

MLOps, meet OPI

Debo Dutta - VP Engineering

About me

Current: VP Engineering, Nutanix (multi cloud platform co)

AI4Infra (P&L engg) & Infra4AI (opensource)

Founding member, MLCommons.org

MLOps/Kubeflow.org (c2017)

Openstack (c2012)

Past:

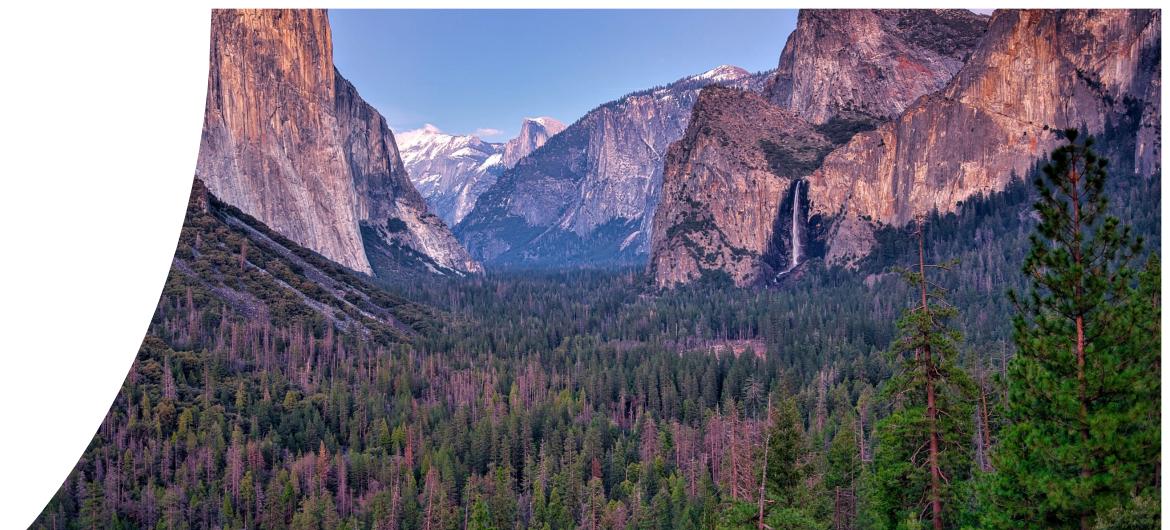
Visiting Scholar, Stanford

Distinguished Engg.@Cisco

Researcher in Computational Biology

PhD CS, USC/ISI (NS2 group)

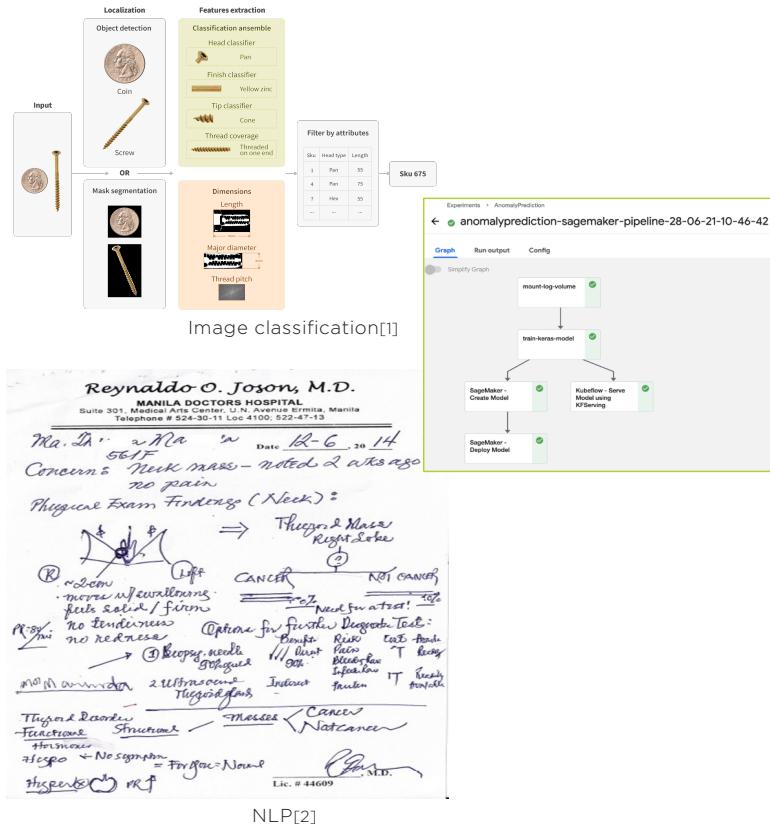
Btech, IIT Kharagpur



AI fueling digital transformation

- Data -> AI -> Insights -> Actions

- Enterprise data explosion – databases, conversations, logs, metrics, knowledge graphs
- Digital transformation: use enterprise data to increase operational efficiency
- AI accelerates digital transformation by providing learned insights which leads to better actions.



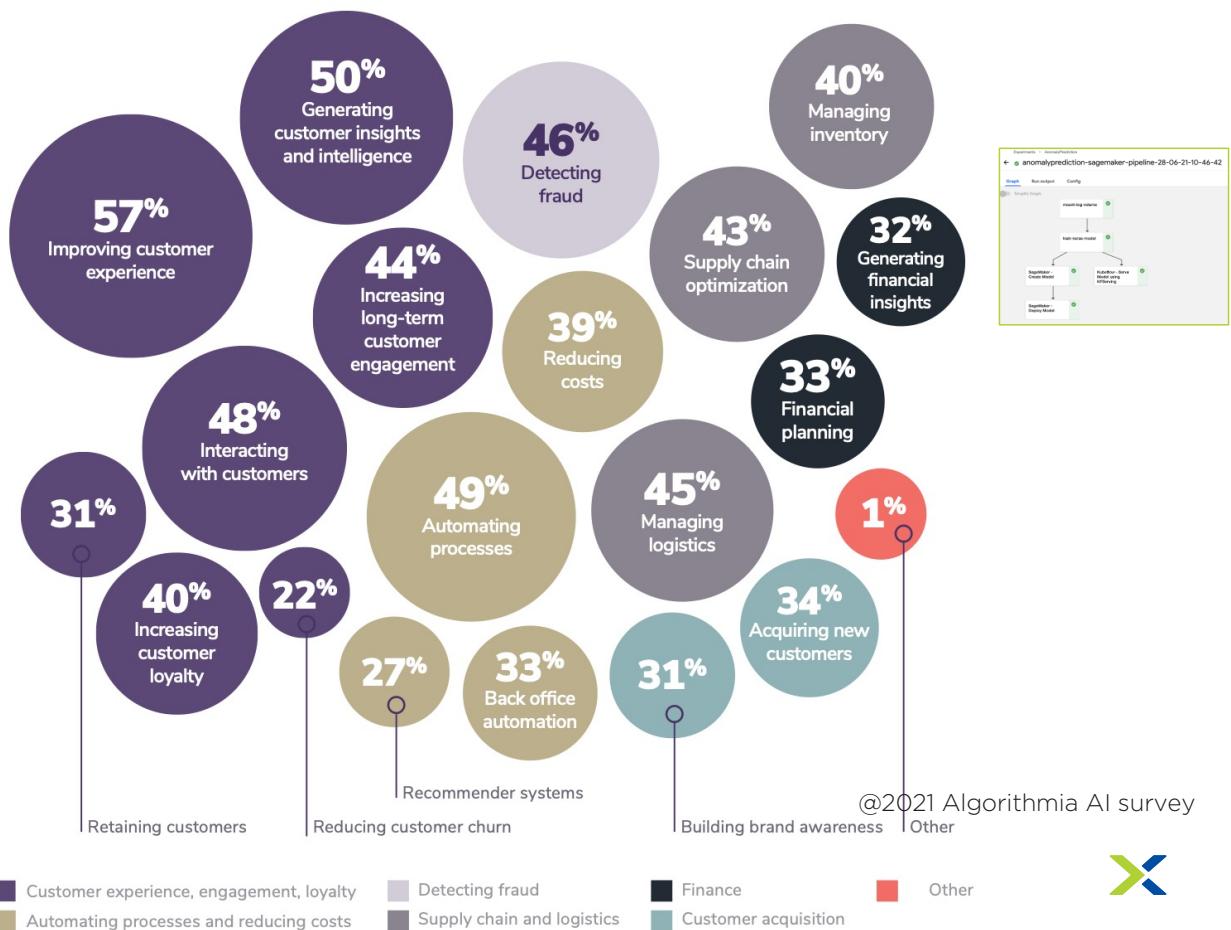
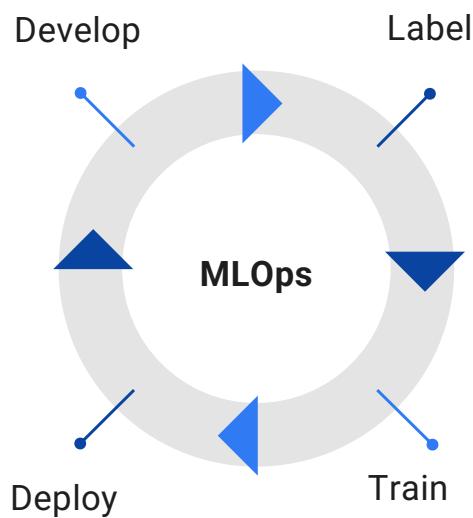
- Key challenge: Accelerating traditional businesses with MLOps

- Finance, retail, pharma, healthcare, public sector
- Manage data and ML

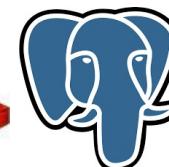
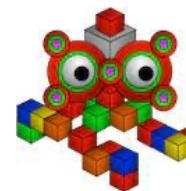
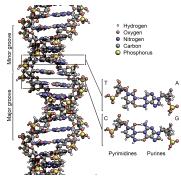


[1] <https://blog.griddynamics.com/identifying-screws-a-practical-use-case-study-for-visual-search/>
 [2] <https://link.springer.com/article/10.1007/s11042-020-10151-w>

Enterprise AI/ML use cases (similar infra needs)

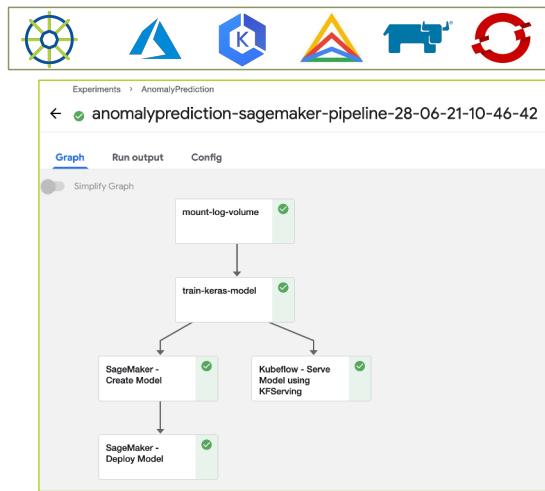


Where is the “*valuable*” enterprise data?



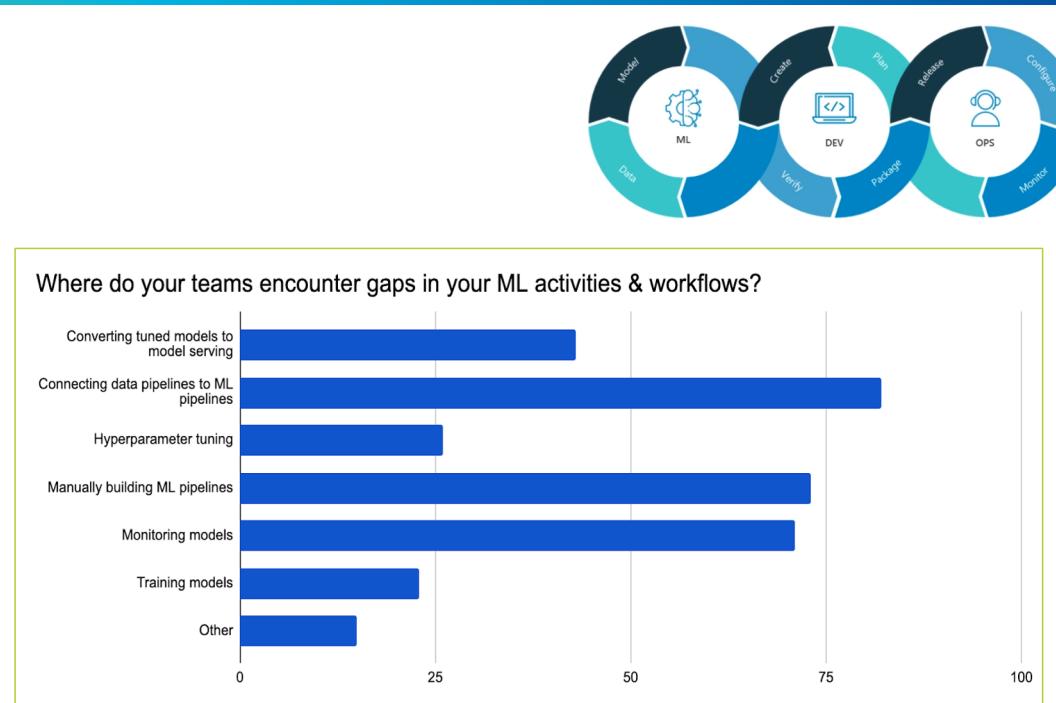
Source: images.google.com

OPI Opportunity: MLOps

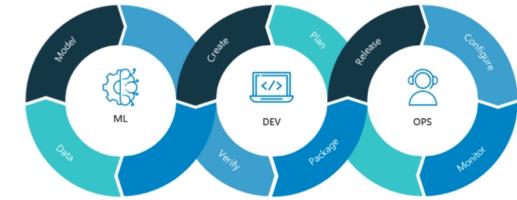


End to end Workflow Orchestration

- Repeatable
- Composable
- Re-usable



MLOps Market Opportunity

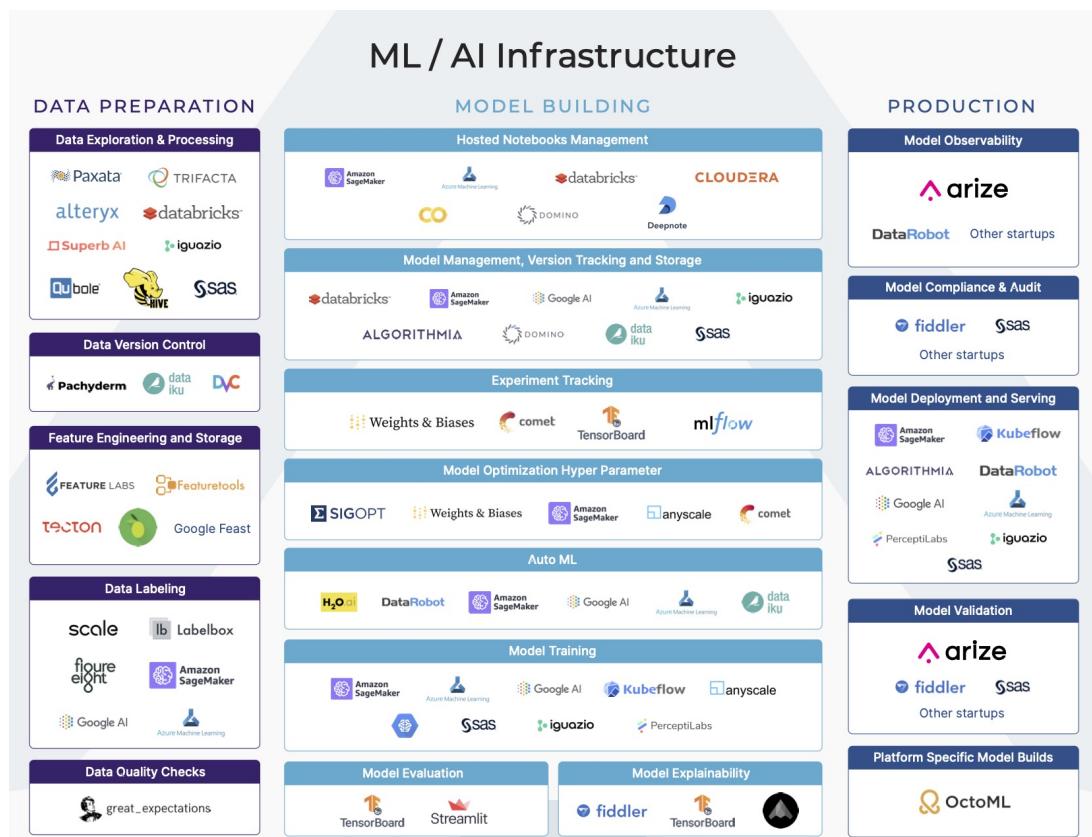
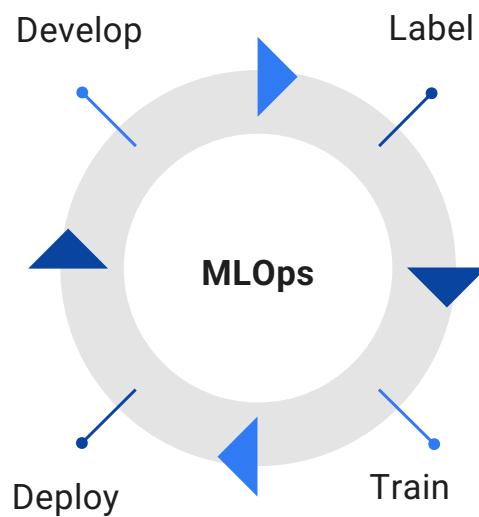


*According to Cognilytica, the global MLOps market will be worth **\$4 billion** by 2025.*

*The industry was worth **\$350 million** in 2019.*

Source: <https://askwonder.com/research/historical-global-ai-ml-ops-spend-uhshsljuf>

MLOps ecosystem: confusion



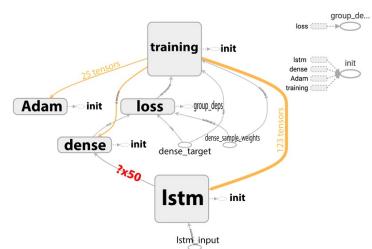
Source: images.google.com



OPI Opportunity: Optimizing MLOps

Current ML Infra is disaggregated

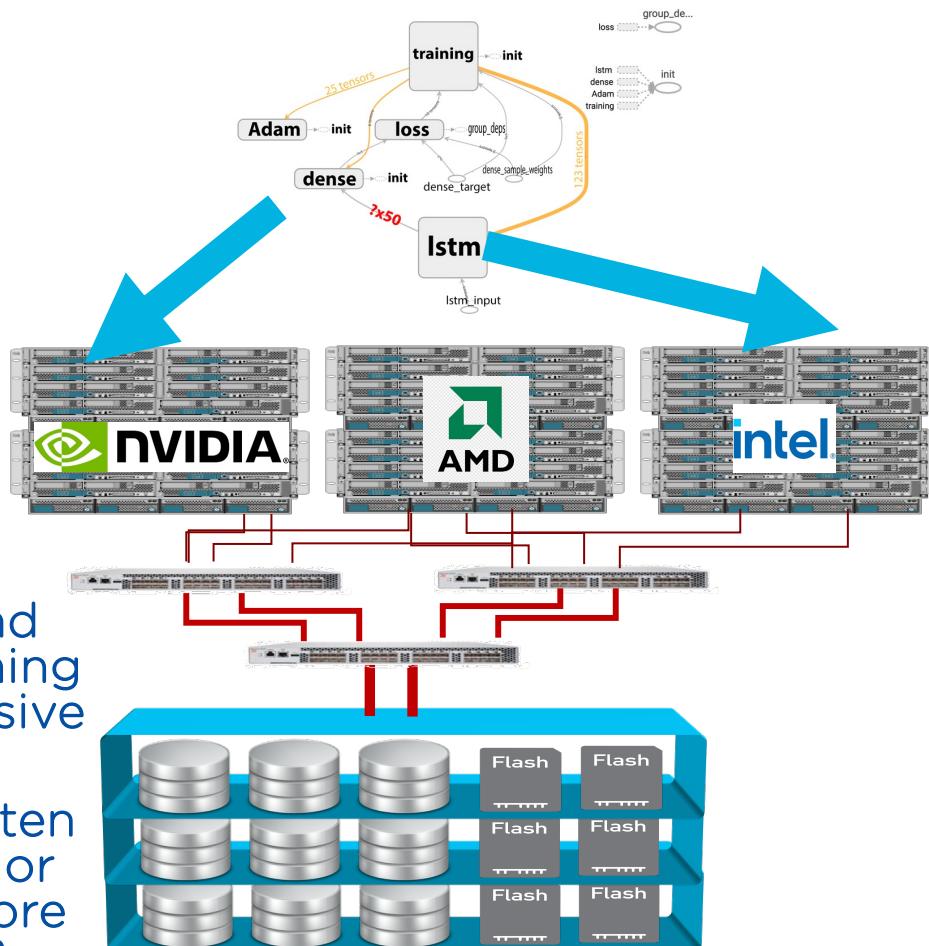
Training is inefficient due
to communication
complexity



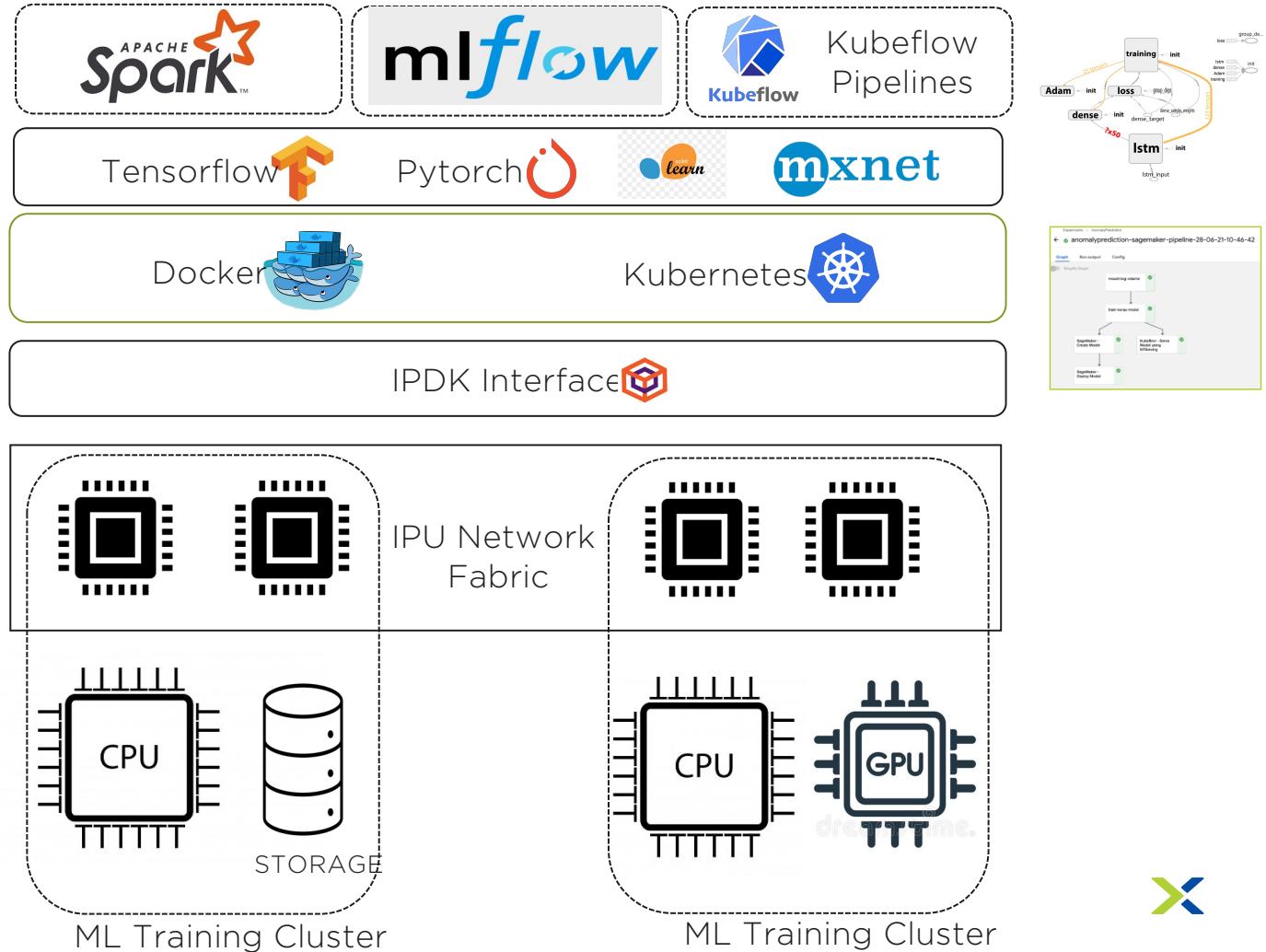
Ref:https://www.researchgate.net/figure/Example-of-TensorFlow-graph-some-parts-of-the-graph-are-enlarged-for-better-visibility_fig3_324362772

ETL and
Pre-training
Is expensive

Data is often
in a filer, or
object store
or in DBs



MLOps + OPI



Kubeflow

@johnugeorge, Kubeflow/Tensorflow



Potential OPI directions:

Reference Arch for ML With OPI

MLPerf workloads

MLCommons “cubes”

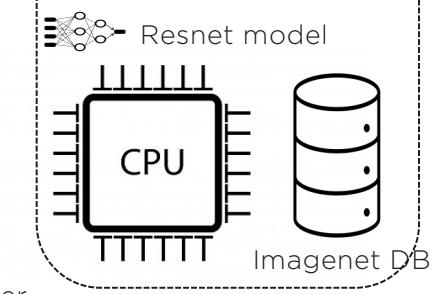
MLPerf Storage-ML



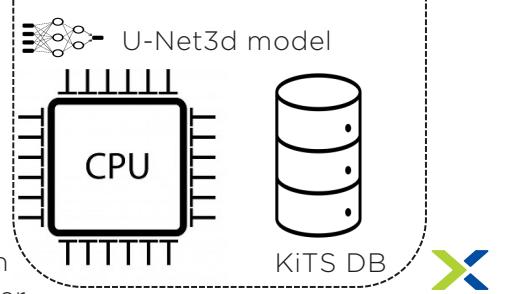
@johnugeorge, co-chair storage/ML



IPDK Interface



IPU Network
Fabric



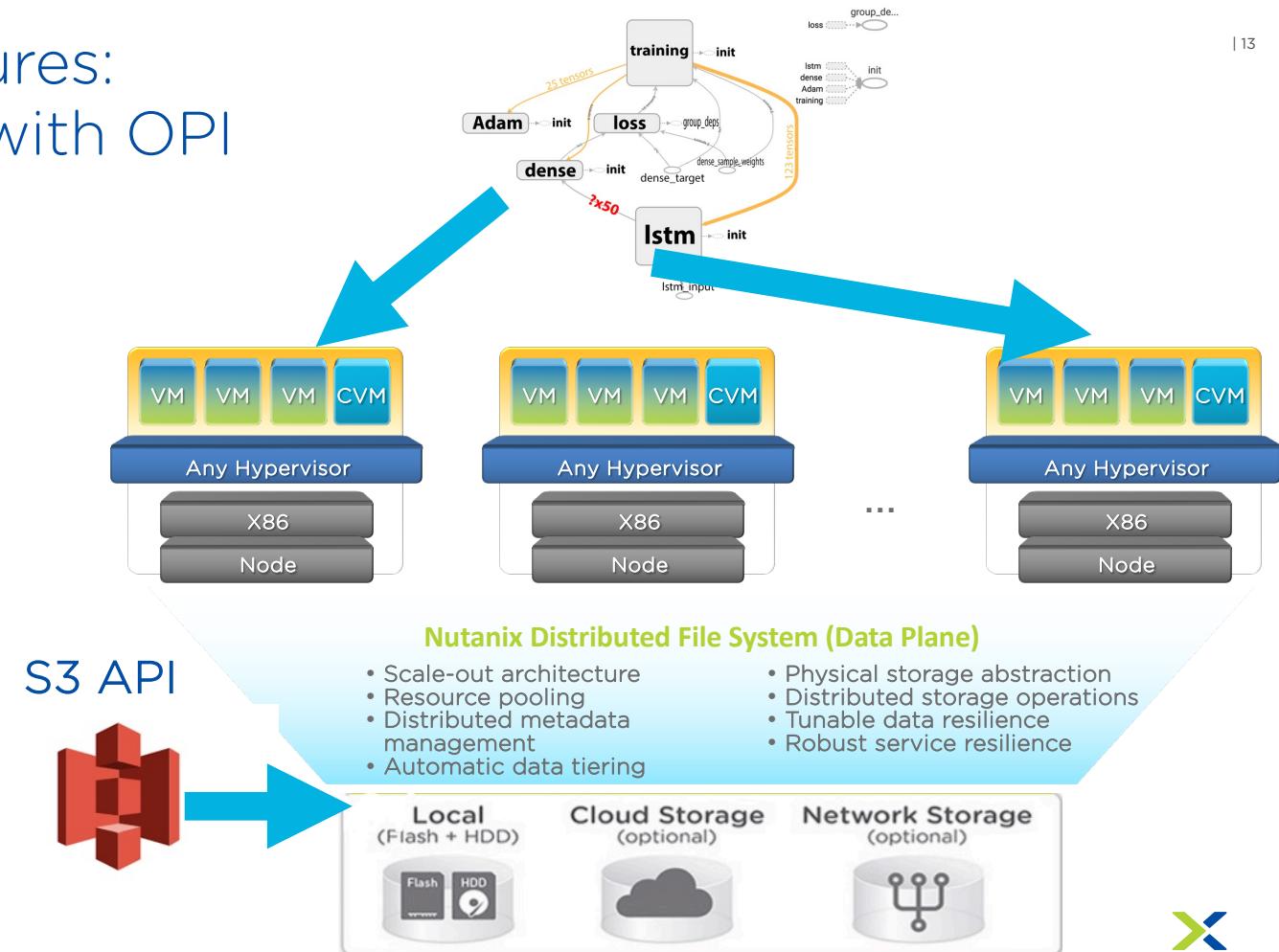
ML Image
segmentation
training cluster

Research-y Futures: Push down ML with OPI

| 13

Collaboration opportunities

- Distributed training
- Split learning
- Federated learning
- SPDK primitives for storage+ML
- Infra optimization - scheduling



NUTANIX™