IEEE COMPUTER SOCIETY

KUNGFU.AI

AI in 2022: A Look at the Power, Potential, and Perils

Dr. Steve Kramer
Chief Scientist, KUNGFU.AI

IEEE Computer Society - Phoenix Chapter
June 2022

Agenda

1. Intro
2. Terminology & Why Now
3. Fundamentals of AI
4. Power & Potential of AI
5. Perils of AI
6. Resources + Q&A

About KUNGFU.AI

KUNGFU.AI provides artificial intelligence strategy and engineering services. We help our clients transform and solve hard problems through custom-built machine learning solutions.

- *Deep background in natural language processing, multispectral computer vision, time series forecasting, and anomaly detection*
- *Frameworks and practices for ethical development of artificial intelligence*

CLIENTS:

BlueCross BlueShield
ABInBev
NAVAL AIR SYSTEMS COMMAND
DARPA
WM
KELLERWILLIAMS
DEFENSE INNOVATION UNIT

PREVIOUS WORK EXPERIENCE:

SPACEX
aws
Expedia
NASA

AI SERVICES

Strategy

We help companies build a roadmap of advanced data capabilities that are most practical to the business.

Engineering

We build custom capabilities in the fields of NLP, Computer Vision, and Predictive Modeling.

Transformation

We help companies become AI self-sufficient by building out DataOps, staffing, and technology foundations

OUR TEAM

We are a group of former SaaS entrepreneurs, venture capitalists, PhD machine learning, and software engineers who are most interested in solving big business problems with the latest in artificial intelligence advances to drive business value.

Team Stats:

- Located in Downtown Austin, TX
- 40 total team members
- Started 10 total SaaS and services companies
- 6 PhDs and 13 total advanced degrees in business, computer science, physics, and statistics
- Over 12 years applying ML in Biotech, Ecommerce, Real Estate, Telco, DoD, IT, CPG, Media, FSS, Energy

CAPABILITIES

- NLP
- Predictive Analytics
- Computer Vision
- Time Series
- Genetic Algorithms
- GANs
- Embeddings
- Robotic Process Automation
- Unsupervised Learning
- Graph Theory
- ML Ops
- Python
- API Development
- Research
- Search
- Recommender Systems
- Dialog Systems

Speaker Background

- Native of Los Alamos, New Mexico
- Ph.D. in computational physics (nonlinear dynamics and chaos theory) in 1993
- 29 years of post-Ph.D. research and high-tech experience
- 13 years as solo data science entrepreneur at Paragon Science
- Principal Investigator on multiple DARPA and DIU contracts
- Reviewer for scientific journals and conferences in intelligence and security informatics since 2011

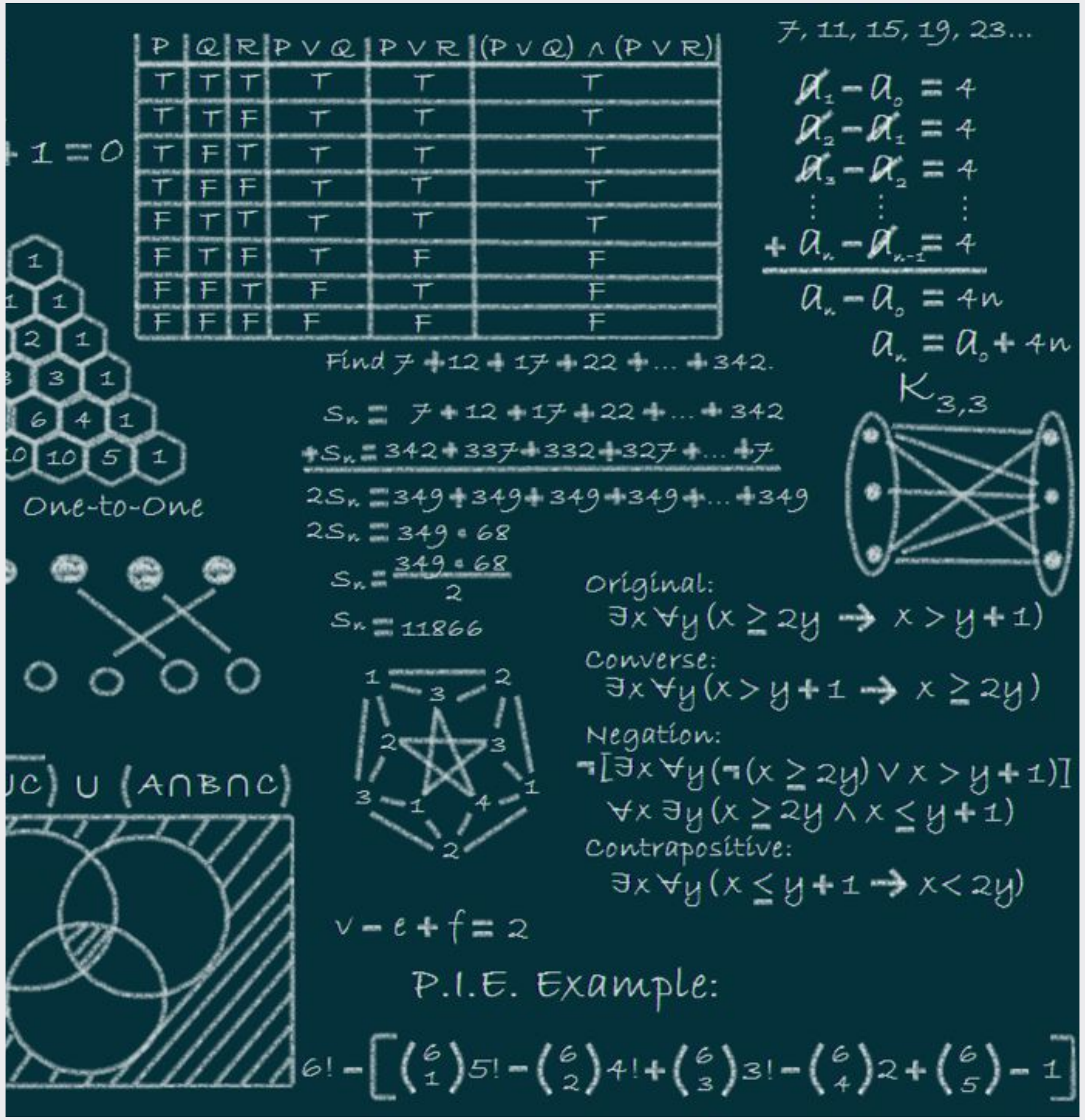
Terminology

Artificial Intelligence

Systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages

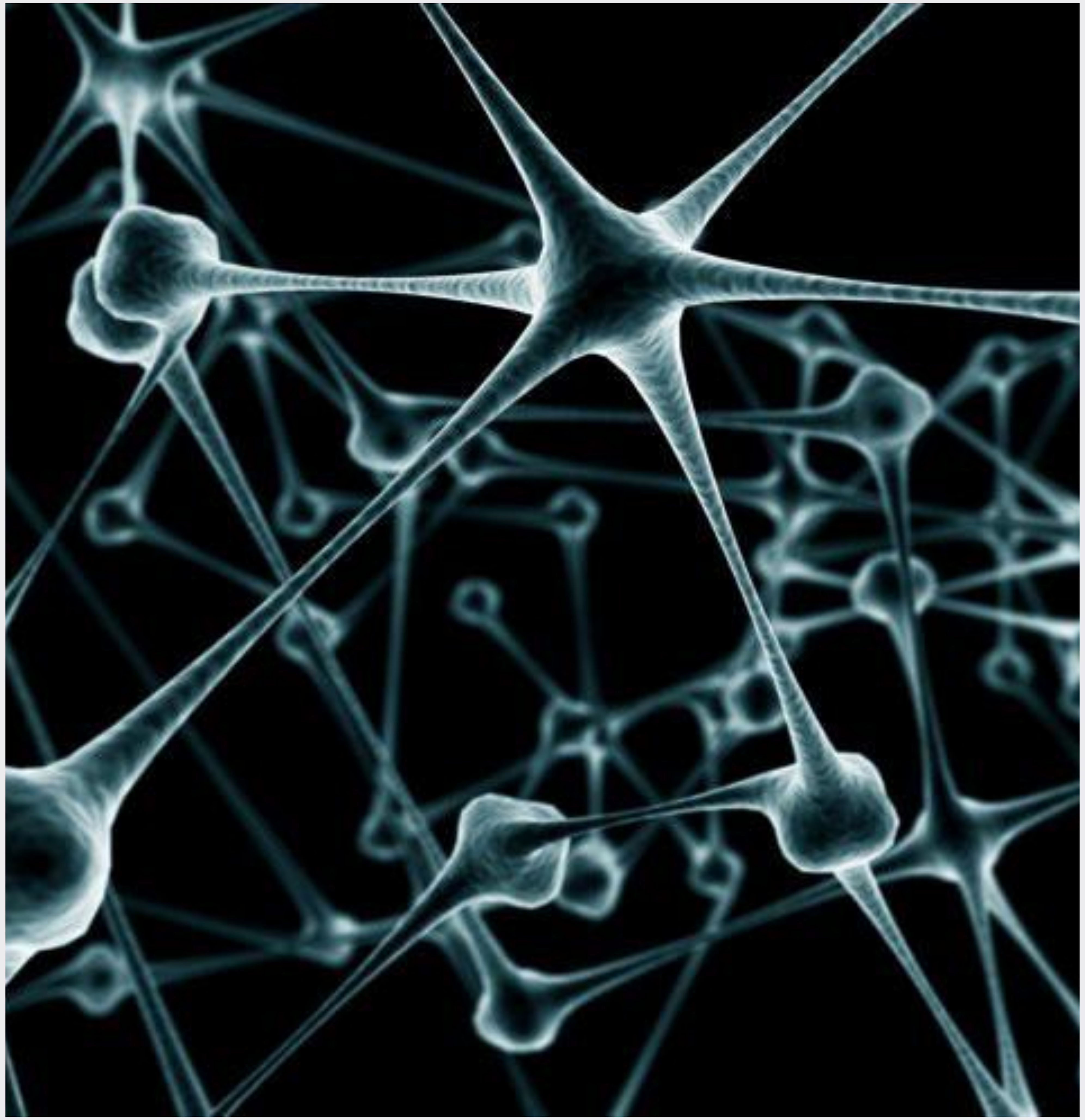
Machine Learning

A subset of Artificial Intelligence that involves algorithms capable of improving their performance when given more data



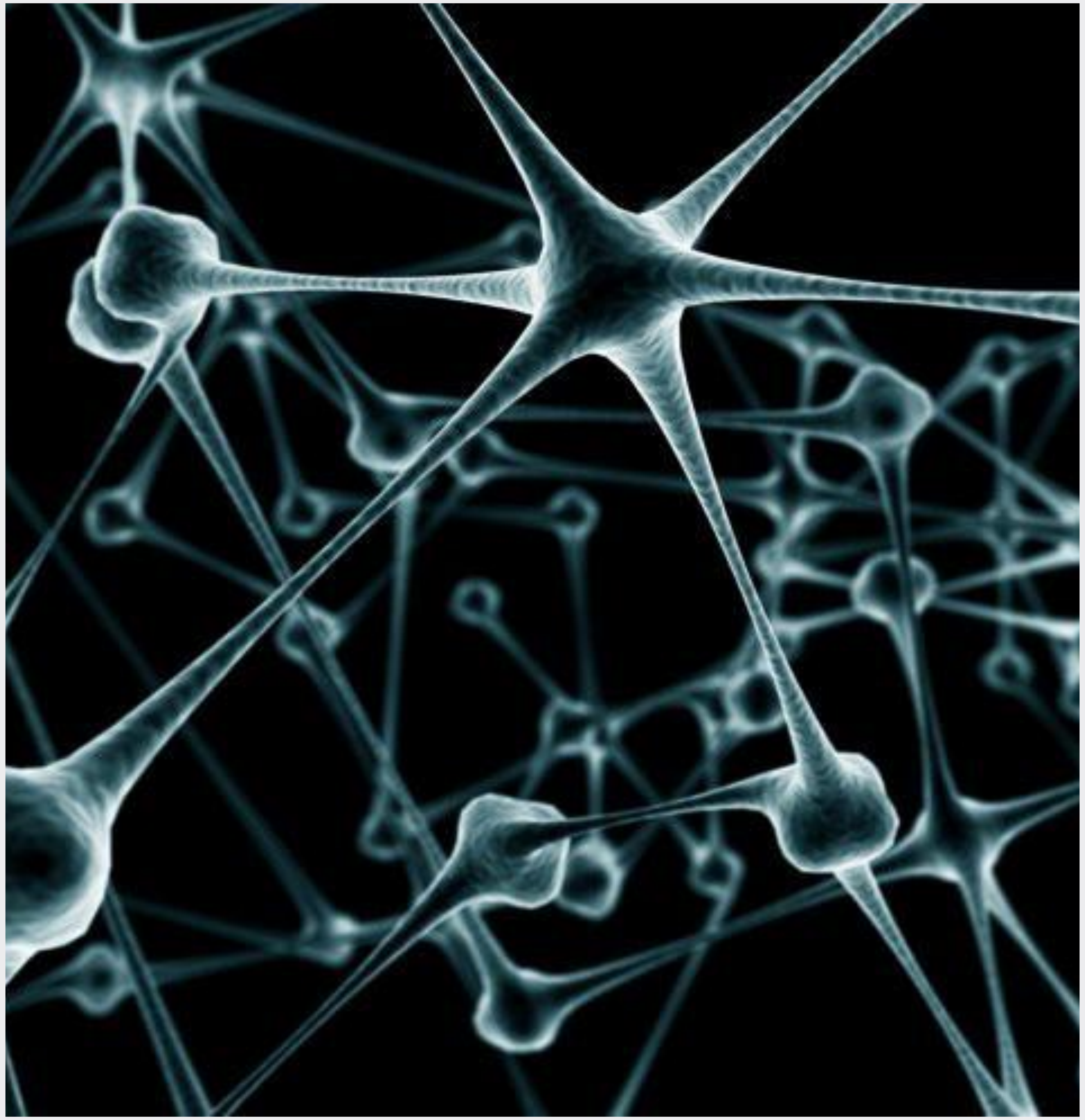
Deep Learning

A subset of machine learning that uses multi-layered artificial neural networks to learn from vast amounts of data



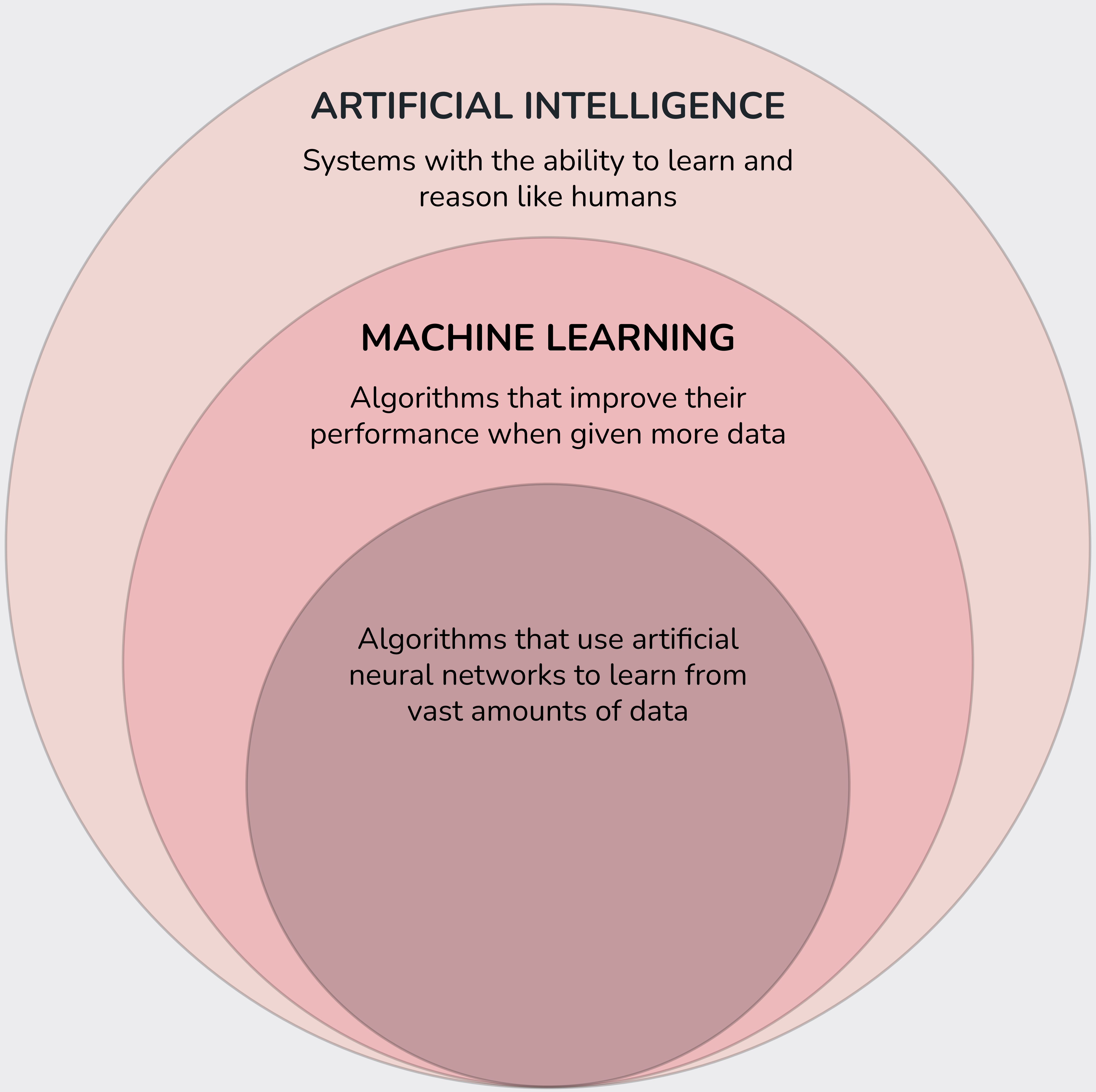
Deep Learning

A subset of machine learning that uses multi-layered artificial neural networks to learn from vast amounts of data



AI Hierarchy

- All machine learning is AI
- Not all AI is machine learning



Overview of Classical Machine Learning

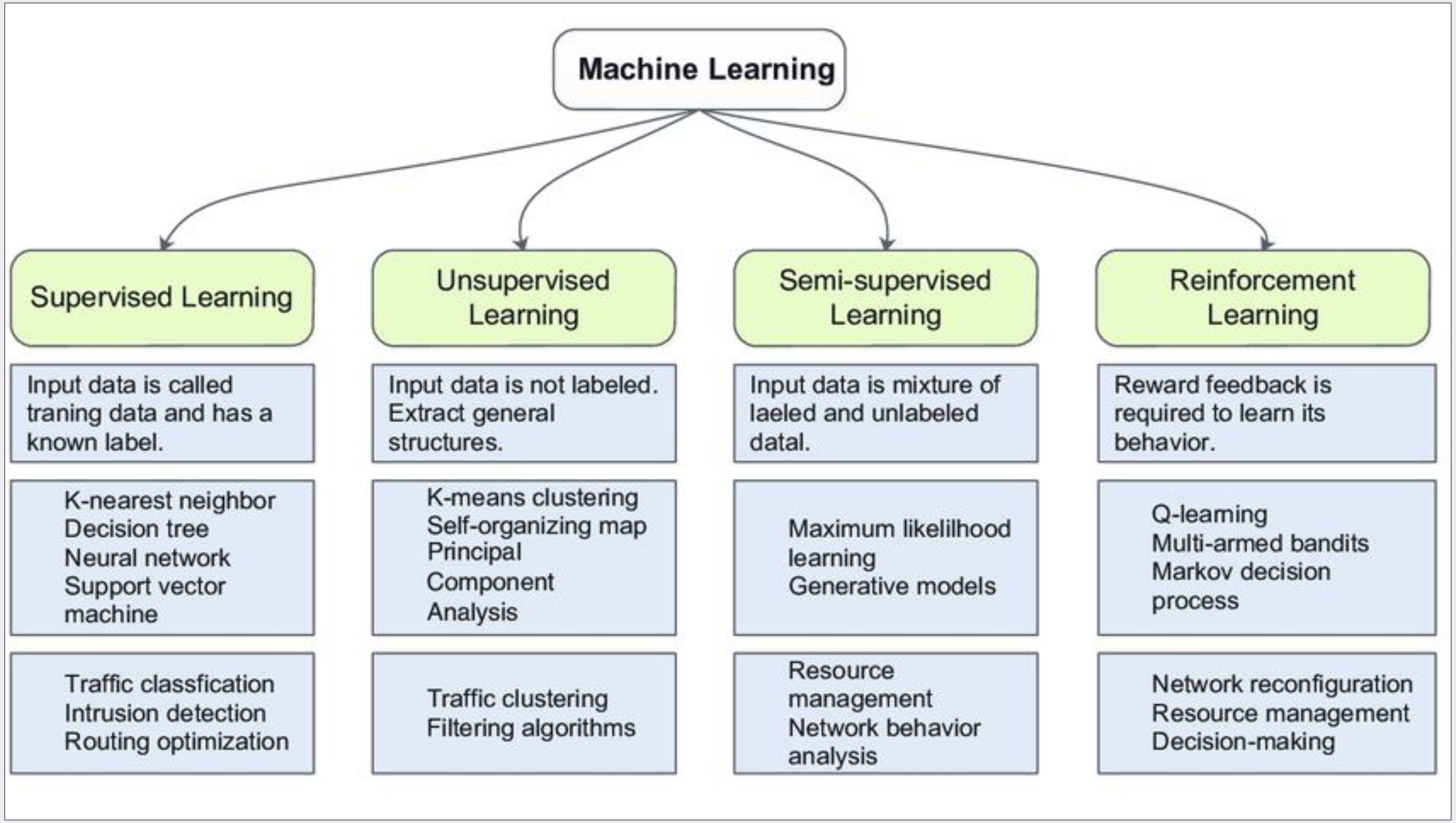
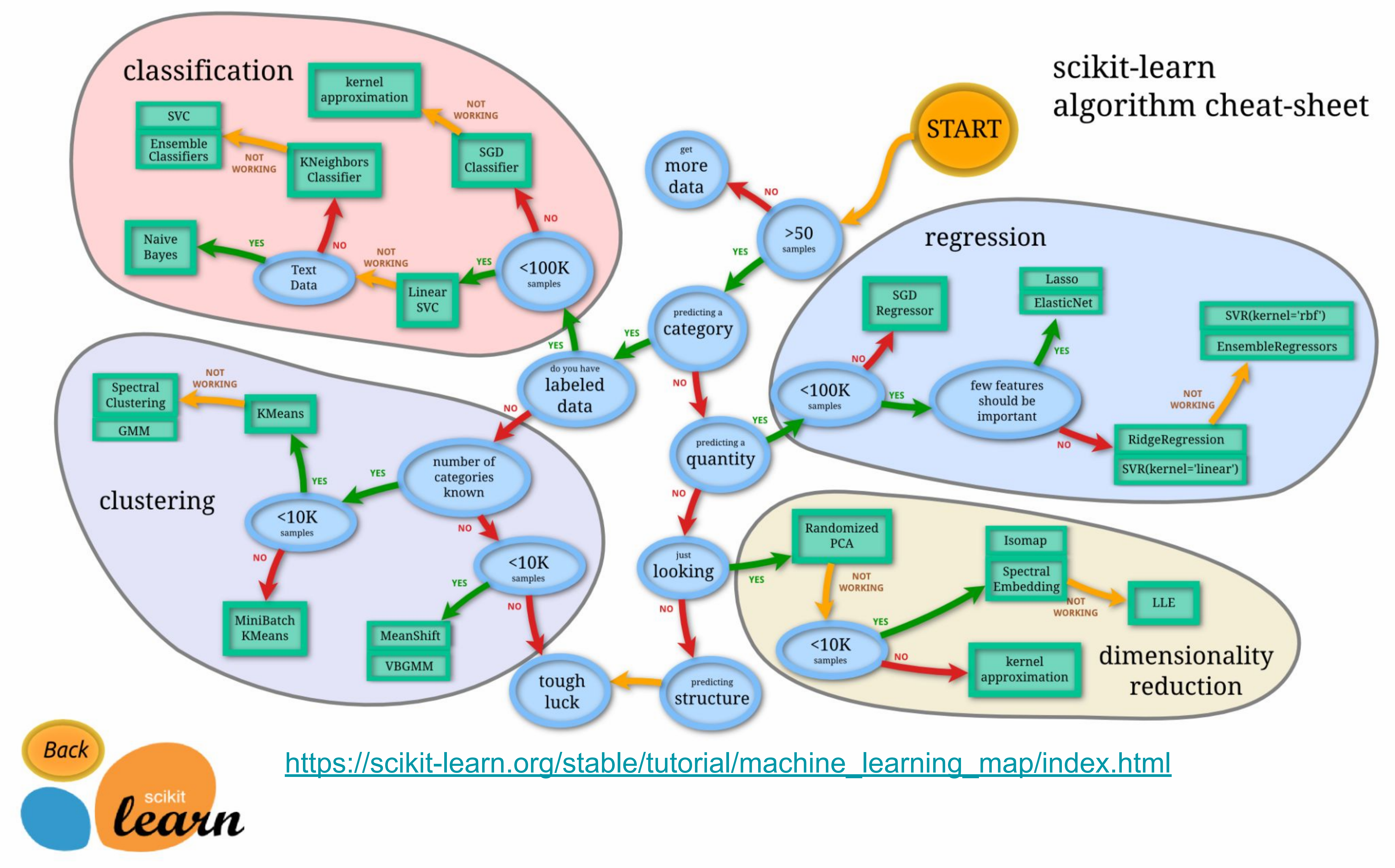


Fig: Liu, Yiming, et al. "Blockchain and machine learning for communications and networking systems." IEEE Communications Surveys & Tutorials 22.2 (2020): 1392-1431.

Classical Machine Learning Tasks



https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

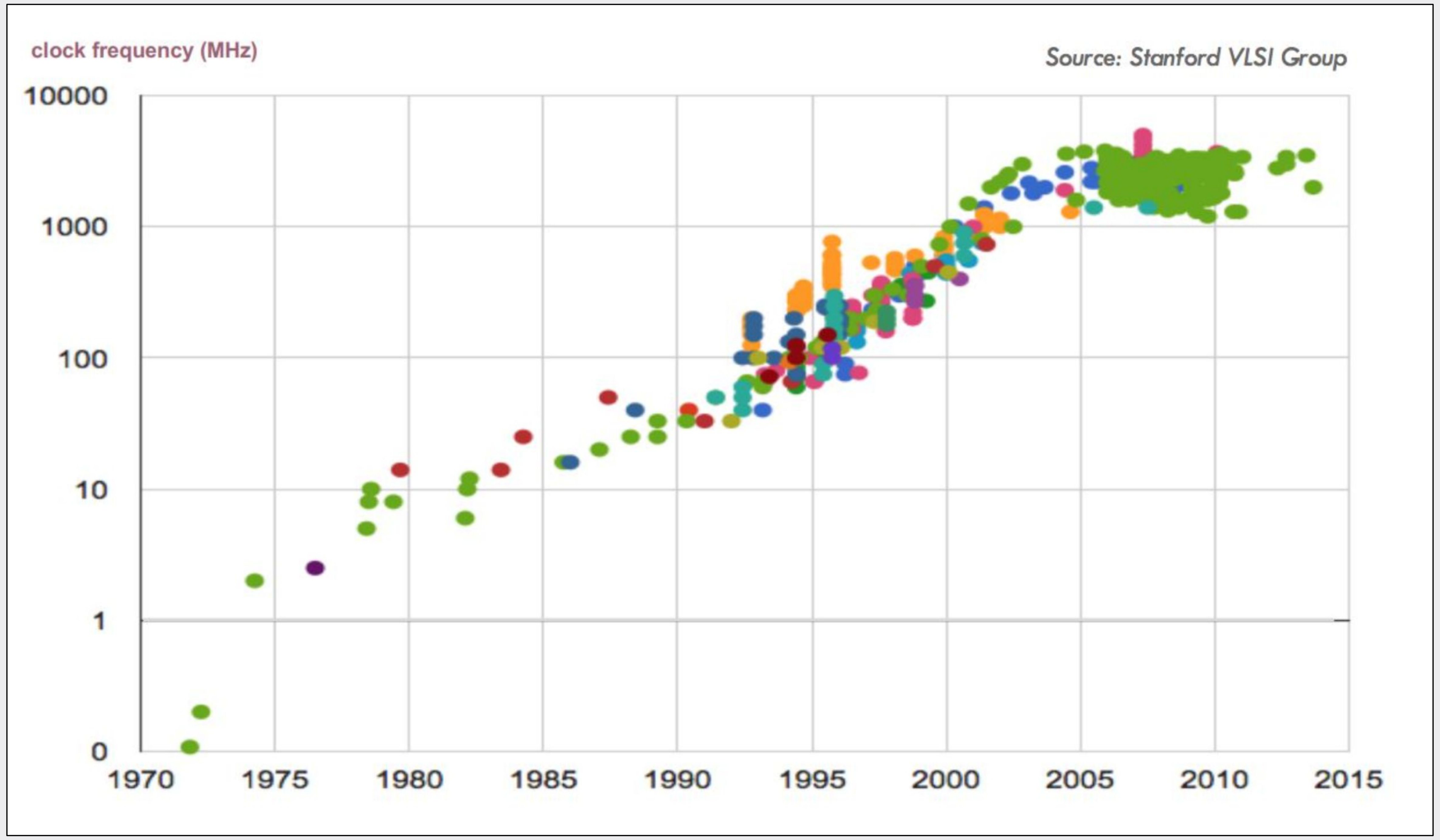
Fig: Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

Why Now?

1. Computation/Hardware

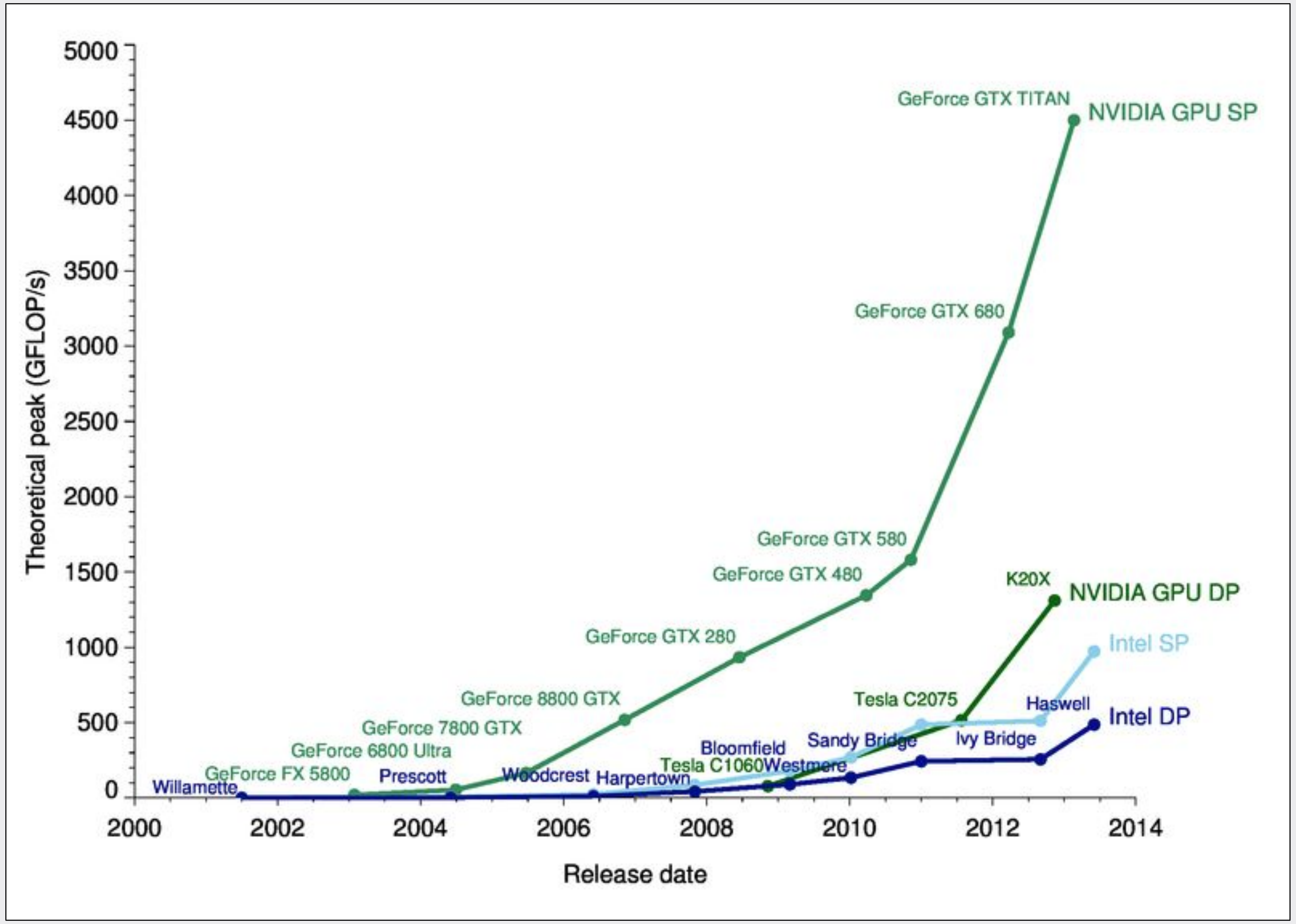
CPU Bottleneck

- CPU performance plateaued
- Clock speeds have experienced minimal increases since 2005
- As transistors shrink, the power required to run them increases



GPU Parallelism

- Graphical Processing Units (GPUs) provide immense computational parallelism
  - Ideal for matrix operations – the heart of AI algorithms
  - 4,000+ cores per chip
  - Workhorse of current AI modeling



Hardware in Perspective

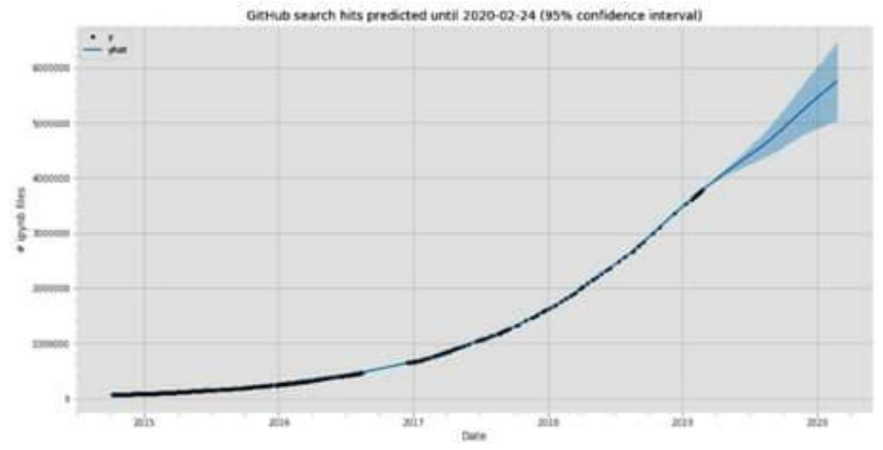
An **emerging trend** disrupts the past 15-20 years of software engineering practice:

hardware > software > process

Hardware is now evolving more rapidly than software, which is evolving more rapidly than effective process

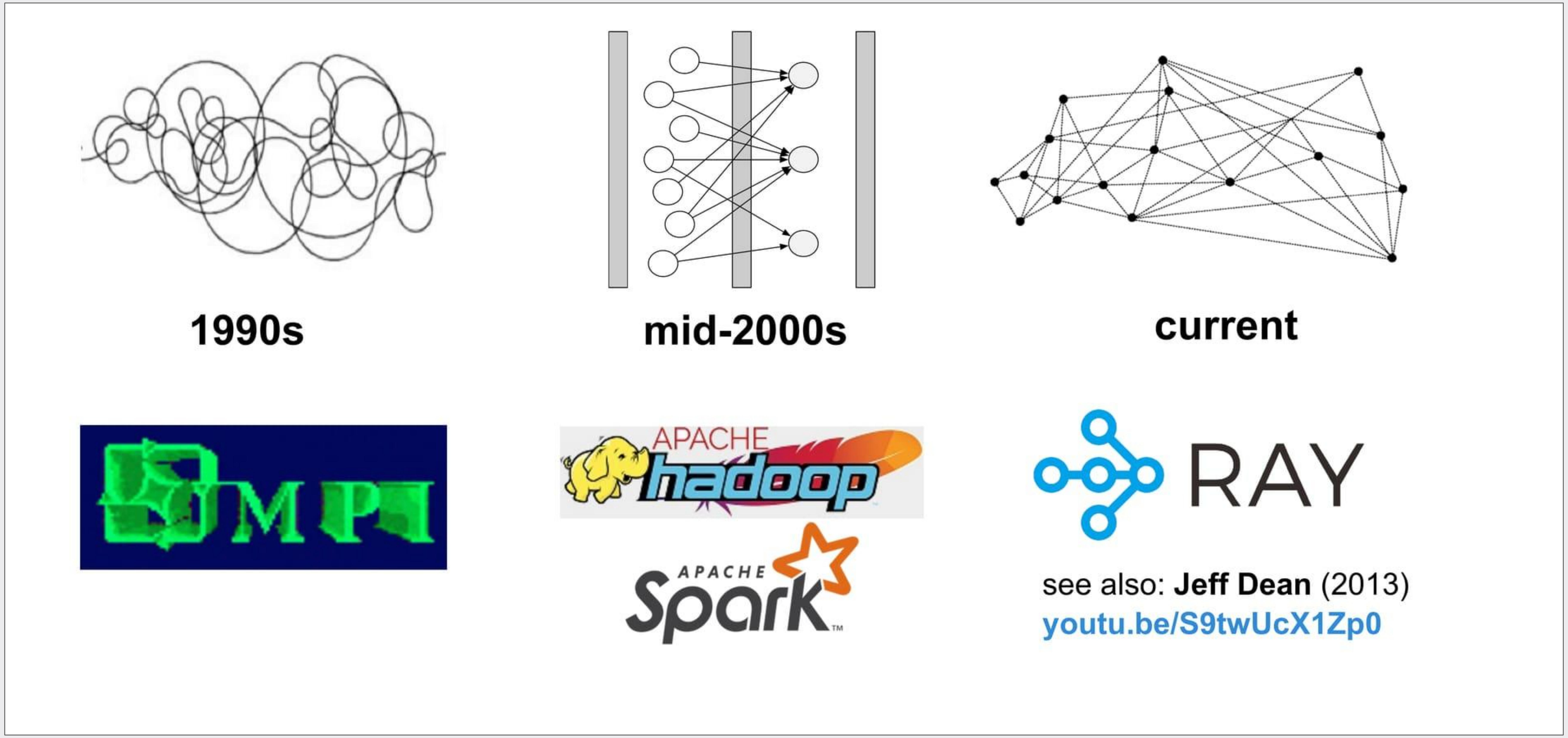
Moore's Law is all but dead, although ironically many inefficiencies grew to be based on it

Project Jupyter, **Apache Arrow**, **NumPyWren** and the related **Ray** are emblematic for data infrastructure transformation in enterprise



Credit: Paco Nathan, "Perspectives on the State of AI" (2020)
https://derwen.ai/s/gw6q#43

Cluster Topologies by Generation

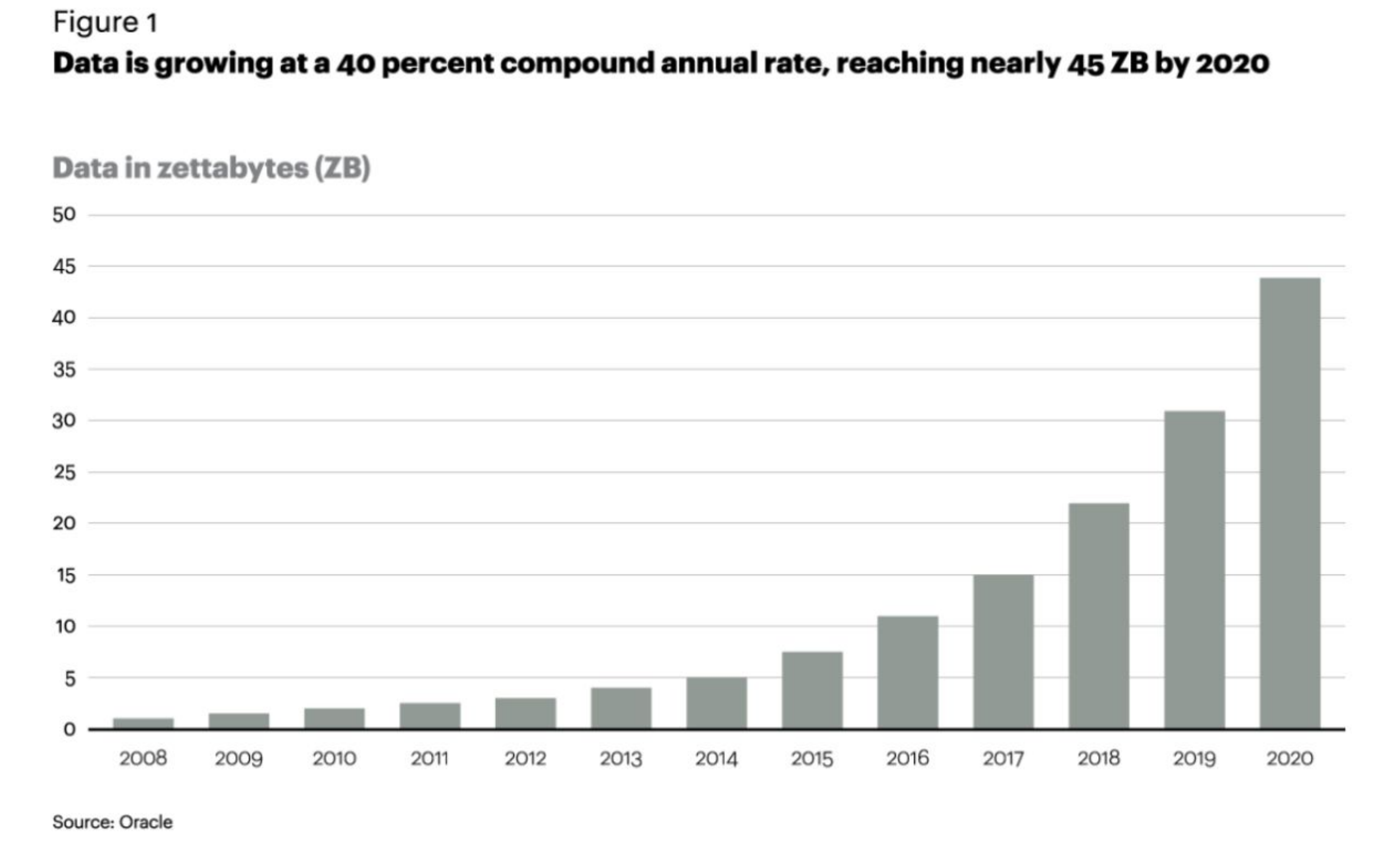


Credit: Paco Nathan, “Perspectives on the State of AI” (2020)
https://derwen.ai/s/gw6q#51

2. Data

Data Growth

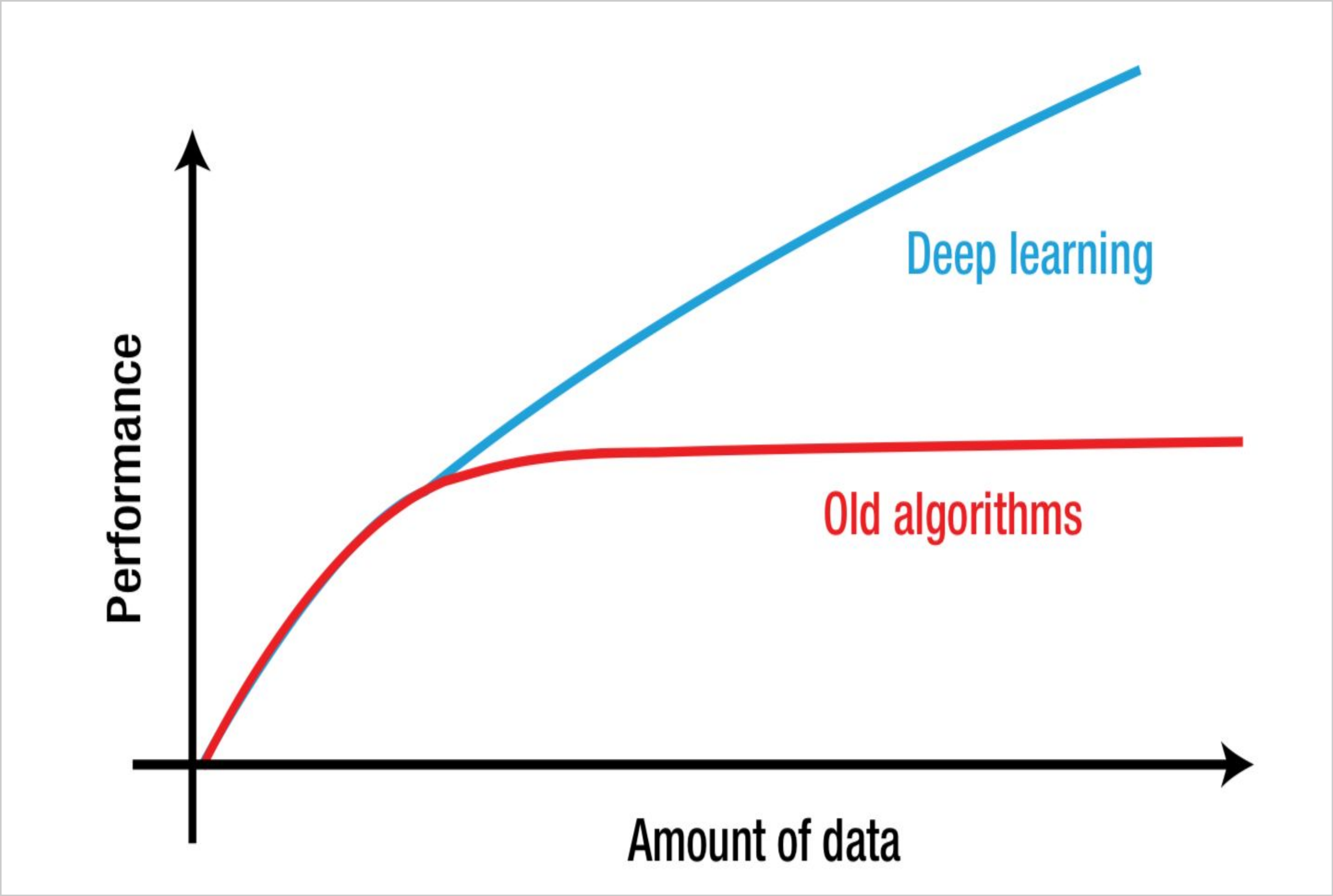
- Approximately 90% of the world's data has been produced in the past two years.
- Electronic-device users generate 2.5 quintillion bytes of data per day.
- Worldwide IP traffic exceeded 20 exabytes (20 billion gigabytes) per month in 2020.



The growth of data in the past 10 years has been enormous, much less the growth since the dawn of BI (source)

Data = The New Oil

- A key feature of AI algorithms is their ability to learn from large amounts of data.
- Most features, if not all, can be learned automatically from the data— provided that enough training data examples are available (sometimes millions).



3. Open-Source Software

Open-Source Deep Learning Software

- Google, Facebook, Microsoft and others have contributed significantly to open source machine learning libraries.
  - Flexible architectures with easy deployment across a variety of platforms
  - State-of-the-art performance

Popularity of Deep Learning Frameworks

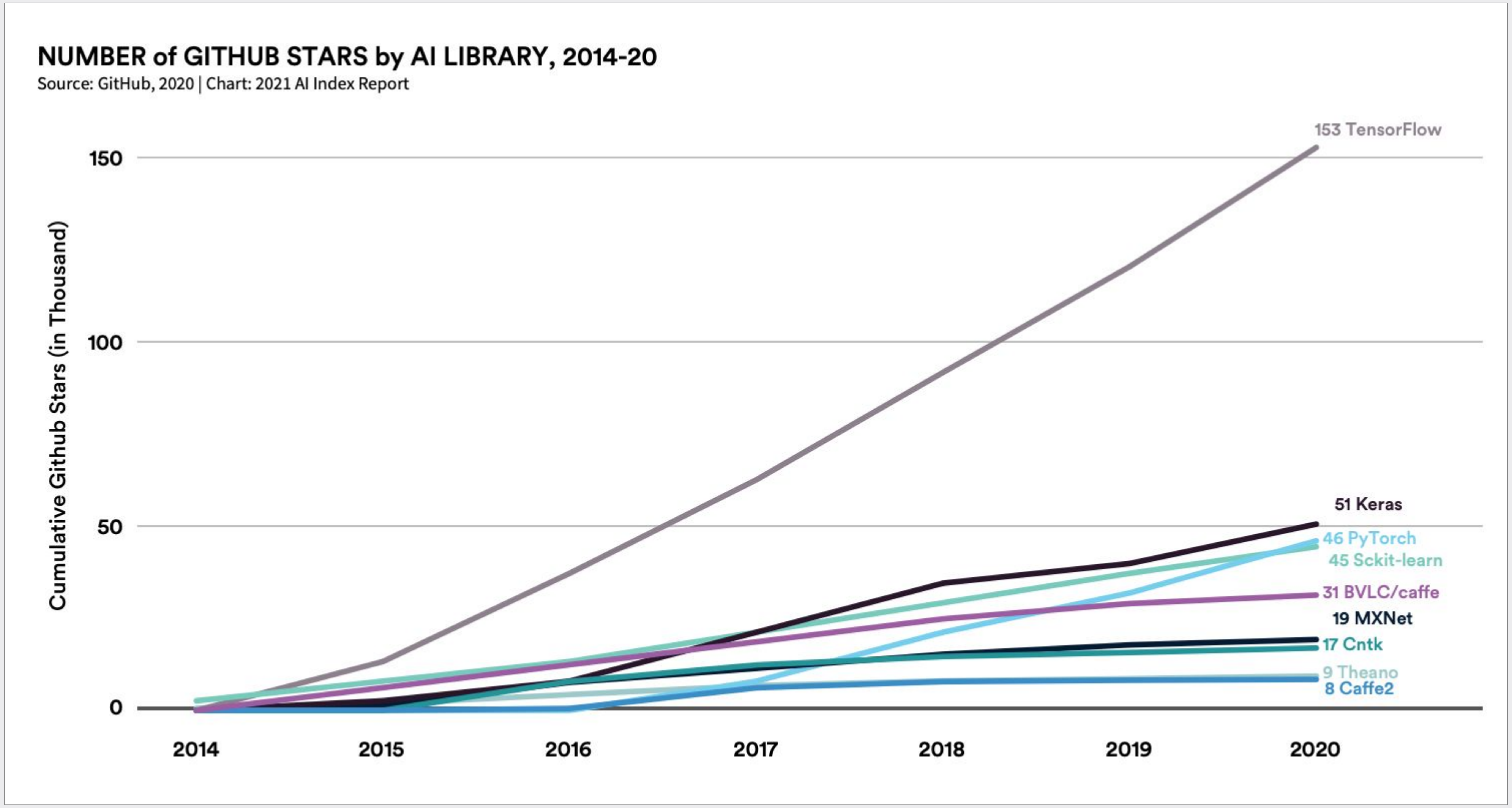
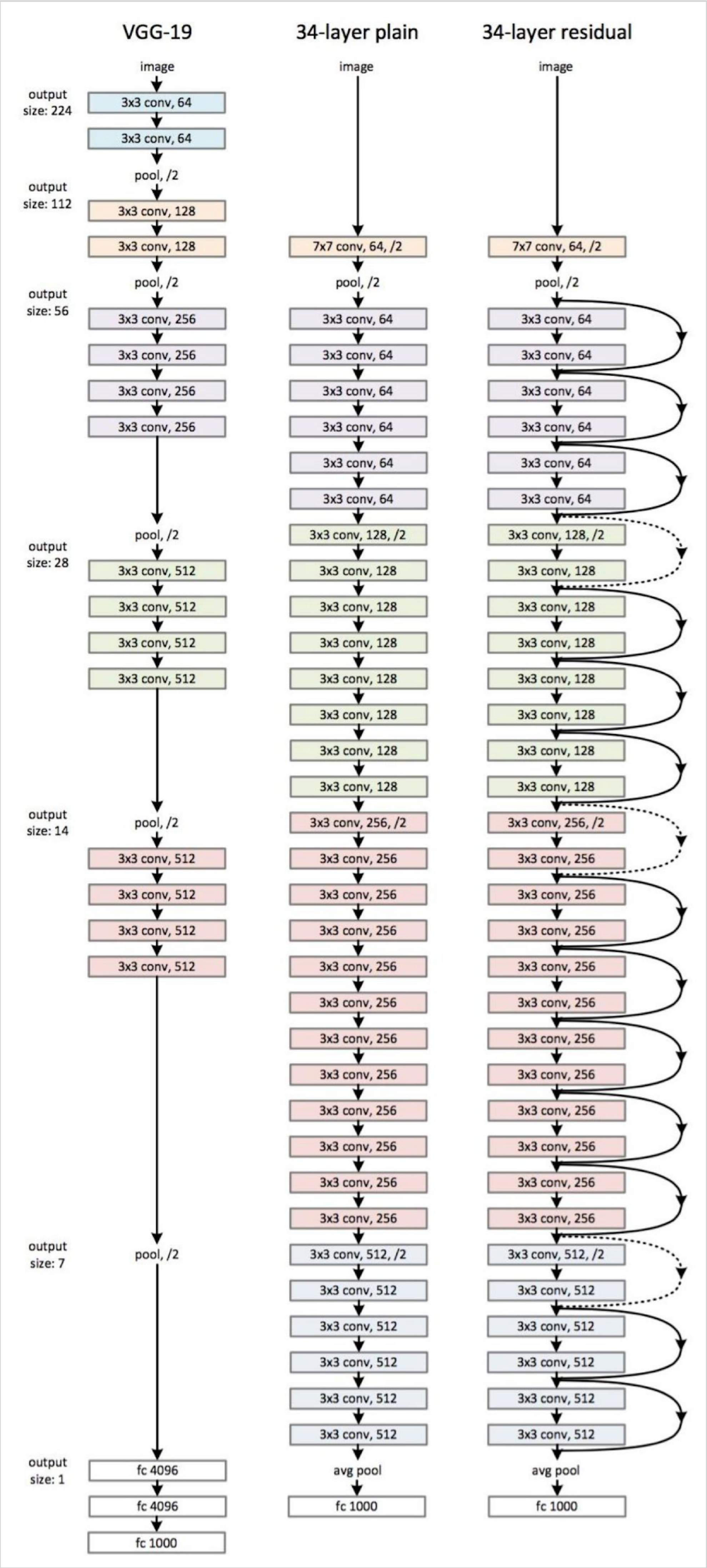


Fig: Daniel Zhang, et al., "The AI Index 2021 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021.

4. Algorithmic Advances

Deep Neural Network Learning Capacity

- Because most DNNs have billions of parameters they don’t saturate easily.
- The more data you have, the more features they can automatically learn.



Typical Deep Learning Architecture

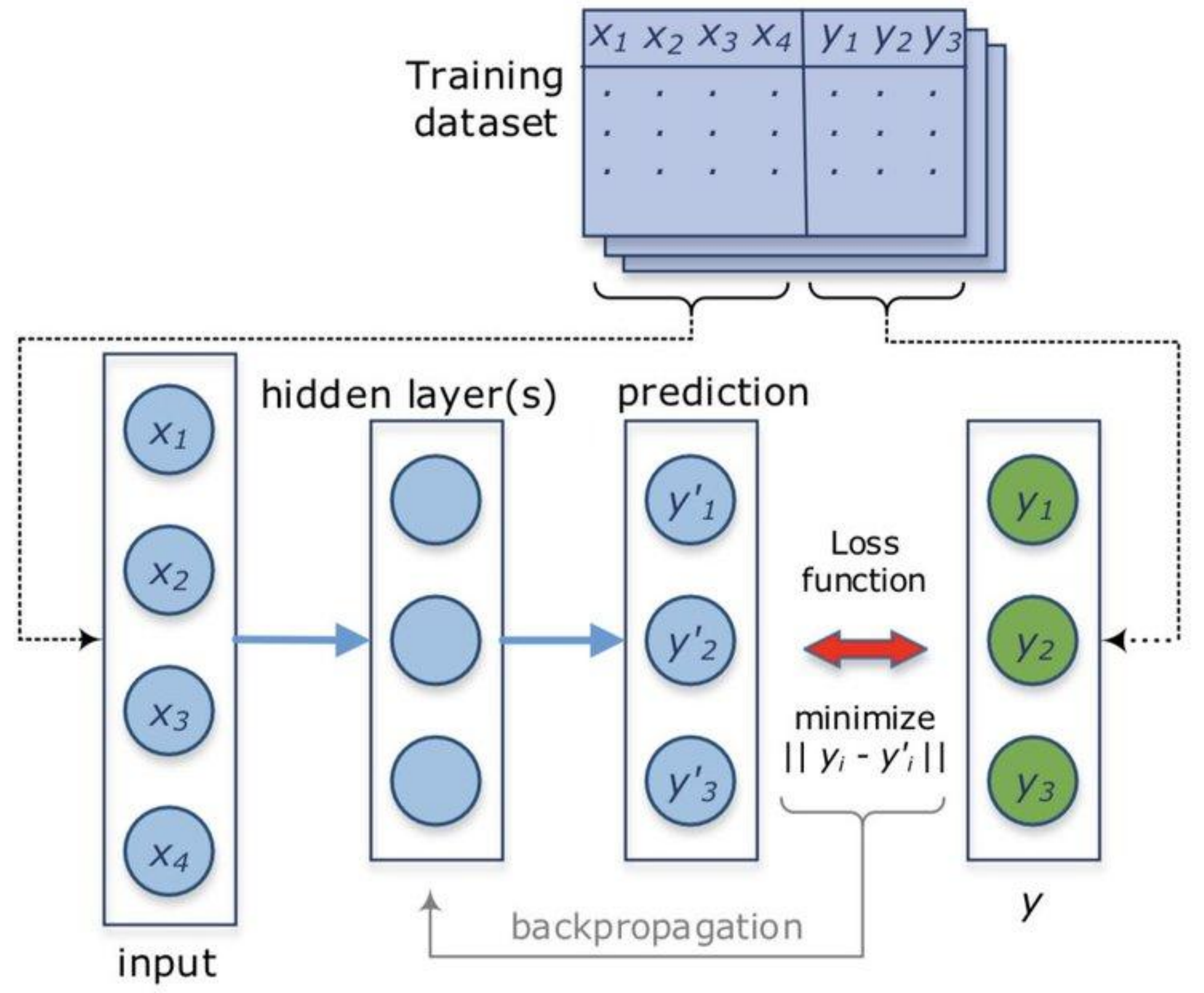


Fig. 6. The overall mechanism of training of a DL model.

Fig: Liu, Mohammadi, Mehdi, et al. "Deep learning for IoT big data and streaming analytics: A survey." IEEE Communications Surveys & Tutorials 20.4 (2018): 2923-2960.

Typical CNN Architecture

Convolutional Neural Networks (CNNs) are frequently used for computer vision problems like image classification.

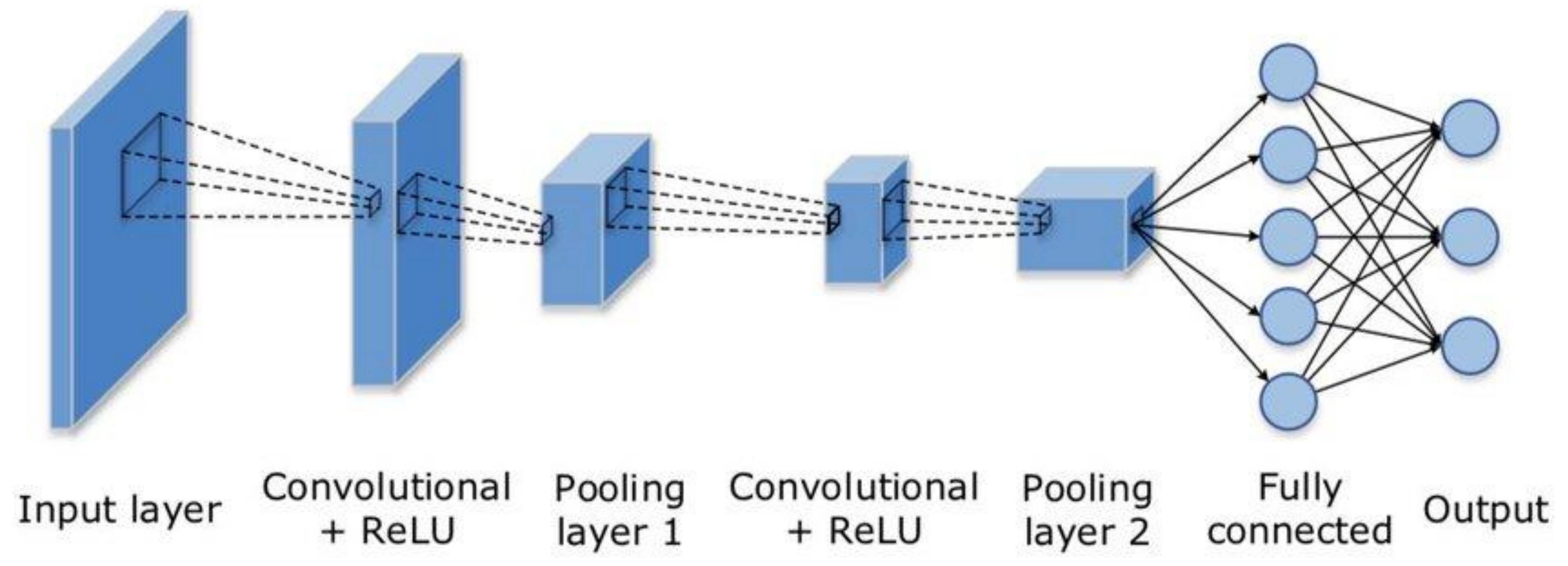
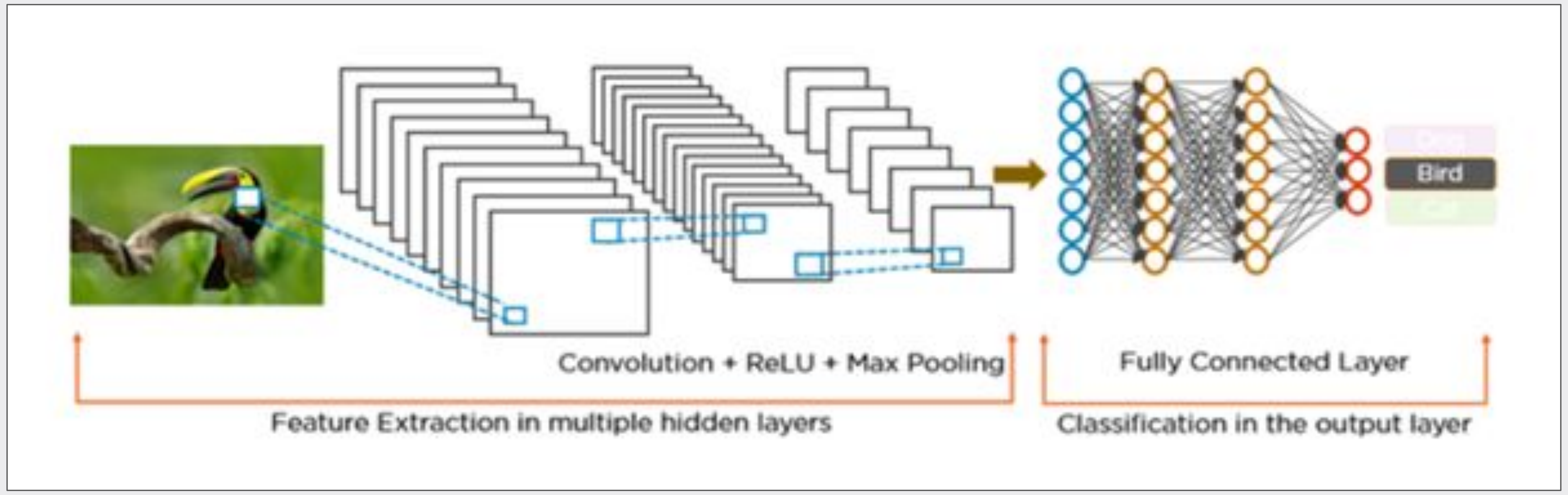


Fig. 7. Architecture of a CNN.

Fig: Liu, Mohammadi, Mehdi, et al. "Deep learning for IoT big data and streaming analytics: A survey." IEEE Communications Surveys & Tutorials 20.4 (2018): 2923-2960.

CNNs vs. RNNs

CNNs are geared towards spatial and image data.



Recurrent Neural Networks (RNNs) are geared towards temporal or sequential data.

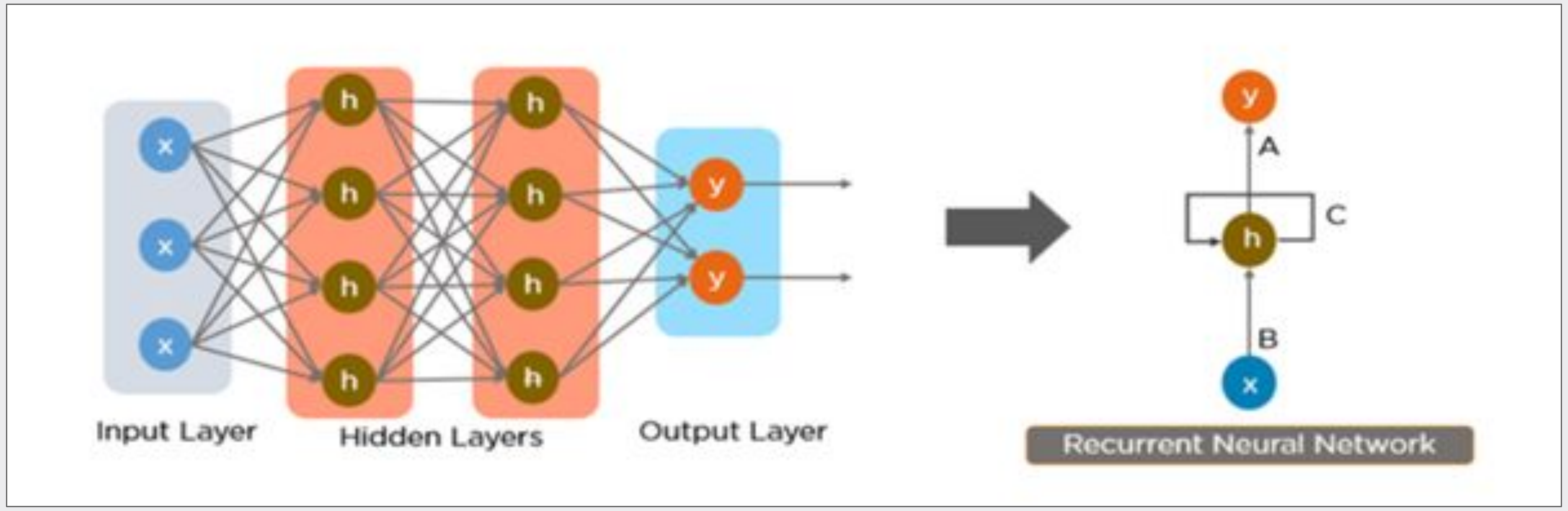


Fig: https://ashutoshtripathi.com/2021/07/12/the-main-difference-between-rnn-vs-cnn-nlp/

Human-Level Performance

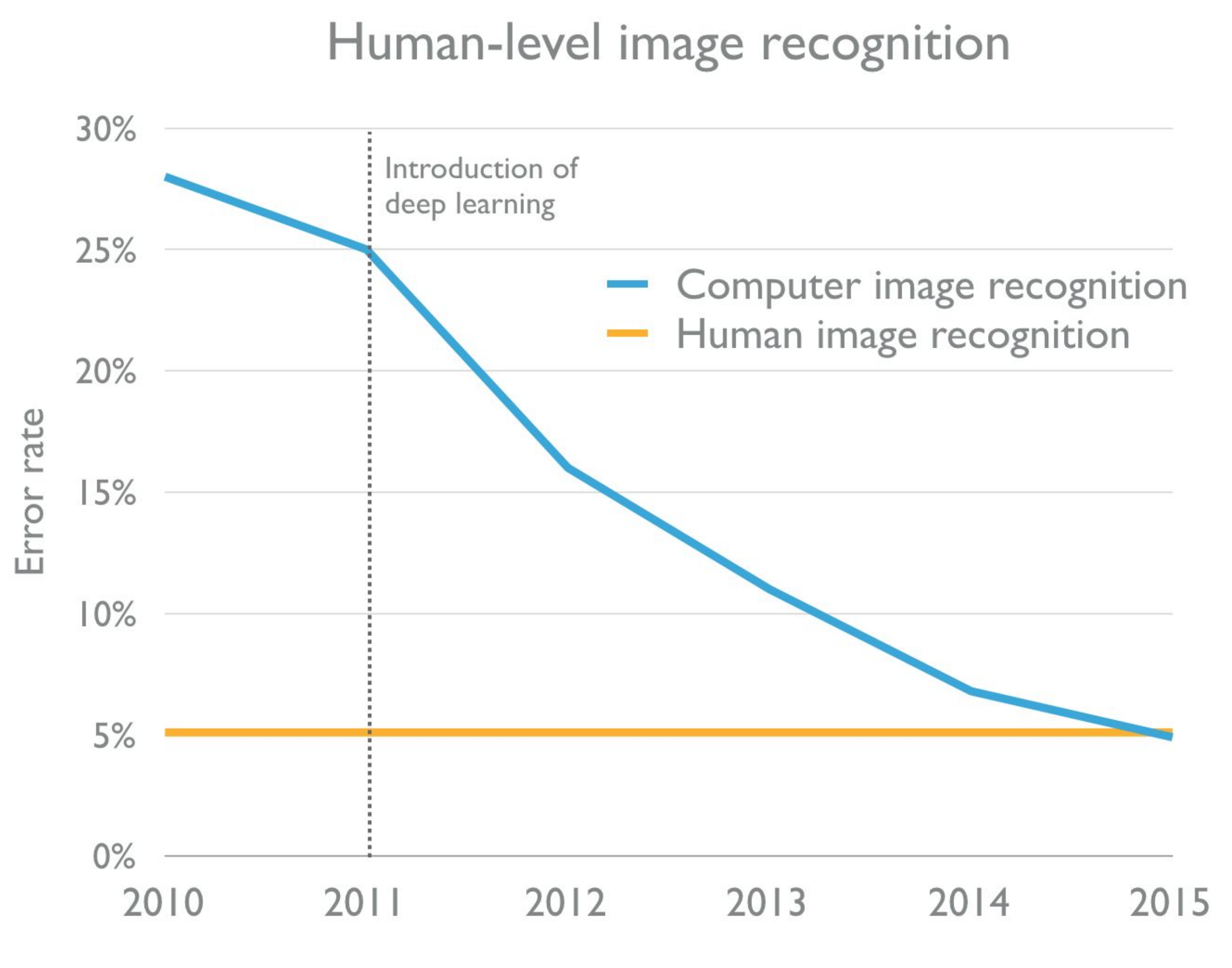


Figure 9

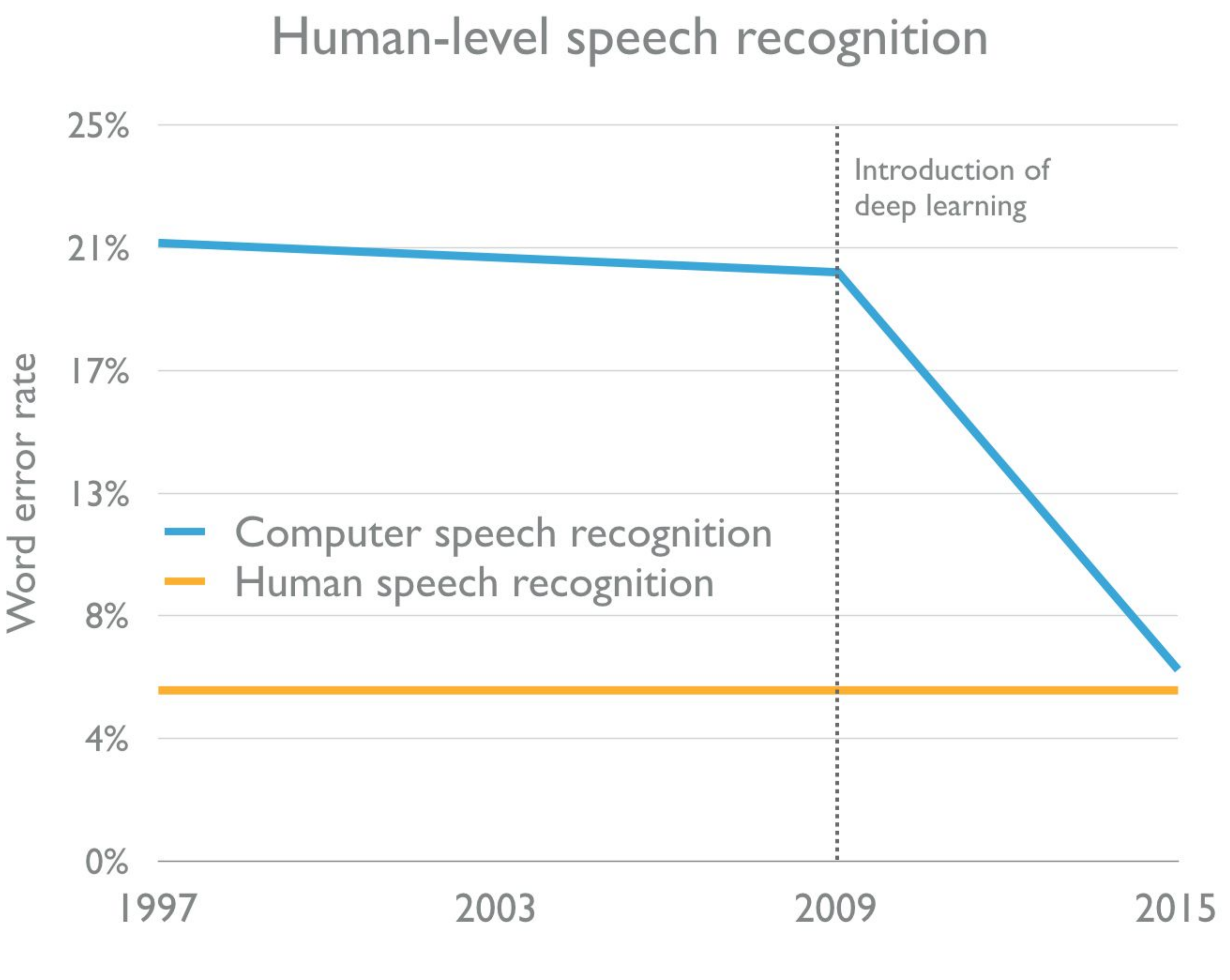


Figure 10

Transformers! *(*not the films)*

- The Transformer model uses **self-attention** to compute the relative importances of input tokens within context and using neither convolution nor recurrence
- Originally developed for NLP, this encoder/decoder architecture is now used in computer vision and other tasks

Attention is all you need
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… the number of **attention** heads and the **attention** key and value dimensions, keeping the amount of computation constant, as described in Section 3.2.2. While single-head **attention** is 0.9 …
☆ Save 99 Cite Cited by 39159 Related articles All 35 versions »

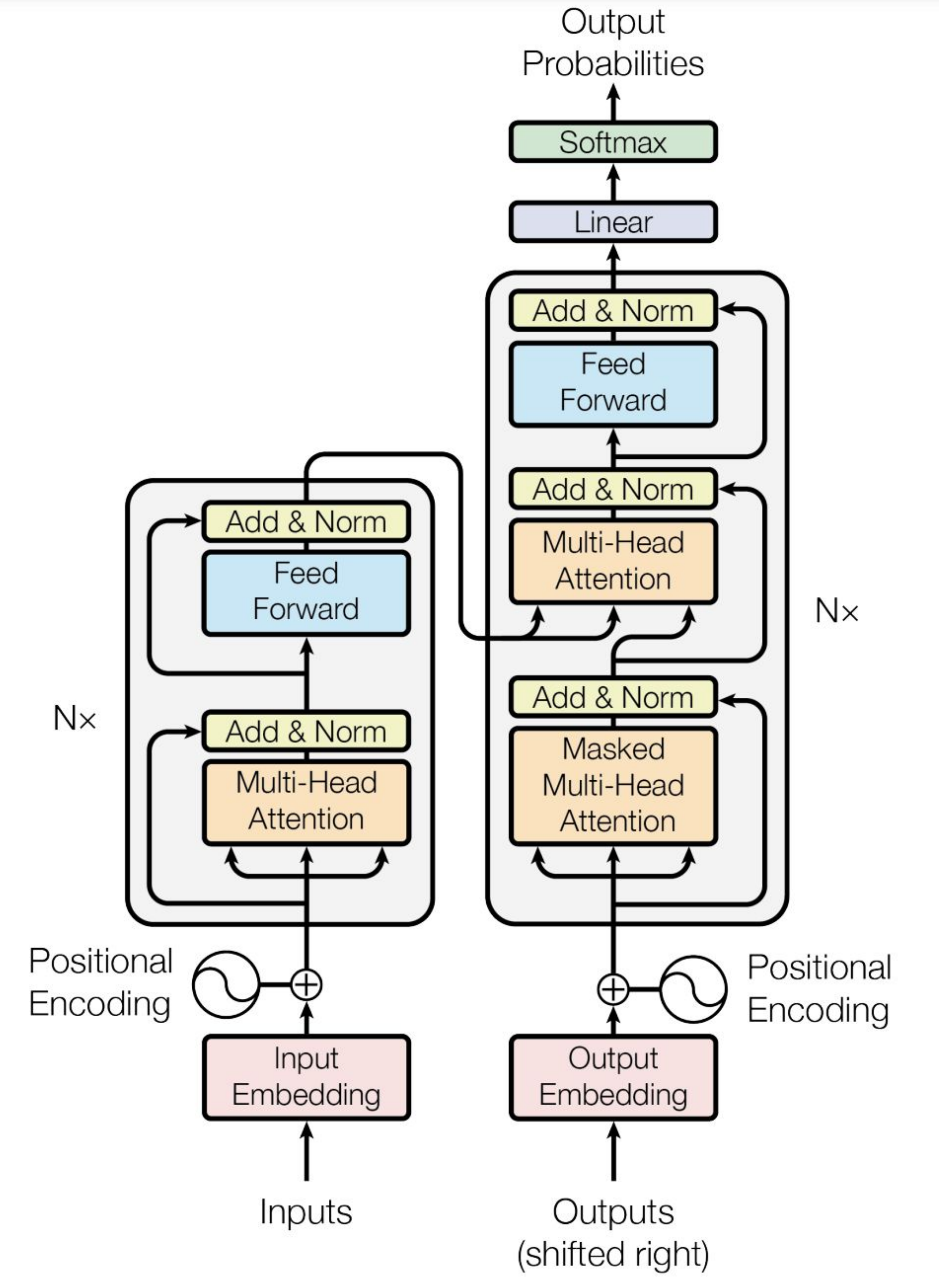
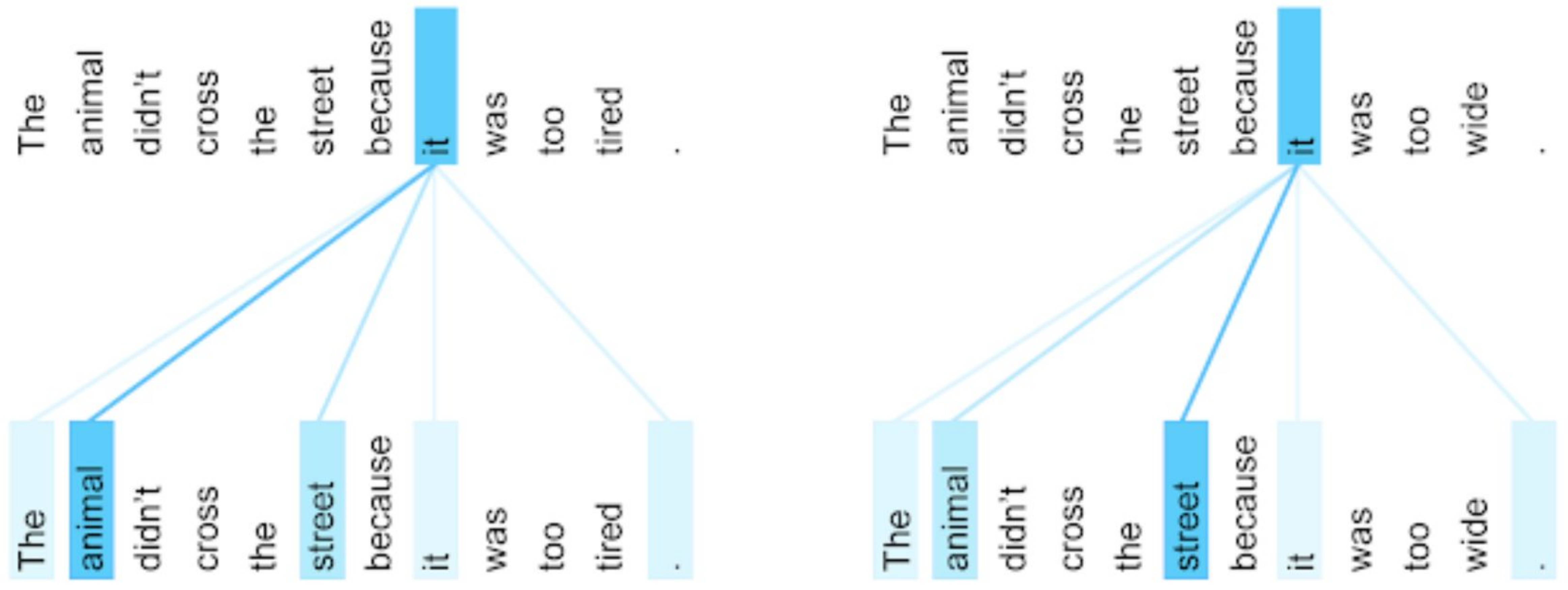


Figure 1: The Transformer - model architecture.

Fig: Vaswani, Ashish, *et al.* "Attention is all you need." Advances in neural information processing systems 30 (2017).

Machine Translation with Attention

The Transformer model “can visualize what other parts of a sentence the network attends to when processing or translating a given word, thus gaining insights into how information travels through the network.”



The encoder self-attention distribution for the word “it” from the 5th to the 6th layer of a Transformer trained on English to French translation (one of eight attention heads).

The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.

Credit: Jakob Uszkoreit, “Transformer: A Novel Neural Network Architecture for Language Understanding,” Google AI Blog (2017)

Comparison of Transformer Models

Encoder only

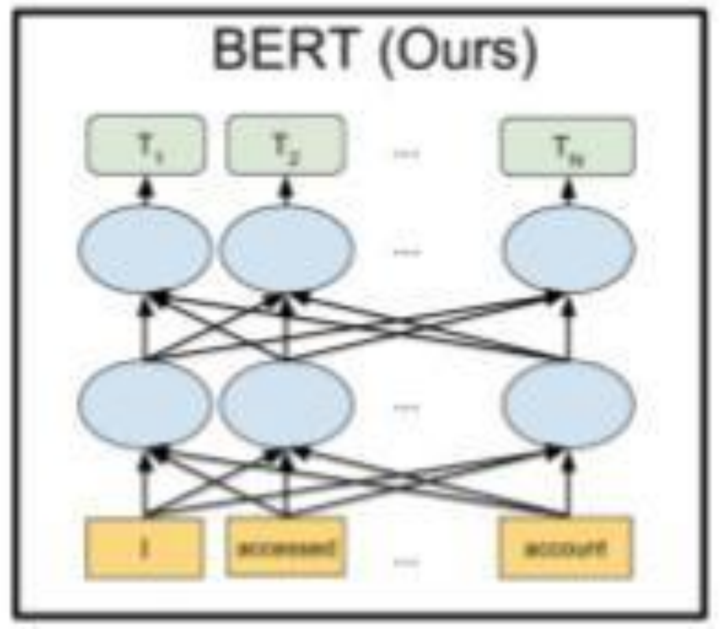
- BERT
- RoBerta
- Reformer
- FlauBERT
- CamemBERT
- Electra*
- MobileBERT
- Longformer

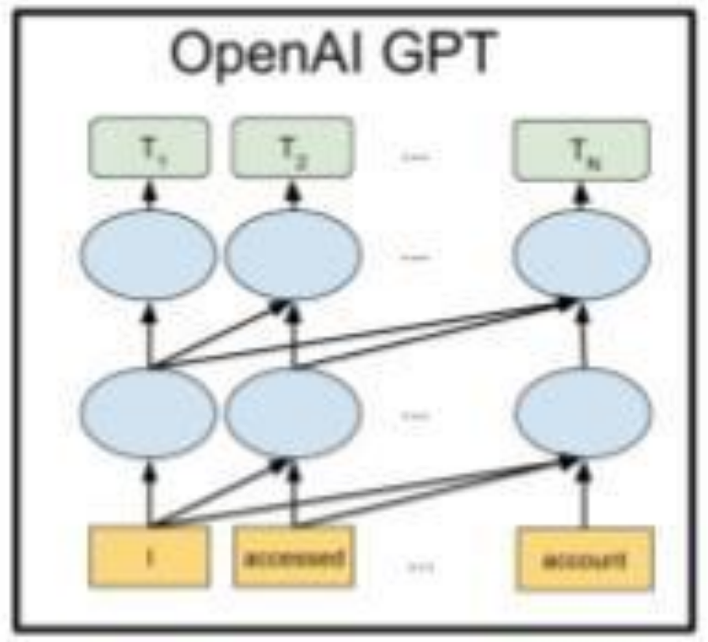
Decoder only

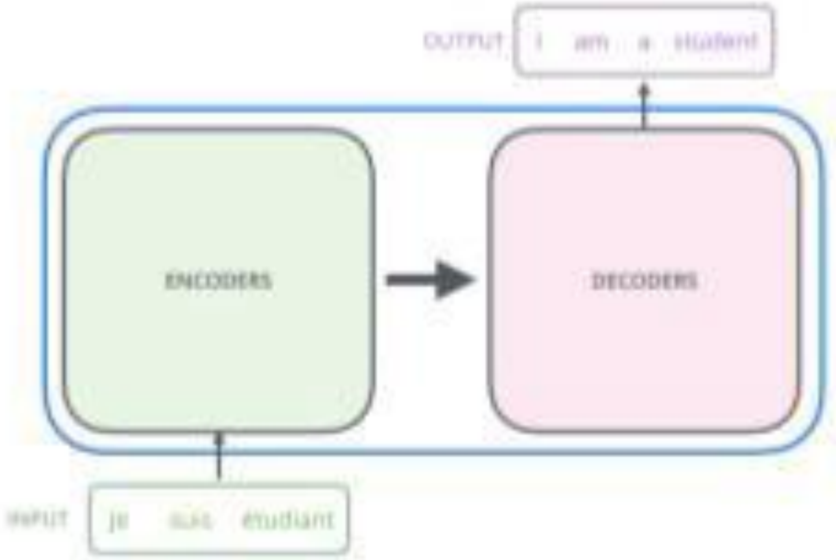
- Transformer-XL
- XLNet
- GPT series
- DialoGPT

Encoder + Decoder

- Transformer
- XLM
- T5
- BART
- XLM-RoBerta
- Pegasus
- mBART







Illustrations are from: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html and http://jalammar.github.io/illustrated-transformer/

Generative Adversarial Networks (GANs)

"Generative adversarial networks (GANs) are algorithmic architectures that use two neural networks, pitting one against the other (thus the "adversarial") in order to generate new, synthetic instances of data that can pass for real data. They are used widely in **image generation, video generation** and **voice generation.**"

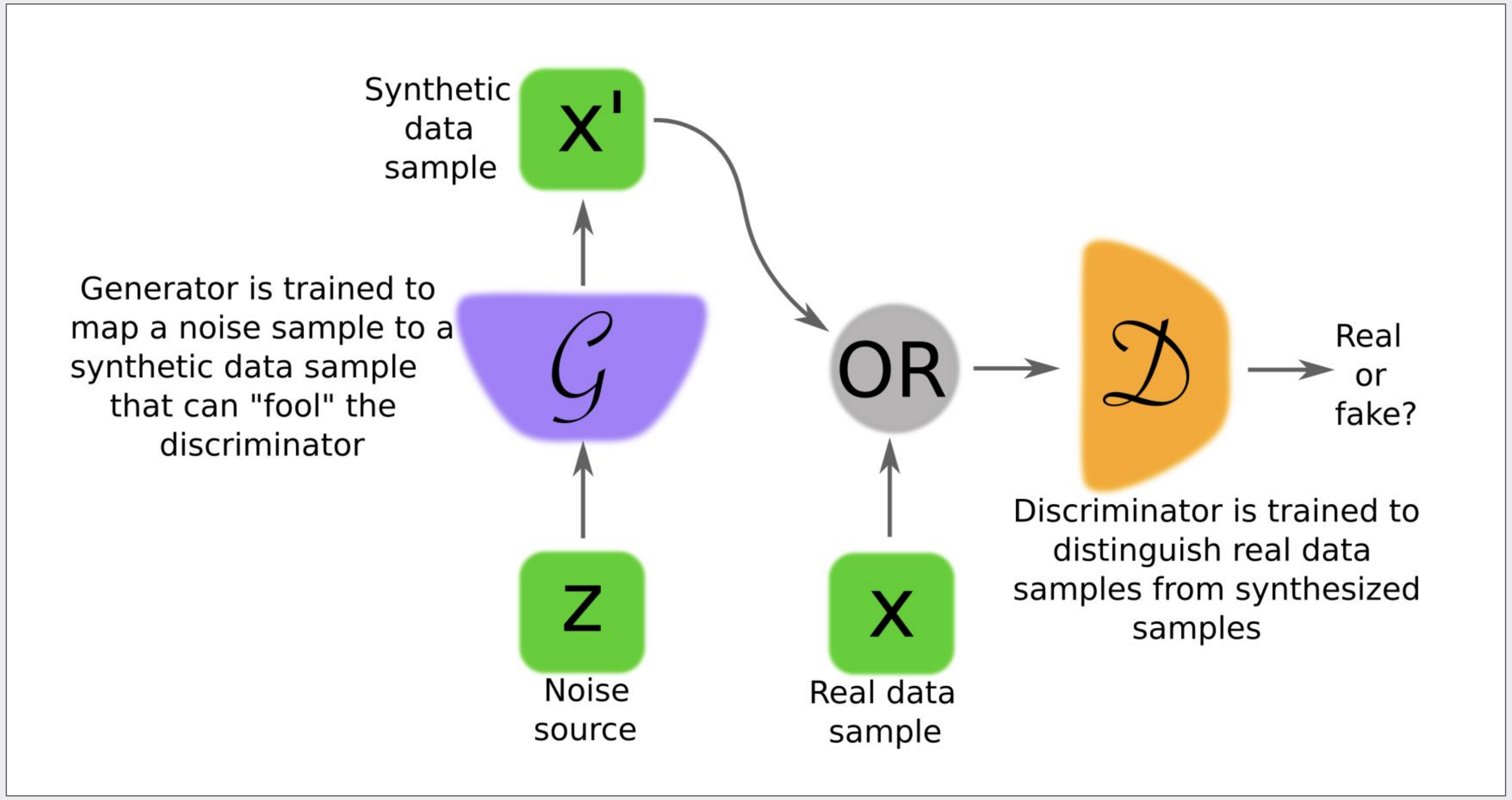
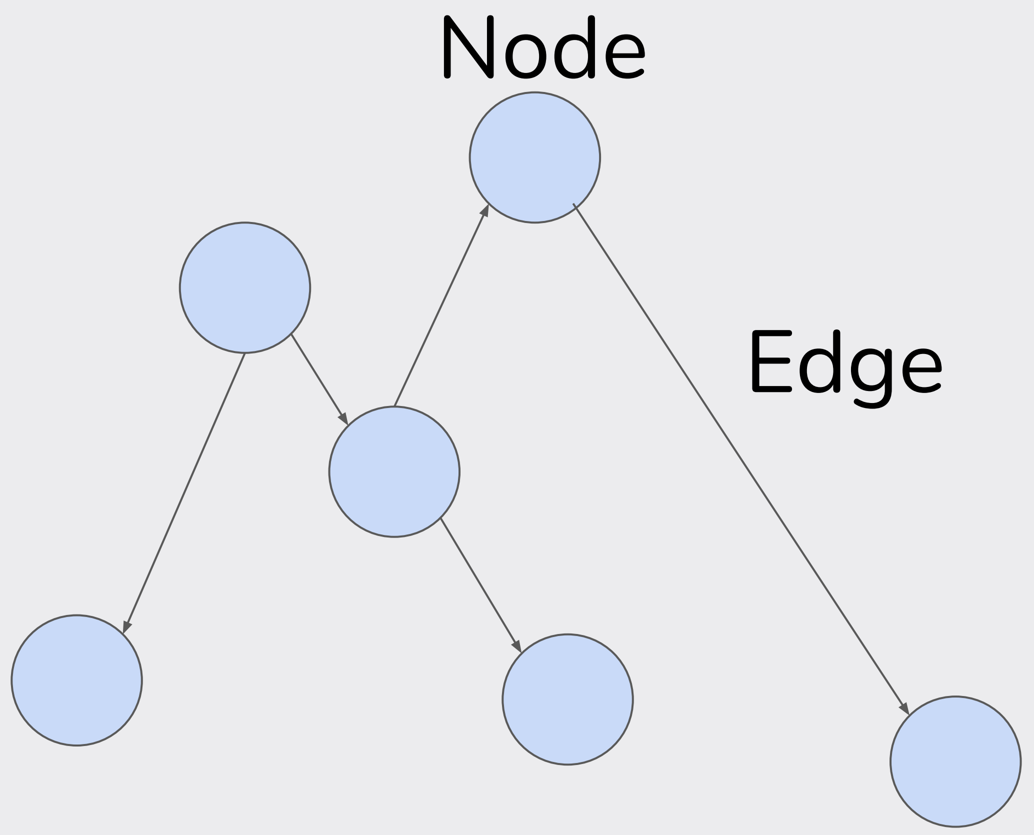


Fig: Creswell, Antonia, et al. "Generative adversarial networks: An overview." IEEE Signal Processing Magazine 35.1 (2018): 53-65.

Credit: "[*Beginner's Guide to Generative Adversarial Networks (GANs)*]" by Pathmind

Graph Representations

- A network graph consists of a set of nodes (or vertices) connected by edges (or links)
- Network graphs arise in many fields
  - Telecommunication networks
  - Computer networks
  - Biological networks
  - Power networks
  - Social networks
- Networks can be
  - Directed or undirected
  - Sparse or dense
  - Static or dynamic



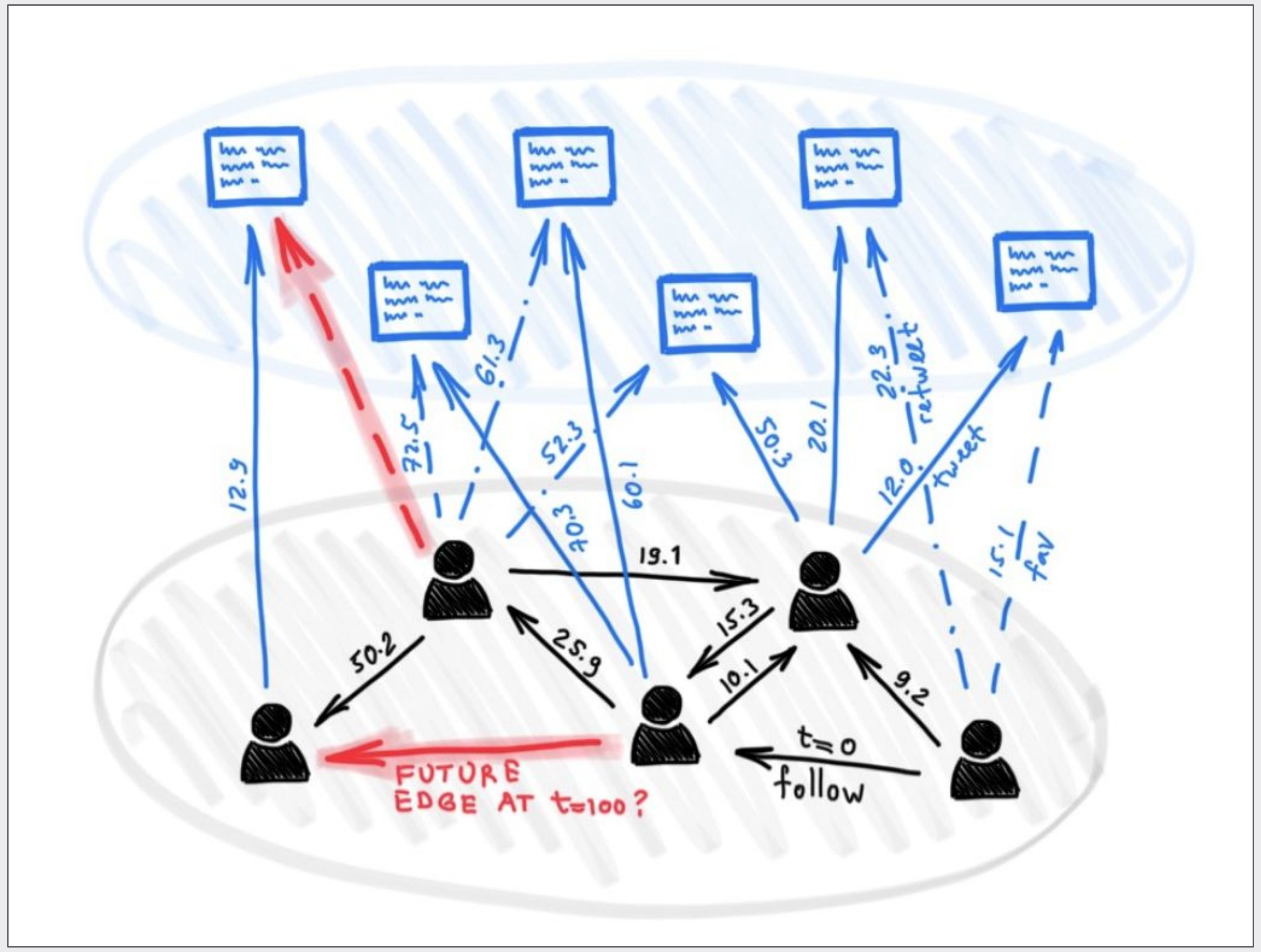


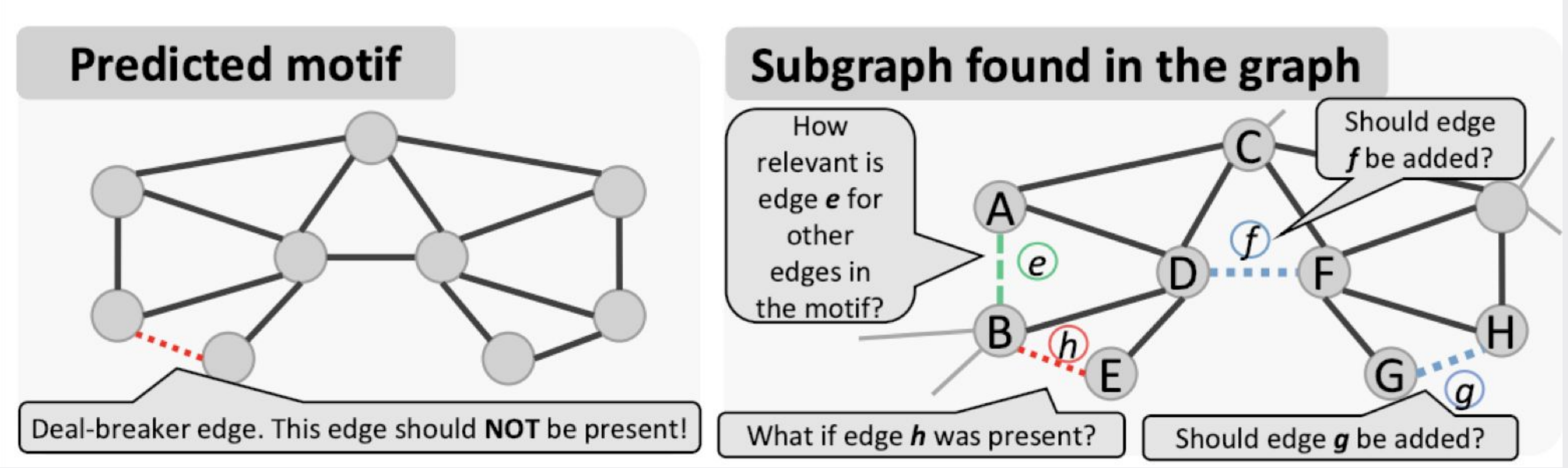
Fig: Rossi, Emanuele, *et al*. "Temporal graph networks for deep learning on dynamic graphs." arXiv preprint arXiv:2006.10637 (2020).

Graph Neural Networks

A substantial thrust in AI toward *graph neural networks:*
geometric deep learning is an umbrella term for emerging techniques that attempt to generalize deep learning models in non-Euclidean domains such as graphs and manifolds, and *motif mining* operates on complex graph patterns:

- “Geometric deep learning: going beyond Euclidean data”
  Michael Bronstein, et al. (2016)
- “Motif Prediction with Graph Neural Networks”
  Maciej Besta, et al. (2021)
- “Machine Learning on Graphs: A Model and Comprehensive Taxonomy”
  Ines Chami, et al. (2021)
- PyG, DGL, GraphGym, etc.

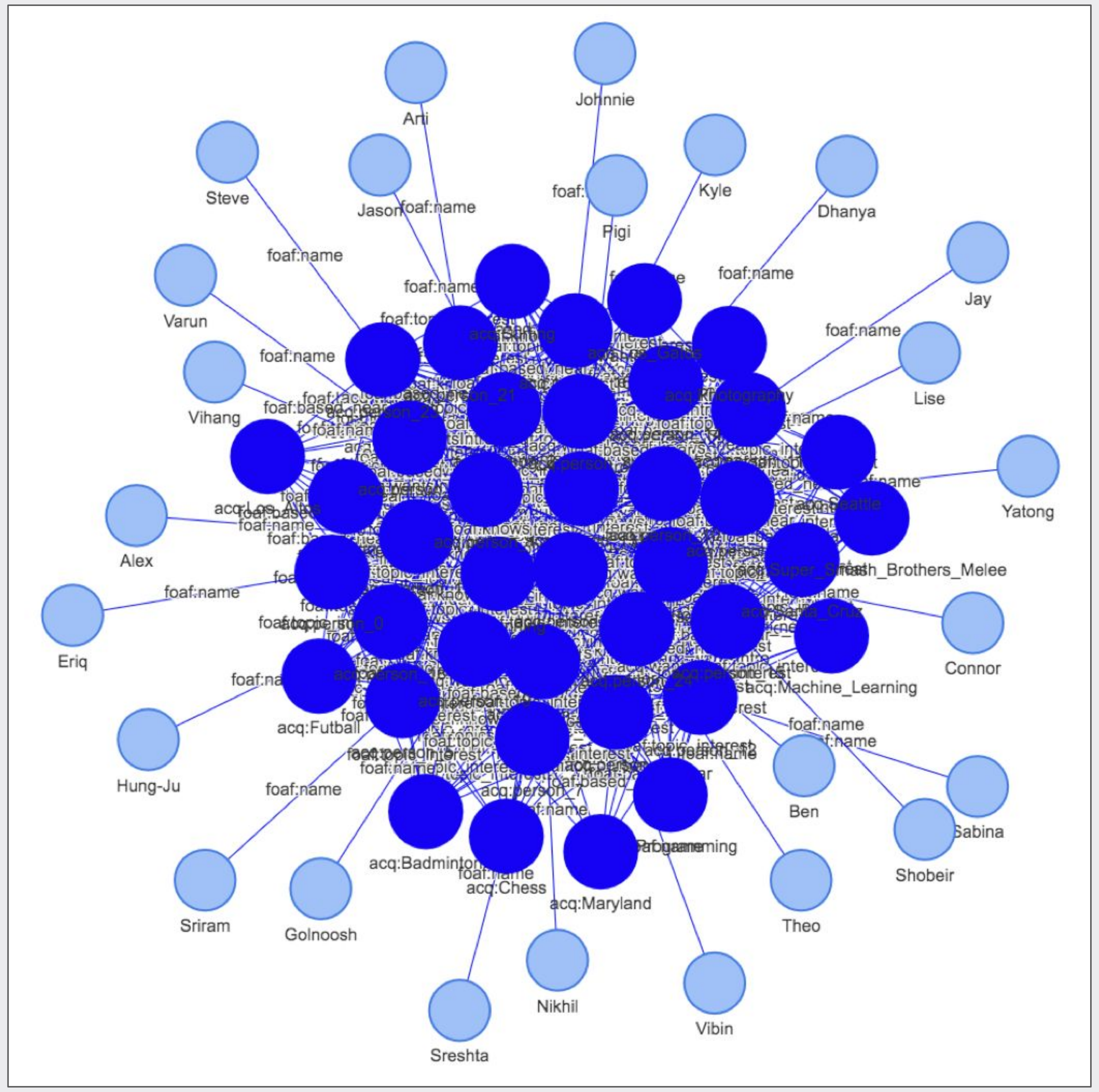
Credit: Paco Nathan, “Graph Thinking” (2021)
https://derwen.ai/s/kcgh#qr



Knowledge Graphs

The gist:

- each entity within a graph has a name and attributes
- some attributes are relations that link to other entities
- other attributes represent values
- use controlled vocabularies to describe the possible kinds of entities, relations, and values
- mix and match vocabularies, or extend per use case



If you've worked with Object Oriented Programming and class hierarchies, you already know this by other names.

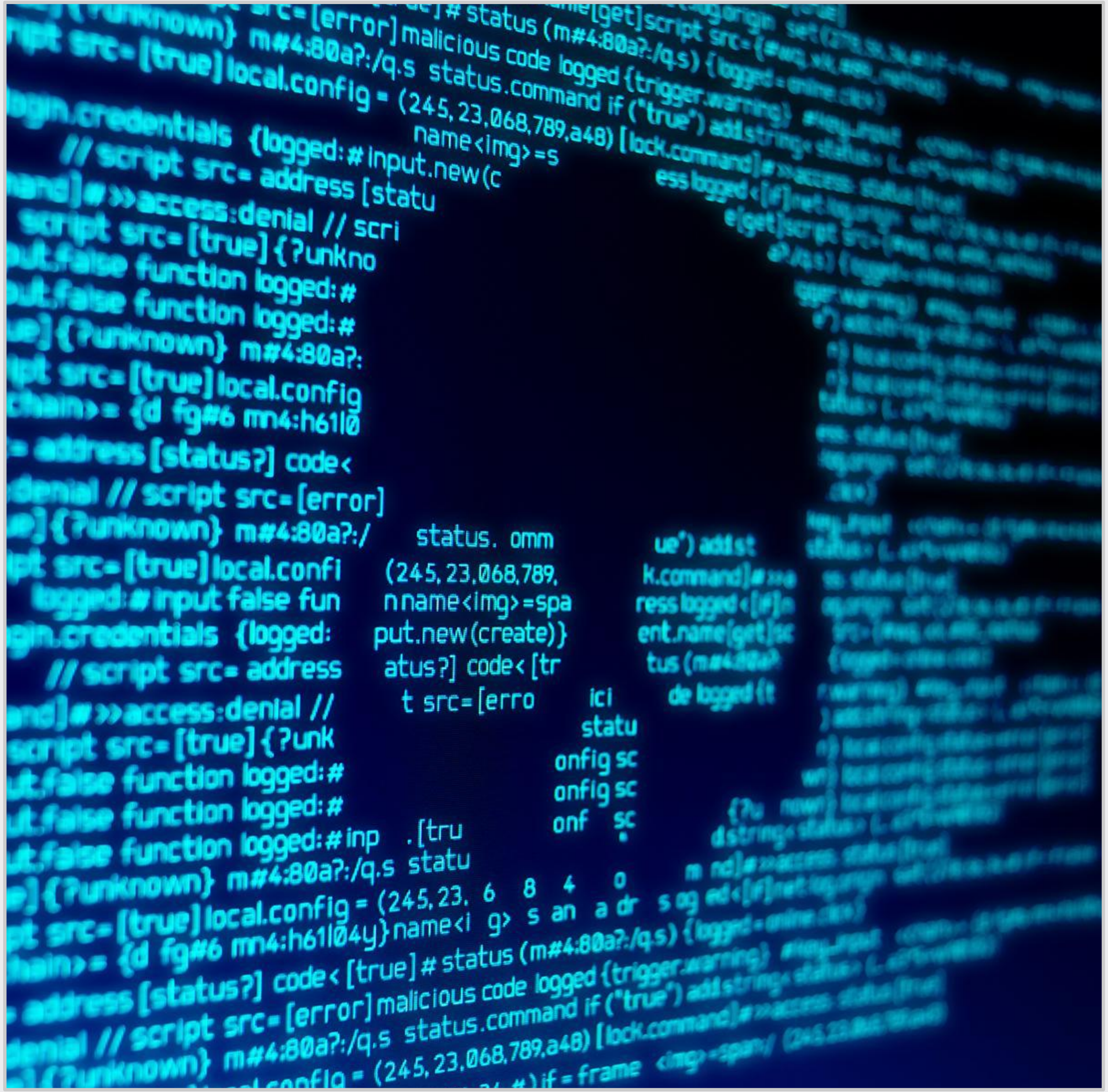
Also, shapes in a graph equate to data objects

Credit: Paco Nathan, "Graph Thinking" (2021) https://derwin.ai

Power and Potential of AI Systems

Reasoning and Discovery

- Fraud and anomaly detection
- Financial market trading
- Legal document assessment
- Financial asset management
- Financial application processing
- Product and media recommendations

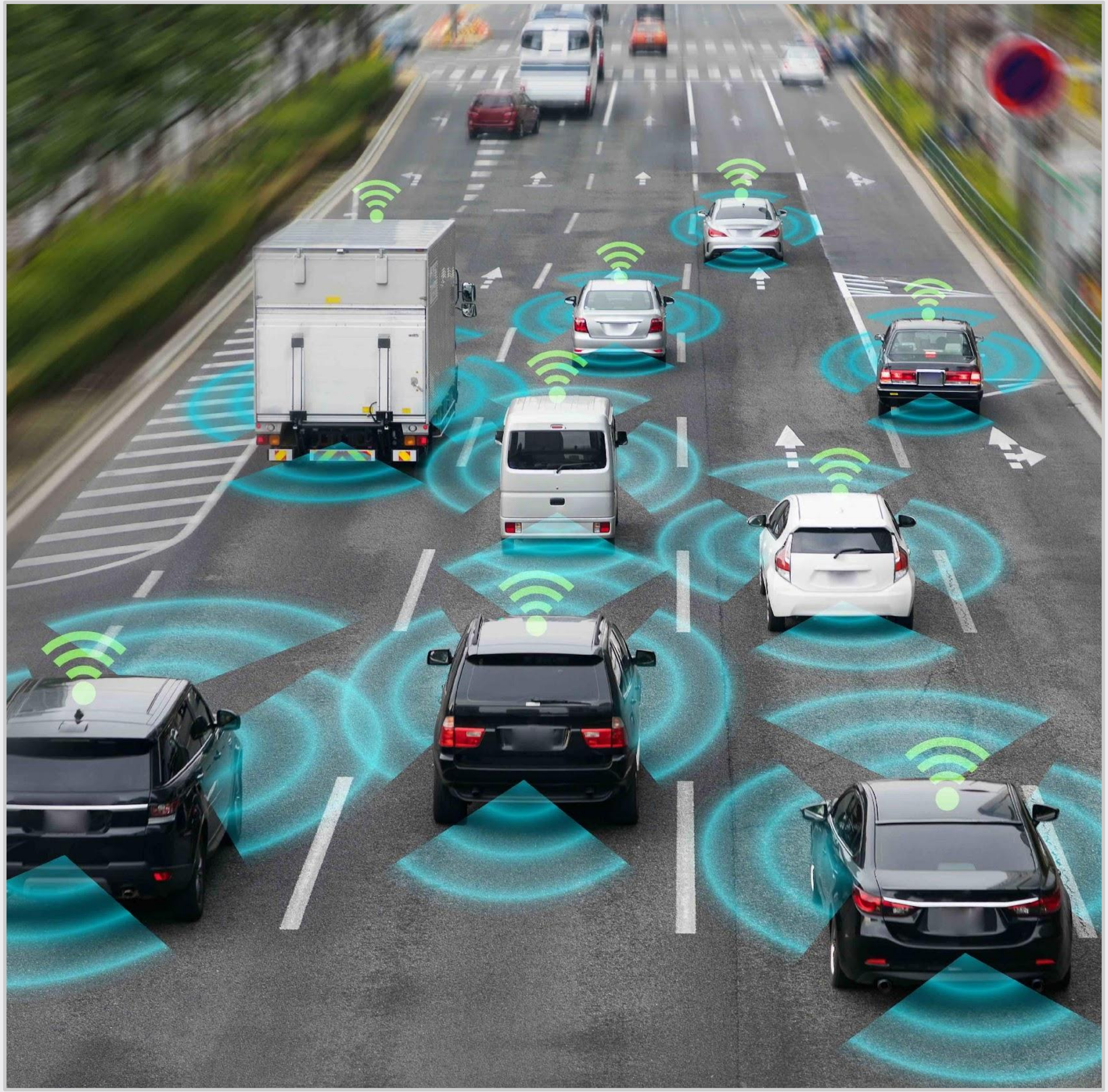


Planning and Optimization

- Logistics and scheduling
- Demand forecasting
- Predictive maintenance
- Inventory optimization
- Sales revenue prediction

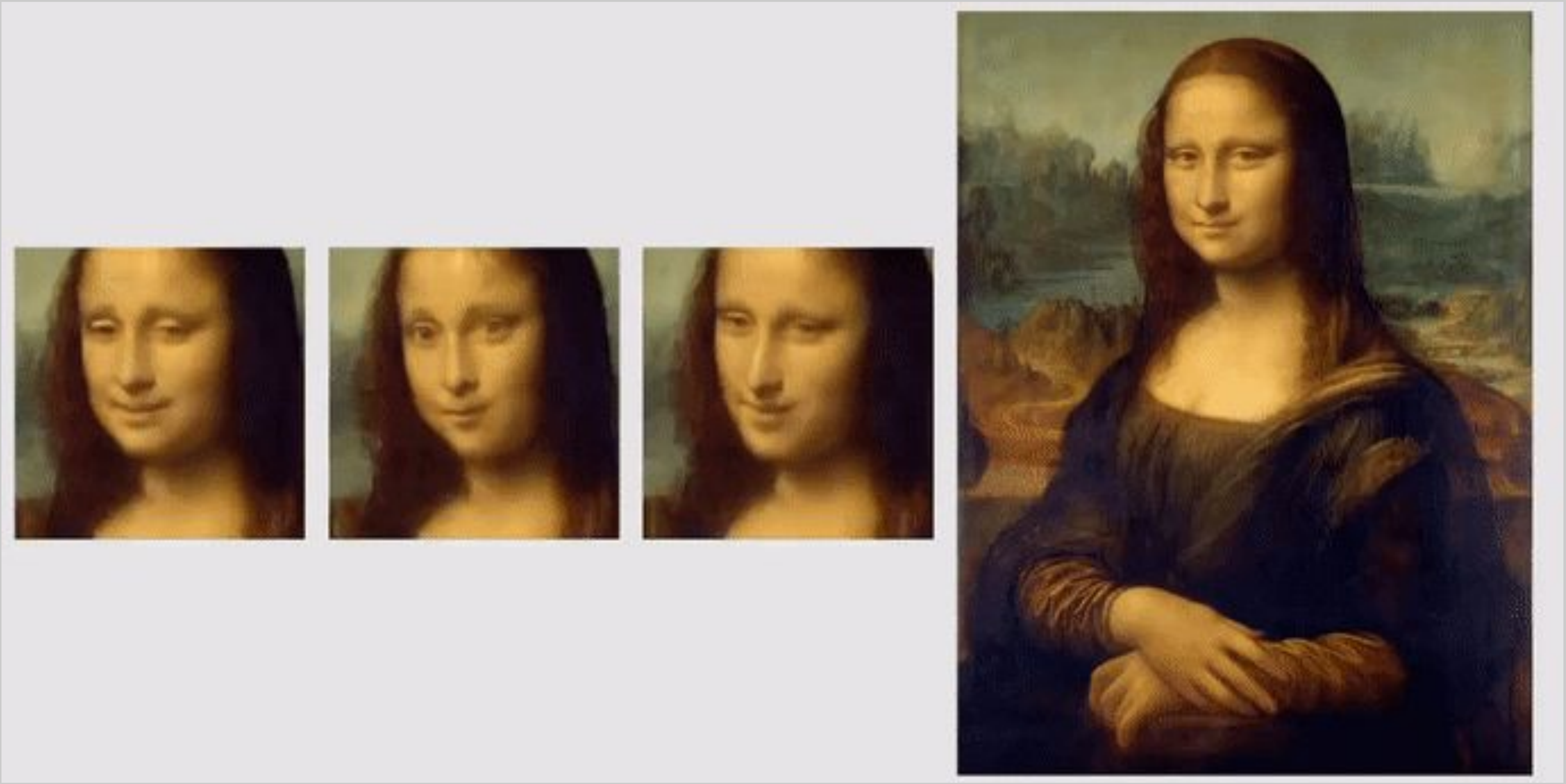
Perception and Communication

- Autonomous vehicles
- Medical imagery analysis
- Intelligent agents
- Voice recognition
- Language translation



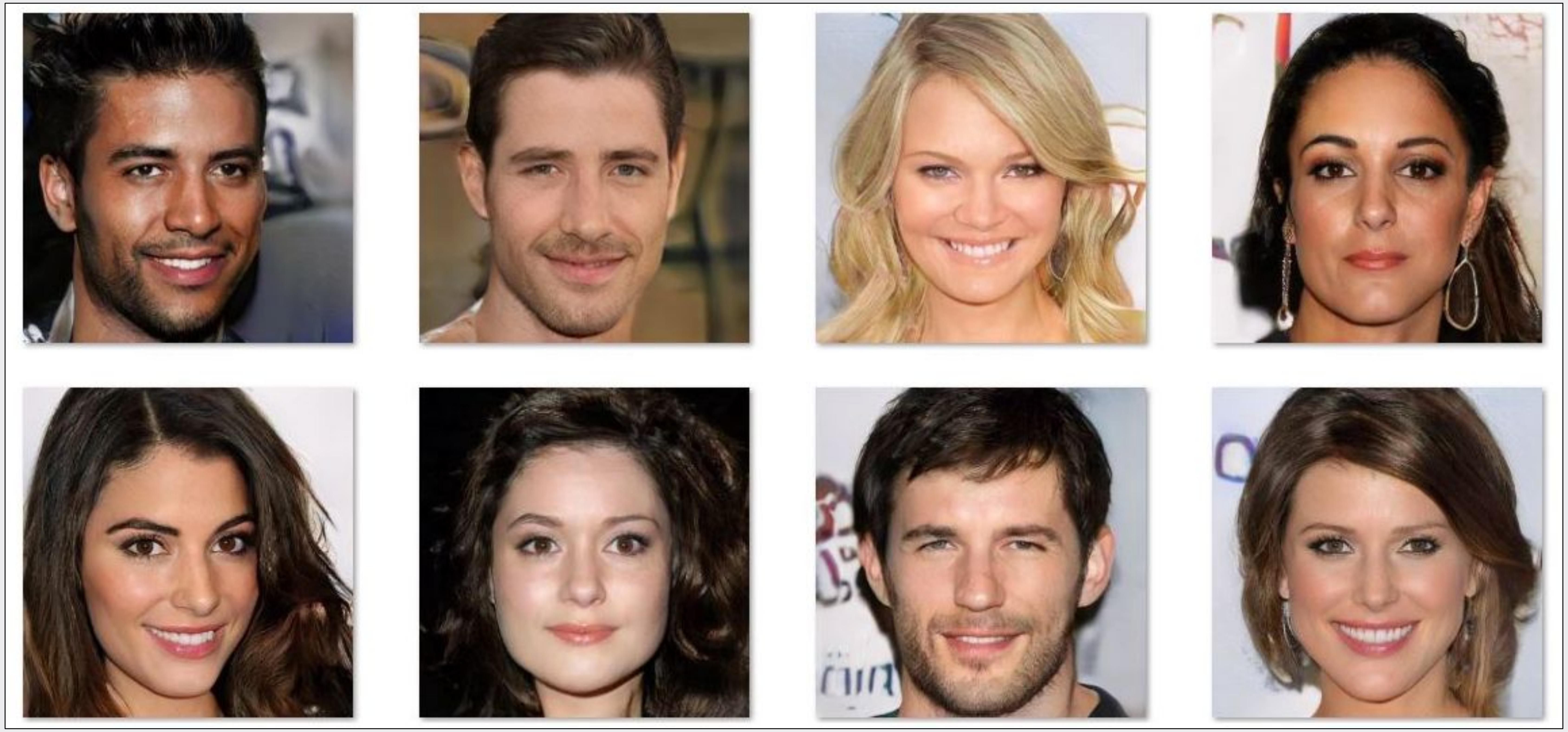
Creativity and Synthesis

- Photo-realistic images
- Text generation
- Music composition
- Text <-> Image
- Single-shot photo animation



AI-Generated Faces

Can you spot the fake?



Vision Transformers (ViT)

- To maximize code and hardware reuse, original Vaswani 2017 encoder used
- Image divided into patches
    a. Projected with learned embedding layer
    b. Fed into the transformer encoder in parallel
- This approach lacks useful inductive biases of CNNs, but seems to work better for large models and (pre) training sets

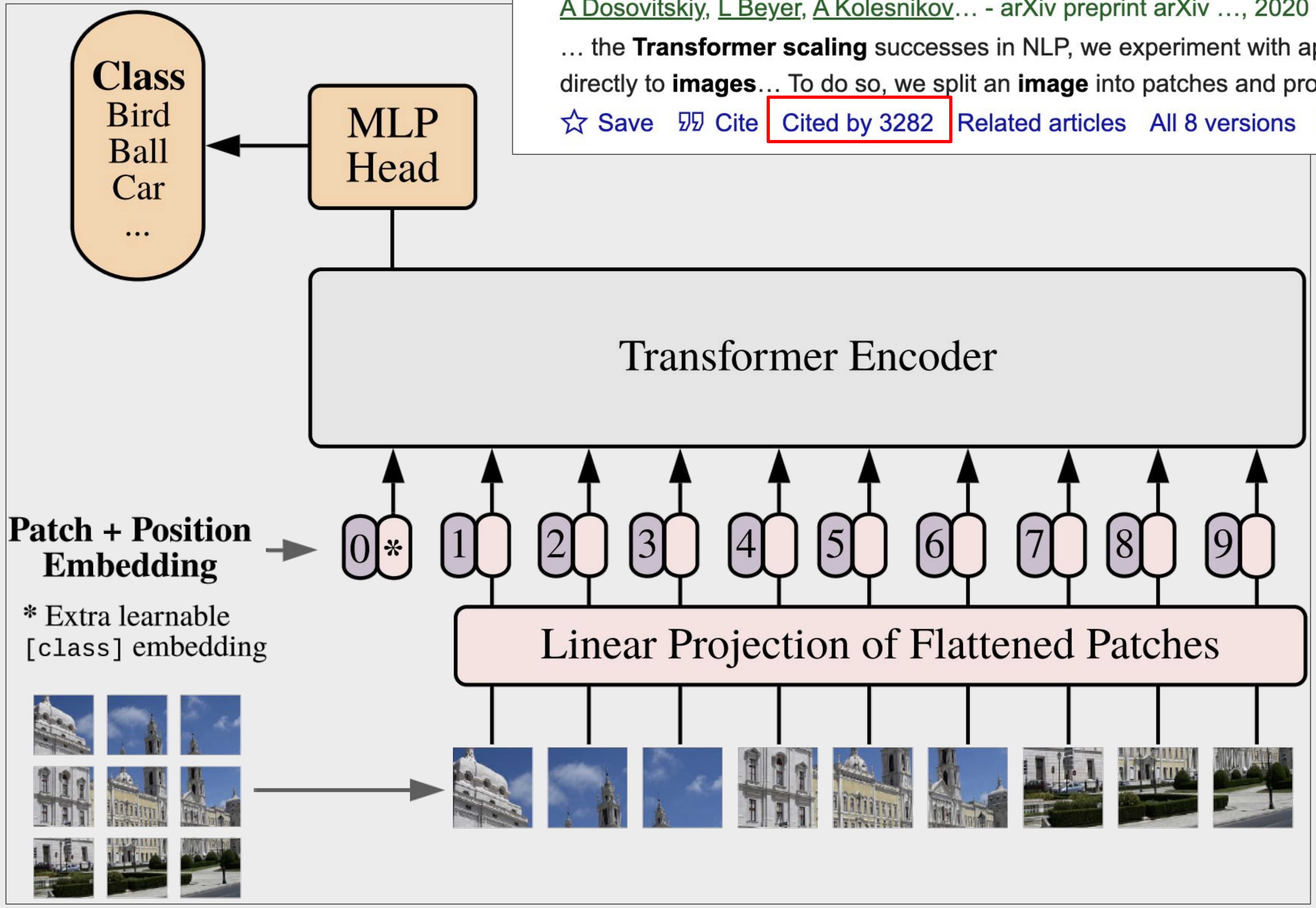
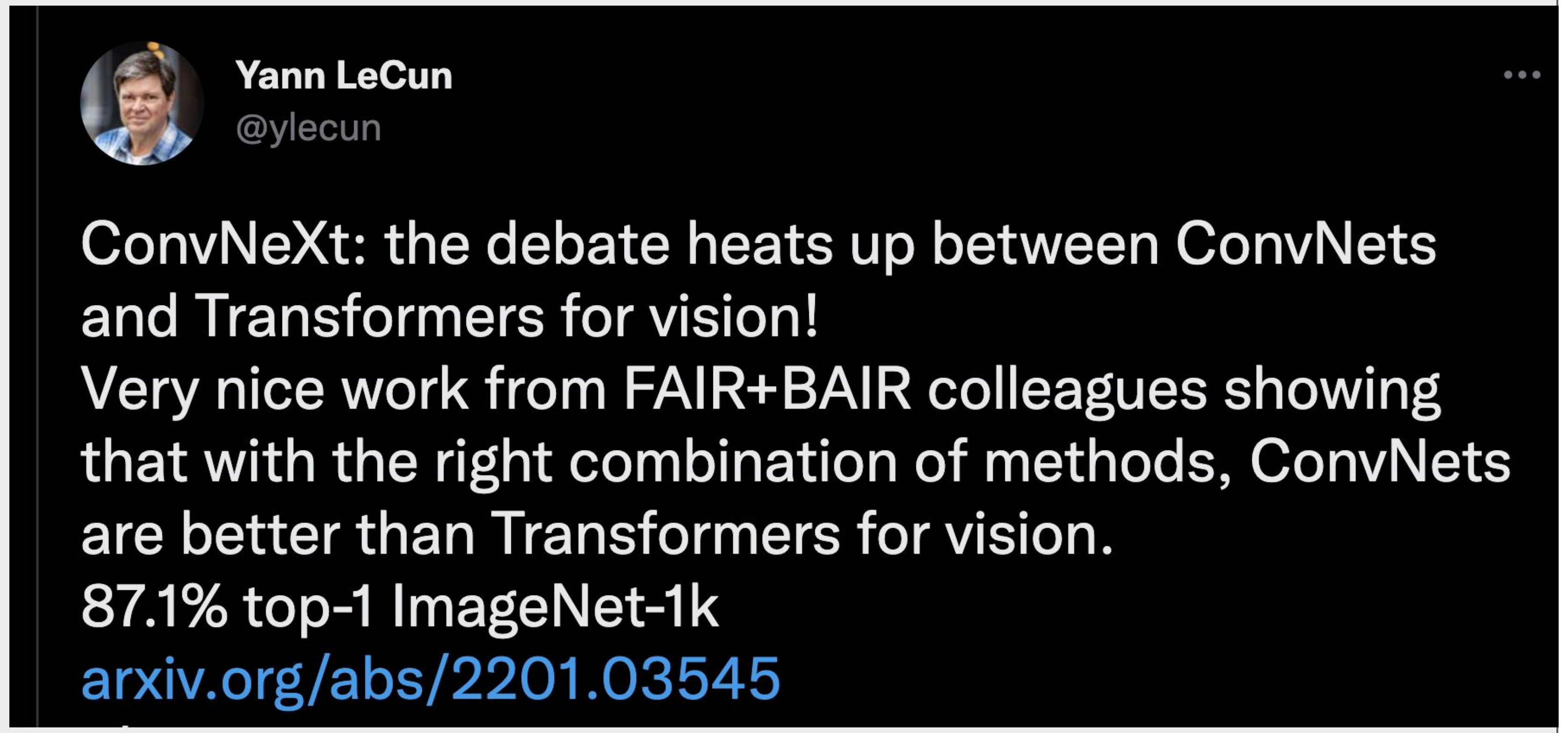


Fig: Dosovitskiy, Alexey, *et al*. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

ConvNeXt: CNNs vs. ViT Ongoing Battle

- Lessons learned from ViT research boost performance of CNNs.
- ConvNeXt outperforms Swin Transformers (hierarchical ViT) in key areas
- Inductive biases, especially translation equivariance, still make CNNs a powerful tool



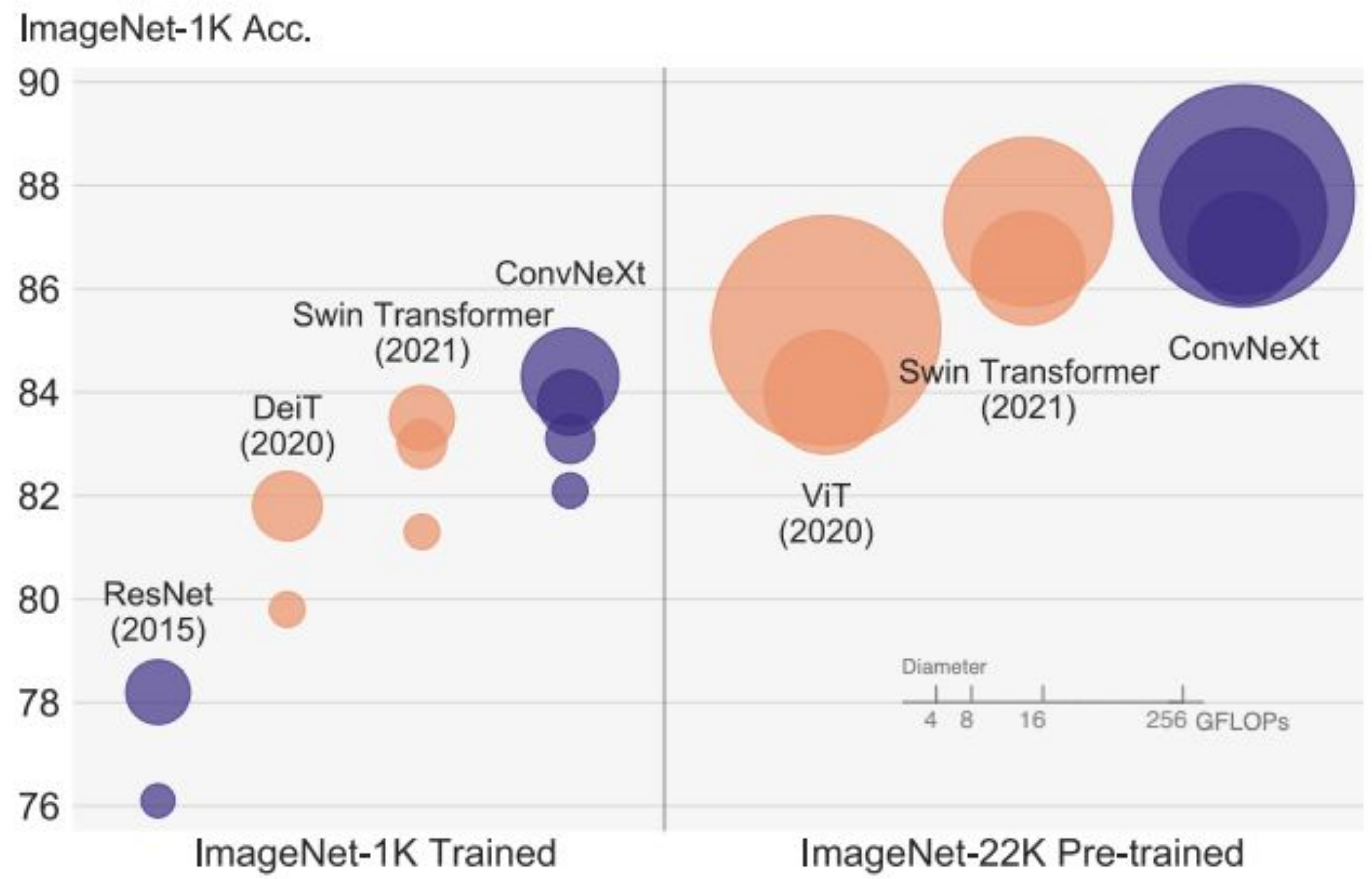
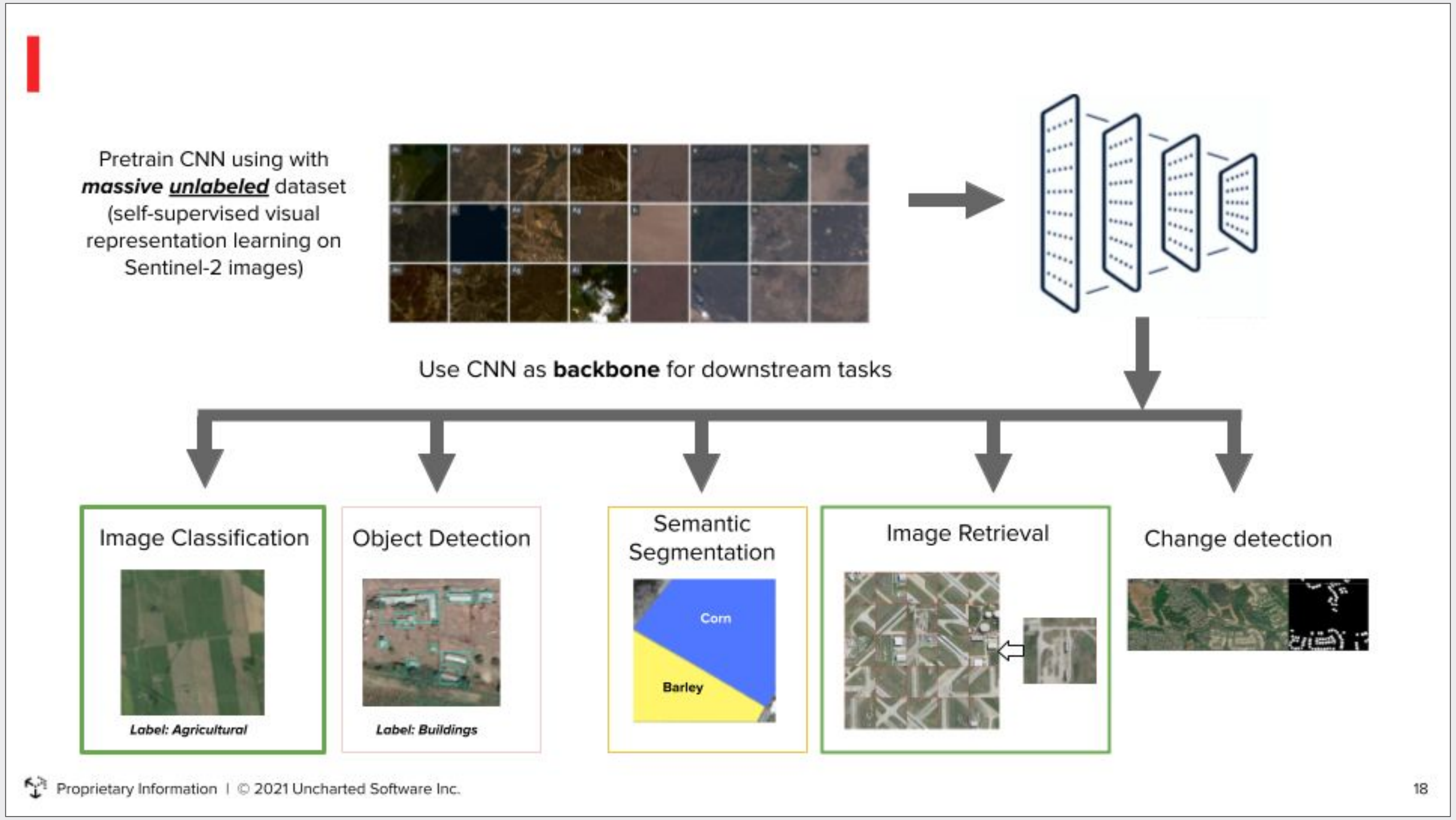


Figure 1. **ImageNet-1K classification** results for ● ConvNets and ○ vision Transformers. Each bubble's area is proportional to FLOPs of a variant in a model family. ImageNet-1K/22K models here take $224^2/384^2$ images respectively. We demonstrate that a standard ConvNet model can achieve the same level of scalability as hierarchical vision Transformers while being much simpler in design.

Fig: Liu, Zhuang, et al. "A ConvNet for the 2020s." arXiv preprint arXiv:2201.03545 (2022).

D3M Remote Sensing with Distil

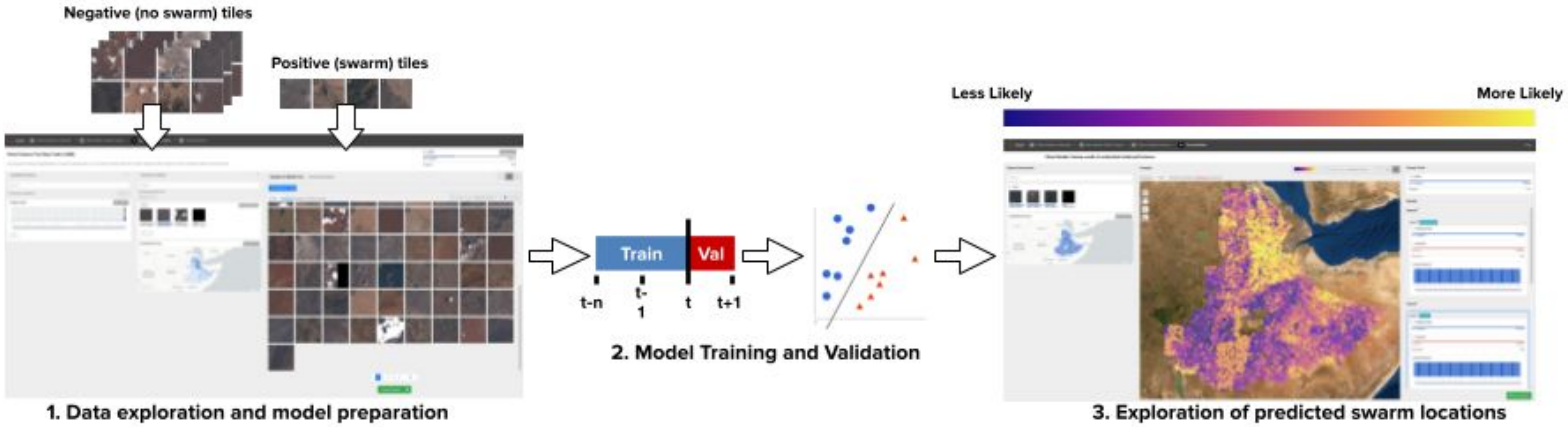
Uncharted Software's Distil system allows users to discover underlying dynamics of complex systems and generate data-driven models. https://d3m.uncharted.software/



Identifying Locust Breeding Grounds from Satellite Imagery

Food security use case: predict agricultural regions in Ethiopia at risk from locust swarms

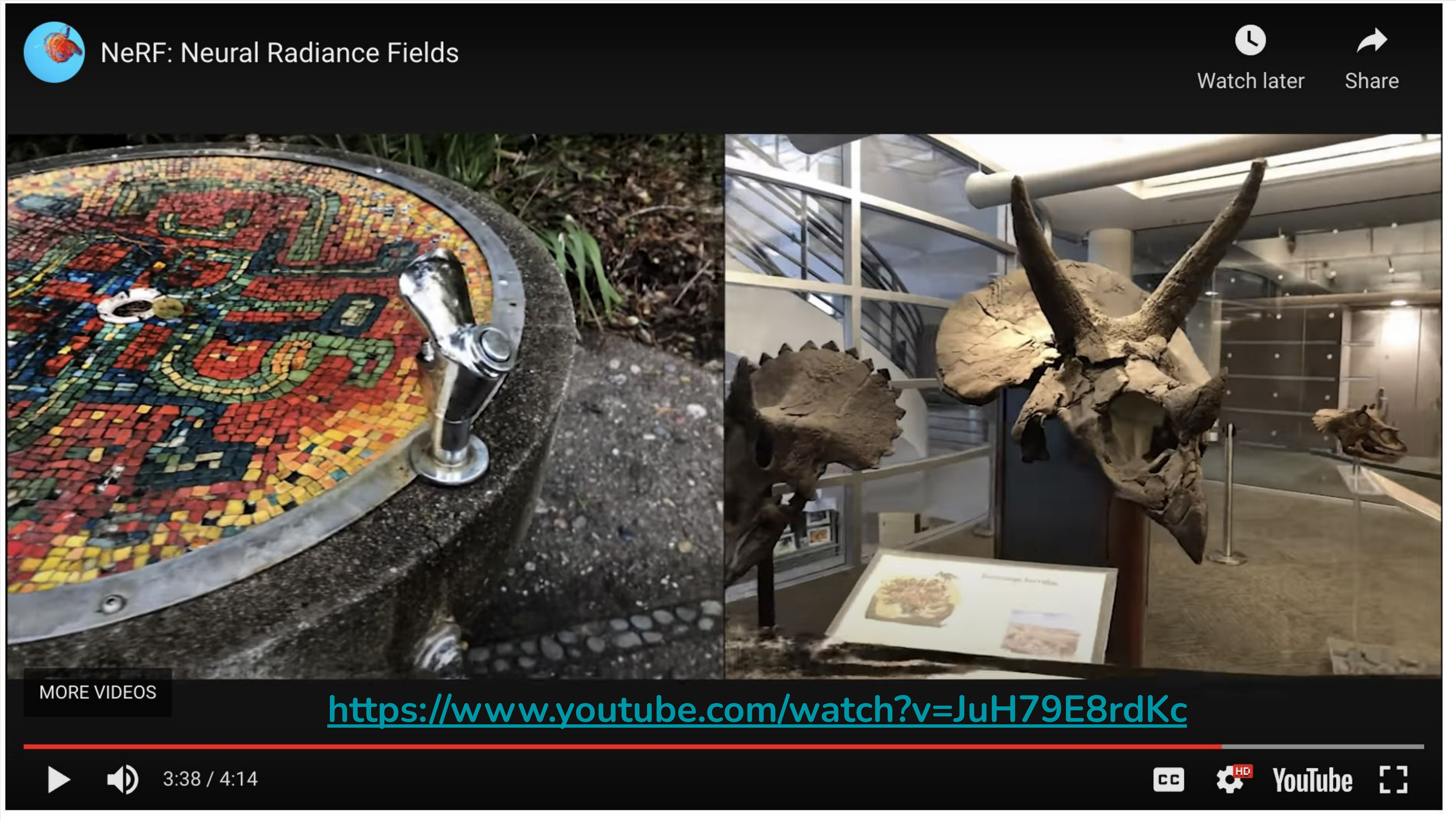
- Collected historical mature swarm sighting locations from FAO Locust Hub to act as ground truth
- Collected geo-temporally corresponding Sentinel-2 imagery to act as positive examples of swarm sighting locations, and a random sample of imagery within Ethiopia to act as a negative examples
- Used previous time-steps to train model to predict sighting locations based on overhead imagery; rank tiles in current time step by model score



Ref: Langevin, Scott, Chris Bethune, Philippe Horne, Steve Kramer, Jeffrey Gleason, Ben Johnson, Ezekiel Barnett, Fahd Husain, and Adam Bradley. "Useable machine learning for Sentinel-2 multispectral satellite imagery." In *Image and Signal Processing for Remote Sensing XXVII*, vol. 11862, pp. 97-114. SPIE, 2021.

NeRF (Neural Radiance Fields)

NeRF is a new method for synthesizing novel views of complex 3-D scenes using automatic unsupervised semantic scene decomposition https://www.matthewtancik.com/nerf

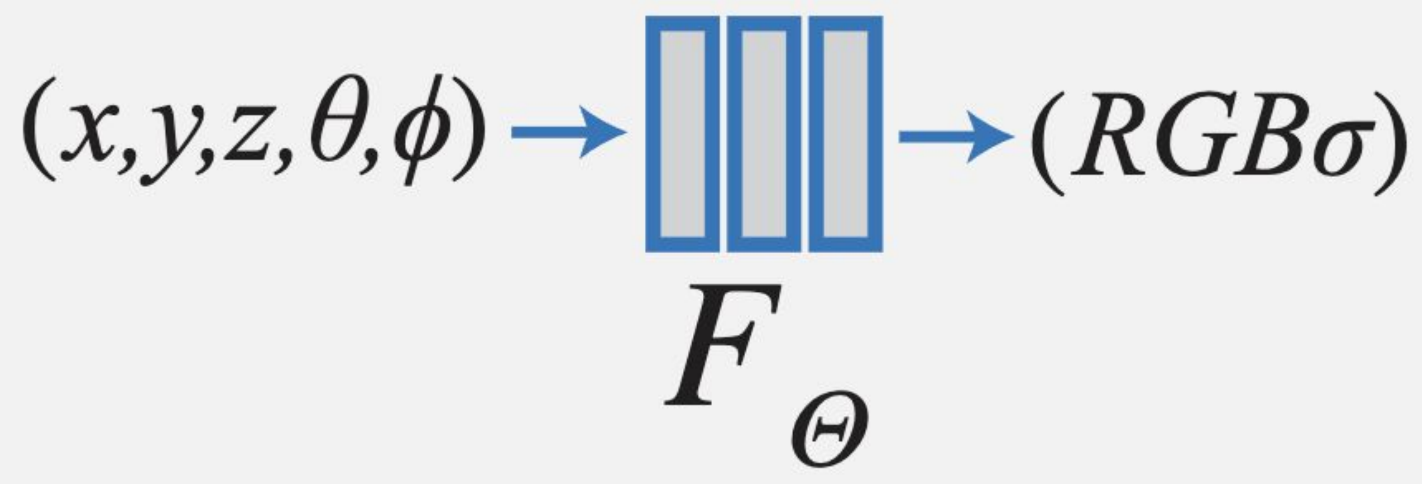


https://www.youtube.com/watch?v=JuH79E8rdKc

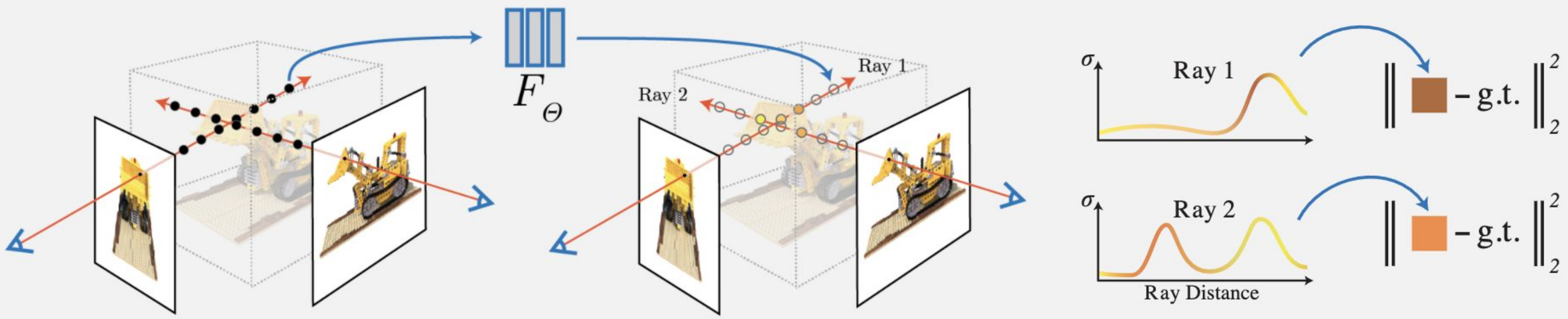
Credit: Mildenhall, B., et al. "Representing scenes as neural radiance fields for view synthesis." Proc. of European Conference on Computer Vision, Virtual. 2020.

NeRF Approach

We present a method that achieves state-of-the-art results for synthesizing novel views of complex scenes by optimizing an underlying continuous volumetric scene function using a sparse set of input views.



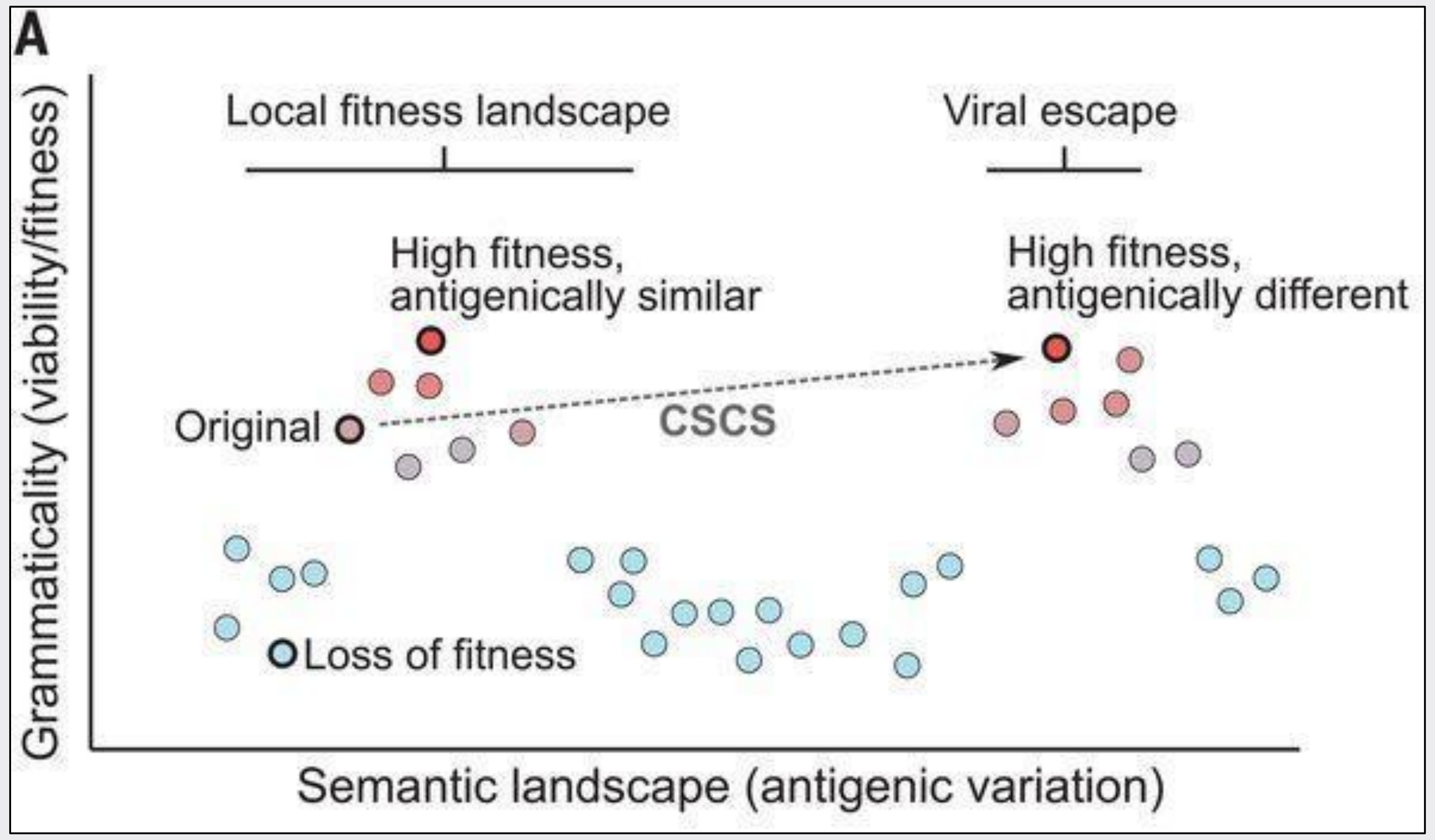
Our algorithm represents a scene using a fully-connected (non-convolutional) deep network, whose input is a single continuous 5D coordinate (spatial location (x, y, z) and viewing direction (θ, φ)) and whose output is the volume density and view-dependent emitted radiance at that spatial location.



We synthesize views by querying 5D coordinates along camera rays and use classic volume rendering techniques to project the output colors and densities into an image. Because volume rendering is naturally differentiable, the only input required to optimize our representation is a set of images with known camera poses. We describe how to effectively optimize neural radiance fields to render photorealistic novel views of scenes with complicated geometry and appearance, and demonstrate results that outperform prior work on neural rendering and view synthesis.

COVID Protein Evolution Prediction

"Language models trained on viral sequences can predict mutations that preserve infectivity but induce high antigenic change, akin to preserving "grammaticality" but inducing high "semantic change".
- Nathan Benaich & Ian Hogarth, State of AI Report (2021)



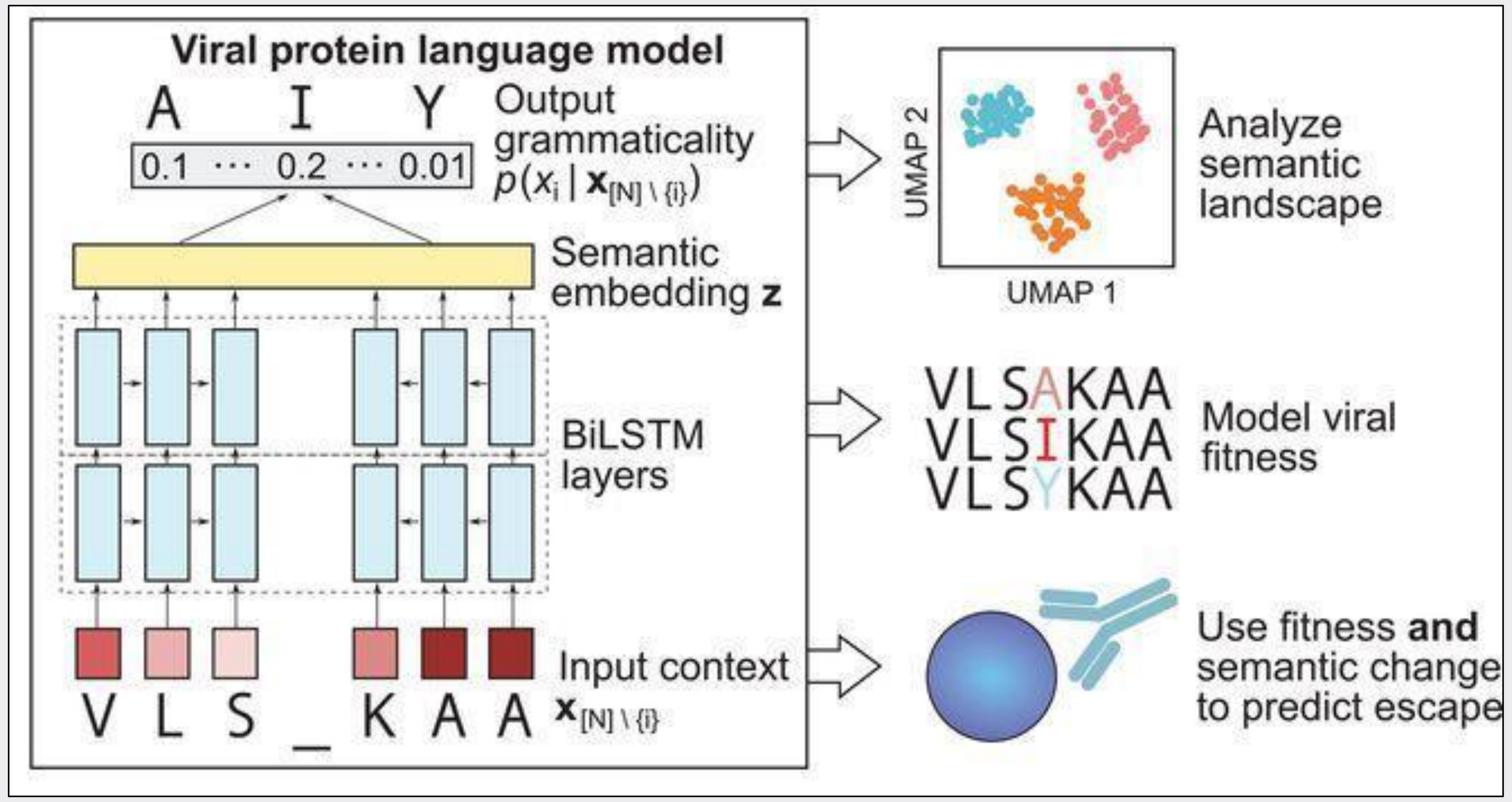
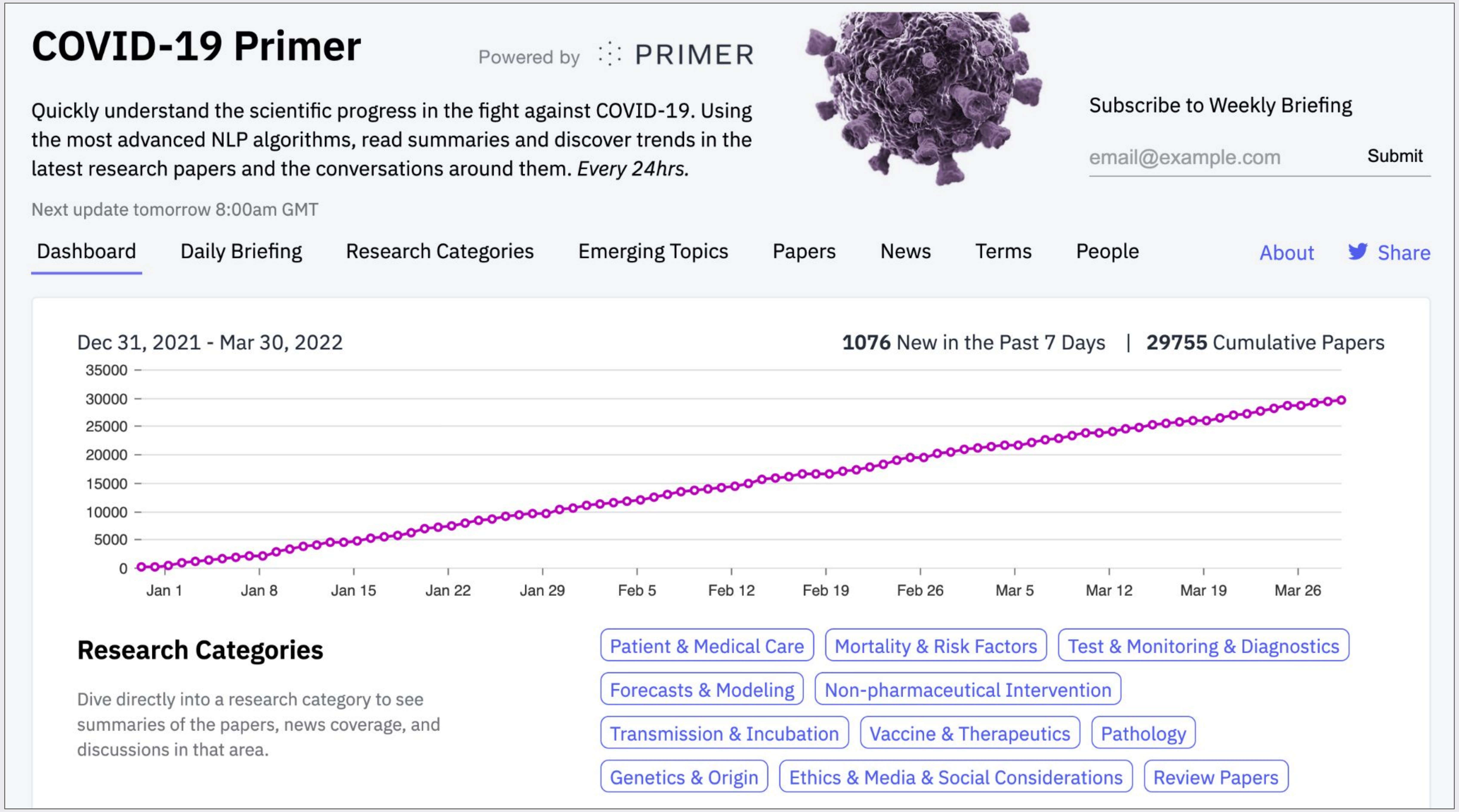


Fig: Maher, M. Cyrus, et al. "Predicting the mutational drivers of future SARS-CoV-2 variants of concern." Science translational medicine (2021): eabk3445.

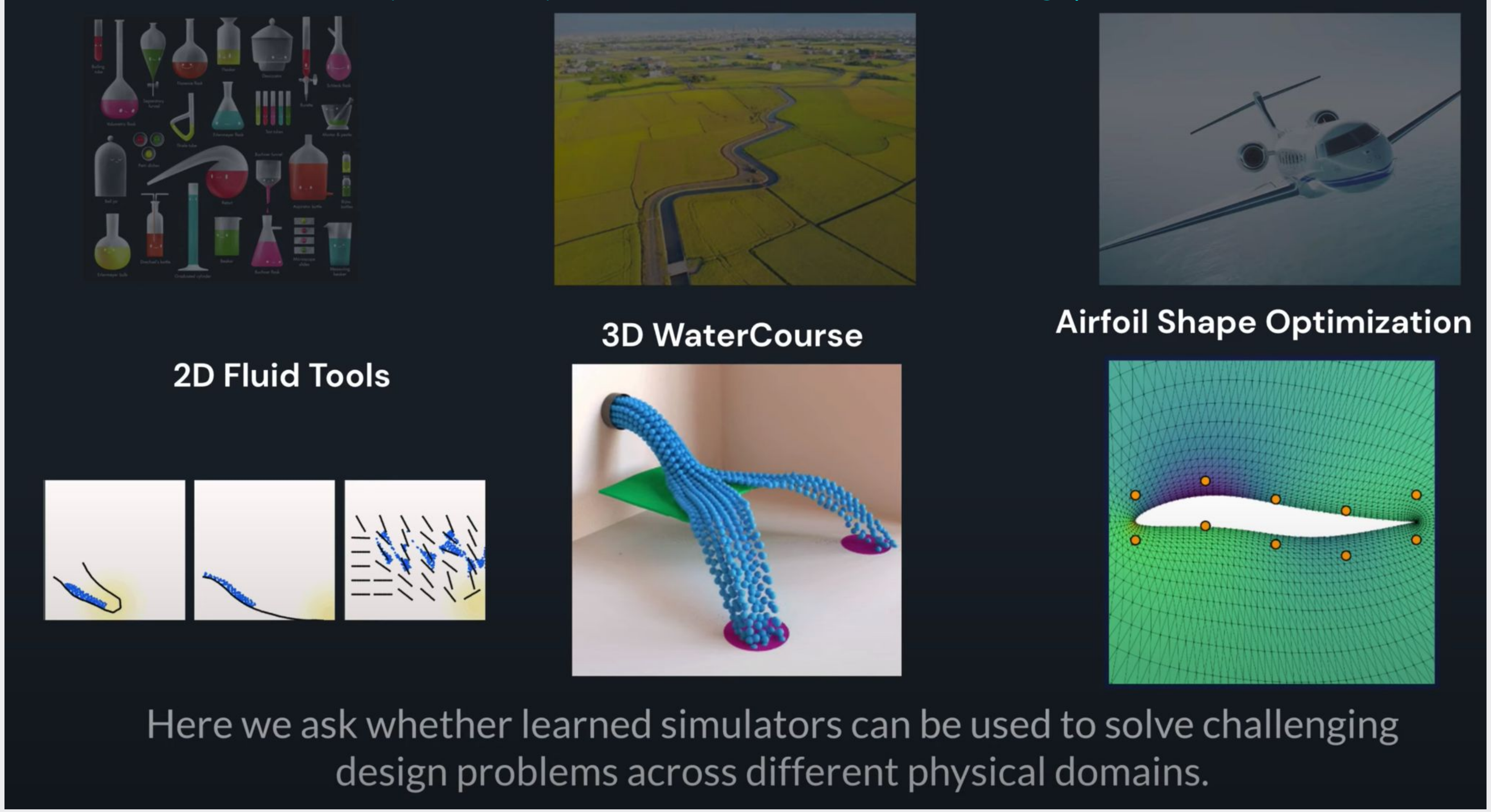
Primer’s COVID Research Summarizer

Primer’s NLP deep learning models analyze and summarize the huge numbers of COVID-related research publications. https://covid19primer.com/



Physical Design using Differentiable Learned Simulators

https://www.youtube.com/watch?v=UX8x2dIAgQw&t=9s



Ref: Allen, Kelsey R., *et al*. "Physical Design using Differentiable Learned Simulators." arXiv preprint arXiv:2202.00728 (2022).

Physical Design using Differentiable Learned Simulators

This approach combines learned forward simulators based on graph neural networks with gradient-based design optimization.

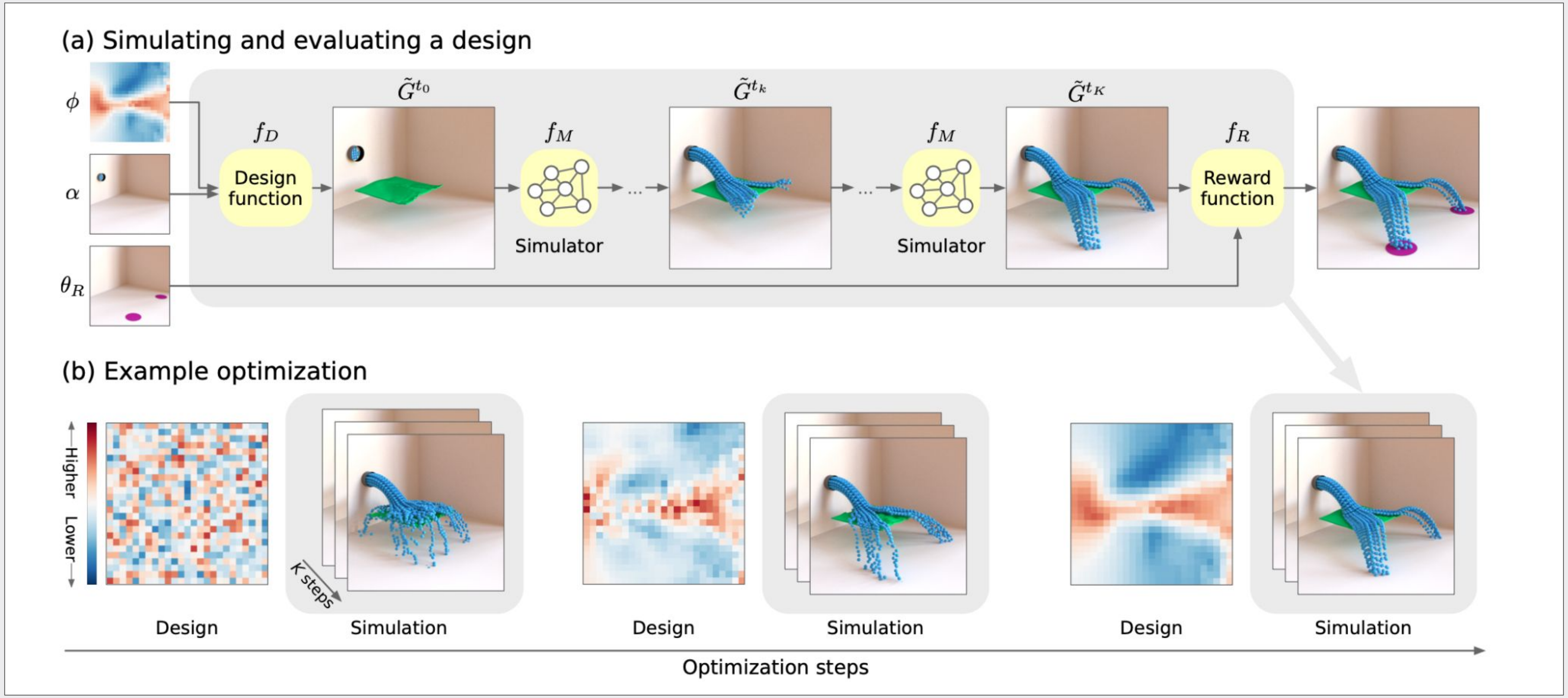
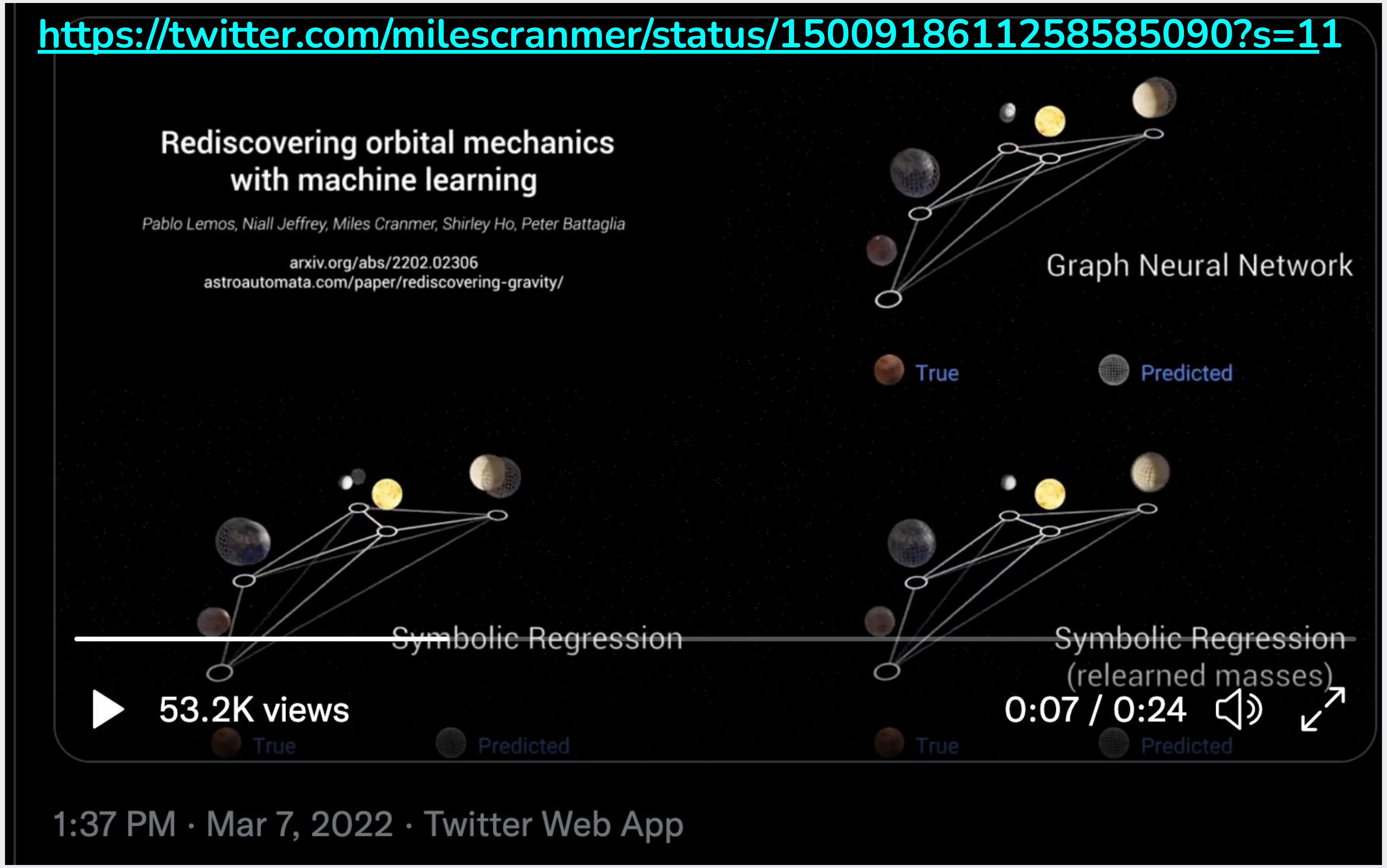


Fig: Allen, Kelsey R., *et al*. "Physical Design using Differentiable Learned Simulators." arXiv preprint arXiv:2202.00728 (2022).

GNNs to Rediscover Physical Laws

Graph neural networks can be combined with symbolic dynamics to model physical systems.



Ref: Lemos, Pablo, et al. "Rediscovering orbital mechanics with machine learning." arXiv preprint arXiv:2202.02306 (2022).

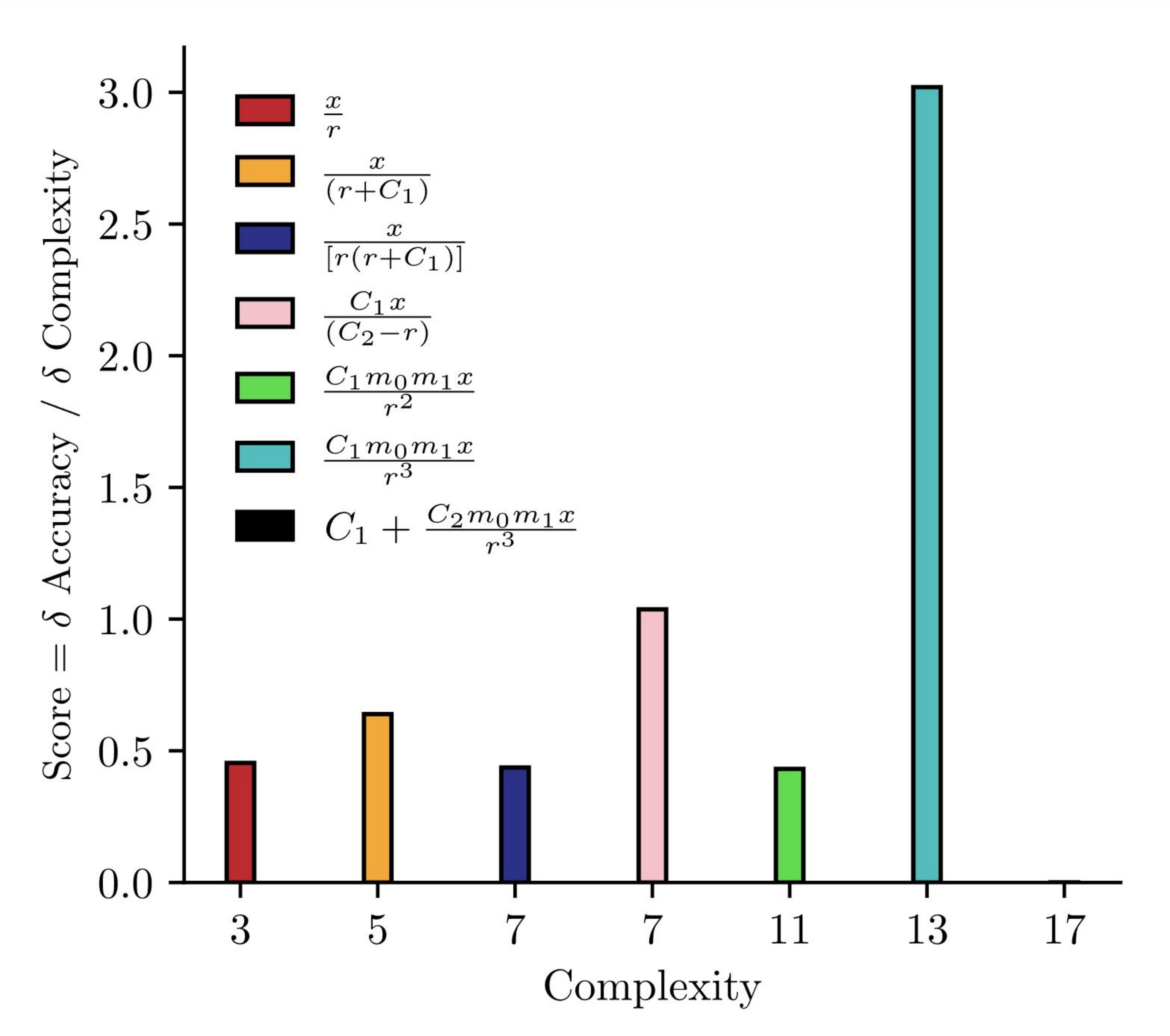
GNNs to Rediscover Physical Laws

Summary of the algorithm:

1. Declare unknown physical properties of a system as trainable parameters in a machine learning model.
2. Update these parameters simultaneously with the model weights.
3. Finally, distill the learned model to a set of symbolic rules.

After training, we use PySR to find the following symbolic forms as approximations of our graph neural network's edge function.

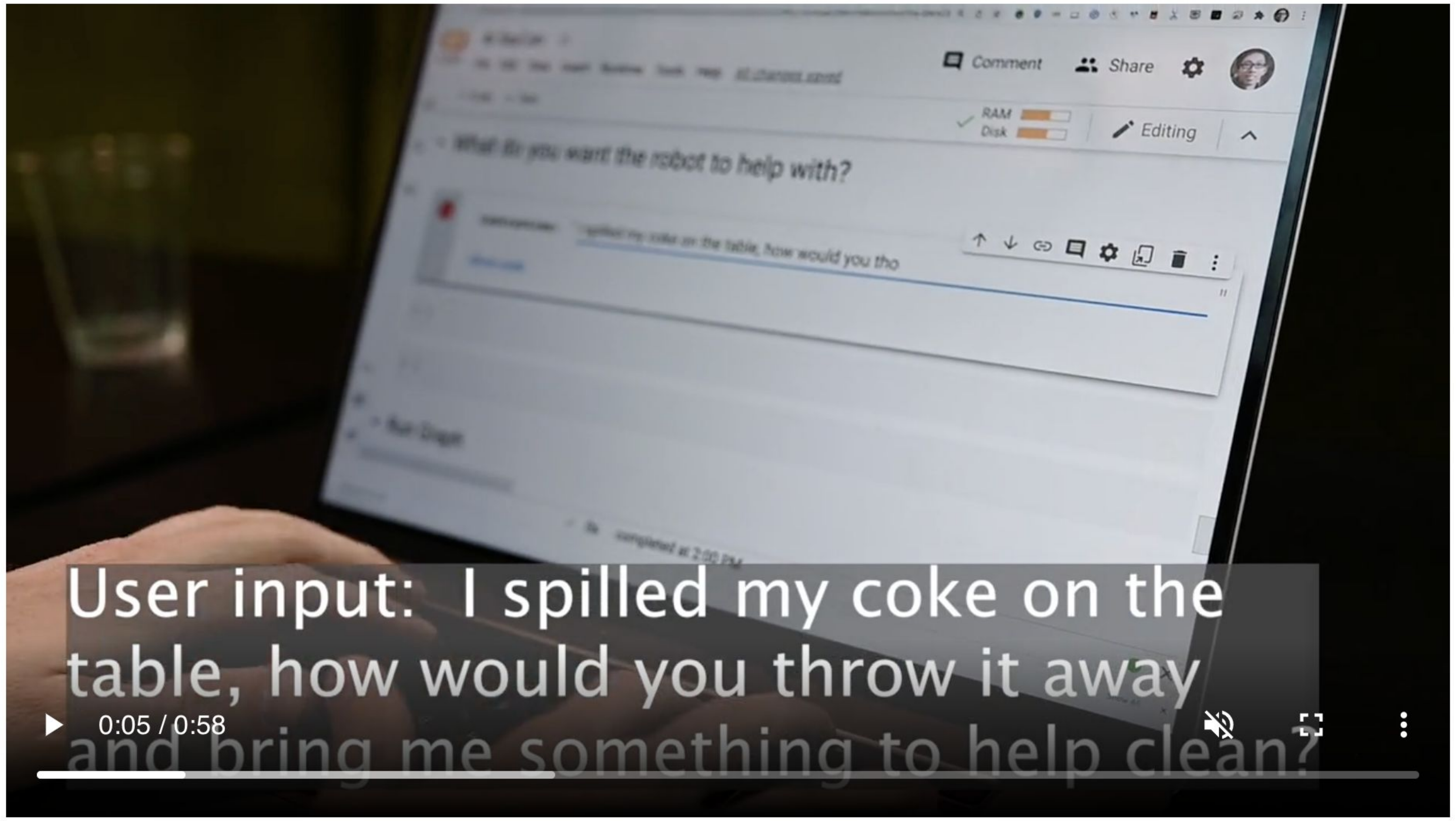
The symbolic rule that best balances accuracy and simplicity is the same as the law of universal gravitation (in teal).



Ref: Lemos, Pablo, et al. "Rediscovering orbital mechanics with machine learning." arXiv preprint arXiv:2202.02306 (2022).

Robotics + Large Language Models

"Large language models can encode a wealth of semantic knowledge about the world. Such knowledge could be extremely useful to robots aiming to act upon high-level, temporally extended instructions expressed in natural language. However, a significant weakness of language models is that they lack real-world experience, which makes it difficult to leverage them for decision making within a given embodiment."



https://say-can.github.io/img/demo_sequence_compressed.mp4

Ref: Ahn, Michael, *et al.* "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances." arXiv preprint arXiv:2204.01691 (2022).

Robotics + Large Language Models

“We propose to provide real-world grounding by means of pretrained skills, which are used to constrain the model to propose natural language actions that are both feasible and contextually appropriate. The robot can act as the language model's ‘hands and eyes,"’while the language model supplies high-level semantic knowledge about the task”

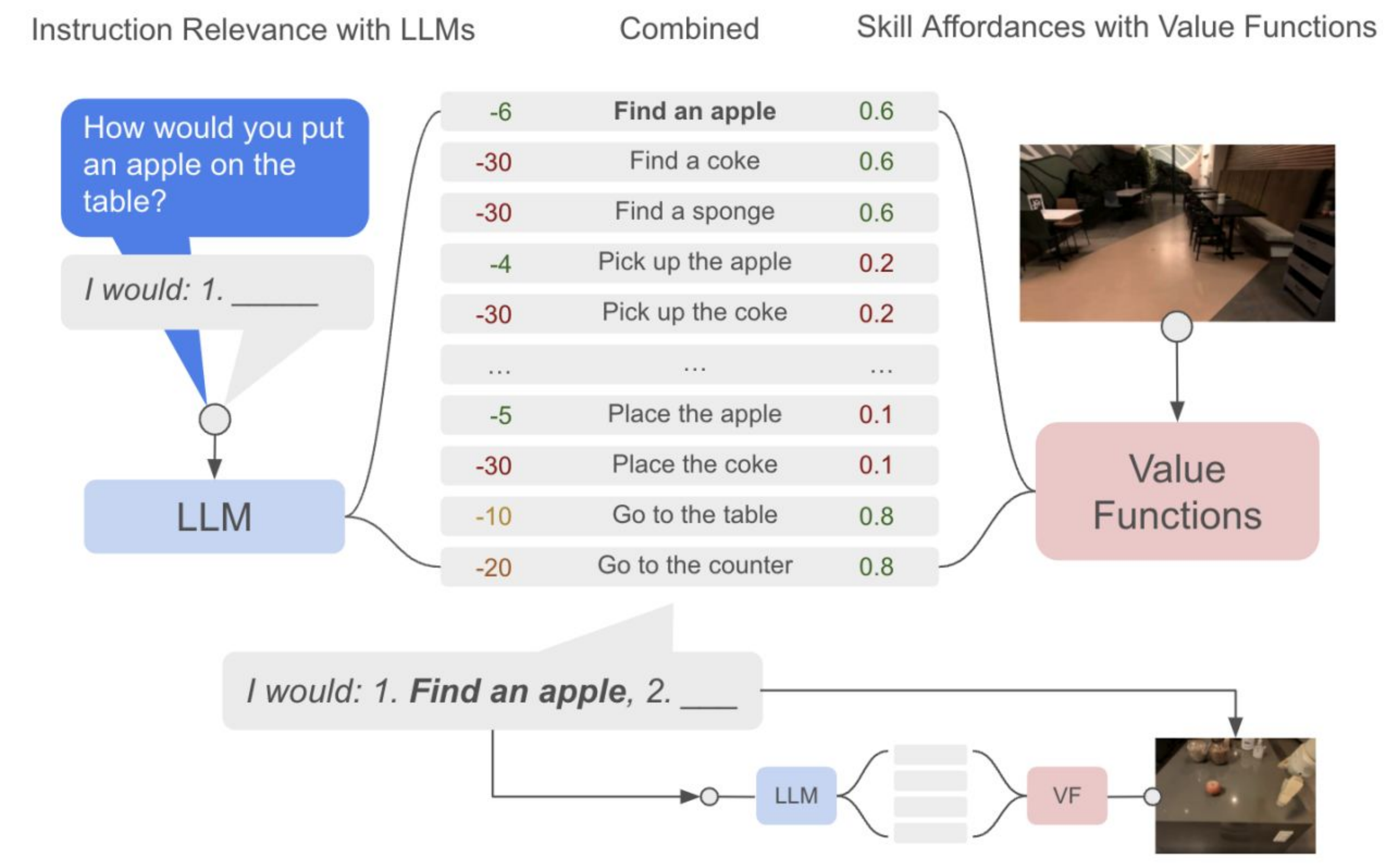
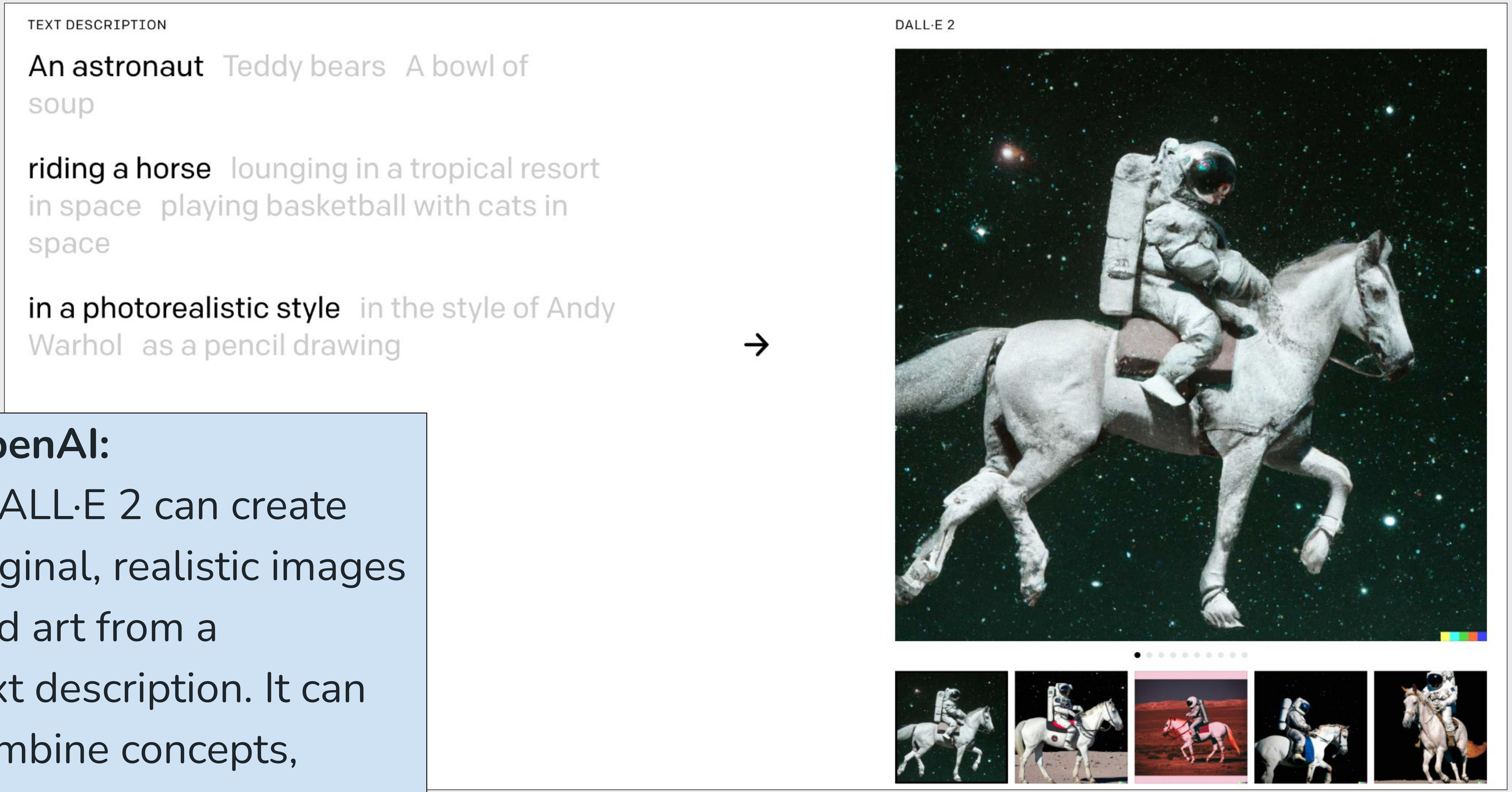


Figure 4: Given a high-level instruction, SayCan combines probabilities from a language model (representing the probability that a skill is useful for the instruction) with the probabilities from a value function (representing the probability of successfully executing said skill) to select the skill to perform. This emits a skill that is both possible and useful. The process is repeated by appending the selected skill to the robot response and querying the models again, until the output step is to terminate.

Ref: Ahn, Michael, *et al.* "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances." arXiv preprint arXiv:2204.01691 (2022).

DALL-E 2: Image Generation from Text



OpenAI:
“DALL·E 2 can create original, realistic images and art from a text description. It can combine concepts, attributes, and styles.“

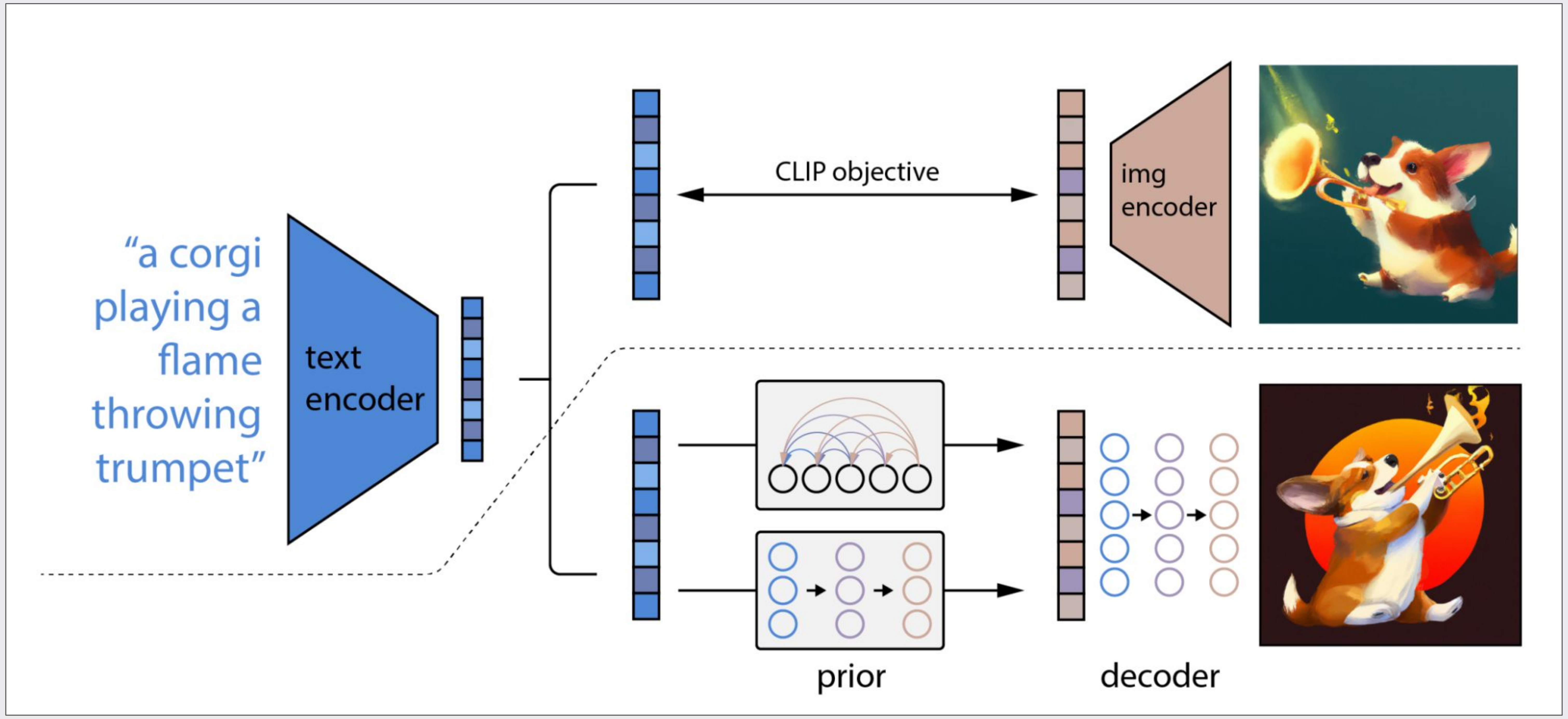
https://openai.com/dall-e-2/

Ref: Ramesh, Aditya, *et al.* "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 (2022)..

DALL-E 2: Image Generation from Text

Abstract Snippet

"We propose a two-stage model: a prior that generates a CLIP image embedding given a text caption, and a decoder that generates an image conditioned on the image embedding."



Ref: Ramesh, Aditya, *et al.* "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 (2022)..

Perils of AI Systems

Areas of Risk for AI Systems

- Security and Robustness
- Privacy
- Fairness
- Bias/Toxicity Reduction and Mitigation
- Ethical Considerations
- Explainability
- Environmental Costs of Large Neural Network Models
- Misinformation/Disinformation

Security and Robustness

- Cybersecurity defenses against hacking and phishing
- Example: OPM hack
- Robustness against adversarial attacks
  - Exploratory attacks attempting to determine how the AI model works
  - Poison attacks that inject incorrect or noisy data during training
  - Evasion/confusion attacks that distort the real-time sensor data to confuse the AI model
- Key research: generative adversarial networks (GANs)

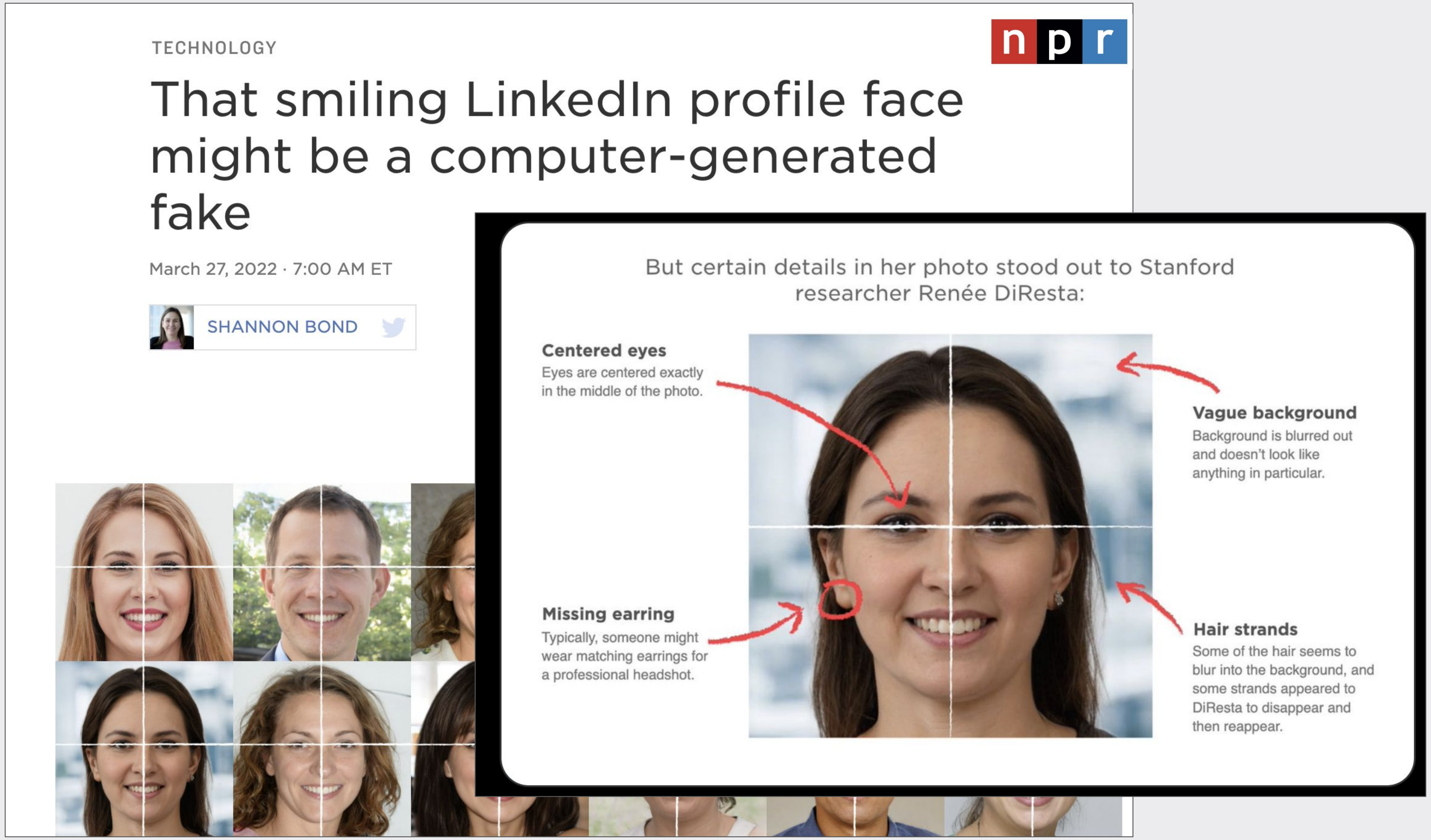
Physical Adversarial Patches

Physical adversarial patches can be generated and printed to confuse computer vision models, e.g., self-driving cars.

Fig: Braunegg, A., *et al.* "Apricot: A dataset of physical adversarial attacks on object detection." European Conference on Computer Vision. Springer, Cham, 2020.

Challenges of Detecting Deep Fakes

Advances in GANs have made it easier for bad actors to create fake images and videos.



Source: Shannon Bond, https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles
Stanford Internet Observatory (Renee DiResta and Josh Goldstein)

New Method for Deep Fake Detection

Existing Deep Fake detectors work well when they are tested against images created by the same GAN model upon which they were trained.

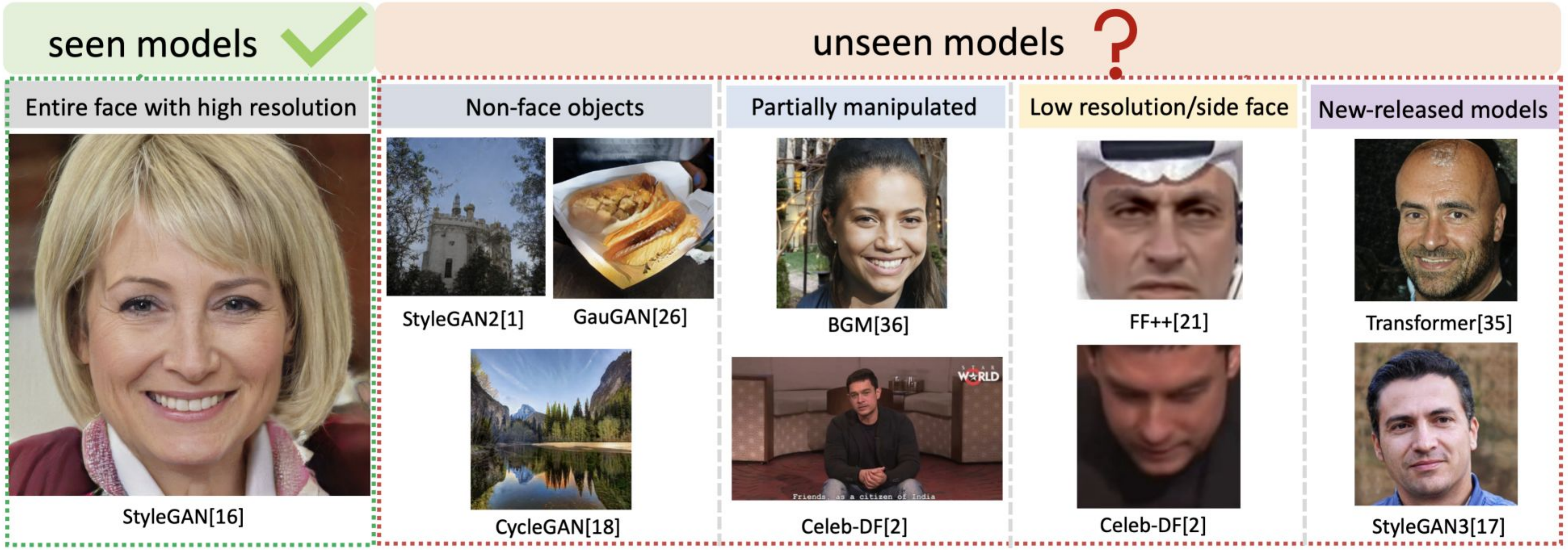


Fig. 1. *Detectors trained with high resolution frontal face images generated by the seen model have high accuracies. What about the images with various resolutions, objects and manipulation types from unseen models?*

Source: Ju, Yan, *et al.* "Fusing Global and Local Features for Generalized AI-Synthesized Image Detection." arXiv preprint arXiv:2203.13964 (2022).

New Method for Deep Fake Detection

This new *(March 26, 2022!)* approach uses both global and local features to detect Deep Fake images.

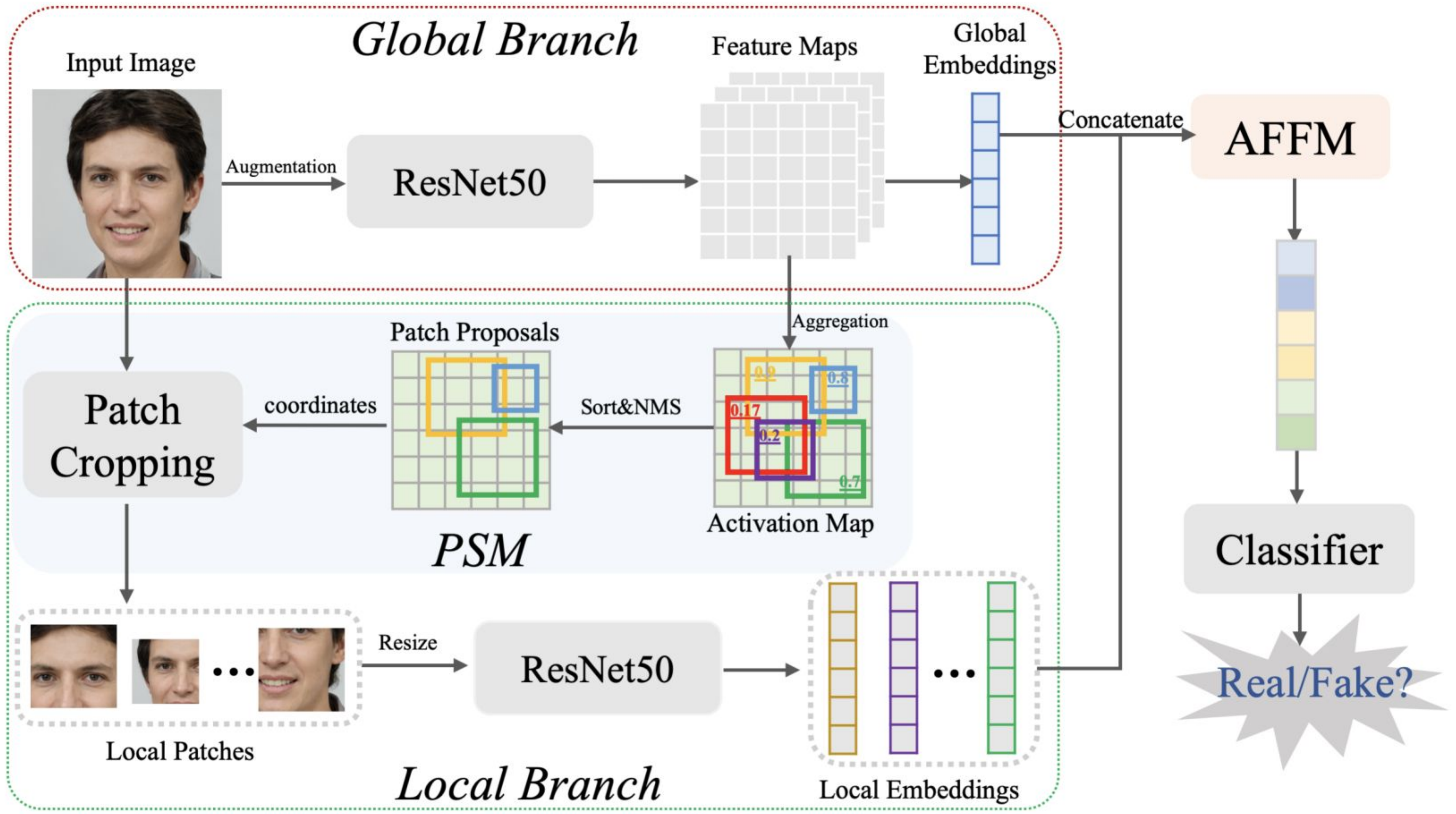


Fig. 2. ***The architecture of our model.*** *The global branch extracts spatial feature and the local branch extracts subtle feature from local patches selected by Patch Selection Module (PSM). The Attention-based Feature Fusion Module (AFFM) fuses global and local features for the classification.*

Source: Ju, Yan, *et al.* "Fusing Global and Local Features for Generalized AI-Synthesized Image Detection." arXiv preprint arXiv:2203.13964 (2022).

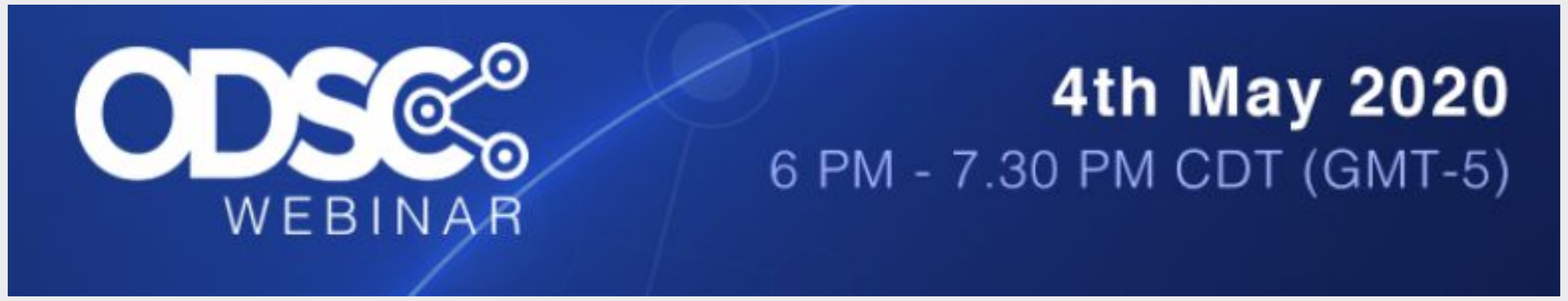
Challenges of Misinformation & Disinformation

There is no question that disinformation is widespread. Research we supported from Professor Jacob Shapiro at Princeton, updated this month, cataloged 96 separate foreign influence campaigns targeting 30 countries between 2013 and 2019. These campaigns, carried out on social media, sought to defame notable people, persuade the public or polarize debates. While 26% of these campaigns targeted the U.S., other countries targeted include Armenia, Australia, Brazil, Canada, France, Germany, the Netherlands, Poland, Saudi Arabia, South Africa, Taiwan, Ukraine, the United Kingdom and Yemen. Some 93% of these campaigns included the creation of original content, 86% amplified pre-existing content and 74% distorted objectively verifiable facts. Recent reports also show that disinformation has been distributed about the COVID-19 pandemic, leading to deaths and hospitalizations of people seeking supposed cures that are actually dangerous.

What we're announcing today is an important part of Microsoft's Defending Democracy Program, which, in addition to fighting disinformation, helps to protect voting through ElectionGuard and helps secure campaigns and others involved in the democratic process through AccountGuard, Microsoft 365 for Campaigns and Election Security Advisors. It's also part of a broader focus on protecting and promoting journalism as Brad Smith and Carol Ann Browne discussed in their Top Ten Tech Policy Issues for the 2020s.

Credit: Tom Burt and Eric Horvitz "New Steps to Combat Disinformation" Microsoft Blog (2020)

My ODSC Talk from May 2020



Identifying Viral Bots and Cyborgs:

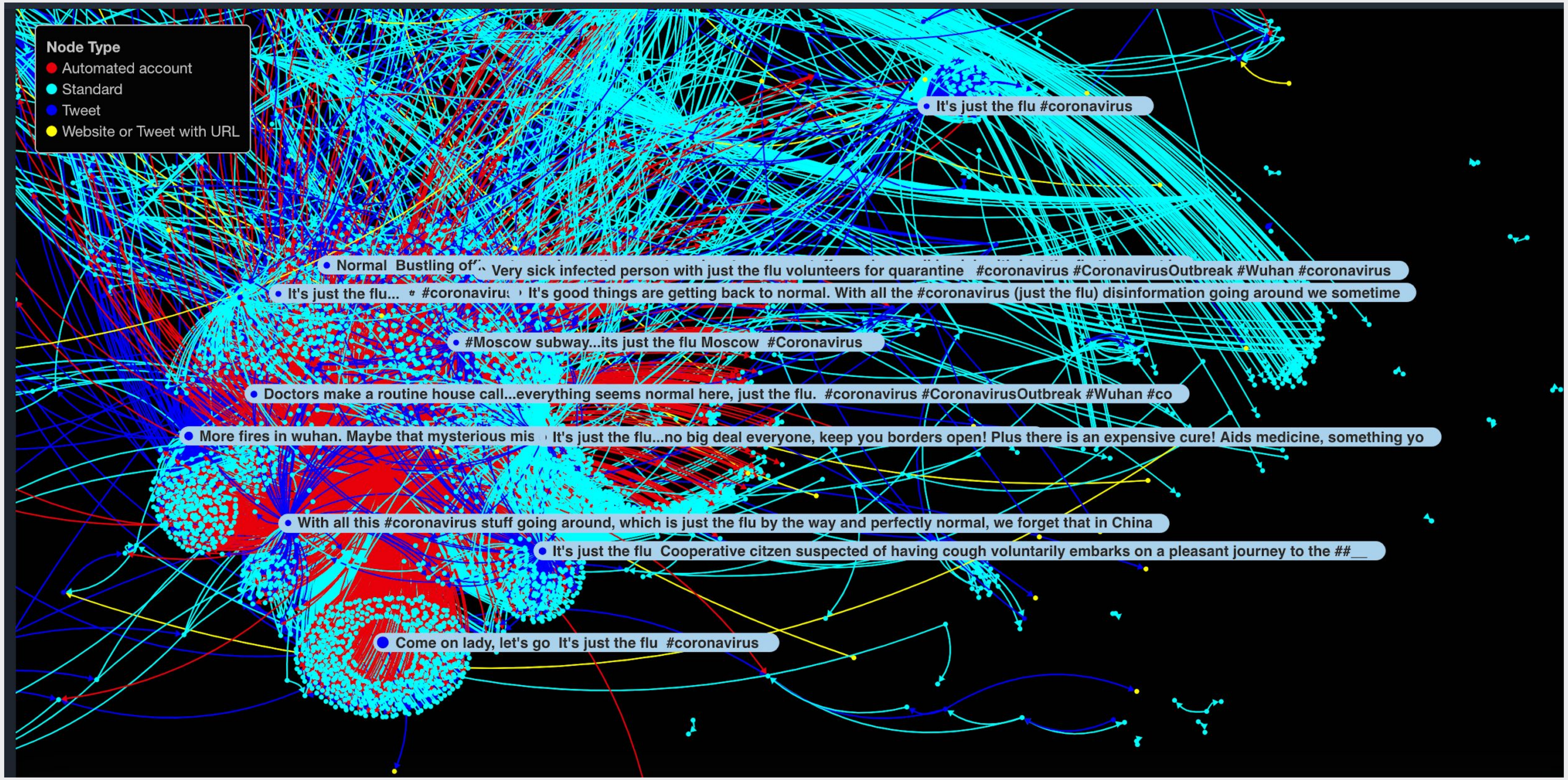
***A Physicist's Journey from
Chaos Theory to
Disinformation Research and AI***

Dr. Steve Kramer
Chief Scientist, KUNGFU.AI

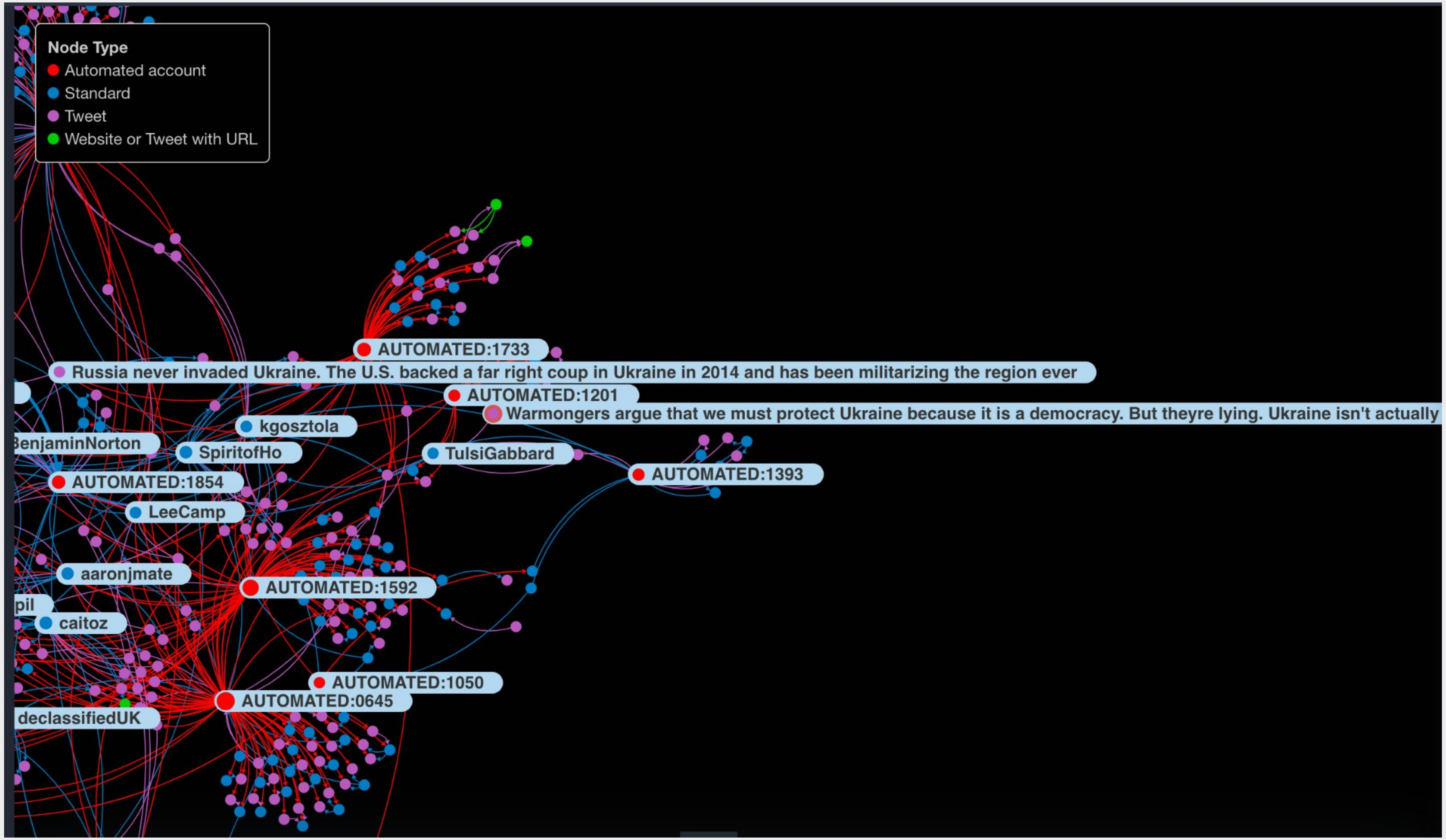
Talk video available at https://bit.ly/KFBotsCyborgsVideo
Slides available at https://bit.ly/KFCOVID19BotsCyborgs

Example of COVID-19 Disinformation: “Just the Flu” from 2020

Interactive Polinode network visualization: http://bit.ly/COVID19BotsKFAI



Russia/Ukraine Twitter Automated Accounts in 2022



Example Tweets by Pro-Russian/Anti-Ukraine Accounts

🥚 | EDUCATE. ORGANIZE. AGITATE. | ABOLISH NATO | 沈 @.. · 27m ···

Ukrops literally can't help themselves
The NATO puppet Ukrainian "State" is literally just a bunch of nazis pretending to be a country

This is why we must #abolishNATO

> **Robert D Skeels, JD, Esq** 🔻 **Free** 🇵🇸 @rdsathene · 9h
> Replying to @NowImNothing_
> 🤔 #NoWar #abolishNATO #nonazis #antiazov

[redacted] Retweeted

Workers Party of Britain
@WorkersPartyGB ···

Russia has every right to defend its own people in the Ukraine against these evil aggressive nazis.

Did our Grandparents pay the ultimate sacrifice for nothing.

The USSR lost 27million lives defeating the fascists in WW2, don't ever forget that.

Privacy Considerations

- Privacy-related laws and regulations
  - HIPAA
  - GDPR in the EU
  - CCPA in California
- Different taxonomies of sensitive data, including PII (personally identifiable information)
- Key challenges
  - Detection
  - Storage, access control, and logging
  - Redaction
- Use in training and testing AI models
- Key research areas: differential privacy and federated learning

Privacy-Preserving Deep Learning

The Private Aggregation of Teacher Ensembles (PATE) method combines the results of Teacher models trained on subsets of confidential through noisy voting that controls the final Student model.

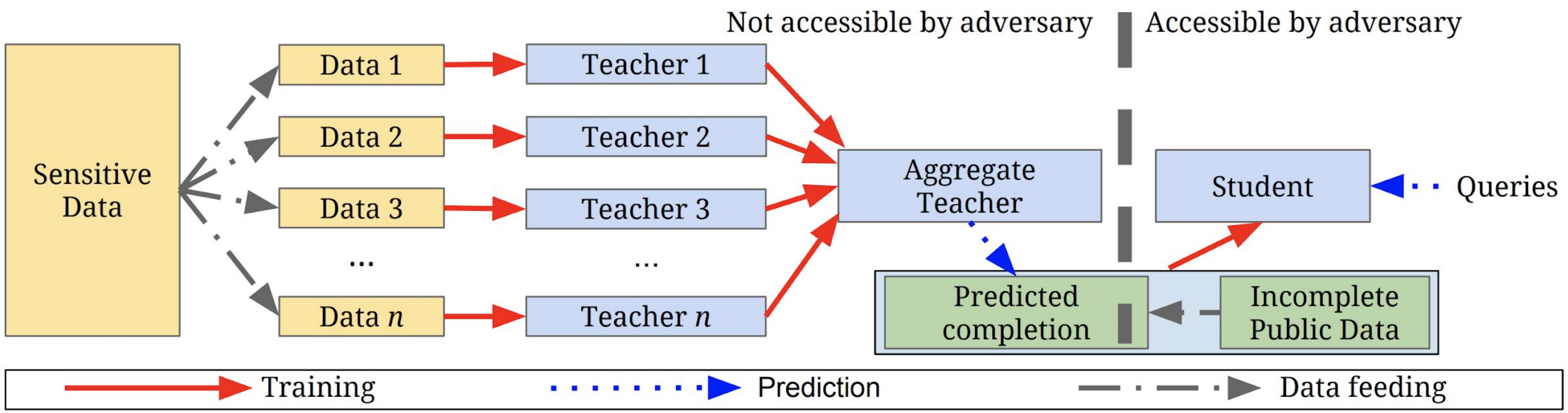


Figure 1: Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble.

Fig: Papernot, Nicolas, *et al.* "Semi-supervised knowledge transfer for deep learning from private training data." arXiv preprint arXiv:1610.05755 (2016).

Fairness Considerations in AI

- Many possible definitions of fairness: 21 fairness definitions and their politics given at ACM FAT* (Fairness, Accountability and Transparency) Conference in 2018 by Prof. Arvind Narayanan (https://www.youtube.com/watch?v=jIXIuYdnyyk)
    - Group fairness
    - Individual fairness
    - Process fairness vs. outcome (utility) fairness
- Applicable metrics depend on the fairness definitions
- Example scenarios:
    - College admission based on SAT scores
    - Mortgage lending decisions
    - Credit ratings

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. "Fairness in machine learning." Nips tutorial 1 (2017).
http://www.fairmlbook.org

Types of Bias in AI Systems

- Stereotyping, prejudice or favoritism towards some things, people, or groups over others
  - automation bias
  - confirmation bias
  - experimenter’s bias
  - group attribution bias
- Systematic error introduced by a sampling or reporting procedure
  - coverage bias
  - non-response bias
  - participation bias
  - reporting bias
  - sampling bias
  - selection bias
- NOT to be confused with prediction bias in machine learning (e.g., bias vs. variance)

Source: Google’s Machine Learning Glossary

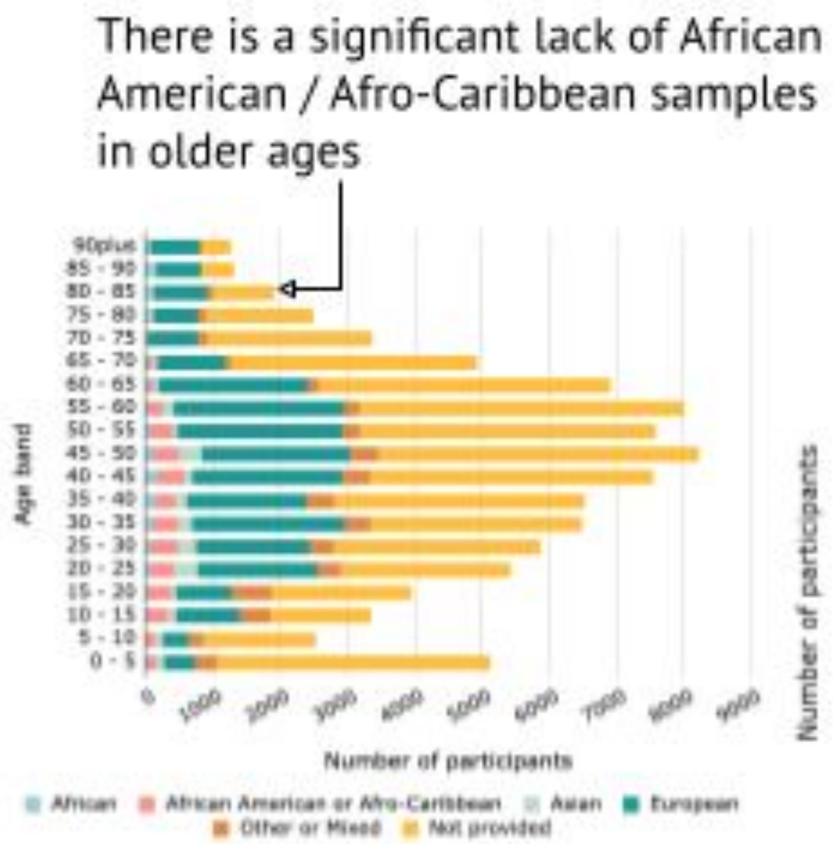
Bias in Healthcare AI Models

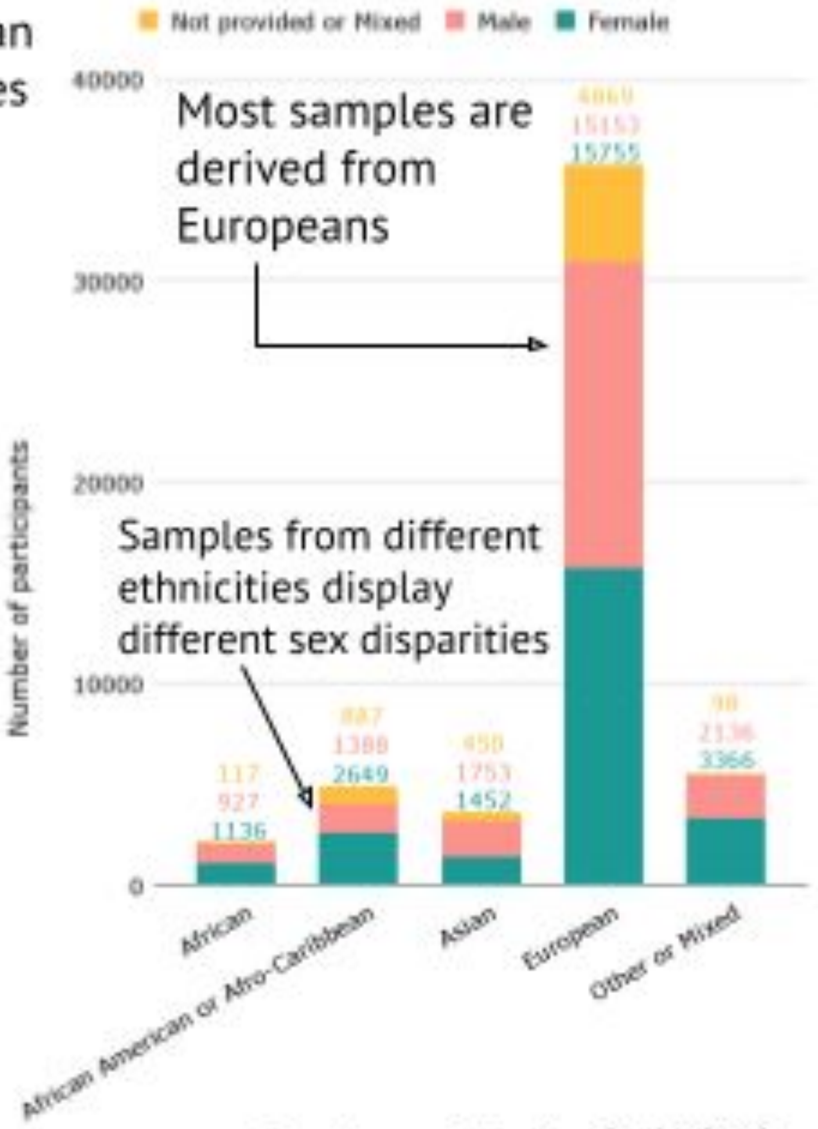
Introduction | **Research** | Talent | Industry | Politics | Predictions **#stateofai | 53**

Measuring bias: a first step towards more inclusive health research outcomes

Missing information and biases in demographic information are widespread in biomedical data that form the basis of the drug discovery process. ML solutions trained on these data need to understand and adapt for these biases to avoid perpetuating health inequities.

- Demographic factors (e.g. age, sex, ethnicity) can influence patient outcomes based on their association with long-standing healthcare and societal inequities or, although less common, can change the efficacy of drugs.
- An analysis of gene expression read-outs from disease relevant tissue samples across 3,000 studies comprising 177,201 individual samples found that many missed information on age (48%), sex (40%) and ethnicity (71%).
- There was a significant lack of non-European samples from older donors, as well as varying sex distributions across different ethnicities.





Benevolent[AI] DAT

stateof.ai 2021

Source: Benaich, Nathan, and Ian Hogarth. "[State of AI Report]." London, United Kingdom (2021).

Bias & Toxicity in Language Models

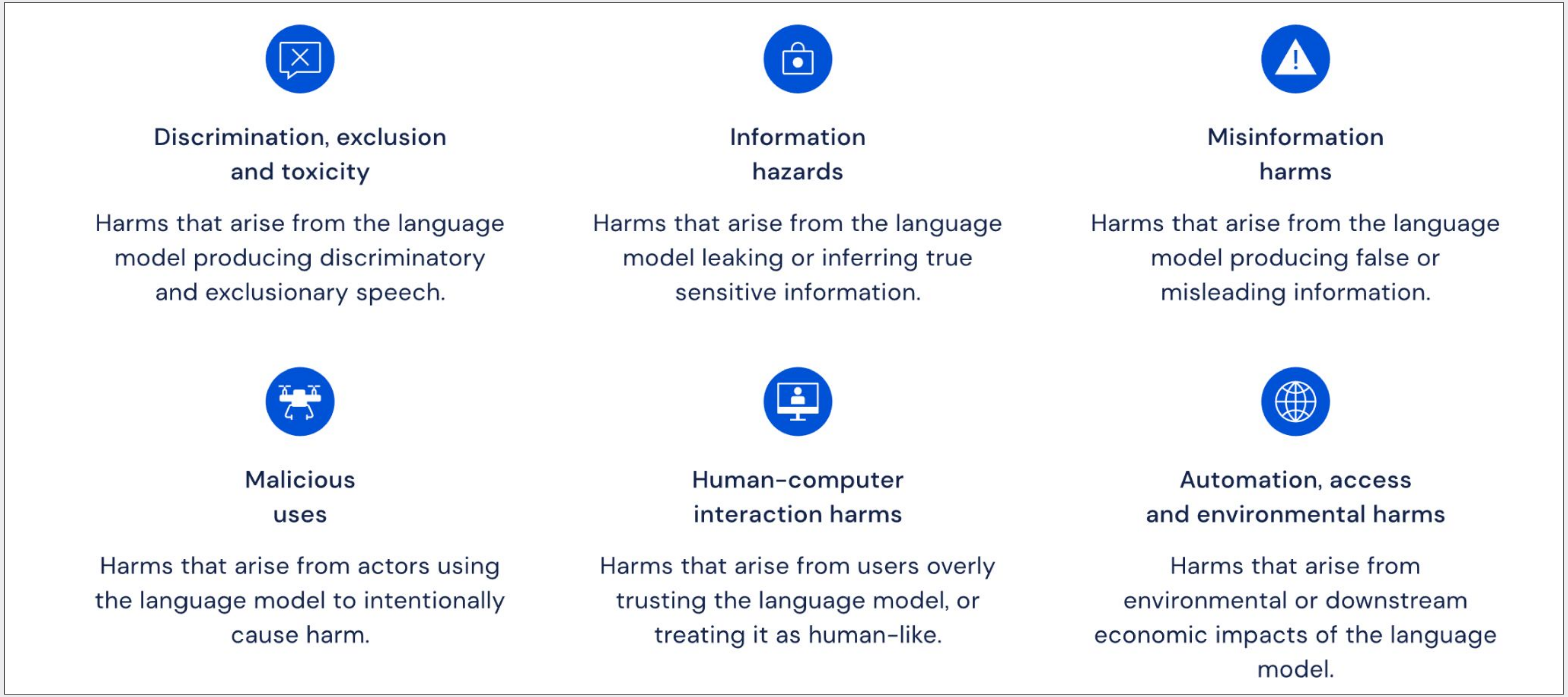
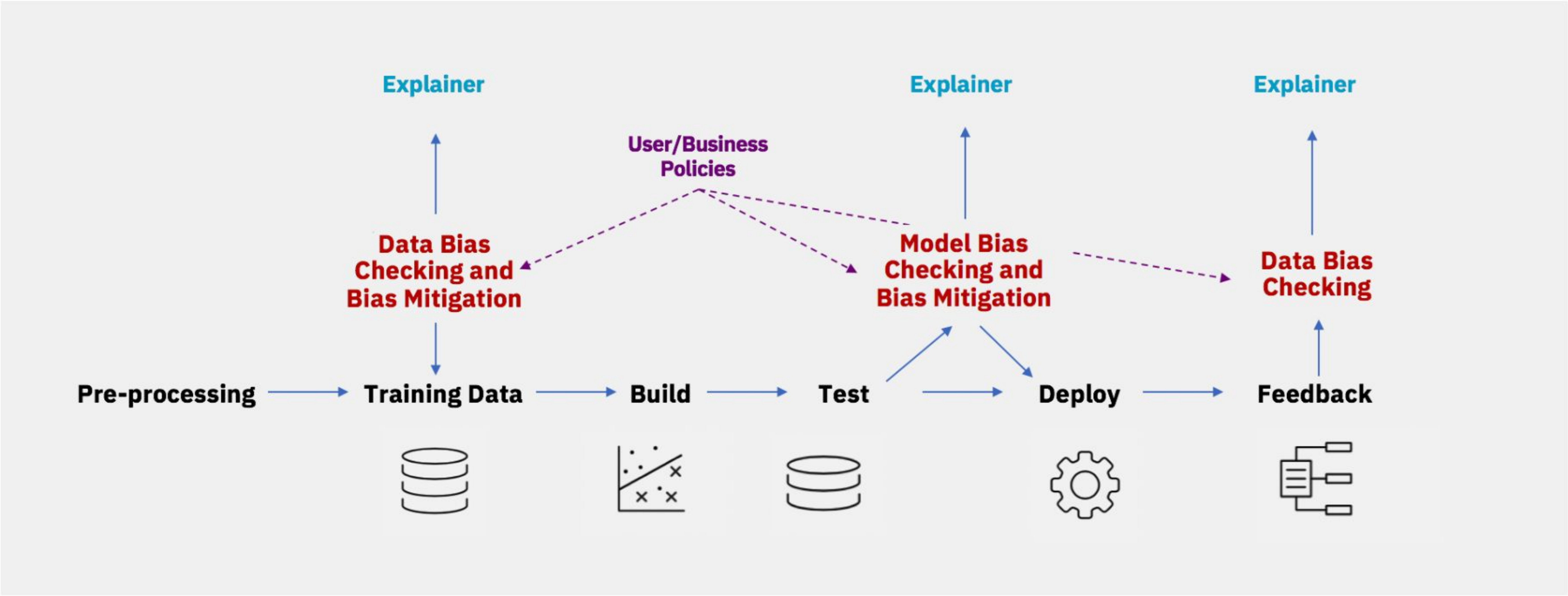


Fig: "Language modelling at scale: Gopher, ethical considerations, and retrieval." Deepmind Blog (2021).

Ref: Weidinger, Laura, *et al*. "Ethical and social risks of harm from Language Models." arXiv preprint arXiv:2112.04359 (2021).

Open-Source Fairness Tools

Example: IBM's AI Fairness 360



Mitigating bias throughout the AI lifecycle

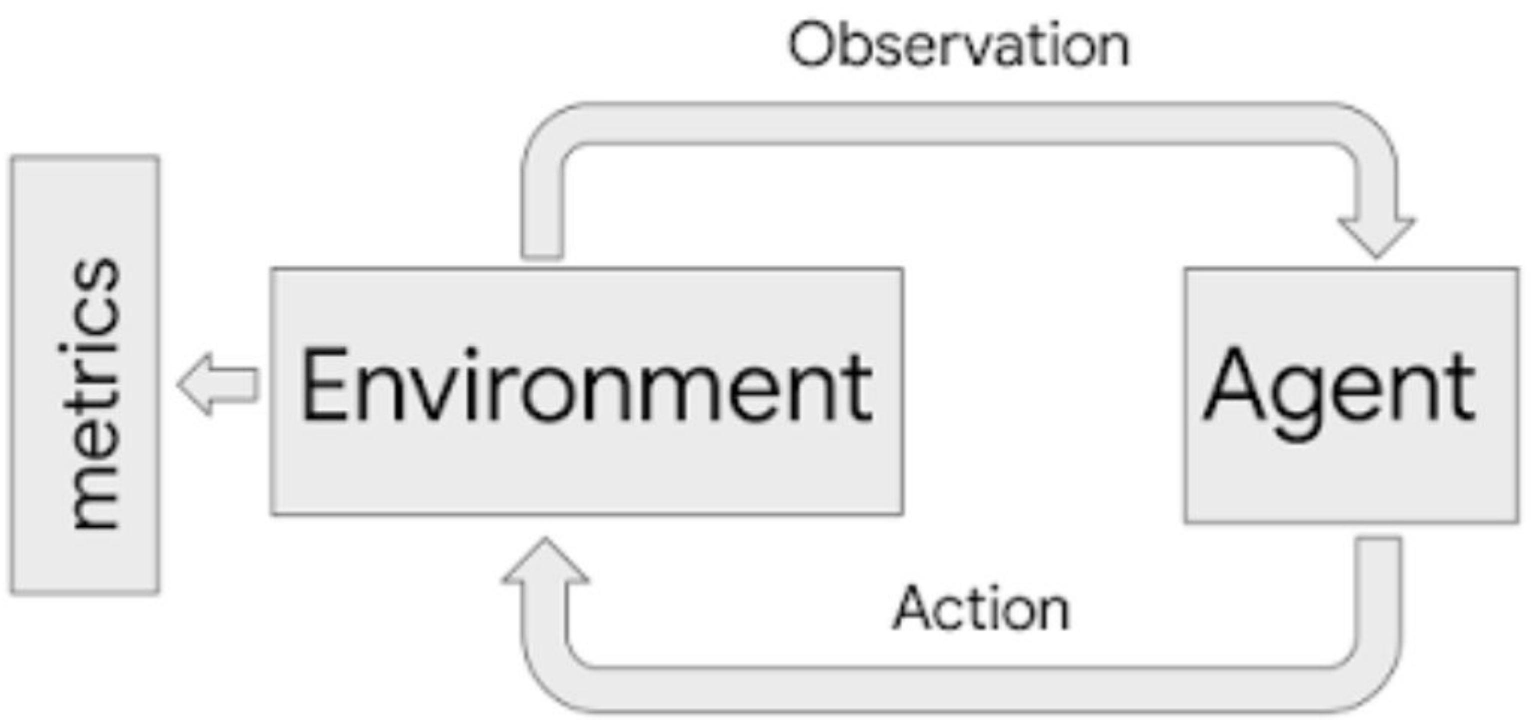
Ref: K. E. Bellamy *et al*., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," in *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1-4:15, 1 July-Sept. 2019, doi: 10.1147/JRD.2019.2942287.

Open-Source Fairness Tools

Example: [Google's ML-fairness-gym]

ML-fairness-gym as a Simulation Tool for Long-Term Analysis

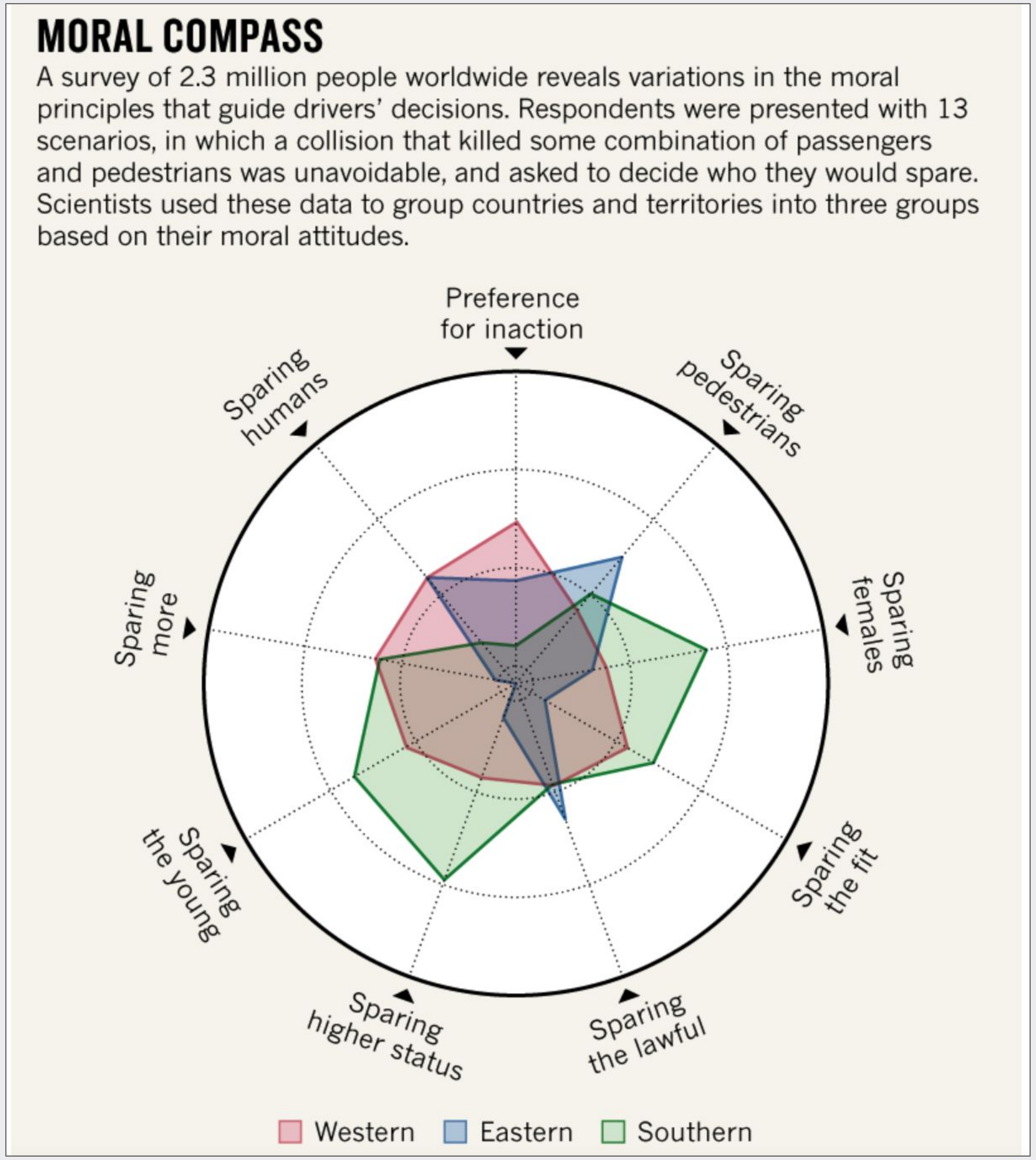
The ML-fairness-gym simulates sequential decision making using Open AI's Gym framework. In this framework, *agents* interact with simulated *environments* in a loop. At each step, an agent chooses an *action* that then affects the environment's state. The environment then reveals an *observation* that the agent uses to inform its subsequent actions. In this framework, environments model the system and dynamics of the problem and observations serve as data to the agent, which can be encoded as a machine learning system.



Ref: D'Amour, Alexander, *et al*. "Fairness is not static: deeper understanding of long term fairness via simulation studies." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020..

Example: Ethics in Autonomous Vehicles

“The largest ever survey of machine ethics1, published today in Nature, finds that many of the moral principles that guide a driver’s decisions vary by country. For example, in a scenario in which some combination of pedestrians and passengers will die in a collision, people from relatively prosperous countries with strong institutions were less likely to spare a pedestrian who stepped into traffic illegally.”



Ref: Maxmen, Amy. "Self-driving car dilemmas reveal that moral choices are not universal." *Nature* 562.7728 (2018): 469-469..

Advances in Algorithmic Accountability

This framework was published in January 2020 as a collaboration between Google and the Partnership on AI and represents a valuable tool in responsible AI efforts.

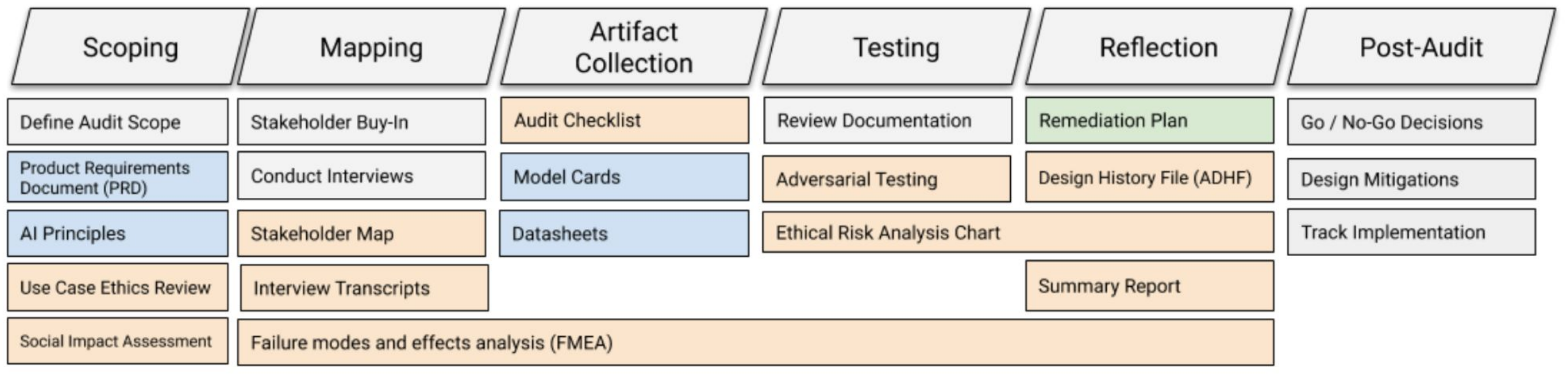


Figure 2: Overview of Internal Audit Framework. Gray indicates a process, and the colored sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed.

Fig: Raji, Inioluwa Deborah, et al. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.

Advances in Algorithmic Accountability

"AI has the potential to benefit the whole of society," the paper reads. "[H]owever there is currently an inequitable risk distribution such that those who already face patterns of structural vulnerability or bias disproportionately bear the costs and harms of many of these systems. Fairness, justice and ethics require that those bearing these risks are given due attention and that organizations that build and deploy artificial intelligence systems internalize and proactively address these social risks as well, being seriously held to account for system compliance to declared ethical principles."

Fig: Raji, Inioluwa Deborah, et al. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.

Regulation of AI Algorithms

FEDERAL TRADE COMMISSION
PROTECTING AMERICA'S CONSUMERS

The question, then, is how can we harness the benefits of AI without inadvertently introducing bias or other unfair outcomes? Fortunately, while the sophisticated technology may be new, the FTC's attention to automated decision making is not. The FTC has decades of experience enforcing three laws important to developers and users of AI:

- **Section 5 of the FTC Act.** The FTC Act prohibits unfair or deceptive practices. That would include the sale or use of – for example – racially biased algorithms.
- **Fair Credit Reporting Act.** The FCRA comes into play in certain circumstances where an algorithm is used to deny people employment, housing, credit, insurance, or other benefits.
- **Equal Credit Opportunity Act.** The ECOA makes it illegal for a company to use a biased algorithm that results in credit discrimination on the basis of race, color, religion, national origin, sex, marital status, age, or because a person receives public assistance.

Among other things, the FTC has used its expertise with these laws to report on big data analytics and machine learning; to conduct a hearing on algorithms, AI and predictive analytics; and to issue business guidance on AI and algorithms. This work – coupled with FTC enforcement actions – offers important lessons on using AI truthfully, fairly, and equitably.

Credit: https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai

Harvard Business Review

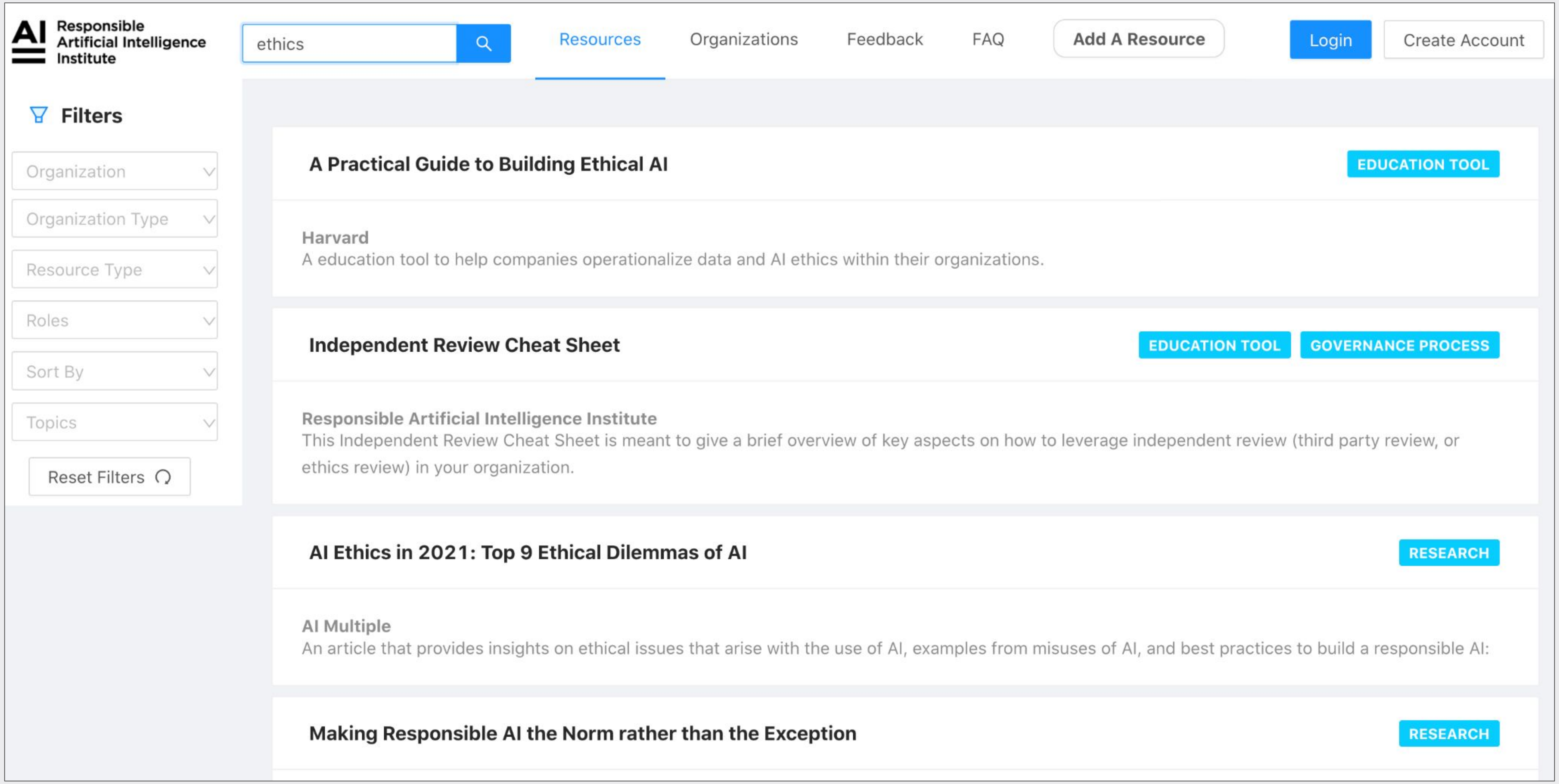
AI And Machine Learning | **AI Regulation Is Coming**

screener, which filtered out female candidates. A recent study published in *Science* showed that risk prediction tools used in health care, which affect millions of people in the United States every year, exhibit significant racial bias. Another study, published in the *Journal of General Internal Medicine,* found that the software used by leading hospitals to prioritize recipients of kidney transplants discriminated against Black patients.

AI increases the potential scale of bias: Any flaw could affect millions of people, exposing companies to class-action lawsuits.

Credit: https://hbr.org/2021/09/ai-regulation-is-coming

Responsible AI Community Portal



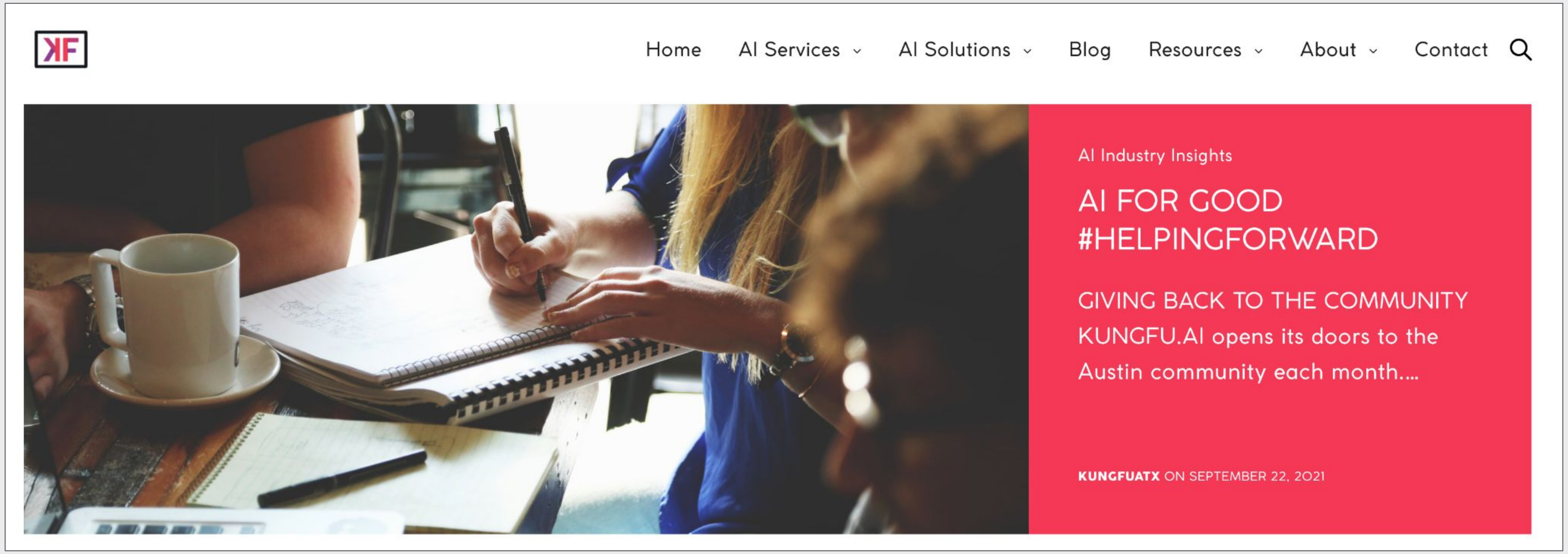
https://portal.ai-global.org/

Thanks,
Tina Lassiter!

Resources

AI for Good Initiatives at KUNGFU.AI

- KUNGFU.AI would love to support community members, nonprofits, and educational institutions that need help with AI.
  https://www.kungfu.ai/ai-for-good/
- Please reach out to us at info@kungfu.ai!



Public Data for Social Good

- COVID-19 Data Resource Hub
  - https://data.world/resources/coronavirus/
  - Swift aggregation of data early on
- Policing in America
  - https://www.datafoundation.org/policing-in-america
  - Evaluating the nexus of open data and perception
  - Legislative work to change how data mandates function
- US Healthcare Pricing
  - https://data.world/ushealthcarepricing
  - Fighting malicious compliance

AI Industry and Ethics Resources and Reports

- Nathan Benaich & Ian Hogarth, "State of AI Report" (2021)
- Daniel Zhang, et al., "The AI Index 2022 Annual Report," AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University (2022)
- Montreal Ethics, "The State of AI Ethics Report" (2021)
- Gradient Flow
  - Newsletter
  - Reports
- Derwen.ai (Paco Nathan)
  - AI in Healthcare 2022
- Paperswithcode.com

Recap

Q&A

KUNGFU.AI

Thank You!

steve.kramer@kungfu.ai