



# Time to Rethink Visual Data Management for Machine Learning

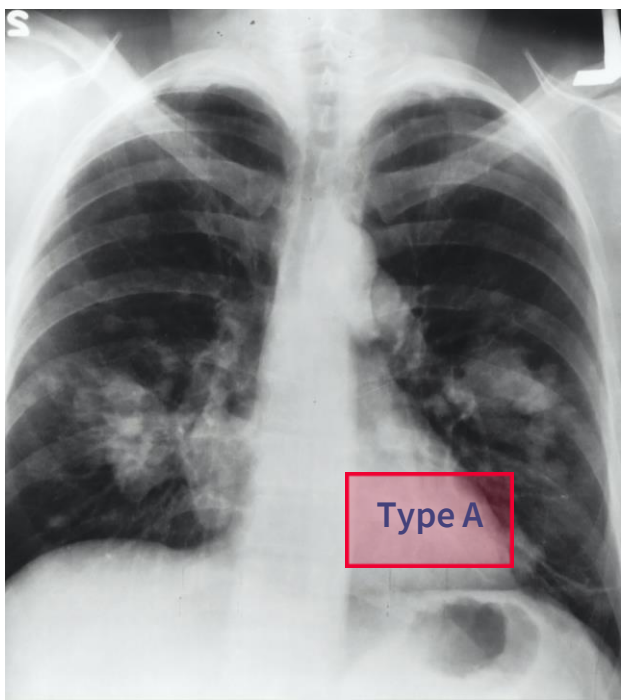
Vishakha Gupta-Cledat and Luis Remis

[team@aperturedata.io](mailto:team@aperturedata.io)

[www.aperturedata.io](http://www.aperturedata.io)

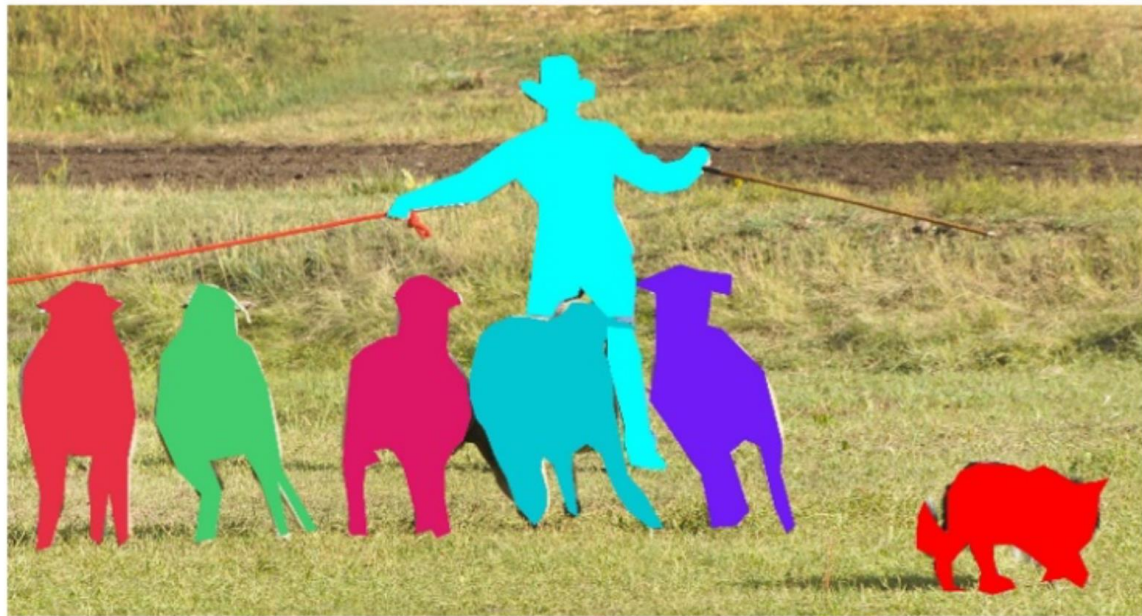
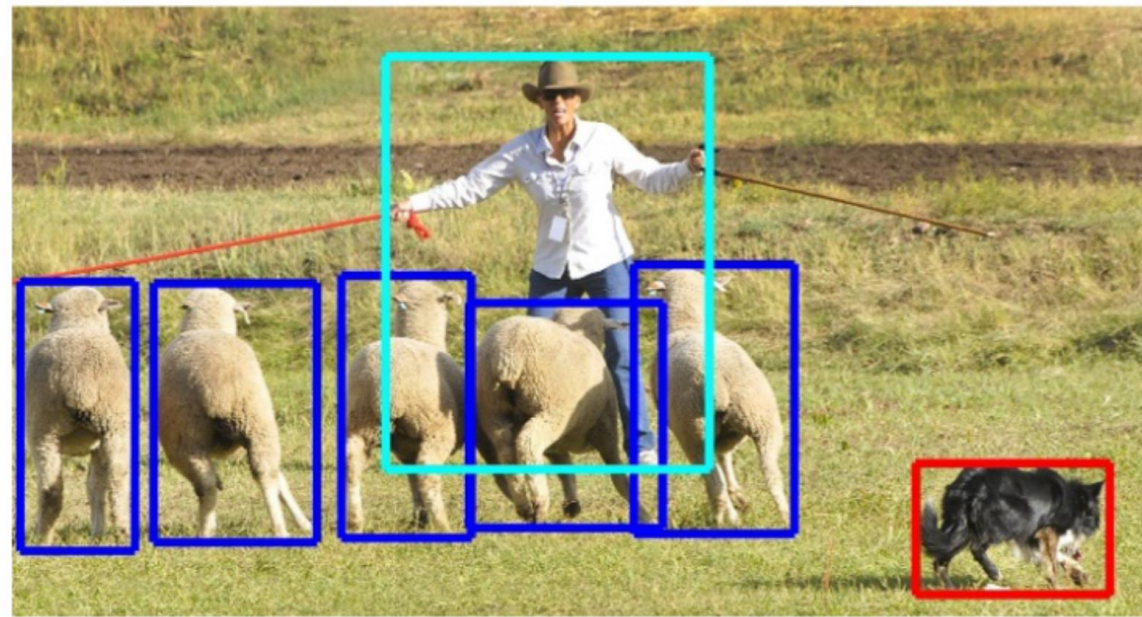


Over 80 Zettabytes by 2025





person, sheep, dog



# Visual Data Management: Compute First

“Data access can be hidden.  
Training takes so long”  
– SysML Professor

“Image datasets can be discarded  
when training is complete”  
– Sr. ML Industry Leader

“We are not worried about  
terabytes of data yet”  
– Data Scientist



# Visual Data Management: Evolving Reality

“It's hard to train on videos. Can I avoid fetching GBs of video for a 2 sec interval?”

- ML Engineer

“Can I store multiple annotations per frame?  
They maybe more complex than boxes”

- Head of Product

“Can I store the embeddings to restart where we left off? or find similar items?”

- Head of Software Engg

“I need to track and snapshot subsets of my dataset to personalize ML models”

- Sr. Data Scientist



# What's Special About Visual Data Now?

## Large and voluminous

Individual image and video can be GB size and start at millions even for small deployments

## Unique transformations

Basic operations like crop / zoom are common and can be computationally expensive

## Extensive metadata

Not just about visual data itself but application context, annotations, features



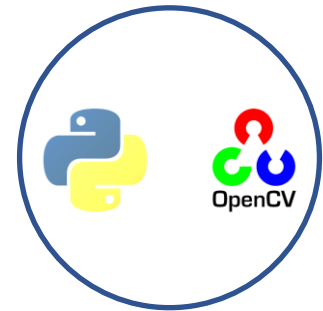
# No Database for Complex Visual Data



Storage buckets for image/video data where they are treated as regular blobs



One among numerous databases for metadata, not always good for all things visual



Numerous libraries for image/video transformation before use

Adapt and glue disparate systems resulting in maintenance nightmare





# Business Impact



Time and resource cost



Hiring cost



Go to market cost

Costs multiple with every new deployment or lead to a new point solution





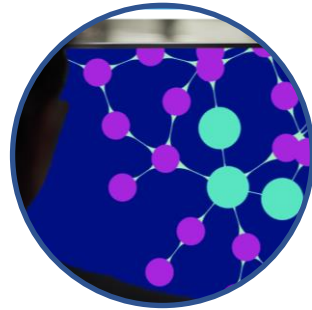
# ApertureDB: A Database for Visual Data



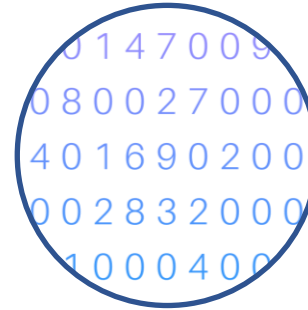
# ApertureDB: Key Features



**Native support for images, videos, bounding boxes, pre-processing and augmentation operations**



**Metadata information as a knowledge graph for easy and more insightful analytics**



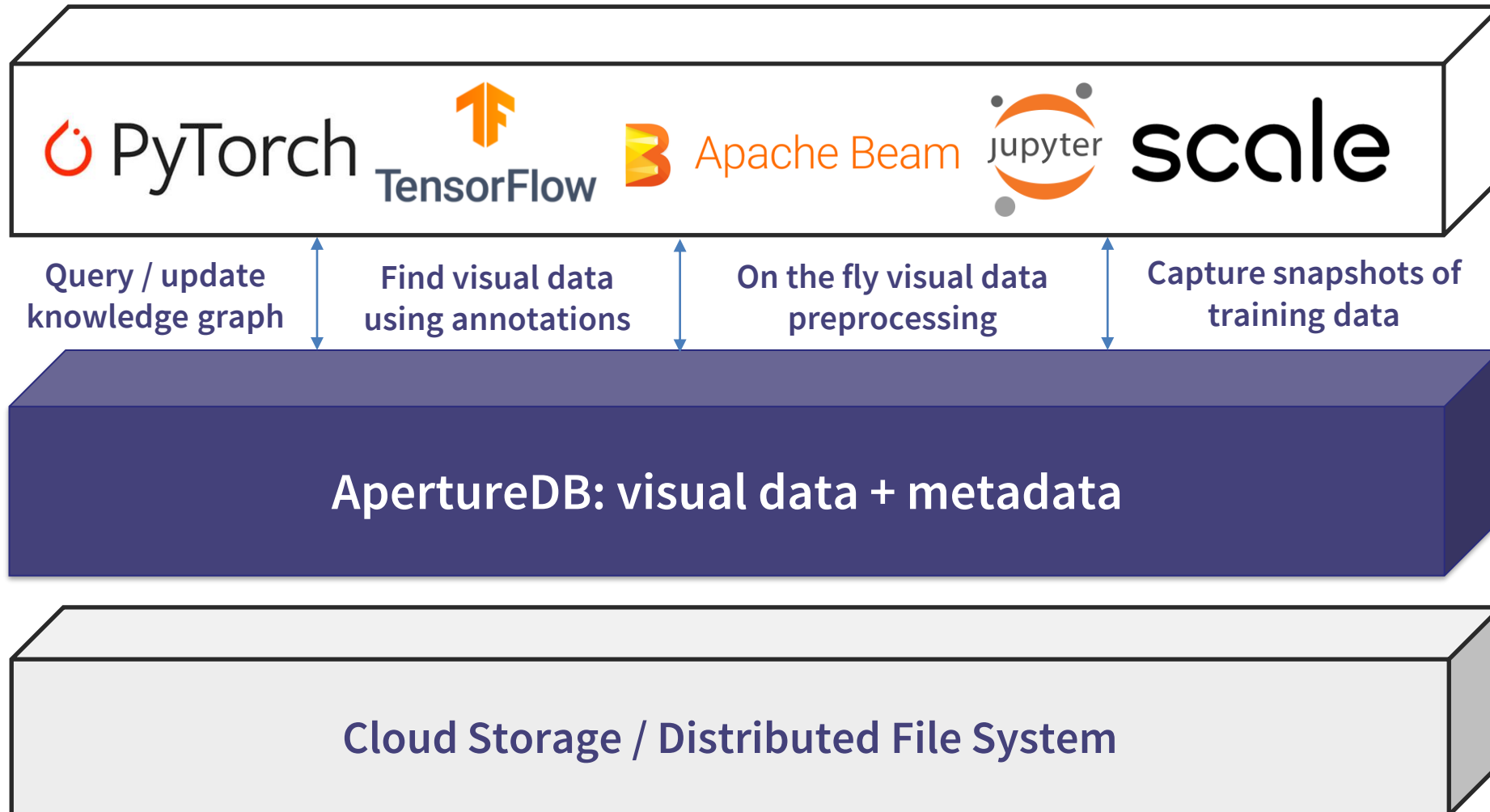
**Built-in similarity matching for high-dimensional feature vectors**



**Save hundreds of hours of data platform engineering leading to faster analytics deployments**



# ApertureDB Ecosystem



Application  
Ecosystem

STORE  
SEARCH  
PRE-PROCESS  
EVOLVE

Data Storage  
Targets





# High Performance: Comparing with State of the Art



vs.



- Metadata-based visual search queries to find right set of images
- Up to **35x** faster and **15x** on average on YFCC100M (~100 million images) dataset



# Resource Efficient: Preprocessing Near Data

**63%**  
**reduction**

in data transferred over the network  
using pre-processing within our API



# Efficient and Scalable: Biotech AI Customer Use Case

## Save weeks

Customer saved weeks of data platform engineering time

## 1B metadata entities

Already loaded over 1B metadata entities and terabytes of data, growing monthly

## 2-3 fewer engineers

Needed 2-3 fewer ML infrastructure engineers

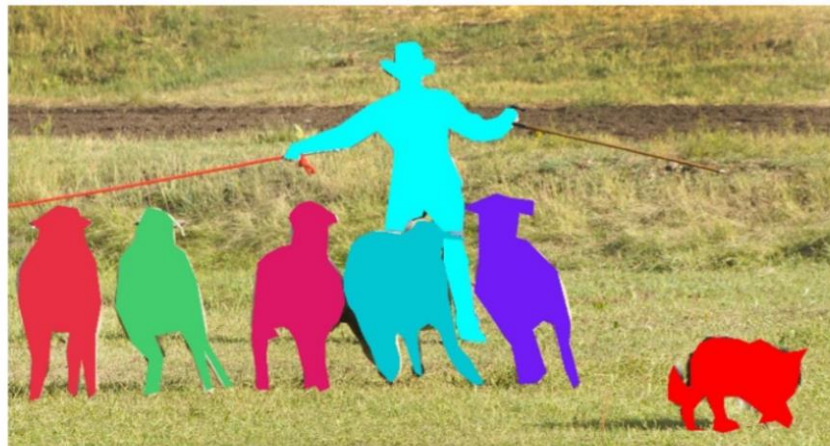
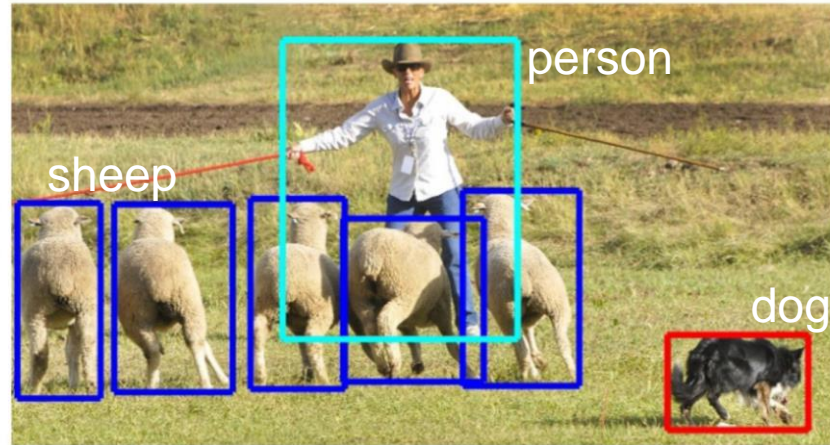




# Demo



# COCO 2017 Dataset



Free cloud trial - let us  
know what you think.

Also, we are hiring!







ApertureData