STATE OF ARTIFICIAL INTELLIGENCE

entelect
everything is possible

FOREWORD

Artificial Intelligence has really woken up as an industry topic in the last two years, and it's clear there is still a lot of confusion about what AI really means to us as consumers and businesses. Are we going to lose our jobs or become subjects of a runaway robotic overmind in our lifetimes?! I think a key challenge we face, especially as technology experts in industry right now is to demystify AI for everybody, to separate applications of it into things we can all understand. For example, we're starting to see that there is a clear difference between an AI that follows rules and an AI that learns.

The rules AI has a strict set framework and leverages formulaic computational strength to win out in an algorithmically intensive battle between a database and a human mind, but it is still very bounded. A well-known example is grand master chess player Garry Kasparov losing to IBM's deep blue AI. While it's despairing to know we may never beat Windows at chess again, in many ways these tools should replace menial tasks and jobs for us, and we should be happy for it because we're then free as a workforce to do things that have more meaning: like planning and interpreting. Businesses should be excited about these, not because it is sexy AI, but because it is efficient. We rejoice as repetitive tasks disappear from our daily work, but as Kasparov himself pointed out, it's one thing to design a computer to play chess at Grand Master level, but it's another to call it intelligence in the pure sense. It's simply throwing computer power at a problem and letting a machine do what it does best.

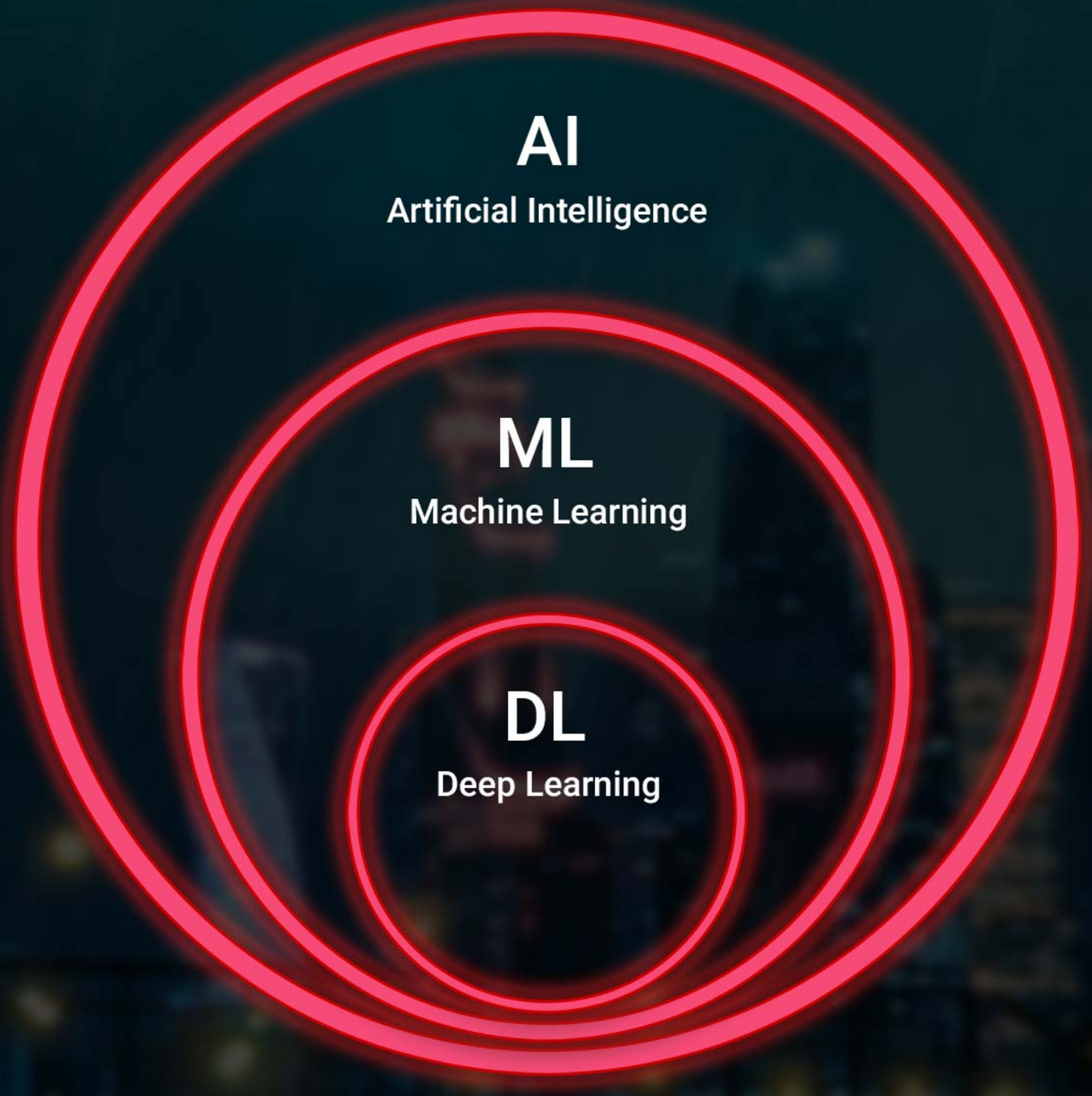
The learning AI is a little more interesting and wild, and we're really at the early stages of how to apply something like this. Also borne from newly accessible computational power, these AI systems actually teach themselves through observation and data. Some famous examples are Microsoft Tay, which famously became racist on Twitter in less than 24 hours, and Google's image recognition platform which we're all training each time we click pictures to ironically prove we're not a robot online. For businesses looking to apply AI like this, it's where things become tricky. First recognising that what you need is an AI that learns rather than an AI that repeats tasks with no experimentation, and then when taking the next step to accept that learning AI is still trained by its inputs. In the software industry we have a widely shared philosophy of "garbage in, garbage out" and this still holds true: just refer back to Microsoft's Tay. This is where Google and Tesla are doing things right, they're sourcing as much data as possible, with huge variety and volume to train their systems, giving autonomous driving and image recognition the best possible chance of succeeding.

Ultimately, as a participant in the global technology industry, we have this responsibility to help our customers and partners cut through the hype, find real applications of what is incredible tech which can make it into production and help their businesses go from good to great.

- Matthew Butler, General Manager at Entelect

entelect
everything is possible

GLOSSARY



Artificial Intelligence is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and other animals...Colloquially, the term "artificial intelligence" is applied when a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving"...
- *Wikipedia*

Machine Learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed...
- *Wikipedia*

Deep Learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised...
- *Wikipedia*

entelect
everything is possible

AI FROM THE TECH GIANTS

entelect
everything is possible

APPLE

Traditionally, Apple is not viewed as an AI powerhouse like many of the other tech giants. Their most public endeavours into the field are exposed through their voice assistant Siri, one of the first of its kind, but these days quite far behind the assistants of Amazon and Google.

Apple recently hired Google's head of Artificial Intelligence, John Giannandrea, to run their Machine Learning and AI strategy division. He reports directly to CEO Tim Cook.

In June 2018, Apple held their annual developers' conference, WWDC. Compared to Google's announcements, Apple were relatively thin in terms of AI, but they did announce a few interesting things.

Apps

Siri will now suggest apps based on one's regular behaviour as well as location and calendar.

The Photos app has been enhanced with some new smarts, including identifying people in photos and suggesting that you share the album with them. They've also significantly improved photo search, with ML used to identify the contents of photos.

Siri Shortcuts use machine learning and AI to learn your behaviour and provide assistance

Autonomous Cars

Apple has been quiet about its efforts in the development of self-driving cars, although CEO Tim Cook has called it "the mother of all AI projects."

It is notable that Apple has yet to release any information to the public about products related to self-driving cars but has made several moves indicating that it is still highly focused on the technology.

Apple has reportedly expanded its fleet of self-driving cars in California, registering an additional two dozen vehicles. It's a significant expansion for a company that has been seen as lagging in the race to build self-driving cars. Apple originally registered three Lexus Rx450h SUVs under its permit to test autonomous vehicles in April 2017. Since then, it has acquired an additional 24 Lexus SUVs.

From various reports, we know that Apple has abandoned its ambitions to build an entirely new vehicle from scratch and has instead shifted focus to building autonomous software it could develop for existing carmakers. Last July, CEO Tim Cook confirmed in an interview, that the iPhone maker is currently "focusing on autonomous systems" — rather than, say, a car stamped with the Apple logo — and that this could be used for many different purposes.

Apple's test vehicles have been spotted a couple of times in the wild, most recently last October. The car appeared to be outfitted with standard third-party sensors and hardware, including six Velodyne-made LIDAR sensors, several radar units, and a number of cameras — all encased in Apple-esque white plastic.

This photo posted on MacRumors is a rare glimpse of an Apple test vehicle

Most recently, Apple has signed a deal with Volkswagen to turn some of the carmaker's new T6 Transporter vans into Apple's self-driving shuttles for employees.

APPLE

Core ML

CoreML allows developers to incorporate trained AI models into their apps.

Core ML is optimized for on-device performance, which minimizes memory footprint and power consumption. In keeping with Apple's privacy policies, these libraries enable and ensure that Artificial Intelligence capabilities run directly on the user's device.

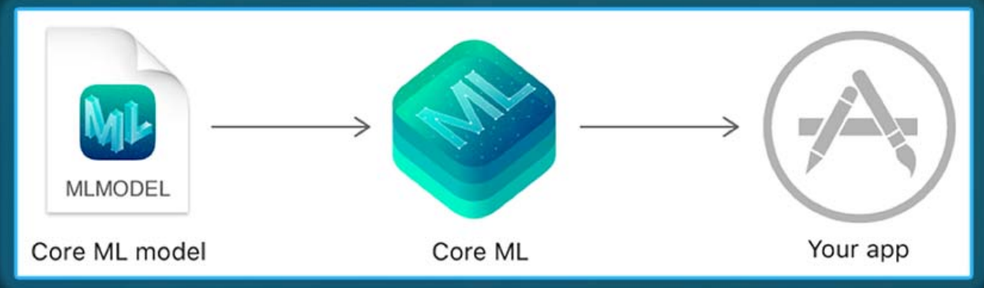


Diagram depicting the basics of Apple's Core ML process

Create ML

CreateML was recently announced at WWDC18 and allows developers to train models to use CoreML in their apps. These days, due to the computational intensitivity of training models, most people train in the cloud. CreateML works differently in that all the hard work is done locally on a Mac. This is most likely due to Apple's privacy policies, as you may not want to upload sensitive data to the cloud for training.

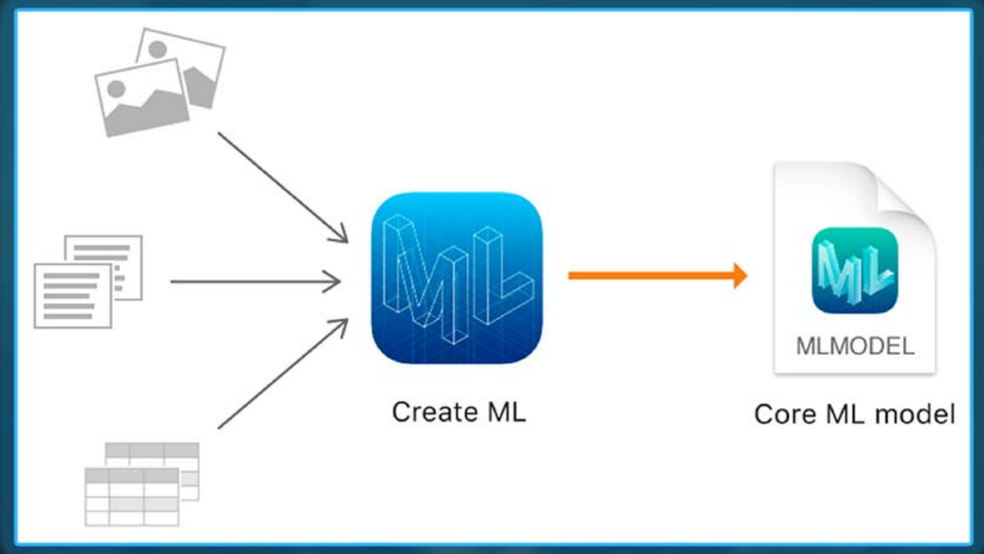


Diagram depicting the basics of Apple's Create ML process

AMAZON

Amazon was one of the early AI adopters, used in areas such as product recommendations, shipping schedules and robots within warehouses. However, in the early part of this decade they were a bit behind compared with the competition, and they recognized it. In 2014 Srikanth Thirumalai proposed a way to use deep learning to revamp the recommendations engine (not the typical image and voice recognition applications of machine learning) to Amazon CEO, Jeff Bezos. This ultimately led to rethinking numerous other areas of Amazon, robotics, data centres and what ending up as the Amazon Echo. AI islands of knowledge started to break down and knowledge-sharing began in the space.

Jeff Bezos says AI, and ML in particular, is a renaissance; a horizontal enabling layer across industries, organizations and even business units, empowering them to improve. Amazon uses these techniques to improve itself, and its products, as well as having the goal to provide these to those who it is not normally accessible to.

Amazon describes its business as a flywheel, where it works as a single perpetual motion machine. The AI flywheel turns and innovates in areas of machine learning, which can be, and are, used within the rest of the business.

Externalizing these services and platforms then also formed a new revenue stream for Amazon, which in turn helps improve the services with enriched data to work from. While there is no central AI unit within Amazon, they do have a unit dedicated to spreading knowledge and support across the organization, as well as some R&D. Amazon is now one of the biggest forces in AI.

Amazon Echo and Amazon Alexa

The Amazon Echo

Project Alexa started as early as 2011 in Amazon. The Echo device was envisaged as “a low-cost, ubiquitous computer with all its brains in the cloud that you could interact with over voice – you speak to it, it speaks to you”. Amazon then worked backwards from there to ultimately achieve the vision. In the early stages Amazon bought several companies to obtain their IP and expertise in speech-to-text, AI driven responses and text-to-speech areas.

They broke ground in far field speech recognition, which at the time was only successfully performed by devices such as nose cones in Trident submarines (billion dollars’ worth), not a kitchen counter device! Amazon already had excellent cloud services, data centres and advance machine learning algorithms with the ability to process large amounts of data, which could be used to solve the problem.

The project resulted in a voice platform, Alexa, that could be used outside of the Echo as well (see “ML in the cloud” below). There are now integrations with other Amazon services (music, video, shopping), as well as the Alexa Skills Kit (ASK) which allows developers to provide additional skills (capabilities) to the platform.

The first generation Amazon Echo, along with Amazon Alexa was launched at the end of 2014, and has enjoyed a lot of success since then. A number of variations of Echo devices and accessories are now available, such as the Dot, Plus, Spot, Show and Look. Partner companies have also developed devices backed by the Alexa platform.

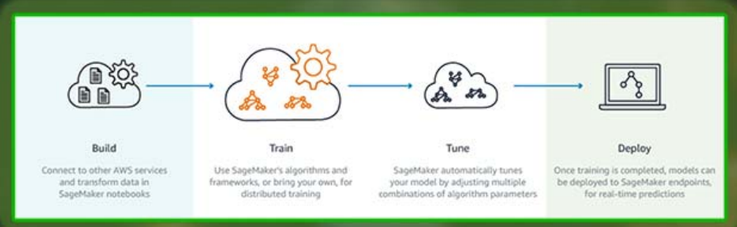
Massive amounts of data from the millions of users of Echo and Alexa have been captured, feeding back into, and spinning the Amazon AI flywheel even more.

AMAZON

Machine Learning in the Cloud

Swami Sivasubramanian drove adding machine learning algorithms, services and frameworks into Amazon's cloud services, namely Amazon Web Services (AWS). His idea was to make what was then only available to well-funded large corporations, accessible to as many as possible.

First offered in 2015, AWS Amazon Machine Learning was built on top of existing libraries such as Google's Tensorflow and Caffé. It was extended in 2016 with the innovations from Alexa, such as text-to-speech service Poly, natural language processing service Lex, and then Rekognition service in the field of vision. SageMaker was announced late 2017, which is a platform for building machine learning into your business.



Example Sagemaker Pipeline

List of services include:

- Amazon SageMaker – platform for building your own machine learning models.
- Amazon Rekognition Image – deep learning-based image analysis.
- Amazon Rekognition Image – deep learning-based video analysis.
- Amazon Lex – build chatbots to engage customers.
- Amazon Comprehend – discover insights and relationships in text.
- Amazon Translate – fluent translation of text.
- Amazon Transcribe – automatic speech recognition.
- Amazon Polly – natural sounding text to speech.

List of machine-learning enabled services/platforms:

- Amazon Deep Learning AMIs (Amazon Machine Images) - Platform-as-a-Service virtual machines with pre-installed Machine Learning Frameworks.
- Amazon Connect – call centre in the cloud backed by Amazon Lex.
- Amazon Macie – security service build on ML to discover, classify and protect data within AWS.

Gluon

Gluon is a Microsoft and Amazon partnership; an open source deep-learning API to be included in several ML frameworks, currently in Apache MXNet and will be within Microsoft Cognitive Toolkit soon. It simplifies the process of prototyping, building, and training deep learning models without sacrificing training speed.

It has been exposed in AWS since 2017, via Amazon Deep Learning AMI within pre-installed Apache MXNet.

AWS DeepLens

AWS DeepLens is a physical HD wireless video camera, enabled by deep-learning capabilities utilizing Amazon SageMaker and other AWS services. Utilizing AWS Lambda (Amazon's Function-as-a-Service offering), developers can customize and program the camera to take-action based on what it sees. With Echo you tell it what to do, with DeepLens it can react to what it sees e.g. integrate with your Smart Home to turn on lights when it gets dark.

AMAZON

AWS DeepLens Video Camera

Prime Air

One of the most exciting projects within Amazon is Prime Air. It is a conceptual drone-based delivery system still currently in development. It's planned to deliver packages 5 pounds (~2kgs) or less in under 30 minutes utilizing unmanned aerial vehicles.

Still in prototype phase, it's unknown when it will be available, since they are still testing and working on getting regulatory support (i.e. how do they integrate into current airspace traffic and regulations). In December 2017 the first successful delivery was made, in a private trial launched in the UK.

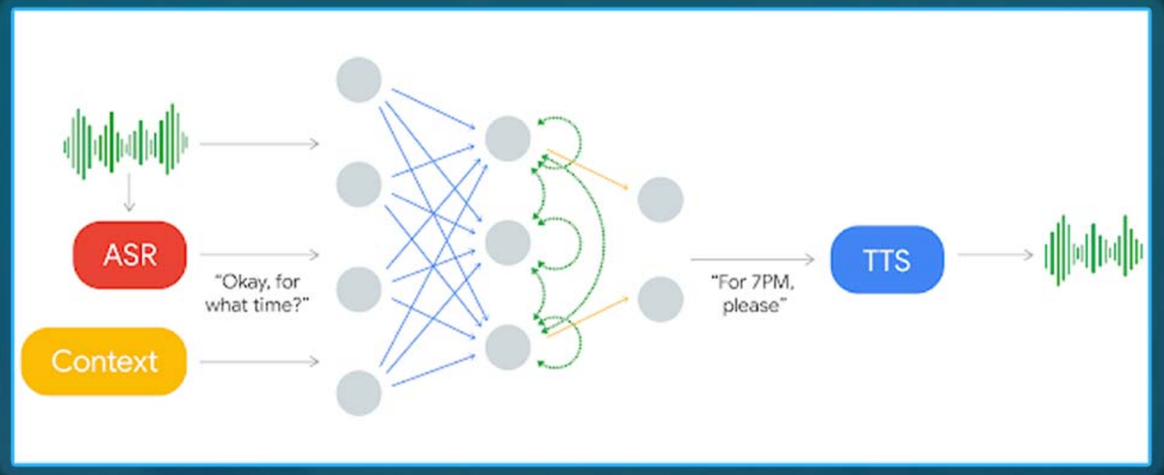
It utilizes a tremendous amount of ML, machine vision systems, natural language understanding and more. One of its challenges is that drones can't count on cloud connectivity.

Amazon Prime Air Drone

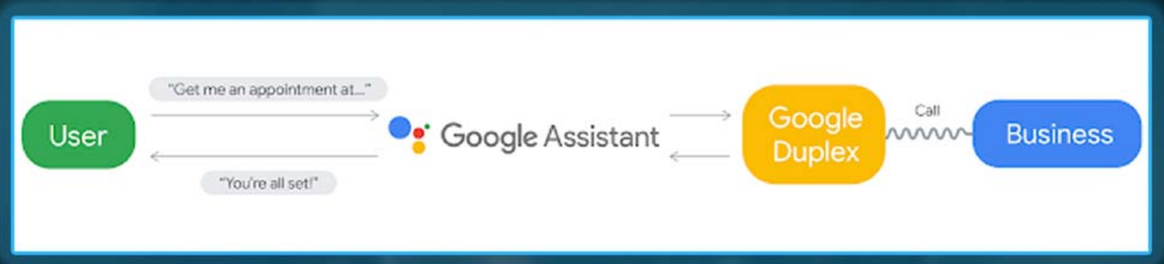
Duplex

Google treats AI as a first class solution, even going so far as to rebrand Google Research to Google AI. Recently, they have made amazing advances in the world of speech recognition and natural language capabilities. The recent demo of Google Duplex, an AI system that can make calls on your behalf, stunned the world with its realistic sounding speech and language patterns. This is the next step in the evolution of the automated assistants that have become commonplace in today's culture. A demo given at I/O showed off the AI's use of conversation flow and adjusting to the caller. Currently, it requires Duplex being constrained to closed domains, which are narrow enough to explore deeply. And as such, it can only carry out natural conversations after being deeply trained in a domain. As a result, it currently cannot carry out general conversations.

"When people talk to each other, they use more complex sentences than when talking to computers. They often correct themselves mid-sentence, are more verbose than necessary, or omit words and rely on context instead; they also express a wide range of intents, sometimes in the same sentence, e.g. "So umm Tuesday through Thursday we are open 11 to 2, and then reopen 4 to 9, and then Friday, Saturday, Sunday we... or Friday, Saturday we're open 11 to 9 and then Sunday we're open 1 to 9."



Example of Automatic Speech Recognition(ASR) and text-to-speech(TTS)



End-to-end example of Google Assistant using Google Duplex

Duplex at its core uses a recurrent neural network designed to deal with the problems of natural language, and uses Tensorflow Extended as the backing solution. The network uses the output of Google's automatic speech recognition (ASR) technology, as well as features of the audio, history of the conversation, reason for the conversation and more. Finally the phrase is passed through a combination of a concatenative text to speech (TTS) and a synthesis TTS engine to sound natural.

The current goal for duplex is to allow a user to have a task (like an appointment) done automatically for them after they ask the device to make the booking.

The demonstration created a lot of controversy, forcing Google to add a message indicating that the call was being made by a machine.

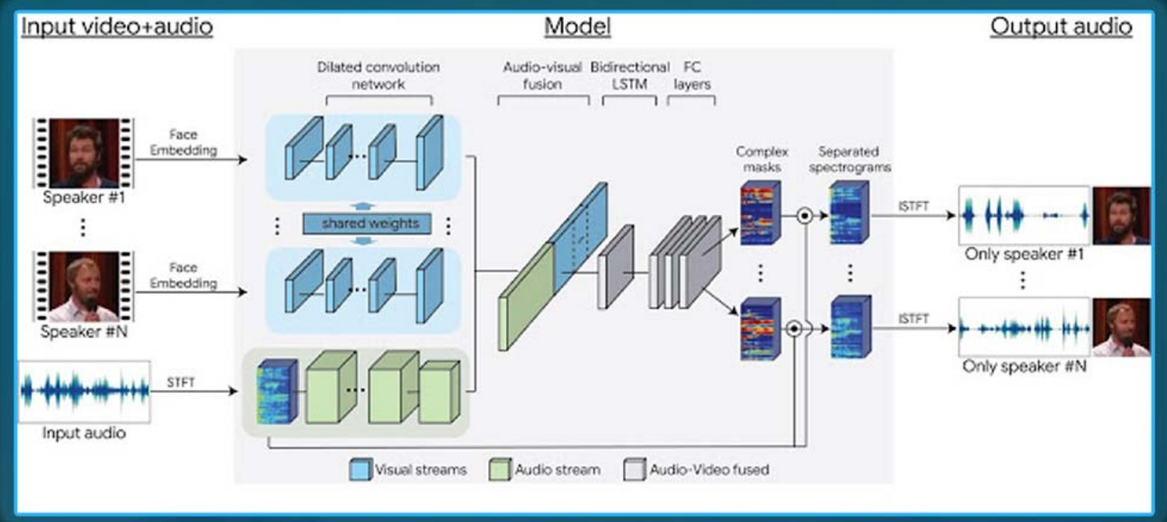
GOOGLE

Look to Listen

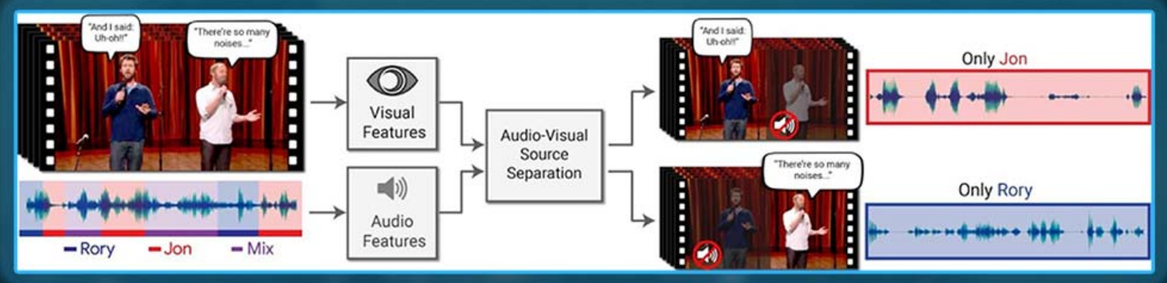
Automatic speech separation (also known as the cocktail party effect) is a well studied but challenging problem for computers. People on the other hand are good at focusing attention in a noisy environment, muting out other sounds and voices.

Google's attempt at this problem is called Looking to Listen at the Cocktail party (to be released at SIGGRAPH 2018), which is a deep learning audio-visual model for isolating single speech signal from a mixture of sounds. This technology allows videos to be produced, in which, specific speech is enhanced, while other sounds are suppressed. The approach works on ordinary videos with a single audio track, and all that is required is that the user selects the face of the person they want to hear, or allows a person to be selected algorithmically based on context.

Unique to this technique is combining both the auditory and visual signals of the video to seperate the speech. For example, movements of a person's mouth should correlate with the sound produced. This helps identify the parts of audio that match to that person. Using the visual signal not only improves the speech separation quality when multiple sources are speaking, but also associates the speech track with visible speakers.



Representation of the Model



Example of Speech Separation

The training examples are generated from a large collection of 100,000 videos of lectures and talks from Youtube. From the videos, segments of clean speech and a single speaker are extracted. Google generated around 2000 hours of video clips, which was used to create synthetic cocktail parties thus allowing the training of a multi-stream convolutional neural network-based model for splitting separate audio streams for each speaker.

By inputting visual features extracted from the face, thumbnails of detected speakers and a spectrogram to the network, it is possible during the training of the network to learn separate encodings for visual and auditory signals. From the learned results, the model is then able to associate the results together to form a joint audio-visual representation of the speaker.

GOOGLE

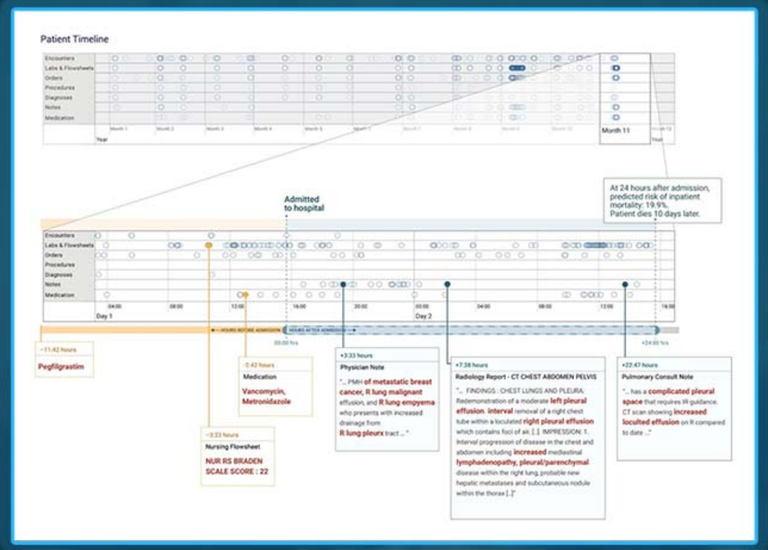
Deep Learning for Health Records

Google have been applying machine learning to help predict what will happen next to patients admitted to hospitals. Joining up with UC San Francisco, Stanford Medicine and the University of Chicago Medicine, they published “Scalable and Accurate Deep Learning with Electronic Health Records” in Nature Partner Journals: Digital Medicine.

They applied deep learning models to make predictions relevant to patients from de-identified electronic health records. Using the data as-is, without performing the manual effort to extract, clean, harmonize and transform variables in the records.

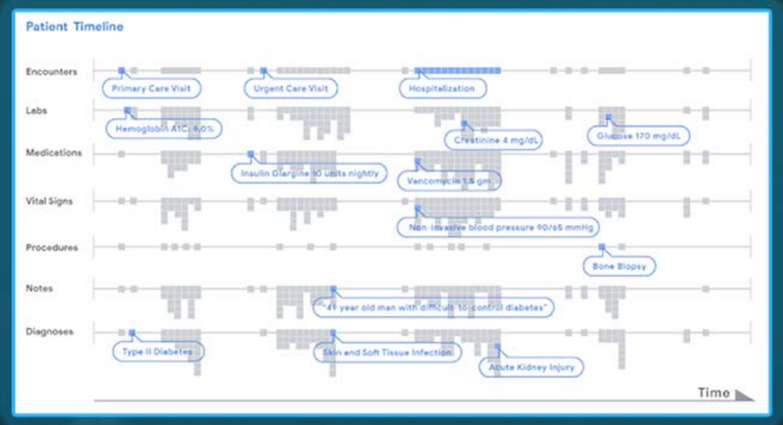
The approach discovered that Electronic Health Records are extremely complicated, since a temperature measurement for example has different meaning according to where it was recorded. In addition, different health systems customizes the Electronic Health Records, meaning a record from one hospital could look different for similar patients. To help resolve this problem, they proposed a consistent way to represent health records which was built on top of the open Fast Healthcare Interoperability Resources standard.

Once in a consistent format, they did not have to harmonize the variables. Instead, a deep learning model can read all the data points from earliest to latest and learn which data predicts the outcome. Due to the amount of data, they developed new types of deep learning modeling approaches based on recurrent neural networks and feedforward networks.



Example of prediction using the model

Google intends to help patients and medical personnel predict what experience the patient is likely to have and from that, help build expectations about how the treatment will progress.

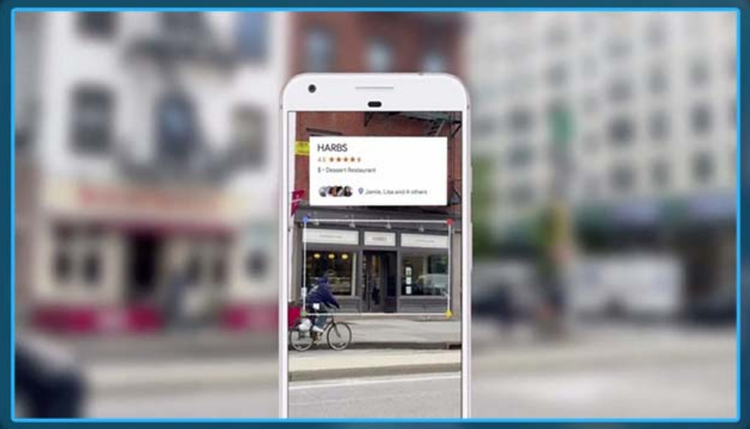


Example of Fast Healthcare Interoperability Resources Standard

To assess accuracy, a measure called the area-under-the-receiver-operator curve is used, which measures the models ability to distinguish a patient with a particular future outcome compared to one who does not. The model is able to determine how likely it is that a patient will need a long stay in a hospital, unexpected re-admissions and inpatient mortality. In addition, they are able to identify the conditions the patient is being treated for. For example, based on the prescribed medication and symptoms, it was able to determine what condition the patient was suffering from. The process is not intended for diagnosis, it can only determine a likely condition based on treatment.

ML Kit

Google have released a machine learning kit focused on mobile development. This Kit is available on firebase and optimized for mobile. The idea is to allow mobile developers the option to include the following features into mobile apps. Image labeling, text recognition, face detection, barcode scanning and landmark detection. An example of the features can be found in Google Lens, allowing a mobile user the ability to use the built in camera to identify landmarks, items and even text. From this, the user could for example get a contextual help popup, popular google searches or select text or barcodes directly from the camera.

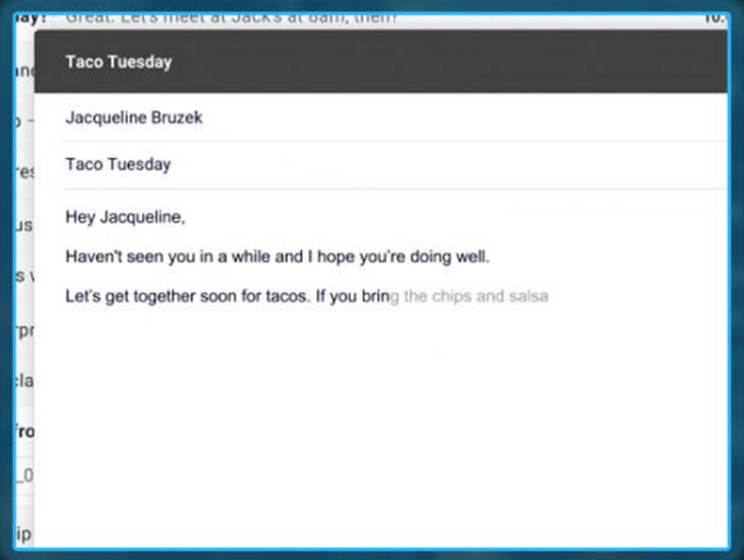


Example of Google Lens

If you want to build your own models instead of using the MLKit defaults or want your model to run on the device, you can use TensorFlow Lite instead. TensorFlow Lite provides a lightweight solution with low latency and a small binary size. It achieves this by optimizing the kernels for mobile apps, pre-fused activations and quantized kernels for smaller and faster (fixed-point math) models.

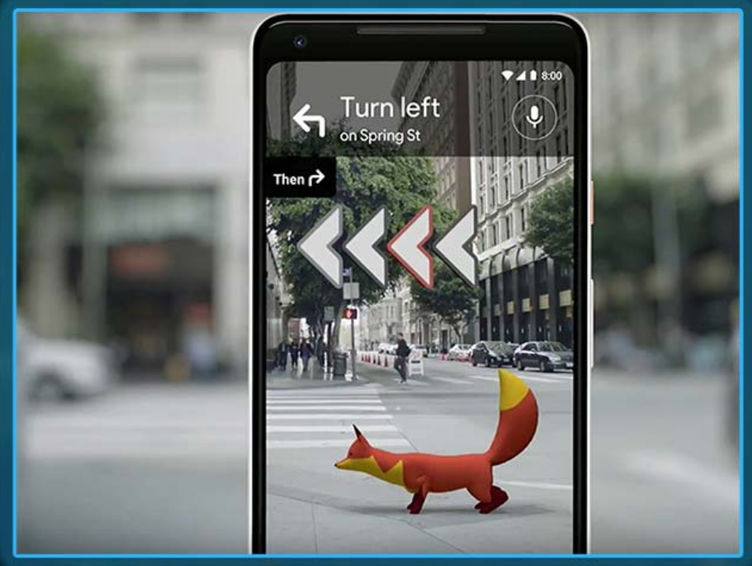
Gmail, Photos and Google Assistant

Gmail has also been updated with smart technology, most notably the feature called smart compose which will help suggest the rest of your email much like predictive text does on a mobile device. It is currently available on the latest version of Gmail.



Example of Smart Compose

Google Maps will soon use Augmented Reality via your device camera to help identify locations and provide suggestions based on learned data, based on your likes and frequently visited locations. Even providing a digital guide to help you navigate to your location via the Augmented Reality on your Camera.



A cute AR(Augmented Reality) guide

The Camera is also getting a massive set of features such as the ability to find similar objects based on what it currently looking at, to being able to copy text directly from images using Google Lens. Google Photos will utilize machine learning to help identify which pictures might mean something extra special to you.

Android P

Google provides a huge selection of AI capabilities for Android P, the next release of Google's popular Operating System. Android P intends to use AI features to help adapt the OS, by learning what applications you use, how you use your device and even adjusts the hardware based on learned preferences.

Android P introduces several battery improvements and one of the key features which uses AI to improve power management is called App Standby Buckets. This feature uses on-device machine learning to understand app usage patterns and uses that data to allocate system resources accordingly. Apps which fall into higher priority buckets get less restrictions than those is lower buckets.

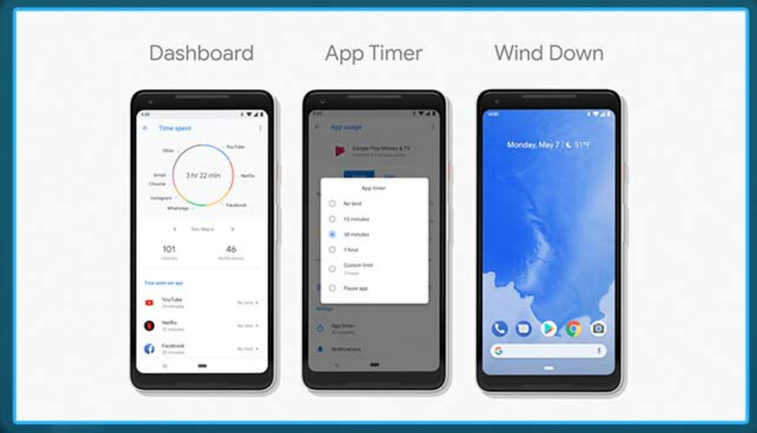
Mobile phone running Android P

Whilst Android P promises improved battery management to ensure we always have technology to help us throughout the day and make our lives much easier, the latest release includes a feature called Android Dashboard. The main purpose of the Android Dashboard feature is to ensure we maintain a healthy balance between life and technology. Dashboard is a digital wellness app, which tracks how much time you spend on top apps, the number of notifications you receive and how many times you unlock your phone.

With Android Dashboard you can set time limits on individual apps using App Timer, giving you the ability to restrict the time you spend on any app per day. Once the limit has been reached, the app will be grayed-out and made inaccessible.

At night, Android P will automatically switch on Night Light and Do Not Disturb mode whilst also cutting down on the amount of blue light emitted by your phone screen. At your set bedtime, your screen will be adjusted to a grayscale mode using a feature called Wind Down.

Using the power of AI, Android P will better understand our addictive habits to technology and suggest we switch off our phones and wind down to make more time for our families.



Dashboard, App Timer, and Wind Down

One of the coolest features which was introduced in Android Nougat was the ability to directly reply to messages and emails from the notification bar. Whilst replying to a message in the notification area without opening the app is very convenient, Android P's AI capabilities go an extra step by giving you quick response options with Smart Reply. On Android P, Google brings its 'Smart Reply' feature (previously made famous on Gmail) to your device notification area. The goal of Smart Reply is to basically make your notification area more useful.

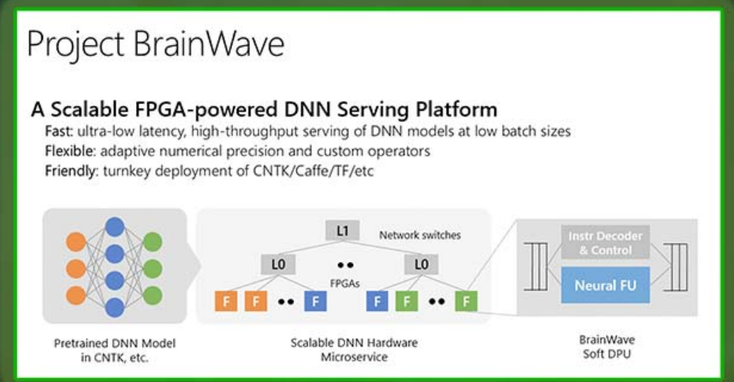
Project Brainwave

Microsoft's main hardware announcement at their Build conference this year was Project Brainwave. An FPGA-powered hardware architecture for real-time AI.

The FPGA (Field-programmable Gate Arrays) has been around for decades, but they haven't been used in large-scale cloud computing before. They're a form of programmable processing unit, where the chip contains a variety of logic blocks and configurable interconnections. These can be reconfigured without needing to redesign the hardware, making them cheaper and more adaptable than specialised hardware.

An example of the hardware used for Project Brainwave

The FPGA-based hardware is already in use running deep neural networks for the Bing search engine, and they're now being deployed across the Azure data centres and integrated into the Azure Machine Learning service. The goal is to allow for real-time AI, for image recognition, speech recognition, language translation and more.



FPGA-based AI architecture

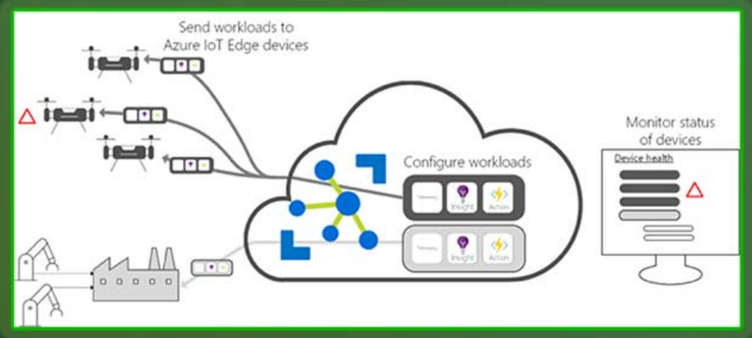
The programming of the FPGAs themselves is not exposed to Azure users. Instead, users will have access to different accelerated models through the Azure ML gallery. These are built upon the Brainwave runtime, which is then deployed onto Neural processing units, which use the FPGAs. This allows for users to gain the advantages of the FPGAs without needing low-level programming skills and it allows for new FPGA architectures to be deployed as there are advances in algorithm design.

Azure IoT Edge

A second announcement, related to Project Brainwave, is the Azure IoT Edge. This is an Azure managed service that interacts with certain IoT devices, such as cameras, drones and similar. It's a convergence of artificial intelligence, cloud, and edge computing.

The core of IoT Edge is moving cloud workloads, such as Azure Stream Analytics and Azure Machine Learning to be deployed down to the IoT device and executed locally via a containerised architecture. This allows for the devices to be managed and code deployed to them from a central location, and for the AI algorithms to be executed locally on the device.
With the intelligence deployed to the IoT device, low latency response and near real-time responses become possible, which allows for uses such as factory safety alerts, or high-speed video-based quality control for manufacturing.

The IoT Edge runtime uses the same programming model as other IoT services, meaning that device-specific programming skills are not needed in order to create custom applications, the runtime supports Java, .NET Core 2.0, Node.js, C, and Python and both Linux and Windows operating systems.



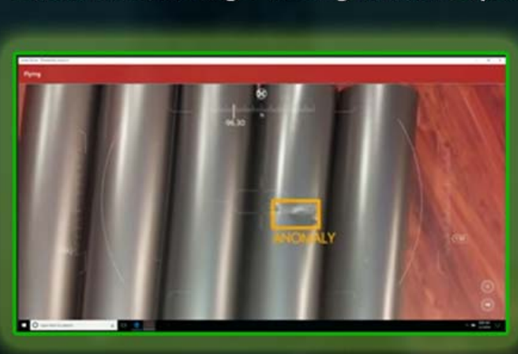
IoT edge workflow

It is in this area that Microsoft has partnered with Qualcomm to create a custom vision developer kit.

Qualcomm IoT cameras

This is a complete hardware and software kit for real-time vision, based on the Qualcomm Snapdragon Neural Processing Engine. This processor is optimised for Machine Learning models, allowing for low-latency image and video recognition.

Similarly, Microsoft has partnered with DJI to add the IoT Edge capabilities to DJI's drones. In a demo at the Build conference, one of the drones was shown identifying anomalies in pipes using on-device image recognition capabilities.



Drone's onboard image recognition identifying an anomaly in a pipe

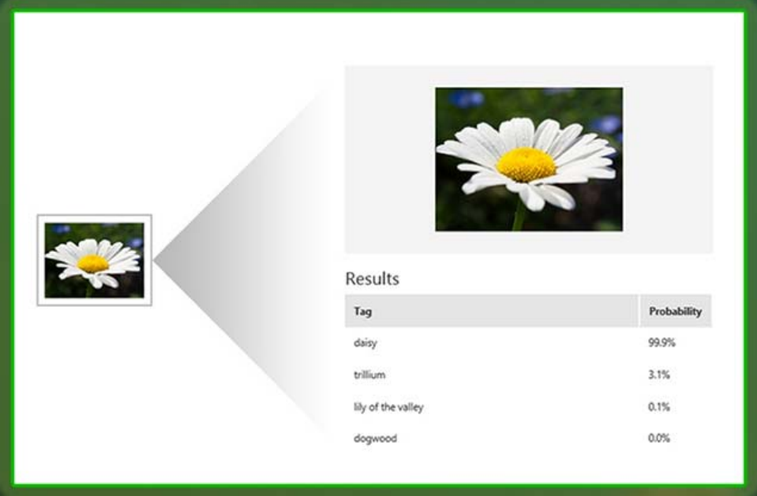
Cognitive Services

Microsoft continues to expand and enhance its cognitive services offerings. Some of the improvements and new features include improvements to their computer vision, content moderation, speech services, text analytics, text translation and search.

The computer vision service now includes OCR for English writing, and adds additional languages to the options for captioning.
The speech APIs have been merged into a unified service that does text to speech, speech to text and translations.

Text Analytics introduces the ability to perform entity identification and linking from raw text, recognising well-known entities and linking to further information on the web. The Bing custom search changes includes custom image search and custom autosuggest. In addition, there are a number of new services, Custom Vision, Visual Search, and the AI for Earth collection of services.

The custom vision API allows for users to submit their own list of labeled images in order to produce a computer vision model tailored for their specific needs. The resulting model can be interrogated via REST API calls, or exported for use locally or offline. The exported model supports Tensorflow, CoreML and ONNX.



Example of CustomVision results

ML.net

The last of the major announcements from Microsoft was for their ML.net libraries for the .Net framework. This allows developers using C#, F# or other .Net languages to easily integrate custom machine learning into their applications. The libraries currently support Light GBM, Accord.NET, CNTK, and TensorFlow, with more coming soon.

The libraries are open source, and run on Windows, Linux and MacOS.

AI for Earth are a new set of cognitive services aimed at solving global environmental challenges. One of the current projects, is land cover mapping, which works out from satellite images, what land is used for and whether the areas are water, grass, trees, buildings or other. Currently this is done by hand, and is a slow process. A machine learning model capable of doing land cover classifications will significantly increase the speed of such identification.

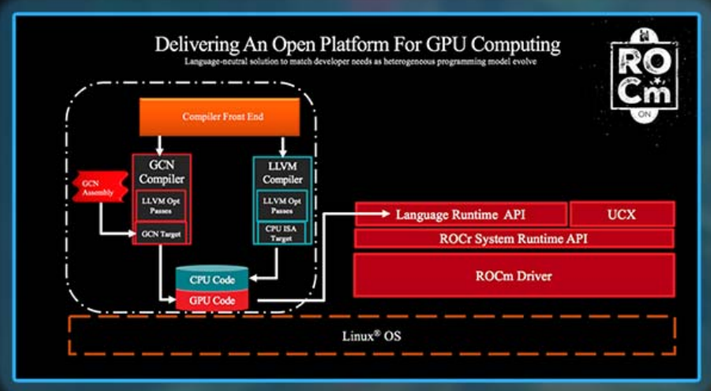
AI FROM THE SILICON GIANTS

entelect
everything is possible

Advanced Micro Devices (AMD) has been one of the biggest players in chip manufacturing for decades. Their foray in AI naturally came about after the increasing need for specialist hardware processing ML data. GPUs architecture lends itself nicely to the general SIMD requirements of ML, and so they are now providing AI capable components for gaming and data centres alike.

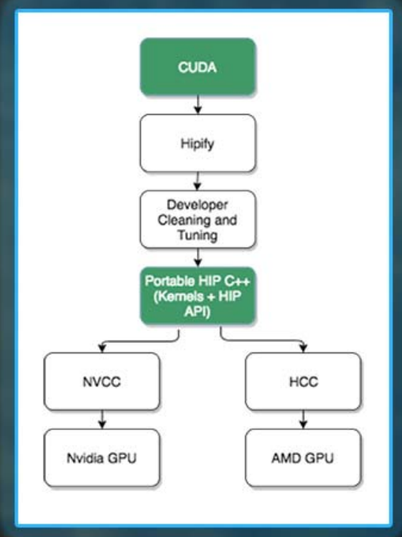
While they are probably still a bit behind their biggest competition, nVidia, in this area, a number of recent product launches and partnerships indicate they are tackling them head on.

ROCm - GPU computing platform



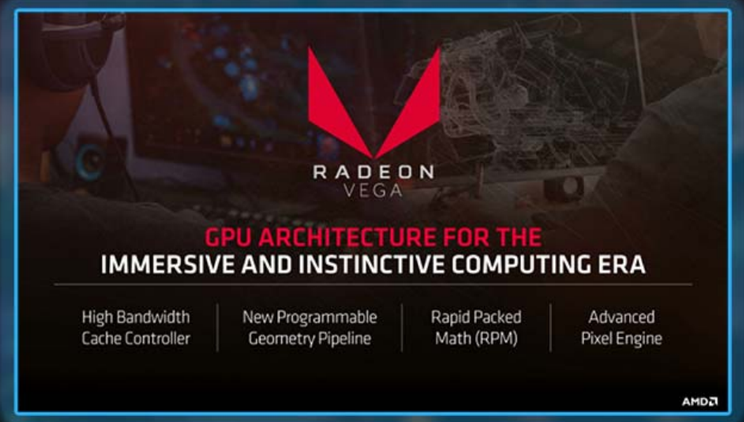
ROCm Architecture

ROCm is AMDs open-source HPC / Hyperscale-class platform for GPU computing that's also programming-language independent, touted as the being first in the world. This was introduced to compete with NVIDIA's CUDA platform, inline with AMDs Heterogeneous System Architecture (HSA) concepts, beliefs and foundation.
Of particular interest for AI and DL is it's inclusion of their library for high performance machine learning primitives, MIOpen.
It introduces support for Convolution Neural Network (CNN) acceleration on top of the ROCm stack. Two programming models are supported, namely, OpenCL - Open Computing Language framework for programming of CPUs, GPUs and others for execution across heterogeneous platforms. And, HIP - For converting CUDA code to portable C++ that can be run on NVIDIA and AMD GPUs.



CUDA to GPU flow

Vega 7nm GPU



Radeon Vega Architecture

At the recent Computex IT tradeshow, AMD announced its next generation Vega Graphics Processing Unit (GPU) based on 7nm technology. Intended for data-centers, it is clear they are aimed at high-performance DL and ML applications. How it will do this is by including 32GB of second-generation High Bandwidth Memory (HBM2) integrated into a multi-chip package, integrated AMD Infinity Fabric interface, and new deep learning instruction set operations. Details of the new supported DL instruction set operations have not been provided, but the intention is for training and inference applications to be accelerated, at lower power. ML models will be able to be directly compiled to Vega 7nm machine code via their GPU computing platform, ROCm.

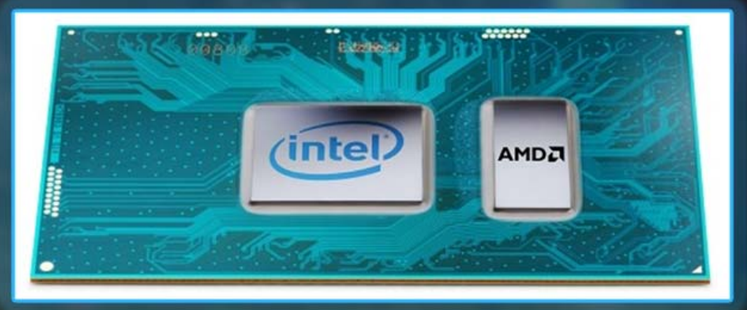
AMD SenseMI Technology

AMD Ryzen processors have a set of learning and adapting features that allow them to customize their performance to their application, thanks to machine intelligence. One of these features is “Neural Net Predication”, where the processor employs an artificial neural network to anticipate the next steps of the workflow in real time. This can result in boosted performance by steering workload down the most efficient paths inside the processor.

AMD SenseMI

Intel & AMD AI Chip

Intel & AMD announced a partnership in November 2017 to develop a new laptop chip in order to take on market leader NVIDIA. This is of marked importance, given the long history of competition and disputes between the two chip giants.
Intel will develop the Central Processing Unit (CPU) of the chip, while AMD will develop the Graphics Processing Unit (GPU).
While it is touted to be “thin and lightweight but powerful enough to run high-end video games”, GPUs are being used more and more for AI due to their highly parallel data processing capabilities.



intel and AMD chip

Tesla & AMD

There were rumours in September 2017 that Tesla were partnering with AMD to develop AI chips for their self driving cars. They were later dispelled, and Elon Musk confirmed they were developing their own chips in December 2017.

IBM's core AI offerings are based around its cloud-based Watson. This is the same basic technology which powered the Watson supercomputer that won a Jeopardy competition several years ago.

The Watson services are focused around language processing and understanding. Among the services that are included are natural language classifier, a natural language understanding engine, text to speech and speech to text, tone analyser, personality insights and language translator.

The natural language classifier can identify the topics in sentences or short pieces of text.


```
{
  "classifier_id": "10D41B-nlc-1",
  "url": "https://gateway.watsonplatform.net/natural-language-classifier/api/
  "text": "How hot will it be today?",
  "top_class": "temperature",
  "classes": [
    {
      "class_name": "temperature",
      "confidence": 0.9998201258549781
    },
    {
      "class_name": "conditions",
      "confidence": 0.00017987414502176904
    }
  ]
}
```

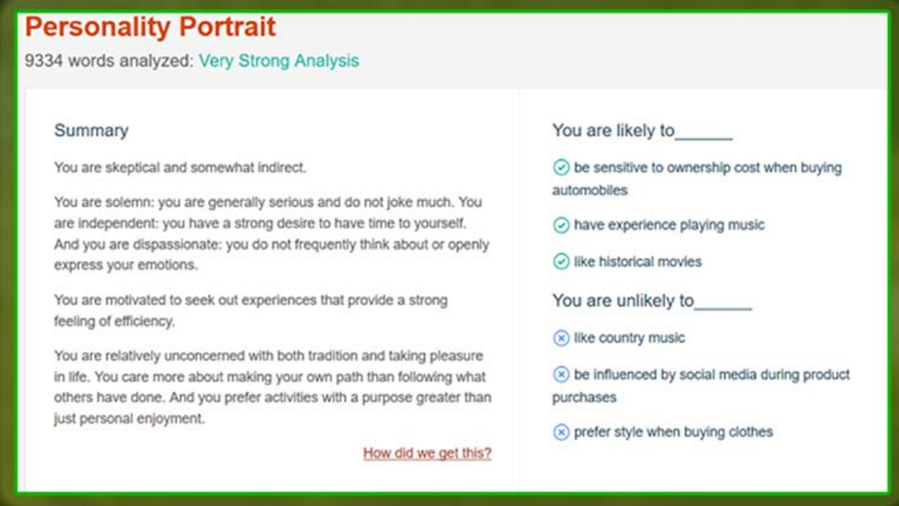

Example of the natural language classifier output

The natural language understanding engine is more flexible. It does sentiment analysis of short sections of text, keyword identification and can classify the topic of the text into a number of categories and subcategories.

The text to speech and speech to text are fairly standard offerings, converting a piece of text into a .wav file or transcribing an audio stream.

The tone analyser is designed to process either generic text or dialog, in text form. In the generic text form, it can return the sentiment of the entire text, or the sentiment of each sentence. In dialog form, it will analyse the sentiment of each piece of dialog.

Personality insight derives insights into the personality of individuals based on their social media postings or other digital communications.



Example output of the tone analyser

And finally, the language translator can translate text between 62 different languages, at the time of writing.

The newest of the Watson services is Watson Assistant, announced at the THINK 2018 conference. Watson Assistant is an enhancement of a previous service called Conversation. It's designed to support conversation flow along with natural language processing. It's not intended as a competitor to Alexa or Cortana, but is instead a service for businesses to create similar experiences across mobile apps, websites and other platforms.

Intel have traditionally been pioneers in designing and manufacturing computer chips, more specifically Central Processing Units(CPUs) and Graphic Processing Units(GPUs) with a wealth of knowledge is this area. Intel have embarked on creating infrastructure and purpose specific chips for training and processing machine learning and deep learning models with the goal of bringing AI to edge devices like autonomous cars, drones, and other IoT hardware.

Intel AI Cloud

Intel have created a developer cloud for AI processing. This is available to developers, data scientists, students, and start-ups for free. The Intel AI Cloud runs on Intel Xeon processors with 24 cores and two-way hyperthreading with access to 96 GB of RAM, and provides 200 GB of storage capacity per user. Jobs are run in a queue so there may be a waiting time.

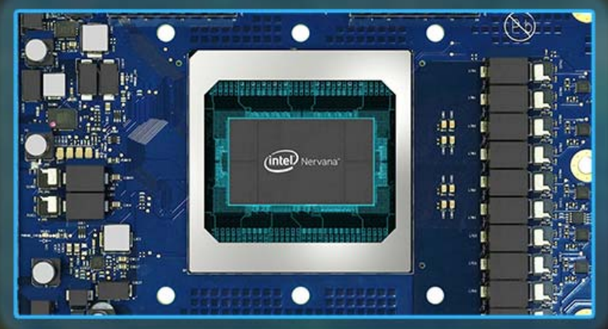
The Intel AI Cloud is based on Python and optimised for Intel architecture, this deep learning framework is designed for ease of use and extensibility on modern deep neural networks, such as AlexNet, Visual Geometry Group (VGG), and GoogLeNet.

The Intel AI Cloud also has optimisations for Theano, Caffe, Tensorflow, and Keras. For simpler AI programs, the Intel AI Cloud also supports higher level abstracted libraries including SciPy, NumPy, and Scikit-Learn.

Image processing using Intel app zoo

Neural Network Processor

The Neural Network ASIC (Application Specific Integrated Circuit) chip was also announced last year. No benchmarks were released, and Intel just said it would be available to select customers. At the time of writing this, we know is that it contains 12 cores based on its “Lake Crest” architecture. It has a total of 32GB memory, has a performance of 40 TFLOPS at an undisclosed precision, and a theoretical bandwidth of less than 800 nanoseconds for 2.4 Terabits per second of high bandwidth for low-latency interconnects.



The Intel Nervana chip

Intel VPU for Edge Deep Learning

Intel's Myriad X VPU is the third generation and most advanced VPU from Movidius. Intel's Myriad X VPU is the first of its class to feature the Neural Compute Engine - a dedicated hardware accelerator for deep neural network inferences.

The Neural Compute Engine in conjunction with the 16 powerful cores provides on-device deep neural networks and computer vision applications. Intel's Myriad X VPU has received additional upgrades to imaging and vision engines including additional programmable SHAVE cores, upgraded and expanded vision accelerators, and a new native 4K ISP pipeline with support for up to 8 HD sensors connecting directly to the VPU.

This processor provides wide capabilities for deep learning computing on the edge for devices in drones, robotics, smart wearables, smart security, augmented reality, and virtual reality.

The Intel Movidius Myriad X

Intel Movidius Neural Compute

The Intel Movidius Neural Compute Stick (NCS) is a tiny fanless deep learning device that can be used to learn AI programming at the edge. NCS is powered by the same low power high performance Intel Movidius Vision Processing Unit (VPU) that can be found in millions of smart security cameras, gesture controlled drones, industrial machine vision equipment, and more. The Intel Movidius stick is aimed at learning and developing code which will be compatible with all other AI hardware offerings from Intel. Intel have also acquired Nervana and Saffron along with Movidius to power the future of neural and cognitive processing. The Movidius stick is compatible with Ubuntu and Raspbian on Raspberry Pi.

Although a single stick is typically used, people have created Intel Movidius neural compute stick clusters

Intel Self-driving Car

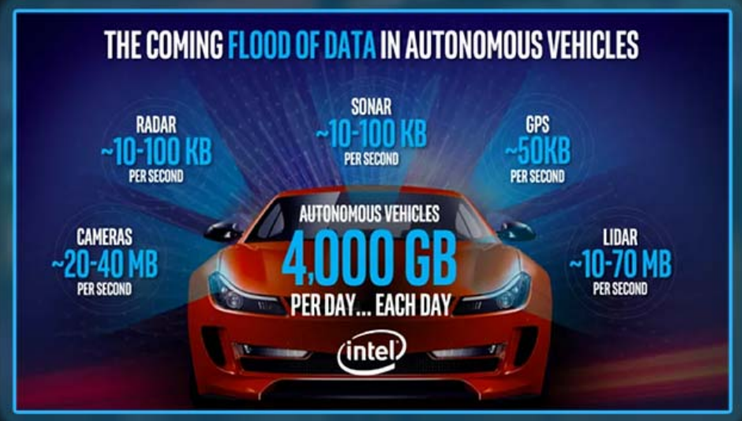
Intel introduced the first autonomous car in its 100-vehicle test fleet during CEO Brian Krzanich's keynote address at the industry expo CES in Las Vegas. The car features 12 cameras, radars, laser scanners and computing technologies from Mobileye and Intel. There are three high-resolution cameras at the front of the vehicle that allow for a 180-degree field of view and lets the car's image processor see at a distance of up to 300 meters.

Mobileye, which makes software for autonomous driving, was purchased by Intel last year for about $15 billion. Back then, Intel also announced that it planned to build a fleet of 100 highly automated vehicles to test in the U.S., Europe and Israel.

Self driving cars are in focus at CES this year. Start-ups, tech companies and automakers are racing to carve out their share of the nascent market. On Sunday, Nvidia and Uber announced that the ride-hailing service will use the former's chips for an artificial intelligence computing system for a fleet of self-driving cars.

Intel added on Monday that about two million vehicles from car makers BMW, Nissan and Volkswagen will use technology from Mobileye to build high-definition maps throughout 2018. Those maps would then be used by autonomous vehicles for navigation.

The U.S. firm also announced partnerships with Chinese automaker SAIC Motor and digital mapping company NavInfo, to develop automated vehicles, and map roads in China. Currently, the chip maker is also part of tech giant Baidu's open source autonomous driving project, called Apollo.



Intel's autonomous car statistics

NVIDIA

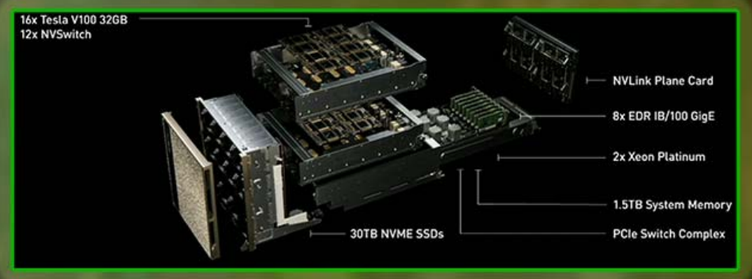
Nvidia has historically been a powerhouse for designing and manufacturing high performance graphic cards to support state of the art games, however with recent developments in computing for artificial intelligence, deep learning, and cryptocurrency mining, the GPU architecture proved to be more suited for computing in these environments in comparison to traditional Central Processing Units(CPUs). Nvidia has also dedicated research teams to tackle problems in computer vision, and image composition spaces using artificial intelligence techniques. Furthermore, a new partnership with ARM will integrate Nvidia Deep Learning Accelerator technology into ARM-based IoT chips.

Graphic Processing Units (GPUs)

In 2018 Nvidia announced several new GPUS - the Quadro GV100 and the DGX-2 workstation.

The Quadro GV100 is a 10 000 core processor. The DGX-2 is a 2 peta FLOP processor with 512 gigabytes of high-bandwidth memory.

The DGX-2 is 10x faster than the DGX-1 that was released 6 months ago. Examples in model training time further demonstrate the impact of GPU advances on the research and practice of deep learning in AI. 5 years ago, it took 6 days to train Alexnet (A popular image recognition model) using 2x Nvidia GTX 580s. With the DGX-2, it takes just 18 minutes.



The Nvidia DGX-2

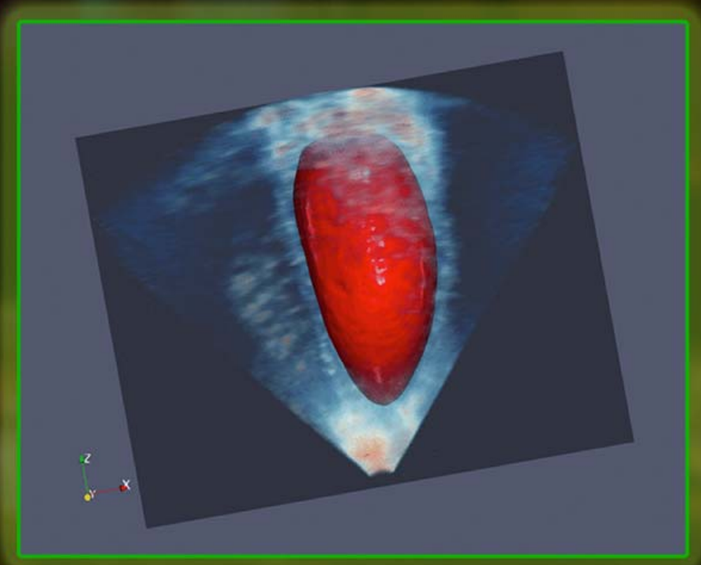
Project Clara

Nvidia's Project Clara is a medical imaging supercomputer. Medical imaging instruments have been vital to early detection and improvement of patient outcomes for more than four decades. Innovation in the field has come from improvements in detector technology and, more recently, parallel computing.

A decade ago, researchers realized Nvidia GPUs provide the most efficient architecture for medical imaging applications and could help reduce radiation exposure, improve image quality and produce images in real time. More recently, deep learning is dominating, with more than half of all new research in medical imaging applications involving AI.

Clara is virtual: it can run many computational instruments simultaneously. Clara is remote: it leverages Nvidia vGPUs to enable multi-user access. Clara is universal: it can perform the computation for any instrument, whether CT, MR, ultrasound, X-ray or Mammography. And Clara is scalable: it uses Kubernetes on GPUs to efficiently scale compute with demand.

Dozens of healthcare companies are working with them, including startups and research hospitals. Their AI applications like AutoMap and V-Net bring intangible value to radiology.



This 3D ultrasound depicts the left ventricle of the heart segmented by V-Net, a fully convolutional 3D neural network, running on a Tesla V100 GPU

Autonomous Driving & Sensor Fusion

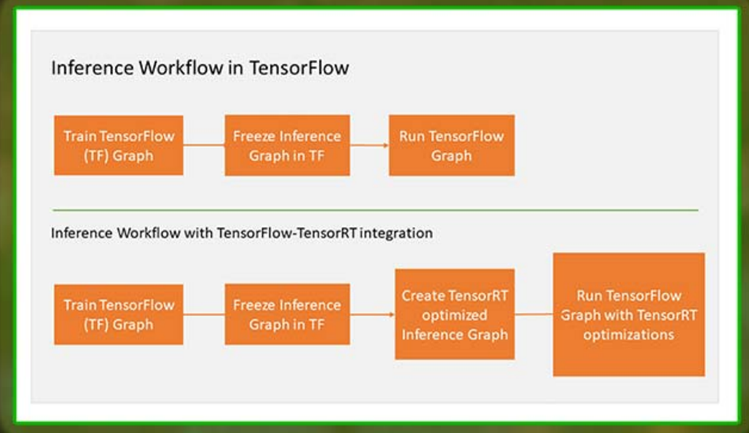
Obtaining data and designing algorithms that fuse input from multiple sensors and sources required for autonomous driving is hard. Nvidia is developing a platform called Nvidia Drive that helps address these issues. This will allow for self driving car researchers to better handle sensory data from varying instruments to create better performing AI in cars.

The Nvidia Drive platform combines deep learning, sensor fusion, and surround vision to change the driving experience. It is capable of understanding in real-time what's happening around the vehicle, precisely locating itself on an HD map, and planning a safe path forward. Designed around a diverse and redundant system architecture, the platform is built to support ASIL-D, the highest level of automotive functional safety.

An example of the NVIDIA DRIVE feature layers

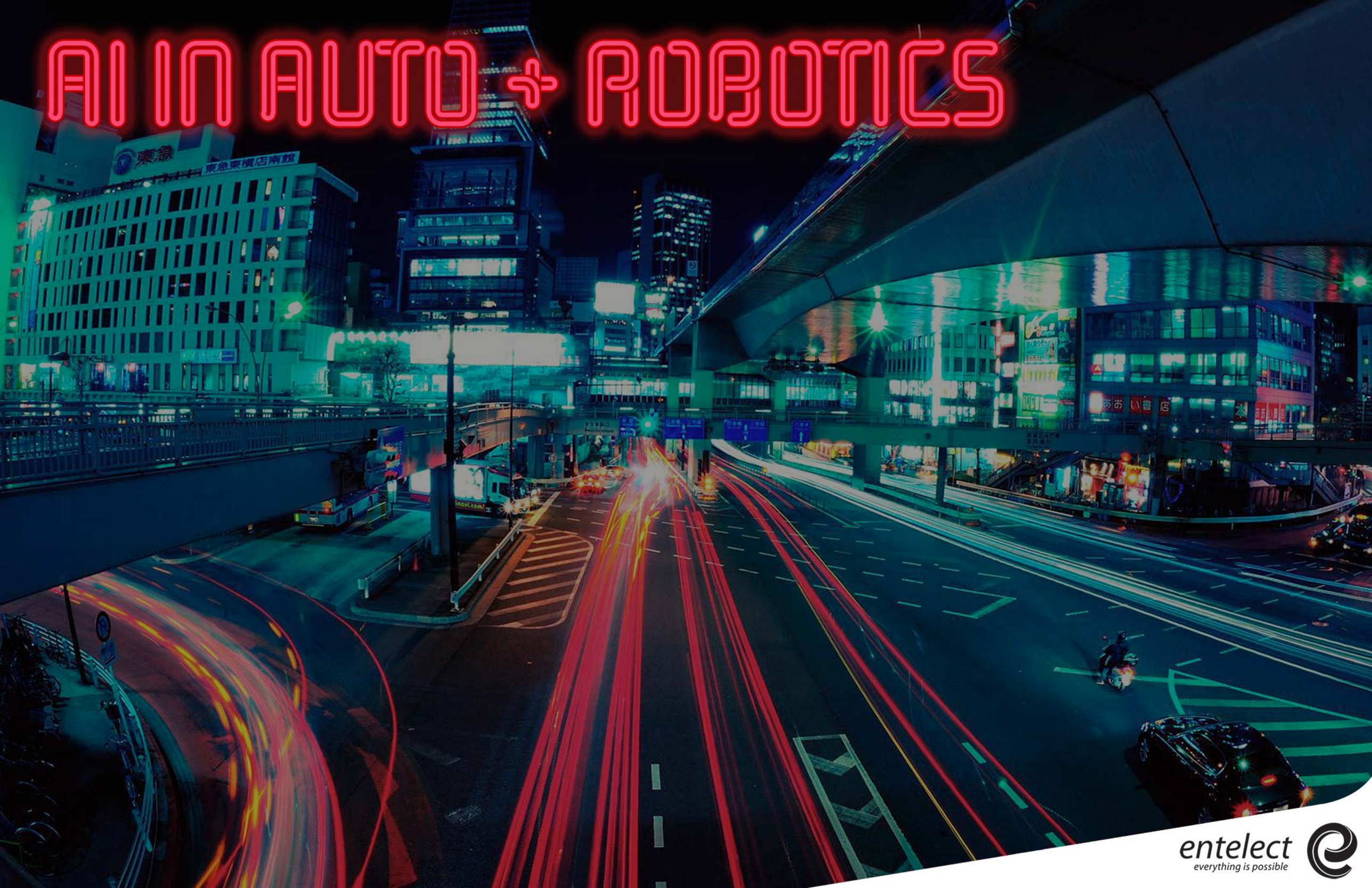
TensorFlowRT

Nvidia announced the integration of the TensorRT inference optimization tool with TensorFlow. TensorRT integration will be available for use in the TensorFlow 1.7 branch. TensorFlow remains the most popular deep learning framework today while Nvidia TensorRT speeds up deep learning inference through optimizations and high-performance runtimes for GPU-based platforms. Nvidia wish to give TensorFlow users the highest inference performance possible along with a near transparent workflow using TensorRT. The new integration provides a simple API which applies powerful FP16 and INT8 optimizations using TensorRT from within TensorFlow. TensorRT sped up TensorFlow inference by 8x for low latency runs of the ResNet-50 benchmark.



A comparison workflow between Tensorflow and TensorflowRT

AI IN AUTO + ROBOTICS



entelect
everything is possible

BOSTON DYNAMICS

Boston Dynamics builds advanced robots with remarkable behavior, mobility, agility, dexterity and speed, with their methodologies often shrouded by secrecy. The company primarily uses sensor-based controls and computation to unlock the capabilities of complex mechanisms with Atlas and Spotmini being two of their most notable robots currently being worked on. It also offers robots such as PETMAN, an anthropomorphic robot for testing equipment; Handle, a robot that combines the rough-terrain capability of legs with the efficiency of wheels and many others.

They have an extraordinary technical team of engineers and scientists who seamlessly combine advanced analytical thinking with bold engineering and boots-in-the-mud practicality. For the most part, their robots are a combination of expertly engineered systems and human operators, with recent focuses on enhancing the autonomy of their robots, more specifically, Atlas and SpotMini.

Left: The Atlas debuted in 2013, Right: Latest look and design of Atlas in 2016

Atlas is the latest in a line of advanced humanoid robots being developed. Atlas' control system coordinates motions of the arms, torso and legs to achieve whole-body mobile manipulation, greatly expanding its reach and workspace. Its ability to balance, while performing tasks, allows it to work in a large volume while occupying only a small footprint. Atlas was unveiled mid 2013 standing tall at 1.9m and tipping the scale at 150kg and was modeled on one of Boston Dynamics previous humanoid robot, PETMAN, along with its BigDog research. Over the years, Atlas has evolved and now stands at a height of 1.5m and weighs 75kg.

Like many of their robots, Boston Dynamics have been showcasing Atlas's amazing agility, balance, and control to the public using brief Youtube videos. Boston Dynamics uploaded a video showing Atlas balancing on one leg, jogging over rocks, and being hit with projectiles. It is designed to be able to operate both indoors and outdoors, across a range of terrain (including snow). Most recently we have seen Atlas flawlessly pull off a gymnastics routine that ends with a picture-perfect backflip.

SpotMini with extendable arm

SpotMini is a small four-legged robot and weighs in at 25kgs. SpotMini is all-electric and can go for about 90 minutes on a charge, depending on what it is doing and is the quietest robot built by Boston Dynamics to date. SpotMini inherits all of the mobility of its bigger brother, Spot, while adding the ability to pick up and handle objects using its 5 degree-of-freedom arm and beefed up perception sensors.

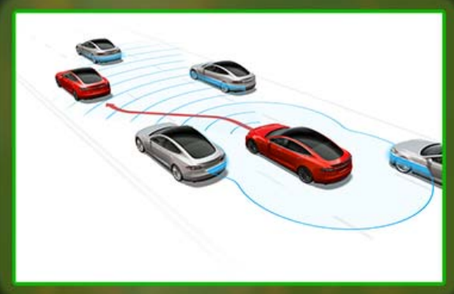
A video released in May 2018, helps explain how Boston Dynamics is getting SpotMini to operate autonomously. It is described in the video's description that an operator first manually drives the robot around its surroundings, as the machine captures the view with cameras on its sides, front and back. This allows SpotMini to map, and understand its environment before being unleashed to walk the same route again, autonomously.

Self-driving car companies get their machines rolling in much the same way. First, they map routes with LIDAR, which sprays its surroundings with lasers to build a 3D model of the world. That helps give the robocar a better understanding of its environment. The difference with SpotMini is that it's using stereo cameras instead of LIDAR. This allows it to visualize the world around it, similar to how we, as humans, see. This is achieved via finely tuned expert systems.

TESLA

Tesla is a company (co-founded by Elon Musk), that produces electric cars and other sustainable energy solutions and who have made amazing advances towards autonomous driving. Although full autonomy is still some time away, their vehicles currently support:

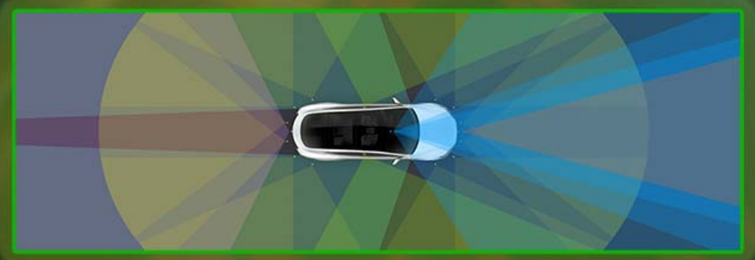
- Autosteer – Automatically keeps the vehicle within a lane (at highway speed).
- Traffic-Aware Cruise Control – Maintains the car's speed in relation to surrounding traffic.
- Auto Lane Change – By engaging the turn signal, the vehicle will automatically transition to an adjacent lane.
- Autopark – It detects parking spots and once activated, it will maneuver into the parking space by controlling vehicle speed, gear changes and steering.
- Summon – You can move the vehicle in and out of a parking space from outside the vehicle using the mobile app.



A visualisation of how a Tesla senses it's environment and performs a lane change

In the future, their vehicles are hoped to be capable of conducting trips with no action required by the person in the driver's seat, to the point of allowing you to rent out your vehicle while not in use - automatically joining a fleet of driverless taxis.

Unfortunately Tesla had some setbacks in the form of fatal accidents while autopilot was engaged, where drivers most likely became complacent and were not paying attention to the road, as the current terms of use require. But data shows that, when used properly, drivers supported by Autopilot are safer than those operating without assistance, and its active safety has prevented many accidents.



New Tesla vehicles have eight surrounding cameras that provide 360 degree visibility around the car, at up to 250 meters of range;welve ultrasonic sensors allowing for detection of both hard and soft objects, and a forward-facing radar with enhanced processing which provides additional data about the world on a redundant wavelength, that is able to see through heavy rain, fog, dust and even the car ahead

To make sense of all of this data, an onboard computer runs the Tesla-developed neural net for vision, sonar and radar processing software. Together, this system provides a view of the world that a driver alone cannot access, seeing in every direction simultaneously and on wavelengths that go far beyond the human senses.

Tesla has a fleet of hundreds of thousands of customer-owned vehicles that test autonomous technology in "shadow-mode" during their normal operation (these are not autonomous vehicles yet).Tesla is able to use billions of miles of real-world driving data to develop its autonomous technology. In "shadow mode," features run in the background without actuating vehicle controls, in order to provide data on how these features would perform in real world and in real time conditions. This data allows Tesla to safely compare self-driving features, not only to the existing Autopilot advanced driver assistance system, but also to how drivers actually drive in a wide variety of road conditions and situations.

The current self driving hardware is on the Nvidia Drive PX2 AI computing platform, but they are also working on custom AI processing chips that Tesla think will enable them to build the best AI hardware in the world.

UBER

Uber started as a simple idea: What if you could request a ride from your phone? More than 5 billion trips later, they are working to make transportation safer and more accessible, helping people order food quickly and affordably, reducing congestion in cities by getting more people into fewer cars, and creating opportunities for people to work on their own terms.

In addition to helping people get from point A to point B, Uber is working to bring the future closer with self-driving technology and urban air transport, helping people order food quickly and affordably, removing barriers to healthcare, creating new freight-booking solutions, and helping companies provide a seamless employee travel experience.

Uber's Advanced Technologies Group is ambitious about transforming the way the world moves. With locations in Pittsburgh, San Francisco, Phoenix and Toronto, the Advanced Technologies Group is comprised of Uber's self-driving engineering team dedicated to self-driving technologies, mapping, and vehicle safety. Their work doesn't end with transporting people, they are also developing self-driving truck technology to move goods more safely and cost effectively around the world.

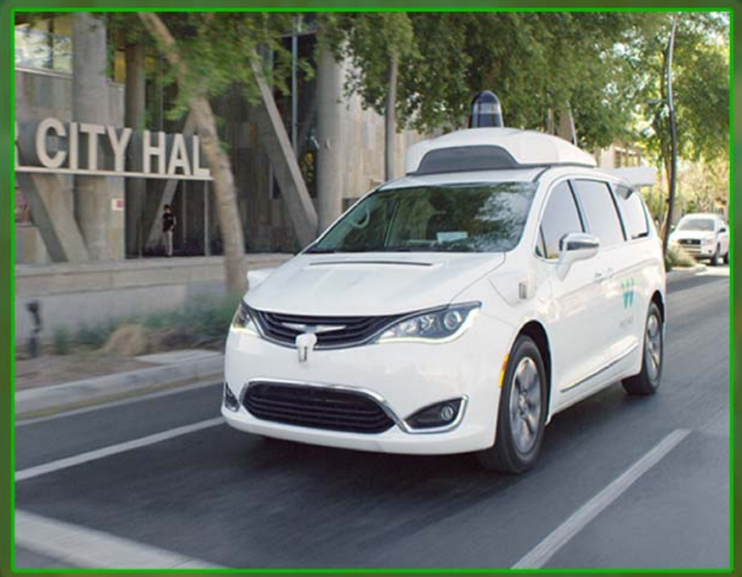
Uber has hundreds of self-driving cars in the world and they are pretty hard to miss. It uses a system of 360 degree cameras, radars and lasers that scan the environment to be aware of traffic signs, people, vehicles etc. Nvidia supply the GPUs that help power Uber's autonomous tech. Currently they require a specially trained vehicle operator to make sure the vehicle does what it is supposed to. The vehicles are used to gather map data that is used to update the software.

Unfortunately apart from minor accidents, an Uber self-driving car had a well publicised fatal accident which resulted in the self-driving fleet operations to be suspended for several months.

An Uber self-driving car

WAYMO

Waymo is a self-driving technology company with a mission to make it safe and easy for people and things to move around. Waymo began as the Google self-driving car project in 2009. Today, it is an independent self-driving technology company without the need for anyone in the driver's seat.



Waymo's fully self-driving Chrysler Pacifica Hybrid minivan on public roads

Their vehicles have sensors and software that are designed to detect pedestrians, cyclists, vehicles, road works and more, from up to three football fields away, in 360 degrees.

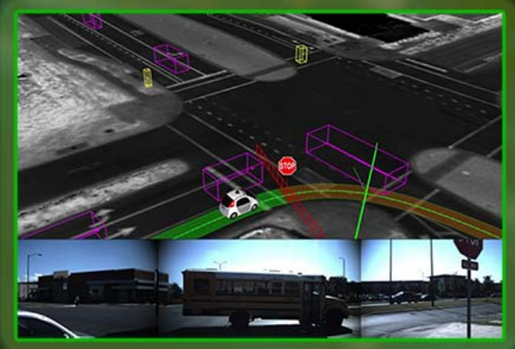
Their cars use LIDAR sensors to scan their environment: it measures distance to a target by illuminating the target with pulsed laser lights and measuring the reflected pulses with a sensor.

They drive more than 25000 autonomous miles (40000 km) each week, largely on complex city streets. They also did 4.3 billion km's of simulated driving in 2017 (akin to unit testing for self driving cars).

Waymo uses an automated process and human labelers to train its neural nets. After they've been trained, these giant datasets also need to be pruned and shrunk so they can be deployed into the real world in Waymo's vehicles.

This process, akin to compressing a digital image, is key in building the infrastructure to scale to a global system.

As part of Alphabet, Waymo uses Google's data centers to train its neural nets. Specifically, it uses a high-powered cloud computing hardware system called "tensor processing units," which underpins some of the company's most ambitious and far-reaching technologies. Typically, this work is done using commercially available GPUs, often from Nvidia. But over the last few years, Google has opted to build some of this hardware itself and optimize for its own software. TPUs are "orders of magnitude" faster than CPUs.



Their vehicles have sensors and software that are designed to detect pedestrians, cyclists, vehicles, road works and more, from up to three football fields away, in 360 degrees

AI RESEARCH

entelect
everything is possible

DEEPMIND

Deepmind is a company that has been acquired by Alphabet (Google) which specializes in research, focussed on AI and Machine Learning.

Google

Deepmind have provided Machine Learning techniques for real-world impact to Google from managing Data Centers to Recommendations on Google Play. Deepmind's work with Google has helped take fundamental research and apply them to real world products in a few months.

The first of the more successful projects Deepmind did for Google, is improving efficiency of the Google data centres. By evaluating billions of possible actions that could be taken and making recommendations for operators based on predicted power usage effectiveness, the project helped make savings up to 40% from cooling and an overall 15% improvements in building efficiency. In addition, the AI helped discover novel methods of cooling which is now being applied at the data centers.

Starcraft 2 feature layer API

In 2016, Deepmind introduced WaveNet, a new deep neural network that can produce realistic-sounding speech. Initially the model was a research prototype and too computationally expensive for consumer products. After 12 months of working with Google Text to Speech and Deepmind research teams, they created a new model that was a 1000 times faster than the original. Wavenet is currently used in the Google Assistant and Google Cloud platform and enables other products to generate voices via Google Cloud text-to-speech.

AlphaGo

AlphaGo is the first computer program to defeat the following Go players, a professional, a world champion and the strongest human player in history. While playing, AlphaGo performed a handful of highly inventive winning moves that were so surprising they overturned hundreds of years of wisdom, and have since been examined by players of all levels. Essentially, AlphaGo taught the world new moves in the game of Go. In January 2017, an improved AlphaGo version was revealed as the online player "Master" which had 60 straight wins in online fast time-control games.

In May 2017, AlphaGo took part in the Future of Go summit in China, to explore the mysteries of Go. Five months later, AlphaGo Zero was published in a Nature paper.

Unlike the earlier versions, which learned how to play from thousands of games played by professional and amateur players, Zero learnt to play by playing games against itself. Using this approach it surpassed the performance of all previous versions, becoming the best Go player of all time.

Starcraft 2

Deepmind have partnered with Blizzard Entertainment in order to open Starcraft 2 as a AI research platform. Testing agents in games, not specifically designed for AI research and where humans play well, provides a benchmark for agent performance. Starcraft and Starcraft 2 are among the biggest and most successful games of all time, with tournaments being played for more than 20 years. Starcraft already provides a environment for AI and ML researchers with the annual AIIDE bot competition.

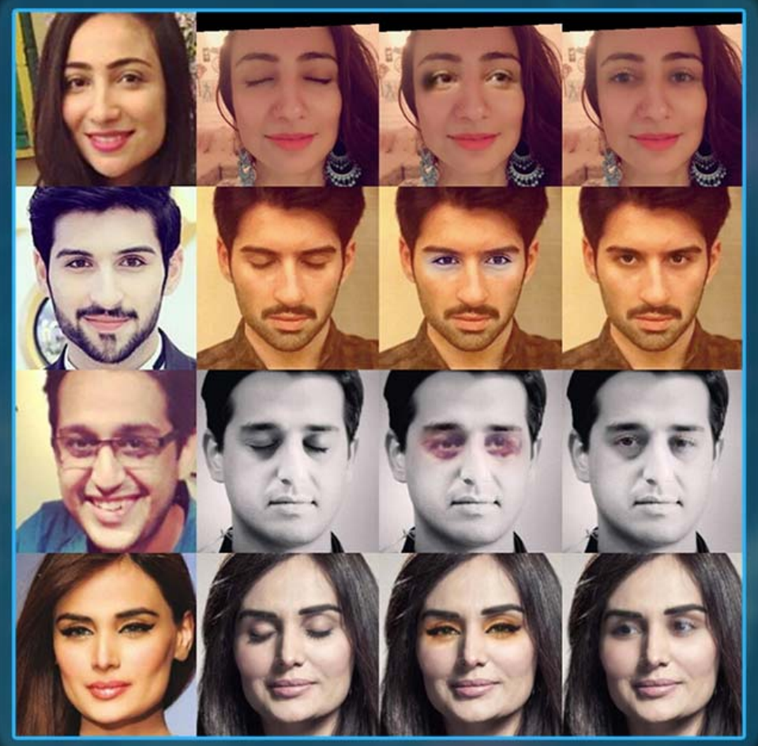
Starcraft 2 is an appealing environment for AI research, with a single objective such as beating the opponent, the player will need to achieve sub goals such as gathering resources and building structures. The length of a match can take from a few minutes to one hour, meaning that early actions may only have long term pay-off. Finally, the map is only partially observed, requiring the agents to use a combination of memory and planning to succeed.

FACEBOOK

AI at Facebook is primarily driven by two complementary teams, the research arm called Facebook AI Research (FAIR), led by one of the founders of the Convolutional Neural Network (CNN), Yann LeCun, who works on pushing the state-of-the-art in the field, and the engineering arm called Applied Machine Learning (AML) who focus more heavily on applications. The teams work across the full spectrum of topics related to AI such as computer vision, natural language processing, reinforcement learning, as well as the creation of production grade AI platforms.

Computer Vision

Their computer vision systems use deep convolutional networks with billions of parameters to process the 1B+ images and video uploaded to Facebook everyday. It's used to predict the content of an image, automatically take down offensive content, generate captions for the blind, and improve search results. Some of their latest work involves improving their image recognition networks by exploiting user-supplied hashtags on images to function as labels. This allows them to scale their training sets into the billions of images. Another interesting bit of work is their solution to replacing closed eyes with open ones because you know, there's always that one person who blinks in a group photo.



From left to right. Exemplar images, source image, Photoshop's eye-opening algorithm, Facebook's ExGAN method. As you can see, Photoshop's method simply pastes in the exemplar images eyes while the ExGAN is able to factor in the source images lighting and colour even if the exemplar image was taken under different conditions

The real break through here is that they're able to achieve this without any obvious colour mismatches or artifacts. They do this by using a Generative Adversarial Network which is essentially a machine learning system that tries to fool itself into thinking it's creations are real. In a GAN, one part of the system learns to recognise some object, let's say eyes in this example, and another part of the system repeatedly creates images that, based on feedback from the recognition part, gradually grows in realism. What Facebook does here is to use a special type of GAN called an Exemplar GAN (ExGAN) where you include exemplar data showing the target person with their eyes open to give the GAN extra information to allow it's output to be personalised, as opposed to creating generic human eyes.

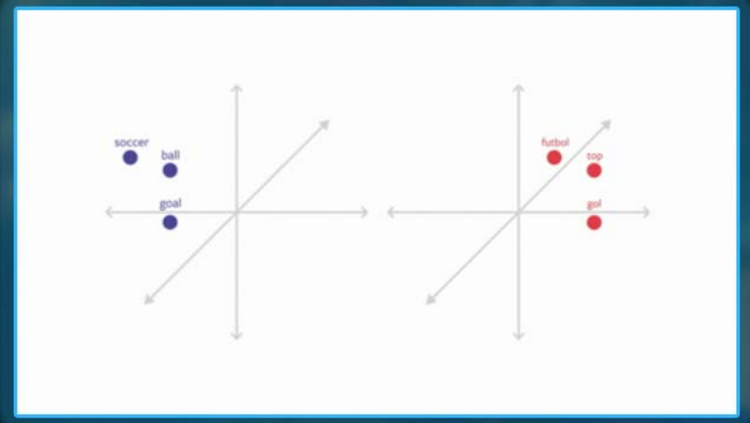
Natural Language Processing

Their language work mainly focuses on trying to eliminate language barriers using deep learning with their systems serving over 2B translations of posts across 48 unique languages every single day. About a year ago though, there was quite the media storm about Facebook panicking and pulling the plug on an AI that supposedly invented it's own secret language.

Unfortunately, the truth was far less remarkable. The research that sparked this frenzy was about creating agents to negotiate in natural language with each other and reach an agreement that maximised their individual rewards. While training, they simply only updated one agents parameters at a time otherwise, as expected, it would lead to a "divergence from human language". An example of this divergence is as follows:

Bob: i can i i everything else . . . . . .
Alice: balls have zero to me to me to me to me to me to me to me to me to
Bob: you i everything else . . . . . .
Alice: balls have a ball to me to me to me to me to me to me to me
Bob: i i can i i i everything else . . . . . .
Alice: balls have a ball to me to me to me to me to me to me to me
Bob: i . . . . . .
Alice: balls have zero to me to me to me to me to me to me to me to me to
Bob: you i i i i i everything else . . . . . .
Alice: balls have 0 to me to me to me to me to me to me to me to me to
Bob: you i i i everything else . . . . . .
Alice: balls have zero to me to me to me to me to me to me to me to me to

Some of their more recent work involves developing multi-lingual word embeddings to enable text classification across languages. Word embeddings (which are basically just vectors) have nice properties that make them easy to operate on. They include the property that words with similar meanings are close together in vector space. Traditionally, word embeddings have been language-specific, with embeddings for each language trained separately and existing in different vector spaces. What this work attempts to do is to make the embeddings exist in the same vector space and maintain the property that words with similar meanings, regardless of language, are close together in vector space. The implication of this work is that you can now train on one or more languages, and learn a classifier that works on languages not seen in training.



The blue dots indicate the word embeddings in English while the red dots indicate those same word embeddings in Spanish. In traditional approaches these embeddings exist in different vector spaces as can be seen in the image above even though these words mean the same thing. This research manages to make the embeddings exist in the same vector space regardless of the language.

FACEBOOK

Reinforcement Learning

Inspired by Deepminds work in creating AlphaGo Zero (which beat AlphaGo Lee 100-0), the team attempted to reproduce that effort with OpenGo using their Extensible Lightweight Framework (ELF)11. Their goal was to create an open source implementation of a system (Deepmind have yet to release their code) that would teach itself how to play Go at the level of a professional human player or better. So far, the ELF OpenGo agent has delivered some remarkable results beating LeelaZero (another open source effort to replicate AlphaGo Zero) 198-2 and 14-0 against four of the top 30 world-ranked human Go players.

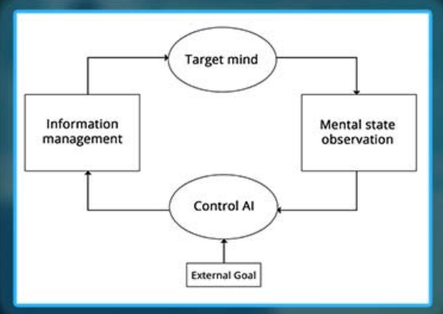
Tooling

On the tooling front, Facebook is developing PyTorch, their answer to Google's TensorFlow. At the moment, at v0.4.0, PyTorch is mainly used and favoured by researchers for it's ease of use, speed, and extensibility. However, it falls short when it comes to running in production at scale, on mobile platforms, or for providing for configurable options, which is where TensorFlow shines but at the expense of usability. However, Facebook has decided to marry PyTorch to Caffe2, their current production-ready platform, in order to get PyTorch to V1 and in a production ready state, all the while maintaining their simple and easy to use API.

Ethics

Now, at this point, you may be wondering why Facebook are investing so much time and effort into developing these technologies and the answer is obvious, to maximise ad revenue. But at what cost? They've already shown that they're not exactly the most responsible or ethical with this technology. Back in 2014, they ran an experiment to see if they could manipulate user's moods by feeding them either mainly positive or negative content. Even though only a small change was observed, they were able to accomplish this. What this demonstrates though is something far more worrying.

As Francois Chollet points out, they're increasingly able to measure everything about us and the information we consume. As shown in the below diagram, this enables them to establish an optimisation loop for human behaviour. Your current mental state is observed and information is continuously fed to you until the desired behaviour (the external goal) is achieved.



In the above diagram, a Control AI feeds information to a user, observes their mental state, and tunes the information to send to them the next time continuously until a prespecified goal is achieved. For example, as shown as possible by Facebook in the past, an AI would continuously feed a user mostly negative content until the user started posting negative thoughts and ideas of their own

What makes this particularly dangerous is that we don't know what these goals are. It could be something benign like showing you a relevant ad that you're likely to click on, or it could start showing you which political articles you should read, nudging you in a particular direction on who you should vote for. Now it's not all doom and gloom, there are positive signs. Recently, lawmakers are making an attempt to reign in companies like Facebook with the summonsing of Mark Zuckerberg to the US Congress and EU parliament earlier this year and the passing of the GDPR. However, it's still not enough. We as consumers need to start demanding that the goals optimised are made transparent. Furthermore, we should also be allowed to tune these goals so that they serve us and align to the goals and objectives we set for ourselves. There's still time to get this right and if we do, this kind of AI can lead to empowering individuals to gain greater control over their lives. If we don't though, will we even realise it?

OPEN AI

OpenAI is a non-profit research company (co-founded by Elon Musk) working to build safe artificial intelligence and ensure that AI's benefits are as widely and evenly distributed as possible. OpenAI's mission is to build safe Artificial General Intelligence (AGI). They expect AI technologies to be hugely impactful in the short term, but their impact will be outstripped by that of the first AGIs.

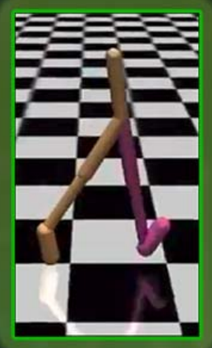
While doing research projects, they often tie in short-term goals to practical projects that allows them to identify new problems and test ideas against measurable, defined objectives.

Some things they work on are Gym, Competitive Multi-Agent Environments and Household robots.

Gym is a toolkit for developing and comparing reinforcement learning algorithms. It supports teaching agents, everything from walking to playing games like Pong or Pinball. It is an open source interface to reinforcement learning tasks.

The gym library provides an easy-to-use suite of reinforcement learning tasks.They provide the environment; you provide the algorithm.

Fork it on github: https://github.com/openai/gym



A demonstration of Deep Deterministic Policy Gradient on the Gym Hopper environment after 2000 iterations, where the algorithm learns how to "walk"

Solving OpenAI Lunar Lander with reinforcement learning

Household Robots

Open AI is working to enable a physical robot (off-the-shelf; not manufactured by OpenAI) to perform basic housework. There are existing techniques for specific tasks, but they believe that learning algorithms can eventually be made reliable enough to create a general-purpose robot. More generally, robotics is a good testbed for many challenges in AI.

To date they have built a robot arm system that can learn a behavior from a single demonstration, delivered within a simulator, then reproduce that behavior in different setups in reality. The system is powered by two neural networks: a vision network and an imitation network.

The vision network ingests an image from the robot's camera and outputs state representing the positions of the objects. The vision network is trained with hundreds of thousands of simulated images with different perturbations of lighting, textures, and objects (The vision system is never trained on a real image).

The imitation network observes a demonstration, processes it to infer the intent of the task, and then accomplishes the intent starting from another starting configuration. Thus, the imitation network must generalise the demonstration to a new setting.

Next, they are releasing eight simulated robotics environments, and a Baselines implementation of Hindsight Experience Replay, all developed for their research over the past year. They used these environments to train models which work on physical robots.

OPEN AI

Competitive Multi-Agent Environments

Open AI created a bot which beats the world's top professionals at 1v1 matches of Dota 2 under standard tournament rules. The bot learnt the game from scratch by self-play, and does not use imitation learning or tree search. This is a step towards building AI systems which accomplishes well-defined goals in messy, complicated situations involving real humans.

Open AI's Dota 2 bot in action

Open AI found that self-play allows simulated AIs to discover physical skills like tackling, ducking, faking, kicking, catching, and diving for the ball, without explicitly designing an environment with these skills in mind. Self-play ensures that the environment is always the right difficulty for an AI to improve on. They have increasing confidence that self-play will be a core part of powerful AI systems in the future.

CONTRIBUTORS

Amrit Purshotam

AI to me is simply Intelligence Augmentation (IA if you will) where data and computation is used to create services that augment human intelligence and creativity and not the overhyped nonsense that's thrown around by the media about various doomsday scenarios and the impending AI apocalypse.

Greg Schroder

AI to me is the next step in automating tasks that humans either don't want to do, or aren't as efficient or precise at as a machine would be. As more and more AI practises become mainstream and routine, they enable the next steps of automation, and aren't considered in the realm of AI any longer. "AI is whatever hasn't been done yet" – Douglas Hofstadter

Privolin Naidoo

AI is a practical approach to understanding the world around us. Creating intelligence and understanding how it works, helps give us a better understanding of our own intelligence. If we can successfully create 'true AI' and understand how it works, then surely we will be capable of understanding our own intelligence and maybe even our own consciousness.

Duane McKibbin

Artificial intelligence is a very overloaded term. I think many companies are jumping on the bandwagon and finding ways to sprinkle some AI and ML into their products, and in many cases it feels a bit forced. But through all the noise, there are definitely some very promising use cases and I'm excited for the future where AI enhances our day to day lives.

Luke Madzedze

AI is the science of mimicking the working of the human brain and at the same time, complimenting it with the versatility, reliability, and accuracy of a computer.

Rishal Hurbans

AI is finding usefulness in the data around us using algorithms to push technology and humanity forward.

Gail Shaw

Artificial intelligence is an ever-changing field, if only because once computers become good at something we stop calling it AI. We call it things like 'automatic captioning' or 'digital assistants'. It's a field that's changed massively in the last few years alone, and I look forward to seeing what will be possible in 5-10 years time.

Marius Kruger

AI is currently an interesting problem to solve: get computers to be more helpful than we can manually program them. Although advanced AI in the wrong hands or with the wrong fitness function could destroy or enslave us, I'm hopeful that we can work towards a peaceful co-existence.

Steven Carter

AI is probably the most exciting and game changing discipline of the 21st century. What we have already learned about our world by teaching Intelligent programs has been beyond our imagination.

REFERENCES

AMD

[1] https://medium.com/intuitionmachine/amds-open-source-deep-learning-strategy-14c228be6248
[2] https://rocm.github.io/dl.html
[3] https://rocmsoftwareplatform.github.io/MIOpen/doc/html/install.html
[4] https://www.khronos.org/opencl/
[5] https://github.com/ROCm-Developer-Tools/HIP
[6] https://www.forbes.com/sites/tiriasresearch/2018/06/06/amd-plugs-ml-into-upcoming-vega-7nm-gpu/#32480e3a7ba7
[7] https://www.amd.com/en/technologies/sense-mi
[8] https://www.cnbc.com/2017/09/21/wall-street-is-gushing-over-amd-on-its-a-i-chip-relationship-with-tesla.html
[9] https://www.reuters.com/article/us-tesla-chips/globalfoundries-says-no-commitment-from-tesla-on-chip-deal-idUSKCN1BW259
[10] https://www.theregister.co.uk/2017/12/08/elon_musk_finally_admits_tesla_is_building_its_own_custom_ai_chips

Apple

[1] https://www.imore.com/apple-and-its-future-artificial-intelligence
[2] https://developer.apple.com/documentation/create_ml
[3] https://developer.apple.com/documentation/coreml
[4] https://www.thecarconnection.com/news/1117308_apple-nabs-ex-google-waymo-engineer-to-boost-its-self-driving-car-efforts
[5] https://www.theverge.com/2018/1/25/16932716/apple-expand-fleet-self-driving-cars-california-lexus
[6] https://www.nytimes.com/2018/05/23/technology/apple-bmw-mercedes-volkswagen-driverless-cars.html
[7] https://www.imore.com/apple-and-its-future-artificial-intelligence
[8] https://www.wired.com/story/apples-plans-to-bring-artificial-intelligence-to-your-phone/
[9] https://www.zdnet.com/article/wwdc-2018-apple-debuts-create-ml-for-simple-machine-learning-training/
[10] https://analyticsindiamag.com/apple-wwdc-2018-7-things-around-ml-announced-this-year/
[11] https://www.cultofmac.com/555331/add-custom-siri-shortcuts/
[12] https://techcrunch.com/2016/06/13/apple-image-and-facial-recognition/

Amazon

[1] https://www.wired.com/story/amazon-artificial-intelligence-flywheel/
[2] https://www.geekwire.com/2017/jeff-bezos-explains-amazons-artificial-intelligence-machine-learning-strategy/
[3] https://developer.amazon.com/alexa-skills-kit
[4] https://aws.amazon.com/machine-learning/
[5] https://docs.aws.amazon.com/aws-technical-content/latest/aws-overview/artificial-intelligence-services.html
[6] https://github.com/gluon-api/gluon-api/
[7] https://www.amazon.com/Amazon-Prime-Air/
[8] https://aws.amazon.com/blogs/aws/introducing-gluon-a-new-library-for-machine-learning-from-aws-and-microsoft/
[9] https://aws.amazon.com/machine-learning/amis/

REFERENCES

Boston Dynamics

[1] https://www.bostondynamics.com/robots
[2] https://www.bostondynamics.com/spotmini
[3] https://www.bostondynamics.com/atlas
[4] https://www.digitaltrends.com/cool-tech/milestones-in-atlas-robot-history/
[5] https://www.youtube.com/watch?v=SD6Okylclb8 (Atlas Update)
[6] https://www.youtube.com/watch?v=fRj34o4hN4I (What's New, Atlas ?)
[7] https://www.youtube.com/watch?v=Ve9kWX_KXus (SpotMini Autonomous Navigation)

Deep Mind

[1] https://deepmind.com/research/alphago/match-archive/master/
[2] https://deepmind.com/research/alphago/alphago-china/
[3] https://www.nature.com/articles/nature24270.epdf

Facebook

[1] https://www.facebook.com/ylecun/posts/494009107472870
[2] http://blog.kaggle.com/2014/12/22/convolutional-nets-and-cifar-10-an-interview-with-yan-lecun/
[3] https://www.quora.com/What-are-the-most-interesting-things-Facebook-is-doing-in-ML-research/answer/Joaquin-Qui%C3%B1onero-Candela
[4] https://code.facebook.com/posts/457605107772545/under-the-hood-building-accessibility-tools-for-the-visually-impaired-on-facebook/
[5] https://code.facebook.com/posts/1700437286678763/advancing-state-of-the-art-image-recognition-with-deep-learning-on-hashtags/
[6] https://www.skynettoday.com/briefs/facebook-chatbot-language/
[7] https://code.facebook.com/posts/1686672014972296/deal-or-no-deal-training-ai-bots-to-negotiate/
[8] https://thenextweb.com/artificial-intelligence/2017/06/19/facebooks-ai-accidentally-created-its-own-language/#.tnw_hJk6Xc8i
[9] https://code.facebook.com/posts/550719898617409/under-the-hood-multilingual-embeddings/
[10] https://research.fb.com/wp-content/uploads/2018/06/Eye-In-Painting-with-Exemplar-Generative-Adversarial-Networks.pdf?
[11] https://research.fb.com/facebook-open-sources-elf-opengo/
[12] https://pytorch.org/2018/05/02/road-to-1.0.html
[13] http://www.pnas.org/content/111/24/8788.full
[14] https://medium.com/@francois.chollet/what-worries-me-about-ai-ed9df072b704

REFERENCES

IBM

[1] https://console.bluemix.net/developer/watson/documentation
[2] https://www.ibm.com/blogs/watson/
[3] https://www.ibm.com/watson/

Intel

[1] https://software.intel.com/en-us/ai-academy/tools/devcloud
[2] https://www.theregister.co.uk/2018/05/23/intels_first_commercial_ai_chips_will_arrive_in_2019/
[3] https://www.intel.co.za/content/www/za/en/automotive/autonomous-vehicles.html
[4] https://ai.intel.com/intel-nervana-neural-network-processor/
[5] https://ai.intel.com/
[6] https://www.movidius.com/myriadx
[7] https://developer.movidius.com/

Microsoft

[1] https://en.wikipedia.org/wiki/Field-programmable_gate_array
[2] https://blogs.microsoft.com/ai/project_brainwave_catapult_moonshot/
[3] https://blogs.microsoft.com/ai/build-2018-project-brainwave/
[4] https://azure.microsoft.com/en-us/services/cognitive-services/custom-vision-service/
[5] https://customvision.ai/
[6] https://labs.cognitive.microsoft.com/
[7] https://www.microsoft.com/en-us/aiforearth
[8] https://www.microsoft.com/net/learn/apps/machine-learning-and-ai/ml-dotnet

Nvidia

[1] https://devblogs.nvidia.com/tensorrt-integration-speeds-tensorflow-inference/
[2] https://blogs.nvidia.com/blog/2018/03/28/ai-healthcare-gtc/
[3] https://towardsdatascience.com/a-review-of-nvidia-gtc-2018-conference-new-gpus-deep-learning-acceleration-data-augmentation-d6d4f638bcda
[4] https://www.nvidia.com/en-au/ai-conference/
[5] https://www.nvidia.com/en-us/design-visualization/quadro-vdws/
[6] https://nvidianews.nvidia.com/in-the-news?year=2018

REFERENCES

OpenAI

[1] https://openai.com/about/#mission
[2] https://gym.openai.com/
[3] https://blog.openai.com/competitive-self-play/
[4] https://blog.openai.com/openai-technical-goals/#goal2
[5] https://blog.openai.com/robots-that-learn/
[6] https://blog.openai.com/ingredients-for-robotics-research/

Tesla

[1] https://www.tesla.com/presskit
[2] http://www.dailymail.co.uk/sciencetech/article-4223450/Dubai-transport-authority-agrees-buy-200-Tesla-vehicles.html
[3] https://jalopnik.com/the-feds-are-unhappy-that-tesla-released-info-on-fatal-1824252520
[4] http://www.theweek.co.uk/electric-cars/93596/elon-musk-tesla-autopilot-system-prevented-severe-injury-in-60mph-crash
[5] https://www.inverse.com/article/45489-zedd-tesla-tweet
[6] https://www.youtube.com/watch?v=FrJ2uPRRtz0
[7] https://www.youtube.com/watch?v=5sicOh6LPBw
[8] https://www.theregister.co.uk/2017/12/08/elon_musk_finally_admits_tesla_is_building_its_own_custom_ai_chips
[9] https://www.theverge.com/2017/12/8/16750560/tesla-custom-ai-chips-hardware
[10] https://en.wikipedia.org/wiki/List_of_autonomous_car_fatalities
[11] https://electrek.co/2018/01/31/tesla-autopilot-autonomous-driving-test/
[12] https://electrek.co/2018/03/28/tesla-autopilot-2-5-computer-model-3-s-x-first-look/
[13] https://www.teslarati.com/tesla-autopilot-free-trials-on-ramp-off-ramp/

Uber

[1] https://www.uber.com/en-ZA/about/
[2] https://www.uber.com/newsroom/media-assets/
[3] https://www.youtube.com/watch?v=27OuOCeZmwI
[4] https://www.theverge.com/2018/5/7/17327682/uber-self-driving-car-decision-kill-swerve
[5] https://www.theverge.com/2018/5/9/17338010/uber-resume-self-driving-car-test-dara-khosrowshahi

Waymo

[1] https://waymo.com/press/
[2] https://www.theverge.com/2018/5/9/17307156/google-waymo-driverless-cars-deep-learning-neural-net-interview
[3] https://en.wikipedia.org/wiki/Lidar

entelect

everything is possible

Enterprise Technology Solutions

Digital • Automation • Data • Mobility • Consulting

www.entelect.co.za | solutions@entelect.co.za

2018