Unlocking the Value of Unstructured Data with Machine Learning

Miroslav Klivansky
Principal Data Architect — AI & Analytics
Pure Storage

What's Happening in the World

Flash Memory Summit

Data is Chaotic

Advanced workloads are growing

Public cloud is not always the answer

Legacy platforms are failing to innovate

Organizations Struggle to Adapt

The future is unpredictable

Inflexible platforms limit value from data

Storage management is too complex

Total unstructured data volumes will almost **quadruple** by 2026.

(GARTNER)

Storage rebuys and refreshes can end up costing organizations **60%** of the original cost of the storage platform.

(GARTNER)

54% of organizations would prefer to apply a consumption-based cloud model in data centers.

(EGG)

35 YEARS OF NAND FLASH MEMORY INNOVATION

Legacy Storage Hinders Digital Transformation

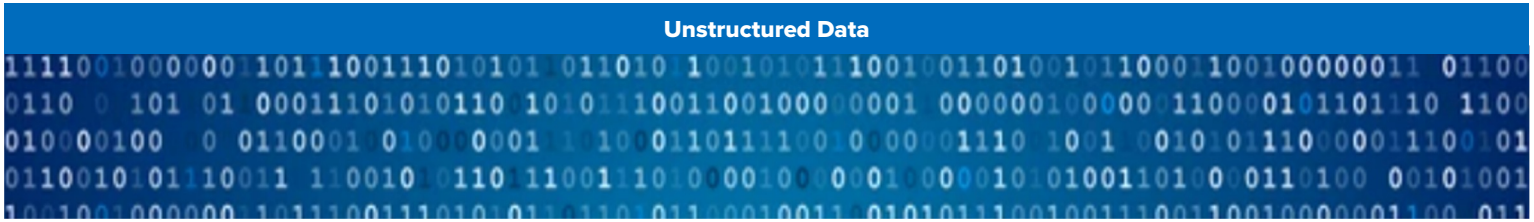
CUSTOMERS ARE HAVING TO SPEND TIME FOCUSING ON THE INS AND OUTS OF DATA STORAGE, RATHER THAN THEIR CORE BUSINESS.

Complexity causes issues with availability, risk, and resources management.

Rigidity is at odds with consumption models, lacking agility and flexibility

Uninspiring solutions aren't innovative enough to support advanced data activities like analytics, AI and ML

Machine Learning Techniques



Supervised Learning

Initial Sample Data Provided

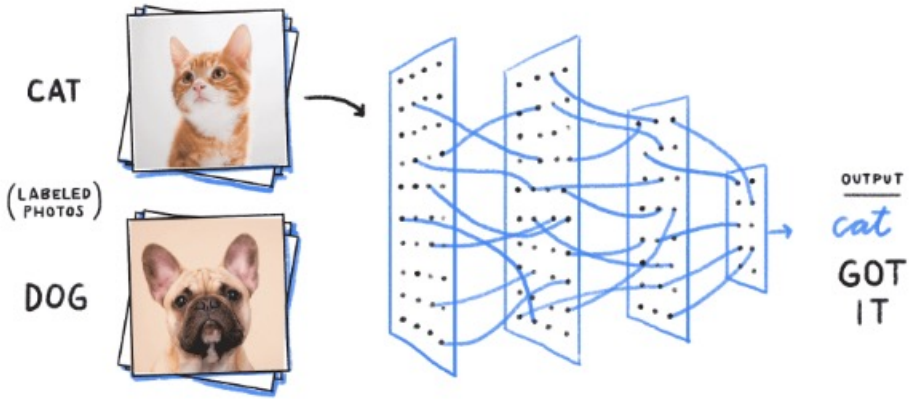
UnSupervised Learning

No human intervention

Reinforcement Learning

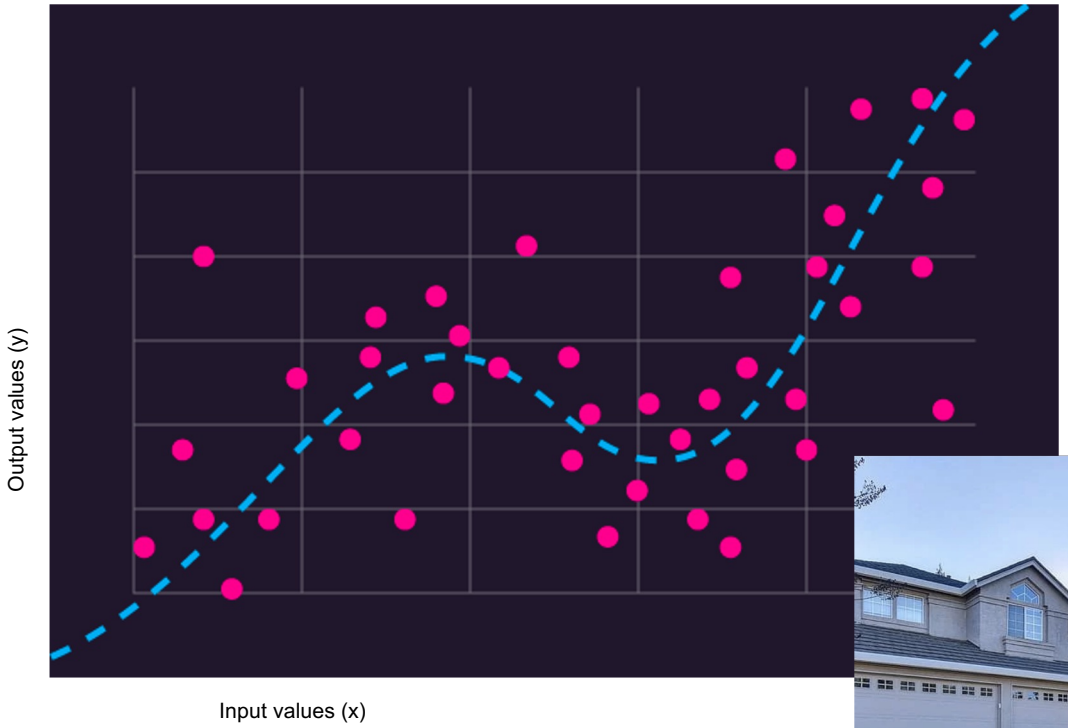
Feedback Driven

Classification (Supervised) - Sorting



Sorting INPUT DATA into one of many classes

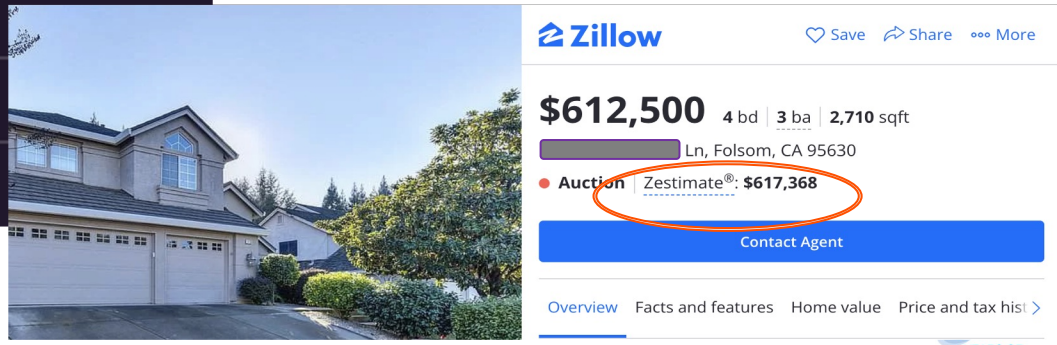
Regression (Supervised Learning) - Predictions



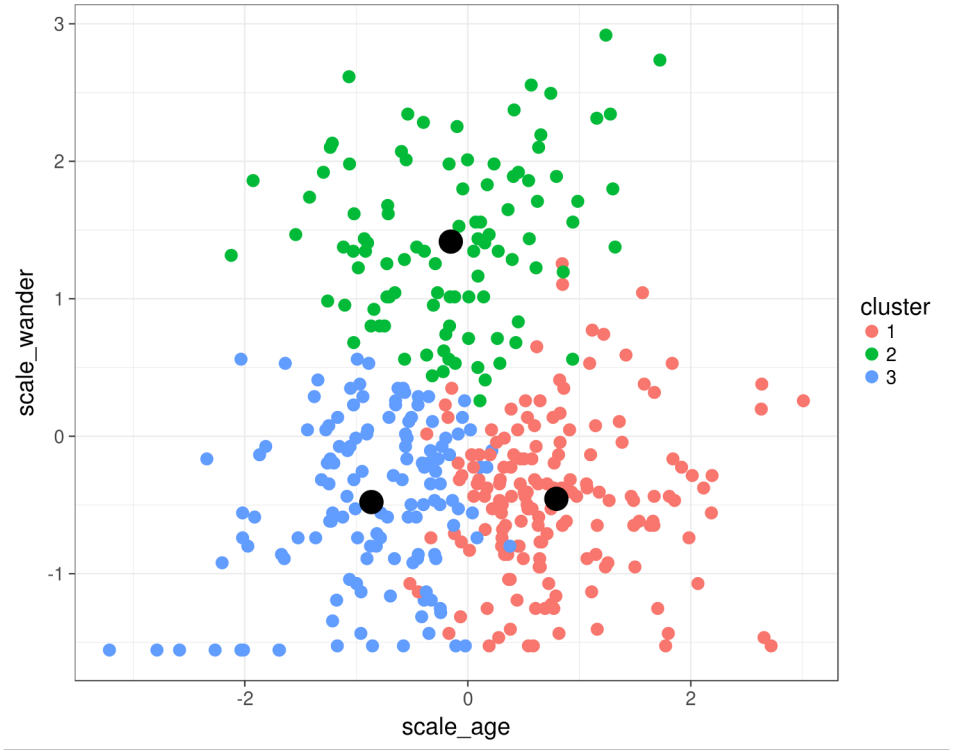
Estimating Function f(x)

Using a few samples, the machine models (deduces) the relationship between INPUT and OUTPUT

Helps predict future values

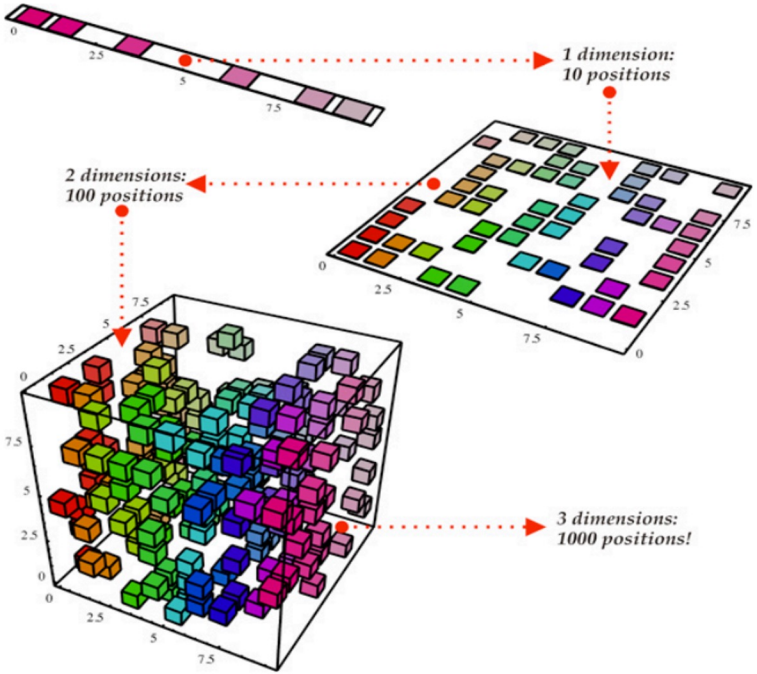


Clustering (UnSupervised) - Segmentation



Without prior knowledge, dividing data into “similar” groups

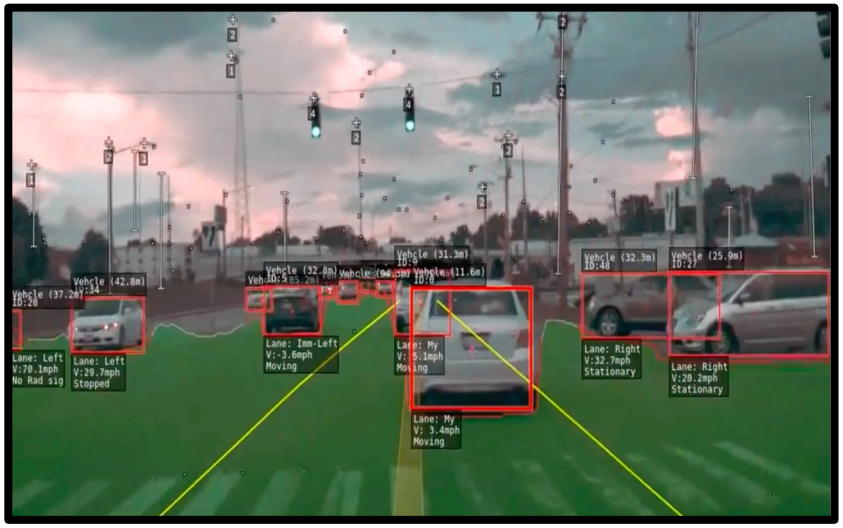
Dimensionality Reduction (UnSupervised)



Compress: Reduce Complexity without losing too much information from the data

Uses a compressed face recognition model that optimizes computation and saves battery!

Reinforcement Learning: "Trial and Error" with feedback

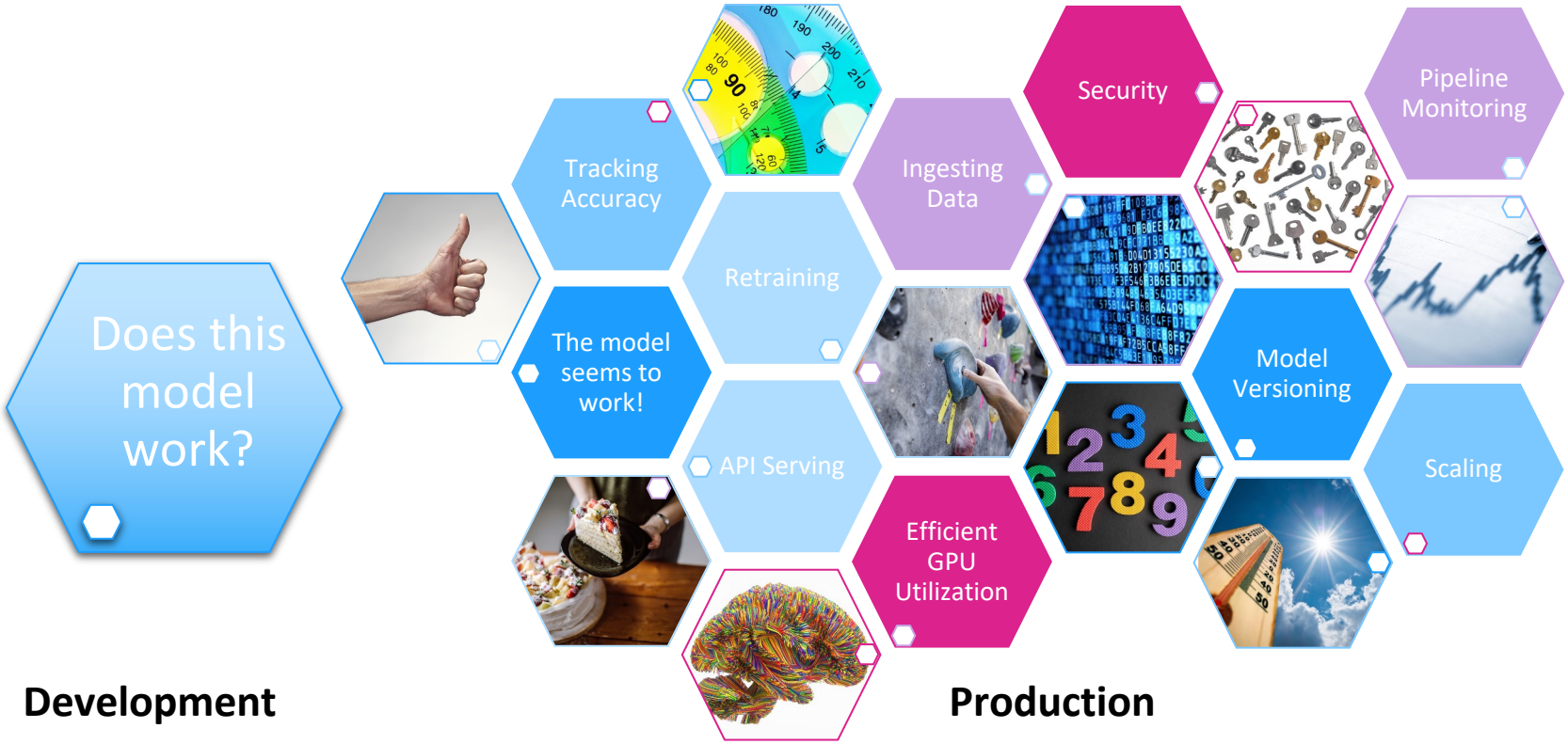


AlphaGo used reinforcement learning to learn the game (GO) better and beat world champion (by playing over 50 million games rapidly and learning how to win)

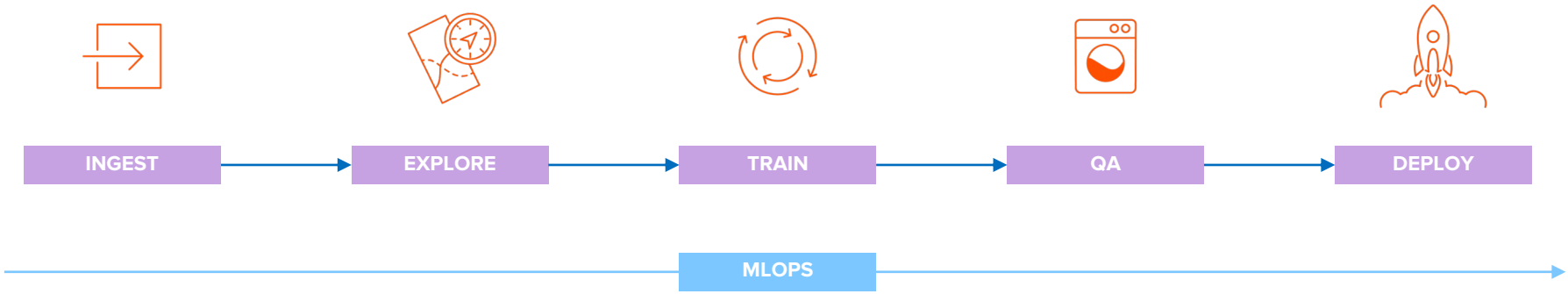
Tesla Self-Driving Models have same experience as a driver who has driven a billion miles!

Machine Learning Operations (MLOps)

Enterprise AI and ML Operations

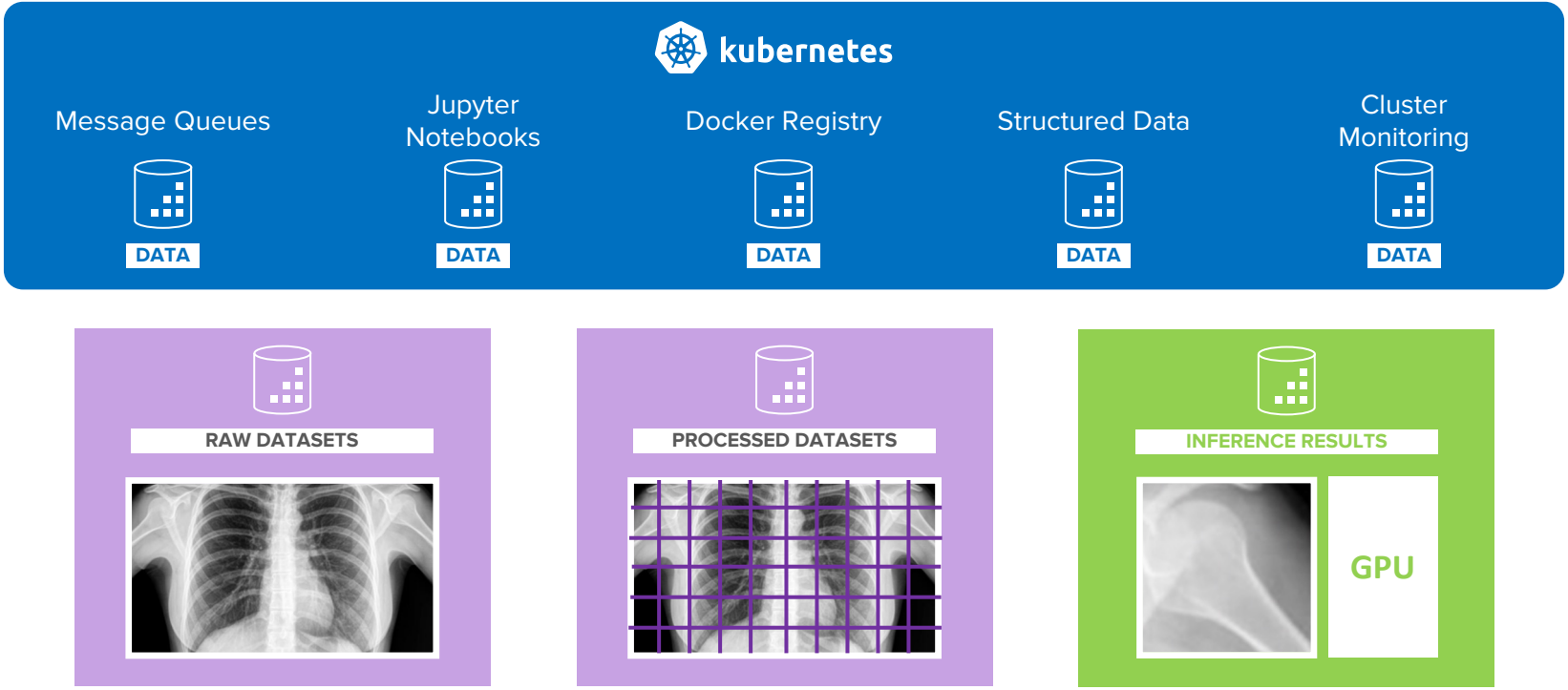


Creating Business Value with ML

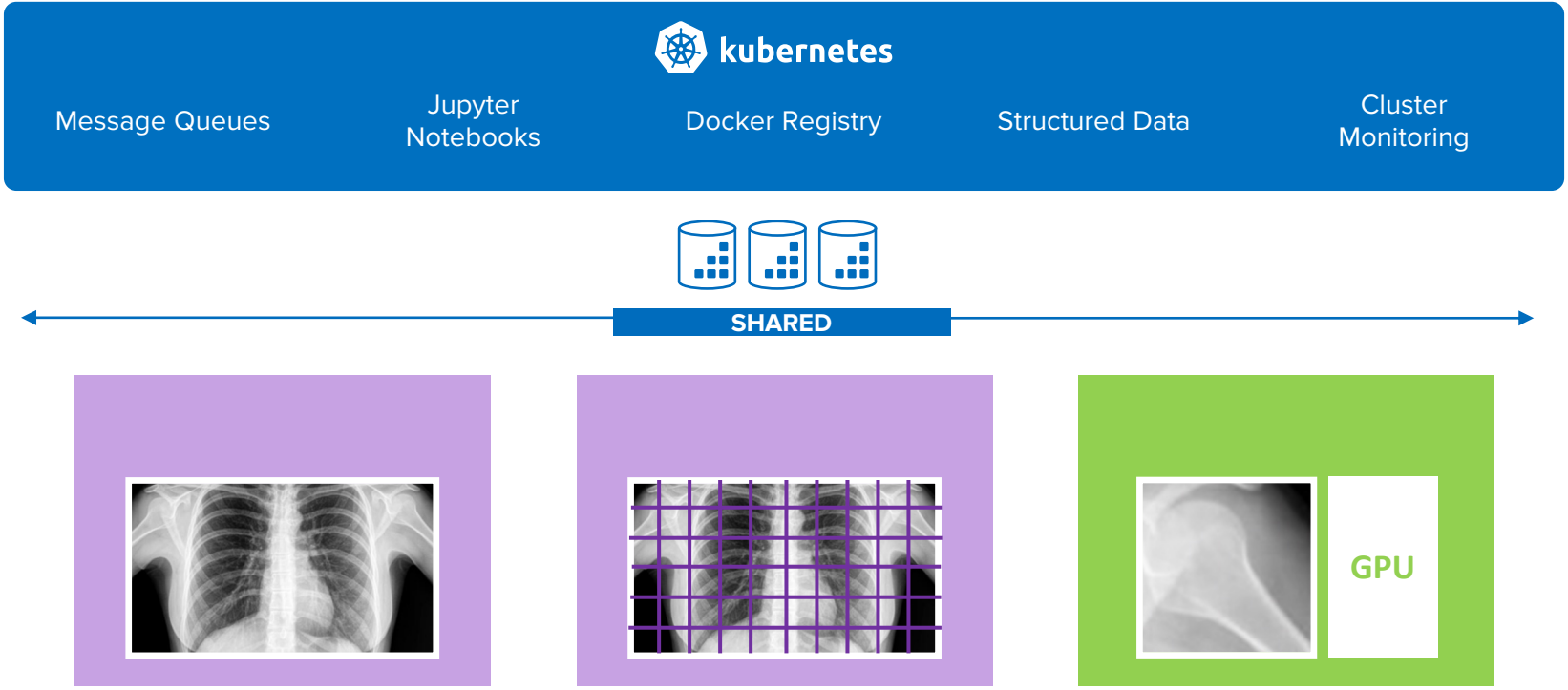


Source: https://towardsdatascience.com/the-rise-of-the-term-mlops-3b14d5bd1bdb

Example Inference Pipeline



Example Inference Pipeline



Unstructured Data Handling Challenges in MLOps

Repeated shared reads for training

Don't think about whether data access performance will fall off a cliff as you add more jobs

Situation

- Training jobs repeatedly reread data, many jobs mixing together to create a constant random read workload.
- Many storage systems hit a performance cliff above a certain randomness percentage.

Ideal

- Access data so it makes sense logically, not to avoid storage quirks.

Solution

- Keep data sets on storage that performs well and scales predictably regardless of workload mix.
- Fast shared storage also eliminates copy steps and don't depend on caching for good performance.

Background read load for data munging

Flash Memory Summit

Don't think about background data exploration & wrangling adding stress to production storage

Situation

- Data needs to be examined, iteratively improved, and cleaned before training even starts.
- The usual ETL, plus fixing class imbalances, adding perturbation, etc.

Ideal

- Explore, wrangle, and munge all you need without worrying about disturbing production training and inference jobs.

Solution

- Make sure your storage has a deep well of random read and sequential write throughput, plus can handle a simultaneous mix of those workloads.

Metadata and many objects

Flash Memory Summit

Don't worry about millions of small objects and random access to metadata

Situation

- Labels and annotations (e.g., bounding boxes) are often in a separate directory or bucket.
- Storage tuned for large file throughput can choke on millions of tiny files and be slow on parts of the workflow.

Ideal

- Use as many labels, annotations, and tags as needed – in whatever format you like – without worrying about storage slowing your jobs.

Solution

- Choose a storage platform that handles billions of small objects and metadata, as well as delivers high throughput for large objects.

Checkpoints at scale

Flash Memory Summit

Don't worry about coding checkpoints at logical intervals to spare the storage overhead

Situation
- Most training jobs write a checkpoint of the model file back to storage, and multiple checkpoints at the same time can overwhelm storage systems tuned for high read performance.

Ideal
- Code jobs to checkpoint at logical intervals, without worrying about overwhelming the storage and interfering with other jobs.

Solution
- Rely on a storage system that's able to handle write workloads at scale and perform predictably as the number of jobs or checkpoints increases.

Enumerating datasets

Don't worry about getting stuck crawling directory trees

Flash Memory Summit

Situation

- Training jobs need to randomize data inputs to prevent overfitting. They need to enumerate all the objects in the dataset to randomize them. When datasets include millions of objects, the process of getting all the metadata can take a very long time and often repeats for every training job.

Ideal

- Enumerating the dataset and randomizing the order happens quickly.

Solution

- Parallelize the metadata operations and use storage that can handle many thousands of metadata requests per second.
- Use parallel tools that can easily plug into existing scripts and workflows.
- Invest in code to generate a manifest of static datasets once, and then randomize the order of the items for each job.

Making the most out of your data for MLOPs

AI-at-scale for every enterprise

AIRI//S extends the power of NVIDIA® DGXA100™ systems

PURESTORAGE® NVIDIA.

Flash Memory Summit

AIRI//S™ AI-READY INFRASTRUCTURE

Industry's first to simplify AI-at-scale

Data scientist teams can focus on algorithms, not infrastructure

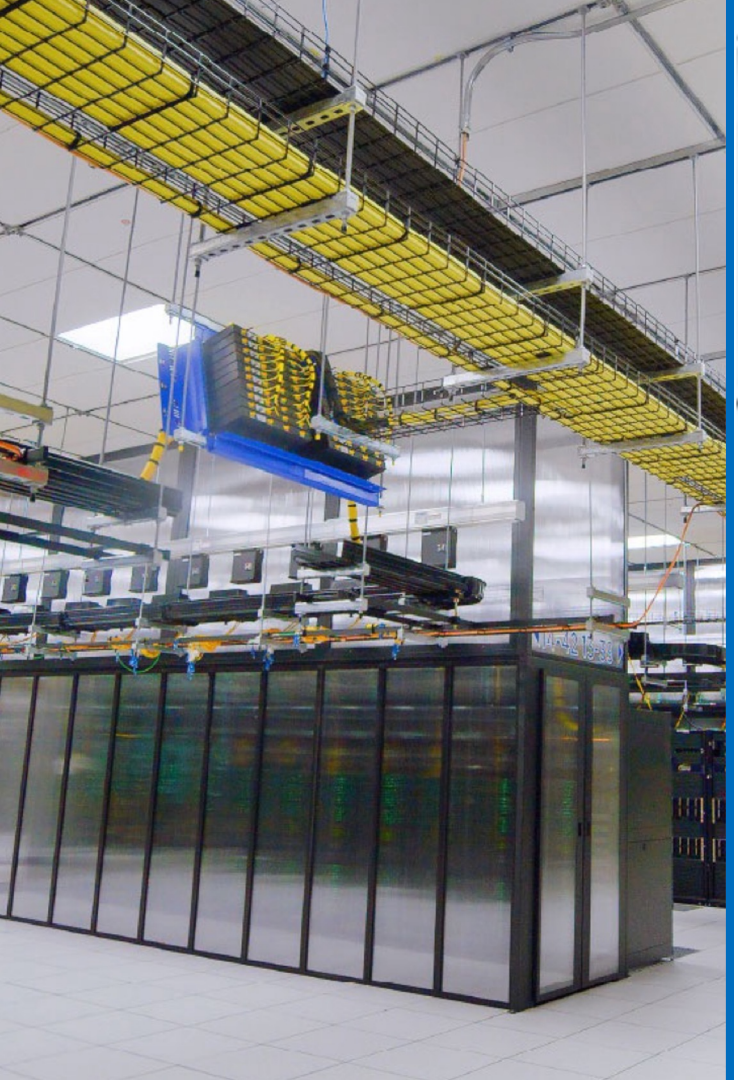
More power-efficient storage paves the way for more GPU power

Ample power is available for additional GPUs as needed

Slash training, deployment, and operational complexity

Only a few experts can run multi-node training, AIRI//S makes it simple

35 YEARS OF NAND FLASH MEMORY INNOVATION



Pure Storage Partners with Meta on AI Research SuperCluster (RSC)

Pure FlashArray and FlashBlade provide a robust and scalable storage solution for the RSC, which Meta believes is among the fastest AI supercomputers now and soon to be the fastest when fully built out by mid-2022

January 24, 2022

Building on first-generation (2017) infrastructure, RSC increases production training speeds by 20x, planned to expand to 1 Exabyte when fully built-out

Key considerations

- **Performance** - 16TB/s of training data to server to GPU compute
- **Scale** - Active datasets of 100's of PB growing to Exabyte-scale
- **Footprint** - Space and Power constraints require high density and efficiency
- **Reliability** - Critical infrastructure supporting long-running production, and ad-hoc researcher use
- **Security** - RSC designed for security and privacy from the ground up

Better, Faster, Smaller, and More efficient than any of the alternatives

Consolidate Data Sets to Maximize ROI

Flash Memory Summit

Key characteristics of a high performance, unified storage platform

Multi-Dimensional Performance

High throughput and low latency for multiple simultaneous workloads

Intelligent Architecture

Built for flash with simple deployment, management, & upgrades; no constant tuning

Cloud-Ready

Cloud-like agility, flexibility, and consumption choices with on-prem control

Always Available

High availability with non-disruptive upgrades and data protection

Dynamic Scalability

Seamless scaling of capacity, performance, metadata, number of files, and objects

Multi-Protocol Support

Outstanding performance and functionality across NFS, SMB, and S3 protocols

Wrapping Up

Takeaways

Flash Memory Summit

- There's a handful of storage challenges in AI and MLOps, and the right infrastructure can help data scientists stop worrying about them.
- To get maximum value out of ML systems, it’s critical that organizations find a way to integrate all of that unstructured data into a unified platform.
- Sustainable Speed and Simplicity at Scale is the key. Pure FlashBlade addresses the AI/ML challenges and unlocks the value of unstructured data.

Thank You!

