Governing the AI Revolution

Allan Dafoe

Governance of AI Program
Future of Humanity Institute
University of Oxford

governance.ai

Outline

- AI Today
- Governance Challenges
- AI Tomorrow
- Canada's AI Grand Strategy

What is AI?

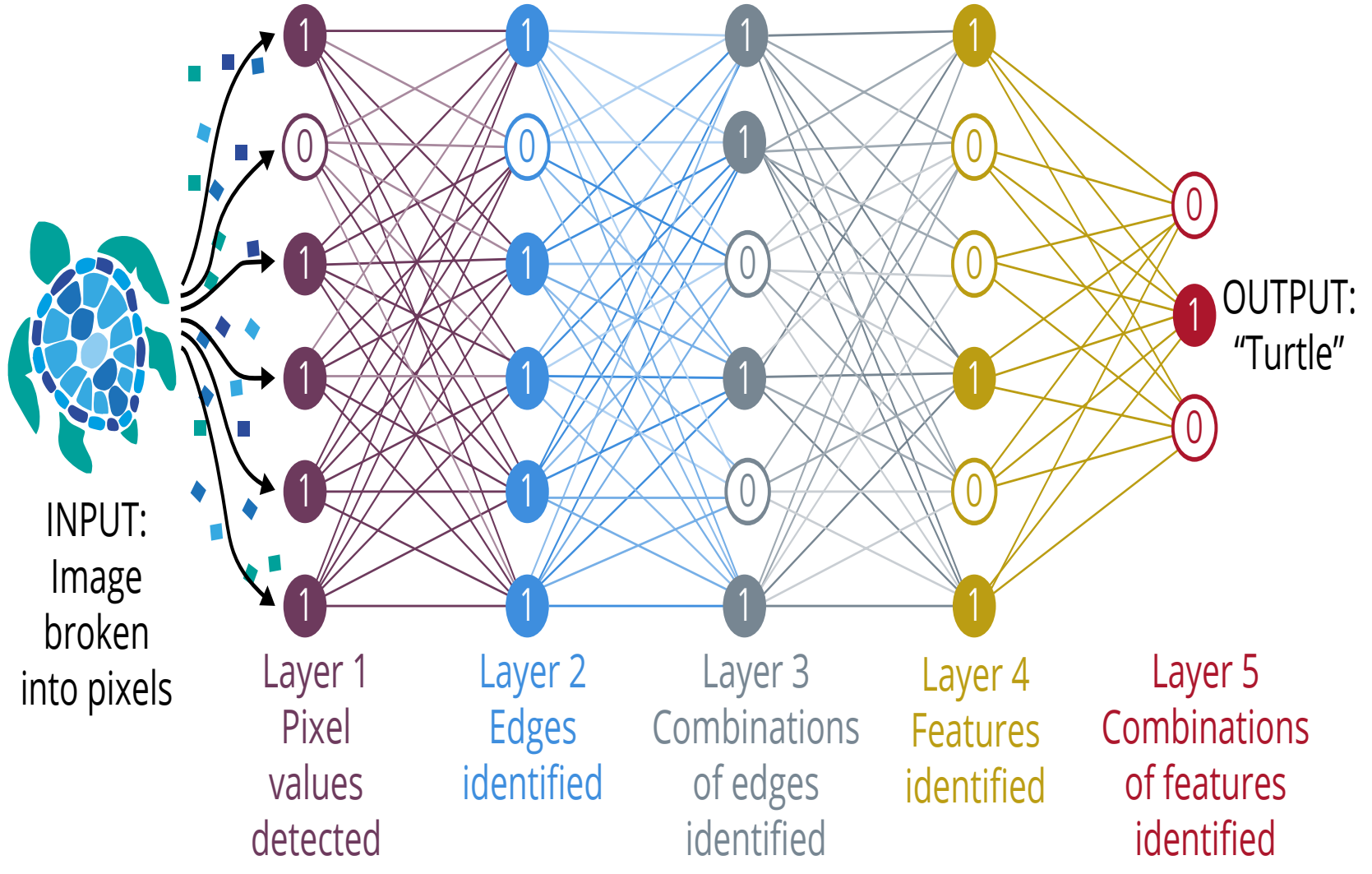
Artificial intelligence (AI): machines able to perform tasks involving sophisticated information processing (eg normally requiring human intelligence).

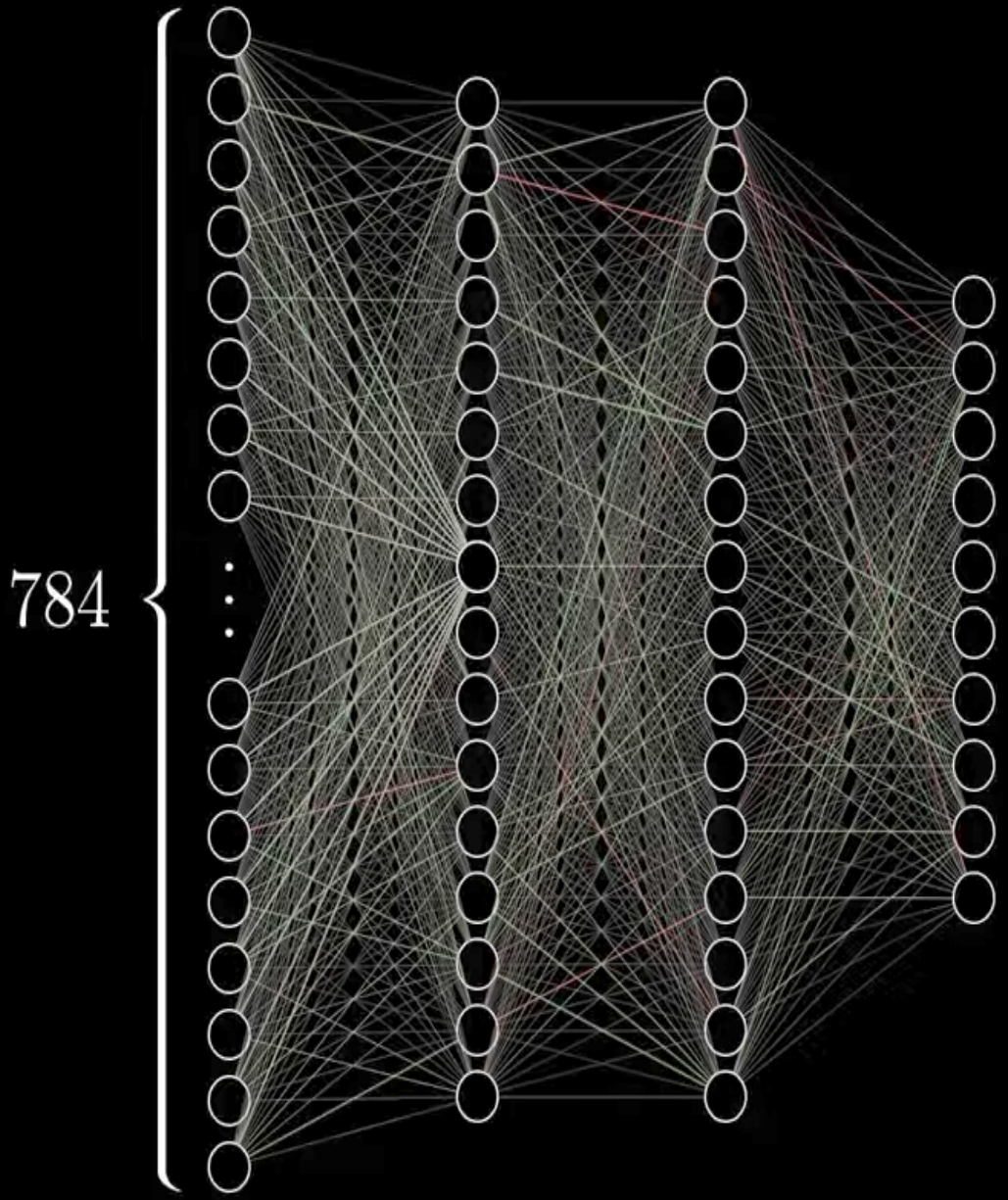
(sensors, input) + algorithms + (actuators/output)

Kasparov vs Deep Blue, 1997

Machine / Deep Learning

Machine / Deep Learning



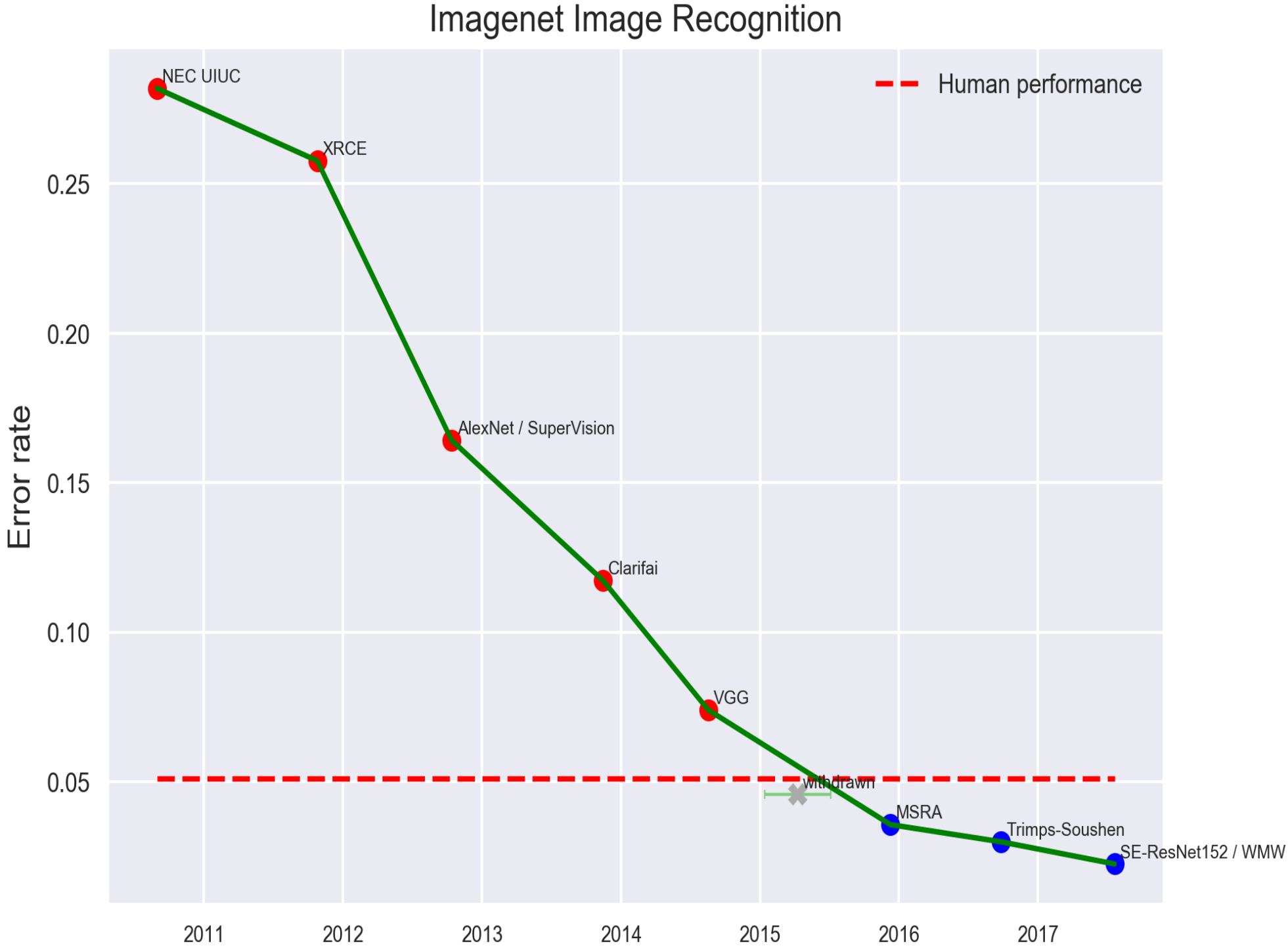


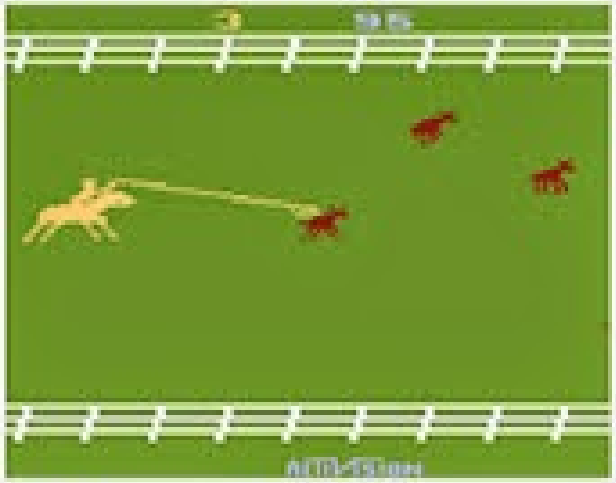
$784 \times 16 + 16 \times 16 + 16 \times 10$
weights

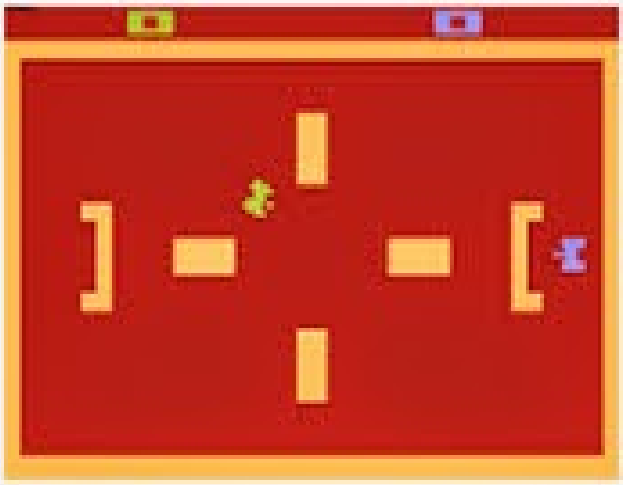
$16 + 16 + 10$
biases

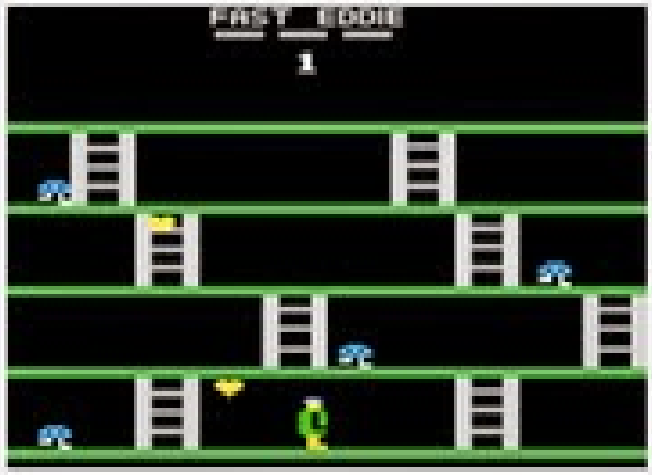
13,002

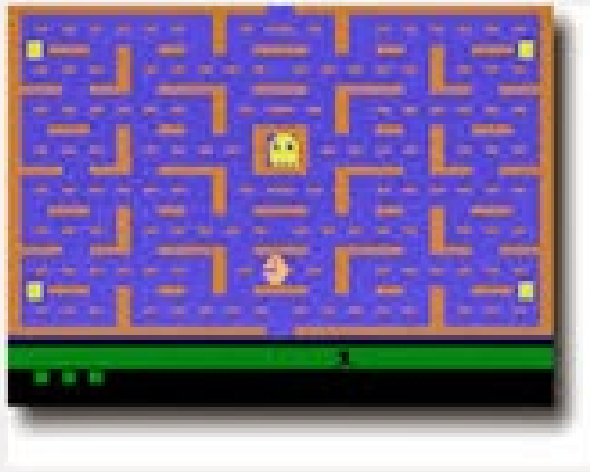
mite	container ship	motor scooter	leopard
mite	container ship	motor scooter	leopard
black widow	lifeboat	go-kart	jaguar
cockroach	amphibian	moped	cheetah
tick	fireboat	bumper car	snow leopard
starfish	drilling platform	golfcart	Egyptian cat


Imagenet Image Recognition
Human performance
NEC UIUC
XRCE
AlexNet / SuperVision
Clarifai
VGG
Withdrawn
MSRA
Trimps-Soushen
SE-ResNet152 / WMW
Error rate
0.25
0.20
0.15
0.10
0.05
2011
2012
2013
2014
2015
2016
2017



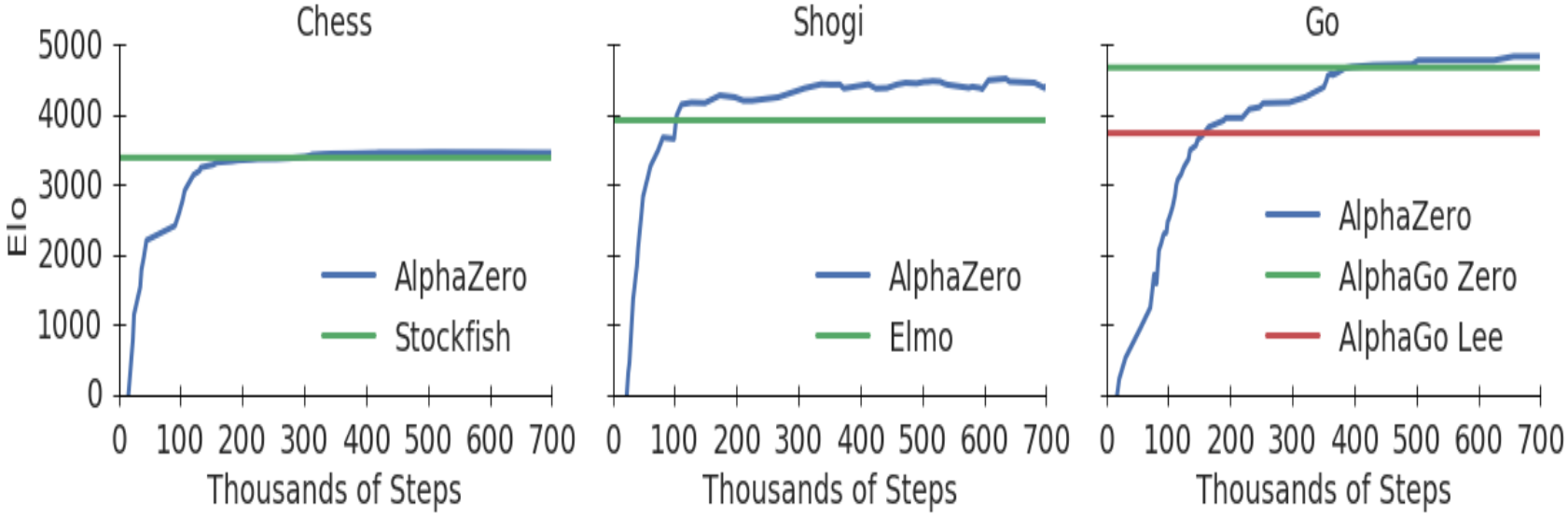






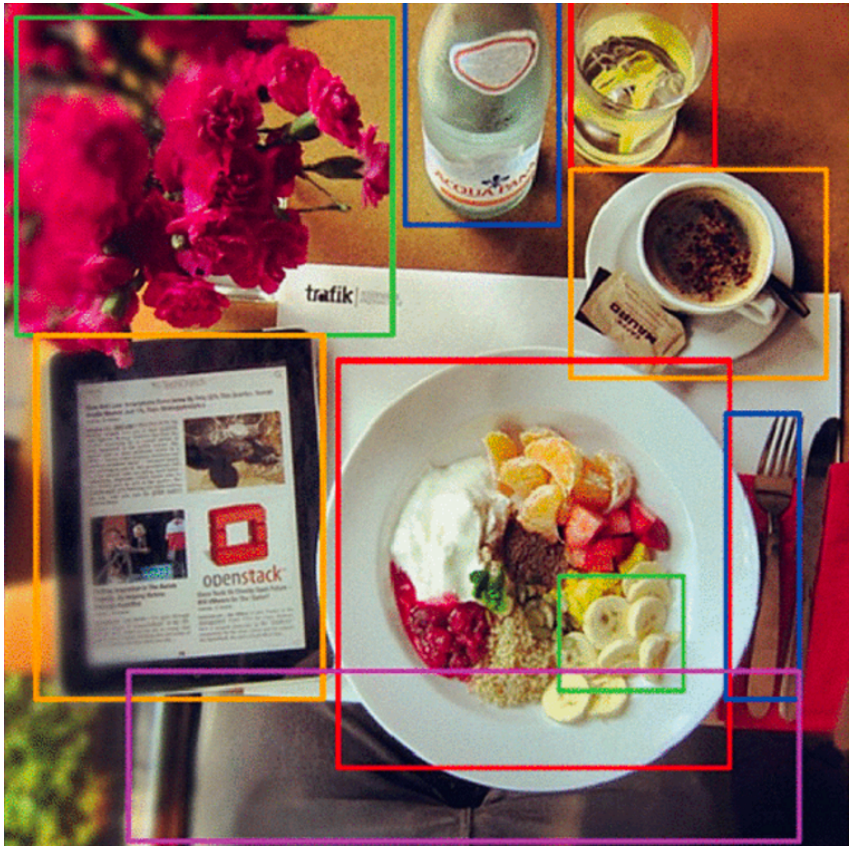
AlphaGo

AlphaGo -> AlphaGo Zero -> AlphaZero



Surpassed Stockfish in four hours, AlphaGo Lee in eight. (Using 5000 1st gen TPUs.)

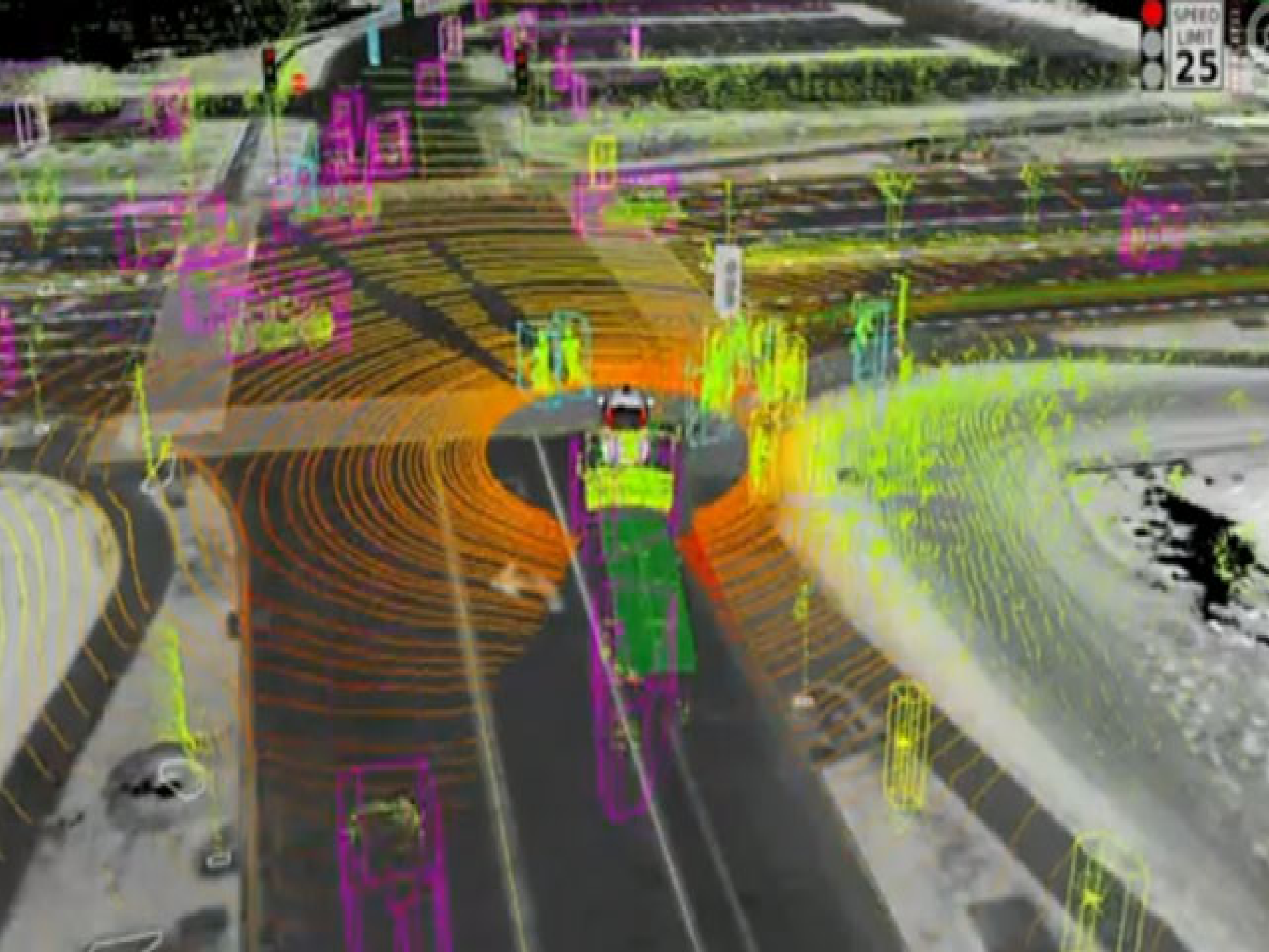
Object Recognition



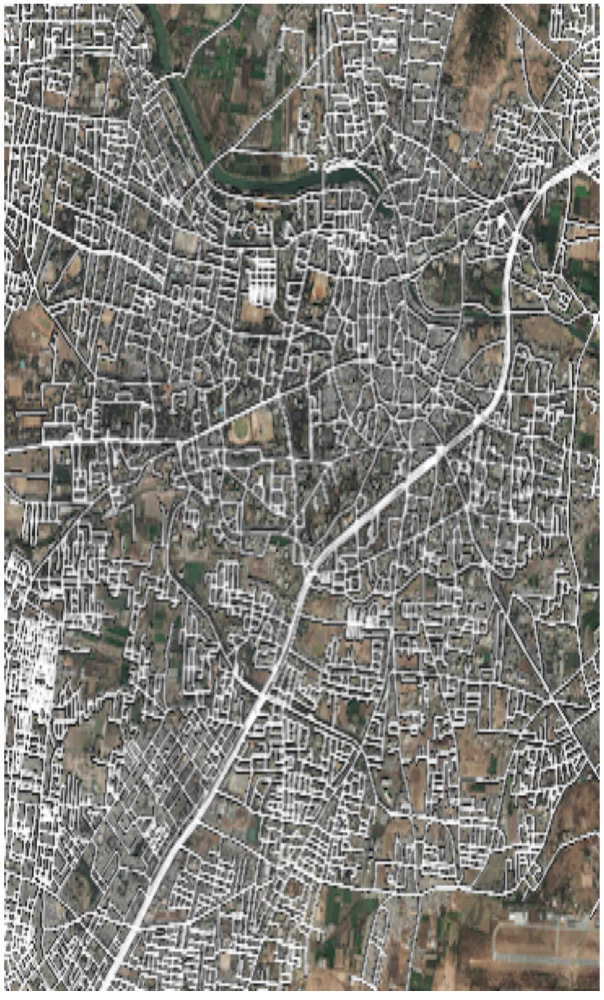
"bouquet of red flowers", "water with ice and lemon", all together: "dining table with breakfast items"

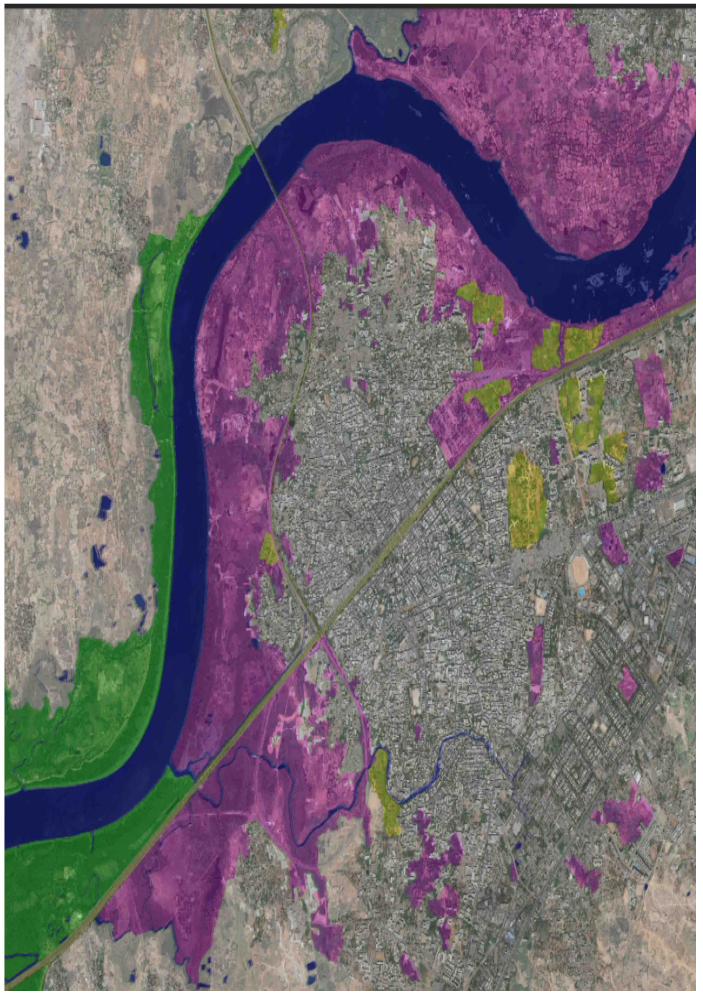
"A group of young people playing a game of frisbee."

"Table of Food"


SPEED
LIMIT
25

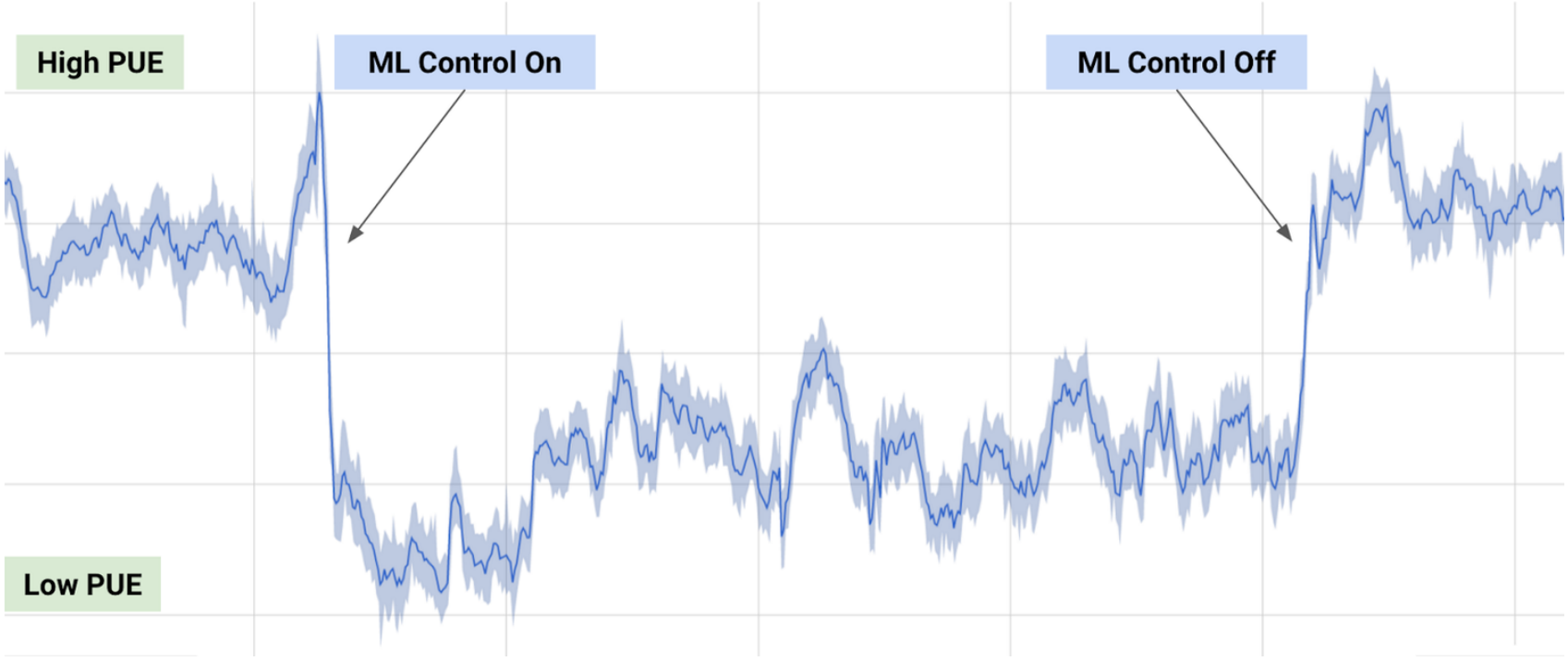

Slip and Fall
Graffiti Detected
Abandoned Object
Facial Recognition in a crowd
Loitering
Counting



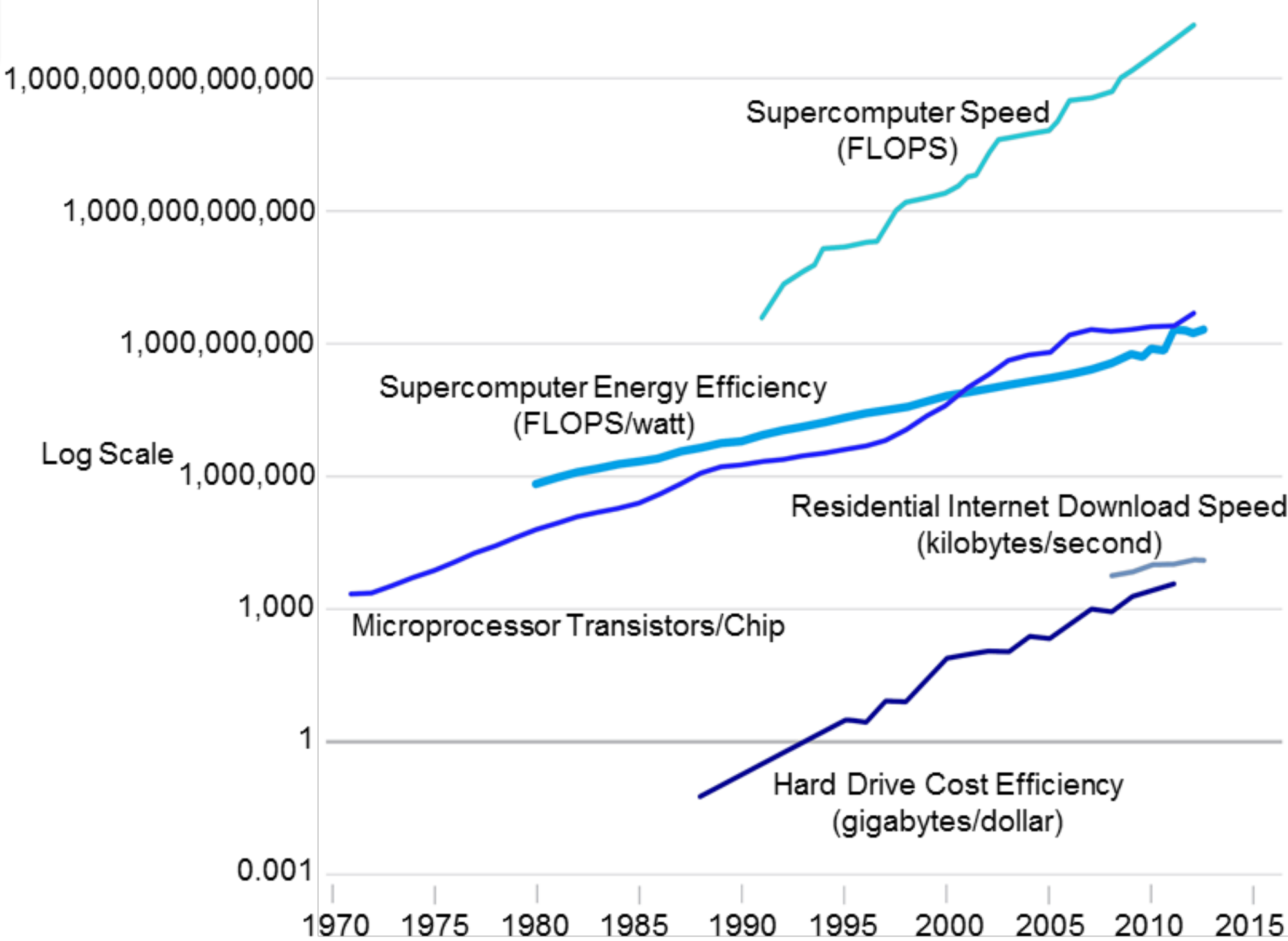


Efficiency


DeepMind AI Reduces Google Data Centre Cooling Bill by 40%

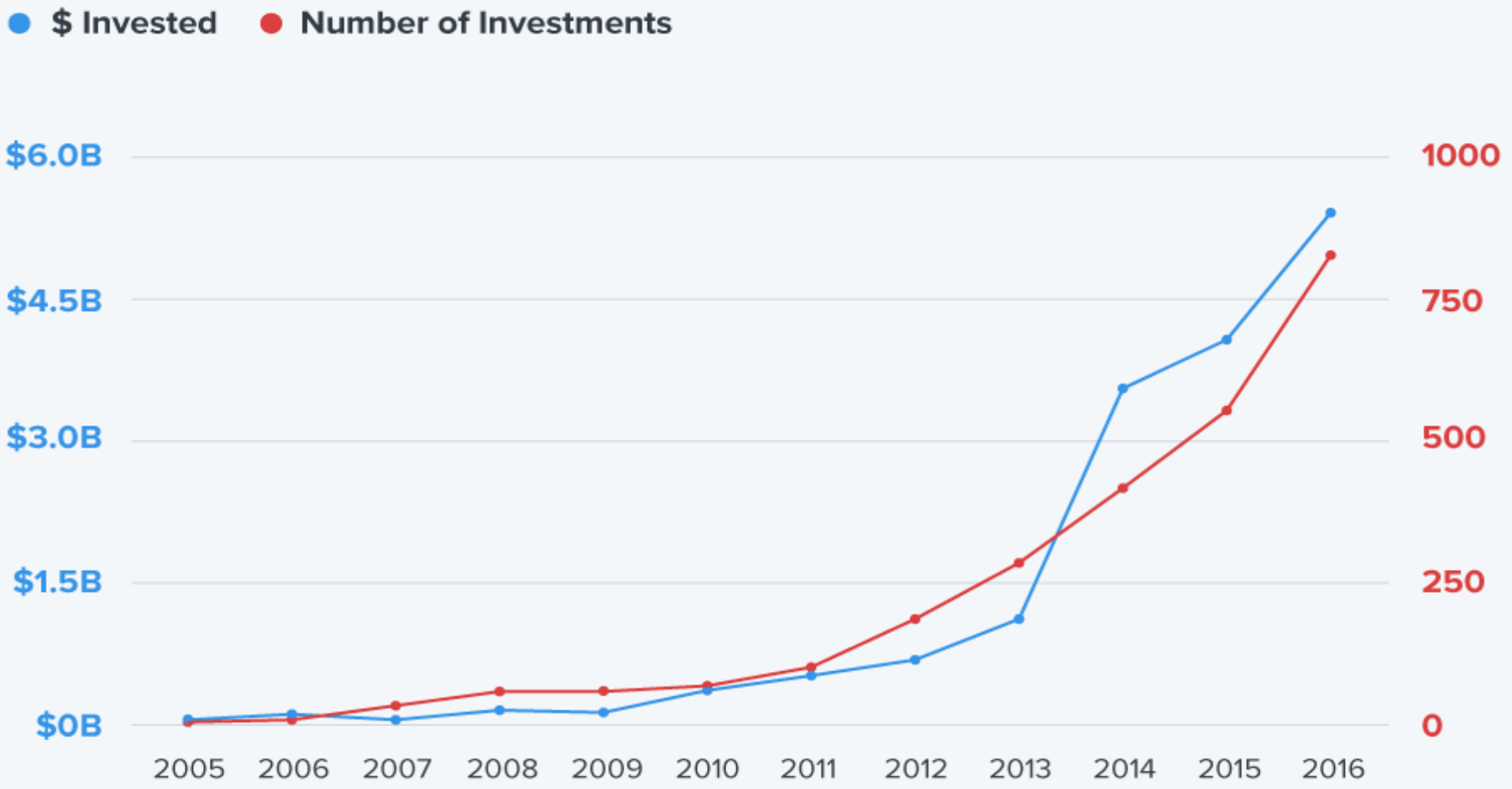

High PUE
ML Control On
ML Control Off
Low PUE


Copyright KMel Robotics
kmelrob

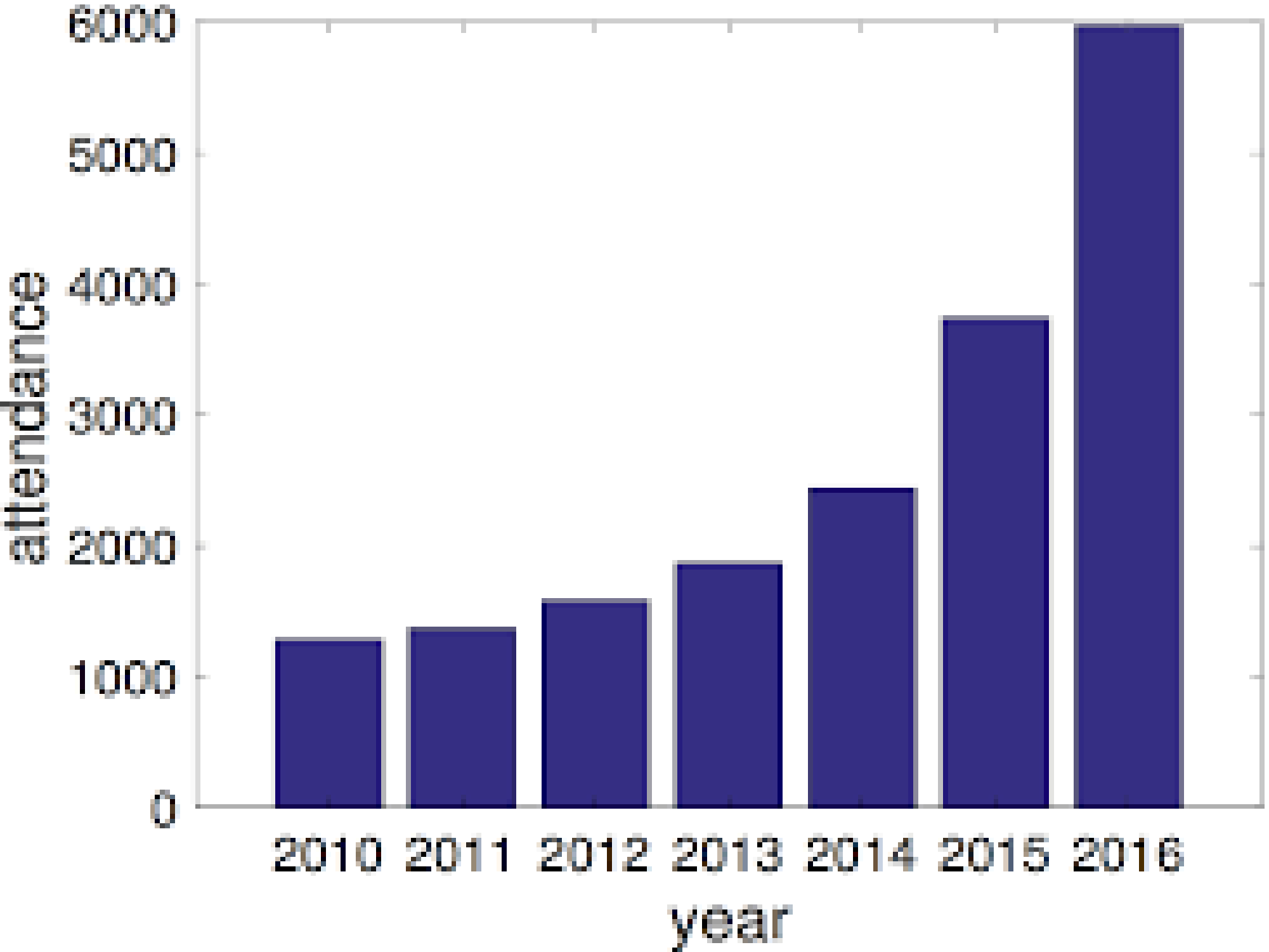

1,000,000,000,000,000
1,000,000,000,000
1,000,000,000
Log Scale
1,000,000
1,000
1
0.001
Supercomputer Speed
(FLOPS)
Supercomputer Energy Efficiency
(FLOPS/watt)
Residential Internet Download Speed
(kilobytes/second)
Microprocessor Transistors/Chip
Hard Drive Cost Efficiency
(gigabytes/dollar)
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015

AI Investments (2005–2016)

Dollars invested and number of investments continue to grow



Source: CB Insights


6000
5000
4000
3000
2000
1000
0
attendance
2010
2011
2012
2013
2014
2015
2016
year

Applications of Narrow AI

- agriculture
- health and medicine; drug discovery
- investment and finance
- logistics
- environment (energy systems, waste, conservation)
- transportation and autonomous vehicles
- manufacturing
- recommendation systems
- personal and professional assistants
- education
- legal services and justice
- social networks
- relationship services
- science and engineering; material design
- policing and security
- military (LAWs, cyber, intel)..

Near-term Governance Challenges

Safety in critical systems, such as finance, energy systems, transportation, robotics, autonomous vehicles.

(Consequential) algorithms that **encode values**, such as in hiring, loans, policing, justice, social network.
Desiderata: fairness ▸ Hardt, accountability, transparency, efficiency, privacy, ethics.

AI impacts on employment, equality, privacy, democracy...

Sufficiency Hypothesis: Even if we stopped fundamental (hardware and algorithmic) progress in AI today, the application of existing AI is sufficient to radically transform wealth, power, and world order.

Some Extreme Challenges from Near-Term AI

- **Mass labor displacement and inequality.** If AI substitutes, rather than complements, labor.
- **AI Oligopolies: strategic industry and trade.** If AI industries are natural global monopolies, due to low/zero marginal costs of AI services, incumbent advantage, high fixed costs from AI R&D.
- **Surveillance and Control**: mass surveillance (sensors, digitally-mediated behavior), intimate profiling, tailored persuasion, repression (LAWs).
- **Strategic (Nuclear) Stability**: autonomous escalation; counterforce vulnerability from AI intel, cyber, drones; autonomous nuclear retaliation (esp w/ hypersonics).
- **Military Advantage**: LAWs, cyber, intel, info operations. Shifts, volatility, common knowledge.
- **Accident/Emergent/Other Risks**, from AI-dependent critical systems and transformative capabilities.

Narrow Transformative Capabilities

Narrow Transformative Capabilities

Most likely where: data rich, can simulate environment, narrow domains, ripe technical problem, fast decisions, many variables, and/or high stakes.

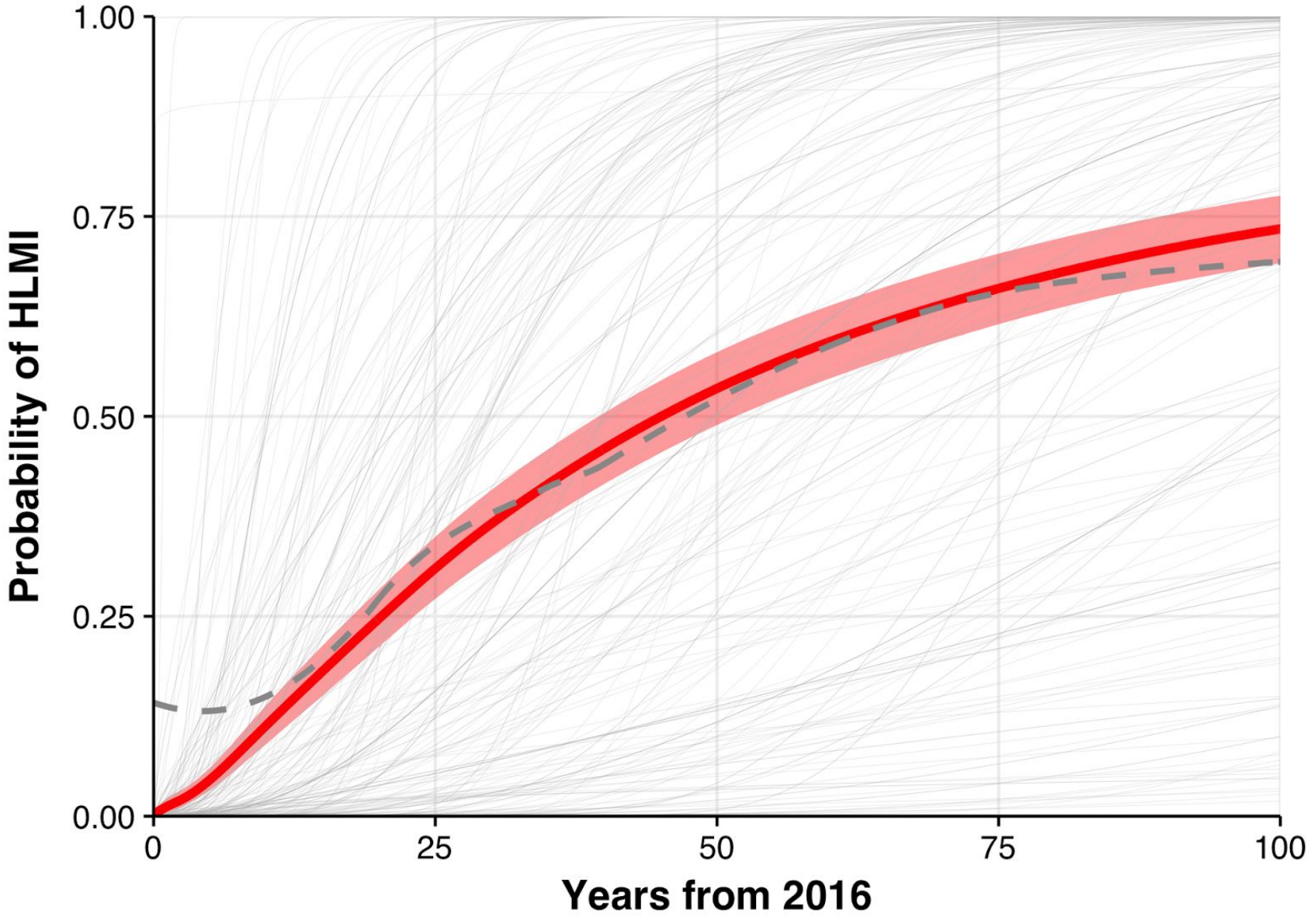
Narrow Transformative Capabilities

Most likely where: data rich, can simulate environment, narrow domains, ripe technical problem, fast decisions, many variables, and/or high stakes.

- Finance. Operations/logistics.
- Engineering, science, math, drug discovery, material science.
- Cyber.
- Surveillance.
- Profiling (lie detection, emotion detection, psychological insight, DNA). Personal assistants/advertising.
- Social network mapping and manipulation.

Survey of NIPS/ICML about HLMI (Grace et al 2017)

Survey of NIPS/ICML about HLMI (Grace et al 2017)



Outstanding AI Challenges / Current Work

Outstanding AI Challenges / Current Work

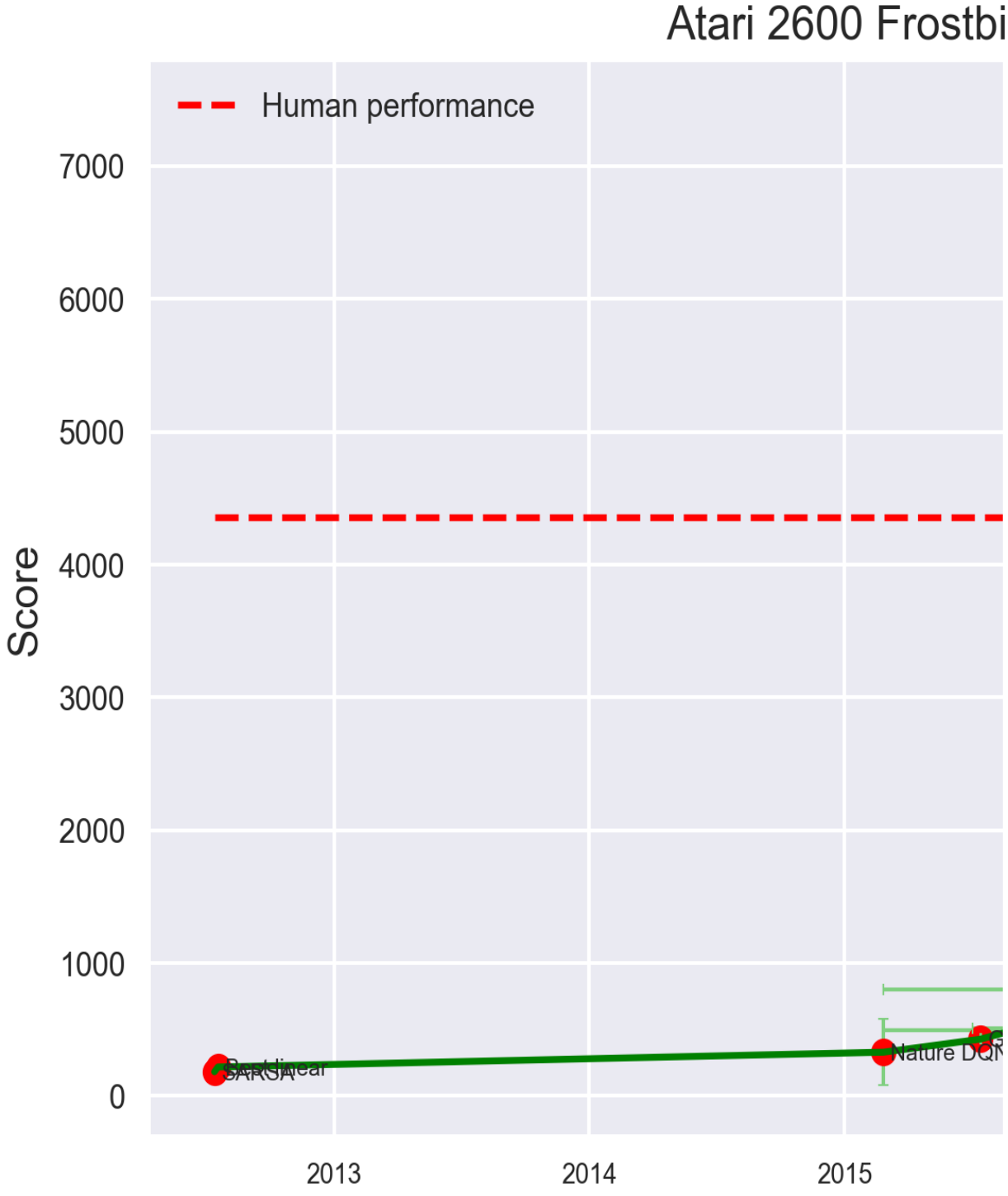
- Imagination-based Planning
- Hierarchical Planning
- Unsupervised Learning
- Memory and One-Shot Learning
- Continual and Transfer Learning
- Language Grounding / Abstract Concepts
- Predictive World Model

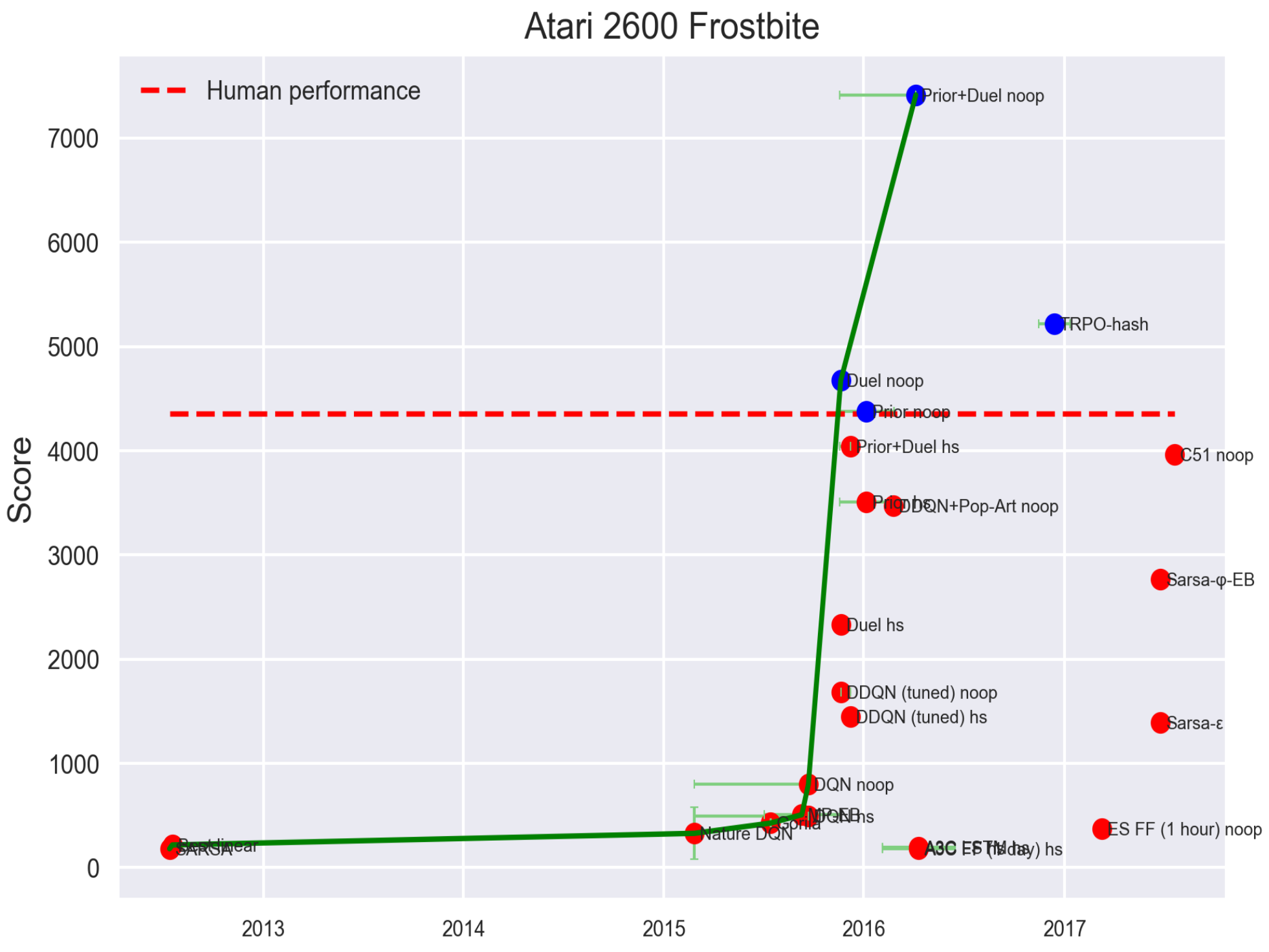
(h/t Hassabis)

Rapid Broad Takeoff? Soon?

Rapid Broad Takeoff? Soon?

- One breakthrough to AGI? (A single missing common factor: common sense/world model/general reasoning?)
- Narrow capabilities unlock cluster of other technologies?
- Highly recursive self-improvement?
- Hardware overhang / insecure compute. Insight overhang. Data/sensor overhang.
- High train to execute costs.


Atari 2600 Frostbite
Human performance
7000
6000
5000
4000
3000
2000
1000
0
Score
2013
2014
2015
Nature DQN


Atari 2600 Frostbite
Human performance
Prior+Duel noop
TRPO-hash
Duel noop
Prior noop
Prior+Duel hs
C51 noop
DDQN+Pop-Art noop
Sarsa-φ-EB
Duel hs
DDQN (tuned) noop
DDQN (tuned) hs
Sarsa-ε
DQN noop
Nature DQN
ES FF (1 hour) noop
Score
0
1000
2000
3000
4000
5000
6000
7000
2013
2014
2015
2016
2017

Accident Risks from Advanced AI

Accident Risks from Advanced AI

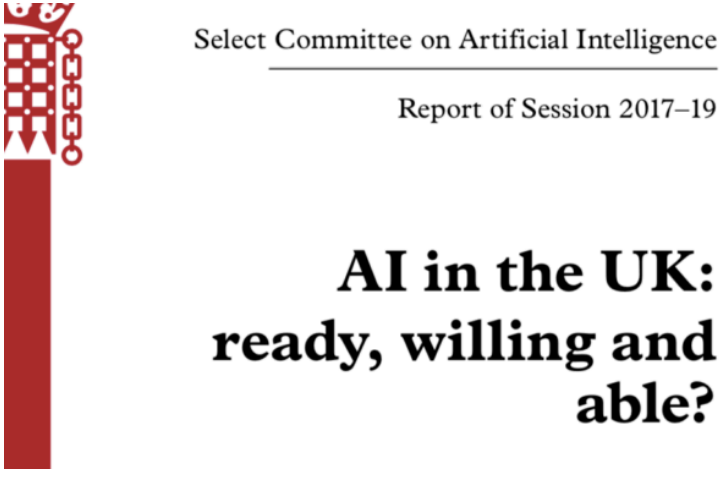
One can imagine such technology outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand. Whereas the short-term impact of AI depends on who controls it, the long-term impact depends on whether it can be controlled at all...

-Stephen Hawking, Stuart Russell, Max Tegmark, Frank Wilczek, 2014

Canada's AI Grand Strategy

- (1) Leading in AI capacity and governance.
- (2) Prepare to co-lead AI-for-common-good coalition.

National Strategies

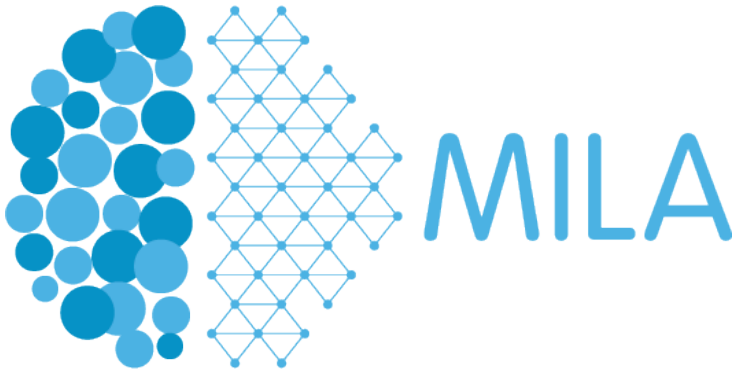


Canada's Advantages

CIFAR

Pan-Canadian Artificial Intelligence Strategy Overview

≡ ELEMENT AI



· Strong AI community; scientific and economic assets.
· National cohesion, identity, and trust.
· Cosmopolitan and non-militaristic values, foreign policy, research culture. Eg Hinton and Sutton both came to Canada in part for this. (cf Google and Project Maven).

The AI Race

The AI Race

The coordination problem is one thing [we should focus on now]. We want to avoid this harmful race to the finish where corner-cutting starts happening and safety gets cut.... That's going to be a big issue on a global scale, and that's going to be a hard problem when you're talking about ***national governments***..

Demis Hassabis, CEO of DeepMind, January 2017


"
Whoever leads in AI will rule the world
Vladimir Putin
"
RT

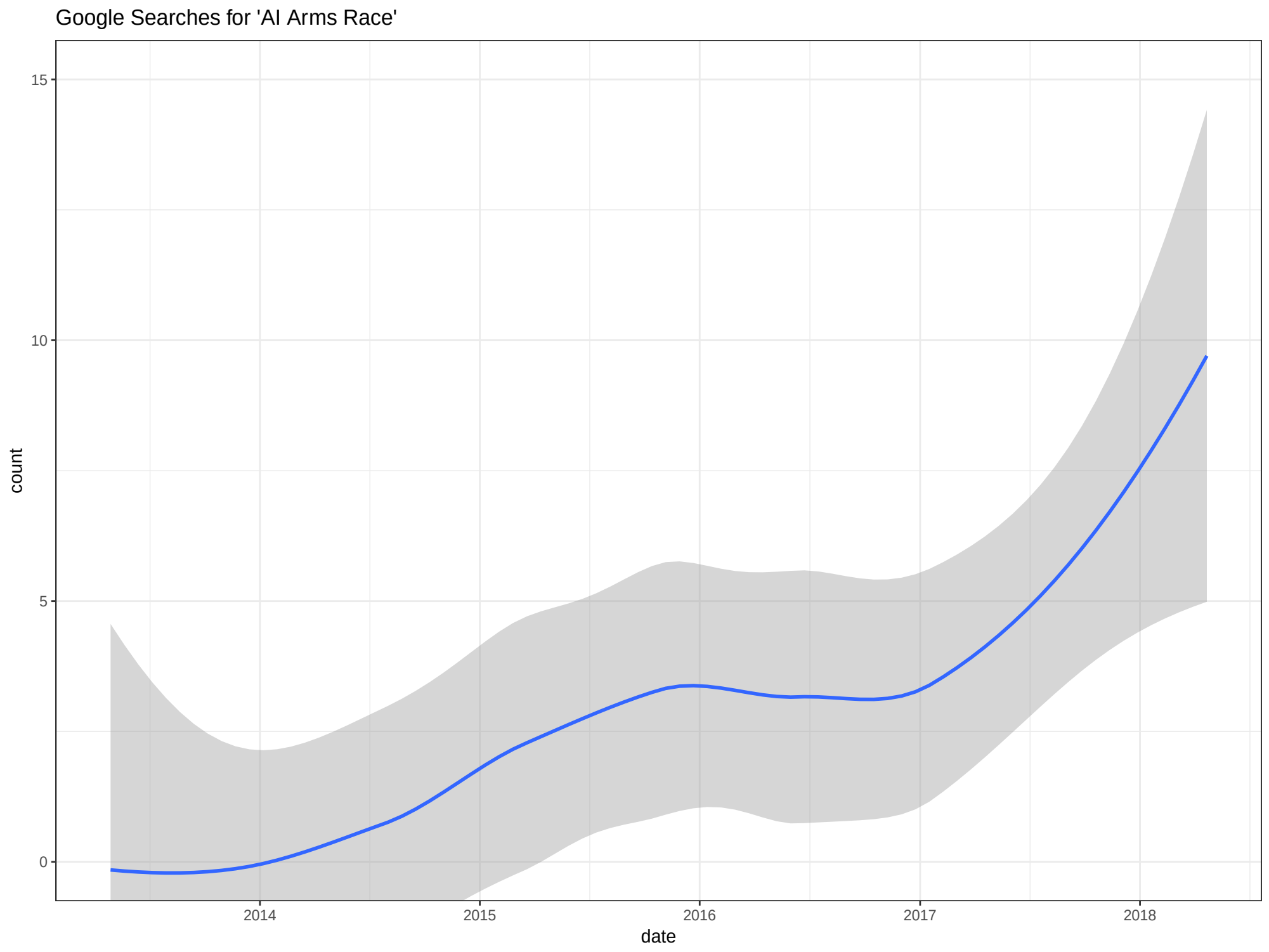
TECHNOLOGY NEWS MARCH 26, 2018 / 8:16 PM / A MONTH AGO

With new plan, Macron wants France to win AI 'arms race'

Michel Rose, Mathieu Rosemain

3 MIN READ

PARIS (Reuters) - French President Emmanuel Macron has set his sights on artificial intelligence as the next technological frontier France cannot afford to miss, and will launch a major “offensive” this week, officials said on Monday.


Google Searches for 'AI Arms Race'
15
10
5
0
count
2014
2015
2016
2017
2018
date

Risks from a Great Power AI Race

- "Corner cutting" on powerful systems, extreme accident risks.
- Breakdown in liberal economic world order from rivalrous strategic industries and securitization of AI.
- Militarization of AI science.
- Strategic instability.
- International hostility.

AI-for-Common-Good Coalition

Canada could co-lead a coalition building **AI for the common good**, an alternative to the US-China AI race. Could channel world resources and legitimacy to non-rivalrous AI project, elevate best practice, ethics, safety, values.

- Explicit commitment to principles: non-threatening, cosmopolitan. Common good.
- Transparency: citizens and allies can ensure compliance with principles.
- Accountability: members have political control over program.
- Openness to inclusion on the basis of rules.

What should Canada's national strategy be for the AI revolution?

Governance: Asilomar 2017 Principles for Beneficial AI

1) **Research Goal**: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.
18) **AI Arms Race**: An arms race in lethal autonomous weapons should be avoided.
19) **Capability Caution**: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
20) **Importance**: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.
21) **Risks**: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.
22) **Recursive Self-Improvement**: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.
23) **Common Good**: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.