



Deep Learning

An Artificial Intelligence Revolution

Published: June 12, 2017

Author

James Wang, Analyst at ARK Invest



EXECUTIVE SUMMARY

Deep learning—a form of artificial intelligence inspired by the human brain—is sweeping across every industry around the world. Thanks to this breakthrough, computer vision, voice recognition, speech synthesis, machine translation, game playing, drug discovery, and robotics are setting game-changing performance records.

ARK believes that deep learning is one of the most important software breakthroughs of our time. Now that software is enabling most industries, deep learning will have a profound impact on autos, robotics, drones, biotech, finance, agriculture, among many others. According to our research, companies founded on deep learning will unlock tens of trillions of dollars in productivity gains, and add \$17 trillion in market capitalization to global equities during the next two decades.

This paper will explore the origins of deep learning, how it works, and how it differs from machine learning. Then, it will examine important use cases, the leading companies in the space, and the algorithmic improvements likely to enter commercial products during the next five years.

RESEARCH HIGHLIGHTS*

*Based on ARK's research

- | \$17 trillion in market capitalization creation from deep learning companies by 2036.
- | \$6 trillion in revenue from autonomous on-demand transportation by 2027.
- | \$6 billion in revenue for deep learning processors in the data center by 2022, growing more than tenfold over 5 years.
- | \$16 billion addressable market for diagnostic radiology.
- | \$100-\$170 billion in savings and profit from improved credit scoring.
- | \$12 trillion in real GDP growth in the US from automation by 2035.



JAMES WANG
ARK Invest

ABOUT THE AUTHOR

James joined ARK as an analyst on the Next Generation Internet team in April 2015, focusing on artificial intelligence and the next wave of the internet. James worked for nine years at NVIDIA where he helped launch GeForce Experience, a PC gaming application with over 80 million users. He founded and edited GeForce.com, NVIDIA's gaming portal with over 15 million visitors. He has written about the technology industry since 2000 for various magazines and publications. James graduated from the University of New South Wales with a Bachelor of Engineering degree in Computer Engineering.



CONTENTS

Why Deep Learning.....	4
The Origins of Deep Learning.....	4
The Rise of Deep Learning	7
The Three Drivers of Growth	9
 The Deep Learning Opportunity.....	 11
Deep Learning Could Be a \$10 Trillion Industry in 20 Years	11
 Deep Learning Use Cases.....	 13
Computer Vision	13
Image Search	14
Auto Face Tagging	14
Computer Vision as a Service	14
AI Assistants & Bots	16
Robotics	20
Radiology.....	23
Credit Assessment.....	27
Autonomous Vehicles	29
 Deep Learning Hardware	 31
From CPU to GPU to ASIC	31
Alternate Deep Learning Processors	33
Deep Learning Data Center Opportunity	34
 The Future of Deep Learning.....	 36
The Limits of Deep Learning	36
Deep Learning With Memory	38
Generative Networks	39
 Conclusion	 40



WHY DEEP LEARNING

This paper focuses on deep learning as opposed to the wider fields of machine learning and artificial intelligence (AI) for four reasons. First, the vast majority of AI breakthroughs in recent years are *thanks to* deep learning. Second, deep learning is not a specific breakthrough; instead, it is a broadly applicable AI technique that is advancing the state of the art. Third, the technology industry is rallying around deep learning because of its step-function increase in performance and broad-based applications. Google, Facebook, Baidu, Microsoft, NVIDIA, Intel, IBM, OpenAI, and various startups have made deep learning a central focus. Fourth, from an investor perspective, deep learning is only five years old and particularly compelling given its low revenue base, large addressable market, and high growth rate.

THE ORIGINS OF DEEP LEARNING

Deep learning is a modern name for an old technology—artificial neural networks. An artificial neural network, or simply neural net, is a computer program loosely inspired by the structure of the biological brain. The brain is made up of billions of cells called neurons connected via pathways called synapses. New observations and experiences alter the strength of the synaptic connections. Through the accumulation of observations and experience, the strength of the connections converges, resulting in “learning”. Neural nets simulate these structures in software, with digital versions of neurons, synapses, and connection strengths. By feeding training examples, or “experience”, to an artificial neural network and adjusting the weights accordingly, a neural net learns complex functions much like a biological brain.

The first neural network built on these biological principles was the Perceptron.¹ This simple network used two layers of connected neurons and could be taught to perform simple image recognition tasks. Improved understanding of the visual cortex (the portion of the brain devoted to visual processing) led to the development of the Neocognitron,² a neural net composed of stacks of smaller, simpler layers. The use of multiple layers makes the network “deep” and allows it to perceive the world across multiple levels of abstraction. As a result, the Neocognitron was able to recognize characters in different positions and of various sizes.

Deep networks excelled at certain perception tasks, but they were difficult to train. In response, computer scientists developed backpropagation,³ a technique that trains deep networks by applying calculus to labeled data sets.

1 “The Perceptron: A Probabilistic Model For Information Storage and Organization In The Brain,” Psychological Review, 1958
2 <http://www.cs.princeton.edu/courses/archive/spr08/cos598B/Readings/Fukushima1980.pdf>
3 <https://www.nature.com/nature/journal/v323/n6088/pdf/323533a0.pdf>



The combination of deep neural networks and backpropagation yielded powerful results. In the early 1990s, a team led by Yann LeCun at AT&T Bell Labs developed a convolutional neural network trained by backpropagation that was able to read handwritten numbers with 99% accuracy at a 9% rejection rate.⁴ Subsequently, this system found its way into the banking system and processed more than 10% of all the checks in the United States.

Despite the early success of neural nets, the broader AI community was skeptical. Researchers generally preferred other machine learning algorithms that were simpler to implement, easier to train, and computationally less demanding. They favored Support Vector Machines, which performed on par with neural nets, for vision tasks in the 1990s.⁵ For voice recognition applications, they preferred hidden Markov models.



4 "Handwritten Digit Recognition with a Back-Propagation Network," Advances in Neural Information Processing Systems 2 (NIPS), 1989
5 "Deep Learning in Neural Networks: An Overview", Jürgen Schmidhuber 2014, <https://arxiv.org/abs/1404.7828>

DEEP LEARNING TIMELINE

1958	Rosenblatt invents the "Perceptron", a machine that can detect shapes through a network of neuron-like hardware units.
1978	Fukushima invents "Neocognitron", a multi-layer neural network capable of detecting different shapes without being affected by shift position or minor distortion. This is the first convolutional neural network and the first "deep" architecture in the modern sense.
1986	Rumelhart, Hinton, and Williams popularizes "backpropagation" as an effective way to train deep neural nets. Backpropagation is the key breakthrough that makes neural nets effective learners and remains the dominant way neural nets are trained as of 2017.
1989	LeCun uses backpropagation to train convolutional neural nets and shows that the resulting network can read handwritten digits with 99% accuracy. This system is later widely deployed in ATMs for automated check reading.
2011	Google Brain, a deep neural network powered by 16,000 CPUs, learns to recognize cats by watching YouTube videos.
2011	Deep neural nets trained GPUs achieve 99% accuracy in street sign recognition, exceeding human performance for the first time in a controlled test.
2012	Deep neural net achieves record performance in classification of natural photographs in the ImageNet 2012 challenge, reducing error rates by 36% relative to existing machine learning programs. Neural nets go on to win the ImageNet contest in each of the next four years. This is considered by many to be the watershed moment in deep learning.
2013	Deep learning achieves record performance in a number of speech and character recognition tests.
2014	Facebook launches DeepFace, a deep neural net that detects faces with 97% accuracy, approaching human performance.
2015	Microsoft's ResNet deep neural net achieves 96% in the ImageNet classification challenge, reaching human level performance for the first time.
2015	Baidu Deep Speech 2 achieves human level performance in certain voice to text transcription for English and Mandarin in various benchmarks.
2016	DeepMind's AlphaGo, an AI program that combines deep learning with Monte Carlo tree search, defeats 18 time international champion Lee Sedol in the game of Go, reaching a major AI milestone ten years ahead of schedule.

Source: ARK Investment Management LLC, Data taken from research paper: "Deep Learning in Neural Networks: An Overview", Jürgen Schmidhuber 2014, <https://arxiv.org/abs/1404.7828>



THE RISE OF DEEP LEARNING

In 2012, neural networks began to deliver astounding performance results that far eclipsed other machine learning algorithms in both visual and audio applications. In image classification, deep learning reduced error rates from 26% to 3%. In voice recognition, deep learning reduced error rates from 16% to 6%. Critically, within the last two to three years, deep learning has surpassed human performance.

FIGURE 1
Image Classification Error Rate

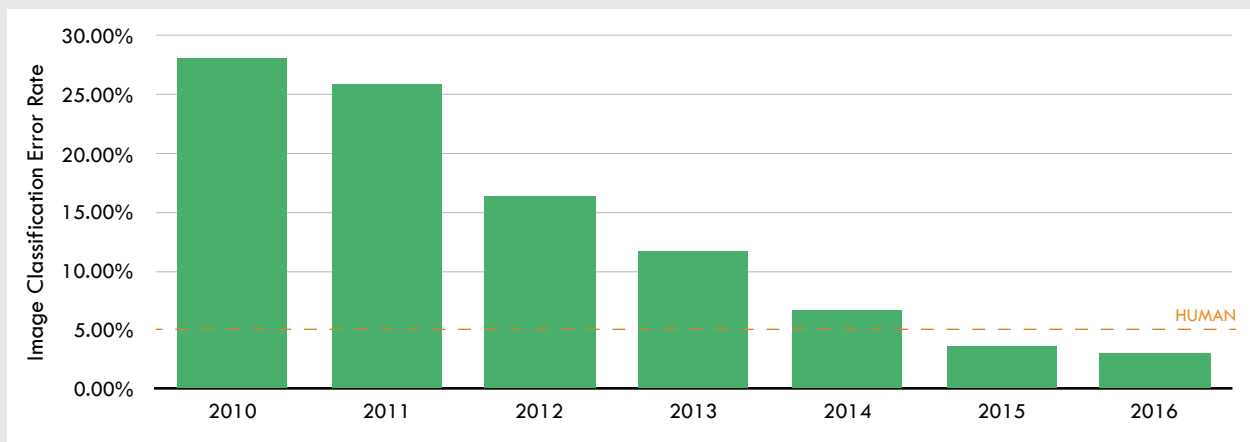
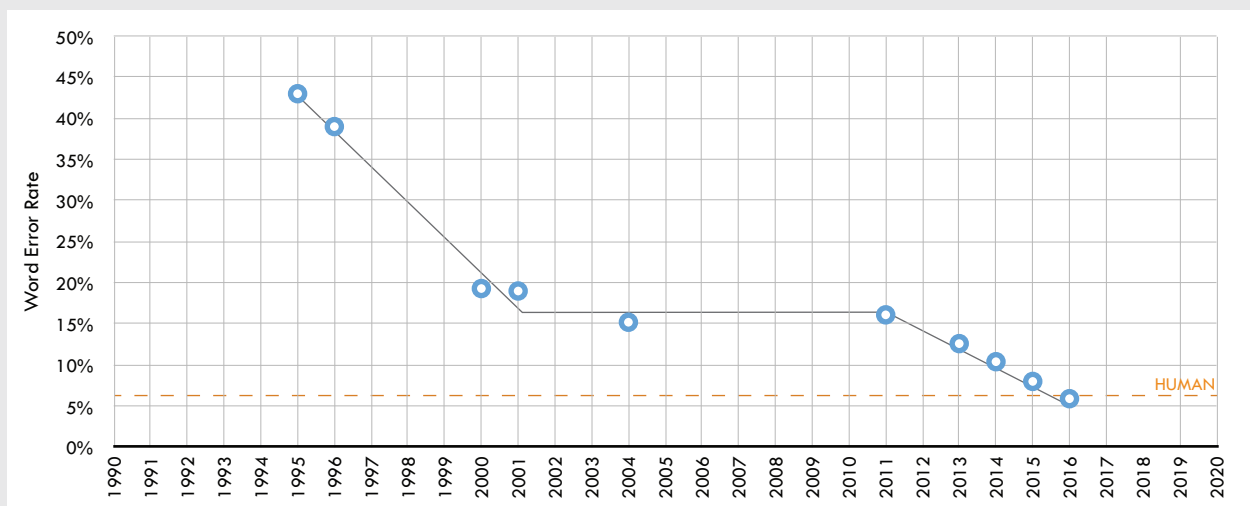


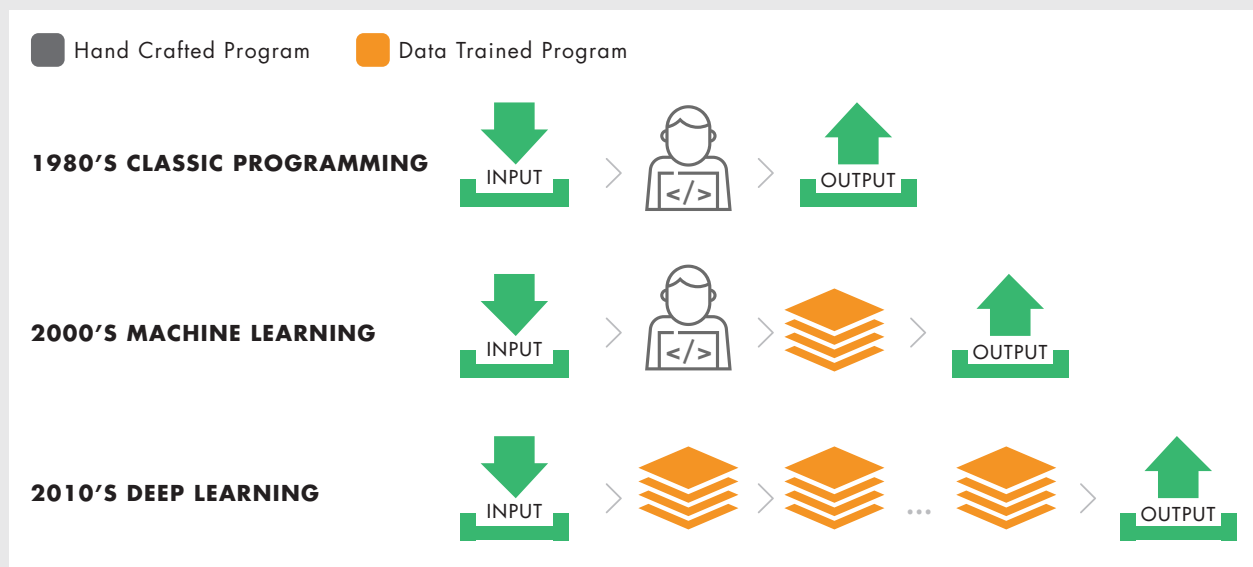
FIGURE 2
Speech to Text Transcription Error Rate (Switchboard)





Deep learning has proven so effective that it has become core to the evolution of the technology industry. Google, Facebook, Microsoft, Apple, Baidu, IBM, and others have gravitated to deep learning for image and voice recognition. Likewise, AI startups like DeepMind, Vicarious, Nervana, OpenAI, Clarifai, and Enlitic, among others have focused on deep learning as their key enabling technology. The fact that companies large and small have embraced the same technology indicates that deep learning truly is distinguished relative to other machine learning algorithms, and explains why it has become foundational for next generation applications.

FIGURE 3
Deep Learning vs. Other Programming Techniques



Source: ARK Investment Management LLC, Yoshua Bengio

What makes deep learning unique is that it uses data to automate the programming process end-to-end, as shown above. In classic programming, humans manually design all parts of the program. This works for simple, well-defined problems but breaks down for more complex tasks. Machine learning improves upon this by replacing some stages of the program with stages that can be trained automatically with data, making it possible for computers to perform more complex tasks such as image and voice recognition.

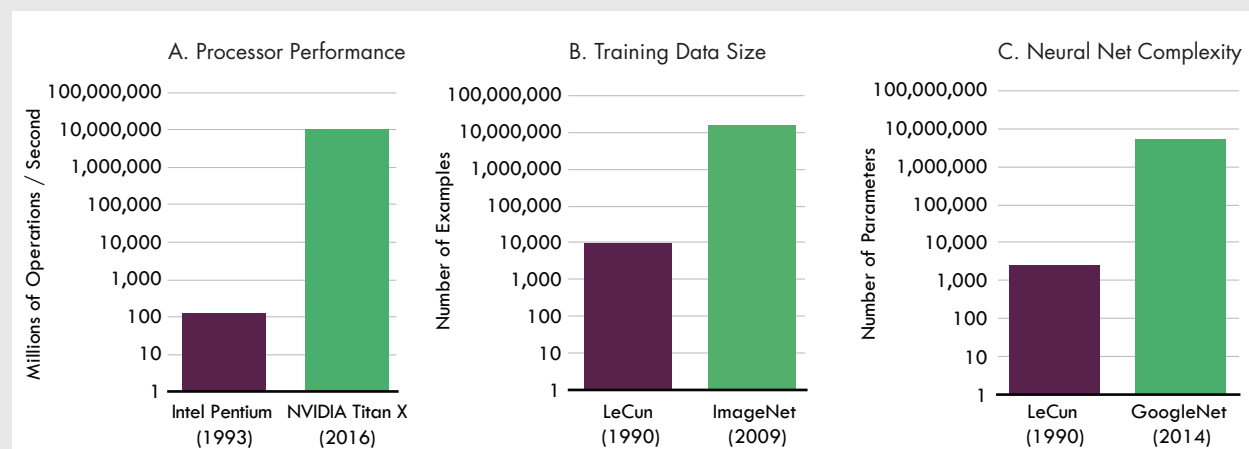
Deep learning takes this idea to its logical conclusion and replaces the entire program with stages that can be trained with data. The result is that programs can be far more capable and accurate while requiring less human effort to create.



THE THREE DRIVERS OF GROWTH

According to ARK's research, three factors have contributed to the recent breakthroughs in deep learning: faster processor performance, larger data sets, and more sophisticated neural nets. In the charts below, we quantify how each of these factors has improved since the 1990s.

FIGURE 4
Drivers of Growth | A. Processor Performance B. Training Data Size C. Neural Net Complexity



Source: Intel, NVIDIA, ImageNet, ARK Investment Management LLC

Processor performance has improved roughly five orders of magnitude since Intel's original Pentium processor for two reasons. First, thanks to 23 years of Moore's Law, the size of transistors has collapsed from 800nm to 16nm,⁶ enabling the creation of processors with billions of transistors operating in the gigahertz range and increasing computational power by five orders of magnitude. Second, graphics processing units (GPUs) emerged as a new class of microprocessors with roughly 10x the performance of central processing units (CPUs) when applied to the types of massively parallel operations required for deep learning.⁷ Consequently, a single processor today can perform over ten trillion calculations per second, making it possible to train highly complex neural networks in a matter of days. During the 1990s, the same computation would have taken multiple lifetimes.

The performance of deep learning programs is correlated highly to the amount of data used for training. While the performance of other machine learning algorithms plateau with more data, those associated with deep learning continue to scale with more training data, as shown below. Thanks to the internet's size and scale, deep learning has thrived with access to very large datasets at a minimal cost.

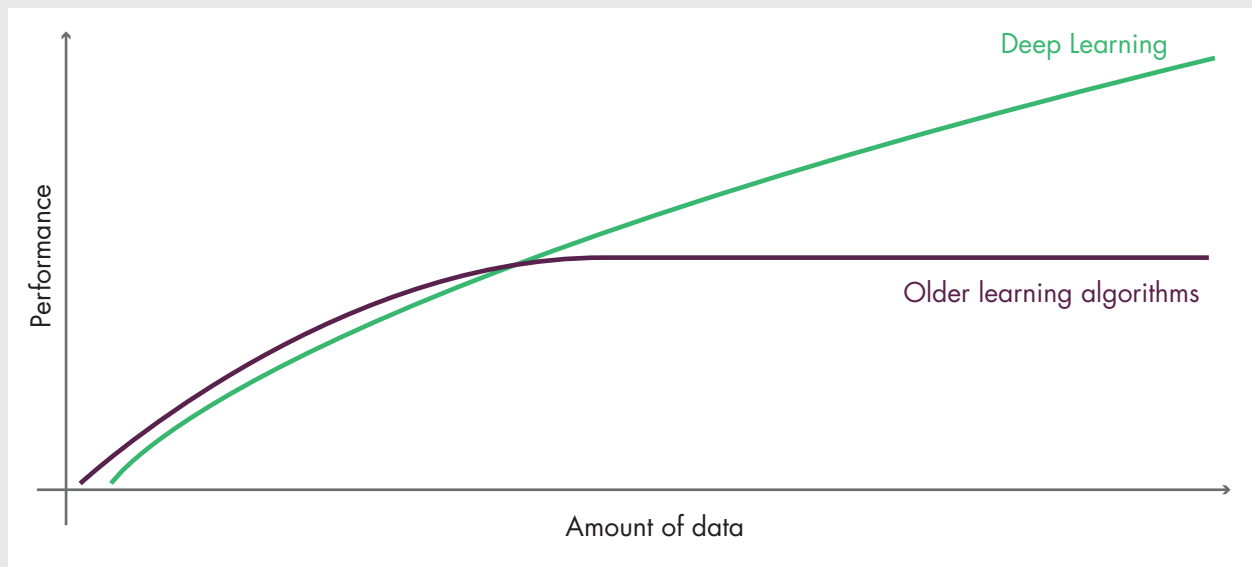
⁶ Historical analysis of prevailing semiconductor manufacturing nodes during this period.

⁷ Overview of deep learning literature shows examples of GPU acceleration from 5x-50x depending on the algorithm.



While LeCun's 1990 handwriting reader used approximately 10,000 samples collected from the US Postal Service, in 2009 ImageNet's dataset contained more than 10 million examples of high resolution photographs. Likewise, Baidu's DeepSpeech 2⁸ draws upon over 10,000 hours of audio data compared to a few hundred hours in legacy data sets.

FIGURE 5
Why Deep Learning



Source: Andrew Ng

Adding to the power of deep learning, the neural nets themselves have become larger and more sophisticated, as measured by their number of free “parameters”. Parameters are dials used to tune the network’s performance. Generally speaking, more parameters allow a network to express more states and capture more data. Today’s deep learning networks have roughly ten million parameters, or four orders of magnitude more than LeCun’s original handwriting reader.⁹

8 “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin,” Baidu, 2015 <https://arxiv.org/abs/1512.02595>

9 “Handwritten Digit Recognition with a Back-Propagation Network,” Le Cun et al, 1989



THE DEEP LEARNING OPPORTUNITY

Andrew Ng, formerly Baidu's Chief Scientist, has called AI the new electricity.¹⁰ We believe its ultimate potential is analogous to the internet. Much like the internet, deep learning will have broad and deep ramifications. Specifically:

- | Like the internet, ***deep learning is relevant for every industry***, not just for the computing industry. Deep learning software is a core differentiator for retail, automotive, health care, agriculture, defense, and many other industries.
- | Like the internet, ***deep learning endows computers with previously unimaginable capabilities***. The internet made it possible to search for information, communicate via social media, and shop online. Deep learning enables computers to understand photos, translate language, diagnose diseases, forecast crops, and drive cars.
- | Like the internet, ***deep learning is an open technology*** that anyone can use to build new applications. While deep learning used to rely on computer scientists and specialized hardware, it now can be done using a laptop and a few lines of Python code.
- | Like the internet, ***deep learning should be highly disruptive***, perhaps far more disruptive than the internet. The internet has been disruptive to media, advertising, retail, and enterprise software. Deep learning could change the manufacturing, automotive, health care, and finance industries dramatically. Interesting to note, those industries have been sheltered to some extent from technological disruption to date.

While sizing new opportunities by their impact on established industries is typical, this paper's angle is from the top down, a more instructive approach when evaluating foundational technologies. In the early days of the internet, for example, analysts anticipated ecommerce and media portals, but could not fathom the existence and power of cloud computing, social media, or sharing economy platforms such as AirBnB. In the same way, deep learning could create new use cases and companies that are impossible to foresee today.

DEEP LEARNING COULD BE A \$17 TRILLION INDUSTRY IN 20 YEARS

If it were to approach the impact of the internet, deep learning could create \$17 trillion in market capitalization during the next two decades. In making this estimate, we identified companies that emerged and flourished because of the internet, while excluding names like Apple, Qualcomm, and Microsoft that benefited from the internet but were founded on different core technologies. Based on this criteria, 12 stocks in the US emerged as direct beneficiaries, representing 8.6% of the S&P 500, or \$1.7 trillion in market capitalization, as shown below.

¹⁰ "Andrew Ng: Why AI is the new electricity," Stanford News, 2017 <http://news.stanford.edu/thedish/2017/03/14/andrew-ng-why-ai-is-the-new-electricity/>



FIGURE 6
Market Capitalization: Internet vs. Deep Learning

INTERNET COMPANY	MARKET CAP IN \$ BILLION	DEEP LEARNING	MARKET CAP IN \$ BILLION
Alphabet	\$541	Global Market Capitalization 2016	\$69,948
Amazon	\$401	Historical Growth Rate	6.9%
Facebook	\$370	Global Market Capitalization 2036	\$264,000
Cisco	\$150	Deep Learning Share*	6.6%
Netflix	\$52		
Salesforce	\$51		
Yahoo	\$40		
Ebay	\$34		
Akamai	\$12		
Juniper Networks	\$11		
Verisign	\$9		
F5 Networks	\$9		
Total	\$1,680	Projected Deep Learning Market Cap Creation By 2036	\$17,000
S&P 500 Market Cap	\$19,622		
New Market Cap Creation from the Internet	8.6%		

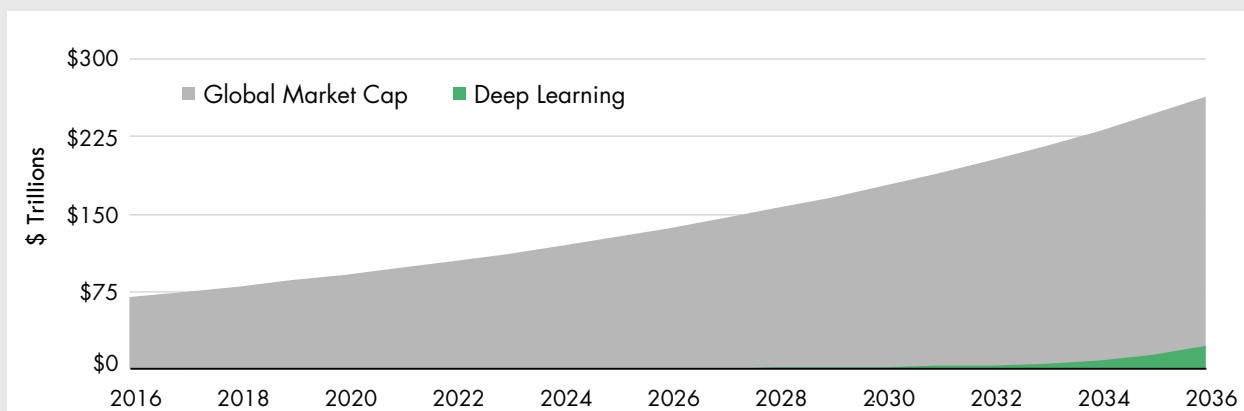
*Estimate uses new market cap creation caused by the internet as a proxy. 8.6% is scaled to 6.6% based on lower share of internet born companies globally versus the US.

Source: ARK Investment Management LLC, Global Federation of Exchanges, MSCI

When scaled to global equity markets, we believe the share of stocks born out of the internet is approximately 6.6%. As shown above, if the global equity market were to continue to appreciate at 6.9% on average per year, its market capitalization would approximate to \$264 trillion two decades from now. If deep learning-based firms were to grow to the same share of stocks born out of the internet over roughly the same time frame, that would imply a potential \$17 trillion in market capitalization by the mid 2030s.

In our view, this estimate is quite conservative, for two reasons. First, it counts only market capitalization creation by new companies; however, deep learning could generate as much value for existing enterprises as for new ones. Second, this estimate assumes deep learning grows at the same rate as the global equity market even though technology is improving at an accelerating rate: deep learning should grow much faster during its first twenty years than will the global equity market.

FIGURE 7
Deep Learning Market Cap Creation



Source: ARK Investment Management LLC



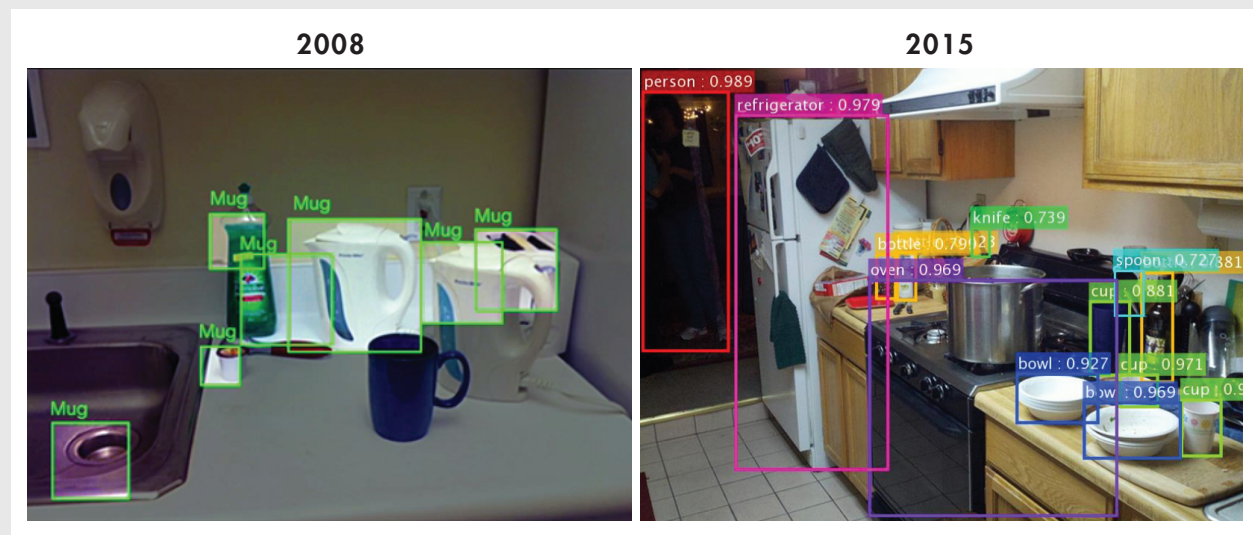
DEEP LEARNING USE CASES

COMPUTER VISION

Deep learning's most impressive achievements to date have been in the field of computer vision. Prior to modern deep learning, computer vision was confined primarily to research, with limited commercial development. Early deployments focused on simple areas or narrow data sets such as optical character recognition, check reading, and diagnostic mammography as opposed to photographs.

Today, deep learning can detect objects in natural photographs with human level accuracy. The images below illustrate the rate of progress during the past eight years. In 2008, computer vision programs struggled to identify a mug in a common kitchen scene. By 2015, Microsoft's "ResNet" deep learning program¹¹ was able to identify various objects in the kitchen, including a dimly lit person in the background, accurately.

FIGURE 8
Computer Vision: 2008 vs. 2015



Source: Stanford, Microsoft

Like human vision, computer vision is a general ability with myriad use cases. The ability to interpret raw photos and videos could prove useful in retail, agriculture, medical imaging, virtual/augmented reality, drones, robotics, autonomous driving, among many other use cases.

This section focuses on consumer applications of computer vision. We will cover other verticals in subsequent sections.

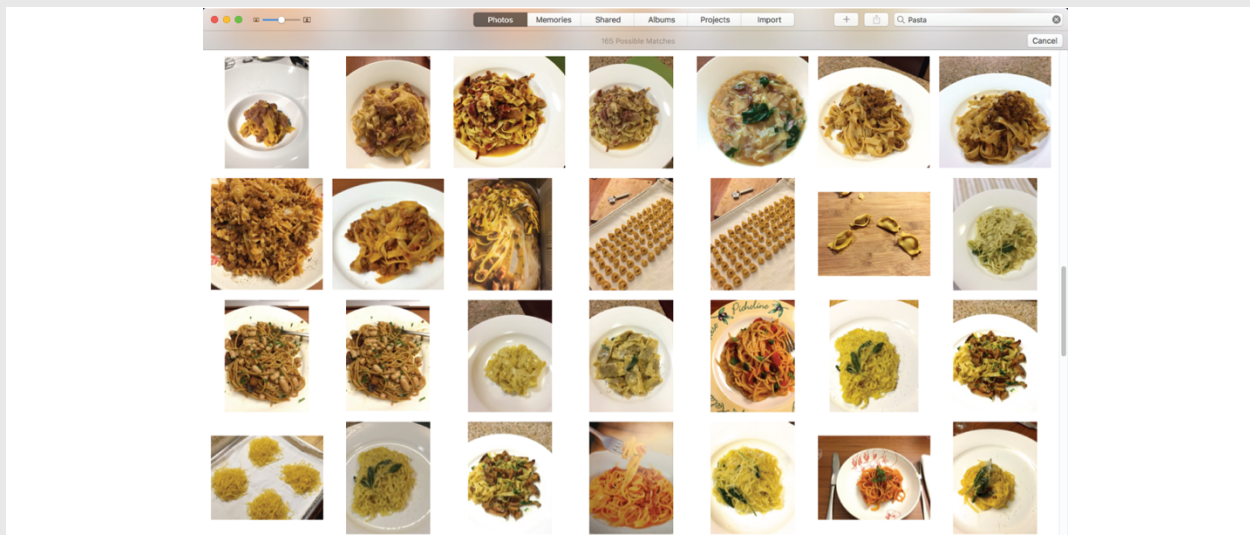
¹¹ Baidu Research, https://www.slideshare.net/ExtractConf/lessons-from-2mm-machine-learning-models?next_slideshow=1



IMAGE SEARCH

Search companies such as Google, Baidu, and Microsoft use deep learning to classify images, improving search accuracy dramatically compared to previous methods that relied on inspecting HTML tags and text. Thanks to recent advancements, deep learning now works on PC and mobile devices. Google Photos and Apple Photos, for example, let users search their camera rolls for keywords such as “cat,” “selfie,” “beach,” or “pasta” without any prior organization or keyword tagging, as shown below.

FIGURE 9
Image Search For “Pasta”



Source: Apple Photos

AUTO-FACE-TAGGING

Facebook users upload over 1 billion photos to its platform every day.¹² Using deep learning, Facebook is able to detect faces and auto-tag friends when a new photo is shared. Facebook’s Moments app uses the same technology to suggest the sharing of group photos through private direct messaging. Facebook also uses machine vision to provide audio descriptions of photos, benefiting people with visual disabilities.

COMPUTER VISION AS A SERVICE

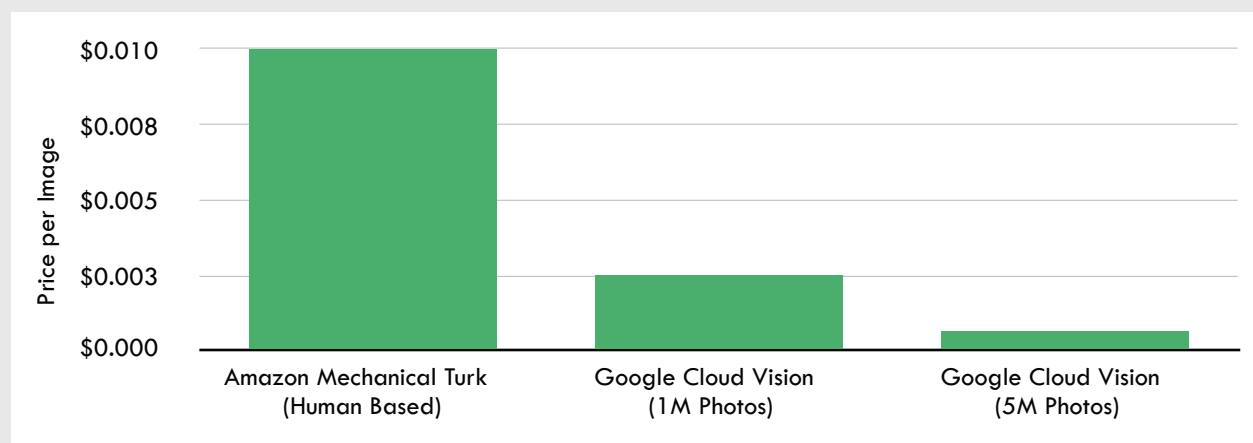
Few companies have the expertise and computing infrastructure to train and deploy machine vision products. To address the enterprise market, Google, Microsoft, Clarifai, IBM, and others have introduced computer vision-as-a-service capabilities. These services allow companies to offload

12 “Accelerating Understanding: Deep Learning, Intelligent Applications, and GPUs” <https://www.youtube.com/watch?v=Qk4SqF9FT-M> (47min)



image processing to the cloud for a fee per image. Services include classification, optical character recognition, facial detection, and logo detection. Compared to manual image reading by a service like Amazon's crowdsourced Mechanical Turk, these cloud based APIs are roughly an order of magnitude cheaper, as shown below. Yet, based on an analysis of the underlying hardware and operating costs, ARK believes cloud based machine vision APIs are vastly overpriced. We would not be surprised if prices come down by several more orders of magnitude during the next few years.

FIGURE 10
Cost for Optical Character Recognition



Source: ARK Investment Management LLC

MACHINE VISION APPLICATIONS ARE BEGINNING TO PROLIFERATE

- | NVIDIA, Mobileye, and Movidius build specialized chips for computer vision. NVIDIA and Mobileye chips enable cars to perceive their surroundings through visual and non-visual sensors. Movidius chips help DJI's Phantom 4 drones avoid obstacles.¹³
- | Snapchat uses computer vision to detect faces and overlay them with playful filters and lenses. Snapchat sells "sponsored lenses" to advertisers for up to \$750,000 per lens per day.¹⁴
- | Descartes Labs uses deep learning to process satellite imagery for agricultural forecasts.¹⁵ It processes over 5 terabytes of new data every day and references a library of 3 petabytes of archival satellite images. By using real time satellite imagery and weather models, Descartes Labs provides highly accurate weekly forecasts of US corn production compared to monthly forecasts provided by the US Department of Agriculture.
- | A number of companies are using deep learning for automated medical imaging diagnosis. See the section: Radiology

¹³ <https://www.movidius.com/news/movidius-and-dji-bring-vision-based-autonomy-to-dji-phantom-4>

¹⁴ "Snapchat Is Asking Brands for \$750,000 to Advertise and Won't Budge", AdWeek, 1/14/2015 <http://www.adweek.com/digital/snapchat-asks-brands-750000-advertise-and-wont-budge-162359/>

¹⁵ "Deep-Learning, Image-Analysis Startup Descartes Labs Raises \$3.3M After Spinning Out Of Los Alamos National Lab," TechCrunch, 5/1/2015; <https://techcrunch.com/2015/05/01/deep-learning-image-analysis-startup-descartes-labs-raises-3-3m-after-spinning-out-of-los-alamos-national-labs/>



AI ASSISTANTS & BOTS

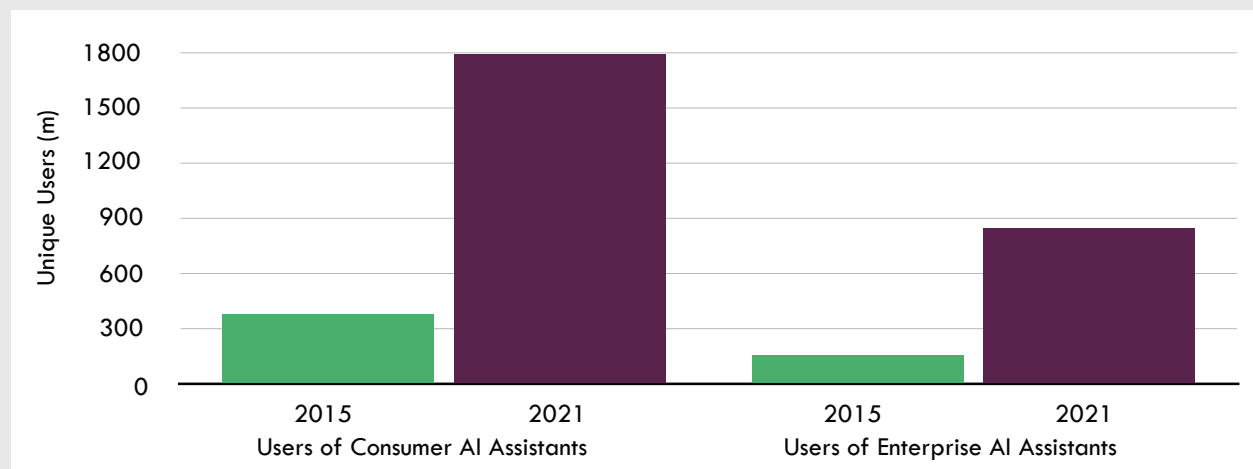
AI assistants - computer programs capable of human level speech and comprehension - have captivated our imagination since the debut of HAL 9000 in *2001: A Space Odyssey*. Needless to say, computer programs that can converse with humans, understand needs, and help with tasks would be a boon to the quality of life and to global productivity. Until recently, such breakthroughs were confined to the realm of science fiction. Now, however, thanks to deep learning, AI assistants are making rapid progress.

AI assistants became mainstream when Apple launched Siri in October 2011.¹⁶ Google Now followed in 2012, and Microsoft Cortana and Amazon Echo in 2014.¹⁷ Today, many other companies are racing to build AI assistants and chat bots that some believe will be larger than the app economy.

Research firm Tractica¹⁸ estimates that the use of consumer AI assistants worldwide will grow 25% per year on average, from 390 million users in 2015 to 1.8 billion, by the end of 2021. Users of enterprise AI assistants are expected to rise at a 33% annualized rate, from 155 million to 843 million, during the same time period, as shown below.

AI assistants generally fall into two camps: voice based and text based. Voice based interfaces like

FIGURE 11
Unique Users of AI Assistants Worldwide



Source: Tractica

Siri, Google Now, Cortana, and Echo have seen solid adoption and usage. Text based AI assistants are nascent and have yet to achieve mainstream adoption.

¹⁶ <https://www.apple.com/newsroom/2011/10/04Apple-Launches-iPhone-4S-iOS-5-iCloud/>

¹⁷ Google, Microsoft, and Amazon product launches

¹⁸ "The Virtual Digital Assistant Market Will Reach \$15.8 Billion Worldwide by 2021," Tractica, 8/3/2016 <https://www.tractica.com/newsroom/press-releases/the-virtual-digital-assistant-market-will-reach-15-8-billion-worldwide-by-2021/>



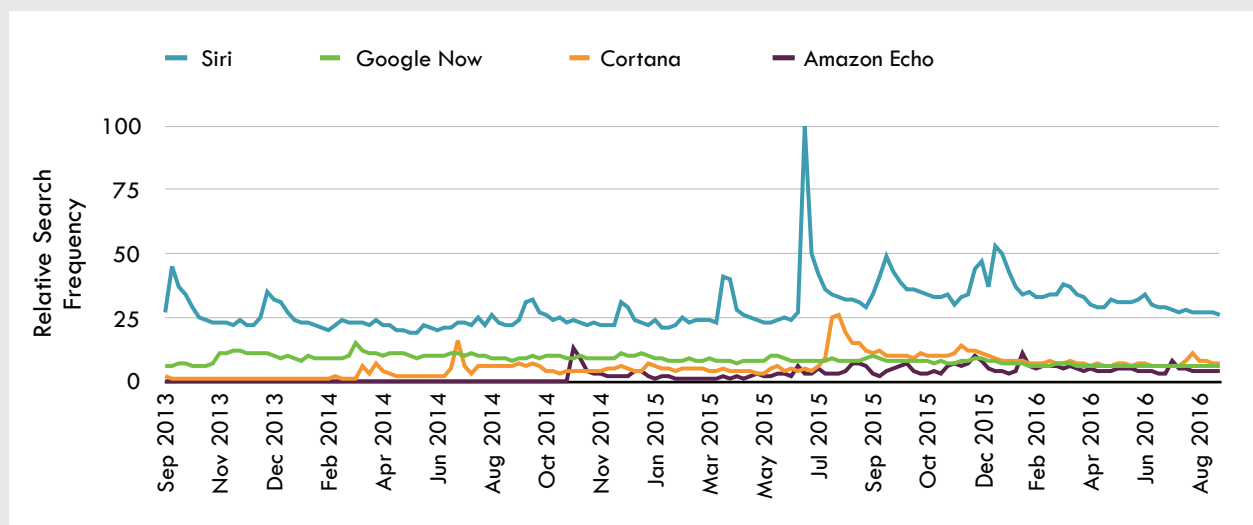
SIRI

Launched in 2011, Siri became a key selling point for Apple devices, despite its modest skillset and spotty voice recognition. Created before the deep learning boon, it relied on statistical techniques like Hidden Markov Models. Early versions of Siri used Nuance's voice recognition technology.

In 2014, Apple migrated many of Siri's core capabilities to deep learning models and achieved a two-fold reduction in error rate.¹⁹ Deep learning also helped Siri speak in a more natural voice and handle more complex queries.

Apple opened Siri up to developers with iOS10, allowing it to integrate with 3rd party services. According to Google Trends, Siri is the most popular personal assistant by far, as shown below. Engagement with Siri also is growing quickly, as Apple has reported a doubling in requests made on average per week, from one billion in 2015 to two billion last year.²⁰

FIGURE 12
Google Search Trends for AI Assistants



Source: Google Trends

GOOGLE NOW / GOOGLE ASSISTANT

Launched in 2012, Google Now is a general purpose assistant focused on predictive results. To deliver personalized results, Google Now leverages Google's search capabilities, its Knowledge Graph database with facts and common knowledge, and personal information such as emails stored in Gmail. For example, Google Now can learn a user's commute hours and provide traffic information proactively. Other proactive notifications include weather alerts, package deliveries, and flight information.

19 "The iBrain Is Here And It's Already Inside Your Phone," Backchannel, 8/24/2016 <https://backchannel.com/an-exclusive-look-at-how-ai-and-machine-learning-work-at-apple-8dbfb131932b#.hk4va5mf3>

20 Apple 2016 Worldwide Developer Conference Keynote

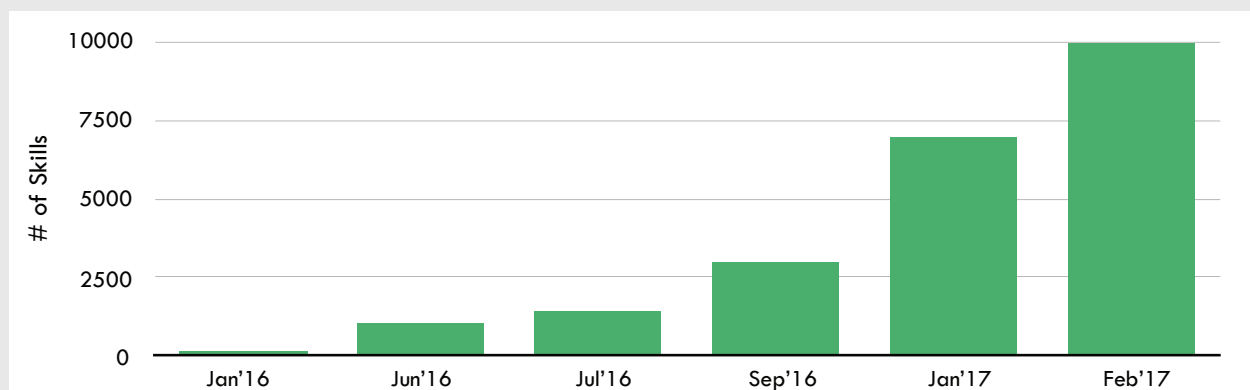


In 2015, Google added a feature called Google Now On Tap,²¹ providing context-aware, in-app assistance. For example, upon activation in a text about movie options, On Tap brings up information about the movie, actors, and show times. In 2016, Google rebranded Google Now as Google Assistant,²² which could become its user-facing AI interface. Over time, that part of Google's ecosystem could become as important to Google's engagement as its core product, Search.

AMAZON ECHO AND ALEXA

While Siri and Google Assistant are tethered to the smartphone, Amazon's Echo is a dedicated device for home use that costs \$179. It is plugged in and always on, picking up voices from across the room through long range microphones. Despite mixed reviews, the Echo has developed a strong fan base. According to CIRP,²³ Amazon has sold 3 million Echo units to date and, according to The Information, it is targeting 10 million units in 2017.²⁴

FIGURE 13
Amazon Alexa "Skills"



Source: Amazon

Powering the Echo is Alexa, Amazon's voice-based assistant. Unlike Siri which Apple has limited to its hardware ecosystem, Alexa is available on almost any device, and thanks to a set of open, easy to use Application Programming Interfaces, or APIs, has vaulted into first place among AI assistants. Alexa has been integrated into a host of devices such as fridges, lights, speakers, smartwatches, and cars, while its skills have scaled to more than 10,000 as of February 2017.²⁵ Through an Echo, Alexa can call an Uber, play songs from Spotify, control smart home devices, and of course order products from Amazon.²⁶

21 Google I/O 2015 Keynote

22 "A Personal Google, Just For You," Google, 10/4/2016 <https://blog.google/products/assistant/personal-google-just-you/>

23 "Amazon Echo sales reach 3M units as consumer awareness grows, research firm says," GeekWire, 4/6/2016 <http://www.geekwire.com/2016/report-amazon-sold-3-million-echo-smart-speakers-awareness-grows/>

24 "Amazon's High Hopes for Echo Sales", The Information, 6/15/2016, <https://www.theinformation.com/amazons-high-hopes-for-echo-sales>

25 "Amazon's Alexa smart assistant reaches 10,000 skills, up from just 1K in June," GeekWire, 2/23/17 <http://www.geekwire.com/2017/amazons-alexa-smart-assistant-reaches-10000-skills-just-1k-june/>

26 Amazon Echo product page https://www.amazon.com/dp/product/B00X4WHP5E/ref=cp_aucc_ods



While Alexa has no native presence on smartphones and thus fragments the user experience inside vs. outside of the home, Alexa's early acceptance suggests that it does have strong potential. In the most optimistic scenario, Alexa could become the de facto operating system for voice based computing, making it the primary interface for home automation, ambient computing, and autonomous cars.

MESSAGING BOTS

As the smartphone market has matured in recent years, developers and investors have intensified their search for the next big platform. Messaging bots – AI assistants that operate primarily through text – could be the answer. Among the reasons for the focus on messaging bots are 1) the success of WeChat in China, 2) the growth in users of and time spent in messaging apps, and 3) the deep learning-related improvements in natural language processing.

Like AI, messaging bots can be narrow or general. Narrow messaging bots perform very specific tasks, such as replying to an email, while in theory general messaging bots can perform any task, much like a personal assistant.

Narrow messaging bots already have been deployed successfully in real applications. The AI assistant Amy by x.ai, for example, can schedule meetings for individuals who do not have access to each other's calendars.²⁷ As shown below, Amy reads the host's calendar and suggests open time slots by writing and sending an email to those invited. Upon receiving a reply, she can read and understand the email, schedule a meeting, or suggest new time slots in response to conflicts.

Google's Smart Reply can read an email and propose three responses.²⁸ According to Google, more than 10% of its Inbox app's email responses are sent via Smart Reply.

A natural home for messaging bots is inside messaging apps. In 2016, Microsoft, Facebook, and Kik all launched chatbot platforms for their respective messaging apps, Skype, Messenger, and Kik chat. So far, more than 11,000 bots have launched on Messenger²⁹ and more than 20,000 on Kik.³⁰ These bots have a range of functions, from ordering flowers to checking the weather, and from recommending books to serving as a personal trainer.

Despite the strong developer interest and backing from popular news portals like CNN on Messenger, the first generation of bots has received a plethora of negative reviews. Not only has the personalization been rudimentary, but chatbots have proved less convenient than a traditional app or website, often involving many steps for simple tasks such as checking the weather. Kik CEO Ted Livingston has

27 "Rise of the bots: X.ai raises \$23m more for Amy, a bot that arranges appointments," TechCrunch, 4/7/2016 <https://techcrunch.com/2016/04/07/rise-of-the-bots-x-ai-raises-23m-more-for-amy-a-bot-that-arranges-appointments/>

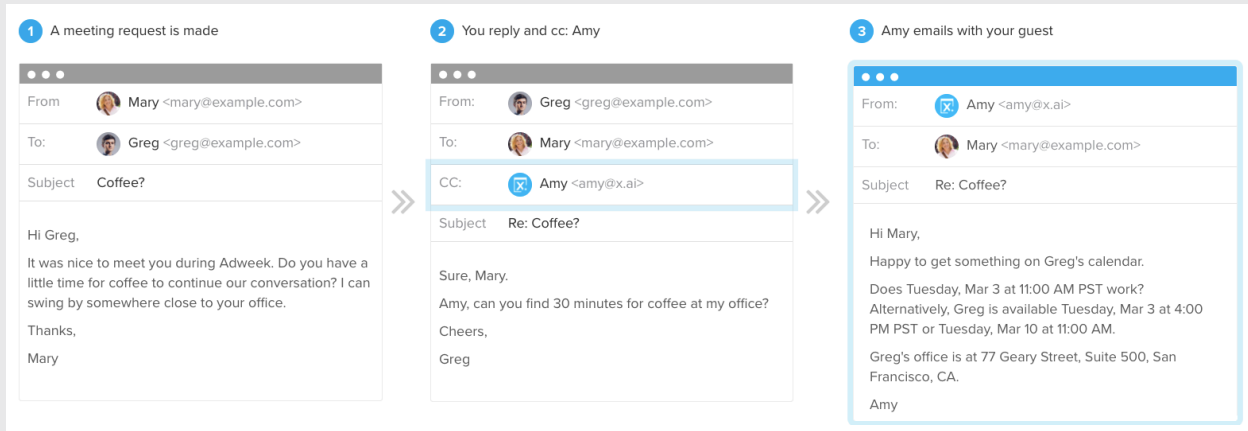
28 Google Blog, 11/3/2015, <https://blog.google/products/gmail/computer-respond-to-this-email/>

29 "11,000 BOTS NOW LIVE ON FACEBOOK MESSENGER AFTER JUST THREE MONTHS," Digital Trends, 7/1/2017 <http://www.digitaltrends.com/computing/facebook-messenger-10000-bots/>

30 "Kik hits 20,000 bot apps four months after store launch," Wired, 8/3/16, <http://www.wired.co.uk/article/kik-bot-store-statistics-messaging-app>



FIGURE 14
Messaging Bots



Source: X.ai

conceded that “so far, there has been no killer bot.”³¹

ROBOTICS

Deep learning’s ability to automate menial processes and boost productivity should have a profound impact on the robotics industry. Despite their widespread use in manufacturing, robots are expensive and difficult to program. For most businesses, robots are not useful, at least not yet. In 2015, global unit sales of industrial robots were only ~250,000, roughly tenfold the number of mainframe computers at their peak.³² By comparison, last year server and PC unit sales totaled ~10 million³³ and ~300 million,³⁴ respectively. Clearly, robotics is at a nascent stage, calling for dramatic improvements in both cost and ease of use before it proliferates.

Cost improvements are well underway. ARK estimates that the cost of industrial robots, which currently are roughly \$100,000,³⁵ will fall by half over the next ten years. Concurrently, a new breed of robots designed for co-operative use with humans will cost on the order of \$30,000.³⁶ Retail assistant robots like SoftBank’s Pepper cost about \$10,000 when service fees are included. Leveraging components from the consumer electronics industry such as cameras, processors, and sensors should drive costs closer to those of consumer products.

31 “Bots are Better Without Conversation,” Medium, 2016 <https://medium.com/@tedlivingston/bots-are-better-without-conversation-fcf9e7634fc4>

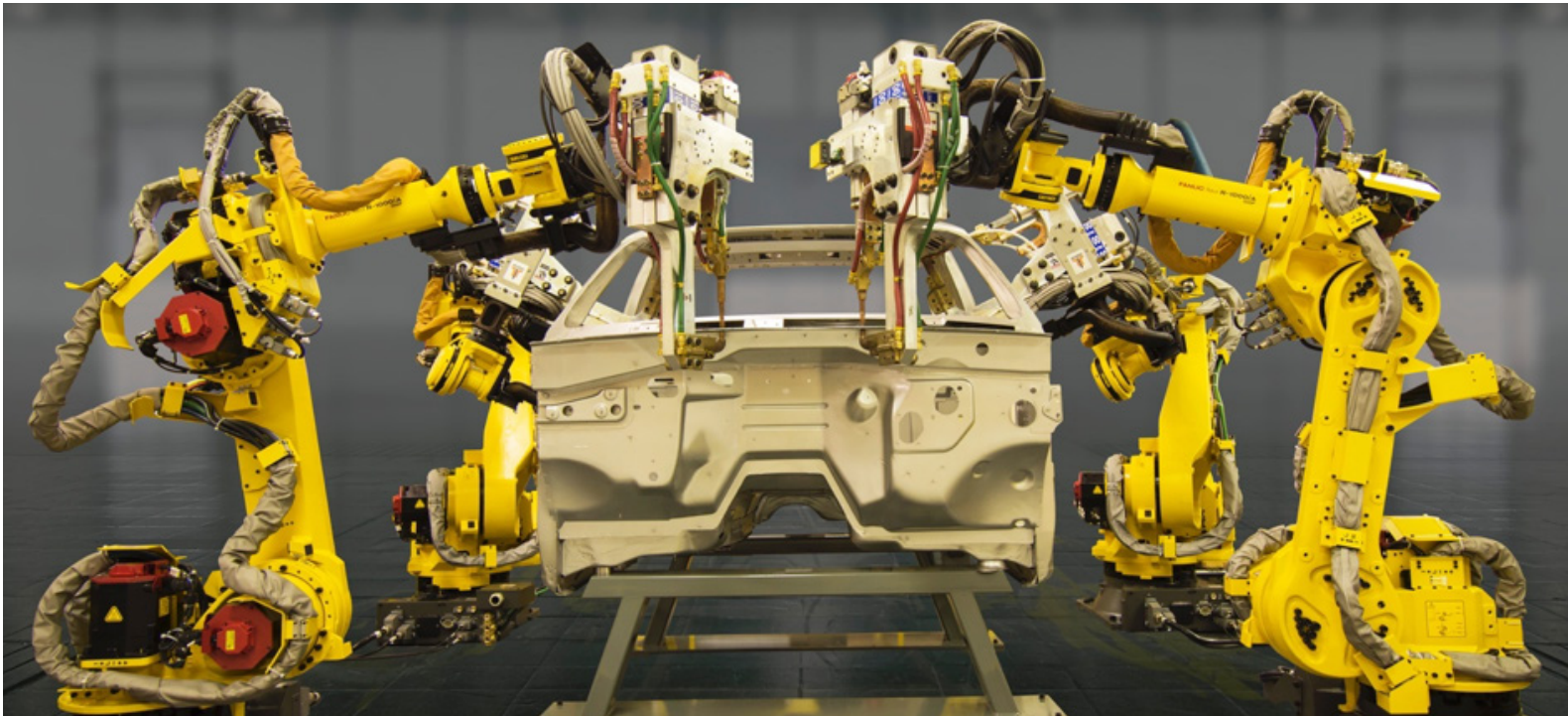
32 “Mobile Ate The World,” March 2016, <http://ben-evans.com/benedictevans/2016/3/29/presentation-mobile-ate-the-world>

33 “IDC Worldwide Quarterly Server Tracker,” 3/9/2016 <https://www.idc.com/getdoc.jsp?containerId=prUS41076116>

34 “IDC Worldwide Quarterly PC Tracker,” 1/12/2016 <https://www.idc.com/getdoc.jsp?containerId=prUS40909316>

35 “How much do industrial robots cost?” RobotWorx, <https://www.robots.com/faq/show/how-much-do-industrial-robots-cost>

36 “Robots rub shoulders with human buddies,” Financial Times, 3/19/2015 <https://www.ft.com/content/ed-7be188-cd50-11e4-a15a-00144feab7de#axzz46Ygbi71C>



Source: Fanuc

The more difficult obstacle to overcome is ease of use. Industrial robots are not designed from a user-centric point of view. They require precise programming using industrial control systems in which each task must be broken down into a series of movements in six dimensions.³⁷ New tasks must be programmed explicitly: the robot has no ability to learn from experience and generalize to new tasks. These limitations have restricted the market for robots to those industrial applications where tasks are predictable and well defined.

Deep learning can transform robots into learning machines. Instead of precise programming, robots learn from a combination of data and experience, allowing them to take on a wide variety of tasks. For example, a warehouse robot capable of picking any item from a shelf and placing it into a box would be highly desirable for many businesses. Yet, until recently, developers couldn't program a robot to recognize and grasp objects that come in an infinite variety of shapes and sizes.

³⁷ "How Robots Are Programmed," Global Robots LTD, <http://www.globalrobots.com/guide.aspx?guide=3>



Source: Beep bop: Rethink Robotics' Baxter model

Deep learning is making the once impossible, possible. In the Amazon Picking Challenge, a robotics competition, robots attempt to pick random items from a shelf and place them in a box. In 2015, the winning robot was able to pick 30 items per hour.³⁸ In 2016, picking performance more than tripled to 100 items per hour,³⁹ with the top two teams using deep learning as the core algorithm for vision and grasp.⁴⁰ Were improvements to continue at that rate, robotics-based item picking would exceed human picking in two years.

Deep learning also is a far more effective way to program robots for simpler, more predictable tasks. According to Preferred Networks, a robotics company based on deep learning, a human programmer must work for several days to teach a robot a new task. By comparison, using deep learning, a robot learns the same task in about eight hours.⁴¹ When eight robots jointly learn the task, training time drops to one hour. Thus deep learning provides a 5x increase in training speed and, via parallel learning, offers firms the ability to compound that improvement as they devote more hardware to the task.

Industry heavyweights have taken notice. In 2015, Fanuc, a leading manufacturer of industrial robots, acquired a 6% stake in Preferred Networks;⁴² and shipped deep learning enabled robots in 2016. ABB Robotics, a Swiss company, has invested in Vicarious, an AI startup with deep learning expertise.

As the costs decline, deep learning has the potential to expand the addressable market for robots dramatically. Today's industrial robots are confined to safety cages and limited to highly programmed, repetitive work. Deep learning robots should be able to operate in a wide variety of settings, learn new

38 "Overview of Amazon Picking Challenge 2015," Amazon, <https://www.amazonrobotics.com/site/binaries/content/assets/amazonrobotics/pdfs/2015-apc-summary.pdf>

39 "Team Delft Wins Amazon Picking Challenge," IEEE Spectrum, 7/5/2016 <http://spectrum.ieee.org/automaton/robotics/industrial-robots/team-delft-wins-amazon-picking-challenge>

40 "Team Delft's Robot Winner of the Amazon Picking Challenge 2016," Delft University of Technology, <https://arxiv.org/abs/1610.05514>

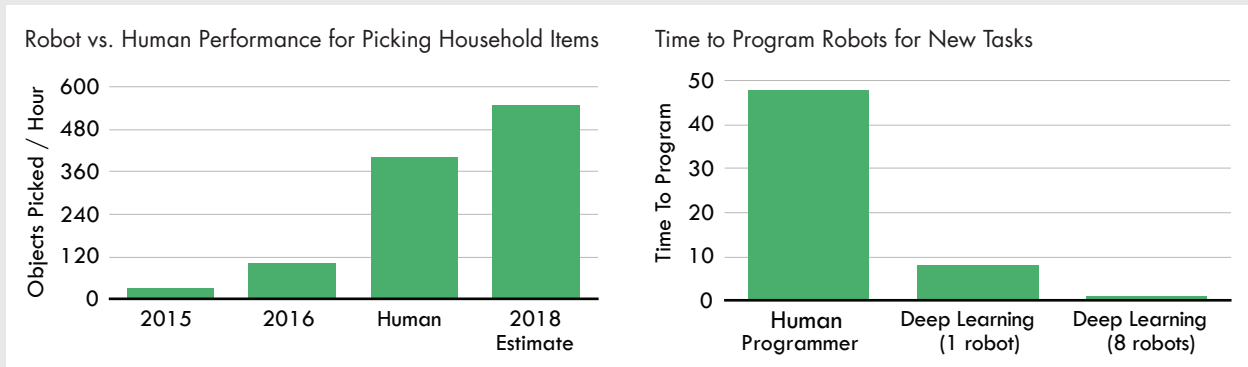
41 "Zero to Expert in Eight Hours: These Robots Can Learn For Themselves," Bloomberg, 3/12/2015, <https://www.bloomberg.com/news/articles/2015-12-03/zero-to-expert-in-eight-hours-these-robots-can-learn-for-themselves>

42 "Fanuc Invests in Startup as Robot Intelligence Race Heats Up," Bloomberg, 8/21/2015, <https://www.bloomberg.com/news/articles/2015-08-21/fanuc-invests-in-startup-as-robot-intelligence-race-heats-up>



tasks on the job, and work safely alongside humans. Consequently, the robotics market could open up to small and medium sized businesses, retail, agricultural, and home applications.

FIGURE 15
Performance and Programming



Source: Amazon, Preferred Networks

ARK's research shows that industrial robots will remain the workhorse of high capacity manufacturing, but their unit volumes will be dwarfed by the newer, nimbler robots, just as mainframe computers were swamped by workstation units, followed by PCs and then smartphones. As a result, robot unit volume shipments could soar 10 to 100 fold. Like drones, many of the new robots will bear little resemblance to the robots that dominate the market today. Smartphones don't look much like mainframes either.

RADIOLOGY

Deep learning is making rapid advances in diagnostic radiology. ARK estimates that the total global addressable market for computer aided diagnostics software could be worth \$16 billion, as shown below. From revenues of \$1 billion today, the growth in medical software companies and imaging device manufacturers could average 20-35% growth per year as deep learning enhances their productivity and creates new products and services during the next ten to fifteen years.

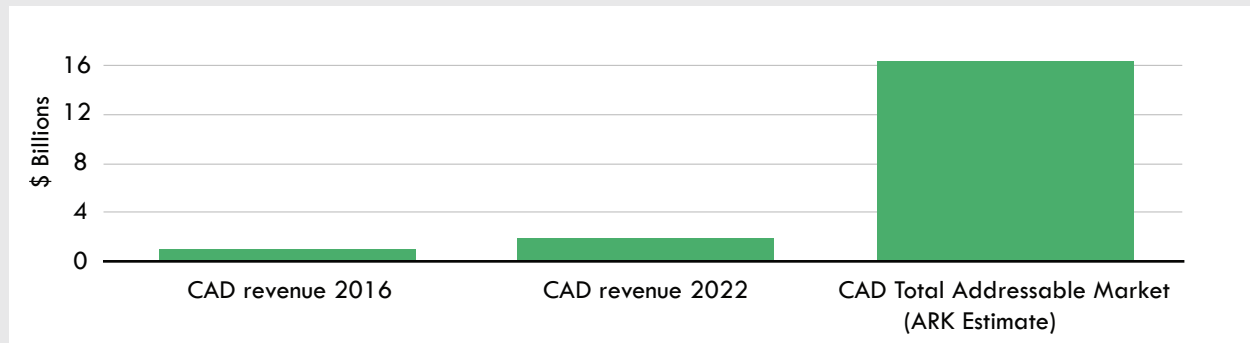
Diagnostic radiology is essential to modern health care; yet, the visual interpretation of medical images is a laborious and error prone process. Historically the average diagnosis error rate among radiologists is around 30%, according to studies⁴³ dating from 1949 to 1992. Because of rudimentary technology, they miss lung cancer nodules routinely, especially at earlier stages of development and miss or misdiagnose 8-10% of bone fractures. Initially, radiologists miss roughly two-thirds of breast cancers in mammograms that are visible in retrospective reviews.⁴⁴

43 "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential," NCBI, 3/8/2007
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1955762/>

44 Computer-aided diagnosis: Breast Imaging," RAI Labs, <http://deckard.mc.duke.edu/breastcad.html>



FIGURE 16
Global Computer Aided Diagnostic Market



Source: Grand View Research, ARK Investment Management LLC

Human error is understandable. Radiological images include overlapping tissue, bones, and organs, making it difficult to identify problem areas accurately. Analyzing up to 200 cases per day,⁴⁵ emergency room radiologists need more powerful tools to make accurate diagnoses in short periods of time.

Intelligent software powered by deep learning has the potential to change the status quo. Early results are promising: the latest deep learning systems already outperform radiologists and existing algorithms in a variety of diagnostic tasks, as illustrated below.

- | Enlitic, a San Francisco based startup, says its deep learning based diagnostic system can detect lung cancer nodules 50% more accurately than a panel of radiologists when benchmarked in an NIH-funded lung image data set.⁴⁶
- | Enlitic's system detects extremity bone fractures, say on the wrist, with 97% accuracy compared to 85% accuracy of radiologists and 71% accuracy of previous computer vision algorithms.
- | Harvard Medical School built a deep learning system that detects breast cancer with 97% accuracy compared to 96% accuracy of a radiologist. When the radiologist was aided by the diagnostic system, accuracy improved to 99%.⁴⁷
- | McMaster University's deep learning system achieves 98-99% accuracy in detecting Alzheimer's disease⁴⁸ in magnetic resonance images. Previous computer vision algorithms achieve only 84% accuracy.

45 "IBM Research Accelerating Discovery: Medical Image Analytics," 10/10/2013 <https://www.youtube.com/watch?v=0i11VC-NacAE>

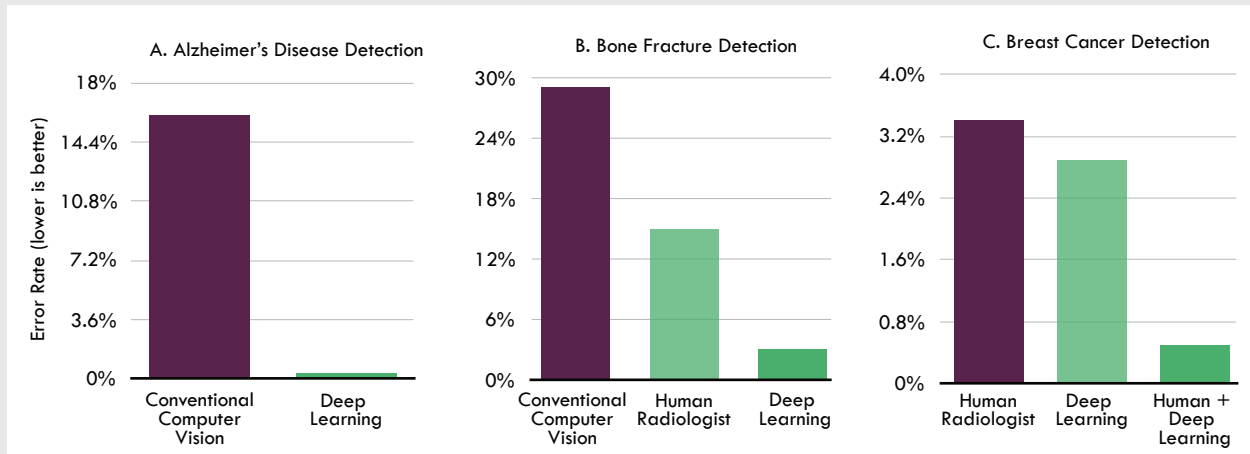
46 "Enlitic and Capitol Health Announce Global Partnership," 10/27/2015 <http://www.enlitic.com/press-release-10272015.html>

47 "Deep Learning Drops Error Rate for Breast Cancer Diagnoses by 85%," NVIDIA Blog, 9/19/2016, <https://blogs.nvidia.com/blog/2016/09/19/deep-learning-breast-cancer-diagnosis/>

48 "DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI," Sarraf et al. 8/22/2016 <http://biorxiv.org/content/early/2016/08/21/070441>



FIGURE 17
Deep Learning Based CAD | A. Alzheimer's Disease Det. B. Bone Fracture Det. C. Breast Cancer Det.



Source: McMaster University, Enlitic, Harvard Medical School, ARK Investment Management LLC

Achieved in a relatively short amount of time, these super-human results have the potential to shake up the traditional computer aided diagnosis (CAD) market. Currently, the CAD market is dominated by companies such as Siemens, Philips, Hologic, and iCad. Global revenues totaled roughly \$1 billion in 2016 and, according to Grand View Research, will grow at a compounded annual rate of 11% to \$1.9B by 2022.⁴⁹ If deep learning based systems can provide better-than human accuracy, the economic benefit could incentivize clinics to make CAD a mandatory purchase decision.

Early diagnosis is key to successful treatment. Each year more than 2 million people worldwide die from lung and breast cancers according to Cancer Research UK.⁵⁰ If 10% of later stage cases could be caught at stage 1 with CAD, ARK estimates it would save 150,000 life years. Valuing human life at \$50,000 per year,⁵¹ breast or lung diagnoses at stage 1 would equate to \$7.6 billion of life value saved. Impacting a wide range of radiology problems from bone fractures to Alzheimer's disease, the value of deep learning would be orders of magnitude greater.

ARK estimates the market size for CAD software could reach \$16 billion. Our estimate is based on 34,000 radiologists⁵² in the US reviewing 20,000 cases⁵³ per year. Given that radiologists pay up to \$2 per case for existing Picture Archiving and Communication Systems (PACS),⁵⁴ a better-than human diagnostic system could be priced at \$10/case. Assuming full adoption, the US market alone would be worth \$6.8 billion. Normalized to incorporate the rest of the world's health care spending brings the total addressable market worldwide to \$16.3 billion.

49 "Computer Aided Detection Market Worth \$1.9 Billion By 2022," Grand View Research, 8/2016 <http://www.grandviewresearch.com/press-release/global-computer-aided-detection-market>

50 Cancer Research UK, <http://www.cancerresearchuk.org/health-professional/cancer-statistics>

51 Standard industry valuation of a human life year

52 "How Many Radiologists? It Depends on Who You Ask!" Health Policy Institute, 4/14/2015, <http://www.neimanhpi.org/commentary/how-many-radiologists-it-depends-on-who-you-ask/>

53 "Radiologists work more to find time for more play," Diagnostic Imaging, 11/2/2009 <http://www.diagnosticimaging.com/articles/radiologists-work-more-find-time-more-play>

54 Perfect Imaging PACS Pricing http://perfect-imaging.com/pacs_star/pacs_features

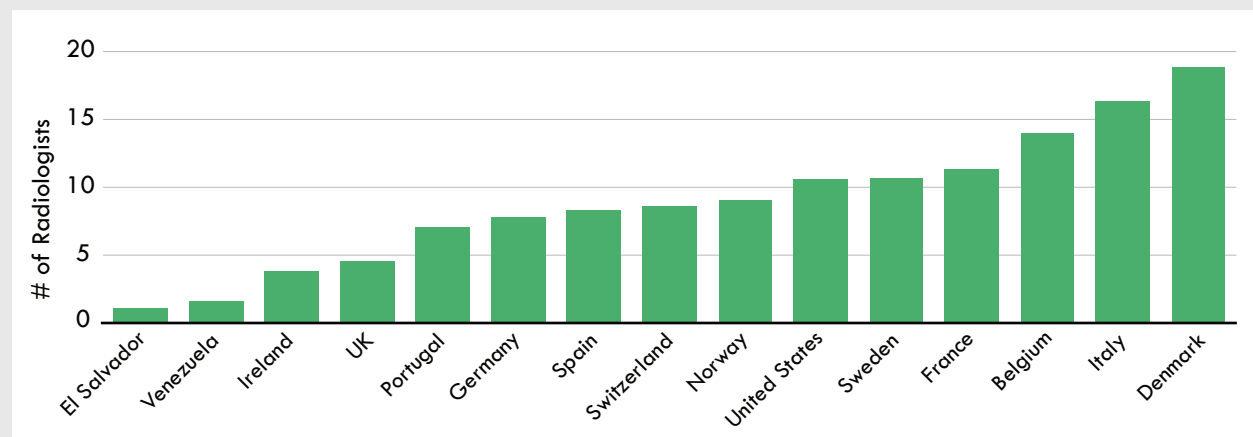


Both market incumbents and new entrants are well positioned to attack this market. The critical variables are deep learning expertise and access to large data sets. Enlitic has partnered with Capitol Health to provide deep learning based diagnostic software to the Australian market.⁵⁵ Samsung has released an ultra-sound machine with built-in deep learning based diagnostics.⁵⁶

IBM has been one of the most aggressive companies in tackling the radiology market with deep learning based solutions. In 2015, IBM acquired Merge Healthcare,⁵⁷ a radiology software provider whose 30 billion patient images were ideal for training Watson AI. The size of this dataset coupled with IBM's deep learning expertise could become a significant barrier to entry for startups with little access to data or incumbents with limited deep learning expertise. That said, Watson's radiology AI still is early stage: IBM says commercial deployment is unlikely for at least five years.⁵⁸

Will deep learning diagnostic systems replace radiologists? We think not. Radiologists spend just a third⁵⁹ of their day analyzing images, the balance in supervising studies, teaching staff, consulting with physicians, and caring for patients. Better CAD tools could improve their productivity, particularly reading more images with greater accuracy. In the developing world, where the availability of radiologists is a tenth the level of developed countries, as shown below, CAD programs could provide much greater access to diagnoses.

FIGURE 18
Radiologists per 100,000 People (2008)



Source: The Royal College of Radiologists, Harvey L. Neiman Health Policy Institute, Inter American College of Radiology.
El Salvador, Venezuela, United States numbers are from 2012

55 "Enlitic and Capitol Health Announce Global Partnership Leveraging Deep Learning to Enhance Physician Care for Millions of Patients," 10/27/2015 <http://www.prnewswire.com/news-releases/enlitic-and-capitol-health-announce-global-partnership-leveraging-deep-learning-to-enhance-physician-care-for-millions-of-patients-300166817.html>

56 "Samsung Applies Deep Learning Technology to Diagnostic Ultrasound Imaging," 4/21/2016 <http://www.samsung.com/global/business/healthcare/insights/news/samsung-applies-deep-learning-technology-to-diagnostic-ultrasound-imaging>

57 "IBM Crafts a Role for Artificial Intelligence in Medicine," The Wall Street Journal, 8/11/2015 <https://www.wsj.com/articles/ibm-crafts-a-role-for-artificial-intelligence-in-medicine-1439265840>

58 "Q&A: Tanveer Syeda-Mahmood on IBM's Avicenna software," 2/16/2016 <http://www.radiologybusiness.com/topics/technology-management/q-ibm-s-tanveer-syeda-mahmood-new-avicenna-software>

59 <http://www.auntminnie.com/index.aspx?sec=ser&sub=def&pag=dis&ItemID=105071>



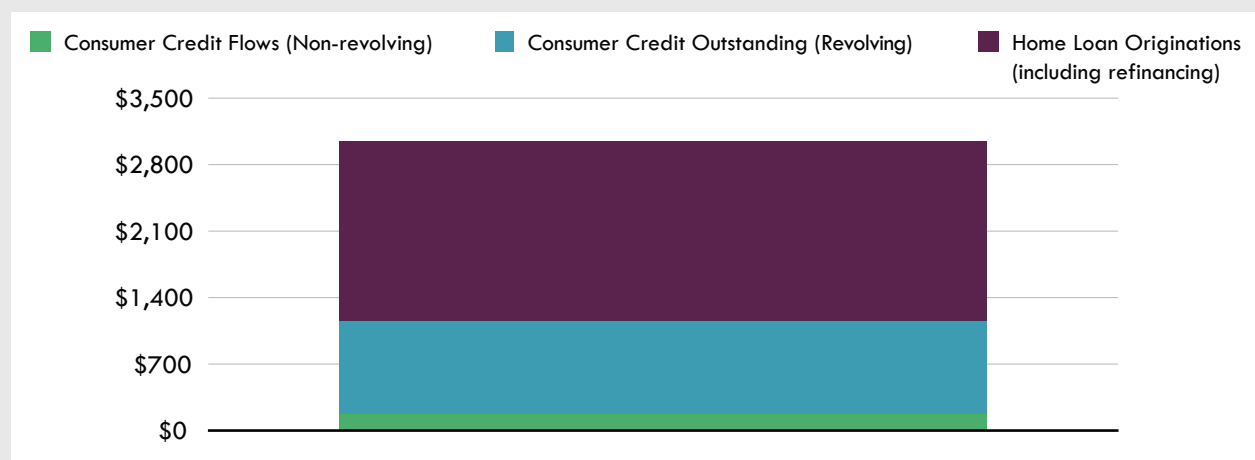
While deep learning based diagnostics offer great promise, deployment will be a gradual process, especially because CAD software is regulated by government health agencies in the US, EU, and China. After regulatory approval, integrating with hospital IT systems, training doctors, and obtaining insurance reimbursements will take more time and effort. Once these obstacles are overcome, radiology could become far more automated, accurate, and accessible.

CREDIT ASSESSMENT

With its vast asset base and large data sets, the financial services industry is in an excellent position to benefit from machine learning and deep learning. In this section, we include both classes of algorithms to reflect the latest benchmark results. These algorithms hold particular promise in assessing the relative risk of prospective borrowers more efficiently and accurately. If applied broadly to the US consumer credit industry, ARK estimates that machine learning could improve the lifetime profits associated with new and revolving loans each year by up to \$170 billion.

To calculate this estimate, we considered three categories of consumer debt in the US: home loans, revolving consumer credit, and non-revolving consumer credit. According to the Mortgage Bankers' Association home loans totaled \$1.9 trillion in 2016,⁶⁰ while revolving consumer credit, primarily credit cards, and non-revolving consumer credit were \$980 billion⁶¹ and \$180 billion, respectively. This \$3 trillion offered financial companies significant opportunities and risks associated with granting new loans requiring a credit rating or revolving loans to be re-rated, as shown below.

FIGURE 19
US Consumer Loans (New + Revolving)



Source: US Federal Reserve, Mortgage Banker's Association

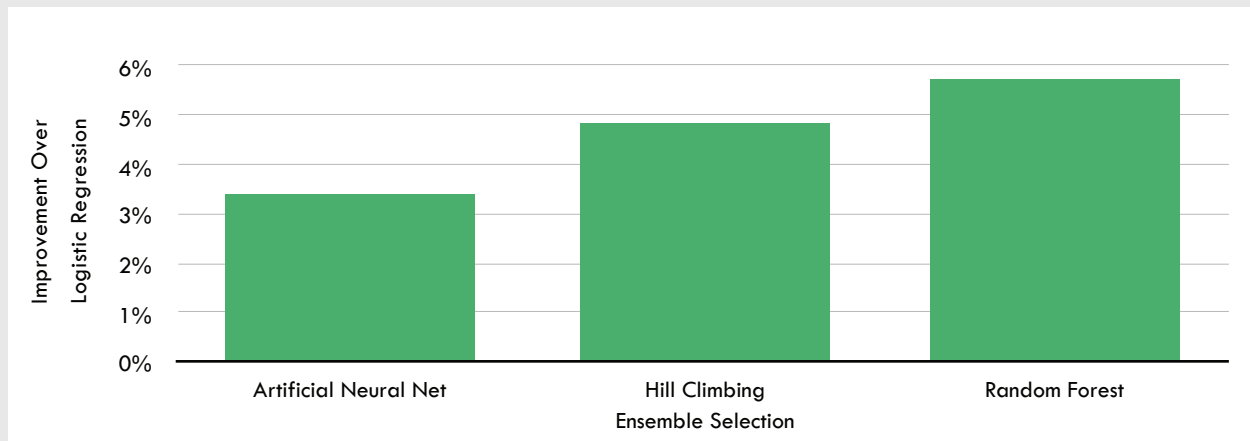
60 The Mortgage Banker's Association, <https://www.mba.org/news-research-and-resources/research-and-economics/forecasts-and-commentary>

61 Federal Reserve Consumer Credit G.19, <https://www.federalreserve.gov/releases/g19/current/>



How much could machine learning improve the rating accuracy of these loans? After reviewing 41 machine learning algorithms in a comprehensive survey⁶² published in 2015, the authors concluded that artificial intelligence outperformed the commonly used logistical regression classifiers significantly, as shown below.

FIGURE 20
Scorecard Profitability Improvement Using Machine Learning



Source: Lessman et al.

The authors reviewed both individual algorithms and ensemble algorithms. An ensemble is a collection of individual algorithms in one system to enhance accuracy and reduce bias. Artificial neural nets, or deep learning algorithms, improved the profitability of a loan by 3.4% compared to a logistic regression. The Hill-Climbing and Random Forest ensembles of algorithms were even more productive, 4.8% and 5.7%, respectively.

More accurate credit assessment improves profitability in two ways, by lowering lending to those more likely to default and by increasing lending to credit worthy borrowers. Applying the improvement rates noted above would impact the profitability of the \$3 trillion in US consumer debt issued in 2016 dramatically. Depending on the algorithm used, total lifetime profitability on these loans could increase by \$100-\$170 billion, as shown below.

The capital investment and technology expertise necessary to integrate machine learning into credit scoring could be daunting to many lenders in the short term, but the potential for a substantial financial payback in the long term will make such an investment imperative. The new algorithms are well suited to ingesting unconventional data sources, such as social media, further improving the accuracy of ratings. Over time, the role of the credit officer could diminish as data-based decision-making becomes the norm. Ultimately, an improvement in the accuracy of credit assessment throughout the financial services industry should reduce the cost of borrowing, a boon to both lenders and consumers alike.

62 "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," 5/2015 https://www.researchgate.net/publication/276280838_Benchmarking_state-of-the-art_classification_algorithms_for_credit_scoring_An_update_of_research



AUTONOMOUS VEHICLES

One of the transformative applications of deep learning is self-driving cars. Navigating a vehicle through streets, weather conditions, and unpredictable traffic is the kind of open ended problem that learning algorithms such as deep learning can solve. ARK believes that deep learning is a fundamental requirement for level 4⁶³ or higher autonomous driving. Indeed, without deep learning, fully autonomous vehicles would be impossible.



Source: Google, Mercedes

Deep learning solves the two key problems facing autonomous driving: sensing and path planning. Neural nets allow a computer to segment the world into drivable and non-drivable paths, detect obstacles, interpret road signs, and respond to traffic lights.⁶⁴ Additionally, with reinforcement learning, neural nets can learn how to change lanes, use roundabouts, and navigate around complex traffic conditions. While self-driving systems have yet to reach the level required for autonomous driving, the observed rate of progress from Google and others⁶⁵ suggests that self-driving technology will be available by the end of this decade.

Fully deployed, self-driving technology will reduce the cost of transport and bring to life Mobility-as-a-Service (MaaS). Based on ARK's research, by 2020 not only will most cars have autonomous driving capabilities but the cost of travel will fall to \$0.35 per mile,⁶⁶ roughly one tenth the cost of human-driven taxis. As a result, transportation will transition primarily to an on-demand model, introducing a flood of new consumers to the point-to-point mobility market, autonomous miles driven will rise dramatically from de minimis to 18 trillion per year by 2027. At \$0.35 per mile, the market for autonomous on-demand transport will approximate a \$6 trillion market in ten years, as shown below.

63 Level 4 is High Automation, before Level 5, Full Automation, https://www.sae.org/misc/pdfs/automated_driving.pdf

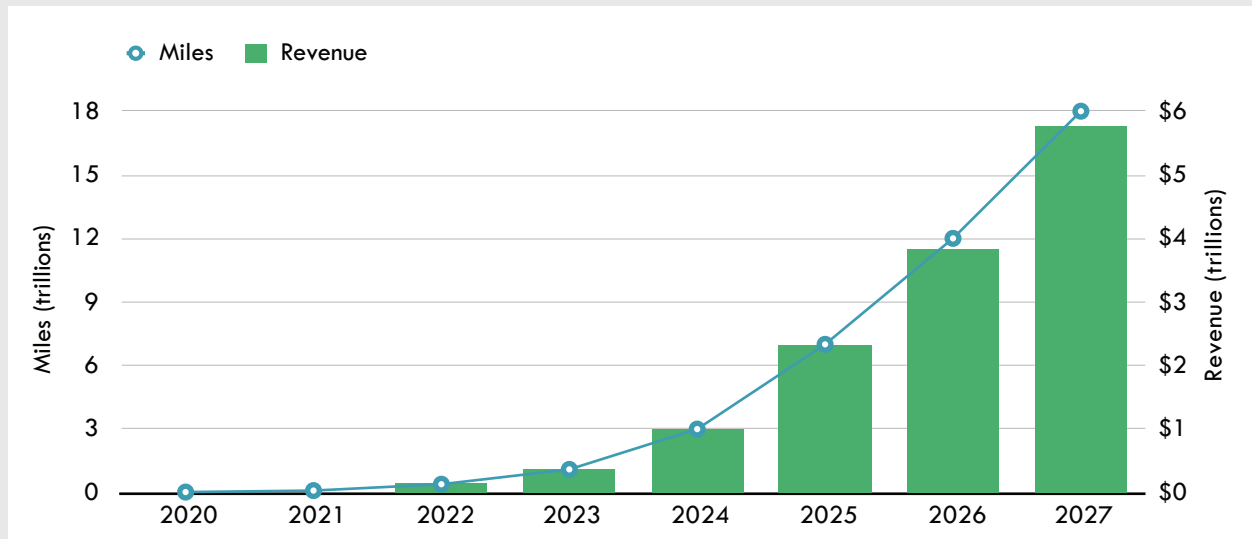
64 "The Three Pillars of Autonomous Driving," Amnon Shashua, 6/20/2016, <https://www.youtube.com/watch?v=GZa9S1MHh-Qc&t=1308s>

65 "DMV Autonomous Vehicle Disengagement Reports," 2016 https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/disengagement_report_2016

66 "Mobility-as-a-Service: Why Self-Driving Cars Could Change Everything," ARK Invest, 2017, <http://research.ark-invest.com/self-driving-cars-white-paper>



FIGURE 21
Mobility as a Service: Miles & Revenue



Source: ARK Investment Management LLC

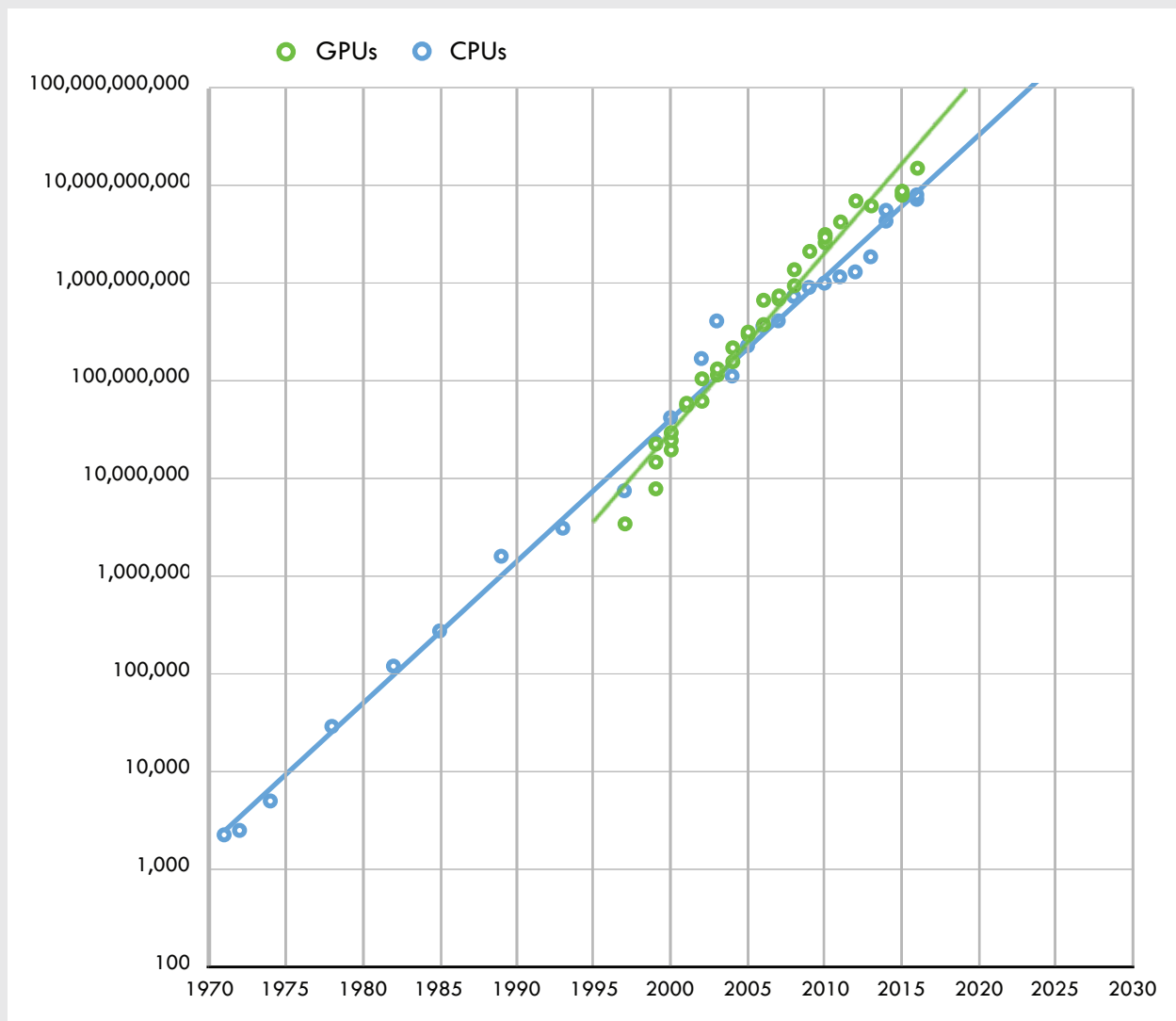


DEEP LEARNING HARDWARE

FROM CPU TO GPU TO ASIC

Deep learning could shift the focus of the microprocessor industry from general application performance to neural net performance. It could impact the entire internet stack, from the cloud to the client device. Computer processors were not designed for deep learning. Both the x86 processors in cloud datacenters and ARM based System-on-Chips (SOCs) in smartphones were designed for general workloads such as internet browsing, mobile apps, and video streaming. These workloads are based on integer operations and tend to be sequential in nature. In contrast, deep learning workloads are based on floating point, or

FIGURE 22
Processor Transistor Count



Source: ARK Investment Management LLC



decimal, operations and are parallel in nature.⁶⁷ As a result, deep learning demands different processor designs, specifically those with high floating point performance.

The highest performing processor for floating point operations is the Graphics Processing Unit (GPU). As shown above, since the mid-2000s GPUs have outstripped Central Processing Units (CPUs) in transistor count, a measure of chip complexity and performance. NVIDIA's Pascal P100 GPU, for example, incorporates 15 billion transistors,⁶⁸ almost double that of Intel's "Knight's Landing" Xeon Phi processor. Further, because GPUs are designed for floating point intensive graphics workloads, most of its transistors are devoted to floating point operations, while CPUs serve a variety of workloads and can devote only a portion of their transistor budget to floating point operations.

When they classify images with deep learning, GPUs outperform CPUs by a factor of 12, as measured by the total fixed and variable cost of ownership, as shown below.

FIGURE 23
Image Classification Performance per Dollar

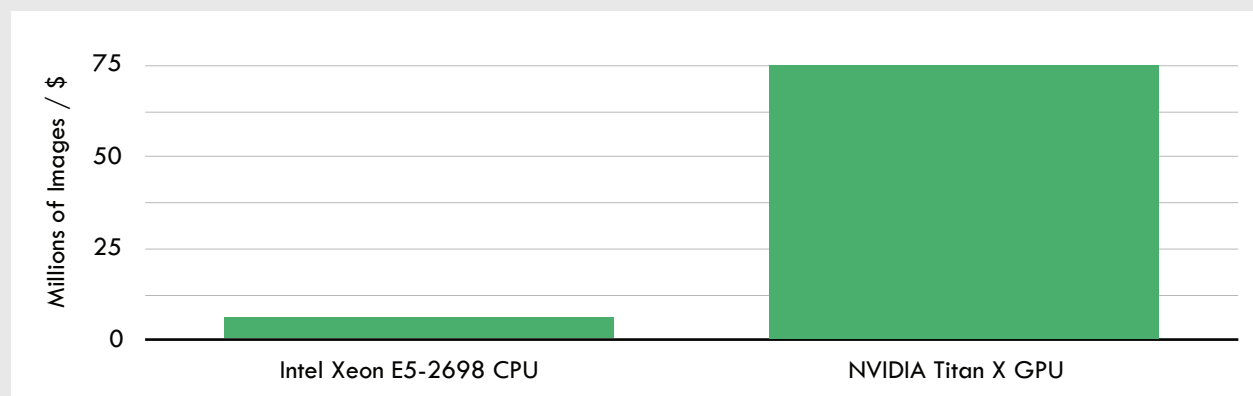


Image classification using AlexNet. Cost includes hardware and power consumption.
Assumptions: 80% utilization and \$0.1 per kWh on both systems.
Source: NVIDIA, ARK Investment Management LLC

In the data center, deep learning already has had a significant impact on infrastructure spending and server designs. Deep learning has two functions, training in which the neural net learns how to do a task by ingesting large amounts of data, and inference in which the trained neural net does actual work. According to the CEO of Nvidia, training is a billion to a trillion times more demanding, perhaps explaining why internet-scale companies are adding many servers to their data centers.

Google, Facebook, Microsoft, and Baidu, among many other companies have created dedicated clusters of servers designed specifically for training neural nets.⁶⁹ Unlike traditional two-socket servers, in which most of the value accrues to CPU and memory manufacturers, deep learning servers require four to eight GPUs.

⁶⁷ The primary operation of neural networks is matrix multiplication and addition of floating point numbers.

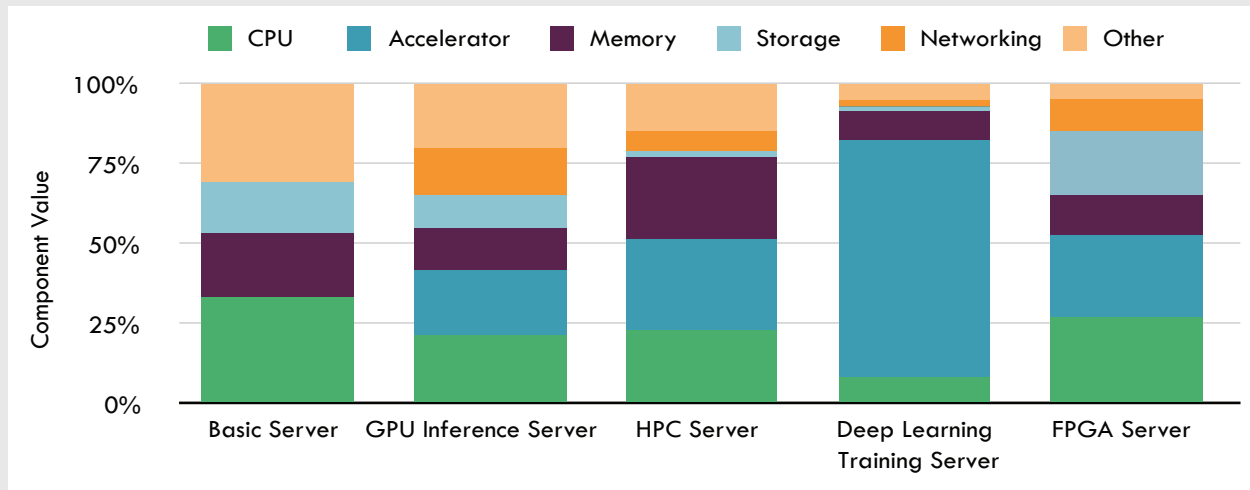
⁶⁸ "Inside Pascal: NVIDIA's Newest Computing Platform," 6/16/2016 <https://devblogs.nvidia.com/parallelforall/inside-pascal/>

⁶⁹ One example: "Inside the GPU Clusters that Power Baidu's Neural Networks," The Next Platform, 12/11/2015 <https://www.nextplatform.com/2015/12/11/inside-the-gpu-clusters-that-power-baidus-neural-networks/>



ARK estimates that up to 75% of the value of deep learning servers accrues to the GPU, as shown below. While they have not been deployed widely for inference, GPUs and other accelerators could account for 25% of the value of future deep learning inference servers.

FIGURE 24
How Servers Components Change with Deep Learning



Source: ARK Investment Management LLC

ALTERNATE DEEP LEARNING PROCESSORS

The GPU is not the only processor capable of accelerating deep learning. Field Programmable Gate Arrays (FPGAs) are another class of processors that offers high performance. Although difficult to program and slower in speed than GPUs, FPGAs can provide power-efficient inference acceleration. Microsoft claims that prospectively almost all of its servers will be coupled with an FPGA to accelerate AI and other internet workloads.⁷⁰ Meanwhile in the cloud, Amazon's EC2 F1 servers provide FPGA based acceleration on demand.

Another way to accelerate deep learning would be a custom chip. A processor designed specifically for deep learning would be an order of magnitude faster than today's GPUs, as it would swap the high precision execution graphics units for lower precision deep learning units, and remove fixed function hardware to save space and power. Companies such as Google,⁷¹ Intel,⁷² and Graphcore⁷³ are following this route.

70 "Microsoft Goes All in for FPGAs to Build Out AI Cloud," Top500.org, 9/27/2016 <https://www.top500.org/news/microsoft-goes-all-in-for-fpgas-to-build-out-cloud-based-ai/>

71 "Google supercharges machine learning tasks with TPU custom chip," 5/18/2016 <https://cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html>

72 Intel is paying more than \$400 million to buy deep-learning startup Nervana Systems," ReCode, 8/9/2016 <https://www.recode.net/2016/8/9/12413600/intel-buys-nervana-350-million>

73 "An Early Look at Startup Graphcore's Deep Learning Chip," The Next Platform, 3/9/2017 <https://www.nextplatform.com/2017/03/09/early-look-startup-graphcores-deep-learning-chip/>



Google's chip, the Tensor Processing Unit, or TPU, has focused specifically on deep learning. Already deployed in Google's data centers, the TPU powered the deep learning program that defeated world Go champion Lee Sedol. Google claims that the TPU enables performance an order of magnitude higher than GPUs and FPGAs. Intel may release a deep learning chip similar to the TPU in late 2017, thanks to its acquisition of deep learning startup Nervana.

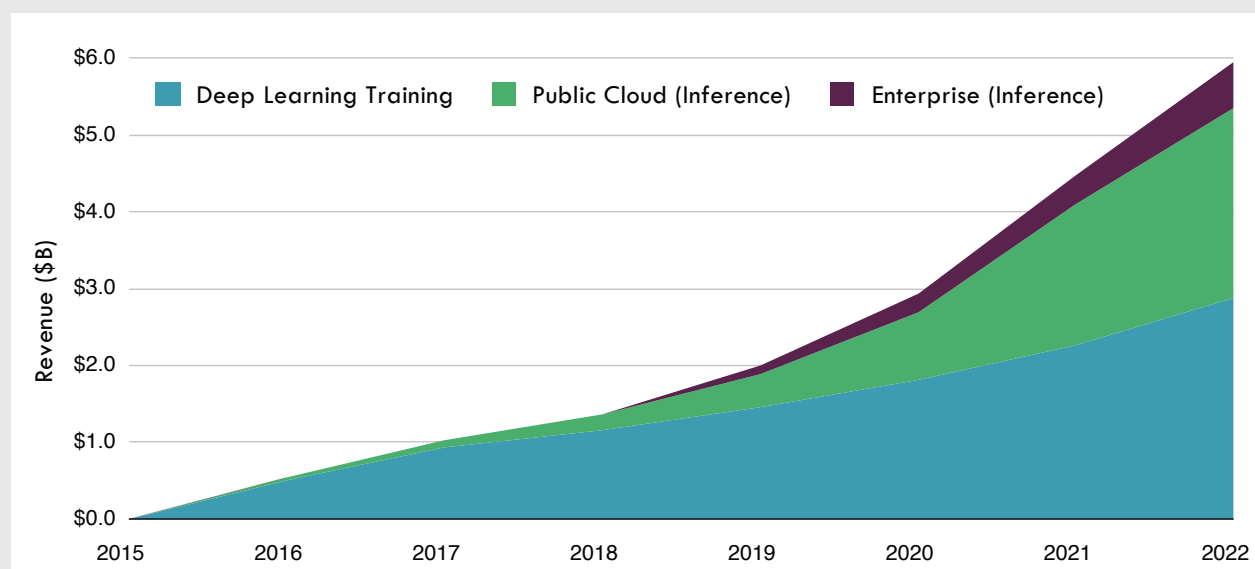
With time, dedicated deep learning processors could become the AI workhorse of the data center, gradually replacing GPUs and FPGAs for training and inference. Deep learning is unique enough as a workload to justify a new architecture and to support the daunting cost of ongoing chip development. With space and power limited on client devices, deep learning eventually could reside on-chip as part of an acceleration block, similar to the way graphics is handled today.

DEEP LEARNING DATA CENTER OPPORTUNITY

The growth of deep learning as a new and demanding workload means that hyper-scale data centers will need to invest aggressively in deep learning accelerators whether they be GPUs, FPGAs, or ASICs.

ARK estimates that deep learning accelerator revenue will grow 70% annually from \$400 million in 2016 to \$6 billion by 2022. At that time, according to our research, roughly half of accelerator revenue will be for training, and half for inference, as shown below.

FIGURE 25
Deep Learning Accelerator Revenue in the Data Center



Source: ARK Investment Management LLC

Note: The information in this chart excludes accelerator revenue for non-deep learning applications.



Training currently makes up the majority of revenue since accelerators are a must-have for efficient training. In contrast, inference can be run on standard servers. Training should grow to a \$3 billion business thanks to continued investment by hyperscale vendors, the increased availability of GPU-based servers in the cloud, and the adoption of deep learning by non-internet industries, particularly automotive where the technology will be key for autonomous vehicles.

As deep learning based services become ubiquitous in web and mobile applications, inference demand should grow and drive demand for accelerators. Microsoft's deployment of FPGAs and Google's rollout of TPUs in their respective data centers suggest that this trend already is under way. We expect hyperscale internet companies to drive the majority of this investment, with on-premise enterprise deployments trailing by roughly two years.

OUR ACCELERATOR REVENUE ESTIMATE ASSUMES THE FOLLOWING:

- | Deep learning will become one of the top workloads at data centers over the next five years.
- | Public cloud servers with accelerators will make up 40% of public cloud server revenue by 2022.
- | Training and inference accelerators will roughly be equal in revenue by 2022.
- | Accelerators will account for 65% of the cost of training servers.
- | Accelerators will account for 25% of the cost of inference servers.

Interesting to note, these estimates focus only on deep learning chips in the data center. The level of deployment in client devices, such as computers, smartphones, cars, cameras, and IoT devices, could be orders of magnitude higher, albeit at lower unit-revenue. Client devices probably will use single chip solutions, with some portion of the chip dedicated to deep learning operations. Their wide adoption, however, will create another source of demand for training and a virtuous cycle of data center and end user adoption.

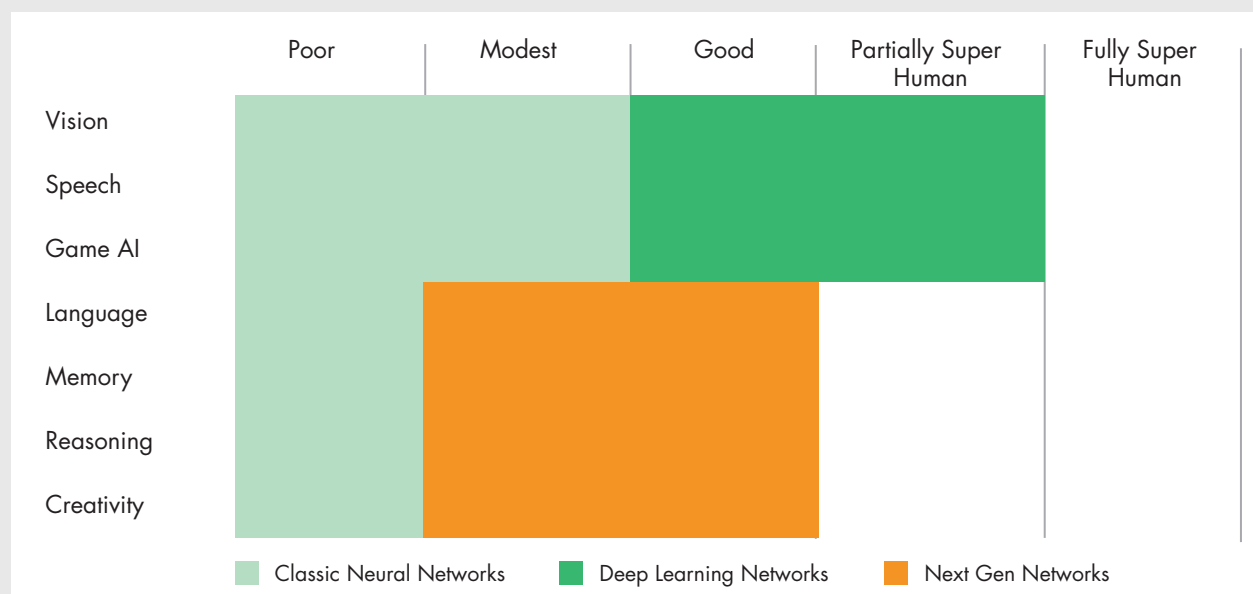


THE FUTURE OF DEEP LEARNING

THE LIMITS OF DEEP LEARNING

While finding applications in most industries, currently deep learning does not come close to general artificial intelligence or human intelligence. In assessing its potential, the following table tries to delineate what class of problems deep learning can and cannot solve.

FIGURE 26
Deep Learning Capability Map



Source: ARK Investment Management LLC

Prior to the current wave of interest in deep learning, neural networks were useful for vision and speech recognition, problems of modest complexity, as shown in the dark green block above. Today, powered by internet scale datasets and compute resources, deep learning is much more advanced, allowing networks to outperform humans in certain vision, speech, and game problems, shown above in light green. Deep learning's primary contribution thus far has been to solve a subset of intelligence problems - those dealing with pattern recognition - with great efficacy.

Unsolved, however, are forms of intelligence beyond mere pattern recognition: the understanding of natural language, memory and recall, multi-step reasoning, and creativity. These problems involve intent, ambiguous sentences, and reasoning about outcomes, often drawing on knowledge about the real world. Today's commercial neural nets have no "general knowledge," no explicit memory system, and no ability to reason through multiple steps. Further, because neural nets are not programmed, a good question is how they arrive at their answers. This inscrutability has been an obstacle to the adoption of deep learning in mission critical applications such as autonomous driving and health care.



Recent improvements, however, have led to neural network architectures with the potential to expand the range of problems solved by deep learning. In the following section, we will review two of these architectures: memory networks and generative networks. If these next-gen networks fulfill their promise, deep learning's capabilities could broaden considerably, as denoted in the orange block above.

COULD DEEP LEARNING LEAD TO ARTIFICIAL GENERAL INTELLIGENCE?

Deep learning has proven surprisingly adaptive but can it take us to full artificial general intelligence? One test that can help answer this question is the Winograd challenge.

The Winograd challenge uses ambiguous sentences to test for natural language understanding and common sense. For example, in the sentence: "The trophy wouldn't fit in the suitcase because *it* was too big," what does *it* refer to? A human knows *it* refers to the trophy but computers currently have no way to solve these problems. Though seemingly simple, Winograd sentences are powerful tests for human level common sense—to solve them reliably, the AI needs to understand language and have a reasonable baseline of empirical knowledge (eg. big things don't fit in small things). If deep learning can be applied to solve Winograd sentences reliably, that would suggest that it has the potential to address artificial general intelligence.

A few more examples of Winograd sentences are listed below. In each case, the AI program must identify what the pronoun (in bold) refers to:

- | The city councilmen refused the demonstrators a permit because **they** feared violence.
- | The man couldn't lift his son because **he** was so weak.
- | Frank was upset with Tom because the toaster **he** had bought from him didn't work.



DEEP LEARNING WITH MEMORY

An important next step in the evolution of deep learning is to incorporate the concept of memory. Convolutional neural nets such as those used for image recognition have no mechanism to reference specific past experience. Recurrent neural networks used for speech recognition provide a short term memory system, but no long term memory. Without a robust memory system, neural nets cannot reason through multiple steps or acquire new skills.

Since 2015, Facebook and DeepMind (owned by Google) have made significant strides in augmenting neural networks with memory. Facebook's implementation, called Memory Networks,⁷⁴ can read short stories and answer questions about them. In the example below, the memory network reads a synopsis of Lord of the Rings and is able to answer questions. An improved version, published in 2016, can read Wikipedia pages and answer questions with 76% accuracy compared to 56% with traditional information extraction techniques.

Bilbo travelled to the cave. Gollum dropped the ring there. Bilbo took the ring.
Bilbo went back to the Shire. Bilbo left the ring there. Frodo got the ring.
Frodo journeyed to Mount-Doom. Frodo dropped the ring there. Sauron died.
Frodo went back to the Shire. Bilbo travelled to the Grey-havens. The End.
Where is the ring? **A: Mount-Doom**
Where is Bilbo now? **A: Grey-havens**
Where is Frodo now? **A: Shire**

Source: Facebook

DeepMind's system, called a Differentiable Neural Computer (DNC), not only can answer questions but also can solve goal-orientated problems.⁷⁵ Notably, DeepMind's DNC is able to ingest a London subway map and provide commuter paths from one station to another. Conventional trip planning software relies on human programming to provide such a path. Learning systems like the DNC understand the subway as a graph, using memory and reasoning to deliver strategies with no human input.

Adding memory to neural networks expands the kind of problems that they can address substantially. Two applications with high market potential are question-answering (QA) and conversational bots. Today Google provides links to answers but not the answers themselves. If neural nets augmented by memory could reference sources like Wikipedia or books, Google would be able to provide complete answers to questions. Likewise, today's AI bots are fragile and forgetful, and they lack common sense. With the addition of memory, AI bots would be able to follow a conversation, remember key facts, and perform simple reasoning to provide answers.

⁷⁴ "Memory Networks," Facebook, 10/15/2014 <https://arxiv.org/abs/1410.3916>

⁷⁵ "Differential Neural Computers," DeepMind, <https://deepmind.com/blog/differentiable-neural-computers/>



GENERATIVE NETWORKS

Today's neural networks are mostly *discriminative*: given an input, they tell you what it is. Generative networks⁷⁶ work the other way around: provide a description and they *generate* an output based on the description. In other words, generative networks don't just provide answers: they create content.

Generative networks are exciting because they expand the addressable market of deep learning dramatically. They can generate handwriting, human voice, photos, art, or even movie scripts.⁷⁷ Researchers at Indico, for example, have trained networks to take a simple text description ("sunrise over the ocean") and create a low resolution photo, as shown below. In a totally different domain, researchers at Sony created a generative network "DeepBach"⁷⁸ to create highly convincing Bach compositions that fooled professional musicians almost 50% of the time.



Source: Indico⁷⁹

Generative networks could make software far more powerful and human workers much more productive. Creative-oriented applications like Photoshop potentially could allow artists to conjure up photos based only on high level descriptions. For example, the artist could ask the application to draw a bedroom with modern furniture, large windows, afternoon sunlight, and two kids. A generative network, having been trained on a large corpus of bedroom photos and interior decoration magazines, would be able to create such a picture in seconds. After reviewing the first render, the artist then could ask for larger windows, a different color of paint on the walls, and so on. Because neural networks understand images at different layers of abstraction and at the object level, they have the ability to make these changes and enable a complete workflow.

76 "Generative Adversarial Networks," Goodfellow et al. 6/10/2014 <https://arxiv.org/abs/1406.2661>

77 "Movie written by algorithm turns out to be hilarious and intense," ArsTechnica, 6/9/2016 <https://arstechnica.com/the-multi-verse/2016/06/an-ai-wrote-this-movie-and-its-strangely-moving/>

78 "DeepBach: a Steerable Model for Bach chorales generation," Sony Computer Science Laboratories, 12/3/2016, <https://arxiv.org/abs/1612.01010>

79 "Deep Advances in Generative Modeling," 3/21/2016, <https://www.youtube.com/watch?v=KeJINHjyzOU>



CONCLUSION

Truly foundational technologies—the steam engine, electricity, the transistor, the internet—are rare, but when they do occur, they impact the world profoundly. They create entirely new industries and lead to products that touch and transform the lives of billions.

ARK believes that deep learning is one of the most important foundational technologies to emerge since the internet. In just a few years, it has moved from academia to production, powering vision, speech, robotics, health care, and various internet services used by billions of people worldwide. In economic terms, deep learning based companies could create over \$17 trillion in new market capitalization over the next 20 years.

As an investment theme, deep learning is particularly attractive. It cuts across industries and sectors, addressing trillions of dollars of revenue. It is young, only five years old as of 2017, and it is growing at a remarkable rate in use cases, startup formation, market adoption, and revenues.

Despite its progress to date, deep learning isn't standing still: new capabilities such as memory networks and generative networks could make deep learning far smarter, possibly providing a bridge to artificial general intelligence. In such a scenario, deep learning could make even the internet look small.



©2017, ARK Investment Management LLC. All content is original and has been researched and produced by ARK Investment Management LLC ("ARK") unless otherwise stated herein. No part of this content may be reproduced in any form, or referred to in any other publication, without the express written permission of ARK.

This material is for informational purposes only and does not constitute, either explicitly or implicitly, any provision of services or products by ARK. Nothing contained herein constitutes investment, legal, tax or other advice and is not to be relied on in making an investment or other decision. Investors should determine for themselves whether a particular service or product is suitable for their investment needs or should seek such professional advice for their particular situation.


All statements made herein are strictly beliefs and points of view held by ARK. Certain of the statements contained herein may be statements of future expectations and other forward-looking statements that are based on ARK's current views and assumptions and involve known and unknown risks and uncertainties that could cause actual results, performance or events to differ materially from those expressed or implied in such statements. In addition to statements that are forward-looking by reason of context, the words "may, will, should, could, expects, plans, intends, anticipates, believes, estimates, predicts, potential, projected, or continue" and similar expressions identify forward-looking statements. ARK assumes no obligation to update any forward-looking information contained herein. Although ARK has taken reasonable care to ensure that the information contained herein is accurate, no representation or warranty (including liability towards third parties), expressed or implied, is made by ARK as to its accuracy, reliability or completeness.

Any reference to a particular company or security is not an endorsement by ARK of that company or security or a recommendation by ARK to buy, sell or hold such security. ARK and clients as well as its related persons may (but do not necessarily) have financial interests in securities or issuers referenced. Past performance does not guarantee future results. The performance data quoted represents past performance and current returns may be lower or higher.

Any descriptions of, references to, or links to other publications, sites, products or services do not constitute an endorsement, authorization, sponsorship by or affiliation with ARK with respect to any such publication, site, product or service or its sponsor, unless expressly stated by ARK. Any such publication, site, product or service have not necessarily been reviewed by ARK and are provided or maintained by third parties over whom ARK exercises no control. ARK expressly disclaims any responsibility for the content, the accuracy of the information, and/or quality of products or services provided by or advertised by these third-party publications or sites.

ARK Invest
155 West 19th, 5th Floor
New York, NY 10011
info@ark-invest.com | www.ark-invest.com

JOIN THE CONVERSATION

 **@ARKinvest**
@jwangARK