Breakout Session 2: Track A

Cloud Strategies for Improving Cost, Scalability, and Accessibility of a Machine Learning System for Pathology Images

Dr. Lee Cooper
Associate Professor, Northwestern University

Dr. Andinet Enquobahrie
Senior Director of Medical Computing, Kitware Inc.

Cloud strategies for improving cost, scalability, and accessibility of a machine learning system for pathology images

Lee Cooper, PhD
Associate Professor of Pathology
Director, Computational Pathology
Director, Center for Computational Imaging and Signal Analytics
Northwestern University Feinberg School of Medicine
Chicago, Illinois, USA
lee.cooper@northwestern.edu

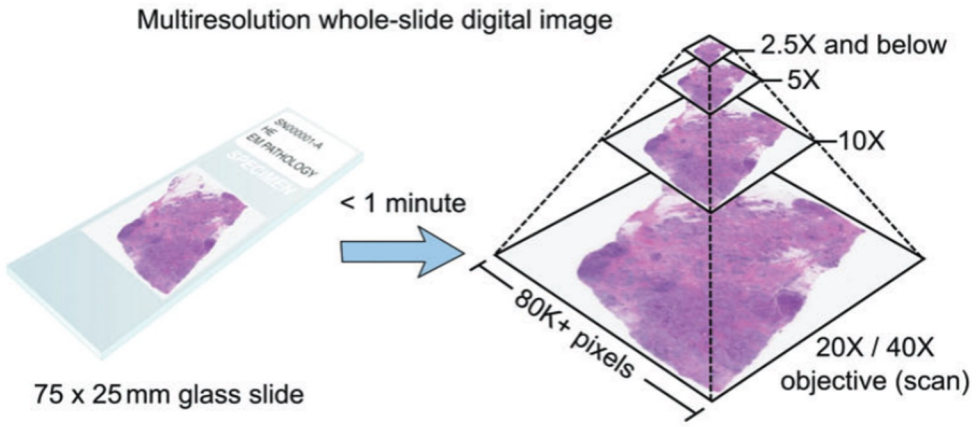
Andinet Enquobahrie, PhD
Senior Director of Medical Computing
Kitware Inc.
Carrboro, NC, USA

R01LM013523-03S1

Parent Project (R01LM013523)

Improve data labeling efficiency and model generalization in computational pathology

3.5 petabytes per year (1.5M slides)



FDA NEWS RELEASE

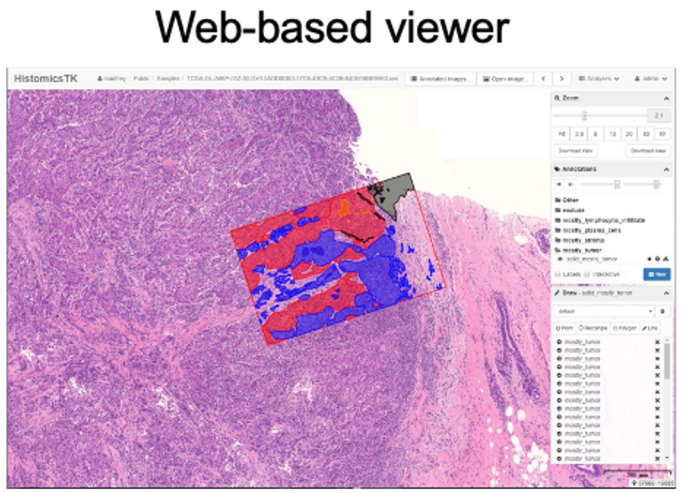
FDA allows marketing of first whole slide imaging system for digital pathology

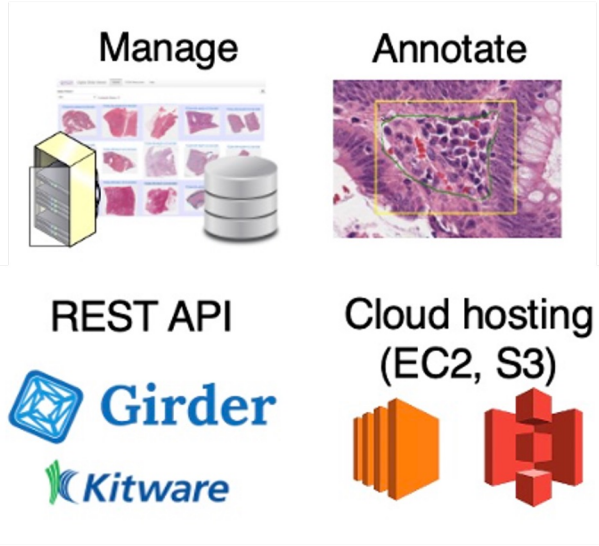
- Massive unlabeled datasets
- Labeling rare instances
- Selection bias in labeling
- Preanalytical variability leads to poor generalization of AI models

Kitware

ResonantACT

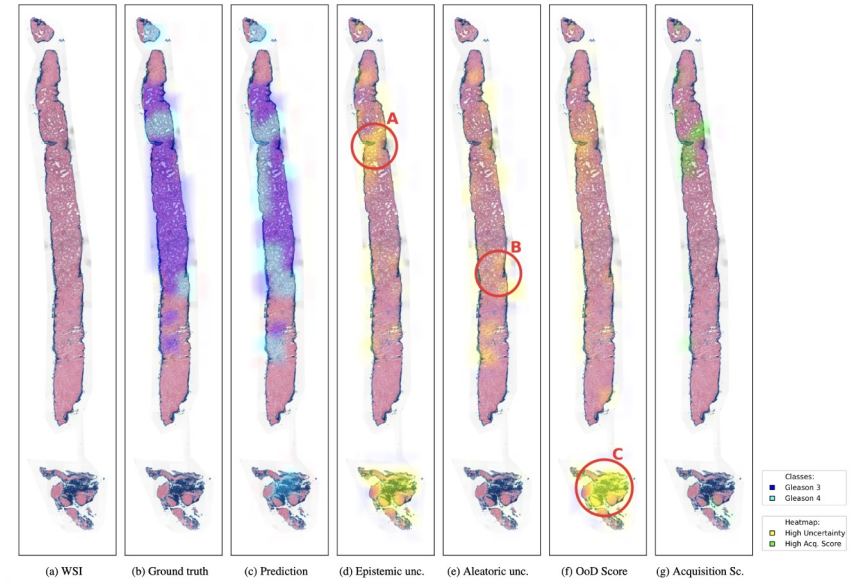
Digital Slide Archive

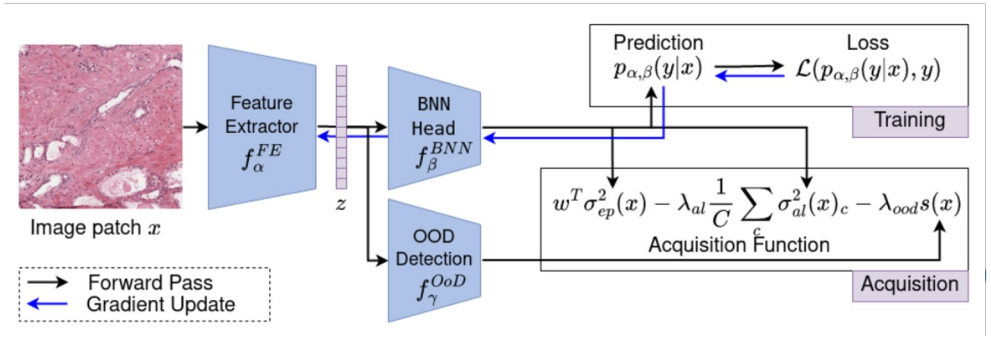




- $12.7M in NIH funding
- 1M+ human annotations generated
- 15K+ Monthly PyPI downloads
- 5 Public challenges with 4000+ participants
- 13 Cancer Center deployments
- 35+ User contributed plugins
- 193+ GitHub contributors
- 2K+ DockerHub pulls

Active learning strategies





Cloud Supplement Goals

Deliver a high-performance cost-effective NVIDIA Triton inference server (TRTIS) solution that is readily deployable on AWS, Azure, and GCP.

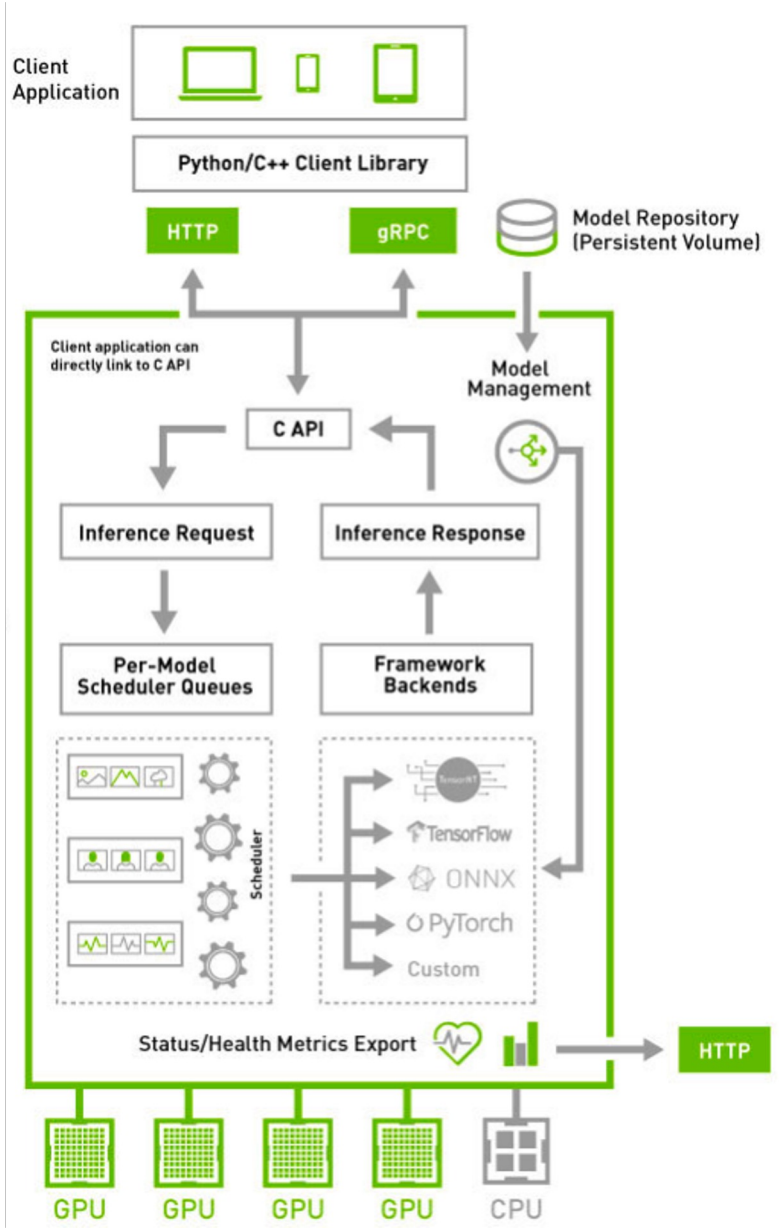
1. Automatic horizontal scaling using NVIDIA Triton inference server (TRTIS)
2. Map the cost : benefit ratio for GPU server asset classes
3. Evaluate impact of data loading strategies and storage asset classes
4. Implement DevOps tools for deployment on AWS, Azure, and GCP.

Northwestern University

Kitware

NVIDIA Triton inference server solution

- Model management, performance metrics, framework support
- Optimizations
  - Model replicates (CUDA streams)
  - Half-precision
  - Scheduling
- Developed a python client for WSI inference (175 MP / sec)
- High performance reader (1.44 GP / sec)



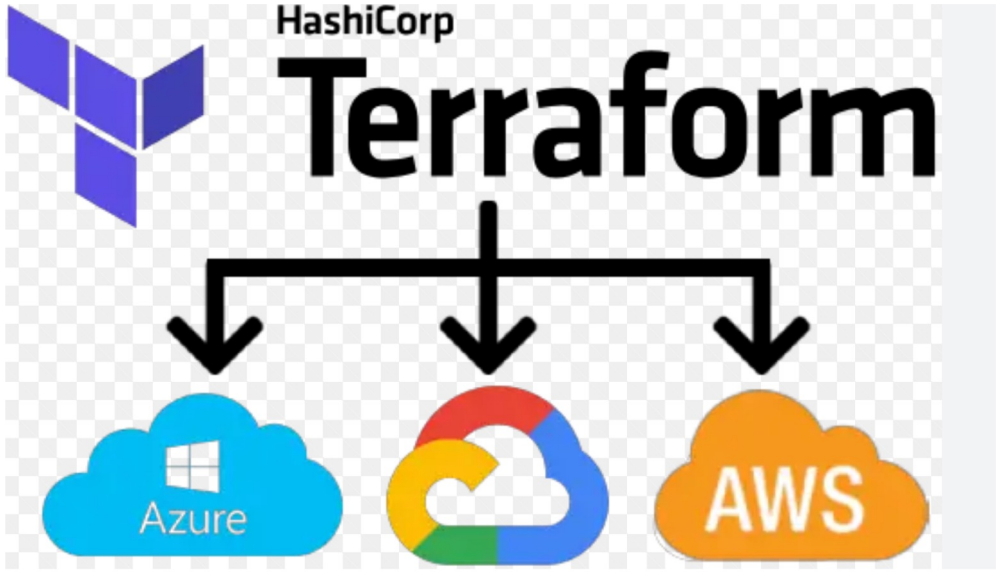
Source: docs.nvidia.com

Multi-cloud Deployment Management

Managing infrastructure and services across diverse cloud platforms

Consistently deploy across multiple clouds

- Modular
- Composable, and
- Flexible



Kitware

Containers and managed environments

- Managed container environments
- Container services
  - Amazon ECS
  - Azure Container Apps, and
  - Google Kubernetes Engine.
- Managed environments
  - CPU
  - GPU
  - Memory

AWS Fargate

Azure Container Apps

Kitware

Thank you!