CSE 573 P:
Artificial Intelligence

Hanna Hajishirzi

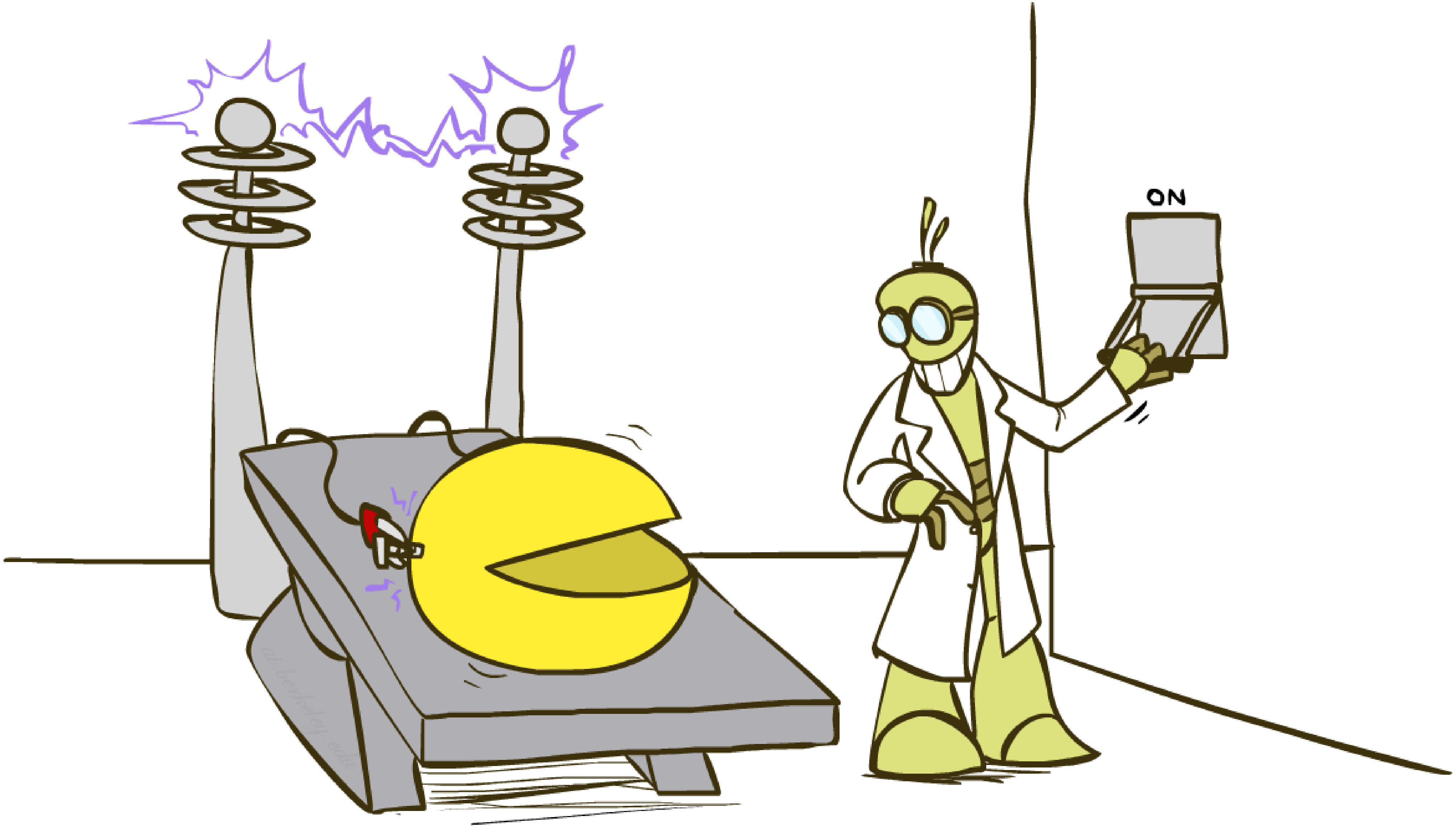
slides adapted from
Dan Klein, Pieter Abbeel ai.berkeley.edu
And Dan Weld, Luke Zettlemoyer

ai.berkeley.edu

Topics in This Course

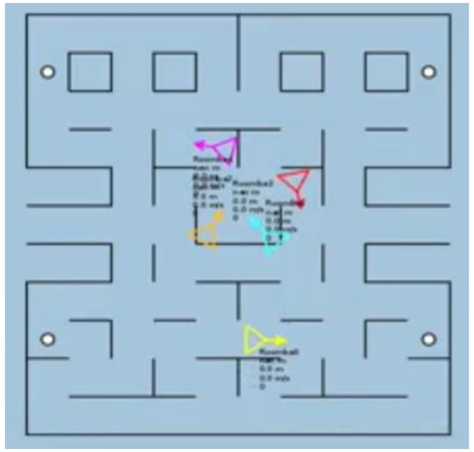
- Part I: Intelligence from Computation
  - Fast search
  - Adversarial and uncertain search
- Part II: Reasoning under Uncertainty
  - Decision theory: Reinforcement Learning, Markov Decision Processes
  - Machine learning
  - Graphical Models - Bayes Nets; HMMs
- Throughout: Applications
  - Natural language, vision, robotics, games, …

Pac-Man Beyond the Game!



Pacman: Beyond Simulation?

ROOMBA PACMAN



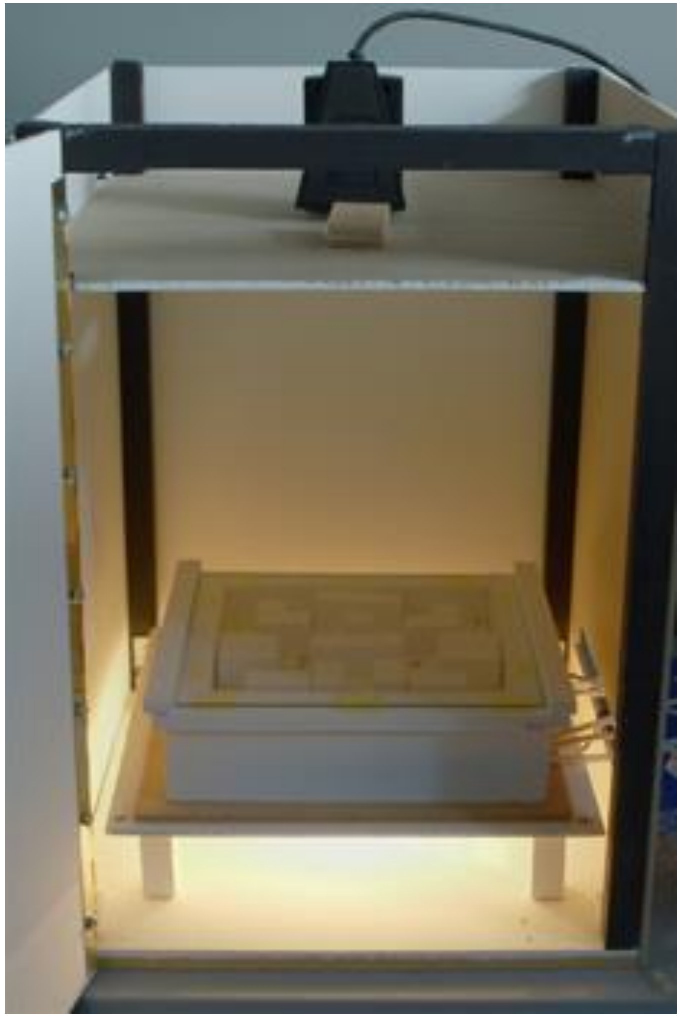
Students at Colorado University: http://pacman.elstonj.com

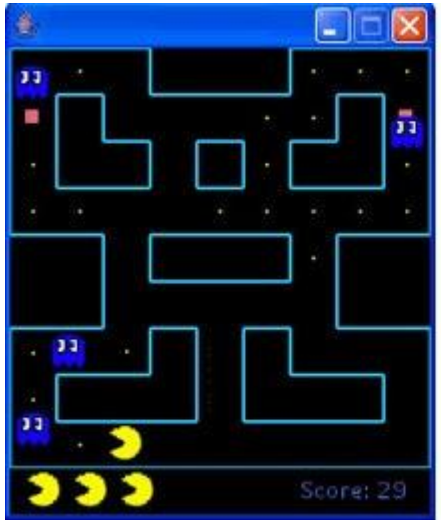
[VIDEO: Roomba Pacman.mp4]

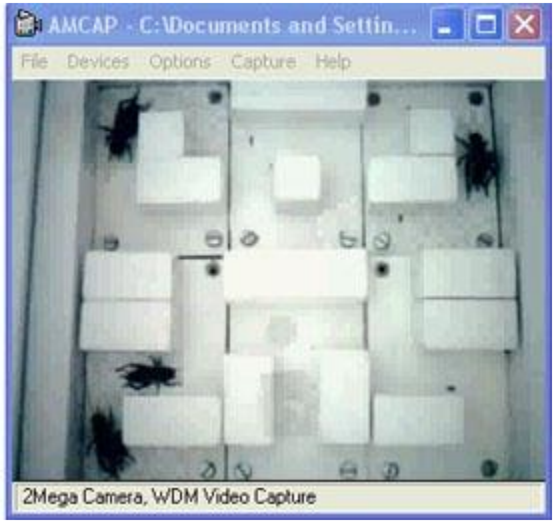
Pacman: Beyond Simulation!

Bugman?

- AI = Animal Intelligence?
  - Wim van Eck at Leiden University
  - Pacman controlled by a human
  - Ghosts controlled by crickets
  - Vibrations drive crickets toward or away from Pacman’s location



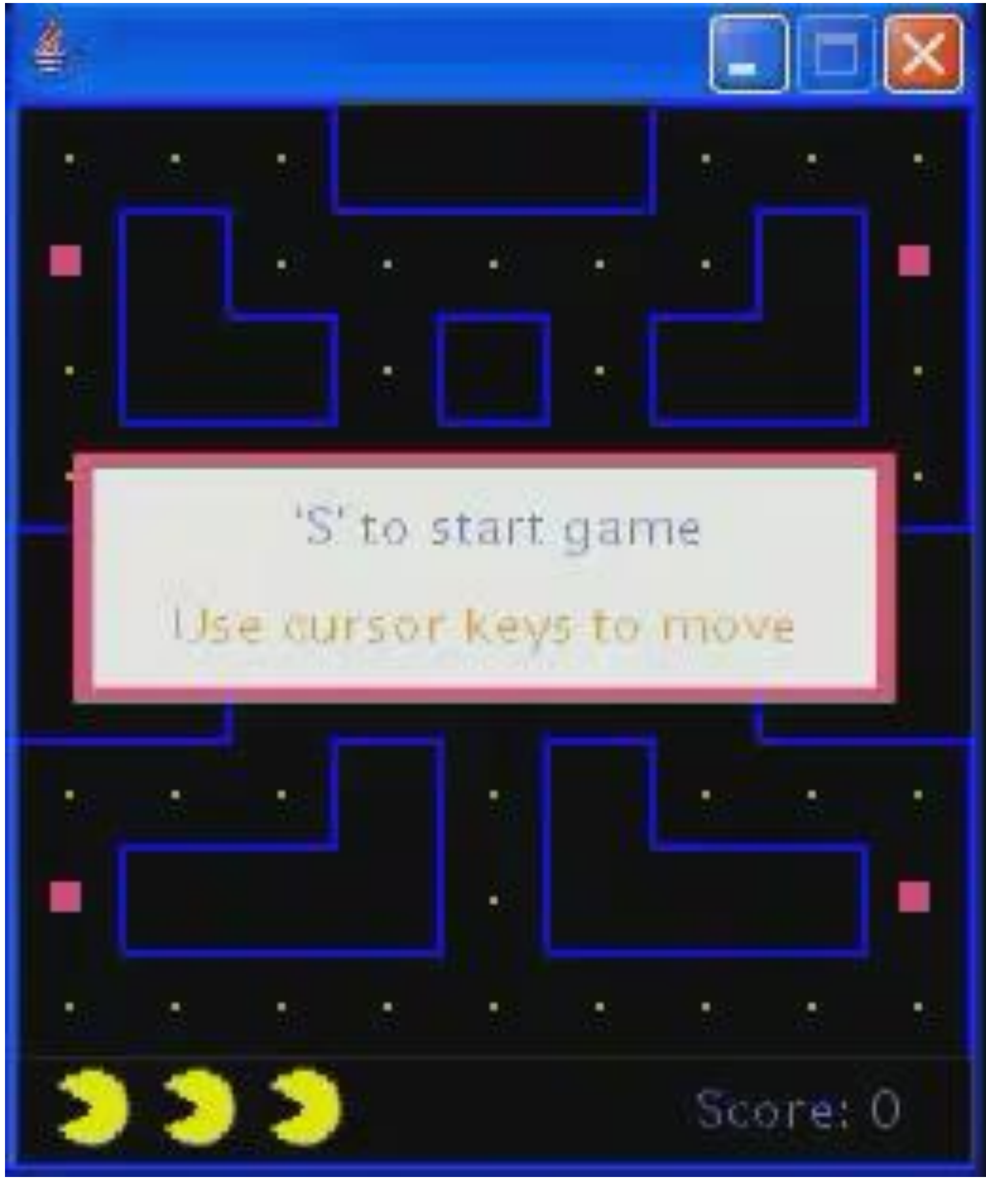



AMCAP - C:\Documents and Settin...
File Devices Options Capture Help
2Mega Camera, WDM Video Capture

http://pong.hku.nl/~wim/bugman.htm

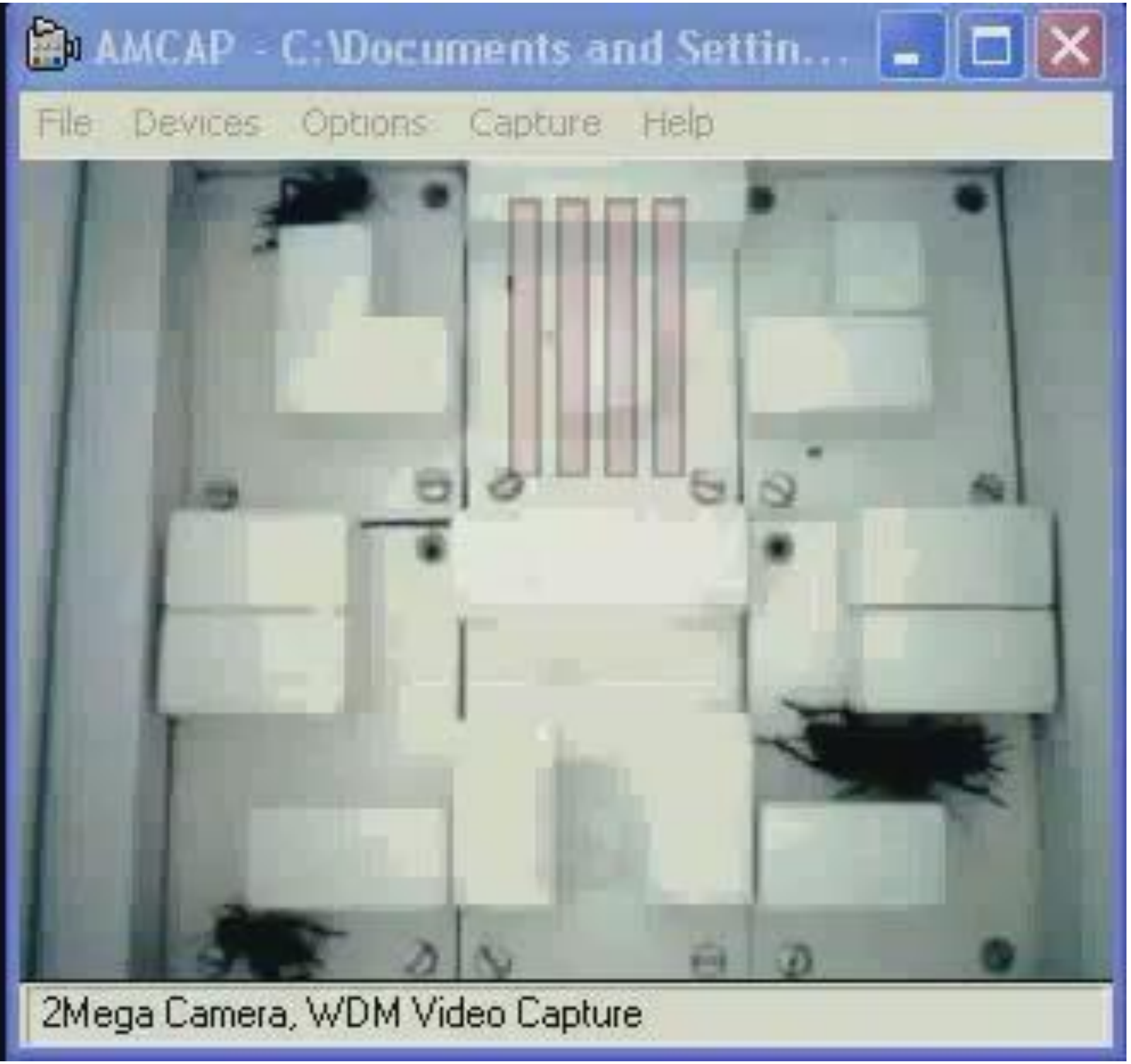
[VIDEO: bugman_movie_1.mov]

Bugman



'S' to start game

Use cursor keys to move



AMCAP - C:\Documents and Settin...
File
Devices
Options
Capture
Help
2Mega Camera, WDM Video Capture

Research Frontiers

- Deep Unsupervised Learning
- AI for Science
- AI and Ethics

Also:

- Unsupervised Deep Reinforcement Learning
- Human-in-the-loop Reinforcement Learning
- …

Research Frontiers

- ***Deep Unsupervised Learning***
- AI for Science
- AI and Ethics

Also:

- Unsupervised Deep Reinforcement Learning
- Human-in-the-loop Reinforcement Learning
- …

Deep Unsupervised Learning

- Key hypothesis:

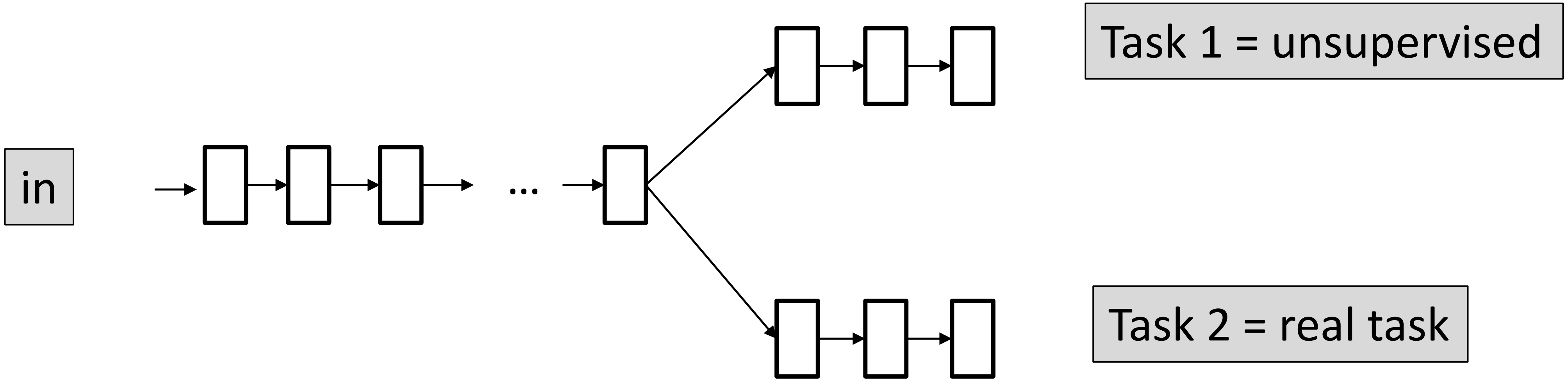
Task 1

  - IF neural network smart enough to predict:
    - Next frame in video
    - Next word in sentence
    - Generate realistic images
    - ``Translate'' images
    - …

Task 2

  - THEN same neural network is ready to do Deep Supervised Learning from very small data-set

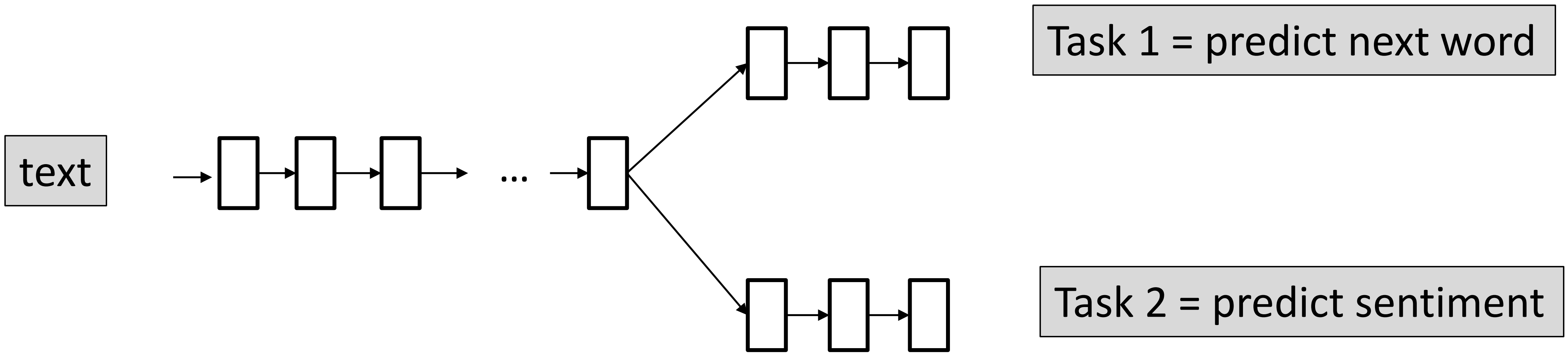
Transfer from Unsupervised Learning



Task 1 = unsupervised

Task 2 = real task

Example Setting



text

…

Task 1 = predict next word

Task 2 = predict sentiment

Next Word Prediction (OpenAI’s GPT-2)

Prompt: A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

GPT-2: *The incident occurred on the downtown train line, which runs from Covington and Ashland stations. In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.*

‘The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,’ said Tom Hicks, the U.S. Energy Secretary, in a statement. ‘Our top priority is to secure the theft and ensure it doesn’t happen again.’

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.

The Nuclear Regulatory Commission did not immediately release any information.

According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

‘The safety of people, the environment and the nation’s nuclear stockpile is our highest priority,’ Hicks said. ‘We will get to the bottom of this and make no excuses.’

Next Word Prediction (OpenAI’s GPT-2)

Prompt: Recycling is good for the world, no, you could not be more wrong.

GPT-2: Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I’m not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world’s most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources.

Text Generati ’s GPT-2)

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking fi
remote, previou
surprising to the
English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist nam
Ovid’s Unicorn. Th
previously unknown

Now, after almost tw
odd phenomenon is fi

Dr. Jorge Pérez, an ev
La Paz, and several co
Mountains when they fou
or humans. Pérez noticed
a natural fountain, surro
snow.

Pérez and the others then
the time we reached the top
with some crystals on top,”

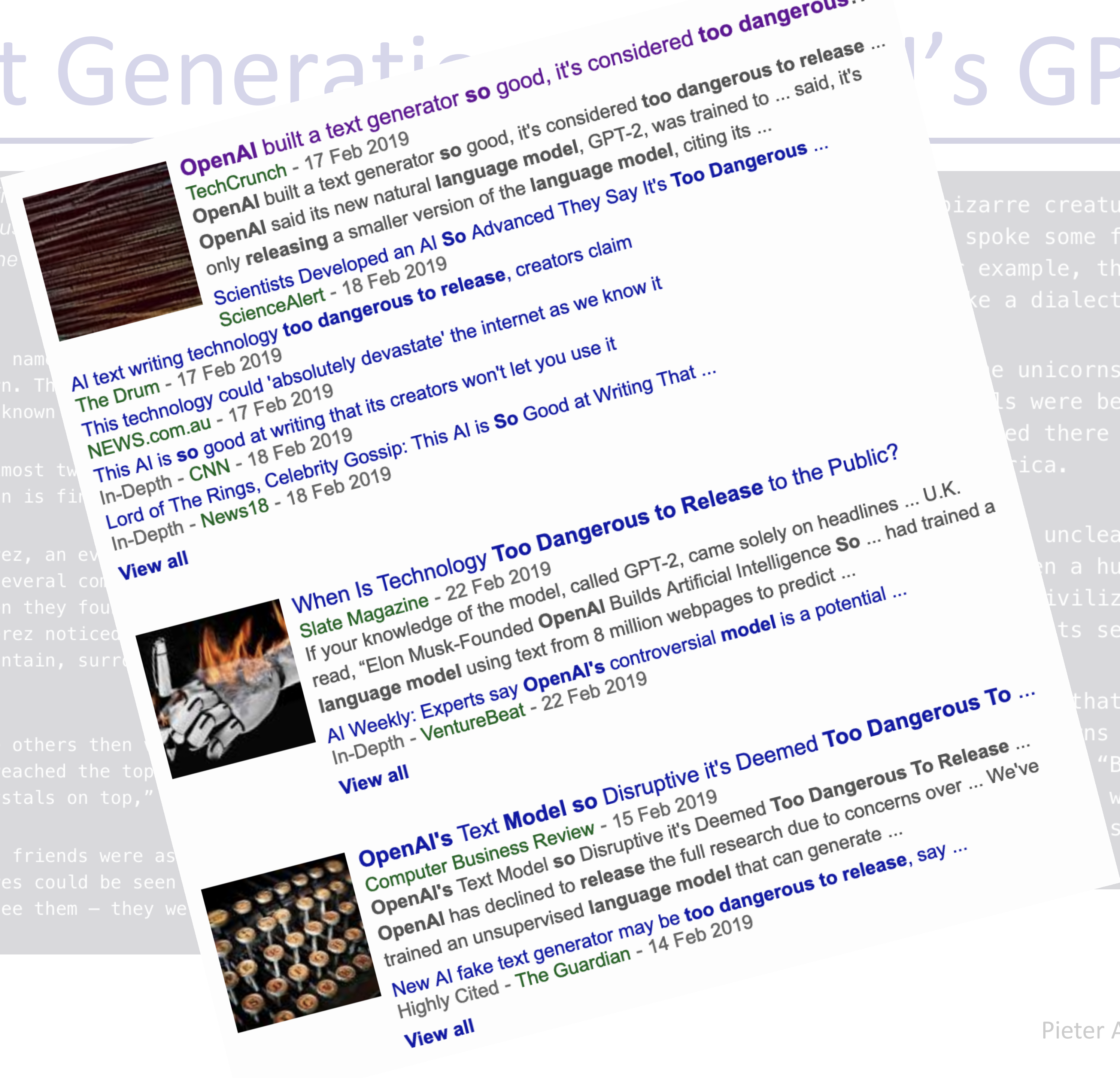
Pérez and his friends were as
These creatures could be seen
too much to see them – they we
horns.

bizarre creatures the scientists discovered
spoke some fairly regular English. Pérez
example, that they have a common
ke a dialect or dialectic.”

e unicorns may have originated in
ls were believed to be descendants of a
ed there before the arrival of humans
ica.

unclear, some believe that perhaps
n a human and a unicorn met each
ivilization. According to Pérez,
ts seem to be quite common.”

that it is likely that the only
ns are indeed the descendants of
“But they seem to be able to
which I believe is a sign of
social organization,” said the



OpenAI built a text generator **so** good, it's considered **too dangerous**...
TechCrunch - 17 Feb 2019
OpenAI built a text generator **so** good, it's considered **too dangerous to release** ... **OpenAI** said its new natural **language model**, GPT-2, was trained to ... said, it's only **releasing** a smaller version of the **language model**, citing its ...

Scientists Developed an AI **So** Advanced They Say It's **Too Dangerous** ...
ScienceAlert - 18 Feb 2019

AI text writing technology **too dangerous to release**, creators claim
The Drum - 17 Feb 2019

This technology could 'absolutely devastate' the internet as we know it
NEWS.com.au - 17 Feb 2019

This AI is **so** good at writing that its creators won't let you use it
In-Depth - CNN - 18 Feb 2019

Lord of The Rings, Celebrity Gossip: This AI is **So** Good at Writing That ...
In-Depth - News18 - 18 Feb 2019

View all

When Is Technology **Too Dangerous to Release** to the Public?
Slate Magazine - 22 Feb 2019
If your knowledge of the model, called GPT-2, came solely on headlines ... U.K. read, “Elon Musk-Founded **OpenAI** Builds Artificial Intelligence **So** ... had trained a **language model** using text from 8 million webpages to predict ...

AI Weekly: Experts say **OpenAI's** controversial **model** is a potential ...
In-Depth - VentureBeat - 22 Feb 2019

View all

OpenAI's Text **Model so** Disruptive it's Deemed **Too Dangerous** To ...
Computer Business Review - 15 Feb 2019
OpenAI's Text Model **so** Disruptive it's Deemed **Too Dangerous To Release** ... **OpenAI** has declined to **release** the full research due to concerns over ... We've trained an unsupervised **language model** that can generate ...

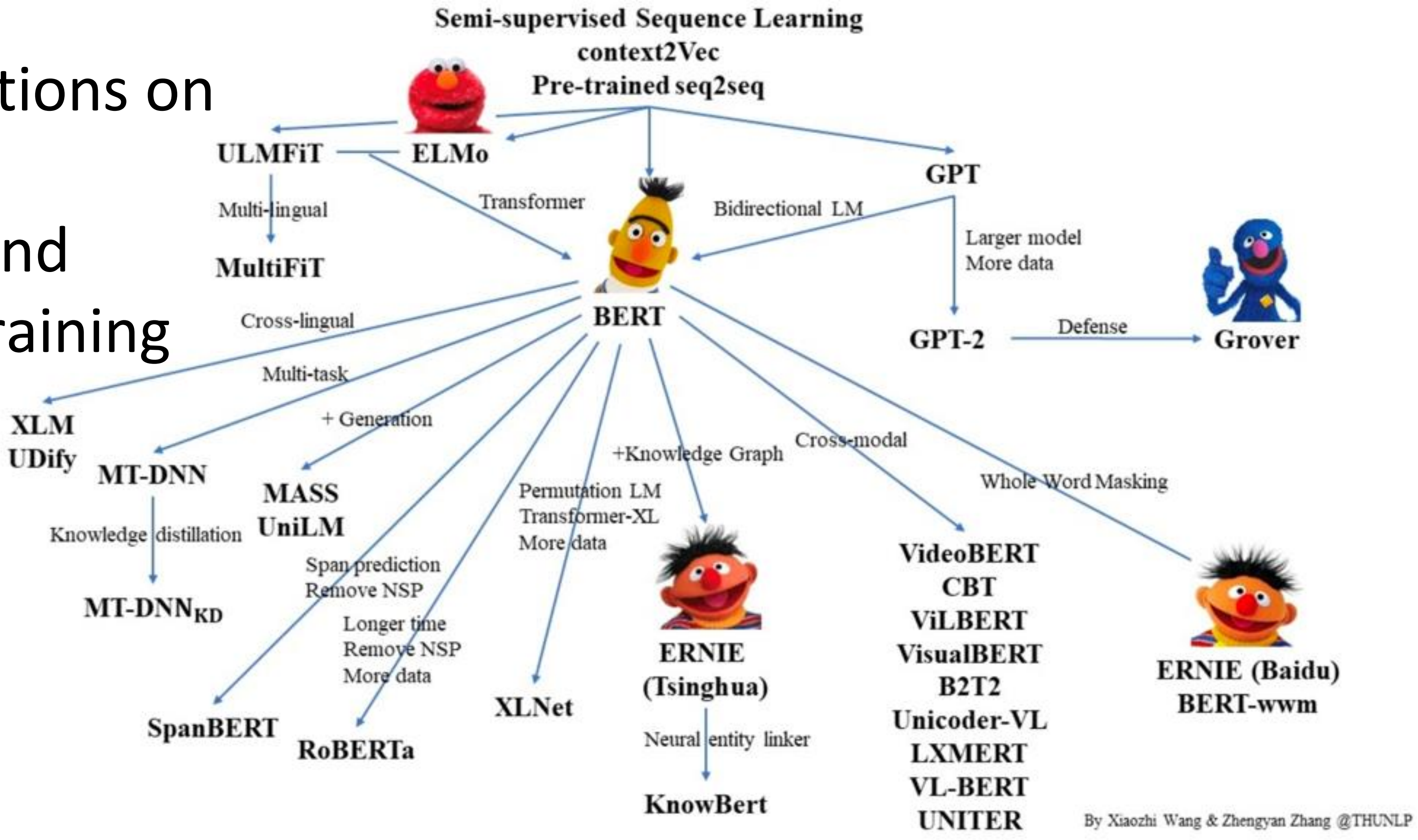
New AI fake text generator may be **too dangerous to release**, say ...
Highly Cited - The Guardian - 14 Feb 2019

View all

[Radford et al, 2019]

BERT and Family

Different Variations on Transformer architectures and different pre-training tasks

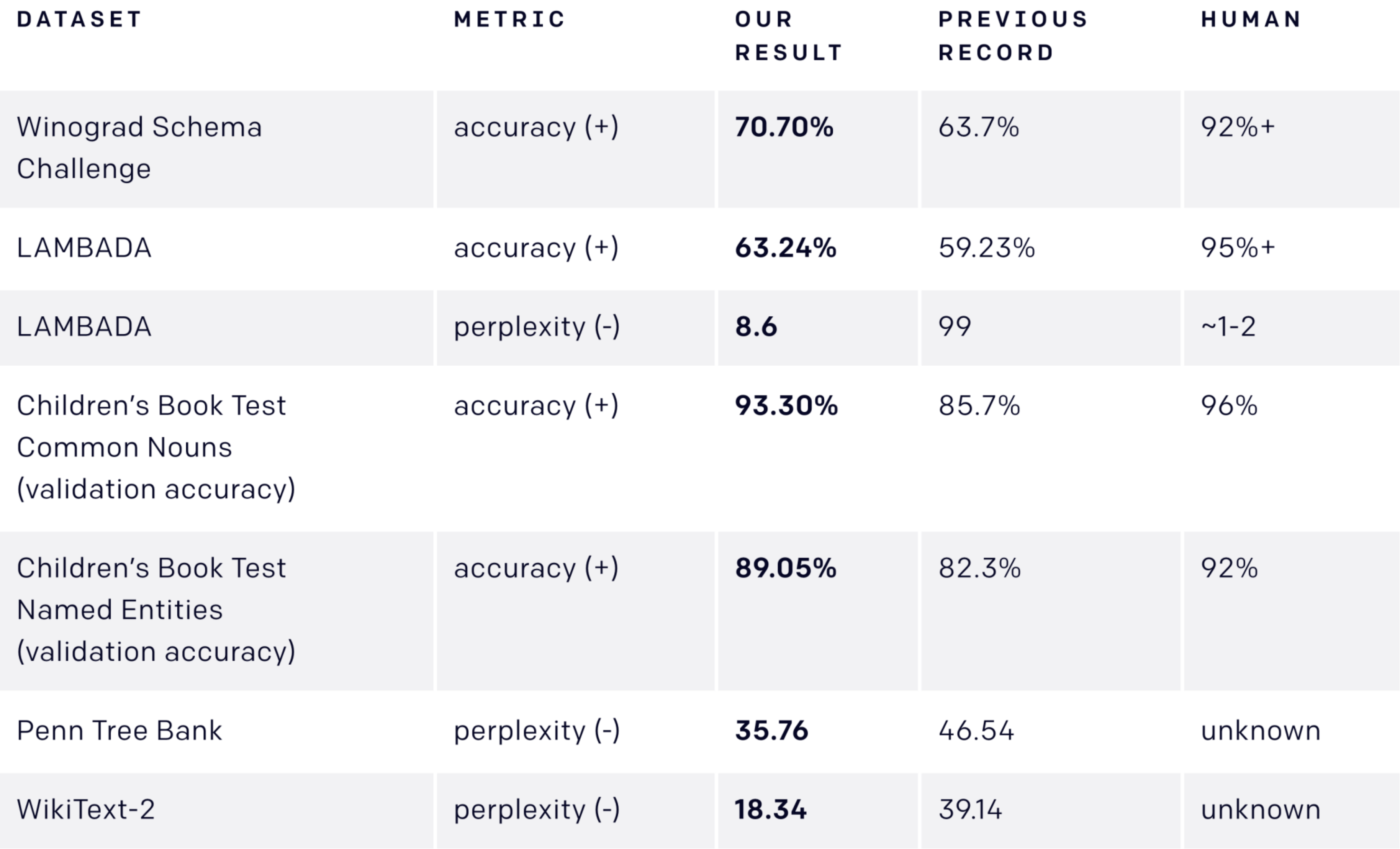


Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq

ULMFiT
ELMo
GPT
Multi-lingual
MultiFiT
Transformer
Bidirectional LM
BERT
Larger model
More data
GPT-2
Defense
Grover
Cross-lingual
XLM
UDify
Multi-task
MT-DNN
Knowledge distillation
MT-DNN$_{KD}$
+ Generation
MASS
UniLM
Span prediction
Remove NSP
SpanBERT
Longer time
Remove NSP
More data
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
+Knowledge Graph
ERNIE (Tsinghua)
Neural entity linker
KnowBert
Cross-modal
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
Whole Word Masking
ERNIE (Baidu)
BERT-wwm

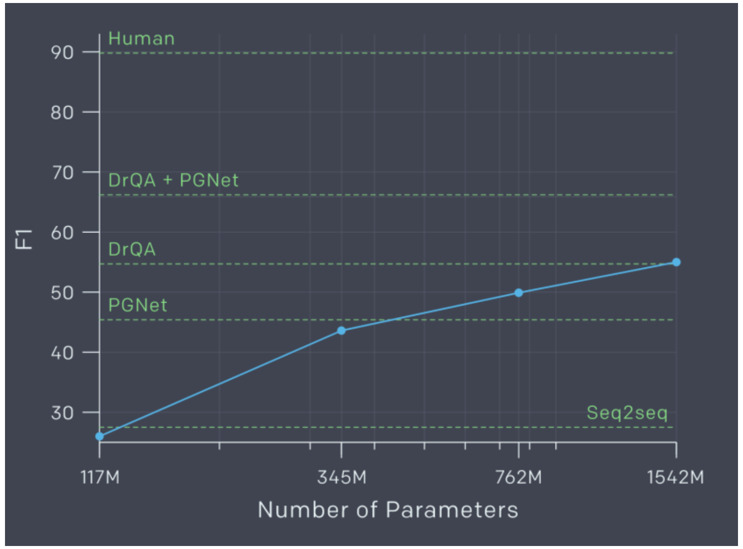
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

Benchmarks

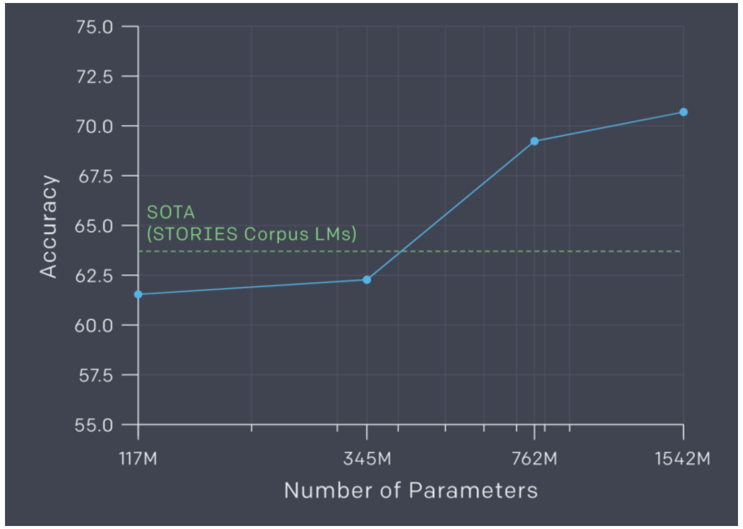


DATASET	METRIC	OUR RESULT	PREVIOUS RECORD	HUMAN
Winograd Schema Challenge	accuracy (+)	**70.70%**	63.7%	92%+
LAMBADA	accuracy (+)	**63.24%**	59.23%	95%+
LAMBADA	perplexity (-)	**8.6**	99	~1-2
Children's Book Test Common Nouns (validation accuracy)	accuracy (+)	**93.30%**	85.7%	96%
Children's Book Test Named Entities (validation accuracy)	accuracy (+)	**89.05%**	82.3%	92%
Penn Tree Bank	perplexity (-)	**35.76**	46.54	unknown
WikiText-2	perplexity (-)	**18.34**	39.14	unknown

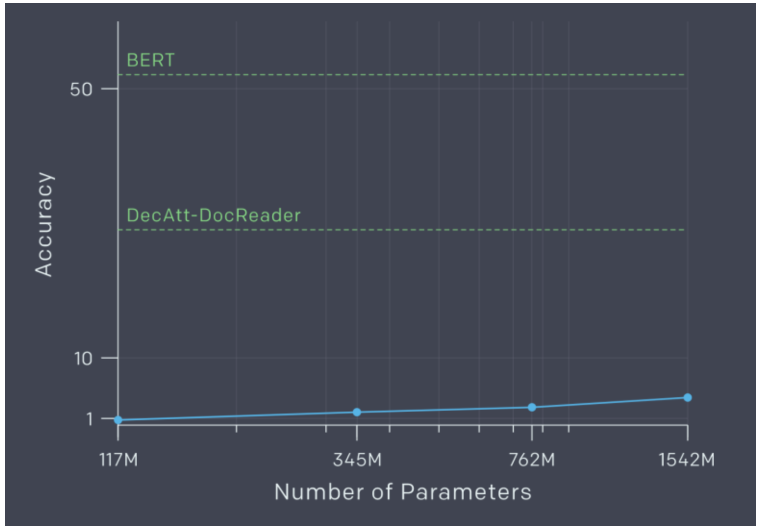
Scaling



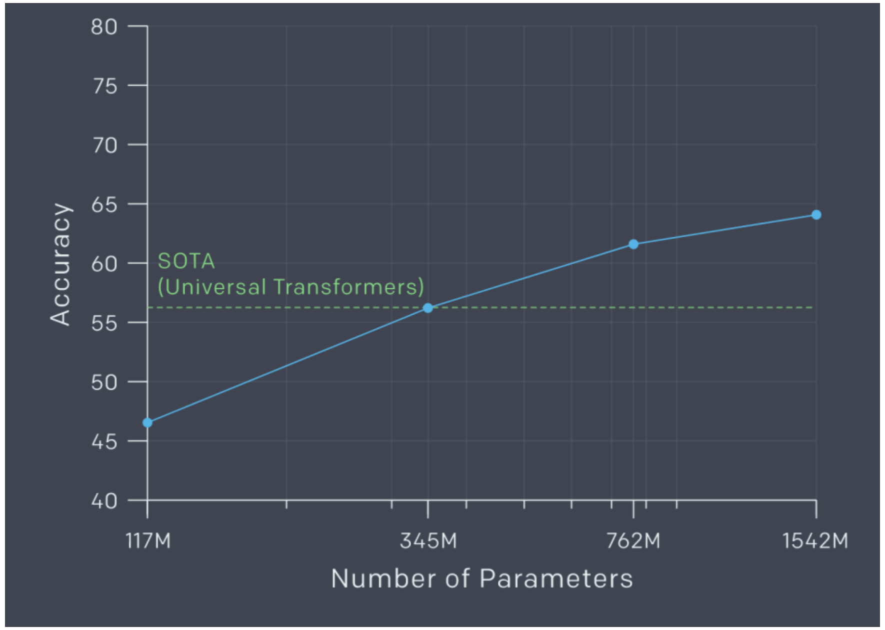
F1
90
80
70
60
50
40
30
Human
DrQA + PGNet
DrQA
PGNet
Seq2seq
117M
345M
762M
1542M
Number of Parameters



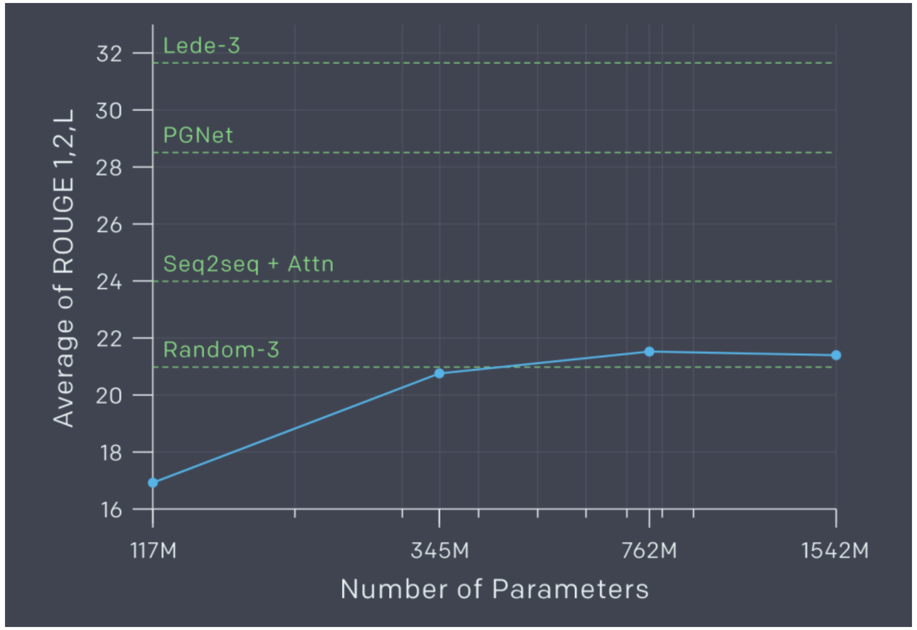
Accuracy
75.0
72.5
70.0
67.5
65.0
62.5
60.0
57.5
55.0
SOTA
(STORIES Corpus LMs)
117M
345M
762M
1542M
Number of Parameters



Accuracy
50
10
1
BERT
DecAtt-DocReader
117M
345M
762M
1542M
Number of Parameters

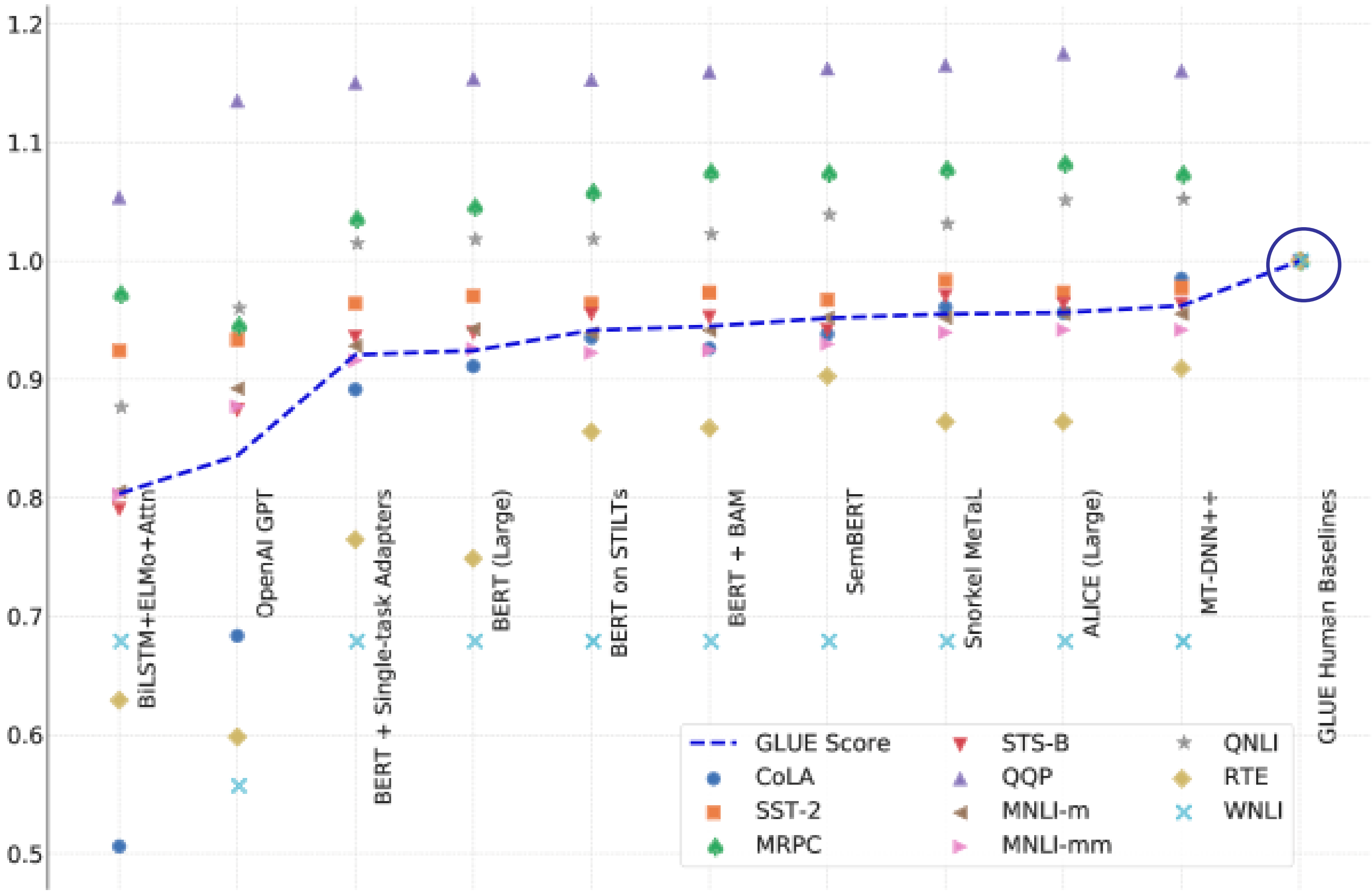


Accuracy
80
75
70
65
60
55
50
45
40
SOTA
(Universal Transformers)
117M
345M
762M
1542M
Number of Parameters



Average of ROUGE 1,2,L
32
30
28
26
24
22
20
18
16
Lede-3
PGNet
Seq2seq + Attn
Random-3
117M
345M
762M
1542M
Number of Parameters

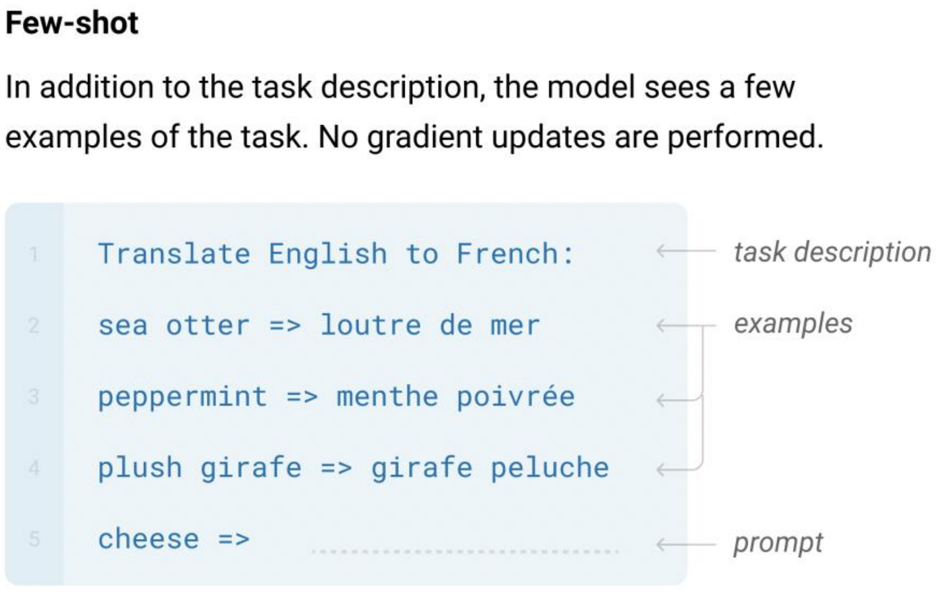
Pretrained Models (BERT) on GLUE Benchmarks



Human Performance on GLUE

Massive Pre-trained models are few-shot learners! (GPT-3)

175B GPT-3 can work without fine-tuning, when it is shown sample **demonstrations** for a task:



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

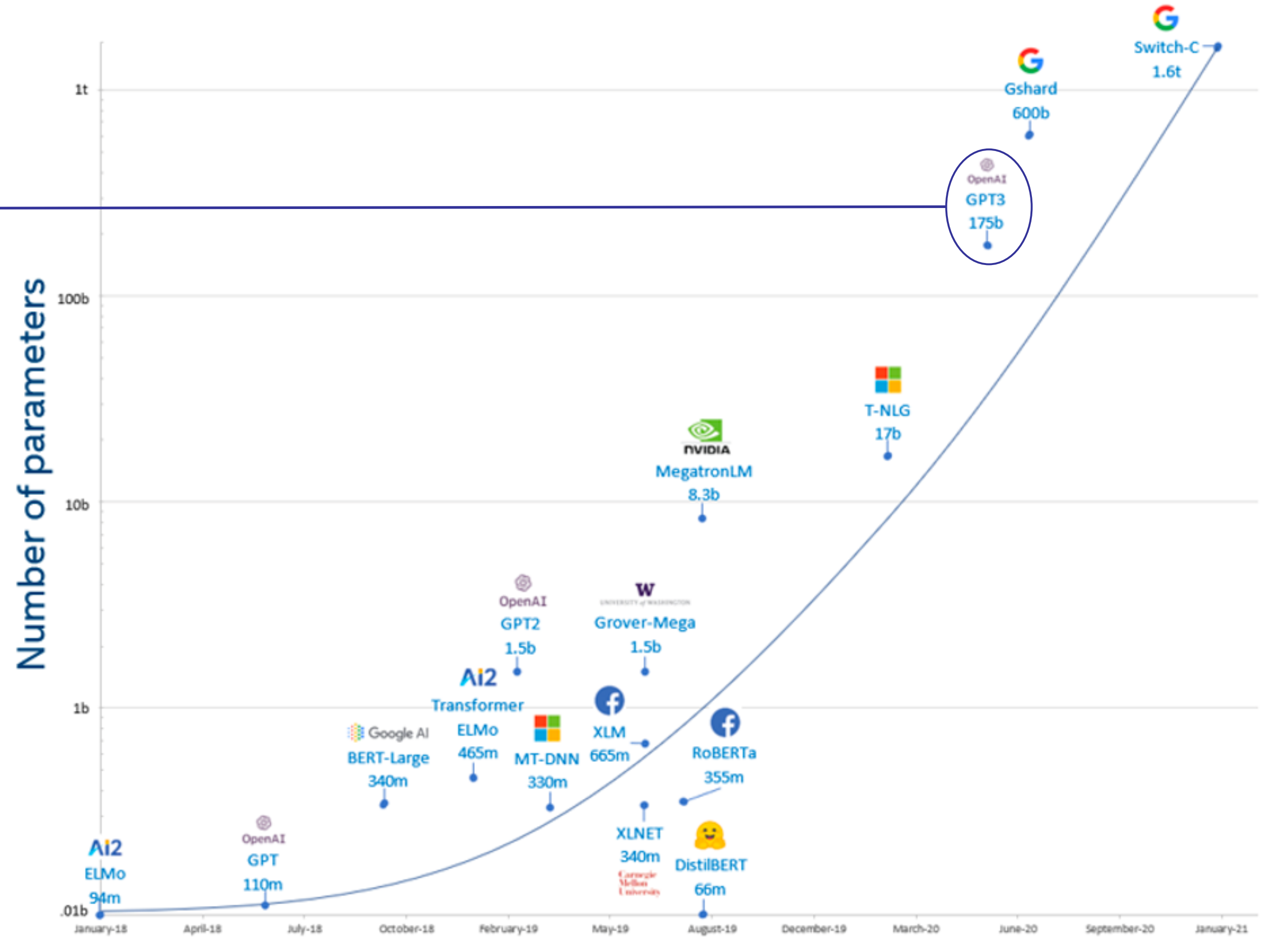
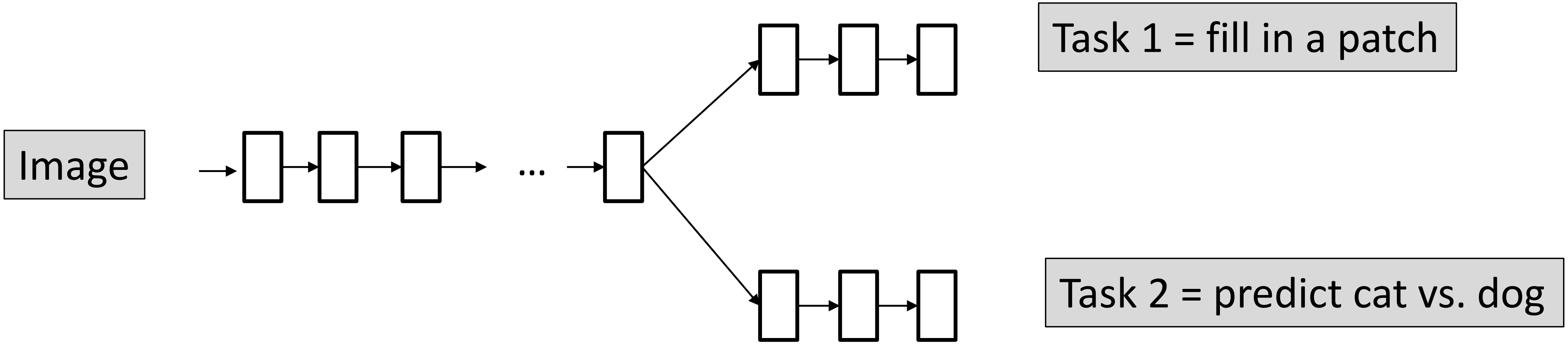


Figure 1: Exponential growth of number of parameters in DL models

Unsupervised Learning in Vision

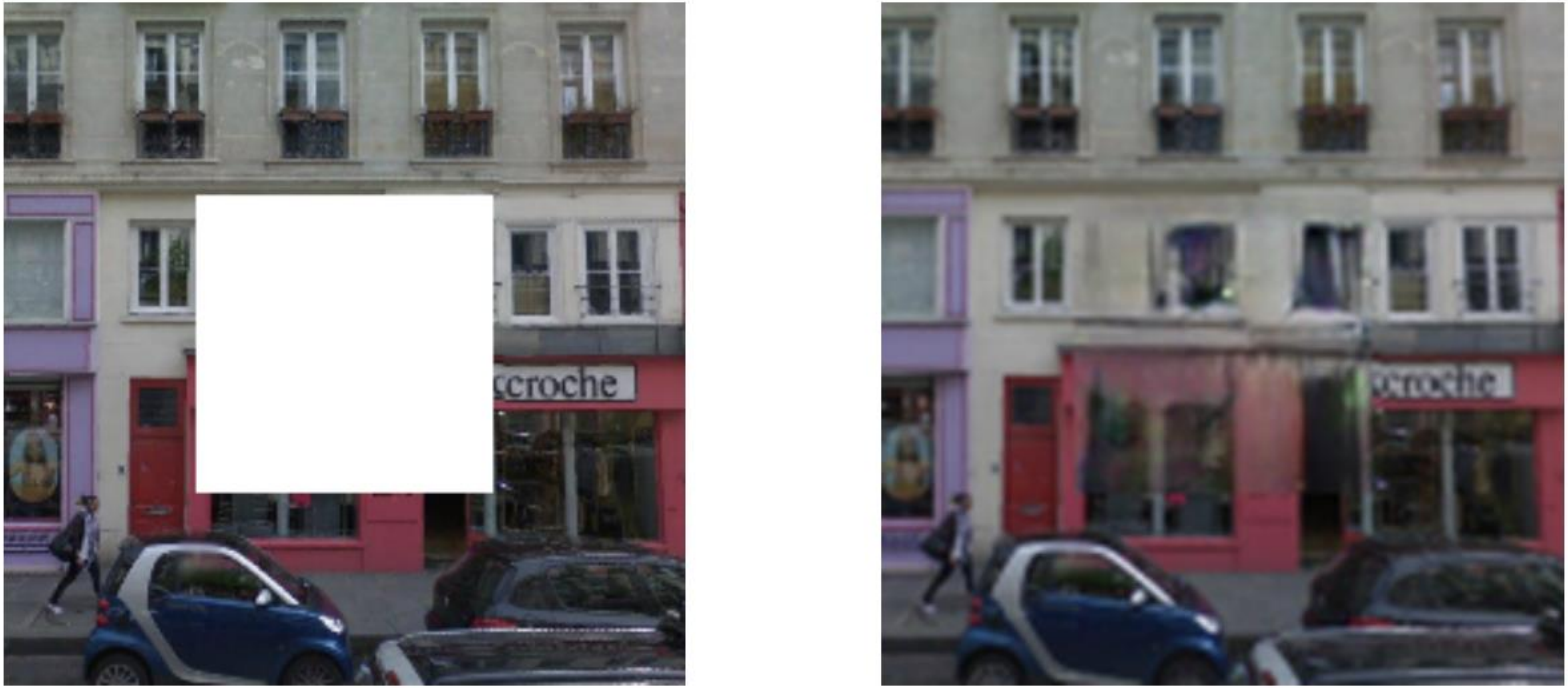


Image

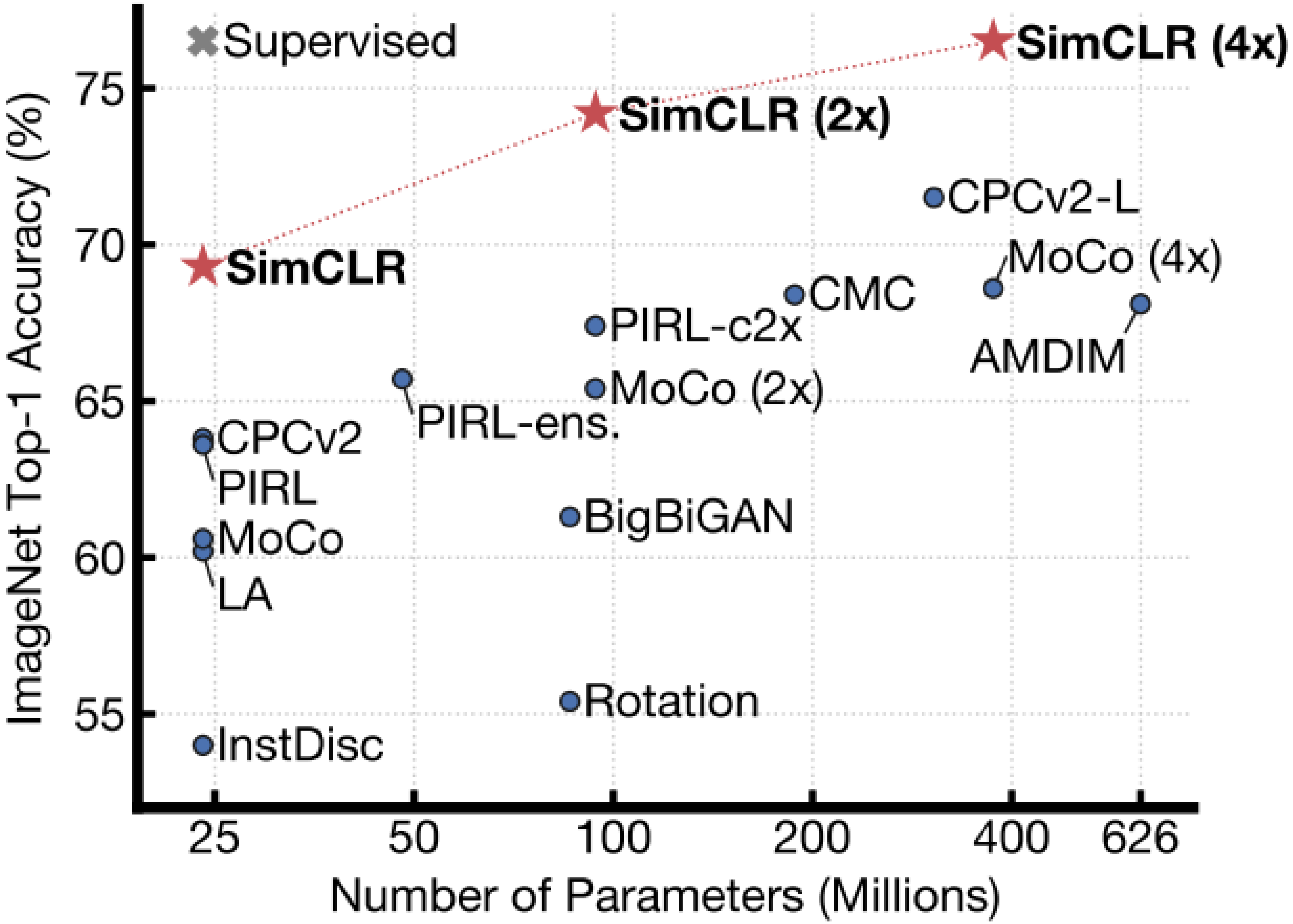
Task 1 = fill in a patch

Task 2 = predict cat vs. dog

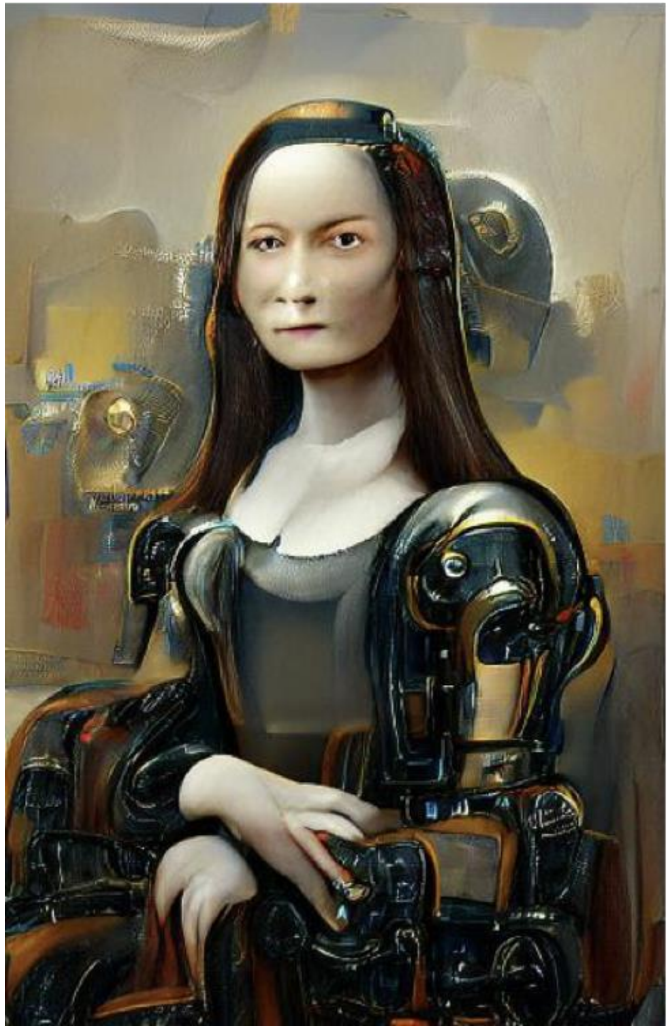
Predict Missing Patch



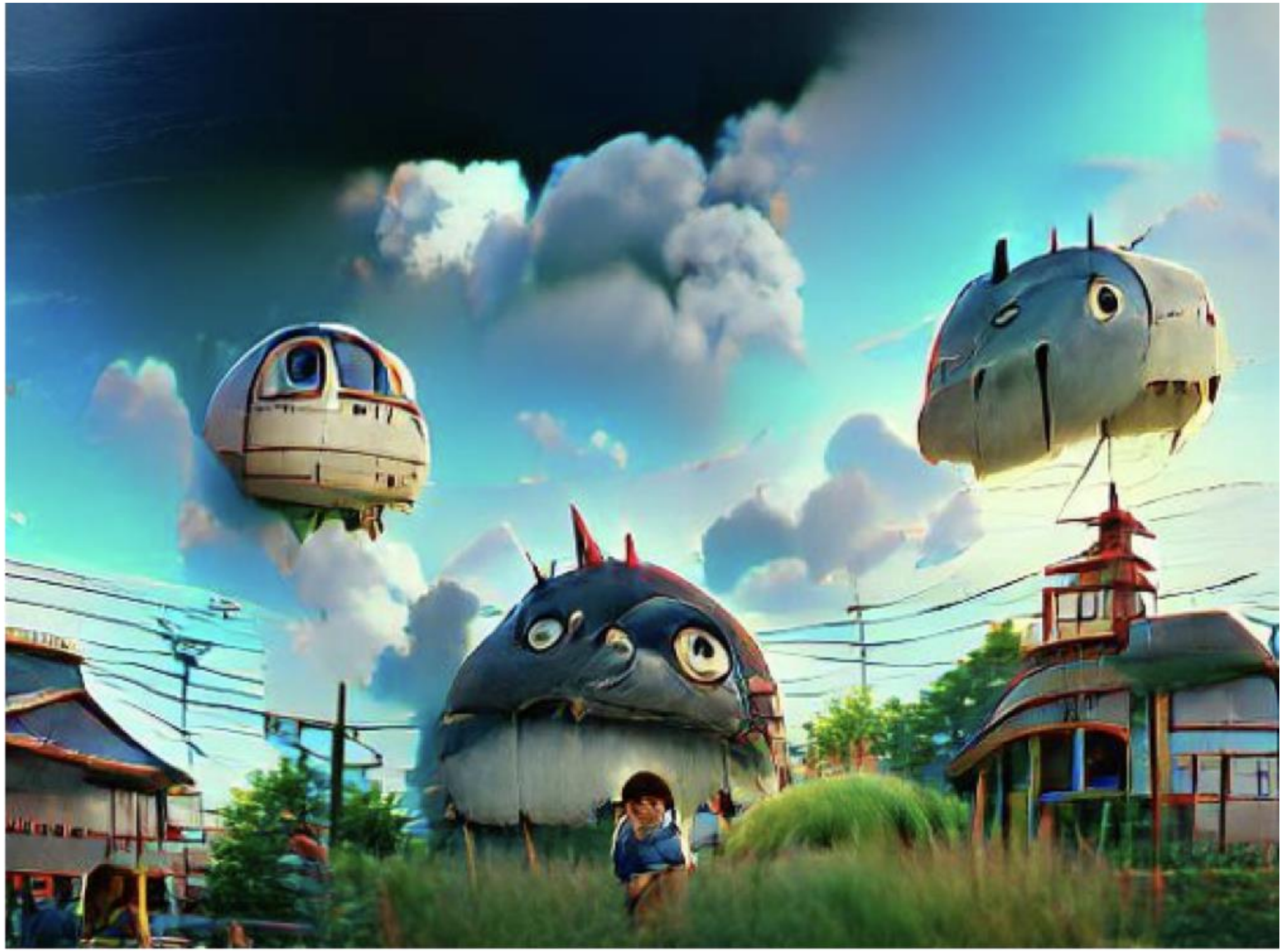
SimCLR + linear classifier



AI for Art Creation

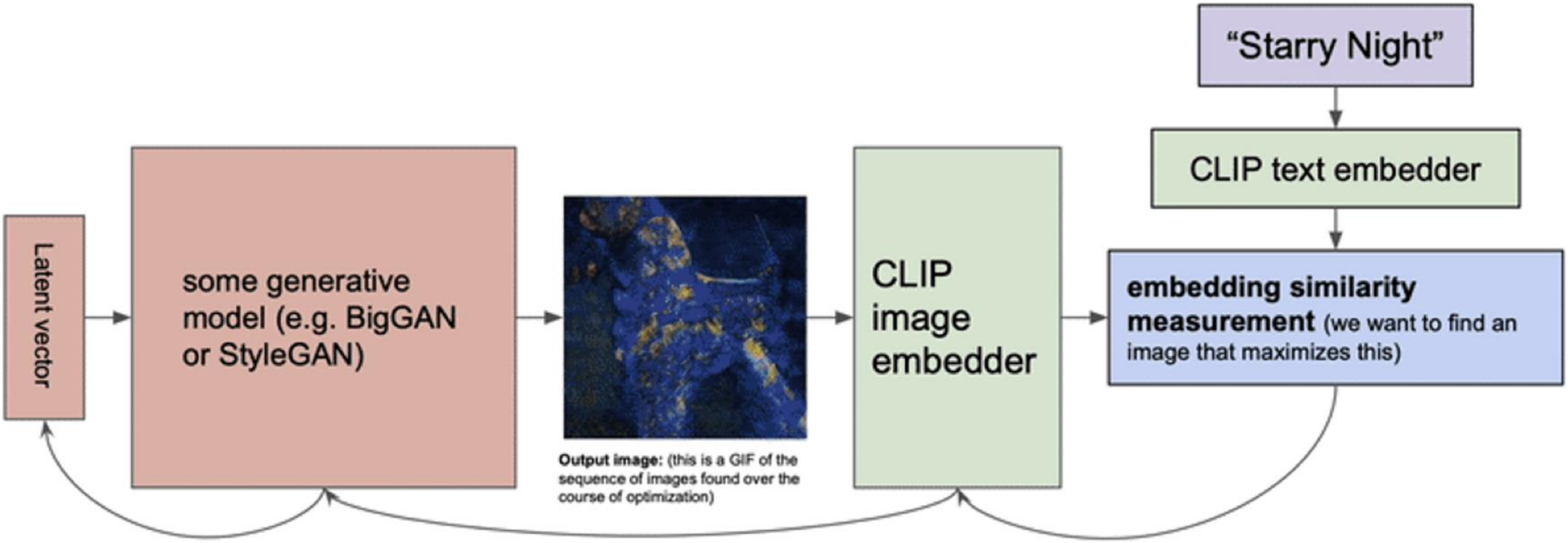


humanoid robot Mona Lisa artstationHQ



studio ghibli trending on artstation | vary

Text-Guided Image Generation

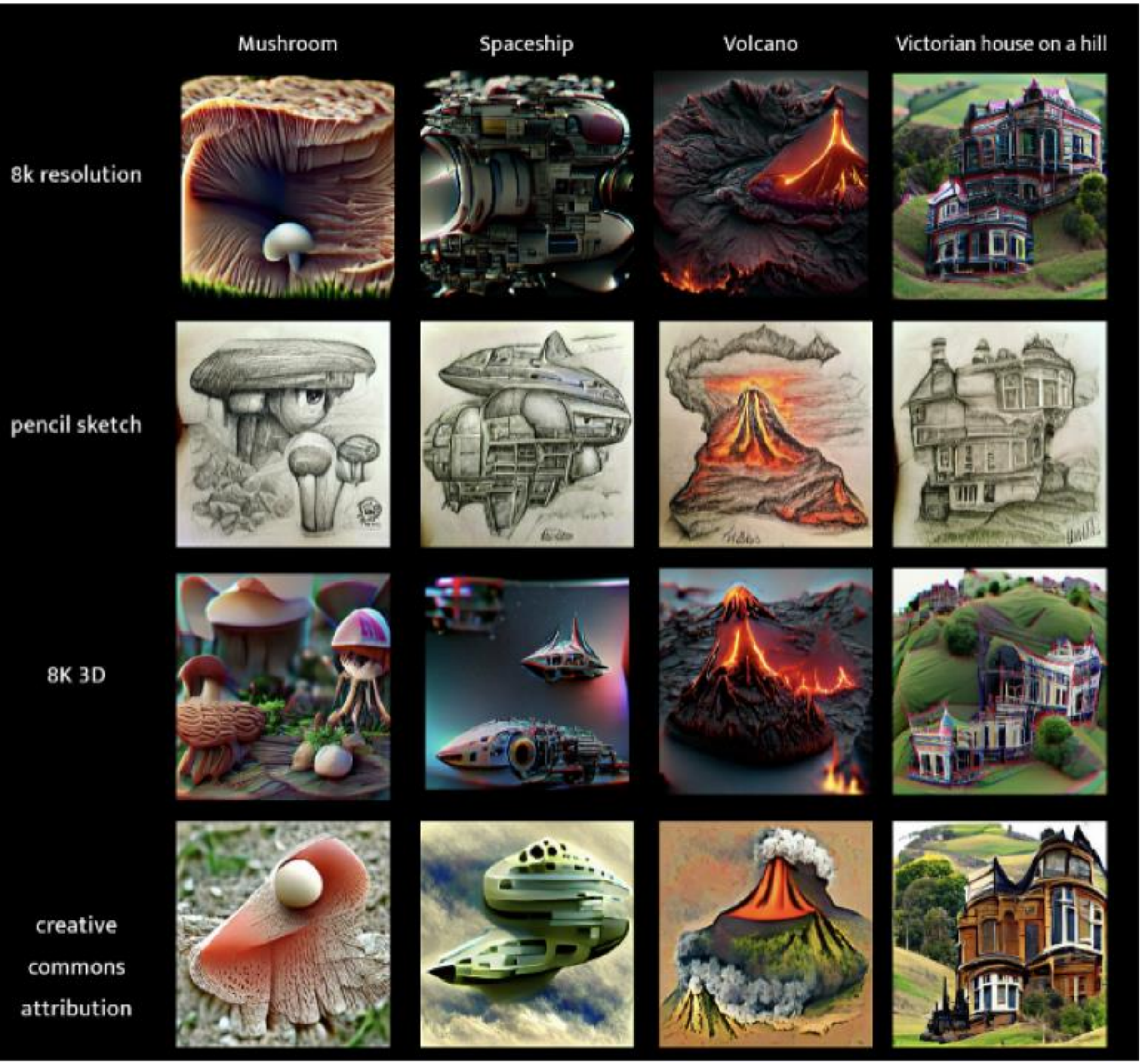


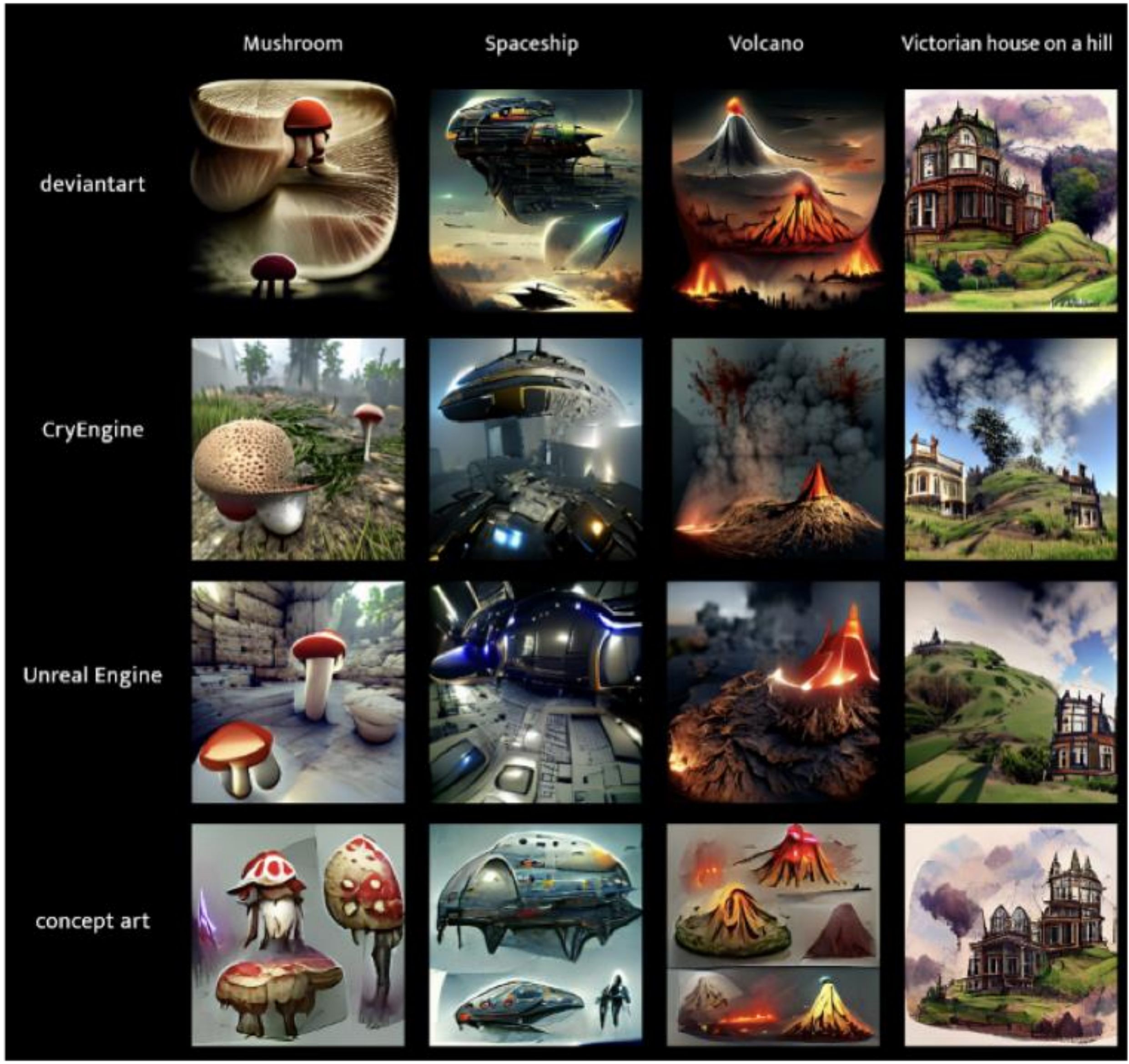
"Starry Night"
CLIP text embedder
Latent vector
some generative model (e.g. BigGAN or StyleGAN)
CLIP image embedder
embedding similarity measurement (we want to find an image that maximizes this)

Output image: (this is a GIF of the sequence of images found over the course of optimization)

via Charlie Snell

Examples (CLIP + VQGAN)



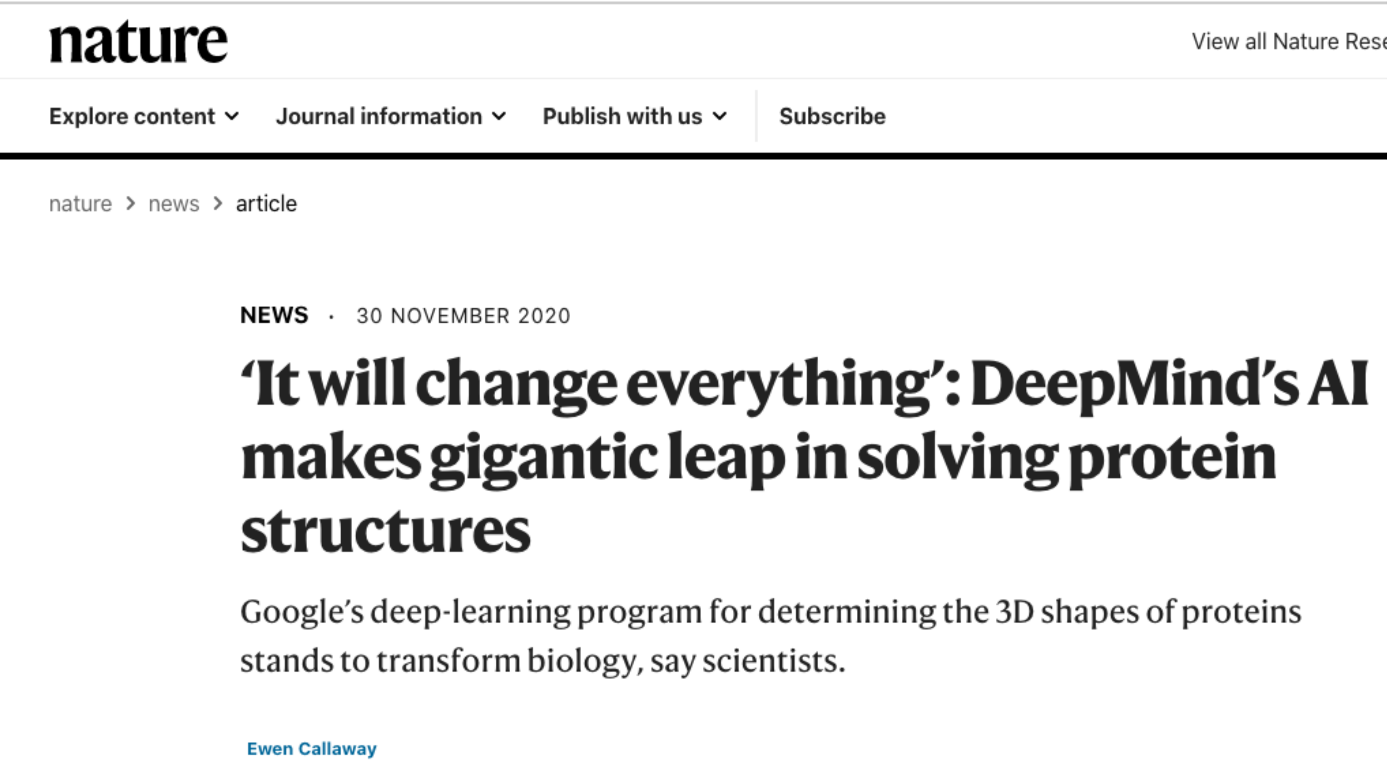


Research Frontiers

- Deep Unsupervised Learning
- AI for Science
- AI and Ethics

Also:

- Unsupervised Deep Reinforcement Learning
- Human-in-the-loop Reinforcement Learning
- …

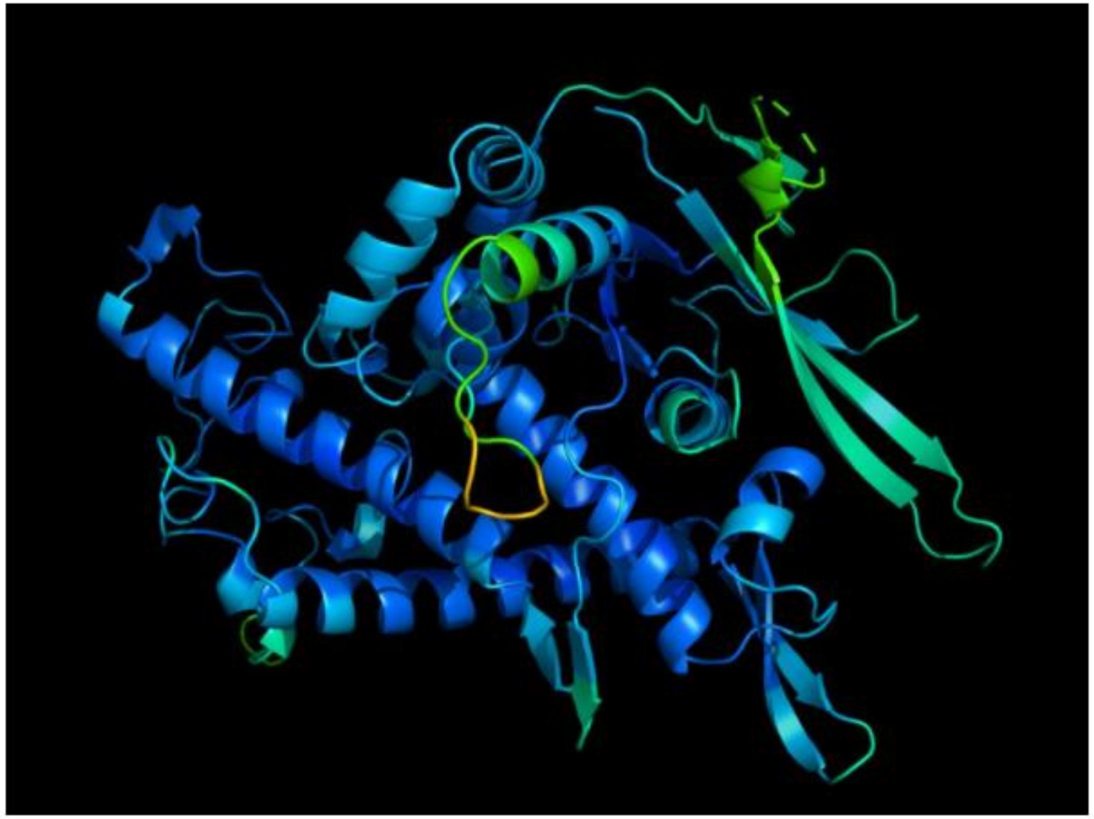


NEWS · 30 NOVEMBER 2020

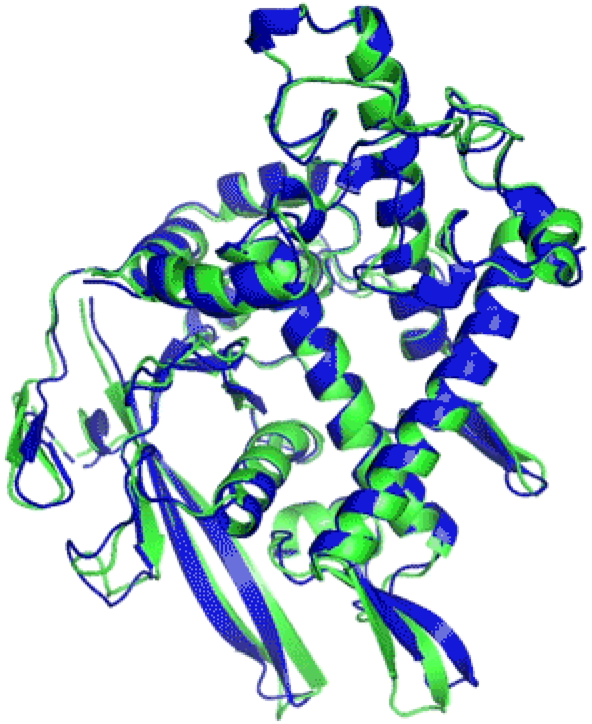
‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures

Google’s deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

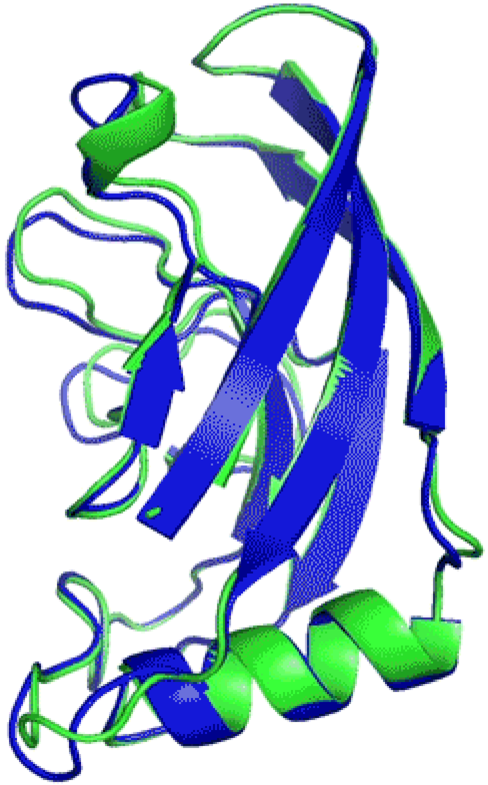
Ewen Callaway



A protein’s function is determined by its 3D shape. Credit: DeepMind



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

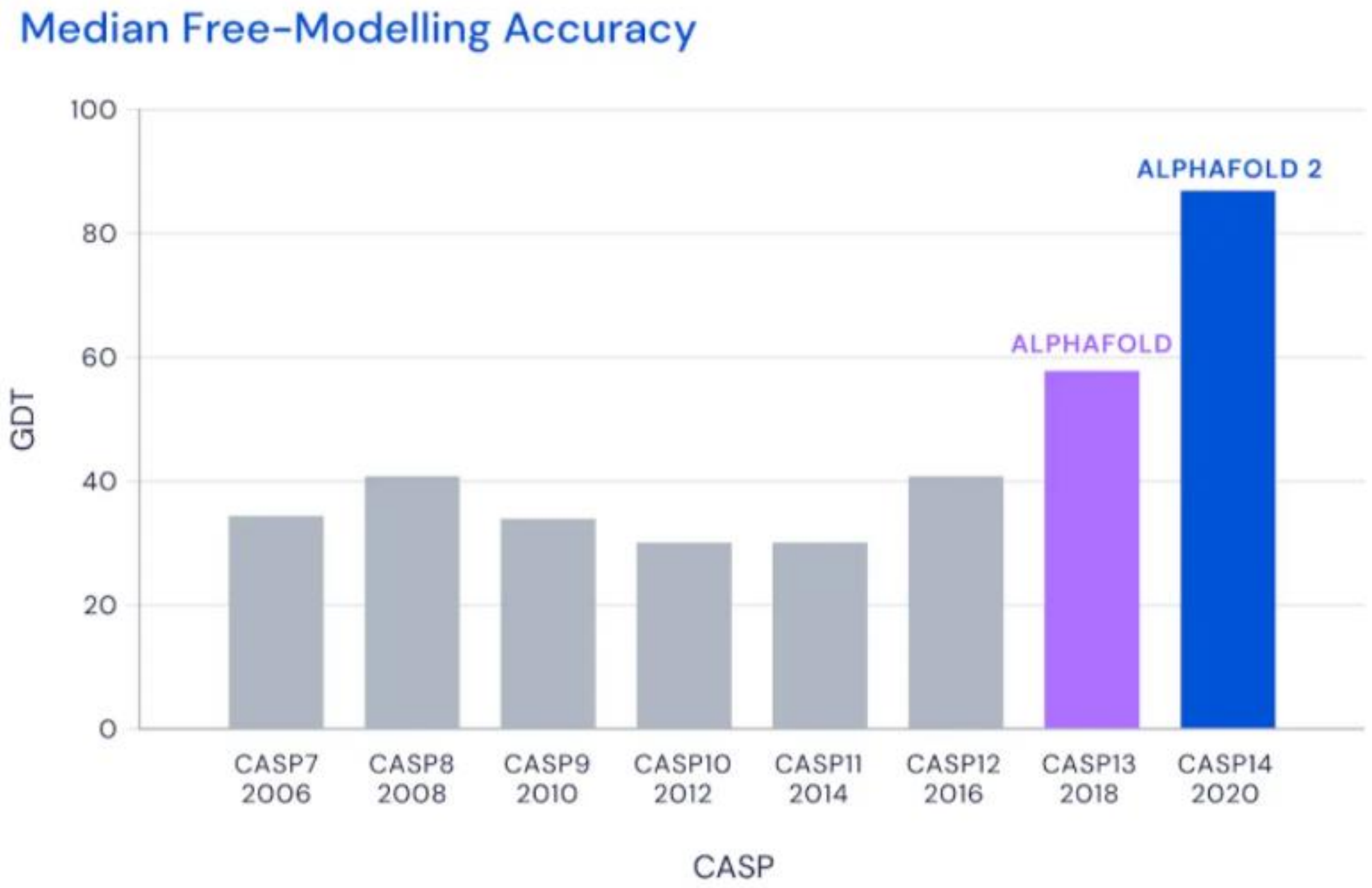


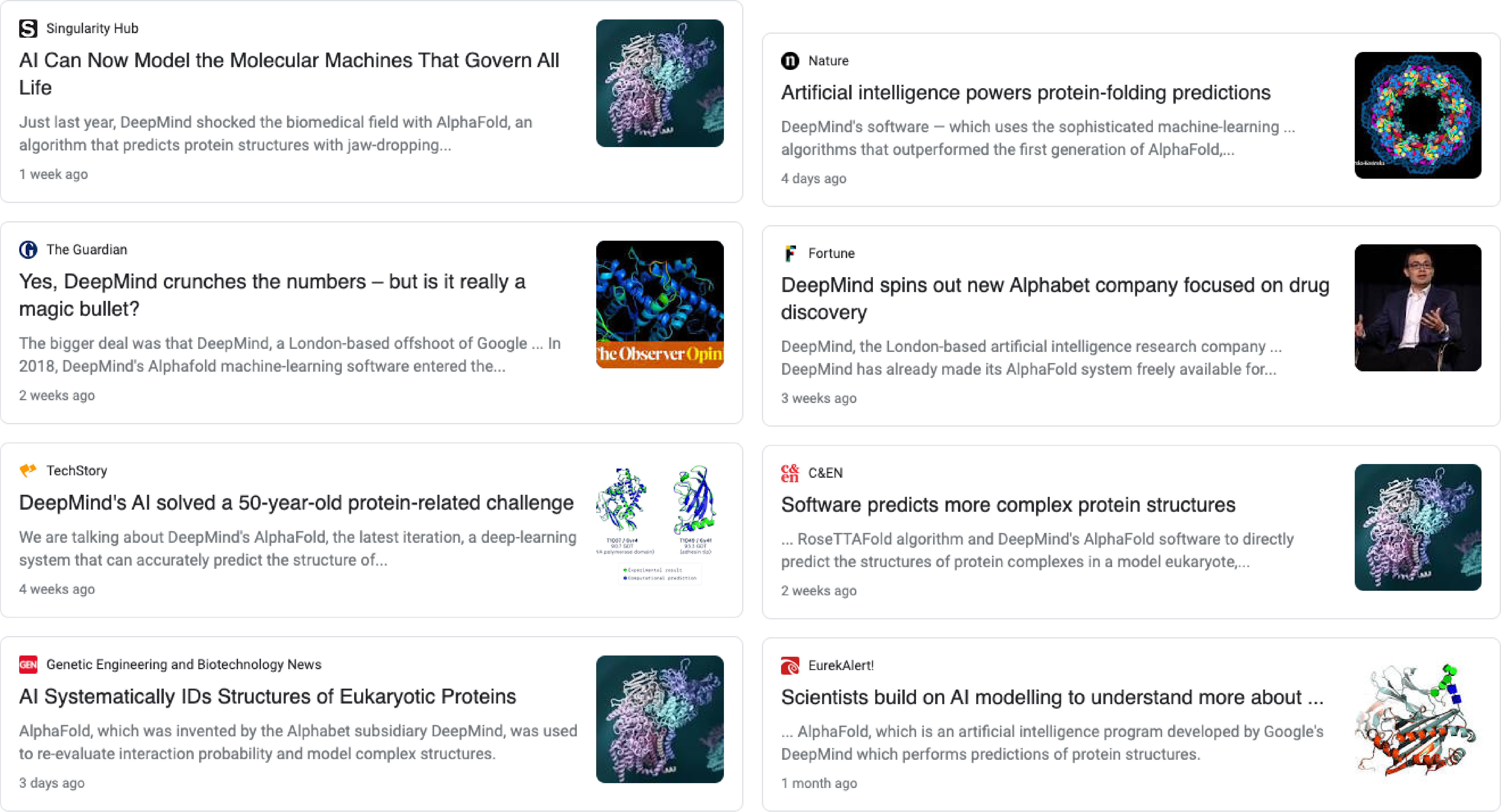
T1049 / 6y4f
93.3 GDT
(adhesin tip)

🟢 Experimental result

🔵 Computational prediction

CASP 2020 Competition





Singularity Hub

AI Can Now Model the Molecular Machines That Govern All Life

Just last year, DeepMind shocked the biomedical field with AlphaFold, an algorithm that predicts protein structures with jaw-dropping...

1 week ago

Nature

Artificial intelligence powers protein-folding predictions

DeepMind's software — which uses the sophisticated machine-learning ... algorithms that outperformed the first generation of AlphaFold,...

4 days ago

The Guardian

Yes, DeepMind crunches the numbers – but is it really a magic bullet?

The bigger deal was that DeepMind, a London-based offshoot of Google ... In 2018, DeepMind's Alphafold machine-learning software entered the...

2 weeks ago

Fortune

DeepMind spins out new Alphabet company focused on drug discovery

DeepMind, the London-based artificial intelligence research company ... DeepMind has already made its AlphaFold system freely available for...

3 weeks ago

TechStory

DeepMind's AI solved a 50-year-old protein-related challenge

We are talking about DeepMind's AlphaFold, the latest iteration, a deep-learning system that can accurately predict the structure of...

4 weeks ago

C&EN

Software predicts more complex protein structures

... RoseTTAFold algorithm and DeepMind's AlphaFold software to directly predict the structures of protein complexes in a model eukaryote,...

2 weeks ago

Genetic Engineering and Biotechnology News

AI Systematically IDs Structures of Eukaryotic Proteins

AlphaFold, which was invented by the Alphabet subsidiary DeepMind, was used to re-evaluate interaction probability and model complex structures.

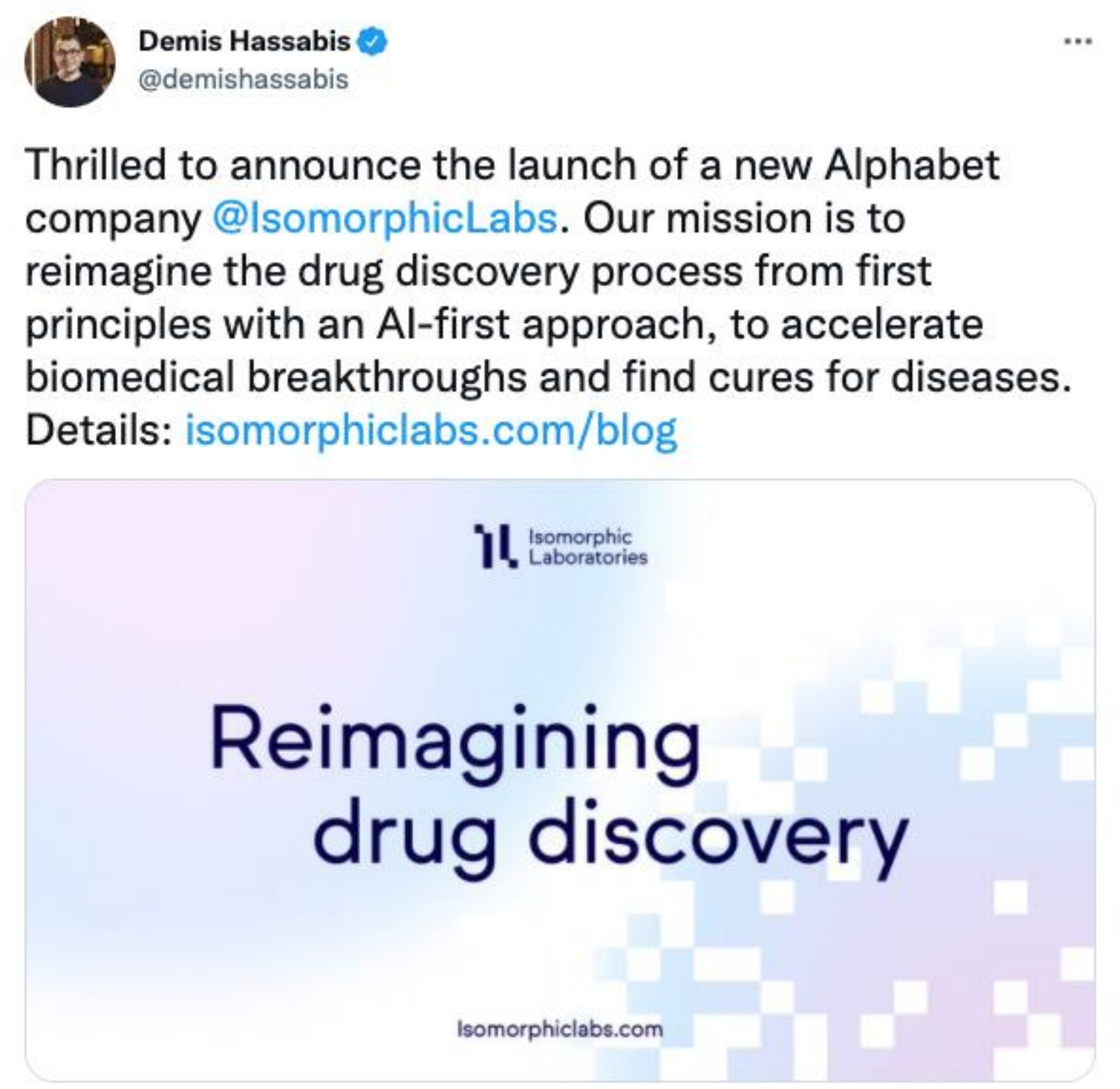
3 days ago

EurekAlert!

Scientists build on AI modelling to understand more about ...

... AlphaFold, which is an artificial intelligence program developed by Google's DeepMind which performs predictions of protein structures.

1 month ago



Demis Hassabis
@demishassabis

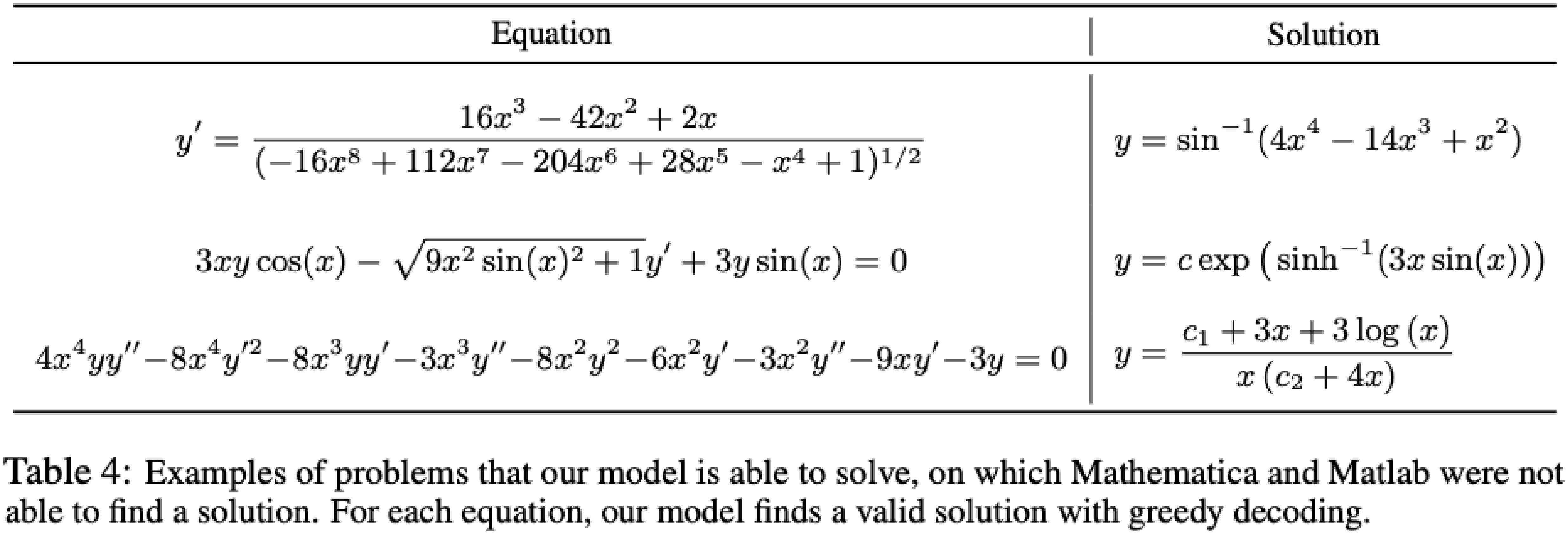
Thrilled to announce the launch of a new Alphabet company @IsomorphicLabs. Our mission is to reimagine the drug discovery process from first principles with an AI-first approach, to accelerate biomedical breakthroughs and find cures for diseases. Details: isomorphiclabs.com/blog

Isomorphic Laboratories

Reimagining drug discovery

Isomorphiclabs.com

Symbolic Math: Integrals and ODEs

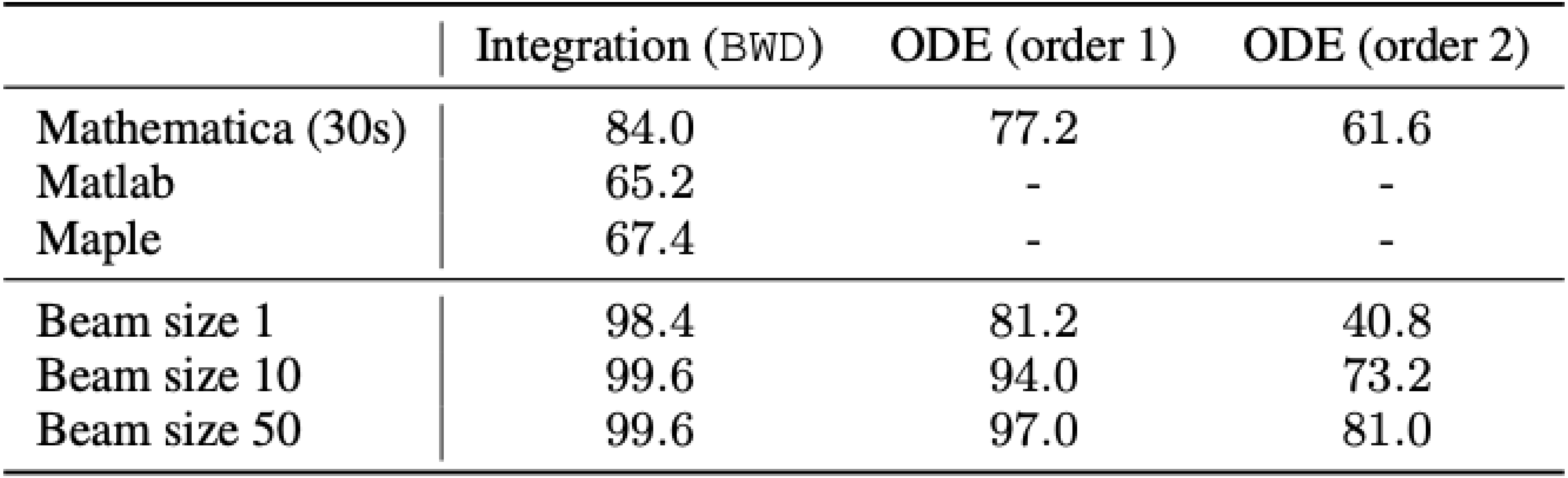


Equation	Solution
$y' = \dfrac{16x^3 - 42x^2 + 2x}{(-16x^8 + 112x^7 - 204x^6 + 28x^5 - x^4 + 1)^{1/2}}$	$y = \sin^{-1}(4x^4 - 14x^3 + x^2)$
$3xy\cos(x) - \sqrt{9x^2\sin(x)^2 + 1}\,y' + 3y\sin(x) = 0$	$y = c\exp\left(\sinh^{-1}(3x\sin(x))\right)$
$4x^4yy'' - 8x^4y'^2 - 8x^3yy' - 3x^3y'' - 8x^2y^2 - 6x^2y' - 3x^2y'' - 9xy' - 3y = 0$	$y = \dfrac{c_1 + 3x + 3\log(x)}{x(c_2 + 4x)}$

Table 4: Examples of problems that our model is able to solve, on which Mathematica and Matlab were not able to find a solution. For each equation, our model finds a valid solution with greedy decoding.

[Lample and Charton, ICLR 2020]

Symbolic Math: Integrals and ODEs



	Integration (`BWD`)	ODE (order 1)	ODE (order 2)
Mathematica (30s)	84.0	77.2	61.6
Matlab	65.2	-	-
Maple	67.4	-	-
Beam size 1	98.4	81.2	40.8
Beam size 10	99.6	94.0	73.2
Beam size 50	99.6	97.0	81.0

[Lampple and Charton, ICLR 2020]

Research Frontiers

- Deep Unsupervised Learning
- AI for Science
- AI and Ethics

Also:

- Unsupervised Deep Reinforcement Learning
- Human-in-the-loop Reinforcement Learning
- …

AI Ethics Ever More Important

- Why?
  - AI is making decisions, at scale
  - Any kind of issues (e.g. bias or malignant use) could significantly affect *people*
- Many open questions:
  - Who is responsible?
  - How to diagnose and prevent?

Some Key AI Ethics Topics

- Disinformation
- Bias and fairness
- Privacy and surveillance
- Metrics
- Algorithmic colonialism

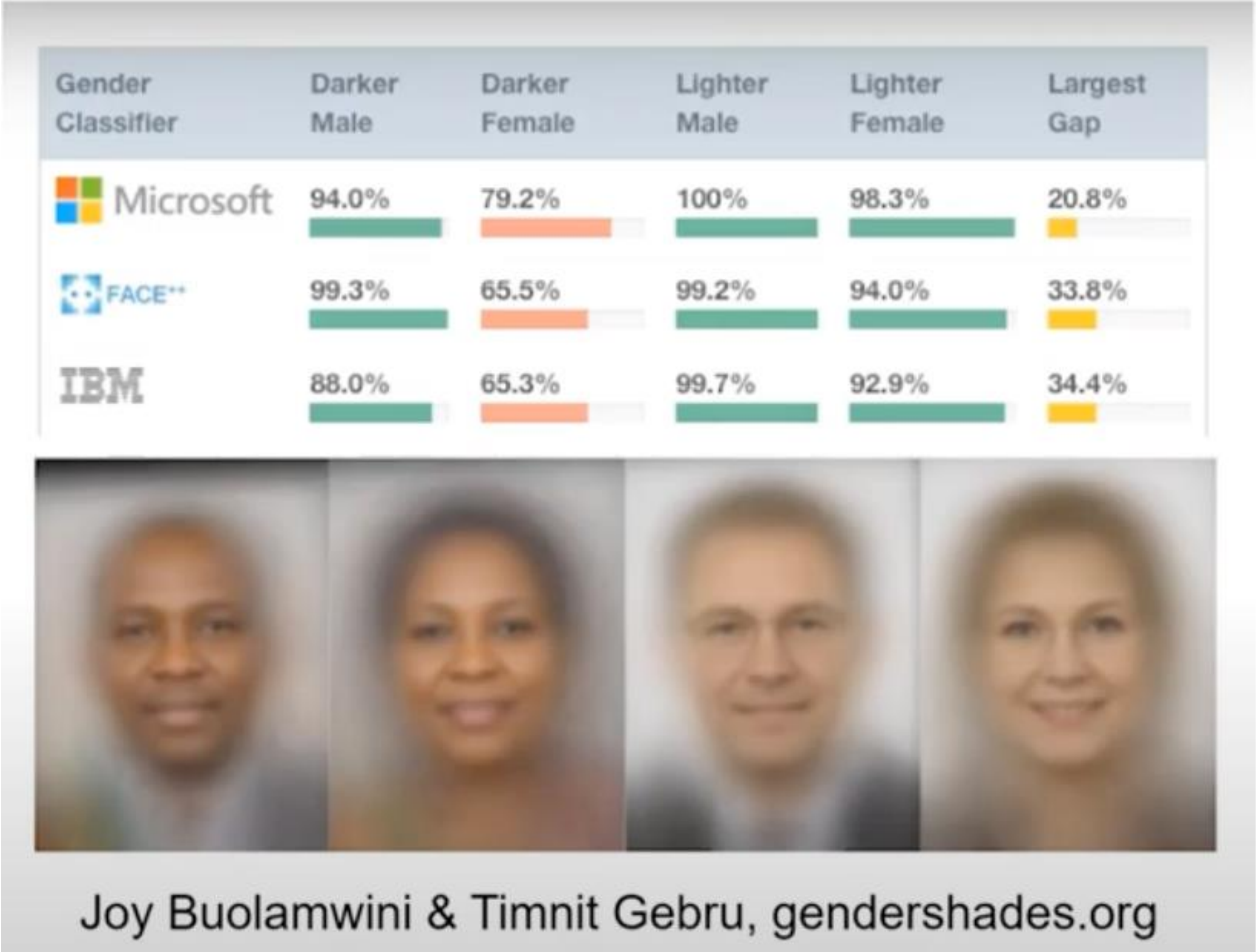
Source: Rachel Thomas (@math_rachel)

Disinformation can be aided by AI generated text

Prompt: Recycling is good for the world, no, you could not be more wrong.

GPT-2: Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I’m not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world’s most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources.

Bias and Fairness – Gender Shades

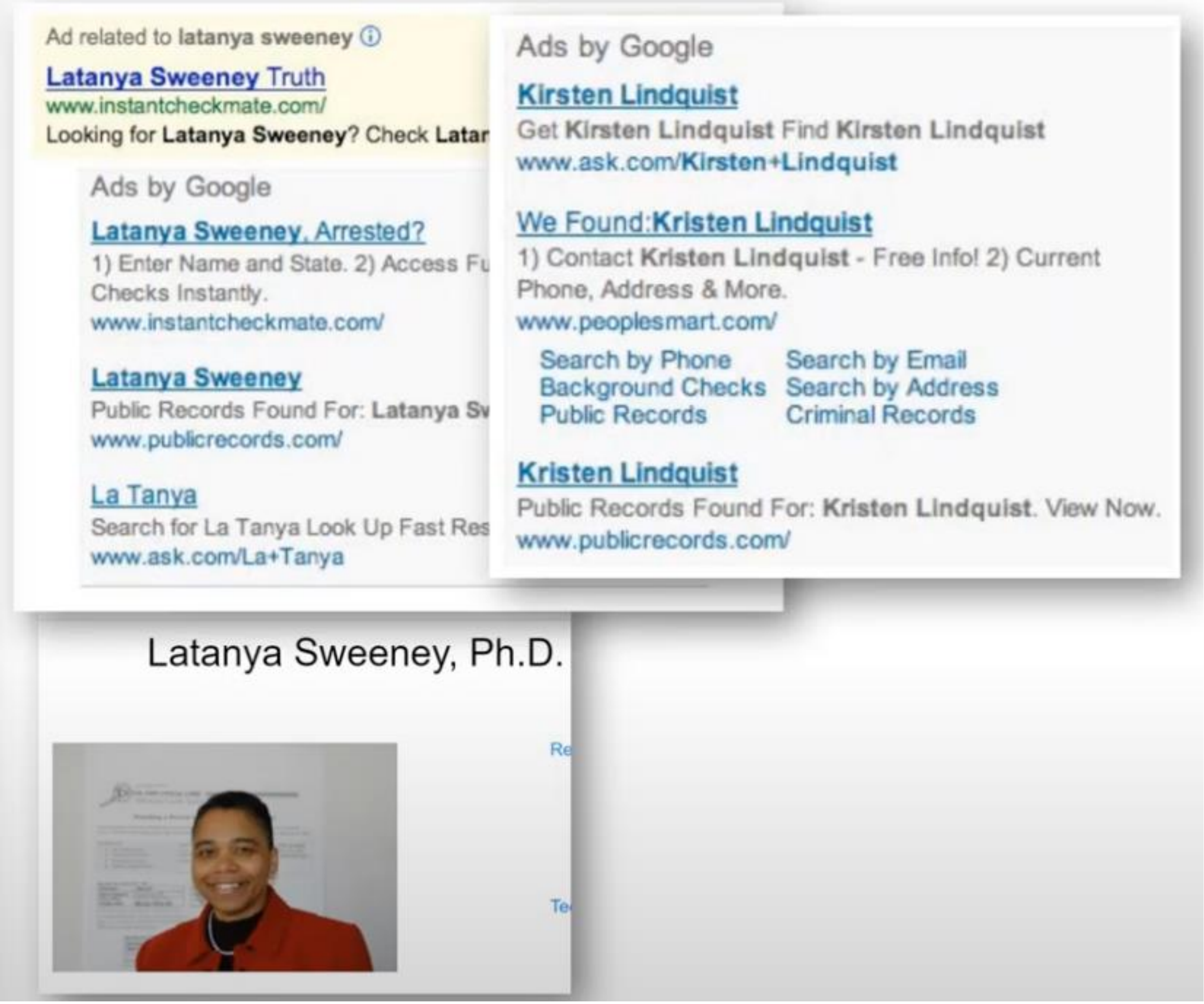


Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Joy Buolamwini & Timnit Gebru, gendershades.org

Source: Rachel Thomas (@math_rachel)

Bias and Fairness – Biased Ads



Ad related to **latanya sweeney**

Latanya Sweeney Truth

www.instantcheckmate.com/

Looking for **Latanya Sweeney**? Check Latar

Ads by Google

Latanya Sweeney, Arrested?

1) Enter Name and State. 2) Access Fu

Checks Instantly.

www.instantcheckmate.com/

Latanya Sweeney

Public Records Found For: **Latanya Sv**

www.publicrecords.com/

La Tanya

Search for La Tanya Look Up Fast Res

www.ask.com/La+Tanya

Ads by Google

Kirsten Lindquist

Get **Kirsten Lindquist** Find **Kirsten Lindquist**

www.ask.com/**Kirsten+Lindquist**

We Found:**Kristen Lindquist**

1) Contact **Kristen Lindquist** - Free Info! 2) Current Phone, Address & More.

www.peoplesmart.com/

Kristen Lindquist

Public Records Found For: **Kristen Lindquist**. View Now.

www.publicrecords.com/

Latanya Sweeney, Ph.D.

Source: Rachel Thomas (@math_rachel)

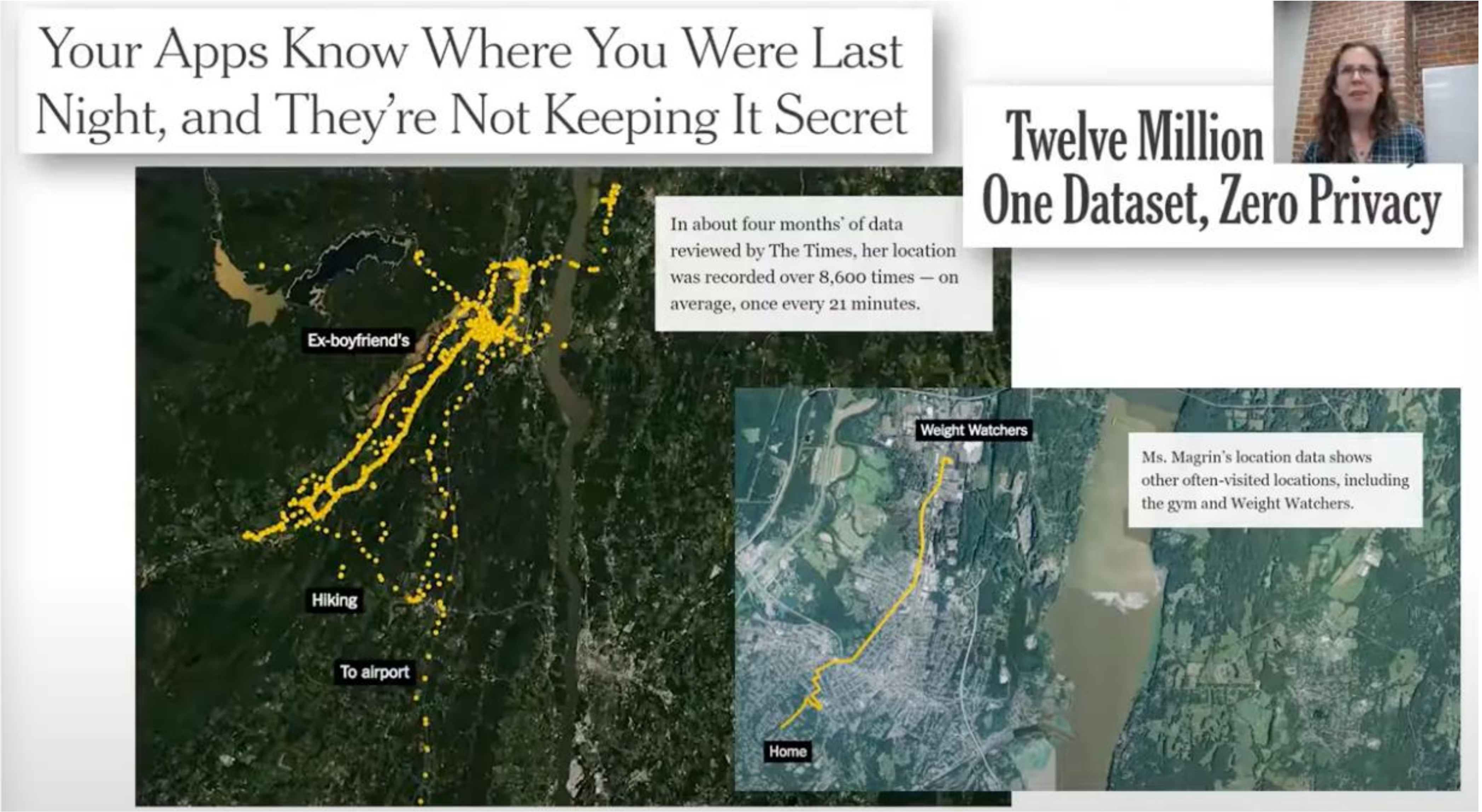
Bias and Fairness

Algorithms are used differently than human decision makers

- People are more likely to assume algorithms are objective or error-free
- Algorithms are more likely to be implemented with no appeals process
- Algorithms are often used at scale
- Algorithmic systems are cheap

Source: Rachel Thomas (@math_rachel)

Privacy and Surveillance



Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret

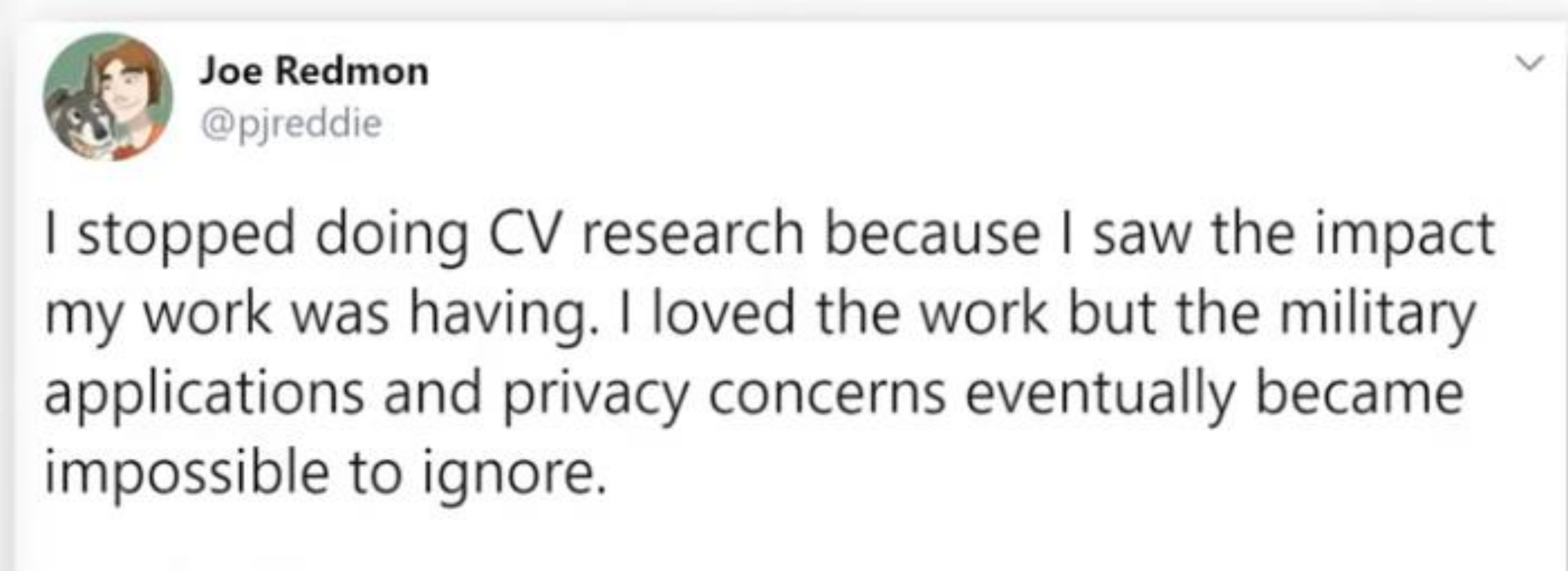
Twelve Million Phones, One Dataset, Zero Privacy

In about four months' of data reviewed by The Times, her location was recorded over 8,600 times — on average, once every 21 minutes.

Ms. Magrin's location data shows other often-visited locations, including the gym and Weight Watchers.

Source: Rachel Thomas (@math_rachel)

Privacy and Surveillance



Joe Redmon
@pjreddie

I stopped doing CV research because I saw the impact my work was having. I loved the work but the military applications and privacy concerns eventually became impossible to ignore.

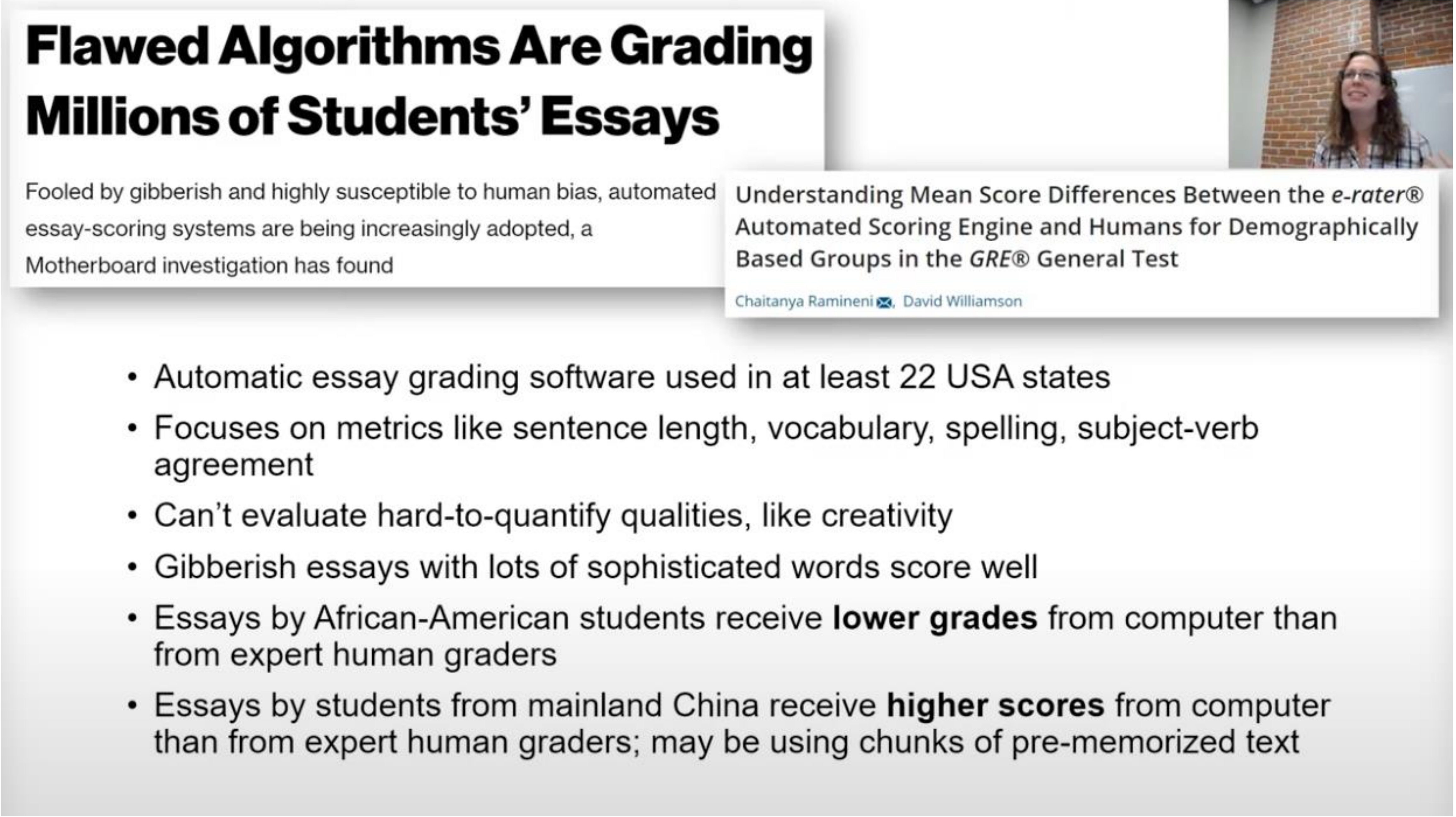
Source: Rachel Thomas (@math_rachel)

Metrics

- Goodhart’s Law:

  When a measure becomes a target, it ceases to be a good measure.

Source: Rachel Thomas (@math_rachel)



Flawed Algorithms Are Grading Millions of Students' Essays

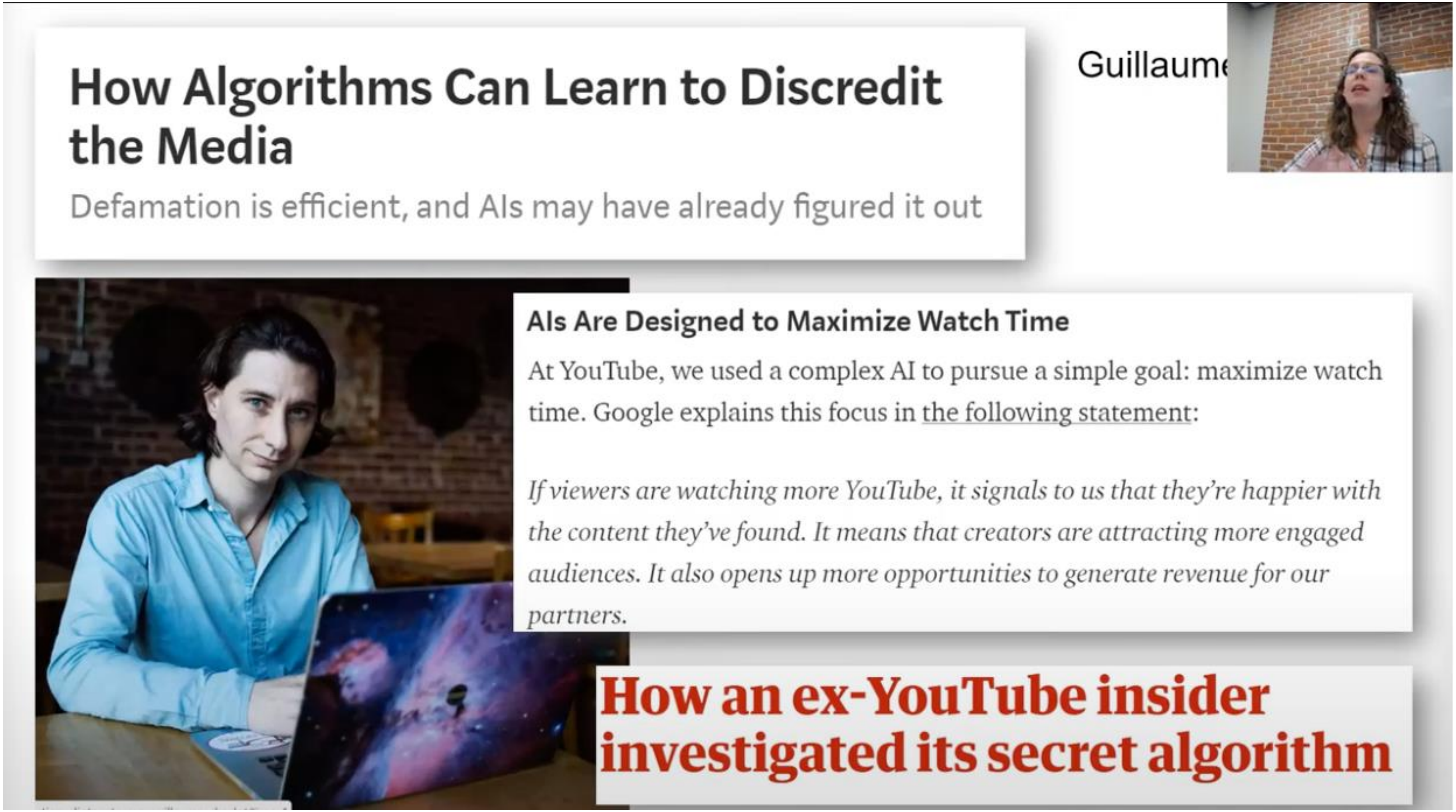
Fooled by gibberish and highly susceptible to human bias, automated essay-scoring systems are being increasingly adopted, a Motherboard investigation has found

Understanding Mean Score Differences Between the *e-rater®* Automated Scoring Engine and Humans for Demographically Based Groups in the *GRE®* General Test

Chaitanya Ramineni, David Williamson

- Automatic essay grading software used in at least 22 USA states
- Focuses on metrics like sentence length, vocabulary, spelling, subject-verb agreement
- Can't evaluate hard-to-quantify qualities, like creativity
- Gibberish essays with lots of sophisticated words score well
- Essays by African-American students receive **lower grades** from computer than from expert human graders
- Essays by students from mainland China receive **higher scores** from computer than from expert human graders; may be using chunks of pre-memorized text

Source: Rachel Thomas (@math_rachel)



Guillaum

How Algorithms Can Learn to Discredit the Media

Defamation is efficient, and AIs may have already figured it out

AIs Are Designed to Maximize Watch Time

At YouTube, we used a complex AI to pursue a simple goal: maximize watch time. Google explains this focus in the following statement:

If viewers are watching more YouTube, it signals to us that they're happier with the content they've found. It means that creators are attracting more engaged audiences. It also opens up more opportunities to generate revenue for our partners.

How an ex-YouTube insider investigated its secret algorithm

Source: Rachel Thomas (@math_rachel)

Where to Go Next?

Where to go next?

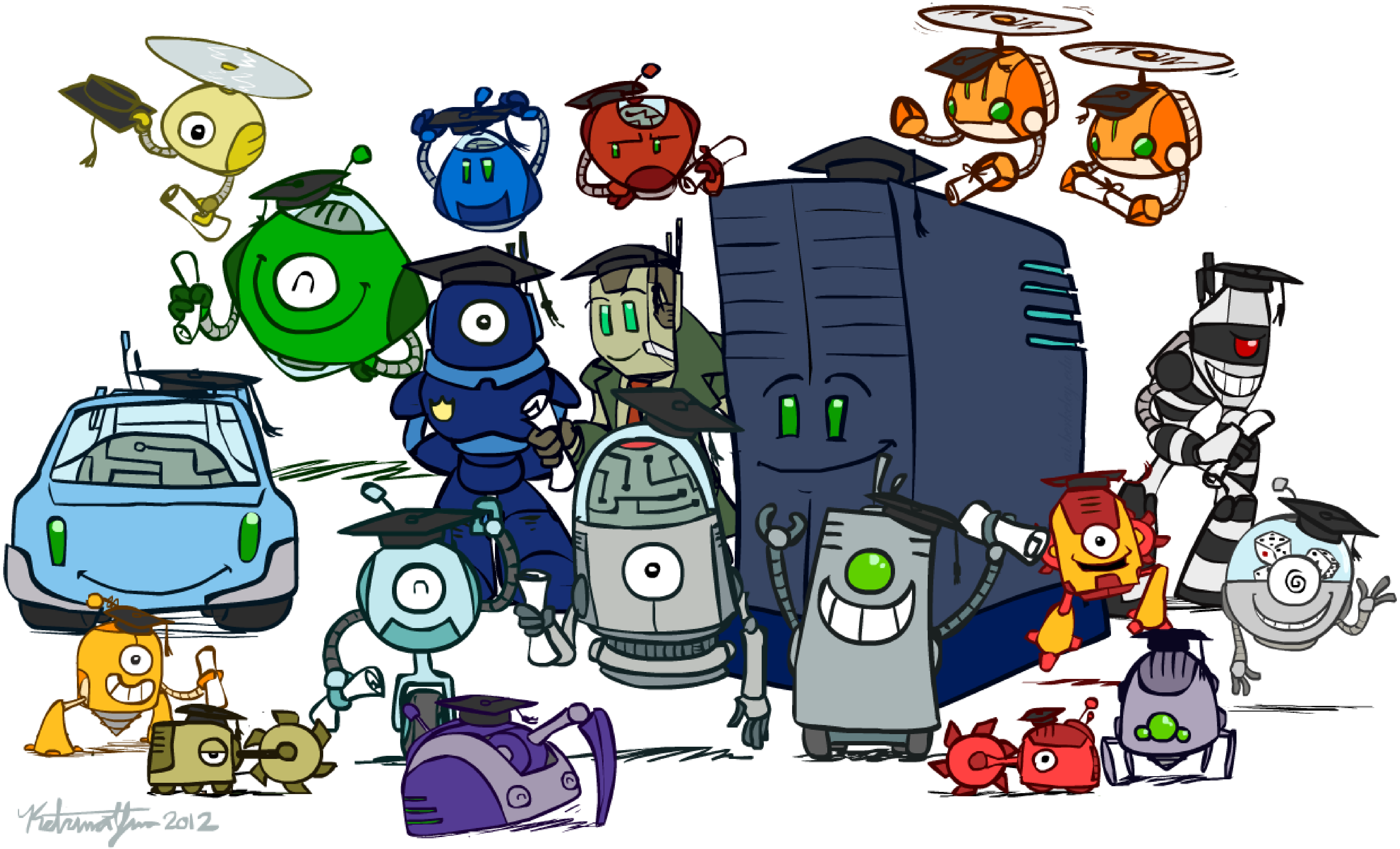
- Congratulations, you’ve seen the basics of modern AI
  - … and done some amazing work putting it to use!

- How to continue:
  - Machine learning:
  - Data Science:
  - Data / Ethics:
  - Probability:
  - Optimization:
  - Computer vision:
  - Reinforcement Learning:
  - Robotics:
  - NLP:
  - … and more; ask if you’re interested

That’s It!

- Help us out with some course evaluations

- Have a great spring break


2012