The Power and Limits of Machine Learning

Dit-Yan Yeung

2020-10-14

Agenda

1 What is Machine Learning?

2 The Power of Machine Learning

3 The Limits of Machine Learning

4 The Journey Ahead

What is Machine Learning?

Artificial Intelligence

Artificial intelligence (AI) enables machines to perform some cognitive functions similar to those attributed to humans,

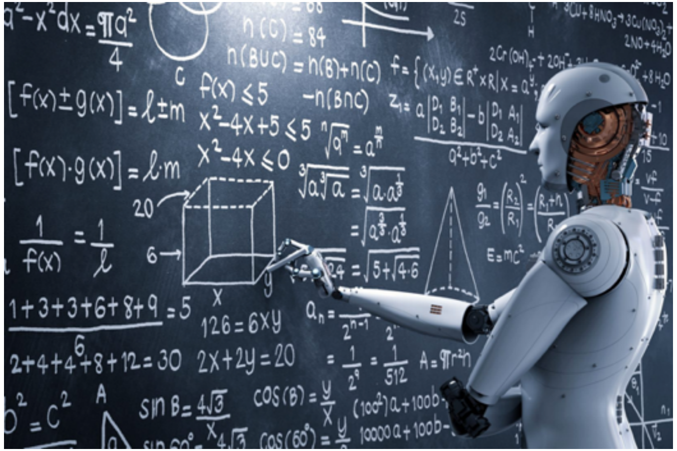
as opposed to conventional machines which act according to how they are programmed to act.

“AI is the new electricity.” – Andrew Ng

History of AI

AI is as old as the field of computer science (CS).

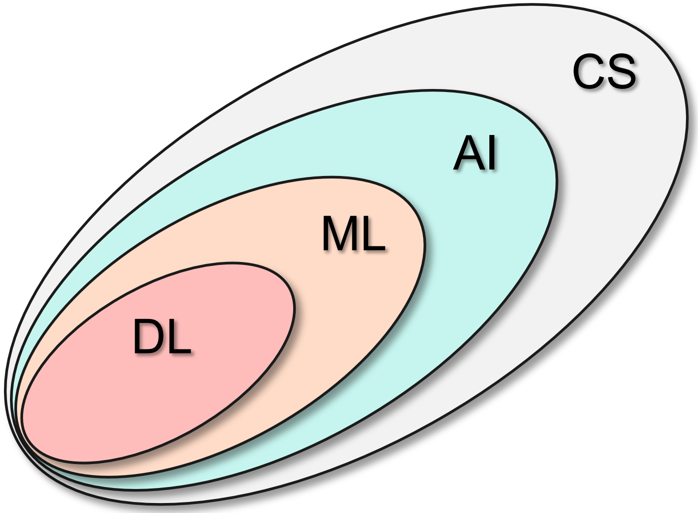
Many pioneers in CS are also pioneers in AI, e.g., Alan Turing, John McCarthy, Herbert Simon, Marvin Minsky.



AI, Machine Learning, and Deep Learning

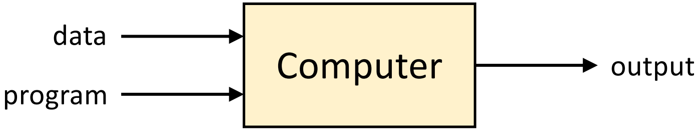
Machine learning (ML) marries algorithms in CS, mathematical and statistical modeling, and learning from data/examples.

Deep learning (DL) is a subarea of ML – representation learning often using relatively deep, layered network architectures.

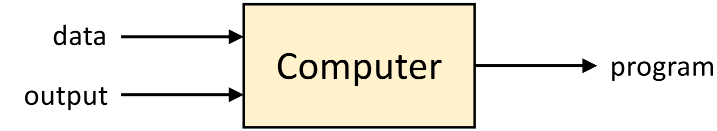


Conventional Programming vs. Supervised Learning

Conventional programming



Supervised learning

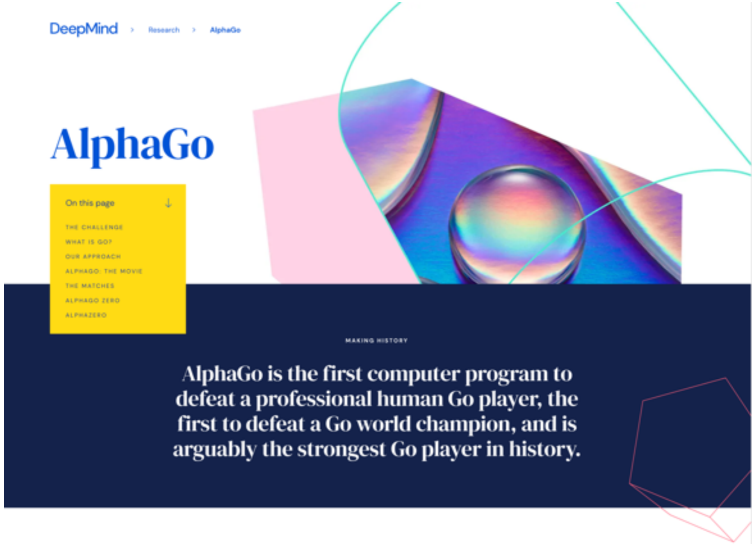


Other learning paradigms:

- Unsupervised learning
- Reinforcement learning
- Semi-supervised learning
- …

The Power of Machine Learning

AlphaGo



DeepMind › Research › AlphaGo

AlphaGo

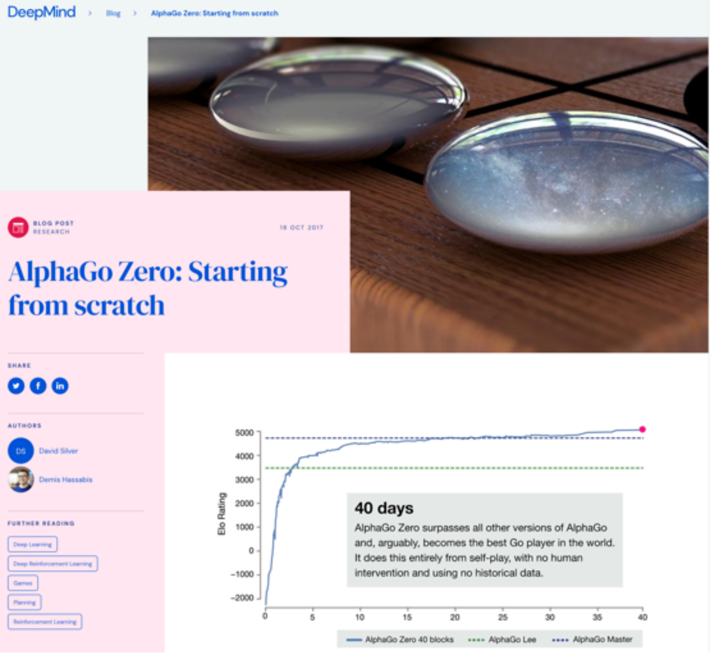
On this page ↓

THE CHALLENGE
WHAT IS GO?
OUR APPROACH
ALPHAGO: THE MOVIE
THE MATCHES
ALPHAGO ZERO
ALPHAZERO

MAKING HISTORY

AlphaGo is the first computer program to defeat a professional human Go player, the first to defeat a Go world champion, and is arguably the strongest Go player in history.

From AlphaGo to AlphaGo Zero and AlphaZero



DeepMind › Blog › AlphaGo Zero: Starting from scratch

BLOG POST
RESEARCH

18 OCT 2017

AlphaGo Zero: Starting from scratch

SHARE

AUTHORS

DS David Silver

Demis Hassabis

FURTHER READING

Deep Learning

Deep Reinforcement Learning

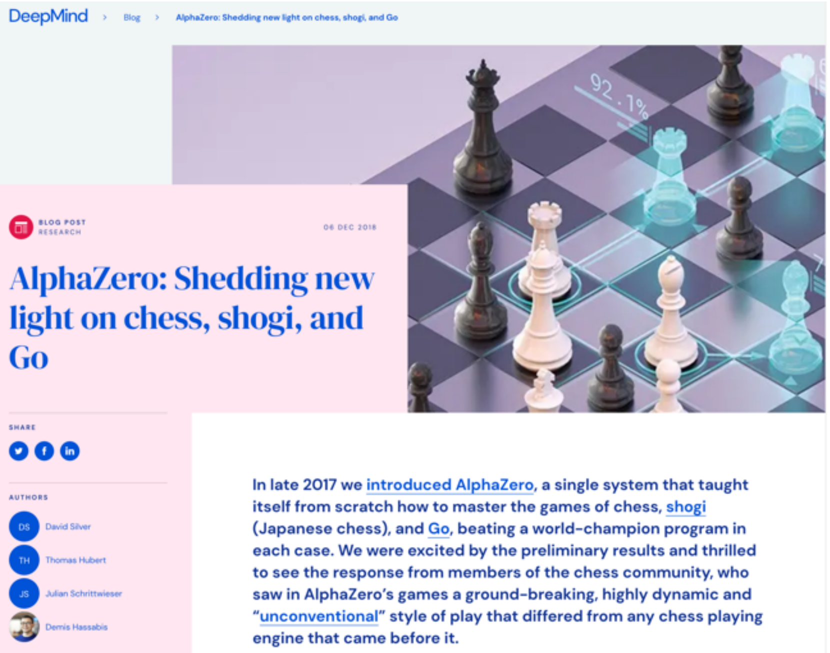
Games

Planning

Reinforcement Learning

40 days

AlphaGo Zero surpasses all other versions of AlphaGo and, arguably, becomes the best Go player in the world. It does this entirely from self-play, with no human intervention and using no historical data.



DeepMind › Blog › AlphaZero: Shedding new light on chess, shogi, and Go

BLOG POST
RESEARCH

06 DEC 2018

AlphaZero: Shedding new light on chess, shogi, and Go

SHARE

AUTHORS

DS David Silver

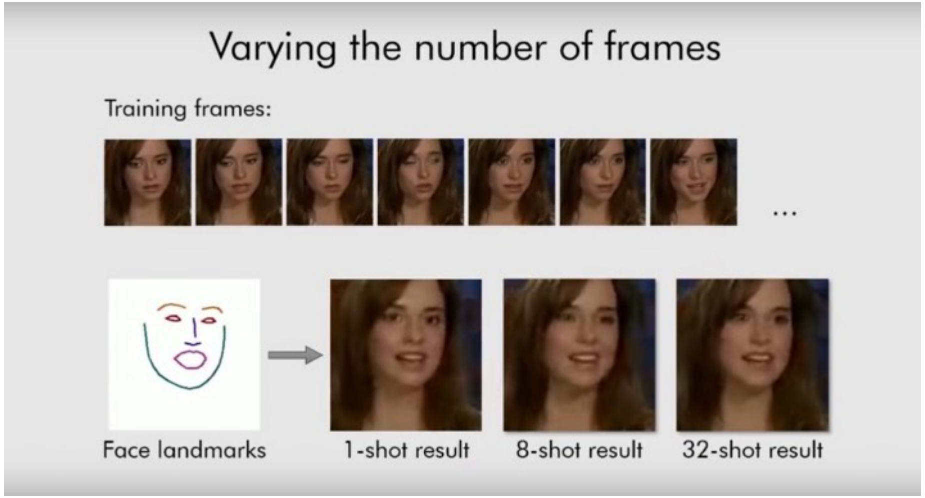
TH Thomas Hubert

JS Julian Schrittwieser

Demis Hassabis

In late 2017 we introduced AlphaZero, a single system that taught itself from scratch how to master the games of chess, shogi (Japanese chess), and Go, beating a world-champion program in each case. We were excited by the preliminary results and thrilled to see the response from members of the chess community, who saw in AlphaZero's games a ground-breaking, highly dynamic and "unconventional" style of play that differed from any chess playing engine that came before it.

Talking Heads with Motion



YouTube video

Zakharov et al., “Few-shot adversarial learning of realistic neural talking head models”, ICCV, 2019.

Generating Photorealistic Deepfakes

YouTube video

Karras et al., “Analyzing and improving the image quality of StyleGAN”, CVPR, 2020.

AI Assistant (Google Duplex)

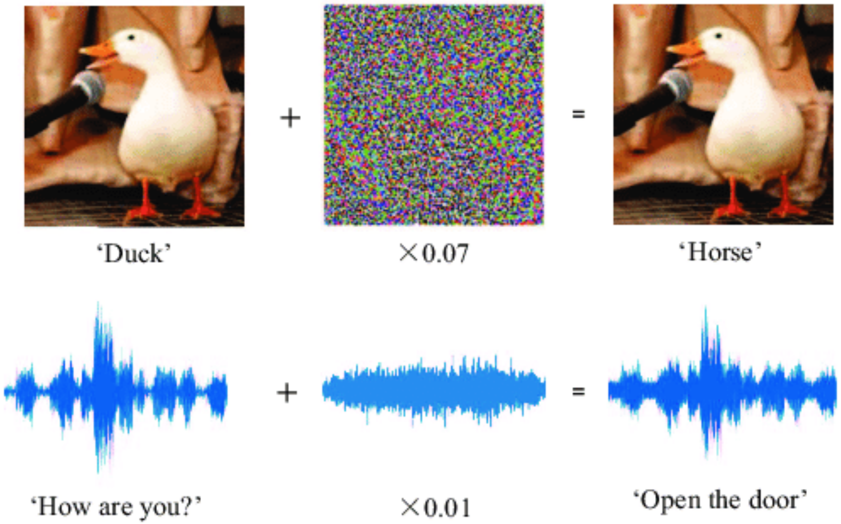
“Hi, I'm calling to book a women's haircut for a client.”

YouTube video

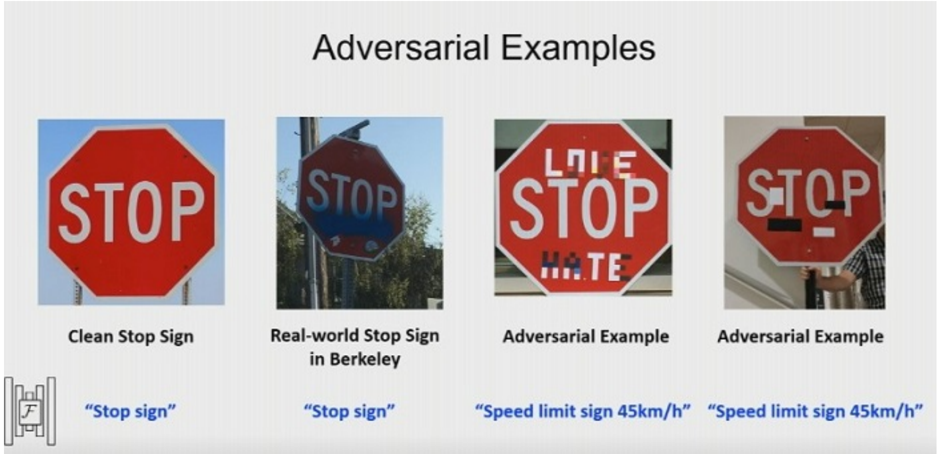
The Limits of Machine Learning

Adversarial Examples (or Adversarial Attacks)

Imperceptible perturbations (which are carefully generated) added to images or audio signals can fool even state-of-the-art classifiers to give incorrect predictions.



Safety Concerns of Adversarial Examples

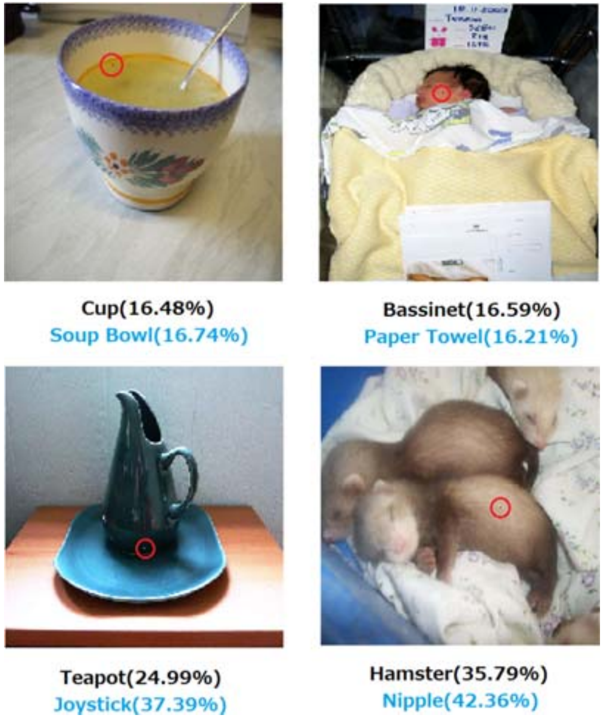


Physical attacks on STOP sign

One-pixel Attacks

Changing just one pixel
(marked by red circle)

Su et al., “One pixel attack for fooling deep neural network”, IEEE T-EC, 2019.



Cup(16.48%)
Soup Bowl(16.74%)

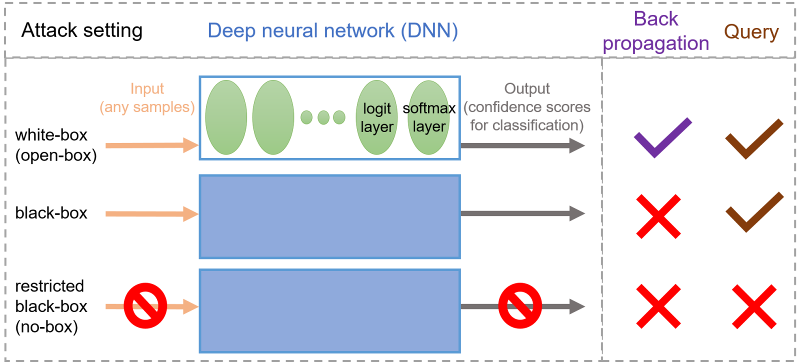
Bassinet(16.59%)
Paper Towel(16.21%)

Teapot(24.99%)
Joystick(37.39%)

Hamster(35.79%)
Nipple(42.36%)

White-box vs. black-box attacks

(Chen et al., AISec 2017)



Attack setting	Deep neural network (DNN)	Back propagation	Query
white-box (open-box)		✓	✓
black-box		✗	✓
restricted black-box (no-box)		✗	✗

White-box attacks

- Have full knowledge of internal structure of target model when generating adversarial attacks
- Worst-case scenario

Black-box attacks

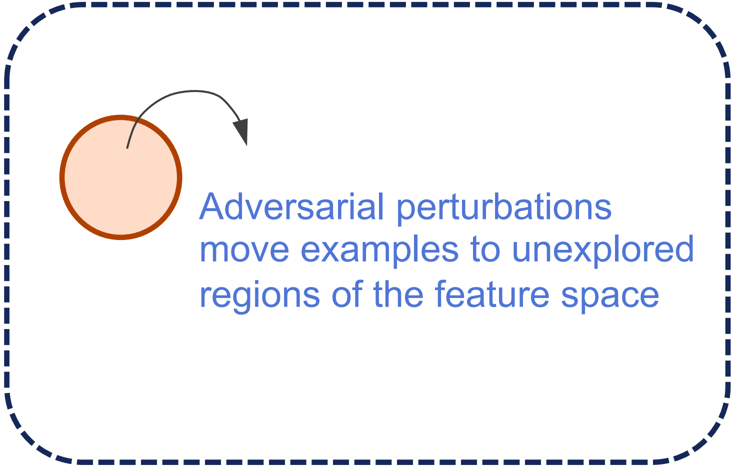
- Have no knowledge of internal structure of target model when generating adversarial attacks
- More realistic scenario

How are Adversarial Examples Generated?

Three major approaches:

1. Optimization-based approach, e.g., $\boxed{\begin{aligned}&\min_{x'}\ c\|\eta\| + J_\theta(x', l')\\ &\text{s.t. } x' \in [0, 1].\end{aligned}}$
2. Gradient-based approach, e.g., $\boxed{\eta = \epsilon \text{sign}(\nabla_x J_\theta(x, l))}$
3. Generative approach, e.g., using a generative model

What Makes Adversarial Attacks Possible?


Adversarial perturbations move examples to unexplored regions of the feature space

Feature space

Theoretical study of underlying reasons for adversarial attacks is still rare and immature – good research topic to work on.

Defenses Against Adversarial Attacks

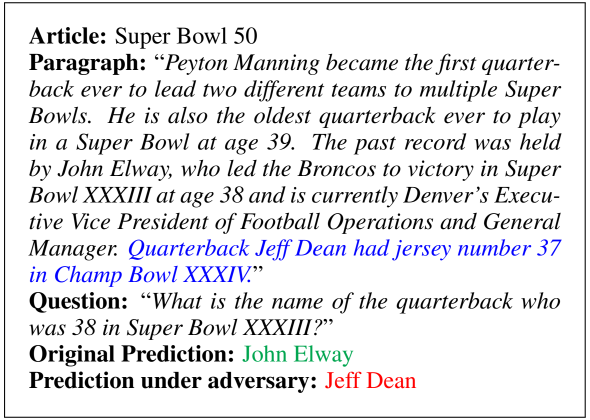
Two major approaches:

1. Retraining the model, e.g., adversarial training, defensive distillation

2. Learning to purify the adversarial examples before feeding them into the model, e.g., MagNet, PixelDefend

Adversarial Attacks Beyond Images and Audio Signals

Attacking reading comprehension systems



Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

(Jia and Liang, EMNLP 2017)

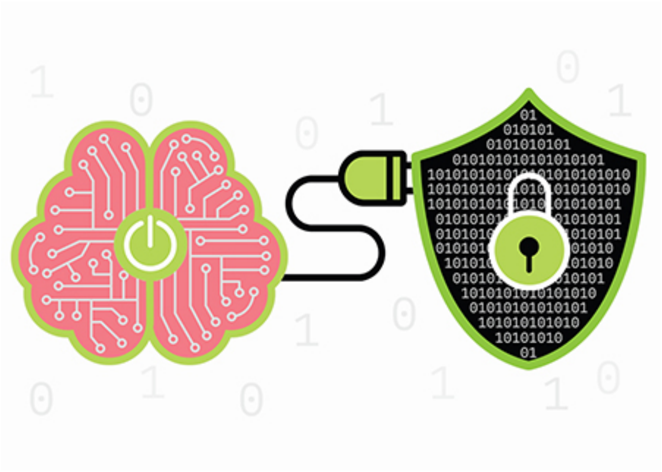
Other adversarial attacks:

- Machine translation
- Text summarization
- Malware detection
- Spam detection
- Reinforcement learning
- ...

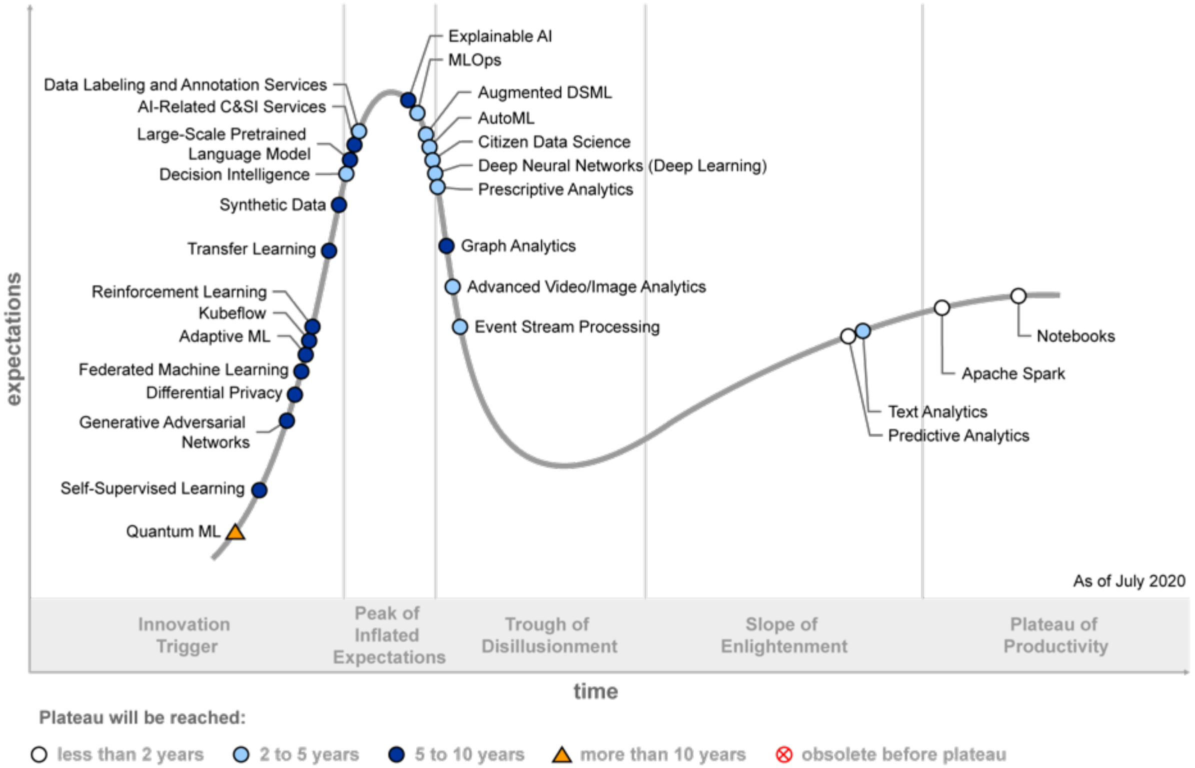
The Journey Ahead

Interplay between ML and Cybersecurity

The study of security, privacy, robustness, resilience, and reliability will be central to the field of machine learning



Hype Cycle for Data Science and Machine Learning, 2020



Source: Gartner
ID: 450404

AI Ethics

executive
officer
purchased
used
within
trade-offs
relatable
organizations
meaningfully
justify
purely
performance
external
feedback
procurement
affected
concerns
seen
tools
able
business
many
evaluate
compliance
like
domain
mechanisms
products
customer
ux
harm
teams
clarity
leadership
impact
uphold
delivering
don't
legal
getting
ethics
processes
time
sharing
better
believe
responsible
committee
demonstrable
new
isn't
view
ai
may
organization
training
team
terms
startup
without
decisions
insight
users
know
concerned
don't
companies
review
data
user
understand
scientists
comprehensive
chief
power
communities
push
clients
engage
principles
applies
practices
vendors
product
revenue
examples
objective
projects
experience
advocacy
shareholders
veto
unbiased
commitments
stakeholders
design
mitigate
developers
consumers
designers

(from Partnership on AI)

Q&A