Live Video Analytics – the “killer app” for edge computing!

Ganesh Ananthanarayanan

Microsoft

Ganesh Ananthanarayanan

Victor Bahl

Peter Bodik

Yuanchao Shu

Shivaram Venkataraman

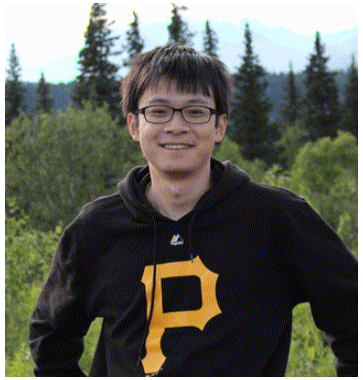
Michael Hung

Kevin Hsieh

Haoyu Zhang

Leana Golubchik

Minlan Yu



Junchen Jiang

Michael Freedman

Phil Gibbons

Onur Mutlu

Cameras are everywhere!

seattle.gov

Seattle Police Receive $600,000 Federal Grant For Body Cameras

theguardian

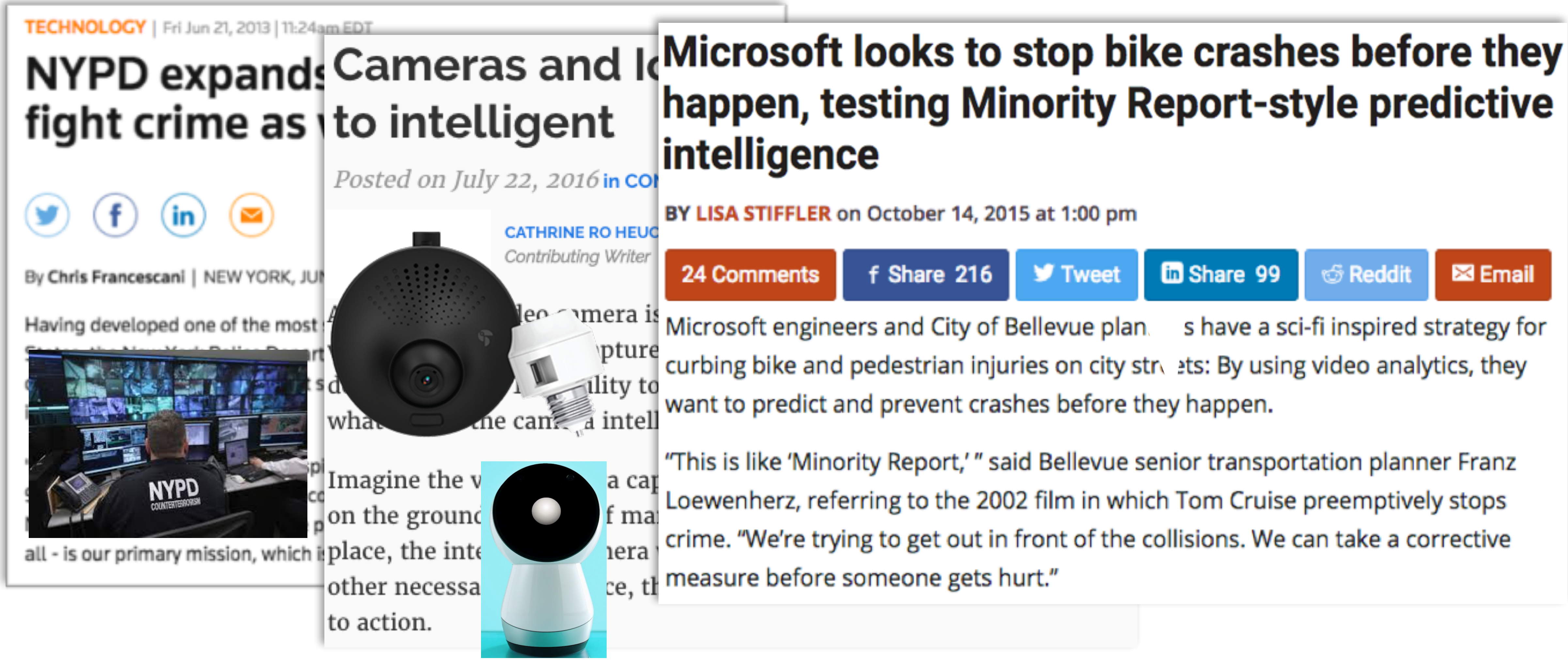
You're being watched: there's one CCTV camera for every 32 people in UK

THE WALL STREET JOURNAL.

REUTERS

There is camera deployed for every 29 people worldwide, and a camera for every eight people in the US!

Cameras are everywhere!



TECHNOLOGY | Fri Jun 21, 2013 | 11:24am EDT

NYPD expands

fight crime as

By Chris Francescani | NEW YORK, JU

Having developed one of the most

all - is our primary mission, which i

Cameras and I

to intelligent

Posted on July 22, 2016 in CO

CATHRINE RO HEUC

Contributing Writer

A ... deo ... mera is ... pture ... d ... ility to ... wha ... the cam ... a intell

Imagine the v ... a cap ... on the ground ... f ma ... place, the inte ... nera ... other necessa ... ce, th ... to action.

Microsoft looks to stop bike crashes before they happen, testing Minority Report-style predictive intelligence

BY LISA STIFFLER on October 14, 2015 at 1:00 pm

24 Comments | f Share 216 | Tweet | Share 99 | Reddit | Email

Microsoft engineers and City of Bellevue plan s have a sci-fi inspired strategy for curbing bike and pedestrian injuries on city str ets: By using video analytics, they want to predict and prevent crashes before they happen.

"This is like 'Minority Report,'" said Bellevue senior transportation planner Franz Loewenherz, referring to the 2002 film in which Tom Cruise preemptively stops crime. "We're trying to get out in front of the collisions. We can take a corrective measure before someone gets hurt."

Video Analytics towards Vision Zero

Traffic fatalities are among the top-ten causes of deaths!

Global Partners

Microsoft
City of Bellevue Washington
University of Washington

WSDOT
SDOT Seattle Department of Transportation
New York City DOT
City of Redmond
Hamilton

Snohomish County Washington
King County METRO
Miami-Dade County
DC .gov
City of Pasadena

SFMTA Municipal Transportation Agency
LADOT Moving Los Angeles Forward
City of Vancouver
Calgary
Gainesville

The University of British Columbia
Lund University
McGill
Polytechnique Montréal
Walk Friendly Communities

Portland State University
UAB Universitat Autònoma de Barcelona
unity
ite A Community of Transportation Professionals
peopleforbikes

ITS America
Vision Zero Network
MetroLab Network
cascade bicycle club

US Department of Transportation
2016
Safer Cities, Safer People Award

Institute of Transportation Engineering
2017
Achievements Award

Democratize video analytics!

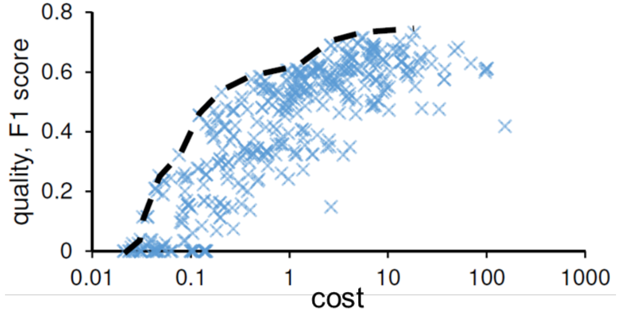
Real-time, low-cost, accurate

video analytics system

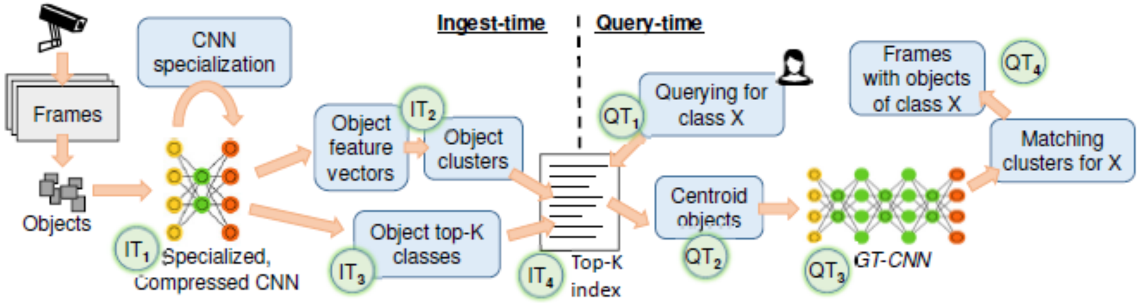
for a collection of cameras

This talk will cover…

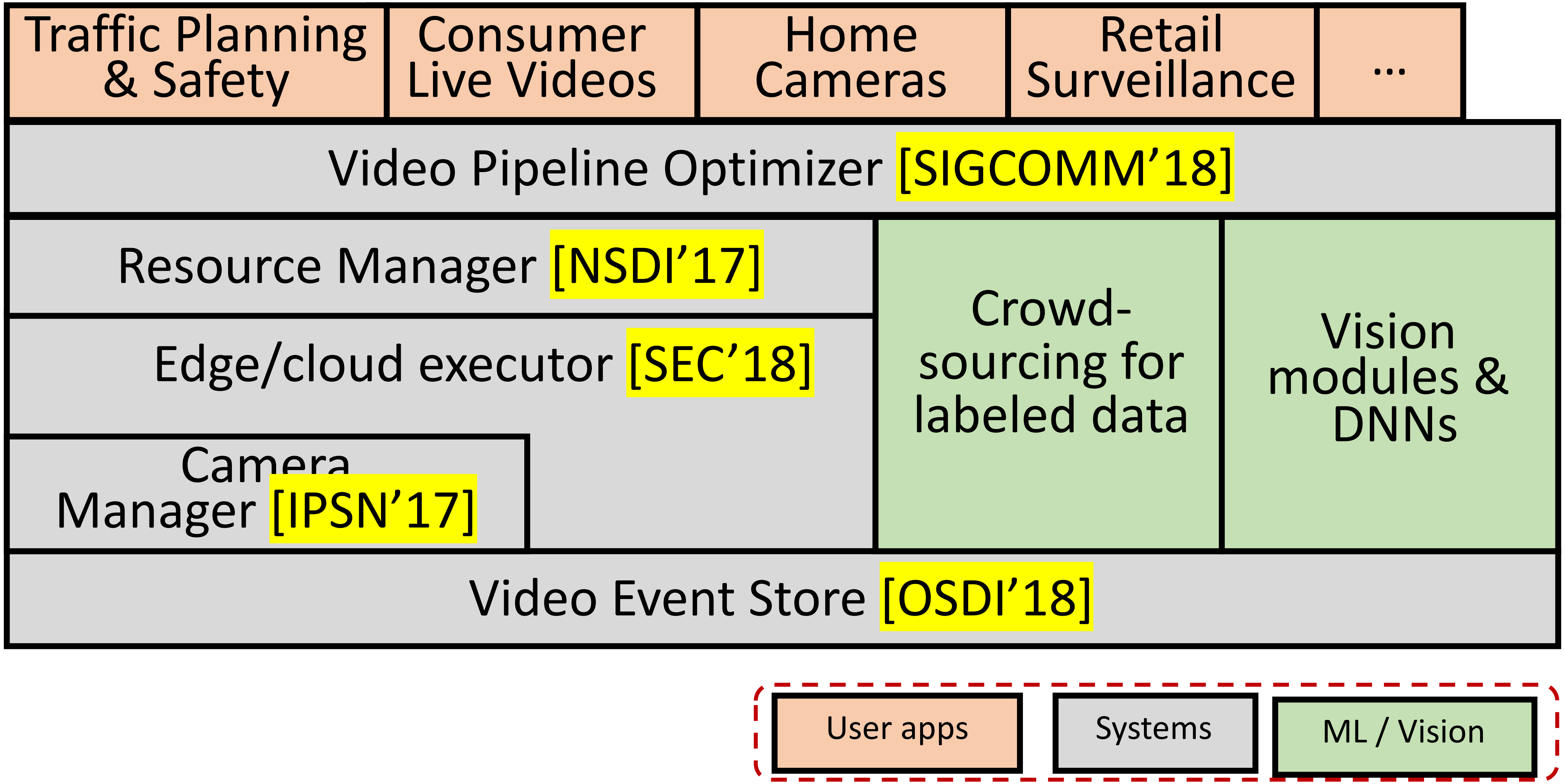
- Video Analytics at scale with *approximation* [NSDI’17, SIGCOMM’18, SEC’18]



- Interactive querying of stored video datasets [OSDI’18]



Rocket: Video Analytics Stack



Traffic Planning & Safety

Consumer Live Videos

Home Cameras

Retail Surveillance

…

Video Pipeline Optimizer [SIGCOMM’18]

Resource Manager [NSDI’17]

Edge/cloud executor [SEC’18]

Camera Manager [IPSN’17]

Crowd-sourcing for labeled data

Vision modules & DNNs

Video Event Store [OSDI’18]

User apps

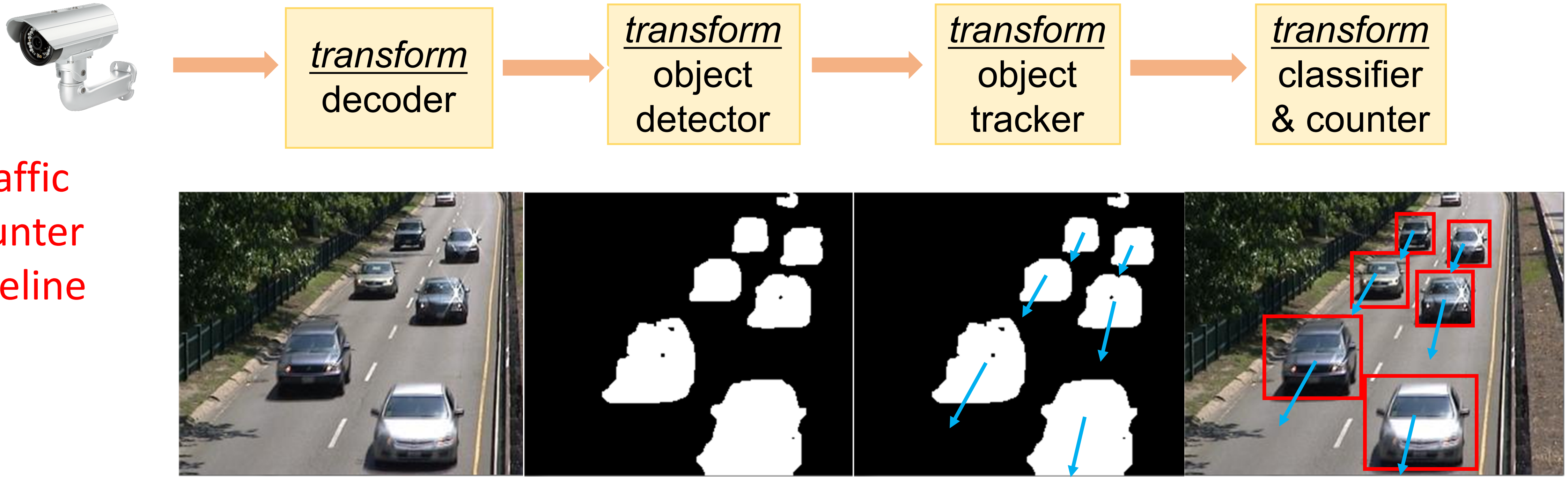
Systems

ML / Vision

Video query: pipeline of *transforms*

Vision algorithms ("*transforms*") chained together

Traffic counter pipeline



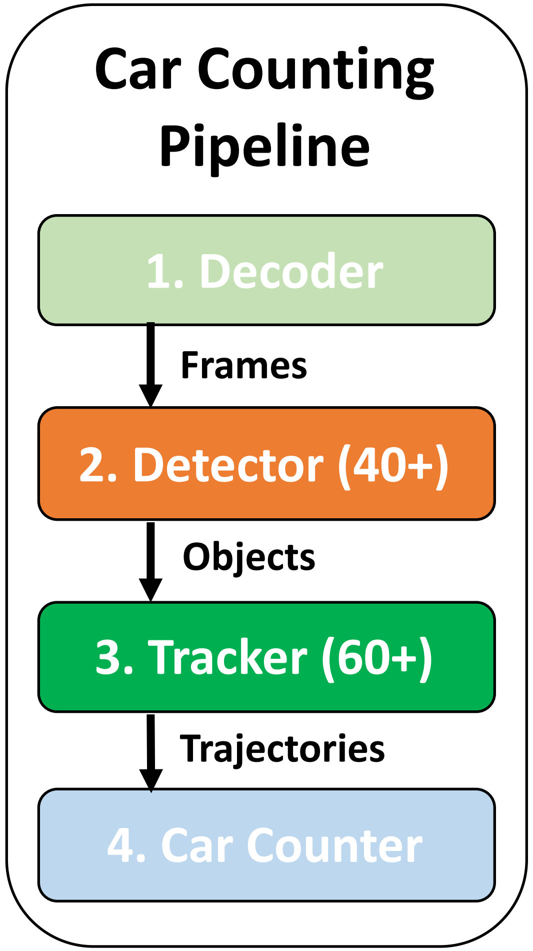
transform decoder

transform object detector

transform object tracker

transform classifier & counter

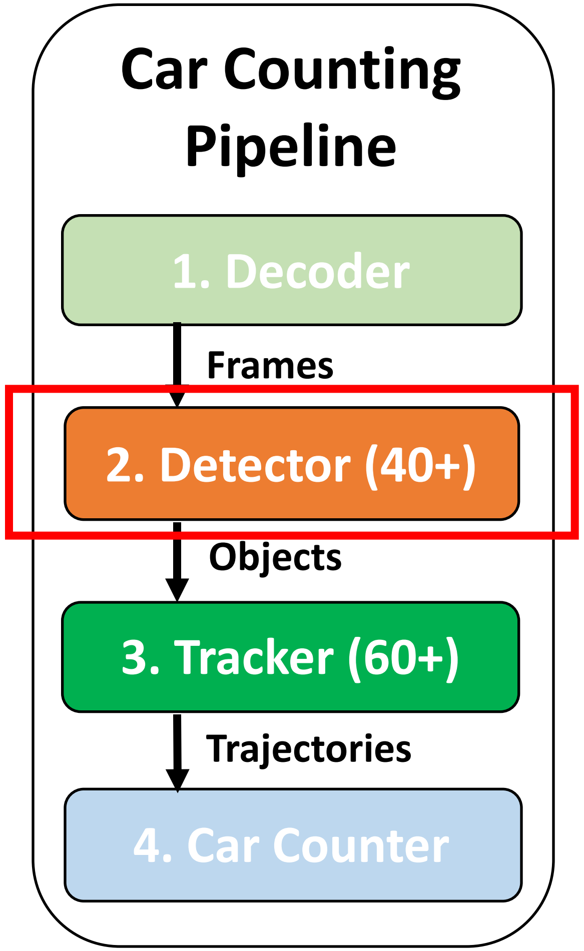
Curse of many choices!



- **Detector implementations (40+)**
  - Motion-based: background subtraction
  - DNN-based: Yolo detection
  - Exhaustive search

- **Tracker implementations (60+)**
  - Moving pattern
  - Color histogram
  - Key-point features: SURF, SIFT

Implementations make different design choices and consume different resources



Car Counting Pipeline

1. Decoder

Frames

2. Detector (40+)

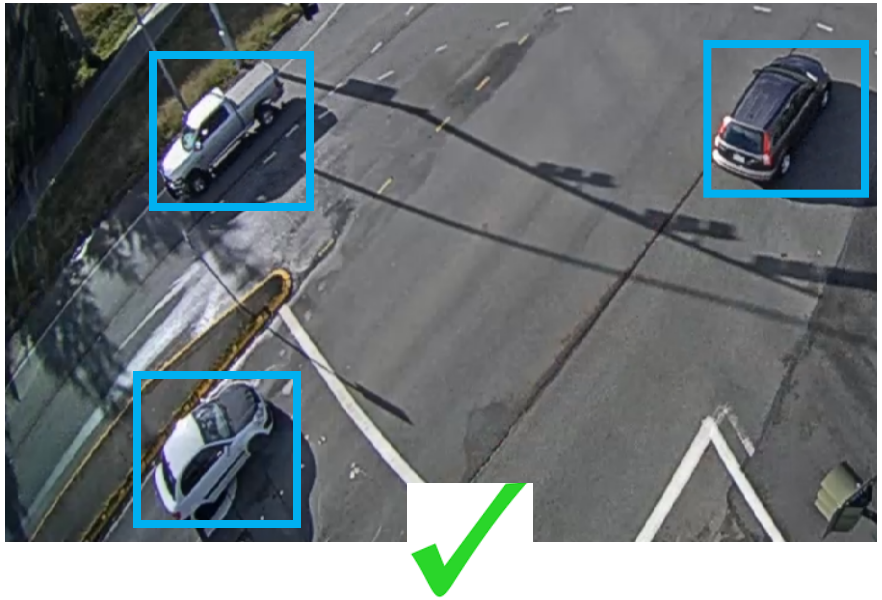
Objects

3. Tracker (60+)

Trajectories

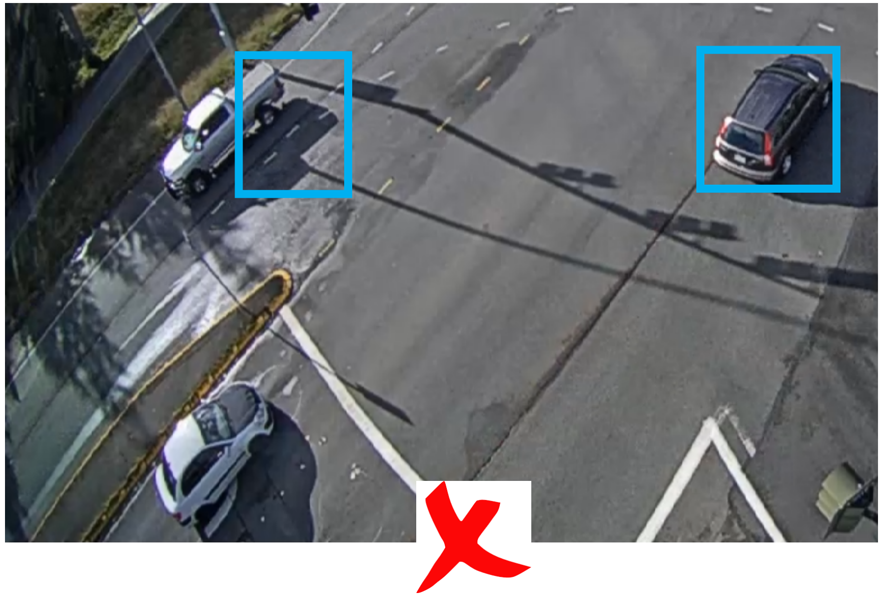
4. Car Counter

Background Subtraction



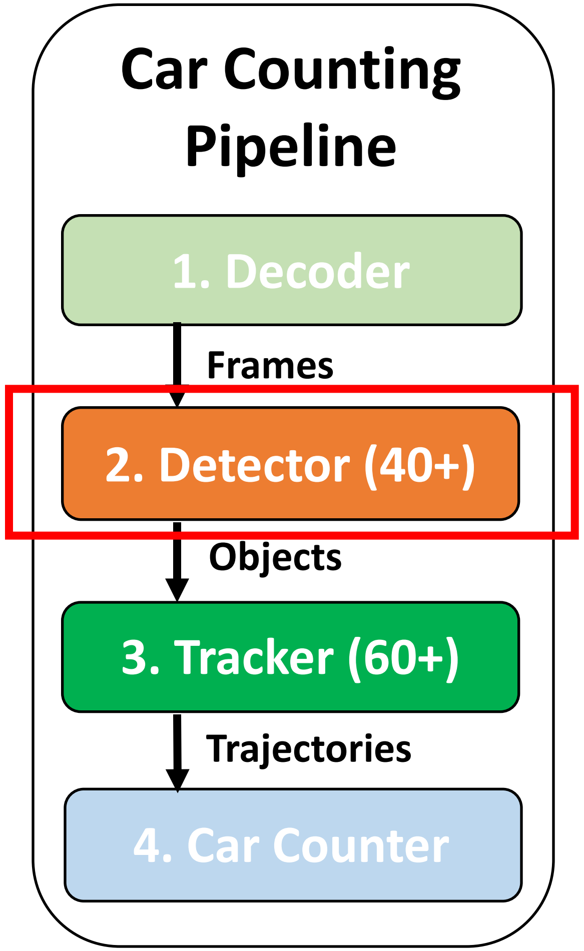
✓

DNN Object Detector



✗

150th NE and Newport Ave
Bellevue, WA



Car Counting Pipeline

1. Decoder

Frames

2. Detector (40+)

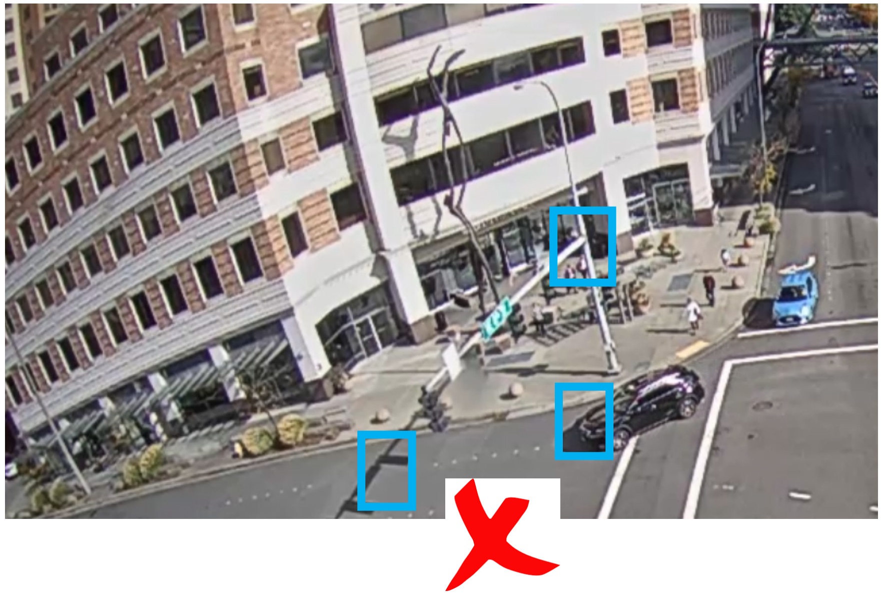
Objects

3. Tracker (60+)

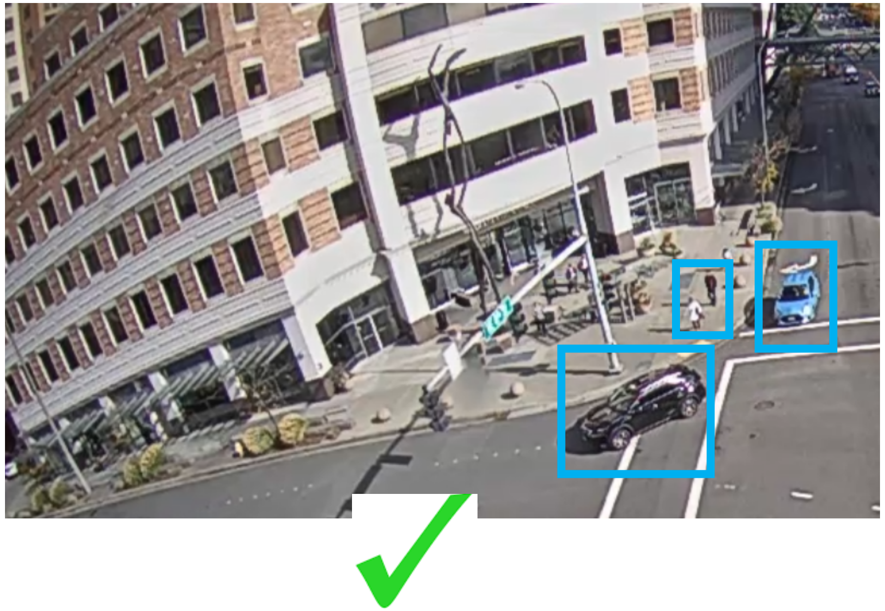
Trajectories

4. Car Counter

Background Subtraction



DNN Object Detector



Bellevue Ave and NE 8th
Bellevue, WA

Vision algorithms have “knobs” to set

Frame Rate

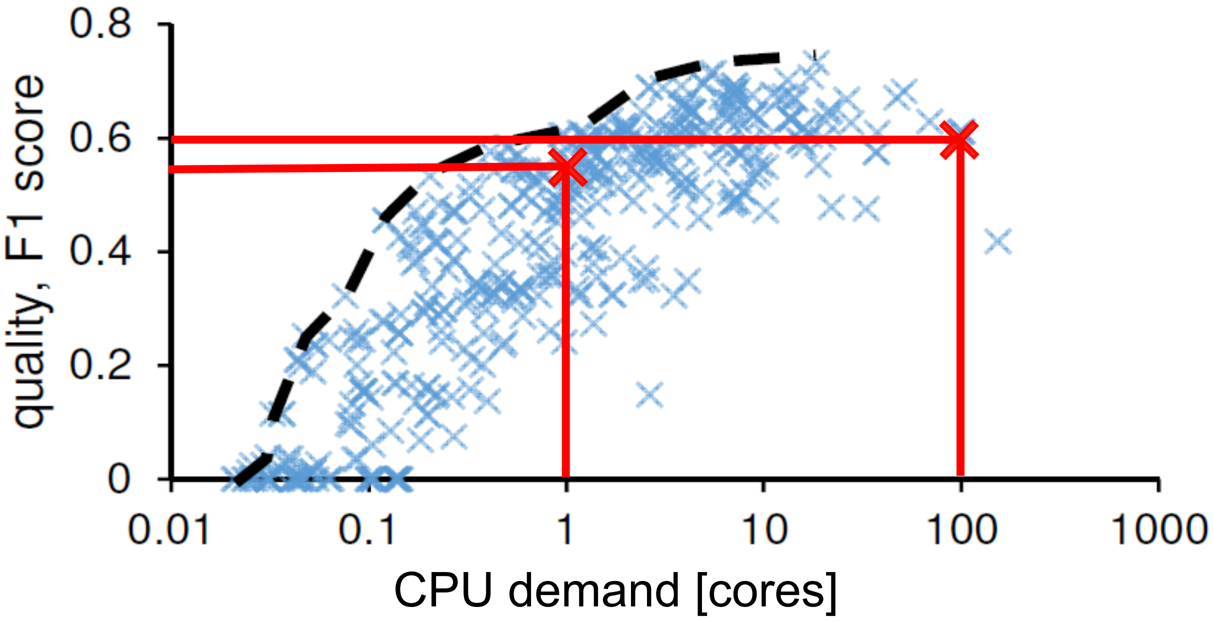
Resolution

30
frames/second
for HD cameras

1080p, 720p,
480p…

How much do the "query plans" – *knobs & implementations* – differ?

License Plate Reader



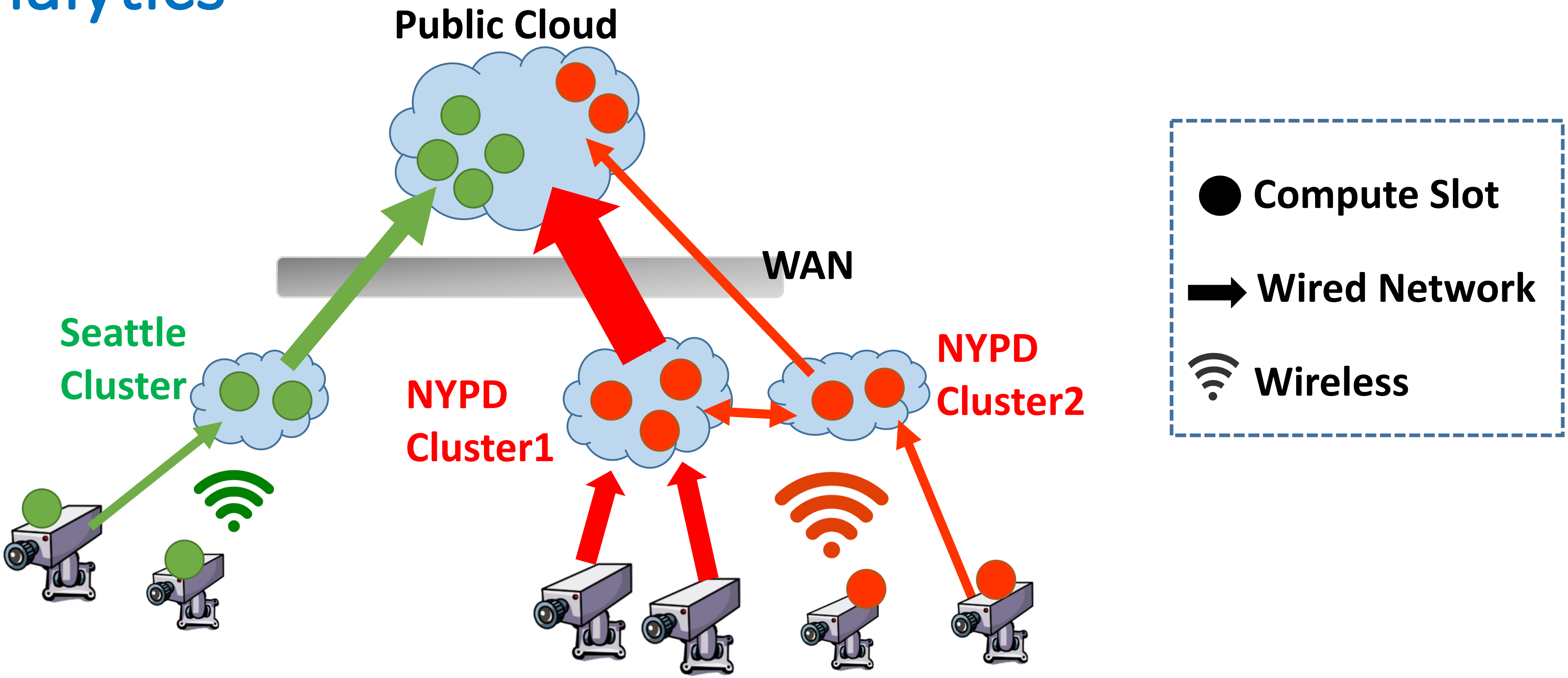
Orders of magnitude cheaper resource demand for little drop in quality

Dependent on the camera, lighting, object color, …

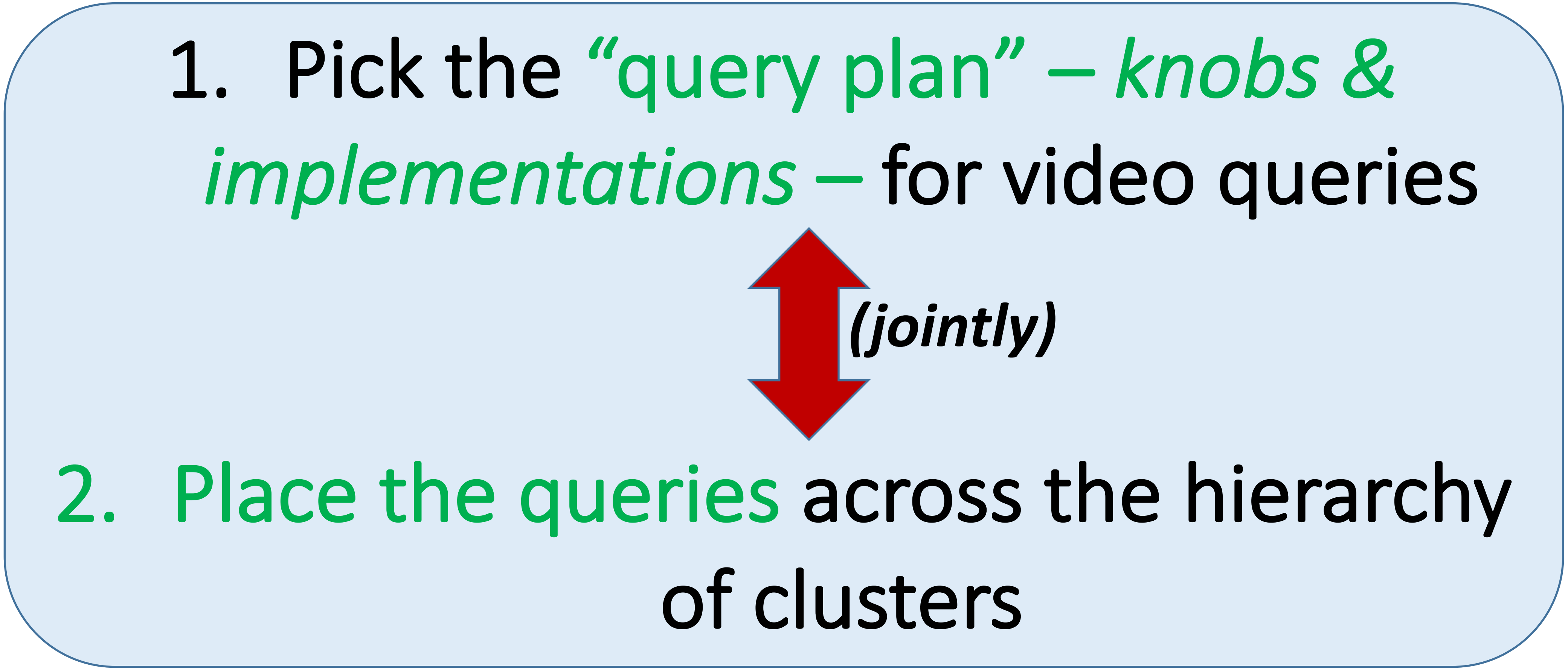
No analytical models to construct resource-quality profiles

- Different from approximate SQL queries

Hierarchy of clusters for video analytics



Edge Computing is a must! ← {Bandwidth, latency, availability}

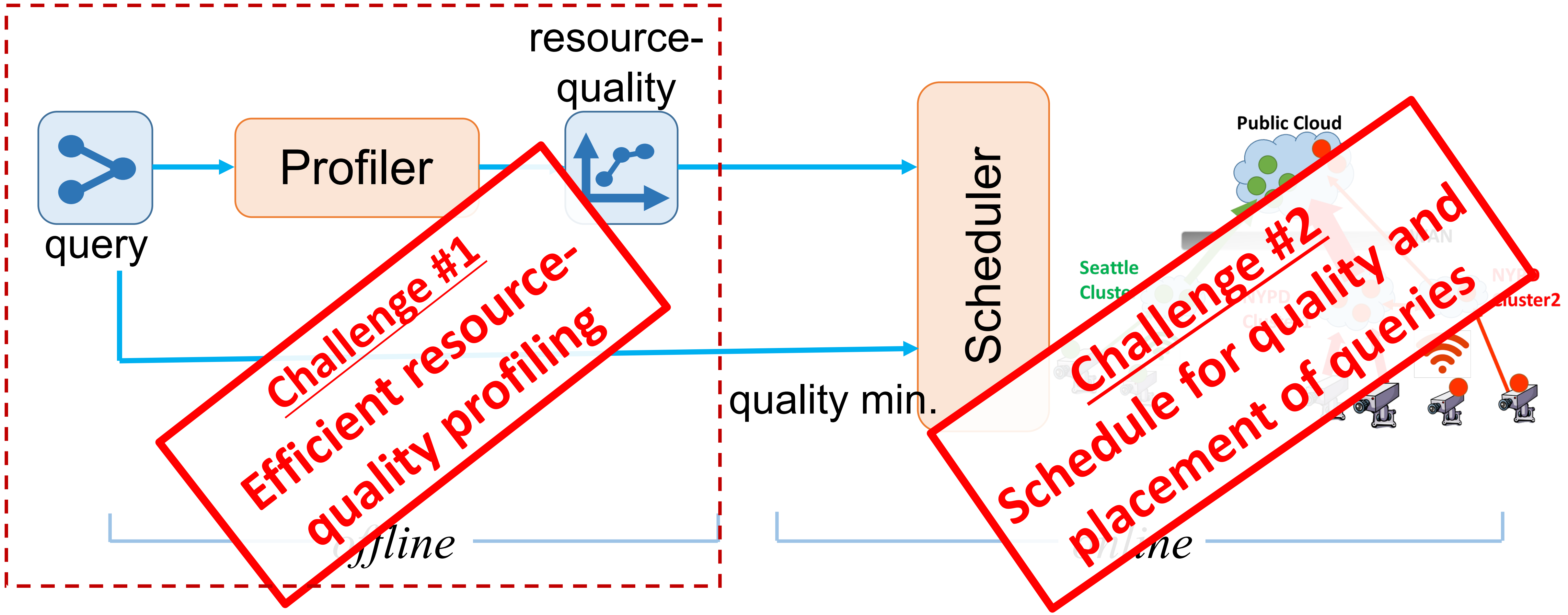


1. Pick the “query plan” – *knobs & implementations* – for video queries

(jointly)

2. Place the queries across the hierarchy of clusters

Solution Overview



Challenge #1

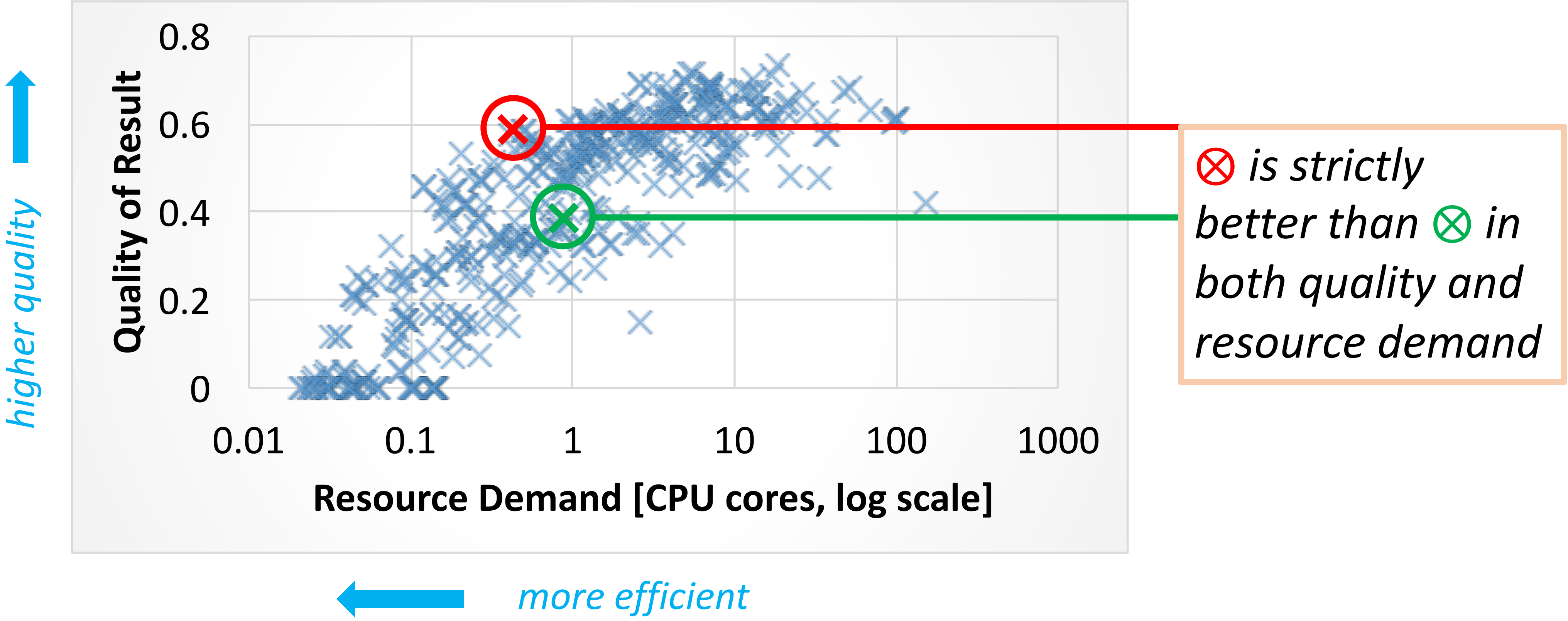
Efficient resource-quality profiling

Challenge #2

Schedule for quality and placement of queries

Offline: Resource-Quality Profiling

- Profile: query plan ⟹ {resource, quality}
  - Ground-truth: labeled dataset or results from *golden* configuration
  - Targeted search for promising query plans

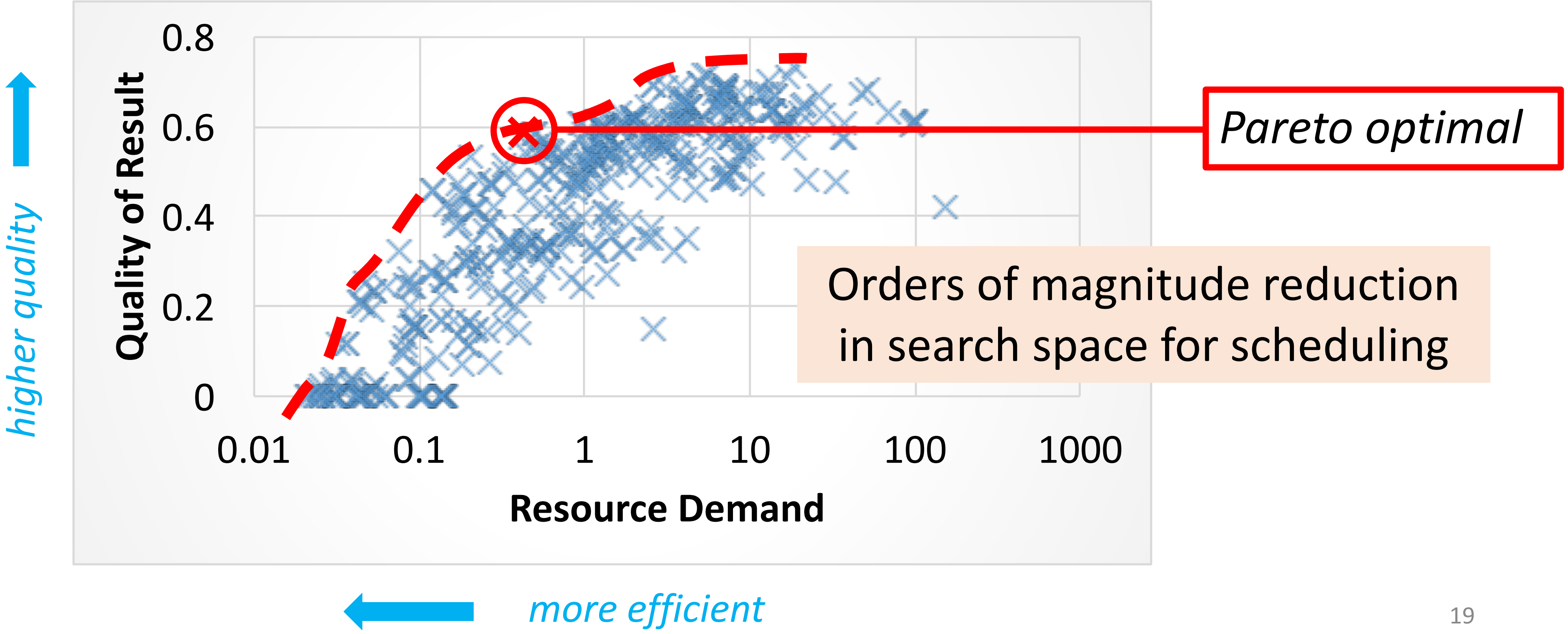


⊗ is strictly better than ⊗ in both quality and resource demand

Offline: Pareto boundary

Pareto boundary: optimal query plans in resource demand and quality

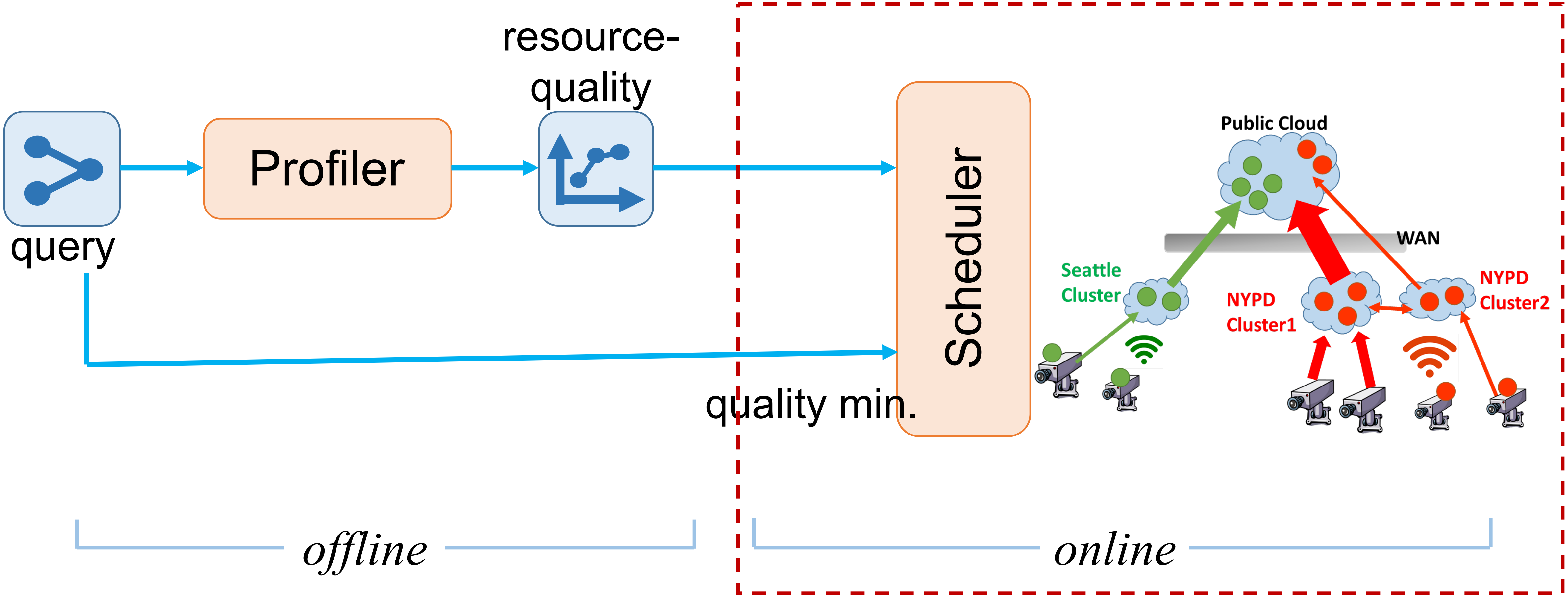
- Non-Pareto plans cannot beat Pareto plans in *both* quality & resources



Pareto optimal

Orders of magnitude reduction in search space for scheduling

Solution Overview



query

Profiler

resource-quality

Scheduler

quality min.

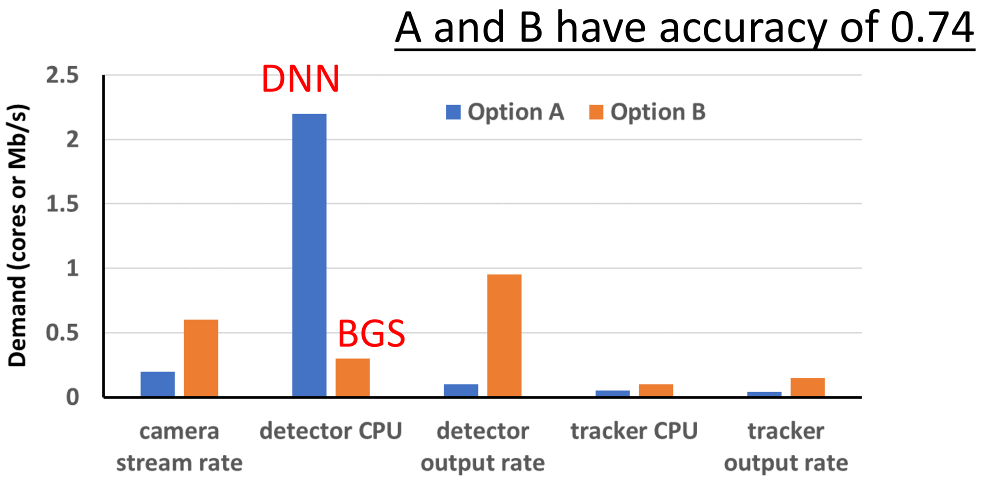
offline

online

Greedy scheduling to max. accuracy of queries

Dominant Resource Demand

- Multi-resource – compute & network


A and B have accuracy of 0.74

- For each (plan, placement) pair, calculate the *fraction* of demand at *each location*

  → calculate the max (or dominant) value

- Avoids lopsided drain of any single resource at any location

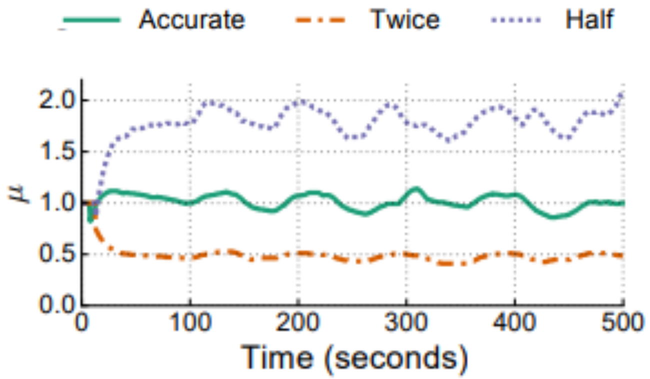
Evaluation Highlights

Workload

- Videos from traffic cameras & surveillance cameras
  - Original frame rate of 14 – 30 fps, resolution 480p – 1080p
- Queries: Object tracker, DNN classifier, Car counter, License plate reader

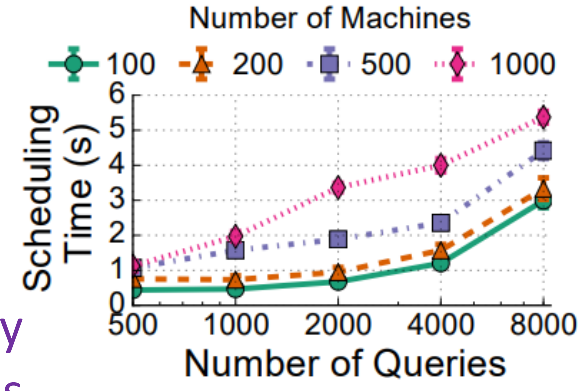
Results

- 25x better accuracy & within 6% of optimal



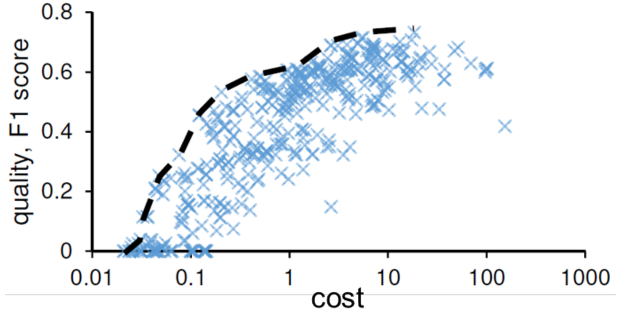
Adapts to errors in the profile

Scales to many 1000's of queries

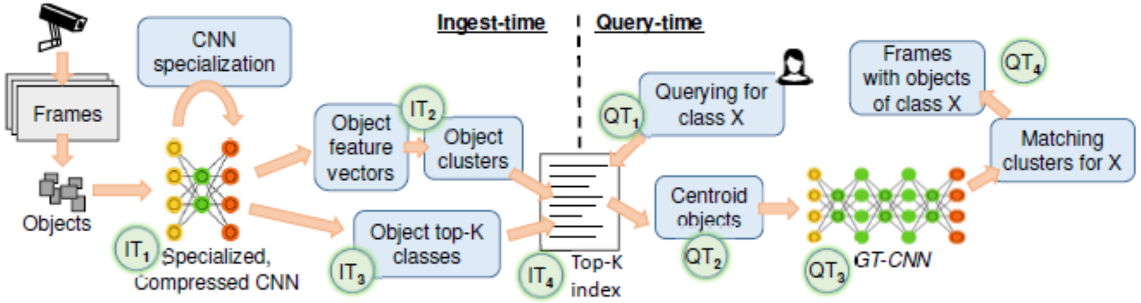


This talk will cover…

- ✓ Video Analytics at scale with *approximation* [NSDI’17, SIGCOMM’18, SEC’18]



- Interactive querying of stored video datasets [OSDI’18]

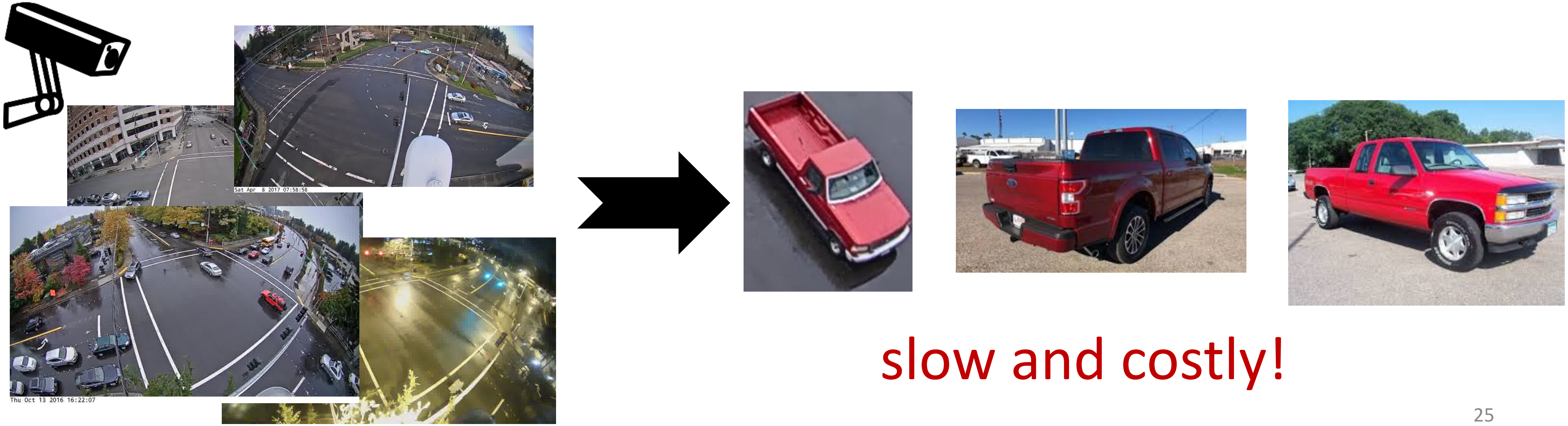


Video Recordings are Ubiquitous

- Massive amounts of video recordings everywhere

Querying on Videos is Useful but Challenging

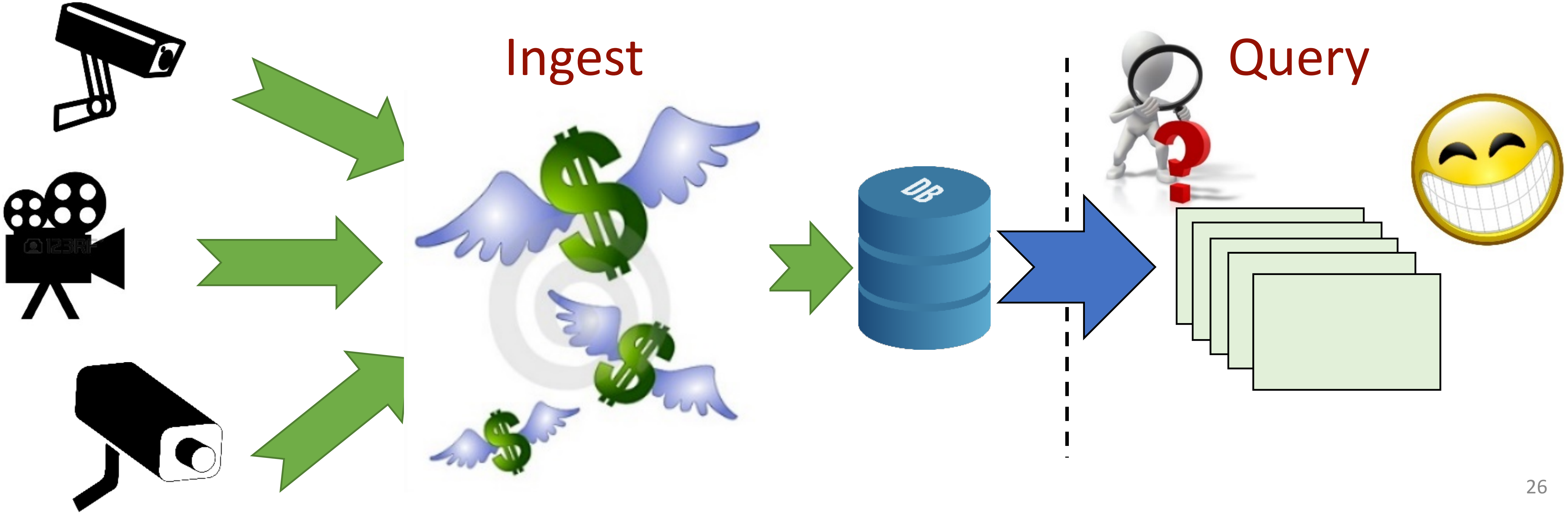
- Querying videos for objects is enabled by Convolution Neural Networks (CNNs)
  - *Find all red trucks in Bellevue traffic videos last week*



slow and costly!

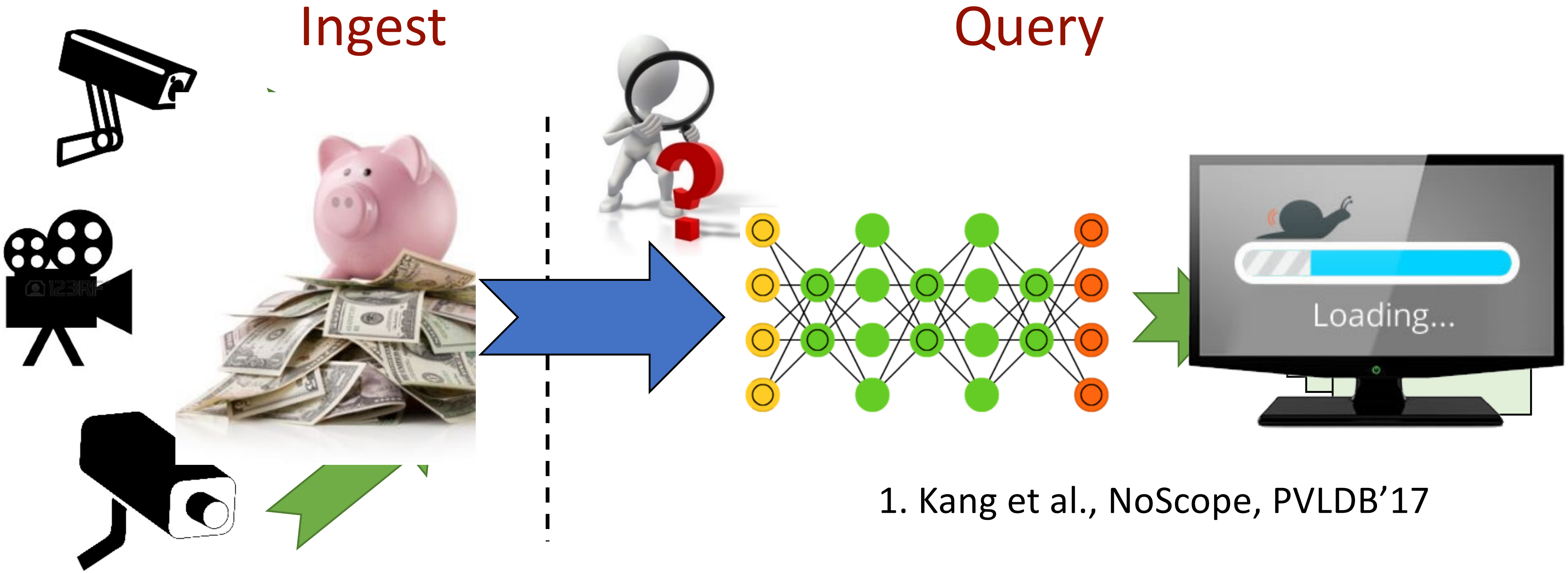
Ingest Time Analysis: Too Costly

- Analyzing all videos at ingest time can make query fast
  - But it is costly and potentially wasteful ($380/month/stream)



Query Time Analysis: Too Slow

- Analyzing videos at query time can save cost
  - But it very slow (5 hr for a month-long video [1])

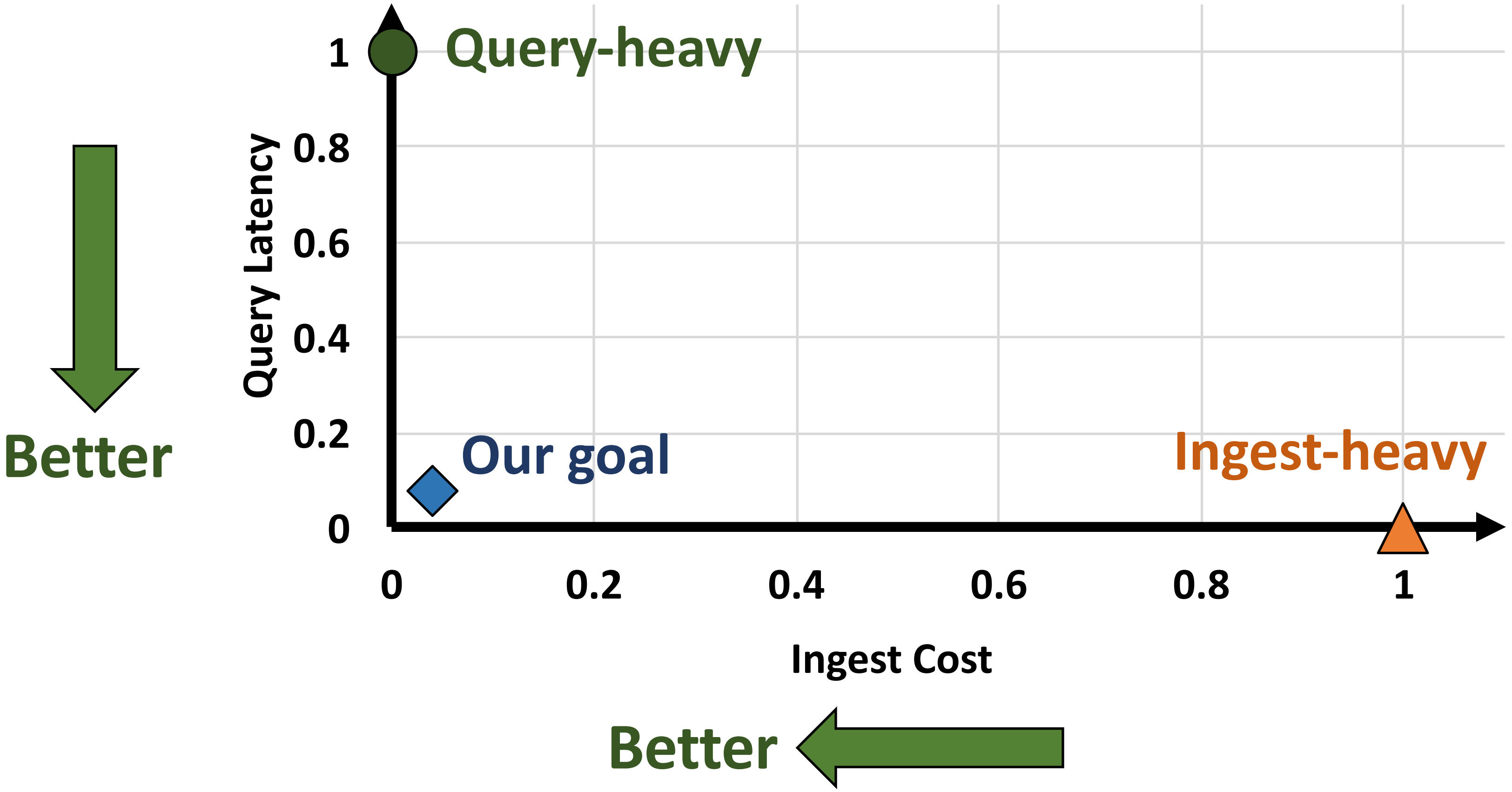


Ingest

Query

1. Kang et al., NoScope, PVLDB'17

Enable low-latency, low-cost, and high-accuracy querying over large historical video datasets



System Objectives

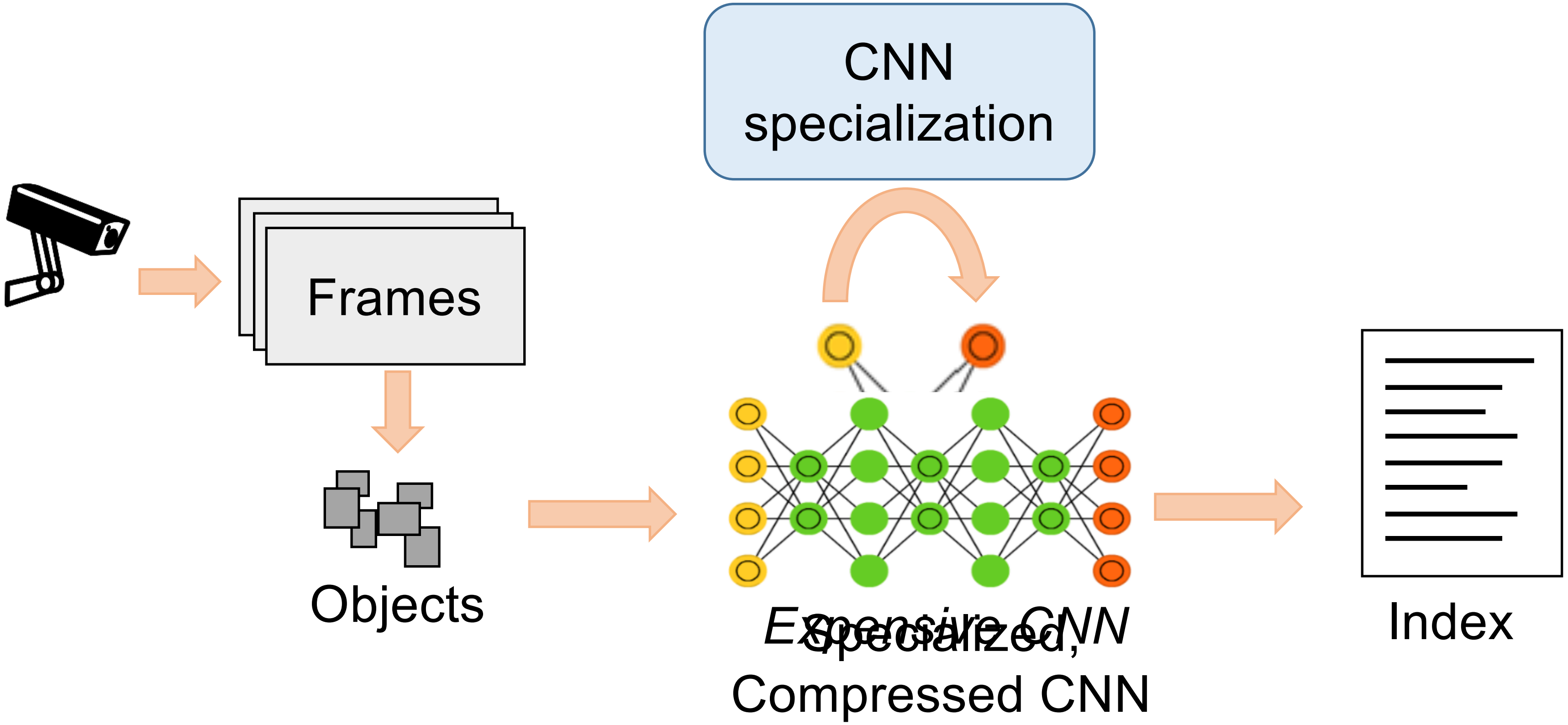
- Provide low-cost indexing at ingest time
- Achieve high accuracy and low latency at query time

System Objectives

- Provide low-cost indexing at ingest time
- Achieve high accuracy and low latency at query time

Low-Cost Ingestion: Cheaper CNNs

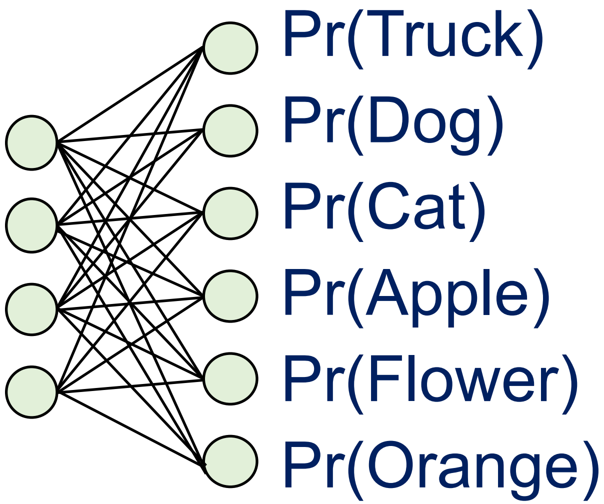
- Process video frames with a cheap CNN at ingest time
  - Compressed and Specialized CNN: fewer layers / weights and are specialized for each video stream



Challenge: Cheap CNNs are Less Accurate

- Cheaper CNNs are less accurate than the expensive CNNs

The best result from the expensive CNN is within the top-K results of the cheaper CNN

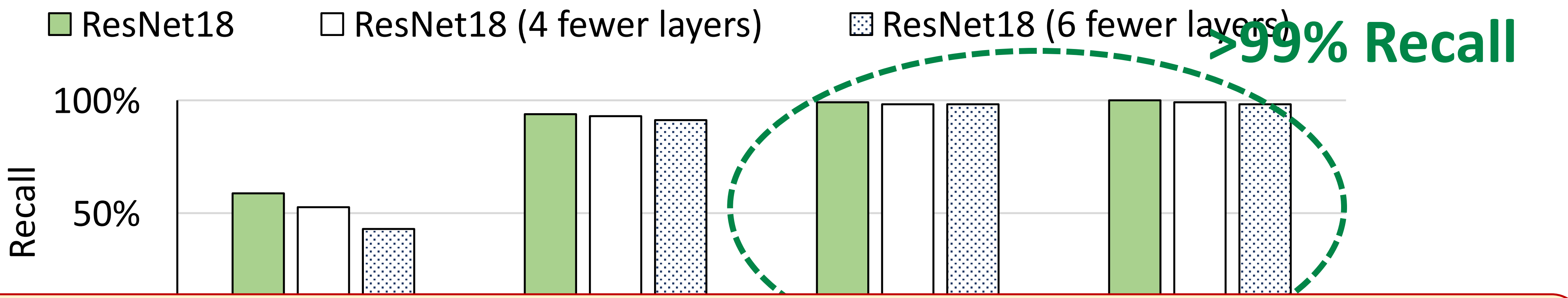


Rank	Expensive CNN	Cheap CNN
1	**Truck**	**Moving Van** ☒
2	Moving Van	**Airplane**
3	Passenger Car	**Truck** ☑
4	Recreational vehicle	Passenger Car

Recall, Precision and Top-K Results

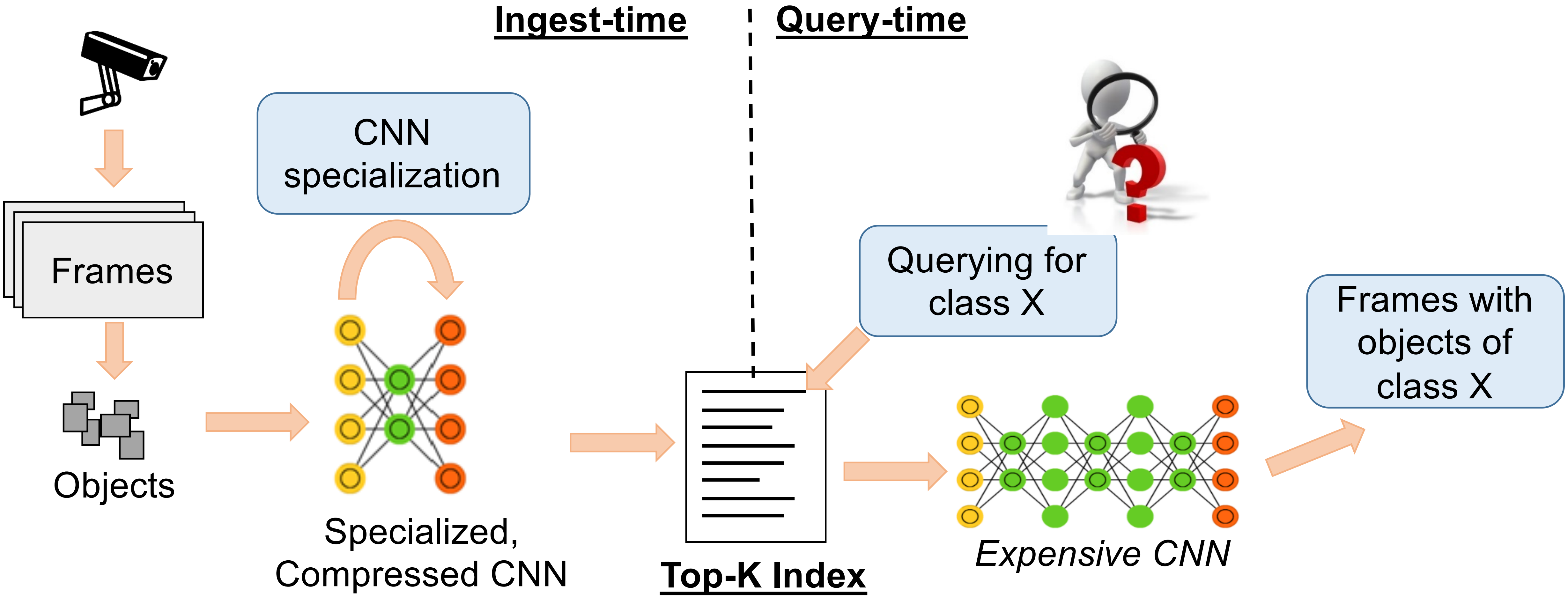
Recall: Fraction of relevant objects that are selected

Precision: Fraction of selected objects that are relevant



Cheap CNNs can achieve high recall
with small top-K results

Solution: Top-K Approximate Index

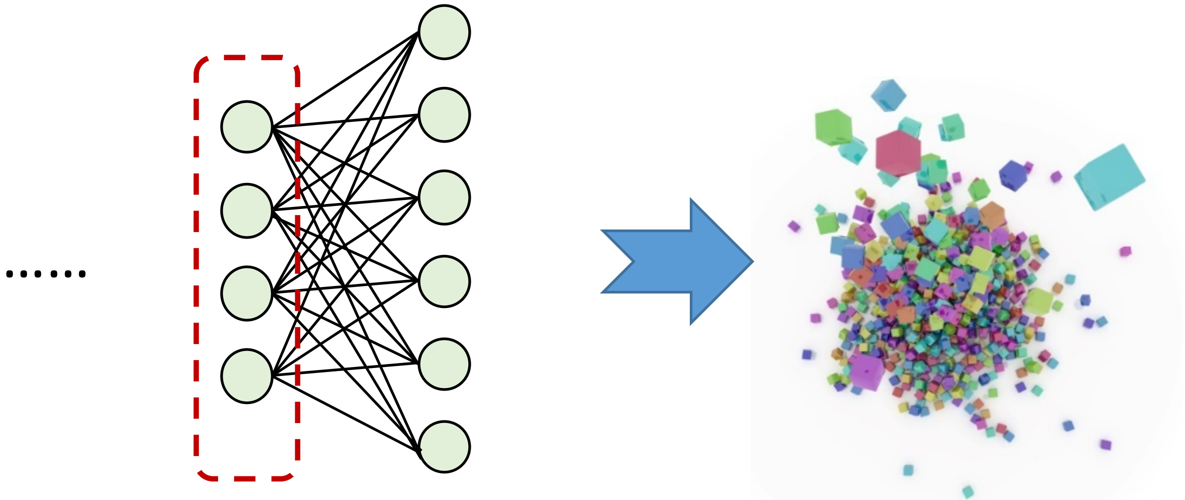


System Objectives

- Provide low-cost indexing at ingest time
- Achieve high accuracy and low latency at query time

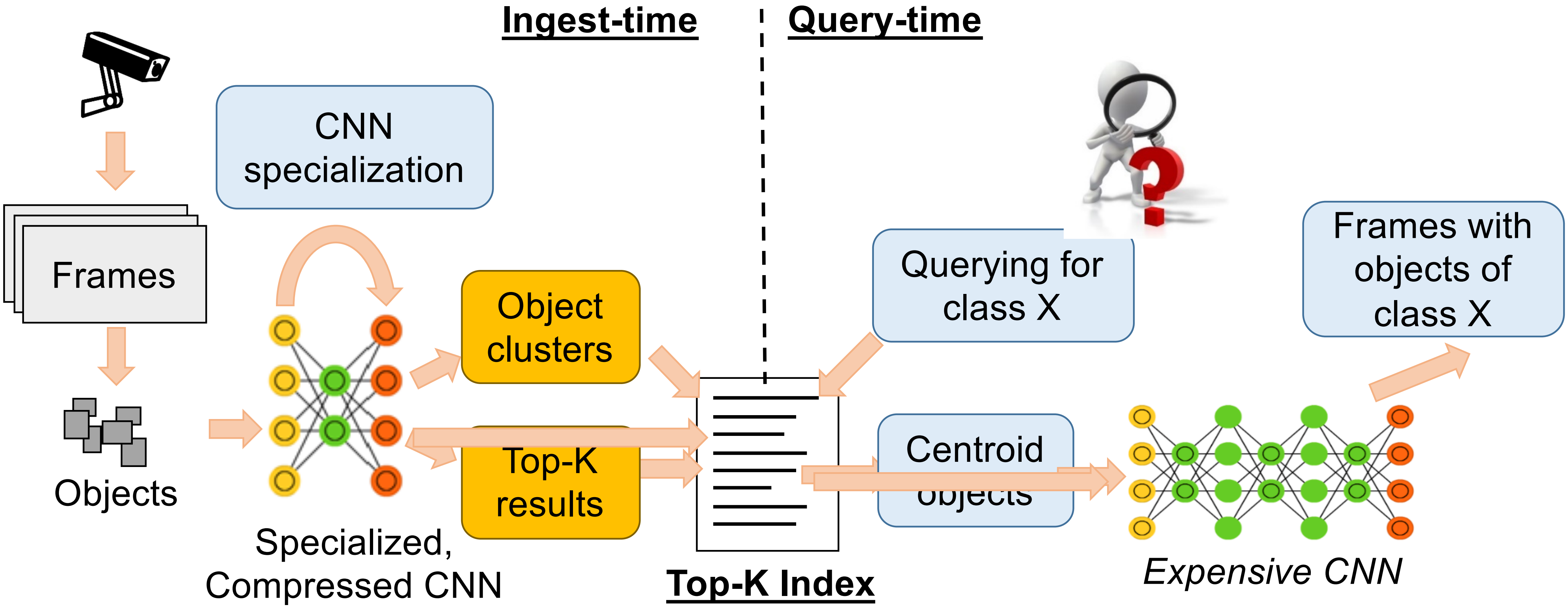
Low-Latency Query: Redundancy Elimination

- Approximate indexing ➔ non-trivial work at query time
- Minimize the work at query time ➔ clustering similar objects based on the extracted features
  - Images with similar feature vectors are visually similar [1, 2, 3]



1. Krizhevsky et al., NIPS’12
2. Babenko et al., ECCV’14
3. Razavian et al., CVPR Workshop’14

Adding Feature-based Clustering



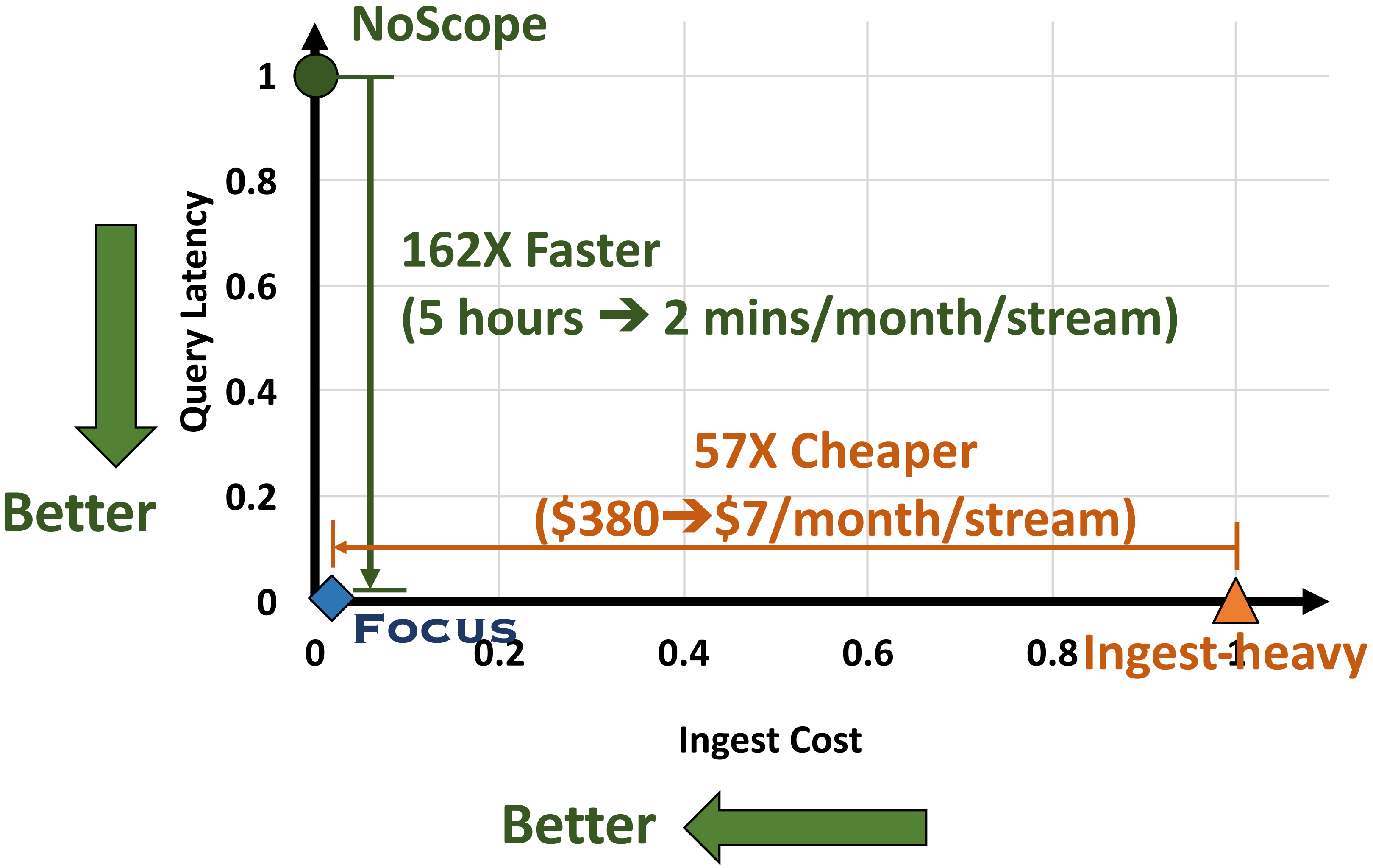
Results Summary

Video Datasets

Traffic & surveillance videos

Accuracy Targets

Recall & precision – 99%
(w.r.t. YOLOv2)



162X Faster
(5 hours → 2 mins/month/stream)

57X Cheaper
($380→$7/month/stream)

Video Analytics & Edge Computing – better together!

- Video Analytics with *approximation* [NSDI’17, SIGCOMM’18, SEC’18]
  - Resource-accuracy tradeoff for multi-dimensional video queries
  - Edge-cloud partitioning
  - ✓25x better accuracy & within 6% of optimal
- Interactive querying of stored video datasets [OSDI’18]
  - Low-cost ingesting of videos for approximate indexing
  - Interactive querying of stored videos
  - 52X cheaper and 162X faster

http://aka.ms/rocket

http://aka.ms/ganesh

Hot Topics in Video Analytics and Intelligent Edges
(co-located with MobiCom 2019 in Los Cabos, Mexico)
Deadline: Jun 14, 2019