Focus: Querying Large Video Datasets with Low Latency and Low Cost

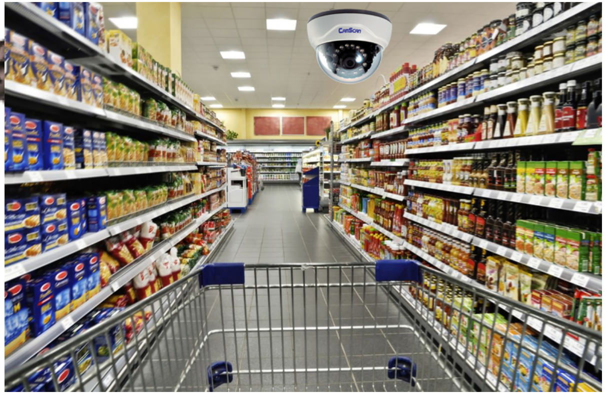
Kevin Hsieh

Ganesh Ananthanarayanan, Peter Bodik, Shivaram Venkataraman, Paramvir Bahl, Matthai Philipose, Phillip B. Gibbons, Onur Mutlu

Carnegie Mellon Microsoft WISCONSIN UNIVERSITY OF WISCONSIN–MADISON ETH zürich

Video Recordings are Ubiquitous

Massive video recordings are happening everywhere

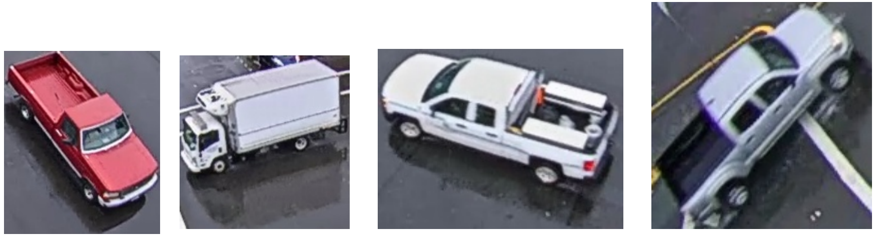


Key Application: Querying Objects in Videos

- Find all trucks among traffic videos in a city last week
- Find all people in garage videos in a company last night

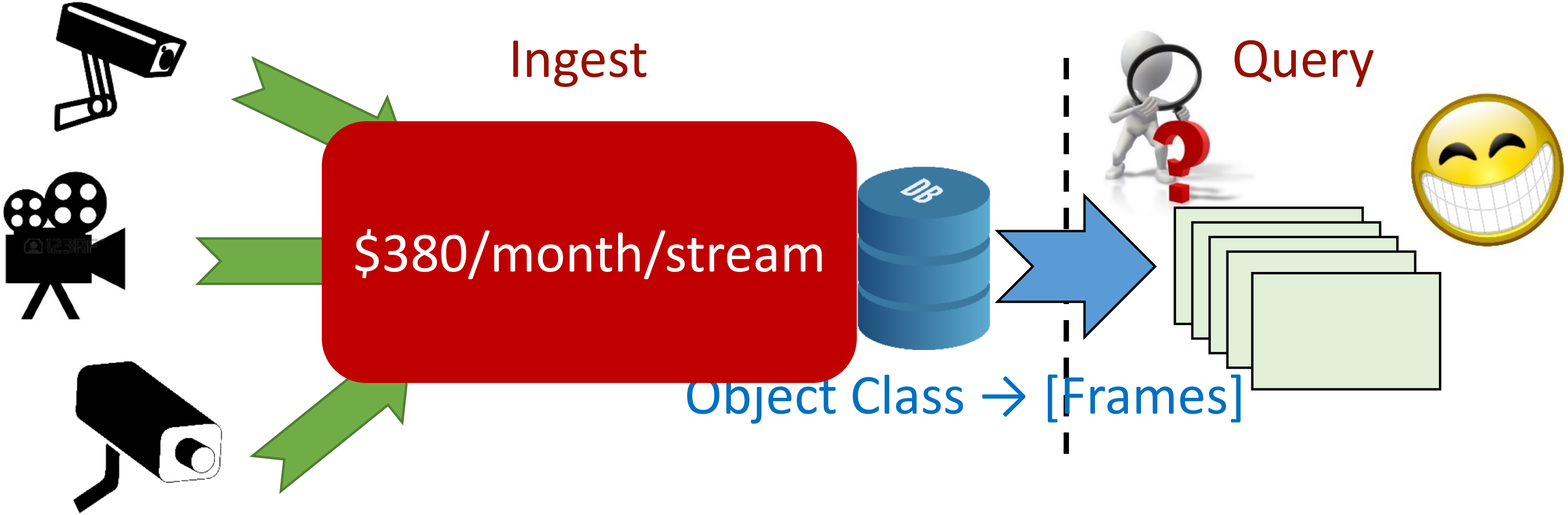
→ *Query execution requires running <u>detector & classifier CNNs</u>*

→ *It is slow and costly on massive videos*



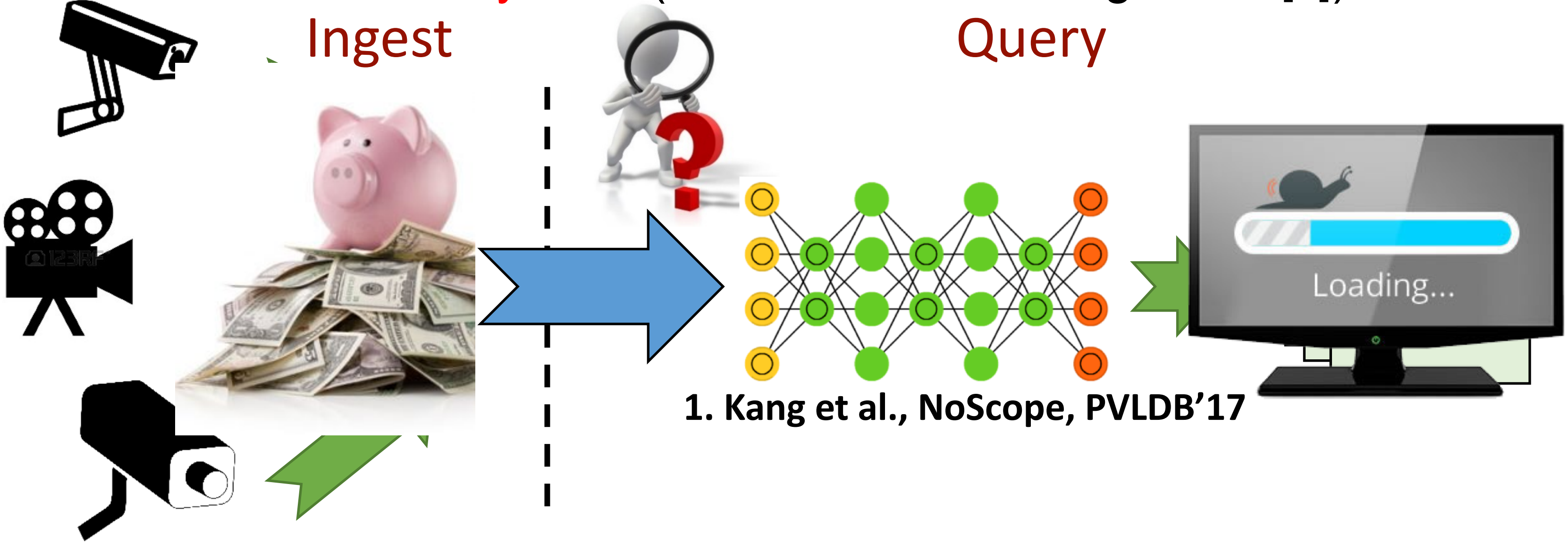
Ingest Time Analysis: Too Costly

- Analyzing live videos at ingest time can make query fast
  - But it is costly
  - Potentially wasteful (ingest all garage cameras vs. query one)



Query Time Analysis: Too Slow

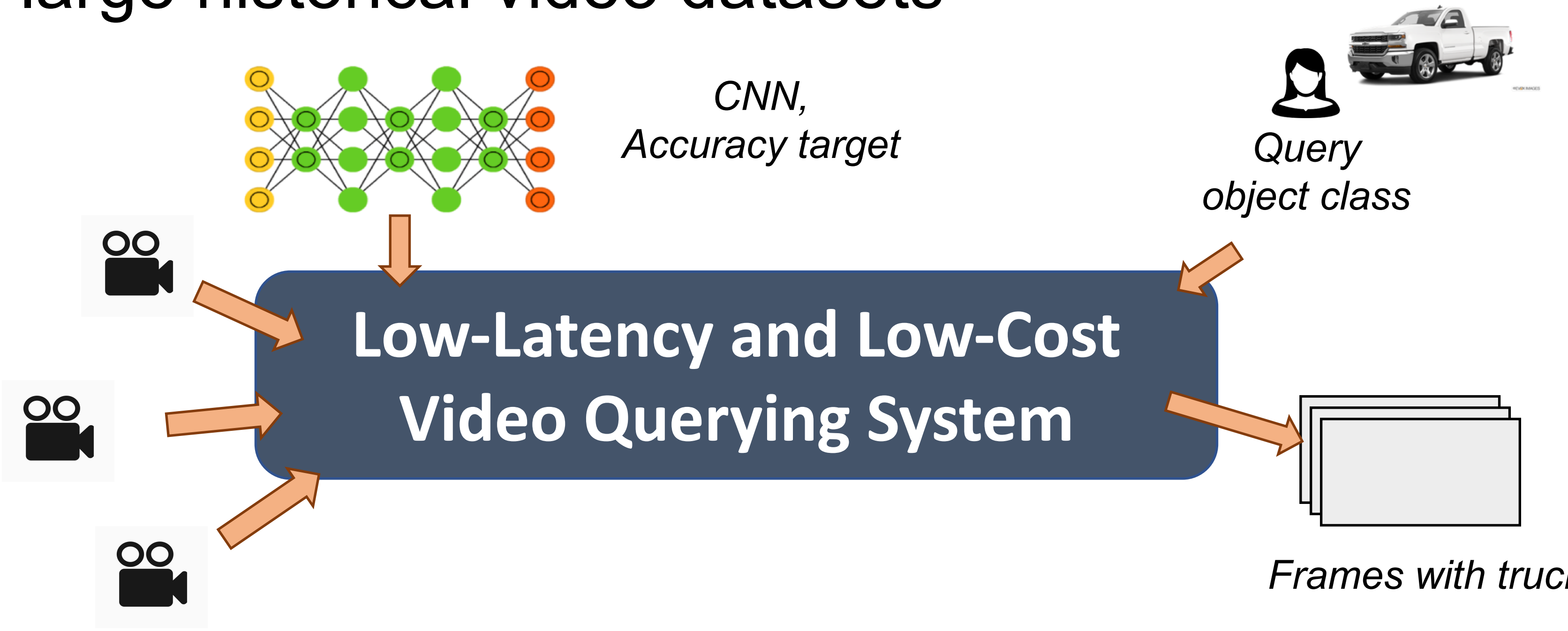
- Analyzing videos at query time can save cost
    - Frame down-sampling / skipping
    - CNN specialization / cascading
    - But it still very slow (5 hr for a month-long video [1])



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

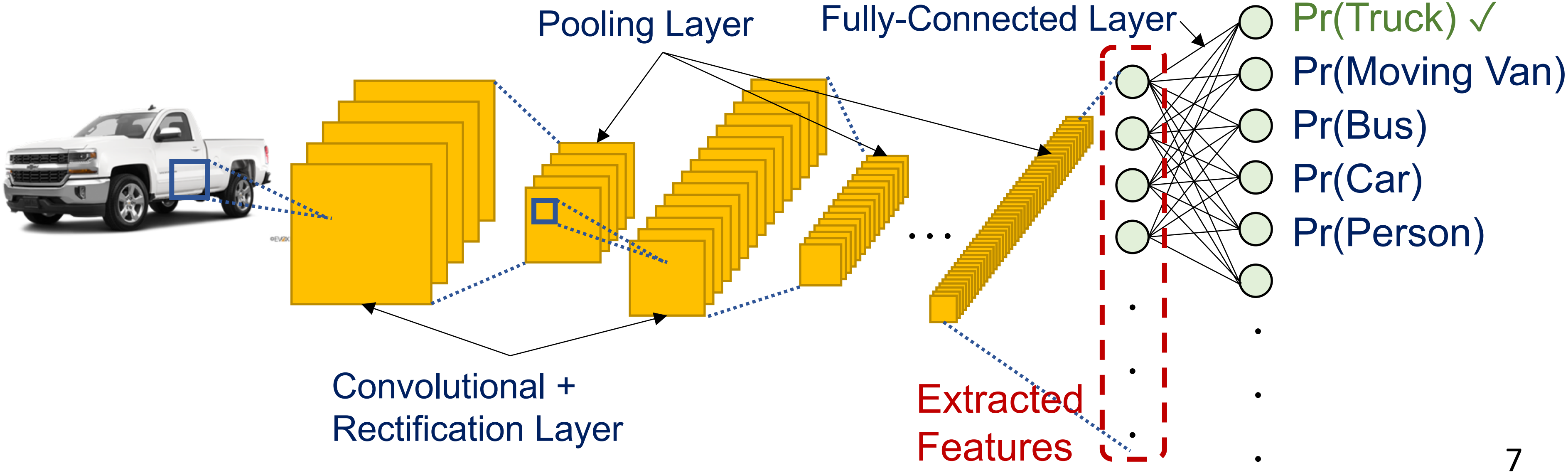
Our Goal

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



Background: Convolutional Neural Networks

- A Convolutional Neural Network (CNN) outputs the probability of each class
- Based on the extracted features (high-level representation)



Focus System: Low-latency query with low-cost ingest

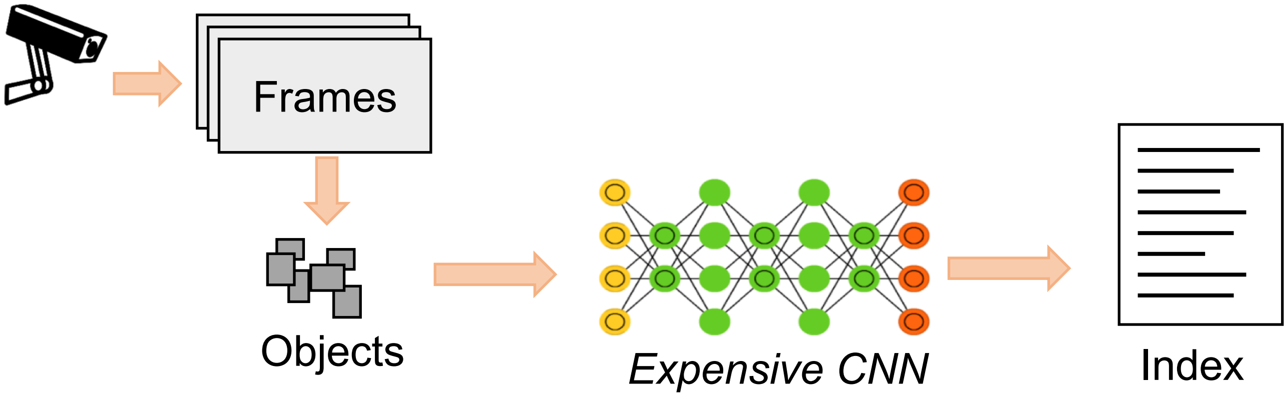
- Approximate indexing via cheap ingest
- Redundancy elimination for fast query
- Trading off ingest cost vs. query latency

Focus System: Low-latency query with low-cost ingest

- Approximate indexing via cheap ingest
- Redundancy elimination for fast query
- Trading off ingest cost vs. query latency

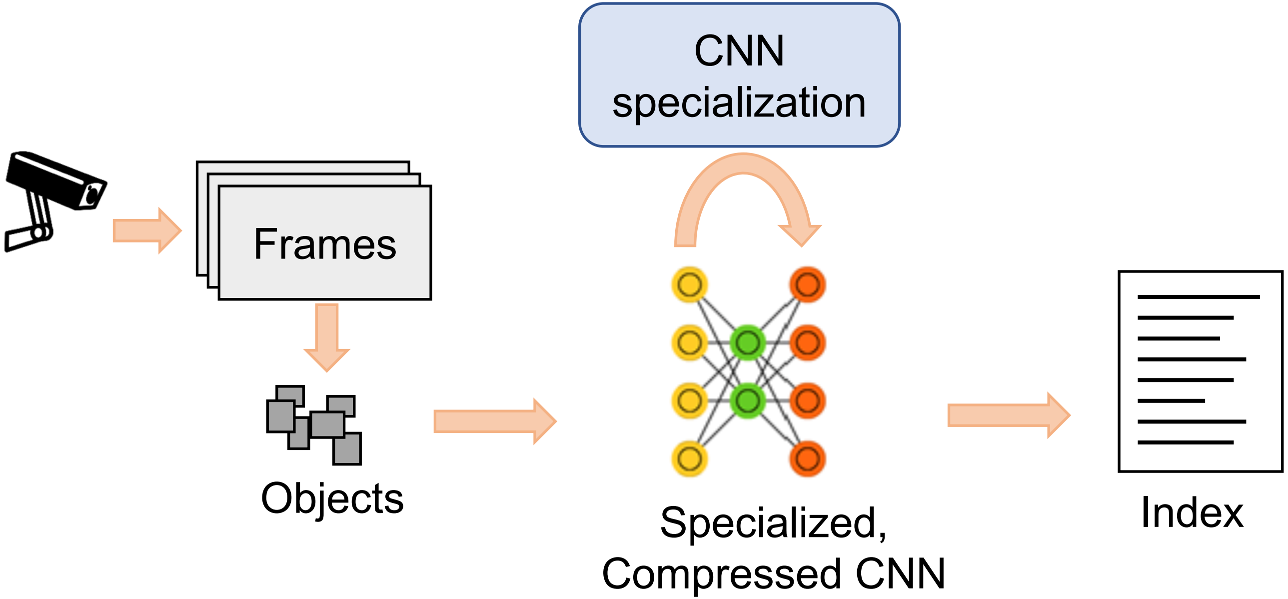
Low-Cost Ingestion: Cheaper CNNs

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



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

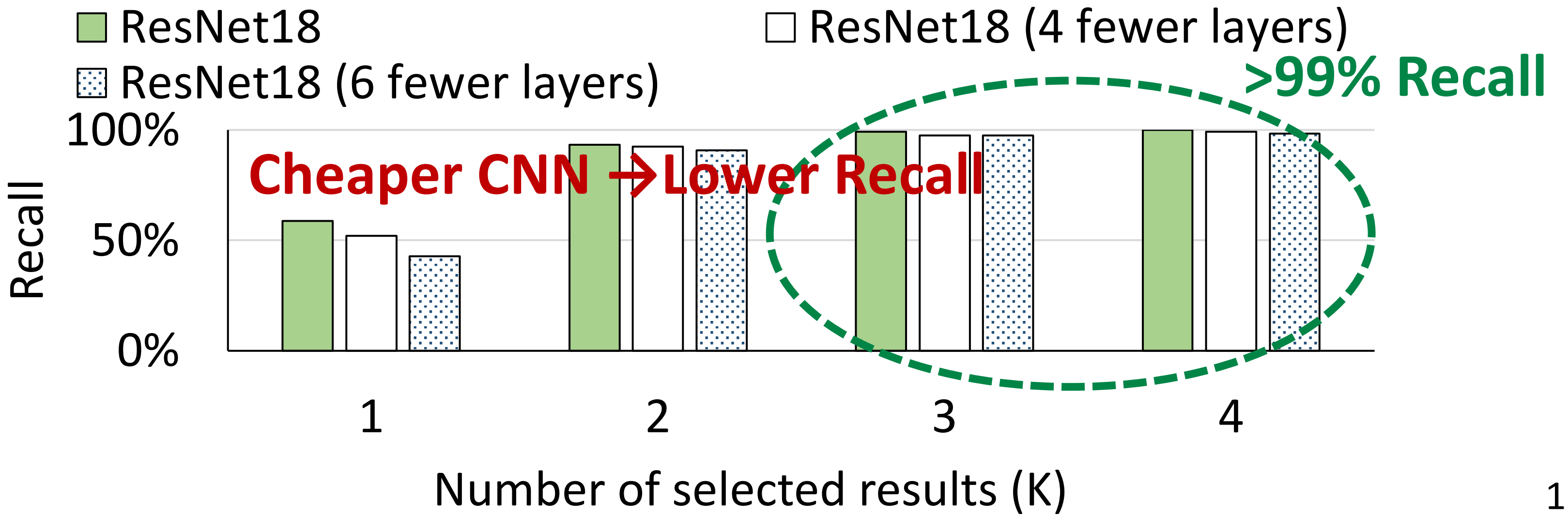
Rank	Expensive CNN	Cheap CNN
1	**Truck**	**Moving Van**
2	Moving Van	**Airplane**
3	Passenger Car	**Truck**
4	Recreational vehicle	Passenger Car

Recall, Precision and Top-K Results

Recall: Fraction of relevant objects that are selected

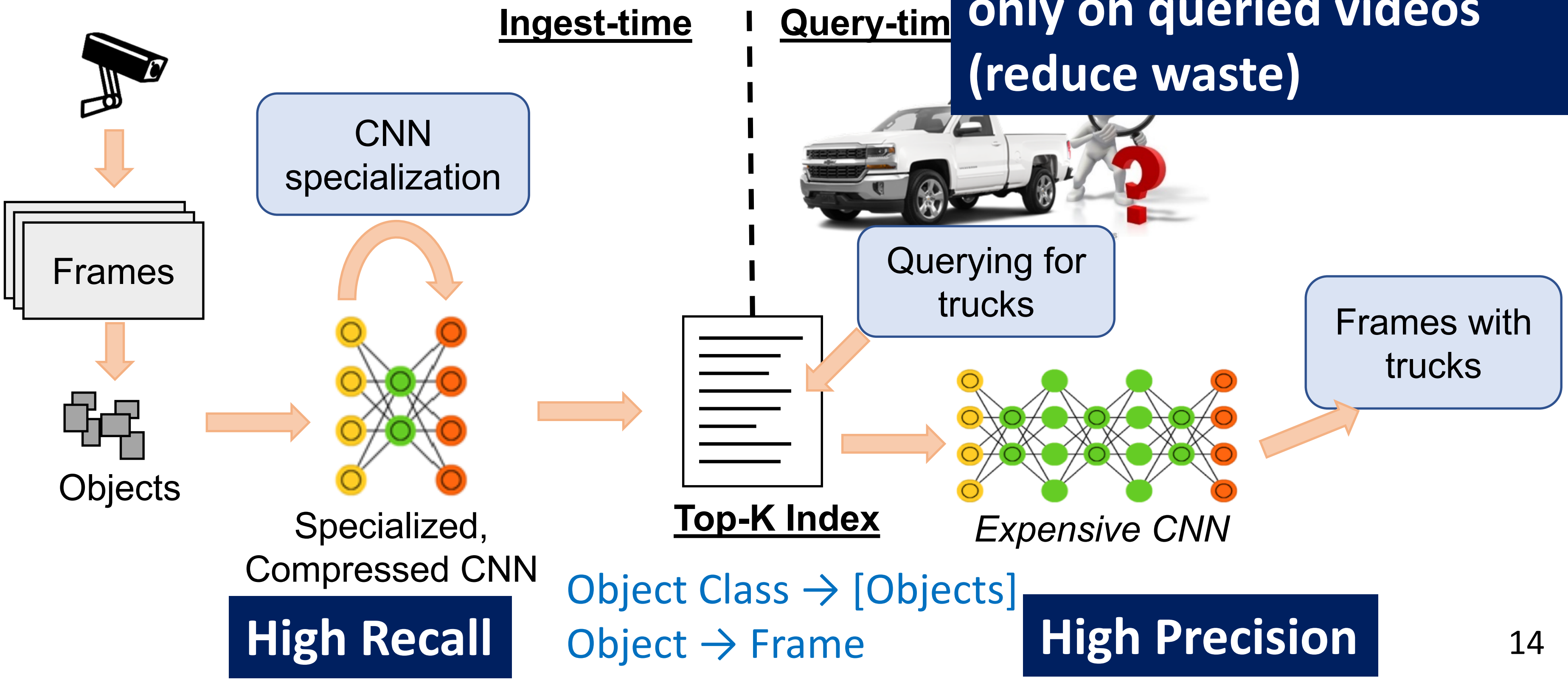
Precision: Fraction of selected objects that are relevant

Ground-truth CNN: YOLOv2



Solution: Split Ingest- and Query-time Work

Query-time work is done only on queried videos (reduce waste)

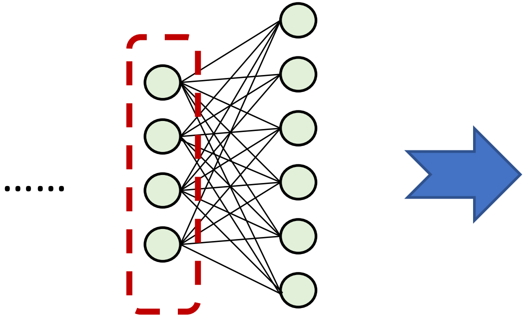


Focus System: Low-latency query with low-cost ingest

- Approximate indexing via cheap ingest
- Redundancy elimination for fast query
- Trading off ingest cost vs. query latency

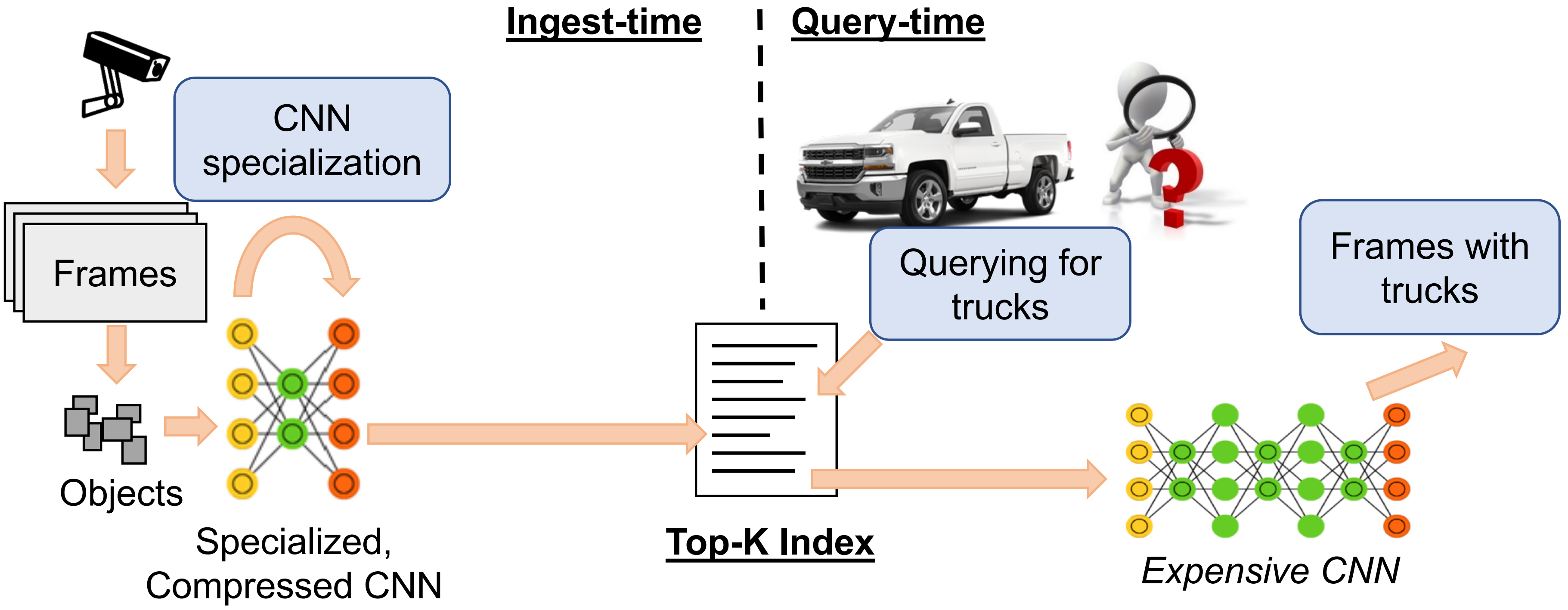
Low-Latency Query: Redundancy Elimination

- Approximate indexing ➔ non-trivial work at query time
  - A larger K ➔ more query-time work
- Images with similar feature vectors are visually similar
- Minimize the work at query time ➔ clustering similar objects based on the extracted features



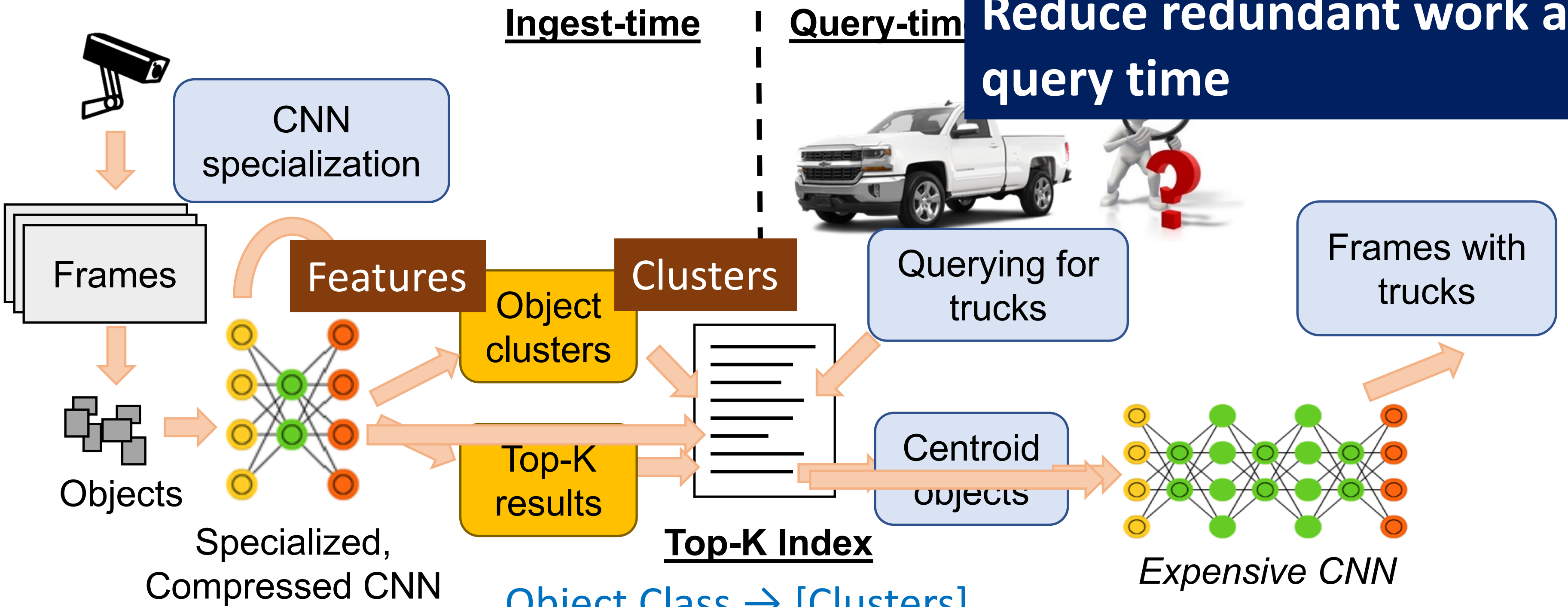
Extracted Features

Adding Feature-based Clustering



Adding Feature-based Clustering

Reduce redundant work at query time



Object Class → [Clusters]

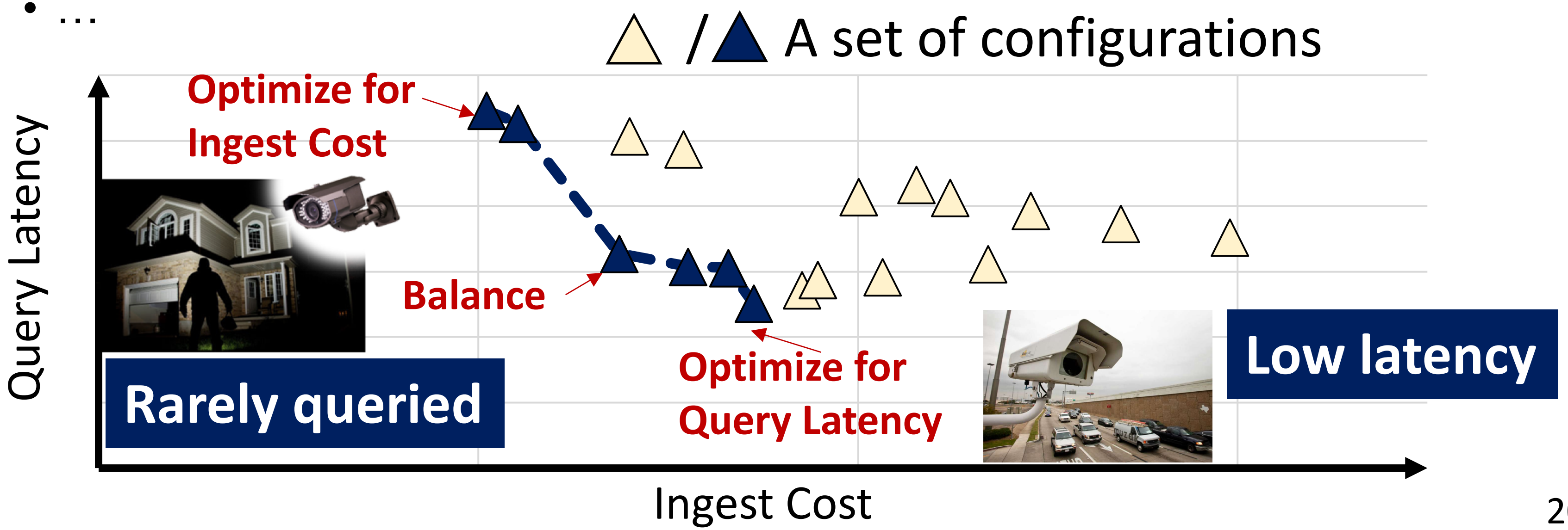
Cluster → Centroid, [Objects], [Frames]

Focus System: Low-latency query with low-cost ingest

- Approximate indexing via cheap ingest
- Redundancy elimination for fast query
- Trading off ingest cost vs. query latency

Ingest Cost vs. Query Latency

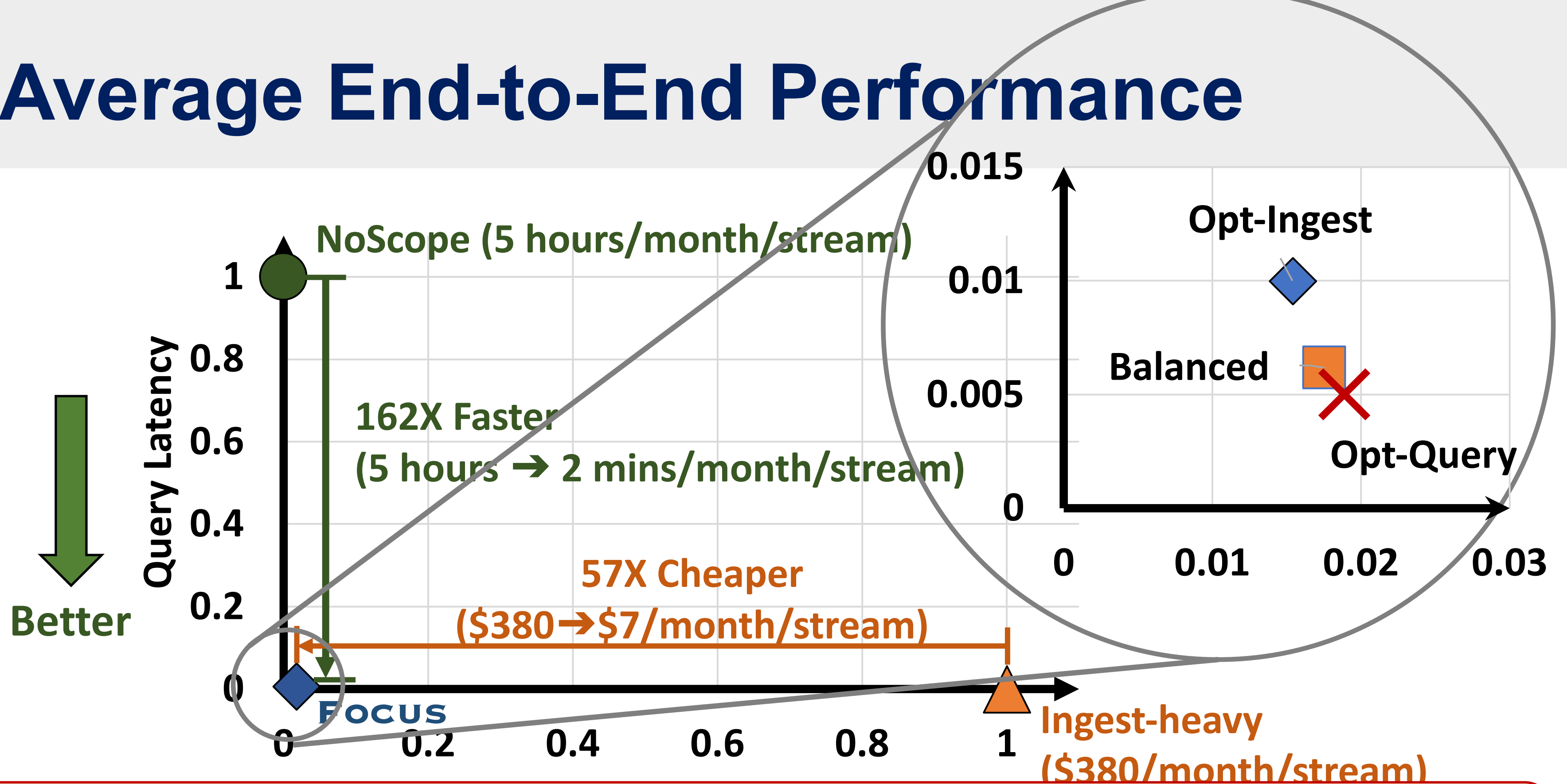
- Parameter selection → trading off ingest cost vs. query latency
  - The cheap CNN at ingest time
  - K in the top-K approximate indexing
  - Clustering threshold for feature-based clustering
  - …



Experimental Setup

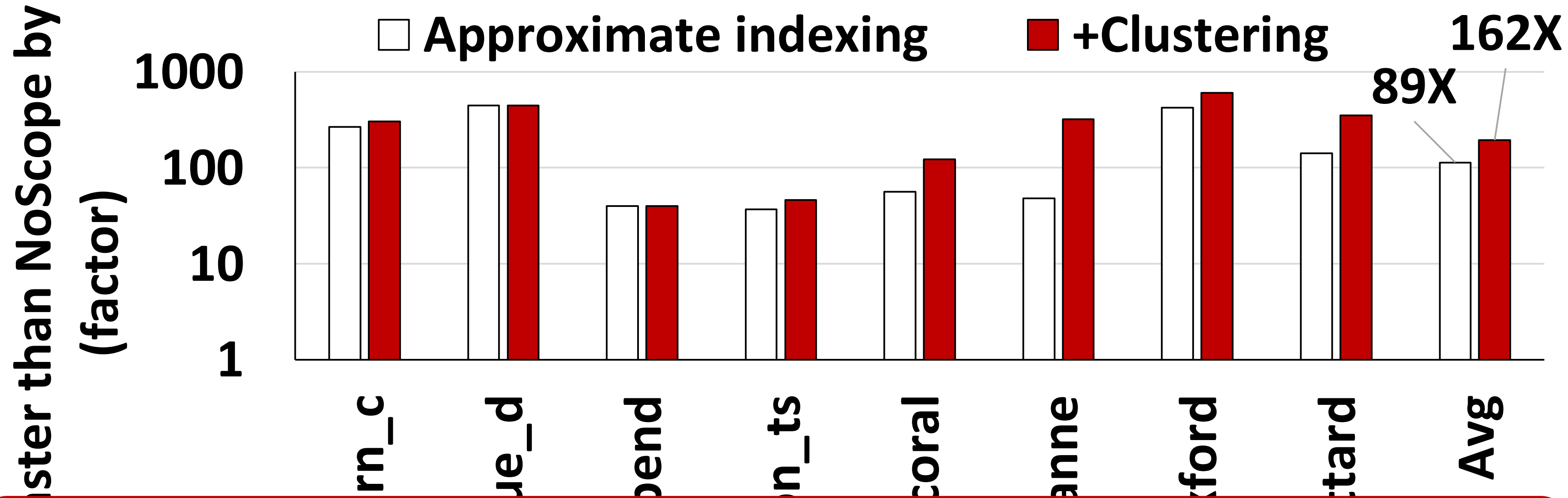
- **Video Datasets**
  - 11 live traffic and enterprise videos
  - Each video stream is evaluated for 12 hours
- **Accuracy Targets**
  - 99% recall and 99% precision w.r.t. YOLOv2
- **Baselines**
  - Ingest-heavy: Analyzes all frames with YOLOv2 at ingest time and stores the inverted index for query
  - NoScope [VLDB'17]: A query-optimized system that analyzes frames only at query time

Average End-to-End Performance



Focus achieves low-latency query with low-cost ingest

Effect of Different Components



Both techniques are important to Focus

Demo

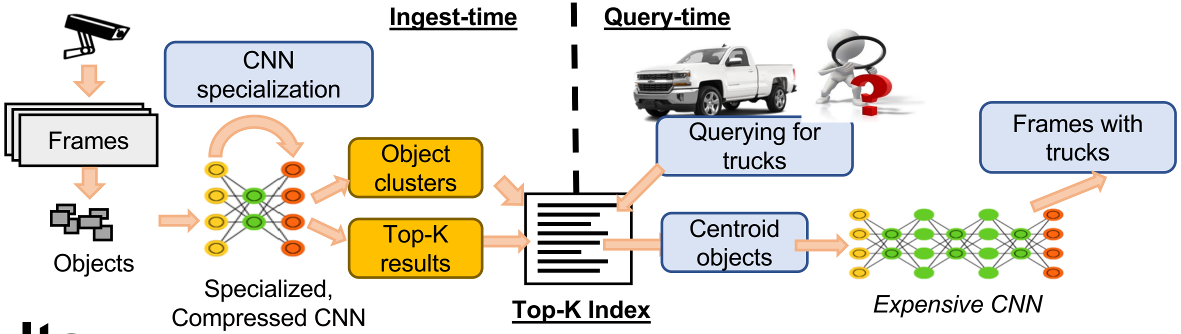
Available at: https://youtu.be/MNCspIm9U38

More in the Paper

- Characterization of real-world videos
- Implementation details
- Other applications
    - Process large and growing data with CNNs, such as audio, bioinformatics, geoinformatics
- More results
    - Trade-off alternatives
    - Sensitivity studies

Key Takeaways

- **Problem**: Querying objects in massive videos is challenging
- **Our Approach**: Low-latency query with low-cost ingest



- **Key Results**
  - 57X (up to 92X) cheaper than ingest-time-only solutions
  - 162X (up to 607X) faster than state-of-the-art, query-time-only solutions

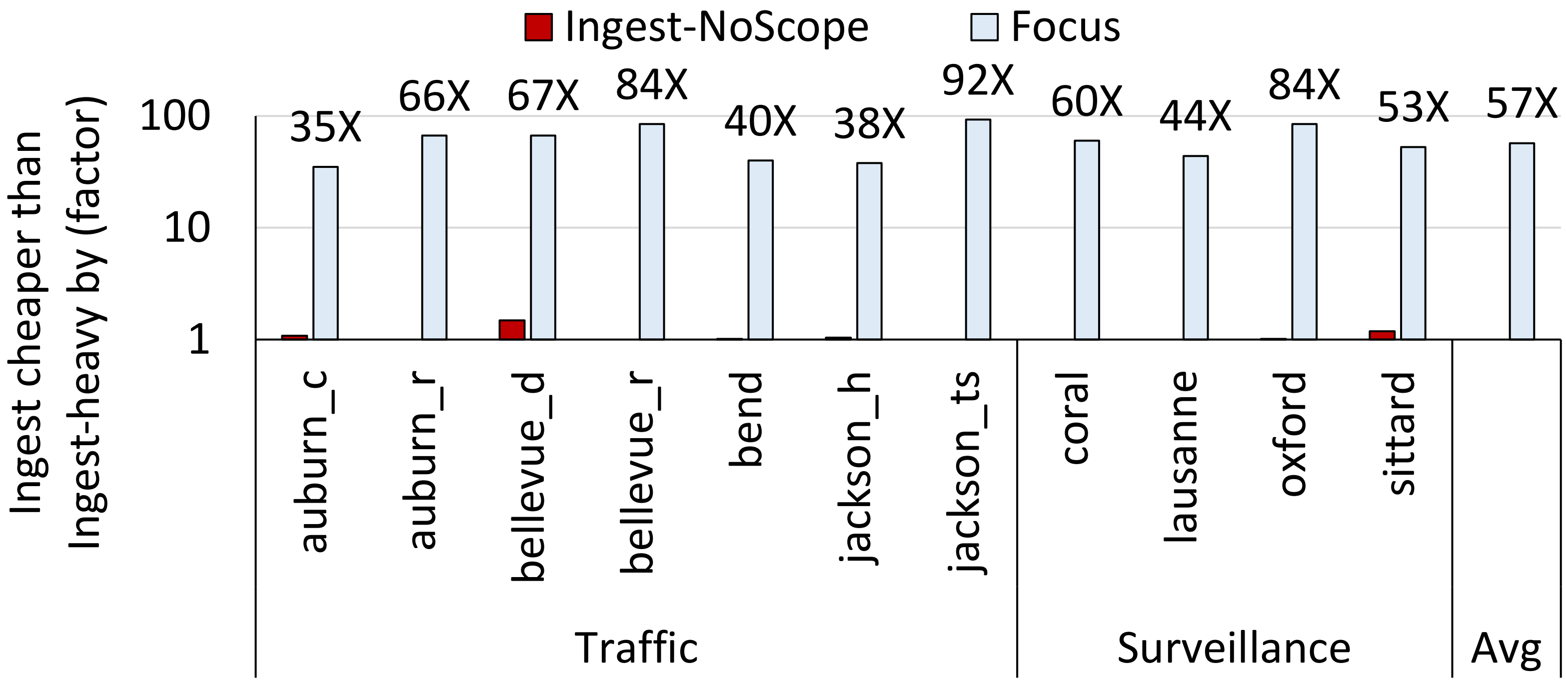
Focus: Querying Large Video Datasets with Low Latency and Low Cost

Kevin Hsieh

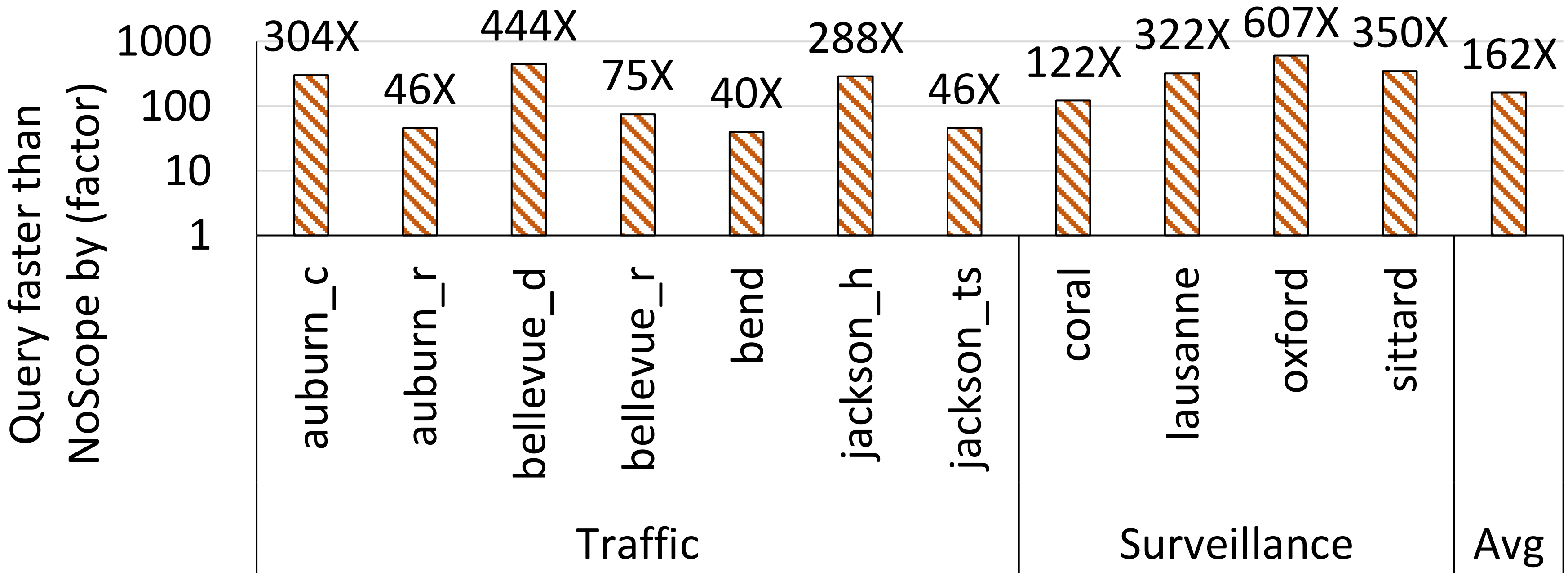
Ganesh Ananthanarayanan, Peter Bodik, Shivaram Venkataraman, Paramvir Bahl, Matthai Philipose, Phillip B. Gibbons, Onur Mutlu

Carnegie Mellon
Microsoft
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON
ETH zürich

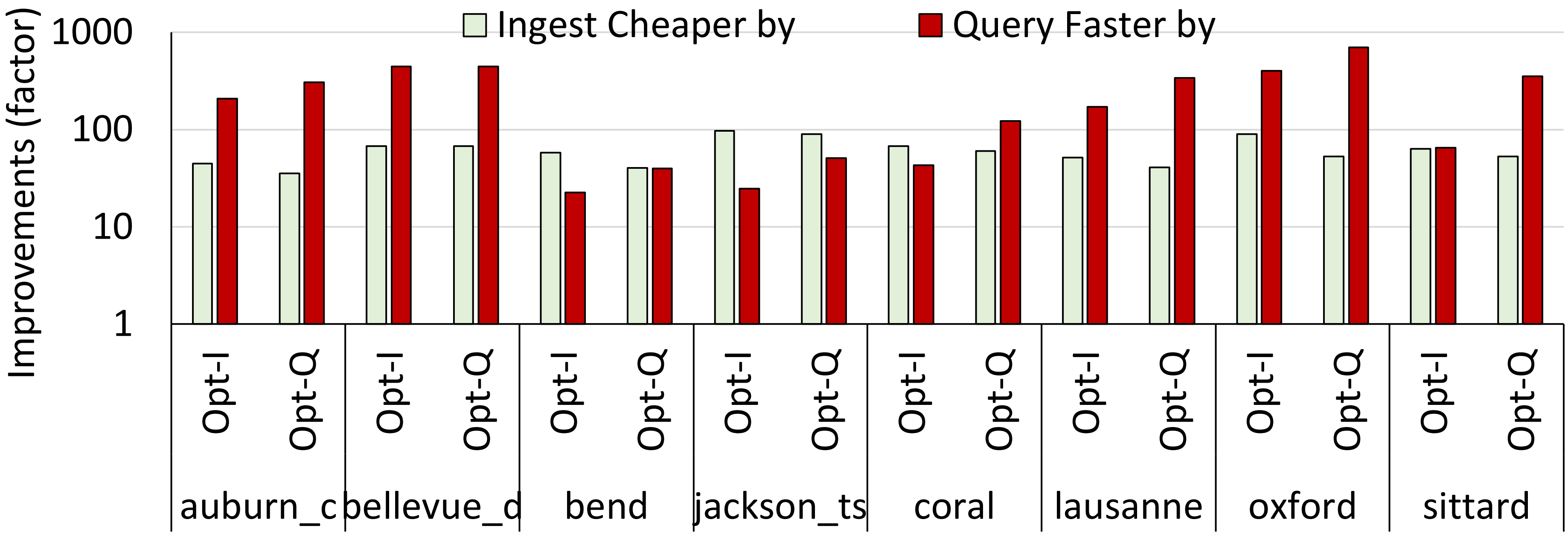
Ingest Cost by Video



Query Latency by Video



Trade-off Alternatives



Sensitivity – Number of Classes

- We study the sensitivity to the number of object class using 1,000 ImageNet classes

- The results show that Focus is
  - 15× faster in query latency
  - 57× cheaper in ingest cost than the baseline systems

Implementation Architecture