MSCI

THEMATIC INSIGHTS

The new computational revolution

Disruptive innovation and the next generation internet

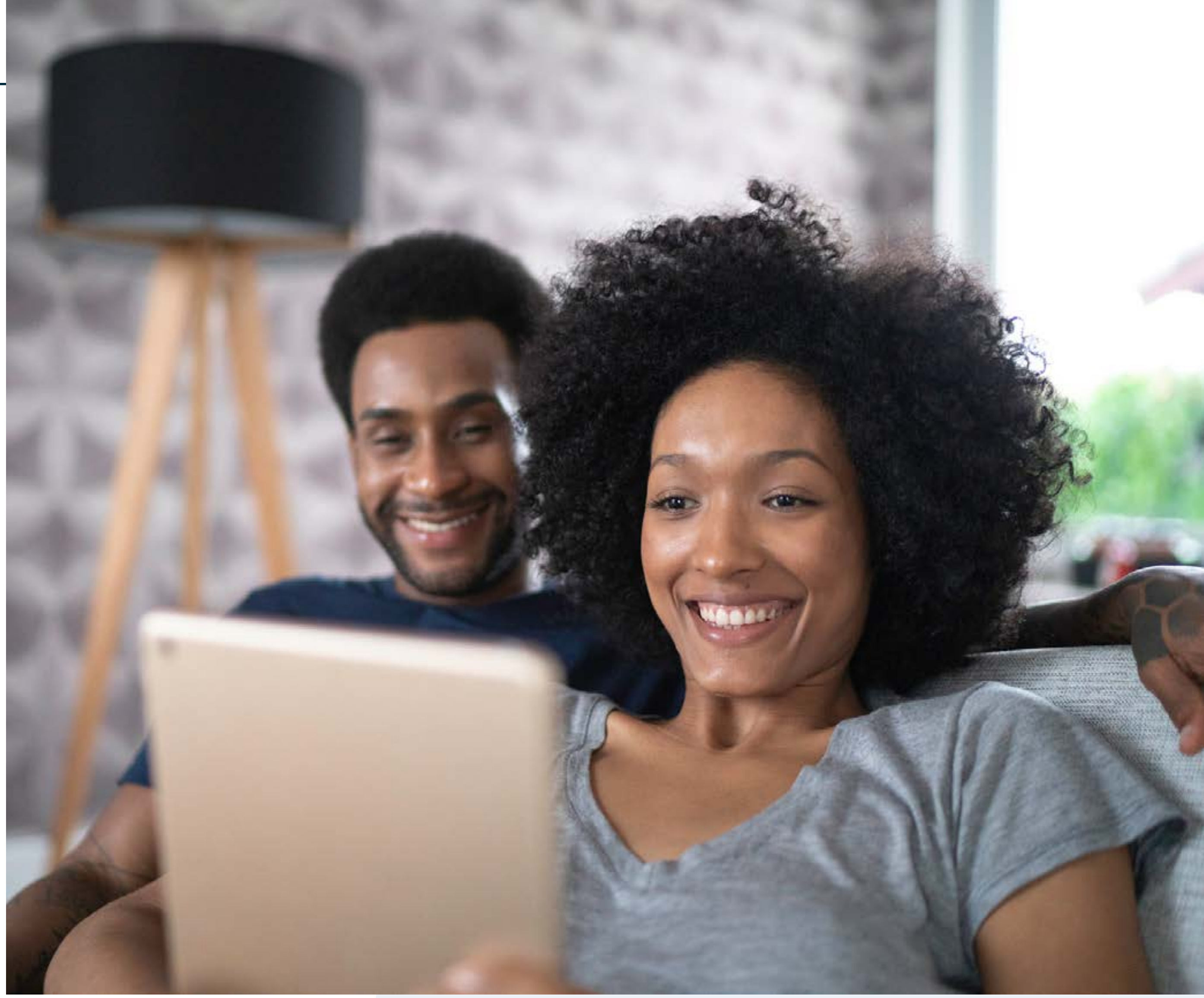
msci.com

Contents


09:41
smart home
OFFICE
lights
electricity
speakers
Wifi
settings

The platform transition in computation

Computers were once servers: affordable only for the enterprise. Then they became personal, small enough to sit easily on a desktop. The profit that had once accrued to the hardware manufacturers shifted to software and at first, the architects of the operating system. Next, computers became less economically important than their interconnectedness — they became portals to the world and shared media — and new, huge enterprises effectively mapped operating systems on top of that network (search engines and social graphs) and built portals that could fit in a pocket or sit on the wrist.[1] After 40 years, we arrive at present day: cloud computing means computers are again servers, affordable and accessible at tremendous scale, intermediated by the phone in your pocket, the watch on your wrist or a tablet consuming content.



Accessing up a supercomputer in the cloud for web services may have become almost as easy as hailing a cab and a highly competitive market for cloud services is well established.[2] However, in this report we examine the idea that another transition has begun, which leverages this cloud technology, and may prove as meaningful as the transition from servers to desktop PCs. Moreover, outside of technology companies, this platform transition may also have the potential to impact every sector and company just as the build-out of the internet (mobile and static) has done.

1 https://en.wikipedia.org/wiki/Client%E2%80%93server_model#Centralized_computing describes this journey; depictions in popular culture include, for example, the series "Halt and Catch Fire".

2 In 2019, it was forecast to reach over USD 600 billion by 2023 https://www.prnewswire.com/news-releases/cloud-computing-market-worth-623-3-billion-by-2023--exclusive-report-by-marketsandmarkets-300802108.html

The age of the learning computer

Deep neural networks provide a new mechanism by which diverse problems can be solved and are the core catalyst for thinking about an age of the learning computer. Historically, society has relied upon software engineers to construct meticulous computational recipes to reliably and quickly convert inputs into appropriate outputs. The idea is that we can now create a new sort of custom and more flexible computational system — based on deep neural networks — by training it smartly with vast quantities of data.

Parallel processing and millions of (often tagged) examples facilitate the creation of a computational system tailored to the particular type of problem that we are trying to solve. That system—or neural model—can be run quickly using relatively cheap chips, and when well-trained, it can help us address problems that seemed intractable within our old computational framework.

This new breed of computational systems have been reported to identify objects in photos with greater accuracy than humans,[3] to answer questions about written paragraphs more accurately than SAT-prepping high schoolers,[4] are able to stitch movie star faces onto body double actors inexpensively and at scale,[5] to create novel faces and stock photography libraries composed of people that never existed,[6] and best the world's best videogame players.[7] They are the potential speed-up mechanism for the entire software-computational system.

Deep learning models (then termed "artificial neural networks" as they drew inspiration from neuroscience) saw their first peak from the start of the 1980s and were deployed against hand-writing recognition for checks by banks in the 1990s.[8] But neural networks were not recognized as a universal tool for solving complex pattern recognition problems until around the start of the 2010s.[9] This second wave began with breakthroughs in the ability to efficiently train multi-layer models efficiently and was coupled with increasing training dataset sizes, increasing model size, increasing computational power, faster networks, and advances in distributed computing (using graphic processing units or GPUs).[10] In 2014, a team of Google researchers won an international computer vision challenge (ImageNet) by using parallel processing to train an unprecedentedly deep neural network.[11] Their success demonstrated that deep neural networks could solve large real-world problems if the training set was sufficiently large. Moreover, building on work from 2009, they also showed — and for commercial applications, this was key — that a deep neural network could also be trained in a reasonable amount of time.[12]

3 http://arxiv.org/pdf/1502.01852

4 https://docs.microsoft.com/en-us/archive/blogs/stevengu/microsoft-achieves-human-performance-estimate-on-glue-benchmark

5 https://www.ft.com/content/9df280dc-e9dd-11e9-a240-3b065ef5fc55

6 https://www.fastcompany.com/90406423/these-ai-generated-people-are-coming-to-kill-stock-photography

7 https://openai.com/projects/five/

8 "Deep Learning", I. Goodfellow, Y. Bengio and A. Courville, MIT Press, 2016. The 1980s saw huge activity with multi-layer perceptron models and the development of key concepts like internal representations as well as implementation technologies like backward propagation - which is now the dominant approach to training deep learning models

9 https://www.nature.com/articles/nature14539

10 "Deep Learning", I. Goodfellow, Y. Bengio and A. Courville, MIT Press, 2016.

11 Deep, here, refers to the number of layers in the network. https://arxiv.org/abs/1409.4842

12 http://www.robotics.stanford.edu/~ang/papers/icml09-LargeScaleUnsupervisedDeepLearningGPU.pdf

The main idea to achieve this fast training was to use a network of graphic processors (GPUs), which in contrast to traditional computer CPUs, have been optimized to run lots of math operations simultaneously.

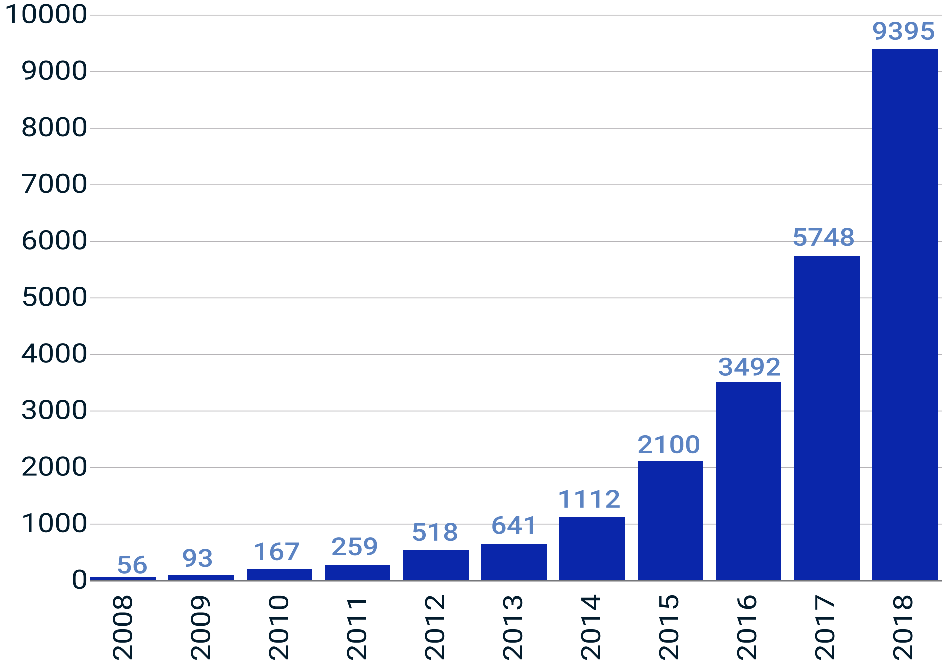
The realization that deep neural networks could prove useful in complex domains and could be cheaply experimented upon and tuned with off-the-shelf computational hardware was a catalyst for a boom of research and huge technical advances.[13] Records fell in virtually every pattern recognition domain. By "inverting" neural networks, researchers demonstrated how they could be used to create — rather than just interpret — digital assets.[14] Incremental reward architectures were derived that enabled neural networks to become world-class in complex puzzles and games.[15] This is the area of so-called reinforcement learning. Neural networks could begin to improve themselves, labeling their own data, playing each other in games, and modifying their own numerical architectures.[16] The past few years have continued these dramatic advances in language processing and generation[17], video and audio generation[18], robotics and automation capability, to name just a few.

Use of these deep learnings models by consumers has become routine. If you have searched on google, you have used a neural network. The video clip you just watched on TikTok was selected by a neural network, as was that next thing you're being told to buy on Amazon. The ability of a car like a Tesla to automatically change lanes depends upon a neural network. And you can call your smartwatch to attention with a single word or make that speaker to play your favorite summer track with a short phrase thanks to a neural network.

A transition to the age of the learning computer has profound implications for how most computation is performed, how we interface with that computation and the scale of the services and entertainment that our technology environment can facilitate.

Exhibit 1:
Deep learning papers published on ArXiv (2008-2018)

Source: https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf



13 https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf

14 https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

15 https://www.nature.com/articles/nature16961

16 See for example https://engineering.fb.com/ml-applications/advancing-state-of-the-art-image-recognition-with-deep-learning-on-hashtags/, https://www.nature.com/articles/nature24270 and https://arxiv.org/abs/1611.01578

17 https://openai.com/blog/better-language-models/

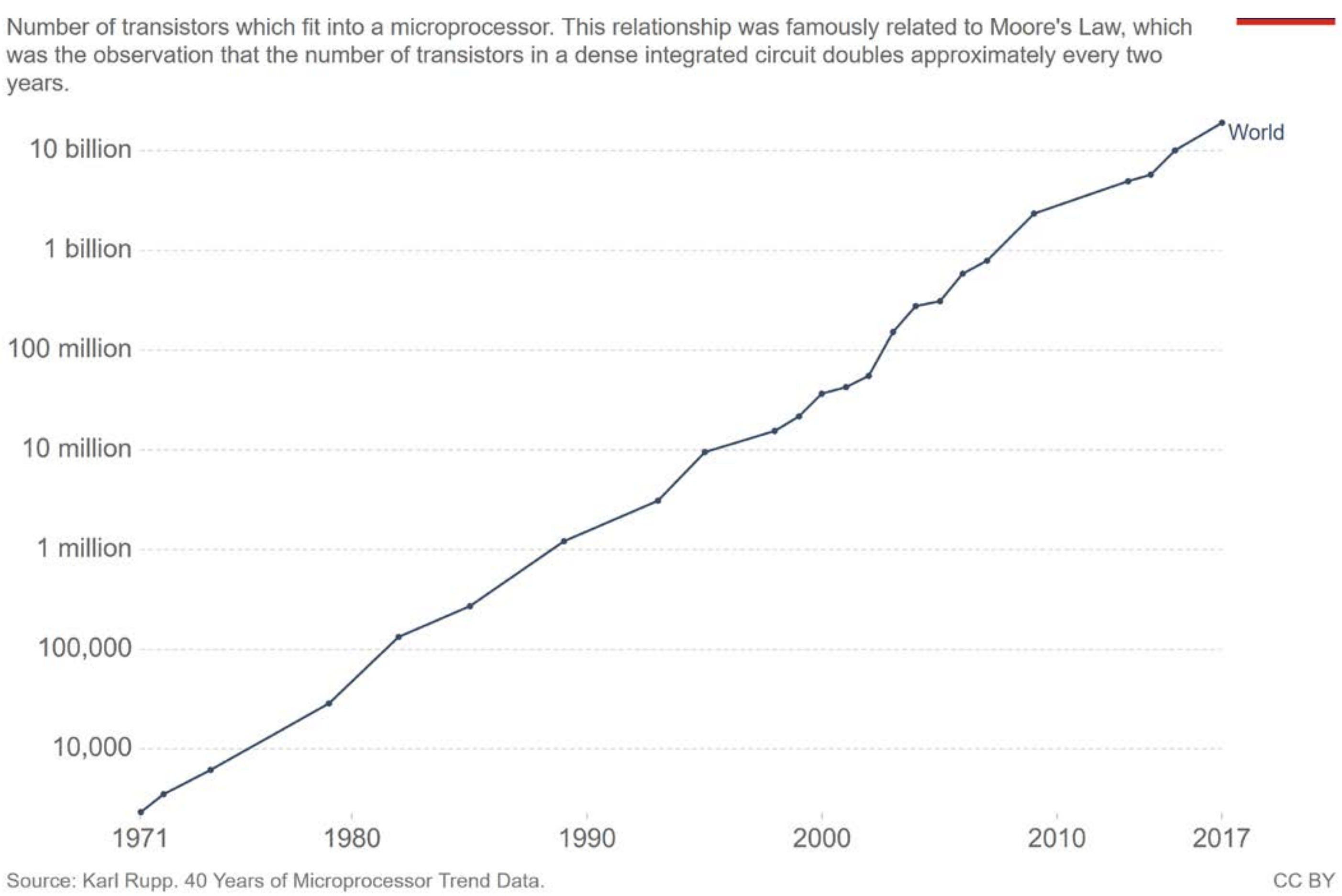
18 https://arxiv.org/pdf/1806.04558.pdf

Technology hardware to support learning computers

The throughput afforded by an average-priced serial processor has traditionally served as a milestone by which we measure our computational progress. This is Moore's law: the prediction about transistor densification that suggests that this double every 2 years.[19] A prediction that has become well-known in part because its advance has been inextricably linked with our ability to conduct computation. With advances in neural networks that relationship is changing.

Exhibit 2:
Moore's Law: Transistors per microprocessor

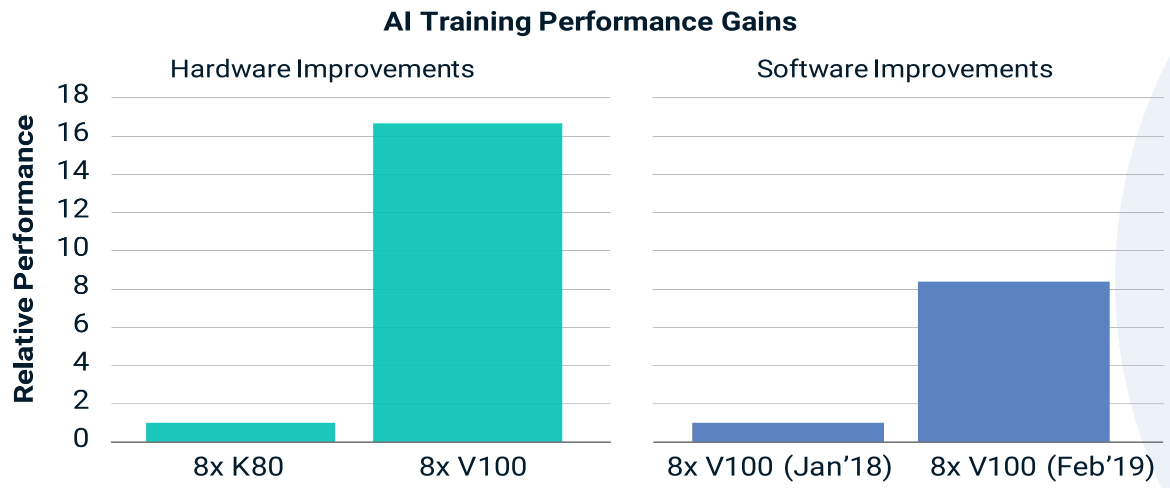
Source: https://ourworldindata.org/grapher/transistors-per-microprocessor



The advances in deep learning models that allow more complex problems to be attacked has been achieved by using a different sort of processor — one that runs its operations simultaneously rather than consecutively — and hence allow those problems to be solved more efficiently. Between advances in distributed computing, parallel processing and advances in neural network implementation, the capabilities of neural networks have improved roughly 140-fold on established benchmarks over the past three years.[20] This disaggregates into more than 16x of performance increase due to processor advances and more than 8x due to architecture and algorithmic improvements, and these two effects compound. Over the same time-scale Moore's Law would be expected to deliver a 2.8x computational improvement. Taken together, neural network compute architecture is advancing at 50x the rate of Moore's Law.[21]

Exhibit 3:
AI training performance gains, 2017 to 2020

Source: Ark Invest, Stanford DAWN Deep Learning Benchmark



19 https://en.wikipedia.org/wiki/Moore%27s_law

20 https://dawn.cs.stanford.edu/benchmark/index.html

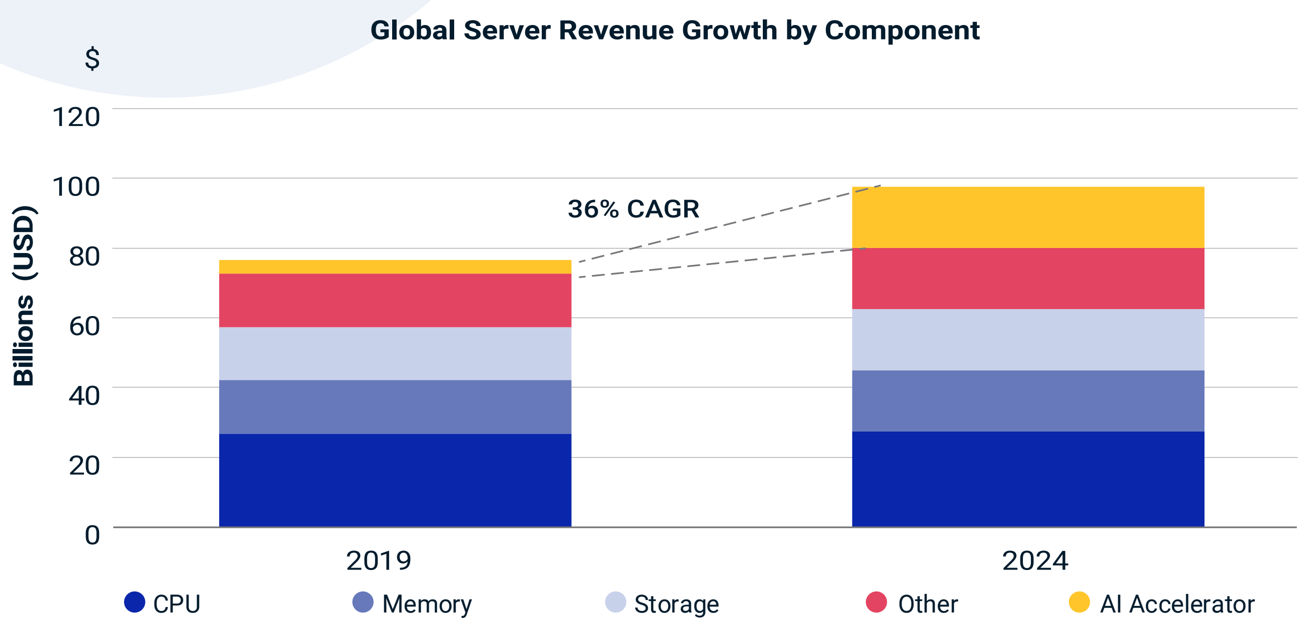
21 "The Cost of AI Training is Improving at 50x the Speed of Moore's Law: Why It's Still Early Days for AI", James Wang, ARK Invest White Paper, May 2020. https://ark-invest.com/analyst-research/ai-training/

This capability-gain and performance speed-up has seen a significant investment cycle. ARK Invest forecast neural network accelerators to significantly outgrow global server spend: in 2019, they commanded approximately USD 4 billion in spend but this is projected to reach USD 18 billion by 2024.This growth and the associated power consumption has led to extensive academic research into faster and more efficient AI chip design.[22]

For the end-user, such investments may lead to important capability-advances for the “computational endpoints” of everyday life. A smartphone can serve as an intuitive miniature computer with a touch interface, but many other internet-connected devices have input-output bandwidth-constraints that currently limit their utility. A smartwatch’s screen is too small, smart TVs have no equivalent to a pointer; while with smart versions of speakers, fire alarms, doorbells or toothbrushes there is no screen at all. Neural networks could provide the vector by which these devices might divorce themselves from the phone.

Exhibit 4:
AI Accelerator share of server spend: 2019 vs 2024

Source: ARK Invest



“Voice interfaces” now depend upon deep learning neural networks to understand the words that we say, construct those commands into meanings and to respond effectively. This machine-learning driven calibration has helped address the three key challenges of speech recognition: discoverability (what commands are allowed?), transcription (what errors are made in listening?) and understanding.[23] Smart speakers have achieved mainstream adoption faster than almost any technological device in history, growing from 10% penetration of US households to more than 50% penetration in 3 years. In part, this is linked to their (still imperfect) neural network-enabled responsiveness: in 2019 Google made its AI assistant 10x faster by siting the underlying neural network directly on the device rather than relying upon a cloud connection.[24]

Potentially more important than an ability to directly respond will be the ability of these computational systems to anticipate needs. Wearables can already use neural networks to recognize both the quotidian — “would you like to start recording an exercise session” or “with current traffic, it’s time to leave the office” — alongside the critical — “your heart seems to be in atrial fibrillation, please call 9-11”.

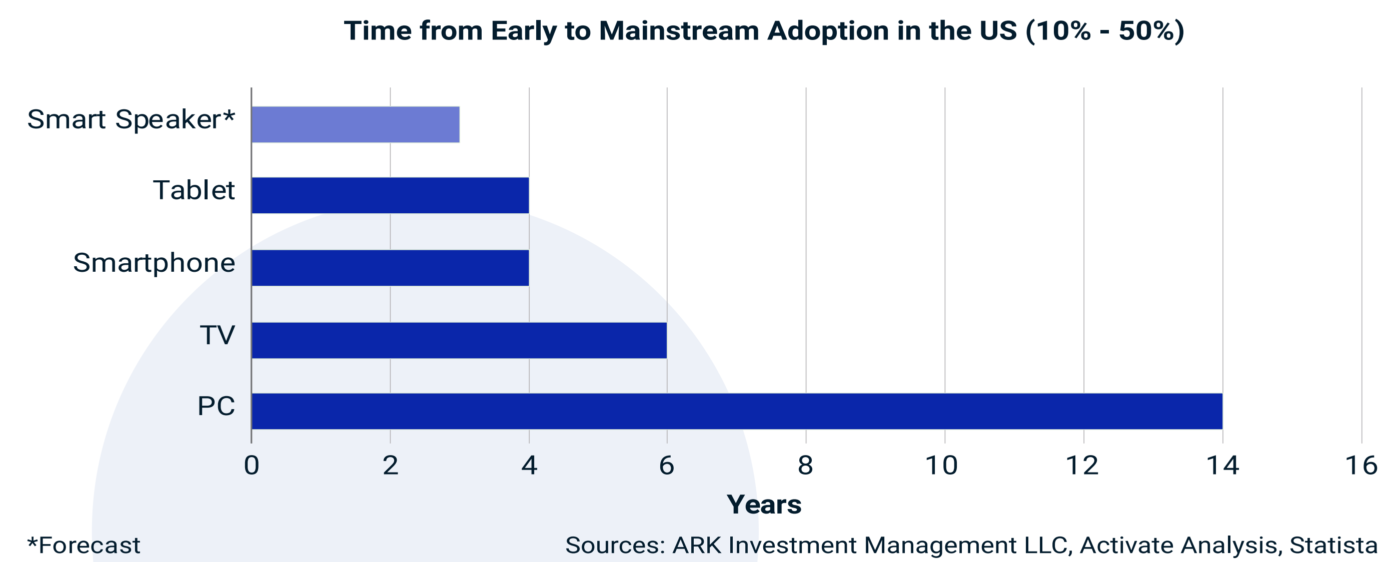
22 For example, http://news.mit.edu/2019/ai-chip-light-computing-faster-0605

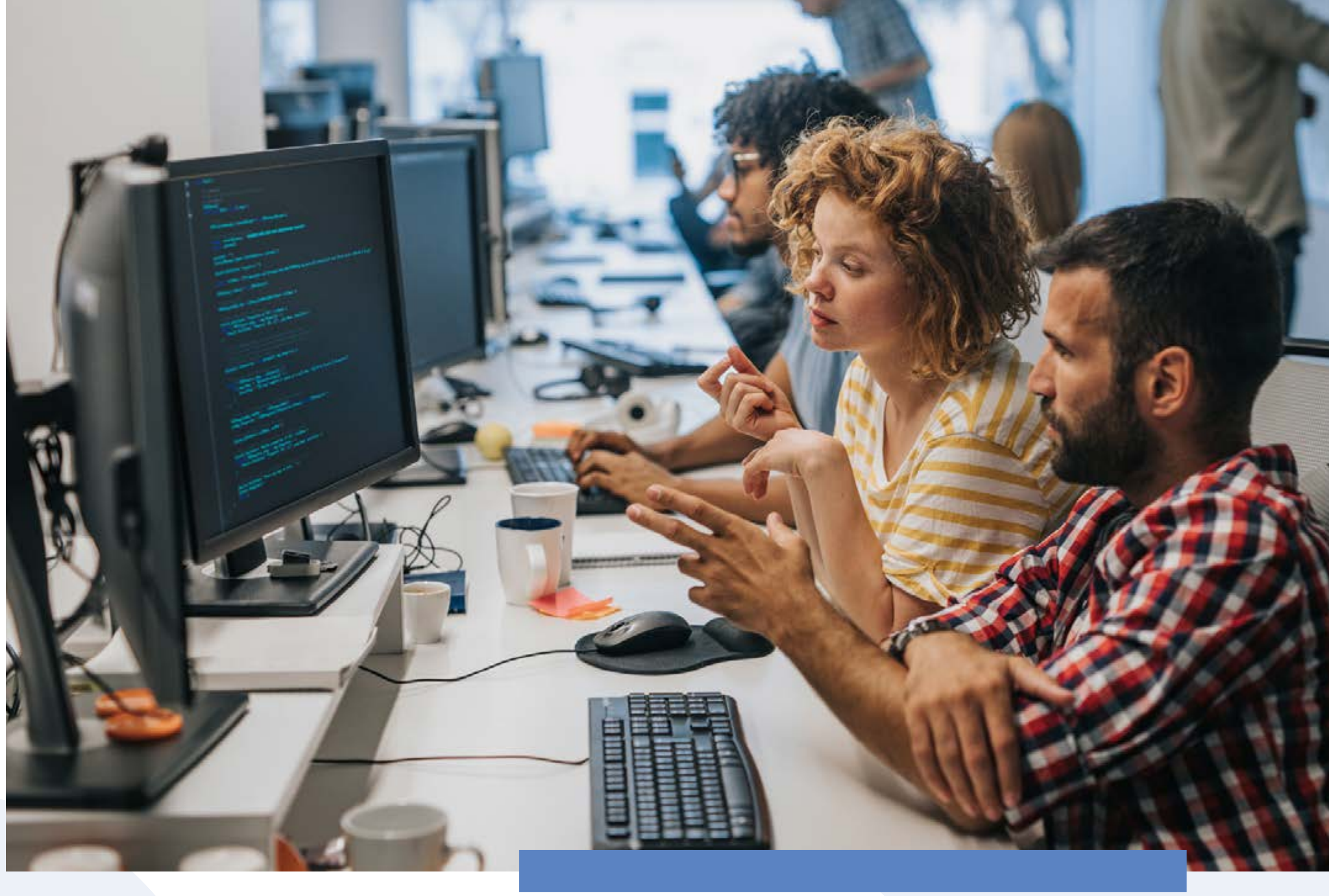
23 https://en.wikipedia.org/wiki/Voice_user_interface#Design_challenges

24 https://www.blog.google/products/assistant/next-generation-google-assistant-io/

Exhibit 5:
Time estimates for mainstream adoption of new technologies

Source: ARK Invest





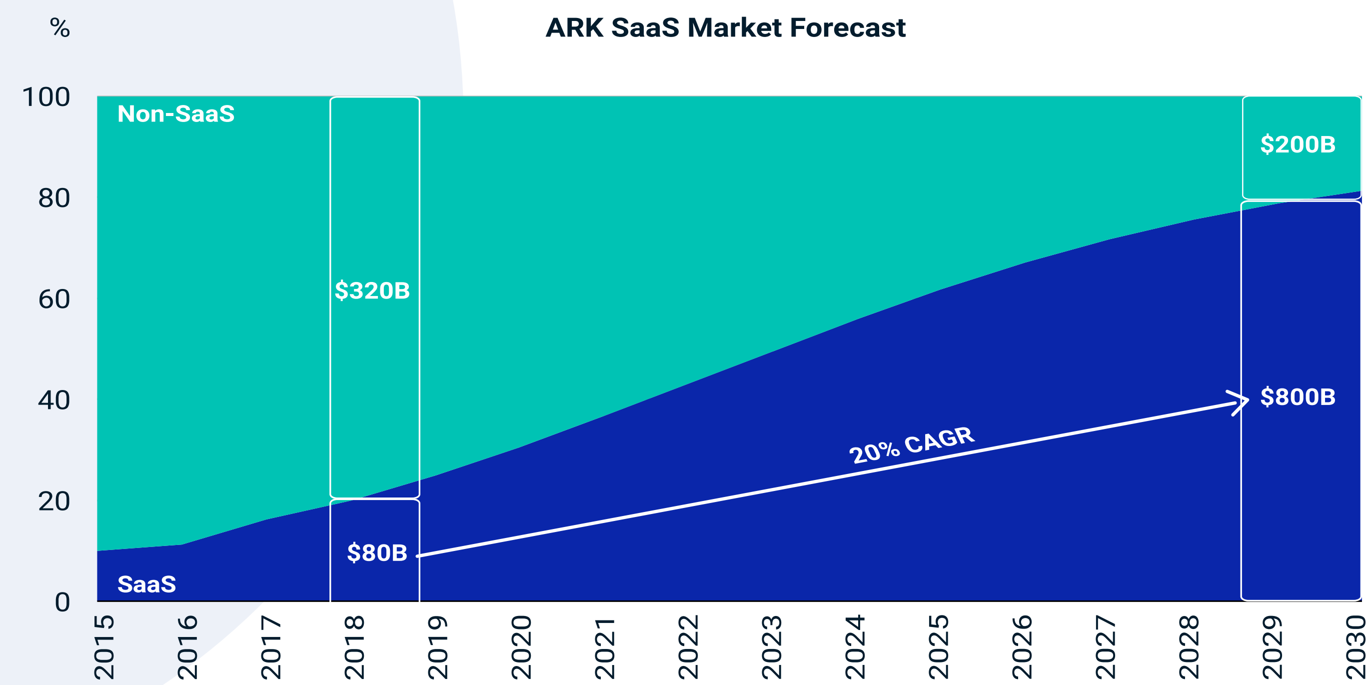
The sensors that we carry and house may become more useful as computational systems when they are better able to make sense of the data. The companies with access to the largest set of orthogonal datasets have a natural advantage when it comes to training the neural networks to achieve this. Deep learning could mean a virtuous circle of user-engagement begetting data begetting more capability begetting more user-engagement is possible.

The changing value of data

To succeed in an age of the learning computer, enterprises need the data for training and updating their neural networks. Business models that incorporate data accrual and the distributed assets that provide those datafeeds then become important to the value-chain. Such business models are already seen in consumer space: smart speakers and smart TVs have been priced to help drive penetration to support hopes of establishing dominant data positions.[25]

Exhibit 6:
Software-as-a-service market share: 2015 to 2030E

Source: ARK Invest



Comparably, in the enterprise space, the dominance of the software-as-a-service business model has emerged. Cloud-hosted software provides developers with a huge window into the world of their customers. The Software-as-a-Service (SaaS) model allows for easier customer acquisition and on-boarding, enables companies to dynamically adjust pricing, and provides for frequent updates while keeping all customers on the same build.

The utility and risks of the data-dependent aspects of the SaaS model, while already well-understood, are likely to gain in importance as deep learning models become more capable.[26] With neural networks, the best practices of some customers could be used to help other customers to operate more productively and efficiently whether in content creation or in the most effective sales process. ARK Invest forecast that SaaS revenues will grow from roughly USD 80 billion in 2018 to USD 650 billion by 2030, while traditional software revenues shrink by more than half (USD 320 billion to USD 150 billion over the period).

25 https://technode.com/2019/07/09/chinas-tech-giants-battle-for-smart-speaker-supremacy-as-price-war-rages-on/
26 https://www.mckinsey.com/business-functions/risk/our-insights/securing-software-as-a-service

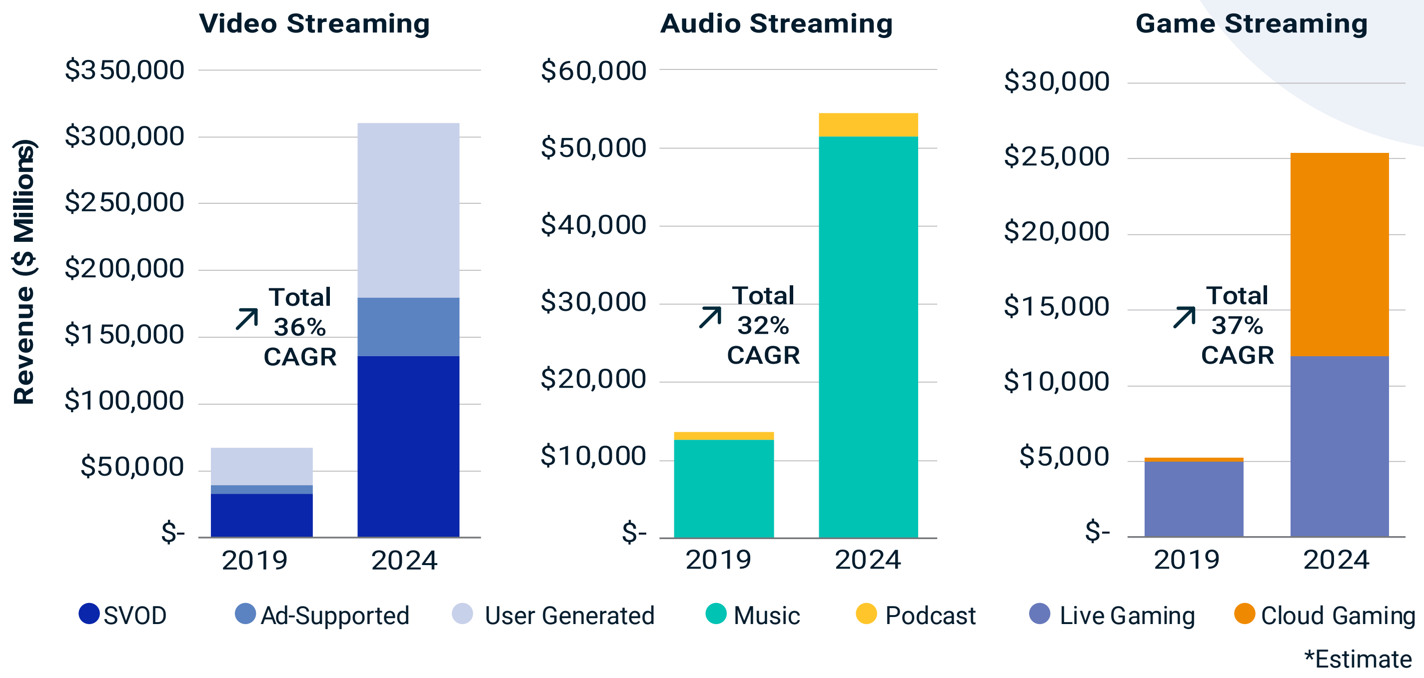
The importance of data is not restricted to software companies, nor are they the only ones with explicit business strategies based on a transformation in the utility of proprietary data. Tesla has stated that it is relying upon its data from its distributed fleet of vehicles to enable its progress to develop autonomous driving capabilities.[27] Amazon's checkout-less Amazon Go retail stores depend upon neural networks to identify the items as they are taken off the shelves; the scope to increase throughput and inventory depends upon that system's improvement.[28] Tiktok has built a business on top of providing users with neural net-enabled content creation tools and a platform that algorithmically serves that content to other users. The company is able to ship more paid content generation capabilities as a function of its ability to measure how different capabilities boost each creator's audience engagement.[29]

Future computation and future services

How might these learning computers disrupt business models and competitive order across different economic sectors? Within the media and content ecosystems, the new computational model has been adopted by on-demand business models that can harness the power of algorithmic content feeds.[30] ARK invest forecast that streaming media platform revenues would grow from less than USD 90 billion in 2019 to USD 390 billion in 2024. All these streaming services, not least new cloud-gaming services, are dependent on neural network models' ability to reduce perceived lag and latency that would otherwise severely disrupt the user-experience.

Exhibit 7:
Media streaming revenues, 2019 to 2024E

Source: ARK Invest



Within the advertising ecosystem, social and algorithmic feed platforms are by nature data-driven and hence are seemingly best positioned to benefit in an age of the learning computer.[31] Digital ad spending has just surpassed 50% penetration of overall ad-spending globally[32] and seems to be relatively less vulnerable than traditional billboard, newspaper and billboard ad-spend during a lockdown-driven economic downturn.[33]

Within logistics, the cost of last-mile delivery may decline by more than an order of magnitude if neural networks can facilitate autonomous delivery platforms.[34] What might the impact be for e-commerce uptake? As the cost of delivery falls, more traditional retail purchases can economically shift to such platforms. By 2030, 60% of global retail sales are forecast to flow through ecommerce platforms, up from 14% in 2019, with more than half delivered by autonomous systems (see Exhibit 8).

27 https://www.youtube.com/watch?v=Ucp0TTmvqOE&feature=youtu.be

28 https://towardsdatascience.com/how-the-amazon-go-store-works-a-deep-dive-3fde9d9939e9

29 https://www.hugheseducation.com/blogs/tiktok-used-business-analytics-to-become-rank-1-social-media-platform

30 https://ark-invest.com/analyst-research/music-streaming-revenue-50-billion/

31 https://ark-invest.com/analyst-research/internet-advertising-growth/ and https://ark-invest.com/analyst-research/tv-ad-spending/

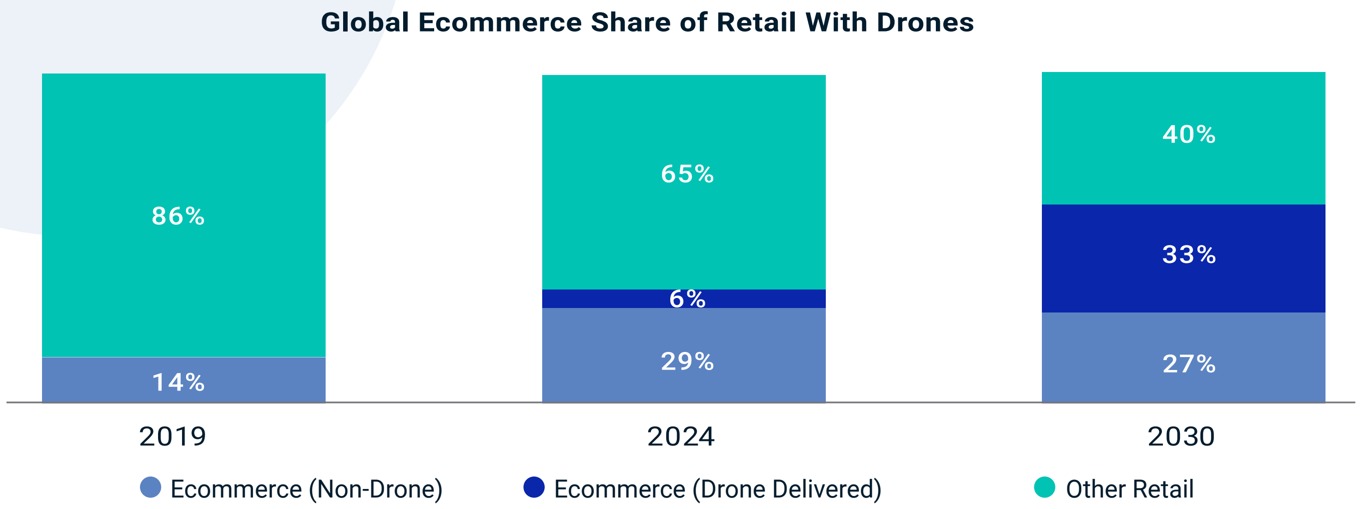
32 https://www.emarketer.com/content/global-digital-ad-spending-2019

33 https://www.iab.com/insights/coronavirus-ad-spend-impact-buy-side/

34 https://ark-invest.com/analyst-research/autonomous-delivery-robots/

Exhibit 8:
E-commerce share of global retail, 2019 and 2030E

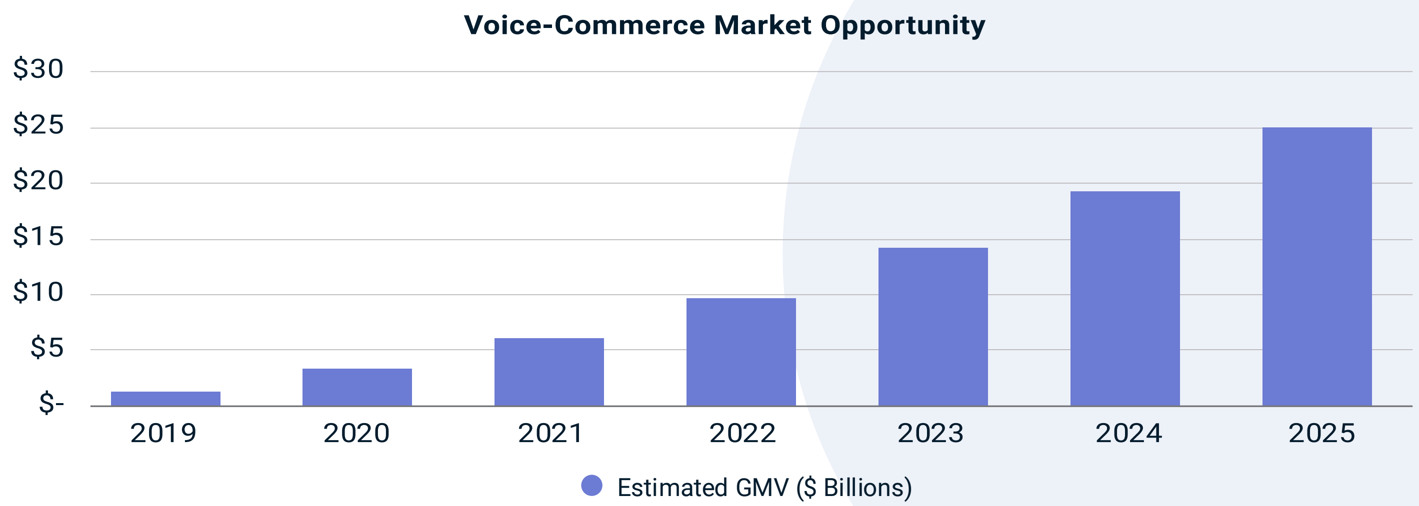
Source: ARK Invest



The proliferation of voice-interface-devices described earlier would facilitate a growth in "voice commerce", with 2025 merchandise volumes expected to reach $25 billion in the US alone.[35] E-commerce platforms could have a natural interest in facilitating this channel over the currently dominant screen-based selection as the consequent lack of transparency/comparison could shift customers into product choices (e.g. private label) where their margins are higher.

Exhibit 9:
Voice-Commerce US Market Forecasts: 2019 to 2025E

Source: ARK Invest

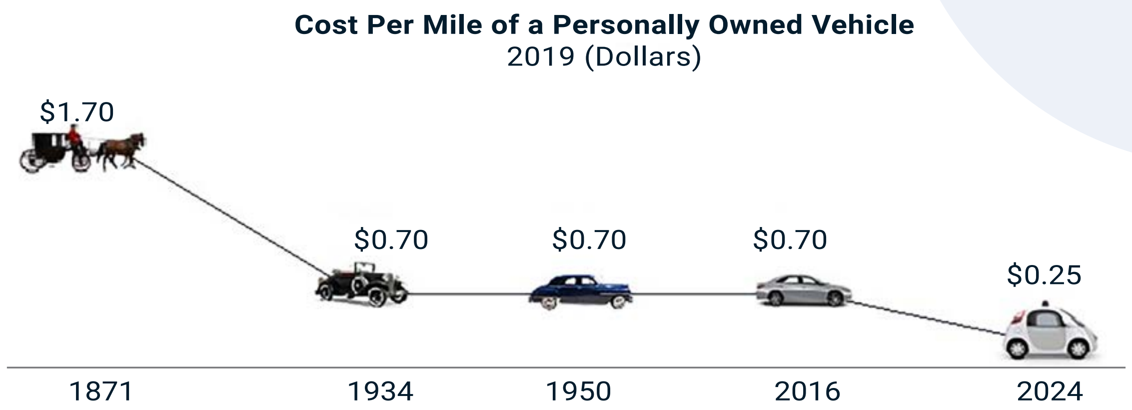


In finance, Amazon Go-style implementations may reduce the need for checkout queues, document signing platforms could use natural language processing to facilitate legal contract creation and standardization, while machine-learning driven risk-assessment could accelerate credit and insurance allocation processes.[36]

In transportation, learning computers are a critical component in the progress towards autonomous taxis. If they arrive, they may deliver a discontinuous reduction in the cost of point-to-point mobility. Cars are expensive machines that US research suggested spend 95% of their time parked.[37] Taxis and ridehailing cabs have more appealing utilization rates. However, the cost of compensating the driver and financing place the final per-mile price at rates that exceed personally driven vehicles by as much as 5x.[38] An autonomous taxi may combine high utilization with low fleet maintenance and zero direct driver costs. Hence point-to-point mobility could be made available at less than half of today's personally driven rates (Exhibit 10).

Exhibit 10:
Cost per mile of personal transportation – 1871 to 2024E

Note: ARK had estimated previously that an autonomous taxi could price at $0.35 per mile. We have refined our estimates and believe that autonomous taxis could be even cheaper, at only $0.25 per mile. Forecasts are inherently limited and cannot be relied upon. | Source: ARK Investment Management LLC, 2019; Morton Salt Company Records, American Automobile Association (AAA).



35 https://ark-invest.com/analyst-research/smart-speaker-apps/

36 For example https://comm.seal-software.com/hubfs/Whitepapers/Seal-Whitepaper-System%20of%20Agreement%20-%20Manage.pdf, ftp://ftp.repec.org/opt/ReDIF/RePEc/ami/articles/14_1_3.pdf and https://www.researchgate.net/publication/23776815_A_neural_network_approach_for_credit_risk_evaluation

37 https://tsrc.berkeley.edu/publications/impacts-car2go-vehicle-ownership-modal-shift-vehicle-miles-traveled-and-greenhouse-gas

38 https://ark-invest.com/analyst-research/autonomous-ridehailing-fees/

Conclusions

The case studies presented here illustrate richly how potentially pervasive the impact on business processes and models might be economywide. From 1997 to 2019, internet-dependent business lines are estimated to have accrued value equivalent to approximately 15% of global market capitalization (USD 10 trillion).[39] If learning computer-based business lines were to command an equivalent share over the next two decades, the figure for a new computational age of deep learning would be USD 30 trillion.

39 https://ark-invest.com/analyst-research/ai-training/

MSCI would like to thank ARK Invest for useful discussions and insightful analysis of this megatrend, which have greatly facilitated the preparation of this document.

MSCI

Contact us

clientservice@msci.com

AMERICAS

Americas	1 888 588 4567 *
Atlanta	+ 1 404 551 3212
Boston	+ 1 617 532 0920
Chicago	+ 1 312 675 0545
Monterrey	+ 52 81 1253 4020
New York	+ 1 212 804 3901
San Francisco	+ 1 415 836 8800
São Paulo	+ 55 11 3706 1360
Toronto	+ 1 416 628 1007

EUROPE, MIDDLE EAST & AFRICA

Cape Town	+ 27 21 673 0100+
Frankfurt	49 69 133 859 00
Geneva	+ 41 22 817 9777
London	+ 44 20 7618 2222
Milan	+ 39 02 5849 0415
Paris	0800 91 59 17 *

ASIA PACIFIC

China North	10800 852 1032 *
China South	10800 152 1032 *
Hong Kong	+ 852 2844 9333
Mumbai	+ 91 22 6784 9160
Seoul	00798 8521 3392 *
Singapore	800 852 3749 *
Sydney	+ 61 2 9033 9333
Taipei	008 0112 7513 *
Thailand	0018 0015 6207 7181 *
Tokyo	+ 81 3 5290 1555

About MSCI

MSCI is a leading provider of critical decision support tools and services for the global investment community. With over 45 years of expertise in research, data and technology, we power better investment decisions by enabling clients to understand and analyze key drivers of risk and return and confidently build more effective portfolios. We create industry-leading research-enhanced solutions that clients use to gain insight into and improve transparency across the investment process.

To learn more, please visit **www.msci.com**