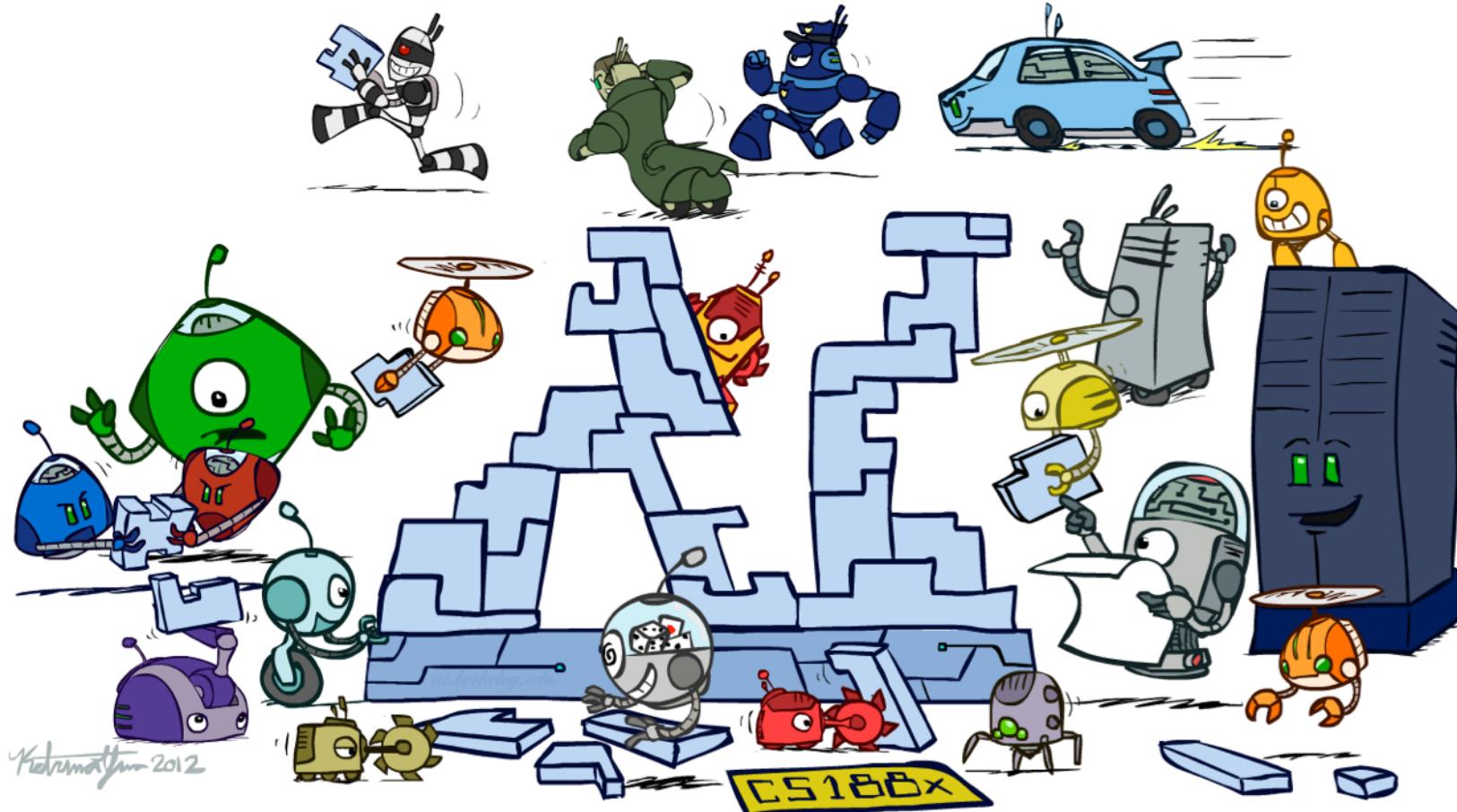
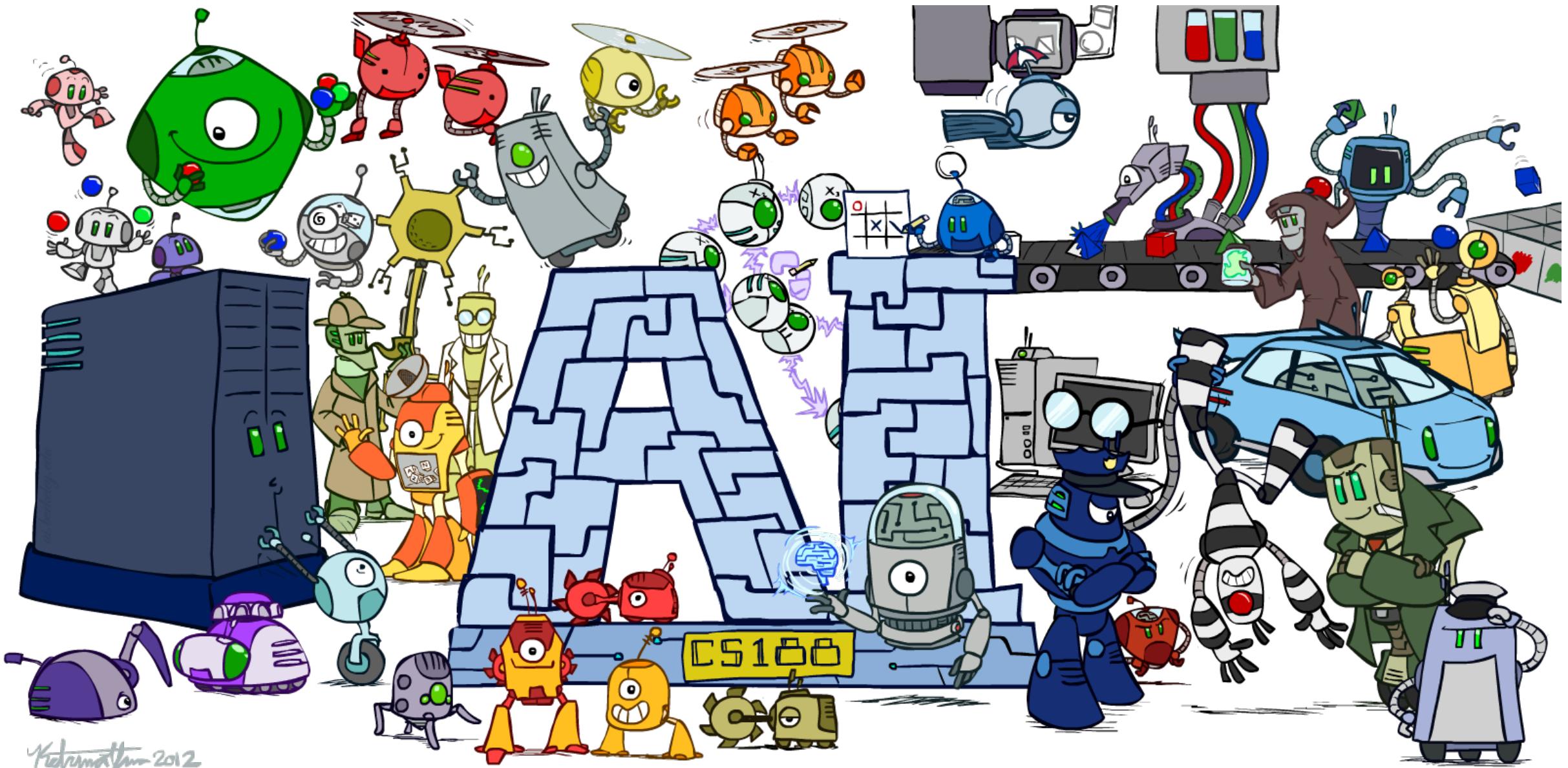


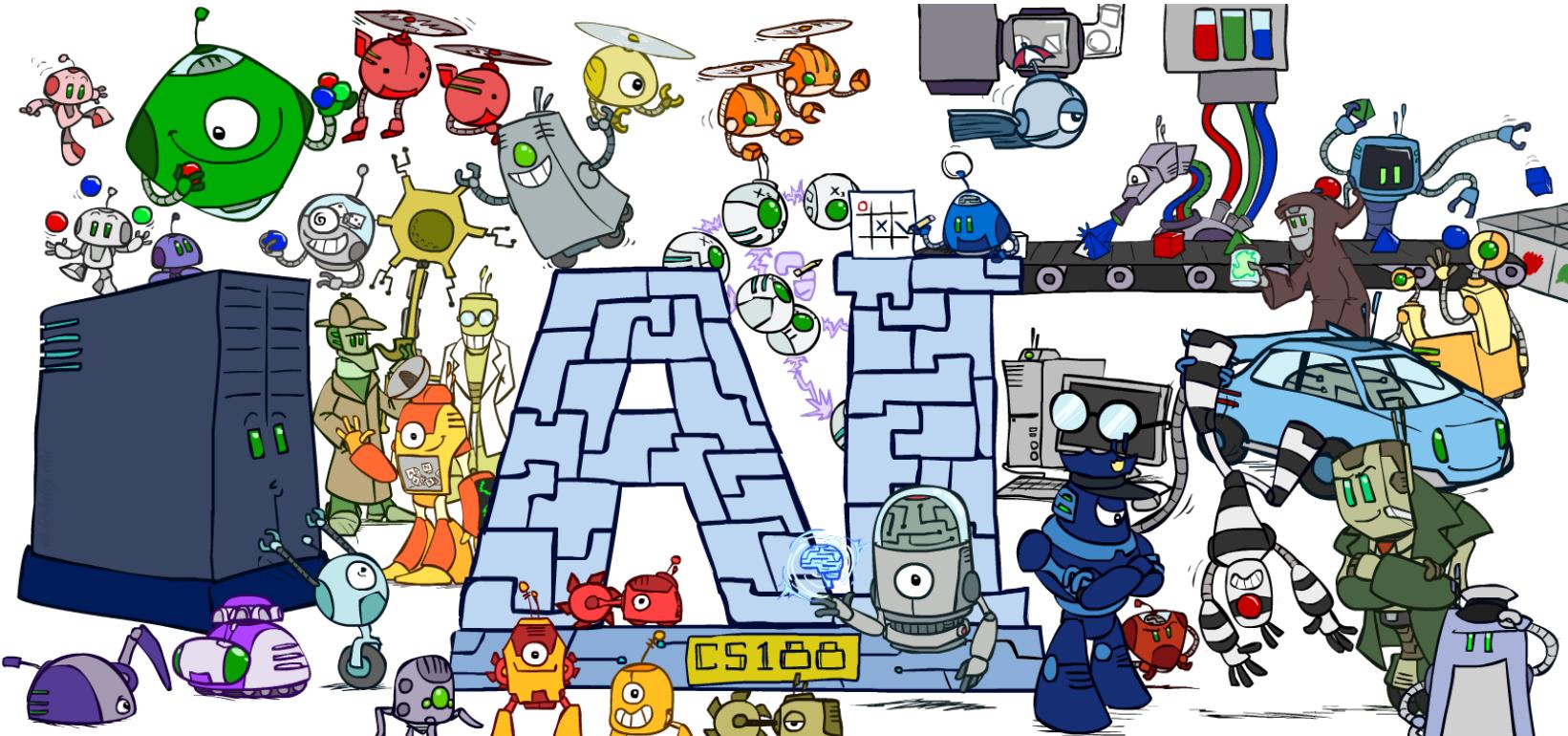
CS 188: Artificial Intelligence

Conclusion



Instructor: Pieter Abbeel --- University of California, Berkeley



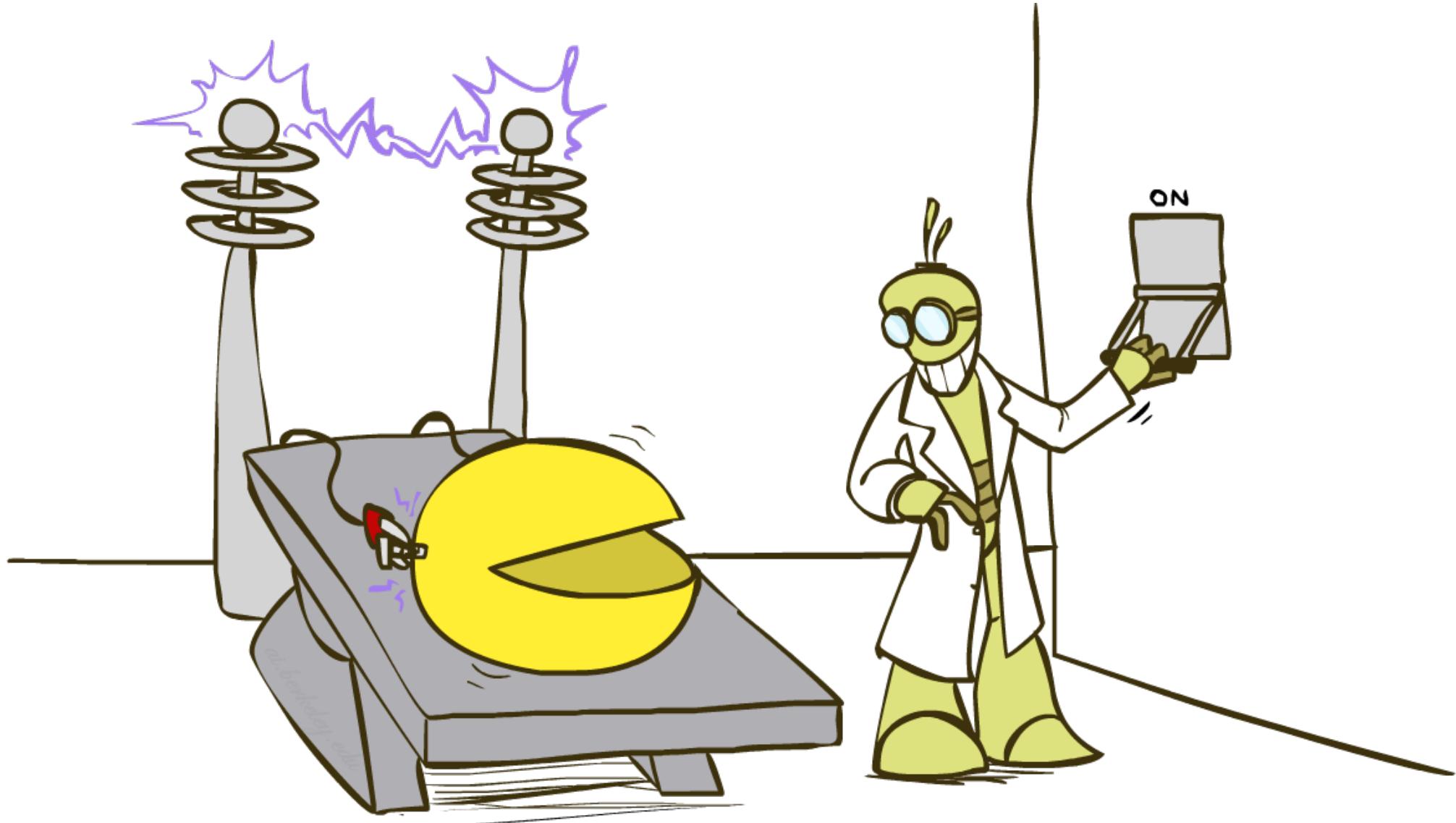


Ketrina Yim
2012



Ketrina Yim
CS188 Artist

Pac-Man Beyond the Game!



Pacman: Beyond Simulation?



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

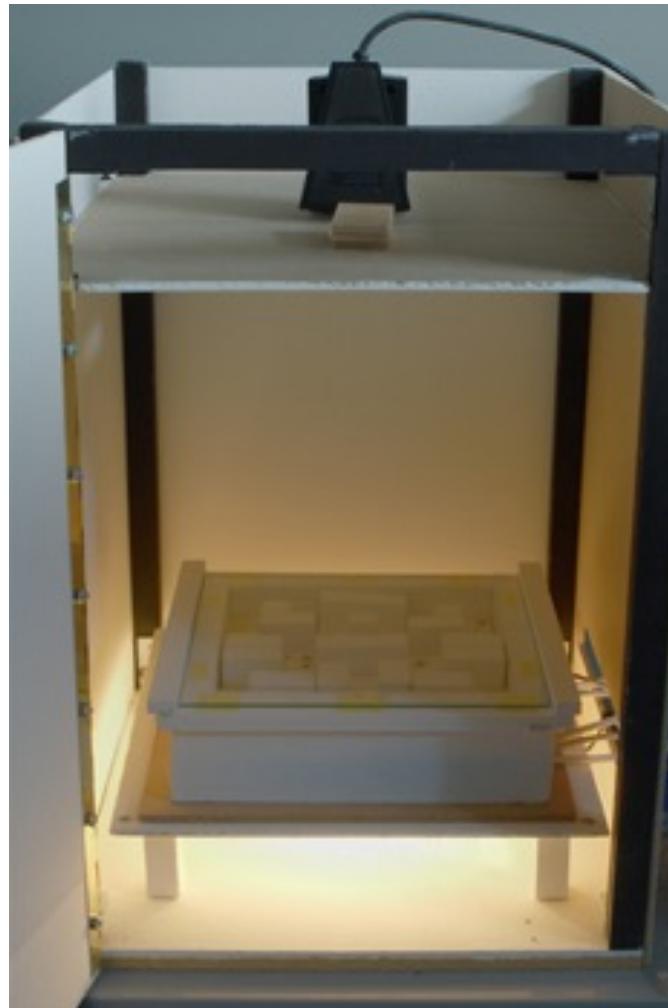
[VIDEO: Roomba Pacman.mp4]

Pacman: Beyond Simulation!

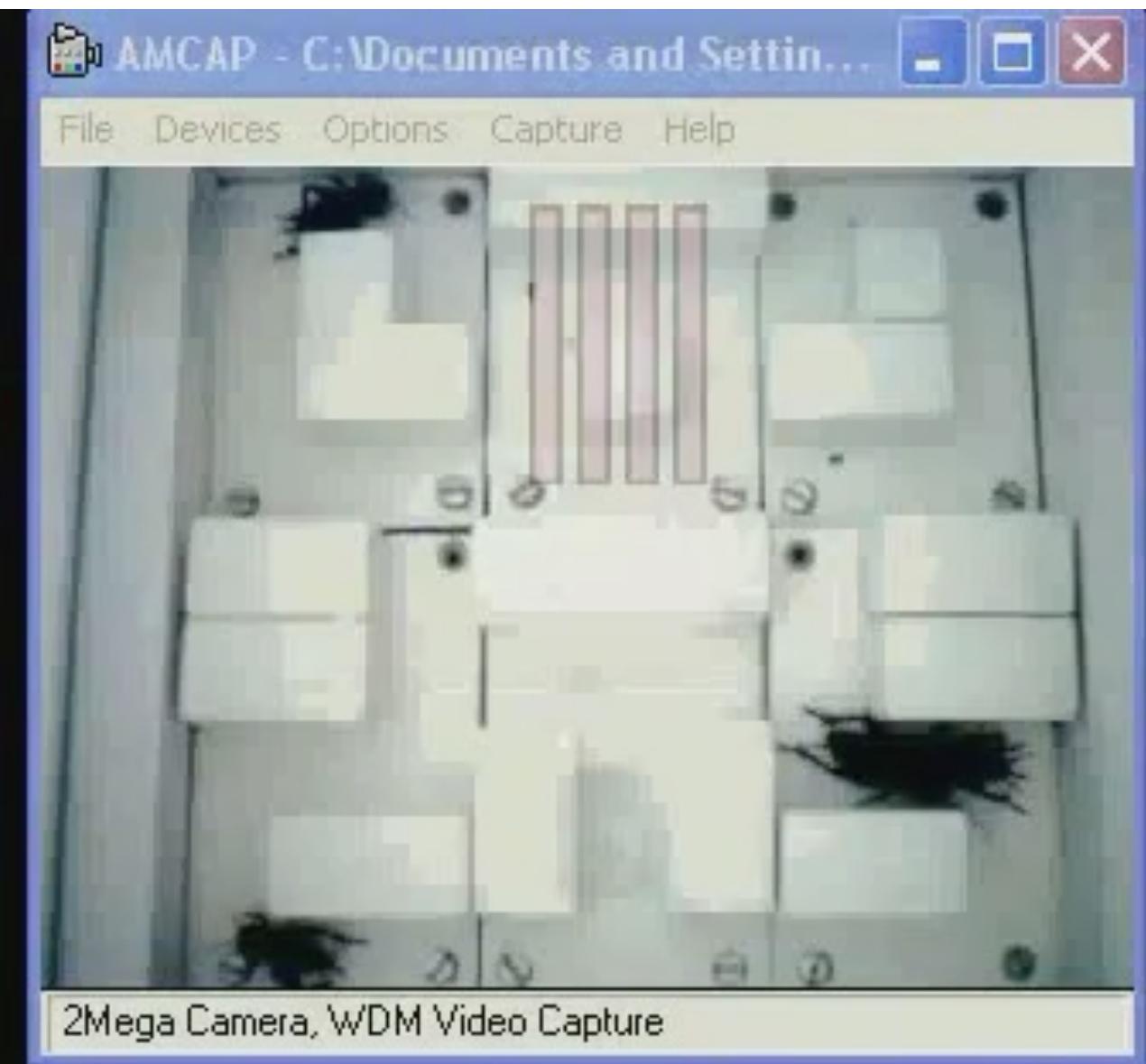


Bugman?

- 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



Bugman



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:

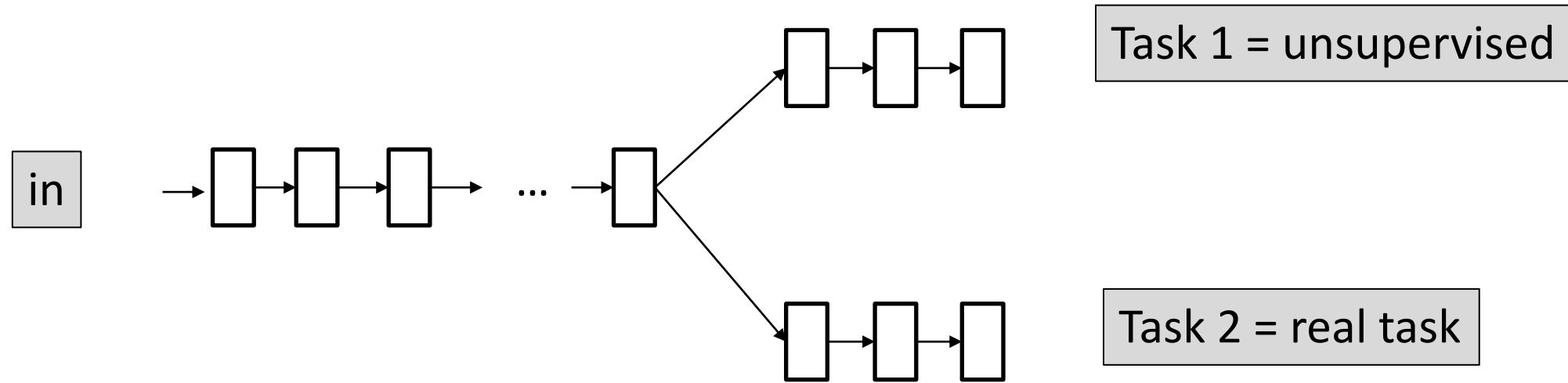
Task 1

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

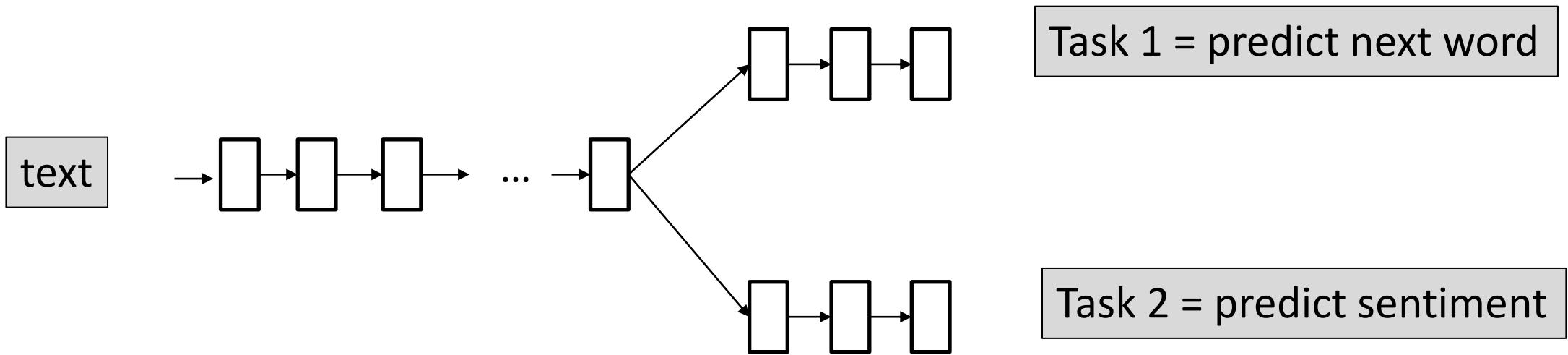
Task 2

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

Transfer from Unsupervised Learning



Example Setting



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 Generation (it generator so good, it's considered too dangerous... considered too dangerous to release ... was trained to ... said, it's ...'s GPT-2)

*In a shocking find, a
remote, previously
surprising to the
English.*

Op
Te
C
S

OpenAI built a text generator so good, it's considered too dangerous..

OpenAI said its
only releasing a smaller ver.
Scientists Developed an AI So A
ScienceAlert - 18 Feb 2019
- dangerous to relea

g technology too as
17 Feb 2019
ology could 'absolutely de
m.au - 17 Feb 2019
so good at writing that its
17 Feb 2019

The Rings, Celebrity Gossip - News18 - 18 Feb 2019

all

A close-up photograph of a hand wearing a dark leather work glove. The hand is gripping a metal object that is engulfed in bright orange and yellow flames, representing intense heat or a hot work environment.

When Is Technology Too Dangerous to Release to the Public?
Magazine - 22 Feb 2019

When Is Technology?
Slate Magazine - 22 Feb 2019
Your knowledge of the model
Jack-Founded O

If you know anything about AI, you've likely heard of OpenAI's language model using text from 2019. Experts say OpenAI's controversial AI system has been trained on data from 2019, which has led to some concerning behavior. The company has since removed the 2019 data from its system.

View all

View all

OpenAI's Text Model so Disruptive it's Deemed Too Dangerous To ...
Business Review - 15 Feb 2019
Disruptive it's Deemed Too Dangerous To Release ...
... full research due to concerns over ... We've ...
... can generate ... say ...

OpenAI's Text Model so Disruptive - Computer Business Review - 15 Feb 2019
Please see the full report at [https://www.computerbusinessreview.com/article/2019/02/openais-text-model-so-disruptive-its-deep-learning-revolution-is-here](#)

OpenAI's Text Model so Dangerous it May be Too Dangerous to Release

 OpenAI has trained an unsupervised New AI fake text generator may be highly Cited - The Guardian - 14 Feb 2019

Highly
now

that it is likely that the only descendants are indeed the descendants of "But they seem to be able to

**Release ...
ver ... We've**

"But they seem to be able to which I believe is a sign of social organization," said the

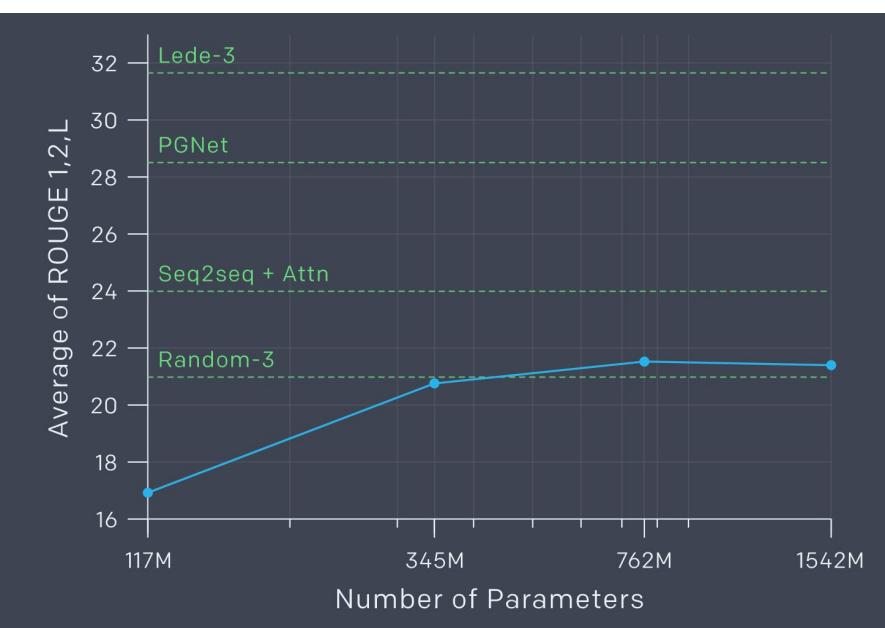
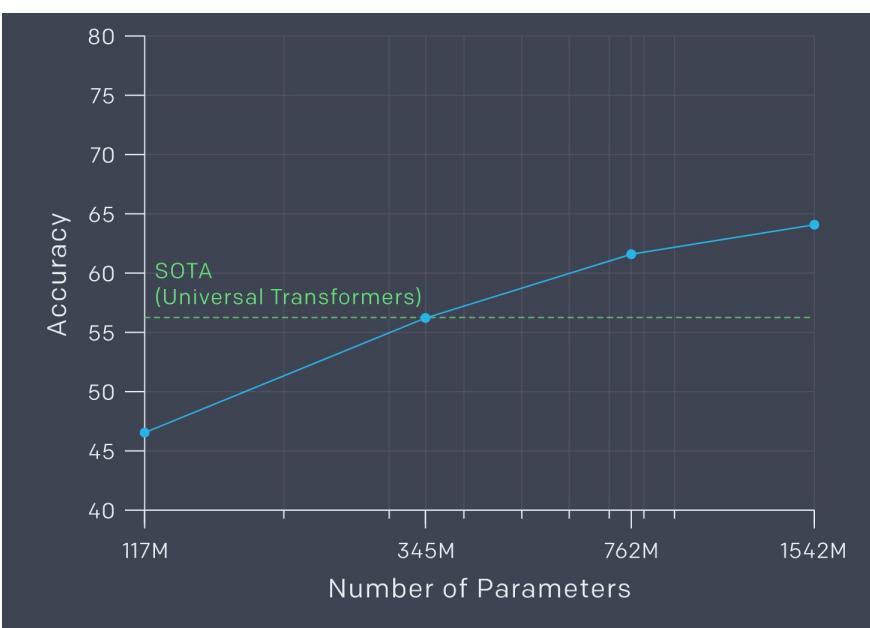
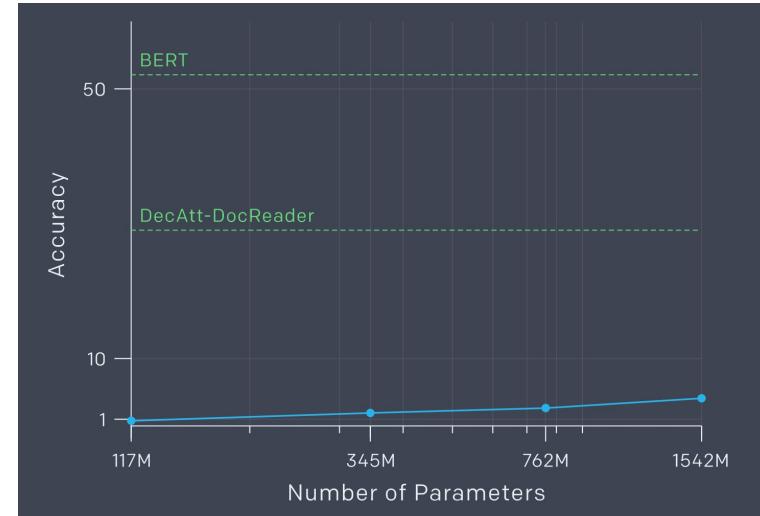
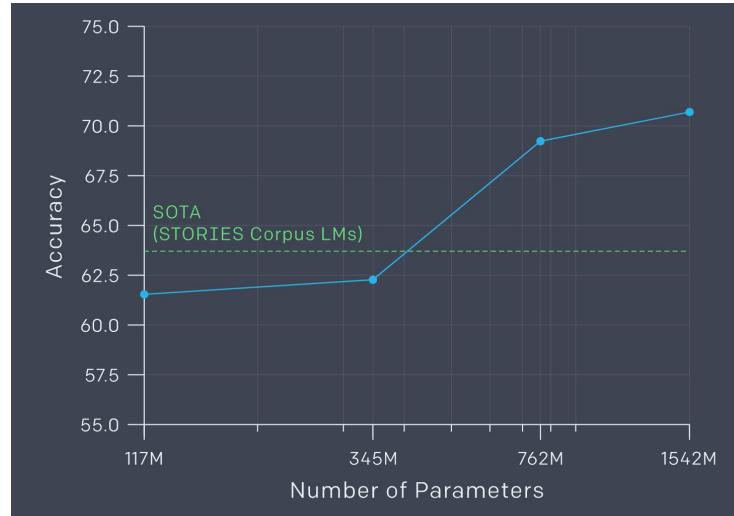
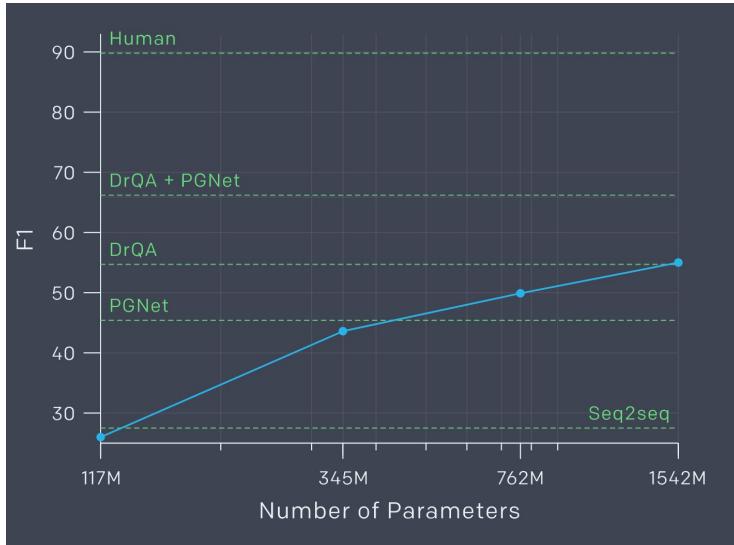
Unsupervised Sentiment Neuron

This is one of Crichton's best books. The characters of Karen Ross, Peter Elliot, Munro, and Amy are beautifully developed and their interactions are exciting, complex, and fast-paced throughout this impressive novel. And about 99.8 percent of that got lost in the film. Seriously, the screenplay AND the directing were horrendous and clearly done by people who could not fathom what was good about the novel. I can't fault the actors because frankly, they never had a chance to make this turkey live up to Crichton's original work. I know good novels, especially those with a science fiction edge, are hard to bring to the screen in a way that lives up to the original. But this may be the absolute worst disparity in quality between novel and screen adaptation ever. The book is really, really good. The movie is just dreadful.

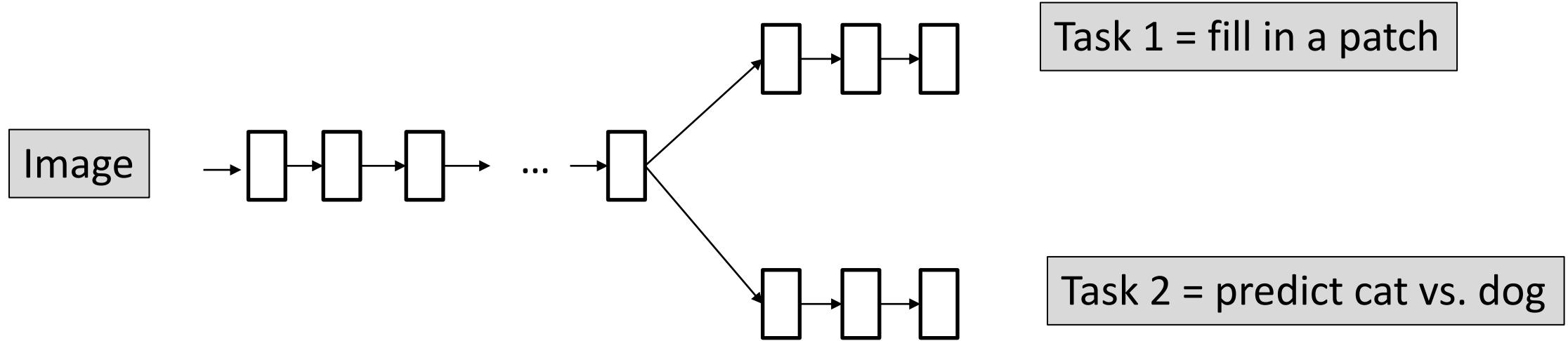
Benchmarks

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



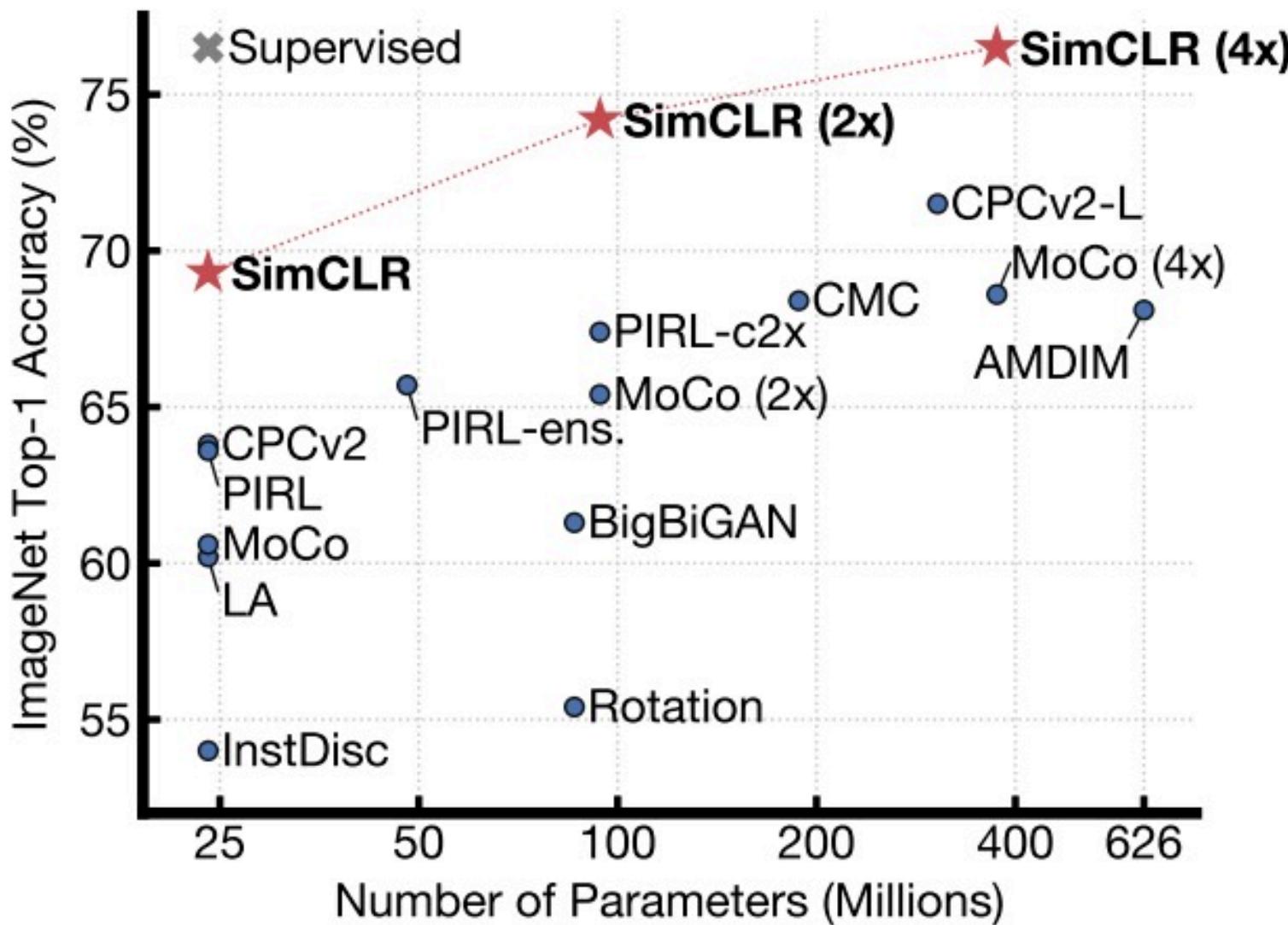
Unsupervised Learning in Vision



Predict Missing Patch



SimCLR + linear classifier



AI for Art Creation

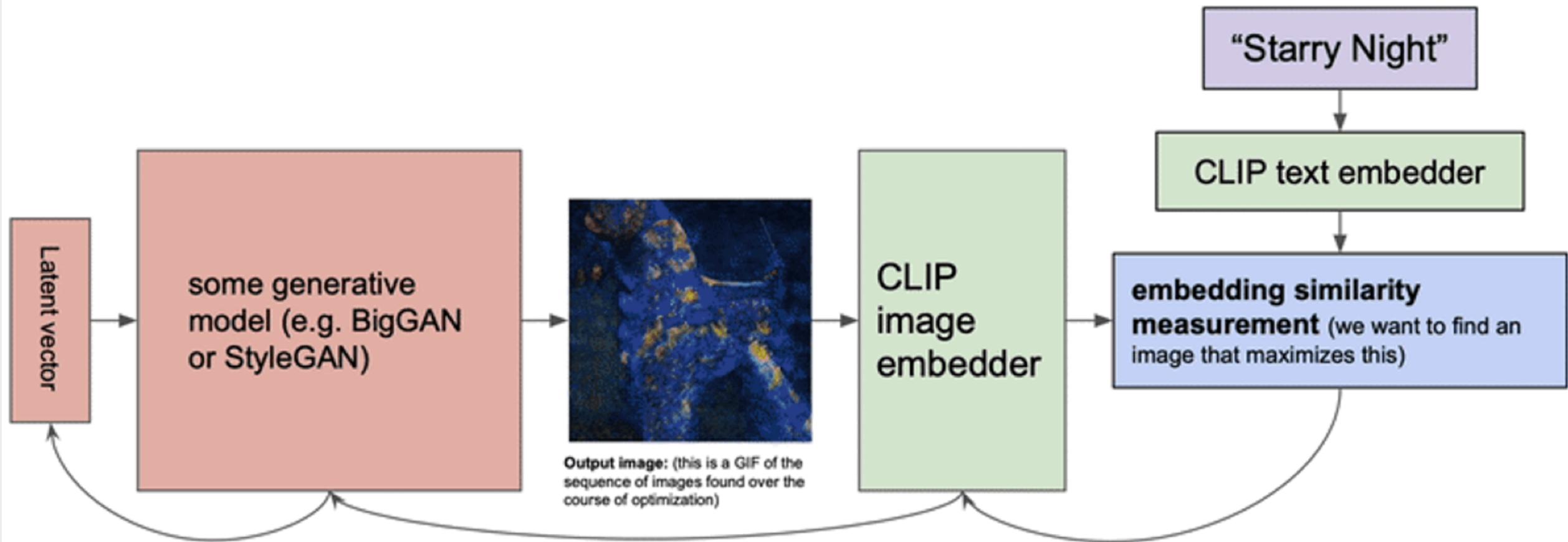


humanoid robot Mona Lisa
artstationHQ



studio ghibli trending on artstation | vary

Text-Guided Image Generation



Examples (CLIP + VQGAN)



Research Frontiers

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

Also:

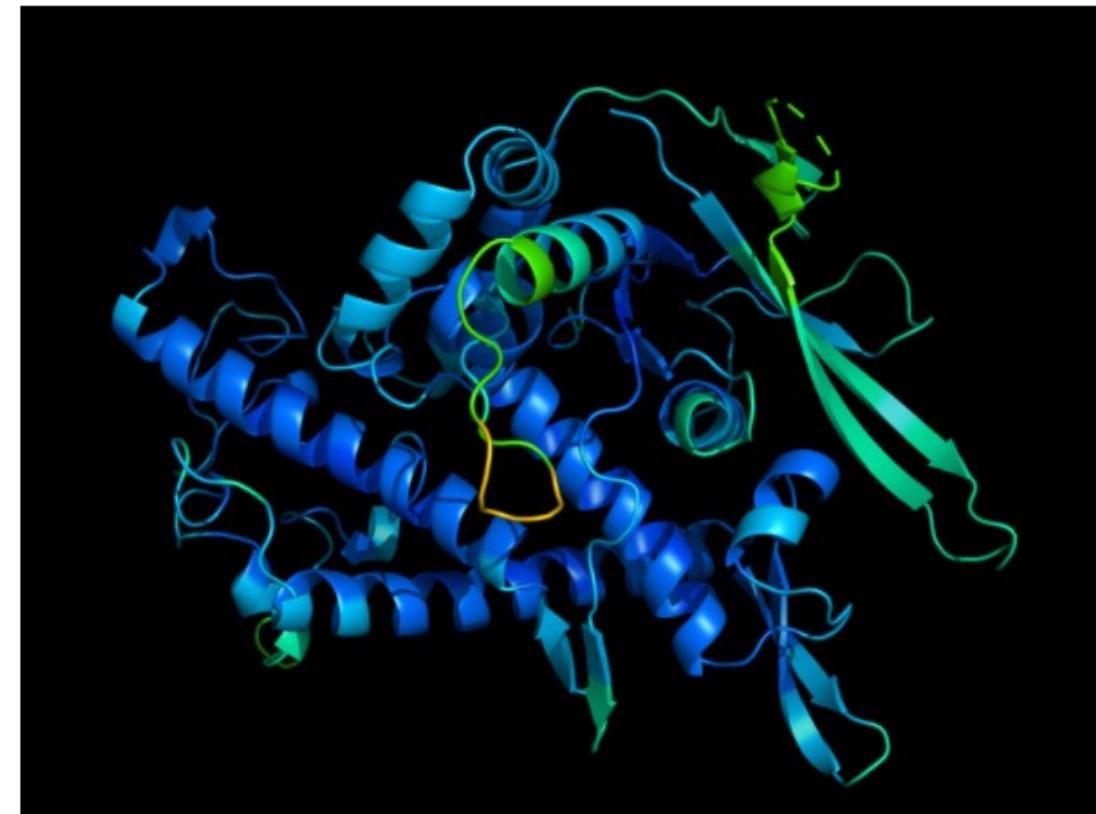
- Unsupervised Deep Reinforcement Learning
- Human-in-the-loop Reinforcement Learning
- ...

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.

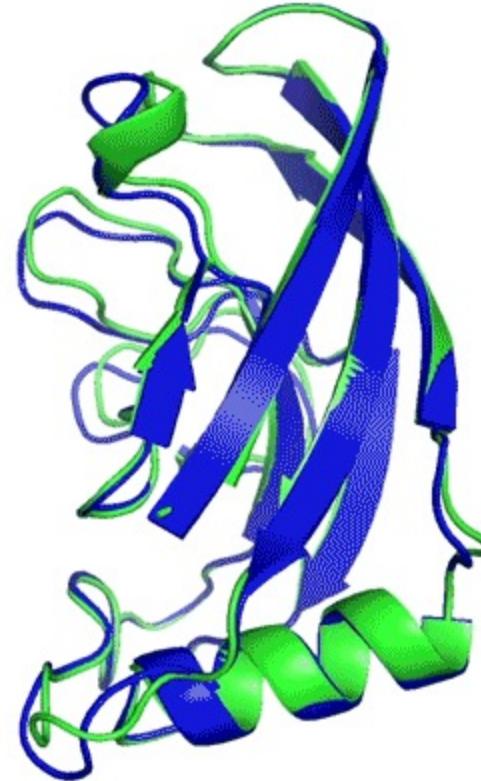
Ewen Callaway



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



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

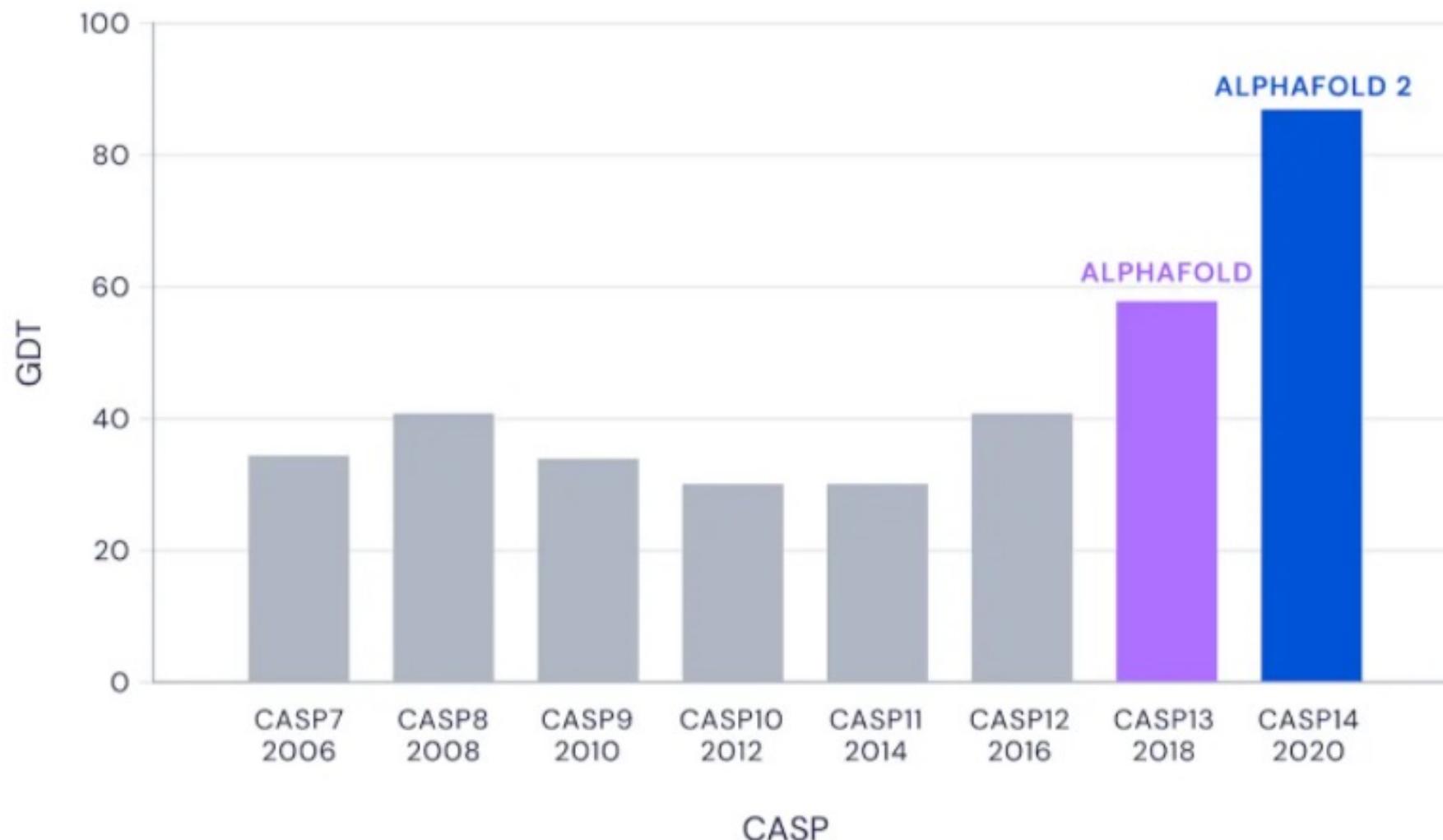


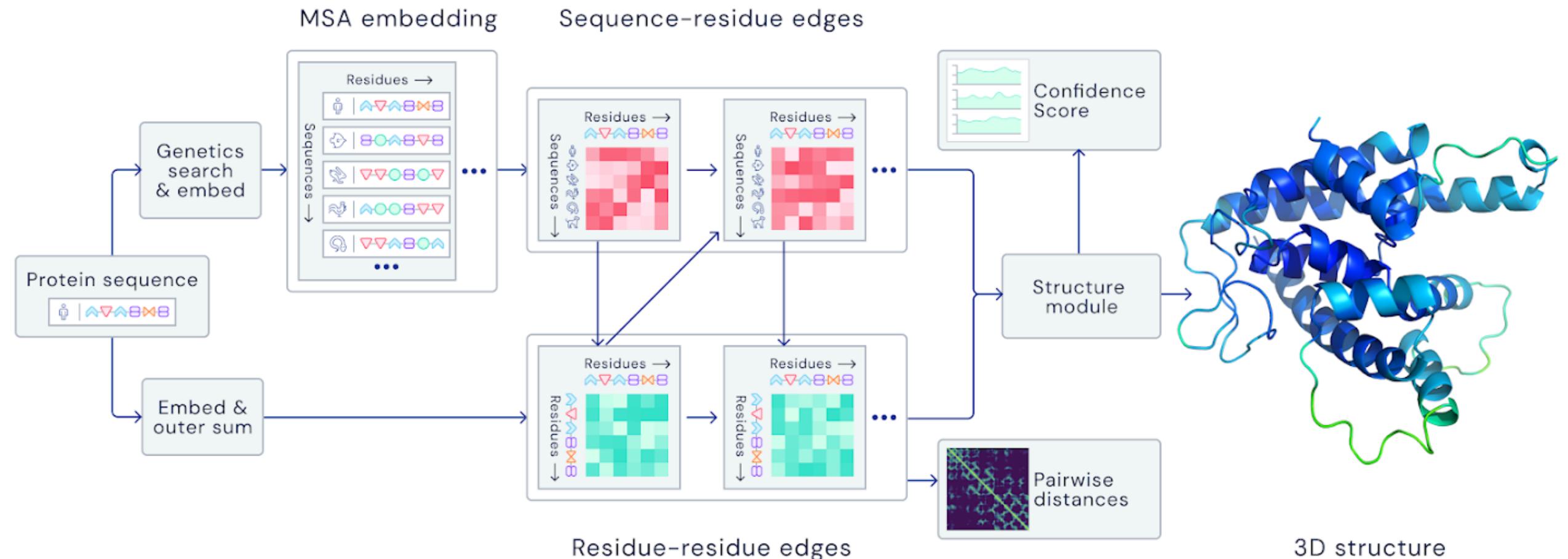
T1049 / 6y4f
93.3 GDT
(adhesin tip)

- Experimental result
- Computational prediction

CASP 2020 Competition

Median Free-Modelling Accuracy



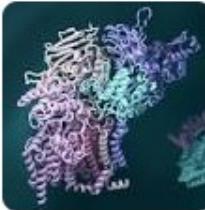


 Singularity Hub

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



 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

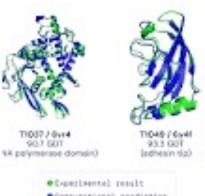


 TechStory

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



 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

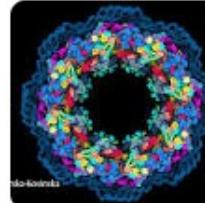


 Nature

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



 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



 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



 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

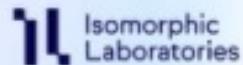




Demis Hassabis 
@demishassabis

...

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



Reimagining
drug discovery

isomorphiclabs.com

Symbolic Math: Integrals and ODEs

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

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.

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

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 experts you'll want to learn from



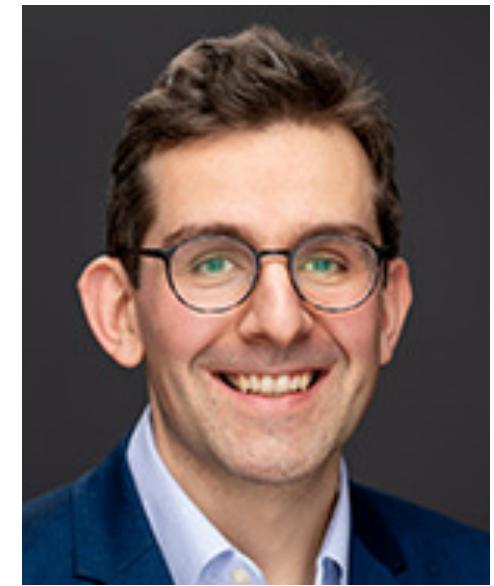
Prof. Rediet Abebe
UC Berkeley



Prof. Rachel Thomas
Univ. of San Francisco



Dr. Timnit Gebru
Google



Prof. Moritz Hardt
UC Berkeley / MPI

Some Key AI Ethics Topics

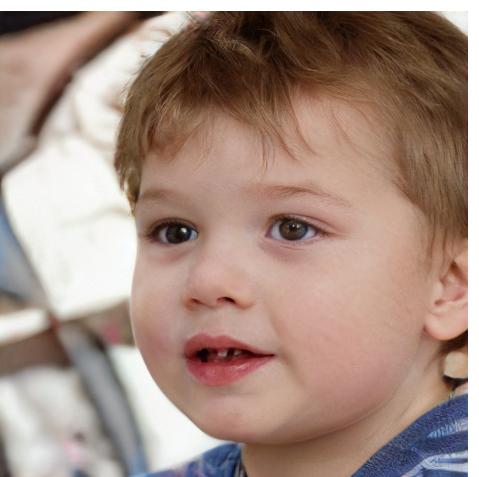
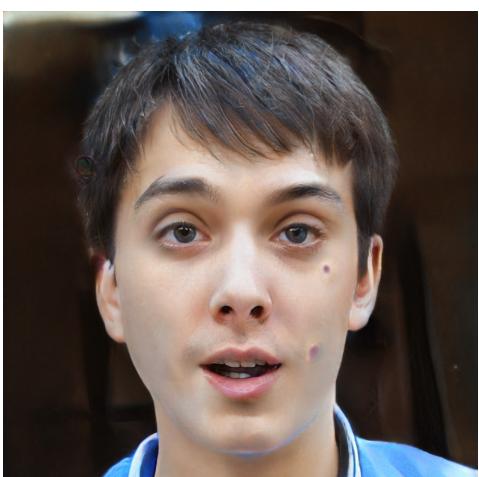
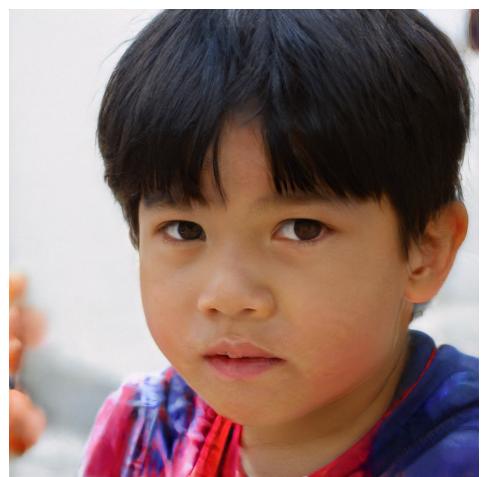
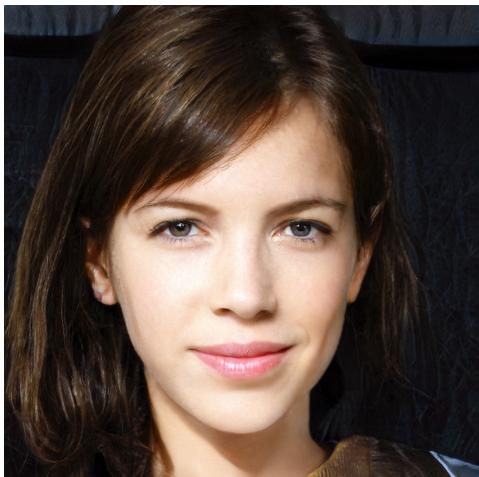
- Disinformation
- Bias and fairness
- Privacy and surveillance
- Metrics
- Algorithmic colonialism

Disinformation can be aided by AI generated text

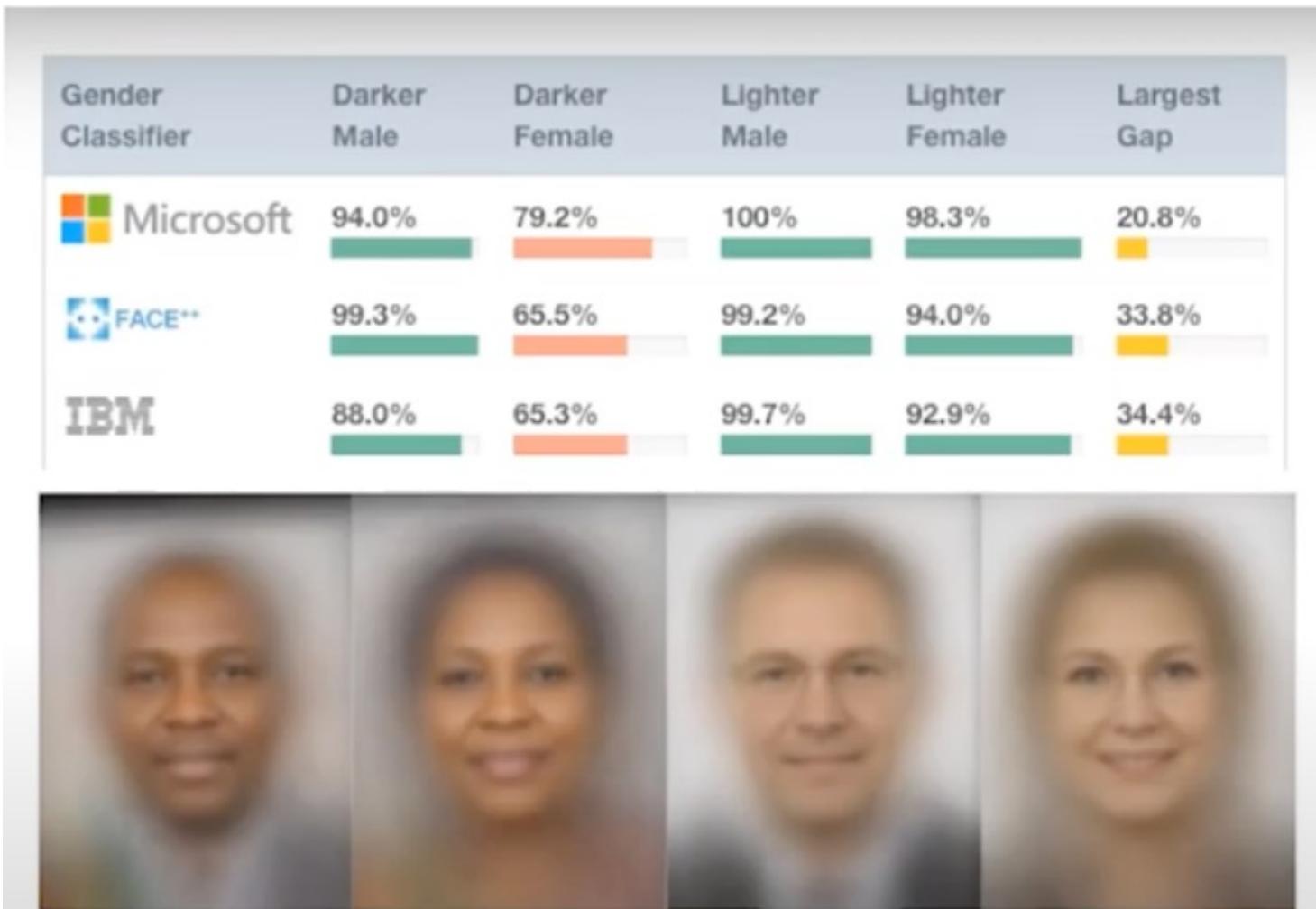
Prompt: Recycling is good for the world, no, you could not be more wrong.

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.

Disinformation can be aided by Deep Fakes



Bias and Fairness – Gender Shades



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 Latanya Sweeney Truth

Ads by Google
[Latanya Sweeney, Arrested?](#)
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

[Latanya Sweeney](#)
Public Records Found For: Latanya Sweeney
www.publicrecords.com/

[La Tanya](#)
Search for La Tanya Look Up Fast Results
www.ask.com/La+Tanya

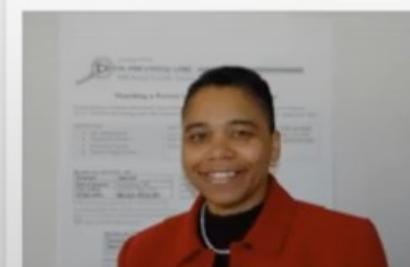
Ads by Google
[Kirsten Lindquist](#)
Get Kirsten Lindquist Find Kirsten Lindquist
www.ask.com/Kirsten+Lindquist

We Found:[Kristen Lindquist](#)
1) Contact Kristen Lindquist - Free Info! 2) Current Phone, Address & More.
www.peoplesmart.com/

[Search by Phone](#) [Search by Email](#)
[Background Checks](#) [Search by Address](#)
[Public Records](#) [Criminal Records](#)

[Kristen Lindquist](#)
Public Records Found For: Kristen Lindquist. View Now.
www.publicrecords.com/

Latanya Sweeney, Ph.D.

A portrait photograph of a Black woman with short dark hair, wearing a red blazer over a black top. She is smiling at the camera. The background is a light-colored wall with some text and logos, possibly from a professional organization.

Source: Rachel Thomas (@math_rachel)

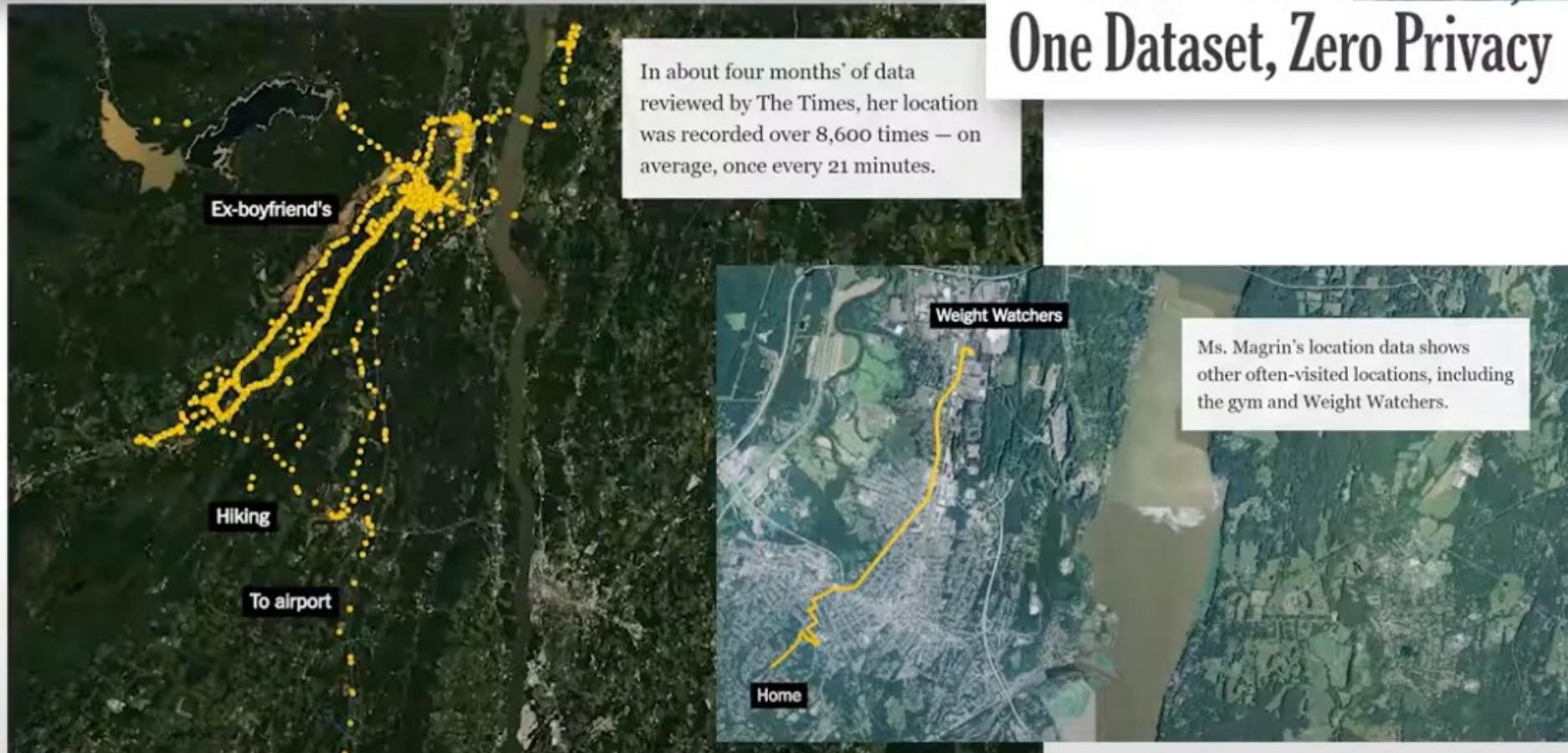
Bias and Fairness

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

Privacy and Surveillance

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



Twelve Million
One Dataset, Zero Privacy

Source: Rachel Thomas (@math_rachel)



Data contains errors

42 babies were < 1 year old at the time they were entered into the database

28 of those were marked as “admitting to being gang members”

California gang database plagued with errors, unsubstantiated entries, state auditor finds

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.

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

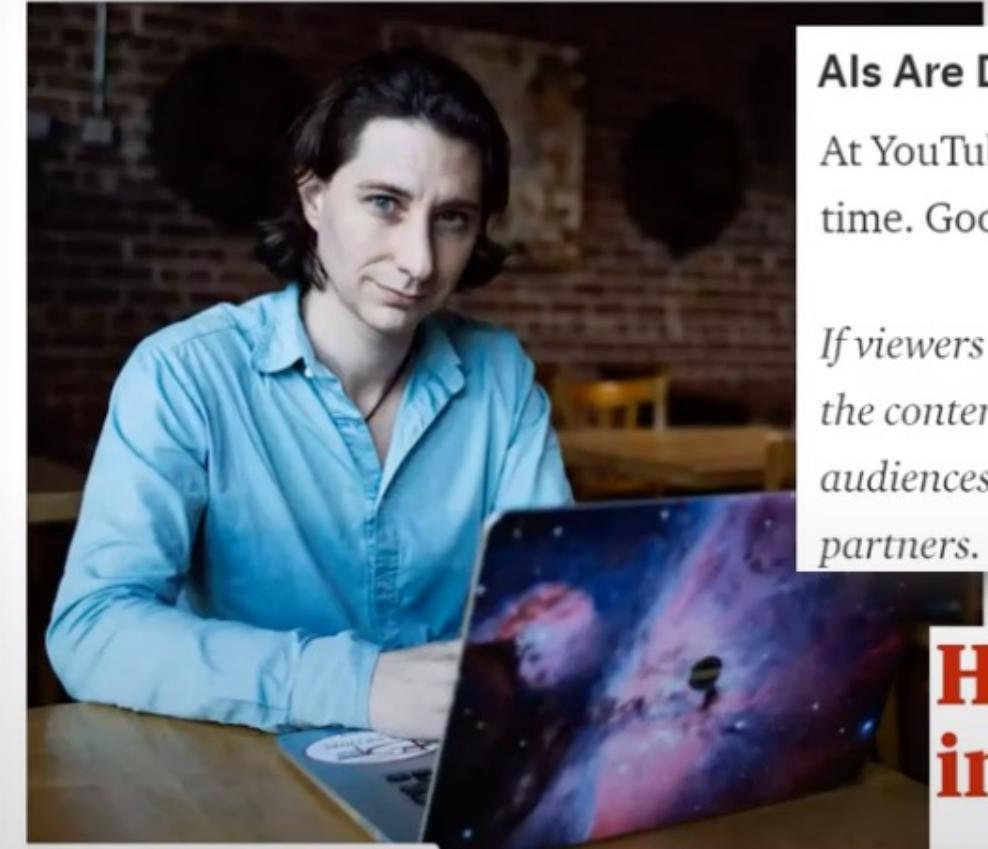
Chaitanya Ramineni ✉, David Williamson

- 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

How Algorithms Can Learn to Discredit the Media

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

Guillaume



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.

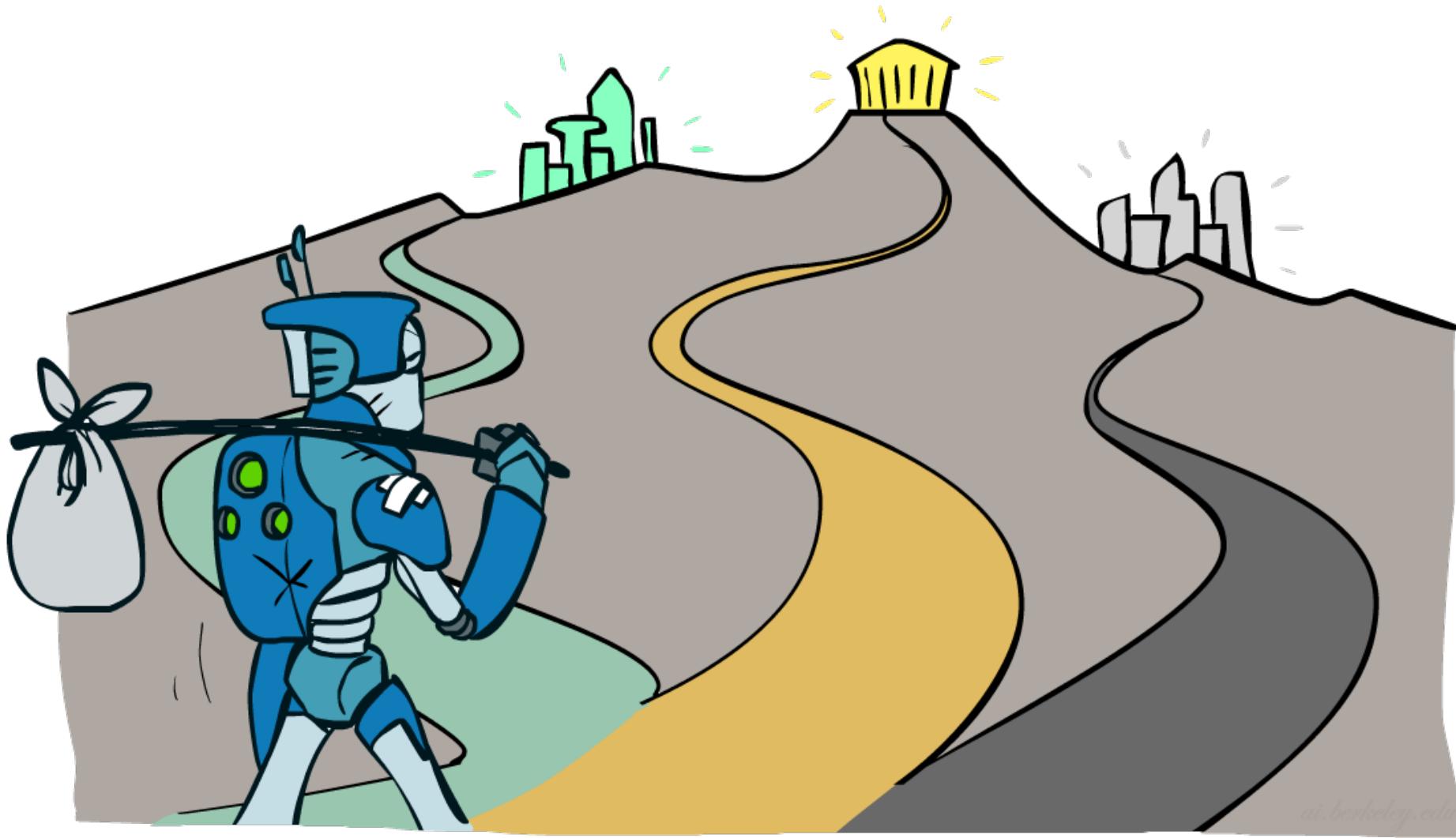
How an ex-YouTube insider investigated its secret algorithm

Source: Rachel Thomas (@math_rachel)

Algorithmic Colonialism

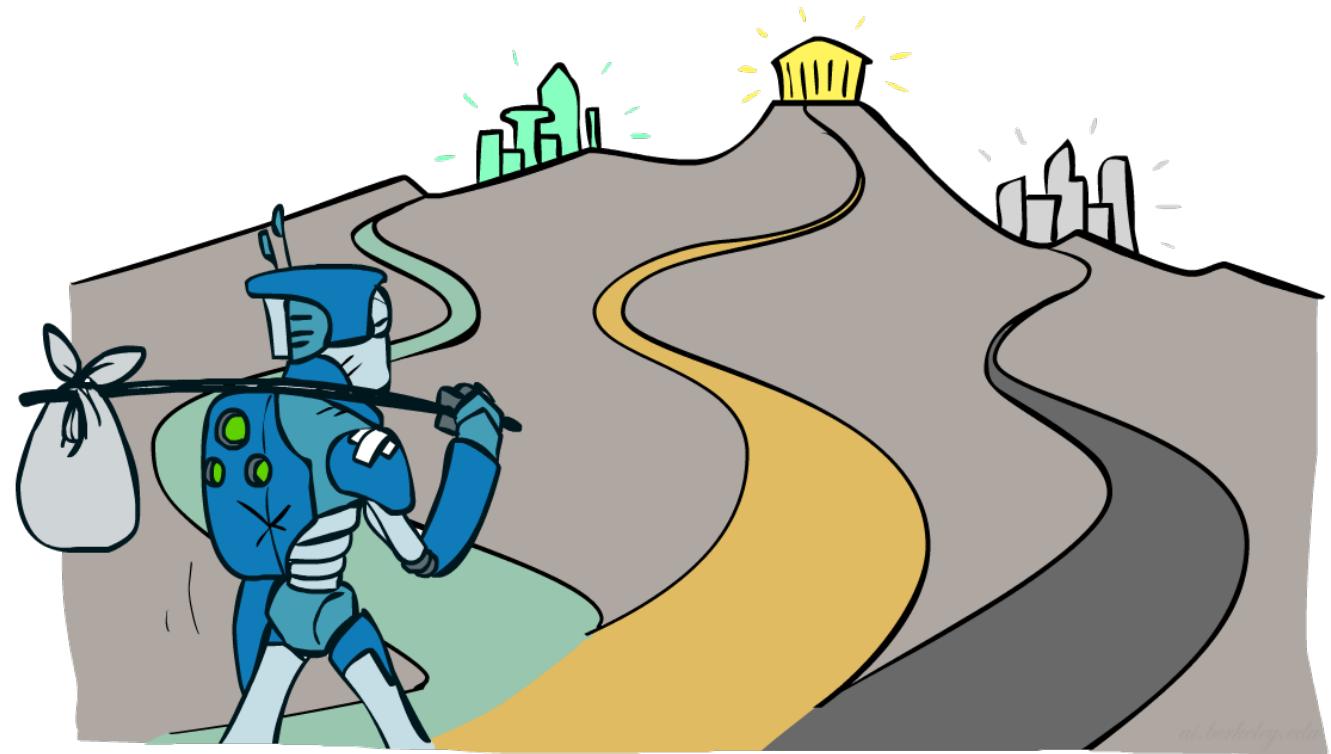
- When corporations from one country develop and deploy technology in many other countries, extracting data and profits, often with little awareness of local cultural issues, a number of ethical issues can arise.

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: cs189, cs182, stat154
 - Data Science: data 100, data 102
 - Data / Ethics: data c104
 - Probability: ee126, stat134
 - Optimization: ee127
 - Cognitive modeling: cog sci 131
 - Machine learning theory: cs281a/b
 - Computer vision: cs280
 - Reinforcement Learning: cs285
 - Robotics: cs287, cs287h
 - NLP: cs288
 - ... and more; ask if you're interested



Lightweight Opportunities to Keep Learning



- Andrew Ng weekly newsletter:

The Batch: <https://www.deeplearning.ai/thebatch/>



- Jack Clark (former Comms Director OpenAI) weekly newsletter:

Import AI: <https://jack-clark.net/>



- Rachel Thomas AI Ethics course:

Course website: ethics.fast.ai



- Pieter Abbeel podcast:

The Robot Brains Podcast: <https://therobotbrains.ai>

That's It!

- Help us out with some course evaluations
- Have a great winter break, and always maximize your expected utilities!



Ketkunath - 2012