A Peek into the World of Machine Learning

Preethi Jyothi, IIT Bombay
ACM ROCS Workshop,
Feb 24, 2024

Image produced by ByteDance's SDXL-Lightning text-to-image generative model at https://huggingface.co/ByteDance/SDXL-Lightning

What is Machine Learning?

Ability of computers to “learn” from “data” or “past experience”

What is Machine Learning?

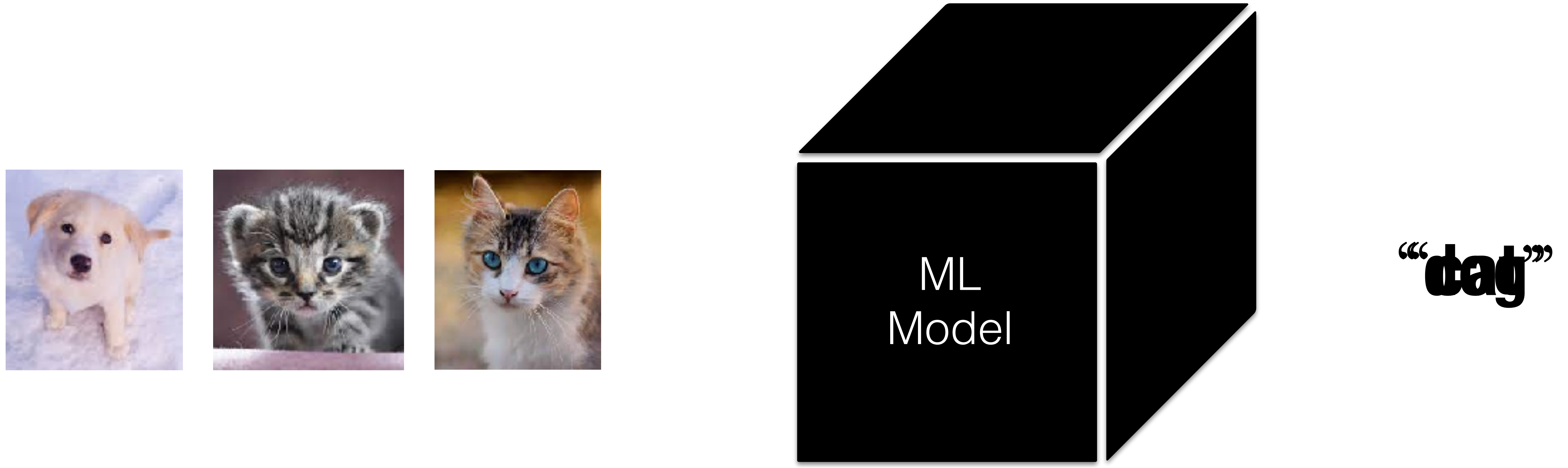
Ability of computers to “learn” from “data” or “past experience”

- data/past experience: Comes from various sources such as sensors, domain knowledge, experimental runs, etc.

What is Machine Learning?

Ability of computers to “learn” from “data” or “past experience”

- data/past experience: Comes from various sources such as sensors, domain knowledge, experimental runs, etc.
- learn: Make predictions or decisions based on data by optimizing a model



What is Machine Learning?

Meme from the Internet

Supervised Learning

- Given a labeled set of input-output pairs, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ objective is to learn a function mapping the inputs x to outputs y
- Inputs can be complex objects such as images, sentences, speech signals, etc. Featurized before being used as inputs.
- Outputs are either categorical (*classification* tasks) or real-valued (*regression* tasks).

Examples:

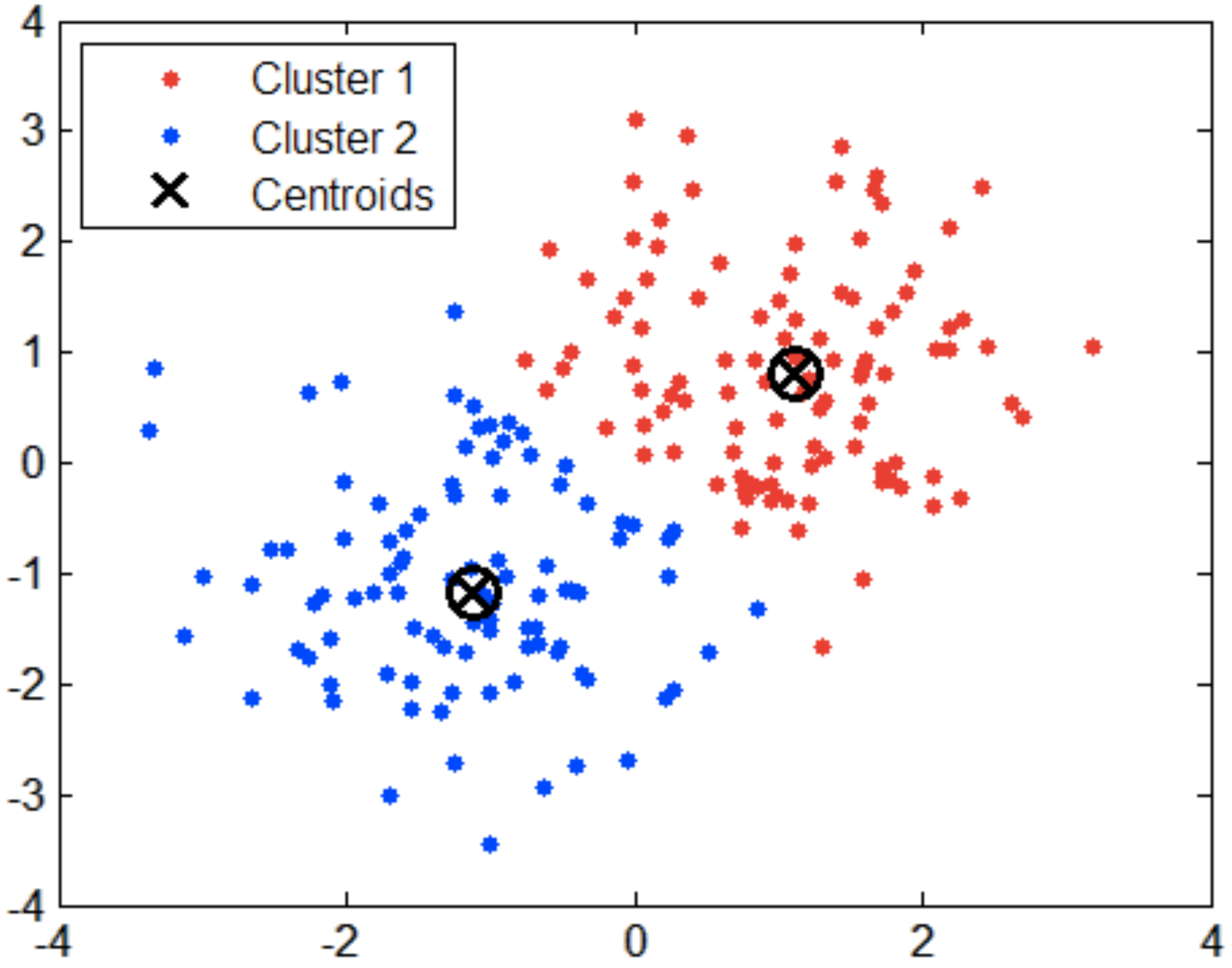
Spam classification: Inputs are emails, $x \in \mathbb{N}^d$, and output $y \in \{0, 1\}$

Predict taxi prices: Inputs are features of the ride, $x \in \mathbb{R}^d$, and output $y \in \mathbb{R}$

Unsupervised Learning

- Given a set of inputs, $\mathcal{D} = \{x_i\}_{i=1}^N$, discover some patterns/groups in the data

- Most common example: Clustering



Relationship between AI, ML, DL

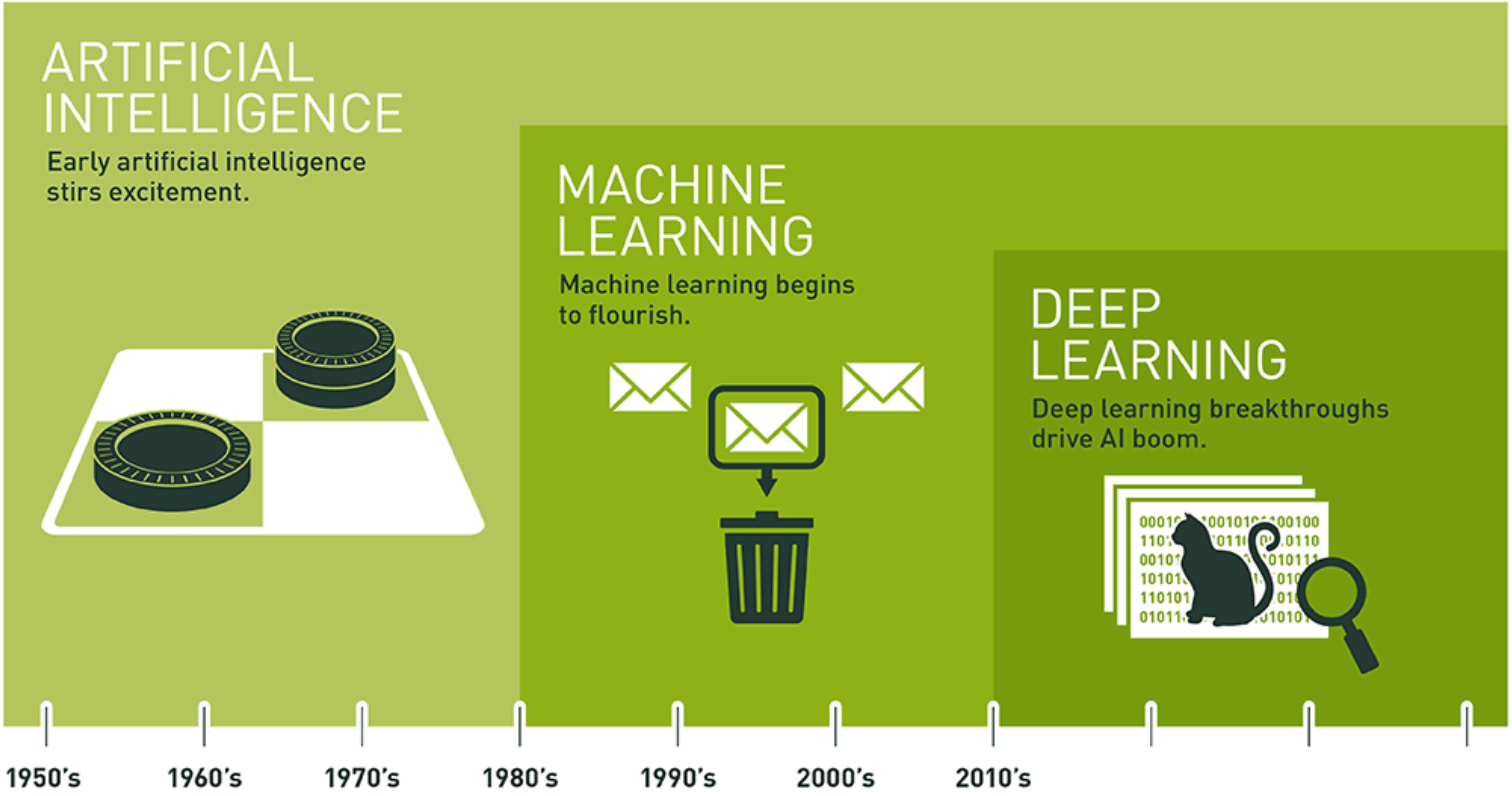
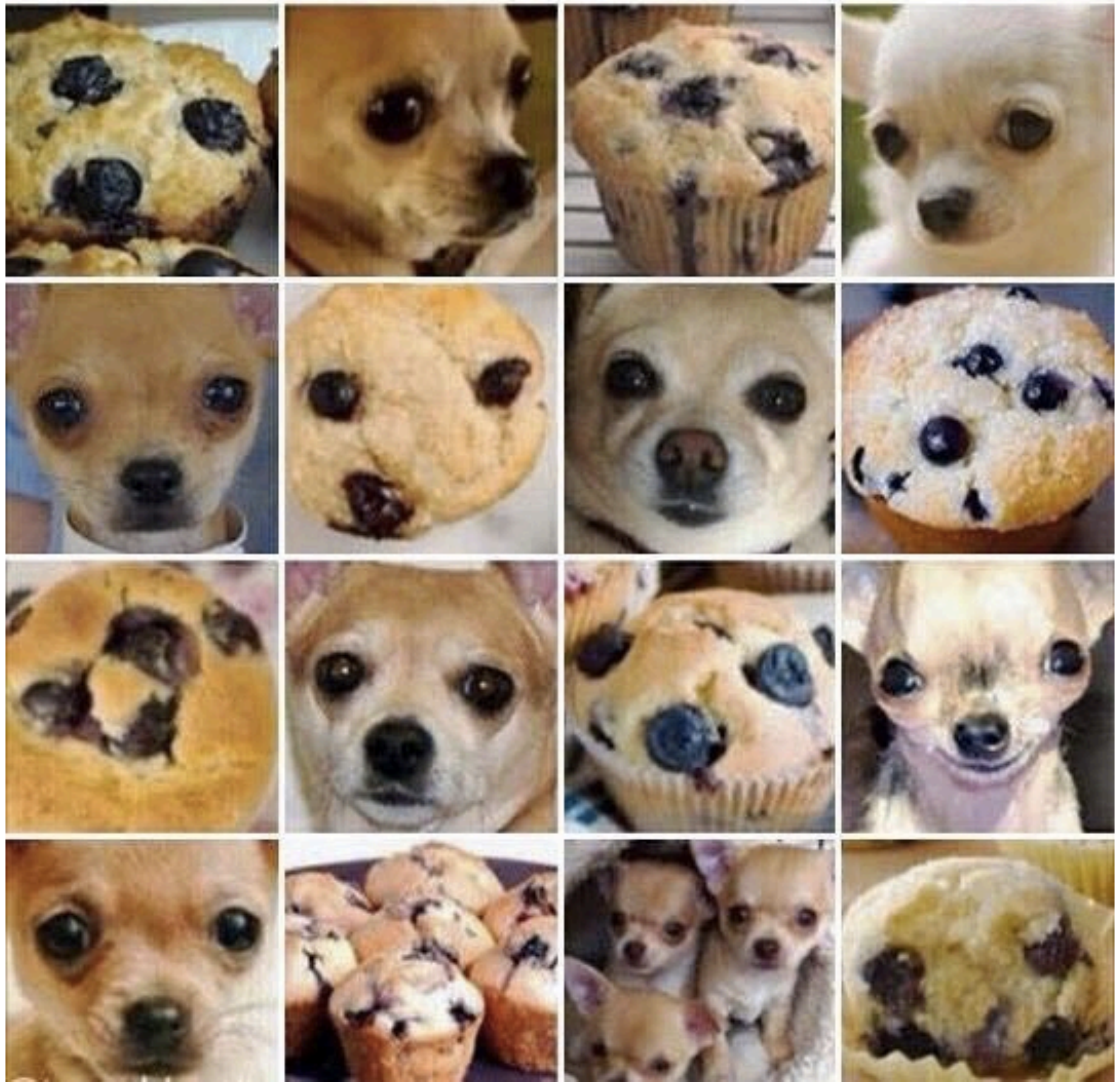

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Image from: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

When do we need ML?

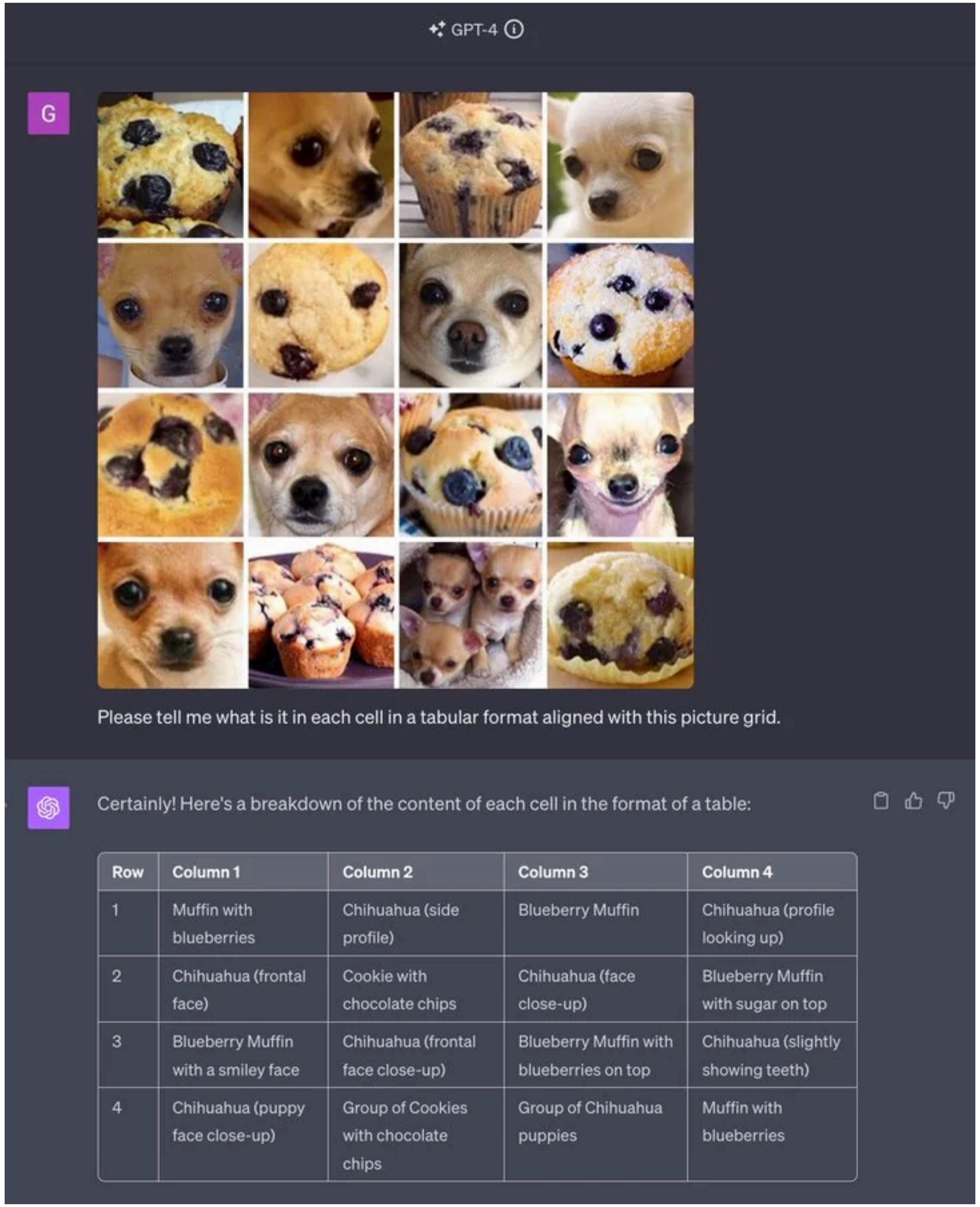
- For tasks that are easily performed by humans but are complex for computer systems to emulate



Dog or Muffin?

Image from Reddit

Solved by GPT-4



Please tell me what is it in each cell in a tabular format aligned with this picture grid.

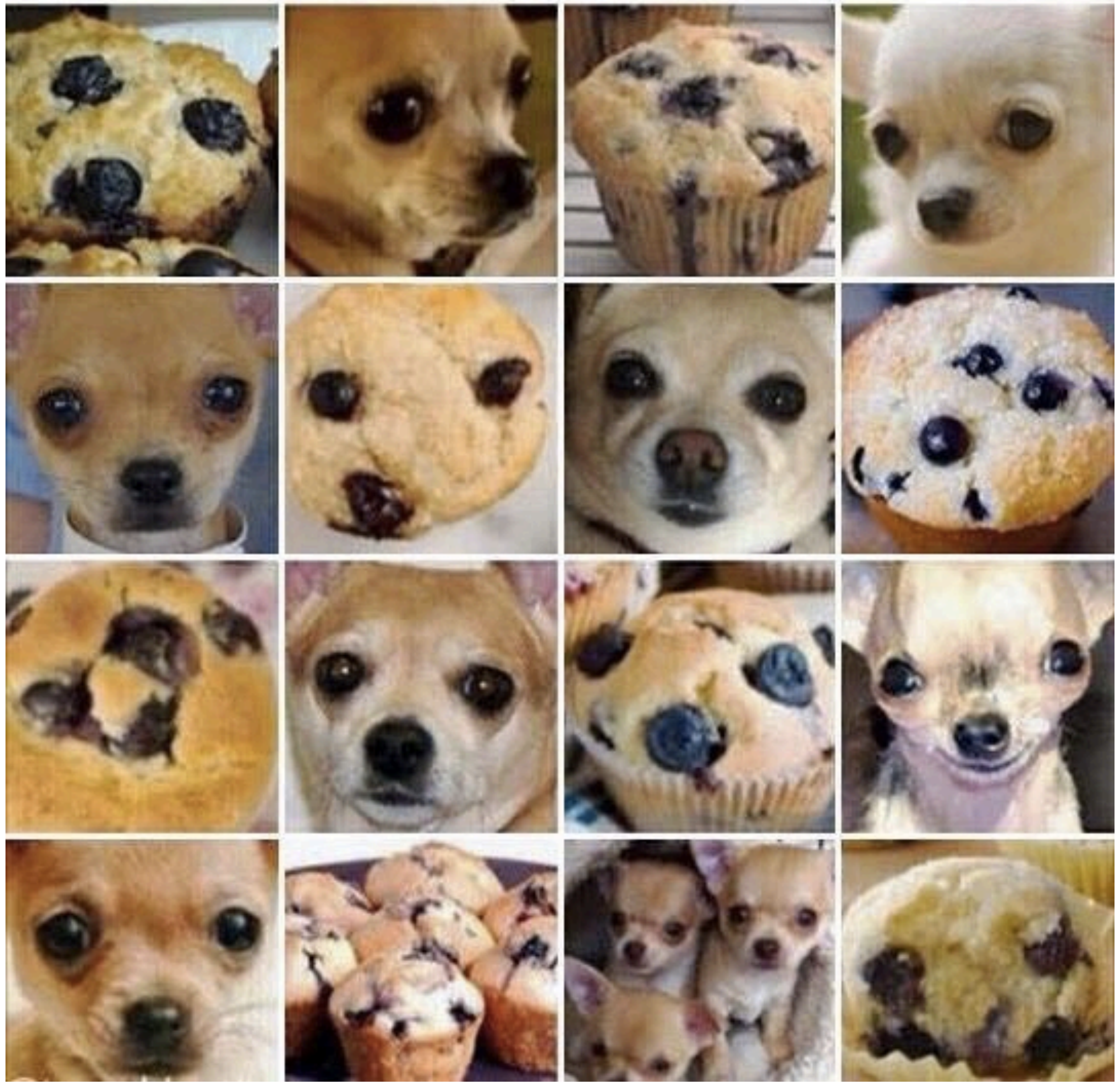
Certainly! Here's a breakdown of the content of each cell in the format of a table:

Row	Column 1	Column 2	Column 3	Column 4
1	Muffin with blueberries	Chihuahua (side profile)	Blueberry Muffin	Chihuahua (profile looking up)
2	Chihuahua (frontal face)	Cookie with chocolate chips	Chihuahua (face close-up)	Blueberry Muffin with sugar on top
3	Blueberry Muffin with a smiley face	Chihuahua (frontal face close-up)	Blueberry Muffin with blueberries on top	Chihuahua (slightly showing teeth)
4	Chihuahua (puppy face close-up)	Group of Cookies with chocolate chips	Group of Chihuahua puppies	Muffin with blueberries

Image from Reddit

When do we need ML?

- For tasks that are easily performed by humans but are complex for computer systems to emulate



Dog or Muffin?

Image from Reddit

When do we need ML?

- For tasks that are easily performed by humans but are complex for computer systems to emulate
  - **Vision:** Identify faces in a photograph, objects in a video or still image, etc.
  - **Natural language:** Translate a sentence from Hindi to English, question answering, identify sentiment of text, etc.
  - **Speech:** Recognise spoken words, speaking sentences naturally
  - **Game playing:** Play games like chess, Go, Dota.
  - **Robotics:** Walking, jumping, displaying emotions, etc.
  - Driving a car, navigating a maze, etc.

When do we need ML?

- For tasks that are beyond human capabilities
  - Analysis of large and complex datasets
  - E.g. IBM Watson’s Jeopardy-playing machine

Image credit: https://i.ytimg.com/vi/P18EdAKuC1U/maxresdefault.jpg

Applications of Machine Learning

Some examples from real life

- Google Search
- Search within articles, e-books, etc.

Information Retrieval

Applications of Machine Learning

Some examples from real life

- Google Search
- Search within articles, e-books, etc.

Information Retrieval

- Face detection in photos
- Detecting cars on roads (e.g., for self-driving cars)

Computer Vision

- Facebook recommending friends/ads
- Netflix/Amazon/Flipkart recommendations

Recommender Systems

- Game playing (chess, Go, poker, etc.)
- Robots (playing soccer, robotic arm, etc.)

Robotics/Game Playing

Applications of Machine Learning

Some examples from real life

• Google Search • Search within articles, e-books, etc.	**Information Retrieval**
• Face detection in photos • Detecting cars on roads (e.g., for self-driving cars)	**Computer Vision**
• Facebook recommending friends/ads • Netflix/Amazon/Flipkart recommendations	**Recommender Systems**
• Game playing (chess, Go, poker, etc.) • Robots (playing soccer, robotic arm, etc.)	**Robotics/Game Playing**
• IBM’s Watson (Jeopardy, etc.) • Medical Diagnosis Systems	**ML in Expert Systems**
• Personal Assistant/LLMs (ChatGPT, Google Assistant, Alexa, etc.) • Closed Captioning (Google Meet, MS Teams, Zoom, etc.)	**Speech & Natural Language Processing**

What is ChatGPT?

ChatGPT is a large language model trained to learn instructions with human feedback

1

What is a language model?

2

How is ChatGPT trained to learn using human feedback?

What is a Language Model?

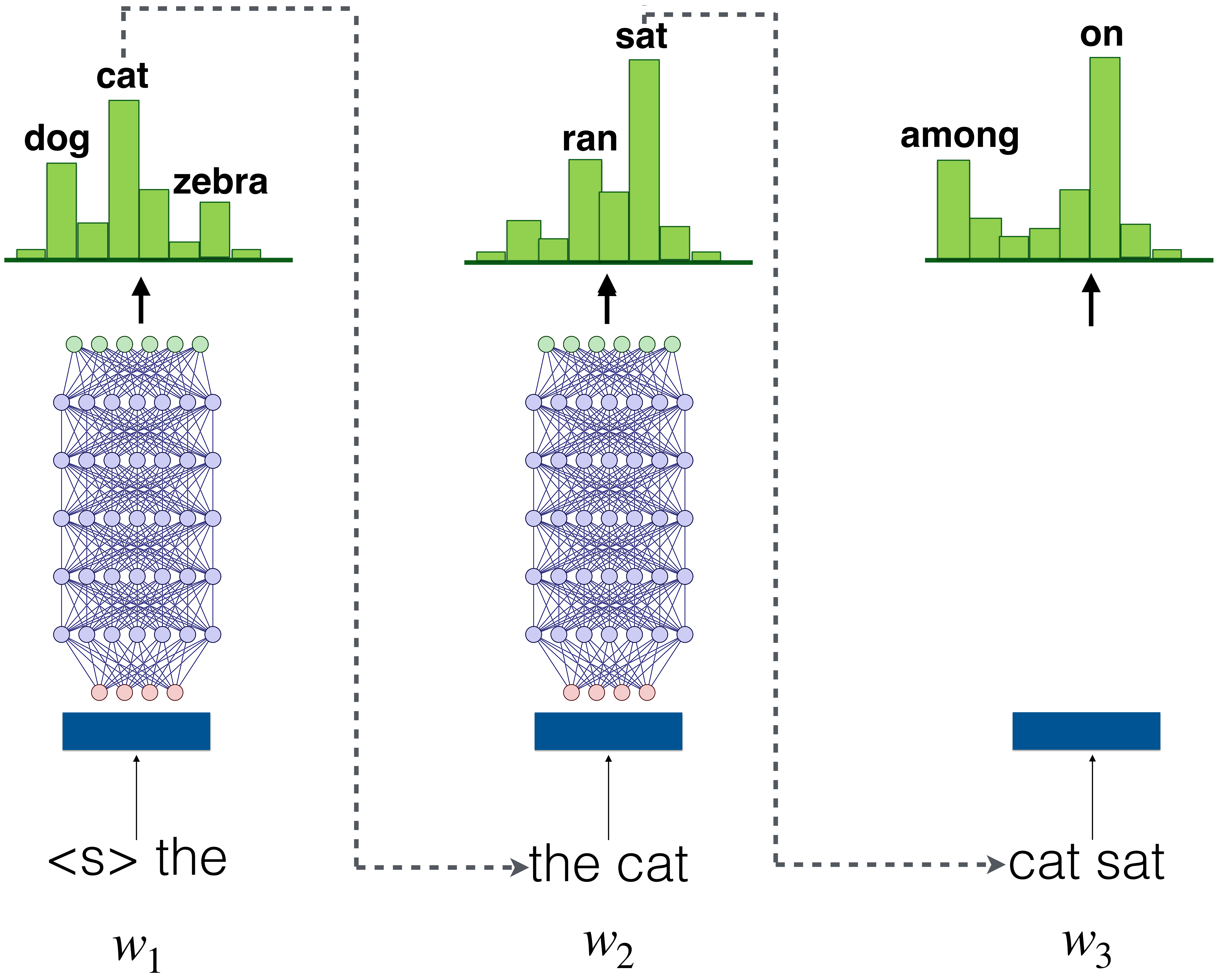
- Given a sequence of words, $w_1, \ldots, w_{t-1}$, what is the most likely next word w_t?
- Given a word sequence $W = \{w_1, \ldots, w_T\}$, what is $P(W)$?
- Language models
  - provide information about the most likely next word

    P(“she delivered a talk”) > P(“she delivered a walk”)

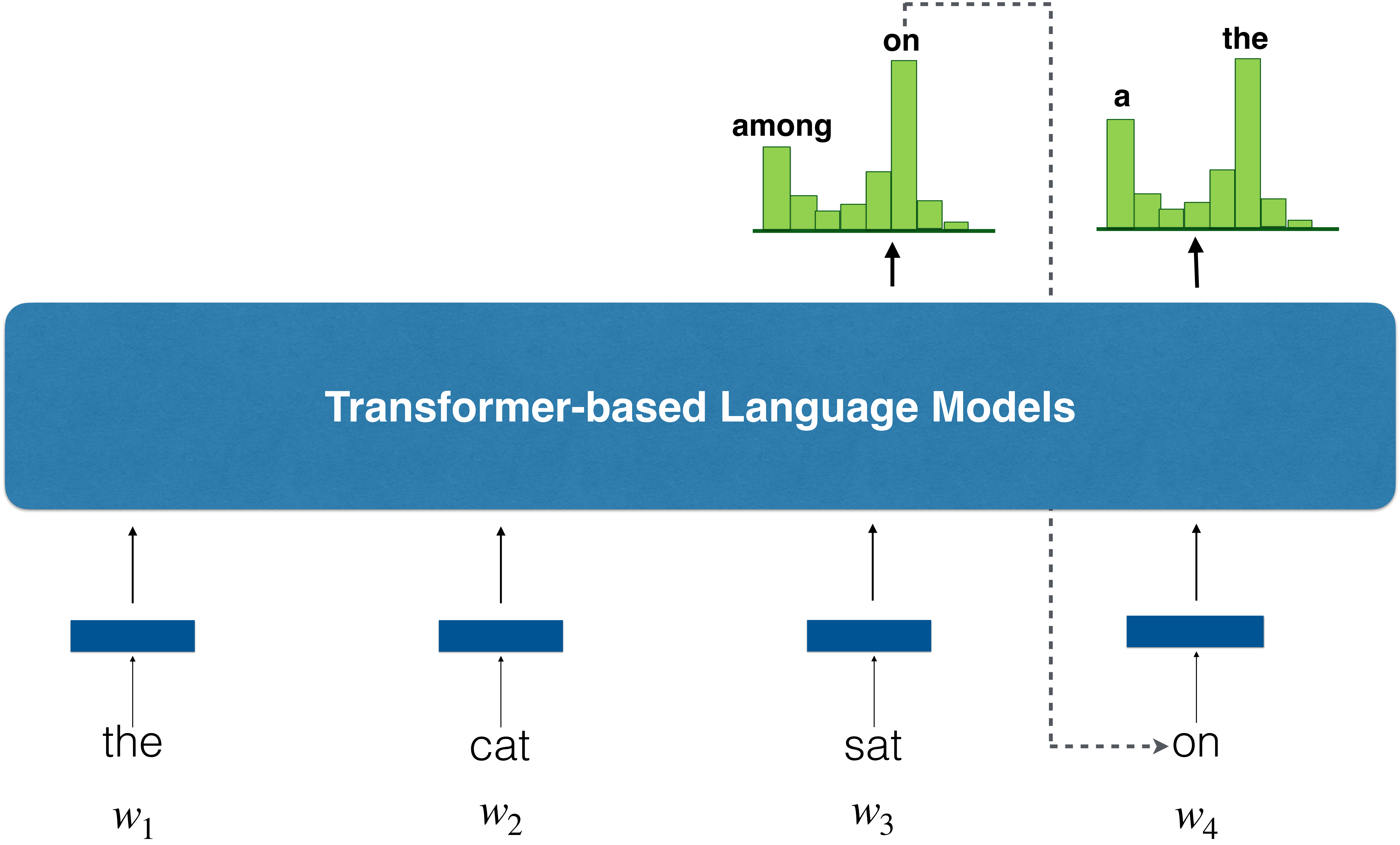
  - provide information about likely word reorderings

    P(“she delivered a talk”) > P(“delivered she talk a”)

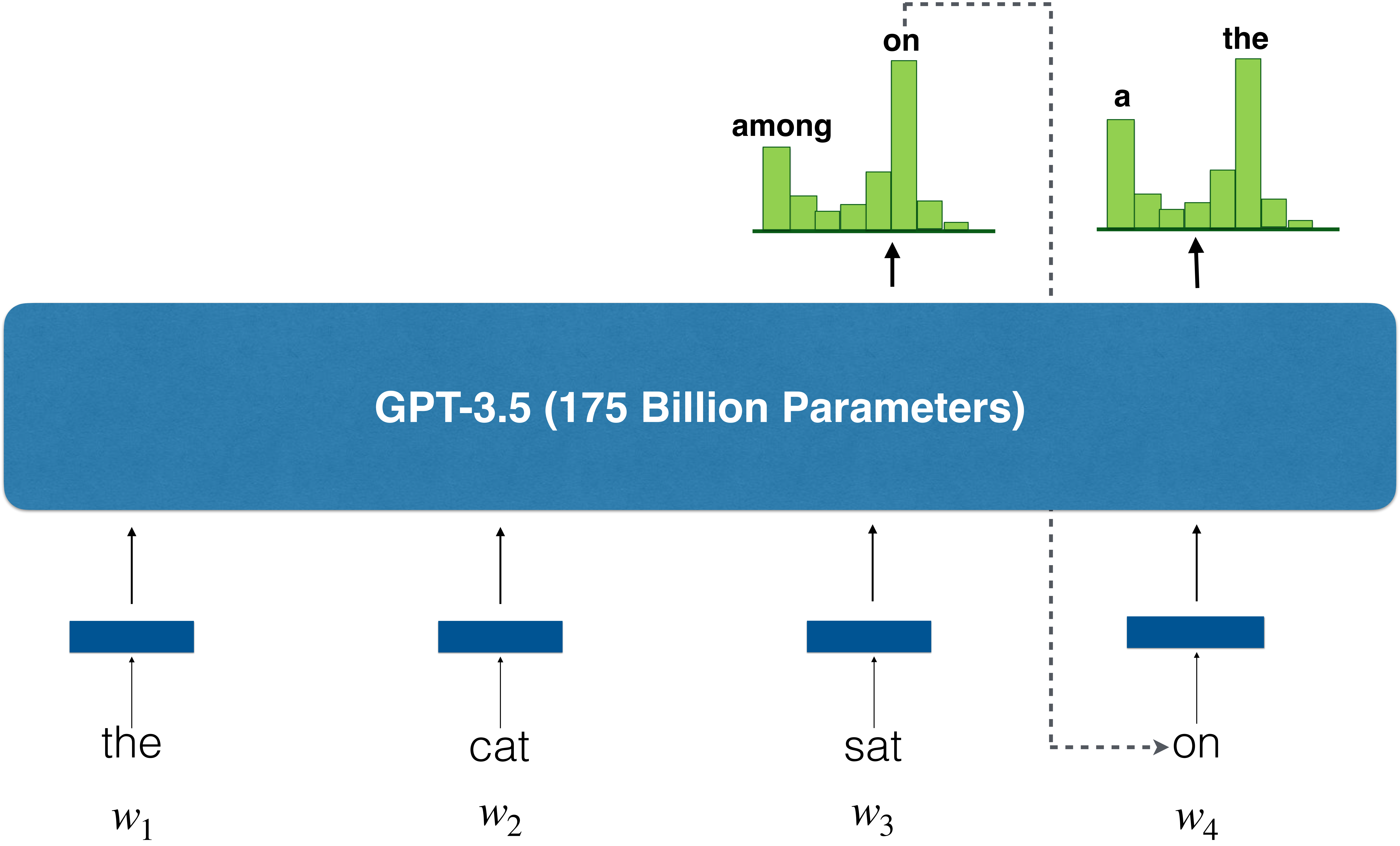
Generating Text with Language Models



Generating Text with Language Models



Generating Text with Language Models



Pretraining is Necessary but not Sufficient

- Pretrained models do well with providing good “continuations” to text but are not necessarily good at conversations or providing good responses to questions/instructions

- Example: On seeing a question such as “Who is India’s Father of the Nation?”, any of the following would be good continuations generated by a pretrained model:
  - Who gave him this honorific title?
  - This is a simple question that was asked during a high-school quiz
  - Mahatma Gandhi

- The pretrained model needs to be further *fine-tuned* or “*aligned*” to behave as we would like

ChatGPT

ChatGPT is a large language model trained to learn instructions with human feedback

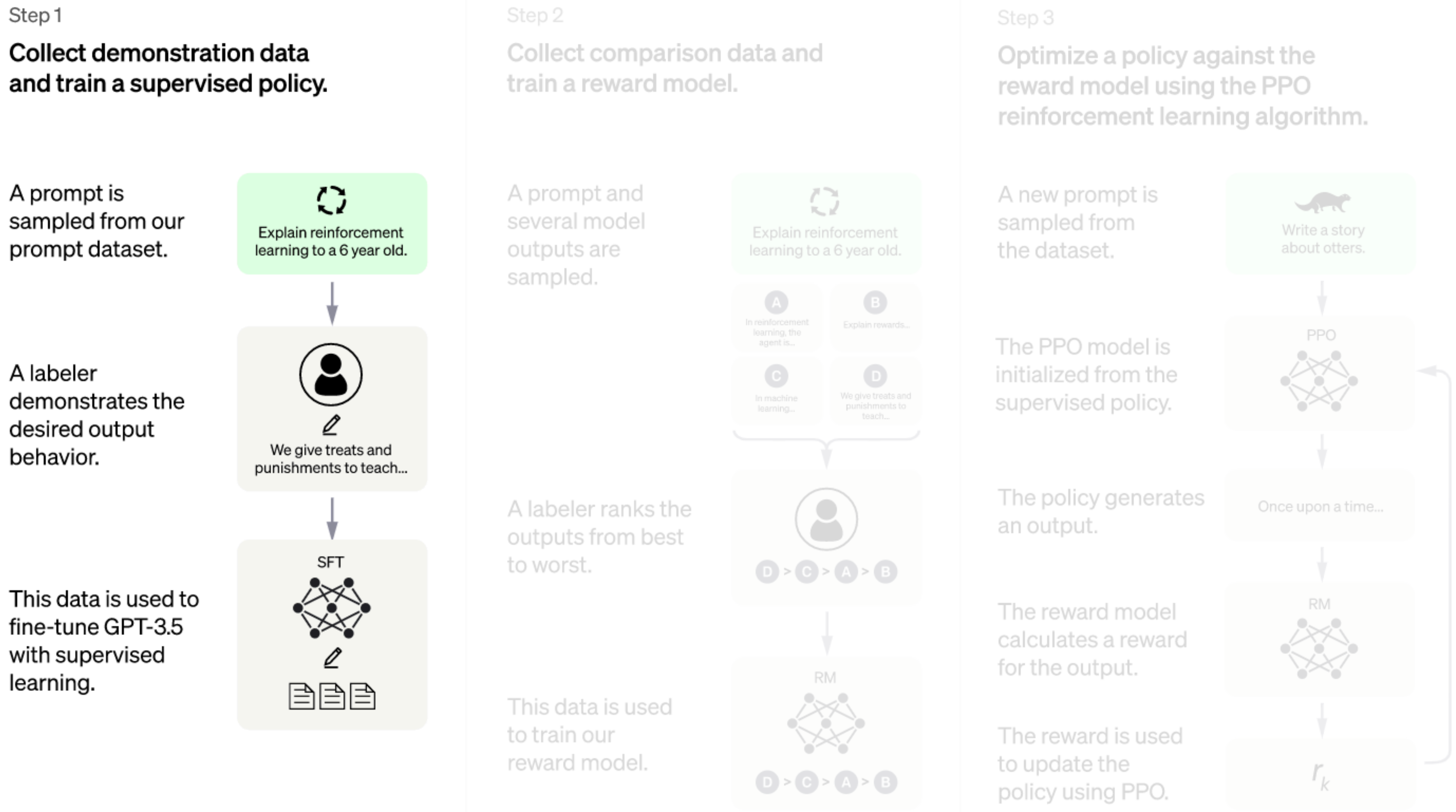
1

What is a language model?

2

How is ChatGPT trained to learn using human feedback?

Reinforcement Learning from Human Feedback (RLHF)



Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

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.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

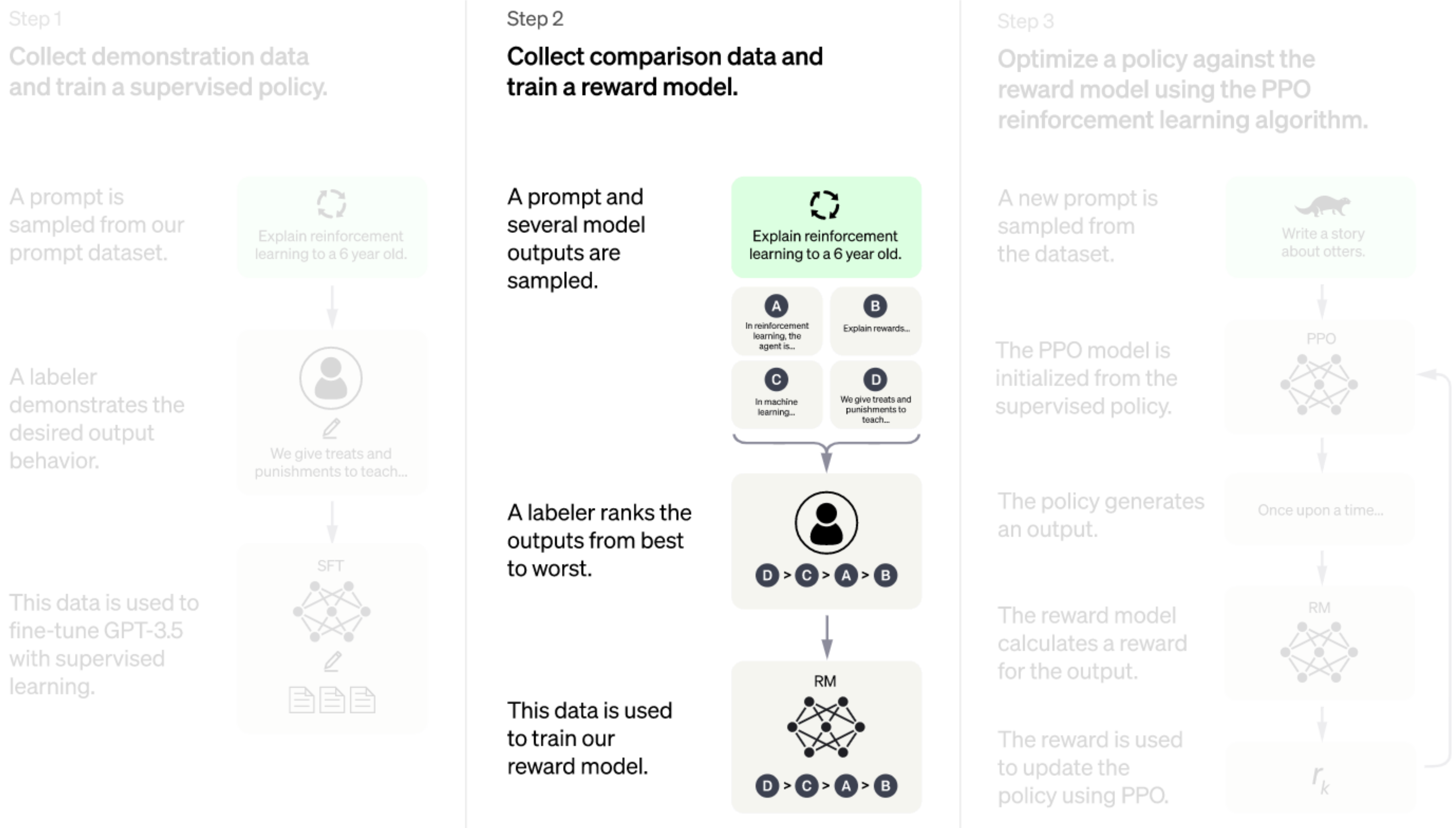
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

Figure reproduced from https://openai.com/blog/chatgpt

Reinforcement Learning from Human Feedback (RLHF)



Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

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.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

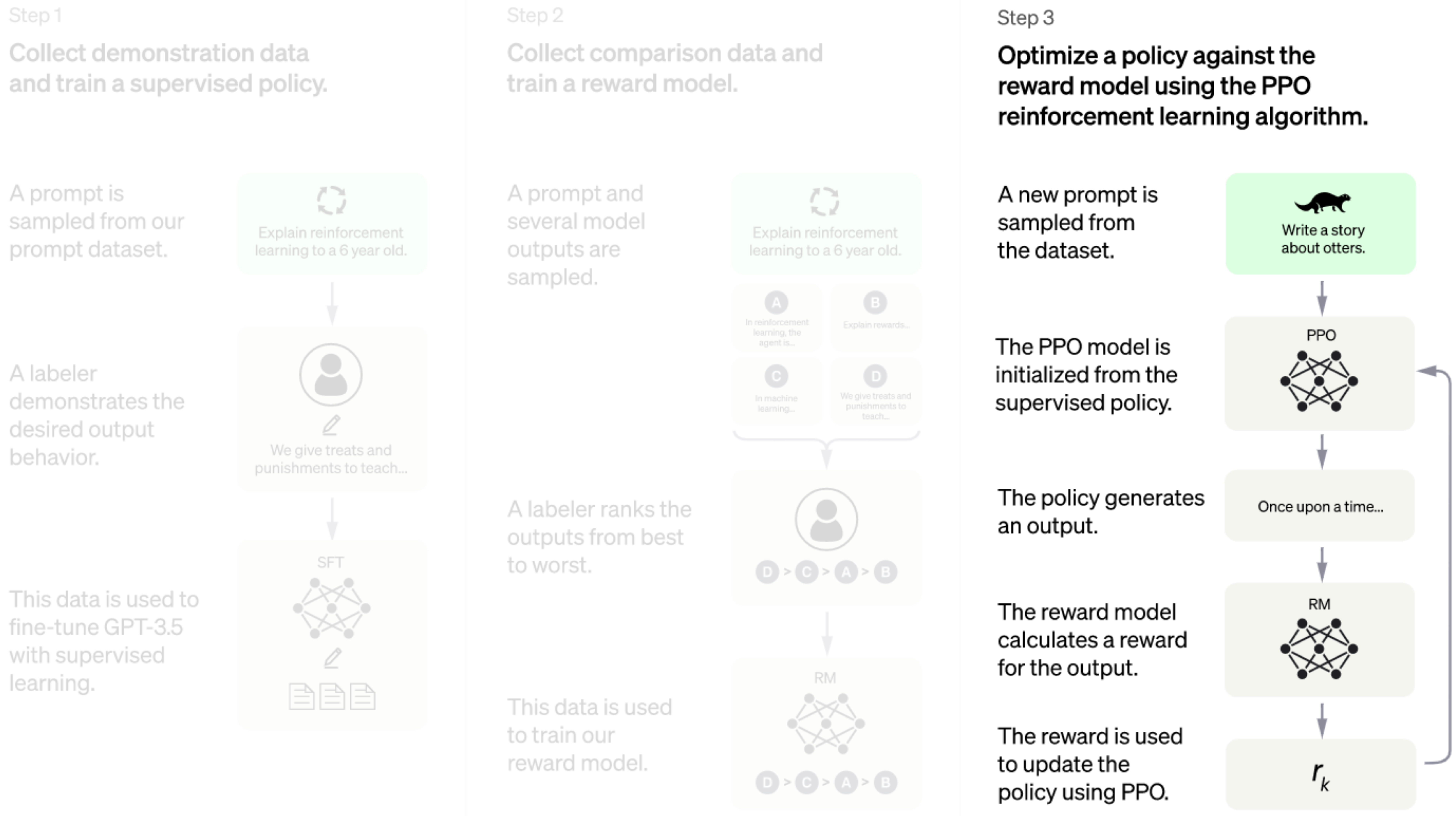
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

Figure reproduced from https://openai.com/blog/chatgpt

Reinforcement Learning from Human Feedback (RLHF)



Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT

Step 2

Collect comparison data and train a reward model.

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.

Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

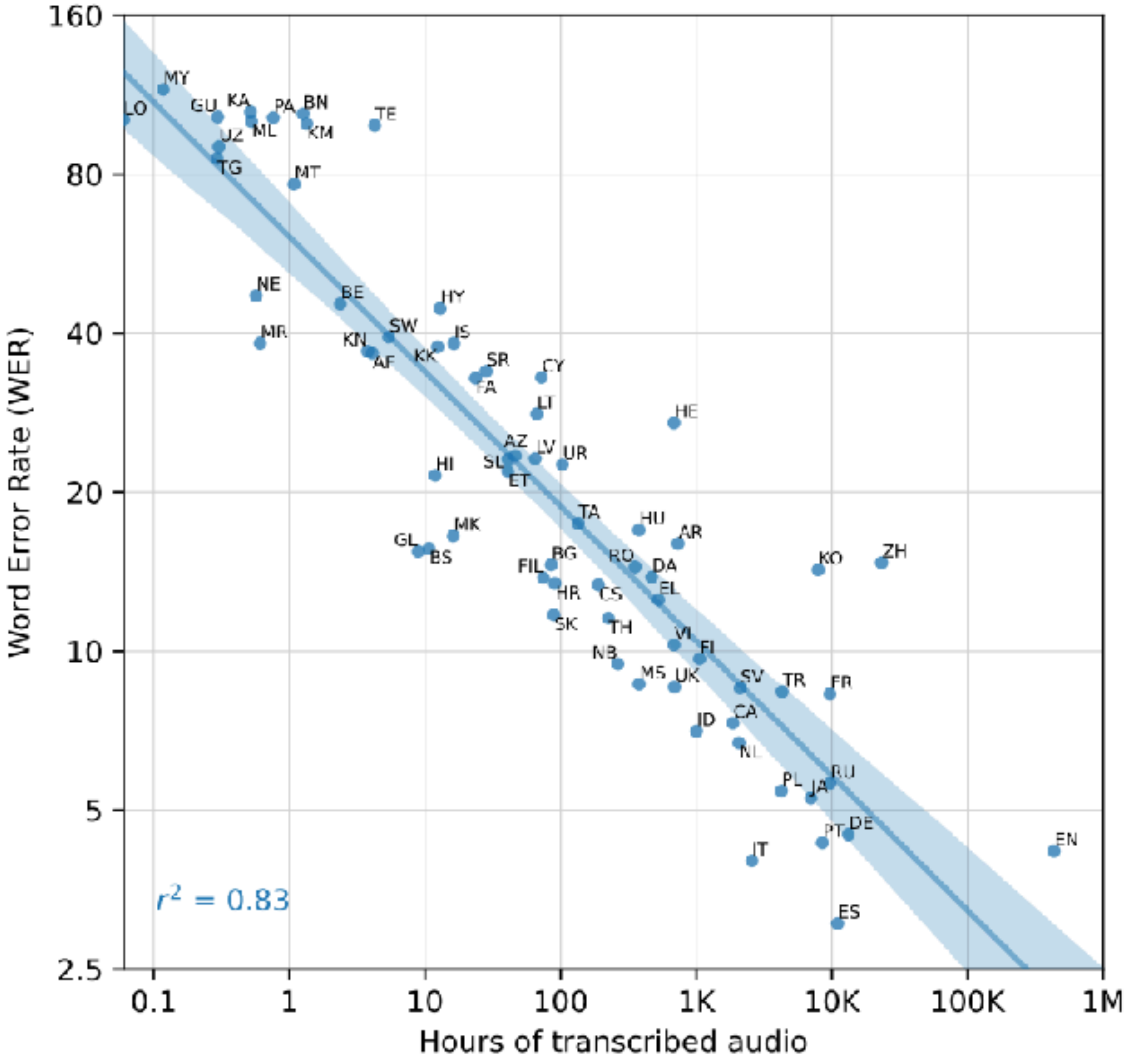
The reward is used to update the policy using PPO.

Figure reproduced from https://openai.com/blog/chatgpt

Speech and Chatbots?

- Add a speech-to-text plugin with LLMs like ChatGPT
  - However, speech technologies face significant challenges in India with 100s of languages, 1000s of dialects*
- State-of-the-art speech recognition systems note high correlation between supervision for a language/accent and final error rates [1]

> "We observe lower accuracy on low-resource and/or low-discoverability languages or languages where we have less training data. The models also exhibit disparate performance on different accents and dialects of particular languages."
>
> https://github.com/openai/whisper/blob/main/model-card.md



* Census 2011: 19,569 raw linguistic affiliations, 1369 rationalized mother tongu

[1] "Robust Speech Recognition via Large-scale Weak Supervision", Radford et al., https://arxiv.org/pdf/2212.04356.pdf, Dec 2022

Chatting using Code-Switching

- Code-switching: Switching between different languages within/across sentences

  Piya Tose Naina Laage का Amazing Rendition Deliver किया इस Audition पे

  - Widely prevalent in multilingual countries like India
  - Hard to get access to large amounts of code-switched data
  - Large diversity in how code-switching manifests

  But laughter therapy ने मेरी life बदल दी really

  But laughter therapy ने really में मेरी life change कर दी

  पर हंसी therapy ने मेरी life बदल दिया वास्तव में

- Computational models still have trouble processing code-switched speech and text

Applications of Machine Learning

Some examples from real life

• Google Search • Search within articles, e-books, etc.	**Information Retrieval**
• Face detection in photos • Detecting cars on roads (e.g., for self-driving cars)	**Computer Vision**
• Facebook recommending friends/ads • Netflix/Amazon/Flipkart recommendations	**Recommender Systems**
• Game playing (chess, Go, poker, etc.) • Robots (playing soccer, robotic arm, etc.)	**Robotics/Game Playing**
• IBM’s Watson (Jeopardy, etc.) • Medical Diagnosis Systems	**ML in Expert Systems**
• Personal Assistant/LLMs (ChatGPT, Google Assistant, Alexa, etc.) • Closed Captioning (Google Meet, MS Teams, Zoom, etc.)	**Speech & Natural Language Processing**

You and ML: AI/ML + X

AI/ML + Language/ Vision

AI/ML + Society

AI/ML + Theory

AI/ML + Systems

AI/ML + Data/Users

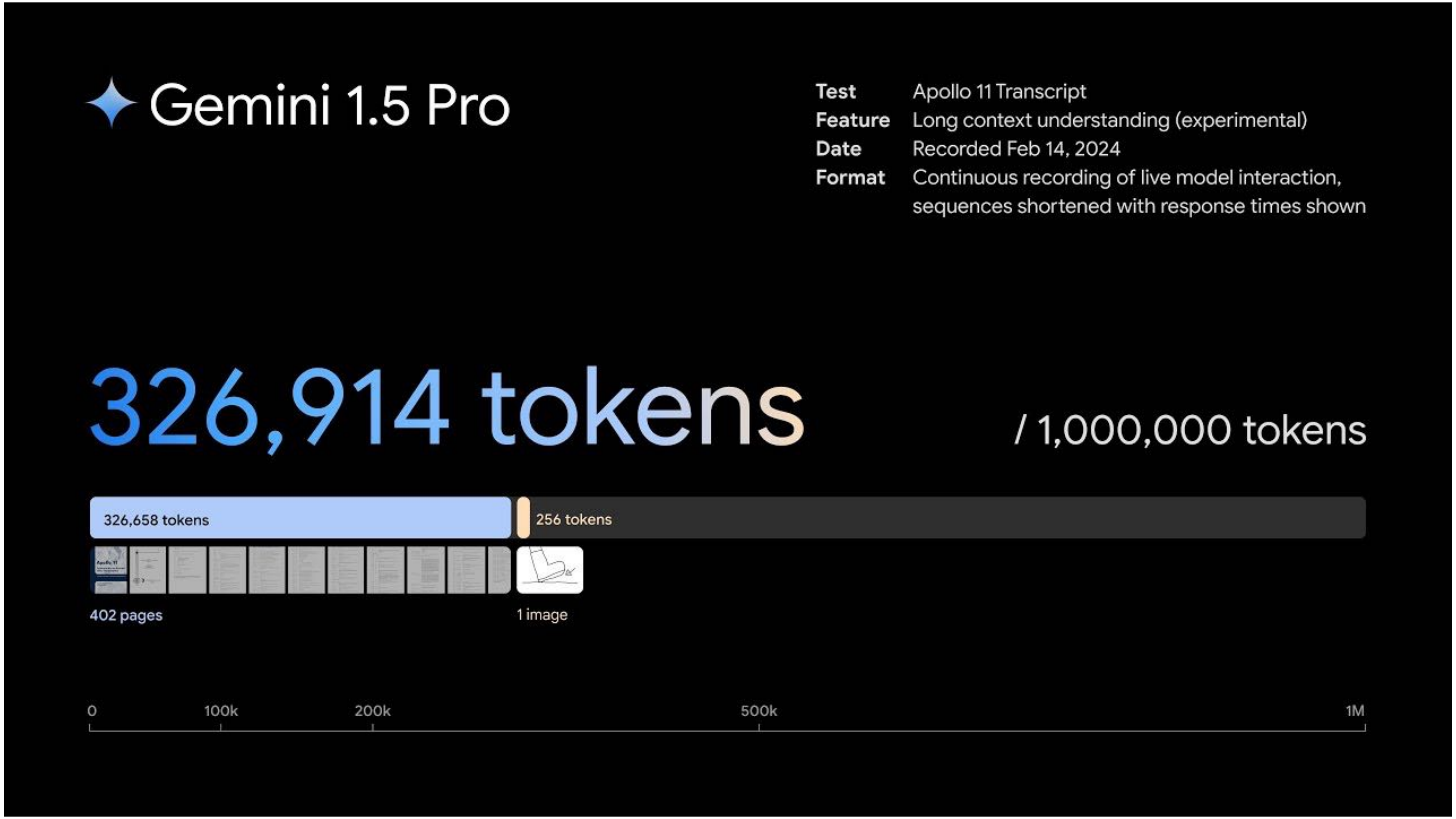
…

AI/ML + Analytics

AI/ML + Sciences

AI/ML + X

AI/ML + Language/ Vision



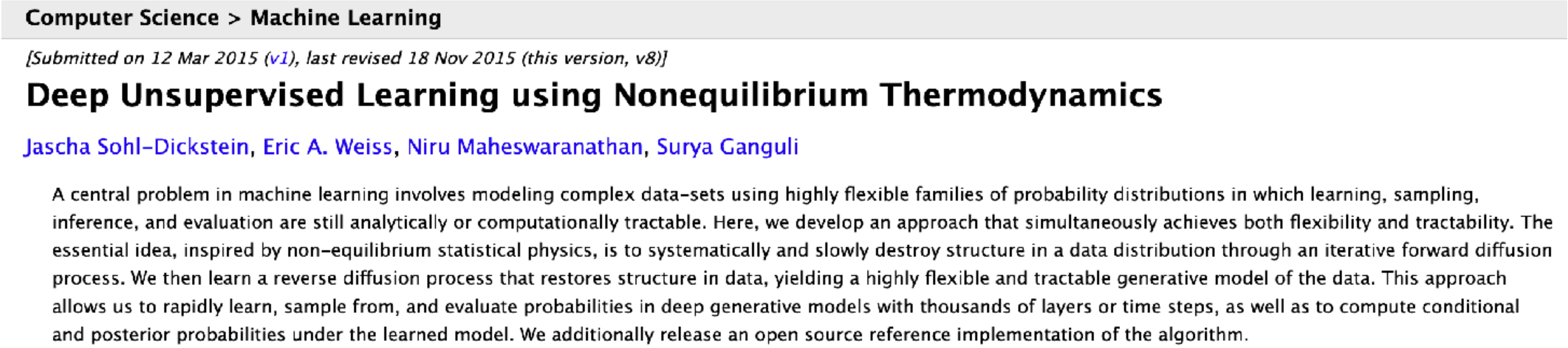
Gemini 1.5 Pro

Test Apollo 11 Transcript
Feature Long context understanding (experimental)
Date Recorded Feb 14, 2024
Format Continuous recording of live model interaction, sequences shortened with response times shown

326,914 tokens / 1,000,000 tokens

Video from: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

AI/ML + Theory



Computer Science > Machine Learning

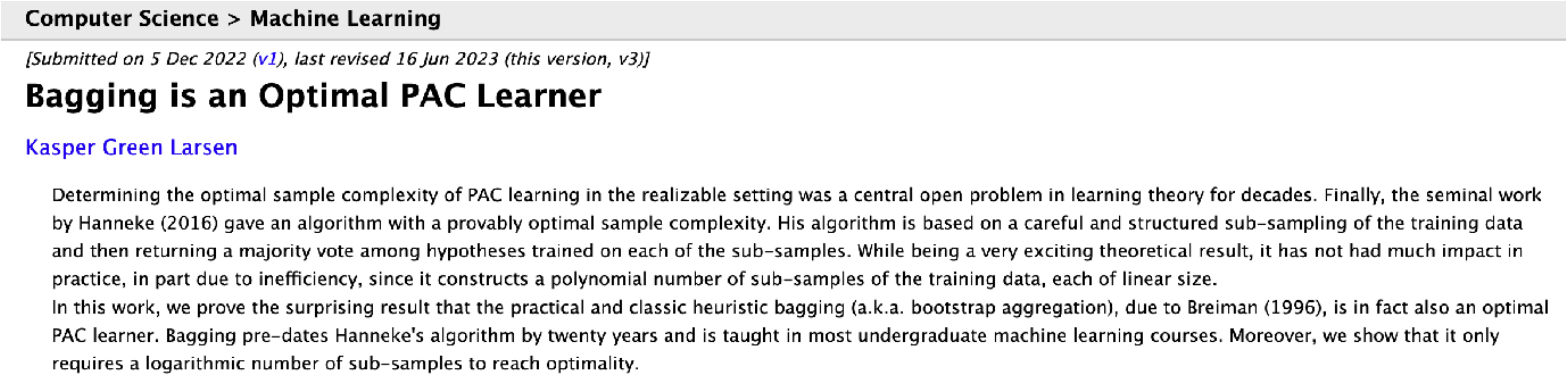
[Submitted on 12 Mar 2015 (v1), last revised 18 Nov 2015 (this version, v8)]

Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, Surya Ganguli

A central problem in machine learning involves modeling complex data-sets using highly flexible families of probability distributions in which learning, sampling, inference, and evaluation are still analytically or computationally tractable. Here, we develop an approach that simultaneously achieves both flexibility and tractability. The essential idea, inspired by non-equilibrium statistical physics, is to systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process. We then learn a reverse diffusion process that restores structure in data, yielding a highly flexible and tractable generative model of the data. This approach allows us to rapidly learn, sample from, and evaluate probabilities in deep generative models with thousands of layers or time steps, as well as to compute conditional and posterior probabilities under the learned model. We additionally release an open source reference implementation of the algorithm.

AI/ML + Theory



Computer Science > Machine Learning

[Submitted on 5 Dec 2022 (v1), last revised 16 Jun 2023 (this version, v3)]

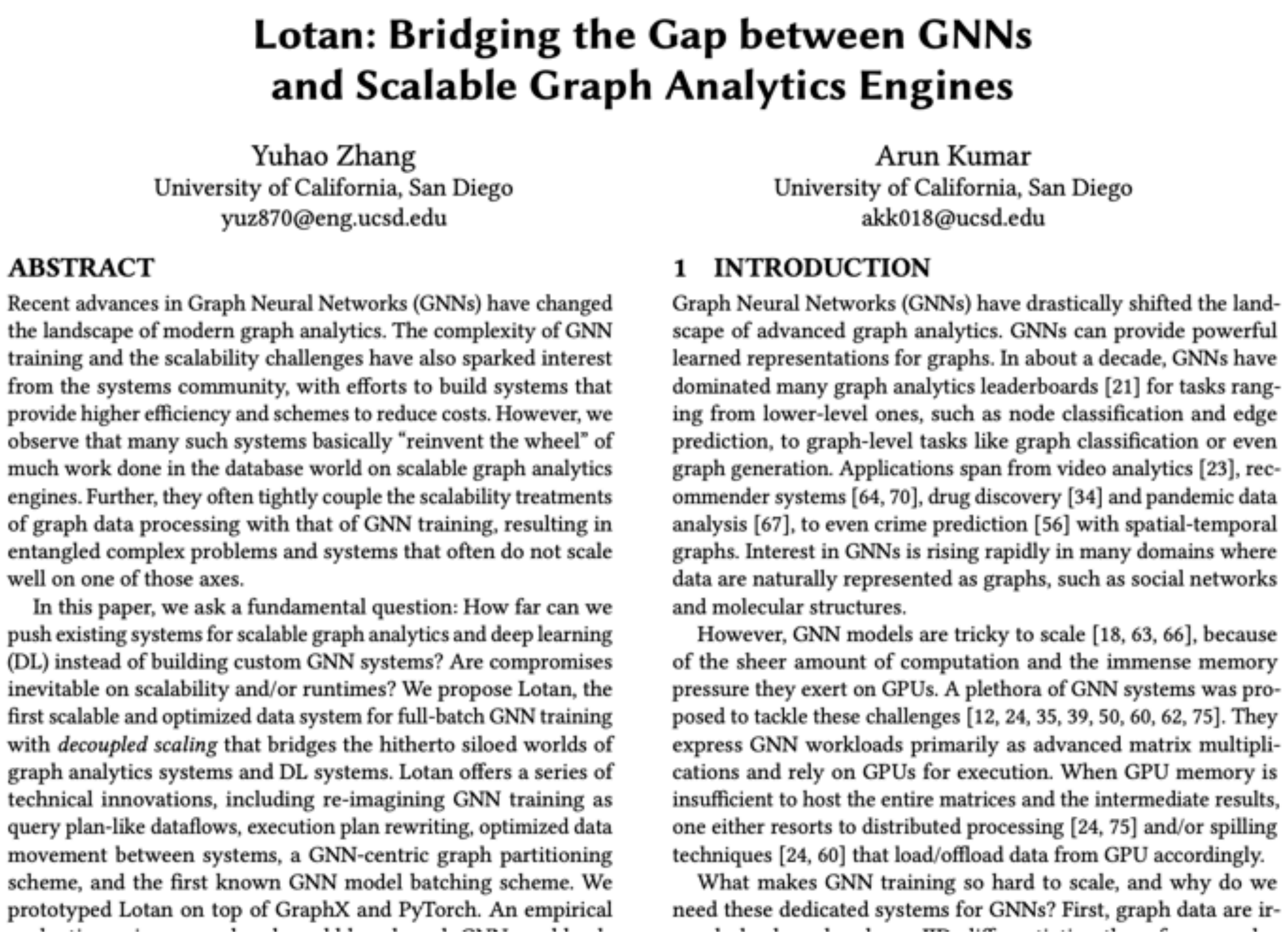
Bagging is an Optimal PAC Learner

Kasper Green Larsen

Determining the optimal sample complexity of PAC learning in the realizable setting was a central open problem in learning theory for decades. Finally, the seminal work by Hanneke (2016) gave an algorithm with a provably optimal sample complexity. His algorithm is based on a careful and structured sub-sampling of the training data and then returning a majority vote among hypotheses trained on each of the sub-samples. While being a very exciting theoretical result, it has not had much impact in practice, in part due to inefficiency, since it constructs a polynomial number of sub-samples of the training data, each of linear size.

In this work, we prove the surprising result that the practical and classic heuristic bagging (a.k.a. bootstrap aggregation), due to Breiman (1996), is in fact also an optimal PAC learner. Bagging pre-dates Hanneke's algorithm by twenty years and is taught in most undergraduate machine learning courses. Moreover, we show that it only requires a logarithmic number of sub-samples to reach optimality.

AI/ML + Analytics



Lotan: Bridging the Gap between GNNs and Scalable Graph Analytics Engines

Yuhao Zhang
University of California, San Diego
yuz870@eng.ucsd.edu

Arun Kumar
University of California, San Diego
akk018@ucsd.edu

ABSTRACT

Recent advances in Graph Neural Networks (GNNs) have changed the landscape of modern graph analytics. The complexity of GNN training and the scalability challenges have also sparked interest from the systems community, with efforts to build systems that provide higher efficiency and schemes to reduce costs. However, we observe that many such systems basically "reinvent the wheel" of much work done in the database world on scalable graph analytics engines. Further, they often tightly couple the scalability treatments of graph data processing with that of GNN training, resulting in entangled complex problems and systems that often do not scale well on one of those axes.

In this paper, we ask a fundamental question: How far can we push existing systems for scalable graph analytics and deep learning (DL) instead of building custom GNN systems? Are compromises inevitable on scalability and/or runtimes? We propose Lotan, the first scalable and optimized data system for full-batch GNN training with *decoupled scaling* that bridges the hitherto siloed worlds of graph analytics systems and DL systems. Lotan offers a series of technical innovations, including re-imagining GNN training as query plan-like dataflows, execution plan rewriting, optimized data movement between systems, a GNN-centric graph partitioning scheme, and the first known GNN model batching scheme. We prototyped Lotan on top of GraphX and PyTorch. An empirical

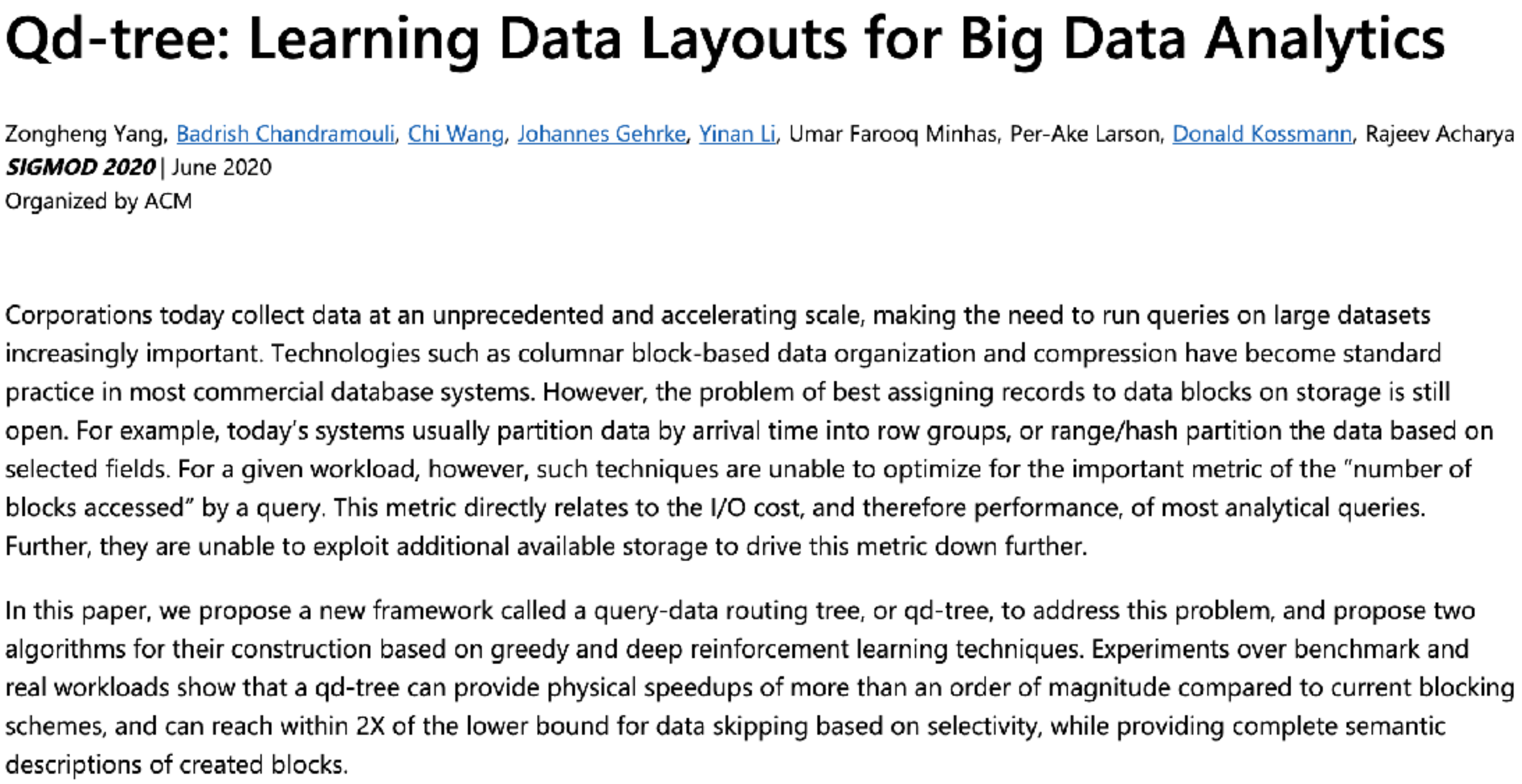
1 INTRODUCTION

Graph Neural Networks (GNNs) have drastically shifted the landscape of advanced graph analytics. GNNs can provide powerful learned representations for graphs. In about a decade, GNNs have dominated many graph analytics leaderboards [21] for tasks ranging from lower-level ones, such as node classification and edge prediction, to graph-level tasks like graph classification or even graph generation. Applications span from video analytics [23], recommender systems [64, 70], drug discovery [34] and pandemic data analysis [67], to even crime prediction [56] with spatial-temporal graphs. Interest in GNNs is rising rapidly in many domains where data are naturally represented as graphs, such as social networks and molecular structures.

However, GNN models are tricky to scale [18, 63, 66], because of the sheer amount of computation and the immense memory pressure they exert on GPUs. A plethora of GNN systems was proposed to tackle these challenges [12, 24, 35, 39, 50, 60, 62, 75]. They express GNN workloads primarily as advanced matrix multiplications and rely on GPUs for execution. When GPU memory is insufficient to host the entire matrices and the intermediate results, one either resorts to distributed processing [24, 75] and/or spilling techniques [24, 60] that load/offload data from GPU accordingly.

What makes GNN training so hard to scale, and why do we need these dedicated systems for GNNs? First, graph data are ir-

AI/ML + Analytics



Qd-tree: Learning Data Layouts for Big Data Analytics

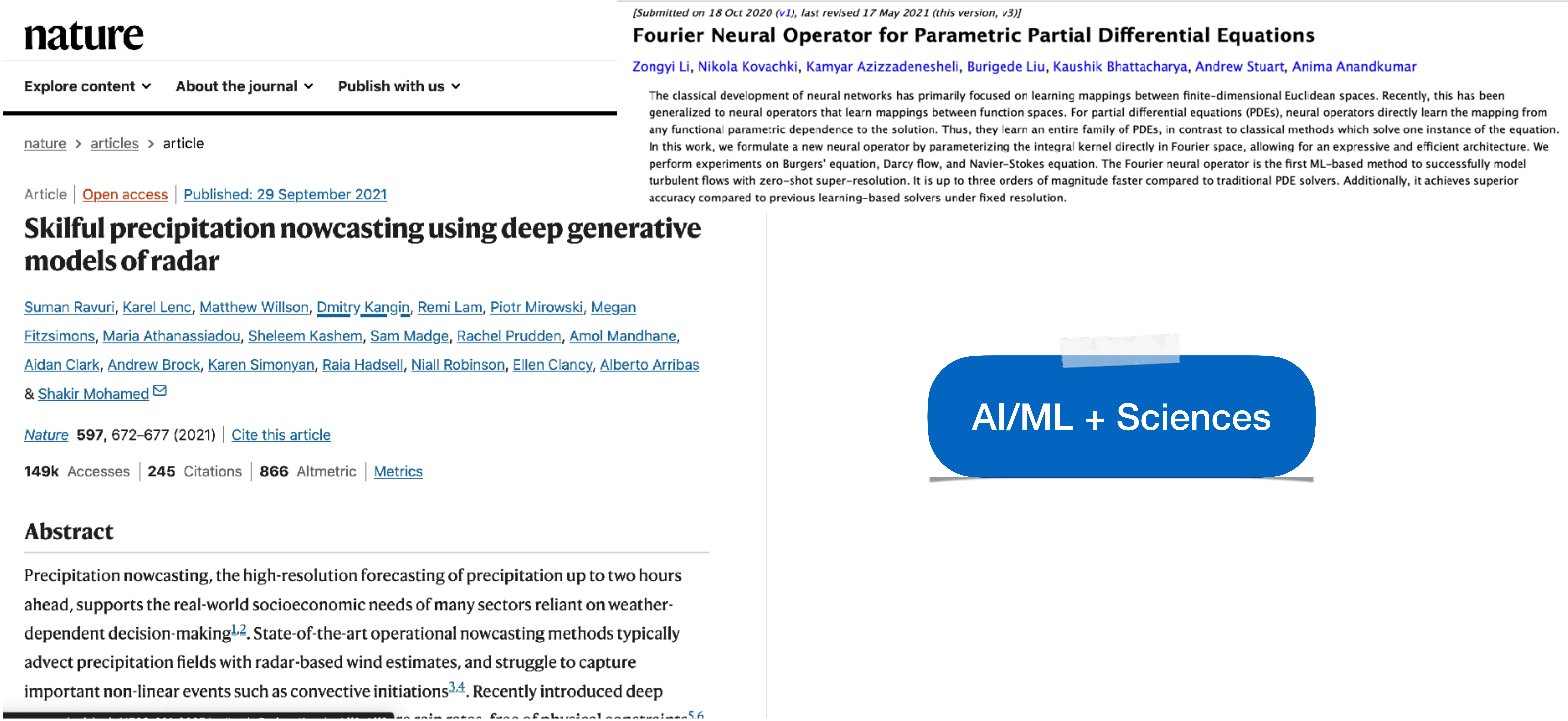
Zongheng Yang, Badrish Chandramouli, Chi Wang, Johannes Gehrke, Yinan Li, Umar Farooq Minhas, Per-Ake Larson, Donald Kossmann, Rajeev Acharya

SIGMOD 2020 | June 2020
Organized by ACM

Corporations today collect data at an unprecedented and accelerating scale, making the need to run queries on large datasets increasingly important. Technologies such as columnar block-based data organization and compression have become standard practice in most commercial database systems. However, the problem of best assigning records to data blocks on storage is still open. For example, today's systems usually partition data by arrival time into row groups, or range/hash partition the data based on selected fields. For a given workload, however, such techniques are unable to optimize for the important metric of the "number of blocks accessed" by a query. This metric directly relates to the I/O cost, and therefore performance, of most analytical queries. Further, they are unable to exploit additional available storage to drive this metric down further.

In this paper, we propose a new framework called a query-data routing tree, or qd-tree, to address this problem, and propose two algorithms for their construction based on greedy and deep reinforcement learning techniques. Experiments over benchmark and real workloads show that a qd-tree can provide physical speedups of more than an order of magnitude compared to current blocking schemes, and can reach within 2X of the lower bound for data skipping based on selectivity, while providing complete semantic descriptions of created blocks.

AI/ML + Sciences



[Submitted on 18 Oct 2020 (v1), last revised 17 May 2021 (this version, v3)]

Fourier Neural Operator for Parametric Partial Differential Equations

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, Anima Anandkumar

The classical development of neural networks has primarily focused on learning mappings between finite-dimensional Euclidean spaces. Recently, this has been generalized to neural operators that learn mappings between function spaces. For partial differential equations (PDEs), neural operators directly learn the mapping from any functional parametric dependence to the solution. Thus, they learn an entire family of PDEs, in contrast to classical methods which solve one instance of the equation. In this work, we formulate a new neural operator by parameterizing the integral kernel directly in Fourier space, allowing for an expressive and efficient architecture. We perform experiments on Burgers' equation, Darcy flow, and Navier-Stokes equation. The Fourier neural operator is the first ML-based method to successfully model turbulent flows with zero-shot super-resolution. It is up to three orders of magnitude faster compared to traditional PDE solvers. Additionally, it achieves superior accuracy compared to previous learning-based solvers under fixed resolution.

nature

nature > articles > article

Article | Open access | Published: 29 September 2021

Skilful precipitation nowcasting using deep generative models of radar

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas & Shakir Mohamed

Nature **597**, 672–677 (2021) | Cite this article

149k Accesses | **245** Citations | **866** Altmetric | Metrics

Abstract

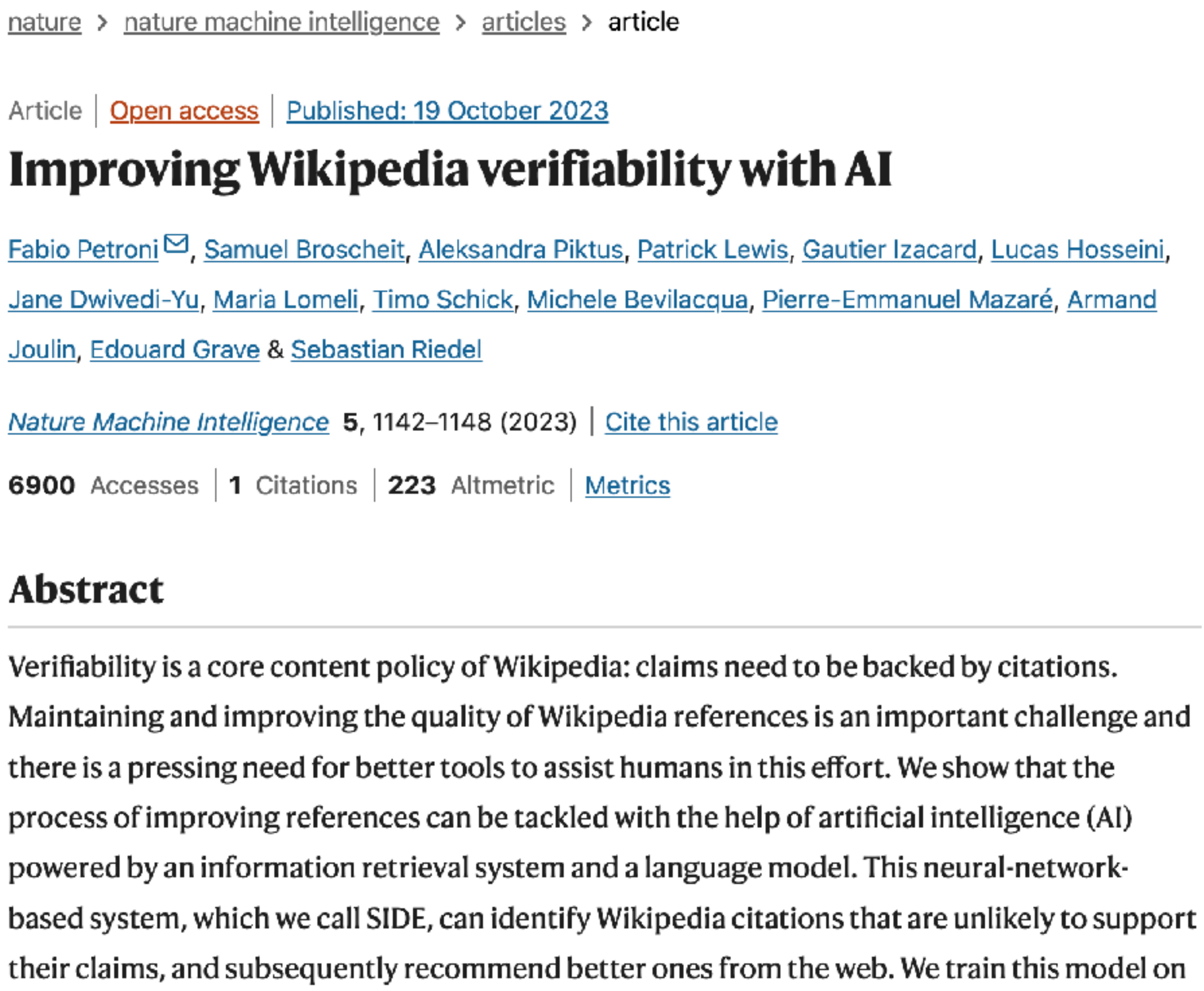
Precipitation nowcasting, the high-resolution forecasting of precipitation up to two hours ahead, supports the real-world socioeconomic needs of many sectors reliant on weather-dependent decision-making[1,2]. State-of-the-art operational nowcasting methods typically advect precipitation fields with radar-based wind estimates, and struggle to capture important non-linear events such as convective initiations[3,4]. Recently introduced deep

re rain rates, free of physical constraints[5,6]

AI/ML + Sciences

AI/ML + Society

AI/ML + Society



Article | Open access | Published: 19 October 2023

Improving Wikipedia verifiability with AI

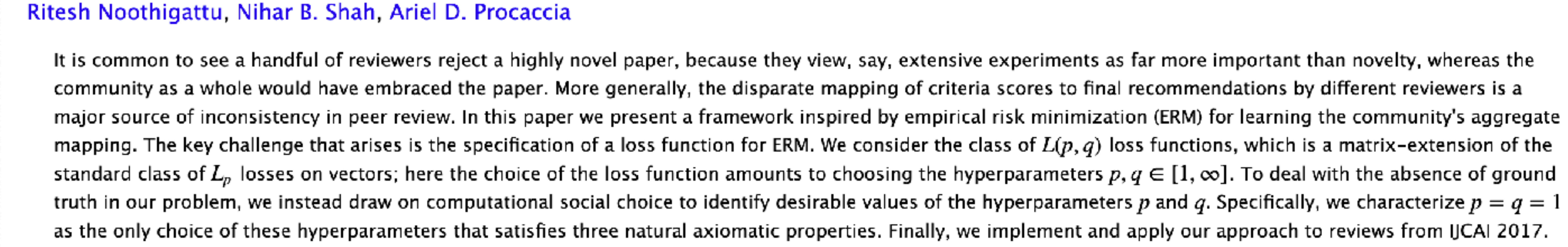
Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Michele Bevilacqua, Pierre-Emmanuel Mazaré, Armand Joulin, Edouard Grave & Sebastian Riedel

Nature Machine Intelligence **5**, 1142–1148 (2023) | Cite this article

6900 Accesses | **1** Citations | **223** Altmetric | Metrics

Abstract

Verifiability is a core content policy of Wikipedia: claims need to be backed by citations. Maintaining and improving the quality of Wikipedia references is an important challenge and there is a pressing need for better tools to assist humans in this effort. We show that the process of improving references can be tackled with the help of artificial intelligence (AI) powered by an information retrieval system and a language model. This neural-network-based system, which we call SIDE, can identify Wikipedia citations that are unlikely to support their claims, and subsequently recommend better ones from the web. We train this model on



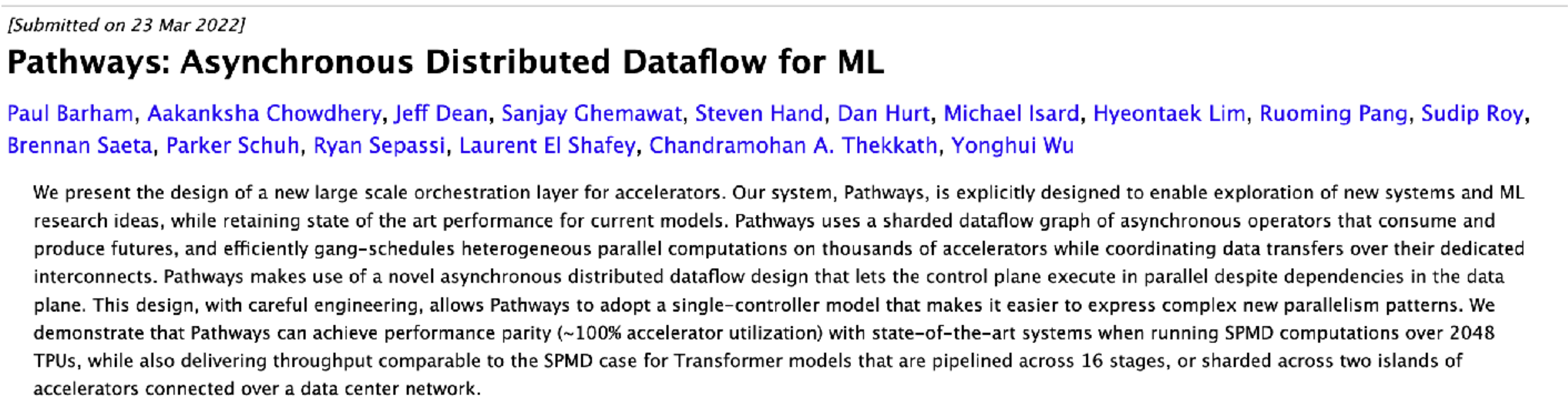
[Submitted on 27 Aug 2018 (v1), last revised 2 Mar 2020 (this version, v2)]

Loss Functions, Axioms, and Peer Review

Ritesh Noothigattu, Nihar B. Shah, Ariel D. Procaccia

It is common to see a handful of reviewers reject a highly novel paper, because they view, say, extensive experiments as far more important than novelty, whereas the community as a whole would have embraced the paper. More generally, the disparate mapping of criteria scores to final recommendations by different reviewers is a major source of inconsistency in peer review. In this paper we present a framework inspired by empirical risk minimization (ERM) for learning the community's aggregate mapping. The key challenge that arises is the specification of a loss function for ERM. We consider the class of $L(p,q)$ loss functions, which is a matrix-extension of the standard class of L_p losses on vectors; here the choice of the loss function amounts to choosing the hyperparameters $p,q \in [1,\infty]$. To deal with the absence of ground truth in our problem, we instead draw on computational social choice to identify desirable values of the hyperparameters p and q. Specifically, we characterize $p=q=1$ as the only choice of these hyperparameters that satisfies three natural axiomatic properties. Finally, we implement and apply our approach to reviews from IJCAI 2017.

AI/ML + Systems

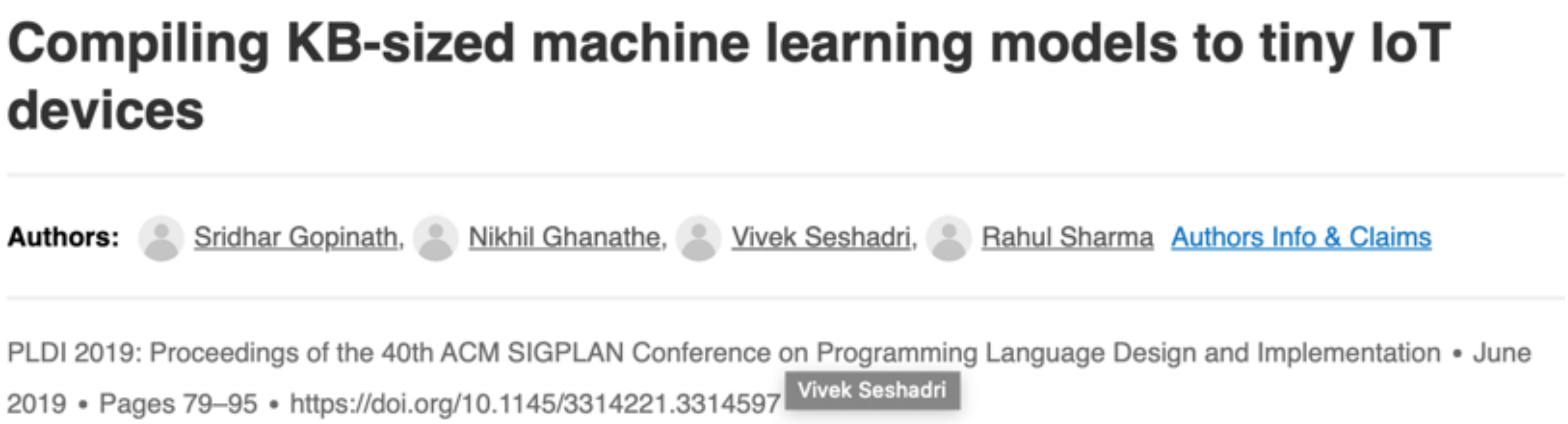


[Submitted on 23 Mar 2022]

Pathways: Asynchronous Distributed Dataflow for ML

Paul Barham, Aakanksha Chowdhery, Jeff Dean, Sanjay Ghemawat, Steven Hand, Dan Hurt, Michael Isard, Hyeontaek Lim, Ruoming Pang, Sudip Roy, Brennan Saeta, Parker Schuh, Ryan Sepassi, Laurent El Shafey, Chandramohan A. Thekkath, Yonghui Wu

We present the design of a new large scale orchestration layer for accelerators. Our system, Pathways, is explicitly designed to enable exploration of new systems and ML research ideas, while retaining state of the art performance for current models. Pathways uses a sharded dataflow graph of asynchronous operators that consume and produce futures, and efficiently gang-schedules heterogeneous parallel computations on thousands of accelerators while coordinating data transfers over their dedicated interconnects. Pathways makes use of a novel asynchronous distributed dataflow design that lets the control plane execute in parallel despite dependencies in the data plane. This design, with careful engineering, allows Pathways to adopt a single-controller model that makes it easier to express complex new parallelism patterns. We demonstrate that Pathways can achieve performance parity (~100% accelerator utilization) with state-of-the-art systems when running SPMD computations over 2048 TPUs, while also delivering throughput comparable to the SPMD case for Transformer models that are pipelined across 16 stages, or sharded across two islands of accelerators connected over a data center network.



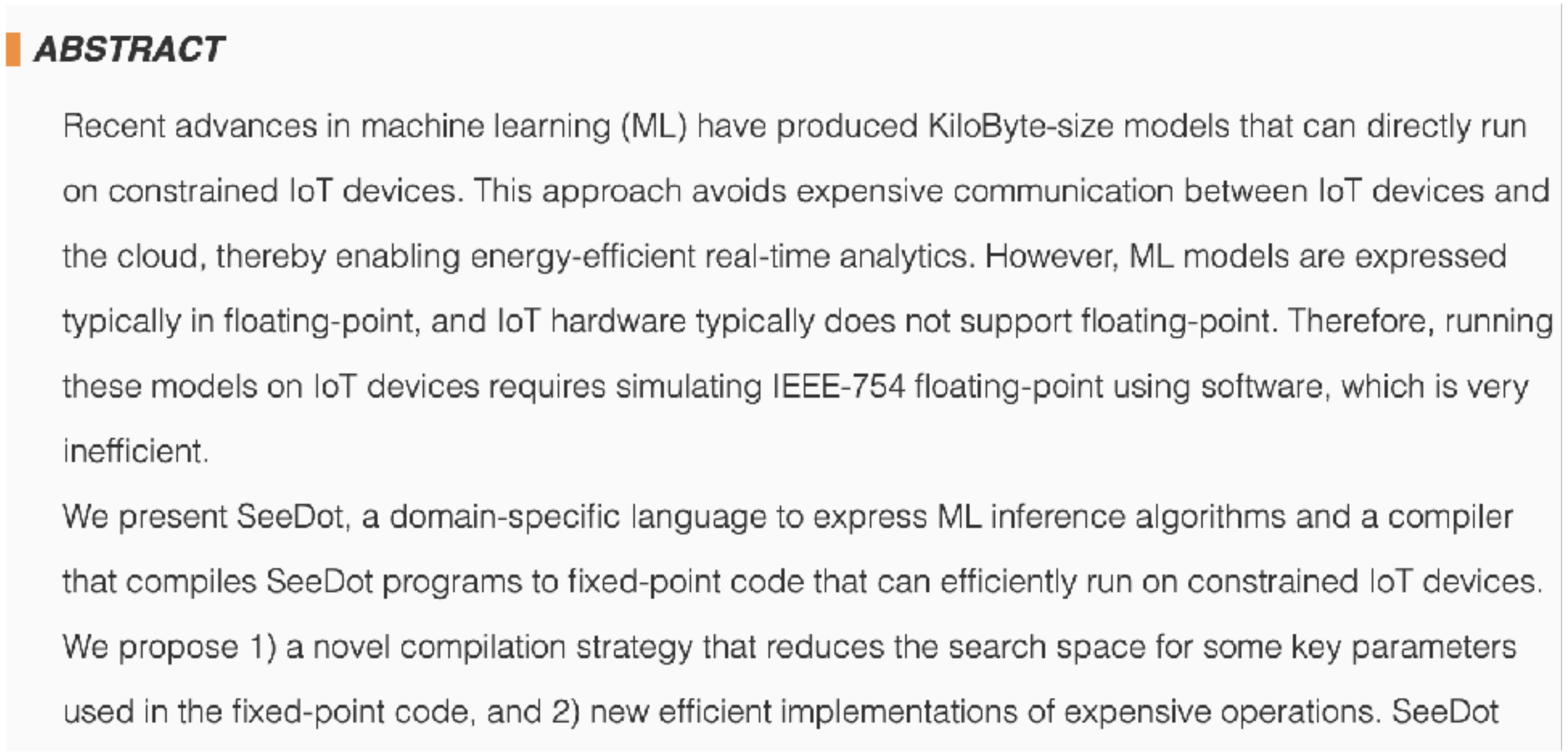
Compiling KB-sized machine learning models to tiny IoT devices

Authors: Sridhar Gopinath, Nikhil Ghanathe, Vivek Seshadri, Rahul Sharma Authors Info & Claims

PLDI 2019: Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation • June 2019 • Pages 79–95 • https://doi.org/10.1145/3314221.3314597

Vivek Seshadri

AI/ML + Systems

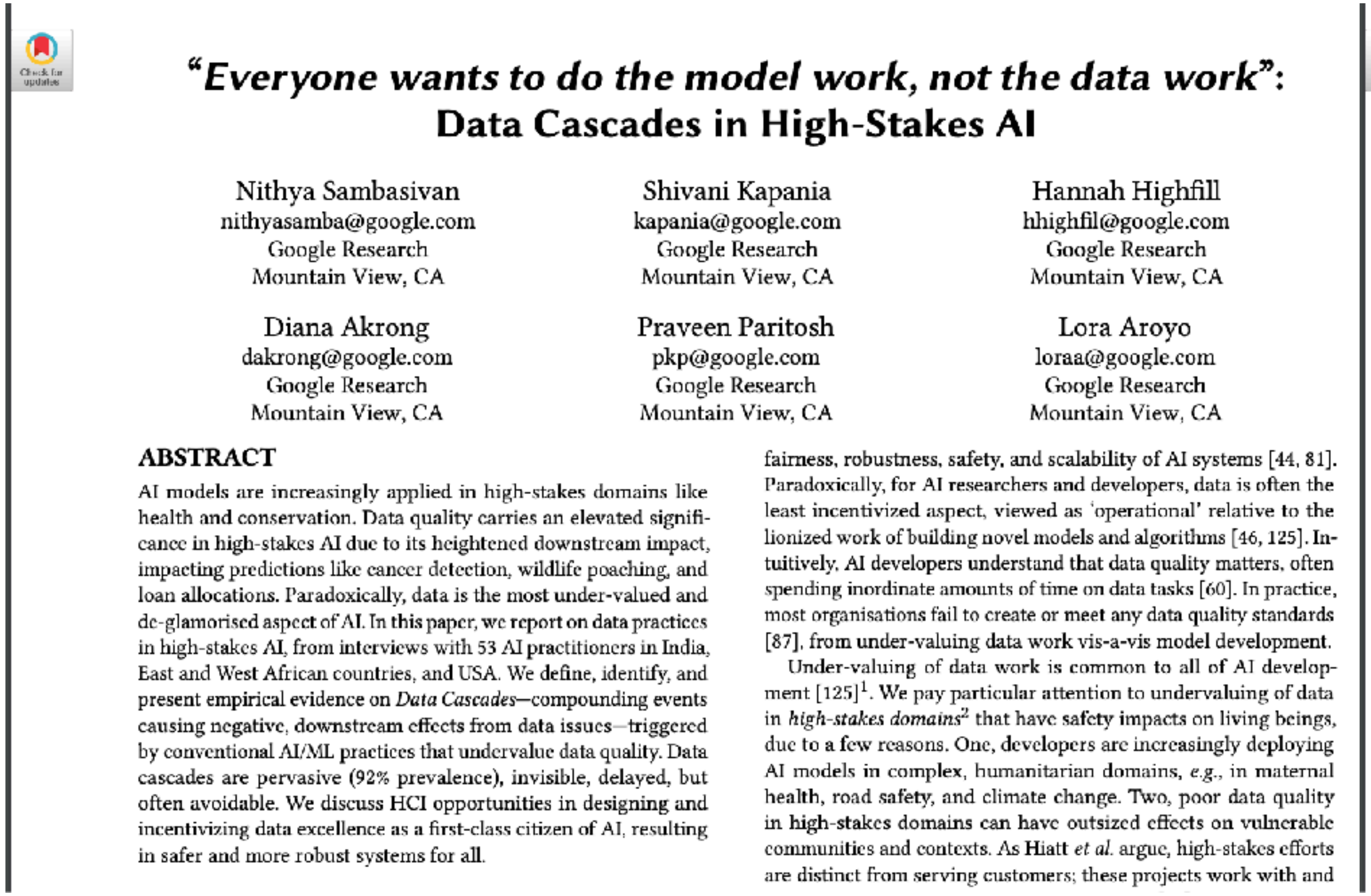


ABSTRACT

Recent advances in machine learning (ML) have produced KiloByte-size models that can directly run on constrained IoT devices. This approach avoids expensive communication between IoT devices and the cloud, thereby enabling energy-efficient real-time analytics. However, ML models are expressed typically in floating-point, and IoT hardware typically does not support floating-point. Therefore, running these models on IoT devices requires simulating IEEE-754 floating-point using software, which is very inefficient.

We present SeeDot, a domain-specific language to express ML inference algorithms and a compiler that compiles SeeDot programs to fixed-point code that can efficiently run on constrained IoT devices. We propose 1) a novel compilation strategy that reduces the search space for some key parameters used in the fixed-point code, and 2) new efficient implementations of expensive operations. SeeDot

AI/ML + Data/Users



"*Everyone wants to do the model work, not the data work*": Data Cascades in High-Stakes AI

Nithya Sambasivan
nithyasamba@google.com
Google Research
Mountain View, CA

Shivani Kapania
kapania@google.com
Google Research
Mountain View, CA

Hannah Highfill
hhighfil@google.com
Google Research
Mountain View, CA

Diana Akrong
dakrong@google.com
Google Research
Mountain View, CA

Praveen Paritosh
pkp@google.com
Google Research
Mountain View, CA

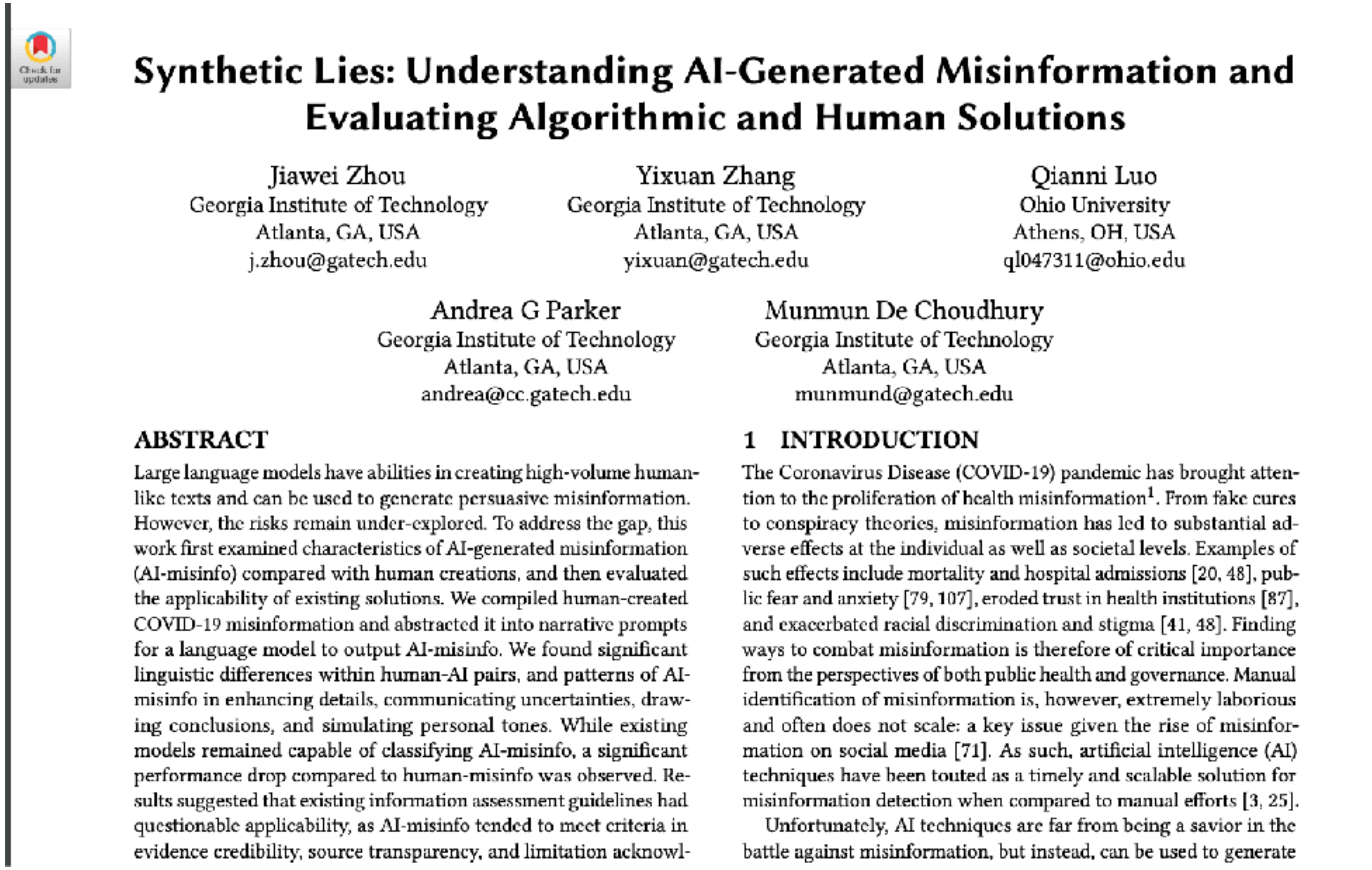
Lora Aroyo
loraa@google.com
Google Research
Mountain View, CA

ABSTRACT

AI models are increasingly applied in high-stakes domains like health and conservation. Data quality carries an elevated significance in high-stakes AI due to its heightened downstream impact, impacting predictions like cancer detection, wildlife poaching, and loan allocations. Paradoxically, data is the most under-valued and de-glamorised aspect of AI. In this paper, we report on data practices in high-stakes AI, from interviews with 53 AI practitioners in India, East and West African countries, and USA. We define, identify, and present empirical evidence on *Data Cascades*—compounding events causing negative, downstream effects from data issues—triggered by conventional AI/ML practices that undervalue data quality. Data cascades are pervasive (92% prevalence), invisible, delayed, but often avoidable. We discuss HCI opportunities in designing and incentivizing data excellence as a first-class citizen of AI, resulting in safer and more robust systems for all.

fairness, robustness, safety, and scalability of AI systems [44, 81]. Paradoxically, for AI researchers and developers, data is often the least incentivized aspect, viewed as 'operational' relative to the lionized work of building novel models and algorithms [46, 125]. Intuitively, AI developers understand that data quality matters, often spending inordinate amounts of time on data tasks [60]. In practice, most organisations fail to create or meet any data quality standards [87], from under-valuing data work vis-a-vis model development.

Under-valuing of data work is common to all of AI development [125][1]. We pay particular attention to undervaluing of data in *high-stakes domains*[2] that have safety impacts on living beings, due to a few reasons. One, developers are increasingly deploying AI models in complex, humanitarian domains, *e.g.*, in maternal health, road safety, and climate change. Two, poor data quality in high-stakes domains can have outsized effects on vulnerable communities and contexts. As Hiatt *et al.* argue, high-stakes efforts are distinct from serving customers; these projects work with and



Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions

Jiawei Zhou
Georgia Institute of Technology
Atlanta, GA, USA
j.zhou@gatech.edu

Yixuan Zhang
Georgia Institute of Technology
Atlanta, GA, USA
yixuan@gatech.edu

Qianni Luo
Ohio University
Athens, OH, USA
ql047311@ohio.edu

Andrea G Parker
Georgia Institute of Technology
Atlanta, GA, USA
andrea@cc.gatech.edu

Munmun De Choudhury
Georgia Institute of Technology
Atlanta, GA, USA
munmund@gatech.edu

ABSTRACT

Large language models have abilities in creating high-volume human-like texts and can be used to generate persuasive misinformation. However, the risks remain under-explored. To address the gap, this work first examined characteristics of AI-generated misinformation (AI-misinfo) compared with human creations, and then evaluated the applicability of existing solutions. We compiled human-created COVID-19 misinformation and abstracted it into narrative prompts for a language model to output AI-misinfo. We found significant linguistic differences within human-AI pairs, and patterns of AI-misinfo in enhancing details, communicating uncertainties, drawing conclusions, and simulating personal tones. While existing models remained capable of classifying AI-misinfo, a significant performance drop compared to human-misinfo was observed. Results suggested that existing information assessment guidelines had questionable applicability, as AI-misinfo tended to meet criteria in evidence credibility, source transparency, and limitation acknowl-

1 INTRODUCTION

The Coronavirus Disease (COVID-19) pandemic has brought attention to the proliferation of health misinformation[1]. From fake cures to conspiracy theories, misinformation has led to substantial adverse effects at the individual as well as societal levels. Examples of such effects include mortality and hospital admissions [20, 48], public fear and anxiety [79, 107], eroded trust in health institutions [87], and exacerbated racial discrimination and stigma [41, 48]. Finding ways to combat misinformation is therefore of critical importance from the perspectives of both public health and governance. Manual identification of misinformation is, however, extremely laborious and often does not scale: a key issue given the rise of misinformation on social media [71]. As such, artificial intelligence (AI) techniques have been touted as a timely and scalable solution for misinformation detection when compared to manual efforts [3, 25].

Unfortunately, AI techniques are far from being a savior in the battle against misinformation, but instead, can be used to generate

AI/ML + Data/Users

Graduate Programs for AI/ML: Groups in India[1]

IIT Bombay

IIT Delhi

IIIT Hyderabad

IISc Bangalore

IIT Madras

IIT Kharagpur

IIT Kanpur

IIIT Delhi

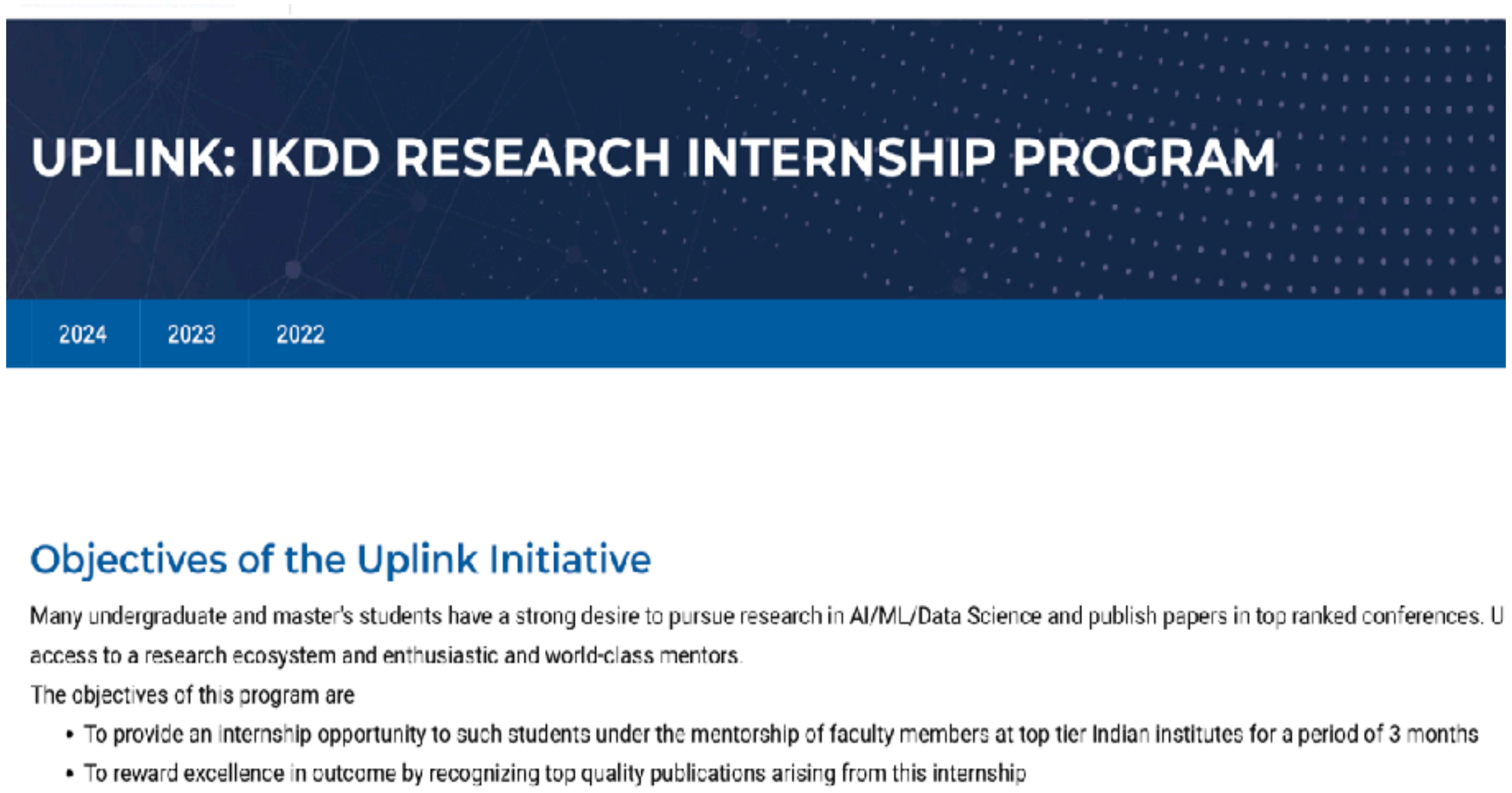
IIT Hyderabad

IIT Jodhpur

[1] Top-10 Indian institutions in AI/ML according to CSRankings: https://csrankings.org/#/index?ai&vision&mlmining&nlp&inforet&in

Getting started

- Many interesting subareas of AI/ML to explore. Identify where your interests lie.
- Apply to research opportunities to get started!



UPLINK: IKDD RESEARCH INTERNSHIP PROGRAM

2024 2023 2022

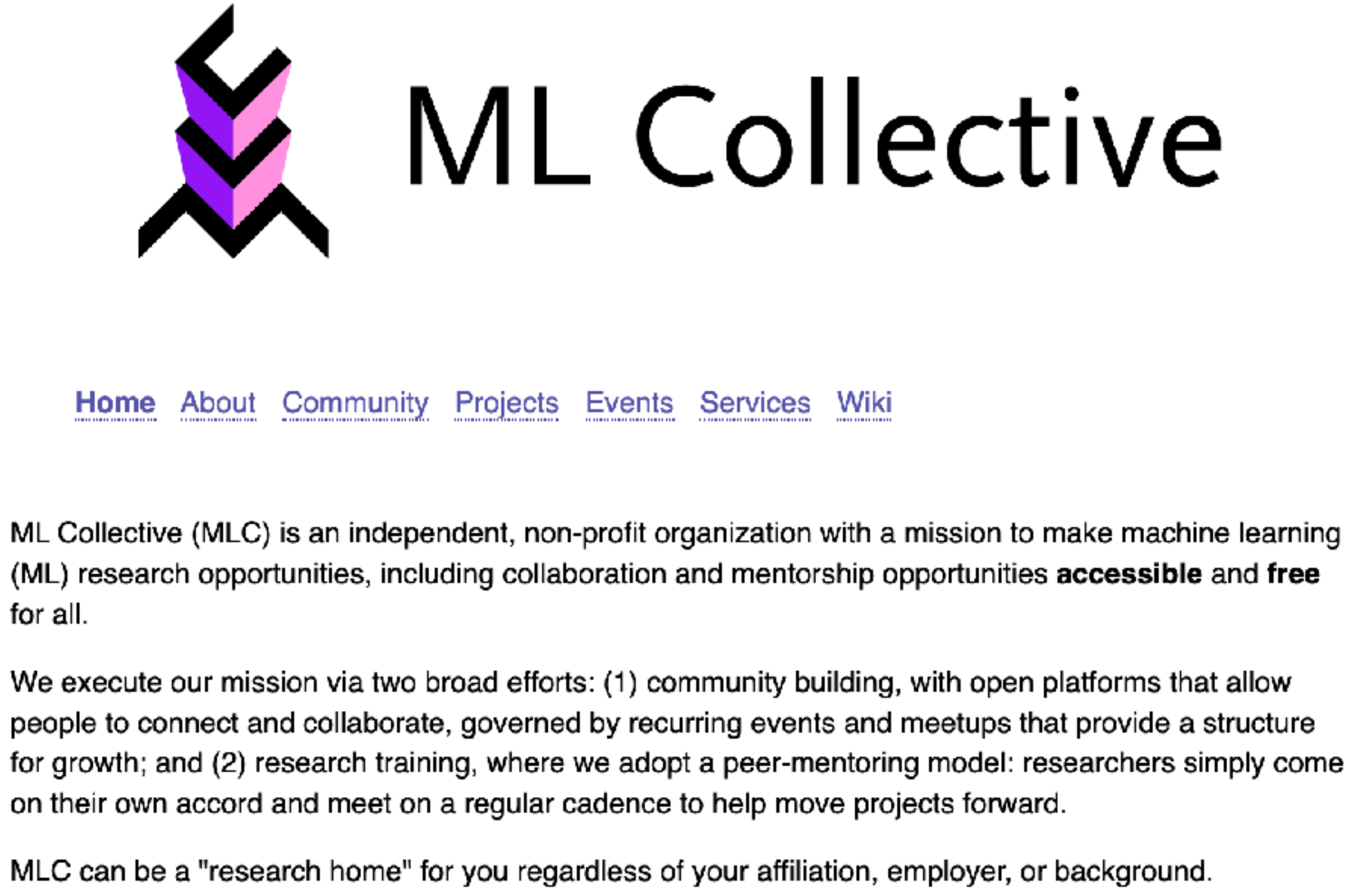
Objectives of the Uplink Initiative

Many undergraduate and master's students have a strong desire to pursue research in AI/ML/Data Science and publish papers in top ranked conferences. U access to a research ecosystem and enthusiastic and world-class mentors.

The objectives of this program are

- To provide an internship opportunity to such students under the mentorship of faculty members at top tier Indian institutes for a period of 3 months
- To reward excellence in outcome by recognizing top quality publications arising from this internship

Deadline: March 1st, 2024



ML Collective

Home About Community Projects Events Services Wiki

ML Collective (MLC) is an independent, non-profit organization with a mission to make machine learning (ML) research opportunities, including collaboration and mentorship opportunities **accessible** and **free** for all.

We execute our mission via two broad efforts: (1) community building, with open platforms that allow people to connect and collaborate, governed by recurring events and meetups that provide a structure for growth; and (2) research training, where we adopt a peer-mentoring model: researchers simply come on their own accord and meet on a regular cadence to help move projects forward.

MLC can be a "research home" for you regardless of your affiliation, employer, or background.

Screenshots from: https://ikdd.acm.org/uplink.php and https://mlcollective.org/