Flash, Storage and Data Challenges for Production Machine Learning

Nisha Talagala

CEO, Pyxeda AI

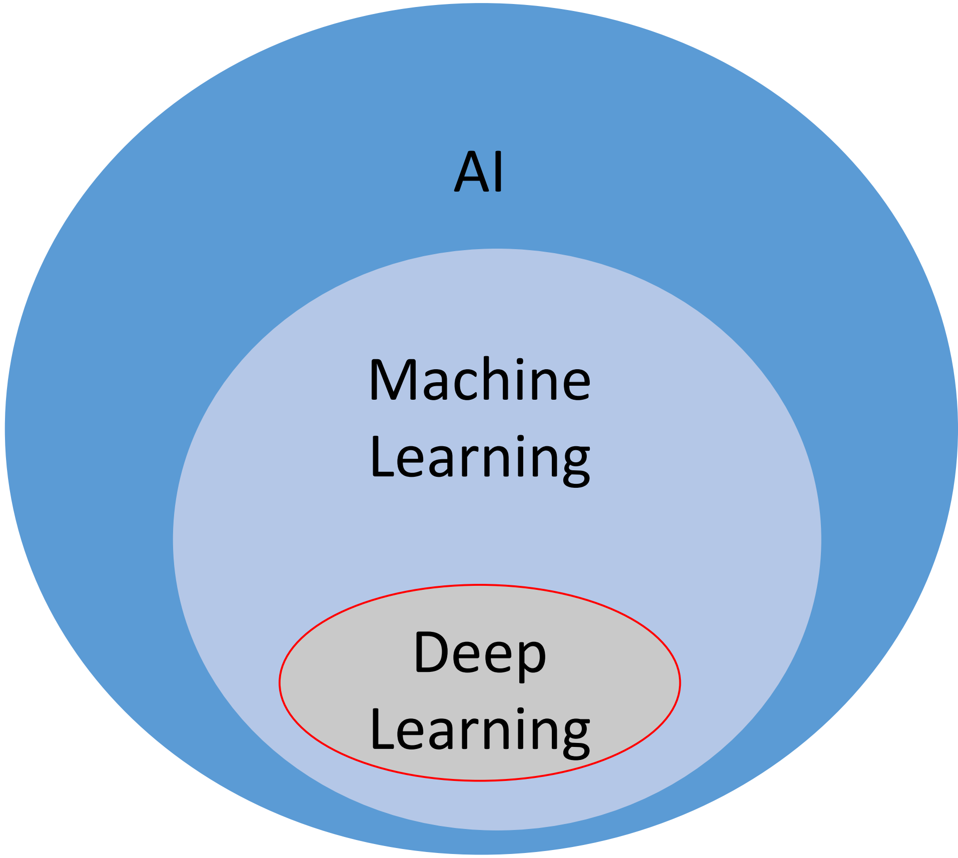
In This Talk:

- AI and ML: A quick overview
- Opportunities for Flash and Storage Systems
  - Workloads
  - Trust, Governance and Data Management
  - Edge
- How Flash and Storage can use ML/DL

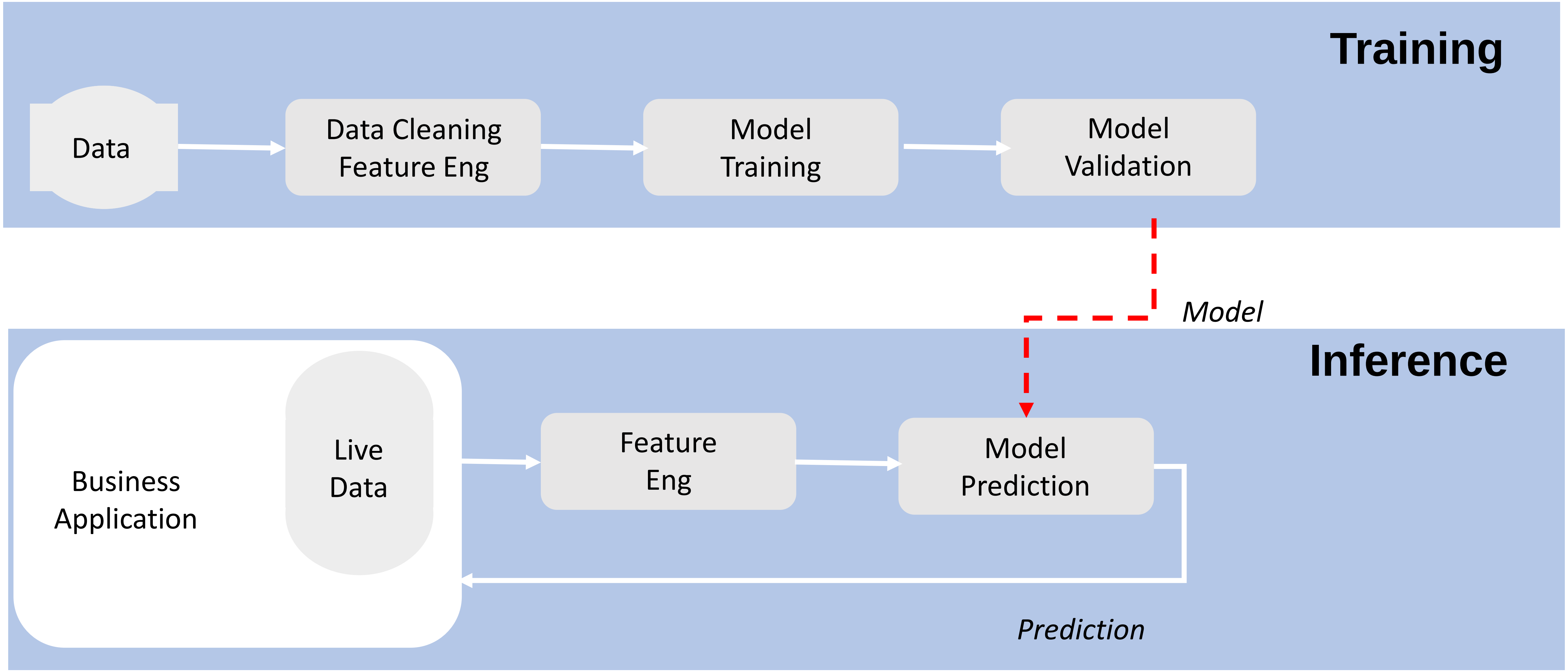
What is Machine Learning and AI?

- AI: Natural Language Processing, Image Recognition, Anomaly Detection, etc.
- Machine Learning: Supervised, Unsupervised, Reinforcement, Transfer, etc.
- Deep Learning: CNNs, RNNs etc.
- Common Threads
  - Training
  - Inference (aka Scoring, Model Serving, Prediction)

Current State: Lots of tools, Lots of experiments, a bit of adoption



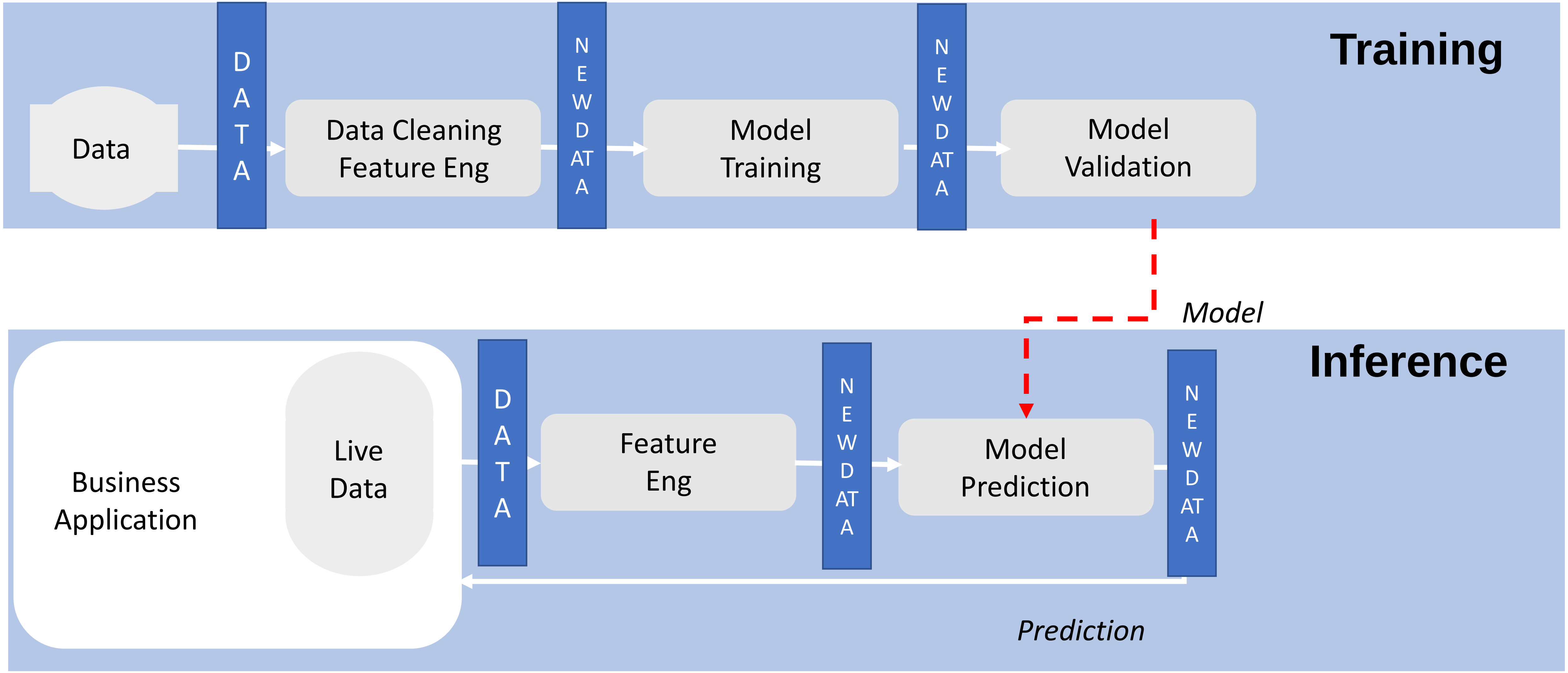
A Typical ML Operational Pipeline



Trend 1: How ML/DL Workloads Think About Data

- Data Sizes
  - Incoming datasets can range from MB to TB
  - Statistical ML Models are typically small. Largest models tend to be in deep neural networks (DL) and range from 10s MB to GBs
  - **Storage and ingest perf is most critical for largest data sets, and with GPUs**
  - **More advanced use cases are also increasing model size – but not common**
- Common Structured Data Types
  - Time series and Streams, Multi-dimensional Arrays, Matrices and Vectors
- Common distributed patterns
  - Data Parallel, periodic synchronization, Model Parallel

What does this mean for data?



Access control, Lineage, Tracking of all data artifacts is critical for AI Trust

Trend 2: Need for Governance

- ML is only as good as its data
- Managing ML requires understanding ***data provenance***
  - *How was it created? Where did it come from? When was it valid?*
  - *Who can access it? (all or subsets)? Which features were used for what?*
  - *How was it transformed?*
  - *What ML was it used for and when?*
- Solutions require both storage management and ML management

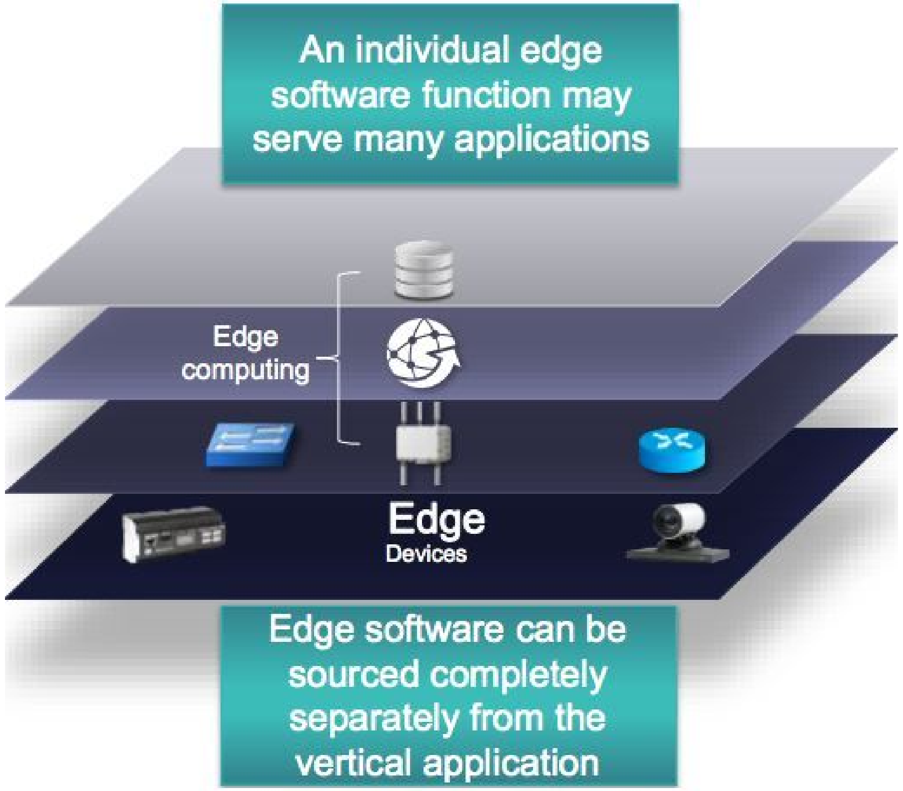
Trend 2: Need for Governance

- Examples
  - Established: Example: Model Risk Management in Financial Services
  - https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf
- Example GDPR/CCPA on Data, Reproducing and Explaining ML Decisions
  - https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/
- Example: New York City Algorithm Fairness Monitoring
  - https://techcrunch.com/2017/12/12/new-york-city-moves-to-establish-algorithm-monitoring-task-force/

Trend 3: The Growing Role of the Edge

Pyxeda

- Closest to data ingest, lowest latency.
  - Benefits to real time ML inference and (maybe later) training
- Varied hardware architectures and resource constraints
- Differs from geographically distributed data center architecture
- Creates need for cross cloud/edge data storage and management strategies


An individual edge software function may serve many applications
Edge computing
Edge
Devices
Edge software can be sourced completely separately from the vertical application

IoT Reference Model

Flash and Other Storage for ML: Opportunities

- Data access Speeds (Particularly for Deep Learning Workloads)
- Data Management
- Reproducibility and Lineage
- Governance and the Challenges of Regulation, Data Access Control and Access Management
- The Edge

In This Talk:

- AI and ML: A quick overview
- Opportunities for Flash and Storage Systems
  - Workloads
  - Trust, Governance and Data Management
  - Edge
- **How Flash and Storage can use ML/DL**

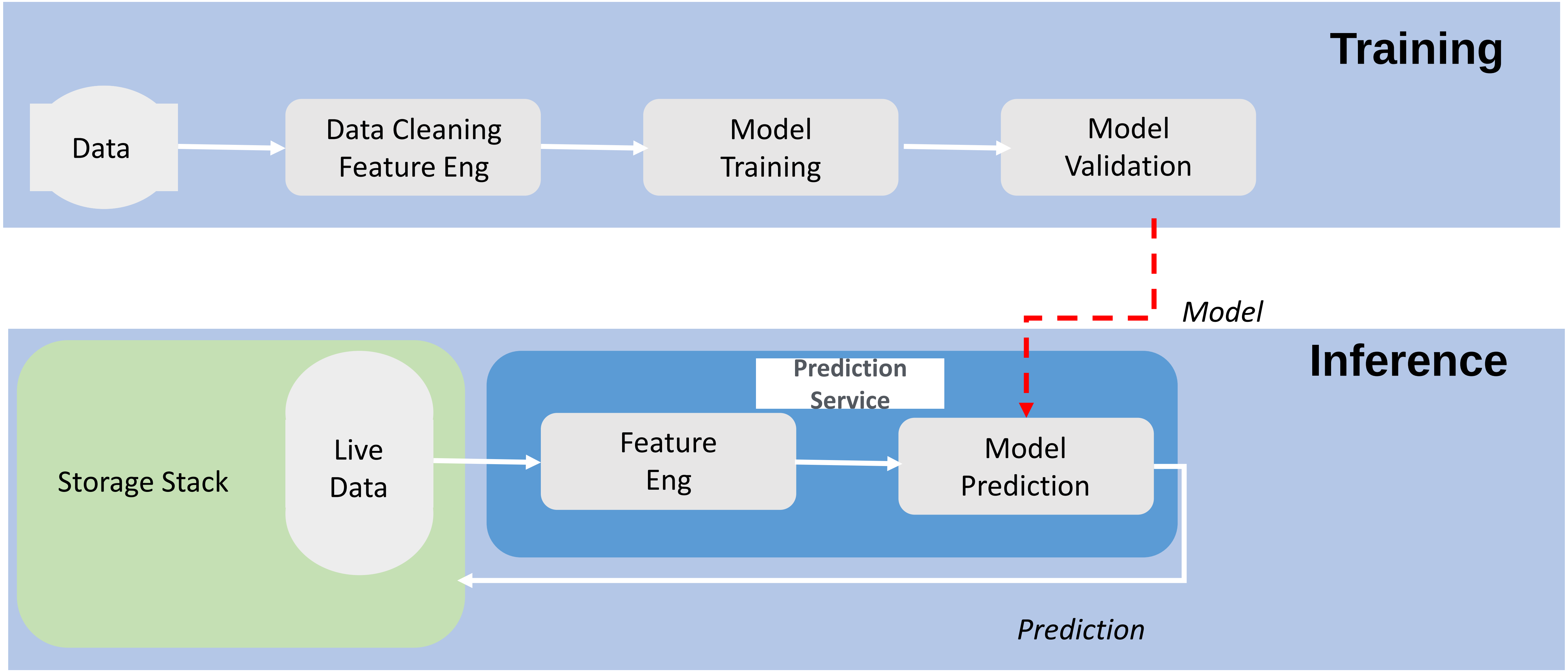
How to Use ML/DL for Storage - Examples

- Caching
  - Adapting caching policy using online learning can have significant benefits
- Workload classification and resource optimization
  - Quantify similarity between workloads
  - Track workload changes
  - Learning workload mixes
- Learning for storage tuning
  - Data distribution / tiering
  - Reconfiguration of parameters, tiers, placement and layout
- Failure Prediction

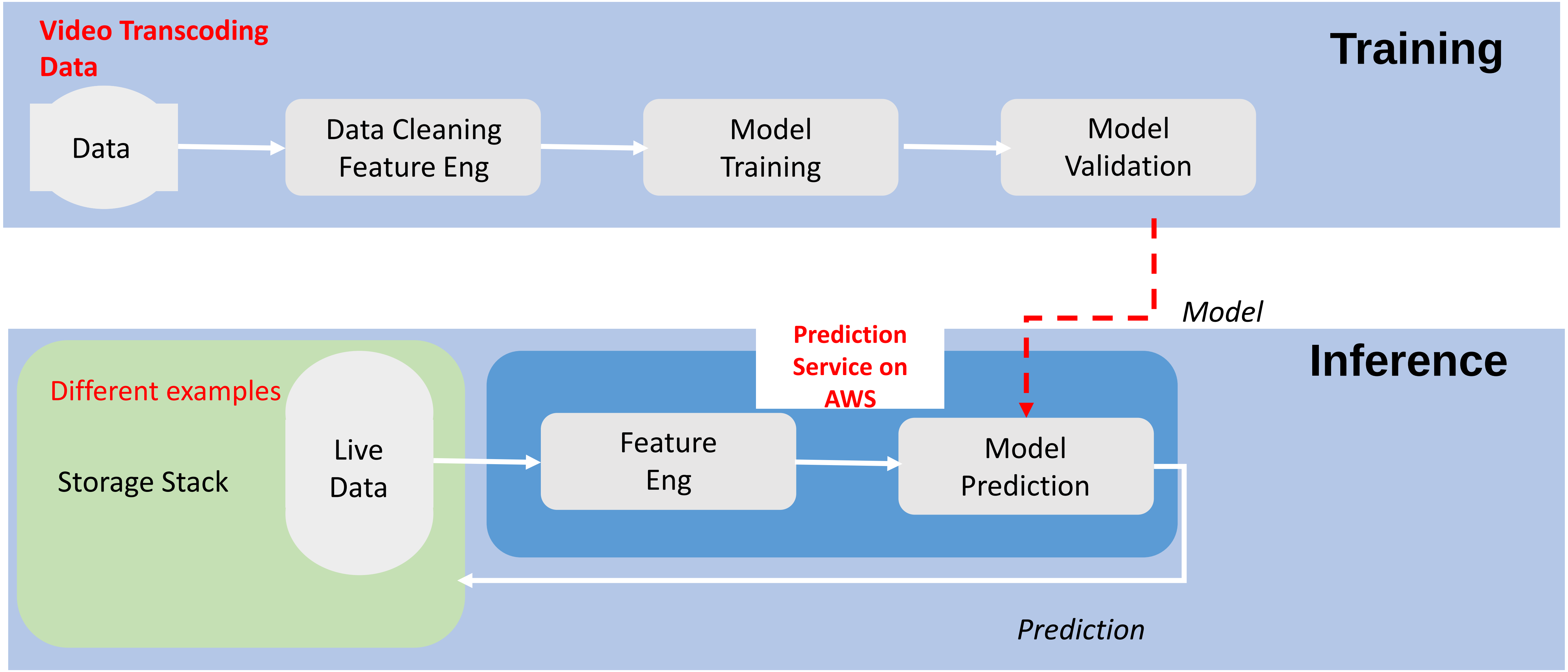
**Taken from NFS Vision Workshop AI and Storage subteam report*

How to add ML/DL to your Storage Stack

Pyxeda



Quick Demo


Video Transcoding Data
Training
Data
Data Cleaning Feature Eng
Model Training
Model Validation
Model
Prediction Service on AWS
Inference
Different examples
Storage Stack
Live Data
Feature Eng
Model Prediction
Prediction

Demo

Takeaways

- The use of ML/DL in enterprise is at its infancy

- Storage/Flash for AI
  - The first and most obvious storage challenge is performance
  - The larger challenge is likely data management and governance
  - Edge and distribution are also emerging challenges
- AI for Storage/Flash
  - Many opportunities exist for systems optimization using ML/DL

Resources

- If you want to build your own ML use case for your storage data, go to http://aiclub.world/signup and get a free account. Send me email if you would like the sample dataset or the video (nisha@pyxeda.ai)
- Examples of Storage for ML and ML for Storage
  - NFS Vision report on Storage for 2025 - See Storage and AI track
  - Proceedings/Slides of USENIX OpML 2019
  - Research at HotStorage, HotEdge, FAST, USENIX ATC
  - Storage Systems for ML: Databricks Delta, Apache Atlas
  - RDMA data acceleration for Deep Learning (Ex. from Mellanox)
  - Time series optimized databases (Ex. BTrDB, GorrillaDB)
  - Memory expansion (Ex. Many studies on DRAM/Persistent Memory/Flash tiering for analytics)
  - RDMA and GPU connectivity (see Mellanox)

Thank You

Nisha Talagala

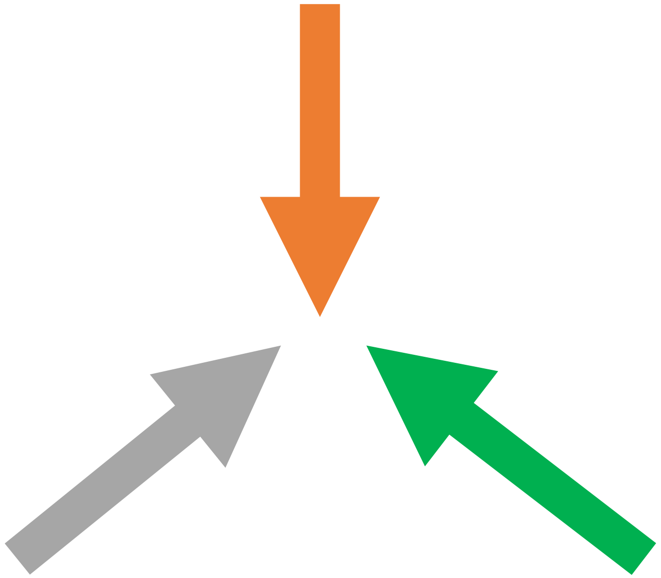
nisha@pyxeda.ai

Trend 1: How ML/DL Workloads Think About Data

- The older data gets – the more its “role” changes
  - Older data for batch- historical analytics and model reboots
  - Used for model training (sort of), not for inference
- Guarantees can be “flexible” on older data
  - Availability can be reduced (most algorithms can deal with some data loss)
  - A few data corruptions don’t really hurt ☺
  - Data is evaluated in aggregate and algorithms are tolerant of outliers
  - Holes are a fact of real life data – algorithms deal with it
- Quality of service exists but is different
  - Random access is very rare
  - Heavily patterned access (most operations are some form of array/matrix)
  - Streaming is starting to gain traction

Machine Learning Growth

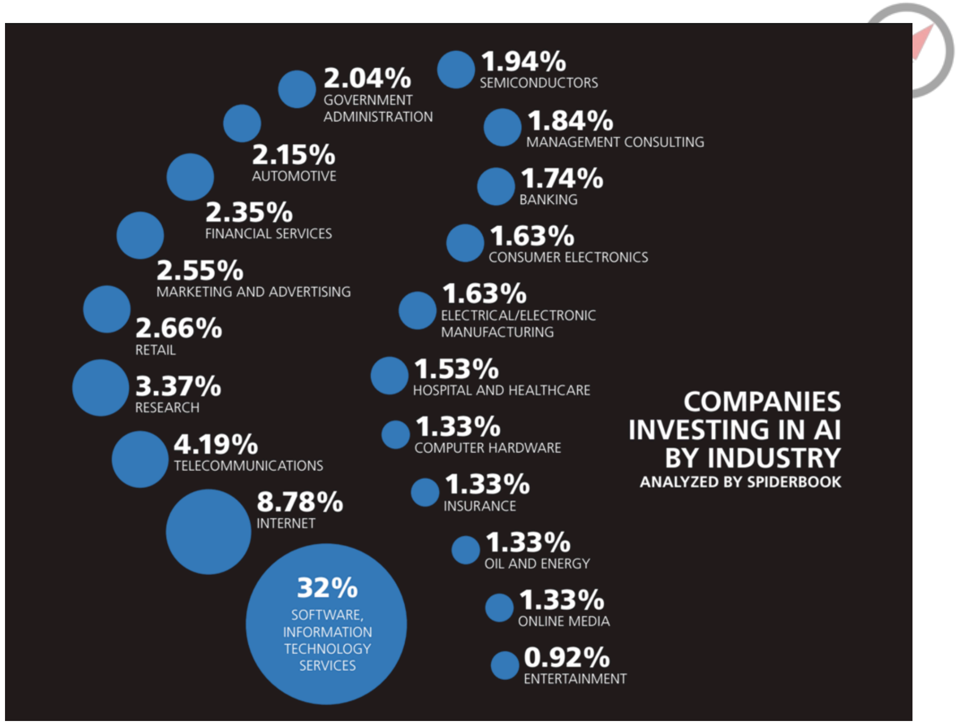
Data: Sources
and Storage

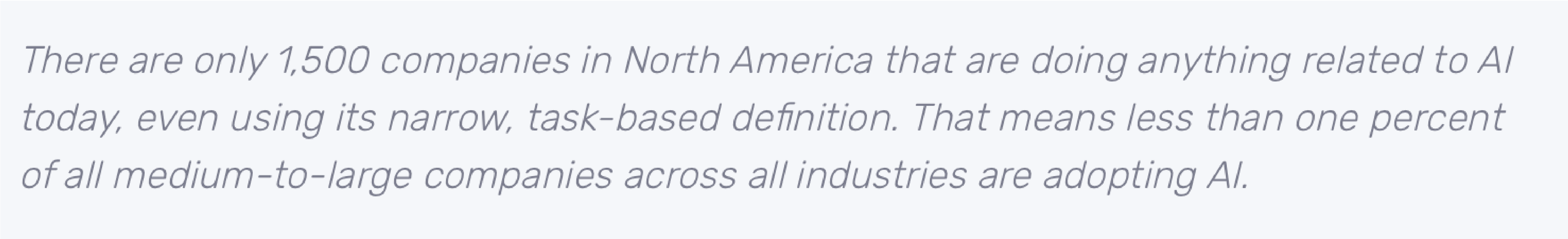


Compute:
Cloud, Hardware
Innovation

Algorithms and
Open Source

Realities of Production Use





There are only 1,500 companies in North America that are doing anything related to AI today, even using its narrow, task-based definition. That means less than one percent of all medium-to-large companies across all industries are adopting AI.

Despite the advanced services available, AI usage still minimal

https://www.oreilly.com/library/view/the-new-artificial/9781492048978/

https://emerj.com/ai-sector-overviews/valuing-the-artificial-intelligence-market-graphs-and-predictions/