vertex

AI-Optimized Chipsets

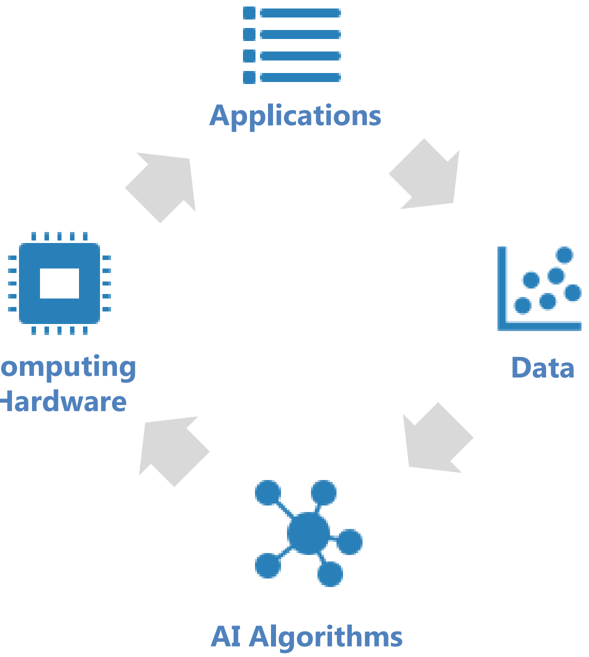
Part I: Key Drivers

Mar 2018

Businesses are increasingly adopting AI to create new applications, driving the development of AI-optimized chips

The ADAC (Applications – Data – Algorithms – Computing Hardware) **Loop**

1 Businesses are increasingly adopting AI to create new applications to transform existing operations. These include connected devices, autonomous vehicles, on-device personal interfaces, voice interactions and AR.



2 Up to 30 billion more IoT devices are coming online by 2020, streaming data that helps build smarter objects, homes, inform consumer lifestyle, enhance security and energy management.

3 Most breakthrough approaches in deep learning use significant computing power. A neural net might have dozens of connected layers and billions of parameters, requiring a step-wise increase in level of computing power.

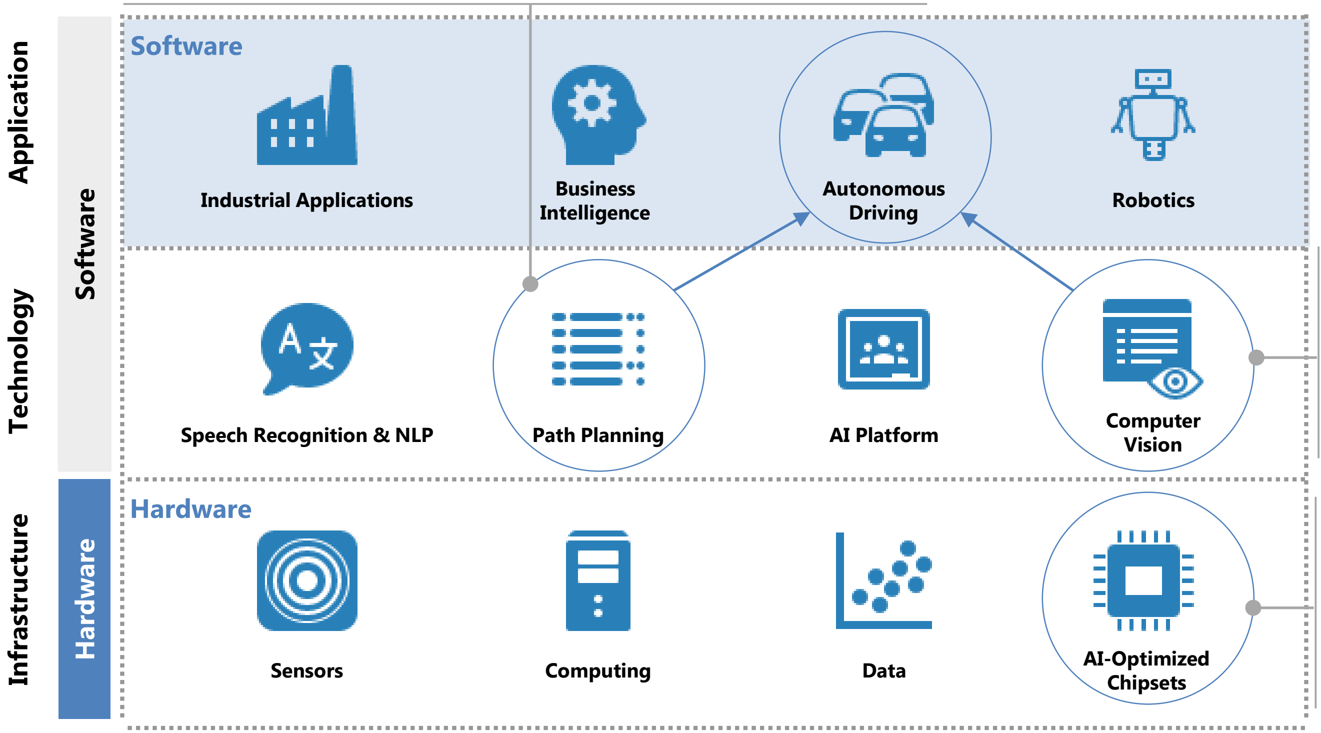
4 This positive, recursive ADAC loop where new applications generate more data, in turn enhancing algorithmic complexity, driving demand for higher computing performance.

Source: When Moore's Law Met AI by Azeem on Medium | icons8

vertex

These new applications are built on other technology and infrastructure layer solutions

Path planning: Simple machine learning algorithms are sufficient to handle driving in high resolution mapped cities or along fixed routes. Deep learning is more suitable in complex situations, (e.g. multiple unknown destinations or changing routes).



Application

Technology

Infrastructure

Software

Hardware

Software

Industrial Applications

Business Intelligence

Autonomous Driving

Robotics

Speech Recognition & NLP

Path Planning

AI Platform

Computer Vision

Hardware

Sensors

Computing

Data

AI-Optimized Chipsets

- **Sensing** uses advanced **computer vision** and **perception**.
- Visual tasks including lane detection, pedestrian detection, road signs recognition and blind-spot monitoring are handled more effectively with deep learning.

- To date, ***deep learning technology has primarily been a software play.***
- Existing processors were not originally designed for these new applications.
- Hence the need to develop AI-optimized hardware.

Examples of Vertex Portfolio Companies that employ deep learning in their solutions

TARANIS

Taranis offers a comprehensive and affordable crop management solution, and the pest and disease prediction algorithms **using deep learning to continually improve accuracy.**

KRYON SYSTEMS

Kryon Systems delivers innovative, intelligent Robotic Process Automation (RPA) solutions using **patented visual and deep learning technologies**.

Horizon Robotics

Horizon Robotics is the leader of embedded AI with leading technologies in autonomous driving perception and decision-making, **deep learning algorithms and AI processor architecture.**

vertex

Source: Vertex | AutomotiveIQ | Icon 8 | Taranis | Kryon Systems | Horizon Robotics

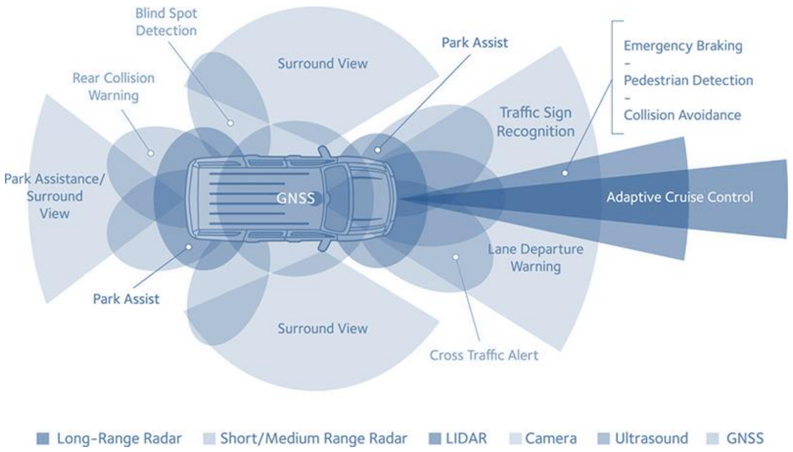
That may reside in the cloud, on edge devices or in a hybrid environment

	Edge Resident	Hybrid Solutions	Cloud Hosted
Consumer | Retail	• Gaming • Smart Displays	• Personal Assistants	• Ad Targeting & E-Commerce
Transportation	• Autonomous Vehicles	• Transportation & Grid Control	• Traffic & Network Analytics
Enterprise	• Delivery Drone • Warehouse Robots	• Cyber Security	• Sales, Marketing & Customer Services
Commodities	• Field Drones & Robots	• Climate, Water • Energy & Flow Control	• Field Sensor Data Analytics
Industrial | Military	• Cobots • Unmanned Systems	• Factory Control & Surveillance	• Factory & Operations Analytics
Healthcare	• Medical Imaging • Surgical Robots	• Medical Diagnostics	• Clinical Analytics

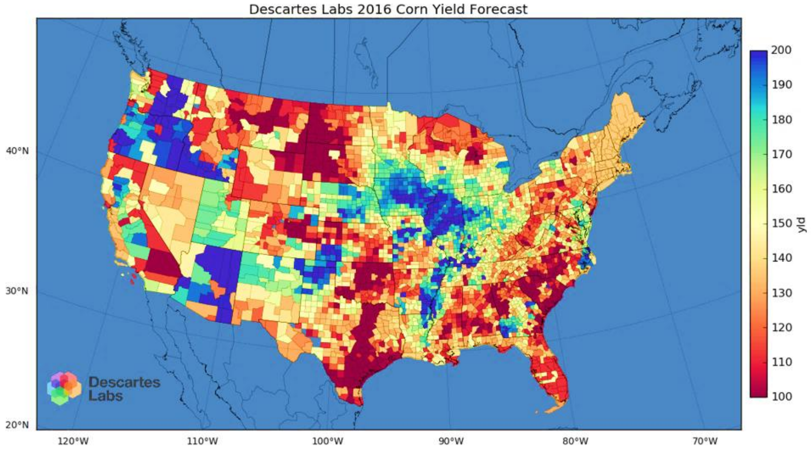
vertex

Source: : Moor Insight & Strategy

And all point to significantly higher data generation



Source: NovAtel



Source: Descartes Labs

Autonomous Vehicles

- In an autonomous car, cameras will generate between 20–60 MB/s, radar upwards of 10 KB/s, sonar 10–100 KB/s, GPS will run at 50 KB/s, and LIDAR will range between 10–70 MB/s.
- Each autonomous vehicle will be **generating approximately 8GB/s, 4TB per day**.
- Autonomous vehicles require a reliable solution with *an* **ultra-low latency of 1ms**.

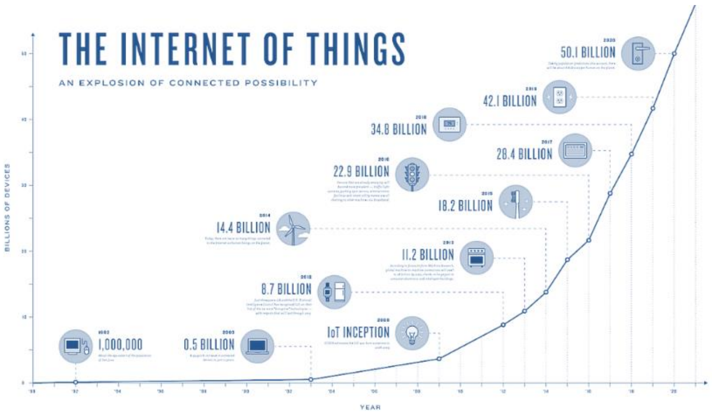
Agriculture

- Descartes Labs uses deep learning to process satellite imagery for agricultural forecasts.
- It **processes over 5TB** of new data every day and **references a library of 3PB** of archival satellite images.
- By using real time satellite imagery and weather models, Descartes Labs provides highly accurate weekly forecasts of US corn production compared to monthly forecasts provided by the US Department of Agriculture.

Source: : Intel | IEEE Spectrum, | Deep Learning: An Artificial Intelligence Revolution by Ark Invest | Descartes Labs | Reducing 5G Latency Benefits Automotive Safety by Bill McKinley | NovAtel

Coupled with the growth of IoT and 5G networks, a data deluge of high volume, velocity and variety is expected

IoT and Exponential Growth in Devices



Source: World Economic Forum

50B
Number of IoT devices by 2020

Top IOT Applications

- Smart Home
- Wearables
- Connected Industries
- Smart City
- Smart Energy
- Connected Car

The 5G Evolution: Latency for Different Generations of Cellular Networks

100KB/s	384KB/s-2MB/s	150KB/s-450MB/s	**10 GB/s**
500-1000 ms	200 ms	100 ms	**1 ms**
2G GSM | GPRS EDGE | CDMA 1990- 2000	3G UMTS CDMA 2000 2000-2010	4G LTE LTE-A 2010-2020	**5G** **>2020**

Source: Wi360

The growth of IoT and 5G networks expected to generate a data deluge of **high volume, velocity and variety**

- Volume
- Velocity
- Variety

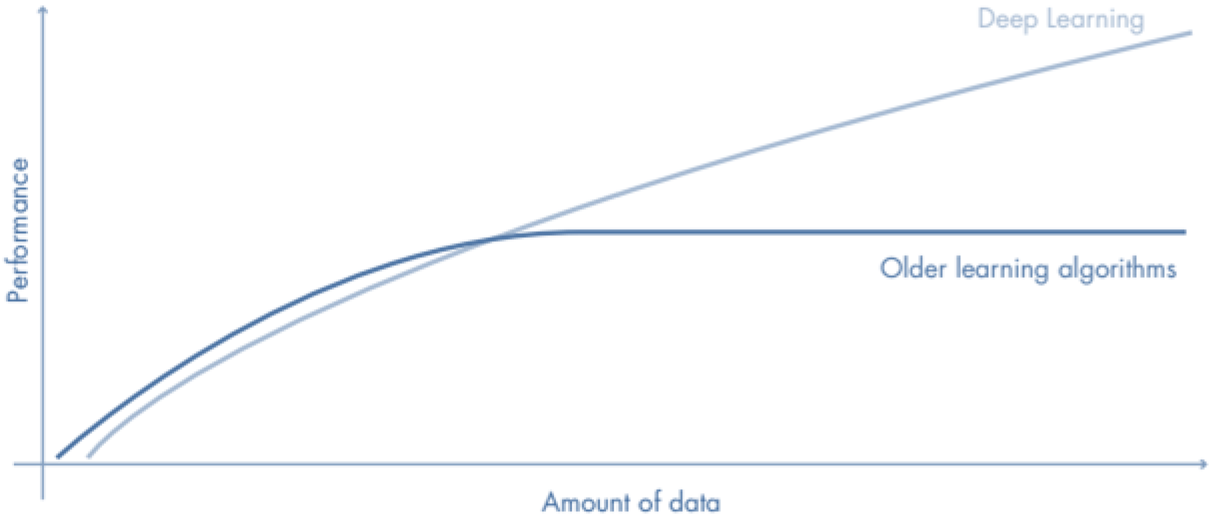
Source: Gartner

vertex

Source: : IoT Analytics | Intel | IEEE Spectrum, | Deep Learning: An Artificial Intelligence Revolution by Ark Invest | Wi360 | icon8

Unlike other machine learning algorithms, those associated with deep learning scale with increasing training data

- Compounding the power of deep learning, the neural nets themselves have become larger and more sophisticated, as measured by their number of free “parameters”.
- Parameters are dials used to tune the network’s performance. Generally, more parameters allow a network to express more states and capture more data.
- It endows computers with previously unimaginable capabilities - understanding photos, translating language, predicting crop yields, diagnosing diseases etc. **Enabling AI to write software to automate business processes that humans are unable to write**.



Source: Andrew Ng, Ark Invest

“The process could be very complicated…As a result of this observation, the AI software writes an AI software to automate that business process. Because we won’t be able to do it. It’s too complicated…

For the next couple of decades, **the greatest contribution of A.I. is writing software that humans simply can’t write**. Solving the unsolvable problems.”

Jensen Huang

CEO | NVIDIA

vertex

Source: Inside Microsoft's FPGA-Based Configurable Cloud by CTO Mark Russinovich | Nvidia | Deep Learning: An Artificial Intelligence Revolution by Ark Invest | Fortune

Given future process complexities, AI will be needed to automate the programming process by coding dynamically

Deep Learning vs. Other Programming Techniques

1980s Classic Programming

- Software developer **codes the solution in software**, which then gets executed in a deterministic and obtuse fashion.
- This works for simple, well-defined problems but breaks down for more complex tasks.

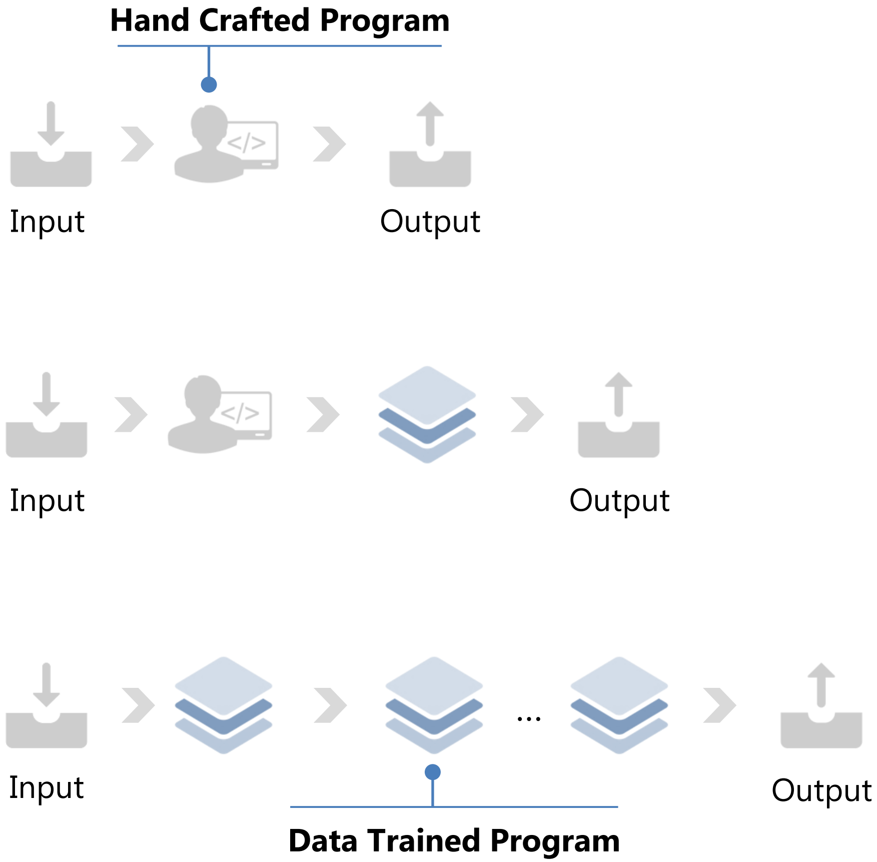
2000s Machine Learning

- Improves upon classic programming by replacing some stages of the program with **stages that can be trained automatically with data**
- Enabling computers to **perform more complex tasks** (e.g. image and voice recognition).
- The software developer focuses less on coding, more on **building models** which require **enormous datasets** to recommend a best output.

2010s Deep Learning

Entire program is replaced with stages that can be trained with data

- Programs can be far more capable and accurate.
- Requires less human effort to create.



Source: Ark Invest Management LLC, Yoshua Bengio

Source: Deep Learning: An Artificial Intelligence Revolution by Ark Invest | icon8

But existing processors were not originally designed for new AI applications. Hence the need to develop AI-optimized hardware

	Strengths	Limitations	Training Rank	Inference Rank	Leading Vendors
CPU	- General-purpose, in servers and PCs - Sufficient for inference	- Serial-processing is less efficient than parallel-processing	N.A.	N.A.	intel
GPU	- Highly parallel, high performance - Uses popular AI framework (CUDA)	- Less efficient than FPGAs - Scalability - Inefficient unless fully utilised	1	3	NVIDIA AMD
FPGA	- Reconfigurable - Good for constantly evolving workloads - Efficient	- Difficult to program, - Lower performance versus GPUs - No major AI framework	2	2	intel XILINX
ASIC	- Best performance, - Most energy and cost efficient - Fully customizable	- Long development cycle - Requires high volume to be practical - Quickly outdated, inflexible	3	1	intel Google

Source: Morningstar | Vertex | icon8

vertex

Looking ahead

This is the end of Part I of a 4-part series of Vertex Perspectives that seeks to understand key factors driving innovation for AI-optimized chipsets, their industry landscape and development trajectory.

In Part II, we review the shift in performance focus of computing from general application to neural nets and how this is driving demand for high performance computing. To this end, some startups are adopting alternative, novel approaches and this is expected to pave the way for other AI-optimized chipsets.

In Part III, we assess the dominance of tech giants in the cloud, coupled with disruptive startups adopting cloud-first or edge-first approaches to AI-optimized chips. Most industry players are expected to focus on the cloud, with ASIC startups featuring prominently in the cloud and at the edge.

Finally in Part IV, we look at other emerging technologies including neuromorphic chips and quantum computing systems, to explore their promise as alternative AI-optimized chipsets.

We are most grateful to Emmanuel Timor (General Partner, Vertex Ventures Israel) and Sandeep Bhadra (Partner, Vertex Ventures US) for their insightful comments on this publication.

Do let us know if you would like to subscribe to future Vertex Perspectives.

Source: Vertex

About Vertex Holdings

Vertex Holdings, a member of Temasek Holdings, focuses on venture capital investment opportunities in the information technology and healthcare markets, primarily through our global family of direct investment venture funds. Headquartered in Singapore, we collaborate with a network of global investors who specialize in local markets. The Vertex Global Network encompasses Silicon Valley, China, Israel, India, Taiwan and Southeast Asia.

Authors

Yanai ORON
General Partner
Vertex Ventures Israel
yanai@vertexventures.com

XIA Zhi Jin
Partner
Vertex Ventures China
xiazj@vertexventures.com

ZHAO Yu Jie
Associate Investment Director
Vertex Ventures China
zhaoyj@vertexventures.com

Brian TOH
Director
Vertex Holdings
btoh@vertexholdings.com

Tracy JIN
Director
Vertex Holdings
tjin@vertexholdings.com

Disclaimer

This presentation has been compiled for informational purposes only. It does not constitute a recommendation to any party. The presentation relies on data and insights from a wide range of sources including public and private companies, market research firms, government agencies and industry professionals. We cite specific sources where information is public. The presentation is also informed by non-public information and insights.

Information provided by third parties may not have been independently verified. Vertex Holdings believes such information to be reliable and adequately comprehensive but does not represent that such information is in all respects accurate or complete. Vertex Holdings shall not be held liable for any information provided.

Any information or opinions provided in this report are as of the date of the report and Vertex Holdings is under no obligation to update the information or communicate that any updates have been made.

vertex

THANK YOU