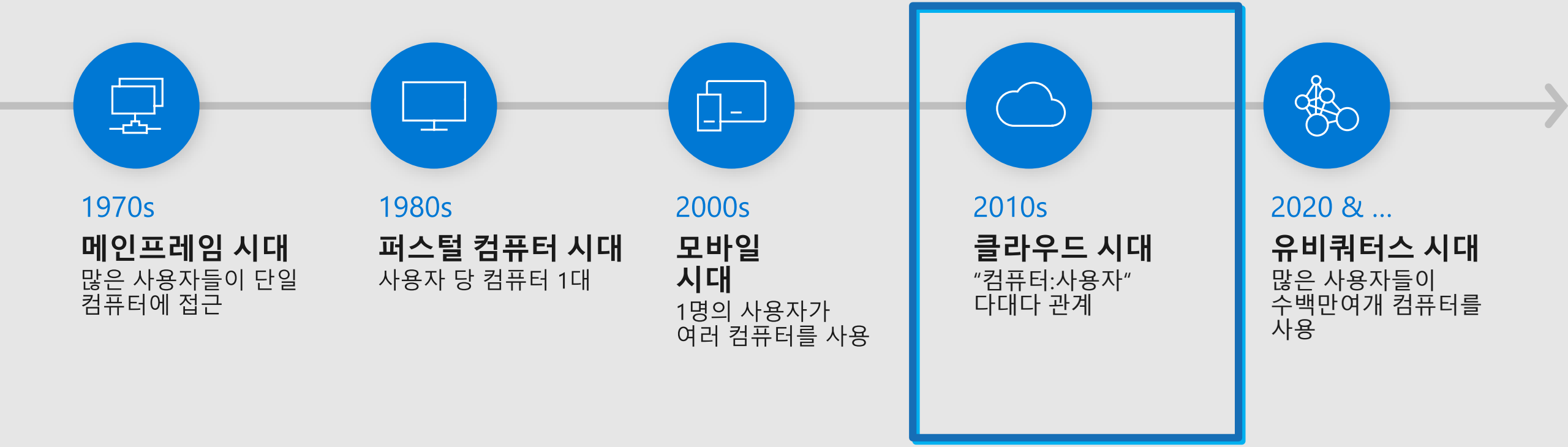
Microsoft

클라우드로 Deep Learning을 다루는 다양한 전략 및 방법

Ian Choi, Developer Product Marketing Manager, Korea

클라우드 컴퓨팅 시대와 함께 하는 변화



Source: Being Human: Human-Computer Interaction in the year 2020, Microsoft Research, April 2008
http://research.microsoft.com/en-us/um/cambridge/projects/hci2020/

A NEW ERA OF COMPUTING

PC INTERNET
WinTel, Yahoo!
1 billion PC users

MOBILE-CLOUD
iPhone, Amazon AWS
2.5 billion mobile users

AI & IOT
Deep Learning, GPU
100s of billions of devices

1995 2005 2015

Microsoft AI 비전

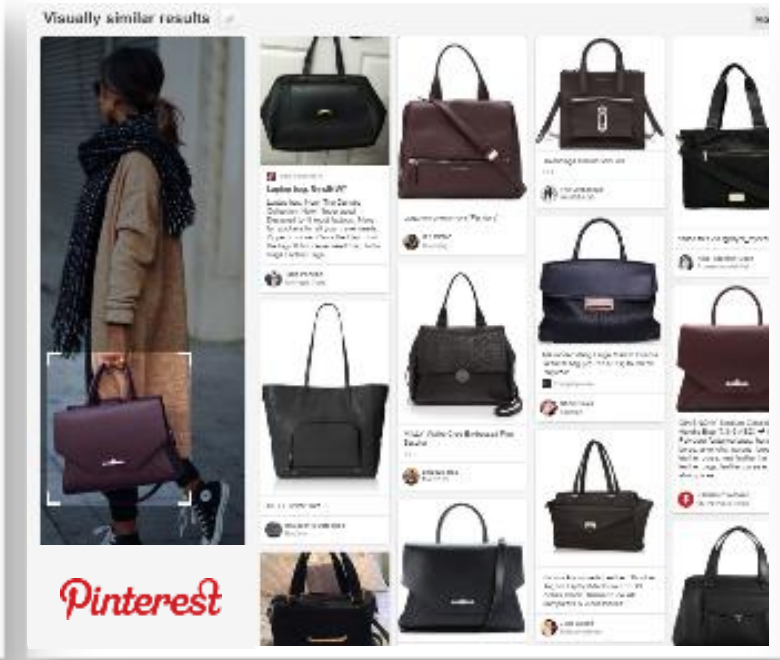
Empower developers
to **innovate**

Empower organizations to
transform industries

Empower people to
transform society

AI IS EVERYWHERE

“Find where I parked my car”



“Find the bag I just saw in this magazine”

“What movie should I watch next?”

“

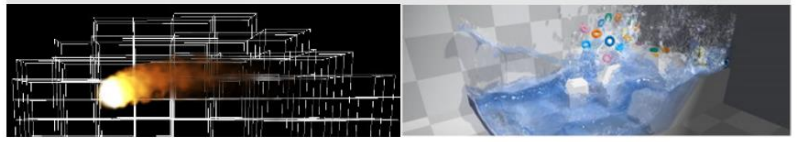
OUR GOAL IS TO **DEMOCRATIZE A.I.**
TO EMPOWER EVERY PERSON
AND EVERY ORGANIZATION
TO ACHIEVE MORE

”

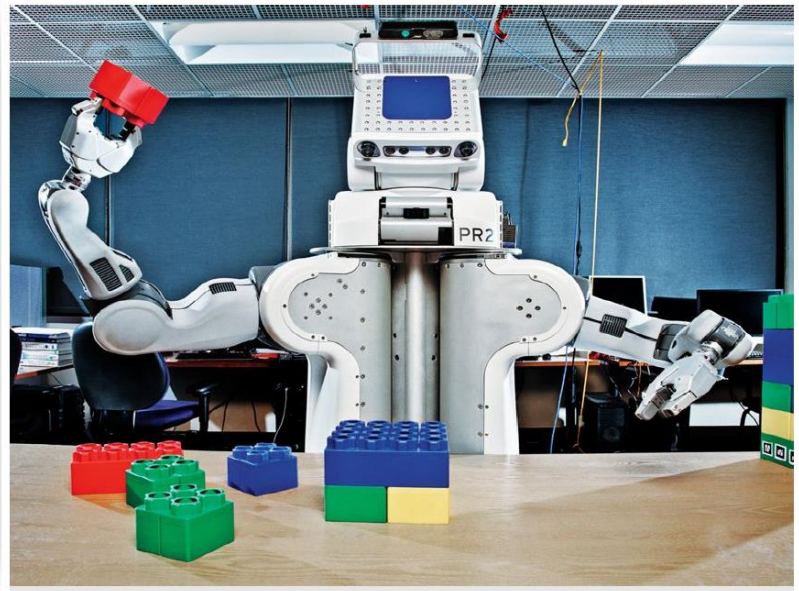
SATYA NADELLA

NVIDIA
“THE AI COMPUTING COMPANY”

GPU Computing



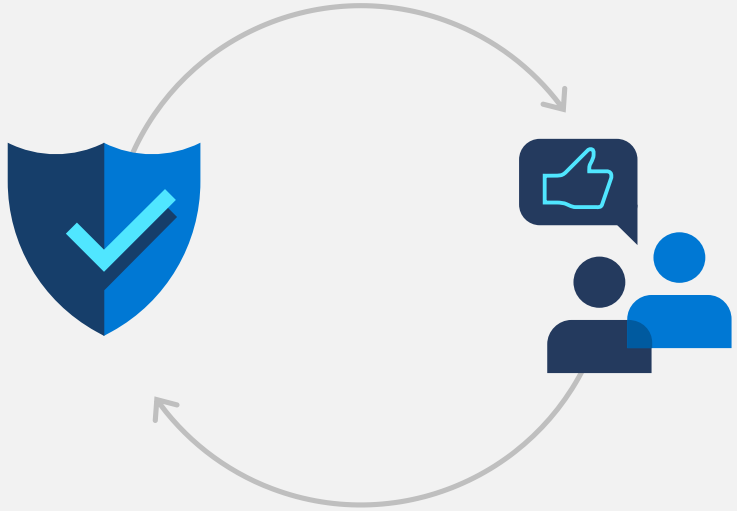
Computer Graphics



Artificial Intelligence

Microsoft’s AI approach

Responsible AI



Trusted AI

Fairness • Reliability • Inclusivity • Privacy • Transparency • Accountability

FUELING ALL INDUSTRIES

Increasing public safety with smart video surveillance at airports & malls

Providing intelligent services in hotels, banks and stores

Separating weeds as it harvests, reduces chemical usage by 90%

NVIDIA/Microsoft Cloud Partnership (2017)

MSFT Services

Cortana

Office 365

Microsoft SQL Server

bing

Translator

Developers

Intelligence
APIs

Cognitive Services

Bot Framework

Cortana

Azure

Public Cloud

N-Series Virtual Machine

CNTK

AI & Research Group

Microsoft/NVIDIA Cloud Partnership (2017)

Artificial Intelligence and Research

Increasing GPU research/engagements

- Cognitive Toolkit (CNTK) joint development and ongoing GPU optimizations
- Joint efforts to simplify GPU use by developers

Azure

Growing N-Series VMs offering in the Public Cloud

- GRID M60 for Visualization and K80 for Compute today
- Pascal (P40 and P100) instances announced
- Volta v100 coming in December

GPU Accelerated Products

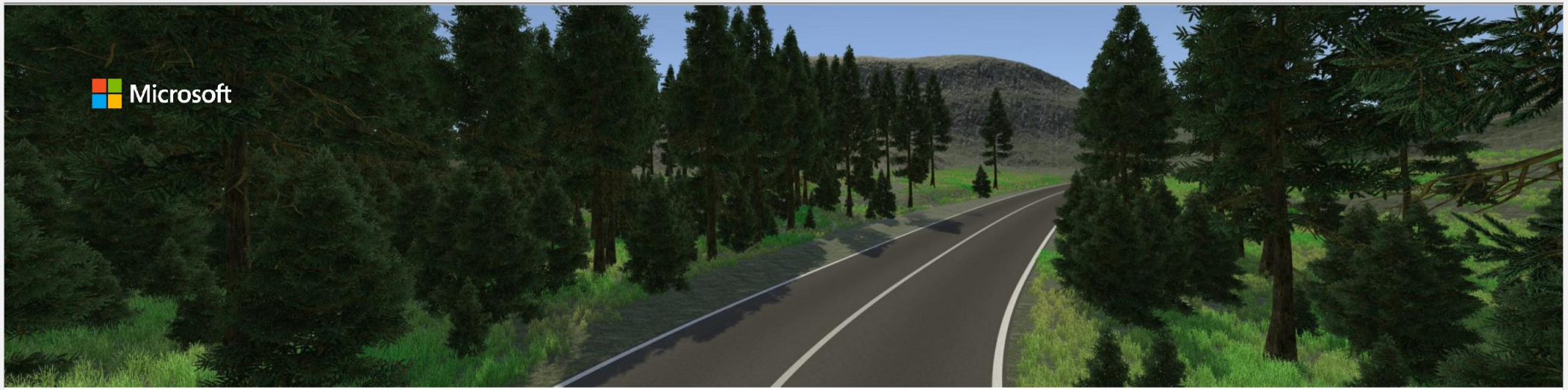
Example research making to commercial products:

- Image within Image Bing Search
- Microsoft Translator
- SQL adding AI capabilities

Open Cloud Server Infrastructure

- HGX-1 Server Design announced at OCP Summit
- Ongoing partnership on next generation designs

클라우드 & GPU: 보통... for IaaS?



Audi technology partner EFS uses deep learning to analyze roads for self-driving vehicles

Based in Gaimersheim, Germany, EFS is the number one partner of Audi in chassis development. It examines and helps implement future-looking technologies, including automated driving. As part of its research efforts, the company used Azure NC-series virtual machines powered by NVIDIA Tesla P100 GPUs to drive a deep learning AI solution that analyzes high-resolution two-dimensional images of roads. The purpose is to give self-driving vehicles a better understanding of those roads. EFS proved that the concept works, and the company can now move ahead with product development.

EFS ELEKTRONISCHE FAHRWERKSYSTEME

Products and Services	Organization Size	Industry	Country	Partner
Microsoft Azure Azure NC-series VMs Azure storage	422 employees	Professional Services	Germany	NVIDIA

클라우드 & GPU: How about PaaS?

Diagnostic services provider uses Azure Machine Learning with NVIDIA GPUs to help end preventable blindness

Diabetes is the leading cause of preventable blindness in the United States, but there was no easy way to diagnose diabetic vision damage through primary care providers. That's why IRIS used Microsoft Azure to help create a platform that can identify diabetic retinopathy before patients suffer from vision loss. Using Azure Machine Learning Package for Computer Vision, the IRIS platform processes images quickly and accurately so doctors can share data with patients and other clinicians, better prevent diabetic blindness, and help reduce healthcare costs.

iris™ Intelligent Retinal Imaging Systems

Products and Services
Microsoft Azure
Azure Functions
Azure Machine Learning
Azure Service Bus
Azure SQL Database

Organization Size
34 employees

Industry
Health Provider

Country
United States

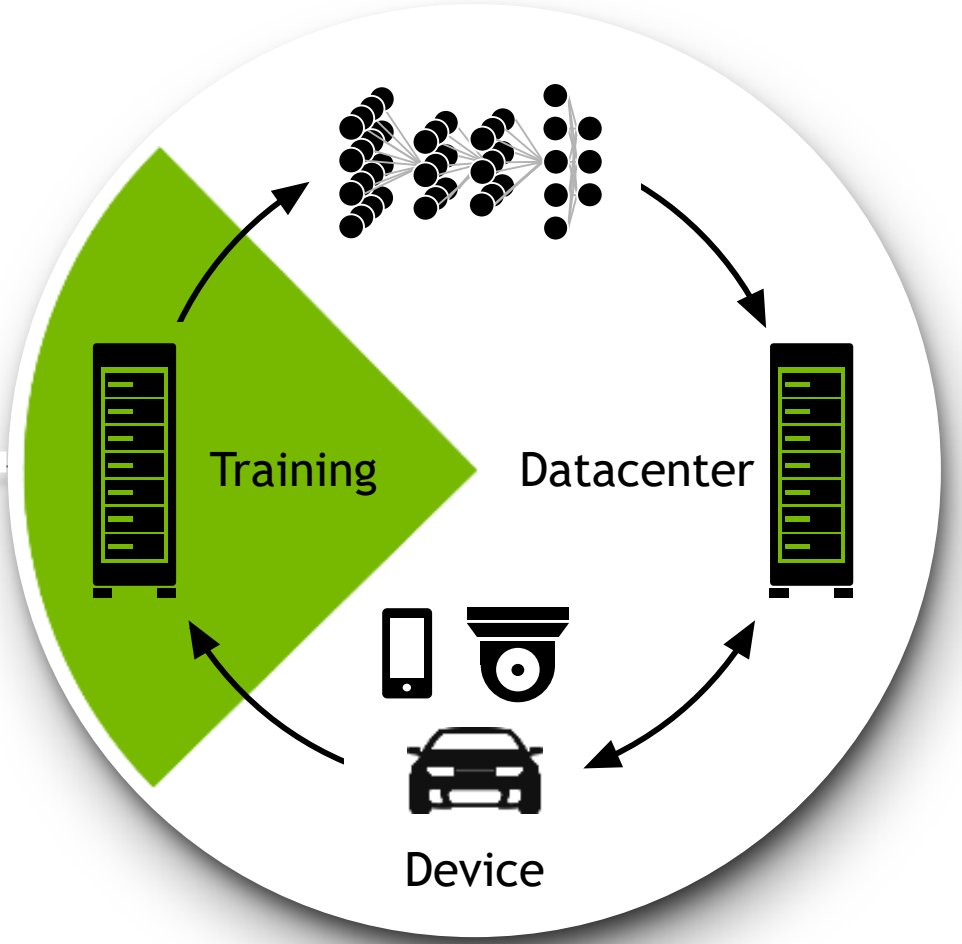
1. GPU와 Deep Learning

GPU DEEP LEARNING IS A NEW COMPUTING MODEL

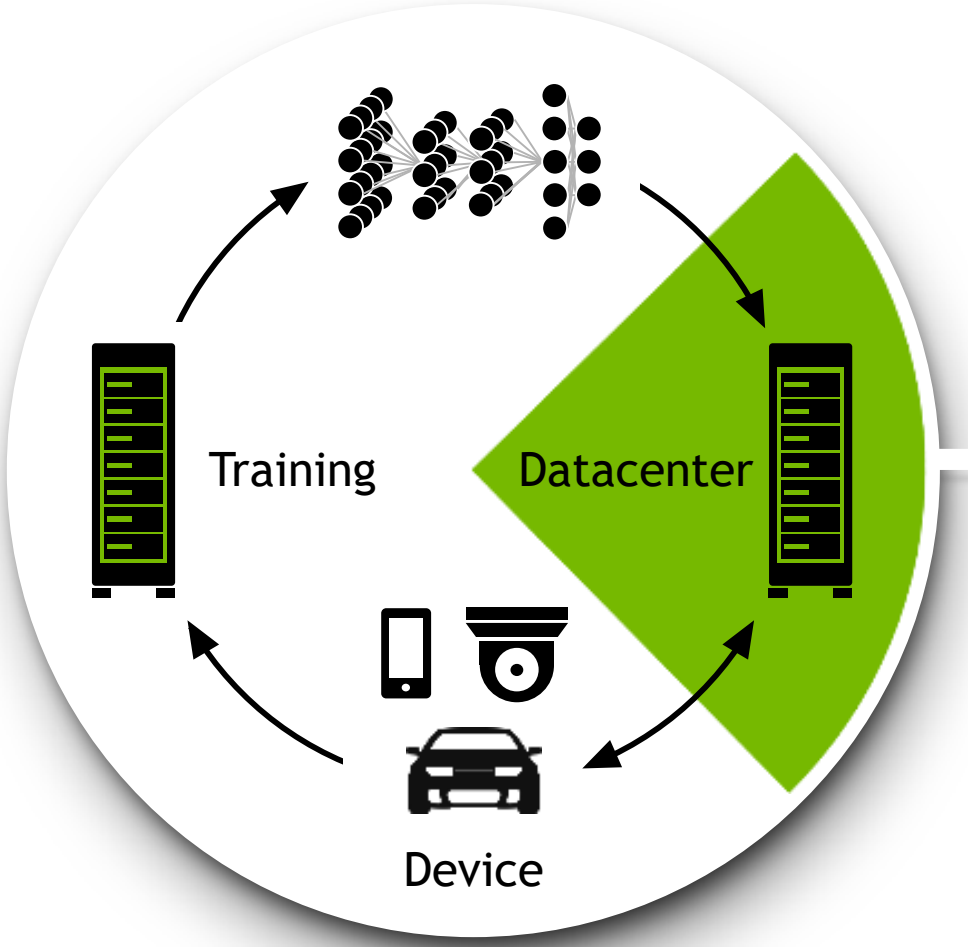
Billions of Trillions of Operations

GPU train larger models, accelerate time to market

TRAINING



GPU DEEP LEARNING IS A NEW COMPUTING MODEL



10s of billions of image, voice, video queries per day

GPU inference for fast response, maximize datacenter throughput

DATACENTER INFERENCING

POWERING THE DEEP LEARNING ECOSYSTEM

NVIDIA SDK accelerates every major framework

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION

SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES

SENTIMENT ANALYSIS

DEEP LEARNING FRAMEWORKS

Caffe

Chainer

DL4J
Deeplearning4j

Mocha.jl
julia

K
KERAS

Microsoft
CNTK

mxnet

Purine

MatConvNet

MINERVA

OpenDeep

Pylearn2

TensorFlow

torch

theano

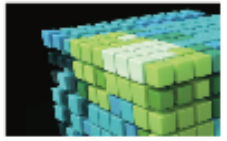
NVIDIA DEEP LEARNING SDK

cuDNN

TensorRT

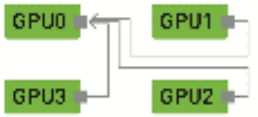
DeepStream SDK

cuBLAS



cuSPARSE

NCCL



developer.nvidia.com/deep-learning-software

CUDA® Deep Neural Network library

NVIDIA cuDNN

Accelerating Deep Learning

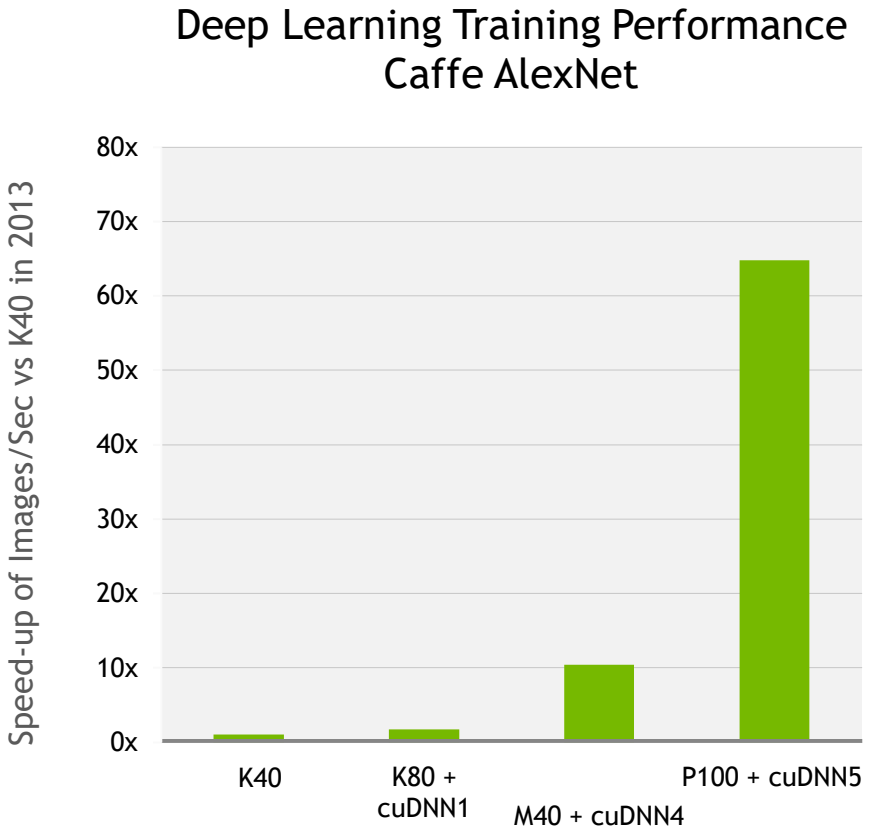
High performance building blocks for deep learning frameworks

Drop-in acceleration for widely used deep learning frameworks such as Caffe, CNTK, Tensorflow, Theano, Torch and others

Accelerates industry vetted deep learning algorithms, such as convolutions, LSTM, fully connected, and pooling layers

Fast deep learning training performance tuned for NVIDIA GPUs

developer.nvidia.com/cudnn



AlexNet training throughput on CPU: 1x E5-2680v3 12 Core 2.5GHz. 128GB System Memory, Ubuntu 14.04
M40 bar: 8x M40 GPUs in a node, P100: 8x P100 NVLink-enabled

"NVIDIA has improved the speed of cuDNN with each release while extending the interface to more operations and devices at the same time."

— Evan Shelhamer, Lead Caffe Developer, UC Berkeley

DEEP LEARNING SUCCESS

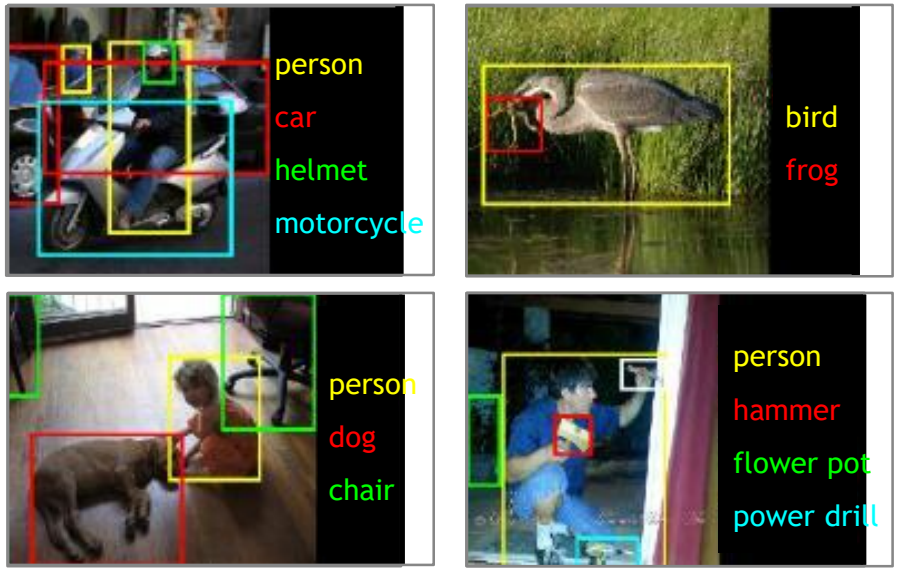
Object classification and localization in images

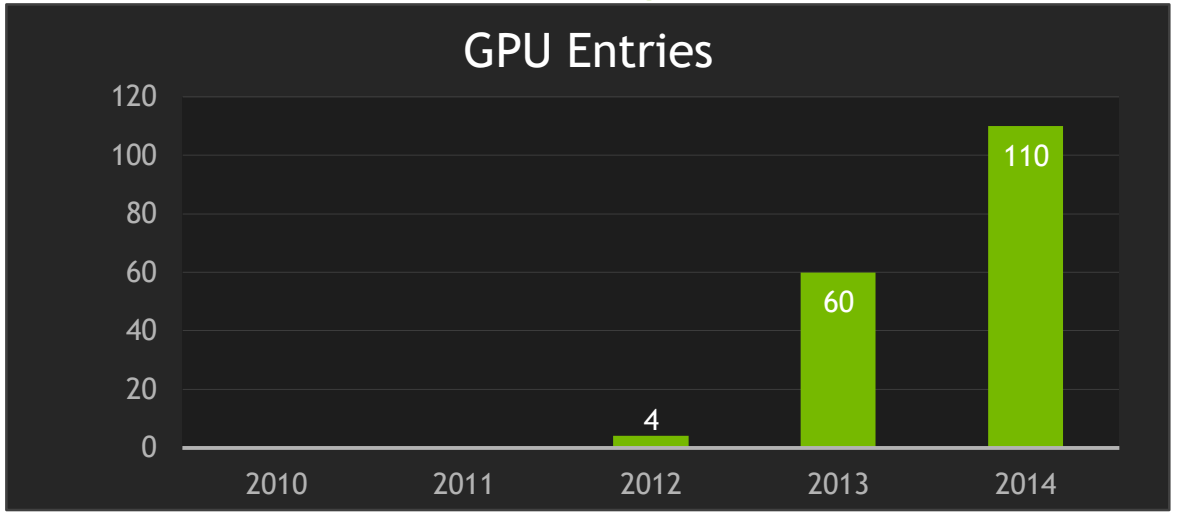
Image Recognition Challenge

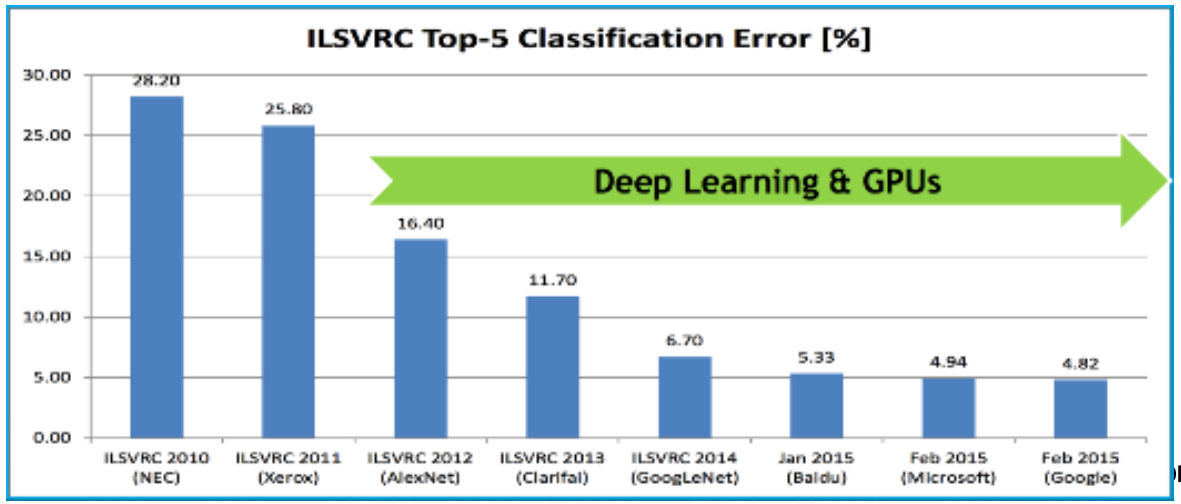
1.2M training images • 1000 object categories

Hosted by

IMAGENET

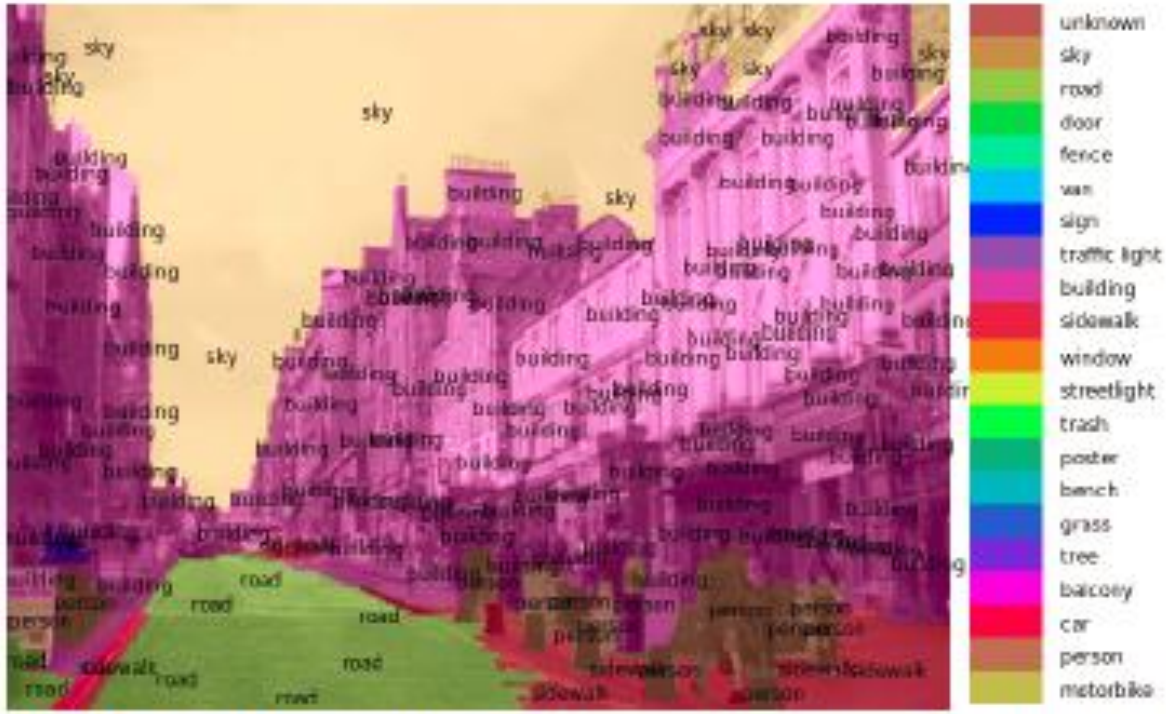






DEEP LEARNING SUCCESS

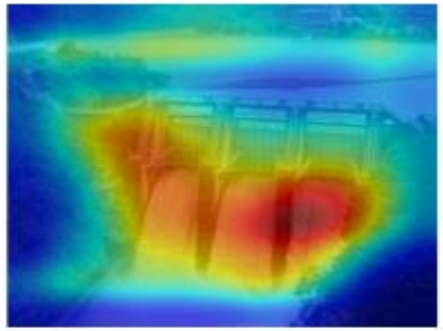
Scene segmentation and classification



Farabet, PAMI 2013

Predictions:

- **Type of environment:** outdoor
- **Semantic categories:** dam:1.00,
- **SUN scene attributes:** naturallight, openarea, man-made, foliage, leaves, moistdamp, runningwater, vegetation, nohorizon, shrubbery
- **Informative region for the category "dam" is:**



http://places.csail.mit.edu/ B. Zhou et al. NIPS 2014

DEEP LEARNING SUCCESS

Activity recognition in video

Google/Stanford

Every Deep Learning Framework is GPU Accelerated

2. 클라우드와 Deep Learning

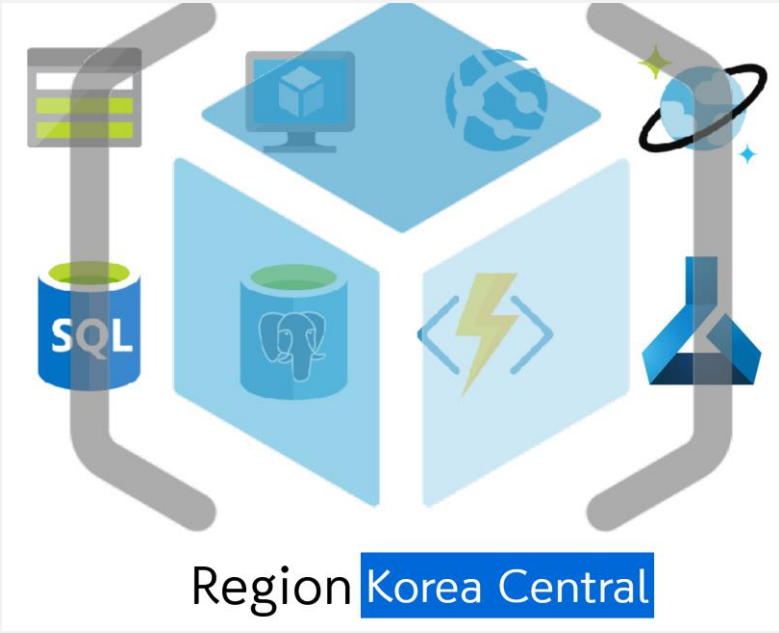
Cifar 10


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

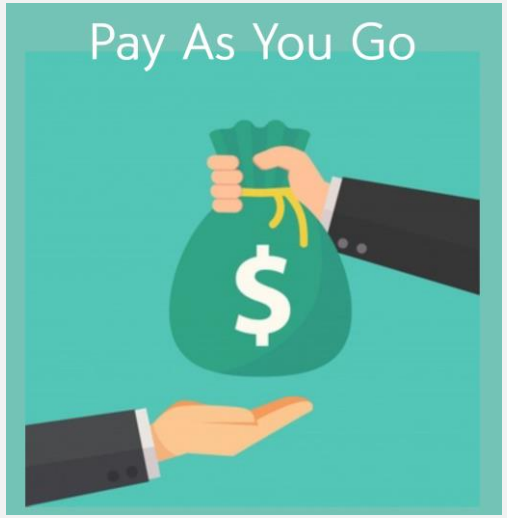
Before we run,

Resource

Resource Group

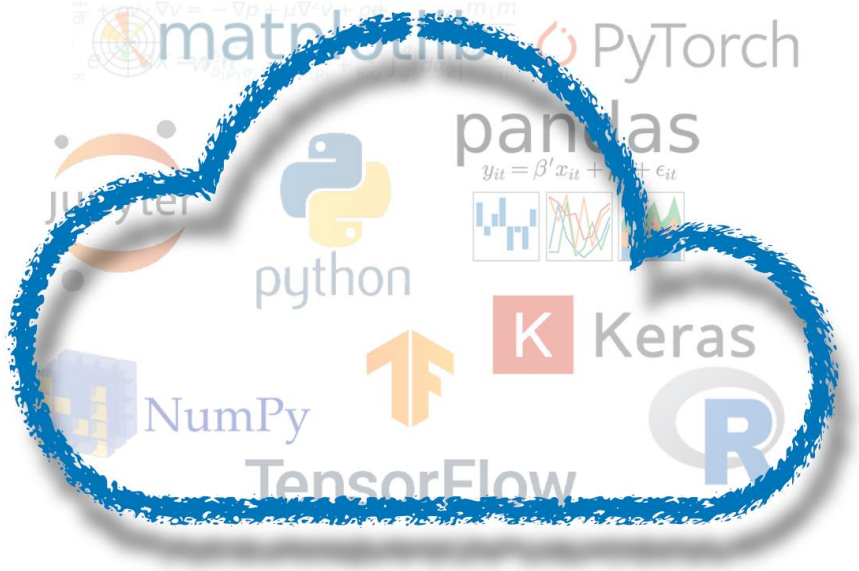


Subscription



CPU based vs. GPU based (가격: 미국 동부 기준)

Notebook Virtual Machine

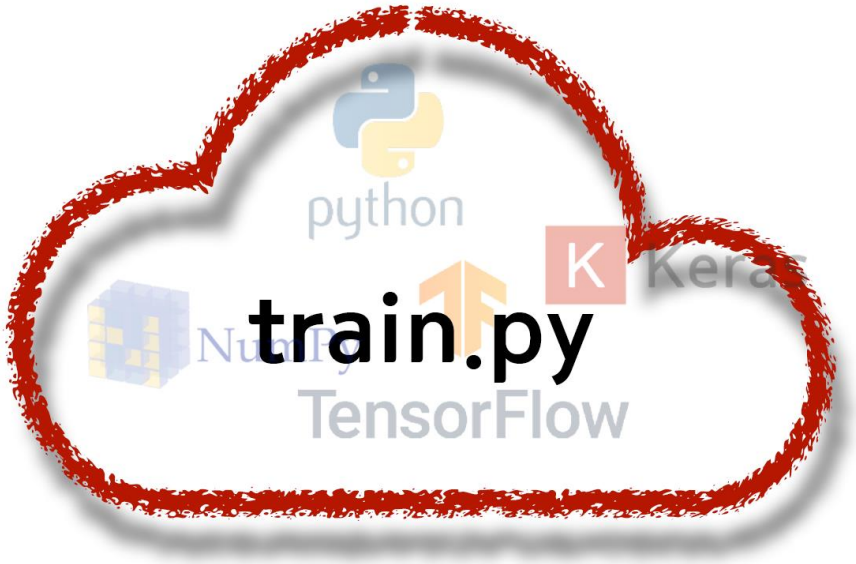


D2_V3 standard

vCPU 2, 8GB RAM, 50GB

😊 $ 0.188/hour

Remote Compute



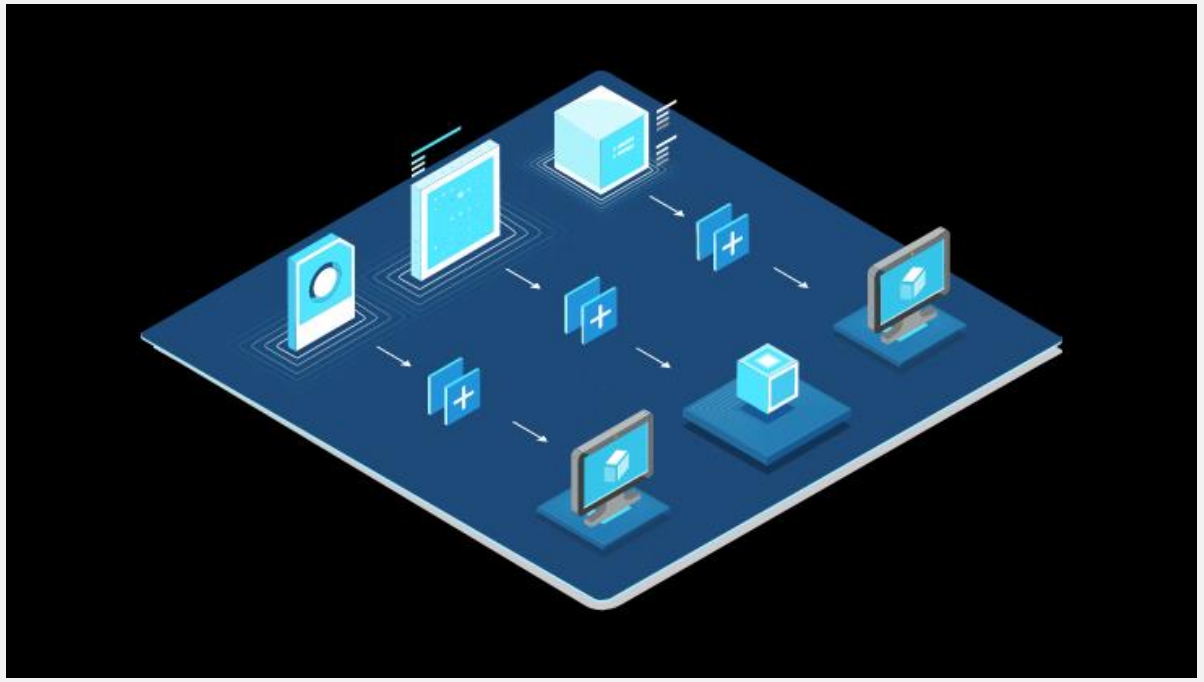
GPU **NC12**

Core 12, 112GB RAM, 680GB

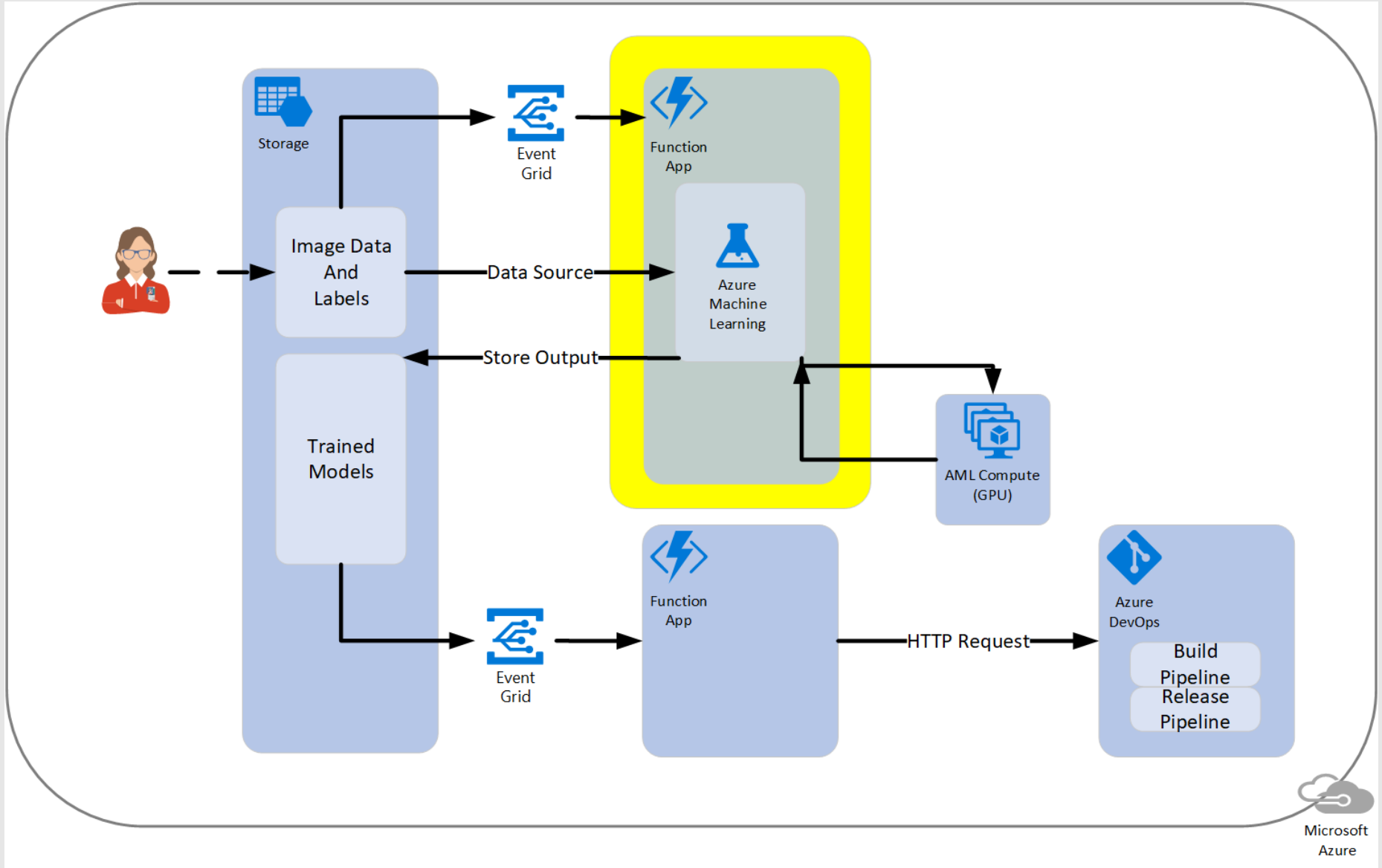
Intel Xeon E5-2690 v3 2.60GHz v3

😱 $ 2.168/hour

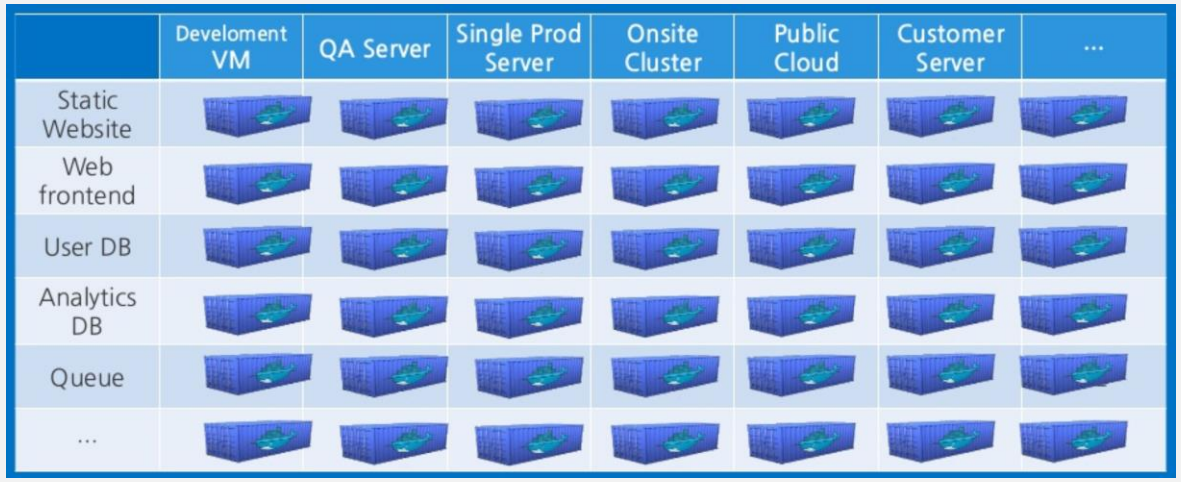
보다 효율적인 방법 1: Spot VM?



보다 효율적인 방법 2: with PaaS (Platform as a Service)

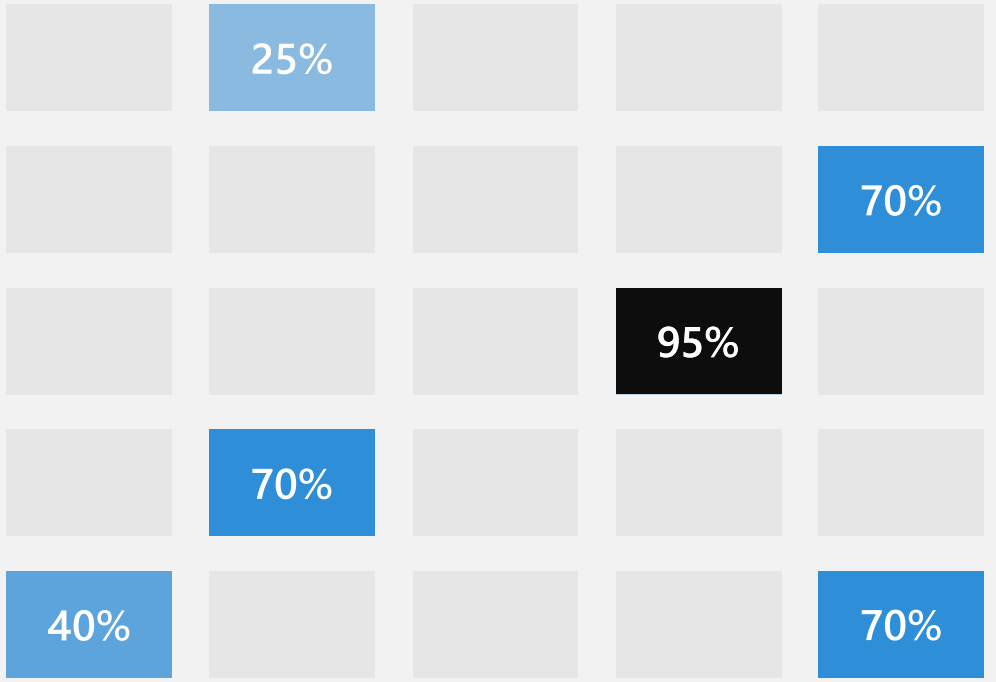


보다 효율적인 방법 3: with containers



Different parameters, variables in deep learning

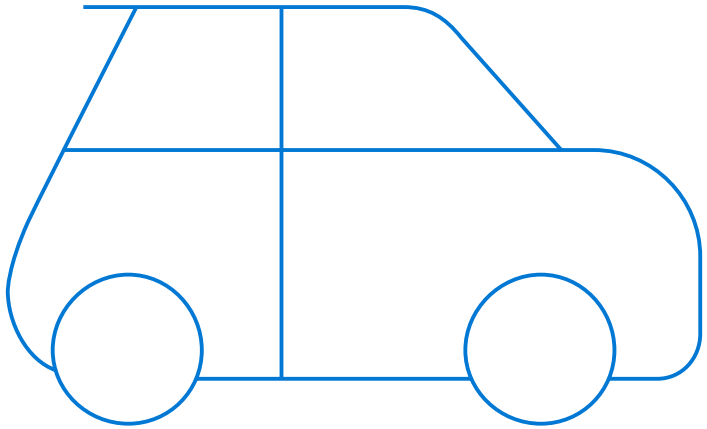
Containers can easier test multiple models in parallel



3. 머신 러닝 서비스를 Production으로 만들 때를 생각해 봅시다

Building your own AI models

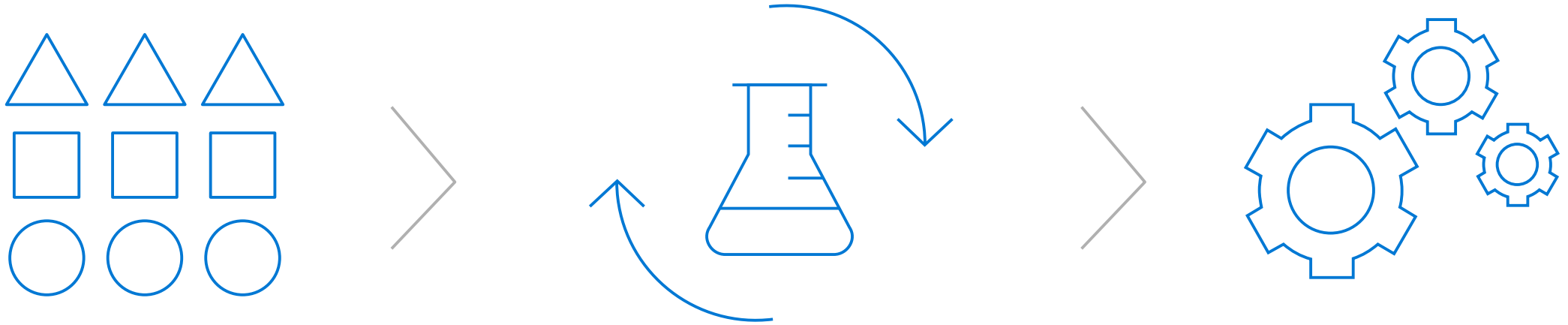
Transforming Data into Intelligence



Q: How much is this car worth?

Building your own AI models

Transforming data into intelligence



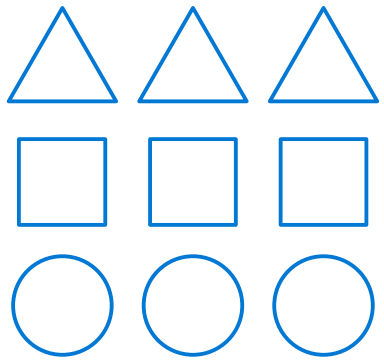
Prepare data

Build and train

Deploy

Building your own AI models

Transforming data into intelligence



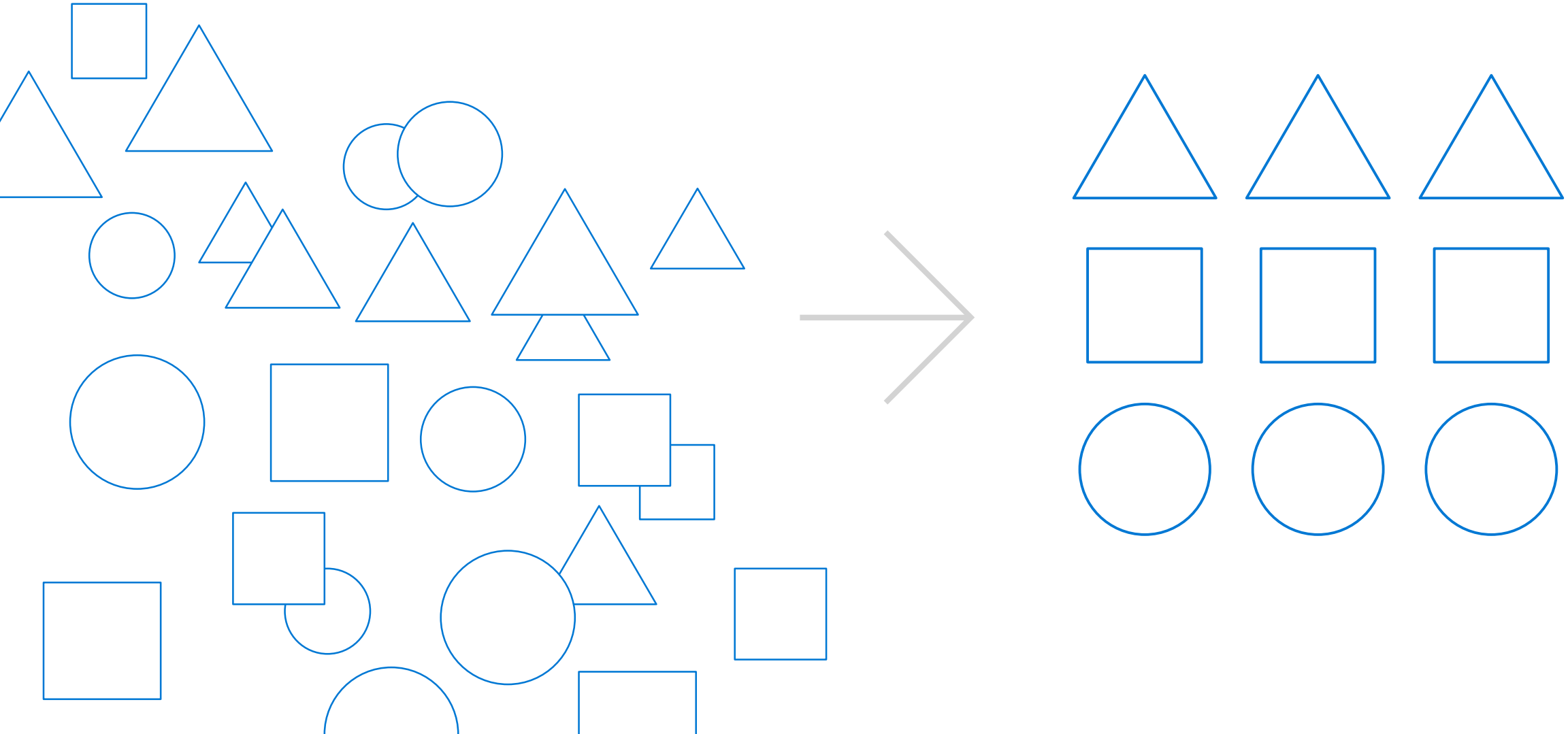
Prepare data

Build and train

Deploy

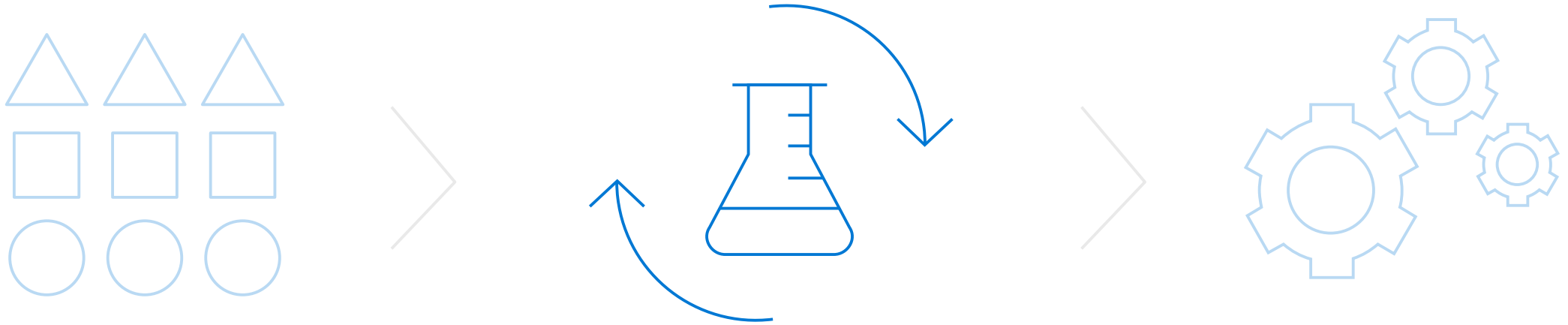
Building your own AI models

Step 1: Prepare data



Building your own AI models

Transforming data into intelligence



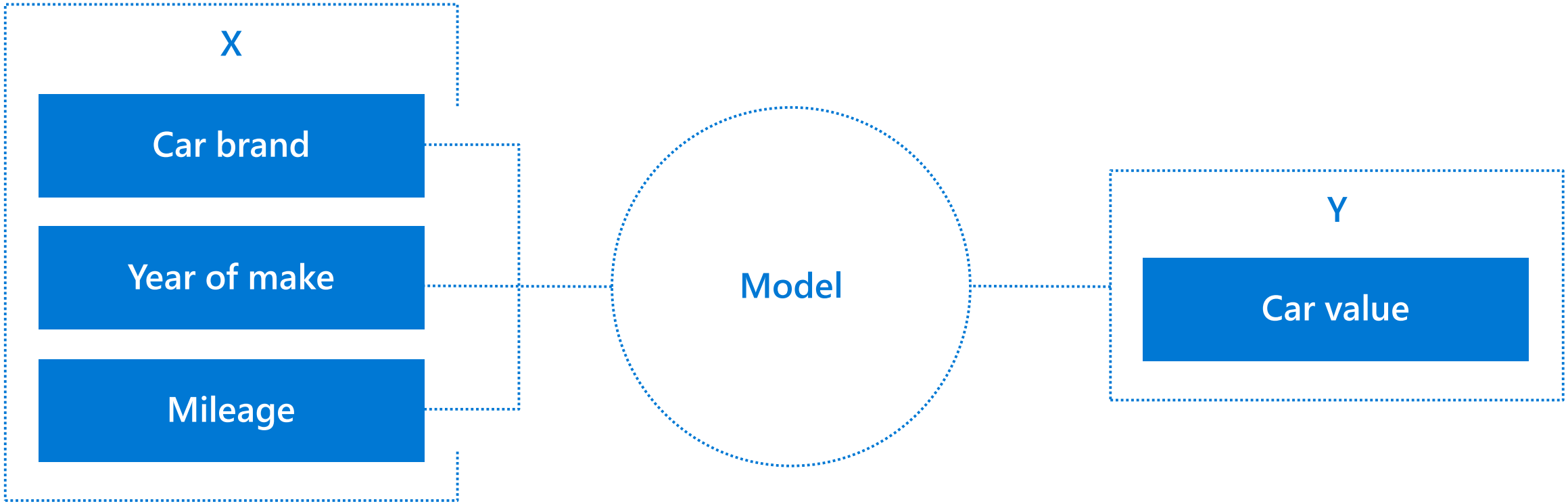
Prepare data

Build and train

Deploy

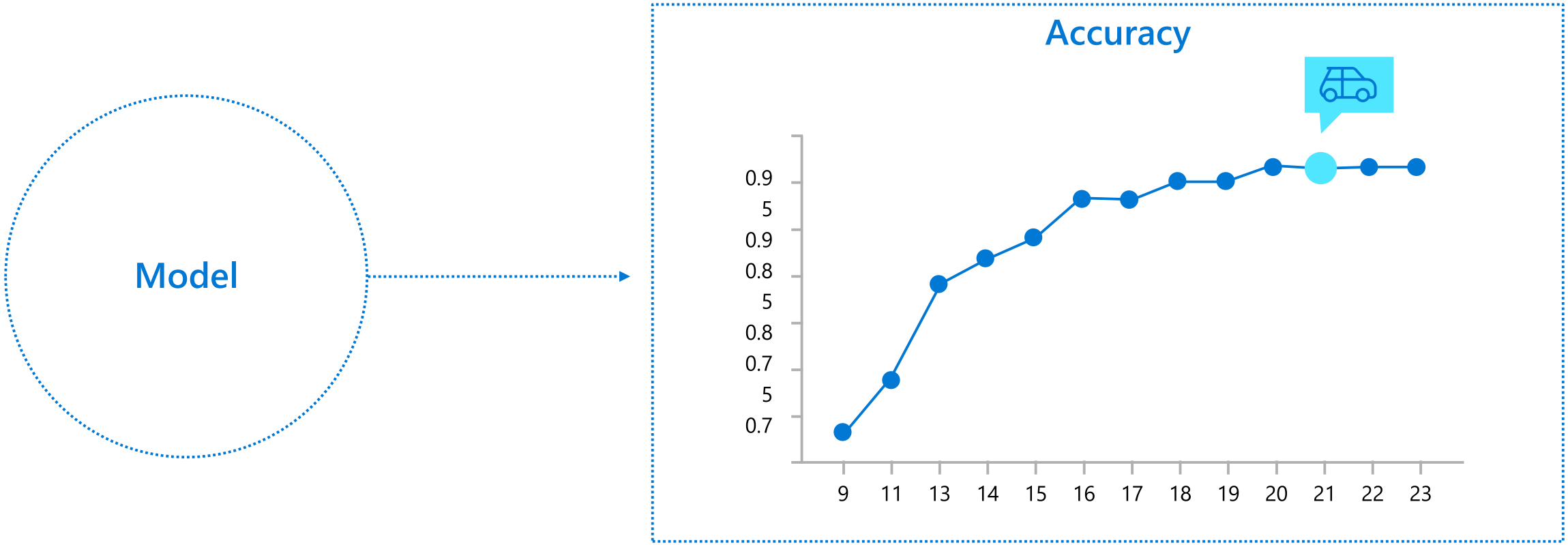
Building your own AI models

Step 2: Build and Train



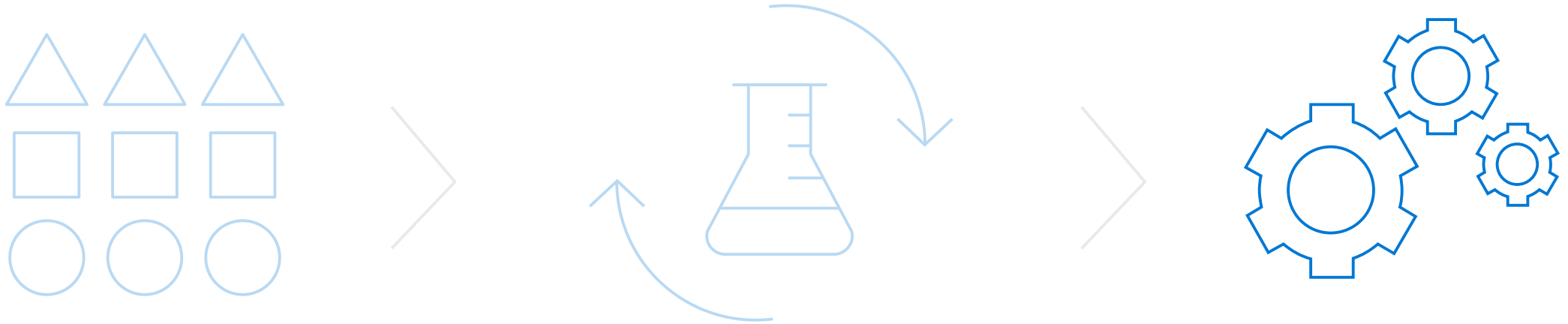
Building your own AI models

Step 2: Build and train



Building your own AI models

Transforming data into intelligence



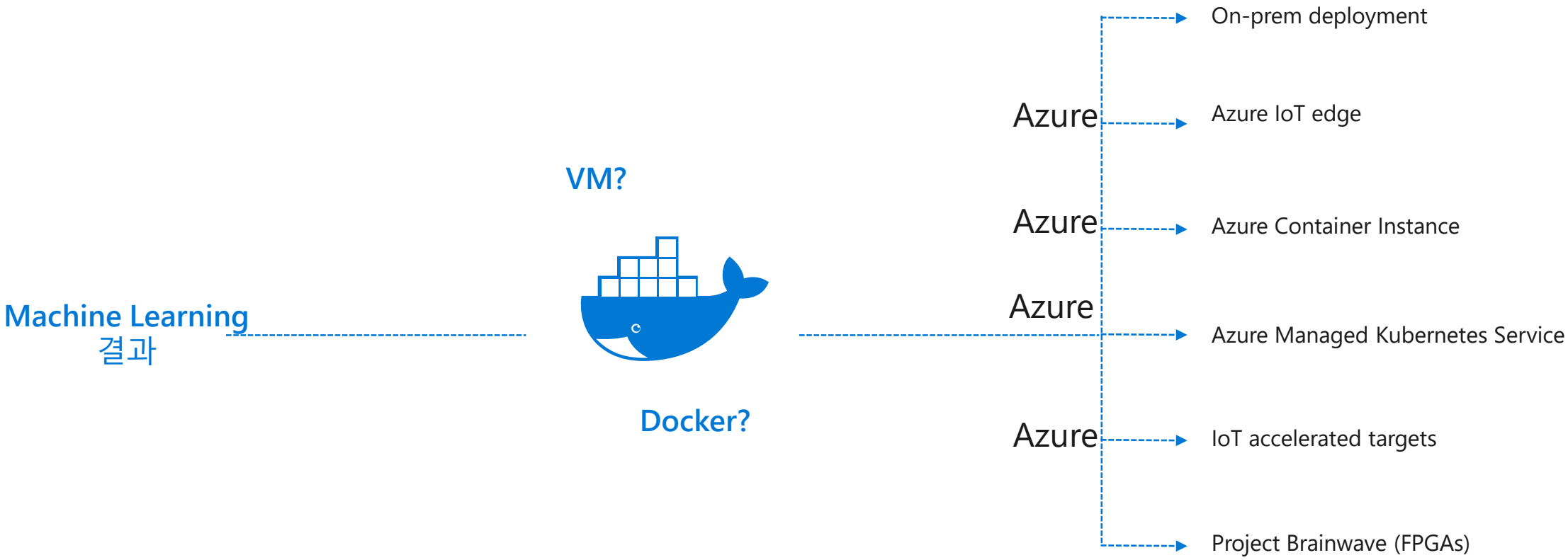
Prepare data

Build and train

Deploy

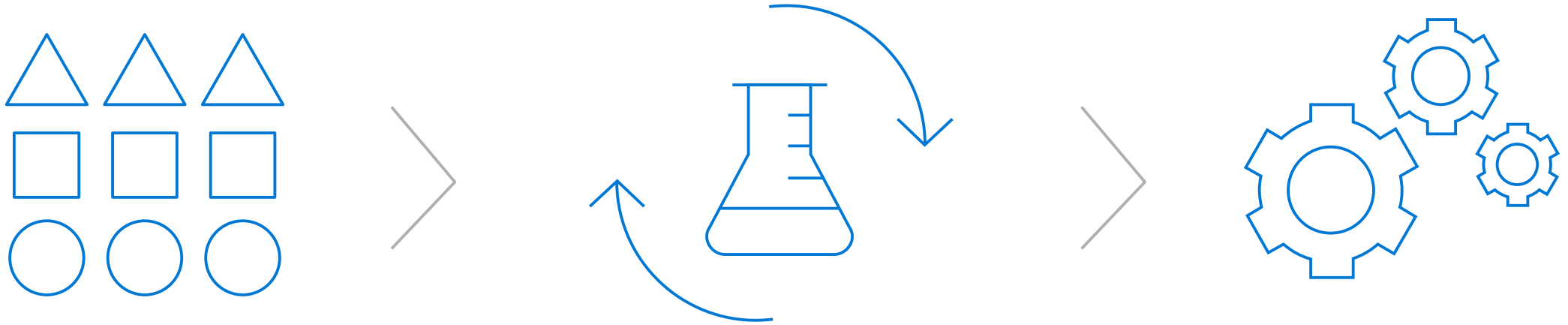
Building your own AI models

Step 3: Deploy



Building your own AI models

Transforming data into intelligence



Prepare data

Build and train

Deploy

Building your own AI models

Transforming data into intelligence

SQL DB

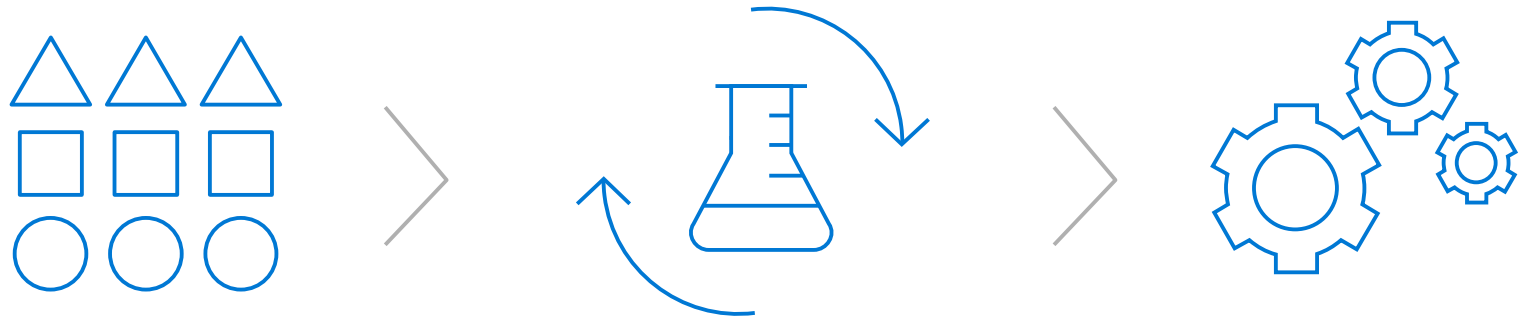
Cosmos DB

Datawarehouse

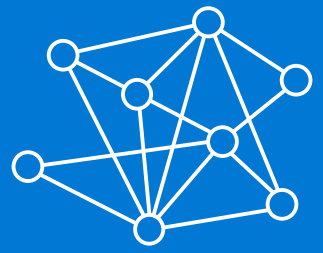
Data lake

Blob storage

...



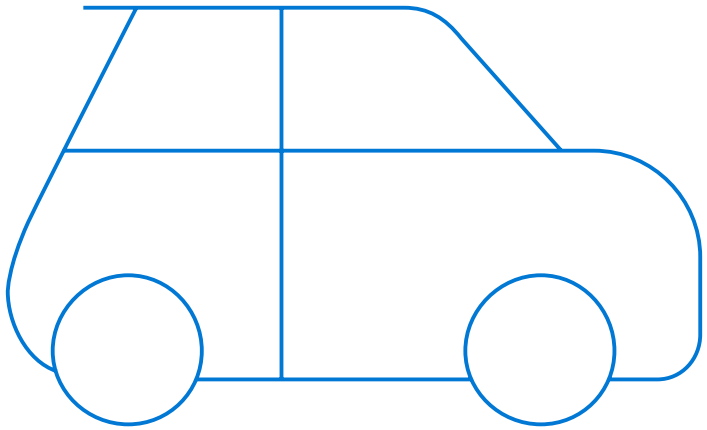
4. 실제 프로덕션에서의 고민



1. 자동화된 머신 러닝/딥러닝

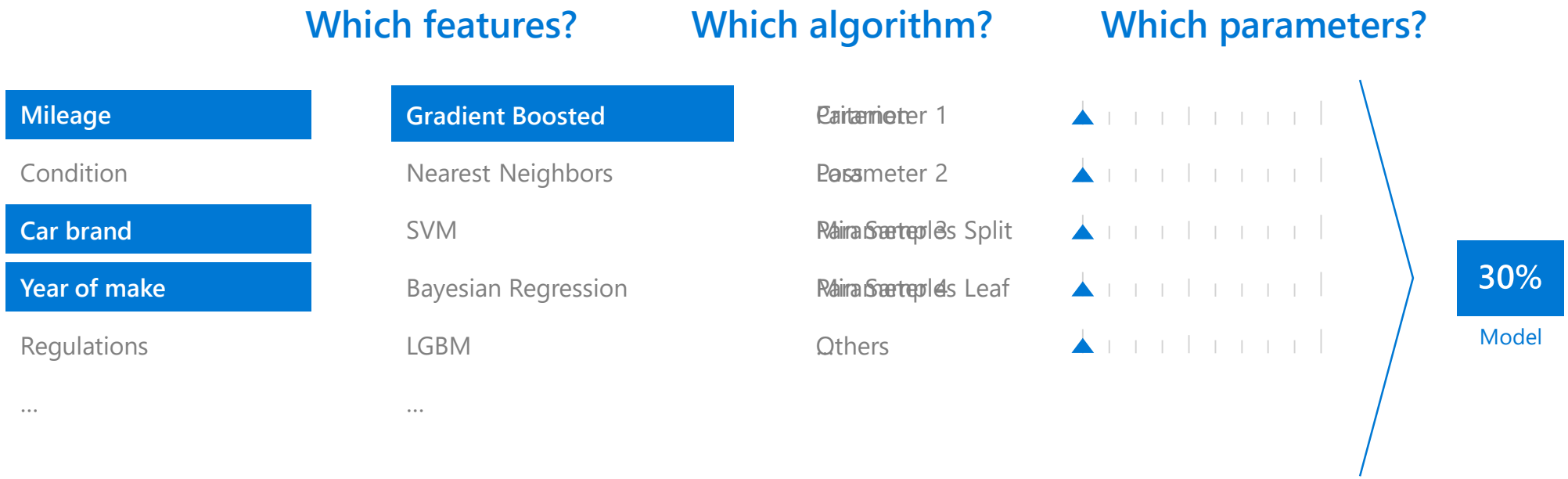
Azure Machine Learning

Automated machine learning



How much is this car worth?

Model creation is typically a time consuming process



Model creation is typically a time consuming process

Which features?

- Mileage
- Condition
- Car brand
- Year of make
- Regulations
- ...

Which algorithm?

- Gradient Boosted
- Nearest Neighbors
- SGD
- Bayesian Regression
- LGBM
- ...

Which parameters?

- Criterion
- Weights
- Min Samples Split
- Min Samples Leaf
- XYX

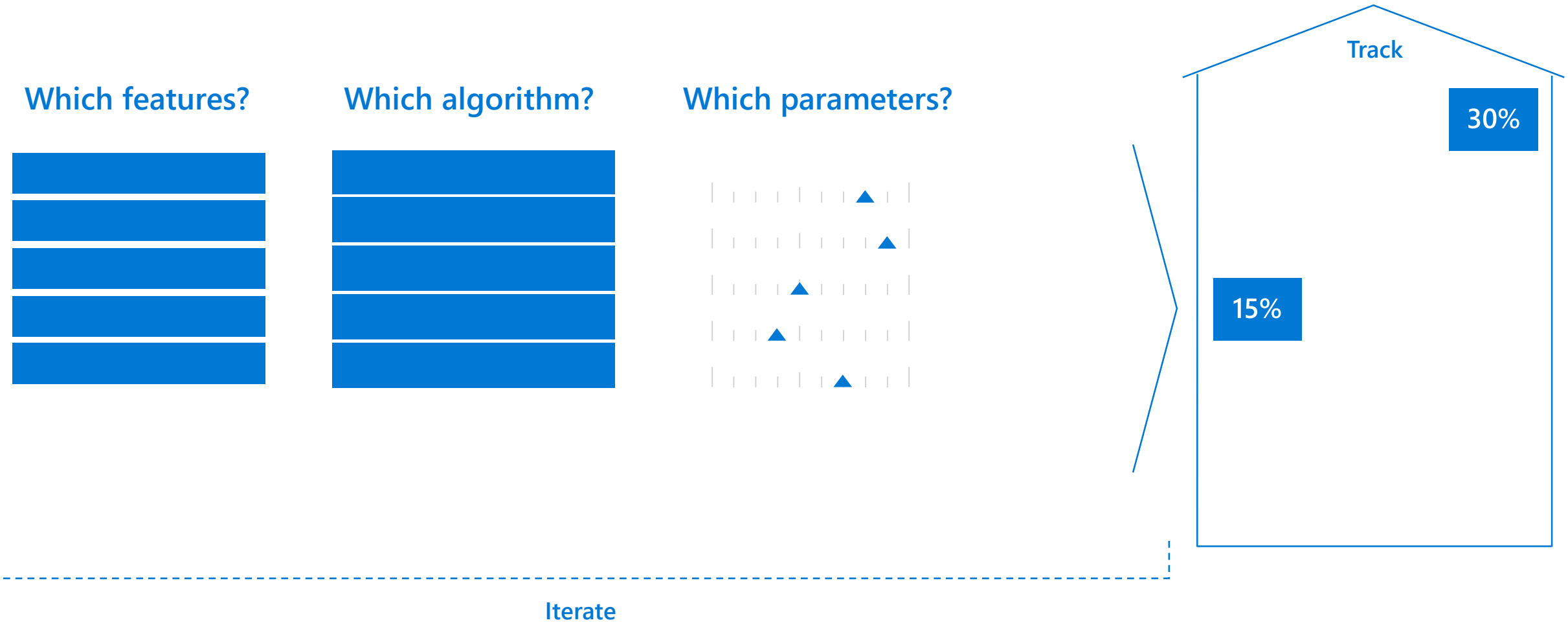
Track

30%

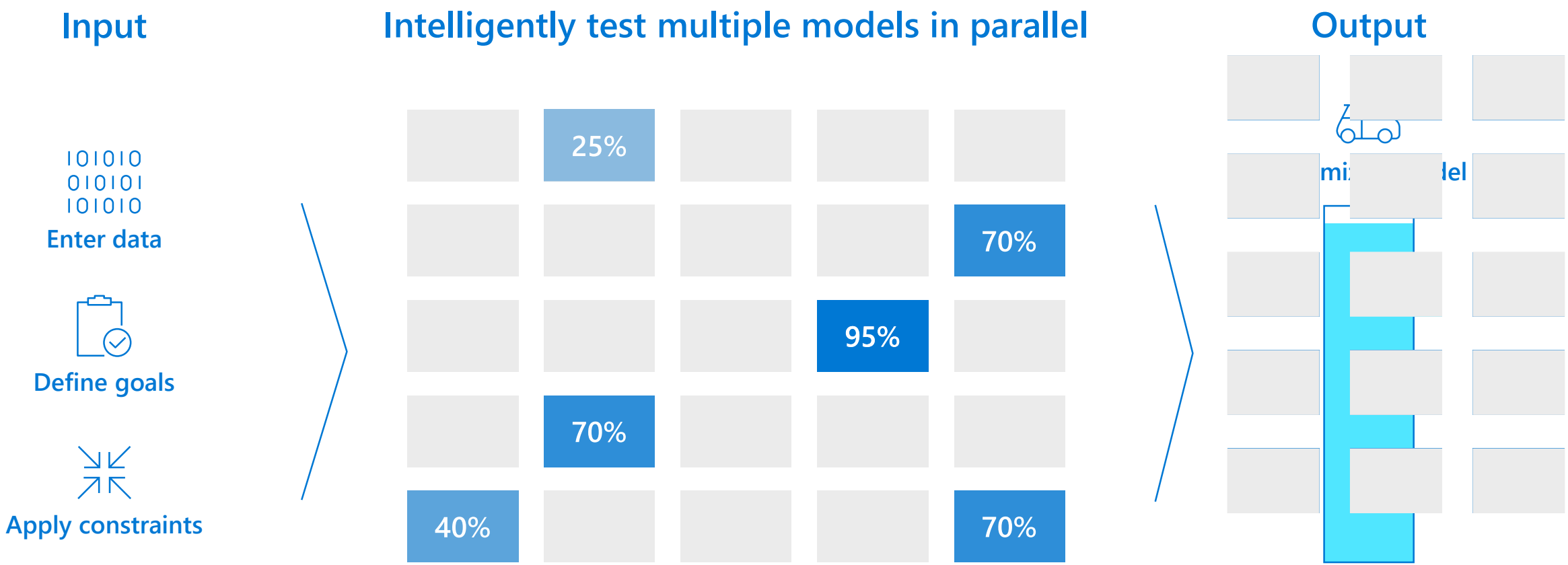
Model

Iterate

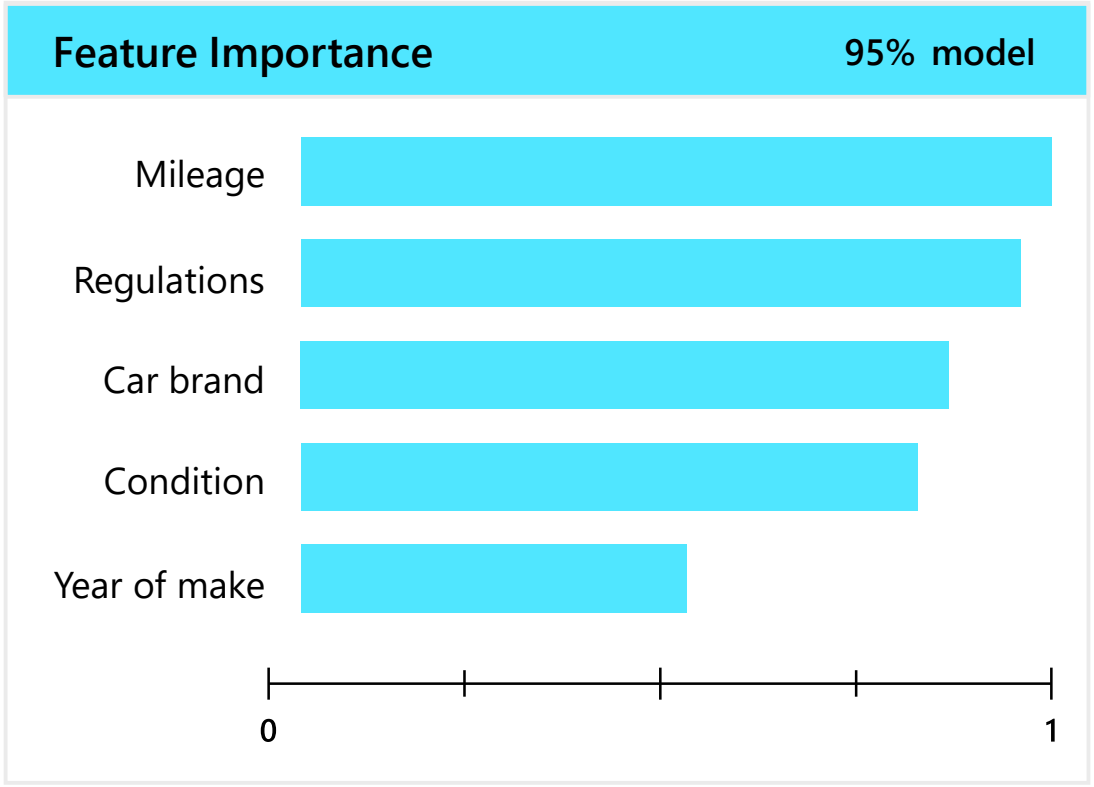
Model creation is typically a time consuming process

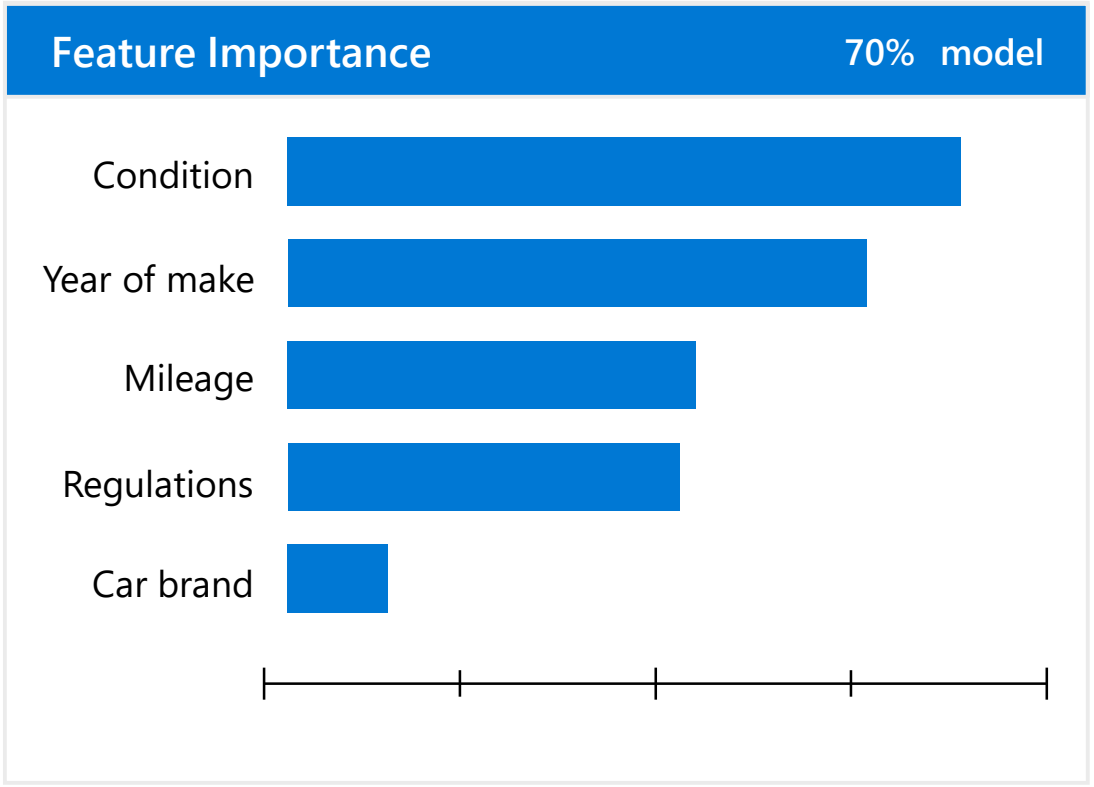


Automated Machine Learning accelerates model development

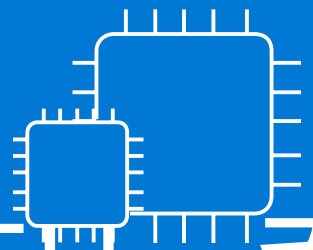


Understand the inner workings of ML by analyzing feature importance



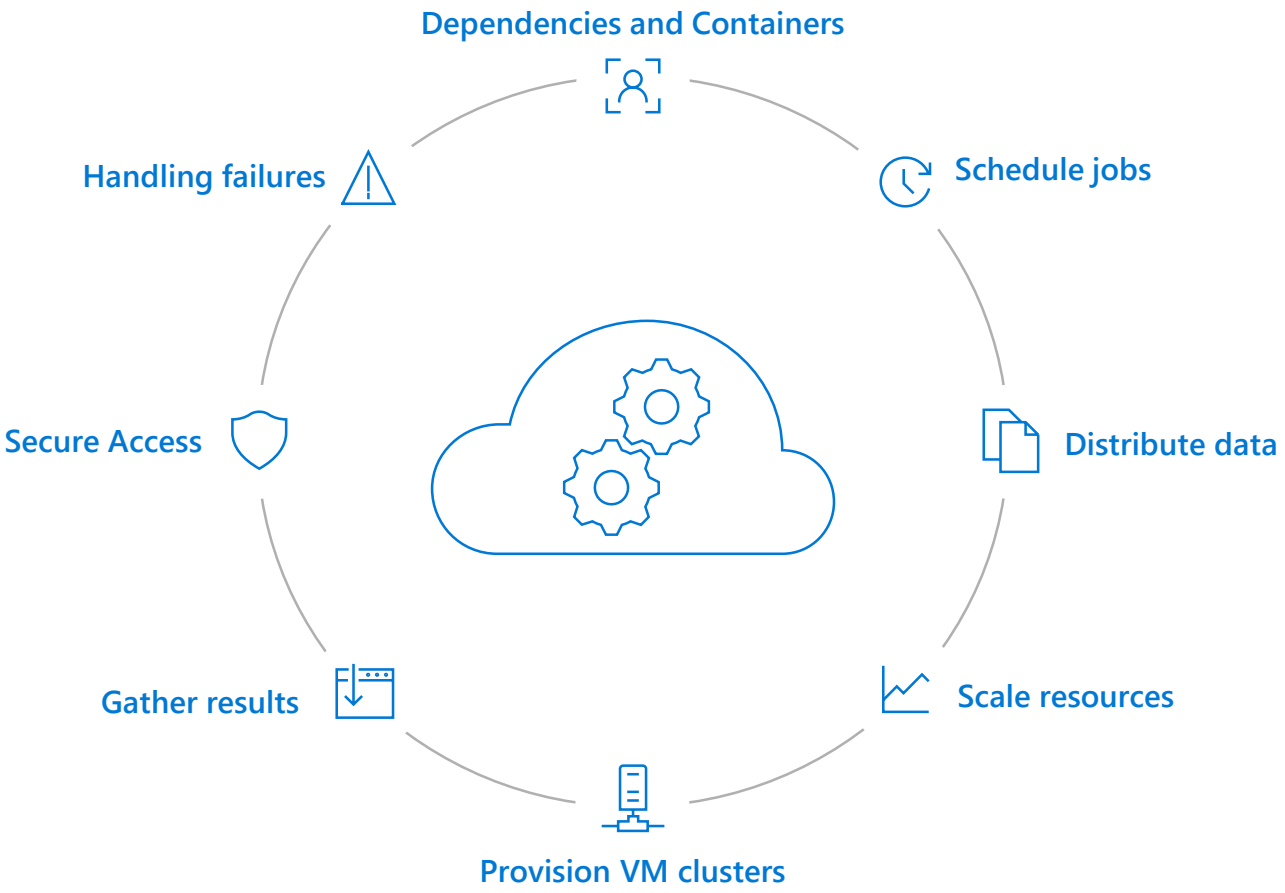


Enable model explain-ability for every automated ML iteration, not just the optimal model



2. 관리가 되는 컴퓨팅 자원 환경

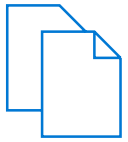
Distributed training on managed compute



Training infrastructure

Dependencies and Containers

Leverage system-managed AML compute or bring your own compute



Distribute data

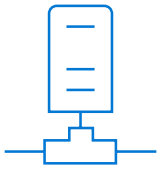
Manage and share resources across a workspace

Schedule jobs

Train at cloud scale using a framework of choice

Scale resources

Autoscale resources to only pay while running a job



Provision clusters

Use the latest NDv2 series VMs with the NVIDIA V100 GPUs

Powerful infrastructure

Accelerate deep learning

CPUs

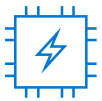
General purpose machine learning

D, F, L, M, H Series

GPUs

Deep learning

N Series



FPGAs

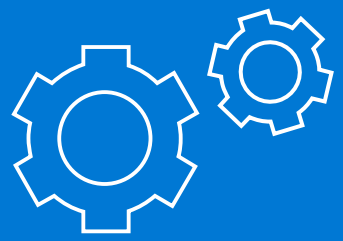
Specialized hardware accelerated deep learning

Project Brainwave

FPGA NEW UPDATES:

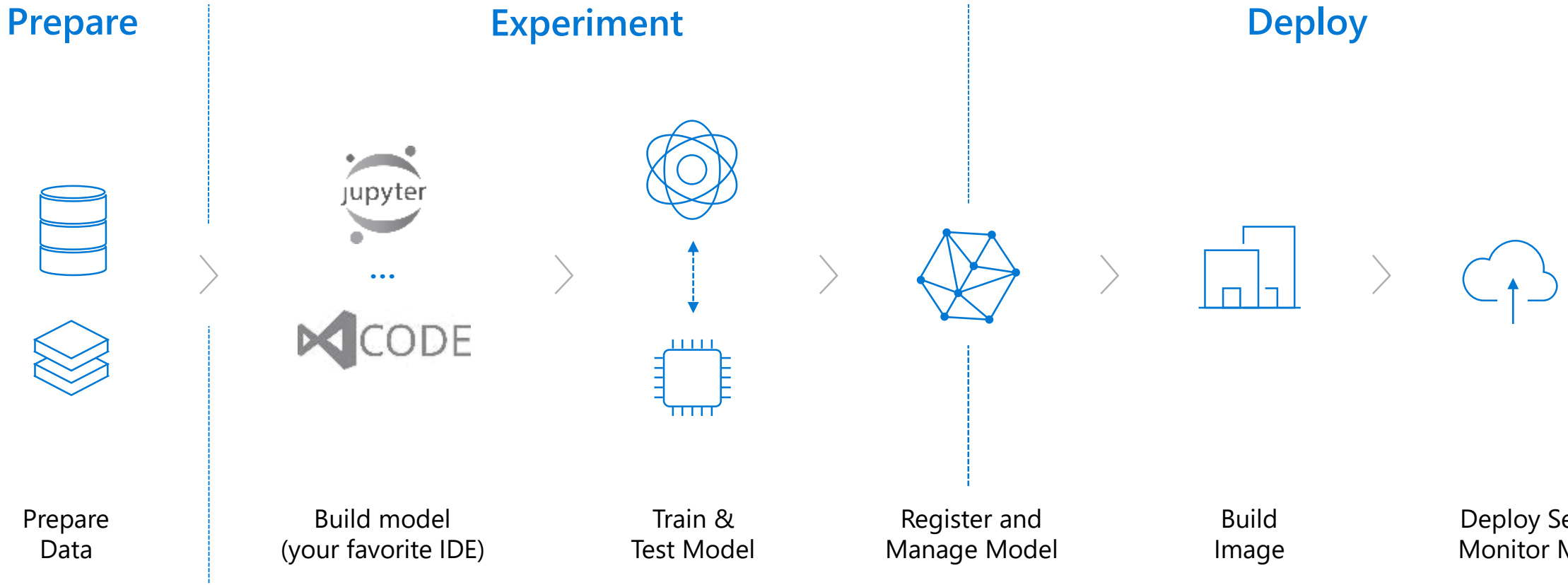
Support for image classification and recognition scenarios

ResNet 50, ResNet 152, VGG-16, SSD-VGG, DenseNet-121

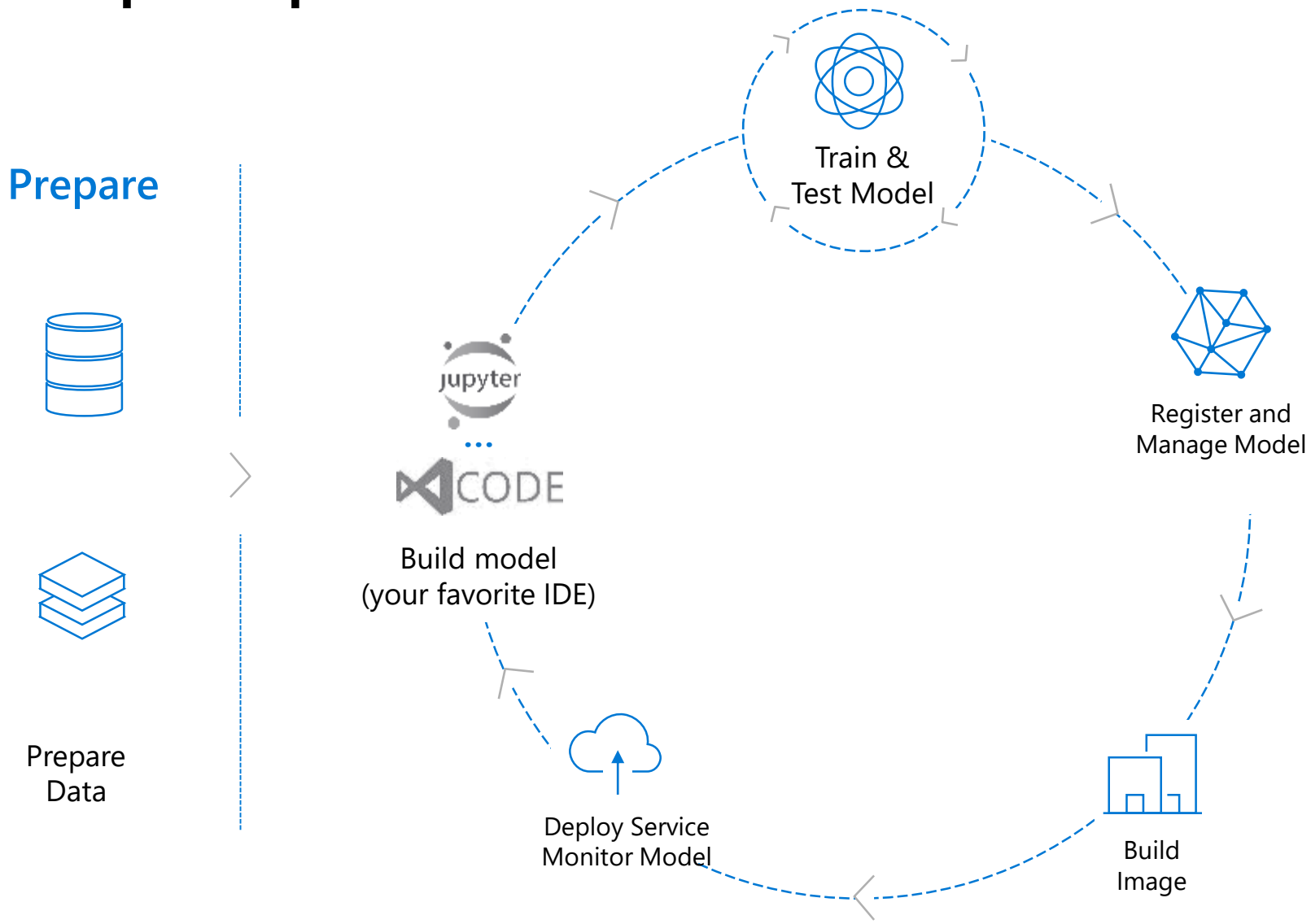


3. DevOps에 대한 고려

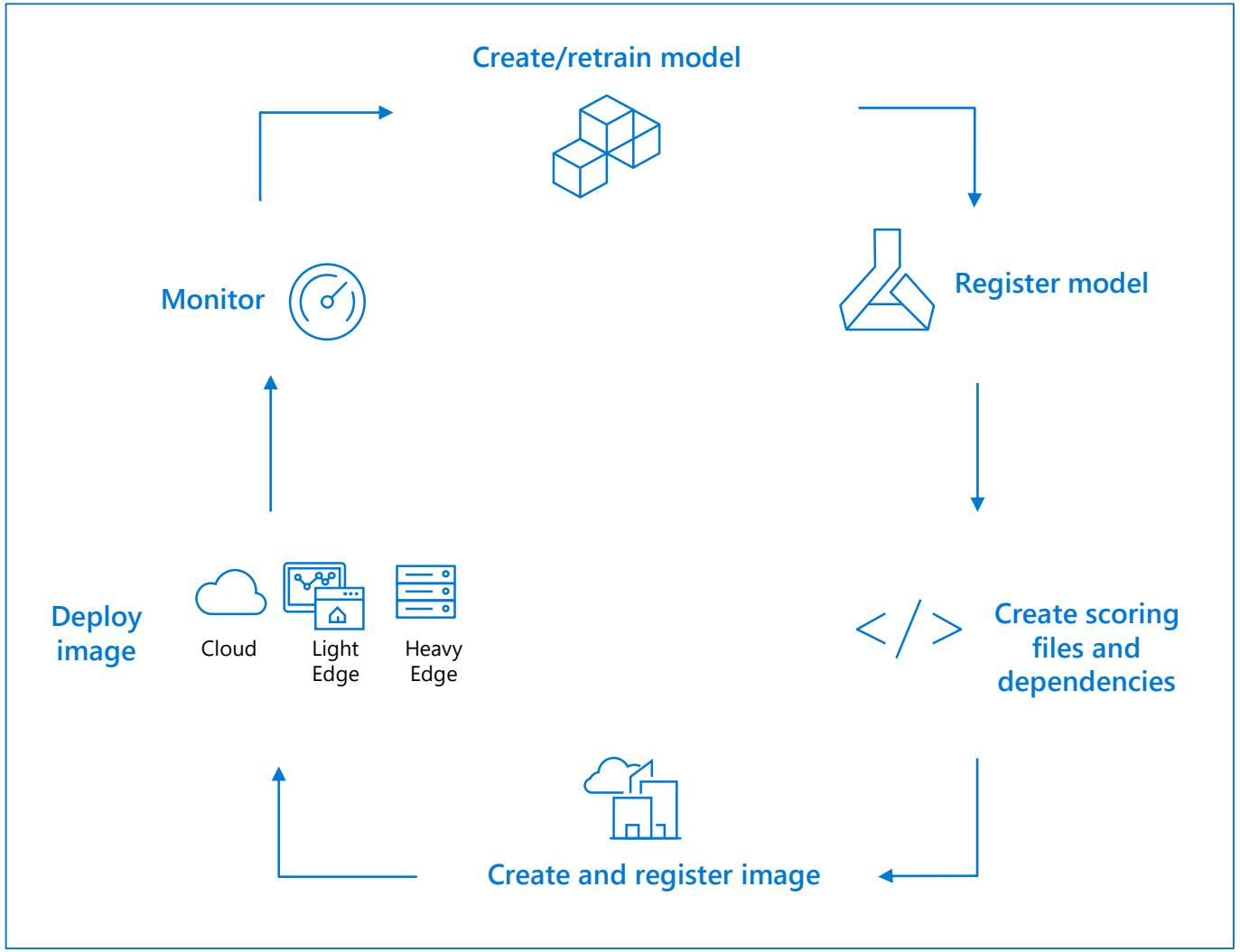
DevOps loop for data science



DevOps loop for data science



Model management in Azure Machine Learning



Model management in detail

Create/Retrain Model

Enable DevOps with full CI/CD integration with VSTS

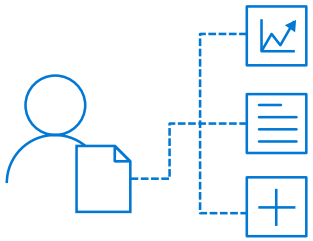
Register Model

Track model versions with a central model registry

Monitor

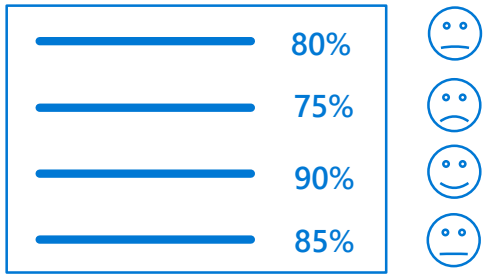
Oversea deployments through Azure AppInsights

Experimentation

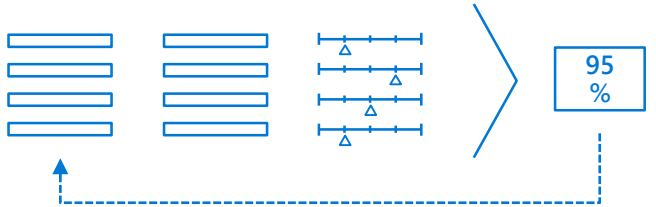


Leverage service-side capture of run metrics, output logs and models

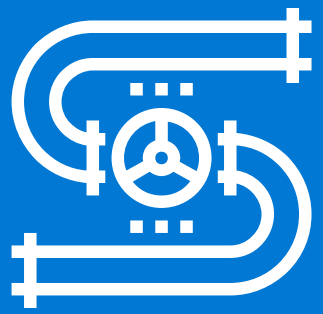
Manage training jobs locally, scaled-up or scaled-out



Use leaderboards, side by side run comparison and model selection

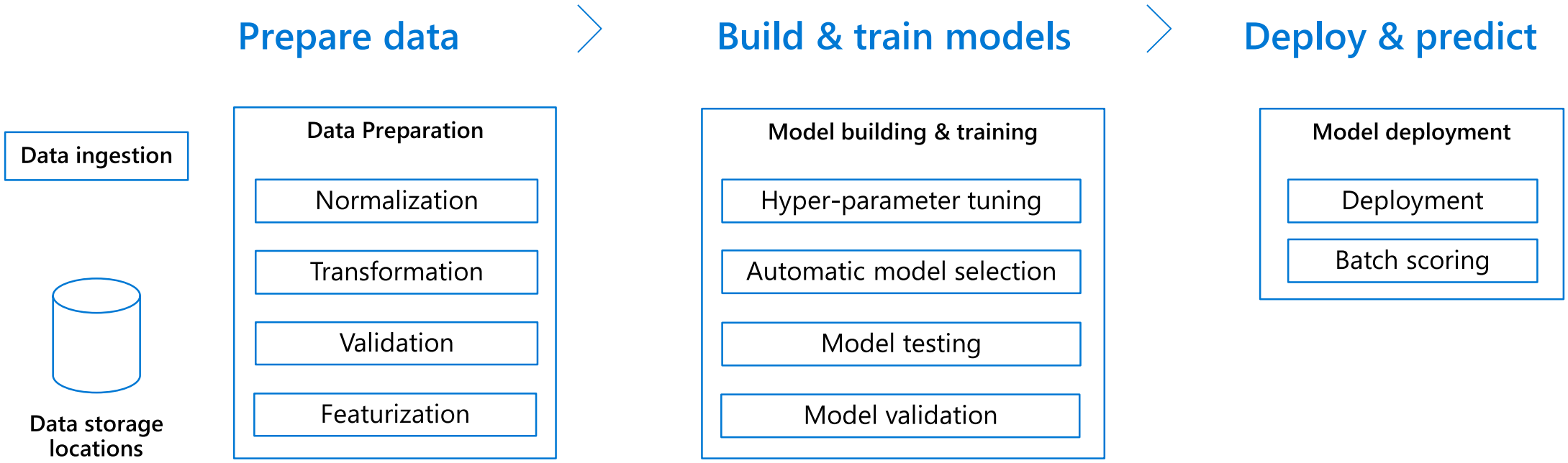


Conduct a hyperparameter search on traditional ML or DNN

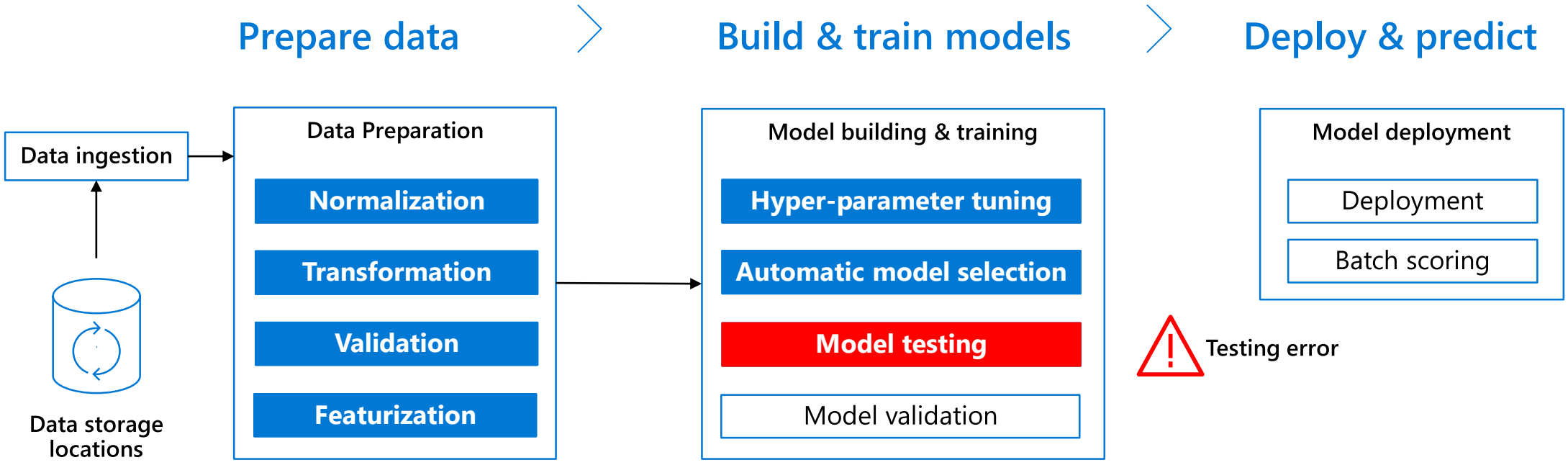


4. 학습 파이프라인에 대한 고려

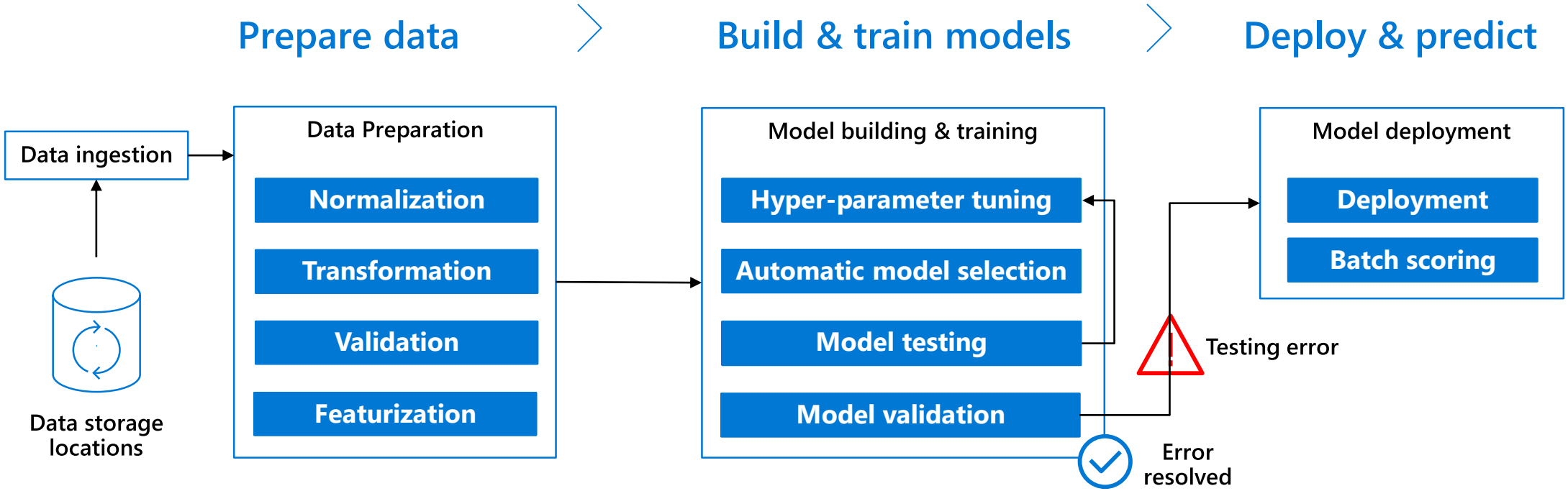
Azure Machine Learning pipelines



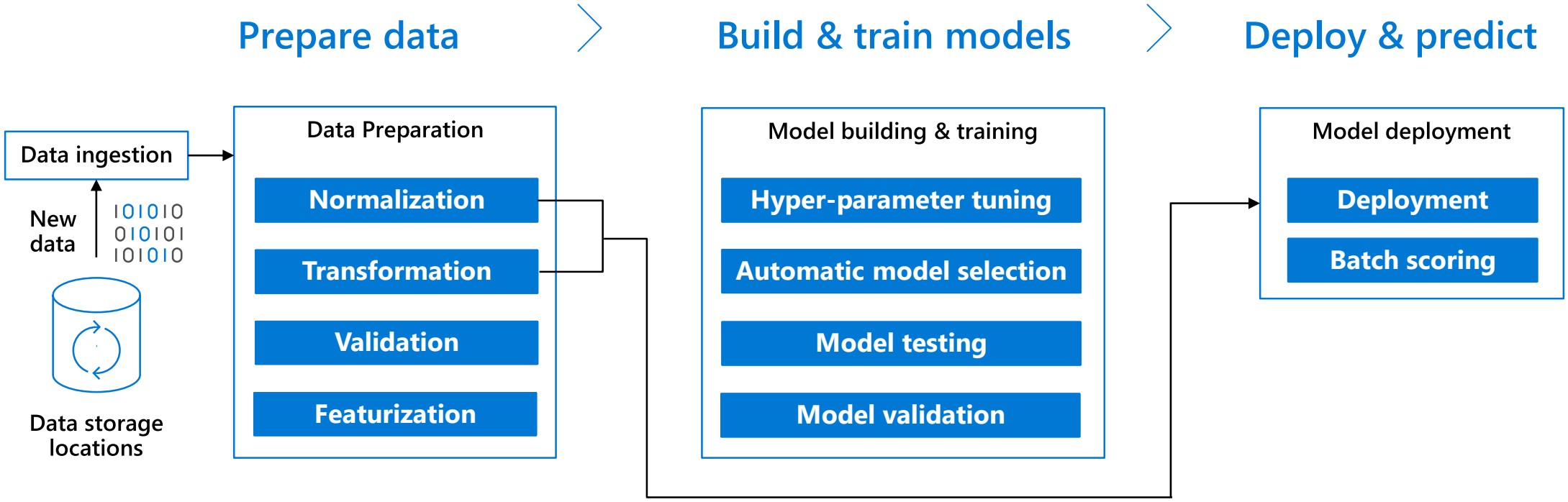
Azure Machine Learning pipelines



Azure Machine Learning pipelines



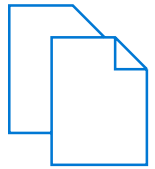
Azure Machine Learning pipelines with new data



Advantages of Azure ML Pipelines

Unattended runs

Schedule a few steps to run in parallel or in sequence to focus on other tasks while your pipeline runs

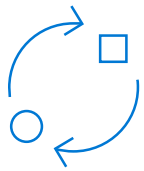


Tracking and versioning

Name and version your data sources, inputs and outputs with the pipelines SDK

Reusability

Create templates of pipelines for specific scenarios such as retraining and batch scoring



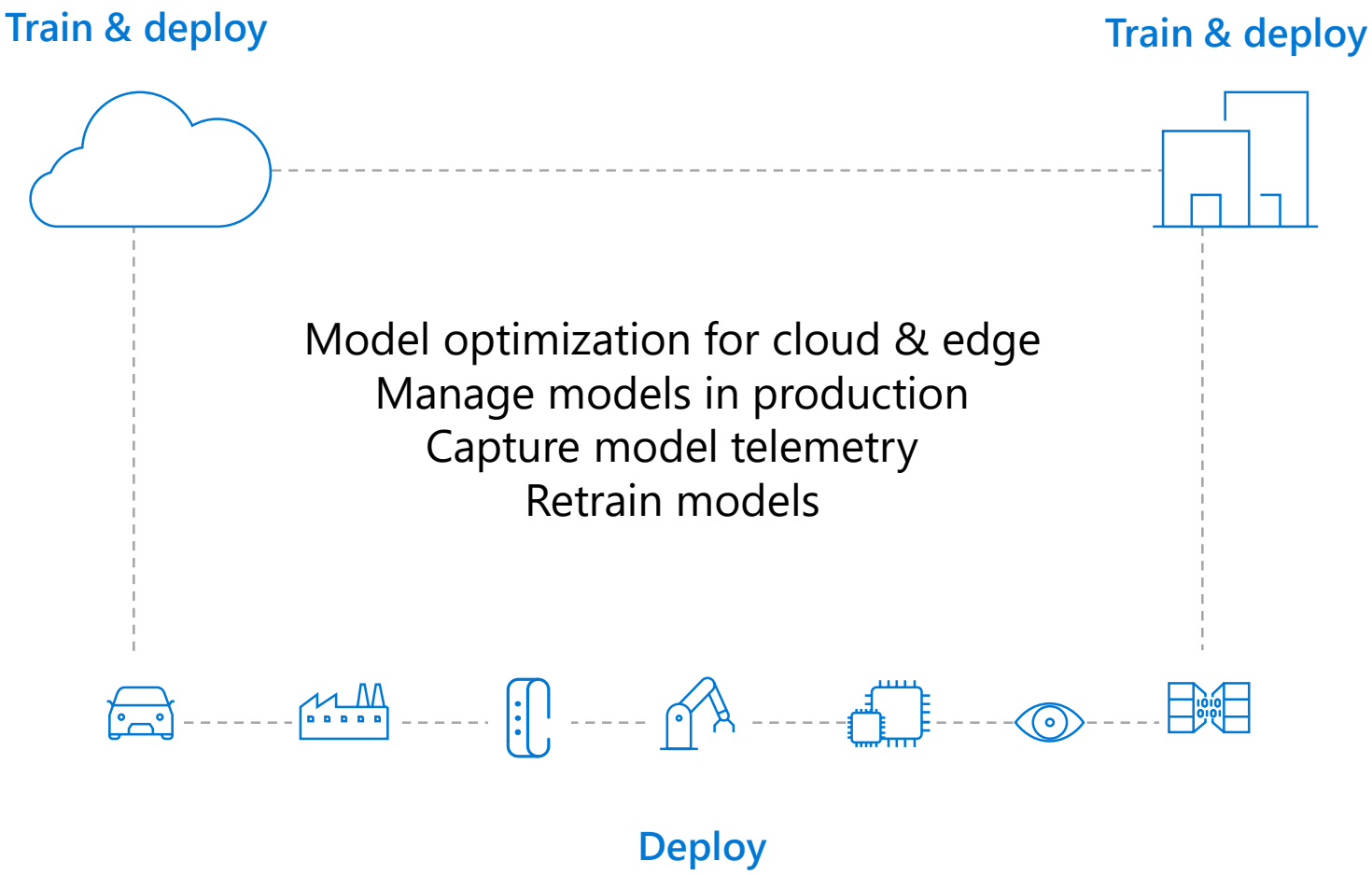
Mixed and diverse compute

Use multiple pipelines that are reliably coordinated across heterogeneous and scalable computes and storages

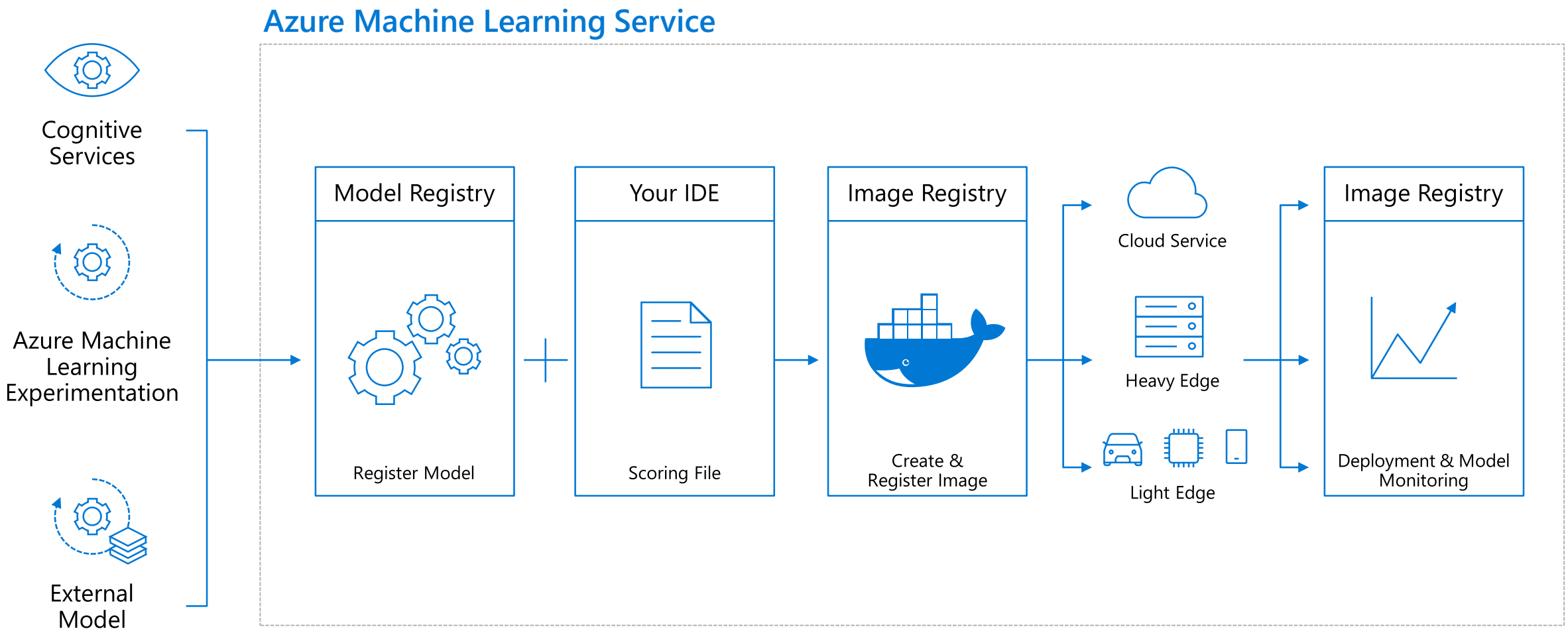
5. 단순한 배포

Flexible deployment

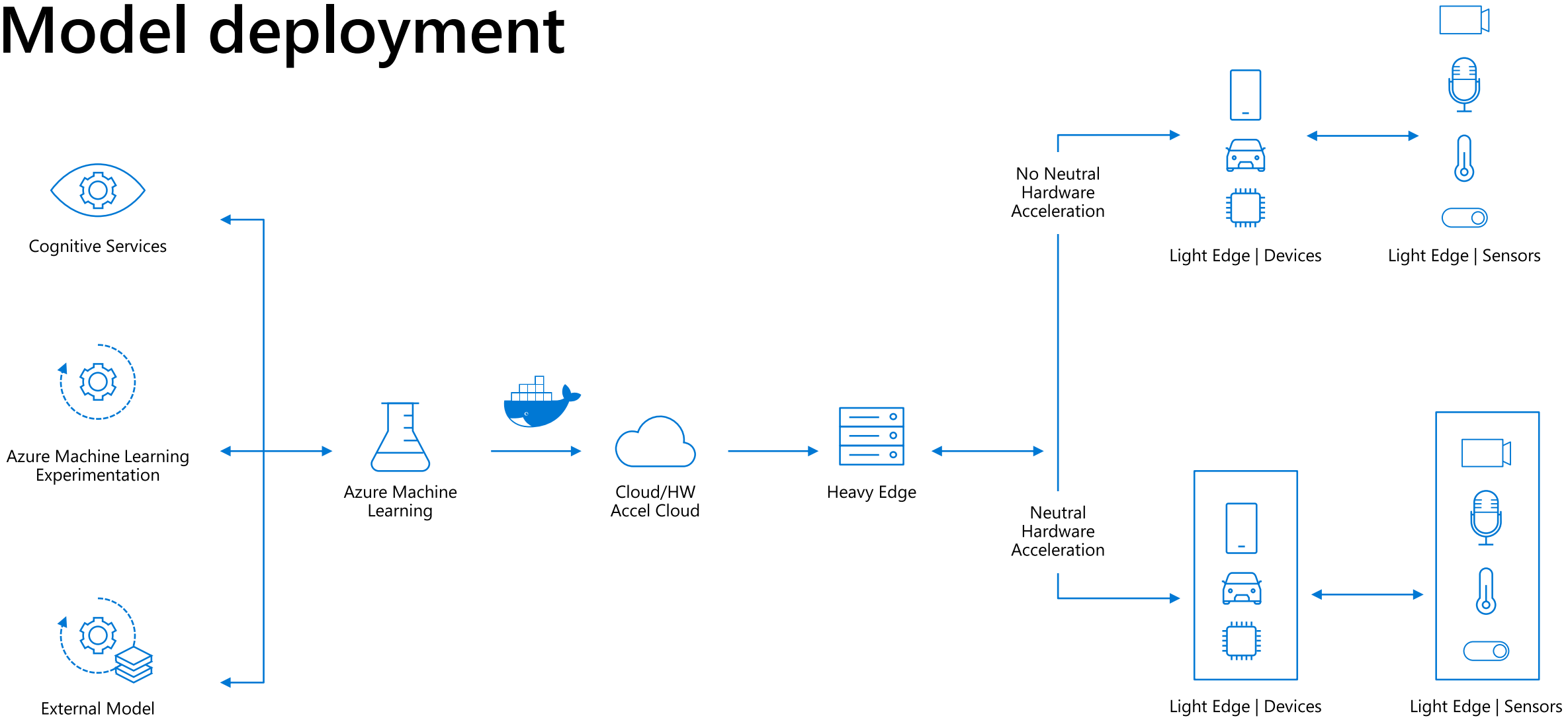
Deploy and manage models on intelligent cloud and edge



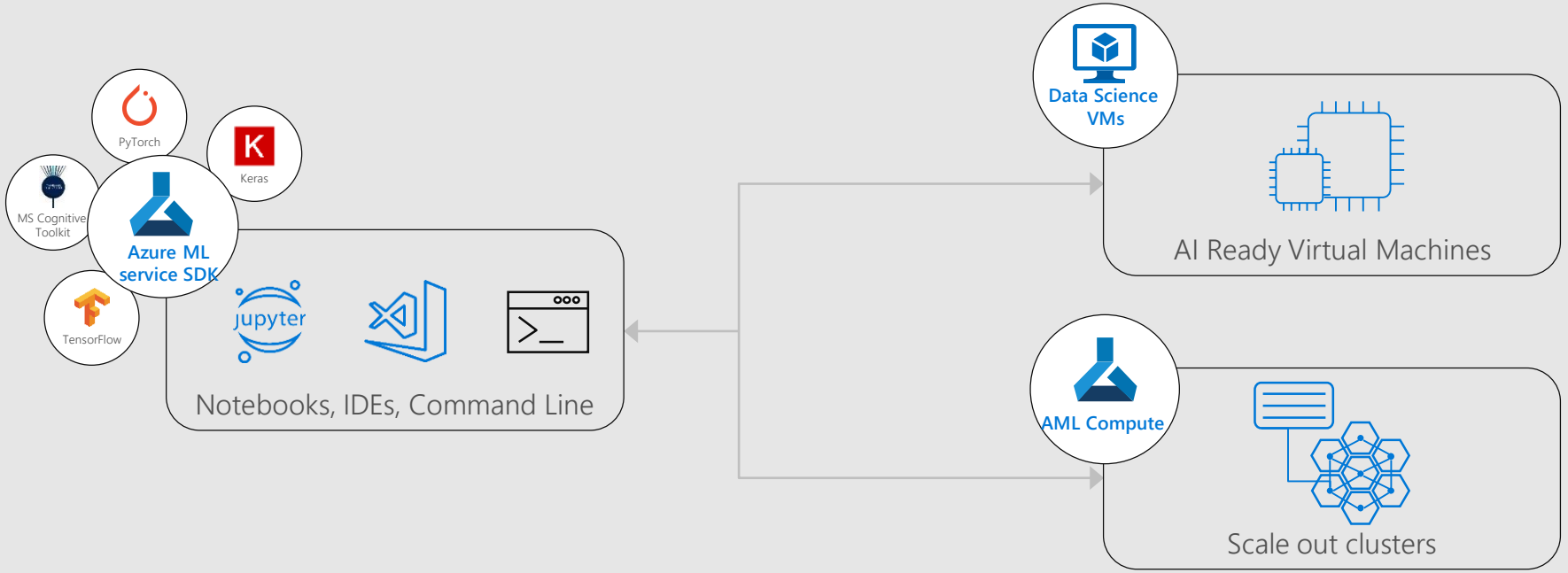
Deploy Azure ML models at scale



Model deployment



Build and deploy deep learning models



Streamline AI development efforts

- Leverage popular deep learning toolkits
- Develop your language of choice

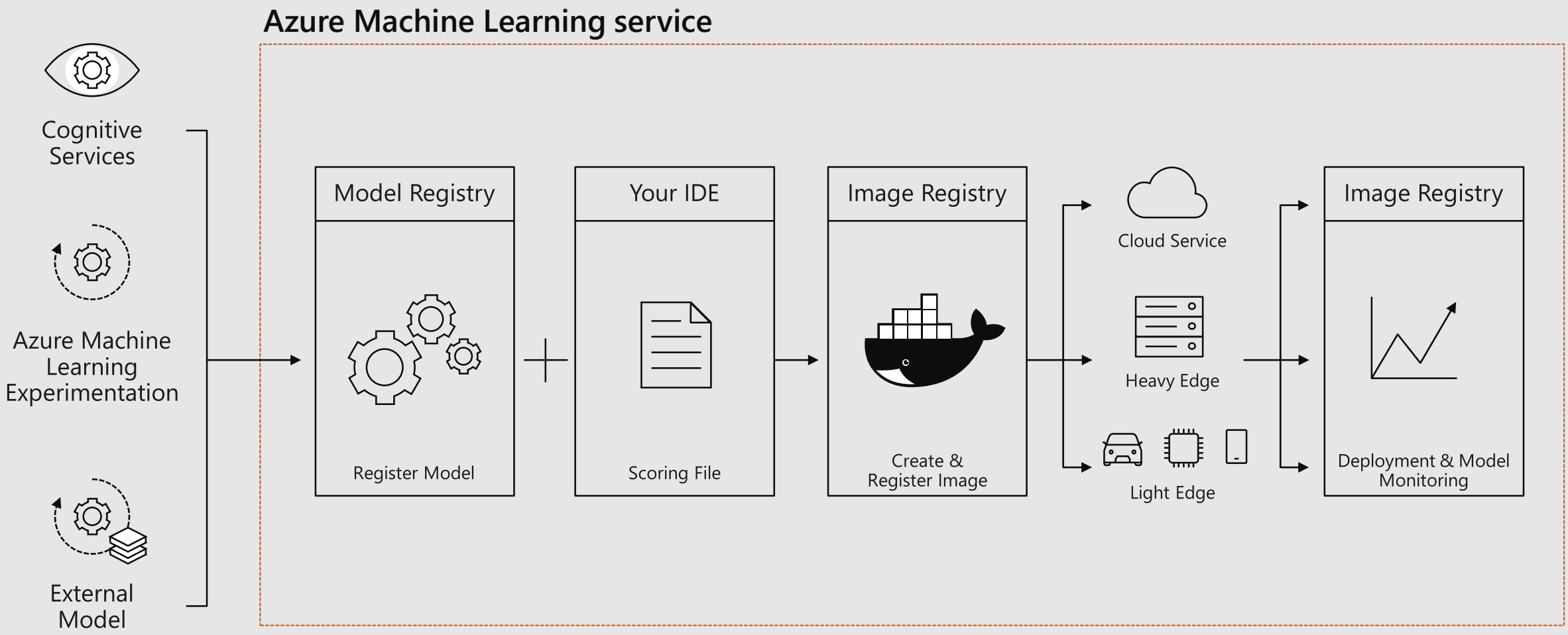
Scale compute resources in any environment

- Choose VMs for your modeling needs
- Process video using GPU-based VMs

Quickly evaluate and identify the right model

- Run experiments in parallel
- Provision resources automatically

Deploy Azure ML models at scale



5. Conclusion

Microsoft AI Platform: Azure + AI

AI SERVICES

CONVERSATIONAL AI

Azure Bot Service

AI MODULES

Cognitive Services

DEVICES

Azure IOT Edge

AI INFRASTRUCTURE

AI ON DATA

Cosmos DB

SQL DB

SQL DW

Data Lake

AI COMPUTE

Spark

DSVM

Batch AI

ACS

Edge

CPU, GPU, FPGA

AI TOOLS

CODING & MANAGEMENT TOOLS

VS Tools for Machine Learning

Azure ML Studio

Azure ML Workbench

Others (Pycharm, Jupyter notebooks…)

DEEP LEARNING FRAMEWORKS

Cognitive Toolkit

TensorFlow

Caffe

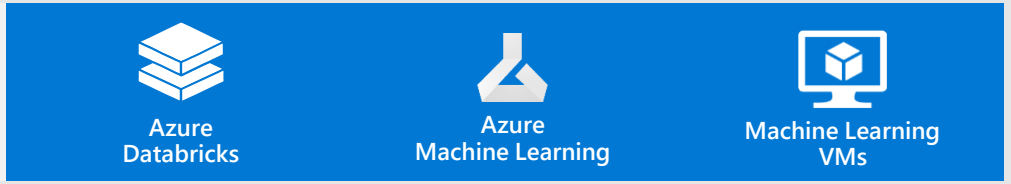
Others (Scikit-learn, MXNet, Keras, Gluon…)

Azure Machine Learning & AI

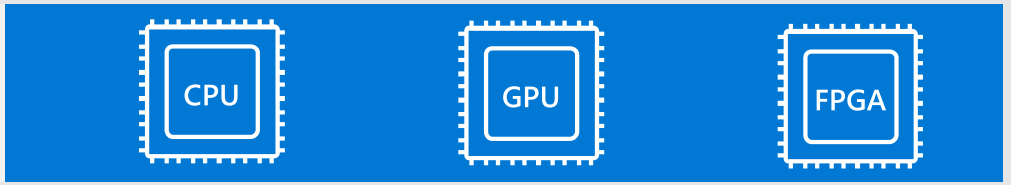
정교하게 미리 학습된 모델
솔루션 개발을 손쉽게 구현하기 위한 방법

유용한 프레임워크 활용
고급 딥 러닝 솔루션을 구축하기 위한 방법

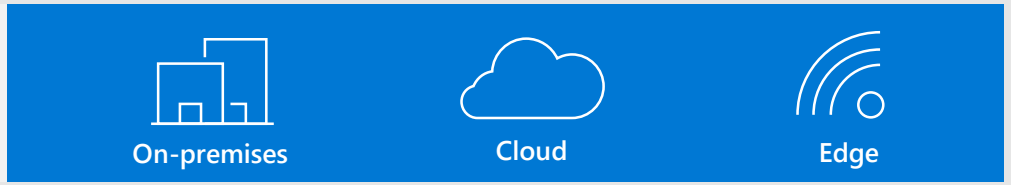
다양한 서비스 활용을 통합 생산성 향상
Data science와 개발팀을 위한 역량 강화



강력한 인프라스트럭처
효율적이고 원활한 딥 러닝 환경제공



유연하고 다양한 배포 모델
인텔리전트 클라우드와 Edge환경 모두 적용가능한 배포 및 관리 지원



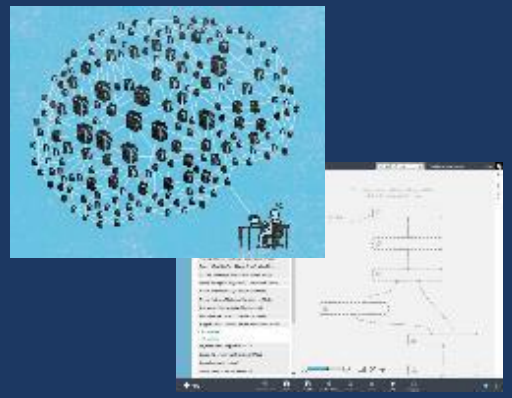
참고: Microsoft Azure에서 GPU 기반 서비스

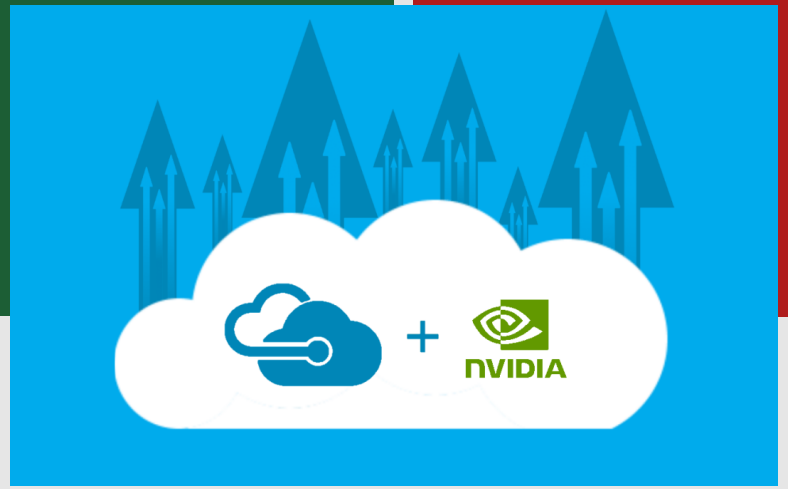
Compute

Machine Learning

Remote Virtualization

Compute AI





요약

- GPU ❤ 클라우드: IaaS를 넘어 PaaS, 그 이상으로 변화 중
- 클라우드에서 Deep Learning을 활용할 때의 주요 전략 및 방법
  - 반복/순차적인 연산 → 병렬로 빠르게 결과 확인 & 적용
  - Scalability를 고려한 Production 레벨에서의 서비스 준비에 용이
  - DevOps & MLOps
  - VM 기반 vs. PaaS 기반
  - 컨테이너 활용 여부
  - 최적화 & Production 준비 & 비용
  - …

Microsoft Azure  NVIDIA

Thank you!