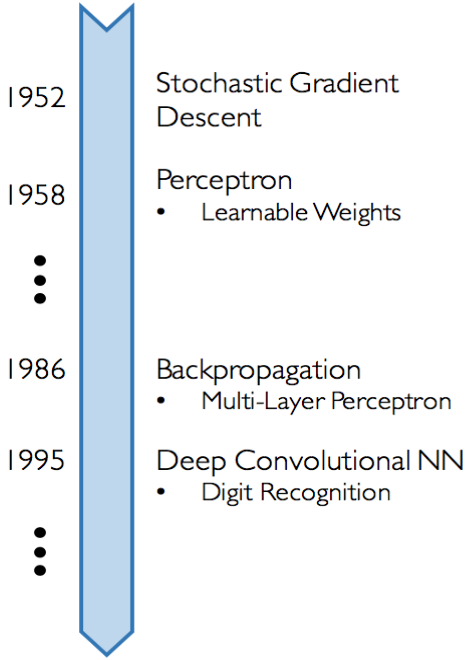
Managing Massive Input Data in Flash for AI and Deep Learning Applications

Dejan Kocic

Netapp

The World Is Changing Fundamentally

Why Now?



1952 Stochastic Gradient Descent

1958 Perceptron

- Learnable Weights

⋮

1986 Backpropagation

- Multi-Layer Perceptron

1995 Deep Convolutional NN

- Digit Recognition

⋮

Neural Networks date back decades, so why the resurgence?

1. Big Data

- Larger Datasets
- Easier Collection & Storage

IMAGENET

WIKIPEDIA
The Free Encyclopedia

2. Hardware

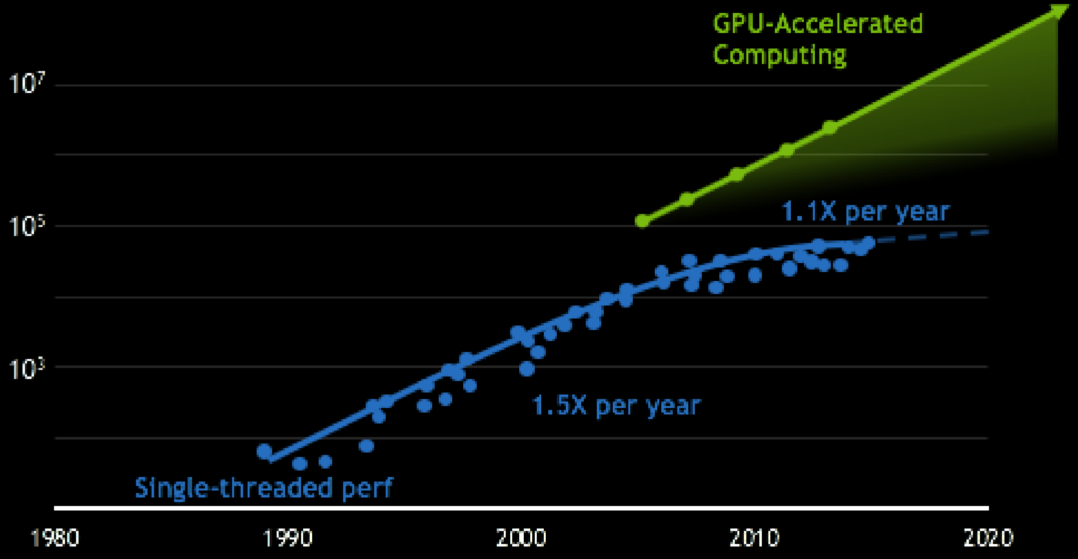
- Graphics Processing Units (GPUs)
- Massively Parallelizable

3. Software

- Improved Techniques
- New Models
- Toolboxes

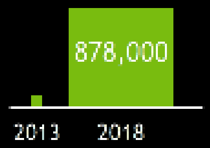
TensorFlow

RISE OF GPU COMPUTING

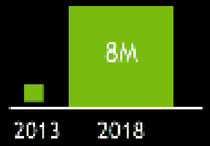


40 Years of CPU Trend Data

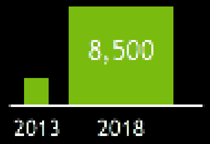
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp



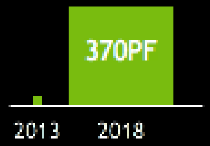
GPU Developers
10X in 5 Yrs



CUDA Downloads
5X in 5 Yrs



GTC Registrations
4X in 5 Yrs



Total GPU FLOPS
of Top 50 Systems
15X in 5 Yrs

What's happening in the world of AI?

AI is becoming a disruptive force impacting nearly every industry

40%

of digital transformation initiatives will use AI services in 2019

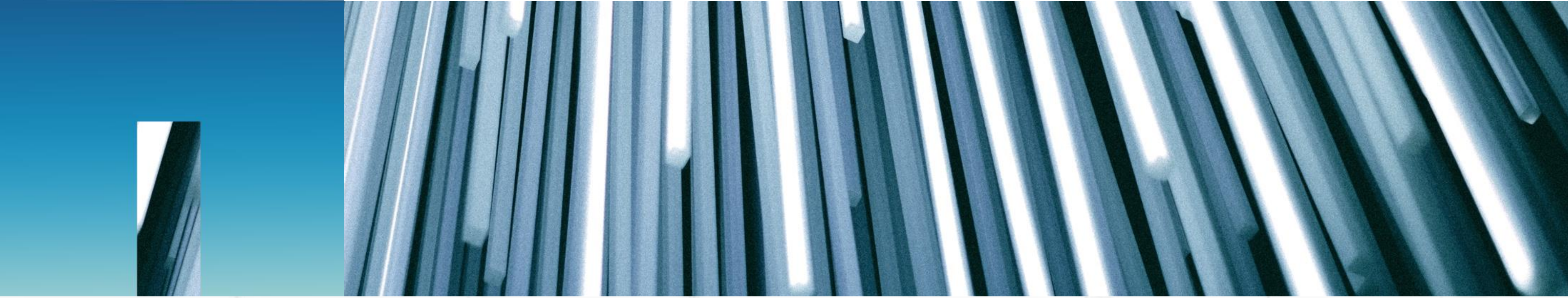
50%

of enterprise infrastructure will employ artificial intelligence by 2021

87%

of global business leaders expect AI to bring better customer experiences within 3 years

Source: IDC Storage Workloads, 2018. AI Business, 2018.



AI is all about data

Data is distributed

Generated and consumed from multiple clouds and on-premises

Data is dynamic

Constantly changing and increasingly cloud-streamed

Data is diverse

Data comes in many forms: video, audio, images, quantitative, logs etc.

Architectural models come and go but data is eternal

Data is critical to AI, but presents significant challenges

(Source IDG research)

51%

Data
silos

37%

Technology
complexity

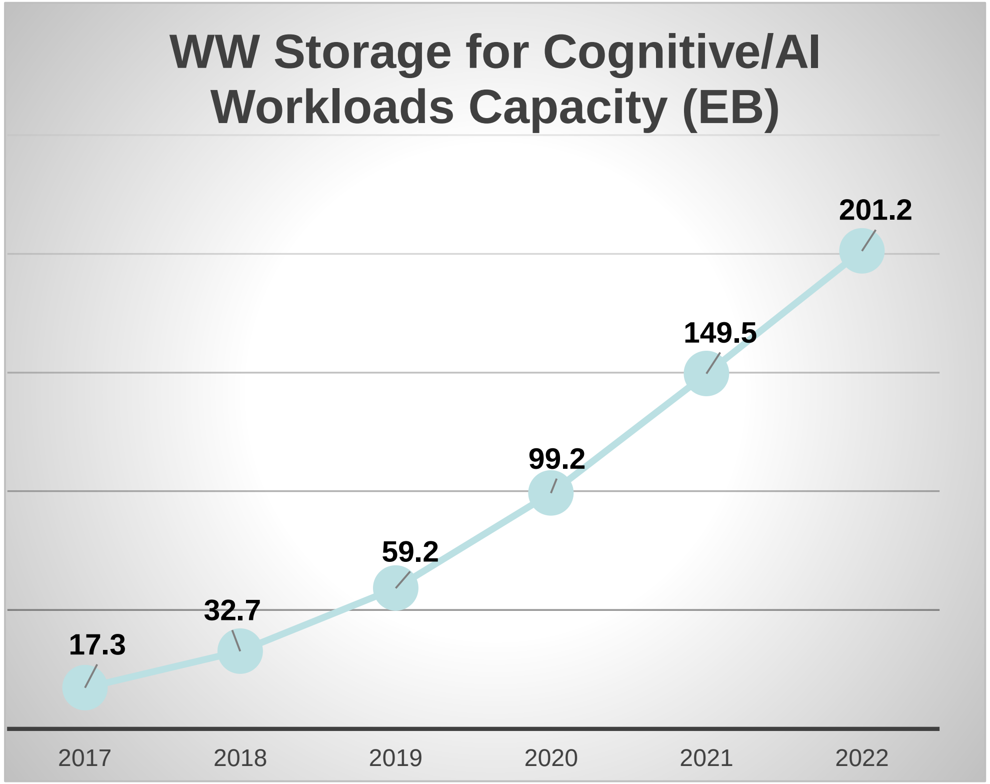
35%

Data
access

35%

Data
preparation

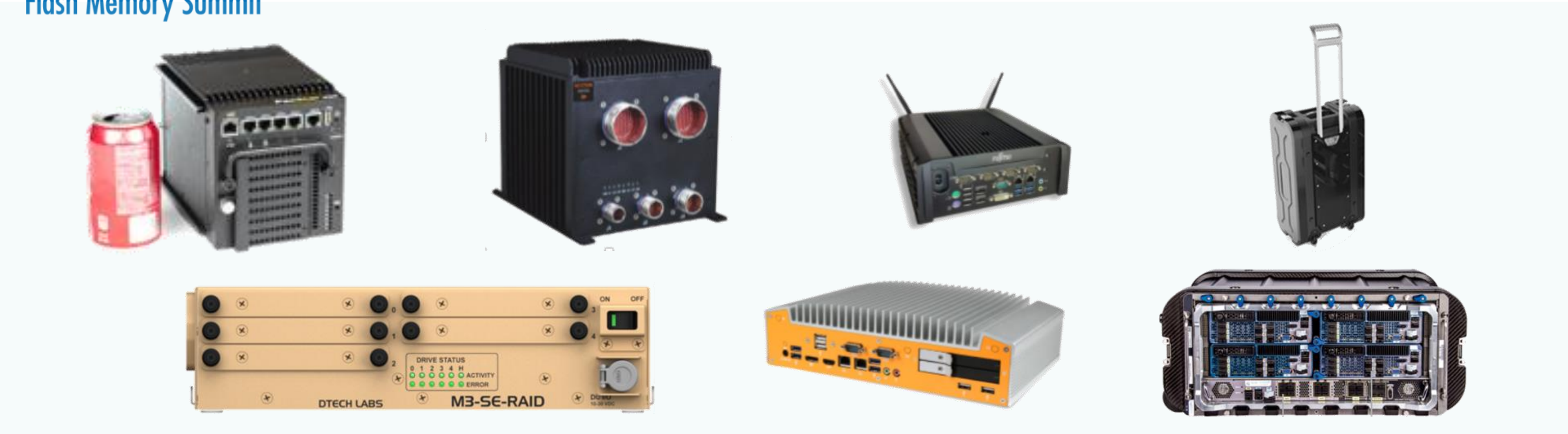
AI Capacity Growth Worldwide



Source: IDC WW Storage for Cognitive/AI Workloads Forecast, 2017-2022

Edge, data comes from various places

Data Collectors Come in Many Form Factors



PacStar®

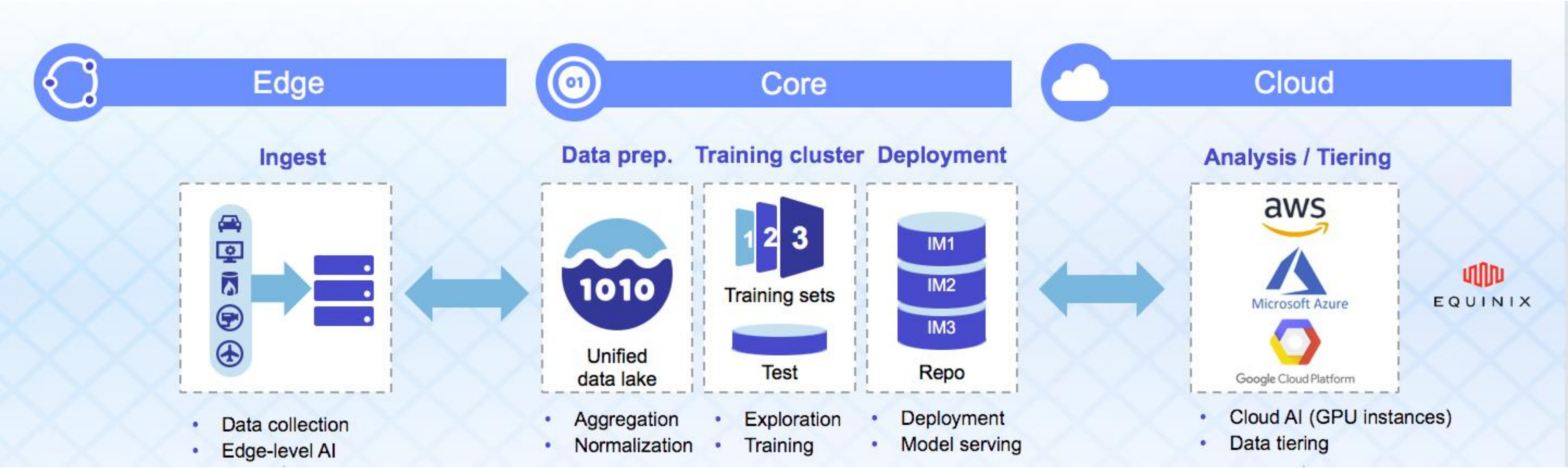
VECTOR DATA™

CUBIC™ | DTECH LABS

GERMANE SYSTEMS™

KLAS TELECOM

Edge to Core to Cloud



Edge

Ingest

- Data collection
- Edge-level AI

Core

Data prep.

- Aggregation
- Normalization

Training cluster

- Exploration
- Training

Deployment

- Deployment
- Model serving

Cloud

Analysis / Tiering

- Cloud AI (GPU instances)
- Data tiering

Data Has Transformative Value

ENABLE
new customer touchpoints

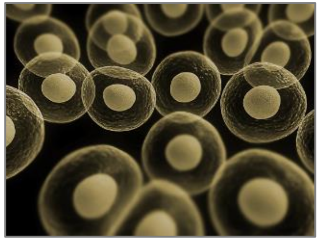
CREATE
innovative business opportunities

OPTIMIZE
operations

AI is Uniquely Enabling a Range of Use Cases

Social, Media,
Internet and Cloud

Cyber
Security



Life
Sciences

Defense
Intelligence

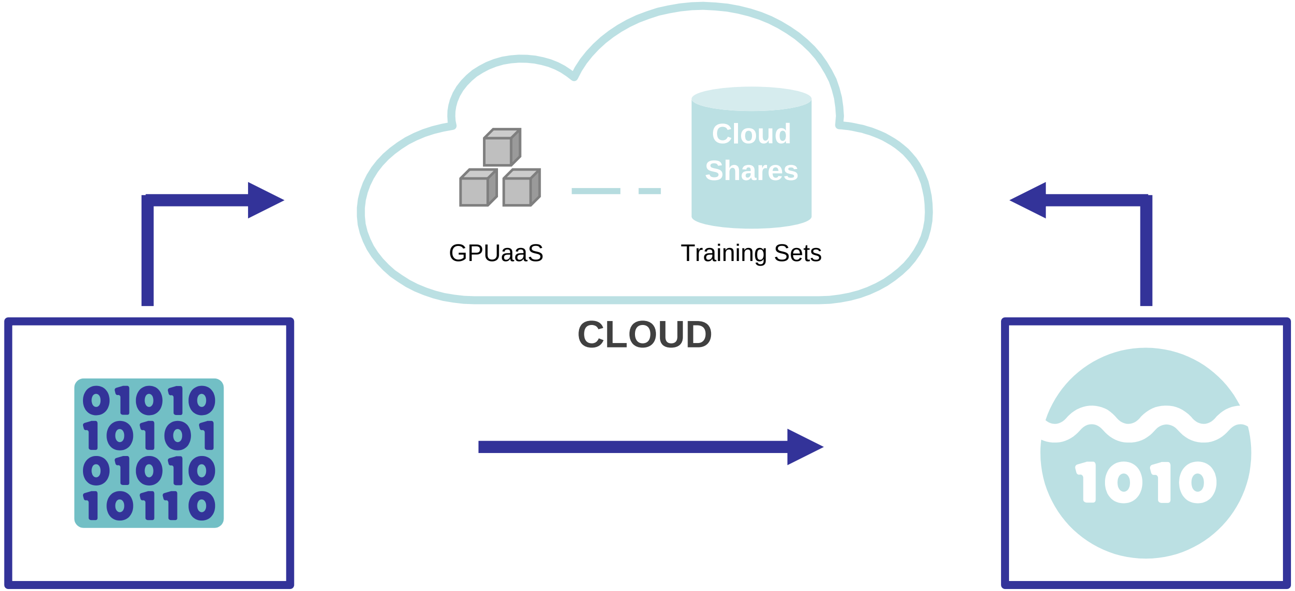
Internet of
Things

Financial
Markets

Autonomous
Machines/Vehicles

Deployment Choices

Using cloud-based GPUs



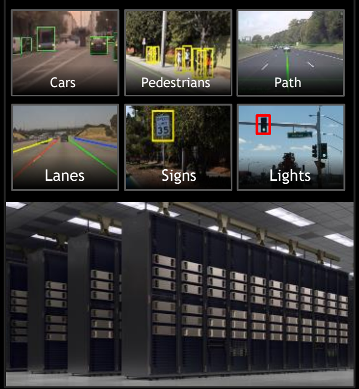
EDGE

CORE

END-TO-END SYSTEM FOR AV

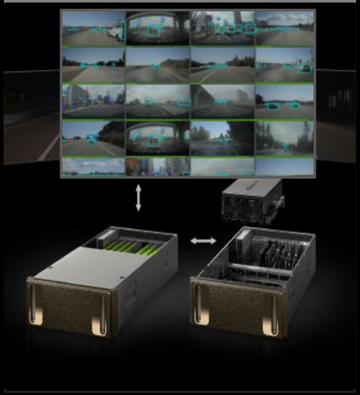
COLLECT DATA

TRAIN MODELS

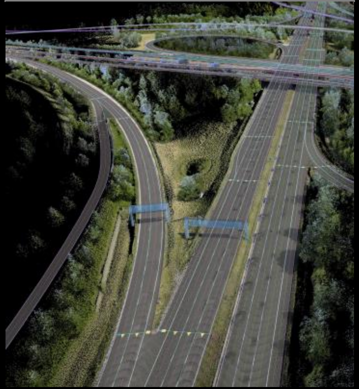


SIMULATE

RE-SIMULATE

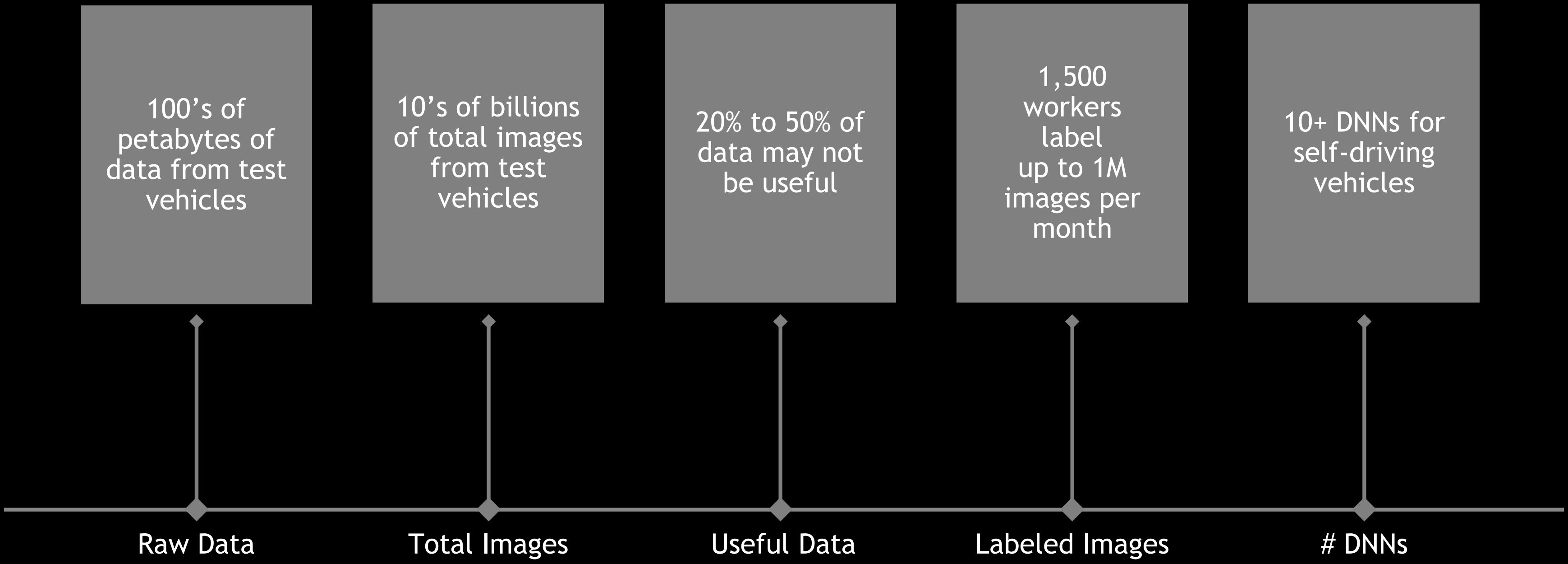


MAPPING



Sources: NVIDIA, RAND Corporation

DATA COLLECTION AND LABELING FOR AI



100’s of petabytes of data from test vehicles

10’s of billions of total images from test vehicles

20% to 50% of data may not be useful

1,500 workers label up to 1M images per month

10+ DNNs for self-driving vehicles

Raw Data

Total Images

Useful Data

Labeled Images

DNNs

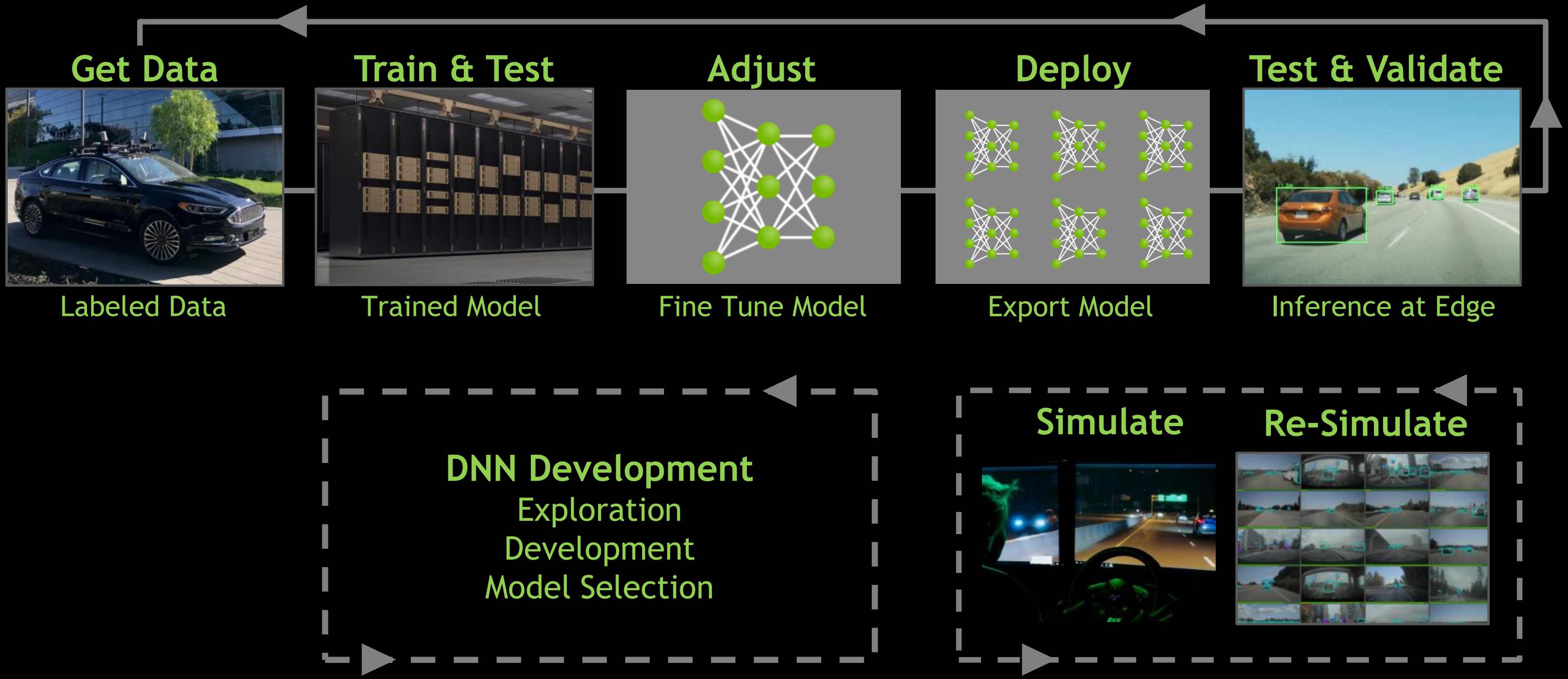
Source: Data from test fleets of 50-100 cars

DATA GENERATION FROM ONE SURVEY CAR

DATA COLLECTED	TOTAL IMAGES	LABELED IMAGES
2 petabytes per car / year	1 billion images / year	3 million images / year

Source: NVIDIA's Self-Driving Cars - BB8

AI FOR SELF-DRIVING WORKFLOW



Get Data

Labeled Data

Train & Test

Trained Model

Adjust

Fine Tune Model

Deploy

Export Model

Test & Validate

Inference at Edge

DNN Development

Exploration

Development

Model Selection

Simulate

Re-Simulate

How Flash Storage compares to other media

Technology	**DRAM**	**3D Xpoint**	**Flash**	**HDD**
Access time	10 ns	7 µs	150-200 µs	6-12 ms
Scale	Baseline in ns	138 times slower than DRAM	20-30 times slower than 3DXP, over 2940 times slower than DRAM	Over 40 times slower than Flash, over 850 times slower than 3DXP, over 120K times slower than DRAM
Bandwidth (seq R/W)	13GB/s / 13GB/s	2.6GB/s / 2GB/s (M.2/NVMe, DIMM FF expected 6GB/s)	3GB/s / 2.6GB/s (M.2/NVMe)	112MB/s / 45MB/s

Where and how flash can help

- Ingesting data from the edge:
  - often lots of small files
  - lots of writes
- Lots of small files create random workload
- In some cases data from edge can be aggregated to reduce the number of IOPS
- Data ingest from Edge to Core can use flash as a landing space at Core to be able to accept huge amounts of data coming from Edge into Data Lakes at Core which then can be tiered to cheaper storage at Core

Where and how flash can help

- When using data from Data Lakes for training, throughput is important and flash can be used to help ingest data faster and make training process run faster

- Speed and low latency are critical for inference, especially when used for real-time, mission critical applications like autonomous vehicles, voice/video recognition, security… and this is the area where flash can also help accelerate data transfer from data collectors to AI inference systems

Questions?