rosenblatt”

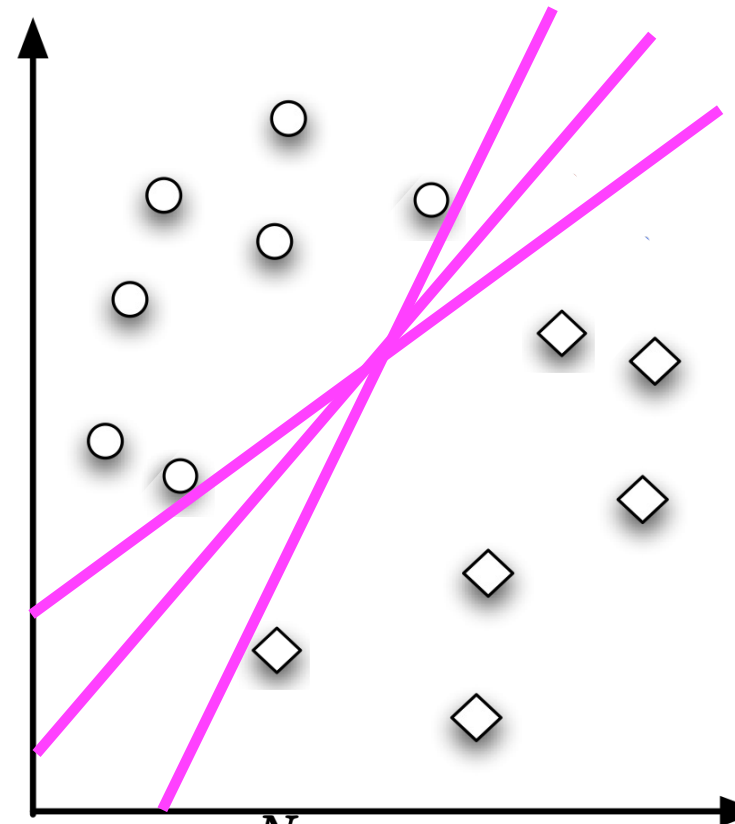
The New York Times

NEW NAVY DEVICE LEARNS BY DOING

July 8, 1958

"The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence... Dr. Frank Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the ets as mechanical space explorers"

Perceptron = Linear Discriminant



$t_i = +1$

$t_i = -1$

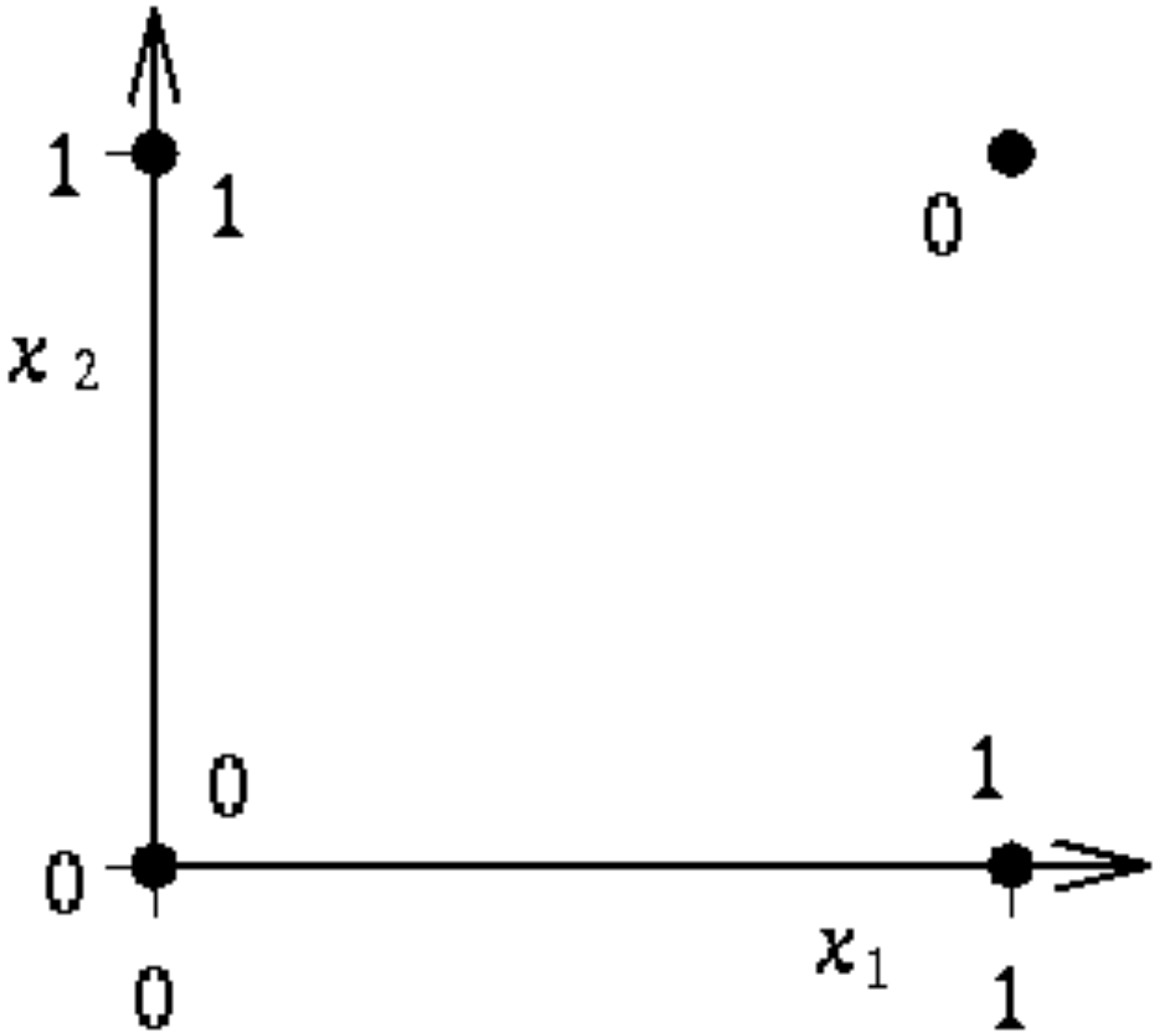
binary labels

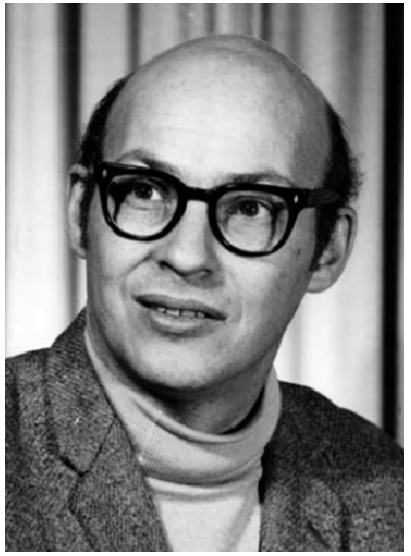
$\mathbf{x}_i = i$-th training example

$\mathbf{w}$ = weight vector

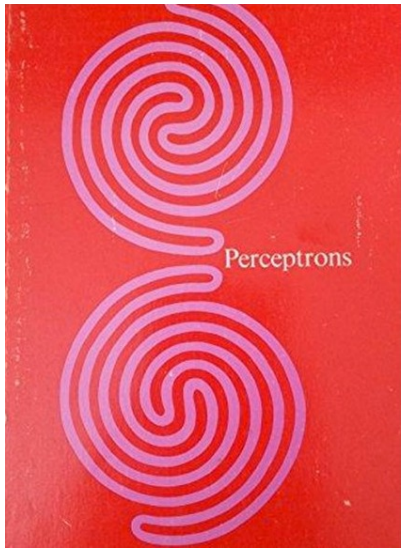
$$\arg\min_{\mathbf{w}} \sum_{n=1}^{N} \max(0, -t_n \mathbf{x}_n^T \mathbf{w})$$

XORs, Minsky and the AI Winter

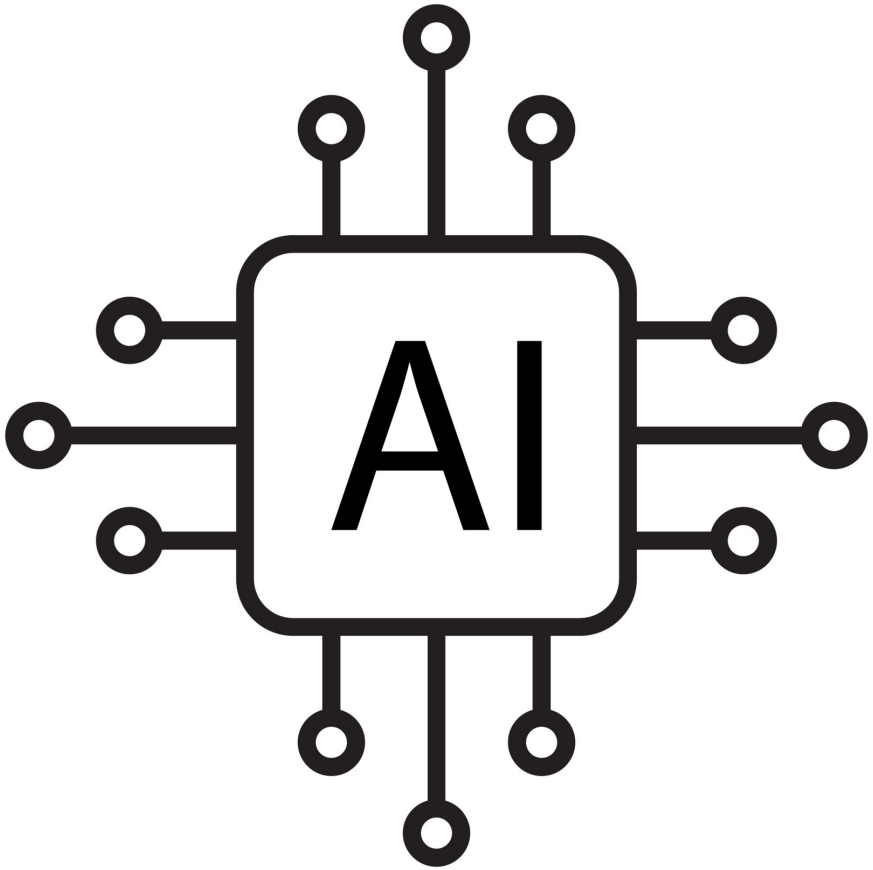




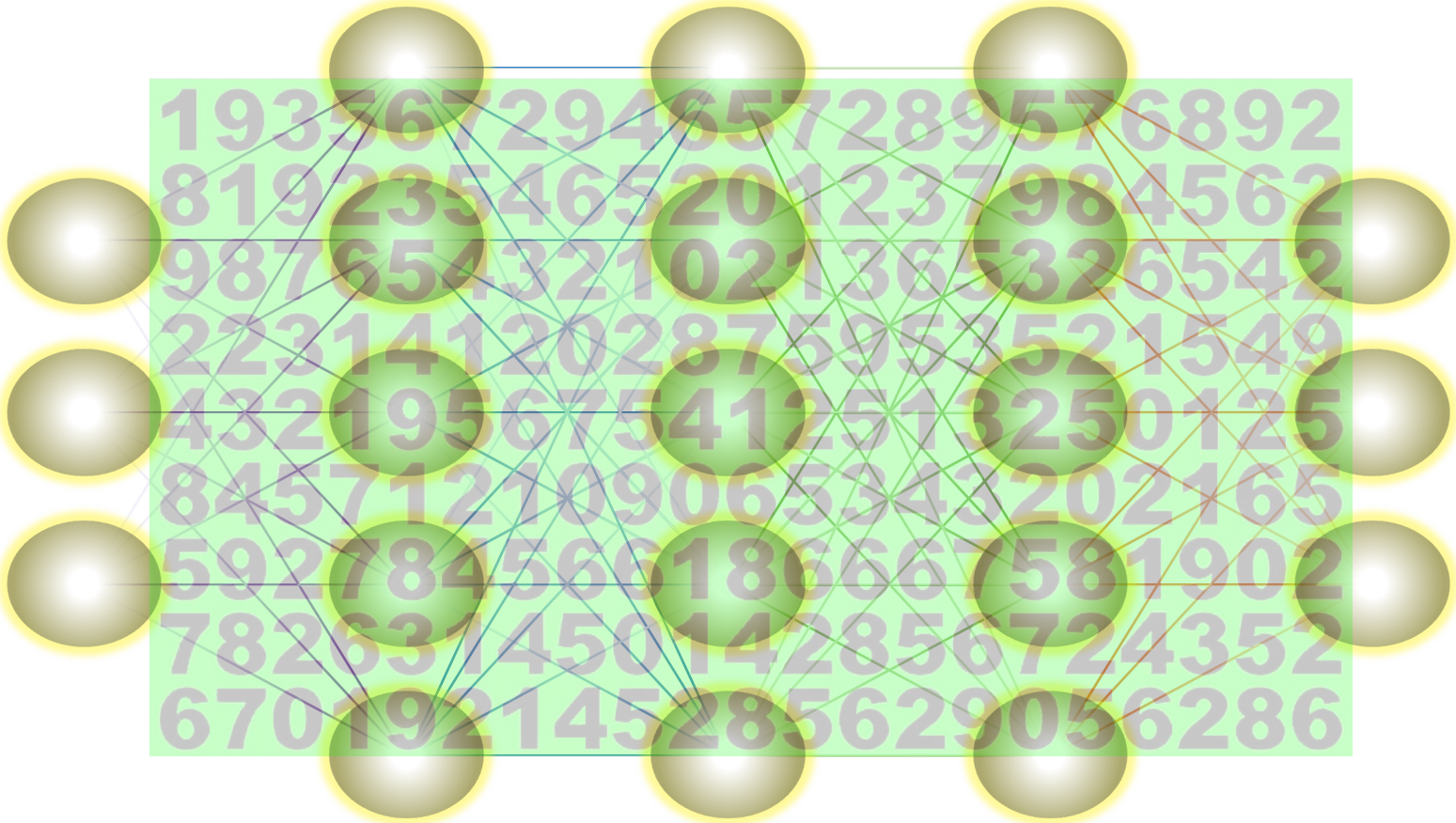
"Marvin Minsky"



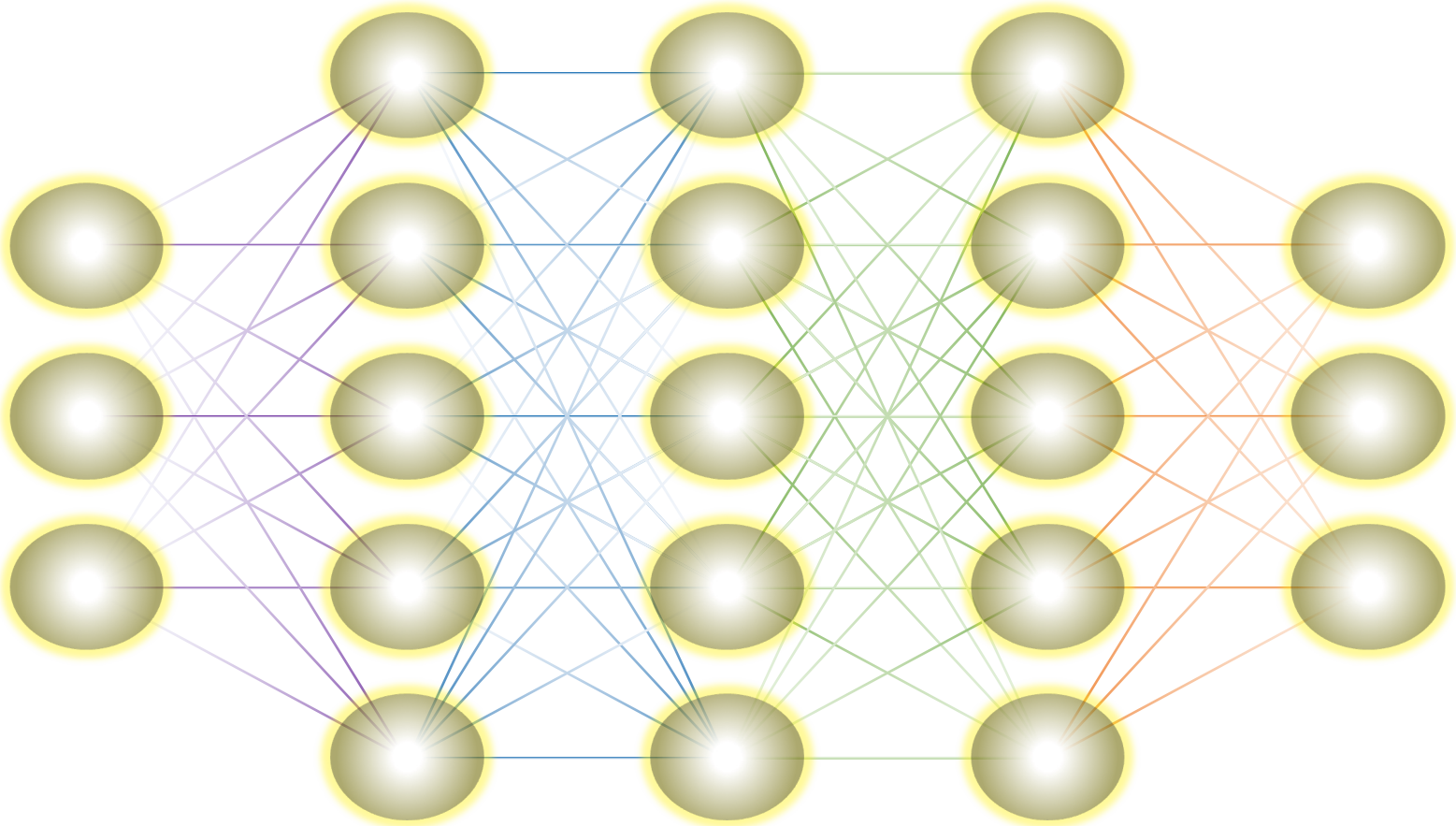
Artificial Intelligence



x

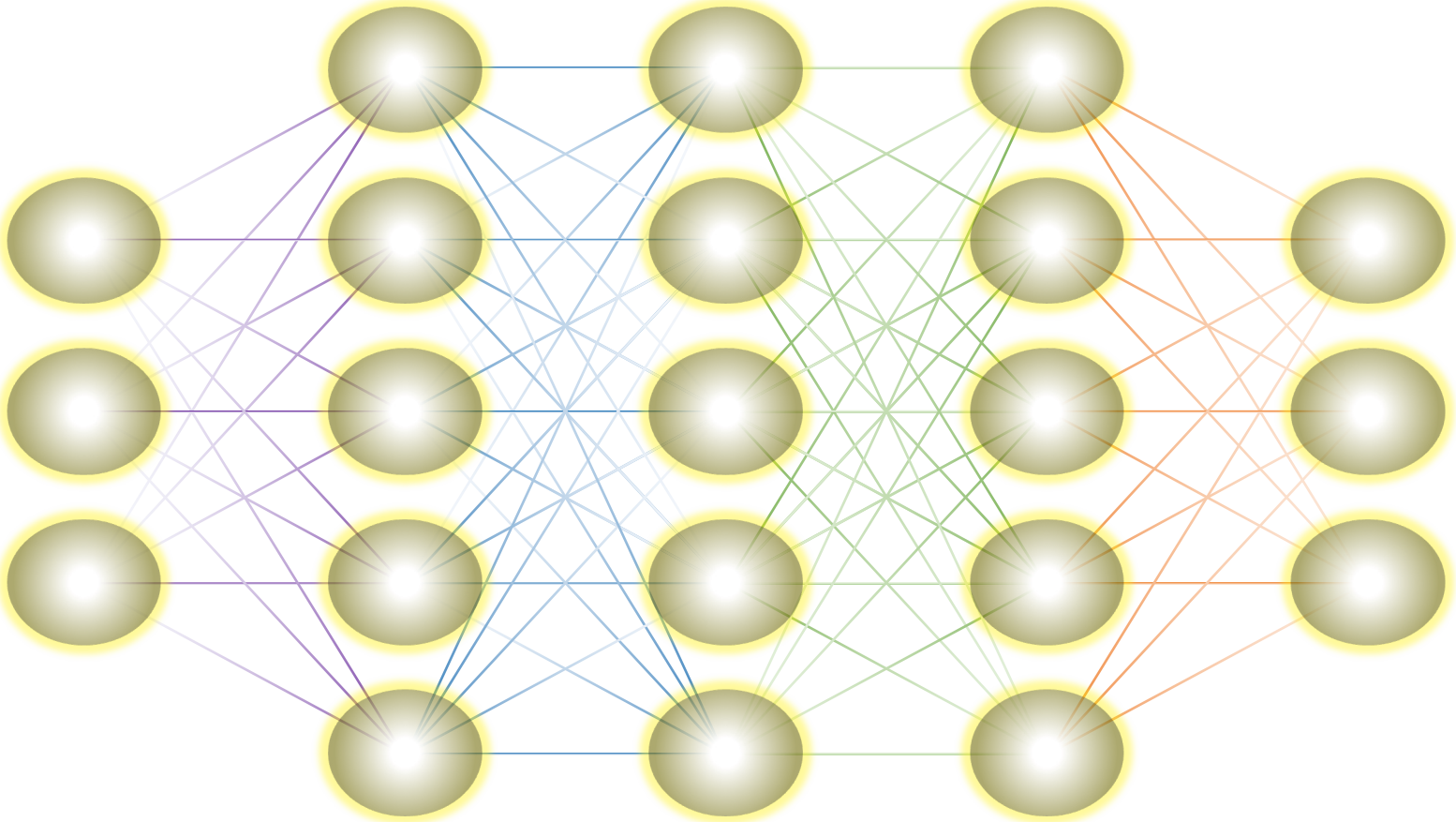


y



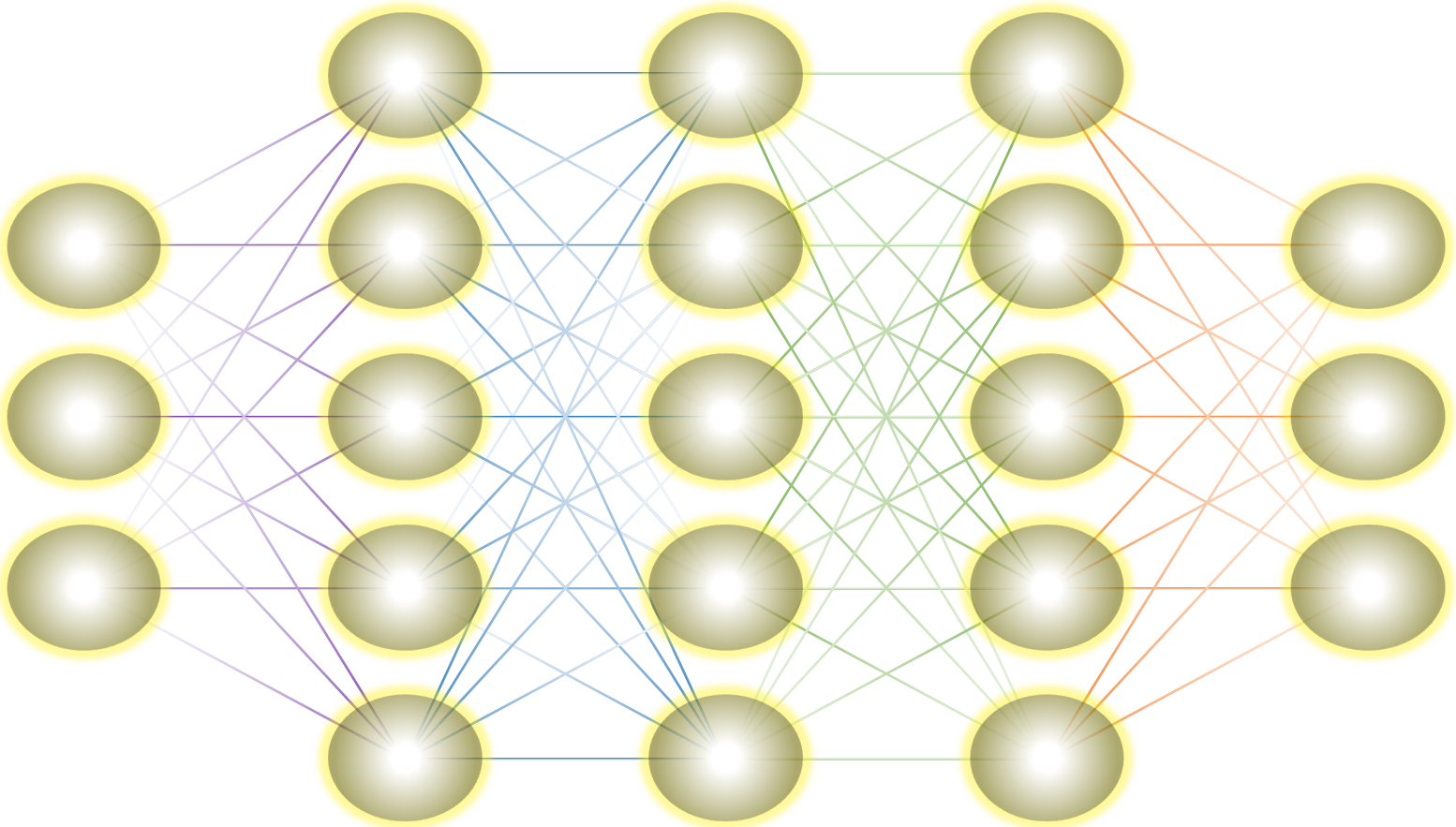
“cat”

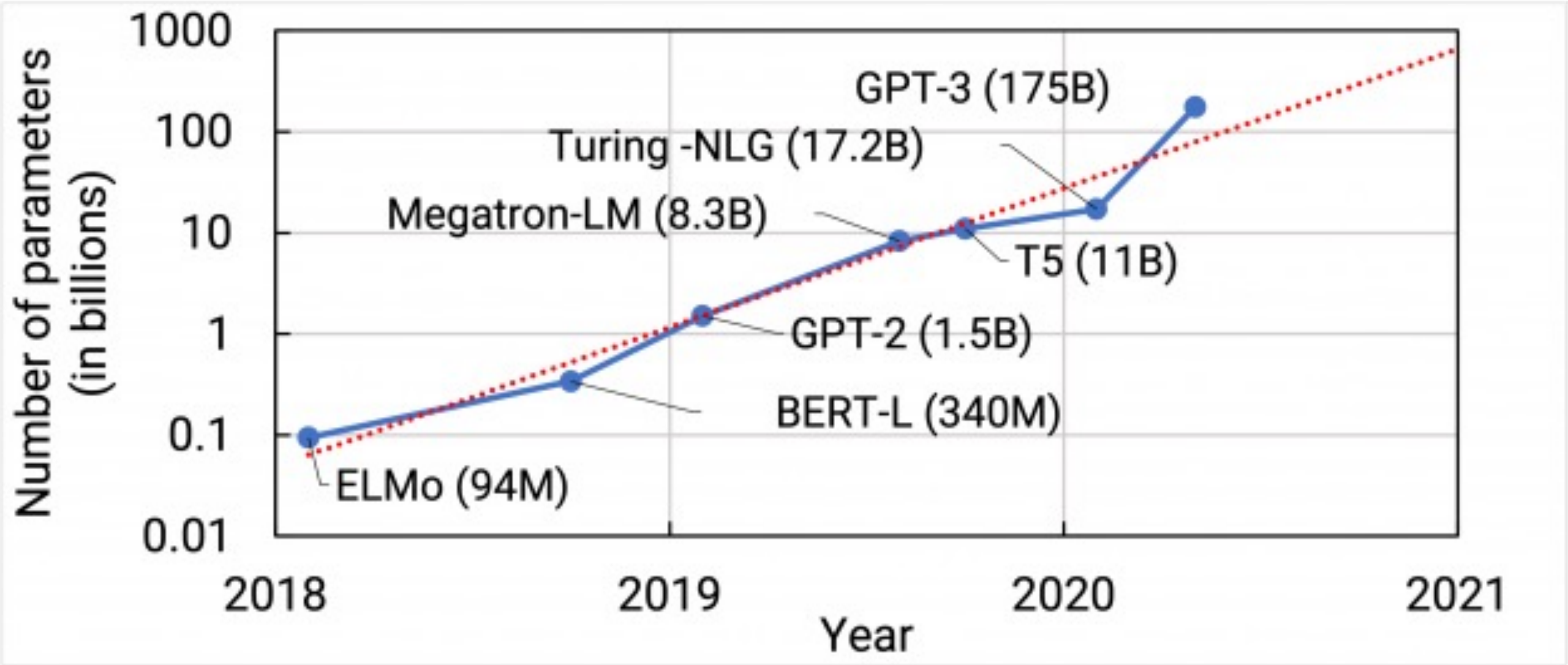
Artificial Neural Networks

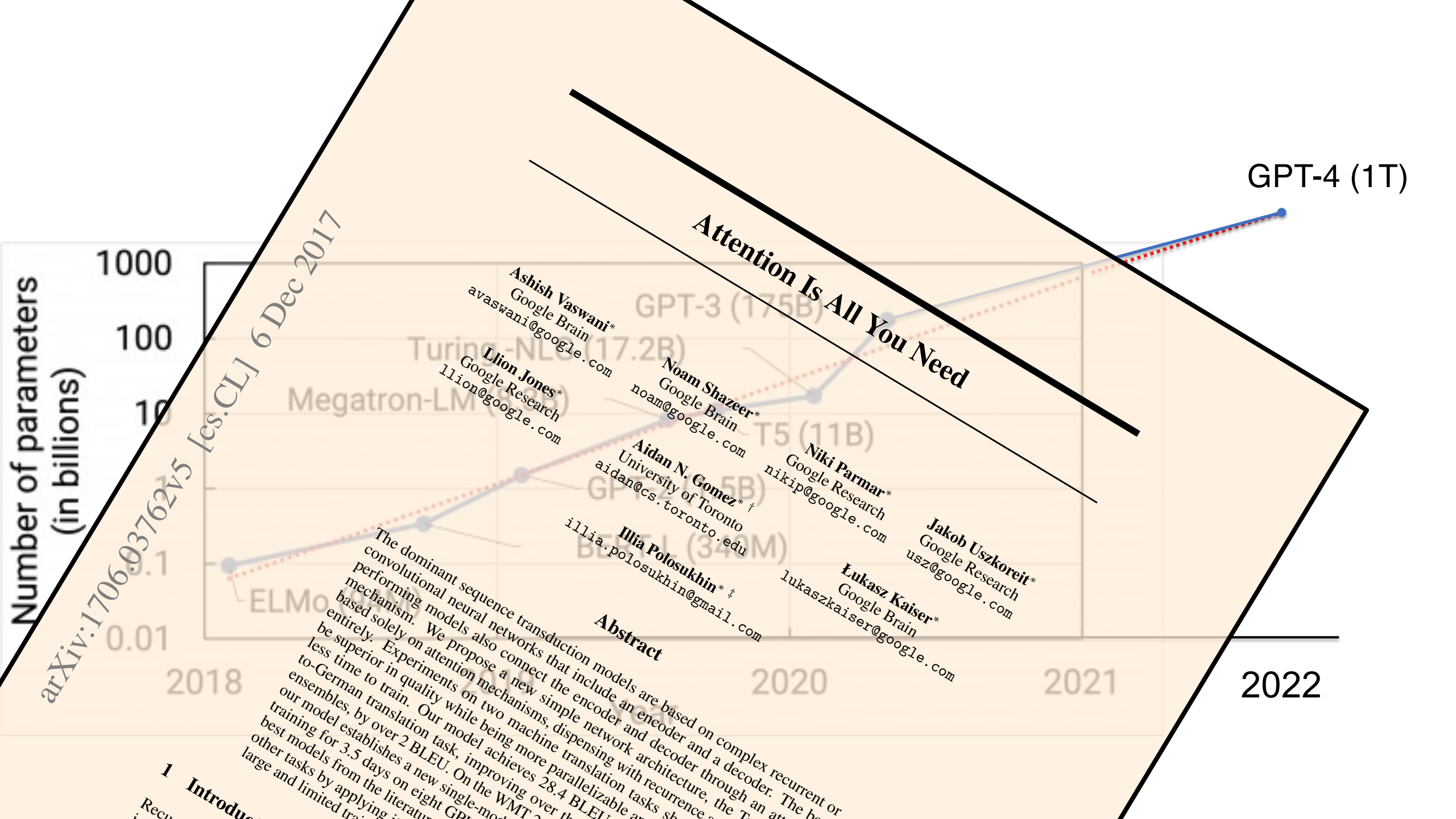


“cat”

“cat”




1000
100
10
1
0.1
0.01
Number of parameters (in billions)
GPT-3 (175B)
Turing -NLG (17.2B)
Megatron-LM (8.3B)
T5 (11B)
GPT-2 (1.5B)
BERT-L (340M)
ELMo (94M)
2018
2019
2020
2021
Year



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

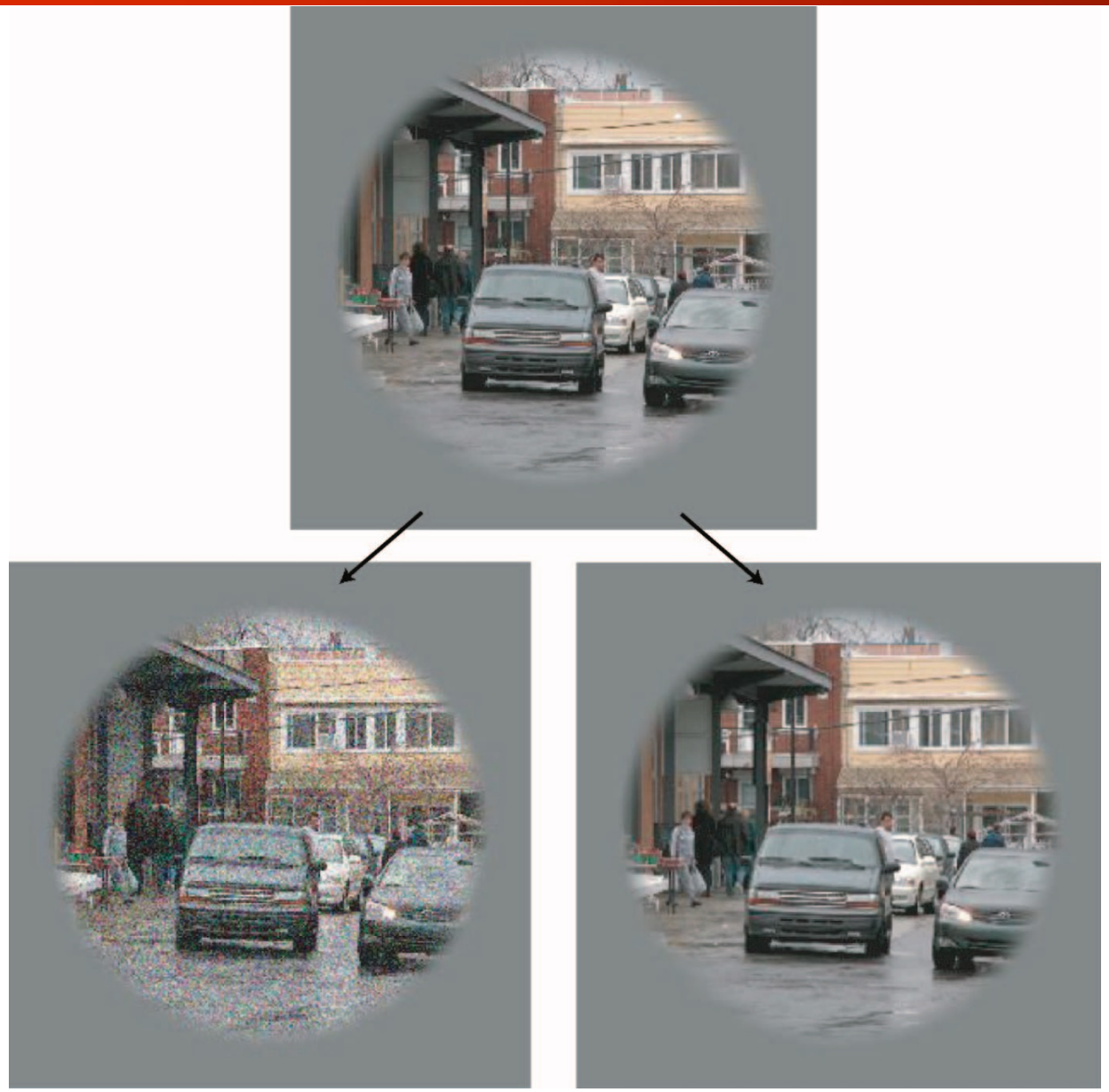
1 Introduction

Recu

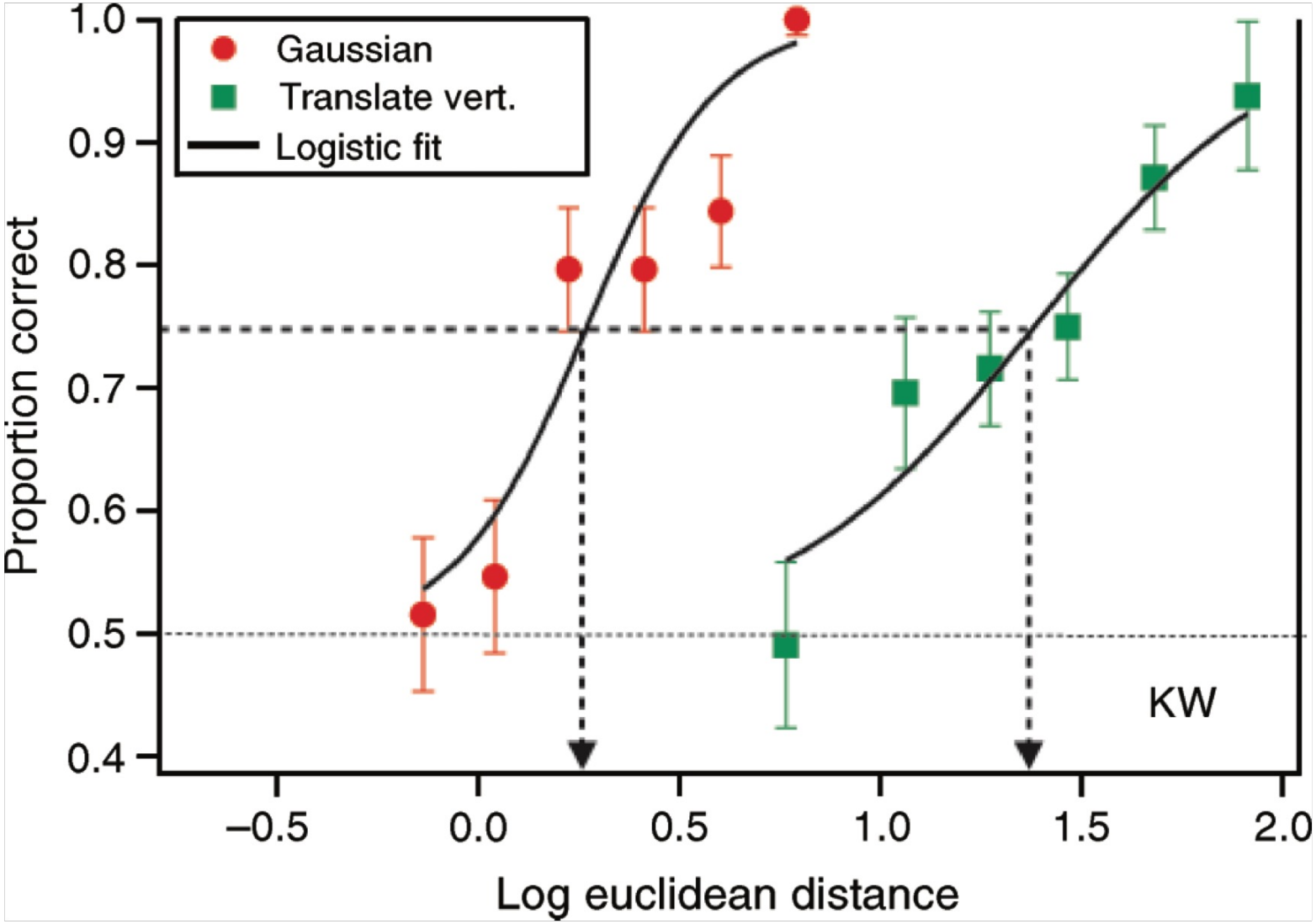
Today

- About the Course
- What is AI?
- **The Rise of Deep Learning**
- Why Math in AI Matters

Question?



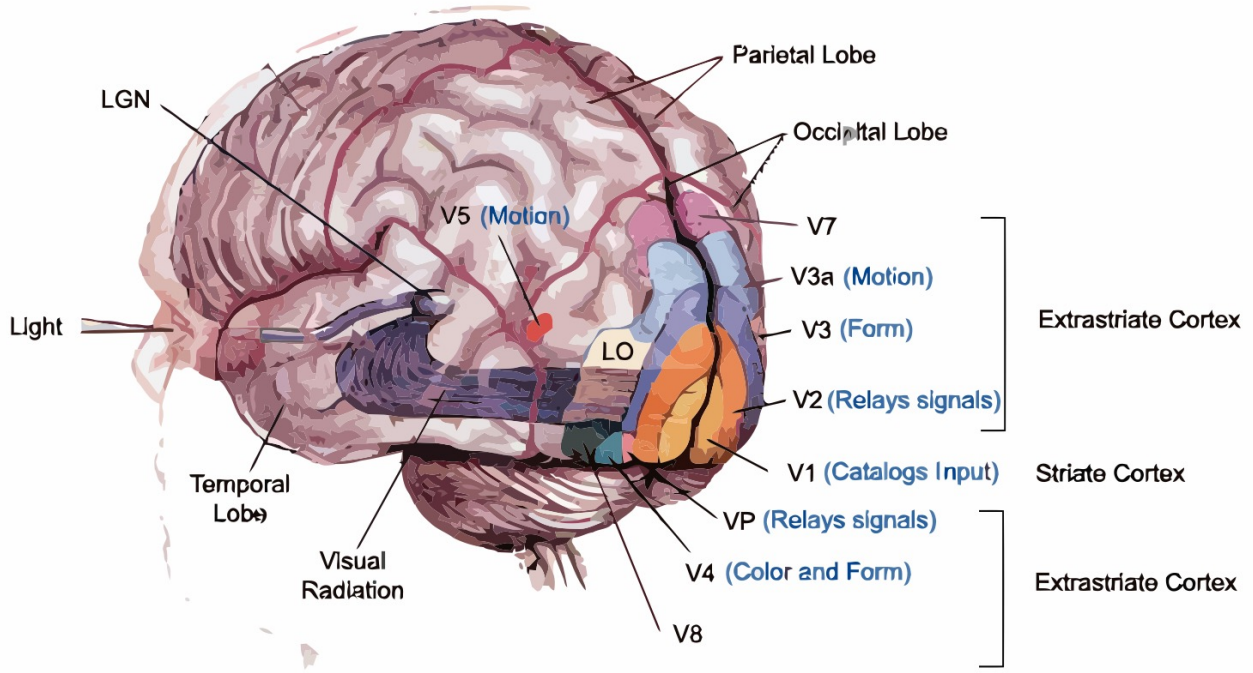
Answer

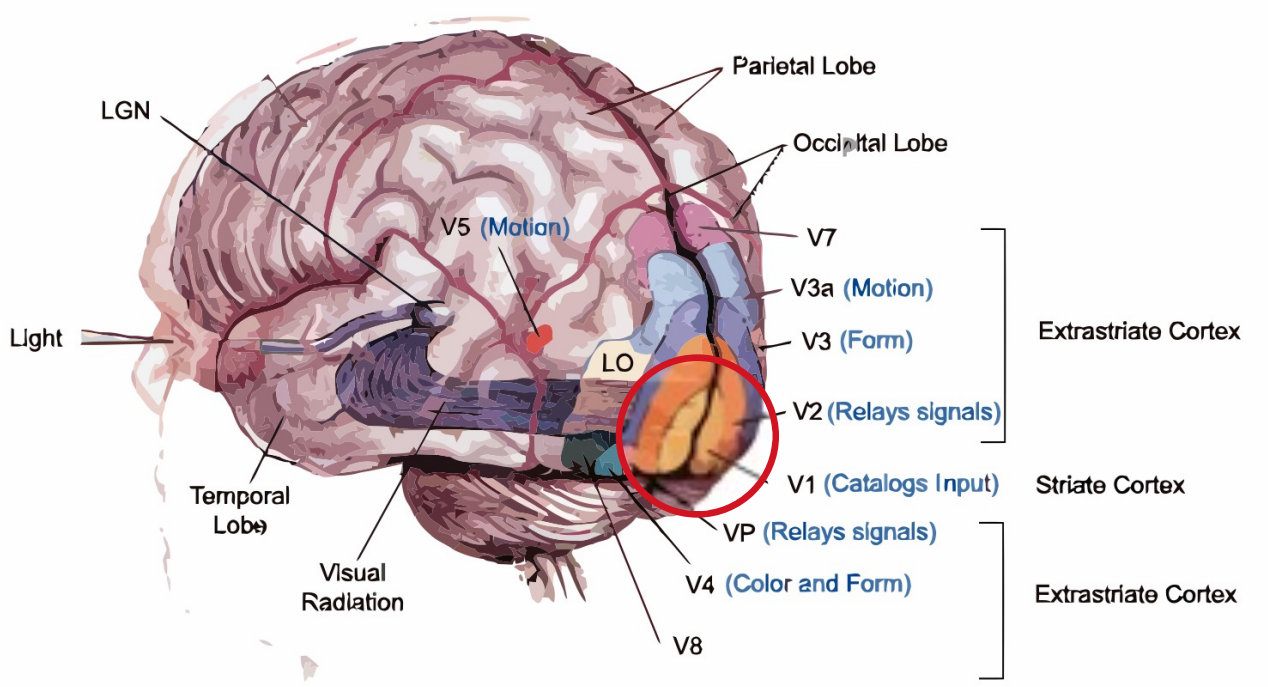


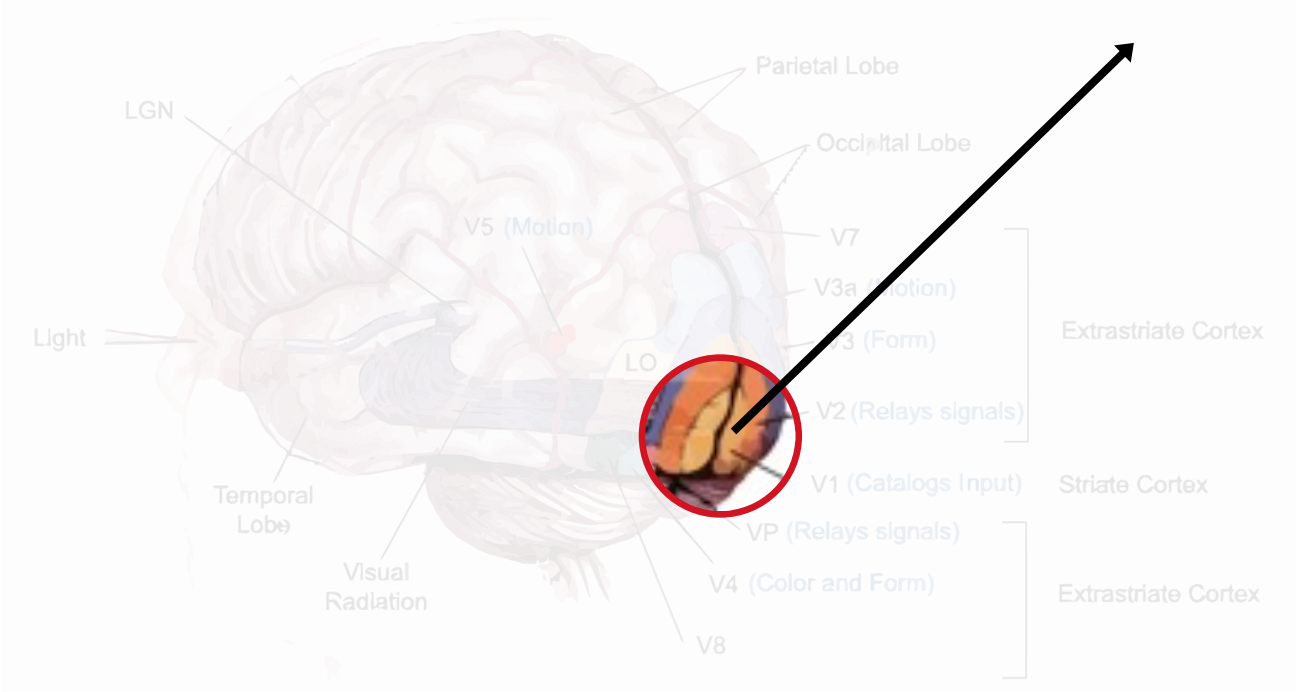
Kingdom, Field, Olmos, 2007

Why?

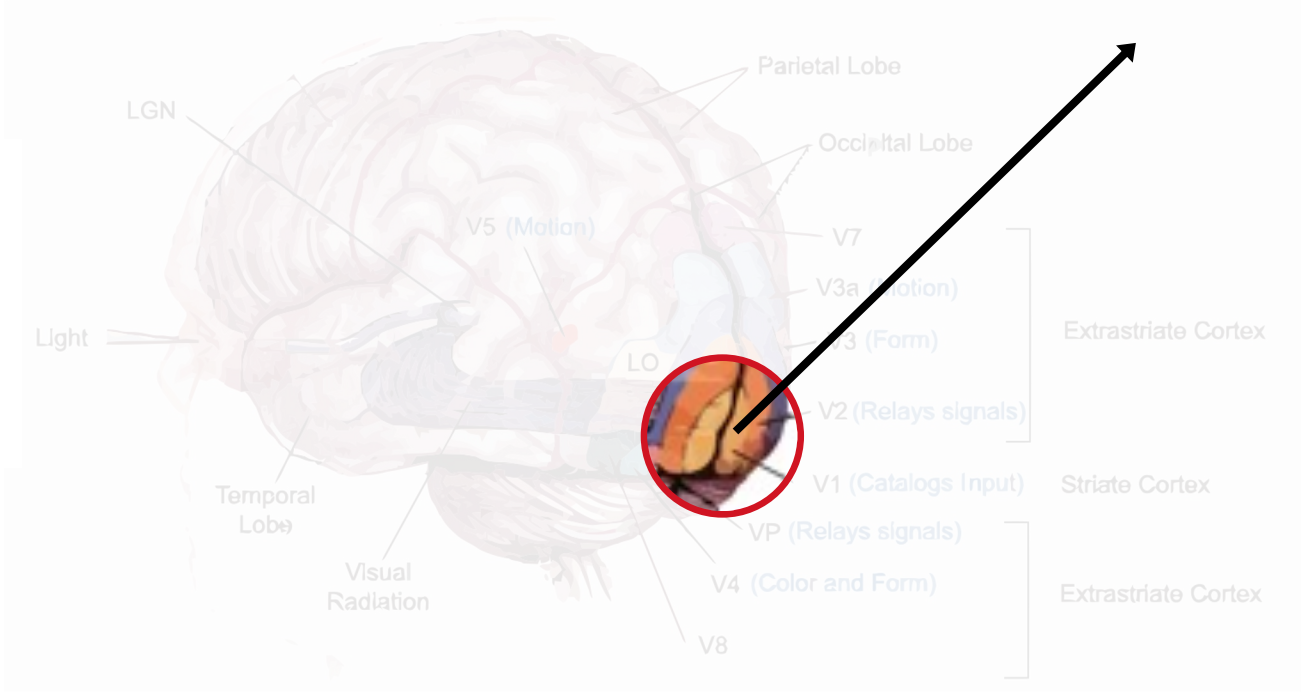
207	245	77	21	247	211	240	1
219	41	58	179	161	154	184	98
215	145	187	71	251	249	65	100
192	2	189	247	166	63	232	213
105	94	66	190	156	61	89	145
159	154	87	184	101	105	72	71
192	111	6	94	60	70	65	226
175	120	210	226	80	183	168	184
134	56	36	240	159	178	76	135
239	244	199	9	132	104	188	185
245	210	78	199	0	92	9	246
5	121	187	122	107	47	12	119
230	171	135	36	82	54	65	37
61	140	79	19	161	96	127	187
56	223	46	6	180	186	142	244
28	20	61	2	178	187	98	220


Parietal Lobe
LGN
Occipital Lobe
V5 (Motion)
V7
V3a (Motion)
Light
V3 (Form)
LO
V2 (Relays signals)
Extrastriate Cortex
V1 (Catalogs Input)
Striate Cortex
Temporal Lobe
VP (Relays signals)
Visual Radiation
V4 (Color and Form)
Extrastriate Cortex
V8

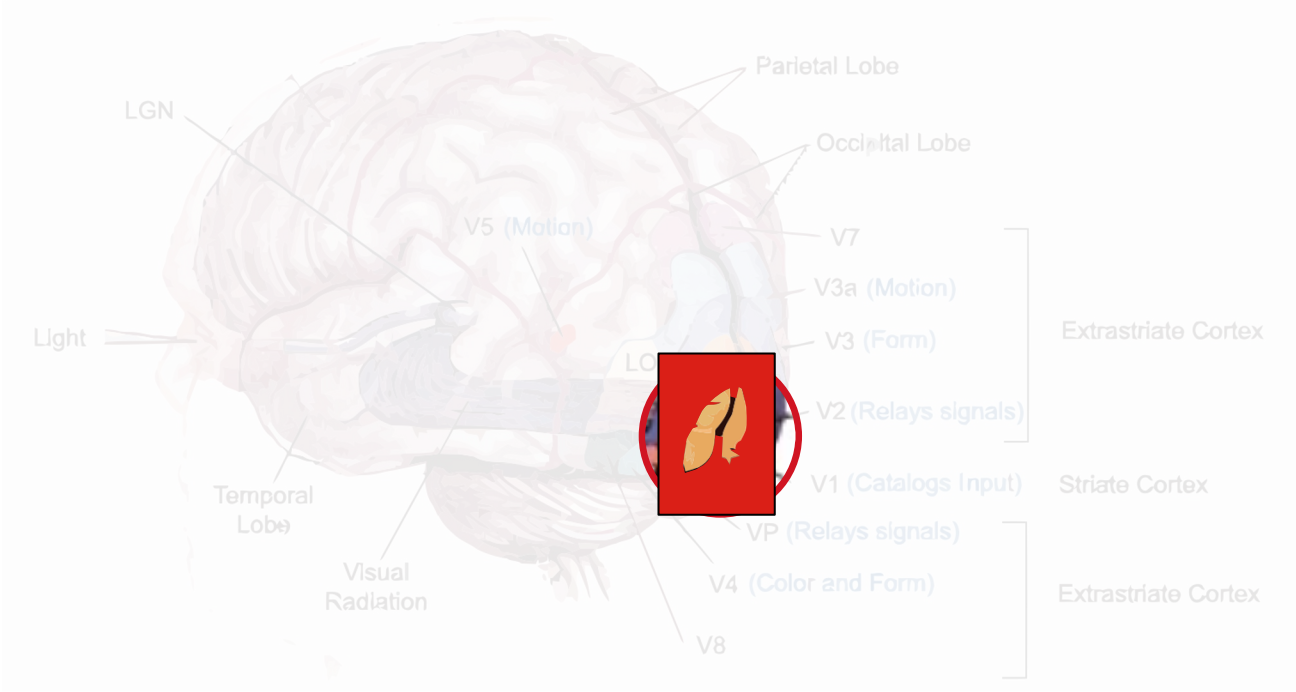

Parietal Lobe
LGN
Occipital Lobe
V5 (Motion)
V7
V3a (Motion)
Light
V3 (Form)
LO
V2 (Relays signals)
Extrastriate Cortex
V1 (Catalogs Input)
Striate Cortex
Temporal Lobe
VP (Relays signals)
Visual Radiation
V4 (Color and Form)
Extrastriate Cortex
V8



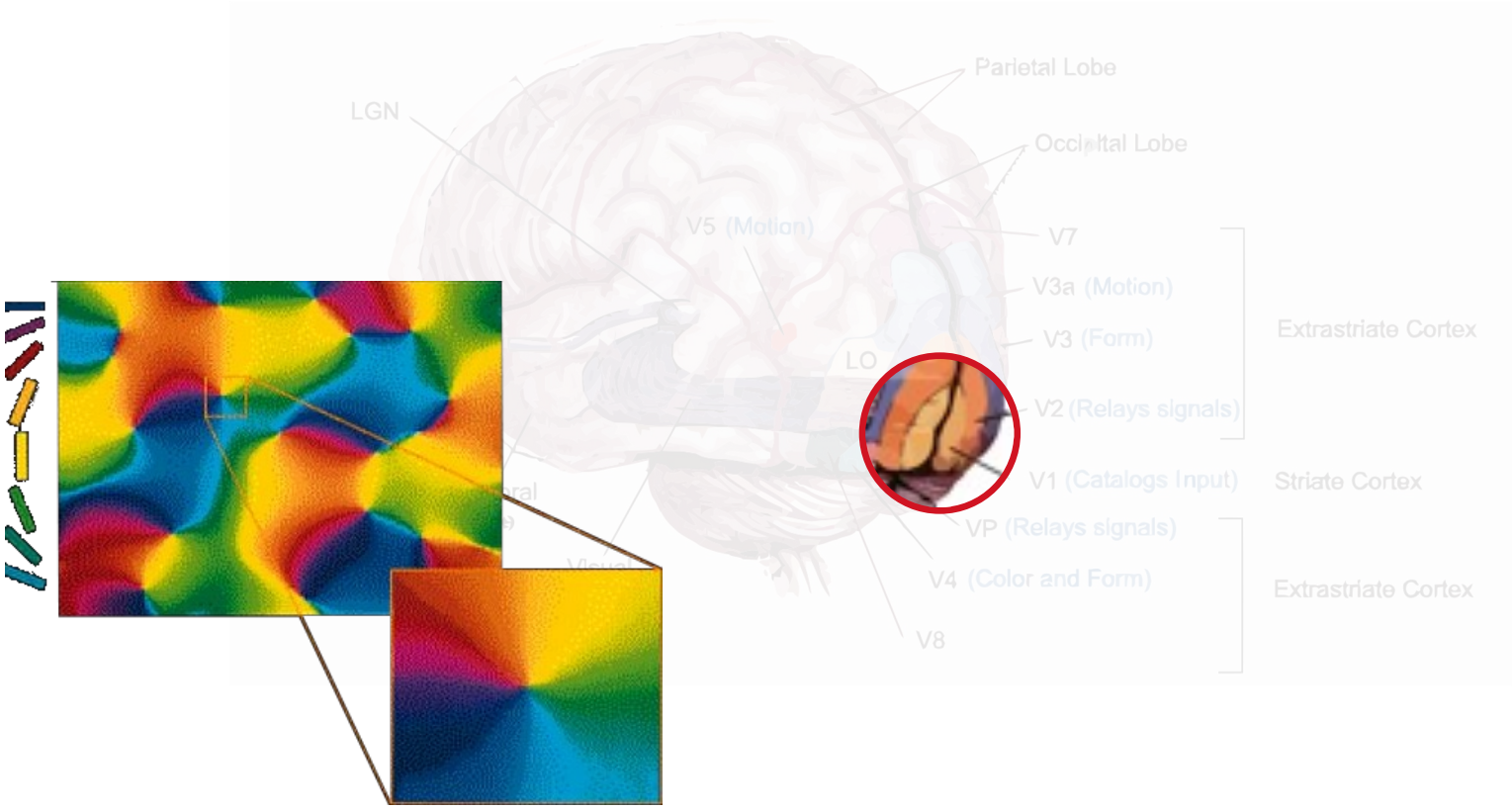
D.H. Hubel & T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160(1):106, 1962.



D.H. Hubel & T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160(1):106, 1962.



D.H. Hubel & T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160(1):106, 1962.



D.H. Hubel & T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160(1):106, 1962.

Emergence of simple-cell receptive field properties by learning a sparse code for natural images

Bruno A. Olshausen* & David J. Field

Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853, USA

THE receptive fields of simple cells in mammalian primary visual cortex can be characterized as being spatially localized, oriented[1–4] and bandpass (selective to structure at different spatial scales), comparable to the basis functions of wavelet transforms[5,6]. One approach to understanding such response properties of visual neurons has been to consider their relationship to the statistical structure of natural images in terms of efficient coding[7–12]. Along these lines, a number of studies have attempted to train unsupervised learning algorithms on natural images in the hope of developing receptive fields with similar properties[13–18], but none has succeeded in producing a full set that spans the image space and contains all three of the above properties. Here we investigate the proposal[8,12] that a coding strategy that maximizes sparseness is sufficient to account for

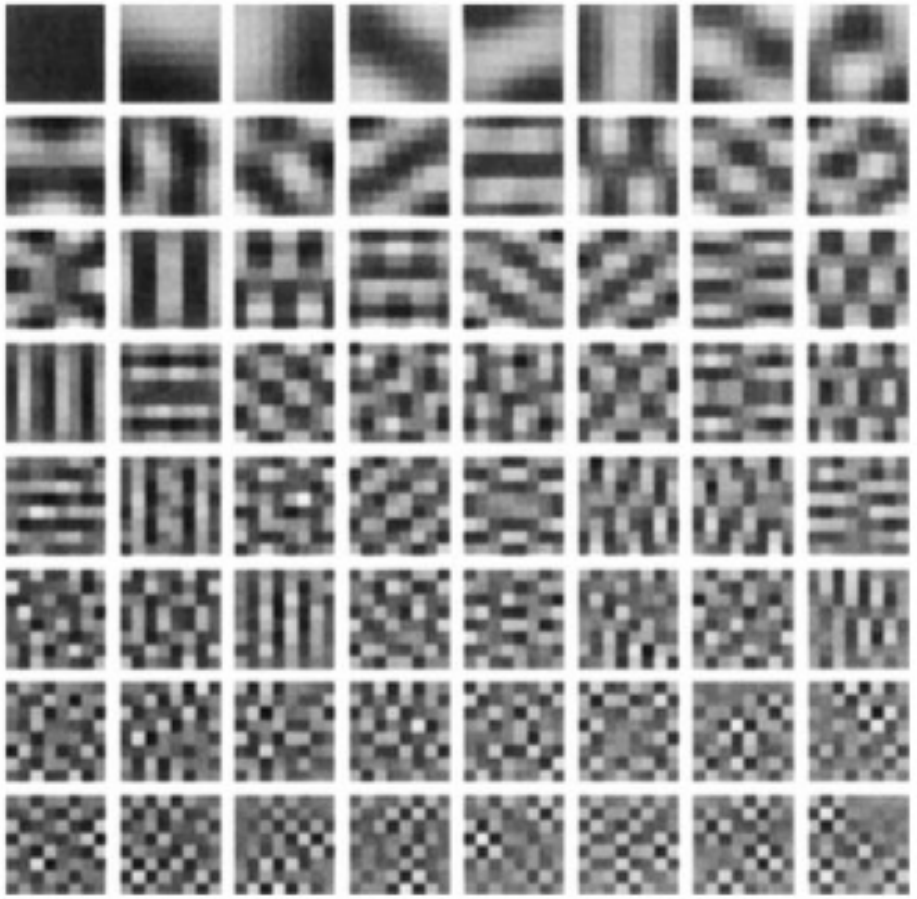
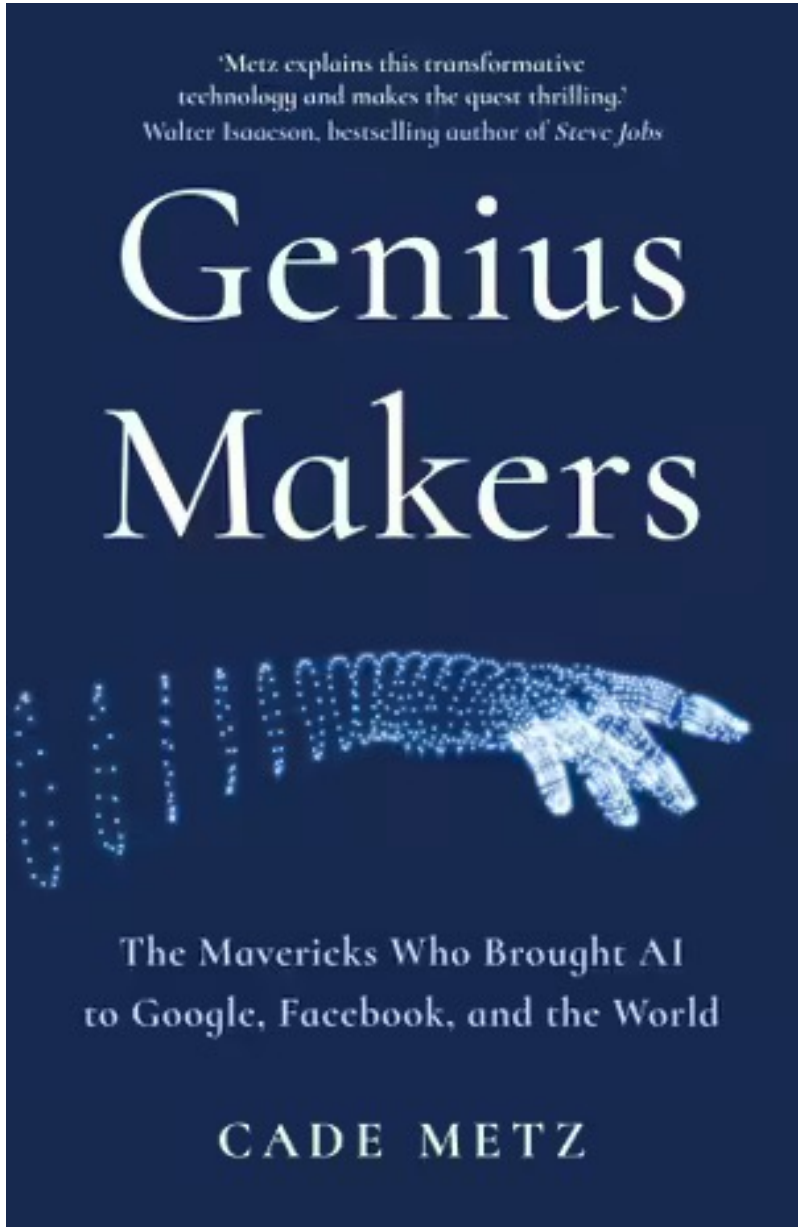


FIG. 1 Principal components calculated on 8 × 8 image patches extracted from natural scenes by using Sanger's rule[14]. The full set of 64 components is shown, ordered by their variance (by columns, then by rows). The oriented structure of the first few principal components does not arise as a result of the oriented structures in natural images, but rather because these functions are composed of a small number of low-frequency components

Olshausen & Field 1996

The Debate Rages On!!!!

Yoshua Bengio



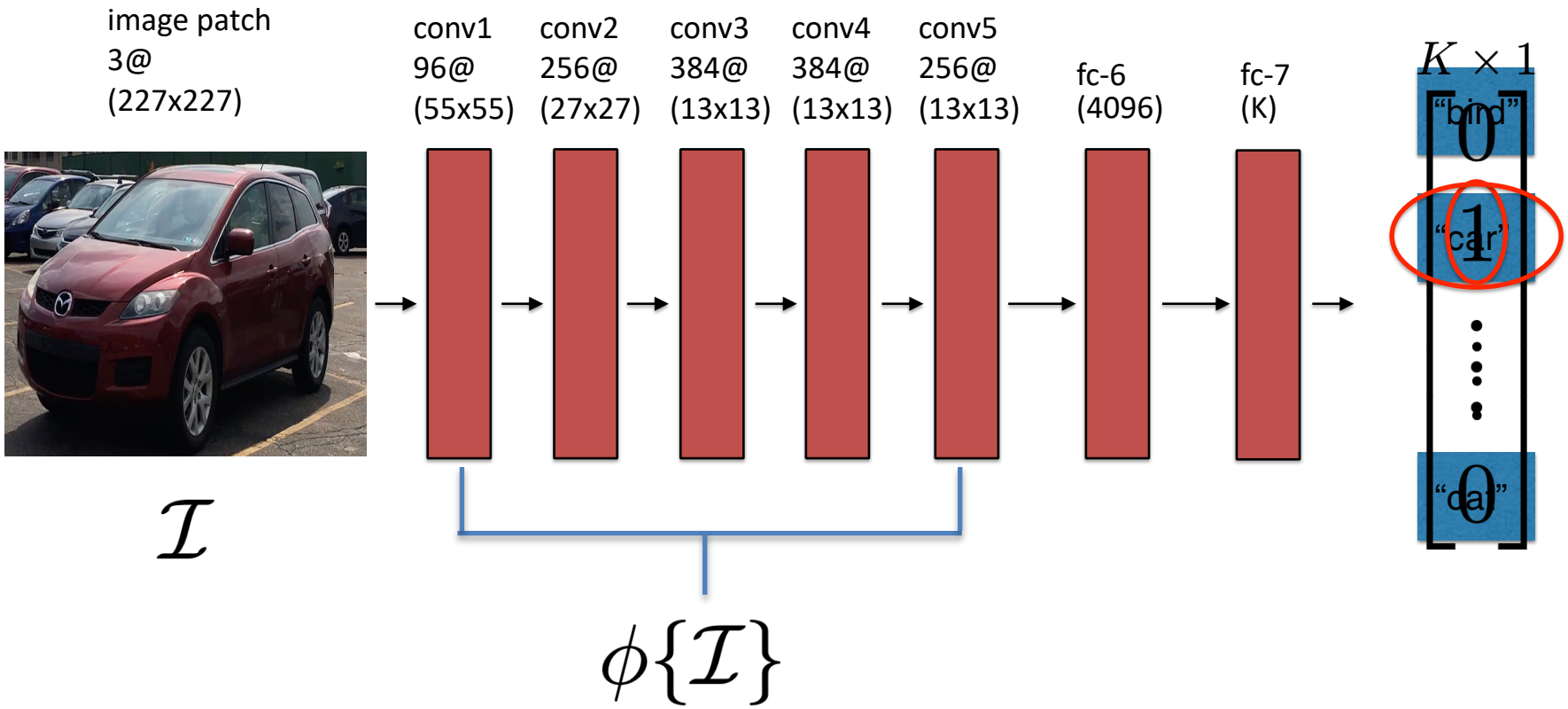
Gary Marcus

Challenge

207	245	77	21	247	211	240	1
219	41	58	179	161	154	184	98
215	145	187	71	251	249	65	100
192	2	189	247	166	63	232	213
105	94	66	190	156	61	89	145
159	154	87	184	101	105	72	71
192	111	6	94	60	70	65	226
175	120	210	226	80	183	168	184
134	56	36	240	159	178	76	135
239	244	199	9	132	104	188	185
245	210	78	199	0	92	9	246
5	121	187	122	107	47	12	119
230	171	135	36	82	54	65	37
61	140	79	19	161	96	127	187
56	223	46	6	180	186	142	244
28	20	61	2	178	187	98	220

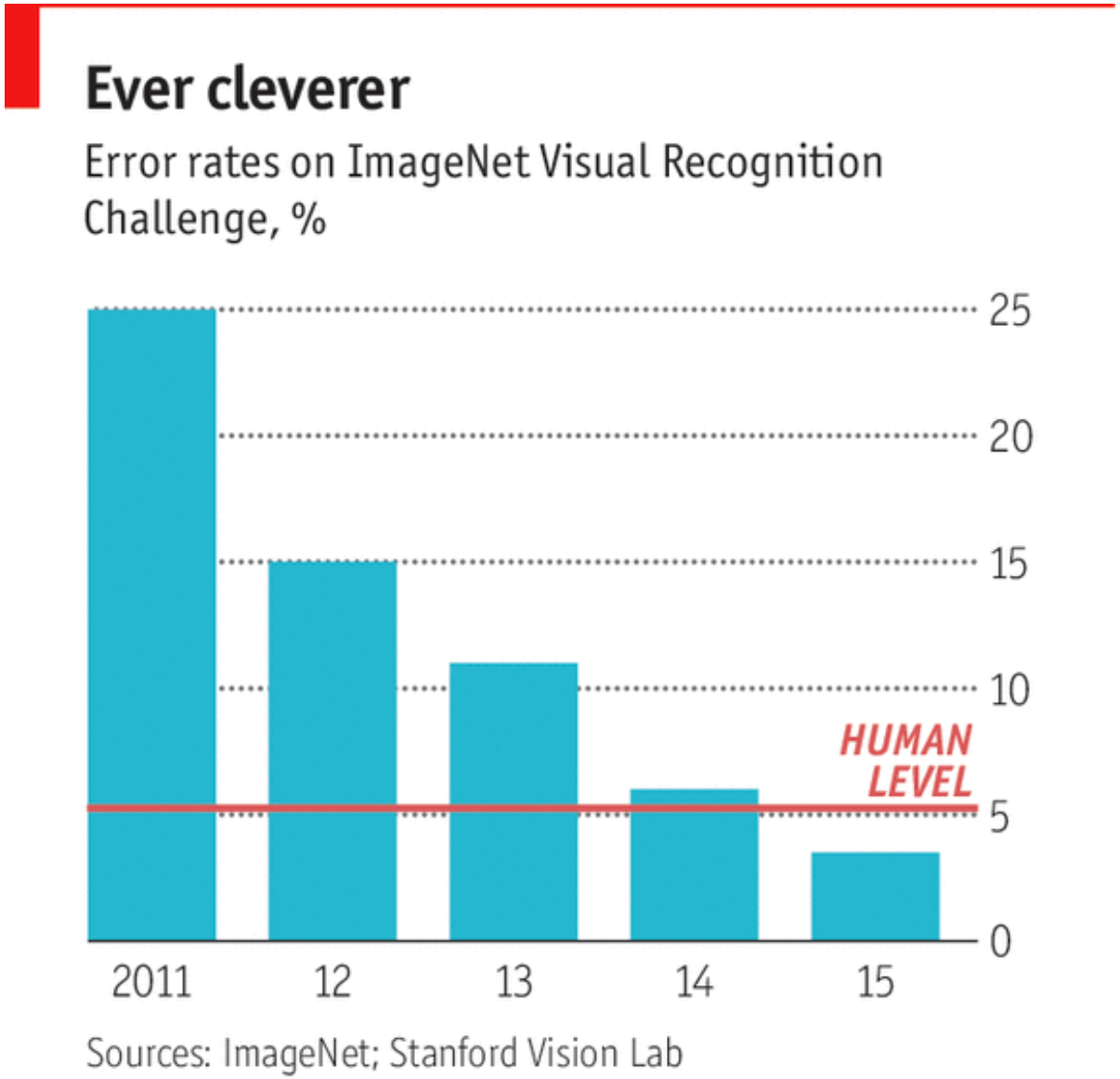
$$\phi\{\mathcal{I}\}$$

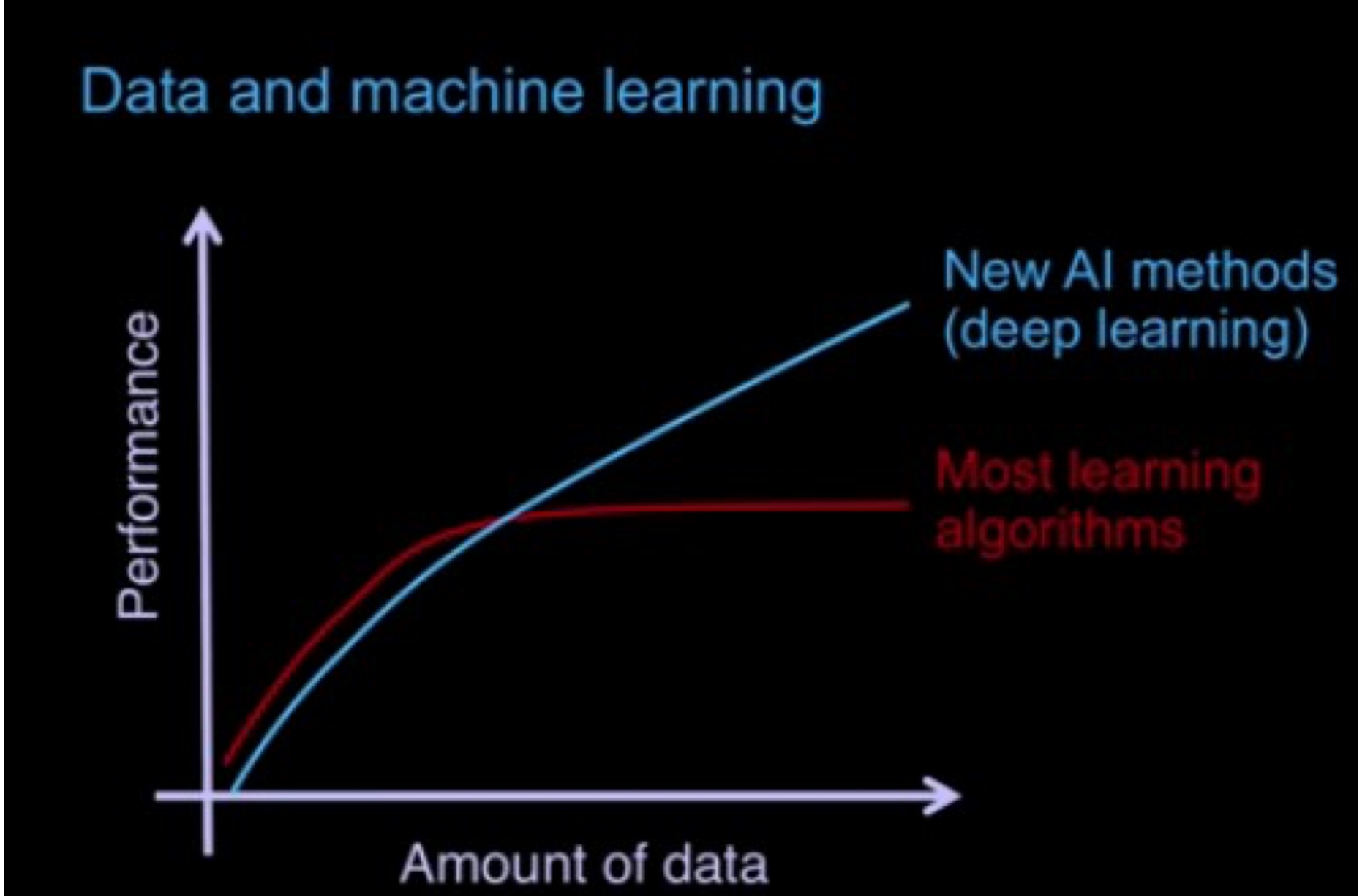
Deep Learning - A Breakthrough!!!



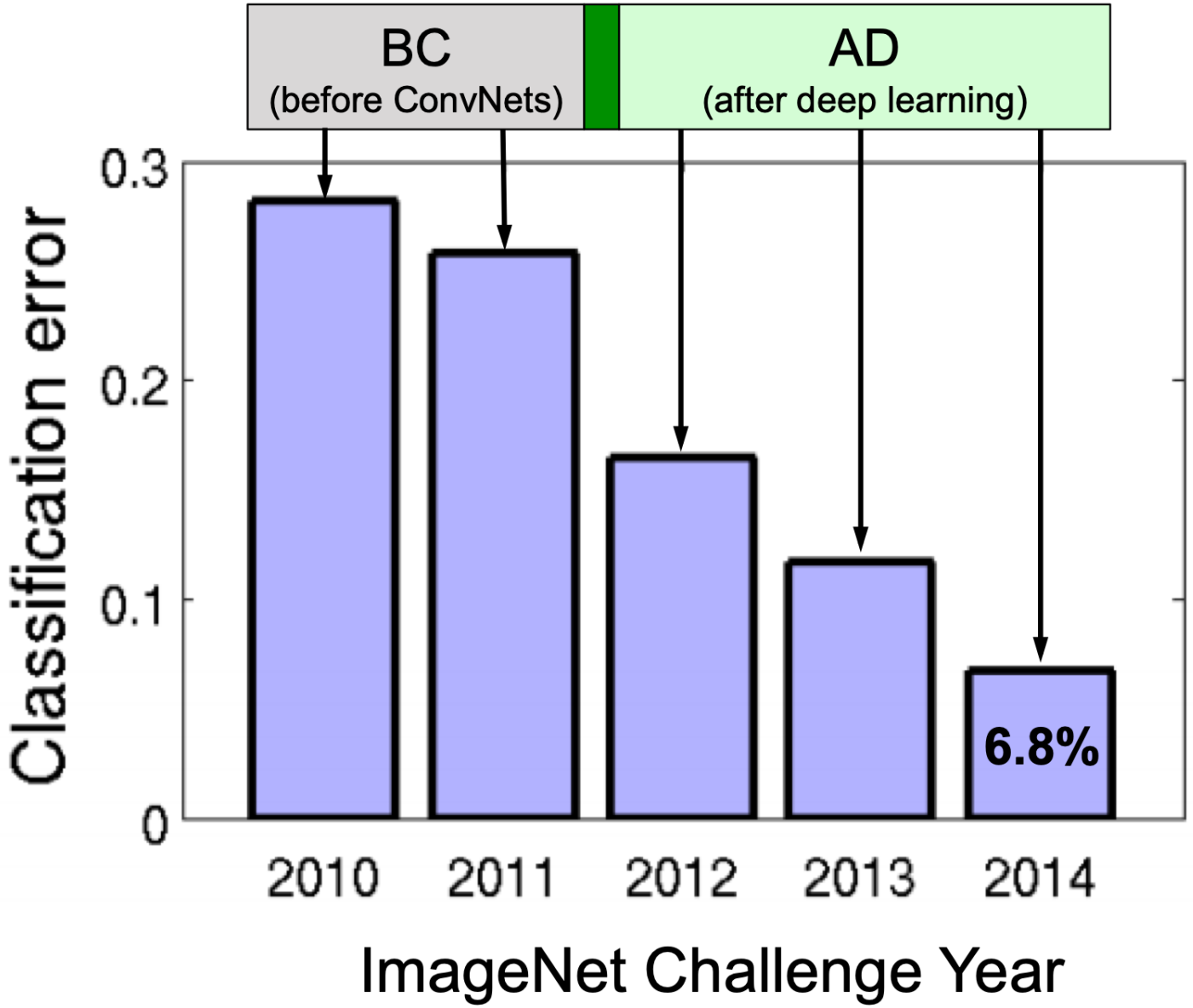
A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.

Deep Learning




Data and machine learning
Performance
New AI methods
(deep learning)
Most learning
algorithms
Amount of data

Impact on Object Recognition



ImageNet

- Over 15M labeled high resolution images.
- Roughly 22K categories.
- Collected from web and labeled by Amazon Mechanical Turk.

Machines are starting to act like us!


GO
STOP
0 MPH
3 FEET
2 MPH
10 FEET

Audio

TIMIT Phone classification	Accuracy
Prior art (Clarkson et al.,1999)	79.6%
Feature learning	**80.3%**

TIMIT Speaker identification	Accuracy
Prior art (Reynolds, 1995)	99.7%
Feature learning	**100.0%**

Images

CIFAR Object classification	Accuracy
Prior art (Ciresan et al., 2011)	80.5%
Feature learning	**82.0%**

NORB Object classification	Accuracy
Prior art (Scherer et al., 2010)	94.4%
Feature learning	**95.0%**

Video

Hollywood2 Classification	Accuracy
Prior art (Laptev et al., 2004)	48%
Feature learning	**53%**

YouTube	Accuracy
Prior art (Liu et al., 2009)	71.2%
Feature learning	**75.8%**

KTH	Accuracy
Prior art (Wang et al., 2010)	92.1%
Feature learning	**93.9%**

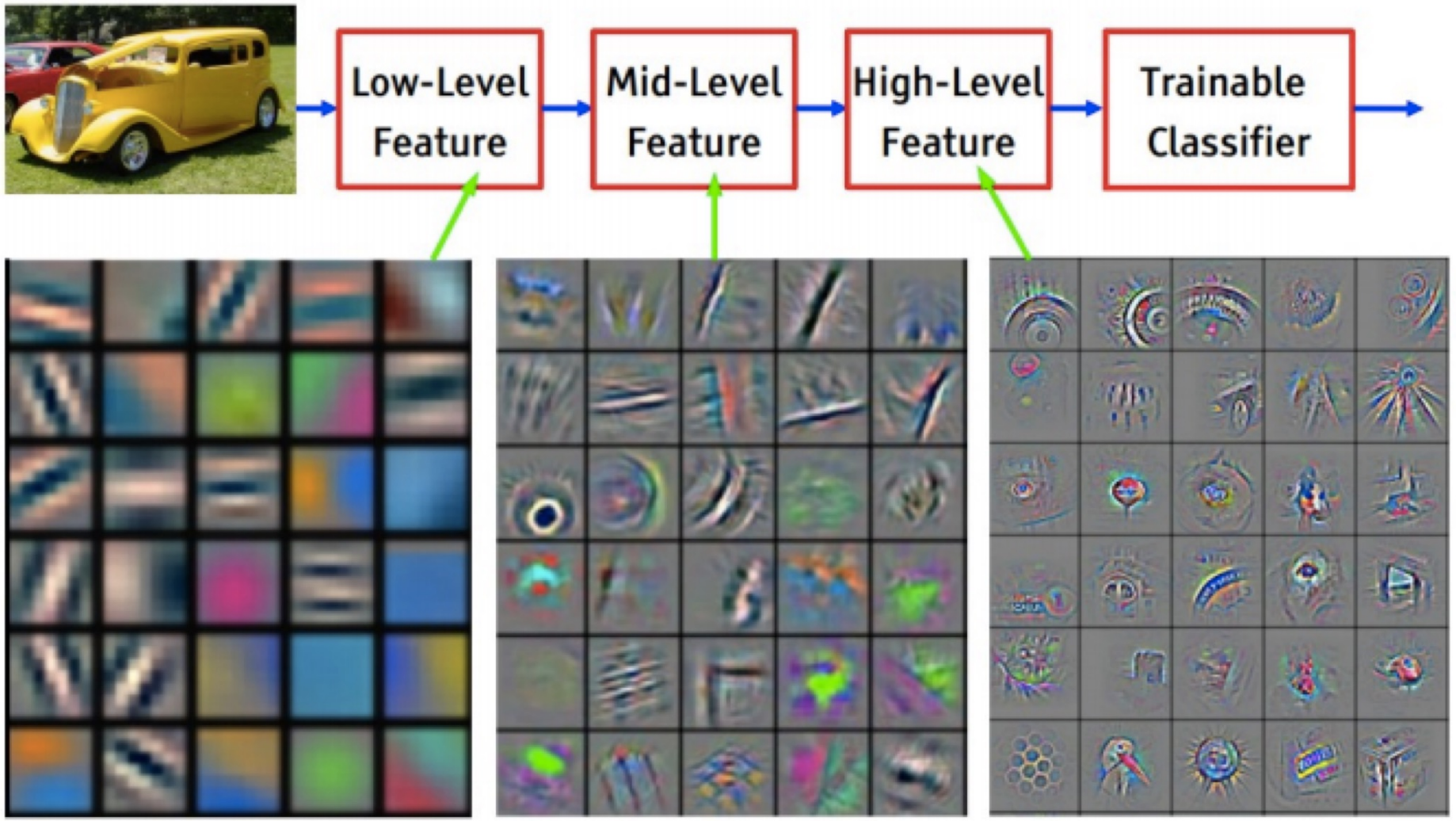
UCF	Accuracy
Prior art (Wang et al., 2010)	85.6%
Feature learning	**86.5%**

Text/NLP

Paraphrase detection	Accuracy
Prior art (Das & Smith, 2009)	76.1%
Feature learning	**76.4%**

Sentiment (MR/MPQA data)	Accuracy
Prior art (Nakagawa et al., 2010)	77.3%
Feature learning	**77.7%**

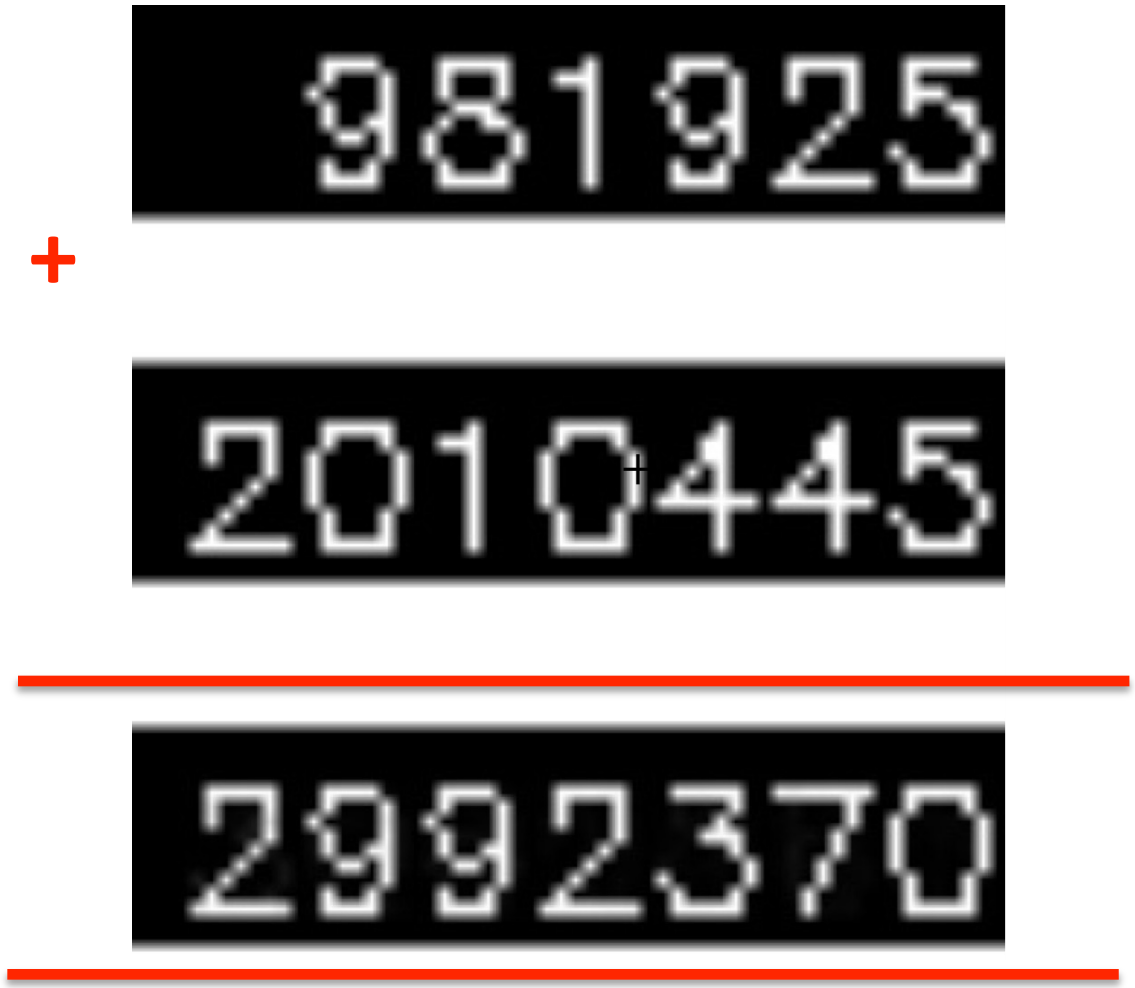
Visualizing CNNs



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

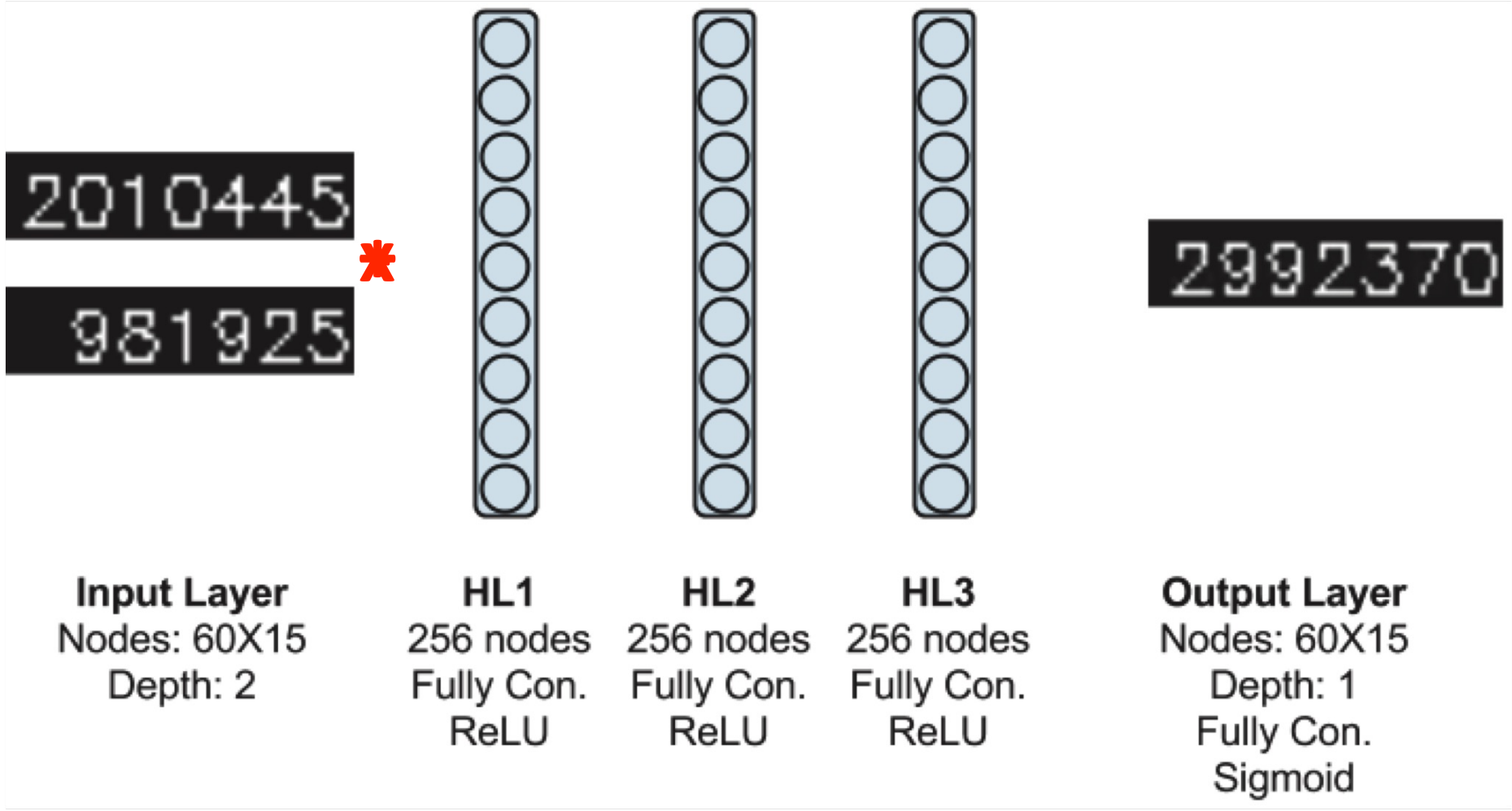
Mimicking ≠ Understanding

Deep Learning & Visual Arithmetic



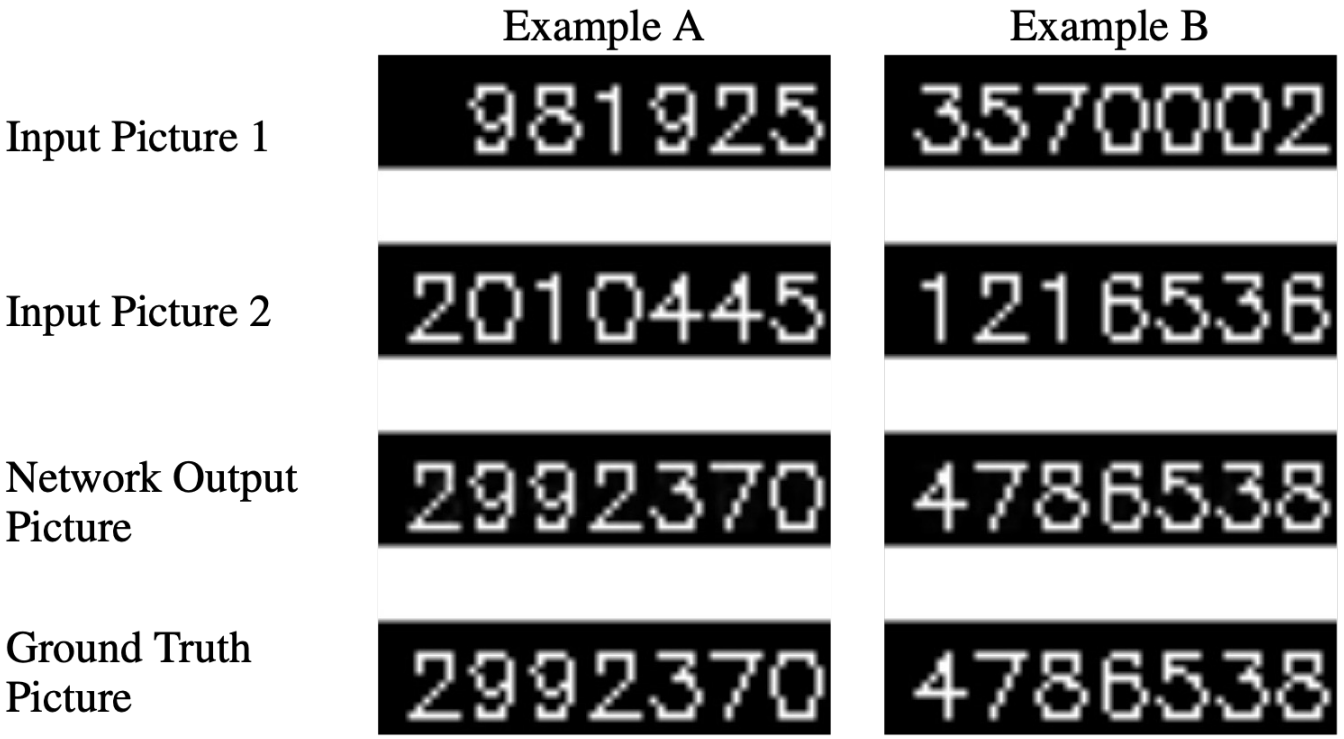
Y. Hoshen and S. Peleg, Visual Learning of Arithmetic Operations, AAAI 2016.

Deep Learning & Visual Arithmetic



Y. Hoshen and S. Peleg, Visual Learning of Arithmetic Operations, AAAI 2016.

Deep Learning & Visual Arithmetic



Y. Hoshen and S. Peleg, Visual Learning of Arithmetic Operations, AAAI 2016.

Deep Learning & Visual Arithmetic

Operation	Pictures		1-hot Vectors	
	No. Layers	% Error	No. Layers	% Error
Add	3	1.9%	1	1.7%
Subtract	3	3.2%	1	2.1%
Multiply	5	71.5%	3	37.6%
Roman Addition	5	74.3 %	3	0.7 %

Y. Hoshen and S. Peleg, Visual Learning of Arithmetic Operations, AAAI 2016.

Multiplying with GPT-4

Multiply 999 by 867

GPT-4

The product of multiplying 999 by 867 is 864,133

The correct answer is 866,133

GPT-4 achieves only 0% accuracy on 5 digit by 5 digit multiplication

Faith and Fate: Limits of Transformers on Compositionality

Nouha Dziri[1*], Ximing Lu[1,2*], Melanie Sclar[2*], Xiang Lorraine Li[1†], Liwei Jiang[1,2†], Bill Yuchen Lin[1], Peter West[1,2], Chandra Bhagavatula[1], Ronan Le Bras[1], Jena D. Hwang[1], Soumya Sanyal[3], Sean Welleck[1,2], Xiang Ren[1,3], Allyson Ettinger[1,4], Zaid Harchaoui[1,2], Yejin Choi[1,2]

[1]Allen Institute for Artificial Intelligence [2]University of Washington
[3]University of Southern California [4]University of Chicago
nouhad@allenai.org, ximinglu@allenai.org, msclar@cs.washington.edu

Abstract

Transformer large language models (LLMs) have sparked admiration for their exceptional performance on tasks that demand intricate multi-step reasoning. Yet, these models simultaneously show failures on surprisingly trivial problems. This begs the question: Are these errors incidental, or do they signal more substantial limitations? In an attempt to demystify Transformers, we investigate the limits of these models across three representative *compositional* tasks—multi-digit multiplication, logic grid puzzles, and a classic dynamic programming problem. These tasks require breaking problems down into sub-steps and synthesizing these into a precise answer. We formulate compositional tasks as computation graphs to systematically quantify the level of complexity, and break down reasoning steps into intermediate sub-procedures. Our empirical findings suggest that Transformers solve compositional tasks by reducing multi-step compositional problems into linearized subgraph matching, without necessarily developing systematic problem solving skills. To round off our empirical study, we provide theoretical arguments on abstract multi-step reasoning problems that highlight how Transformers' performance will rapidly decay with increasing task complexity.

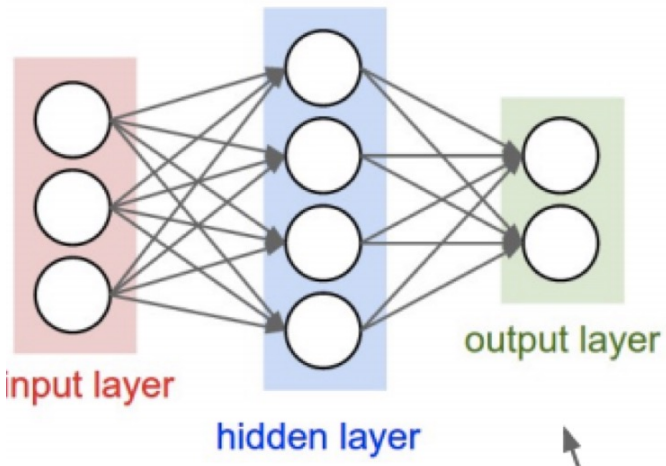
1 Introduction

"It was the epoch of ...

Large-scale Transformer ... bilities [41] ... some...

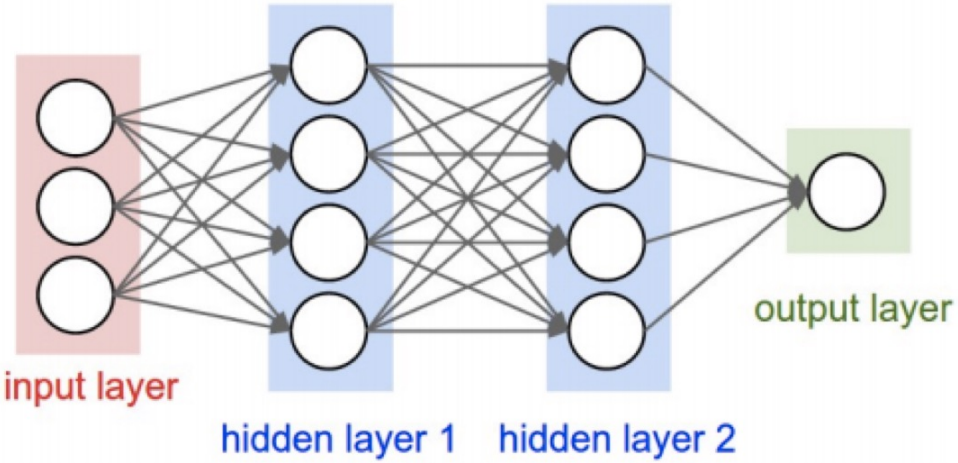
Shallow vs. Deep Layers

“shallow”



“2-layer neural net,” or
“1-hidden-layer neural net”

“deep”



“3-layer neural net,” or
“2-hidden-layer neural net”

“Fully-connected” layers

Why deepness?

Shallow vs. Deep Learning

- A shallow learner has **only one** hidden unit,

$$\mathbf{W}^{(2)} h(\mathbf{W}^{(1)} \mathbf{x})$$

- A deep learner **has more than one** hidden unit,

$$\mathbf{W}^{(3)} h(\mathbf{W}^{(2)} h(\mathbf{W}^{(1)} \mathbf{x}))$$

"Geoffrey Hinton"

Today

- About the Course
- What is AI?
- The Rise of Deep Learning
- **Why Math in AI Matters**

AI: Bakery vs. Chemistry?

Many mysteries still remain?

DOI:10.1145/3446776

Understanding Deep Learning (Still) Requires Rethinking Generalization

By Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals

Abstract

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small gap between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family or to the regularization techniques used during training.

Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice.

We interpret our experimental findings by comparison with traditional models.

We supplement this republication with a new section at the end summarizing recent progresses in the field since the original version of this paper.

underwrites the generalization ability of a model has occupied the machine learning research community for decades.

There are a variety of theories proposed to explain generalization.

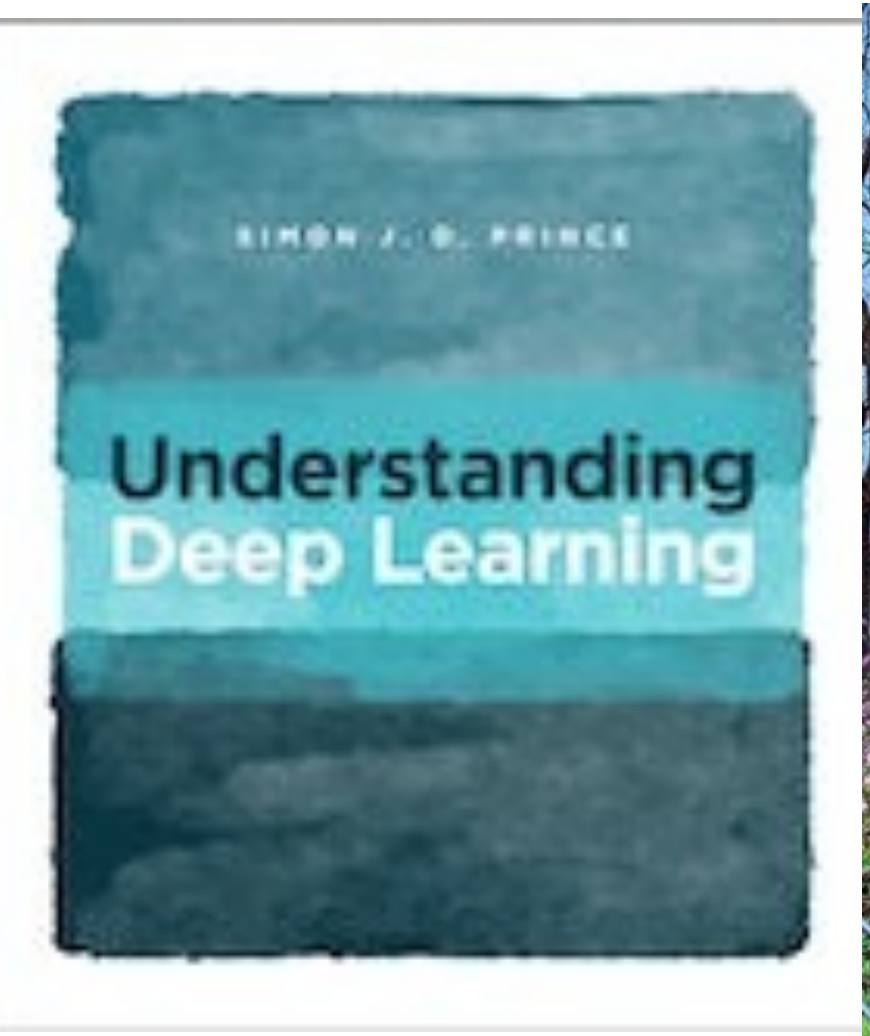
Uniform convergence, margin theory, and algorithmic stability are but a few of the important conceptual tools to reason about generalization. Central to much theory are different notions of *model complexity*. Corresponding generalization bounds quantify how much data is needed as a function of a particular complexity measure. Despite much significant theoretical work, the prescriptive and descriptive value of these theories remains debated.

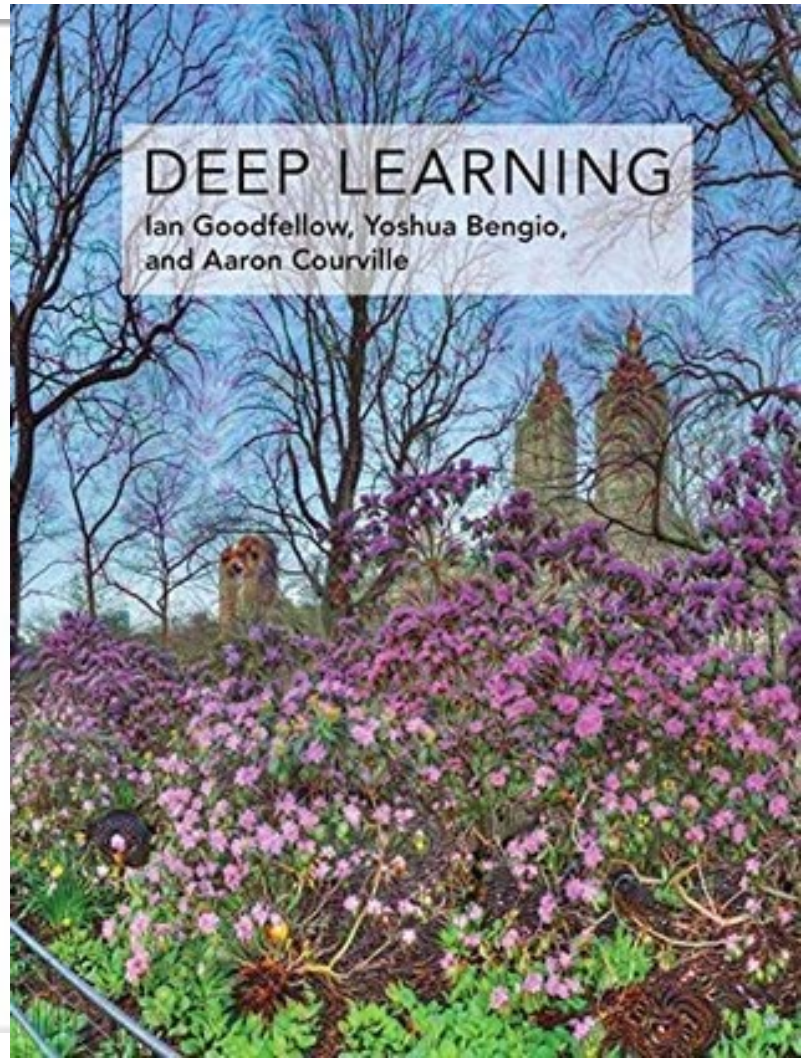
This work takes a step back. We do not offer any new theory of generalization. Rather, we offer a few simple experiments to interrogate the empirical import of different purported theories of generalization. With these experiments at hand, we broadly investigate what practices do and do not promote generalization, what does and does not measure generalization?

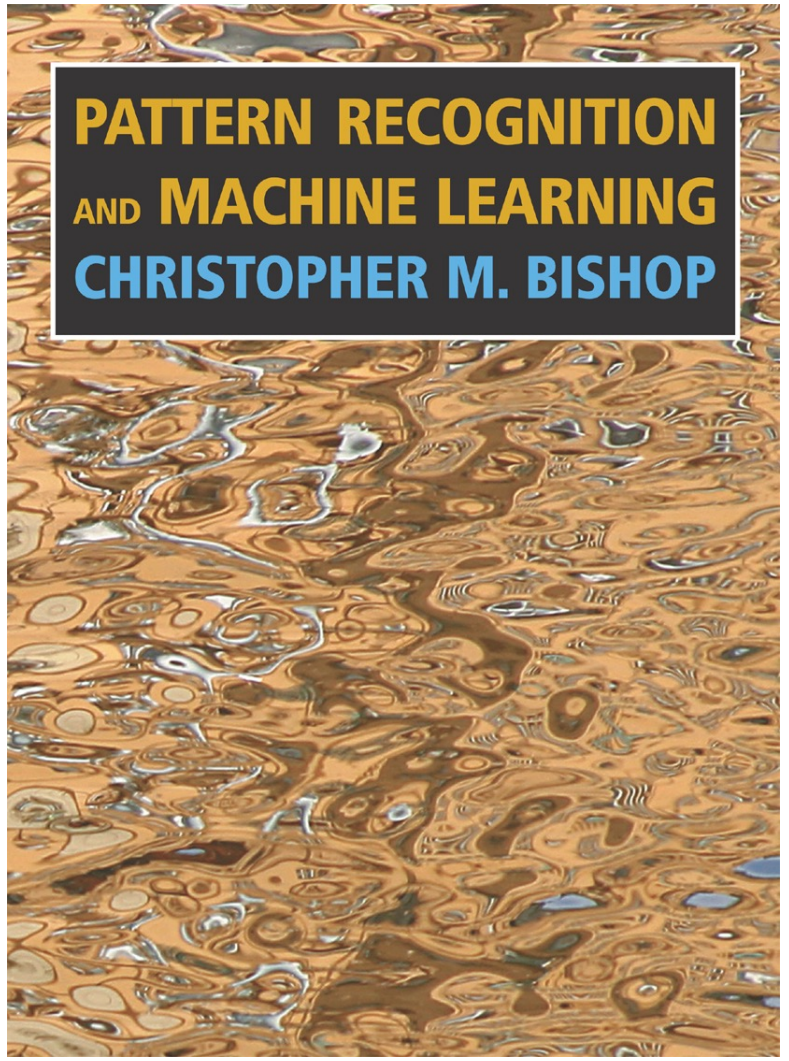
1.1. The randomization test

In our primary experiment, we create a copy of the training data where we replace each label independently by a random label chosen from the set of valid labels. A dog picture labeled "dog" might thus become a dog picture labeled "air-

More to read…


SIMON J. D. PRINCE
Understanding
Deep Learning


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio,
and Aaron Courville


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP