Week 3

Artificial Intelligence Program

Infrastructure and Architecture

> Agenda // Program

Assignments [50%]
EXAMS [40%]
C# Challenges [10%]

WEEK	SUBJECT	ASSIGNMENT / TO BE DELIVERED	CHALLENGES
2	Intro / AI Function / Enablers		
3	Infra and Architecture / On-prem vs. Cloud / CSPs		C 1
4	Data Pipeline / Processes / Framework / AutoML	#1 Image Classifier **[5%]**	
5	Data Pipeline / Processes / Framework / AutoML		C 2
6	More Data / SSIS / ADF / Data Quality	#2 Machine Learning Studio **[10%]**	
7	Azure services – Intro **EXAM 1 [20%]**		C 3
8	**READING WEEK**	**NO CLASSES**	
9	Azure Cognitive Services 1		4.1
10	Azure Cognitive Services 2	#3 Draw your own Architecture **[5%]**	4.2
11	Azure Cognitive Services 3		4.3
12	Azure Cognitive Services 4	#4 Azure Pipeline / Sentiment Analysis **[10%]**	4.4
13	AWS Academy // Cloud Foundations		
14	AWS Academy // Machine Learning	#5 AWS Academy // Cloud Foundations **[10%]**	
15	Enterprise Architecture **EXAM 2 [20%]**	#6 AWS Academy // Machine Learning **[10%]**	

aws academy

> Agenda (3)

- On prem vs. cloud
- SaaS, PaaS, IaaS
- AI Workstations / Servers
- AI Processing (CPU / GPU / ASIC)
- Storage Solutions
- Embedded Devices
- Edge Devices
- Infra Requirements

- Assignment #1 – Image Classifier

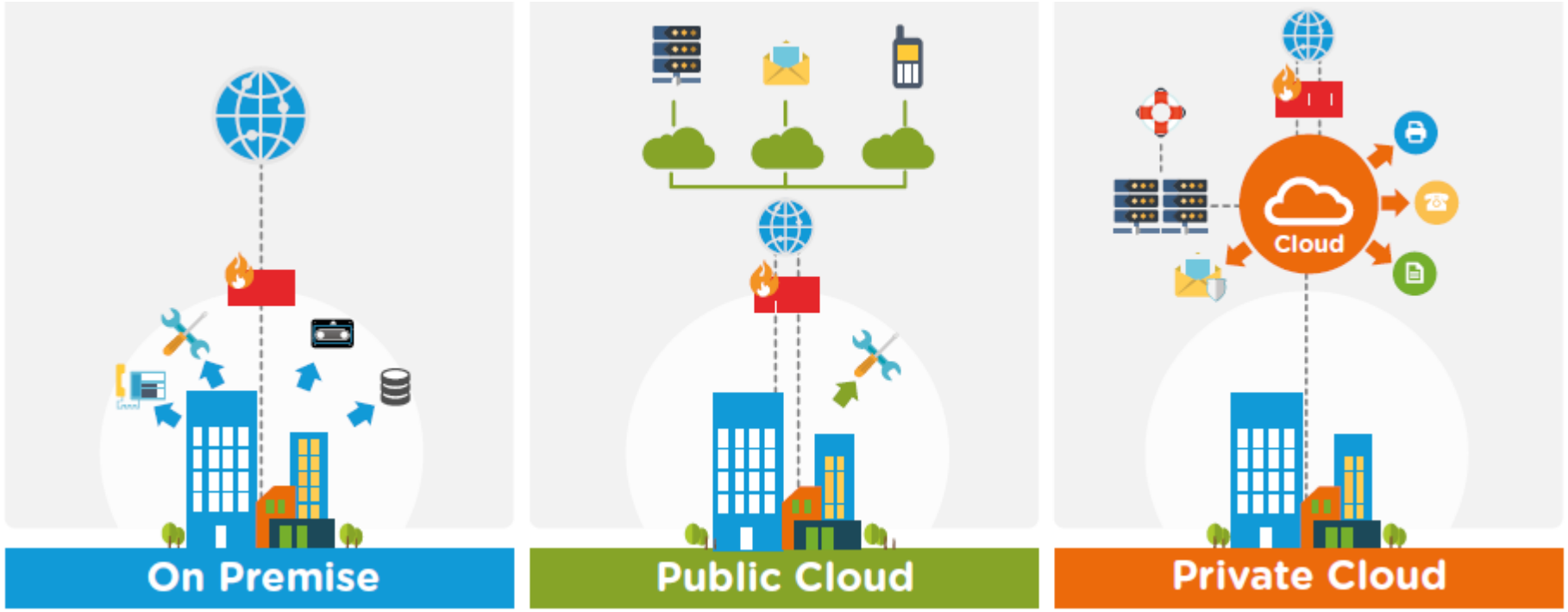
Programmer on vacation, still thinking about the bug

Infra and Architecture

On prem *vs.* Cloud

> On Premises vs. Cloud

> On Premises vs. Cloud



> On Premises vs. Cloud

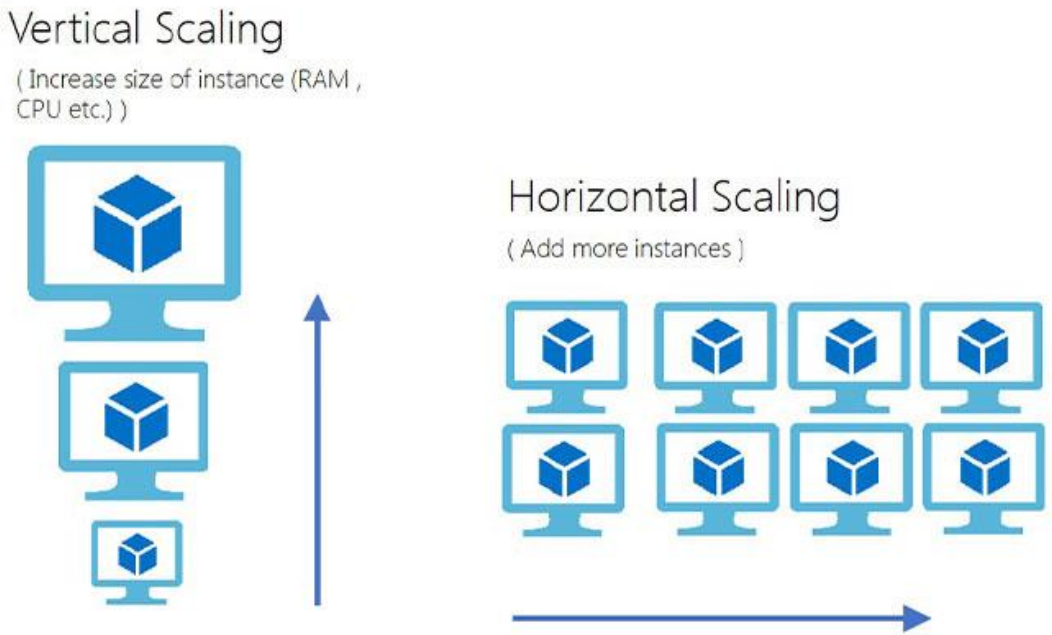
- **Computing Environment** —cloud-based systems offer more complete packaging in **provided physical and logical infrastructure to host services, virtual servers**, **intelligent applications**, and containers for their subscribers.
- **Licensing** — cloud-based systems offer **Pay per use** (Subscription Model) vs. on-premises need continuous maintenance of OS and software licenses.
- **Maintenance** —cloud-based systems offer 360 degree of key infrastructure services such as physical hardware, computer networking, firewalls and network security, data-center fault tolerance, compliance, and physical security of the buildings.

• **Availability** — cloud-based systems prove to be better because many services and platforms use SLAs to ensure that customers know the capabilities of the platform they are using.

• **Support** — cloud-based systems offer a variety of available data type support + standardization of the environment.

• **Total Cost of Ownership (TCO)** — cloud-based systems yield substantial economies of scale and lowers total cost of ownership (TCO).

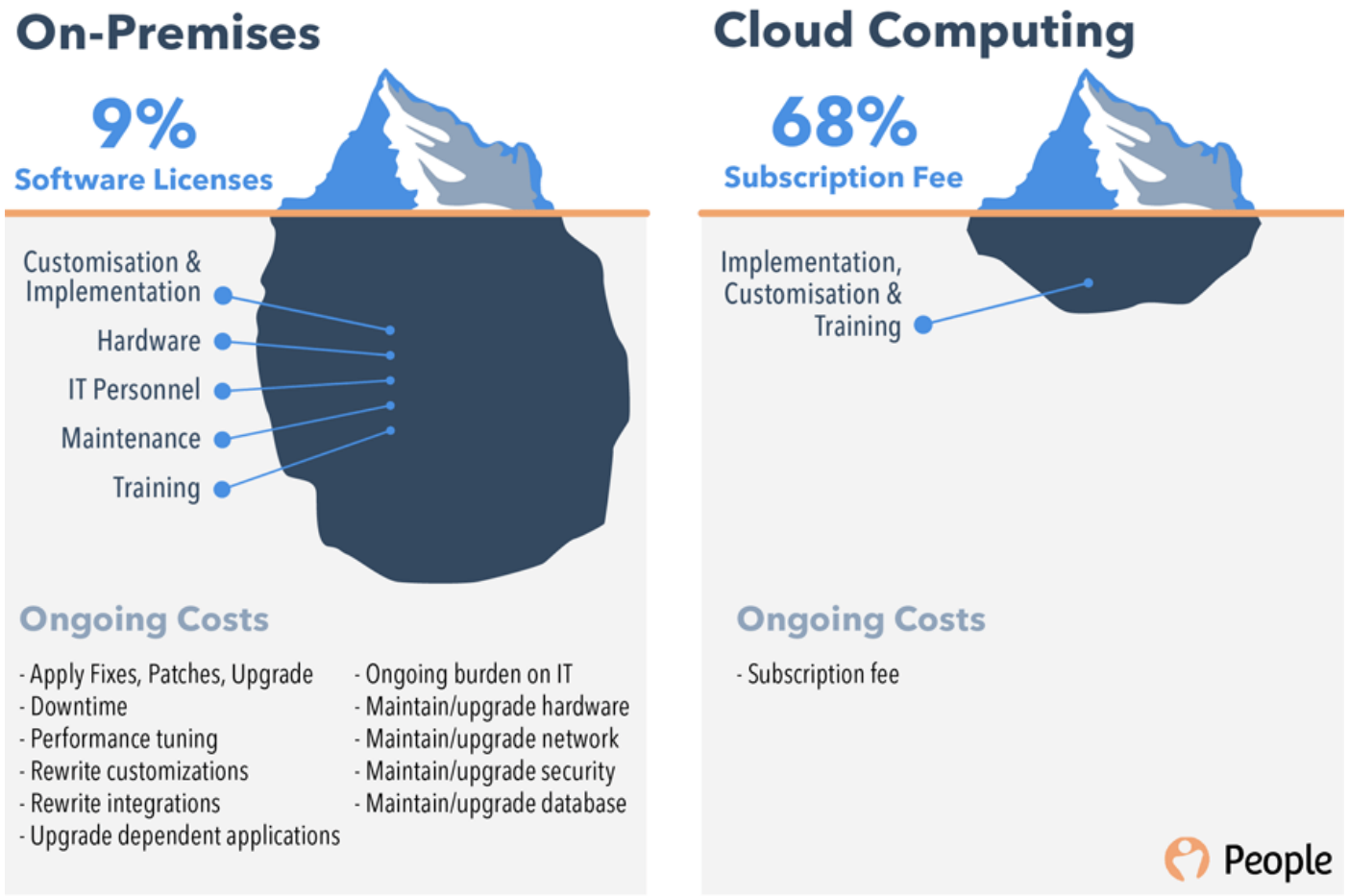
> On Premises vs. Cloud



• **Scalability** — Horizontal scaling means that you scale by adding more machines into your pool of resources whereas Vertical scaling means that you scale by adding more power (CPU, RAM) to an existing machine. On-premises are generally the scaling horizontally; server administrators add another server node to a cluster. The main disadvantage is that software has to handle all the data distribution + parallel processing complexities, and worse than that, limited number of software are available to take advantage of the horizontal scaling option. The vertical scaling, on the other hand, can easily manage and install hardware within a single machine despite the system to be incredibly powerful.

> On Premises vs. Cloud | Costs

+CAPEX



+OPEX

EXAM

> On Premises vs. Cloud | SaaS

SaaS: Software as a Service

Software as a Service, also known as cloud application services, represents the most utilized option for businesses in the cloud market. SaaS utilizes the internet to deliver applications, which are managed by a third-party vendor, to its users. Most SaaS applications run directly through your web browser, which means they do not require any downloads or installations on the client side.

Examples of SaaS

Popular examples of SaaS include:

- Google Workspace (formerly GSuite)
- Dropbox
- Salesforce
- Cisco WebEx

EXAM

> On Premises vs. Cloud | PaaS

PaaS: Platform as a Service

Cloud platform services, also known as Platform as a Service (PaaS), provide cloud components to certain software while being used mainly for applications. PaaS delivers a framework for developers that they can build upon and use to create customized applications. All servers, storage, and networking can be managed by the enterprise or a third-party provider while the developers can maintain management of the applications.

Examples of PaaS

Popular examples of PaaS include:

- AWS Elastic Beanstalk
- Windows Azure
- Azure SQL Server (depends)
- Heroku
- OpenShift

EXAM

> On Premises vs. Cloud | IaaS

IaaS: Infrastructure as a Service

Cloud infrastructure services, known as Infrastructure as a Service (IaaS), are made of highly scalable and automated compute resources. IaaS is fully self-service for accessing and monitoring computers, networking, storage, and other services. IaaS allows businesses to purchase resources on-demand and as-needed instead of having to buy hardware outright.

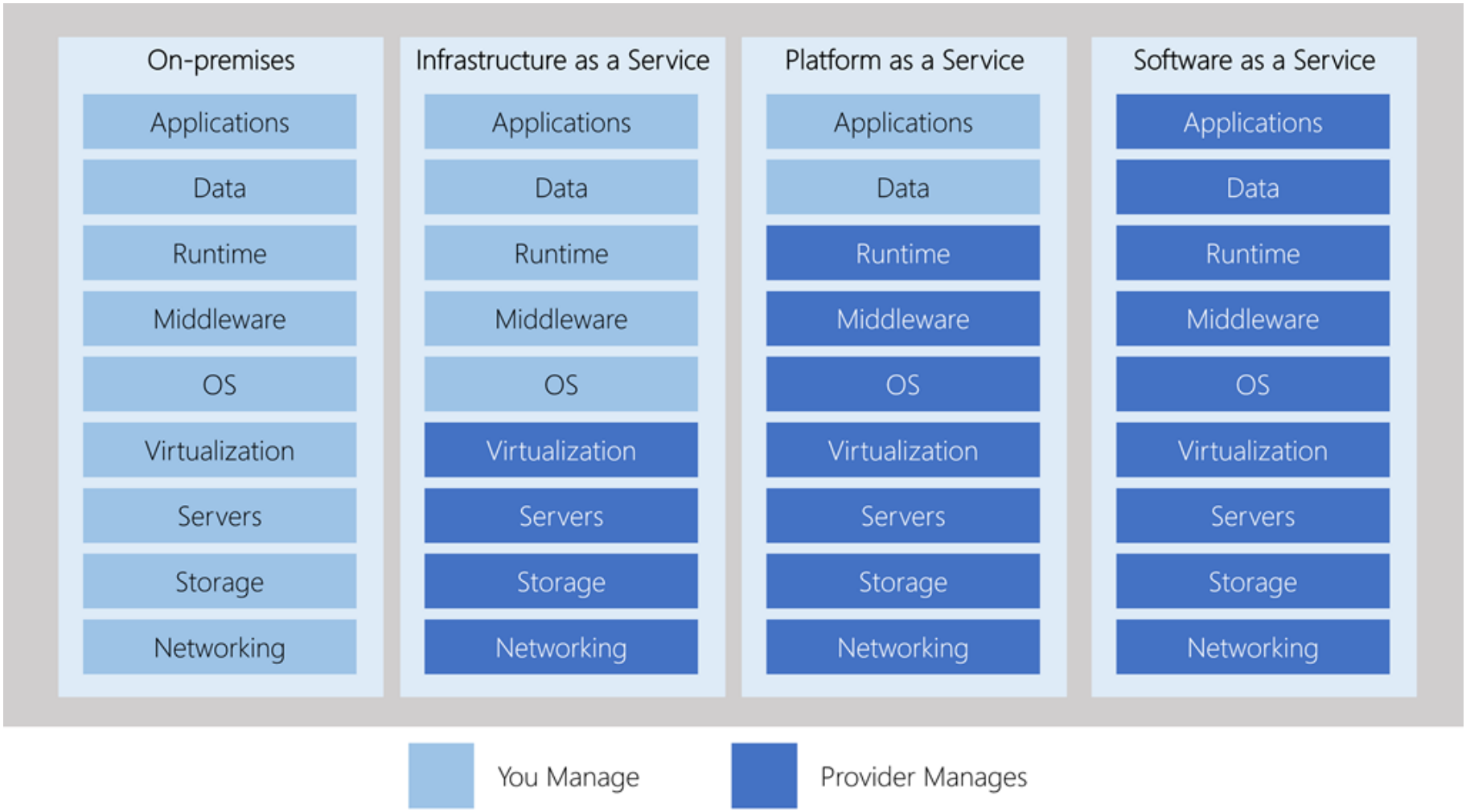
Examples of IaaS

Popular examples of IaaS include:

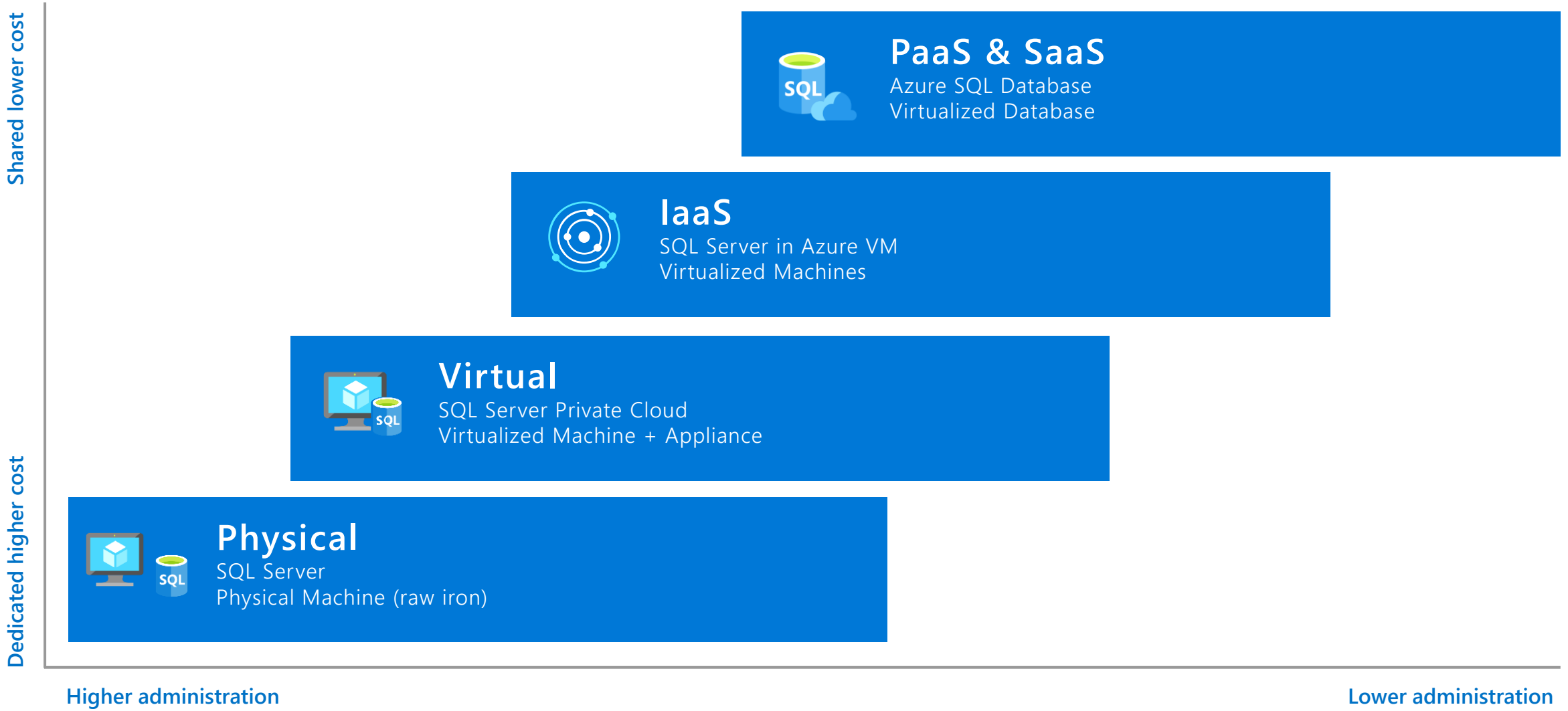
- DigitalOcean
- Linode
- Amazon Web Services (AWS)
- Cisco Metacloud
- Microsoft Azure

EXAM

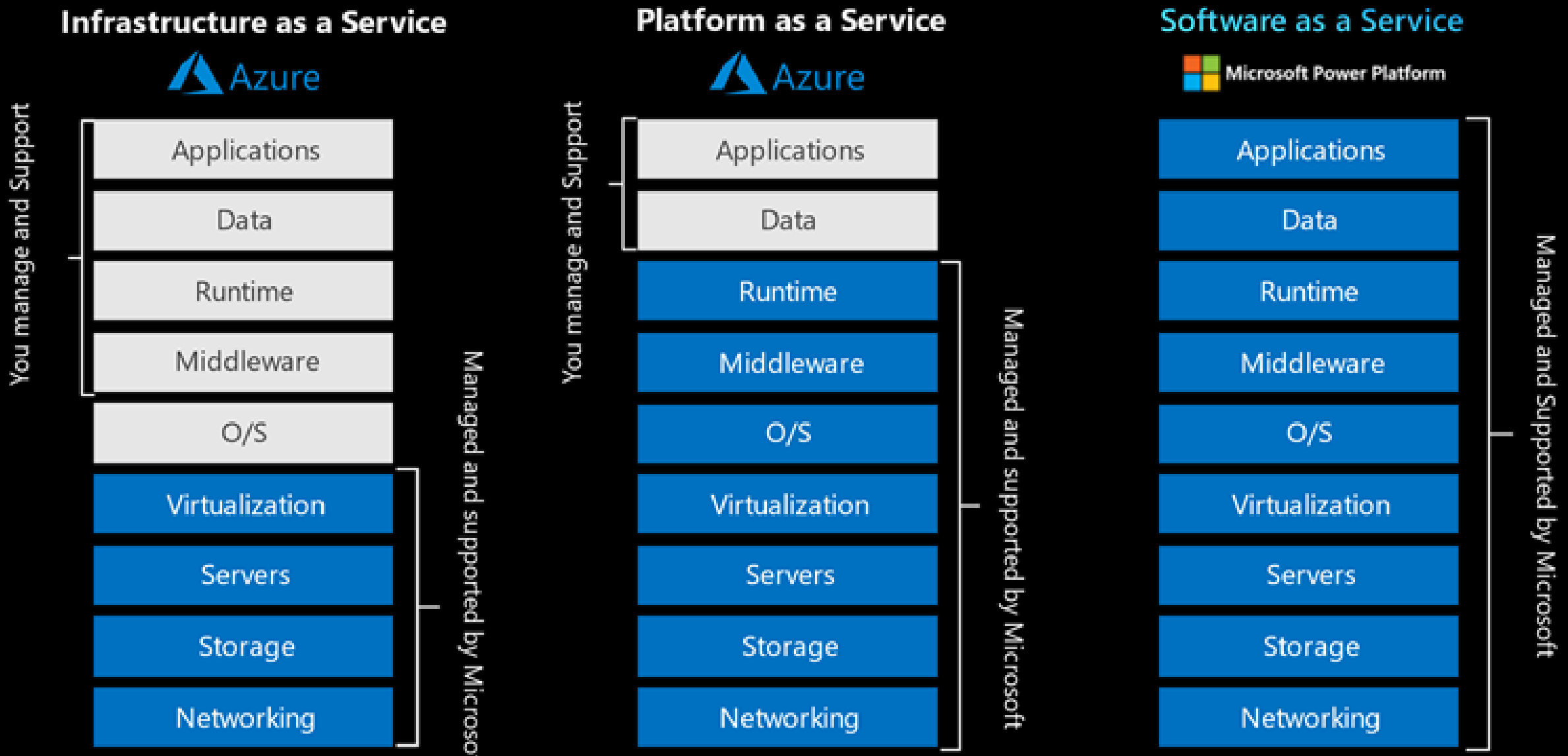
> On Premises vs. Cloud



Data platform continuum



Which tool for the job?



EXAM

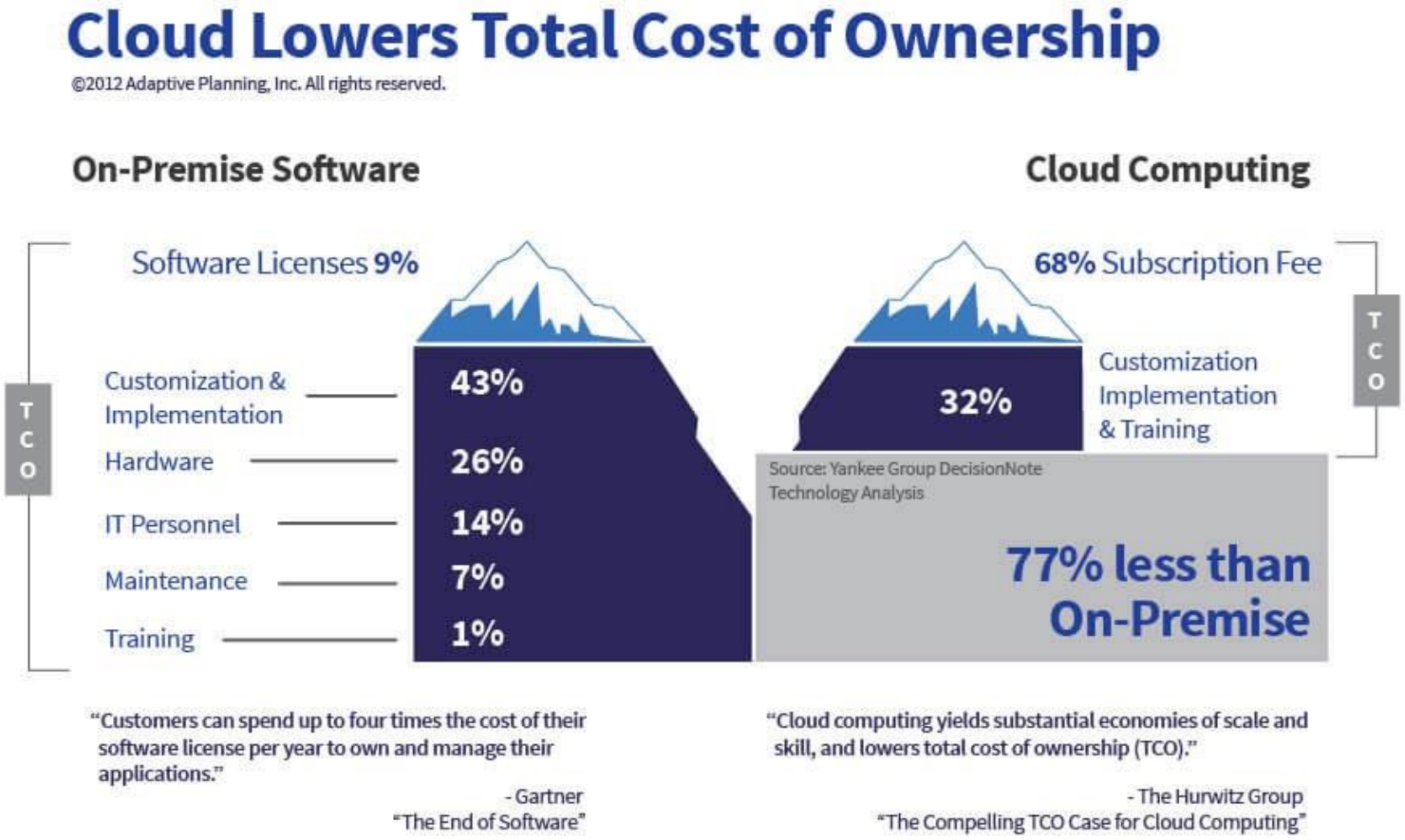
> Example | SQL instance

IaaS vs. PaaS

SQL Server on a virtual machine is considered IaaS. The other deployment options in the Azure SQL platform, Azure SQL Managed Instance and Azure SQL Database, are platform as a service (PaaS) deployments. These PaaS Azure SQL deployment options contain a fully managed database engine that automates most of the database management functions, like upgrading, patching, backups, and monitoring. Here are some key features of SQL Managed Instance and SQL Database:

- **Business continuity** enables your business to continue operating in the face of disruption.
- **High availability** guarantees your databases are up and running 99.99% of the time. No need to worry about maintenance or downtimes.
- **Automated backups** are created and use Azure read-access geo-redundant storage (RA-GRS) to provide geo-redundancy.
- **Long-term backup retention** enables you to store specific full databases for up to 10 years.
- **Geo-replication** creates readable replicas of your database in the same datacenter (region) or a different one.
- **Scalability**. You can easily add more resources (CPU, memory, storage) without long provisioning.
- **Network security** features protect your data over the network. These features include firewalls to restrict connectivity, Azure Private Link to ensure your data isn't exposed to the internet, and integration with virtual networks for connectivity to on-premises environments.
- **Advanced security** detects threats and vulnerabilities in your databases and enables you to secure your data.
- **Automatic tuning** analyzes your workload. It provides recommendations that can optimize performance of your applications by adding indexes, removing unused indexes, and automatically fixing query plan problems.
- **Built-in monitoring** capabilities enable you to get insights into the performance of your databases and workload and troubleshoot performance problems.
- **Built-in intelligence** automatically identifies potential problems in your workload and provides recommendations that can help you to fix those problems.

> Considerations



On Prem

AI Workstations and Servers

> AI Workstations

Lambda

Servers GPU Workstation TensorBook GPU Cloud Resources +1 (866) 711-2025

Lambda Workstation

From $5,258.00
Free 30-day returns

Customize now

New! Cloud GPU servers from $1.25 per hour >

Deep learning workstation with up to 4 GPUs

NVIDIA RTX 3090, RTX 3080, RTX 3070, RTX A6000, RTX 5000, RTX 6000, and RTX 8000 options. Pre-installed with Ubuntu, TensorFlow, PyTorch, CUDA, and cuDNN.

Customize now

1000+ research groups trust Lambda

Technical specifications

GPU	Up to 4x NVIDIA GPUs RTX 30XX (Ampere), Quadro RTX A6000, RTX 8000, and RTX 6000 options
Processor	AMD Ryzen or Intel Core i9 Configurable up to 64 cores, 128 threads, and 256 MB cache
Memory	Up to 256 GB Fits up to eight 32 GB DIMMs at 3200 MHz
OS drive	Up to 2 TB 3,200 MB/s seq. read and 2,000 MB/s seq. write
Extra storage	Up to 61 TB Fits up to eight 7.68 TB SATA SSDs.
Power supply	Up to 1600 watts of maximum continuous power at voltages between 100 and 240V
Size & weight	**Width:** 13.1" (332 mm) **Height:** 16.3" (415 mm) **Depth:** 18.4" (458 mm) **Weight:** 38 pounds (17.2 kg)

> AI Servers

Lambda

Servers GPU Workstation TensorBook GPU Cloud Resources +1 (866) 711-2025

Lambda Blade

From $26,512.00
Free 30-day returns

Customize now

New! Cloud GPU servers from $1.25 per hour

GPU server built for deep learning

Up to ten customizable GPUs with AMD EPYC or Intel Xeon processor. Pre-installed with Ubuntu, TensorFlow, PyTorch, CUDA, and cuDNN.

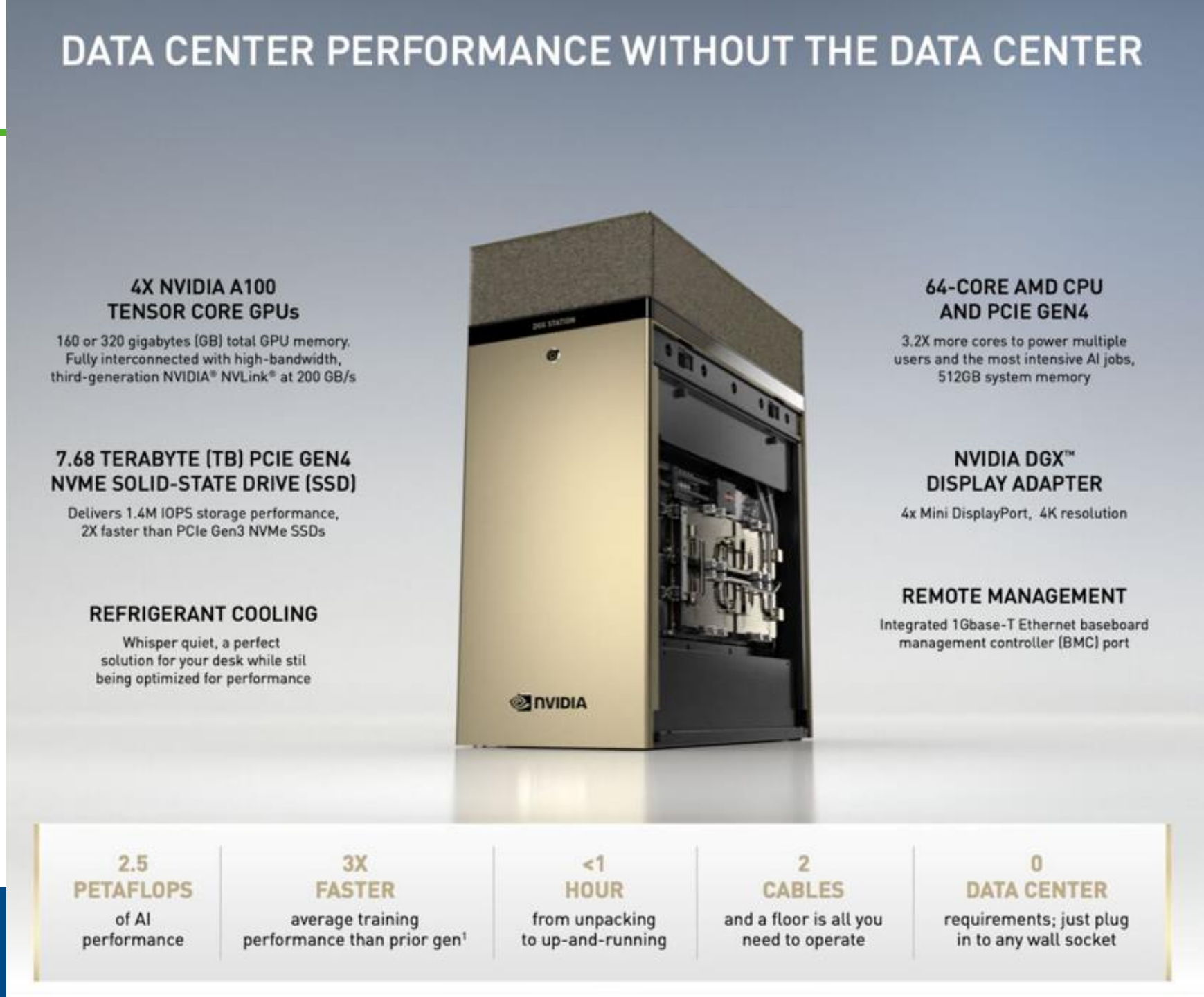
Customize now

1000+ research groups trust Lambda

Apple MIT WELLS FARGO Caltech Los Alamos NATIONAL LABORATORY amazon Anthem

Easy system administration

> AI Servers



Georgian

> AI Servers and Workstations

DELL Technologies AI Precision › AI Solutions & Technologies Search Contact Sign In EN/CA

Dell Precision 5820 Tower

The new Dell Precision 5820 Tower is ideal for cognitive solution development and inference applications.

- Up to 6x Hard Disk Drive (HDD)/Solid State Drive (SSD) Storage
- Ubuntu 18.04 Factory Installed
- NVIDIA NGC-Ready for install of NVIDIA Data Science Software, powered by RAPIDS

up to 18	up to 256GB	up to 2x
cores with Xeon ®	DDR4 Memory	NVIDIA Quadro ® RTX™ 8000 48GB GPUs

Dell Precision 7920 Tower

The Dell Precision 7920 Tower handles learning model training and larger solution frameworks with ease.

- Up to 10x Hard Disk Drive (HDD)/Solid State Drive (SSD) Storage
- Ubuntu 18.04 Factory Installed
- NVIDIA NGC-Ready for install of NVIDIA Data Science Software, powered by RAPIDS

up to Dual 56	up to 3TB	up to 3x
cores with Xeon-SP	DDR4 Memory	NVIDIA Quadro ® RTX™ 8000 48GB GPUs

https://www.delltechnologies.com/en-ca/ai-technologies/index.htm

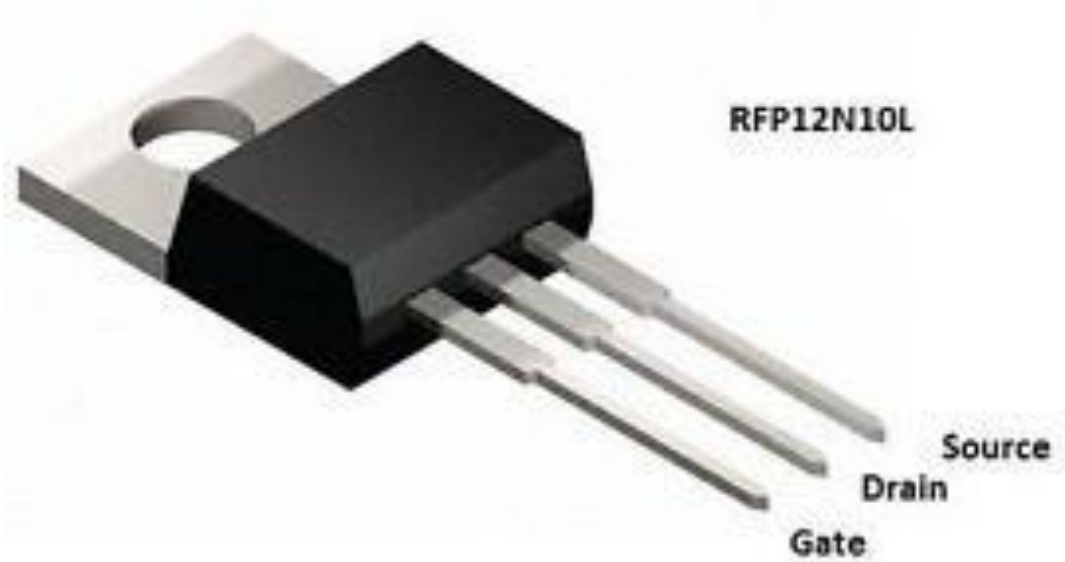
AI Accelerator

> AI Processing | AI accelerator

An **AI accelerator** is a class of specialized hardware accelerator or computer system designed to accelerate artificial intelligence applications, especially artificial neural networks, machine vision and machine learning.

Typical applications include algorithms for robotics, internet of things and other data-intensive or sensor-driven tasks. They are often manycore designs and generally focus on low-precision arithmetic, novel dataflow architectures or in-memory computing capability.

As of 2018, a typical AI integrated circuit chip contains billions of MOSFET transistors.



https://en.wikipedia.org/wiki/AI_accelerator

> AI Processing | GPU's

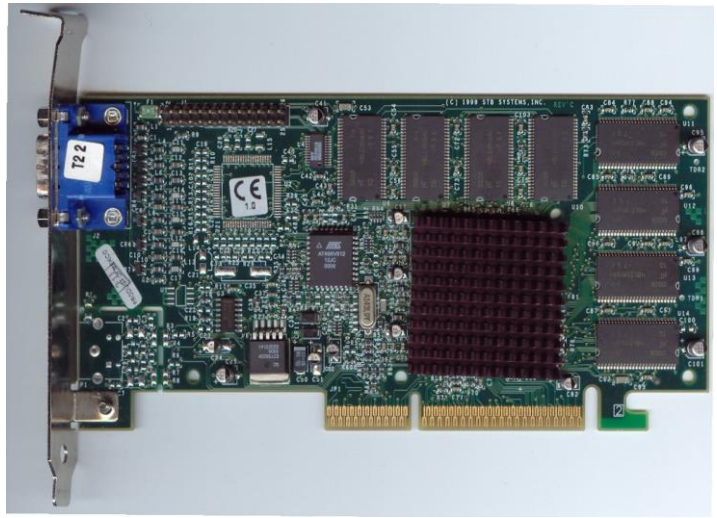
A **graphics processing unit (GPU)** is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device. GPUs are used in embedded systems, mobile phones, personal computers, workstations, and game consoles. Modern GPUs are very efficient at manipulating computer graphics and image processing. Their highly parallel structure makes them more efficient than general-purpose central processing units (CPUs) for algorithms that process large blocks of data in parallel. In a personal computer, a GPU can be present on a video card or embedded on the motherboard.

https://en.wikipedia.org/wiki/AI_accelerator

> AI Processing | GPU's | Examples

Atari ANTIC microprocessor on an Atari 130XE motherboard (1977 / 1978)

dedicated to generating 2D computer graphics on a television screen or computer display.

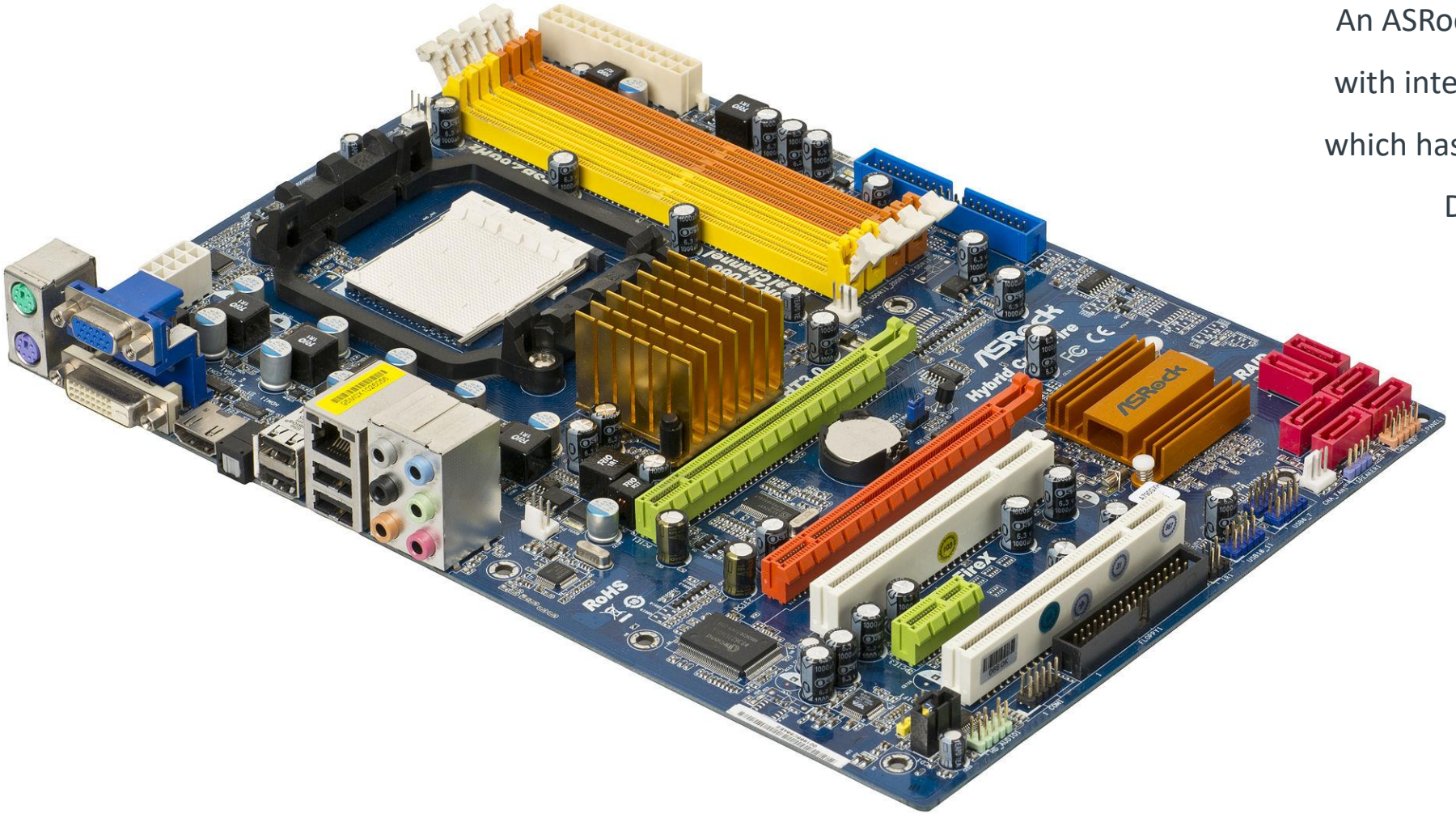


Voodoo3 2000 AGP card (1998 / 1999)

computer gaming video cards manufactured and designed by 3dfx Interactive.

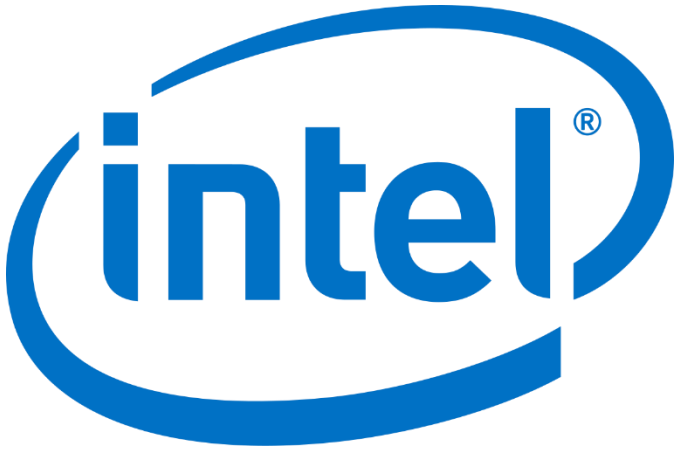
MSI GeForce RTX 2080 TI GAMING X TRIO Video Card

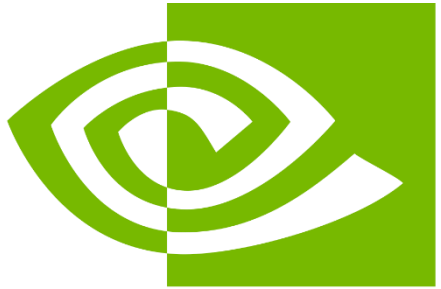
> AI Processing | GPU's | Examples



An ASRock motherboard with integrated graphics, which has HDMI, VGA and DVI outs.

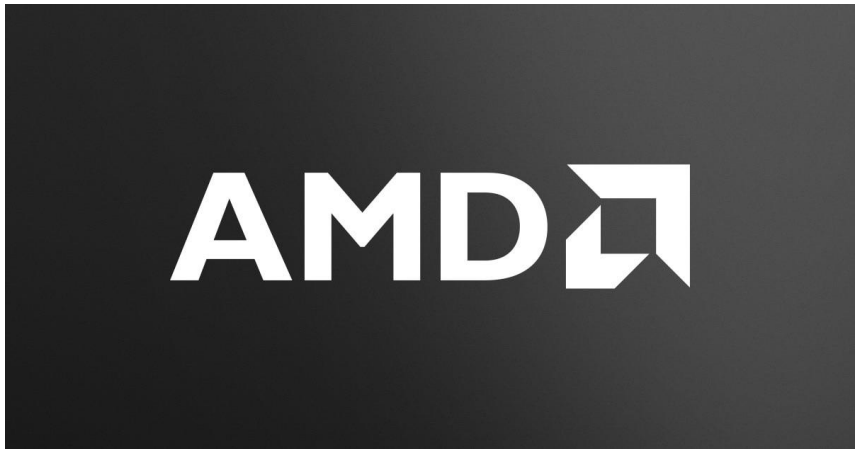
> AI Processing | GPU's | Companies

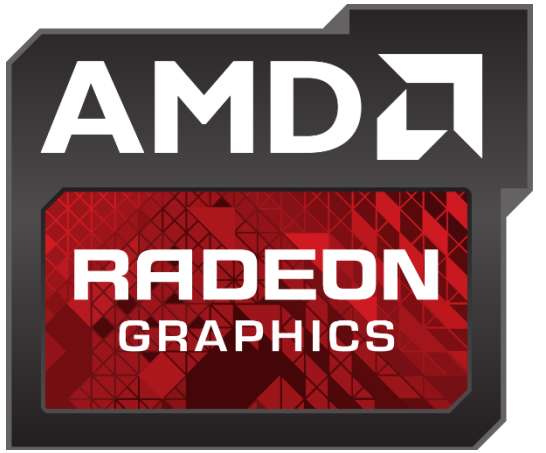




NVIDIA®

https://youtu.be/e2_hsjpTi4w





> AI Processing | GPU's | Specific Use

Most GPUs are designed for a specific usage, real-time 3D graphics or other mass calculations:

1. Gaming

1. GeForce GTX, RTX
2. Nvidia Titan
3. Radeon HD, R5, R7, R9, RX, Vega and Navi series

2. Cloud Gaming

1. Nvidia Grid
2. AMD Radeon Sky

3. Workstation (Video editing, encoding, decoding, transcoding and rendering (digital content creation), 3D animation and rendering, VFX and motion graphics (CGI), videogame development and 3D texture creation, product development/3D CAD, structural analysis, simulations, CFD analysis and scientific calculations...)

1. Nvidia Quadro
2. AMD FirePro
3. AMD Radeon Pro
4. AMD Radeon VII

4. Cloud Workstation

1. Nvidia Tesla
2. AMD FireStream

5. Artificial Intelligence training and Cloud

1. Nvidia Tesla
2. AMD Radeon Instinct

6. Automated/Driverless car

1. Nvidia Drive PX

NVIDIA RTX A6000

First Look for Data Science 48GB Ampere GPU

https://m.youtube.com/watch?v=85-K7qTSvS8

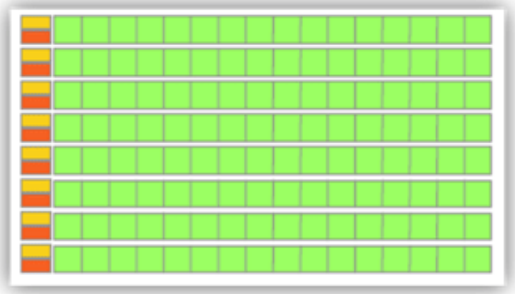
EXAM

> AI Processing | GPU's | Specific Use

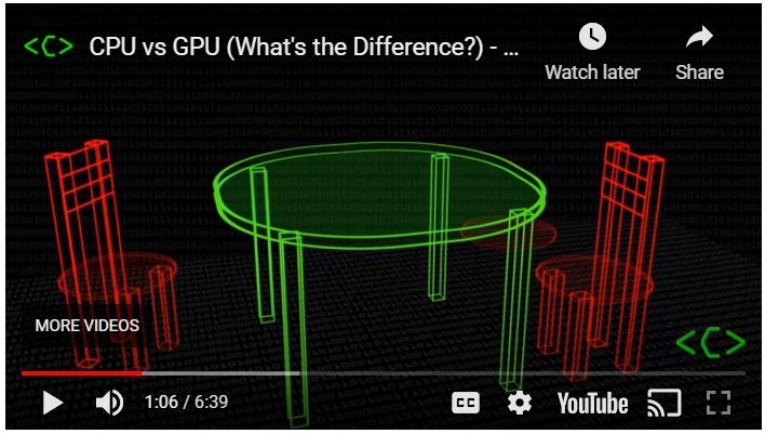
CPU

- Low compute density
- Complex control logic
- Large caches (L1$/L2$, etc.)
- Optimized for serial operations
  - Fewer execution units (ALUs)
  - Higher clock speeds
- Shallow pipelines (<30 stages)
- Low Latency Tolerance
- Newer CPUs have more parallelism

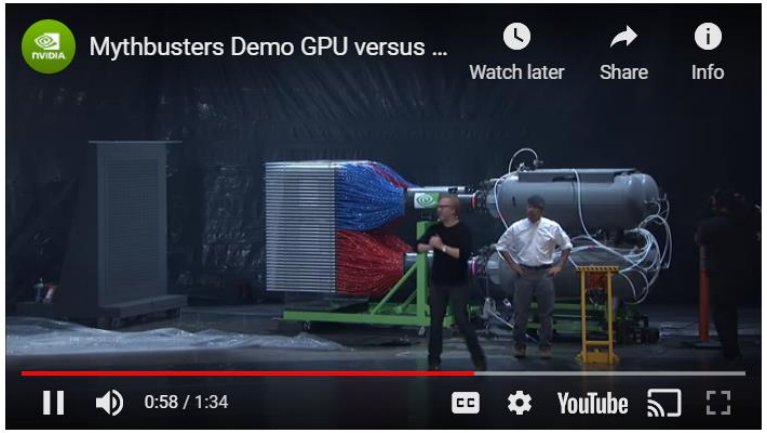
GPU



- High compute density
- High Computations per Memory Access
- Built for parallel operations
  - Many parallel execution units (ALUs)
  - Graphics is the best known case of parallelism
- Deep pipelines (hundreds of stages)
- High Throughput
- High Latency Tolerance
- Newer GPUs:
  - Better flow control logic (becoming more CPU-like)
  - Scatter/Gather Memory Access
  - Don't have one-way pipelines anymore



https://youtu.be/_cyVDoyI6NE



https://youtu.be/-P28LKWTzrI

https://en.wikipedia.org/wiki/AI_accelerator

> AI Processing | FPGA's

Deep learning frameworks are still evolving, making it hard to design custom hardware. Reconfigurable devices such as **field-programmable gate arrays (FPGA)** make it easier to evolve hardware, frameworks and software alongside each other.

Microsoft has used FPGA chips to accelerate inference. The application of FPGAs to AI acceleration motivated Intel to acquire Altera with the aim of integrating FPGAs in server CPUs, which would be capable of accelerating AI as well as general purpose tasks.

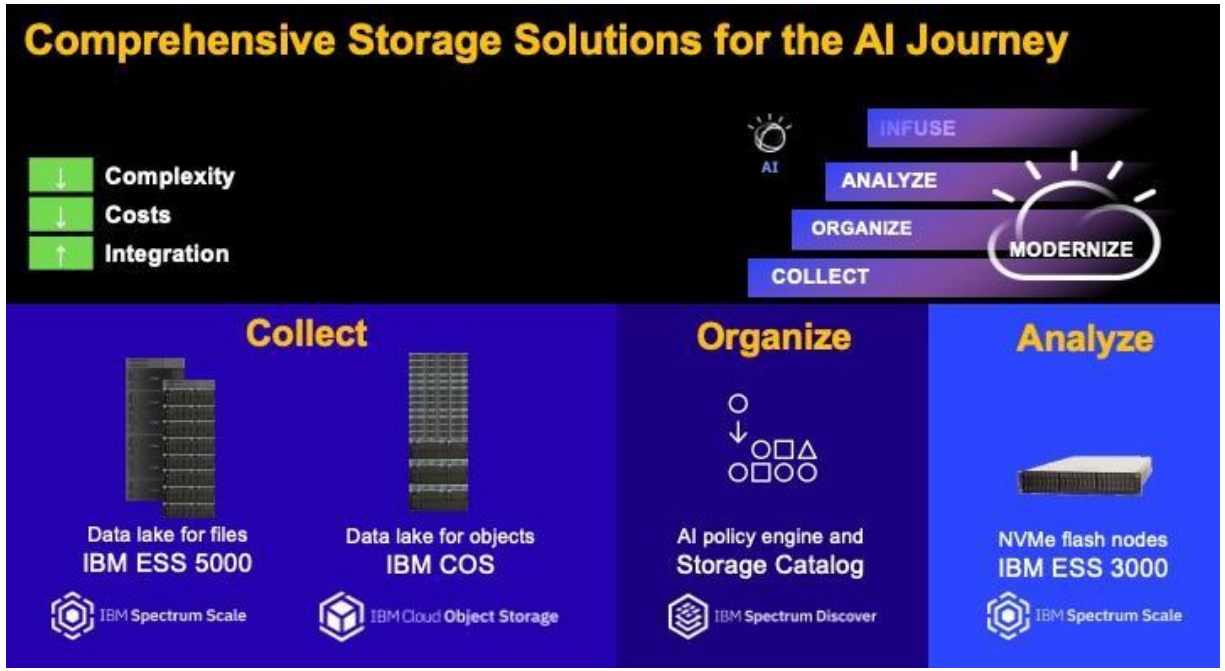
https://en.wikipedia.org/wiki/AI_accelerator

> AI Processing | Storage Solutions

Georgian

https://www.groovypost.com/reviews/what-is-nvme-m2-ssd-drive-how-fast-is-it/

> AI Processing | Storage Solutions



NVMe (nonvolatile memory express) is a new storage access and transport protocol for flash and next-generation solid-state drives (SSDs) that delivers the highest throughput and fastest response times yet for all types of enterprise workloads.

To help deliver a high-bandwidth, low-latency user experience, the NVMe protocol accesses flash storage via a PCI Express (PCIe) bus, which supports tens of thousands of parallel command queues and thus is much faster than hard disks and traditional all-flash architectures, which are limited to a single command queue.

The NVMe specification takes advantage of nonvolatile memory in all kinds of computing environments. And it's future-proof, extendable to work with not-yet-invented persistent memory technologies.

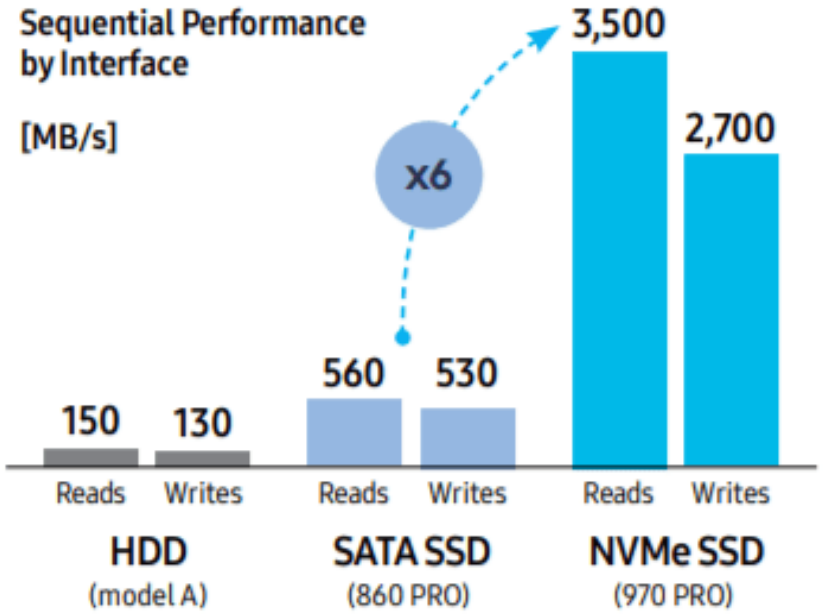
https://www.netapp.com/data-storage/nvme/what-is-nvme/

> AI Processing | Storage Solutions

The NVMe Difference

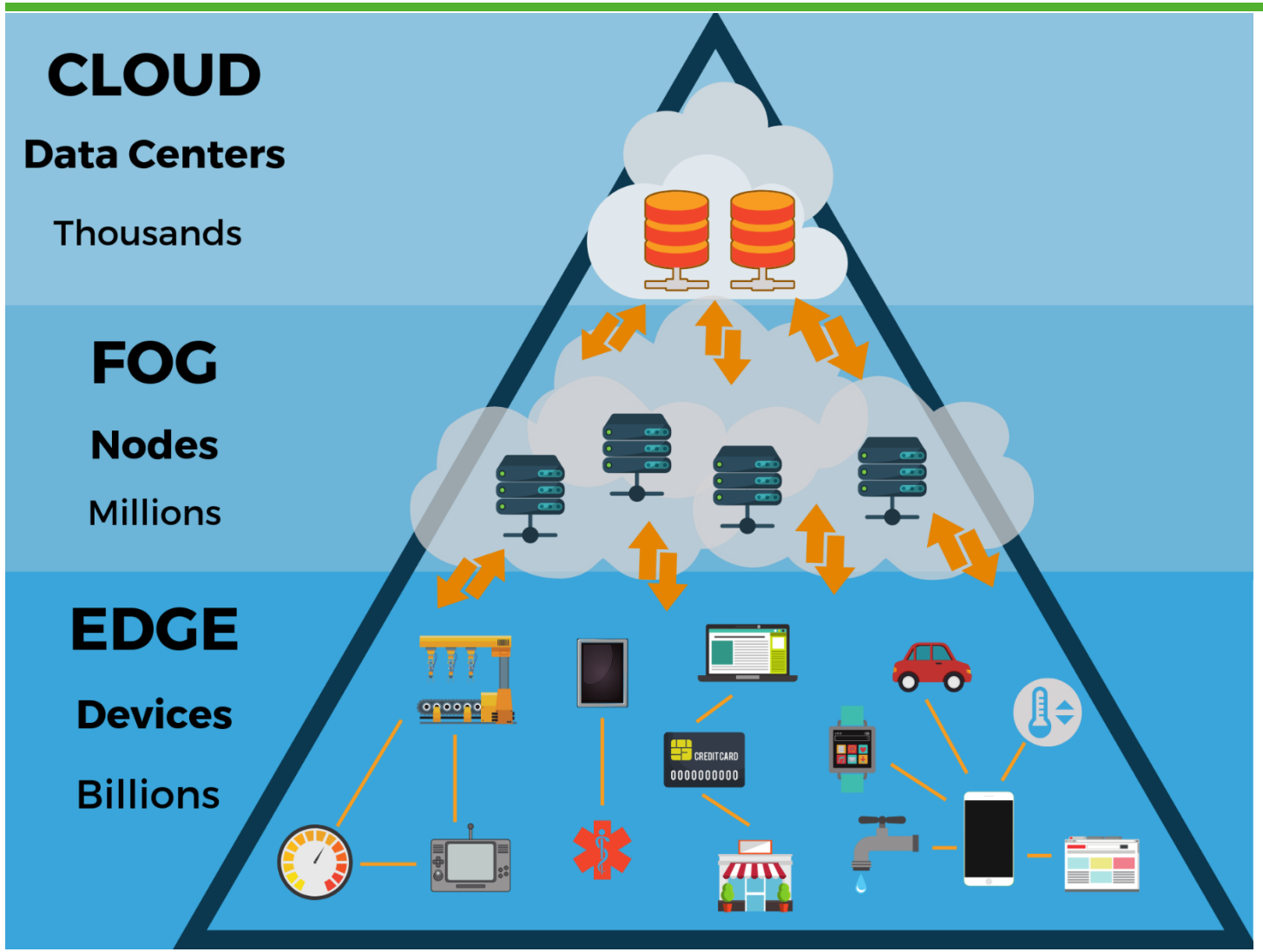
The Non-Volatile Memory express (NVMe) interface is specifically engineered to overcome constraints of older interfaces by increasing NAND flash storage transfer speeds. Sequential speeds are approximately 6 times faster than SATA SSDs as NVMe enables your drive to reach its full potential.

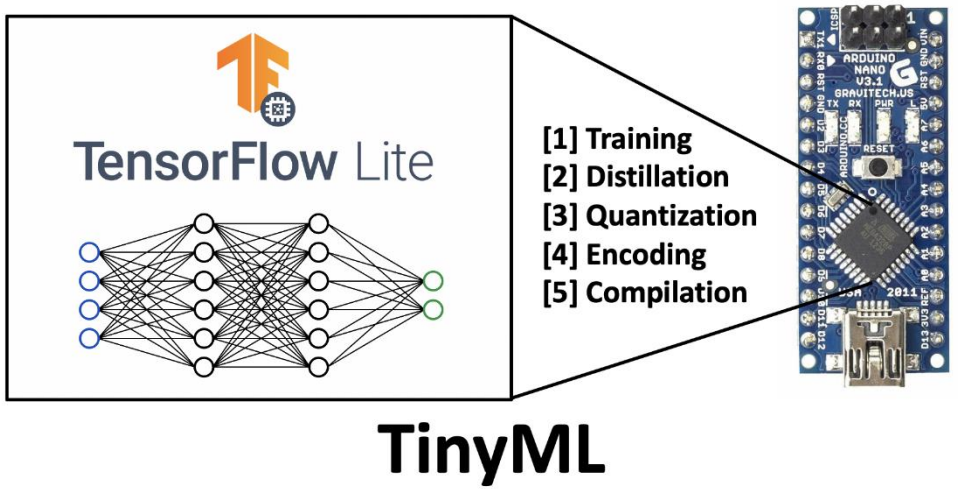
* Performance may vary based on SSD's firmware version, system hardware and configuration.



https://www.netapp.com/data-storage/nvme/what-is-nvme/

> AI Processing | Embedded Devices





TinyML

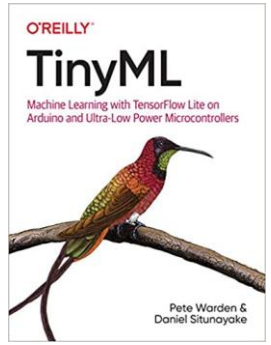
https://www.groovypost.com/reviews/what-is-nvme-m2-ssd-drive-how-fast-is-it/

EXAM

> AI Processing | Embedded Devices

Tiny machine learning (tinyML) is the intersection of machine learning and embedded internet of things (IoT) devices. The field is an emerging engineering discipline that has the potential to revolutionize many industries.

The main industry beneficiaries of tinyML are in edge computing and energy-efficient computing. TinyML emerged from the concept of the internet of things (IoT). The traditional idea of IoT was that data would be sent from a local device to the cloud for processing. Some individuals raised certain concerns with this concept: privacy, latency, storage, and energy efficiency to name a few.



https://tinymlbook.com/

https://towardsdatascience.com/tiny-machine-learning-the-next-ai-revolution-495c26463868/

> AI Processing | Embedded Devices | Benefits

Energy Efficiency - Transmitting data (via wires or wirelessly) is very energy-intensive, around an order of magnitude more energy-intensive than onboard computations (specifically, multiply-accumulate units). Developing IoT systems that can perform their own data processing is the most energy-efficient method. AI pioneers have discussed this idea of “data-centric” computing (as opposed to the cloud model’s “compute-centric”) for some time and we are now beginning to see it play out.

Privacy - Transmitting data opens the potential for privacy violations. Such data could be intercepted by a malicious actor and becomes inherently less secure when warehoused in a singular location (such as the cloud). By keeping data primarily on the device and minimizing communications, this improves security and privacy.

Storage - For many IoT devices, the data they are obtaining is of no merit. Imagine a security camera recording the entrance to a building for 24 hours a day. For a large portion of the day, the camera footage is of no utility, because nothing is happening. By having a more intelligent system that only activates when necessary, lower storage capacity is necessary, and the amount of data necessary to transmit to the cloud is reduced.

Latency - For standard IoT devices, such as Amazon Alexa, these devices transmit data to the cloud for processing and then return a response based on the algorithm’s output. In this sense, the device is just a convenient gateway to a cloud model, like a carrier pigeon between yourself and Amazon’s servers. The device is pretty dumb and fully dependent on the speed of the internet to produce a result. If you have slow internet, Amazon Alexa will also become slow. For an intelligent IoT device with onboard automatic speech recognition, the latency is reduced because there is reduced (if not no) dependence on external communications.

https://towardsdatascience.com/tiny-machine-learning-the-next-ai-revolution-495c26463868/

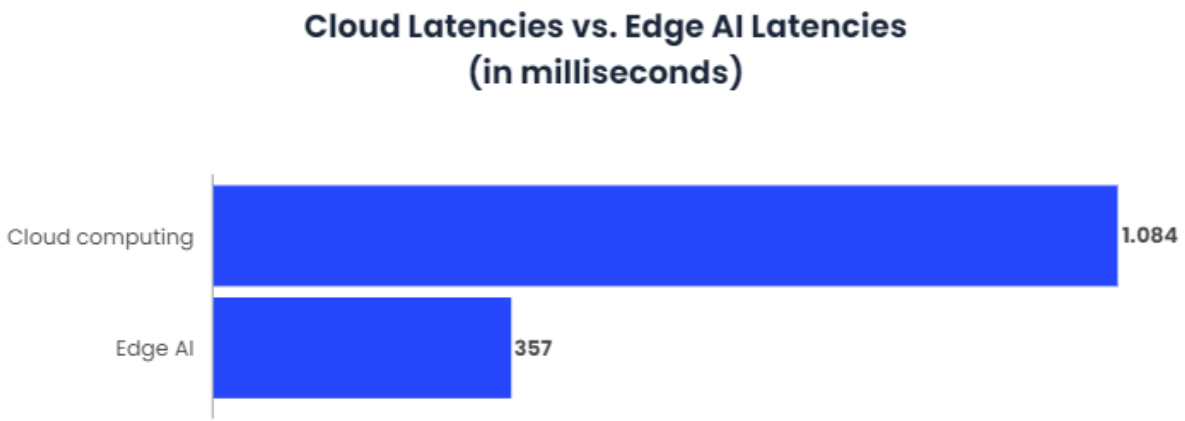
> Edge AI Devices

Edge AI is a system that uses Machine Learning algorithms to process data generated by a **hardware device at the local level**. The device does not need to be connected to the Internet to process such data and make decisions in real time, in a matter of milliseconds. This considerably reduces the communication costs derived from the cloud model. In other words, Edge AI takes the data and its processing to the closest point of interaction with the user, whether it is a computer, an IoT device or an Edge server.

An example of this technology can be seen in the speakers of **Google, Alexa or the Apple Homepod**, which have learned words and phrases through Machine Learning and then stored them locally on the device. When the user communicates something to applications such as Siri or Google, they send the voice recording to an Edge network where it is passed to text via AI and a response is processed. Without an Edge network the response time would be seconds, with Edge the times are reduced to less than 400 milliseconds.

EXAM

> Edge AI Devices | Latency



Edge AI eliminates the privacy issue of transmitting millions of data and storing it in the cloud, as well as the bandwidth and latency limitations that reduce data transmission capacity.

Edge technology is essential for many industries, for example, for autonomous cars will help reduce power consumption by increasing battery durability. It will also be applicable to robots, surveillance systems and others. As a result, the market for Edge AI software is expected to grow in value from $355 million in 2018 to $1.12 trillion by 2023.

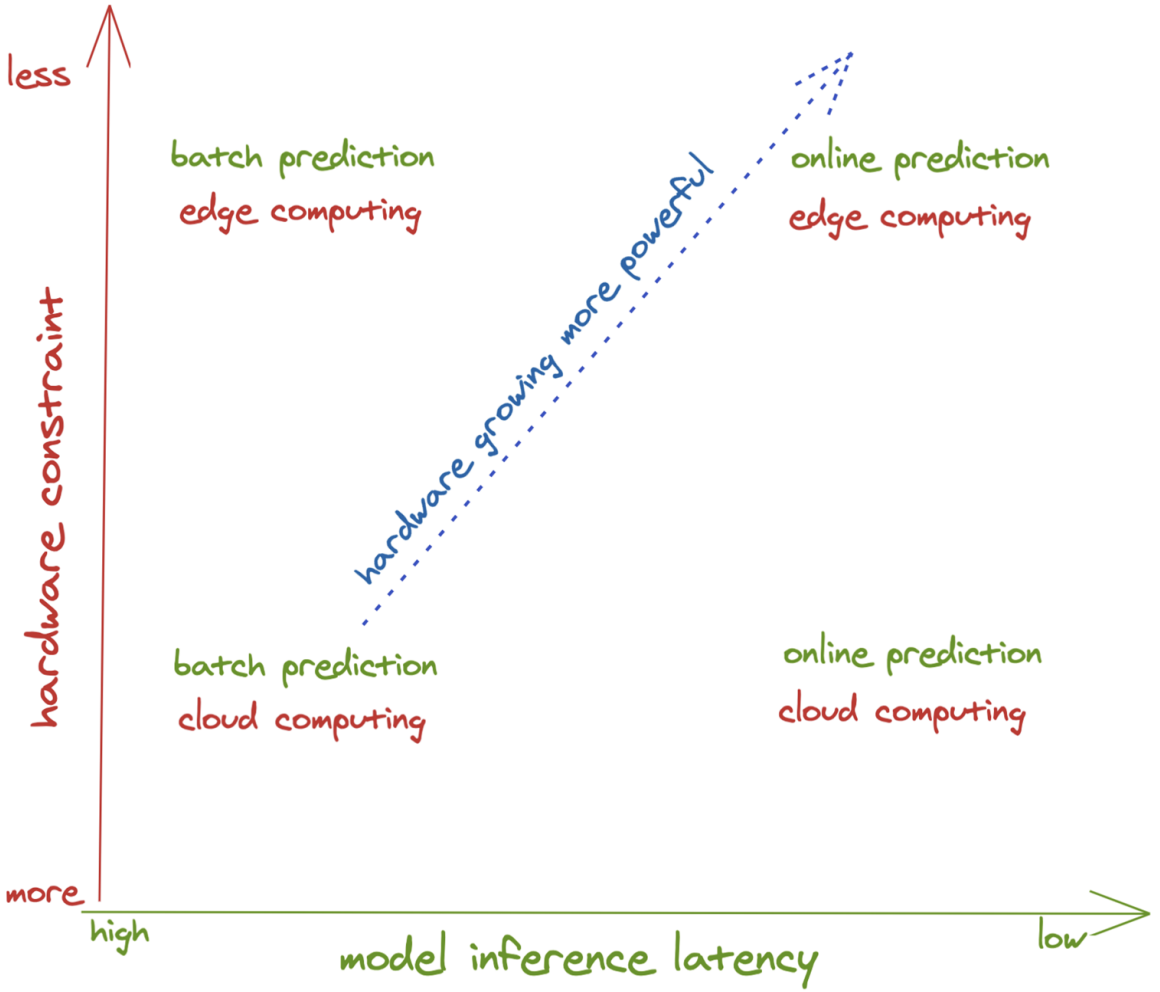
> Edge AI Devices | Benefits

- **Reduces costs and latency times** for an improved user experience This facilitates the integration of wearable technologies focused on the user experience, where you interact in real time to make payments, or where bracelets monitor your exercise and sleep patterns.
- **It increases the level of security** in terms of data privacy through local processing. Data is no longer shared in a centralized cloud.
- Technically, the reduction in required bandwidth should lead to a **reduction in the costs of the contracted internet service**.
- Edge technology devices **do not require specialized maintenance** by data scientists or AI developers. The graphic data flows are automatically delivered for monitoring; therefore, it is an autonomous technology.

> Cloud vs. Edge Computing

	Cloud computing	Edge computing
Computations	Done on cloud (servers)	Done on edge devices (browsers, phones, tablets, laptops, smart watches, activity watchers, cars, etc.)
Requirements	**Network connections**: availability and speed for data transfer	**Hardware**: memory, compute power, energy for doing computations
Examples	• Most queries to Alexa, Siri, Google Assistant • Google Translate for rare language pairs (e.g. English - Yiddish)	• Wake words for Alexa, Siri, Google Assistant • Google Translate for popular language pairs (e.g. English - Spanish) • Predictive text • Unlocking with fingerprints, faces

> Future of ML: online and on-device



> Questions…

What would be the **capacity requirements** for a given AI system?

Performance/cost benefit of hardware. Ie. we can spend more, and have it crunch models faster but what are the costs and benefits.

What are the **costs of buying server hardware vs using Cloud resources**?

> Infra Requirements…

1. High computing capacity

To fully take advantage of the opportunities presented by AI, organizations need sufficient performance computing resources, including CPUs and GPUs. A CPU-based environment can handle basic AI workloads, but deep learning involves multiple large data sets and deploying scalable neural network algorithms. For that, CPU-based computing might not be sufficient. For example, GPUs can accelerate deep learning by 100 times compared to traditional CPUs. Computing capacity and density will also grow, as will demand for high-performance networks and storage.

2. Storage capacity

It's fundamental that your infrastructure has the ability to scale storage as the volume of data grows. Figuring out what kind of storage an organization needs depends on many factors, including the level of AI an organization plans to use and whether they need to make real-time decisions. For example, a FinTech company that uses AI systems for real-time trading decisions may need fast all-flash storage technology, while for other companies slower but very large storage will be the most suitable solution. Businesses need to factor in how much AI data applications will generate. AI applications make better decisions when they're exposed to more data. As databases grow over time, companies need to monitor capacity and plan for expansion.

https://blog.leaseweb.com/2019/07/04/infrastructure-requirements-ai/

> Infra Requirements…

3. Networking infrastructure

Networking is another key component of AI infrastructure. Deep learning algorithms are highly dependent on communications, and networks will need to keep stride with demand as AI efforts expand. That's why scalability must be a high priority, and that will require a high-bandwidth, low-latency network. The best choice for expansive service is a global infrastructure provider who can ensure the service wrap and technology stack are consistent in all regions.

4. Security

AI can involve handling sensitive data such as patient records, financial information, and personal data. Having this data breached will be a disaster for any organization. Also, the infusion of bad data could cause the AI system to make incorrect inferences, leading to flawed decisions. The AI infrastructure must be secured from end to end with state-of-the-art technology.

https://blog.leaseweb.com/2019/07/04/infrastructure-requirements-ai/

> Infra Requirements…

5. Cost-effective solutions

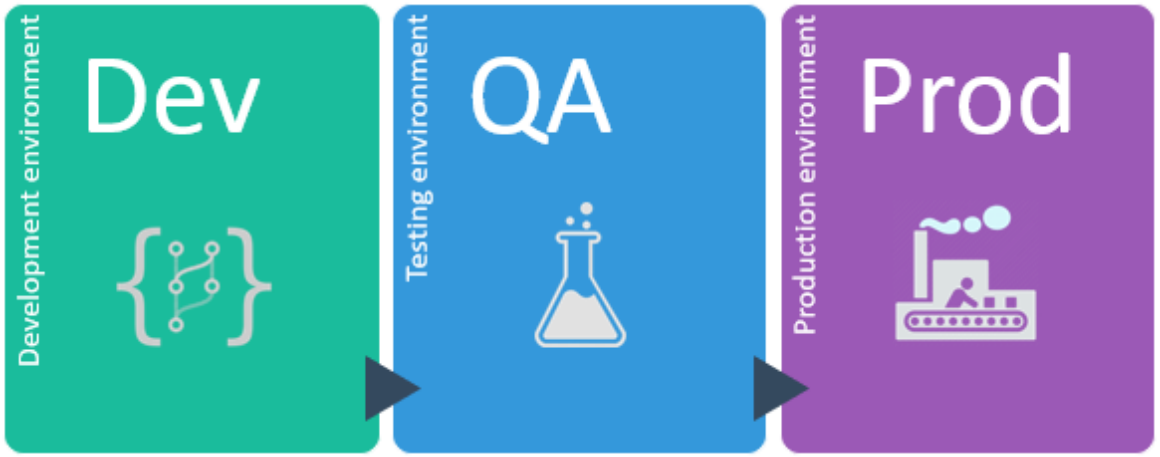
As AI models become more complex, they become more expensive to run, so getting extra performance from your infrastructure is pivotal to corralling costs. Over the next few years, we can expect continued growth in companies using AI, placing heavier burdens on the network, servers, and storage infrastructures to enable the use of this technology.

By making careful choices and identifying providers who can offer cost-effective dedicated servers, there is an opportunity to boost performance. This will enable companies to continue investing in AI without an increase in budget.

https://blog.leaseweb.com/2019/07/04/infrastructure-requirements-ai/

EXAM

> Infra Requirements | Environments



CONSULTING COMPANY

CLIENT CLOUD

Practice #1

Image Classifier

> #1 Image Classifier Assignment

INSTRUCTIONS:

Record a video with 10-15 minutes explaining how to use Image Classifier and your discoveries.

- In the Image Classifier select 3 or more different categories of images
- Tag each image
- Train your model
- Validate the results
- Evaluate the results
- Analyze all the steps you performed and think how should be the behavior (processes) of an enterprise solution with the same purpose. What are the differences?

In your video explain how and why you performed each step and explain the final result.

EVALUATION:

Mark: 5 points

- Ensure that you recorded yourself using the tool
- Ensure that you showed all the performed steps
- Ensure that you analyse the results
- Explain what kind of ML you are using in this exercise and why

Will be considered: Your results, explanations, level of details, clarity to explain and presentation / video quality (preparation).

Due date: Week 3 class

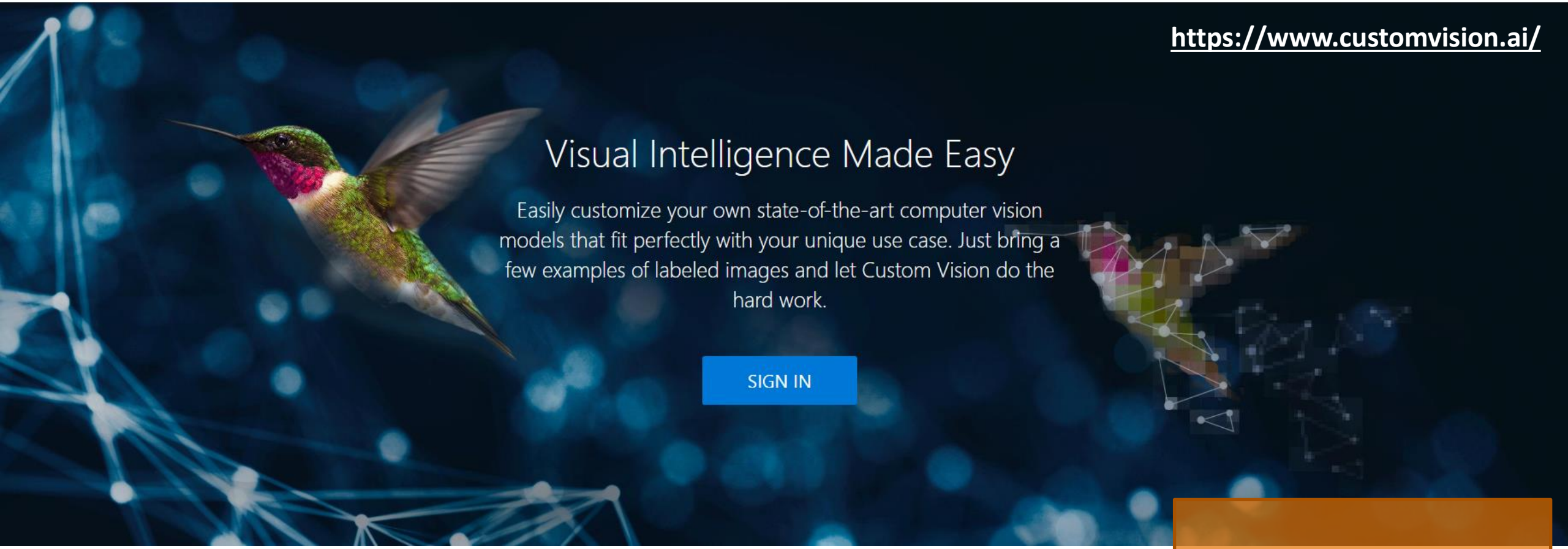
SCREENCAST MATIC

To record your video you can use:
https://screencast-o-matic.com/

> Image Classifier | Practice

Microsoft | Cognitive Services | Custom Vision

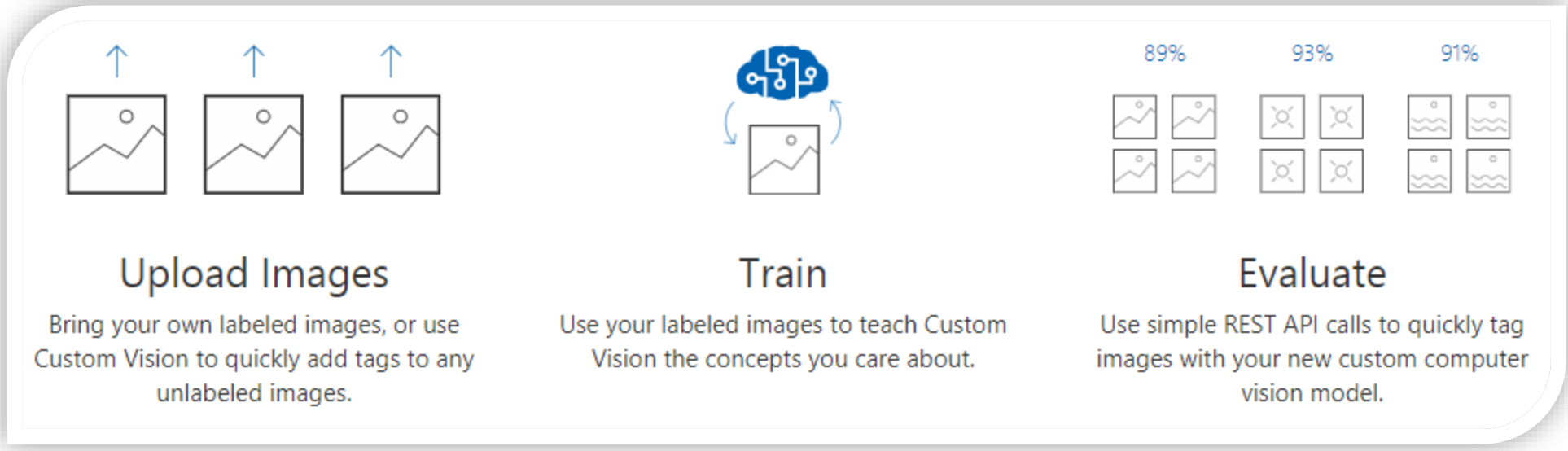
https://www.customvision.ai/



This is an example for academic purposes

> Image Classifier | Practice

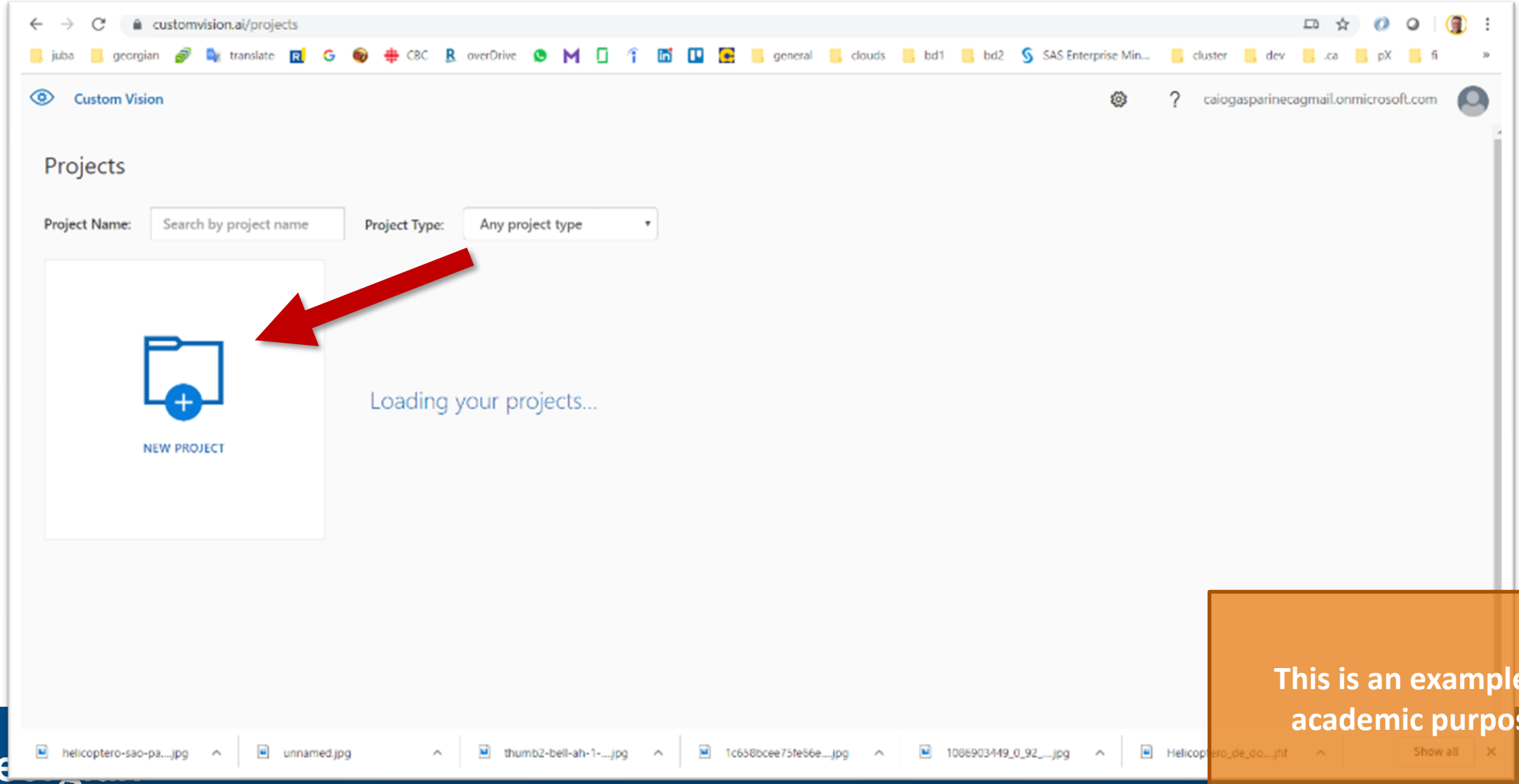
HOW IT WORKS?



This is an example for academic purposes

> Image Classifier | Practice

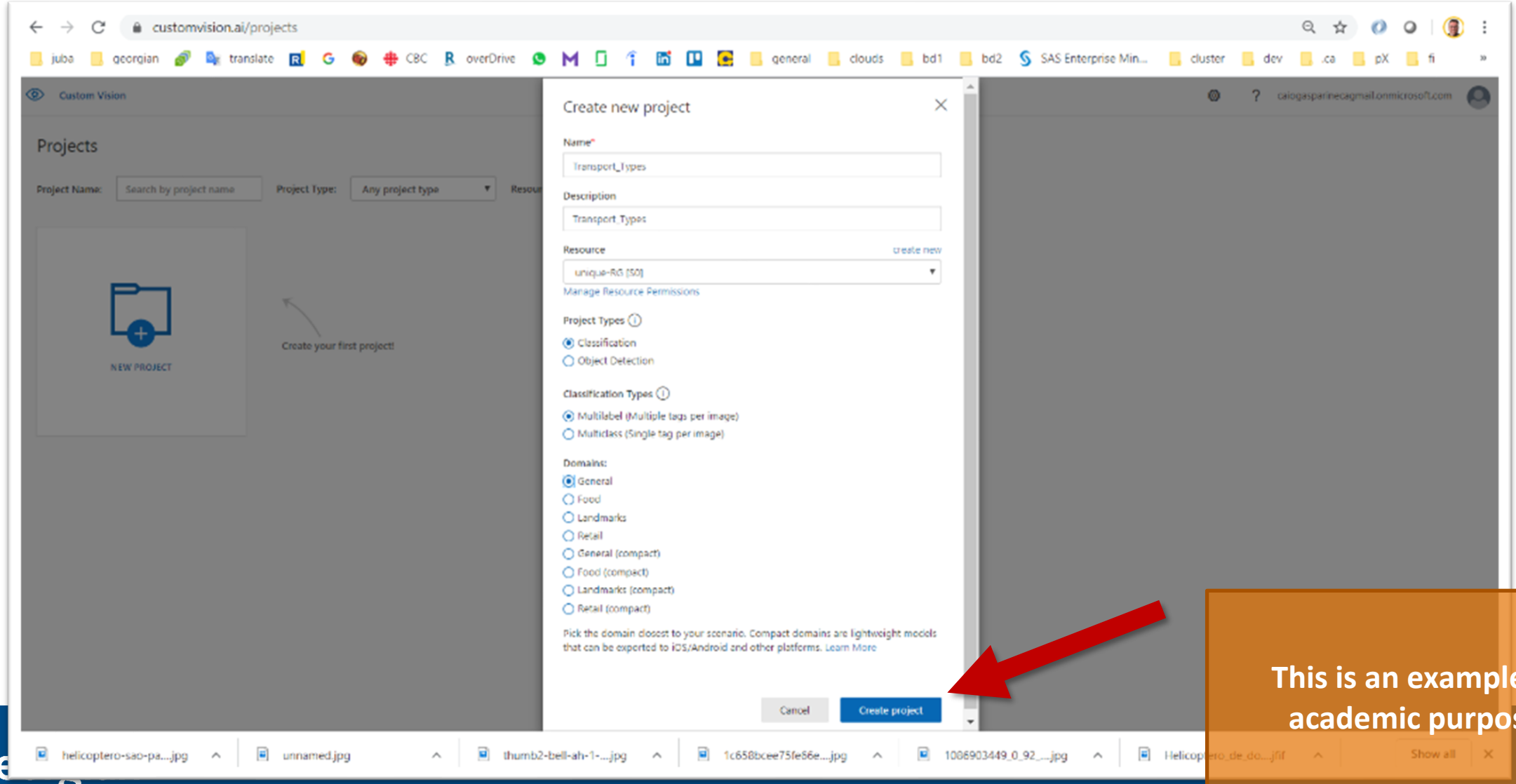
FIRST STEP - START A NEW PROJECT



This is an example for academic purposes

> Image Classifier | Practice

SET THE PARAMETERS FOR THE NEW PROJECT



This is an example for academic purposes

> Image Classifier | Practice

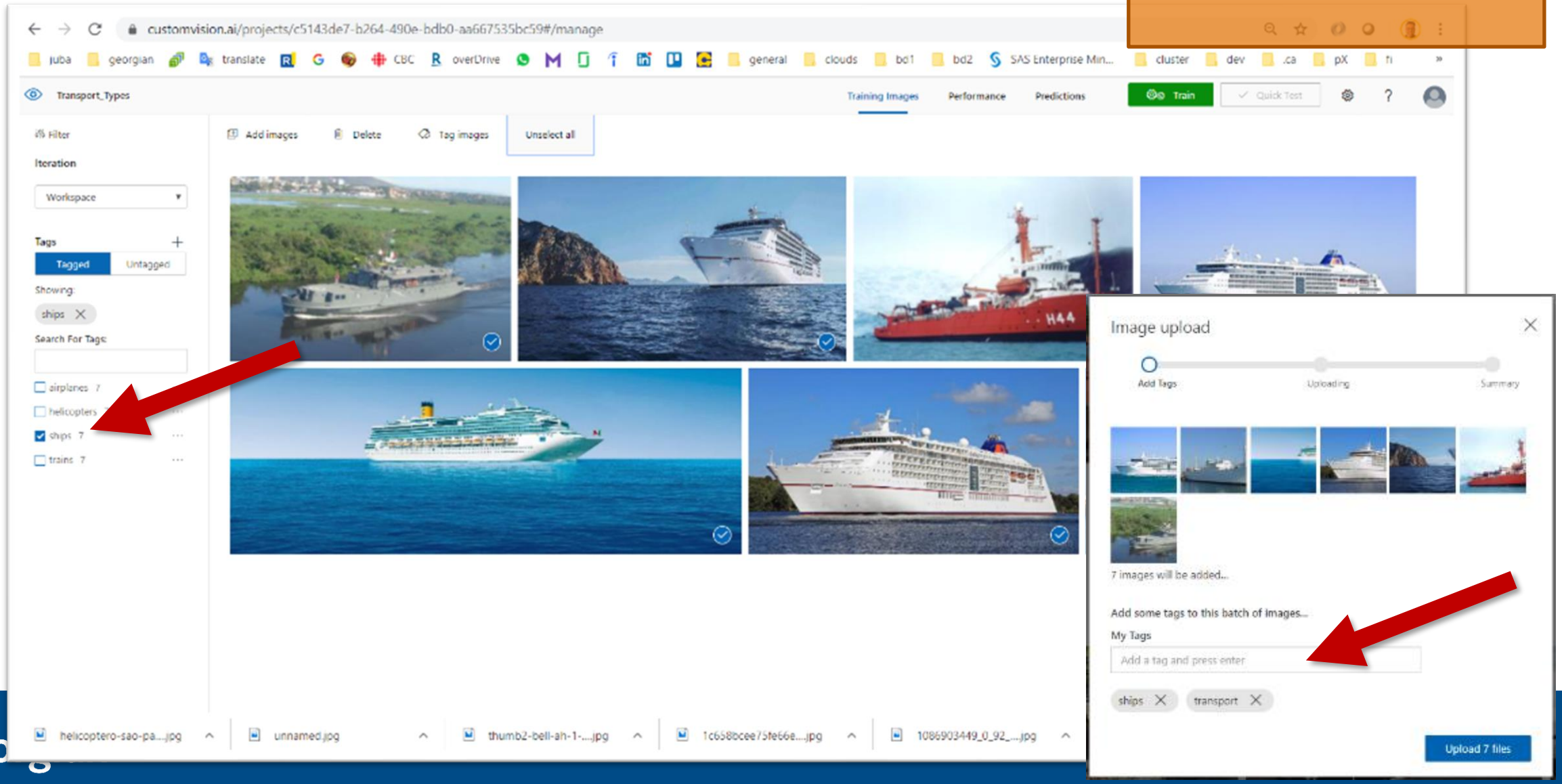
SELECT THE IMAGES TO TRAIN YOUR MODEL

This is an example for academic purposes

> Image Classifier | Practice

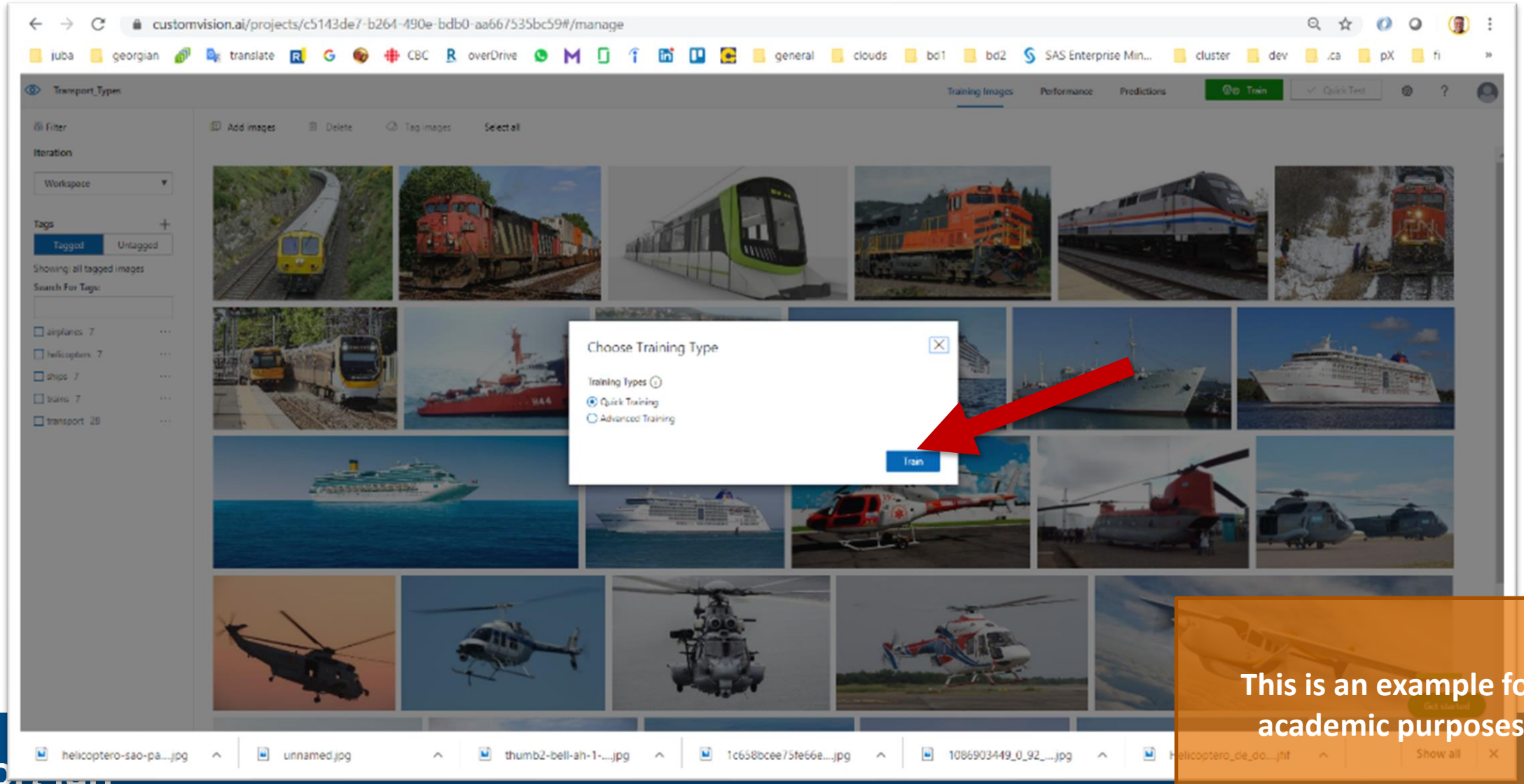
This is an example for academic purposes

TAG EACH IMAGE WITH A LABEL



> Image Classifier | Practice

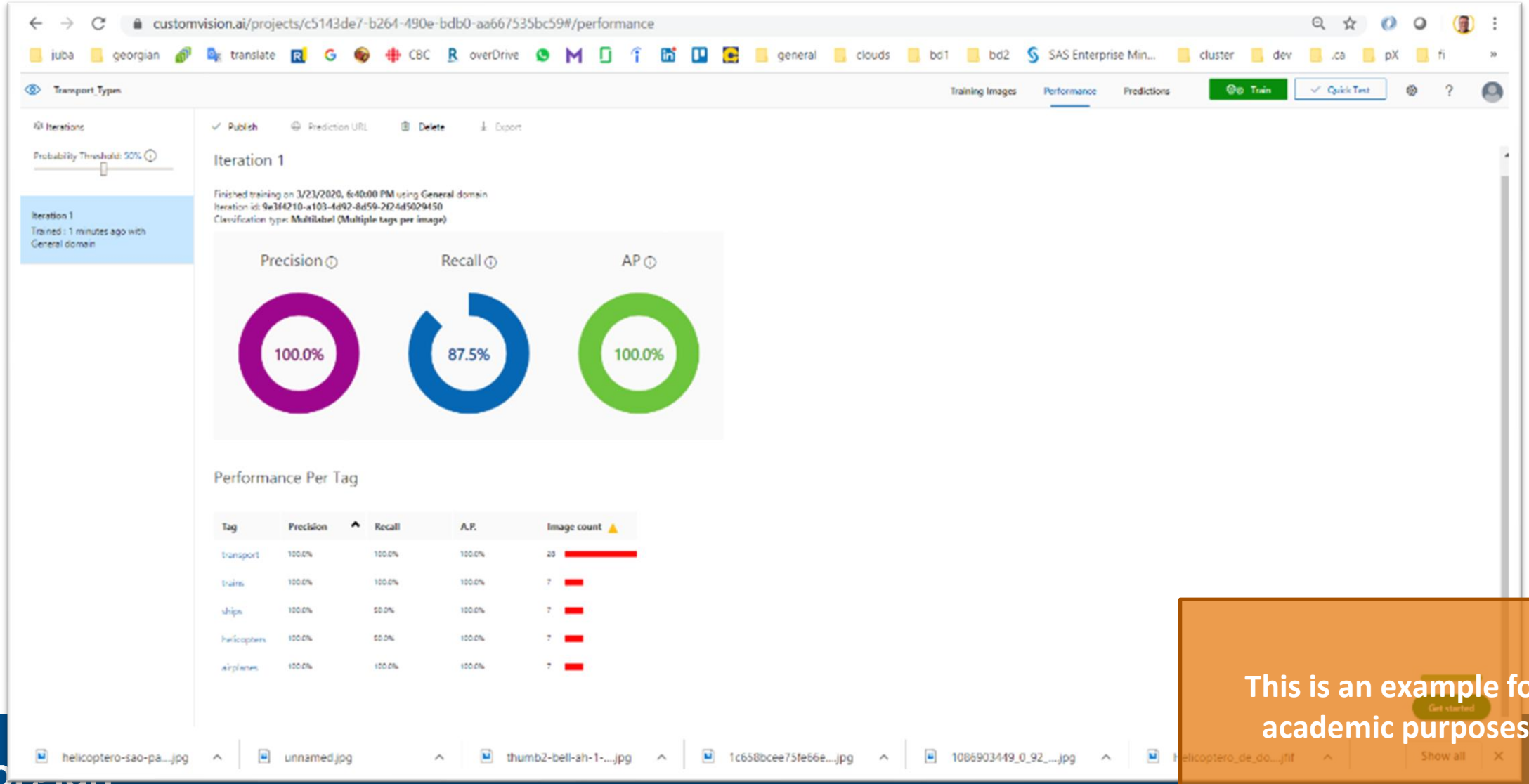
TRAIN YOUR MODEL WITH THE IMAGES AND TAGS



This is an example for academic purposes

> Image Classifier | Practice

CHECK THE RESULTS AND ACCURACY OF YOUR MODEL



This is an example for academic purposes

Georgian

References

> References

- Big Data Analytics Program, 2019/2020 – Georgian College, Barrie, Ontario
- Fair Learn – GitHub Repository – https://github.com/fairlearn/fairlearn
- A Tutorial on Fairness in Machine Learning - Ziyuan Zhong - https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb
- GDPR Regulation - (EU) 2016/679 (General Data Protection Regulation) - https://gdpr.eu/
- AWS, Gartner Report, 2020 Magic Quadrant for Cloud, https://pages.awscloud.com/GLOBAL-multi-DL-gartner-mq-cips-2020-learn.html
- Microsoft, Azure Portal, https://portal.azure.com/#home
- Microsoft, Custom Vision, https://www.customvision.ai/
- NY Times, https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html
- Wisetrend, https://www.wisetrend.com/on-premise-vs-cloud-ocr-data-capture-licensing/
- BMC Software website, https://www.bmc.com/blogs/saas-vs-paas-vs-iaas-whats-the-difference-and-how-to-choose/
- Cleo.com, https://www.cleo.com/blog/knowledge-base-on-premise-vs-cloud
- Wikipedia, AI Accelerator, https://en.wikipedia.org/wiki/AI_accelerator
- Wikipedia, Application-specific integrated circuit, https://en.wikipedia.org/wiki/Application-specific_integrated_circuit
- Lambda, website, https://lambdalabs.com/gpu-workstations/vector
- Microsoft, Azure Data Platform End-to-End, Implement a Modern Data Platform Architecture, Official Material

> References (2)

- Big Data Analytics Program, 2019/2020 – Georgian College, Barrie, Ontario
- Lambda Architecture, Databricks, website, https://databricks.com/glossary/lambda-architecture
- NetApp, What is NVMe?, website, https://www.netapp.com/data-storage/nvme/what-is-nvme/
- Gpost, What is a NVMe M.2 SSD and How Fast is it? Website, https://www.groovypost.com/reviews/what-is-nvme-m2-ssd-drive-how-fast-is-it/
- AI Benchmark.com, website, https://ai-benchmark.com/
- Juniper Networks, What is AI for networking?, website, https://www.juniper.net/us/en/products-services/what-is/ai-networking/
- Leaseweb blog, website, https://blog.leaseweb.com/2019/07/04/infrastructure-requirements-ai/
- Data Iku, ebook, 2021 Trends: Where Enterprise AI is headed next?
- PwC, Responsible AI Survey 2020
- Stanford University, Machine Learning Systems Design, Chip Huyen, cs329s.stanford.edu
- Microsoft, Success by Design Implementation Guide, First Edition, 2021
- Microsoft Learn – Courses & Certifications – website, https://docs.microsoft.com/en-us/learn/

END OF DAY 3