

A CASUAL INTRODUCTION TO AI, ML, DL, & LLMs

2024.09.03 Guest Lecture

Ethical Challenges in AI, Ethics Lab @ Georgetown University

Yo Joong "YJ" Choe

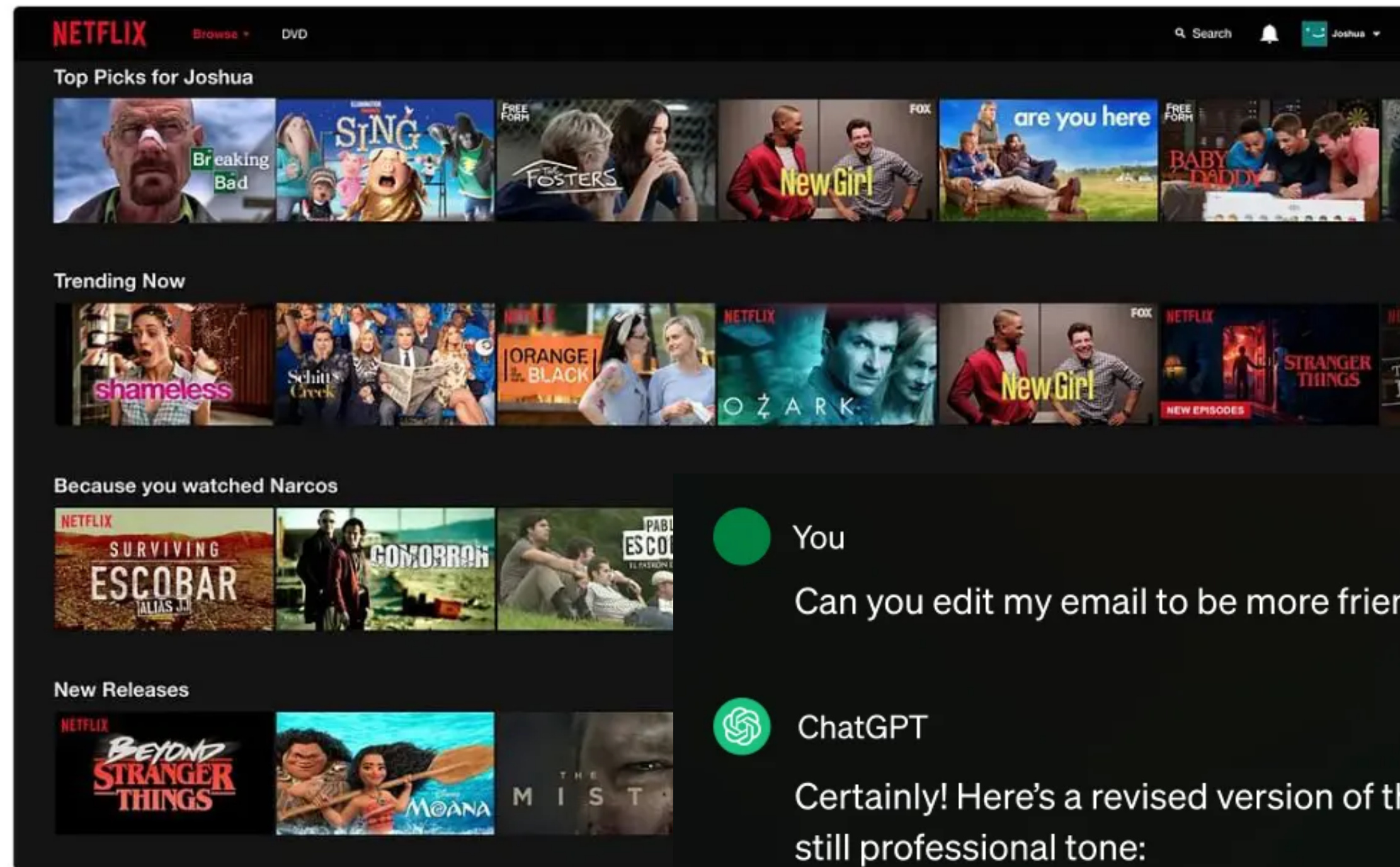
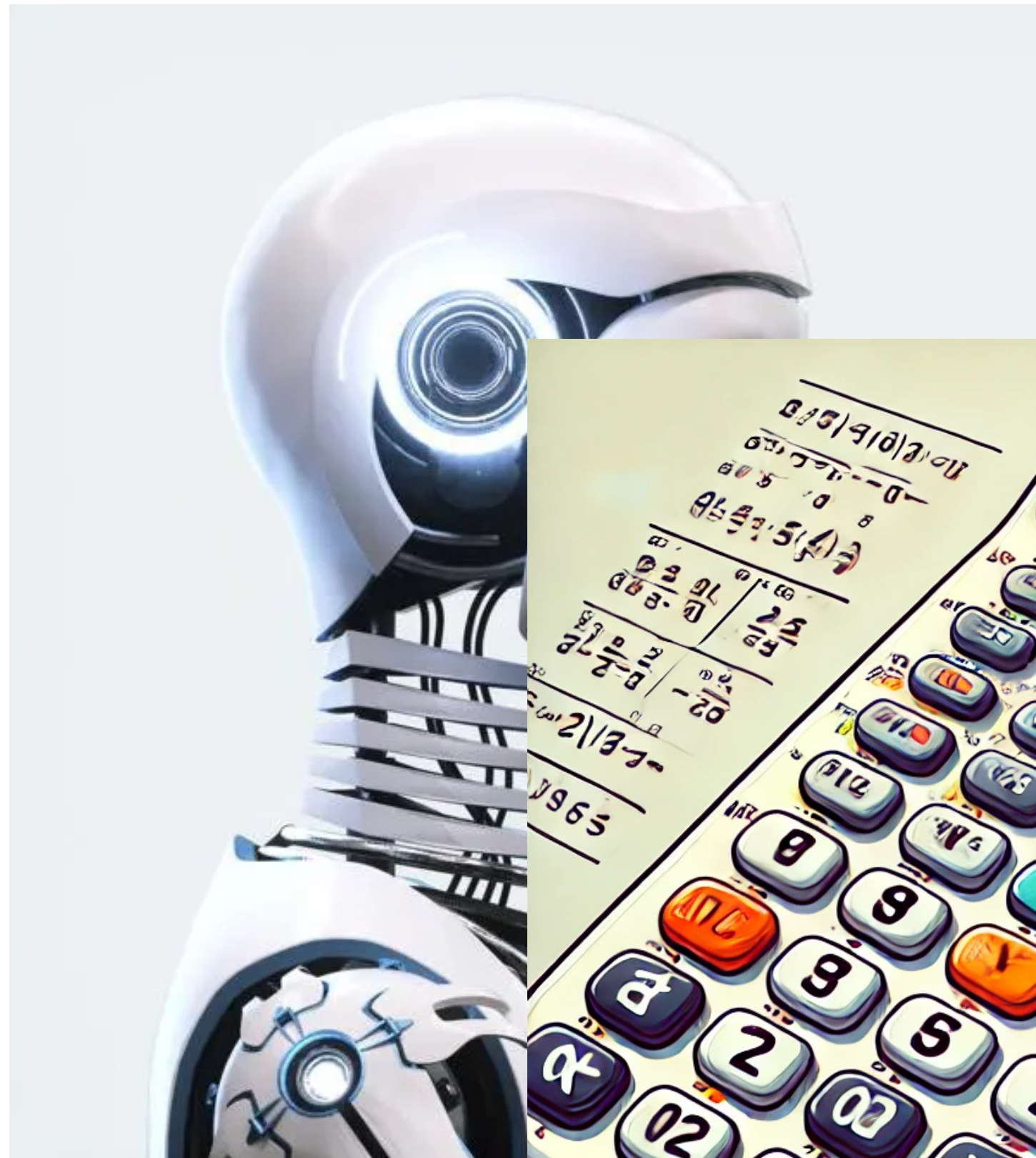
Data Science Institute @ University of Chicago


GOALS


1. Explain what AI, ML, DL, and LLMs are.
2. Explain how machines actually “learn.”
3. Introduce some key applications of AI in the past, present, and future.
(& Help you think about where the ethical challenges may arise!)

WHAT IS AI?

WHAT IS AI?



 You
Can you edit my email to be more friendly, but still professional?

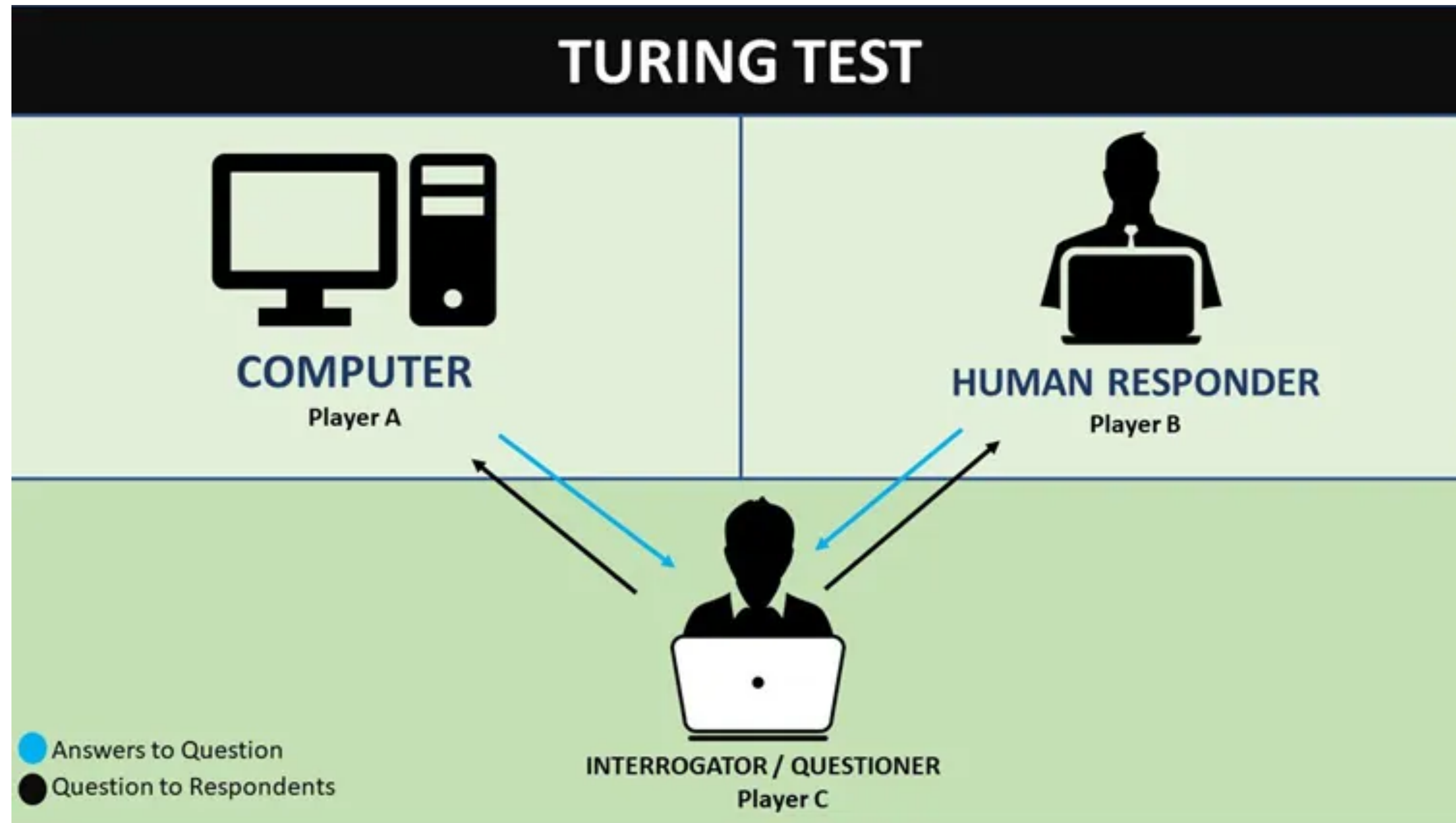
 ChatGPT
Certainly! Here's a revised version of the email with a more friendly yet still professional tone:

Subject: Excited to Move Forward and Connect Further!

Hi Judy,

I hope this email finds you well. I wanted to express our genuine excitement ●

WHAT IS AI?



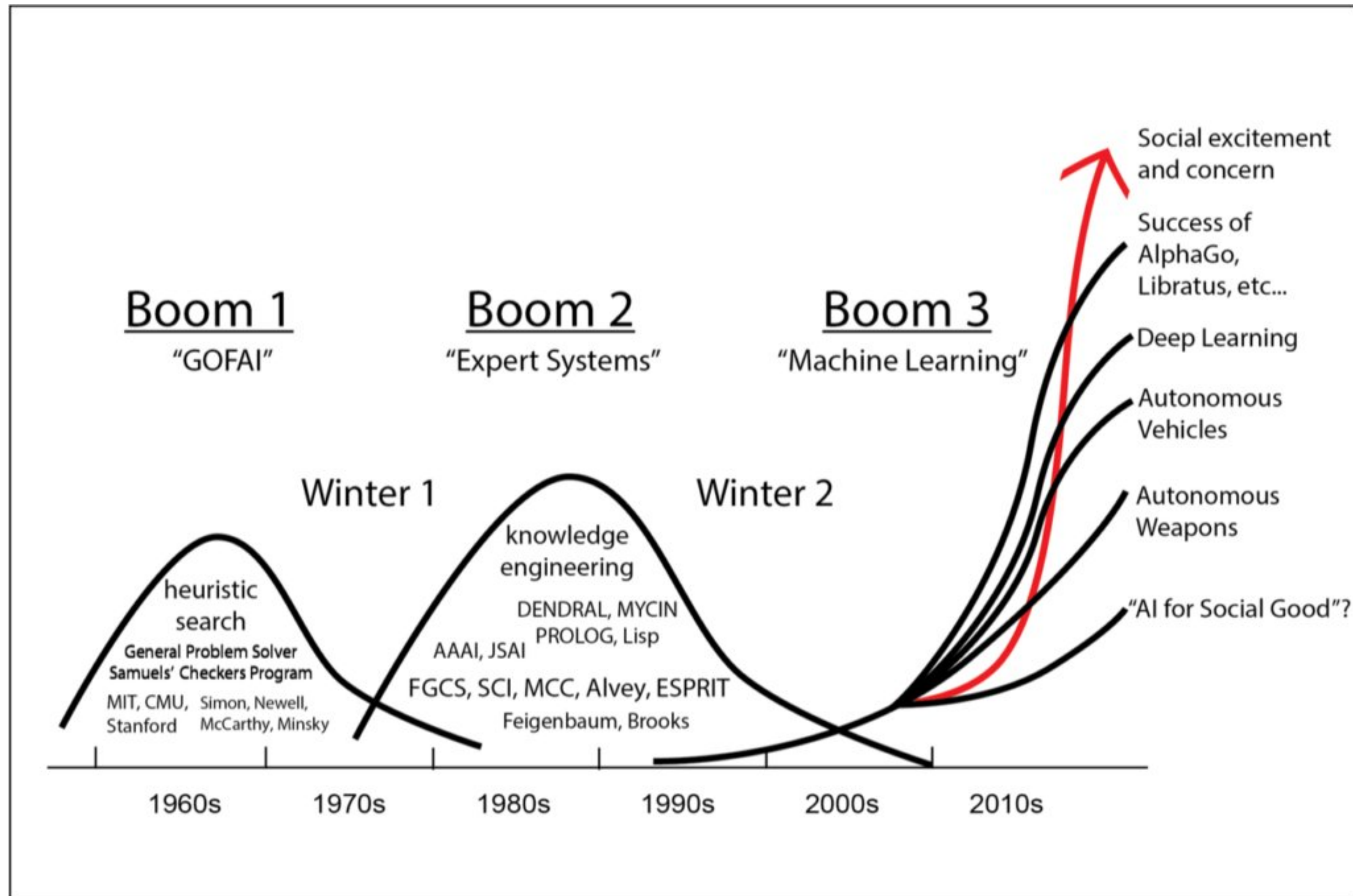
A. Turing (1950)

WHAT IS AI?

“We call programs intelligent if they exhibit **behaviors that would be regarded intelligent if they were exhibited by human beings.**”

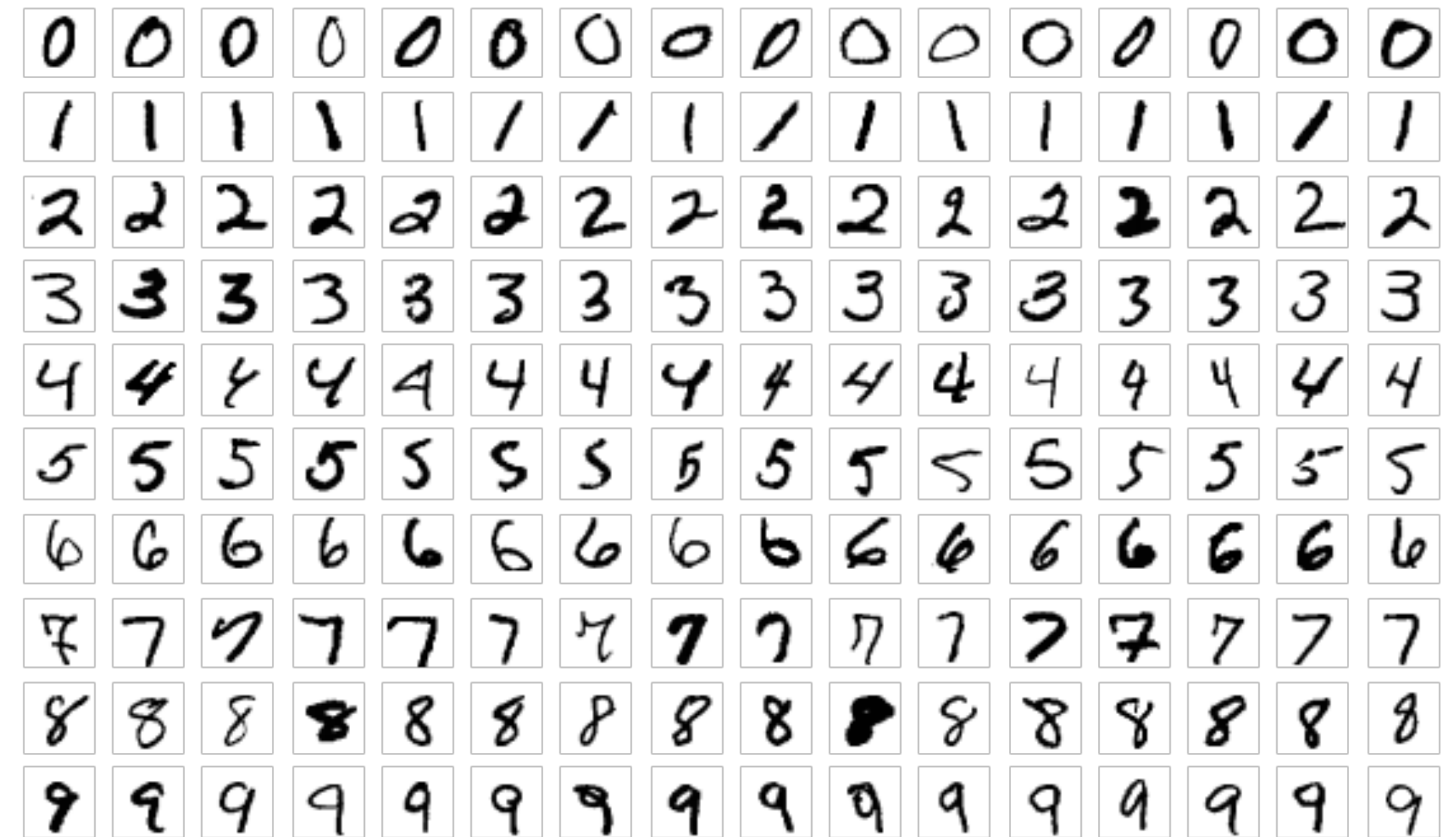
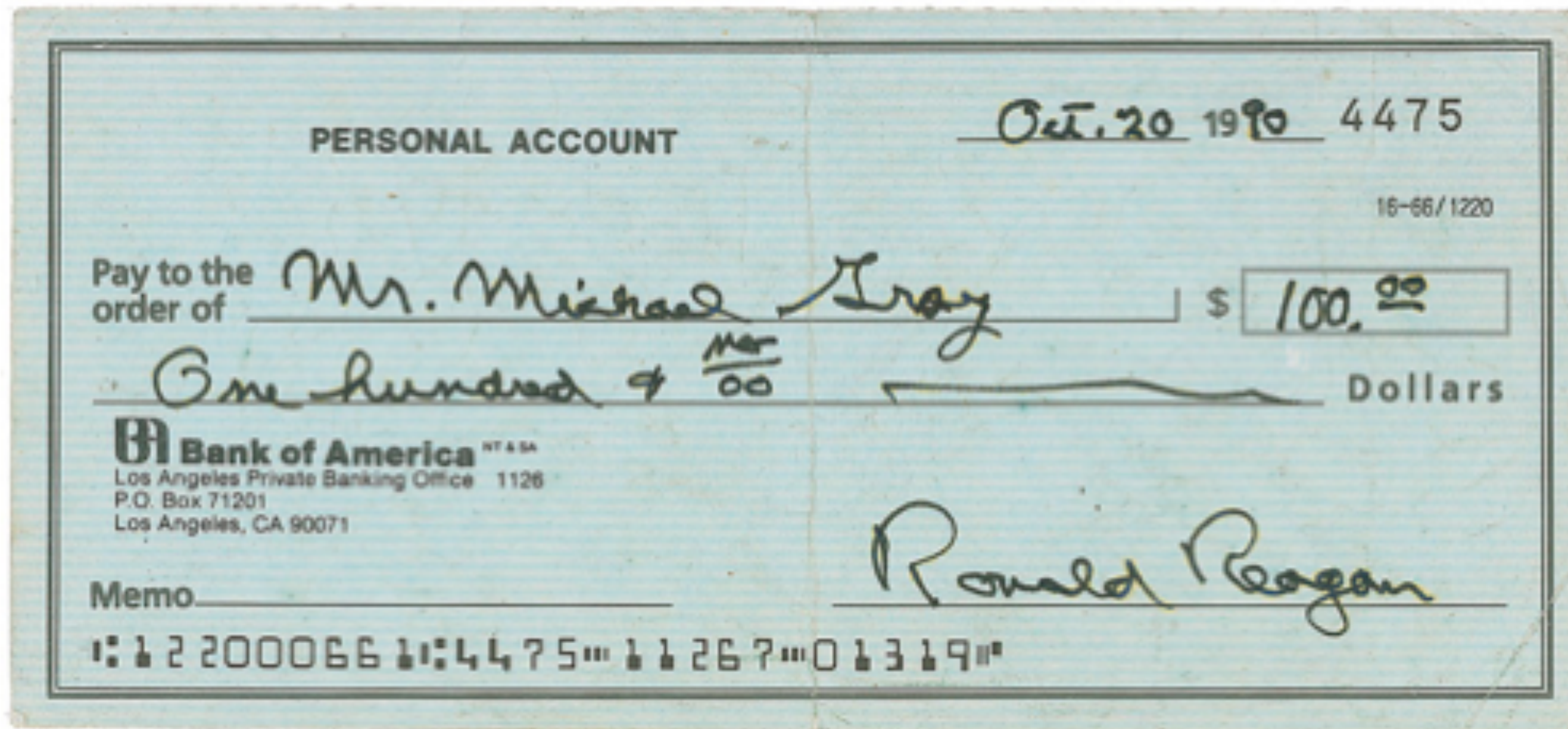
- Simon (see also: Minsky, 1968)

AI: A HISTORY OF **BOOMS** & **BUSTS**

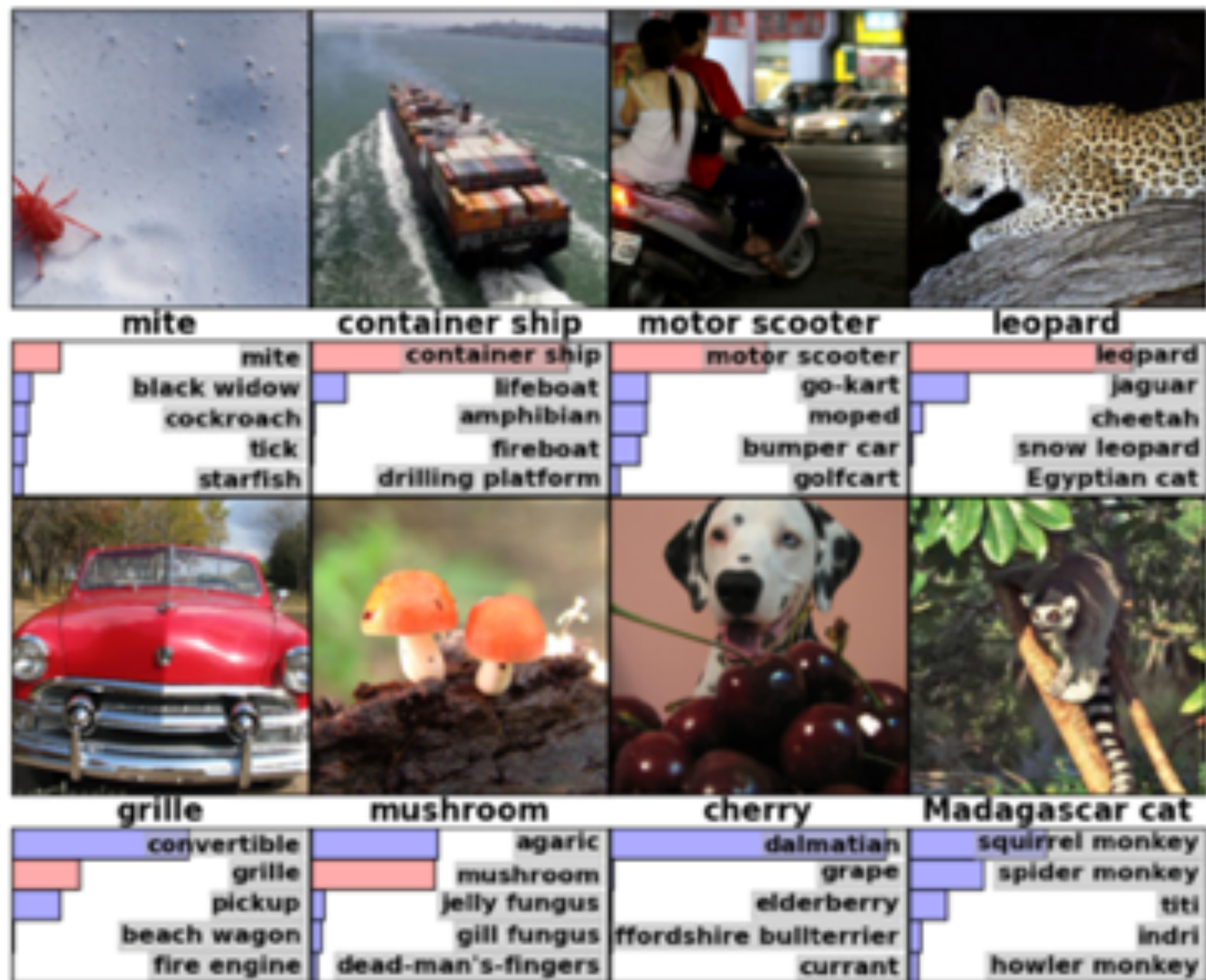


SUCCESS STORIES OF AI

HANDWRITING RECOGNITION



OBJECT RECOGNITION & DETECTION



Krizhevsky et al. (2012)



He et al. (2017)

MACHINE TRANSLATION



(1954)

Google Translate



Text Images Documents Websites

Detect language English French Spanish

Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data and thus perform tasks without explicit instructions.[1] Recently, artificial neural networks have been able to surpass many previous approaches in performance.[2]

ML finds application in many fields, including natural language processing, computer vision, speech recognition, email filtering, agriculture, and medicine.[3][4] When applied to business problems, it is known under the name predictive analytics. Although not all machine learning is statistically based, computational statistics is an important source of the field's methods.

The mathematical foundations of ML are provided by mathematical optimization (mathematical programming) methods. Data mining is a related (parallel) field of study, focusing on exploratory data analysis (EDA) through unsupervised learning.[6][7]



980 / 5,000

French English Spanish

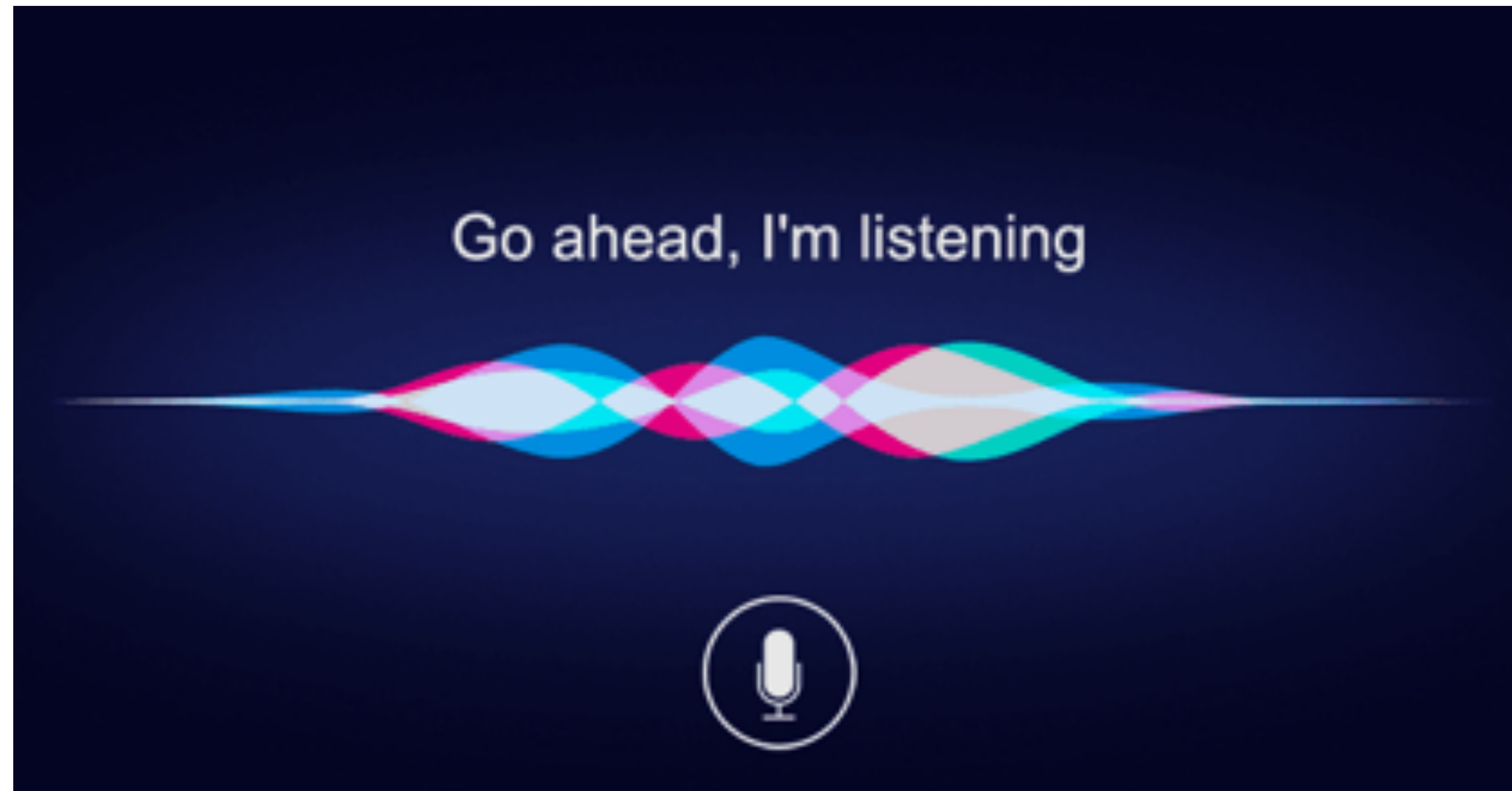
L'apprentissage automatique (ML) est un domaine d'étude de l'intelligence artificielle qui s'intéresse au développement et à l'étude d'algorithmes statistiques capables d'apprendre à partir de données et de généraliser à des données invisibles et ainsi d'effectuer des tâches sans instructions explicites.[1] Récemment, les réseaux neuronaux artificiels ont pu surpasser de nombreuses approches précédentes en termes de performances.[2]

Le ML trouve des applications dans de nombreux domaines, notamment le traitement du langage naturel, la vision par ordinateur, la reconnaissance vocale, le filtrage des e-mails, l'agriculture et la médecine.[3][4] Lorsqu'il est appliqué à des problèmes commerciaux, il est connu sous le nom d'analyse prédictive. Bien que l'apprentissage automatique ne soit pas entièrement basé sur des statistiques, les statistiques informatiques constituent une source importante des méthodes du domaine.

Les fondements mathématiques du ML sont fournis par des méthodes d'optimisation mathématique (programmation mathématique). L'exploration de données est un domaine d'étude connexe (parallèle), axé sur l'analyse exploratoire des données (EDA) par le biais d'un apprentissage non supervisé.[6][7]



SPEECH RECOGNITION & SYNTHESIS



RECOMMENDER SYSTEMS

The image shows a screenshot of the YouTube homepage. On the left is a dark sidebar with navigation options: Home, Trending, Subscriptions, Library, History, Watch later, Liked videos, and Math. Below these are subscription channels: The Office, Antastesia, and Supreme Bana... The main content area is titled "Recommended" and displays a grid of video thumbnails. Each thumbnail includes a video preview, a duration timer in the bottom right corner, a title, a channel name, and view/viewer information.

Video Title	Channel	Views	Time Ago	Duration
How craving attention makes you less creative Joseph...	TED	661K views	6 days ago	13:16
Pride and Prejudice: Marriage Proposal	BBC Studios	2.3M views	11 years ago	4:05
I Ate Like BILLIE EILISH For 24 Hours..	Supreme Banana	97K views	1 week ago	17:13
D'Arcy Carden Talks About Impersonating Her...	Late Night with Seth Meye...	797K views	8 months ago	6:25
TECHNIQUE CRITIQUE S1 • E12	WIRED	2.2M views	9 days ago	25:20
Accent Expert Breaks Down 17 Actors Playing Real...	WIRED	2.2M views	9 days ago	25:20
La littérature est-elle morte?	Antastesia	6.2K views	20 hours ago	21:09
Kardashians at Area 51	SimgmProductions	235K views	5 days ago	5:31
73 Questions With Olivia Colman Vogue	Vogue	587K views	1 week ago	10:52

IBM DEEP BLUE (CHESS)



The 1997 "Rematch" between Garry Kasparov & IBM Deep Blue.



ALPHAGO & ALPHASTAR



Go (2016)



Starcraft II (2019)

LIBRATUS (POKER)



Libratus, the poker AI (2017)

Name	Rank	Results (in chips)
Dong Kim	1	-\$85,649
Daniel MacAulay	2	-\$277,657
Jimmy Chou	3	-\$522,857
Jason Les	4	-\$880,087
Total:		-\$1,766,250

*Results of 4 human players against Libratus
after 20 days of competition*

THE EXPLOSION OF GENERATIVE AI: CHATGPT, MIDJOURNEY, COPILOT, SORA, ...

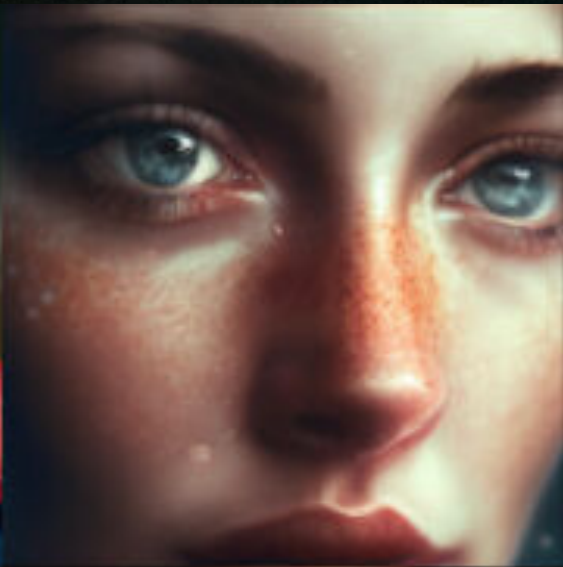
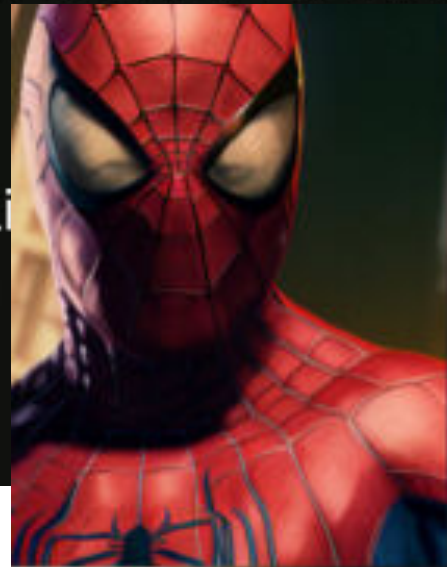
You
Can you edit my email to be more friendly, but still professional?

ChatGPT
Certainly! Here's a revised version of the email with a more friendly yet still professional tone:

Subject: Excited to Move Forward and Connect Further!

Hi Judy,

I hope this email finds you well and brings you a sense of excitement



```
GITHUB COPILOT: CHAT
monalisa
Write a unit test for this function

GitHub Copilot

import unittest
import datetime

class TestParseExpenses(unittest.TestCase):
    def test_parse_expenses_with_valid_input(self):
        expenses_string = '2023-01-02 -34.01 USD\n2023-01-03 2.59 DKK\n2023-01-03 -2.72 EUR'
        expected_output = [(datetime.datetime(2023, 1, 2), -34.01, 'USD'), (datetime.datetime(2023, 1, 3), 2.59, 'DKK'), (datetime.datetime(2023, 1, 3), -2.72, 'EUR')]
        self.assertEqual(parse_expenses(expenses_string), expected_output)

    def test_parse_expenses_with_empty_input(self):
        expenses_string = ''
        expected_output = []
        self.assertEqual(parse_expenses(expenses_string), expected_output)

if __name__ == '__main__':
    unittest.main()

The code assumes that the datetime module is installed and that the input string is formatted correctly. The test cases cover different scenarios such as valid input, input with comments, invalid date format, and invalid currency.

Ask a question or type '/' for commands
```

```
module.py
import datetime

def parse_expenses (expenses_string):
    """Parse the list of expenses and return the list of triples (date, amount, currency)
    Ignore lines starting with #.
    Parse the date using datetime.
    Example expenses_string:
    2023-01-02 -34.01 USD
    2023-01-03 2.59 DKK
    2023-01-03 -2.72 EUR
    """
    expenses = []

    for line in expenses_string.splitlines():
        if line.startswith("#"):
            continue
        date, value, currency = line.split(" ")
        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"), float(value), currency))
```



**WHAT IS MACHINE LEARNING?
(HOW MACHINES ACTUALLY “LEARN”)**

WHAT IS MACHINE LEARNING?

A. L. Samuel

Some Studies in Machine Learning Using the Game of Checkers

Abstract: Two machine-learning procedures have been investigated in some detail using the game of checkers. Enough work has been done to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote more, it can learn to do this in a remarkably short period of time (8 or 10 hours of when given only the rules of the game, a sense of direction, and a redundant c parameters which are thought to have something to do with the game, but whose co weights are unknown and unspecified. The principles of machine learning verified are, of course, applicable to many other situations.

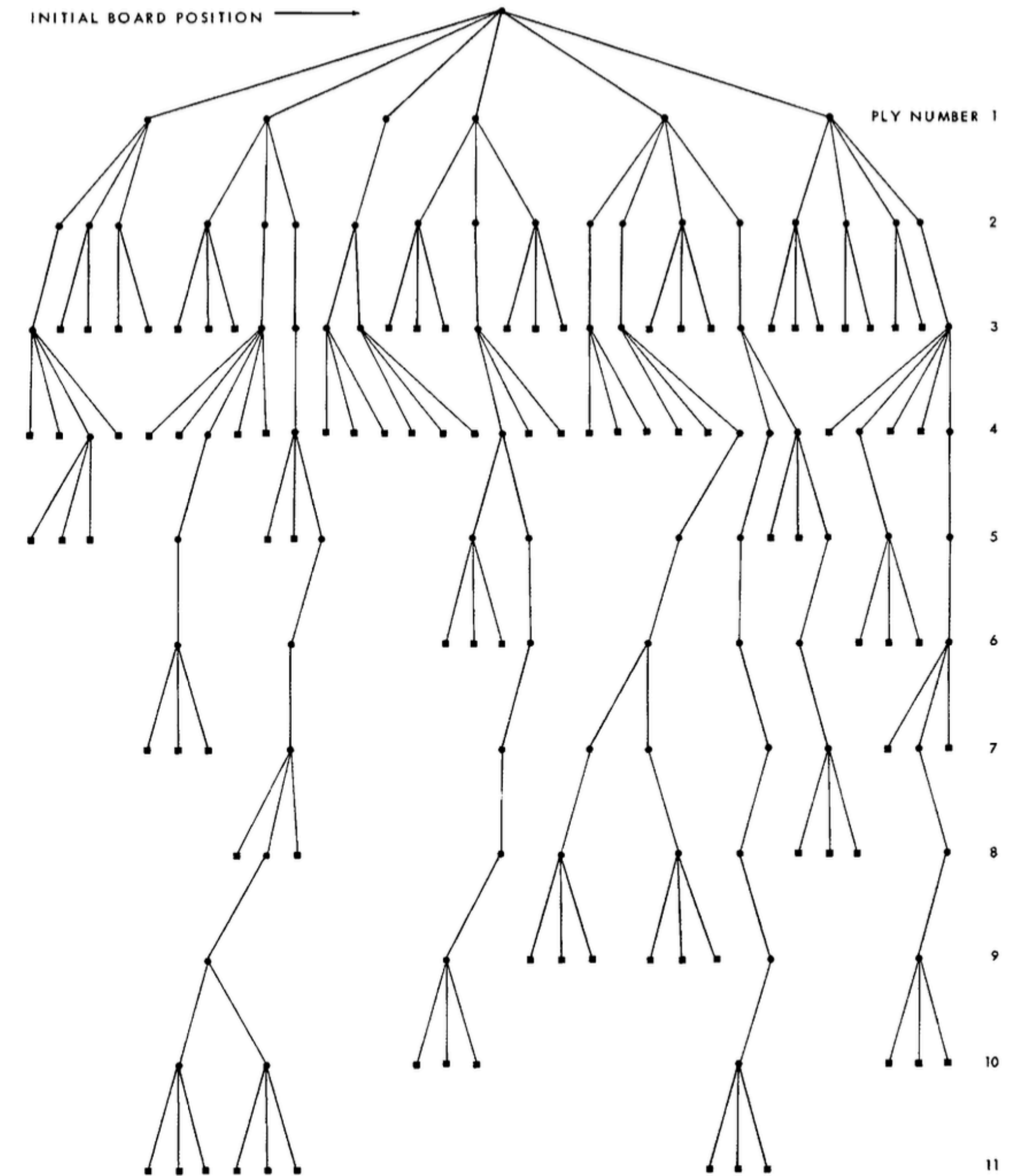


Figure 1 A "tree" of moves which might be investigated during the look-ahead procedure. The actual branchings are much more numerous than those shown, and the "tree" is apt to extend to as many as 20 levels.

WHAT IS MACHINE LEARNING?

“Programming computers to learn from experience should eventually **eliminate the need for much of this detailed programming effort.**”

- A. Samuel (1959)

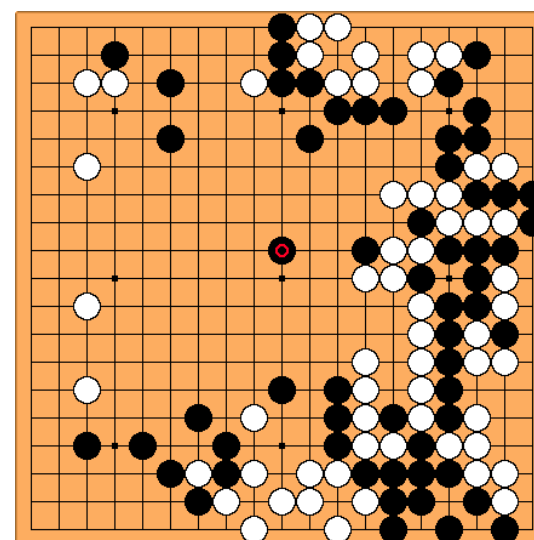


Weekday,
Afternoon,
Sunny, 85F,
10mph Wind,
Nearby Georgetown,
During the Semester,
...



Bikeshare
Demand:
80

“Thank you for helping me
maintain my weight.”



C16

HISTORY OF AI = AUTOMATION

**Artificial
Intelligence**

Extracted features +
Hand-coded rules

1950s

**Machine
Learning**

Extracted features +
Learned rules

1980s

**Representation
Learning**
(Deep Learning)

Learned features +
Learned rules

2010s

A (SLIGHTLY MORE) MODERN DEFINITION

“A computer program is said to learn [...] if its performance at [given] **tasks**, as measured by some **performance measure**, improves with **experience**.”

- T. Mitchell (1997)



KEY OBJECTIVES IN ML

- **Generalization**

- How well can the algorithm process previously unseen examples?

- **Representation**

- How well can the algorithm capture relevant features of (raw) data?

EXAMPLE: SPAM CLASSIFICATION

Sun 3/13/2016 9:15 AM
[Redacted]
Debt #69677 , Customer Case Nr.: 492
To: [Redacted]

Message confirm_92585533.zip

Dear Customer,

Despite our constant reminders, we would like to note that the mentioned debt #69677 for \$910,62 is still overdue for payment.

We would appreciate your cooperation on this case and ask you to make the payment as soon as possible.

Unless the full payment is received by April 1st, 2016 this case will be transferred to the debt collection agency, will seriously damage your credit rating.

Please, find the attachment enclosed to the letter below.

We hope on your understanding.

Kind regards,
Finance Department
[Redacted]
[Redacted],
[Redacted]
Phone nr: [Redacted]

No Items



ML



REPRESENTING EMAILS

- **Examples of features in an email:**
 - Number of words
 - Term frequencies
 - The domain of sender's email address
 - Whether the text contains keywords:
“debt”, price (“\$3.99”), special characters, all caps, ...
- **Automatically learned** features for the task

LEARNING FROM EXPERIENCE

ID	# Words	Domain	\$	“debt”	“FREE”	Label
0	120	<u>gmail.com</u>	Yes	Yes	Yes	Spam
1	500	<u>gmail.com</u>	No	No	No	Ham
2	400	<u>cmu.edu</u>	Yes	Yes	No	Ham
3	1200	<u>suspicio.us</u>	Yes	No	Yes	Spam

LEARNING FROM EXPERIENCE

1. Define a **loss function** (the “performance measure”).
 - 0-1 Loss: 1 if incorrect, 0 otherwise.
2. Define a **model** between features (“input”) and the label (“output”).
 - *The model often has parameters that can be learned, e.g.:*
 - $\text{Odds}(\text{Spam}/\text{Ham}) = \exp[w_1 \cdot (\#\text{Words}) + w_2 \cdot (\text{KnownDomain}) + w_3 \cdot (\text{'FREE'})]$
3. Make a prediction, and **update** your model according to the loss.

LEARNING FROM EXPERIENCE

- Initial Guess:

$$\text{Odds}(\text{Spam}/\text{Ham}) = \exp[(0.0) \cdot (\# \text{Words}) + (-1.0) \cdot (\text{KnownDomain}) + (0.0) \cdot (\text{'FREE'})] \approx 0.36$$

- Loss: 1 (incorrect)
- **Feedback (“gradient”): Add 0.001 to w_1 , 0.5 to w_2 , 0.5 to w_3**

ID	# Words	Domain	“FREE”	Label
0	120	<u>gmail.com</u>	Yes	Spam

LEARNING FROM EXPERIENCE

- Initial Guess:

$$\text{Odds}(\text{Spam}/\text{Ham}) = \exp[(0.001) \cdot (\# \text{Words}) + (-0.5) \cdot (\text{KnownDomain}) + (0.5) \cdot (\text{'FREE'})] \approx 1.65$$

- Loss: 1 (incorrect)
- **Feedback (“gradient”): Add -0.002 to w_1 , 0.25 to w_2 , 0.25 to w_3**

ID	# Words	Domain	“FREE”	Label
0	120	<u>gmail.com</u>	Yes	Spam
1	500	<u>gmail.com</u>	No	Ham

LEARNING FROM EXPERIENCE

- Initial Guess:

$$\text{Odds}(\text{Spam}/\text{Ham}) = \exp[(-0.001) \cdot (\# \text{Words}) + (-0.25) \cdot (\text{KnownDomain}) + (0.75) \cdot (\text{'FREE'})] \approx 0.52$$

- Loss: 0 (correct!)
- **Feedback (“gradient”): no change**

ID	# Words	Domain	“FREE”	Label
0	120	<u>gmail.com</u>	Yes	Spam
1	500	<u>gmail.com</u>	No	Ham
2	400	<u>cmu.edu</u>	No	Ham

LEARNING FROM EXPERIENCE

- Initial Guess:

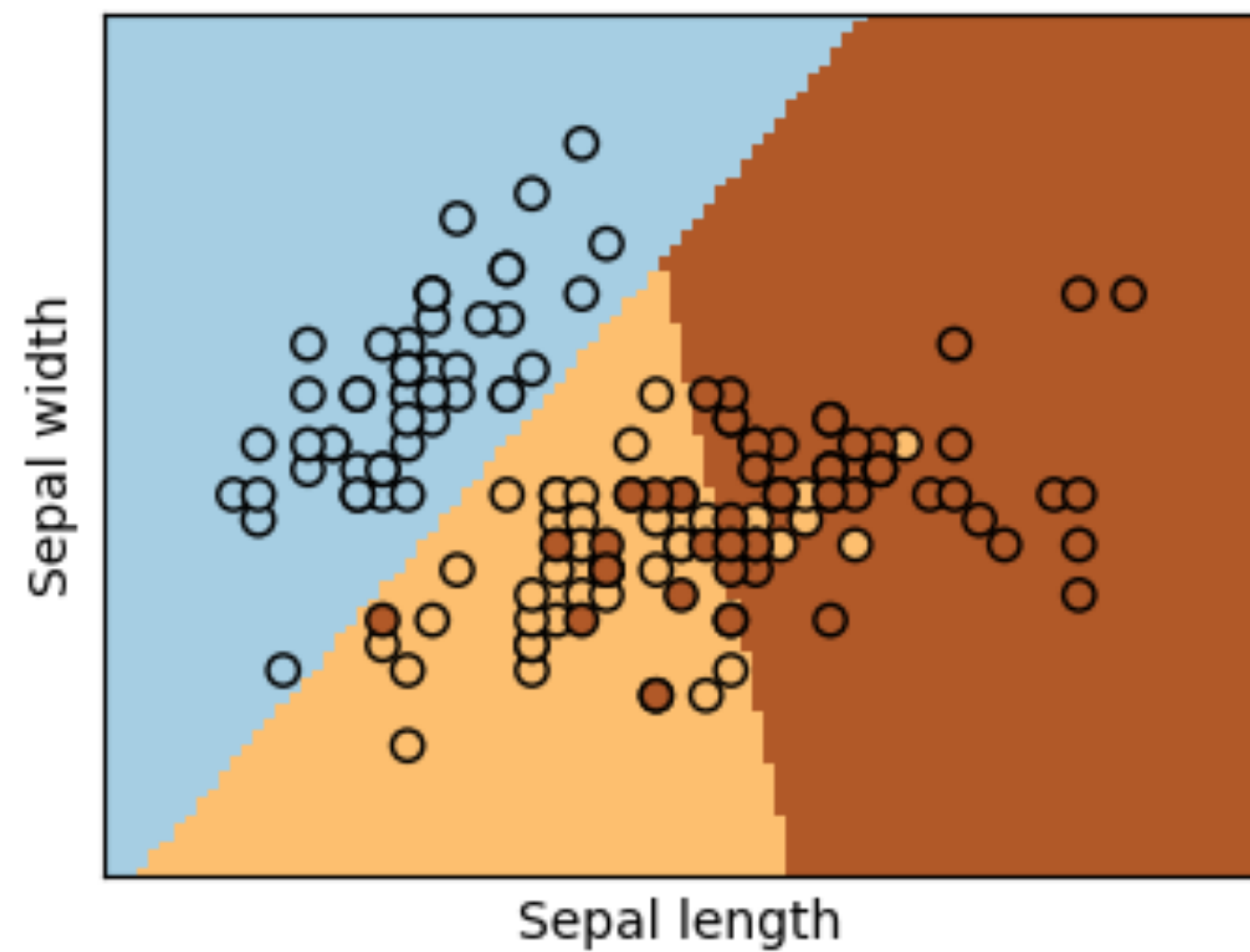
$$\text{Odds}(\text{Spam}/\text{Ham}) = \exp[(-0.001) \cdot (\# \text{Words}) + (-0.25) \cdot (\text{KnownDomain}) + (0.75) \cdot (\text{'FREE'})] \approx 0.64$$

- Loss: 1 (incorrect)
- **Feedback (“gradient”): Add 0.0005 to w_1 , -0.1 to w_2 , +0.1 to w_3**

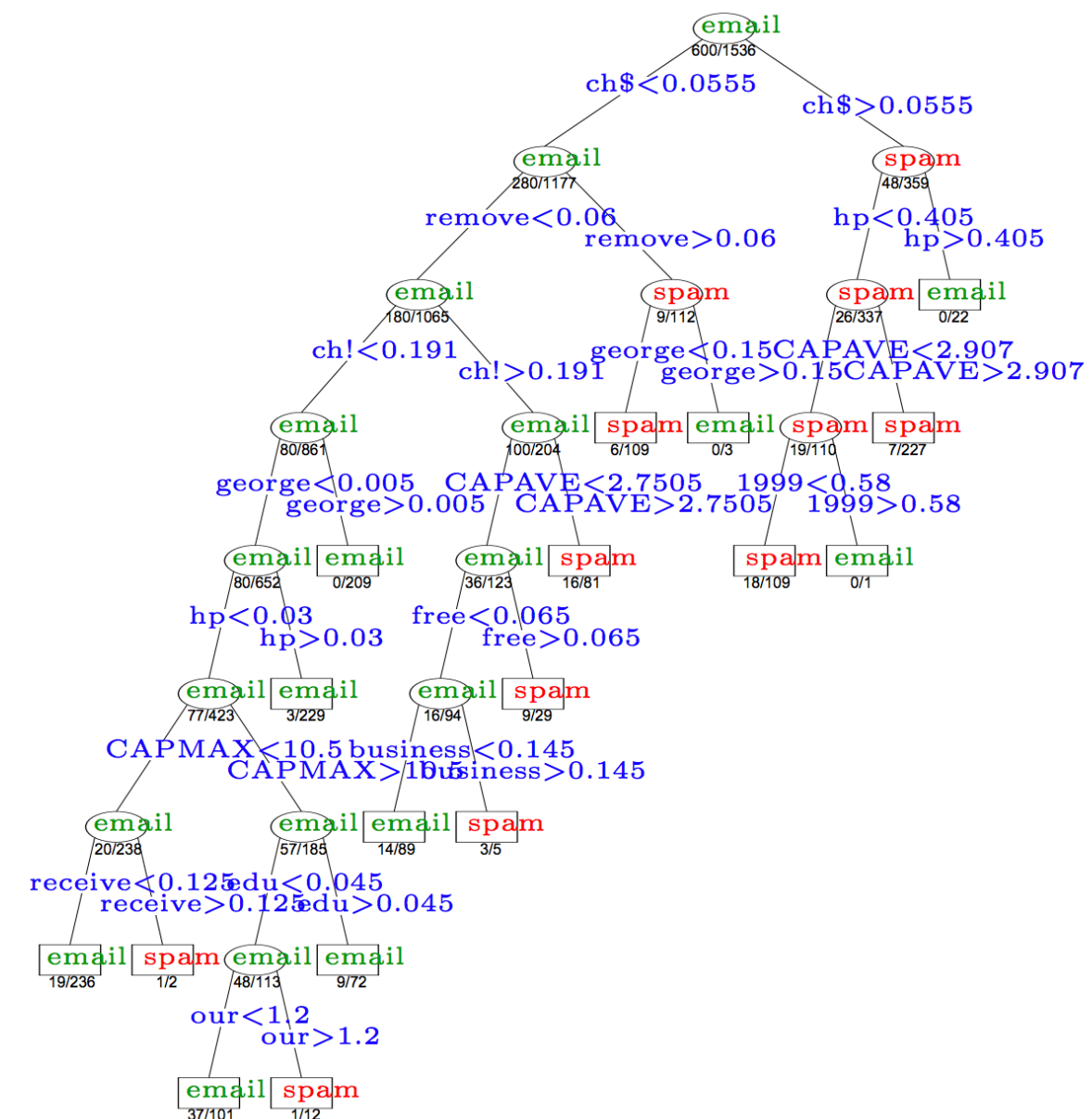
ID	# Words	Domain	“FREE”	Label
0	120	<u>gmail.com</u>	Yes	Spam
1	500	<u>gmail.com</u>	No	Ham
2	400	<u>cmu.edu</u>	No	Ham
3	1200	<u>suspicio.us</u>	Yes	Spam

SUPERVISED ML METHODS

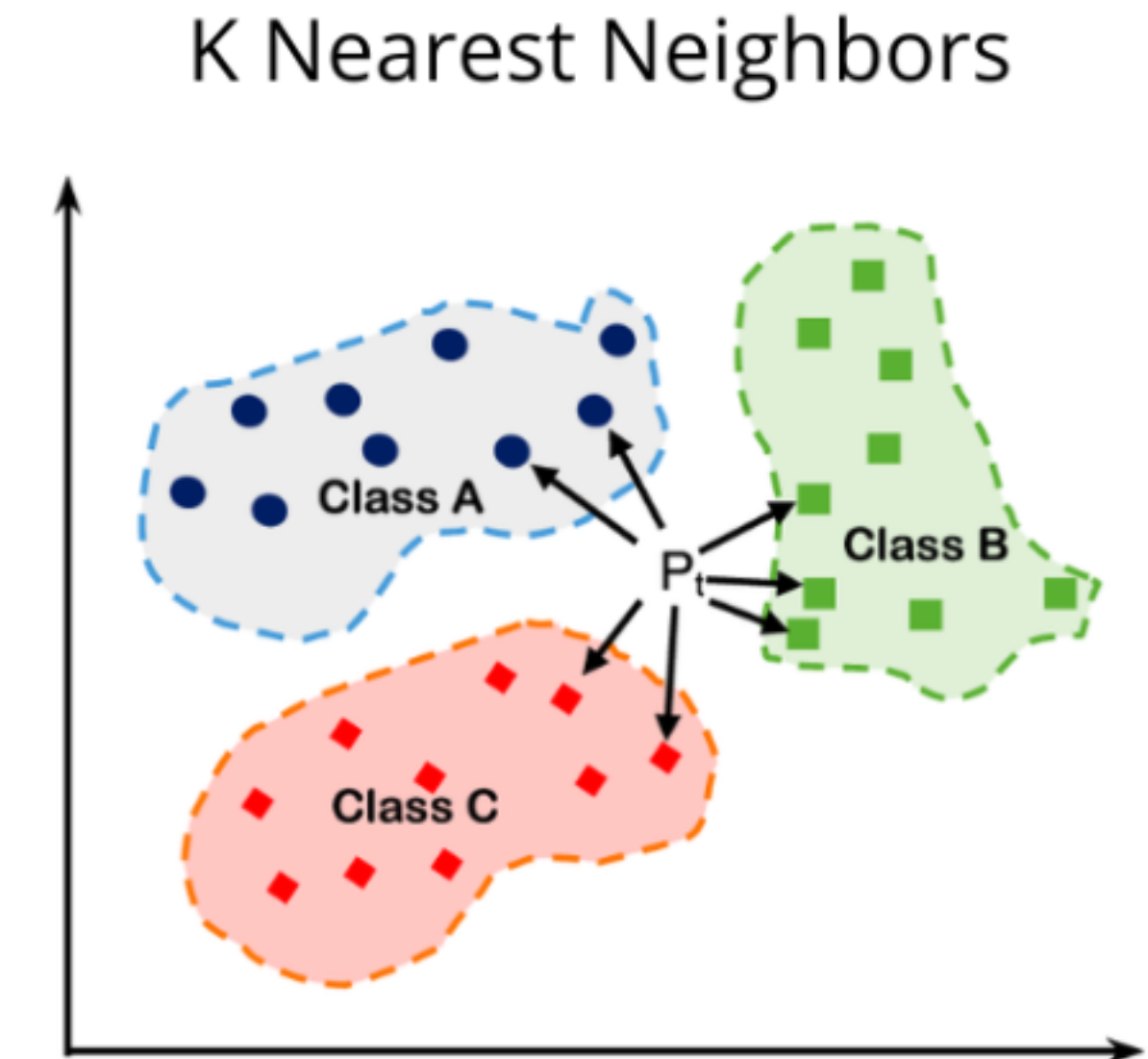
Linear/Logistic Regression



Decision Trees



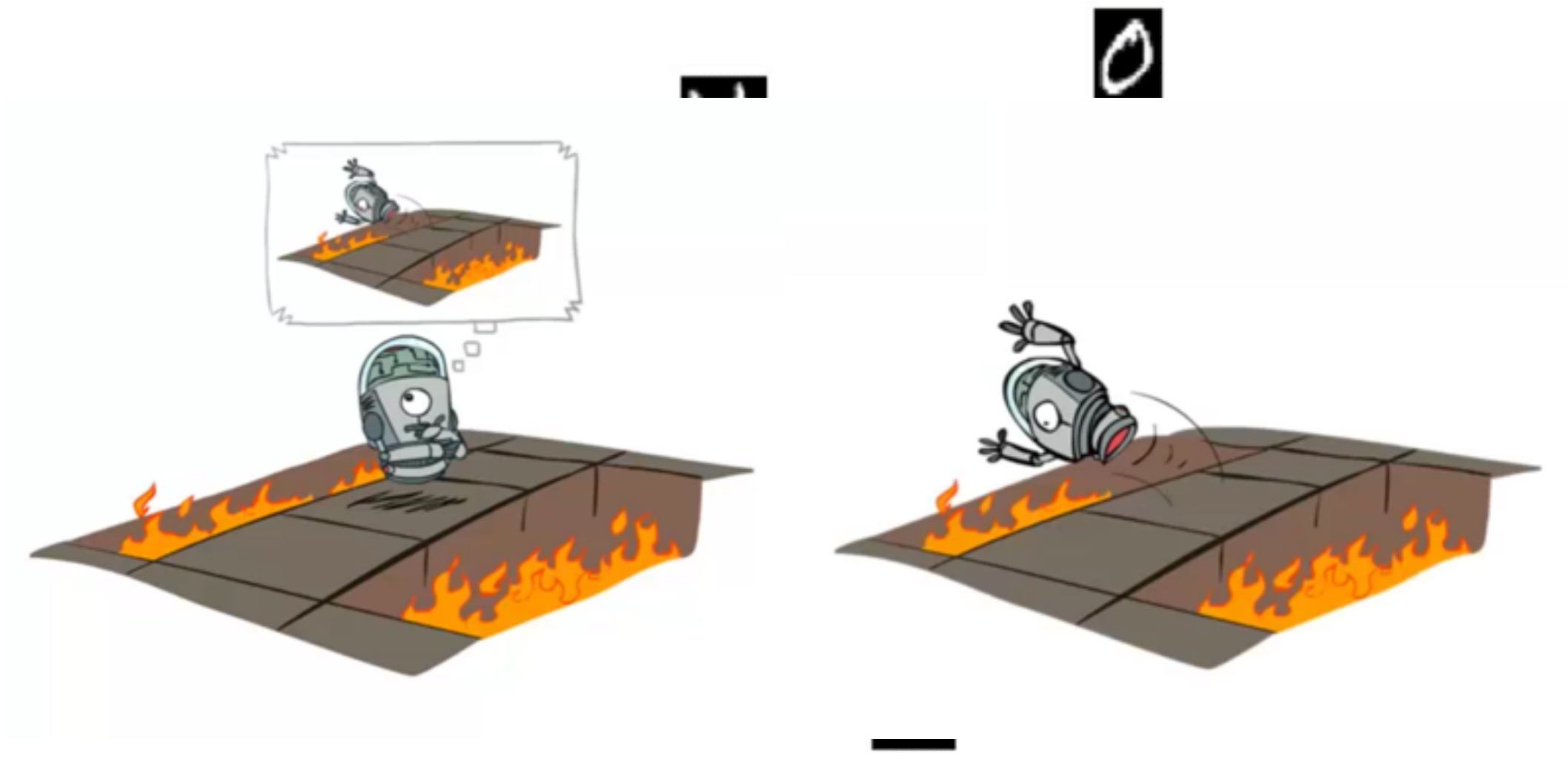
Nearest Neighbors



(...and many, many more)

OTHER TYPES OF ML TASKS

- **Unsupervised Learning:** no labels
- **Semi-supervised Learning**
- **Pre-training & Fine-tuning**
- **Reinforcement Learning:** (input, state) \rightarrow (action, next_state)
- **Online Learning, Active Learning, Meta-Learning, ...**



CASES WHEN CLASSICAL ML MIGHT BE USEFUL

- **“Tabular data” with meaningful hand-crafted features**
 - Predicting an individual’s credit/loan, political leanings, hiring decisions, etc.
 - Predicting the likelihood of cancer using key health measures
 - Predicting the chance of recidivism using demographic/socioeconomic information
- **Small data scenario**
 - Medical data analysis / genomics / neuroscience / rare events
- **(Arguably) whenever transparency/explanability/interpretability is needed**
 - Would you use AI for policing, hiring, loan decisions, etc.?

WHAT IS DEEP LEARNING?



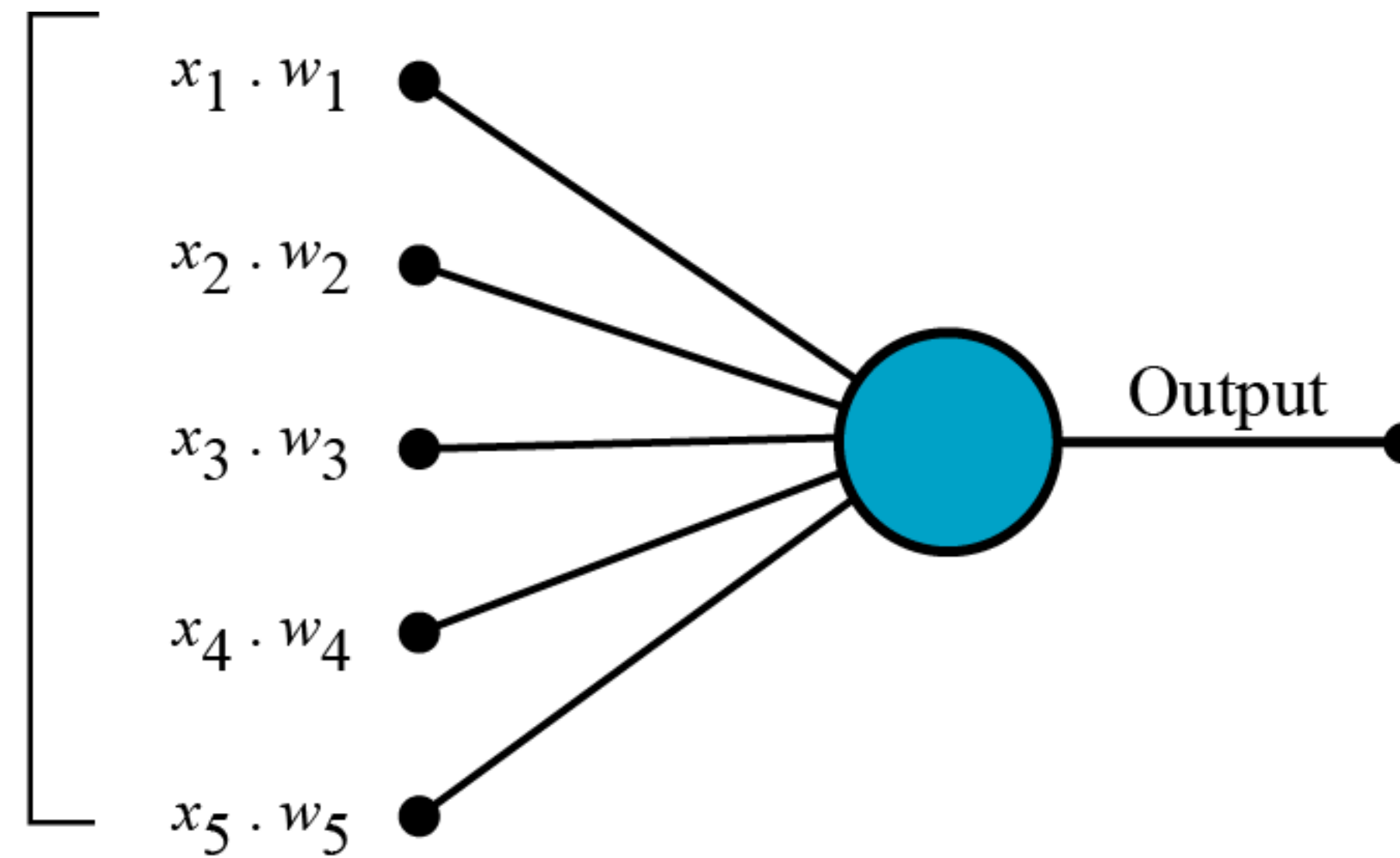
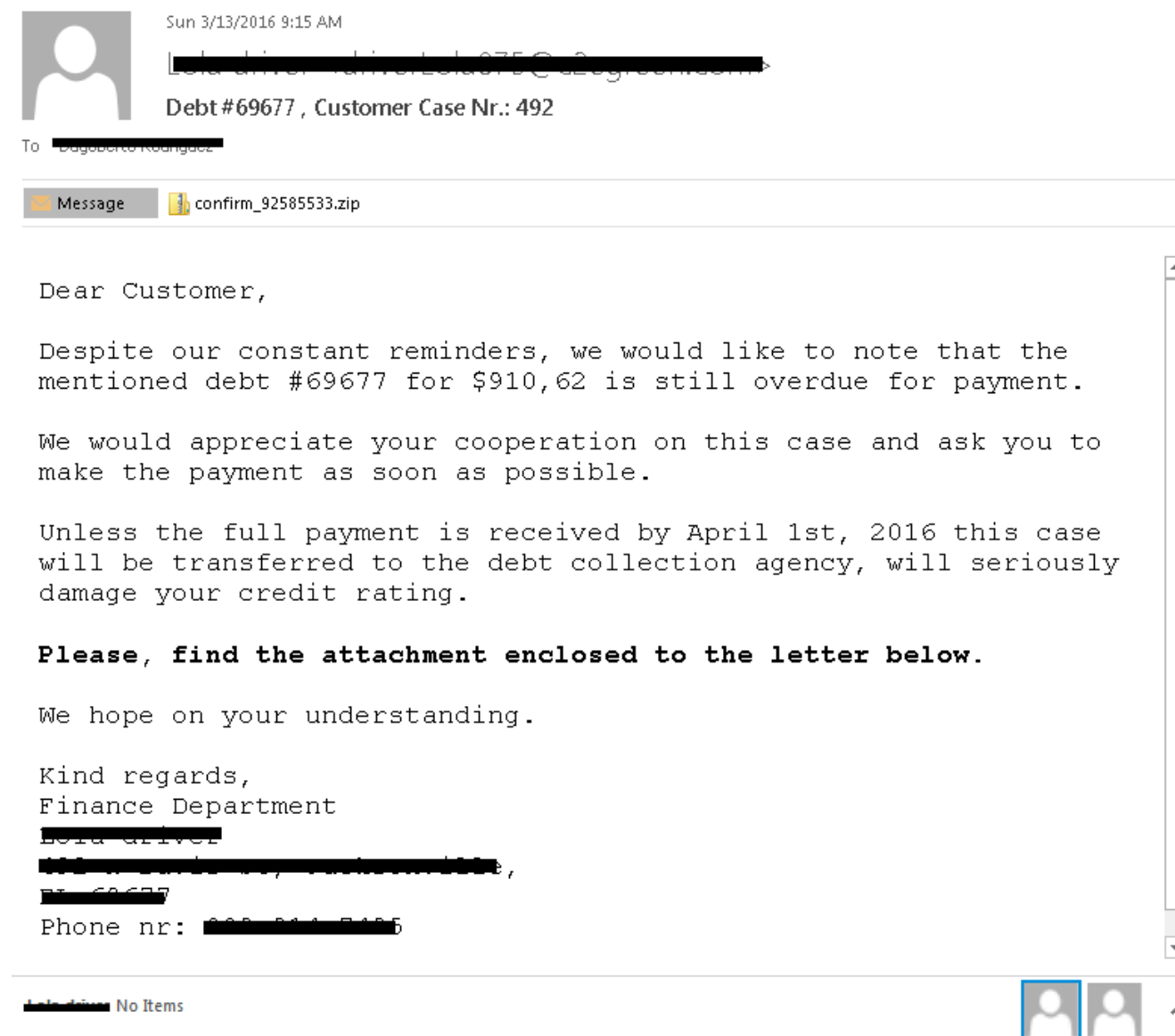
*These five years are
well past us now!
(Roughly, 2012-2017)*

IN CS, IT CAN BE HARD TO EXPLAIN
THE DIFFERENCE BETWEEN THE EASY
AND THE VIRTUALLY IMPOSSIBLE.

<https://xkcd.com/1425/>

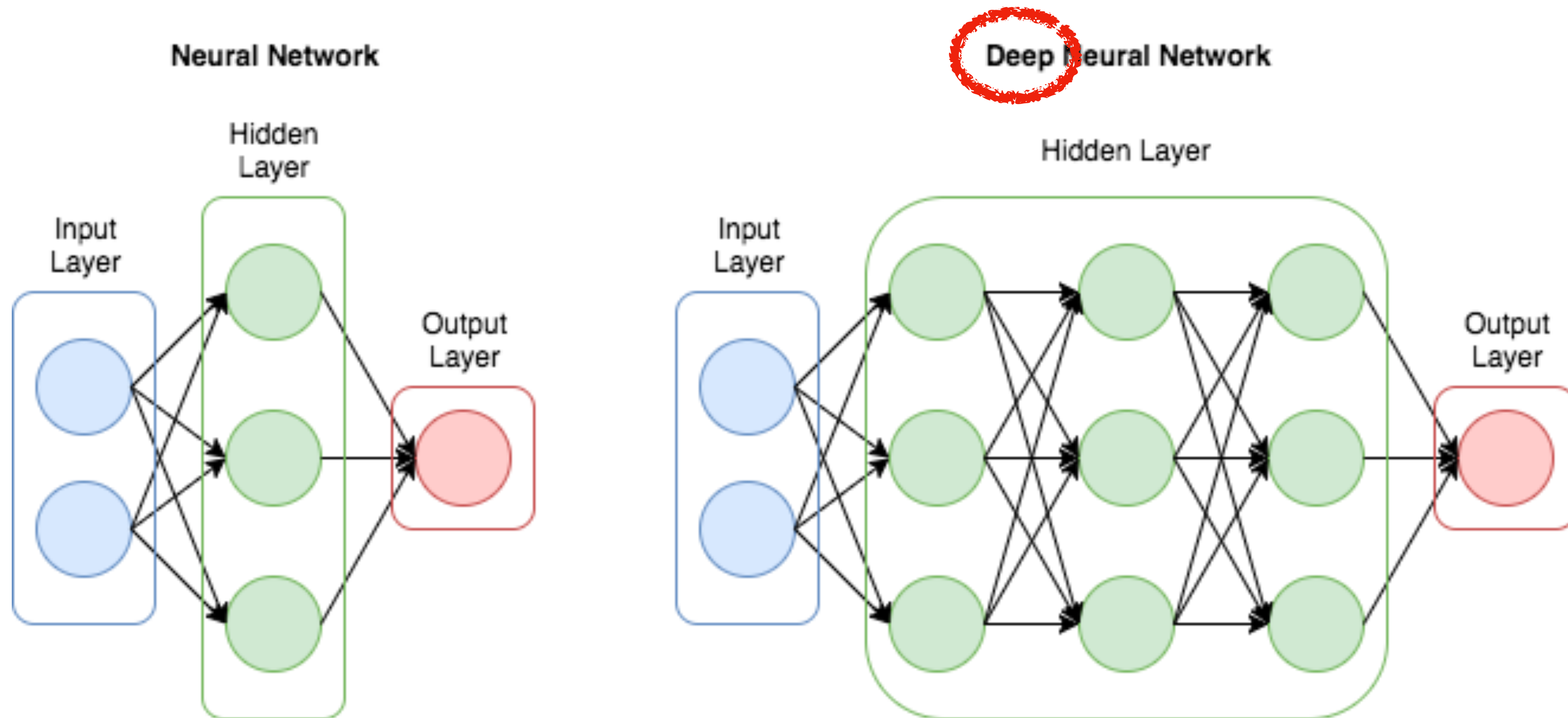
THE PERCEPTRON

(Rosenblatt 1959)



<https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53>

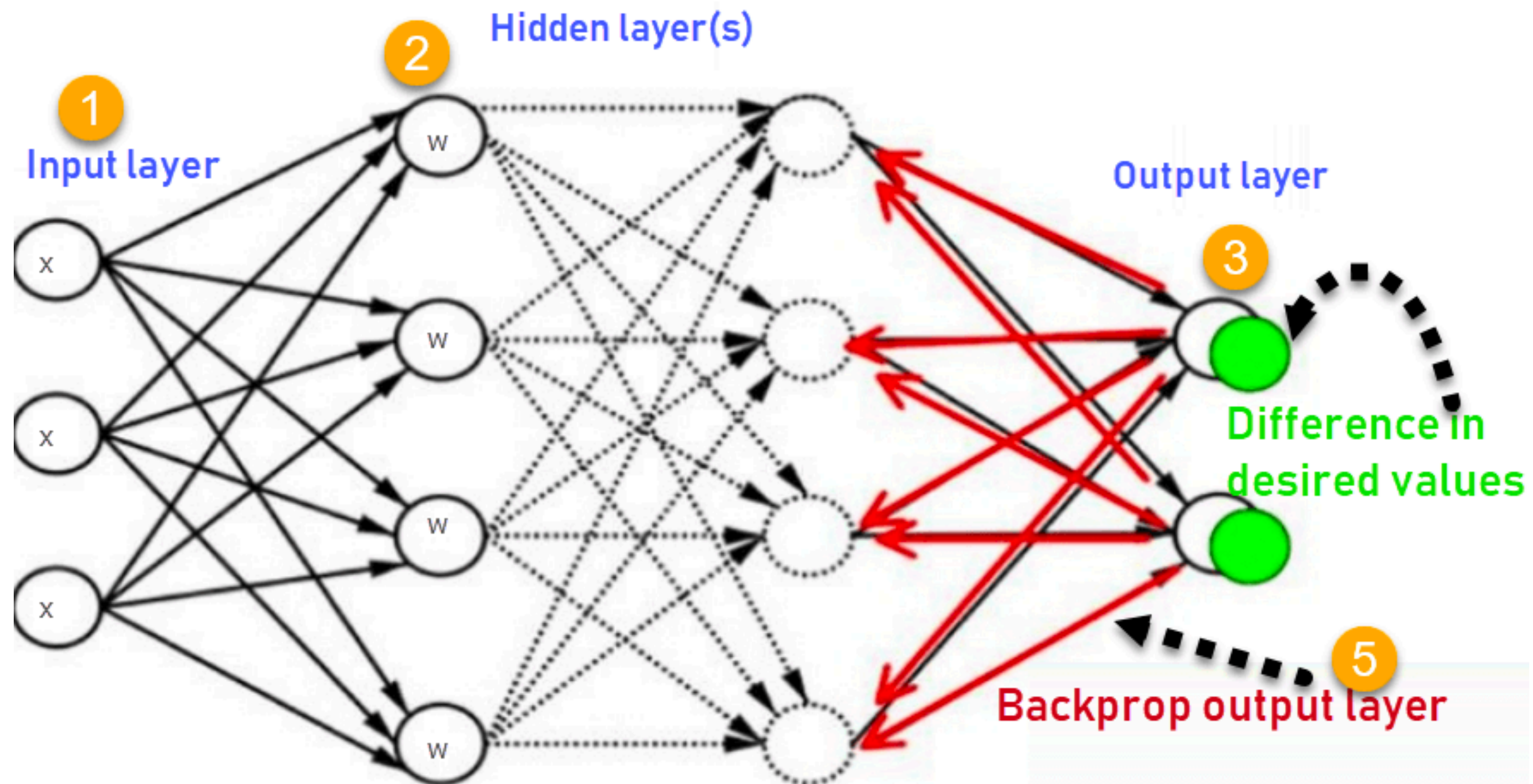
ARTIFICIAL NEURAL NETWORKS



<http://marubon-ds.blogspot.com/2017/09/simple-tutorial-to-write-deep-neural.html>

BACKPROPAGATION

(Rumelhart et al., 1986)



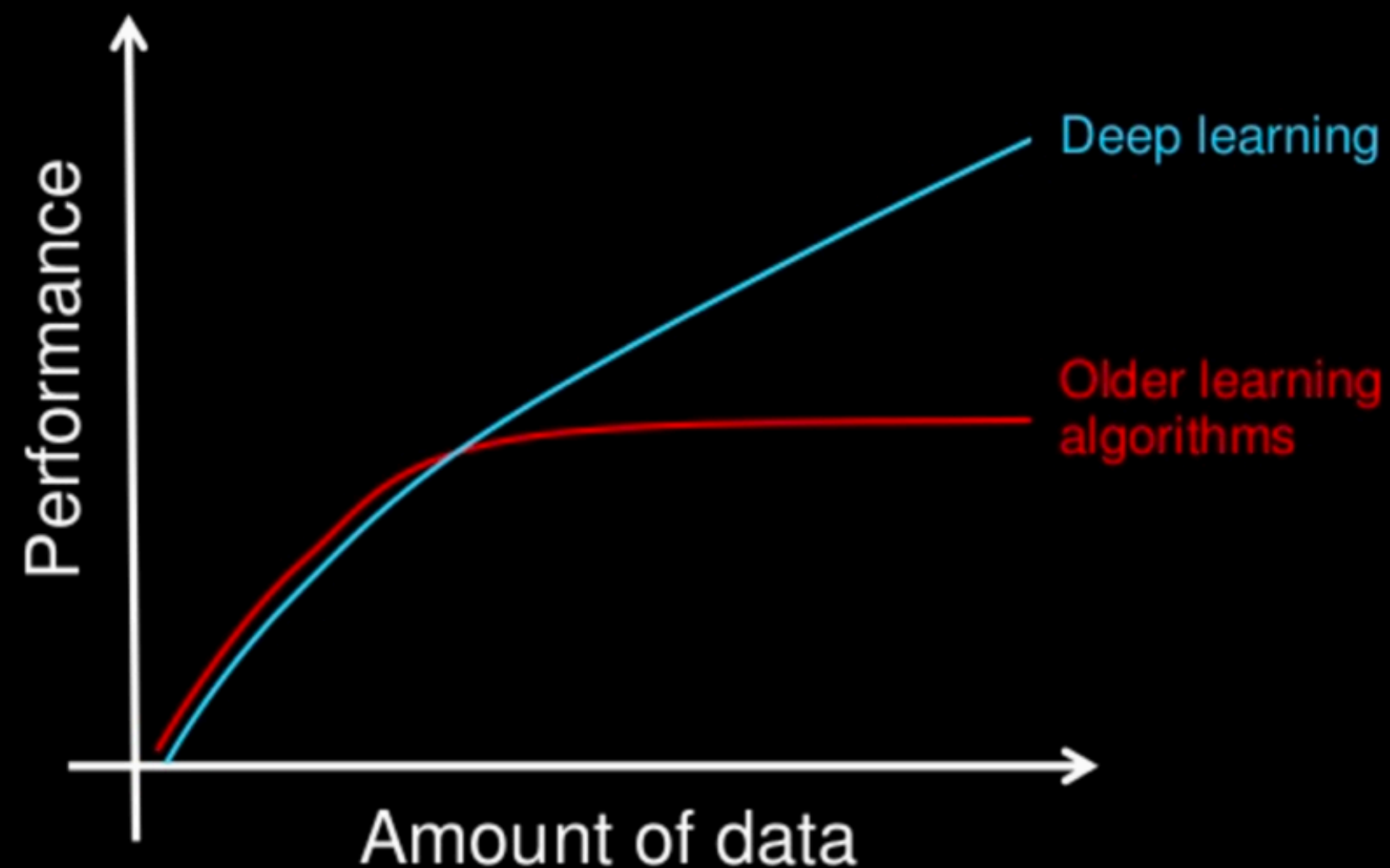
<https://www.guru99.com/backpropogation-neural-network.html>

WHY “DEEP” LEARNING?

Deep learning is **data-efficient**.

Deep learning is **modular**, and thus it can take advantage of **GPUs** that can efficiently process billions of simple math operations quickly.

Why deep learning



How do data science techniques scale with amount of data?

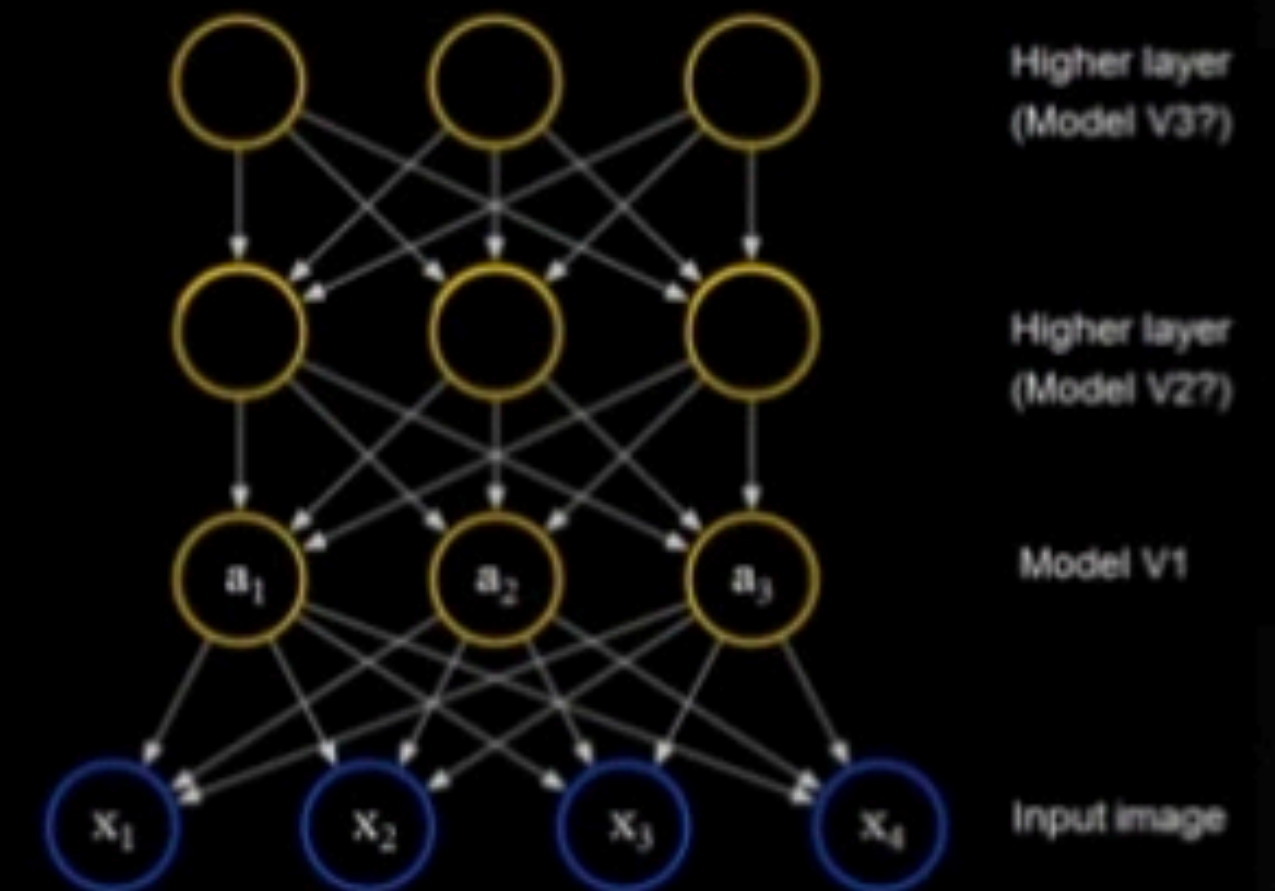
Andrew Ng's Slides

WHY ARE GPUS GOOD FOR DEEP LEARNING?

	Neural Networks	GPUs
Inherently Parallel	✓	✓
Matrix Operations	✓	✓
FLOPS	✓	✓
Bandwidth	✓	✓

GPUs deliver --

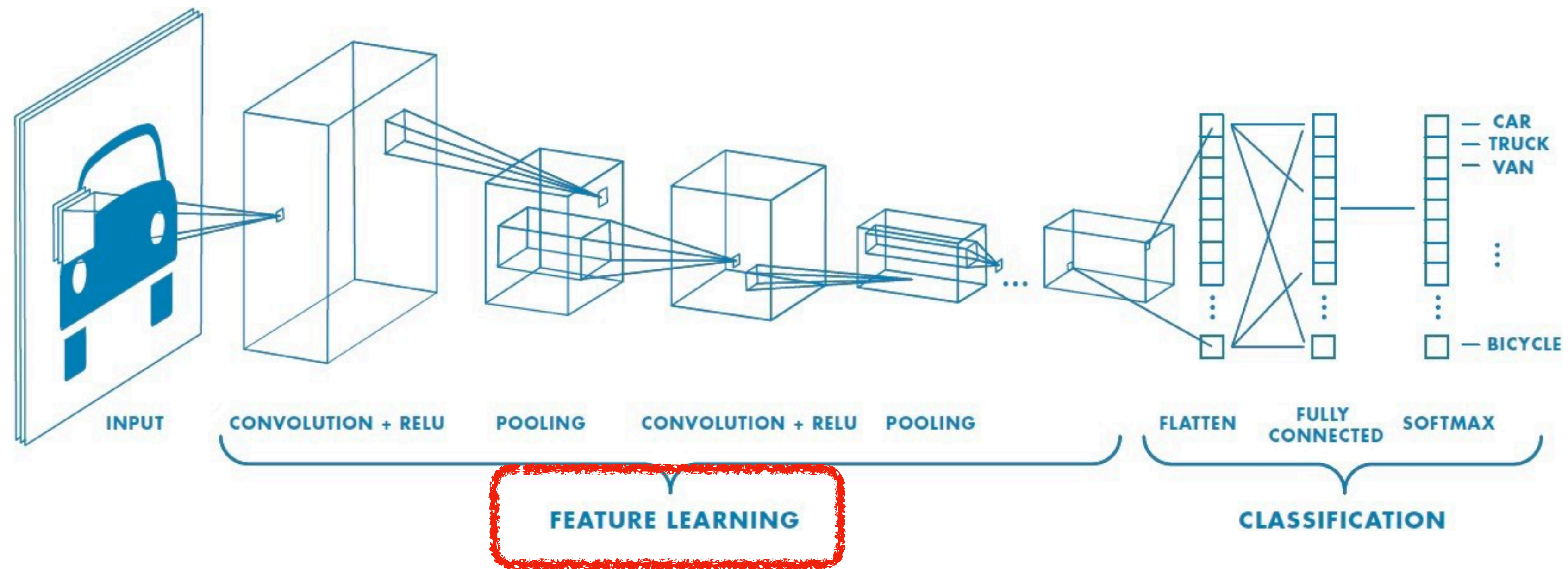
- same or **better** prediction accuracy
- **faster** results
- **smaller** footprint
- **lower** power
- **lower** cost



[Lee, Ranganath & Ng, 2007]

NVIDIA's Slides

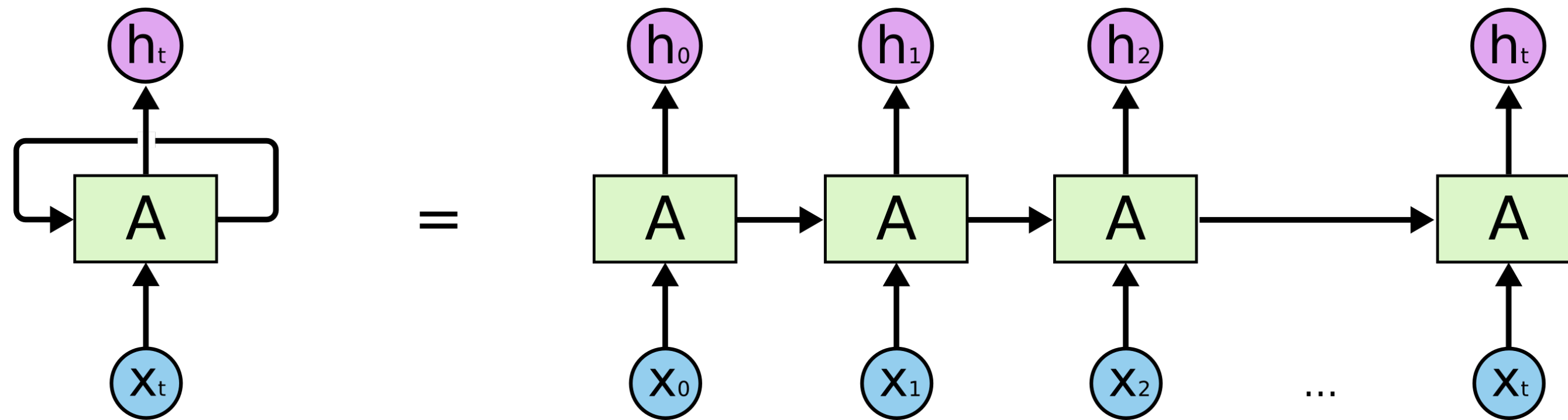
CONVOLUTIONAL NEURAL NETWORKS (CNNs)



Deep learning models learn representations!

<https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>

RECURRENT NEURAL NETWORKS (RNNs)



<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

For $\bigoplus_{n=1, \dots, m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset \mathcal{H} in \mathcal{H} and any sets \mathcal{F} on X , U is a closed immersion of S , then $U \rightarrow T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \text{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \rightarrow V$. Consider the maps M along the set of points Sch_{fppf} and $U \rightarrow U$ is the fibre category of S in U in Section, ?? and the fact that any U affine, see Morphisms, Lemma ???. Hence we obtain a scheme S and any open subset $W \subset U$ in $\text{Sh}(G)$ such that $\text{Spec}(R') \rightarrow S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S . We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \rightarrow \mathcal{O}_{X',x'}$ is separated. By Algebra, Lemma ?? we can define a map of complexes $\text{GL}_{S'}(x'/S'')$ and we win. \square

To prove study we see that $\mathcal{F}|_U$ is a covering of \mathcal{X}' , and \mathcal{T}_i is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and \mathcal{F}_p exists and let \mathcal{F}_i be a presheaf of \mathcal{O}_X -modules on \mathcal{C} as a \mathcal{F} -module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\text{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1} \mathcal{F}$$

is a unique morphism of algebraic stacks. Note that

$$\text{Arrows} = (\text{Sch}/S)_{fppf}^{opp}, (\text{Sch}/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \mapsto (U, \text{Spec}(A))$$

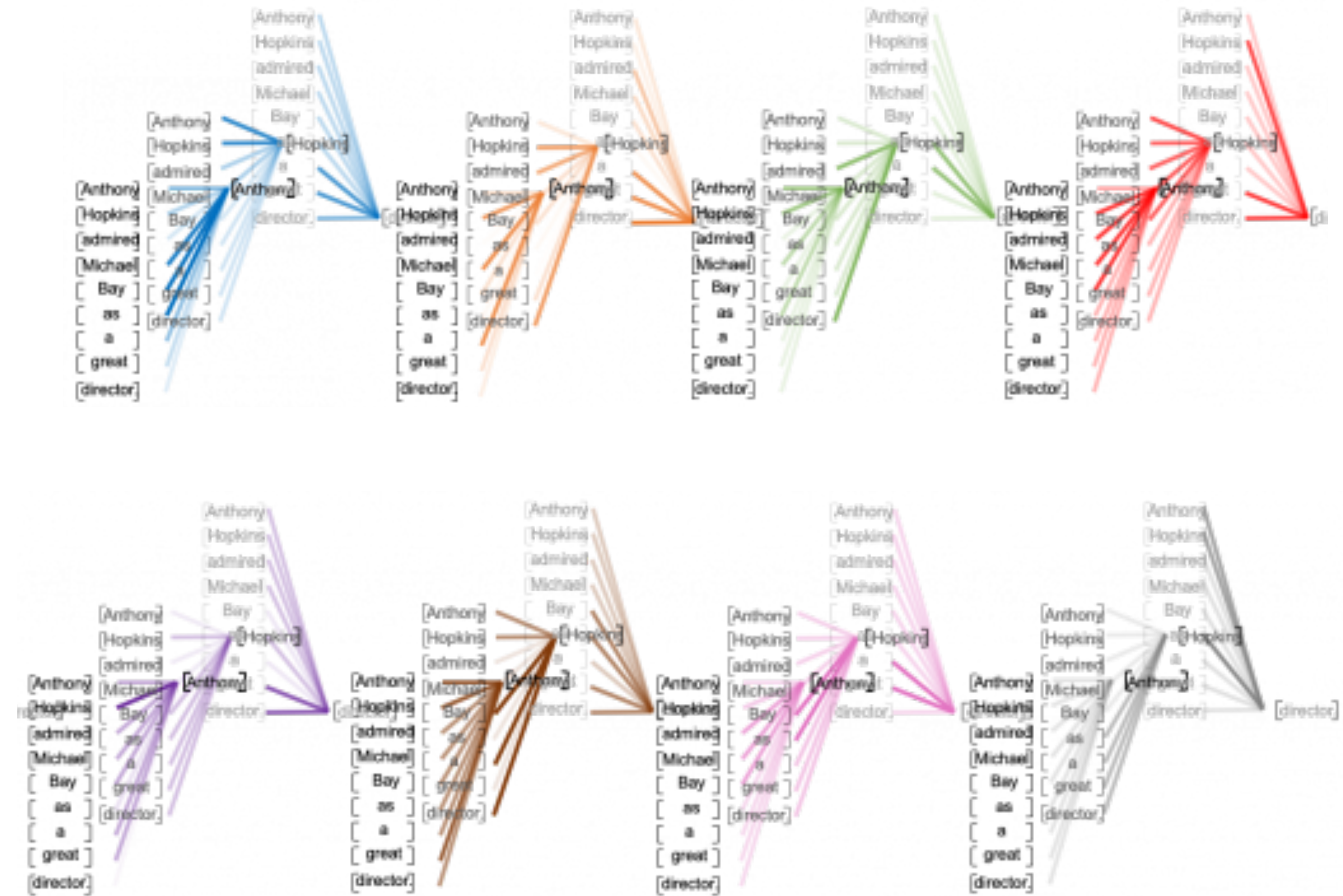
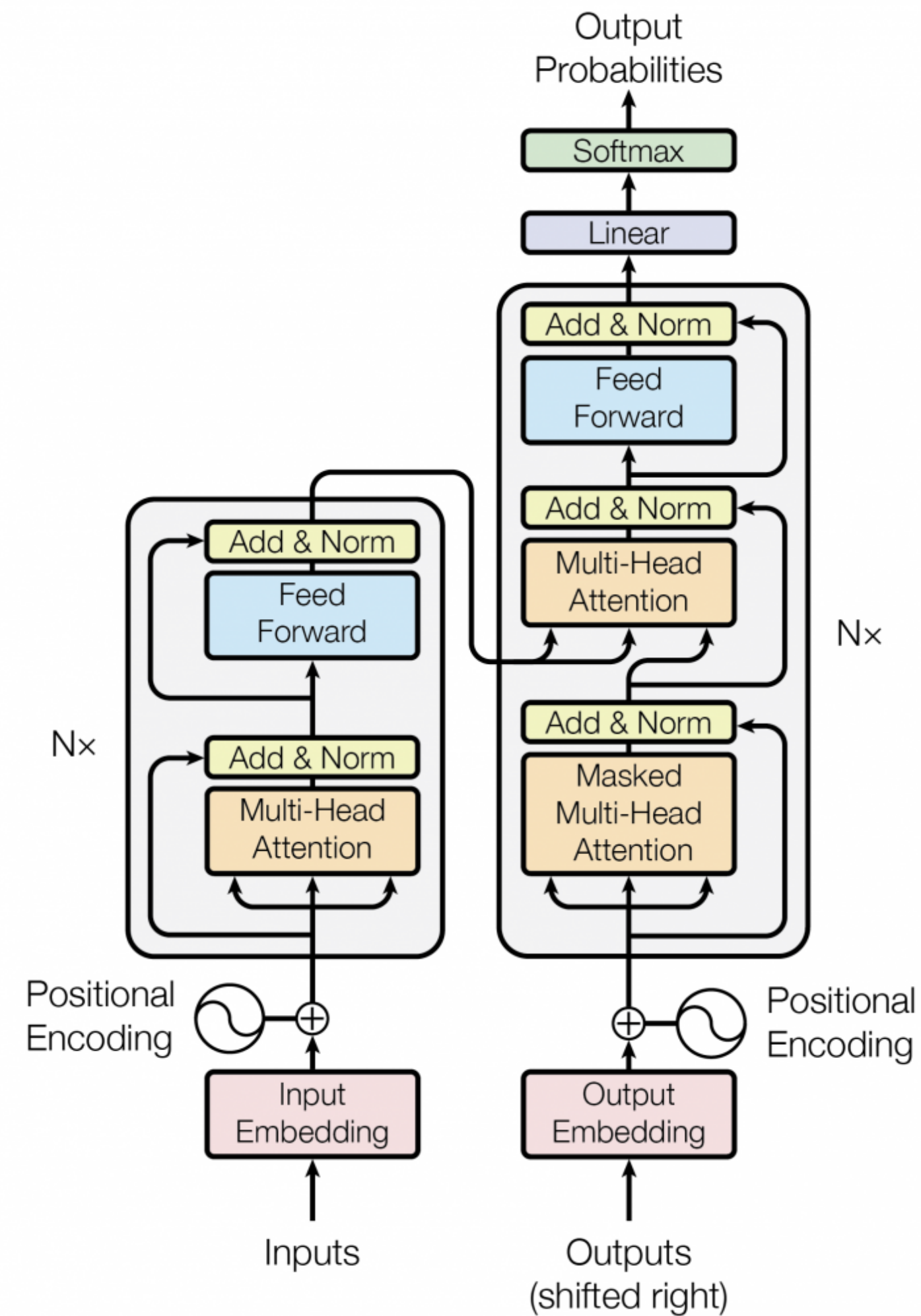
is an open subset of X . Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S .

Proof. See discussion of sheaves of sets. \square

The result for prove any open covering follows from the less of Example ??. It may replace S by $X_{spaces, \acute{e}tale}$ which gives an open subspace of X and T equal to S_{Zar} , see Descent, Lemma ??. Namely, by Lemma ?? we see that R is geometrically regular over S .

<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

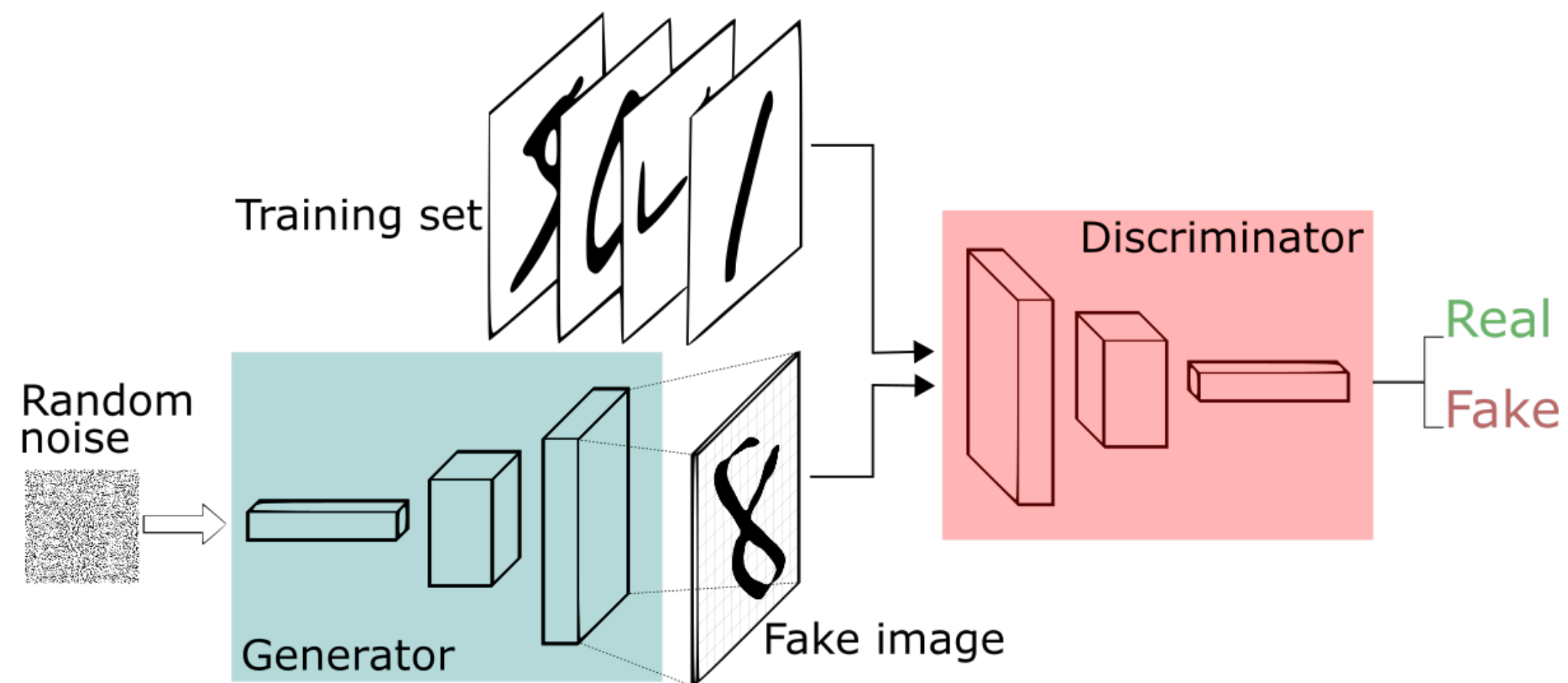
TRANSFORMERS (SELF-ATTENTION NNs)



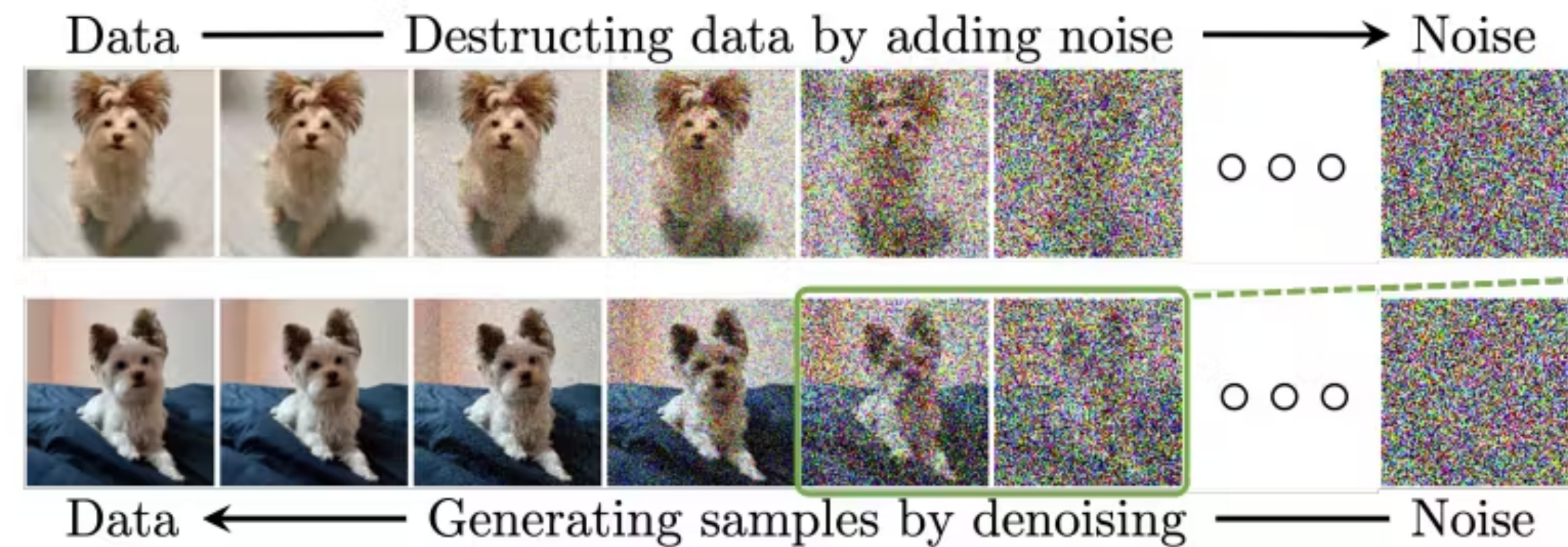
<https://data-science-blog.com/blog/2021/04/07/multi-head-attention-mechanism/>

DEEP GENERATIVE MODELS

Generative Adversarial Networks (GANs)



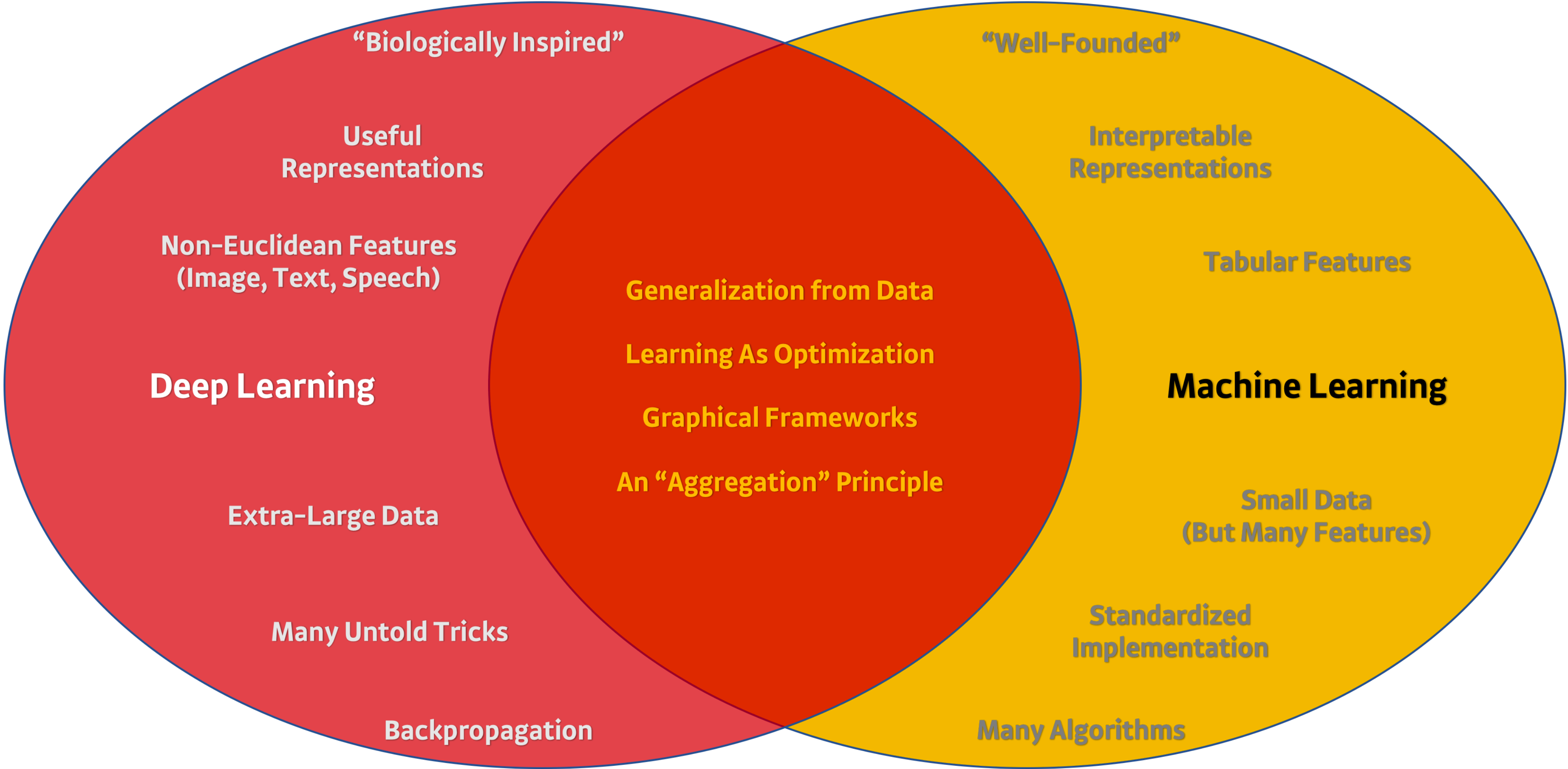
Diffusion Models



<https://medium.freecodecamp.org/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394>

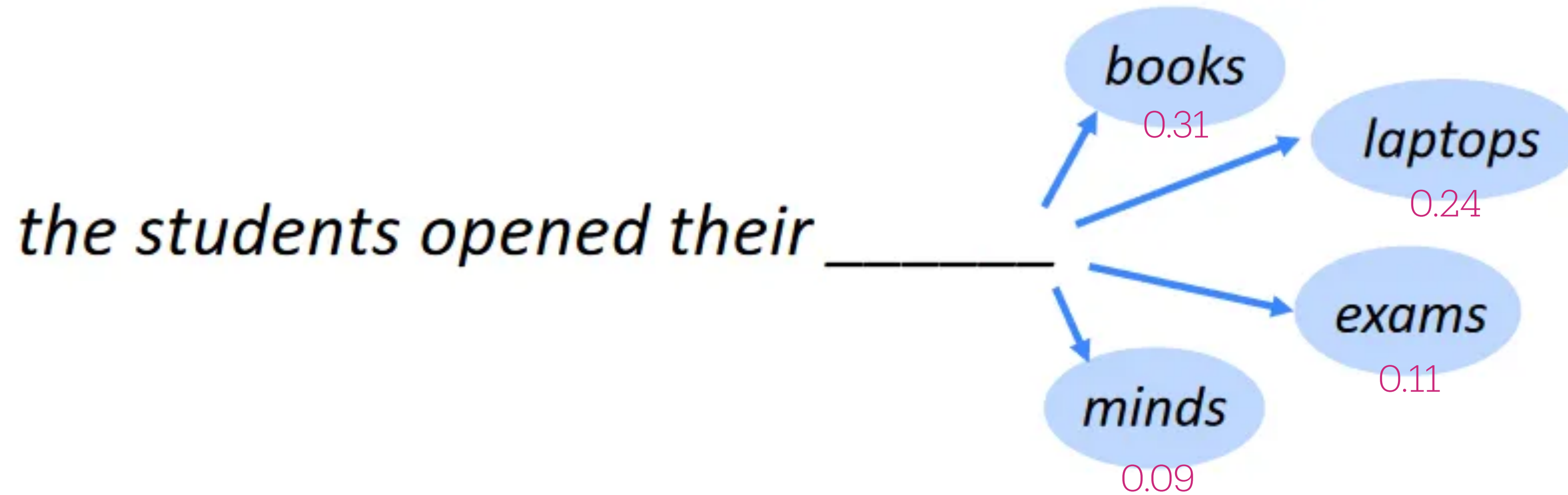
Yang et al. (2022). "Diffusion Models: A Comprehensive Survey of Methods and Applications"

DEEP LEARNING VS. CLASSICAL MACHINE LEARNING



LARGE LANGUAGE MODELS (LLMs)

LANGUAGE MODELING*



Task: Estimate the probability of the next word given each context

LM ISN'T *JUST* A SUPERVISED TASK

- On the surface, language modeling appears to be a **supervised** task.
(Input: context, Output: next word)
- But we don't need to manually provide labels (a **self-supervised** task).
- A sentence consisting of N words turns into (essentially) N data points.

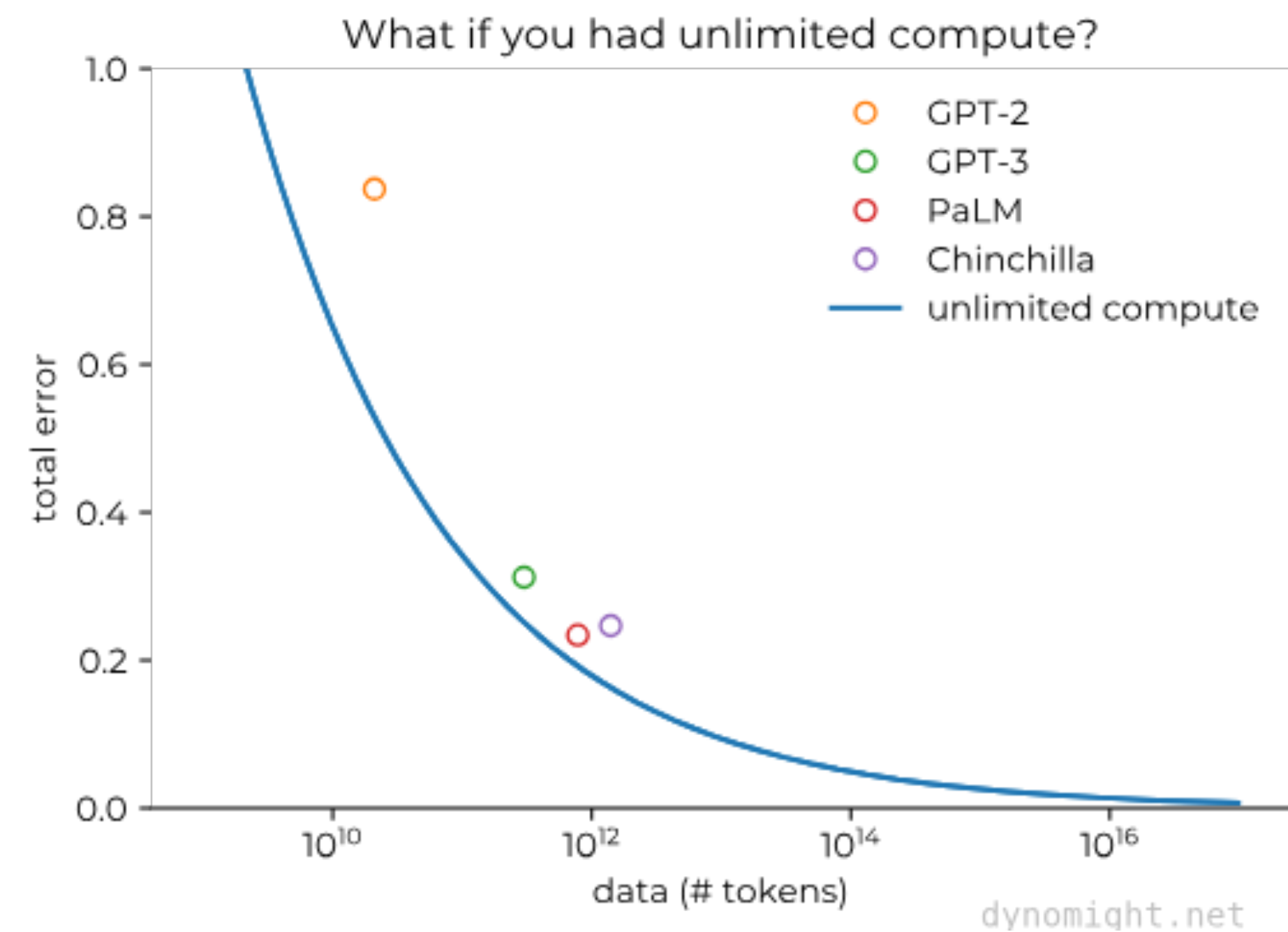
The	students				
The	students	opened			
The	students	opened	their		
The	students	opened	their	books	
The	students	opened	their	books	.

*This also depicts the **generative** process (after training)!*

This one sentence contains 5 examples for the “next word prediction” task.

“LARGE” LANGUAGE MODELING: A RECIPE

1. Get a LOT of text-form **data** (trillions of words/symbols).
2. Set up a large *Transformer* model (billions of **parameters**).
3. Train it on the language modeling task with many, many **compute-hours** on GPU.



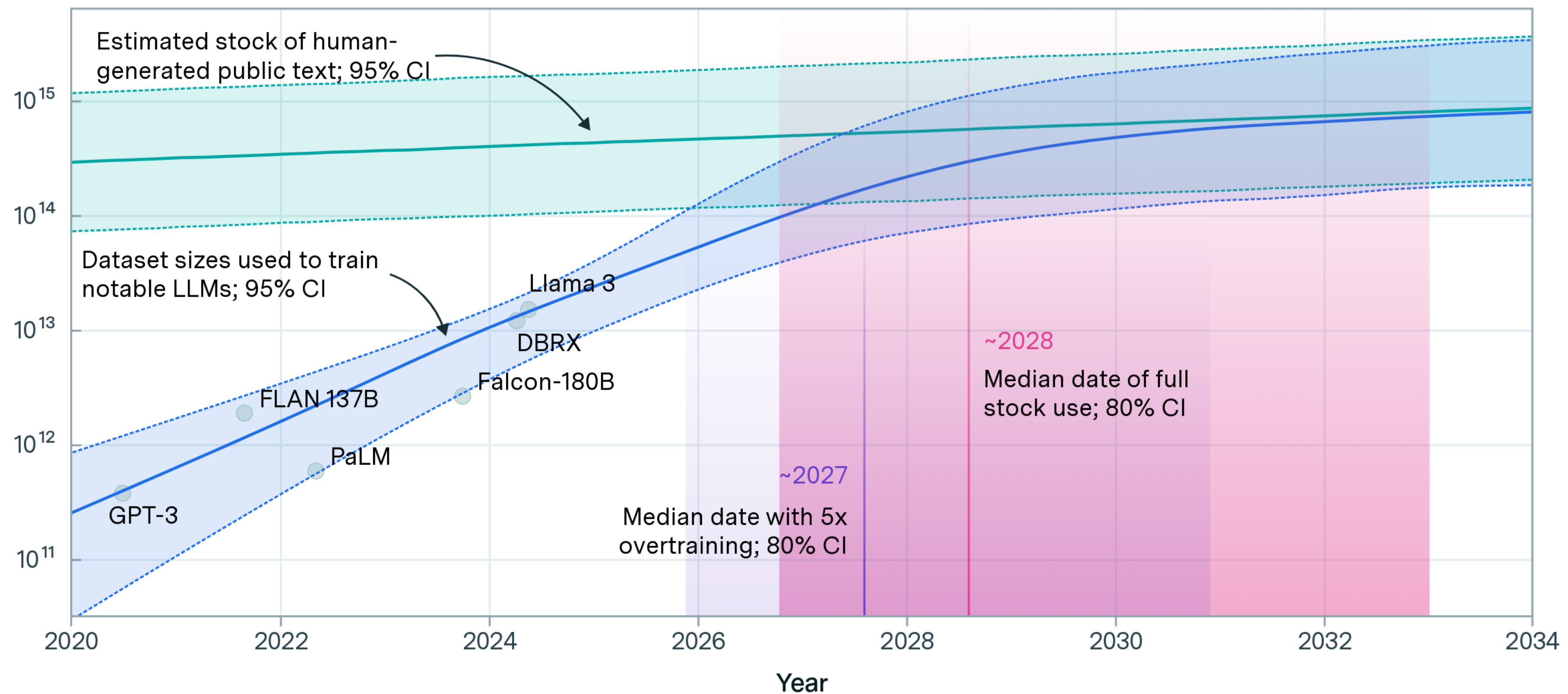
<https://dynomight.net/scaling/>

WHERE WE MIGHT BE AT IN TERMS OF DATA

Projections of the stock of public text and data usage

EPOCH AI

Effective stock (number of tokens)



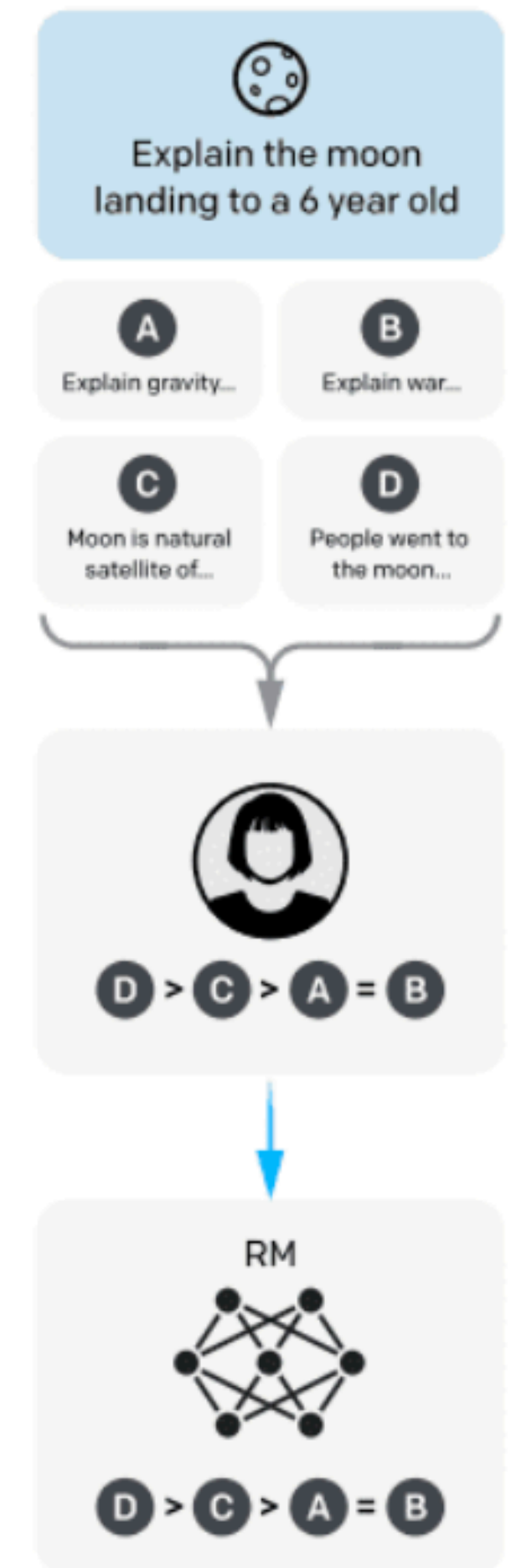
ARE LLMs “INTELLIGENT”?

- Recall, a language model is a **generative** model that can predict entire sequences of words, one-word-at-a-time, via next word prediction.
- *Maybe*, a good chunk of human intelligence **can** be posed as (a sequence of) next word prediction problems:
 - *“The first president of the United States was [____].”*
 - *“If I compose two hydrogen atoms with one oxygen atom, I get [____].”*
- ***So, is language modeling merely a (self-)supervised learning task, or is it a task of (crudely) modeling human intelligence?***

ALIGNMENT TO HUMAN PREFERENCES (RLHF)

- It turns out that, when you train a model on loads of random text from the Internet, it will learn to say all kinds of things, **good and bad**.
 - *It can be helpful, or it can be sarcastic.*
 - *It can be polite/nice, or it can be hurtful/harmful.*
 - *It can make things up when it doesn't "know" the answer.*
- An *ad hoc* strategy: fine-tune LLMs using **human preferences**.
- **To think about:** Which human values are we aligning to?

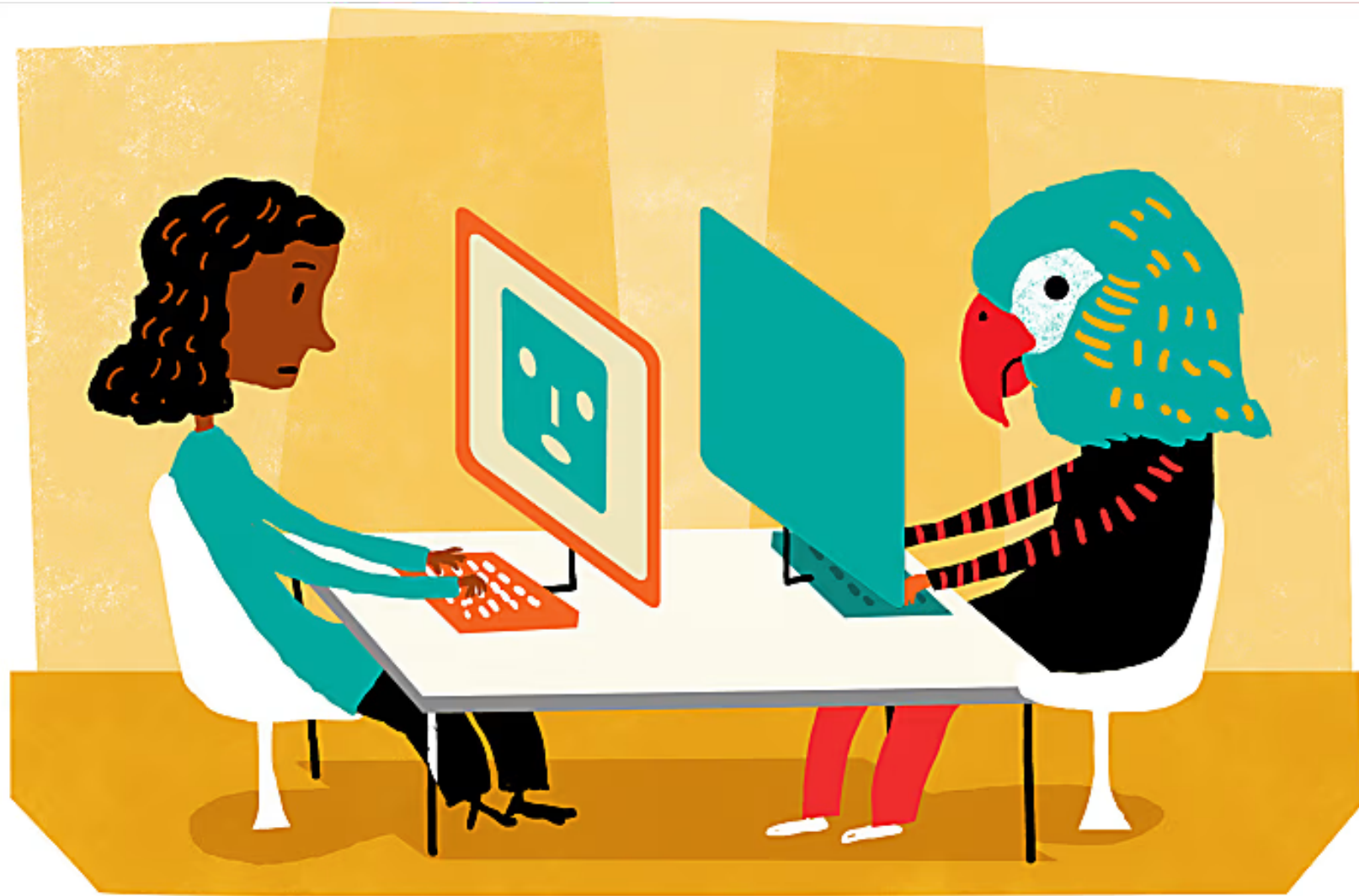
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

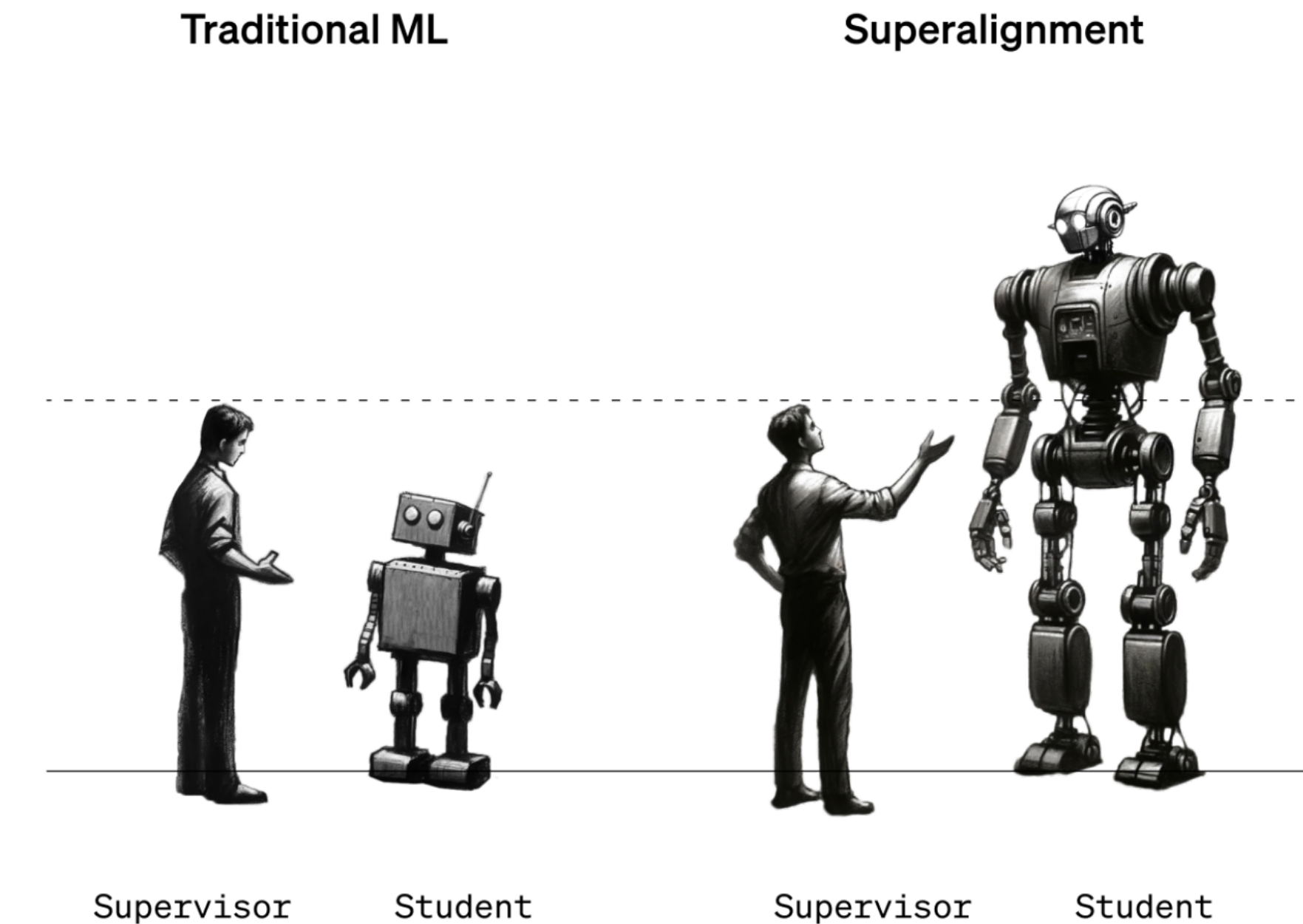
SOME (VERY) DIFFERENT VIEWS OF LLMs



“Stochastic Parrots”
(Bender et al., 2021;
image from Wall Street Journal)



“Einstein in Your Basement”
(Henrik Kniberg, 2024;
recreated image from DALL-E)



“Superintelligence”
(OpenAI, 2023)

HALLUCINATION

What are some papers that discuss feasibility of nuclear fusion?



Research on the feasibility of nuclear fusion has been a topic of significant interest, and there are several key papers and reviews that discuss the challenges, progress, and potential of this technology. Here are some influential papers and sources:

1. "The Physics of Inertial Fusion: BeamPlasma Interaction, Hydrodynamics, Hot Dense Matter" by Stefano Atzeni and Jurgen Meyer-ter-Vehn (2004)



This book provides a comprehensive overview of inertial confinement fusion (ICF), discussing the physics involved and the challenges faced in achieving practical fusion energy. It's highly regarded in the field and covers the feasibility of fusion from a technical standpoint.

2. "Overview of the ITER Physics Basis" (Nuclear Fusion, 1999)



This paper provides an extensive review of the ITER project, which is one of the most significant global efforts toward achieving controlled nuclear fusion. It discusses the scientific and engineering challenges involved in making fusion a feasible energy source.

3. "Nuclear Fusion as a Sustainable Energy Source" (Nature Physics, 2014)



This article reviews the current state of nuclear fusion research, particularly focusing on magnetic confinement fusion and the development of tokamak reactors. It discusses the progress made and the remaining challenges that need to be addressed.

- Generally speaking, neural networks are very good **pattern matchers**.
 - *An undertrained LLM may know how to put together what looks like a citation.*
 - *But then, it may not know what name to put as the next token, so it just predicts the token with the highest probability.*
- **Retrieval-Augmented Generation (RAG)** is the go-to remedy, although it's clearly not perfect.

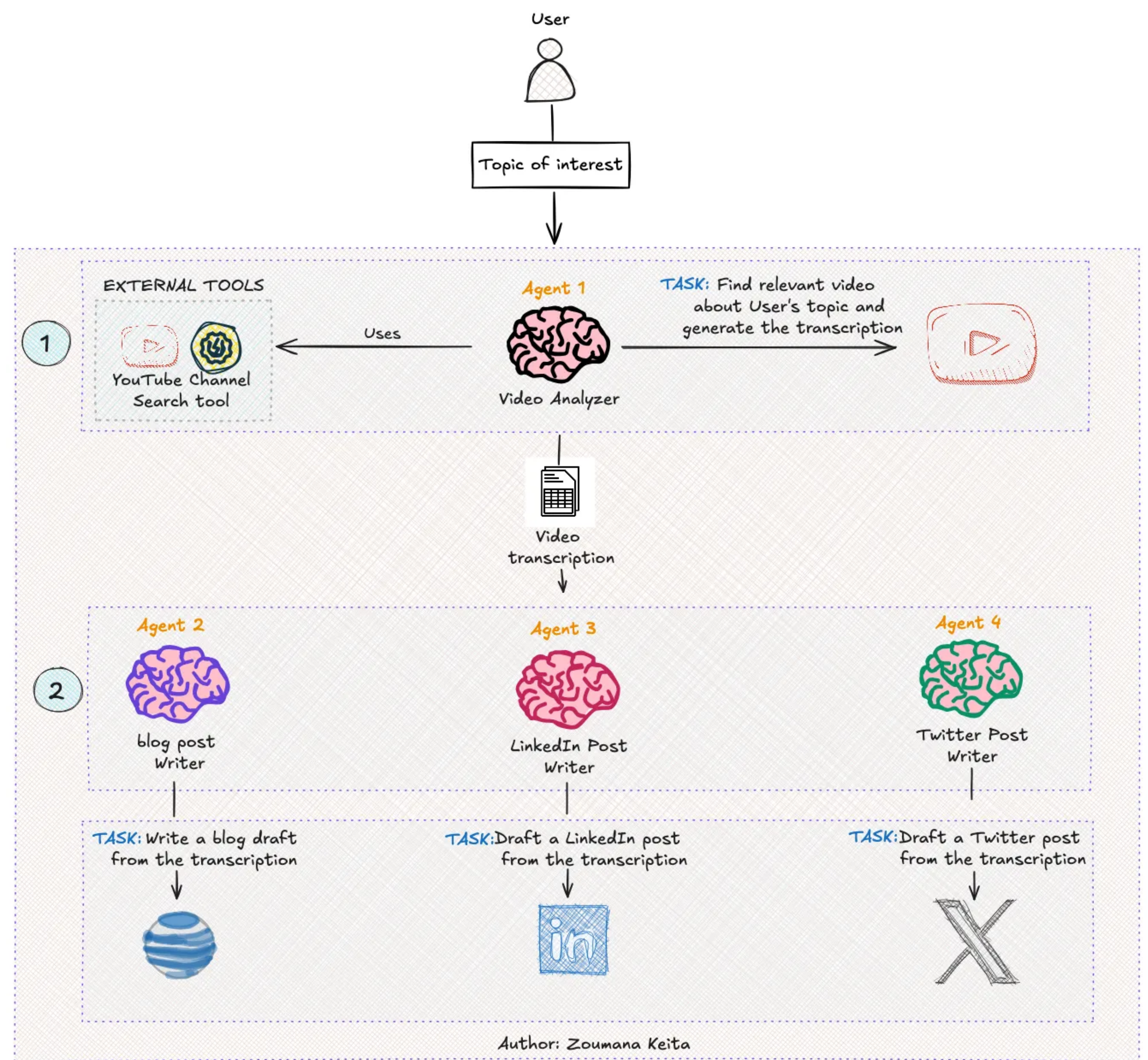
FROM PATTERN MATCHERS TO REASONERS?

- LLMs generally require lots of examples to learn.
- Yet, humans can often learn from only few examples; this is arguably because we can **reason**.
- **Chain-of-Thought (CoT) prompting** is an *ad hoc* technique that leads the model toward a correct reasoning path.

Standard Prompting	Chain-of-Thought Prompting
<p>Model Input</p> <p>Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?</p> <p>A: The answer is 11.</p> <p>Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?</p>	<p>Model Input</p> <p>Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?</p> <p>A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.</p> <p>Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?</p>
<p>Model Output</p> <p>A: The answer is 27. ❌</p>	<p>Model Output</p> <p>A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅</p>

Wei et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models."

“AGENTIC” AI



<https://towardsdatascience.com/ai-agents-from-concepts-to-practical-implementation-in-python-fb26789b1560>

AUTONOMOUS AI AGENTS INTERACTION FROM ANALYZING USER'S REQUEST TO CREATING SOCIAL MEDIA CONTENT

WHAT'S NEXT?

- Many, many AI applications: autonomous driving, protein/drug modeling, medical AI, personalized assistants/companions, ...
- Multi-modal learning: videos & sensory data; “World Models”
- Explainability & interpretability
- Long-term decision making
- “AGI” & Superalignment
- Unlimited potential and/or risk?

TAKEAWAYS

- **Machine Learning** finds generalizable patterns from data with trial & error.
- **Deep Learning** finds useful representations from complex data sources.
- **Large Language Models** (& other “GenAI” models) leverage massive DL models & computation, allowing them to learn from massive amounts of text/image/video/etc.

TO THINK ABOUT:

- **(General) Intelligence**
- **Understanding** (what does it mean to “know” something?)
- **Alignment** (whose values are we aligning to?)
- **Agency/Responsibility/Morality/Honesty/...**
- ...

THANK YOU

Any questions?