AWS re:Invent

AIM329

Using deep learning to track wildfires and air quality

Sanjay Padhi
Head of AWS Research, US Education
Amazon Web Services

Feng Yan
Assistant Professor
University of Nevada, Reno

AWS re:Invent

aws

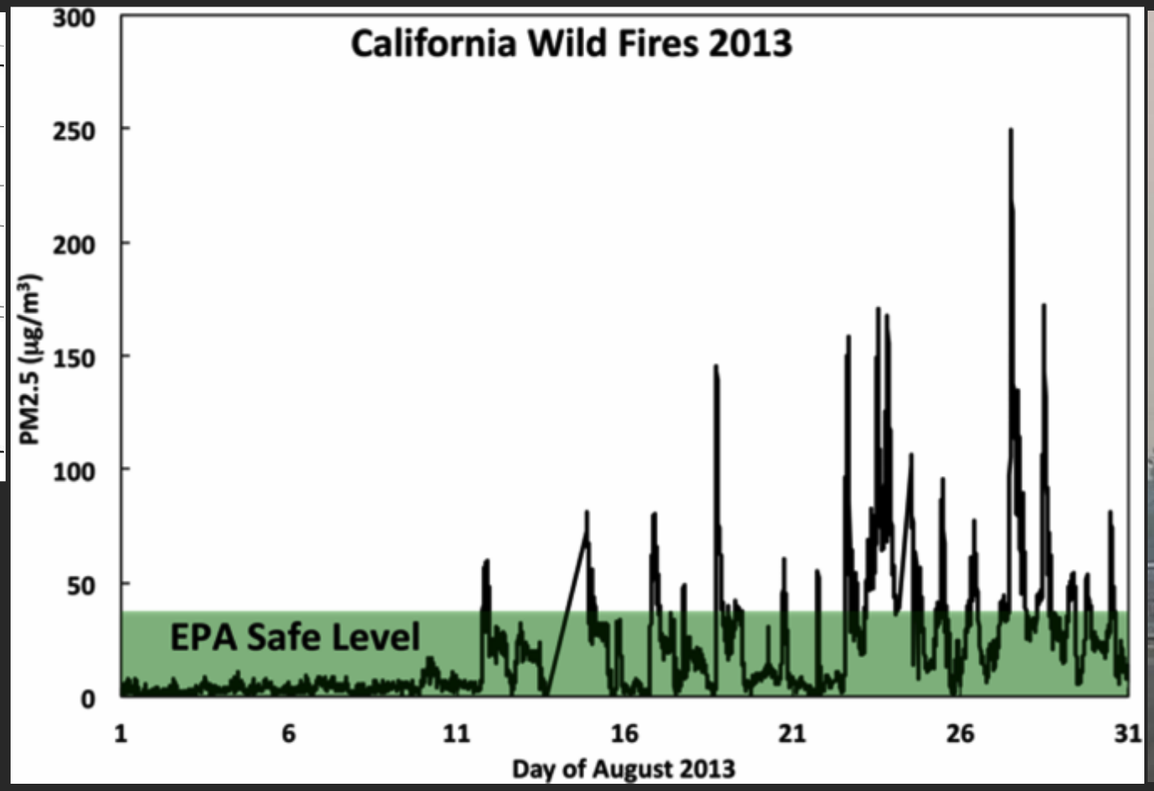
NSF BIGDATA: IA: Collaborative Research: Protecting Yourself from Wildfire Smoke: Big Data Driven Adaptive Air Quality Prediction Methodologies, 2019–2022

Feng Yan, Evgenia Smirni, Lei Yang, Heather Holmes,

Sponsored by NSF and AWS, supported by ALERTWildfire, UNR OIT

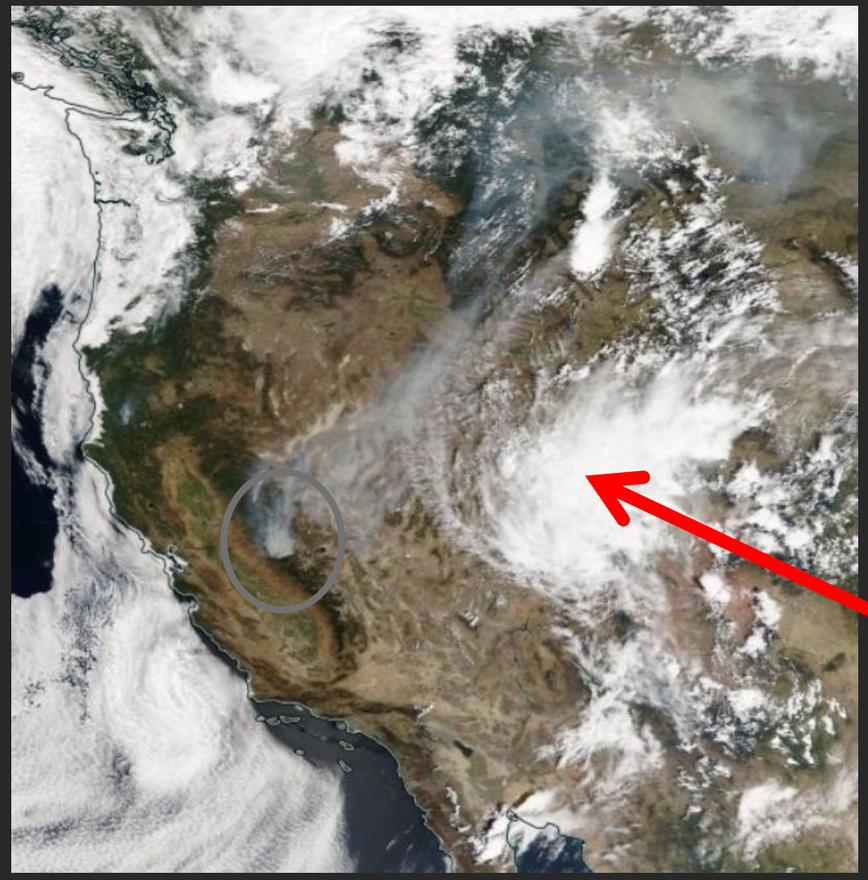
Why is wildfire smoke dangerous?

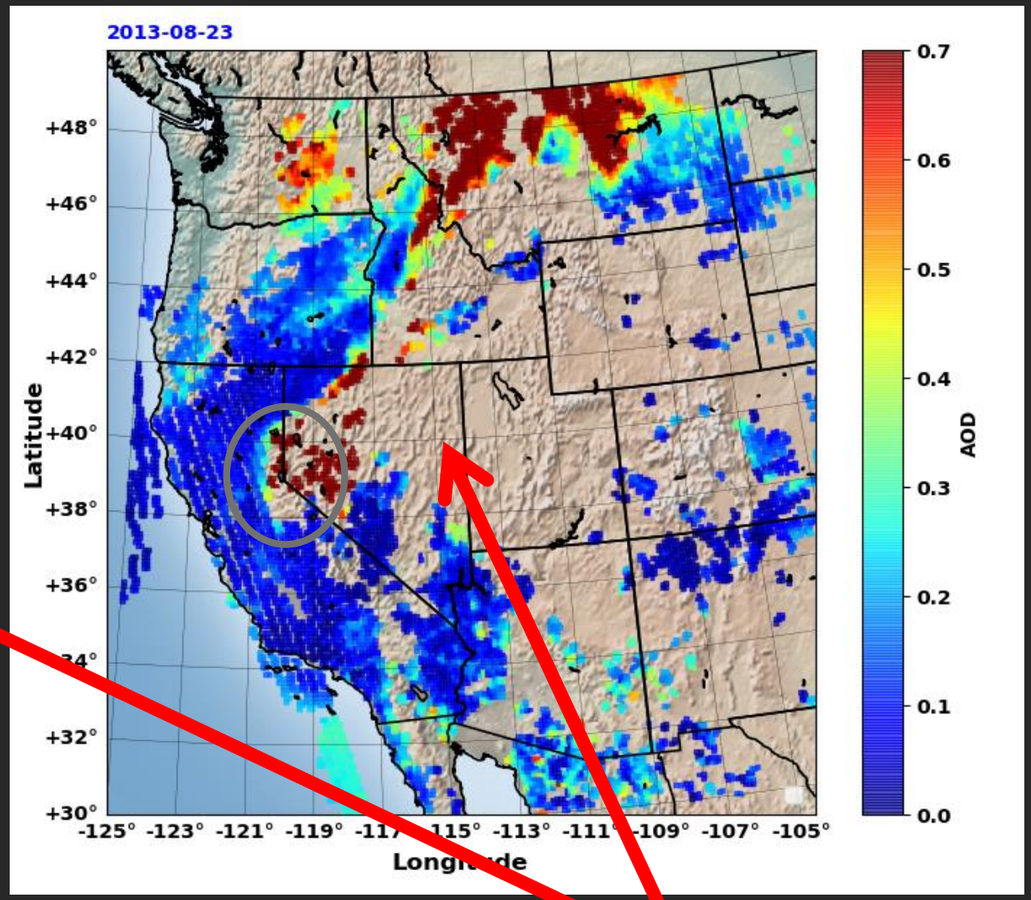
- Human health impacts result from wildfire smoke exposure
- Visibility and radiative forcing impacts for climate
- Increasing drought conditions in western US can lead to more fires



Fire/smoke and clouds: Missing data

August 23, 2013, Yosemite Rim fire in California





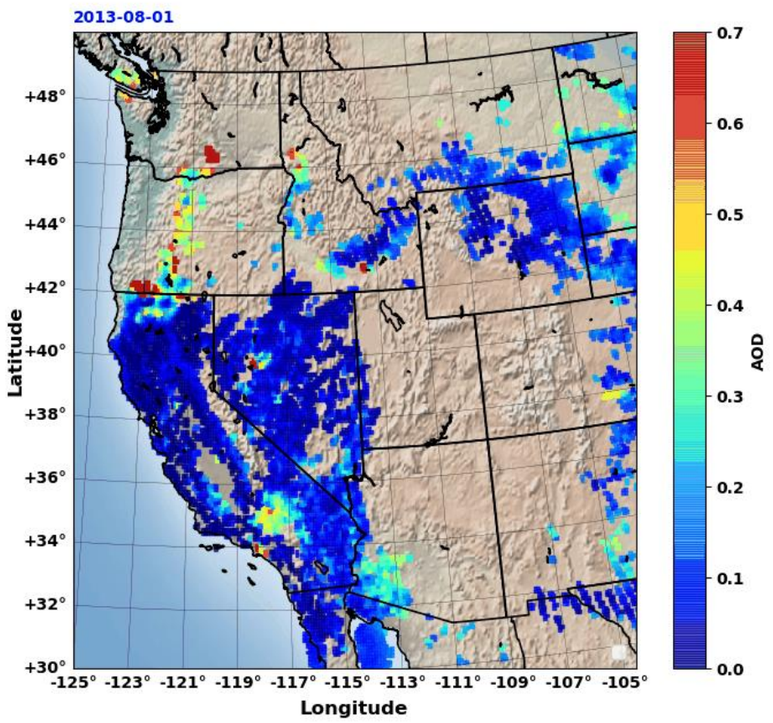
Satellites cannot measure the air pollution below clouds

Missing predictions at a fine-grained level

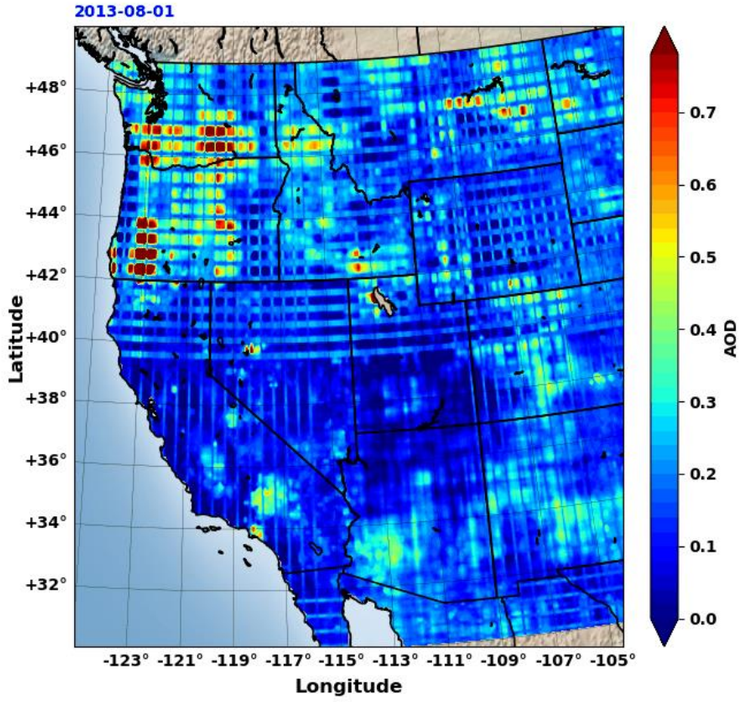
MODIS satellite aerosol optical depth

- Example California fire in August 2013
- Daily, 10-km resolution, air pollution estimates from satellites
- Use machine learning to recover missing satellite data

Daily MODIS Aqua



Daily Gap Filled



Smoke spreading quickly

Reno smoke on August 18, 2013

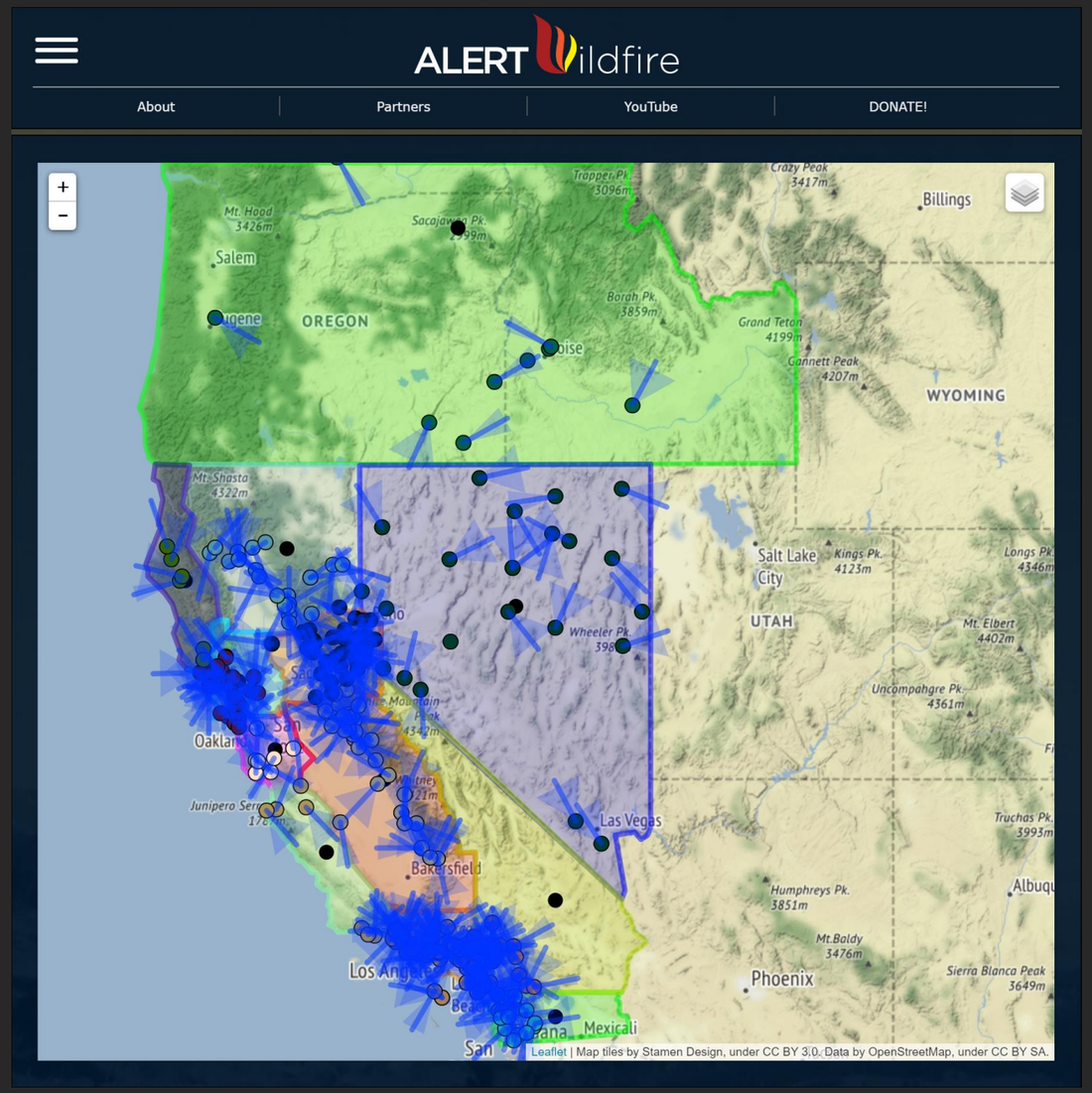
The smoke was from the American River fire near Sacramento, CA

Wildfire smoke can travel very fast!

Real-time, fine-grained monitoring and prediction

This time-lapse video is captured by the DRI Western Regional Climate Center (provided by Dr. Patrick Arnott)

ALERTWildfire camera network



ALERTWildfire is a fire camera network and associated tools

Use pan-tilt-zoom (PTZ) cameras

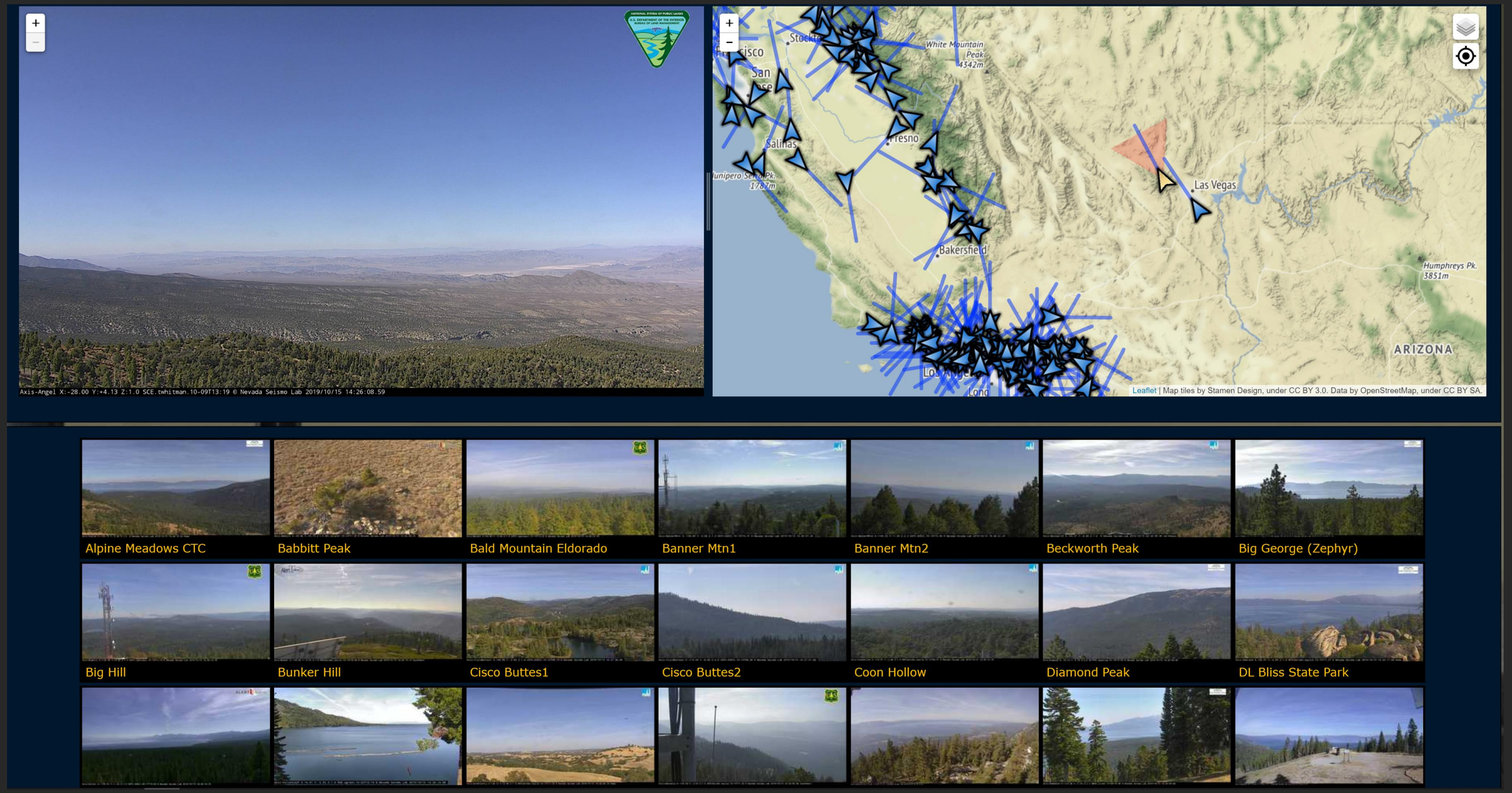
Help firefighters and first responders

Consortium of University of Nevada, Reno (UNR), University of California, San Diego (UCSD), and University of Oregon (UO)

NEVADA SEISMOLOGICAL LABORATORY

http://www.alertwildfire.org/

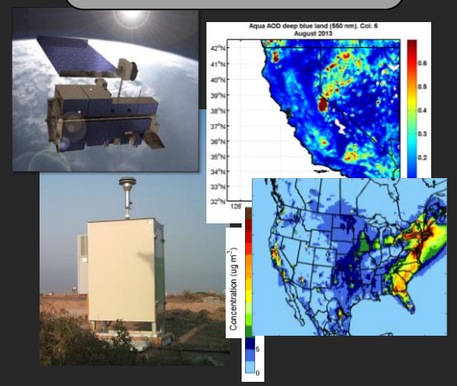
ALERTWildfire camera network



Architecture

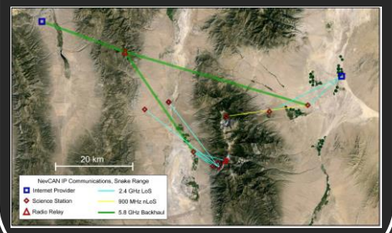
Big data

Existing data sources

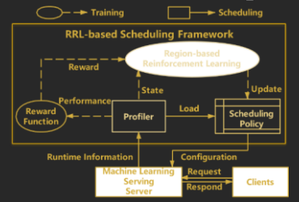


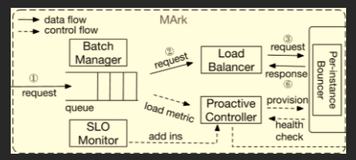
New camera network data

Edge computing & data transfer



Machine learning as a service (MLaaS)





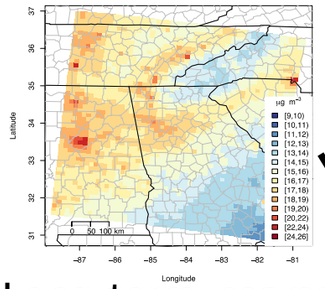
Latency, scaling, efficiency

Amazon SageMaker

Amazon EC2

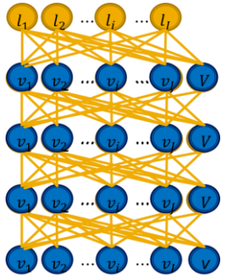
AWS Lambda

Prediction methodology



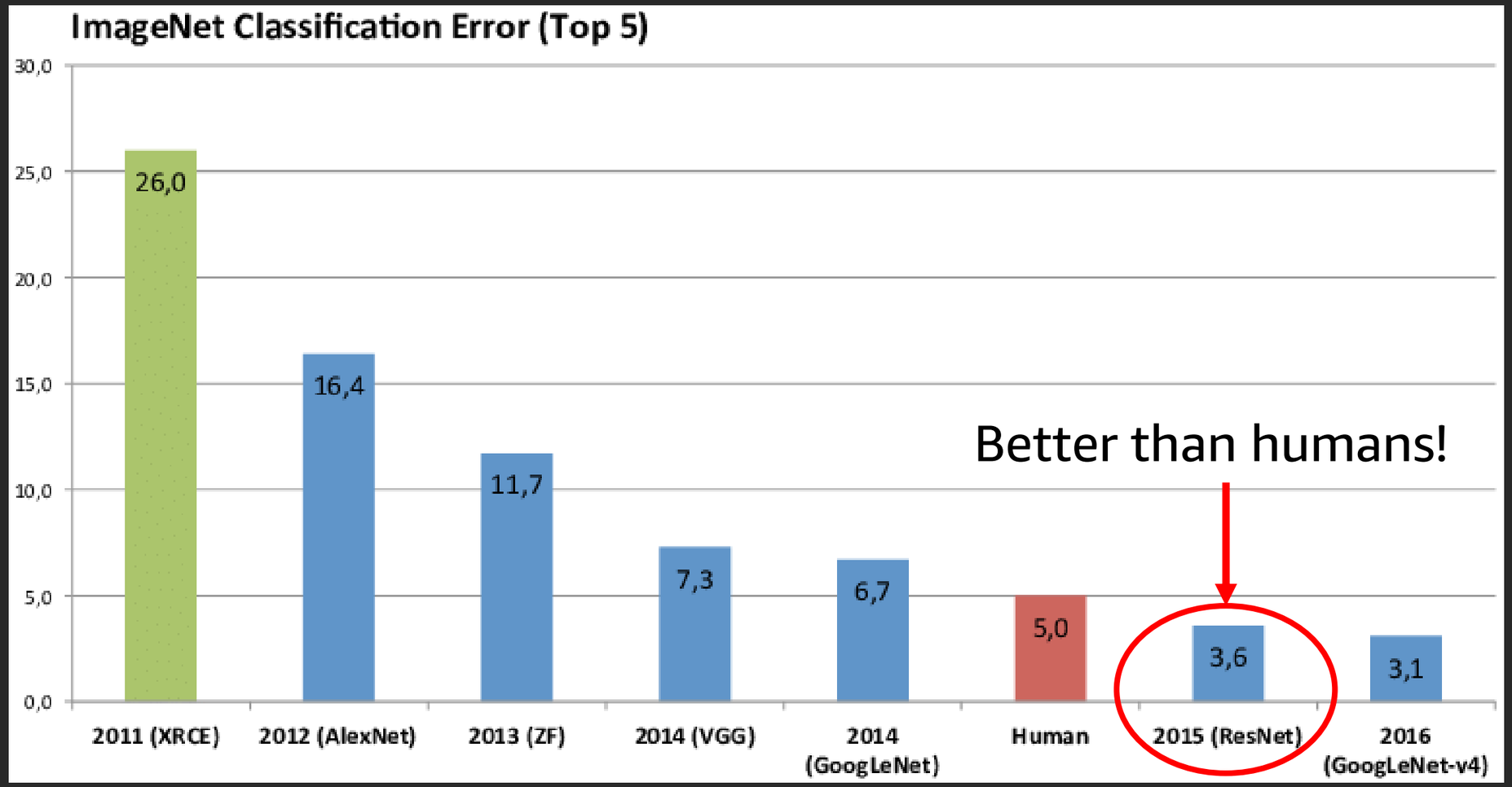
Long-term, coarse-grained model

Real-time, fine-grained model



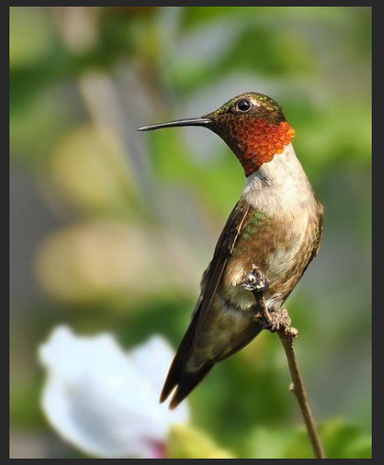
Deep learning

Deep neural network for image classification



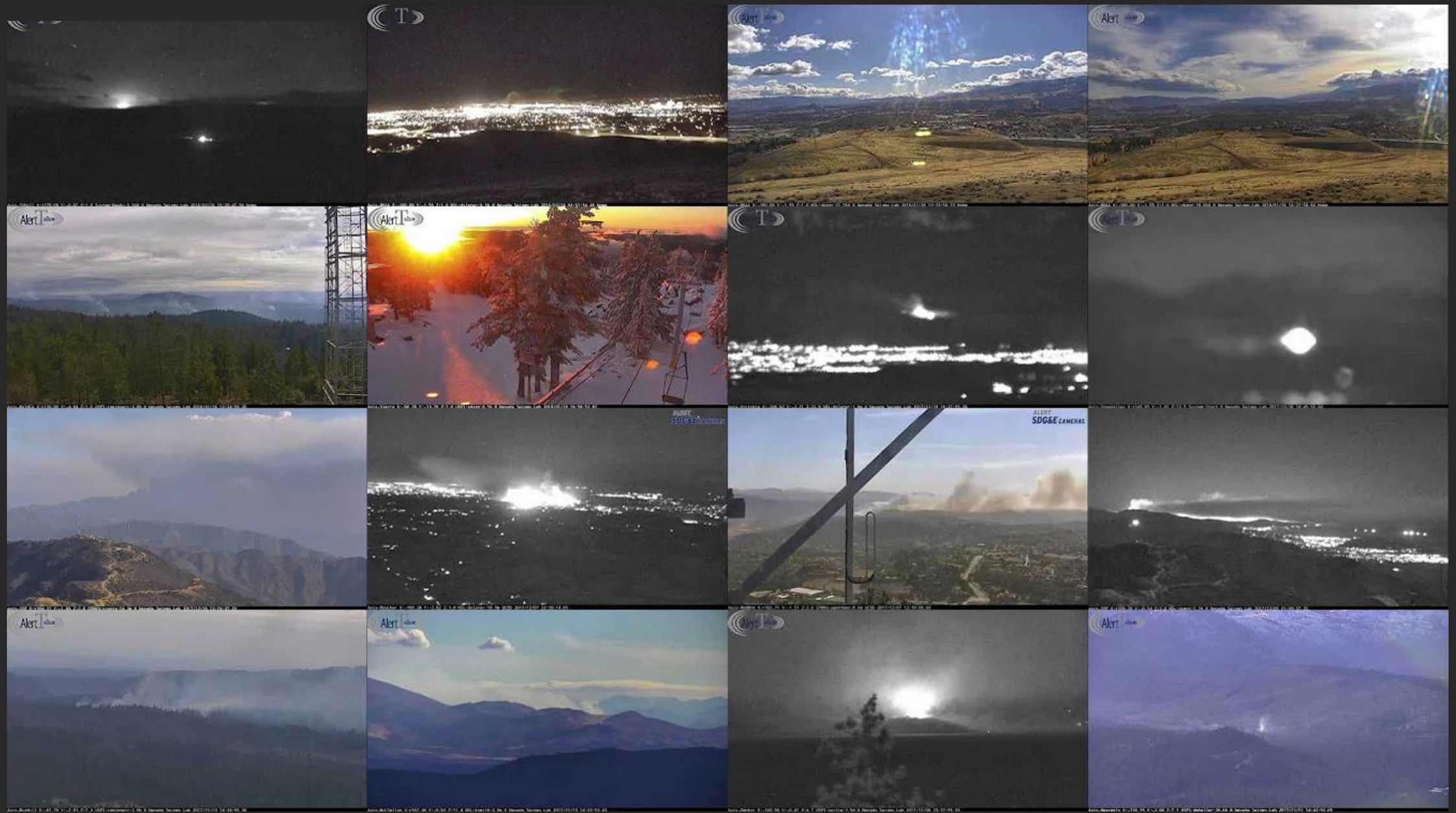
Modified from Hung-yi Lee's introduction v8. Source of image: https://www.researchgate.net/figure/Winner-results-of-the-ImageNet-large-scale-visual-recognition-challenge-LSVRC-of-the_fig7_324476862

Data used in machine learning benchmarks



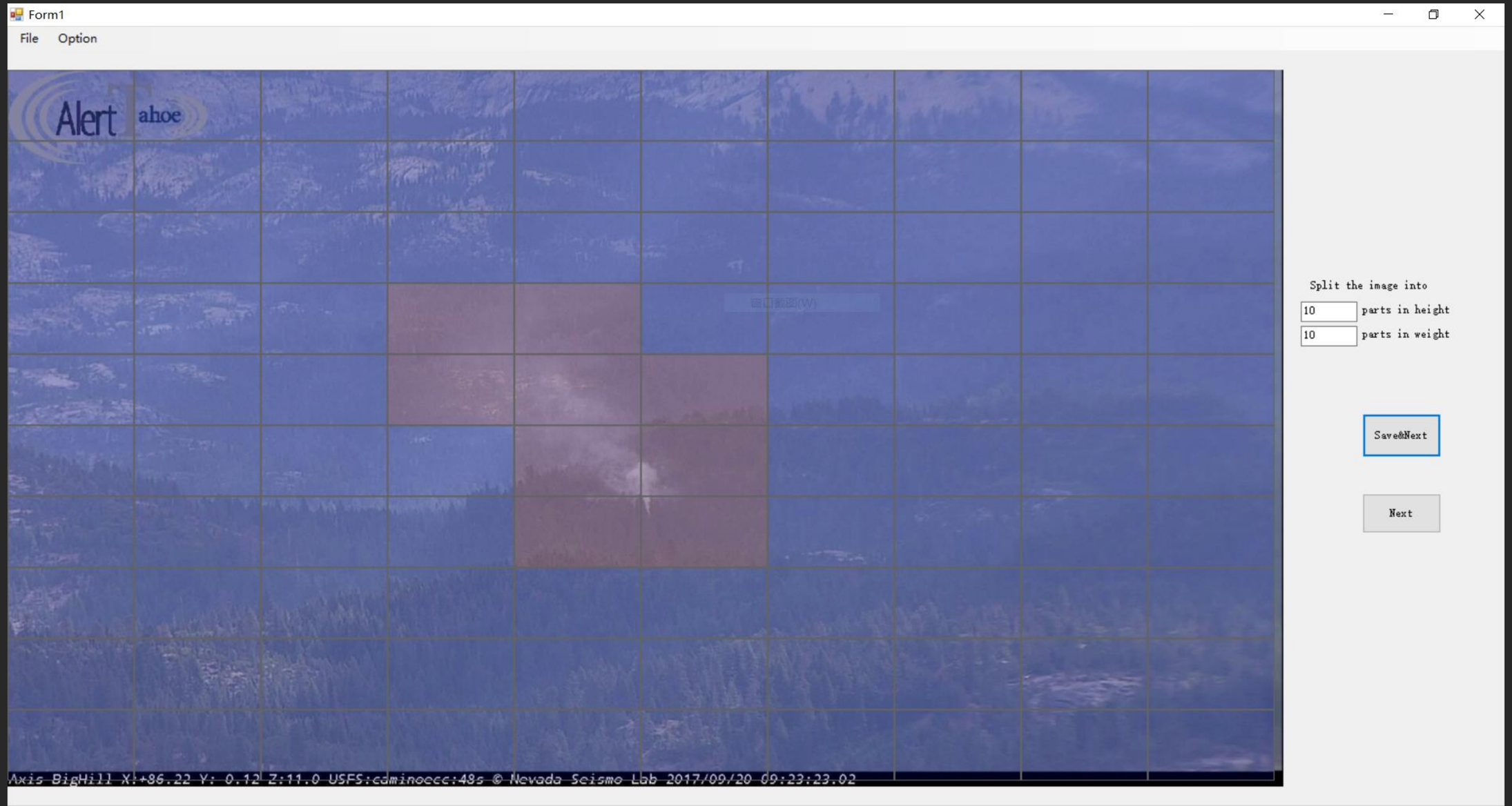
Pictures from MNIST and ILSVRC2012 datasets

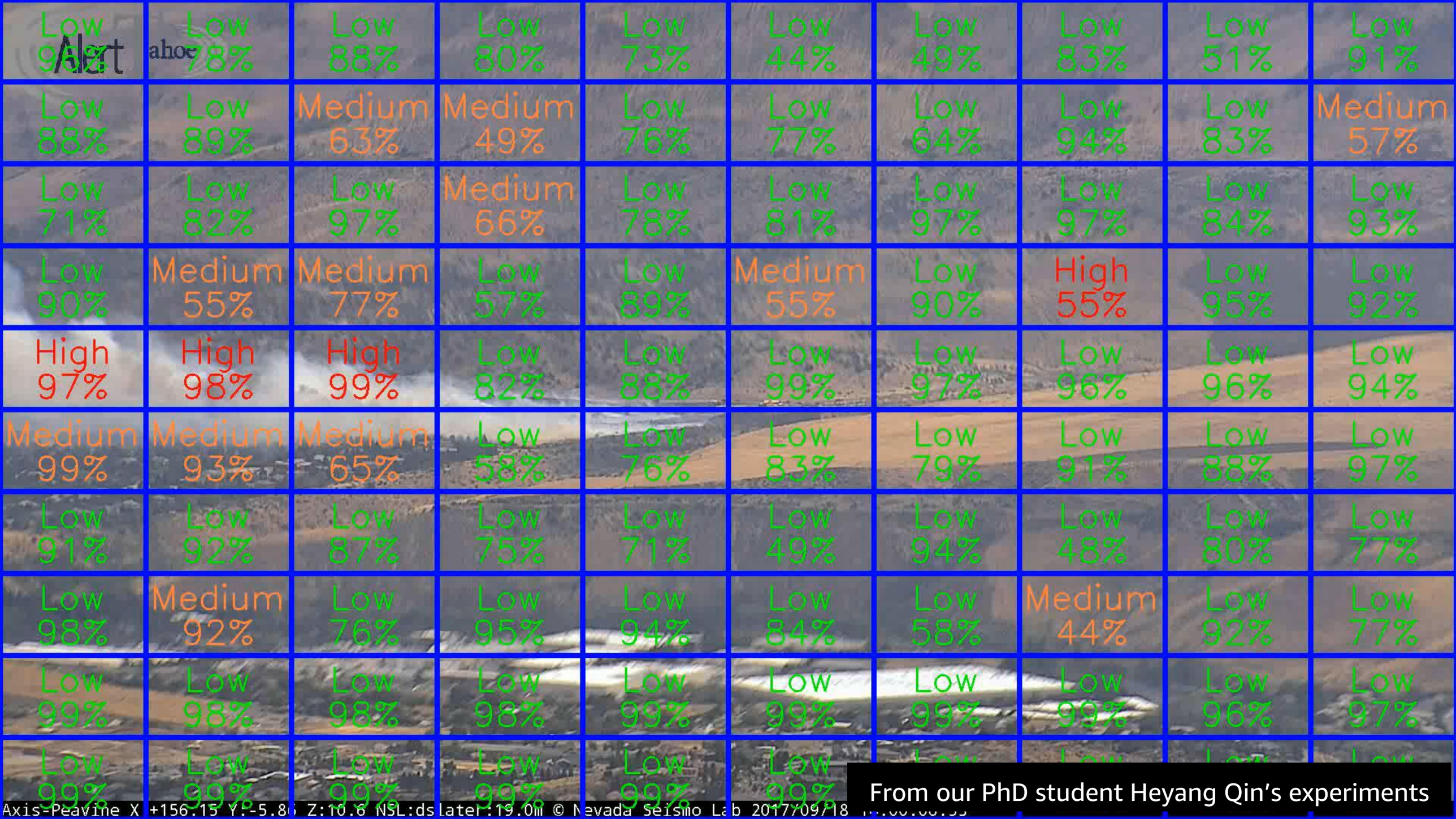
The camera data we need to deal with

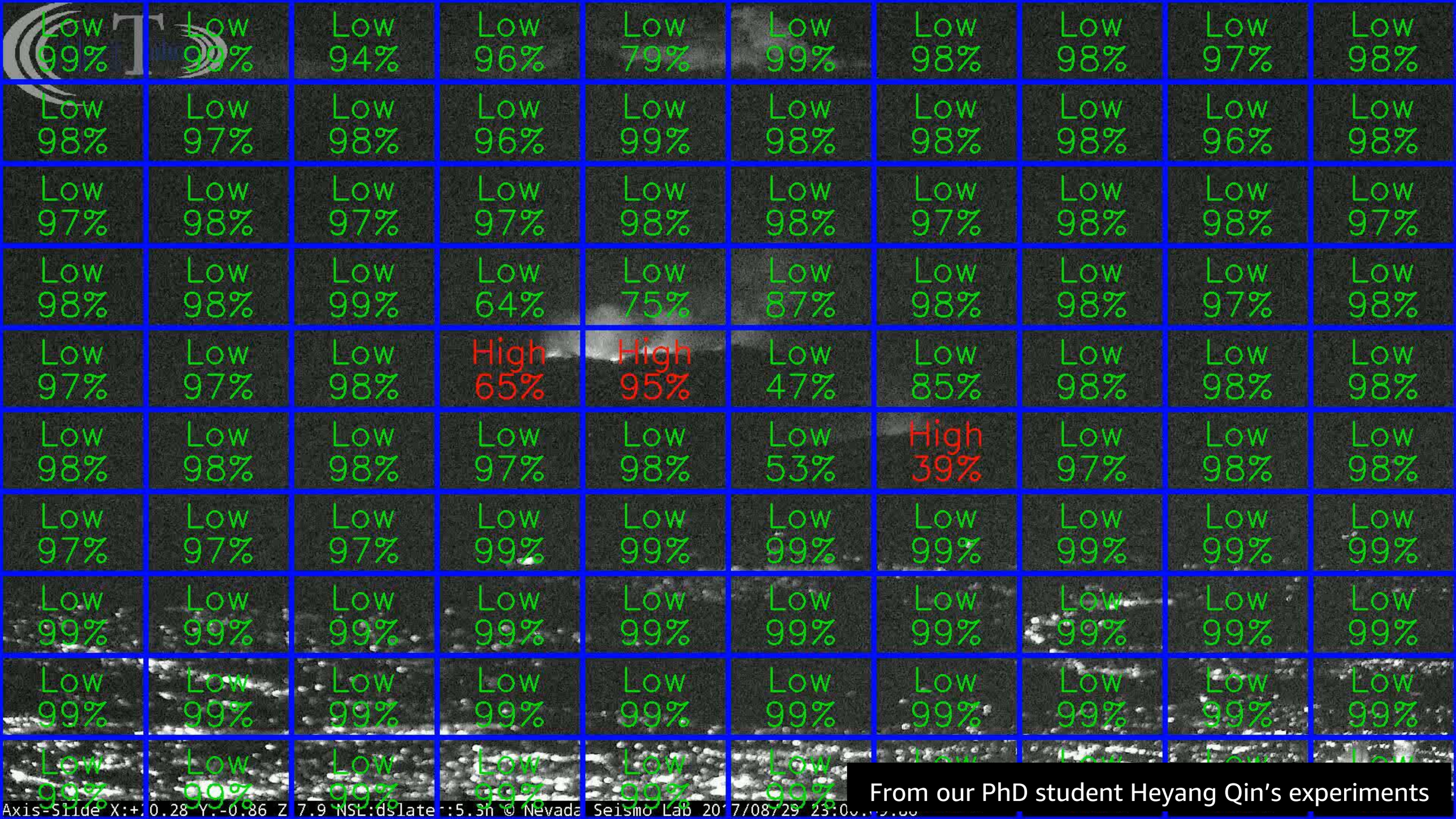


Many different shapes, changing all the time

Identify the smoke regions on images and add labels

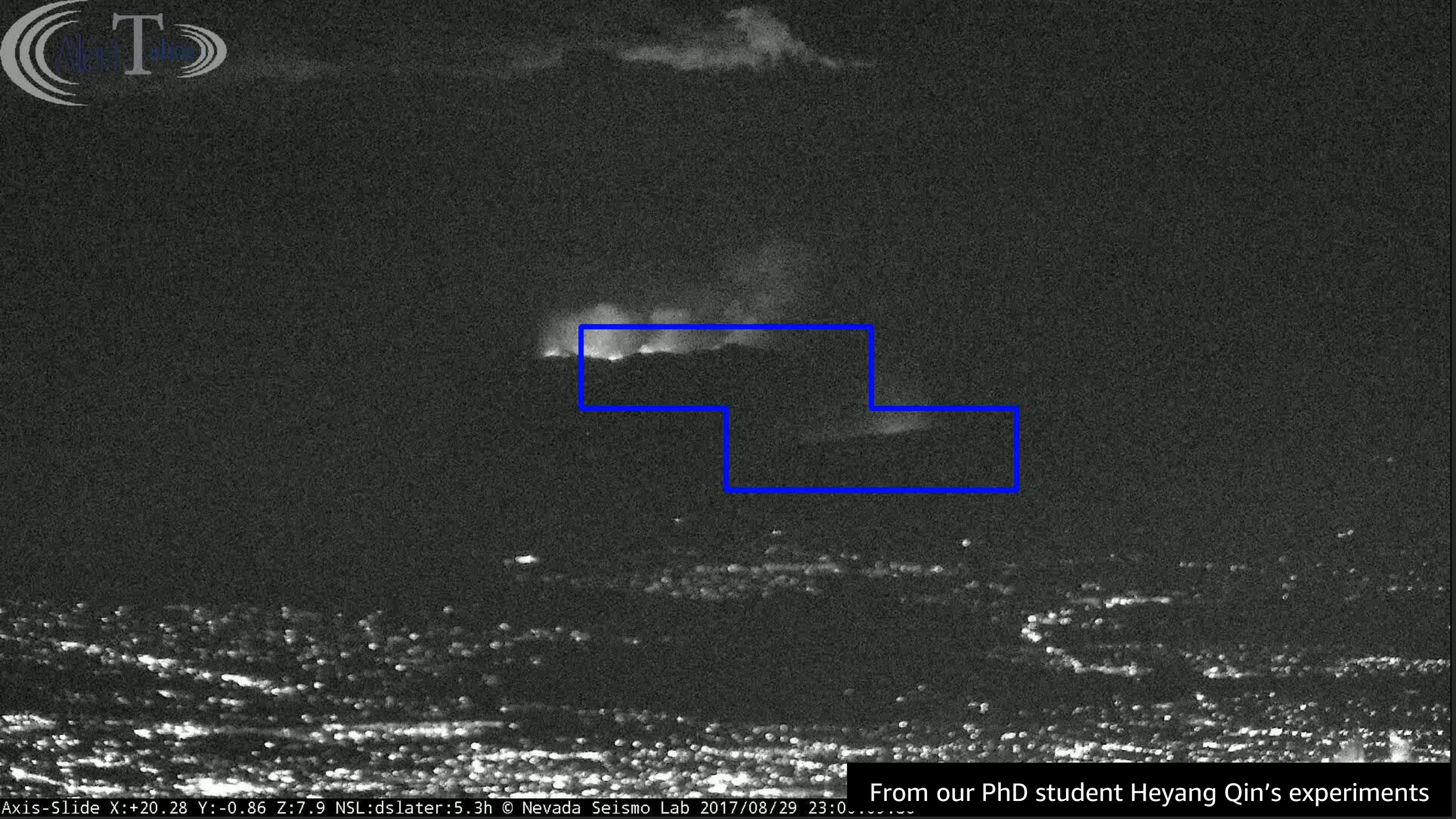



Low 98%
Low 78%
Low 88%
Low 80%
Low 73%
Low 44%
Low 49%
Low 83%
Low 51%
Low 91%
Low 88%
Low 89%
Medium 63%
Medium 49%
Low 76%
Low 77%
Low 64%
Low 94%
Low 83%
Medium 57%
Low 71%
Low 82%
Low 97%
Medium 66%
Low 78%
Low 81%
Low 97%
Low 97%
Low 84%
Low 93%
Low 90%
Medium 55%
Medium 77%
Low 57%
Low 89%
Medium 55%
Low 90%
High 55%
Low 95%
Low 92%
High 97%
High 98%
High 99%
Low 82%
Low 88%
Low 99%
Low 97%
Low 96%
Low 96%
Low 94%
Medium 99%
Medium 93%
Medium 65%
Low 58%
Low 76%
Low 83%
Low 79%
Low 91%
Low 88%
Low 97%
Low 91%
Low 92%
Low 87%
Low 75%
Low 71%
Low 49%
Low 94%
Low 48%
Low 80%
Low 77%
Low 98%
Medium 92%
Low 76%
Low 95%
Low 94%
Low 84%
Low 58%
Medium 44%
Low 92%
Low 77%
Low 99%
Low 98%
Low 98%
Low 98%
Low 99%
Low 99%
Low 99%
Low 96%
Low 97%
Low 99%
Low 99%
Low 99%
Low 99%
Low 99%
Low 99%
From our PhD student Heyang Qin’s experiments

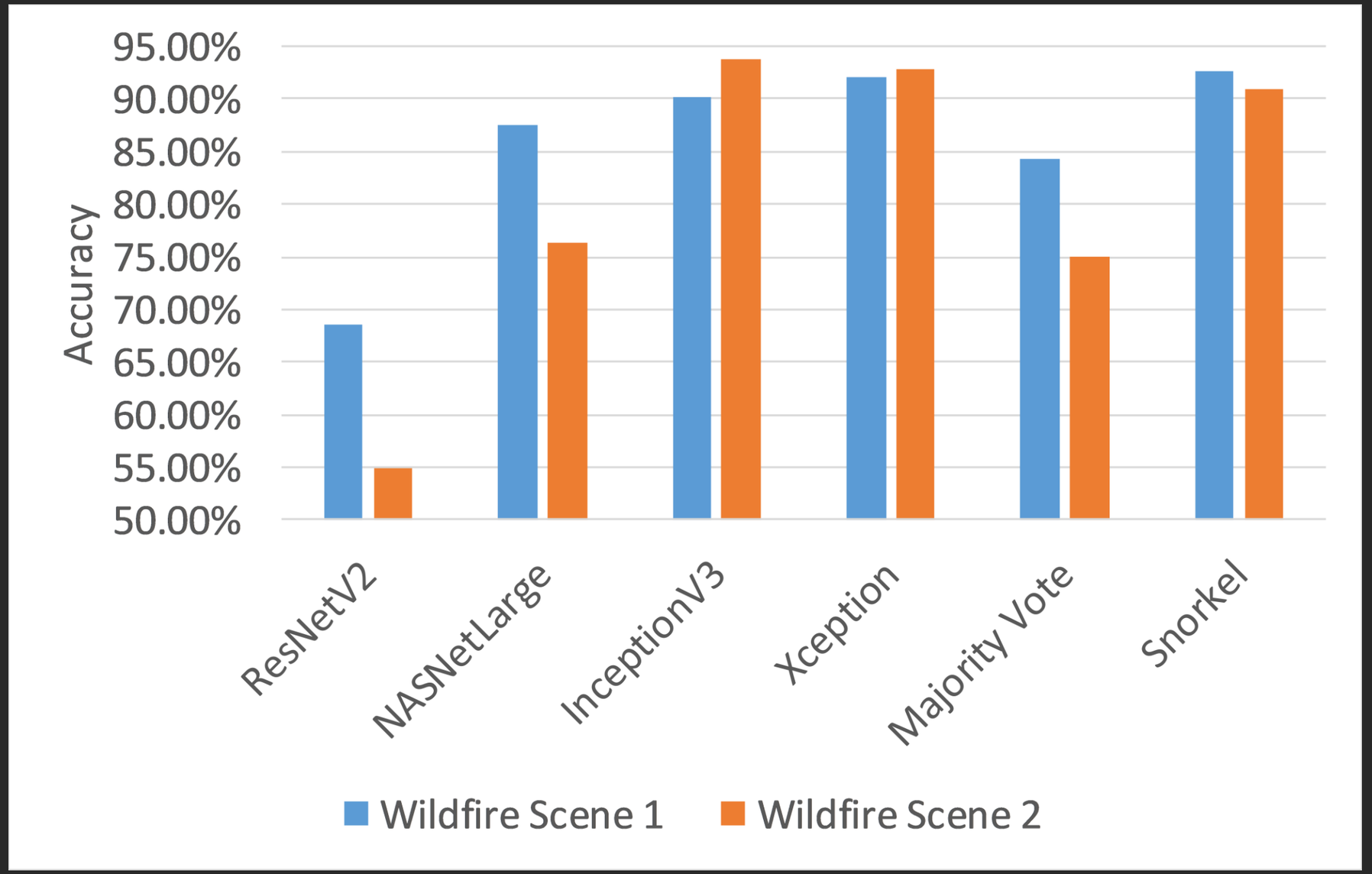

Low 99% Low 99% Low 94% Low 96% Low 79% Low 99% Low 98% Low 98% Low 97% Low 98%
Low 98% Low 97% Low 98% Low 96% Low 99% Low 98% Low 98% Low 98% Low 96% Low 98%
Low 97% Low 98% Low 97% Low 97% Low 98% Low 98% Low 97% Low 98% Low 98% Low 97%
Low 98% Low 98% Low 99% Low 64% Low 75% Low 87% Low 98% Low 98% Low 97% Low 98%
Low 97% Low 97% Low 98% High 65% High 95% Low 47% Low 85% Low 98% Low 98% Low 98%
Low 98% Low 98% Low 98% Low 97% Low 98% Low 53% High 39% Low 97% Low 98% Low 98%
Low 97% Low 97% Low 97% Low 99% Low 99% Low 99% Low 99% Low 99% Low 99% Low 99%
Low 99% Low 99% Low 99% Low 99% Low 99% Low 99% Low 99% Low 99% Low 99% Low 99%
Low 99% Low 99% Low 99% Low 99% Low 99% Low 99% Low 99% Low 99% Low 99% Low 99%
Low 99% Low 99% Low 99% Low 99% Low 99% Low 99%
From our PhD student Heyang Qin’s experiments

Use bounding box for detection

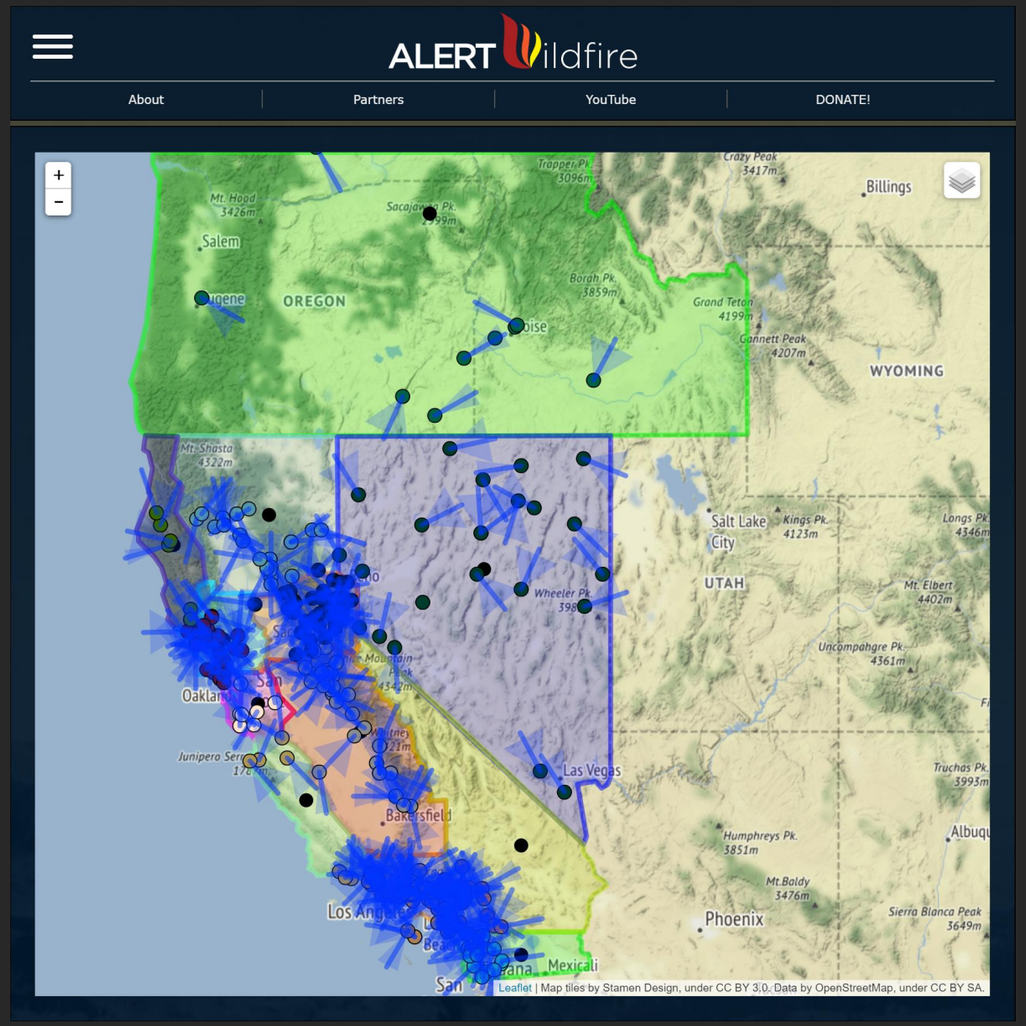

Alert
ahoe
Axis-Peavine X:+156.15 Y:-5.86 Z:10.6 NSL:dslater:19.0m © Nevada Seismo Lab 2017/09/18
From our PhD student Heyang Qin’s experiments


From our PhD student Heyang Qin’s experiments
Axis-Slide X:+20.28 Y:-0.86 Z:7.9 NSL:dslater:5.3h © Nevada Seismo Lab 2017/08/29 23:0

Prediction accuracy using different models



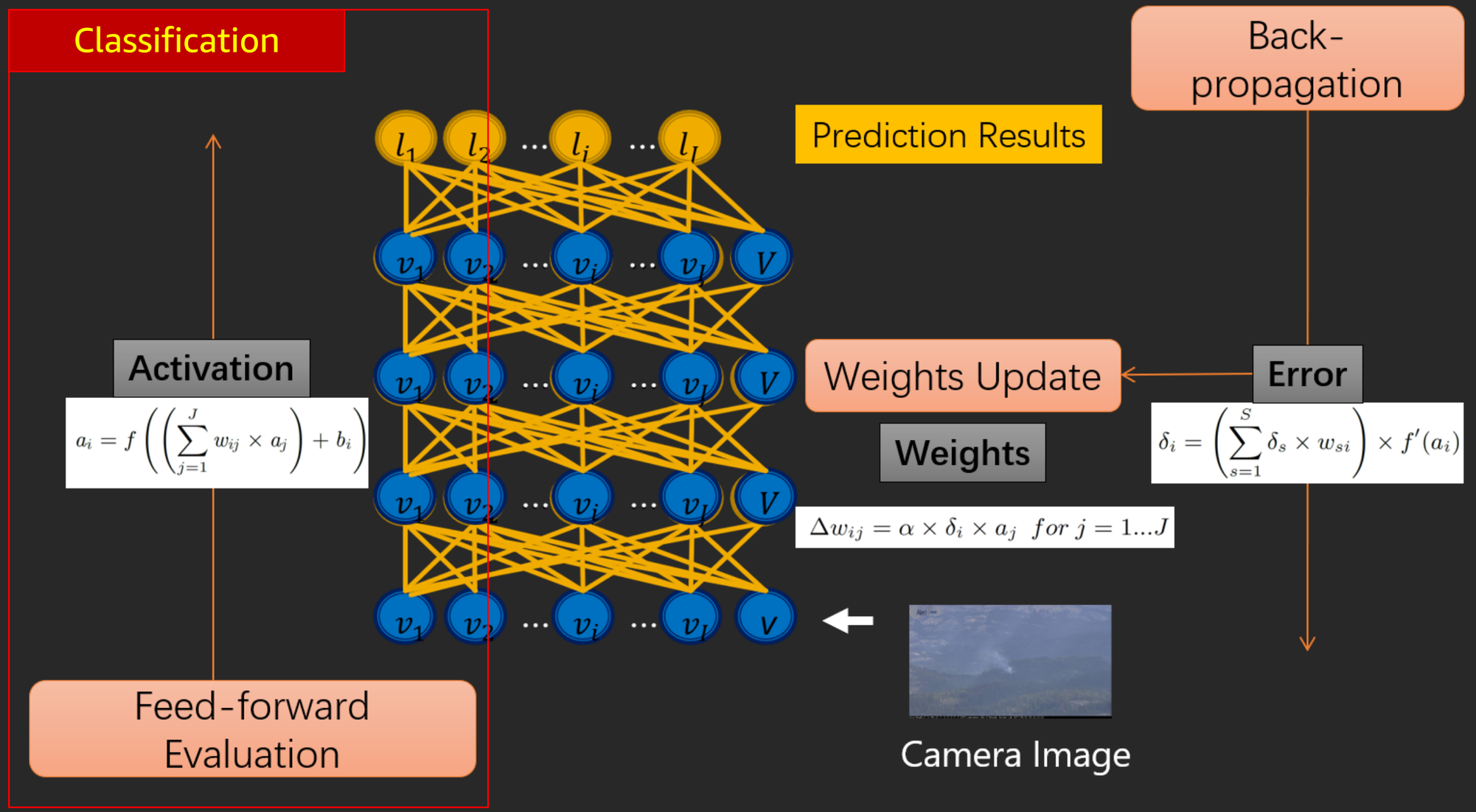
Real-time camera image classification



Many cameras dynamically join and leave during wildfire

How to provide scalable and real-time image/video classification?

Classification vs. training



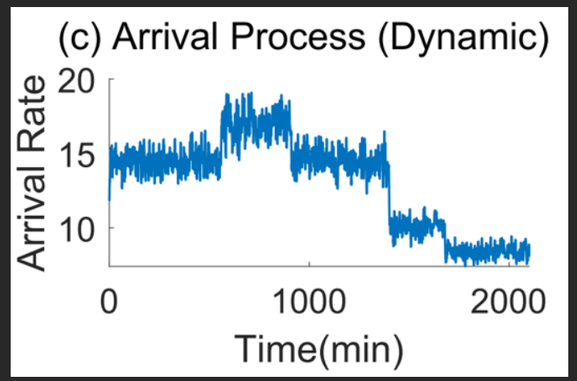
Use machine learning as a service (MLaaS)

Camera network

Smoke/wildfire detection

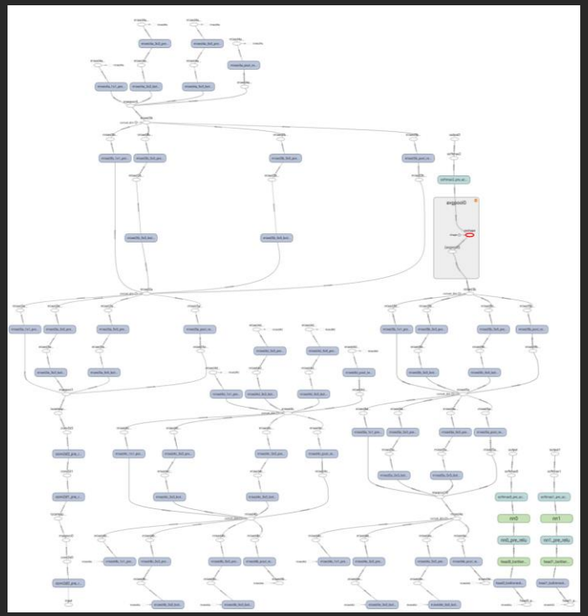
API

Machine learning model is a callable API

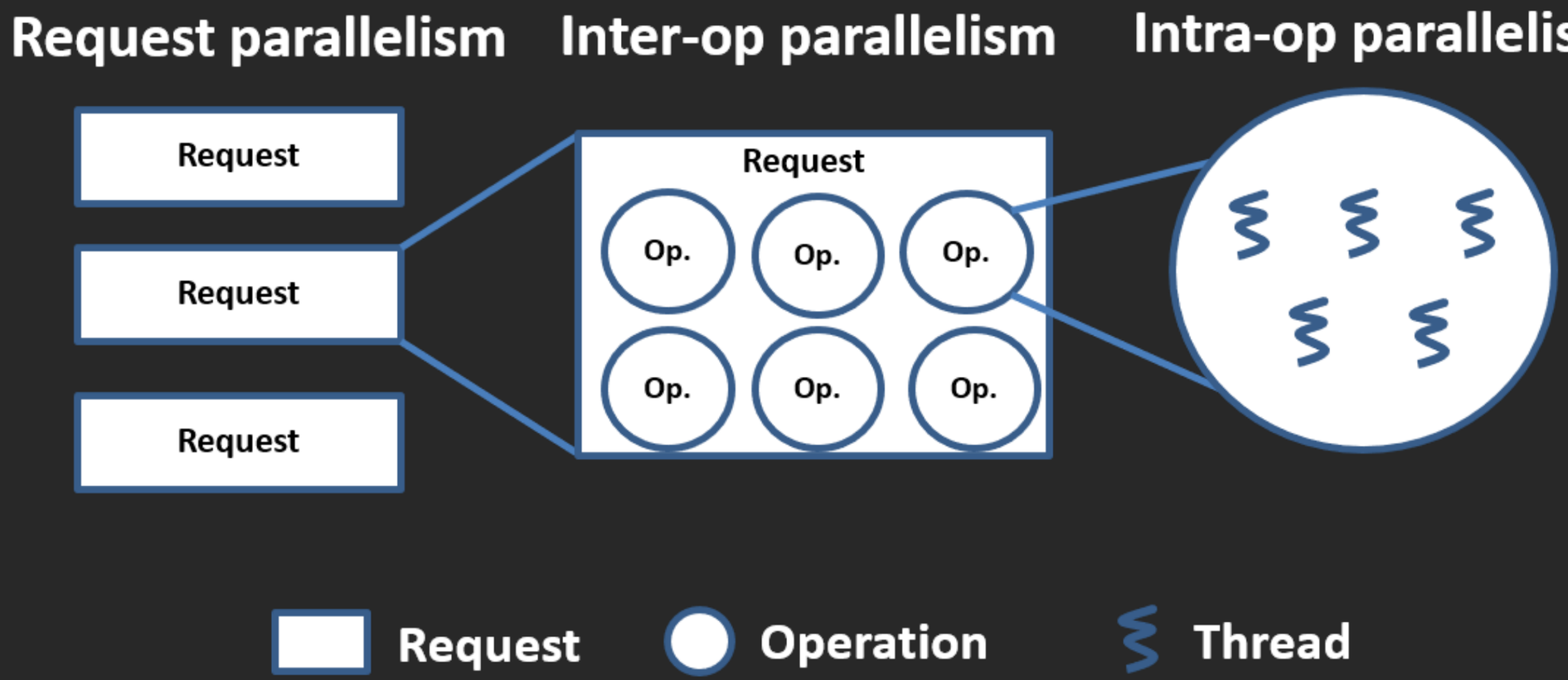


Two major challenges

1. Sequential execution is too slow
2. Dynamic classification workload



Use parallelism to speed up classification

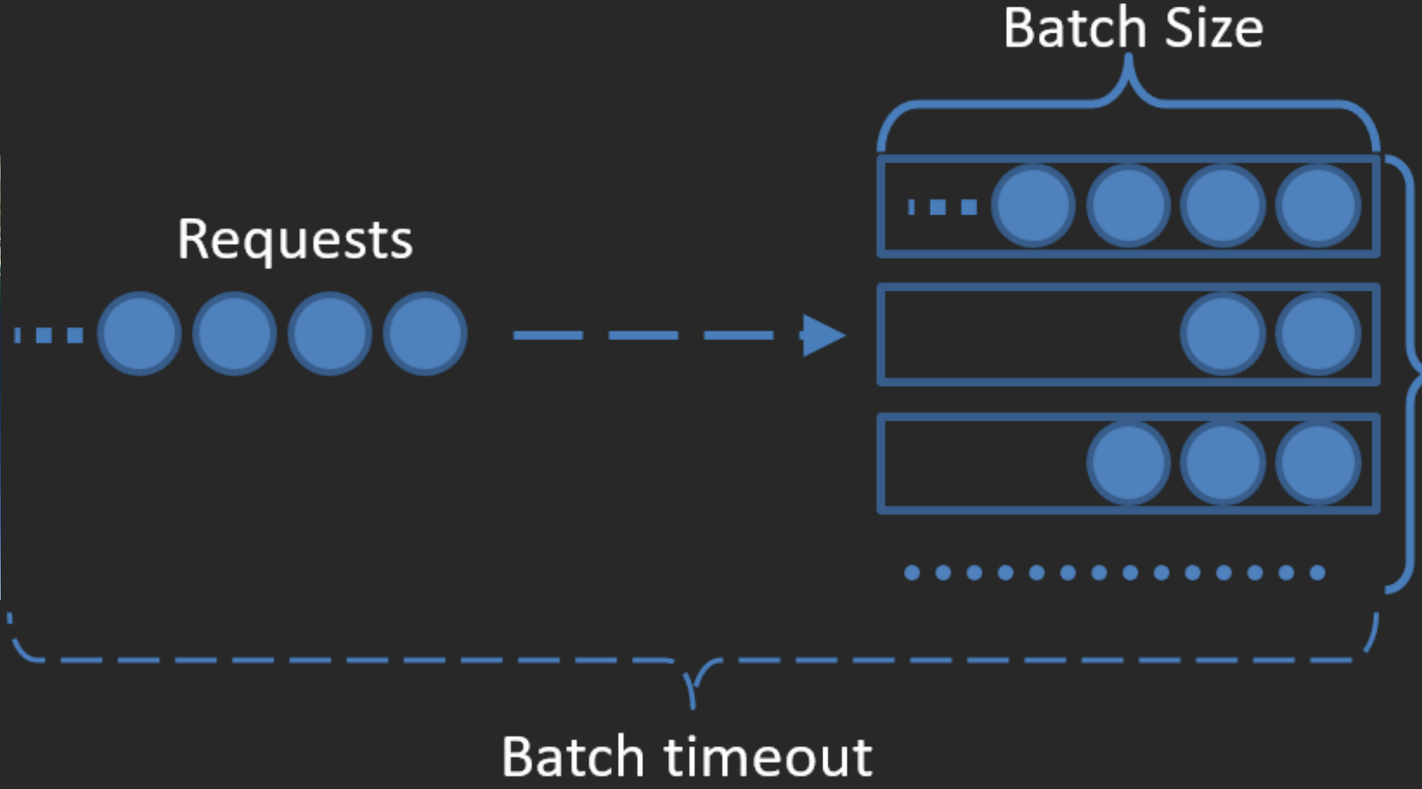


[CoRR '15] *MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems*

This work published at SC'19

Use batching to speed up classification

Camera network



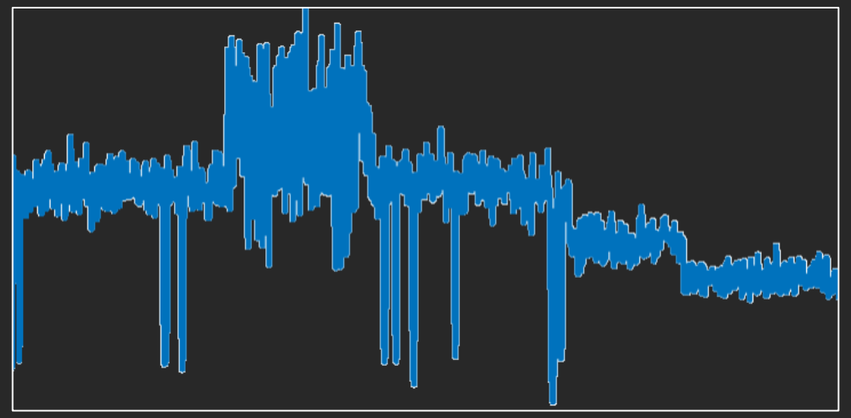
[CoRR '15] *MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems*

This work published at SC'19

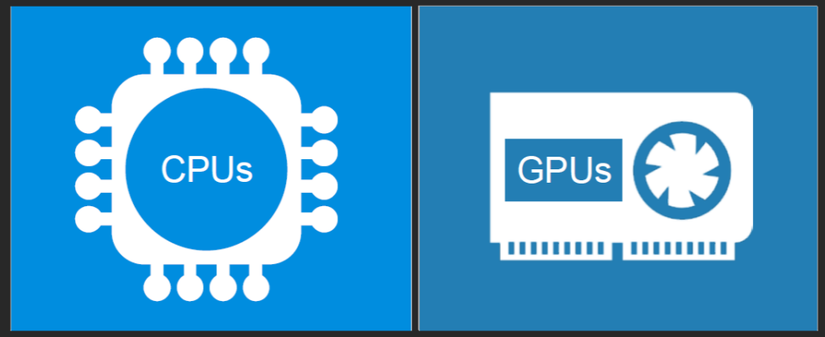
Problem

Camera network

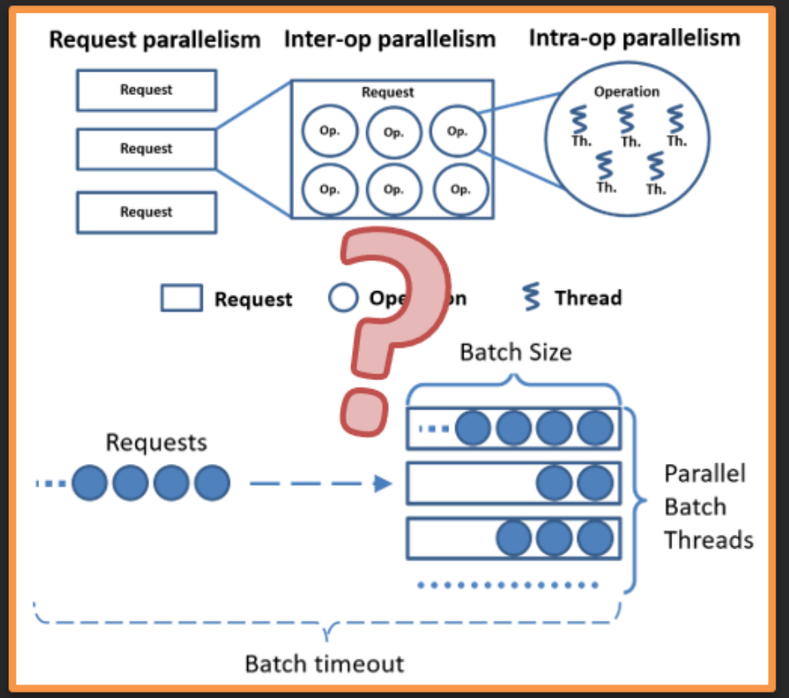
Classification requests form Dynamic Arrival Processes



Different infrastructures

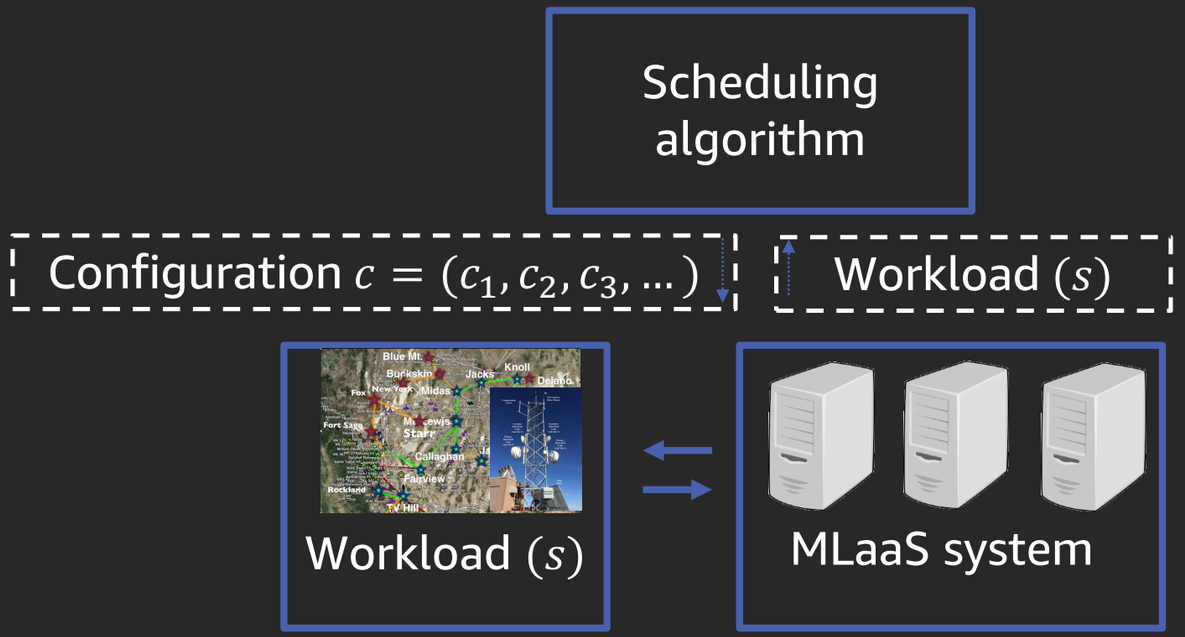


How to tune the system configurations?



This work published at SC'19

Problem definition



Minimize $r(s, c)$

Subject to $s \in \mathcal{S}, c \in \mathcal{C}$

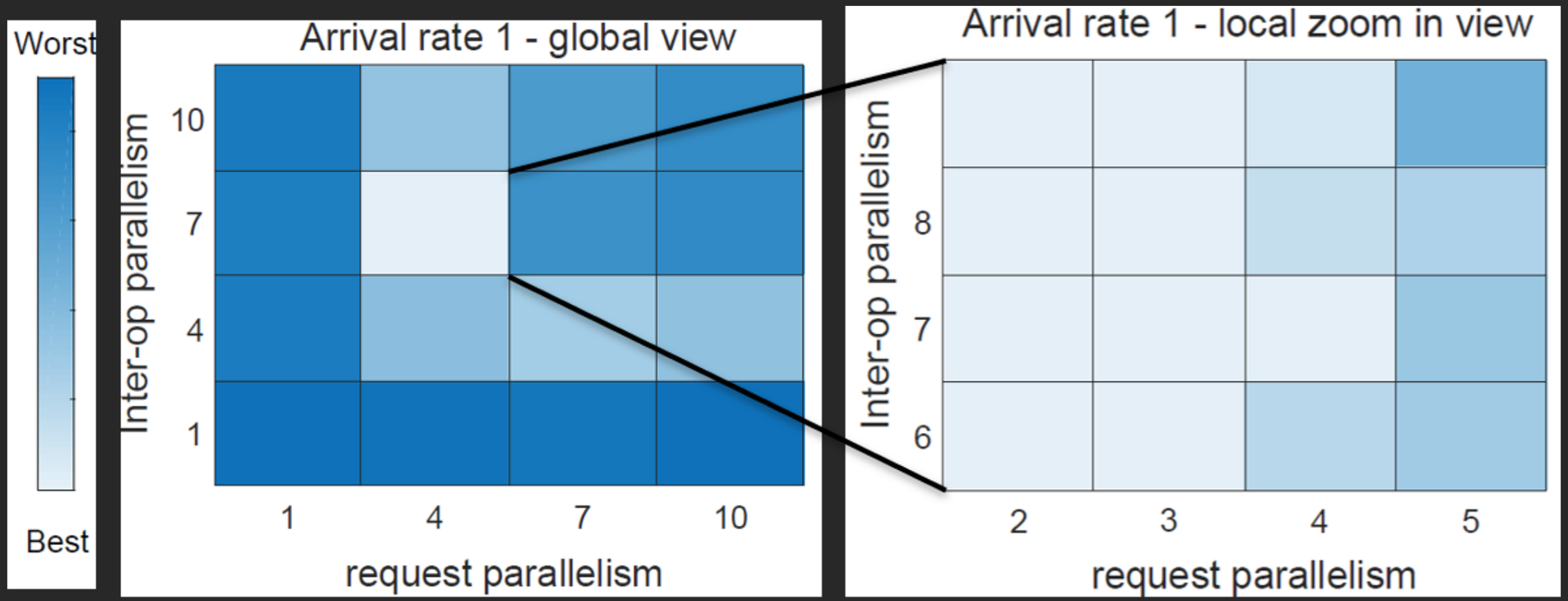
$r(s, c)$ - Average request latency

$\mathcal{S}$ - Possible workloads

$\mathcal{C}$ - Possible system configs

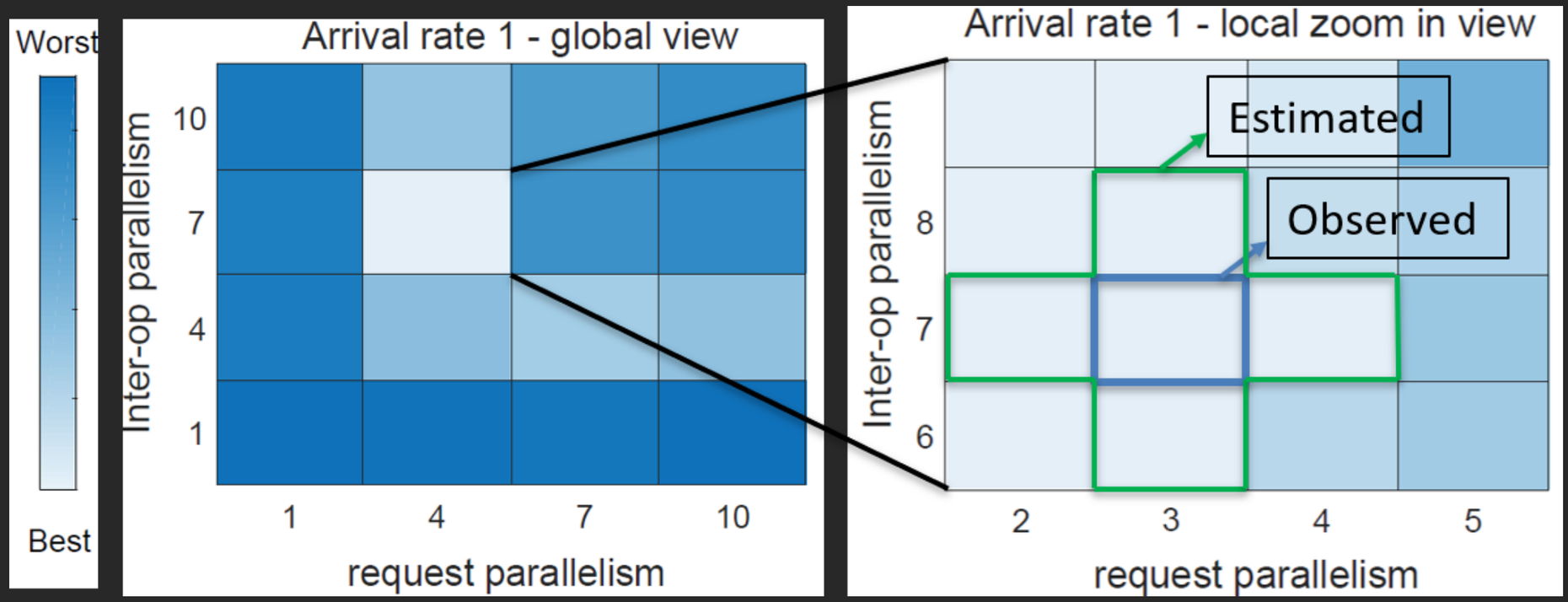
This work published at SC'19

Key observations



versatile globally vs smooth locally

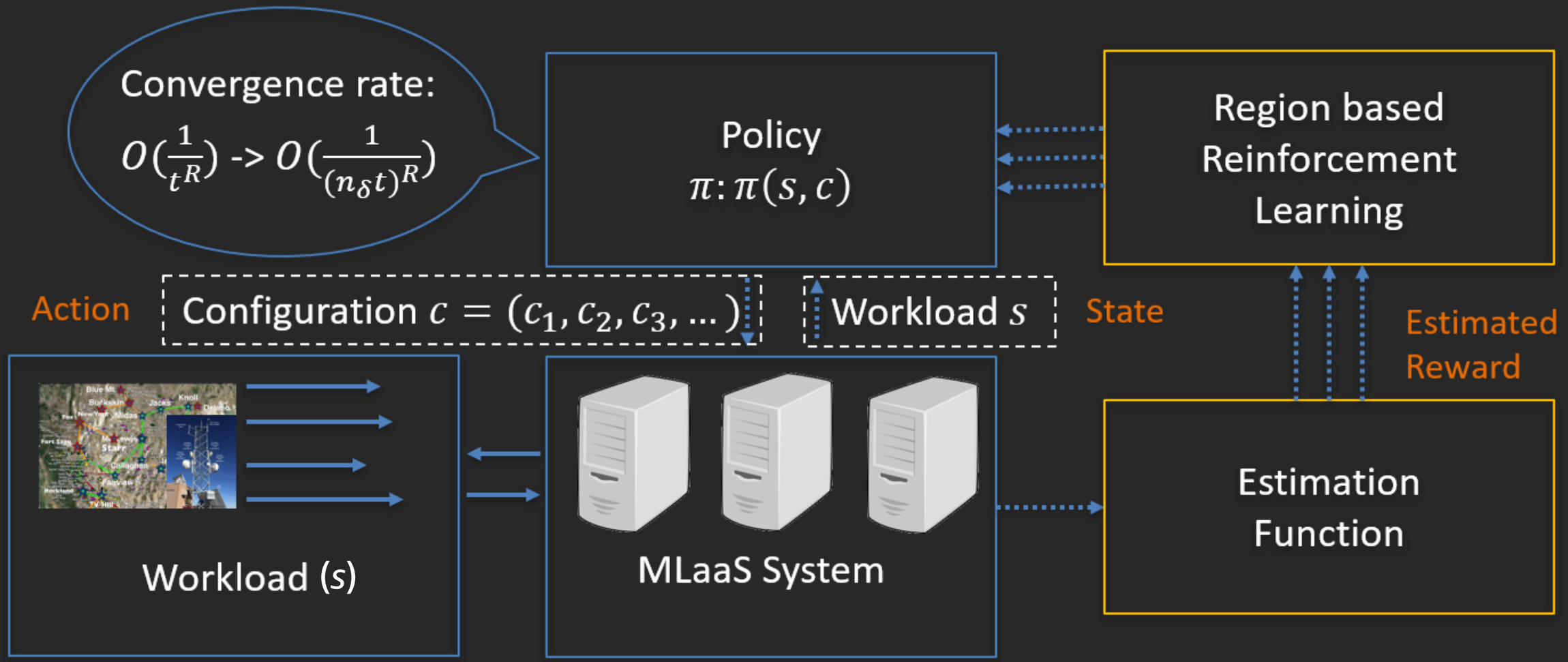
Key idea



Use a single observation to update the configurations in a nearby region

This work published at SC'19

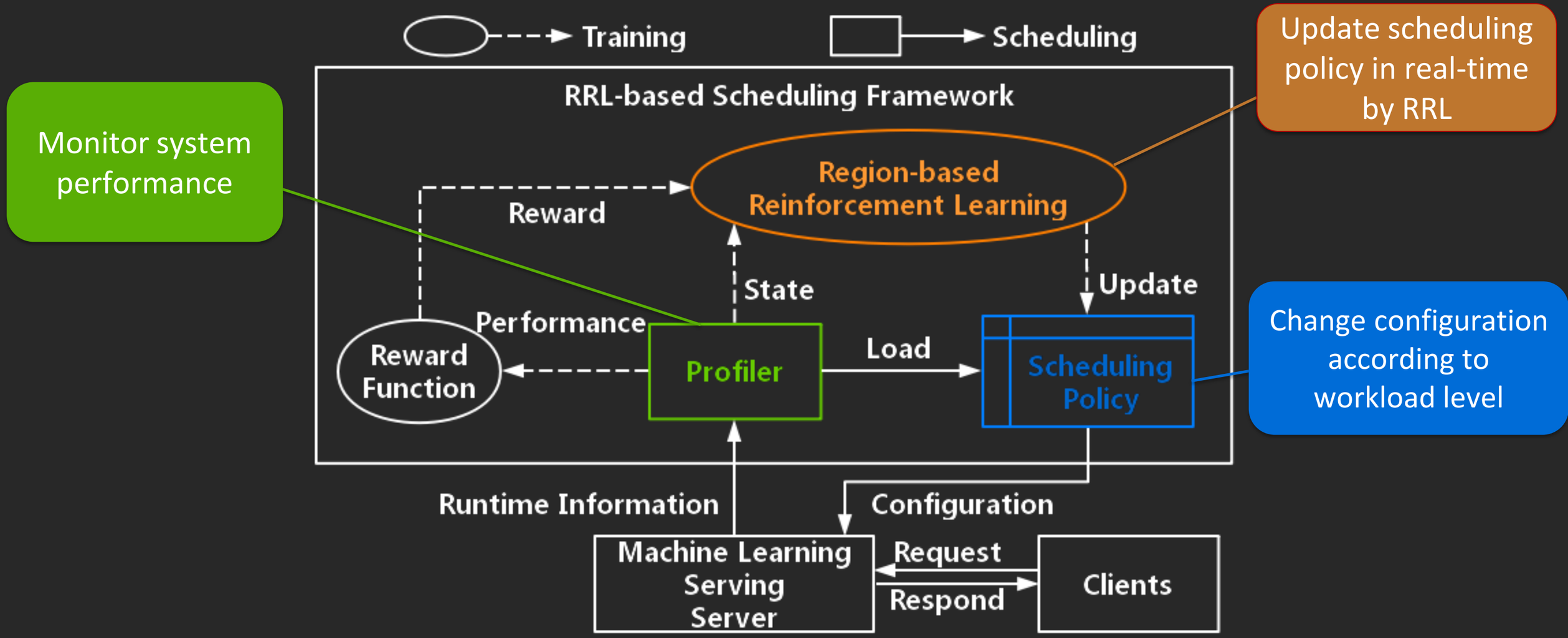
Region based Reinforcement Learning (RRL)



Swift Machine Learning Model Serving Scheduling: A Region Based Reinforcement Learning Approach
https://dl.acm.org/citation.cfm?id=3356164

This work published at SC'19

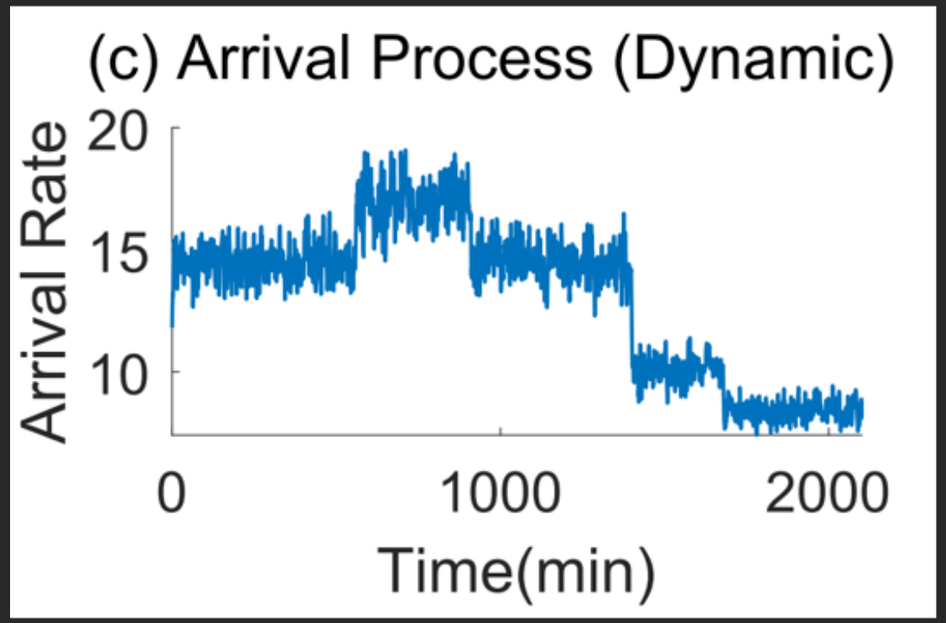
System design

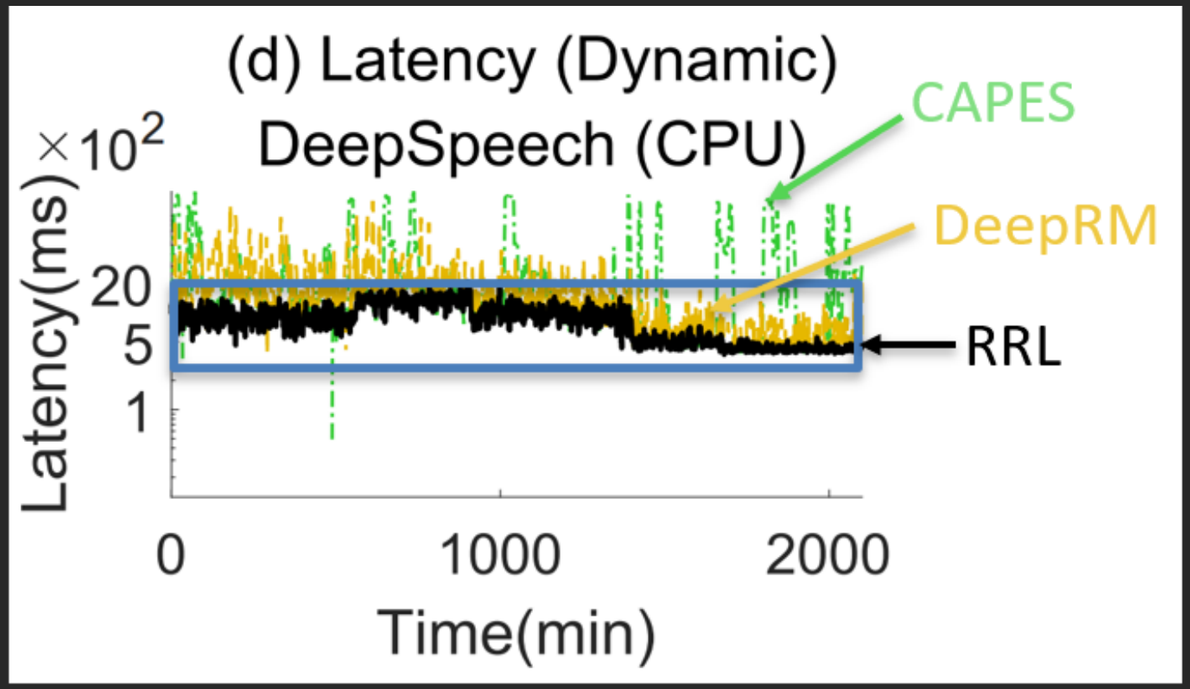


Swift Machine Learning Model Serving Scheduling: A Region Based Reinforcement Learning Approach
https://dl.acm.org/citation.cfm?id=3356164

This work published at SC'19

Results



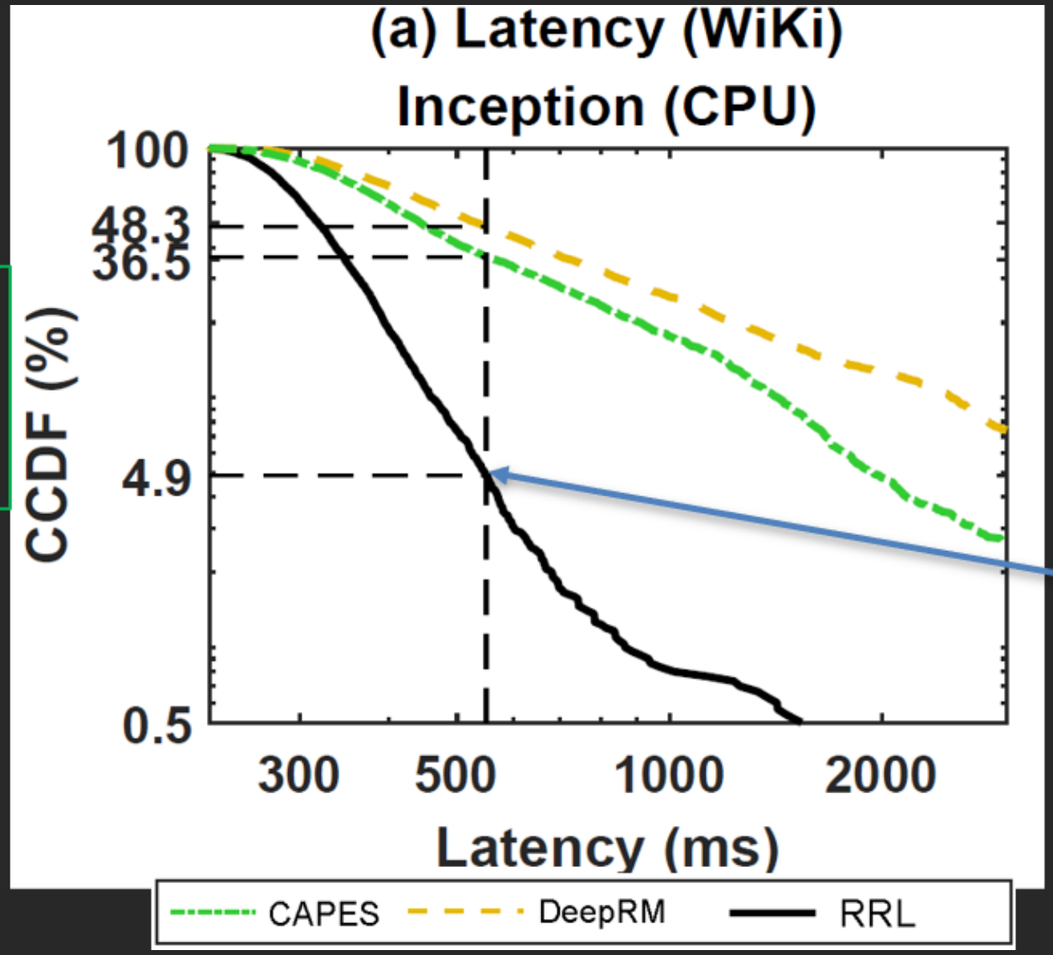


RRL has 63% less average latency than CAPES and 49% than DeepRM.

This work published at SC'19

Results

Complementary cumulative distribution function

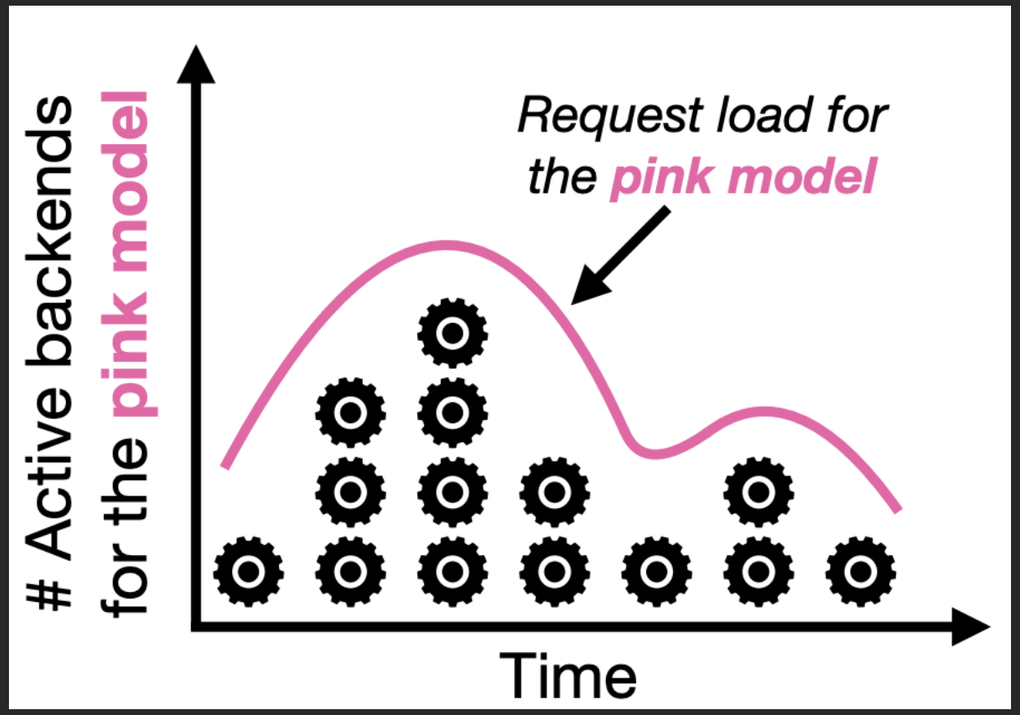


Please note that the Y-axis is log scale

The 95th percentile of RRL is 550 ms

This work published at SC'19

Conventional autoscaling: Amazon SageMaker



Reactive scaling: based on current load

Provisioning time (minutes) >> Execution time (< 1s)

Hide provisioning time → overprovisioning
How much overprovisioning?

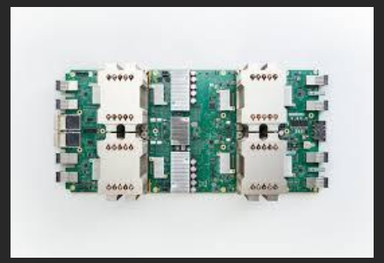
[1] https://people.mpi-sws.org/~arpanbg/pdfs/middleware2017_slides.pdf

Joint work with HKUST (Chengliang, Wei et al.) and published at USENIX ATC'19

ML accelerators: GPU, TPU, FPGA

- Mass parallel support
- Essential for training complex models
- Expensive

CPU: m5.xlarge: $0.192 per hour
GPU: p2.xlarge: $0.9 per hour
TPU v2: $4.5 per hour



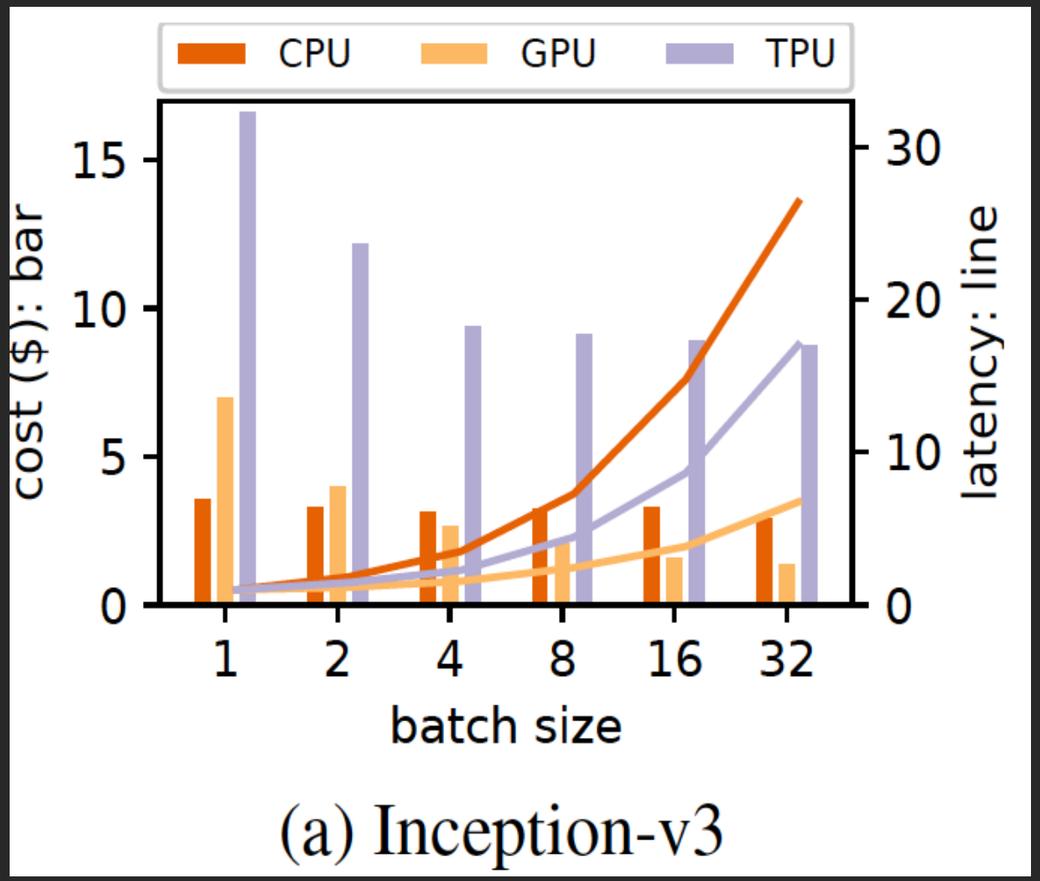
Inference
- Run comfortably without them
- Way less parallelism

- Choose between CPU and accelerators
- Justify the price tag

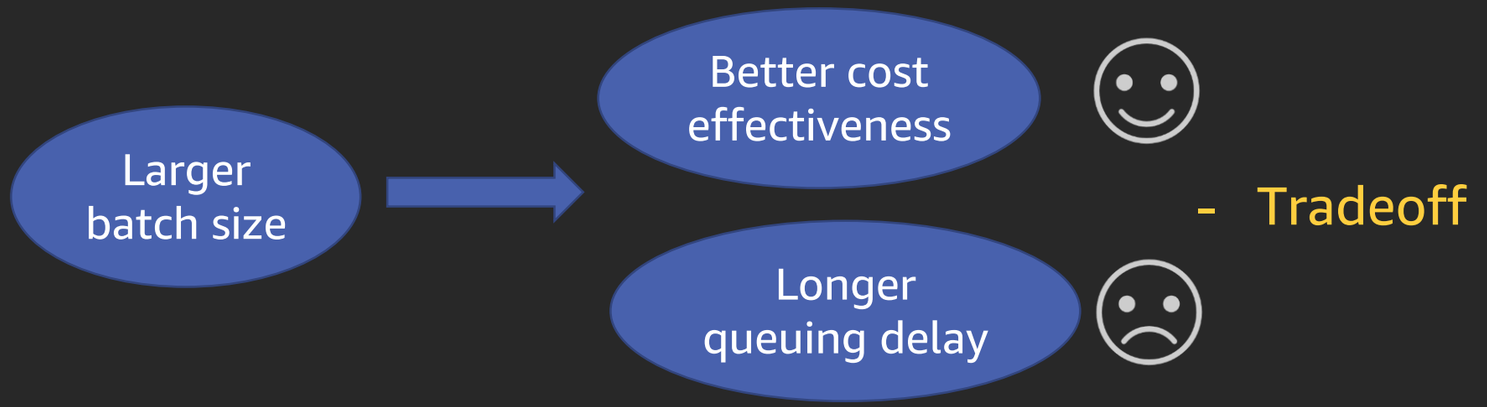
Joint work with HKUST (Chengliang, Wei et al.) and published at USENIX ATC'19

Characterization: CPU vs. GPU

CPU: 1 vCPU, 2 GB mem; GPU: K80; TPU: TPU-v2



- CPU: no significant benefits for small instances
- GPU: substantial benefit
- GPUs can be cheaper, but only with batching and high utilization



Joint work with HKUST (Chengliang, Wei et al.) and published at USENIX ATC'19

Design consideration: Cloud services for model serving

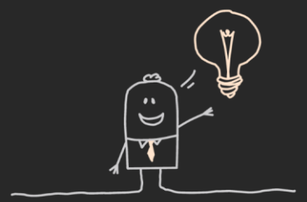
Infrastructure as a Service

Container as a Service

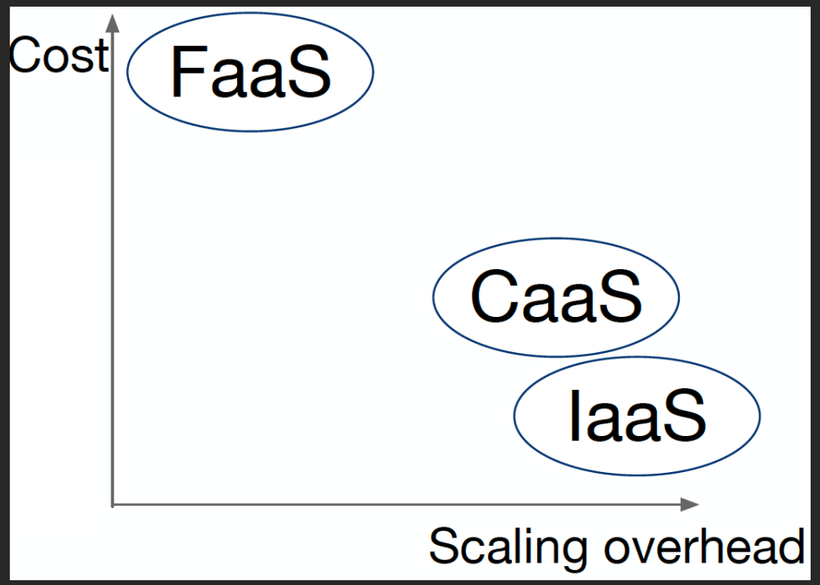
Function as a Service (FaaS, serverless comp.)

ML Model	EC2		ECS		Lambda	
	$	t (ms)	$	t (ms)	$	t (ms)
Inception-v3	5.0	210	9.17	217	19.0	380
Inception-ResNet	9.3	398	16.4	411	39.3	785
OpenNMT-ende	51.5	2180	96.3	2280	155	3100

EC2: c5.large; ECS: 2vCPU, 4GB mem; Lambda: 3008MB mem

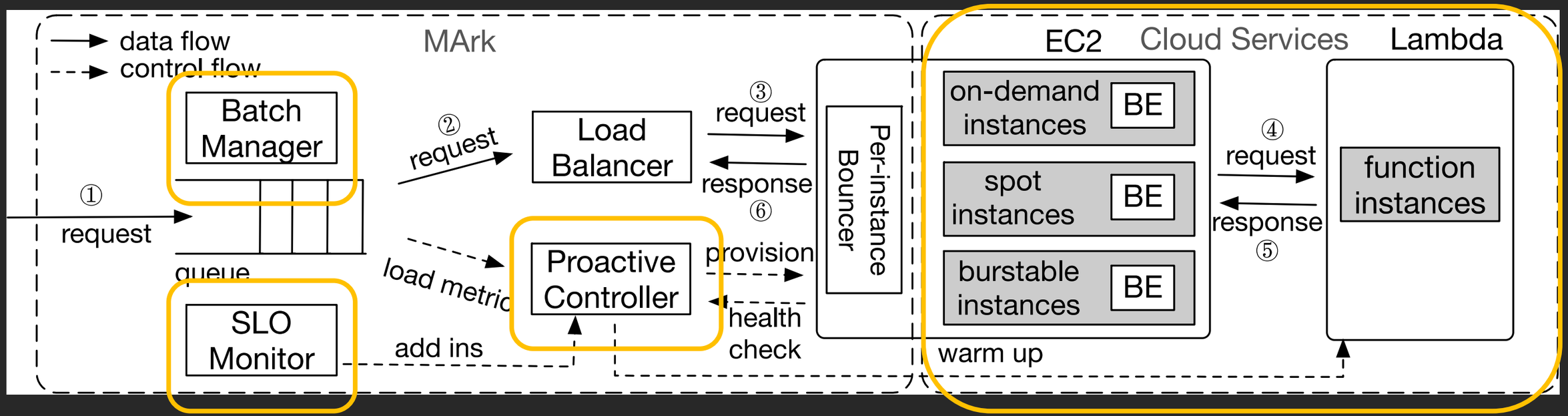


- Combine IaaS's cost advantage with FaaS's scalability
- Instead of overprovisioning IaaS, use FaaS to handle demand surge and spikes



Joint work with HKUST (Chengliang, Wei et al.) and published at USENIX ATC'19

MArk (Model Ark) using AWS Lambda



- Weighted round robin for load balancing
- Server front implemented with Sanic framework
- Support different serving frameworks

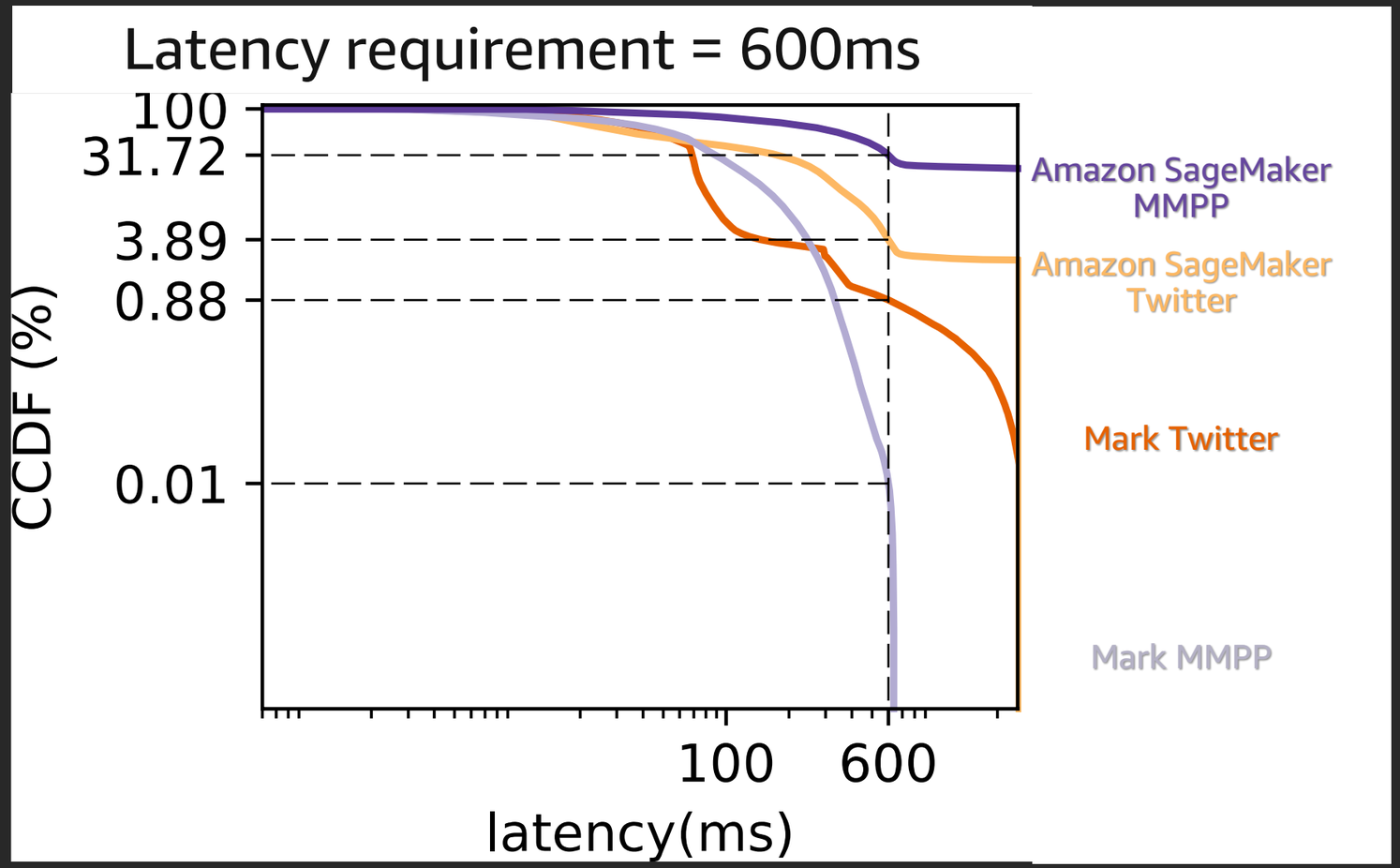
- Nginx and Gunicorn for admission and parallelism control
- Support for Amazon EC2 Spot Instances

MArk: Exploiting Cloud Services for Cost-Effective, SLO-Aware Machine Learning Inference Serving
https://www.usenix.org/system/files/atc19-zhang-chengliang.pdf

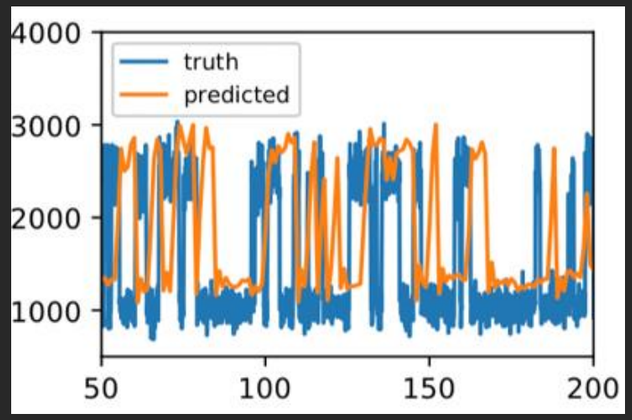
Joint work with HKUST (Chengliang, Wei et al.) and published at USENIX ATC'19

Results

Latency complementary cumulative distribution function



What if workload is unpredictable?



Markov-modulated Poisson Arrivals (MMPP)

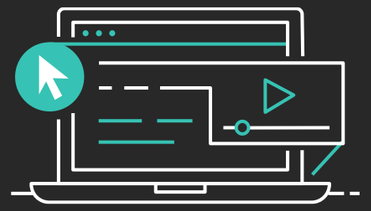
MMPP: unpredictable, highly dynamic workload

Joint work with HKUST (Chengliang, Wei et al.) and published at USENIX ATC'19

Learn ML with AWS Training and Certification

The same training that our own developers use, now available on demand

Role-based ML learning paths for developers, data scientists, data platform engineers, and business decision makers



70+ free digital ML courses from AWS experts let you learn from real-world challenges tackled at AWS

Validate expertise with the
AWS Certified Machine Learning - Specialty exam

Visit https://aws.training/machinelearning

aws training and certification

Thank you!

Sanjay Padhi

sanpadhi@amazon.com

Feng Yan

fyan@unr.edu

AWS re:Invent

aws

Please complete the session survey in the mobile app.