Agua: A Concept-Based Explainer for Learning Enabled Systems

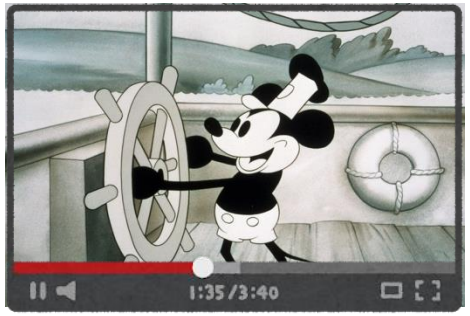
Sagar Patel, Dongsu Han, Nina Narodystka, Sangeetha Abdu Jyothi

UC Irvine

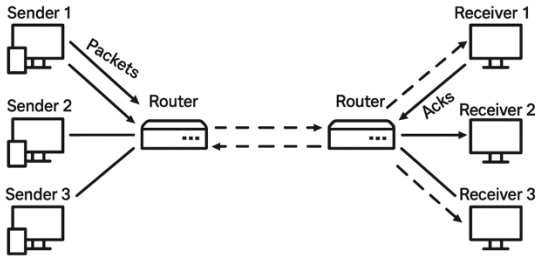
KAIST

vmware by Broadcom

Promise of Learning Enabled Systems

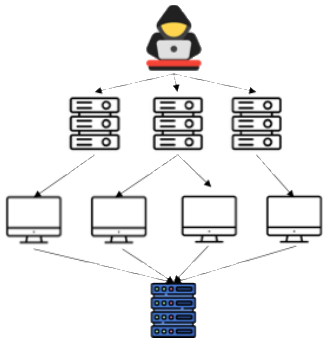


Adaptive Video Streaming:
Pensieve [SIGCOMM'17], Fugu [NSDI'19], …
Gelato [CoNEXT'24]

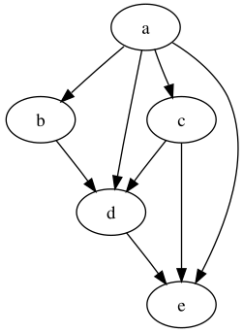


Congestion Control:
Aurora [ICML'19], Orca [SIGCOMM'20]…
Sage [SIGCOMM'22]

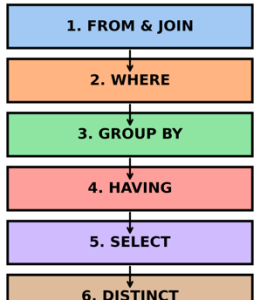
…



DDoS Detection:
GRU-RNN [NetSoft'18], …
LUCID [IEEE TNSM'20]



Cluster scheduling:
Decima [HotNets'18], Gavel [SIGCOMM'20]…
Sinan [SIGCOMM'22]

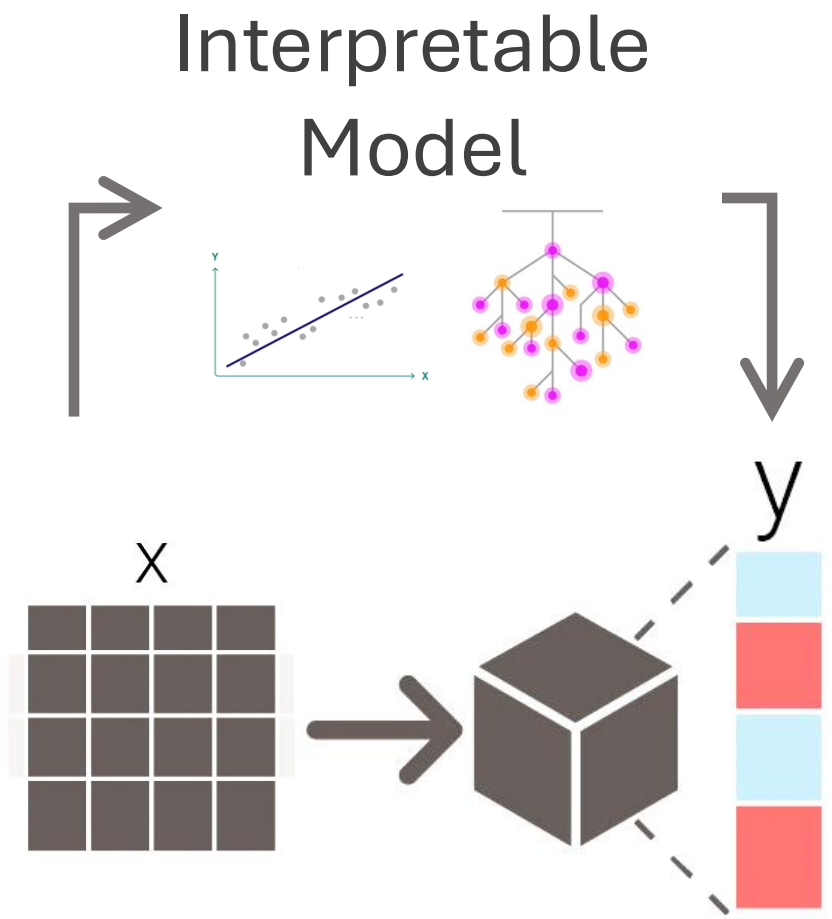


Query Optimization:
Neo [VLDB'18], Bao [SIGMOD'21]…
Lero [VLDB'23]

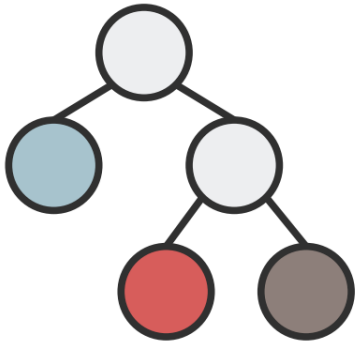
The Need for Explainability

- Operators are hesitant to deploy ML solutions despite high performance in lab
- Difficulty:
  - Understanding
  - Debugging
  - Trusting

Surrogate Functions as Explanations

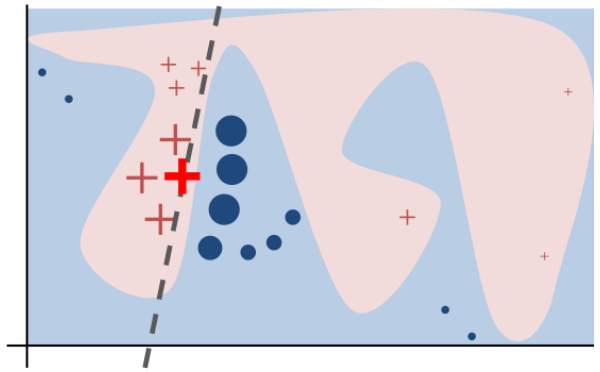


Feature-level Explainers



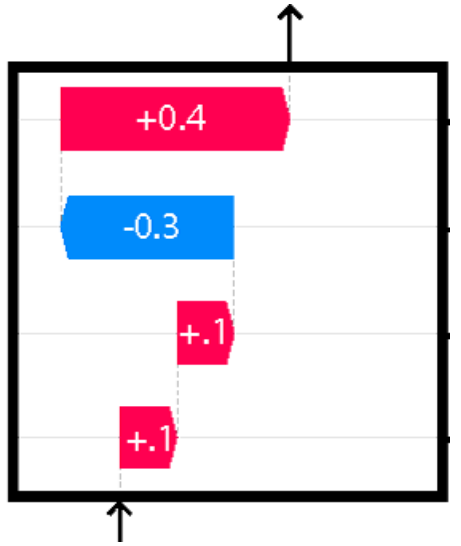
Decision Trees

Metis [SIGCOMM'20]
Trustee [CCS'22]



Local Linear Regression

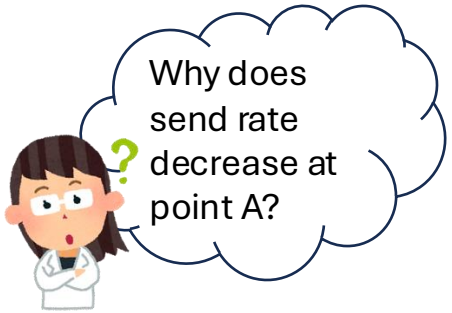
LIME [SIGKDD'16]

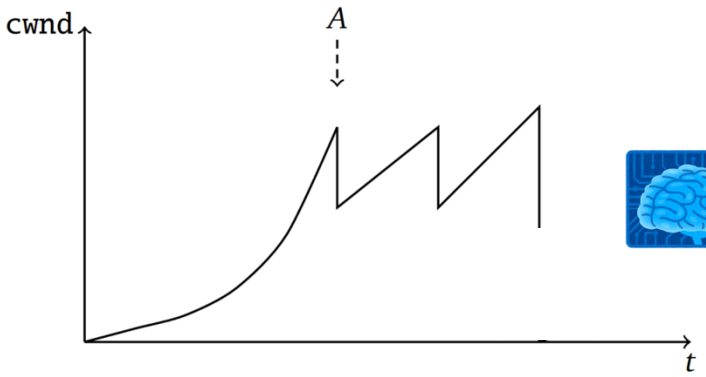


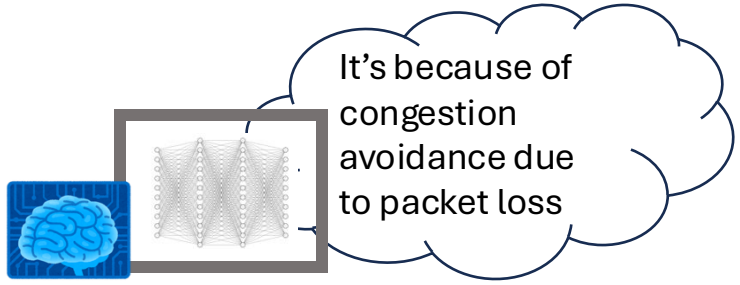
Game-theoretic

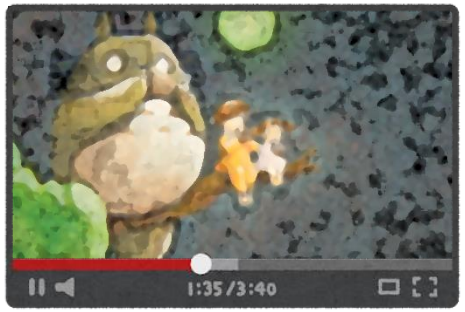
SHAP [KAIS'14]

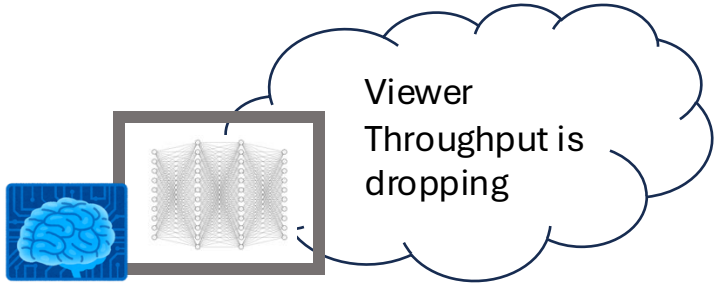
The Vision: Intuitive Understanding



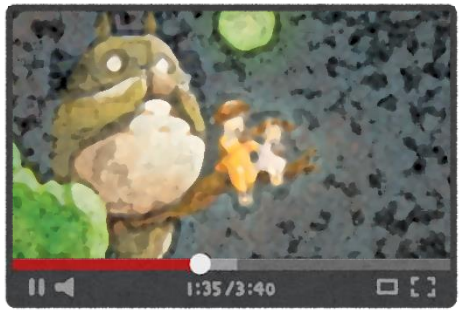


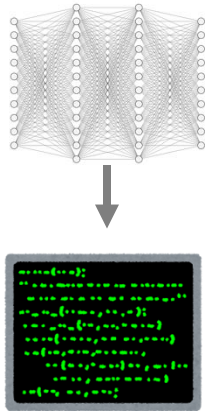


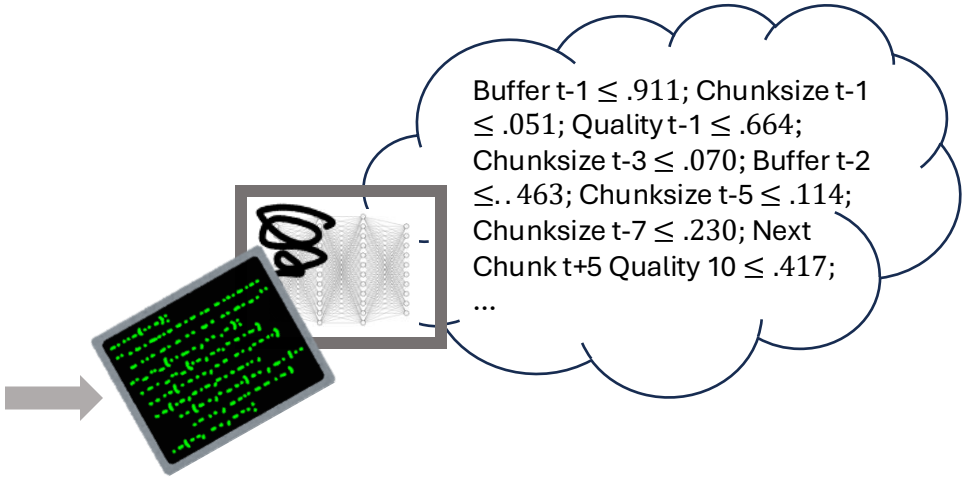




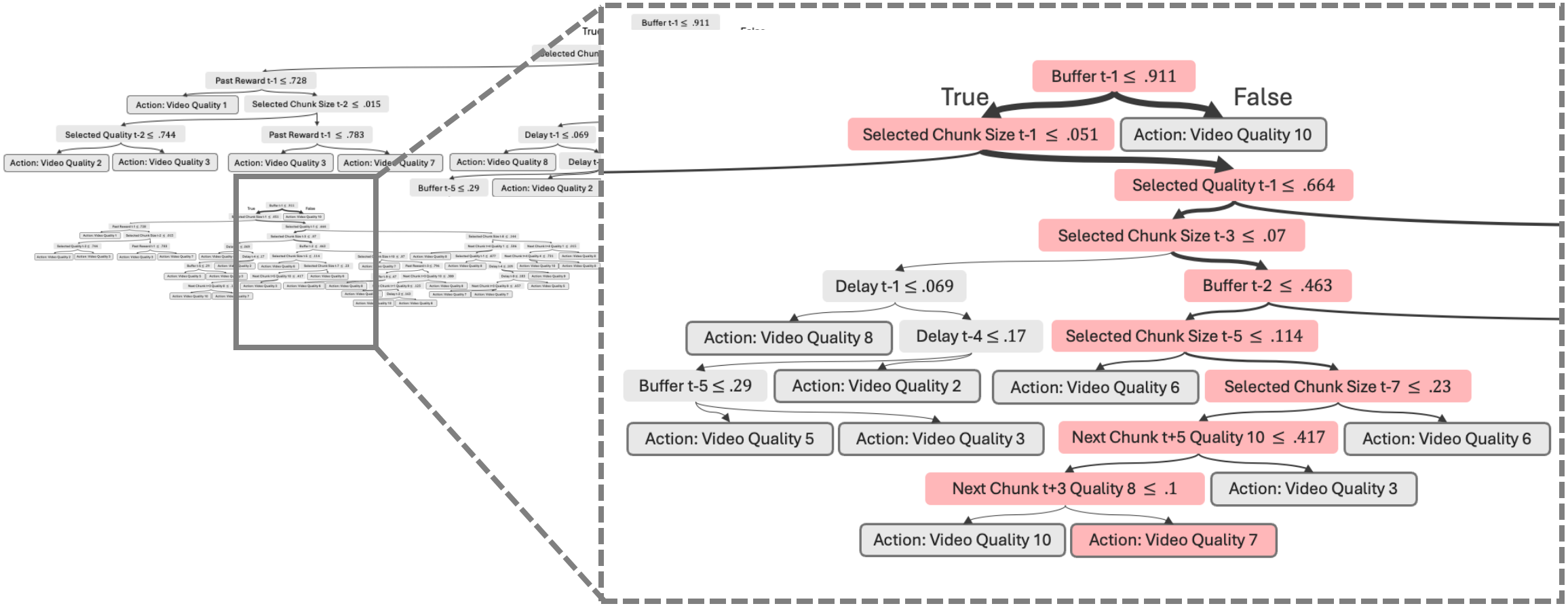
Reality of Current State-of-the-Art







Reality of Current State-of-the-Art

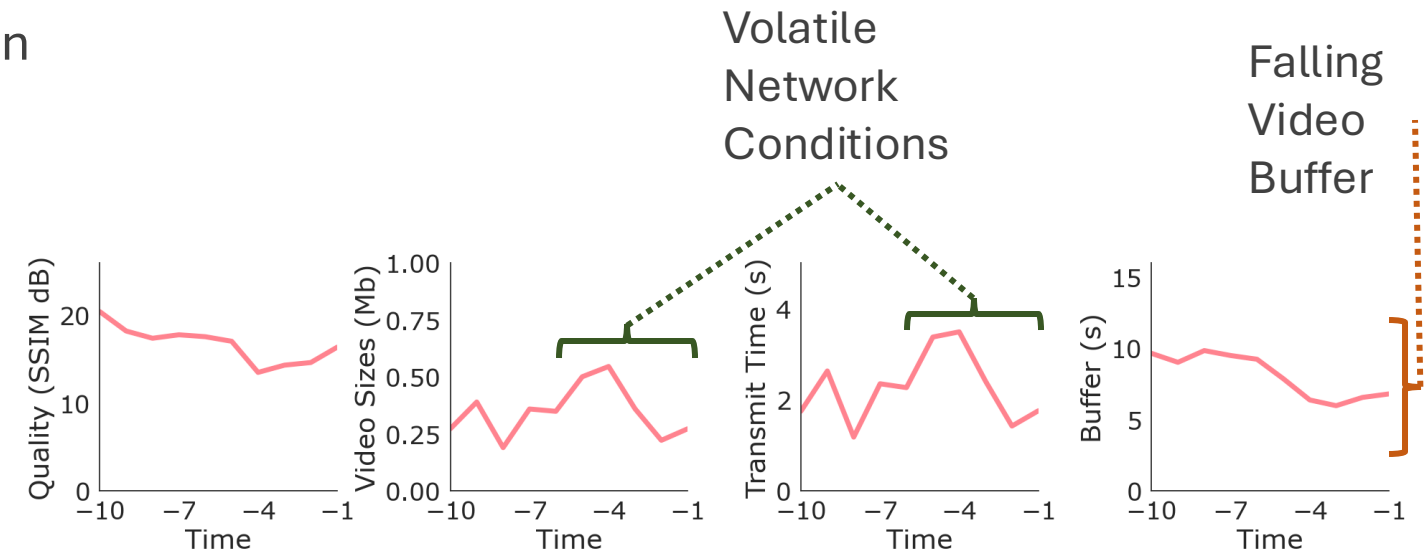


Limitations of Feature-level Explainers in Systems

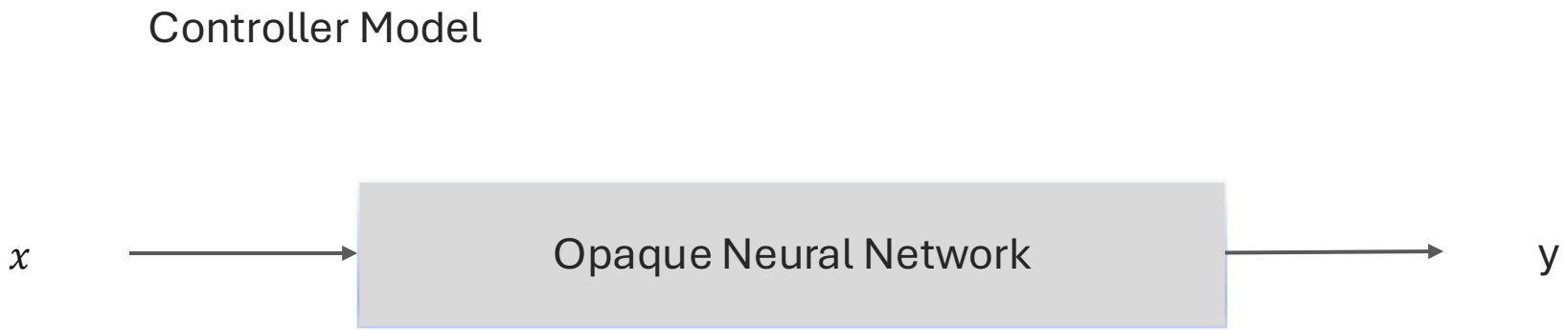
- Salient features as explanations
  - Can involve low-level system features
  - Miss trends and inter-feature interactions
  - Difficult to decode

Key Idea: Concepts as Units of Explanation

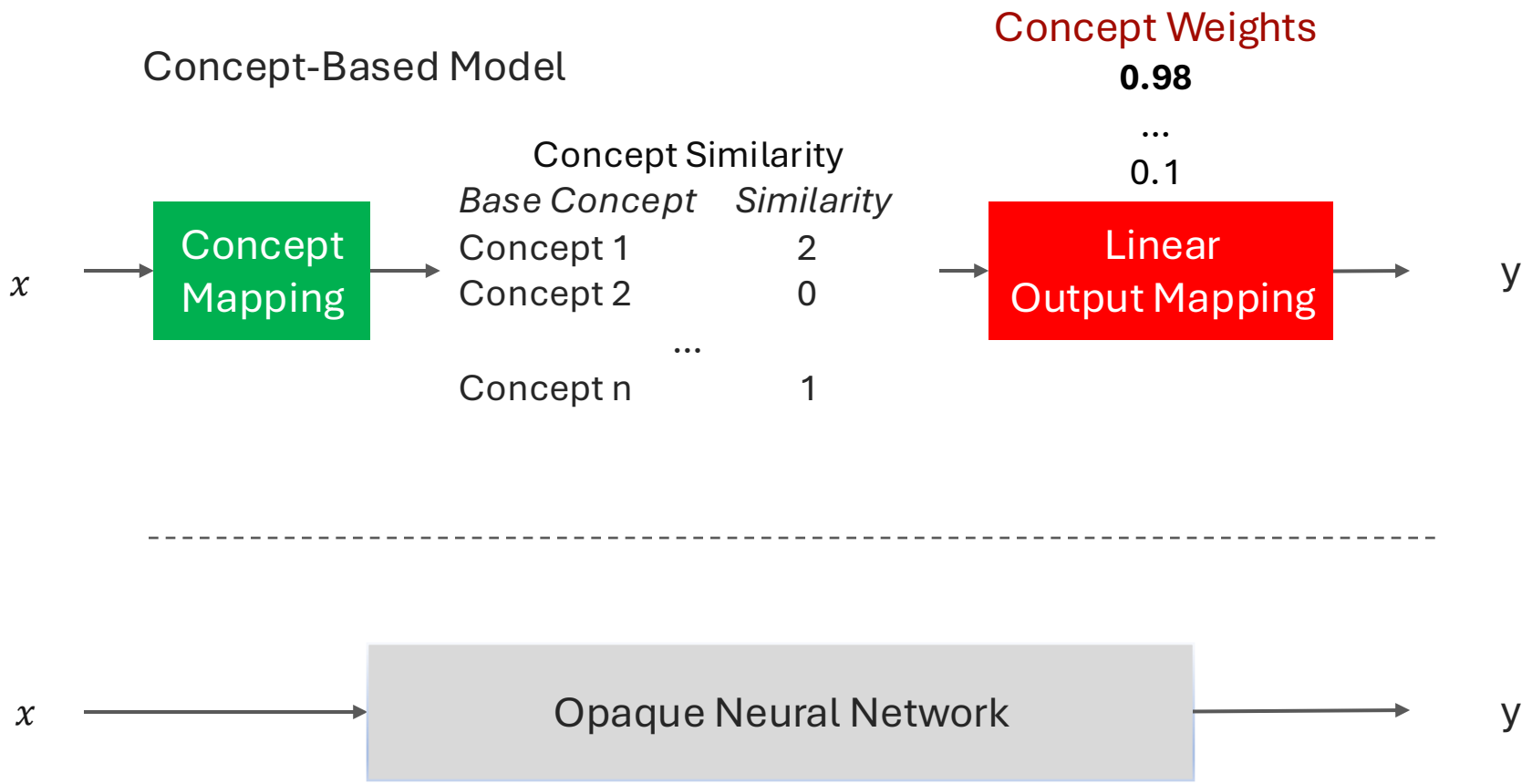
- Concepts
  - Human-understandable attributes that capture controller and environment characteristics
  - Can capture intricate patterns, trends, and behaviors in systems
  - Align with human intuition



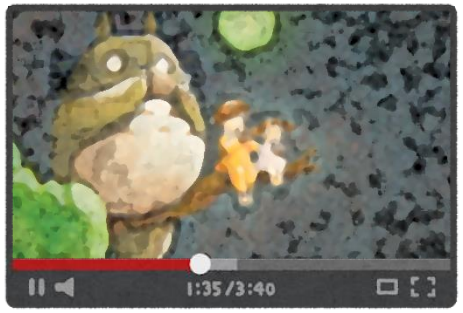
Key Idea: Surrogate Concept-Based Model

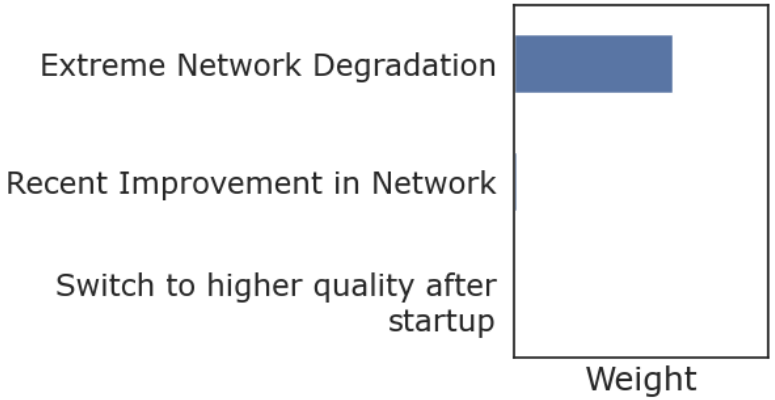


Key Idea: Surrogate Concept-Based Model

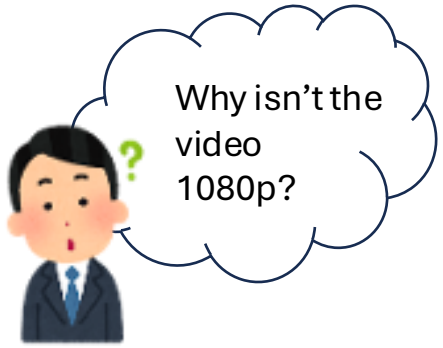


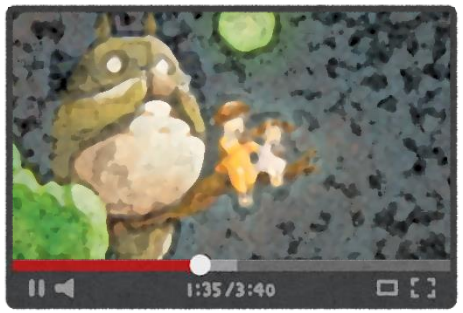
Concept-Based Explanation

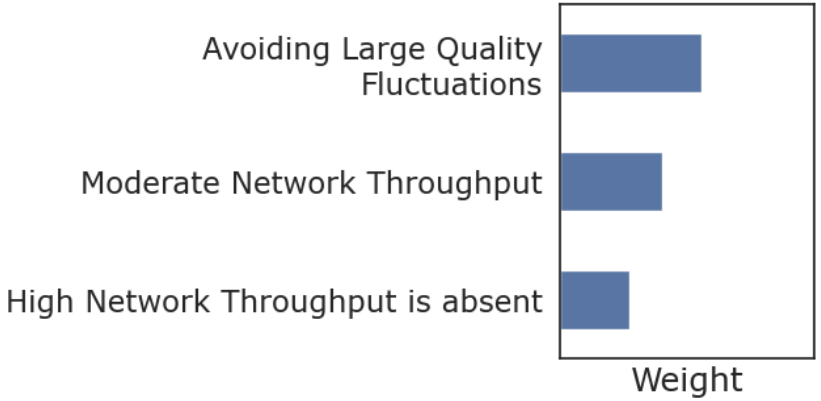




Factual Explanation

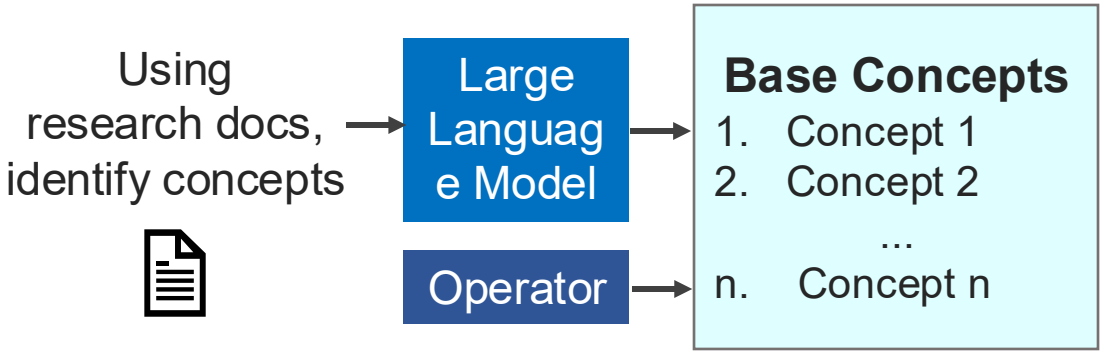






Counterfactual Explanation

Training Pipeline: 1. Base Concept Generation



Adaptive Bitrate Selection: A Survey

Yusuf Sani, Andreas Mauthe, and Christopher Edwards

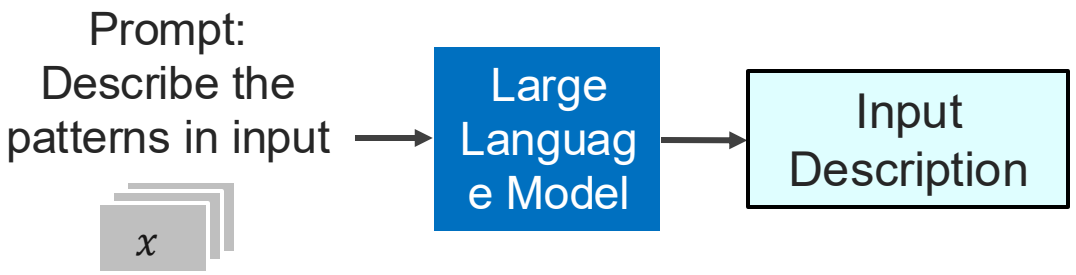
A comprehensive survey on machine learning for networking: evolution, applications and research opportunities

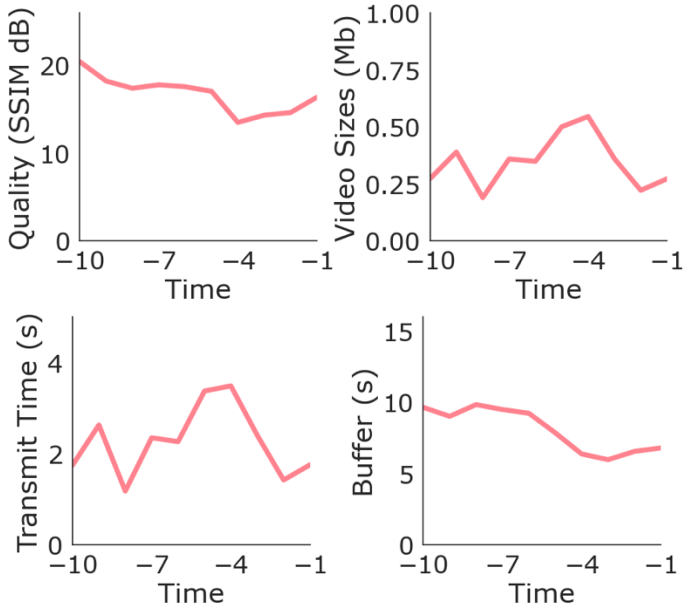
Raouf Boutaba[1*], Mohammad A. Salahuddin[1], Noura Limam[1], Sara Ayoubi[1], Nashid Shahriar[1], Felipe Estrada-Solano[1,2] and Oscar M. Caicedo[2]

Base Concepts

1. Volatile Network Conditions
2. Falling Buffer
...
n. Medium Throughput

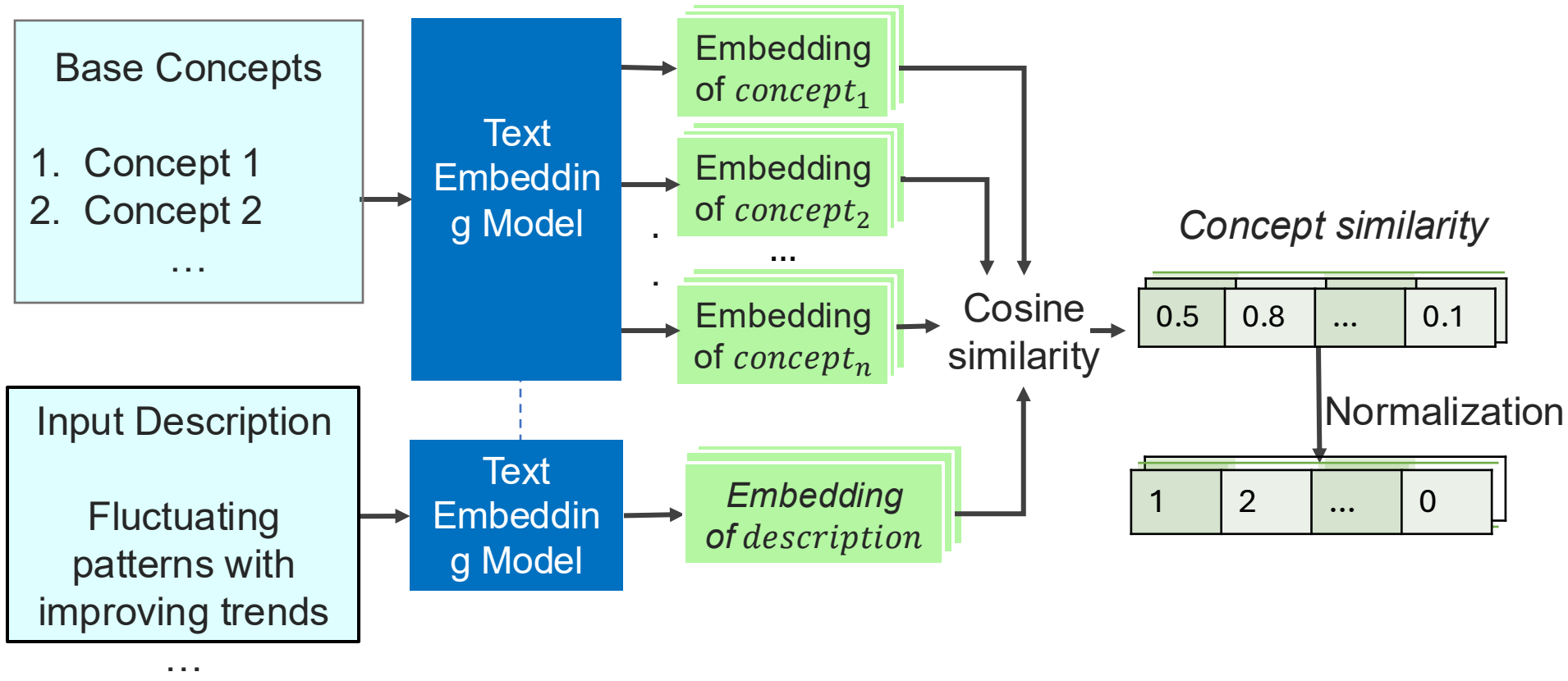
Training Pipeline: 2. Input Description Generation





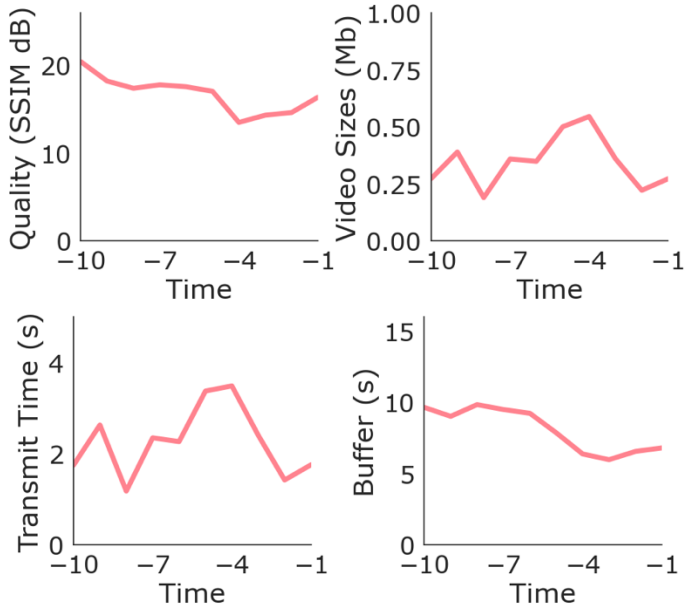
Network conditions:
- Initially starts off with a fluctuating pattern, as observed from the features "Transmission Time of Chunk."
- In the middle…
- Overall, the trend is improving, indicating the presence of stable network conditions.

Training Pipeline: 3. Target Concept Similarity Generation



Training Pipeline: 4. Training Concept Mapping

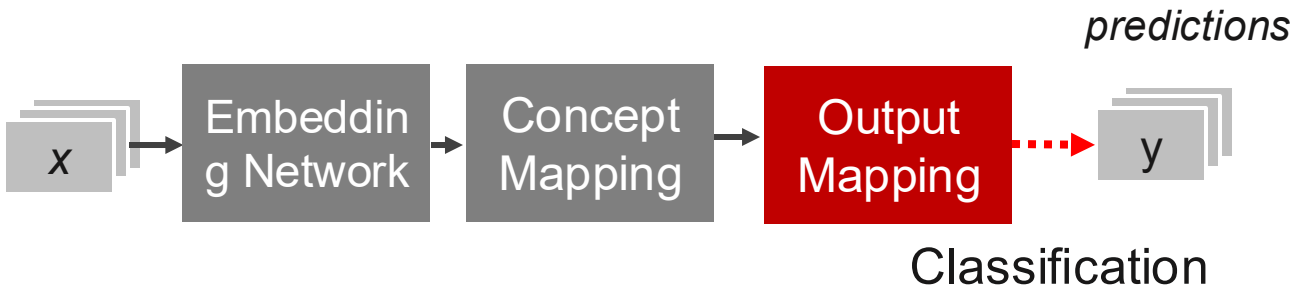
Multi-dimensional classification

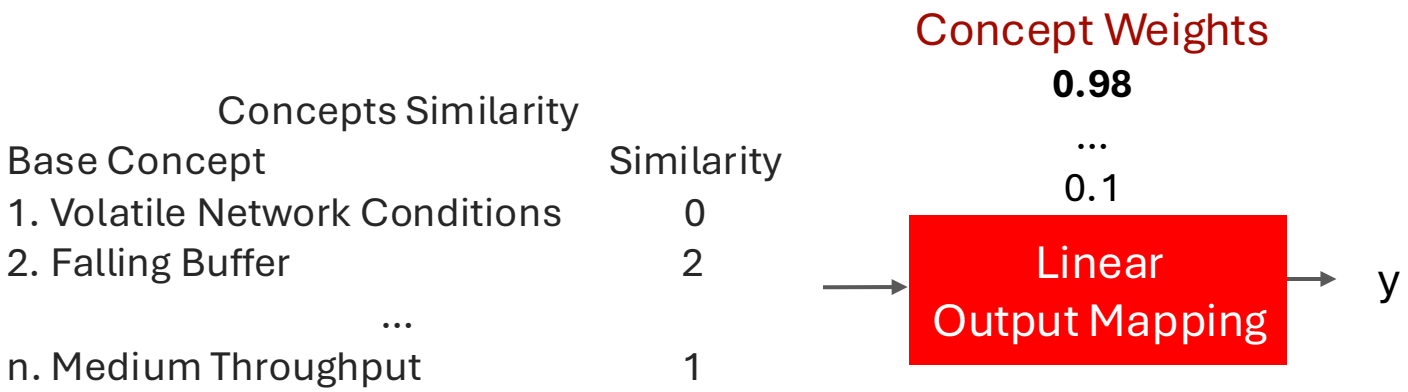


Concepts Similarity

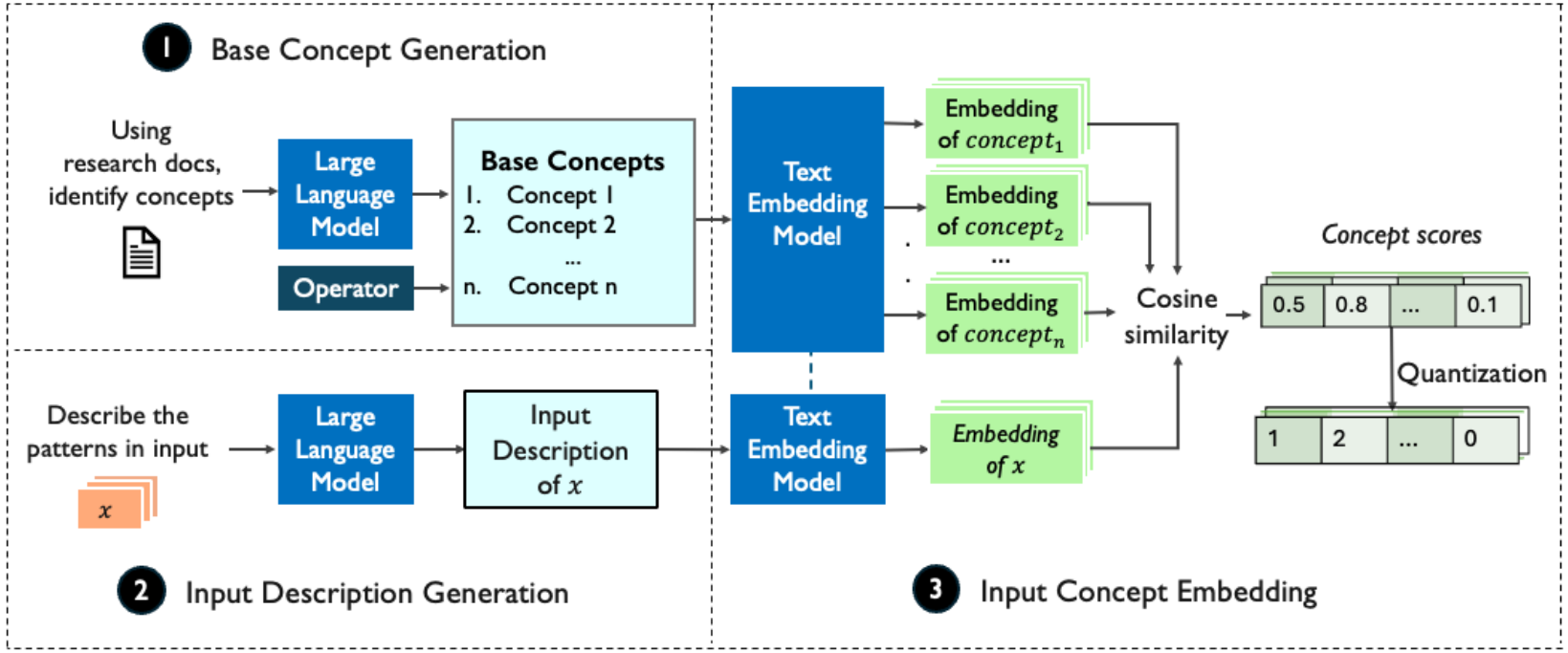
Base Concept	Similarity
1. Volatile Network Conditions	0
2. Falling Buffer	2
...	
n. Medium Throughput	1

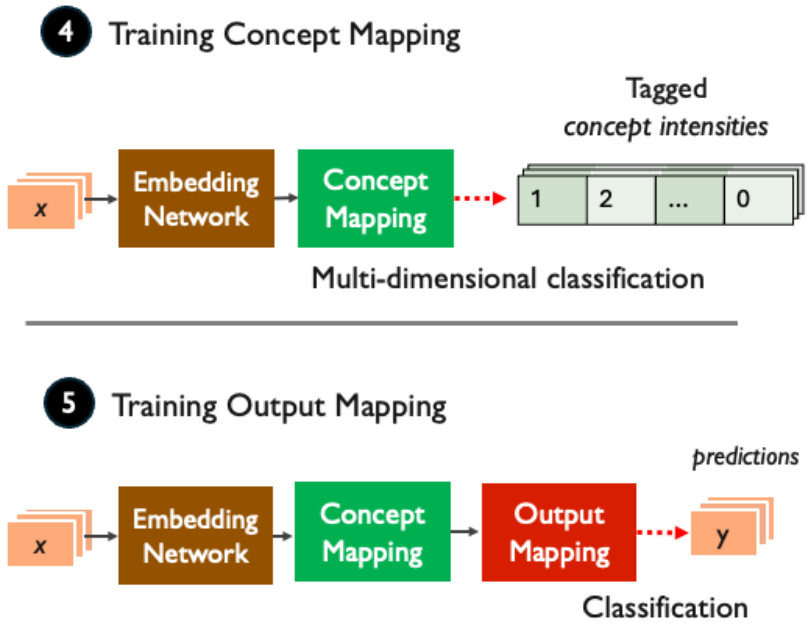
Training Pipeline: 5. Training Linear Output Mapping





Training Pipeline





Experimental Settings

- LLMs evaluated:
  - Closed-source: OpenAI GPT 4o
  - Open-source: Llama 3.3 70B
- Baseline:
  - Trustee [CCS'22]

Application	Controller	Learning Paradigm
Adaptive Bitrate Streaming	Gelato [CoNEXT'24]	Reinforcement Learning
Congestion Control	Aurora [ICML'19]	Reinforcement Learning
DDoS Detection	LUCID [IEEE TNSM'20]	Supervised Learning

Fidelity

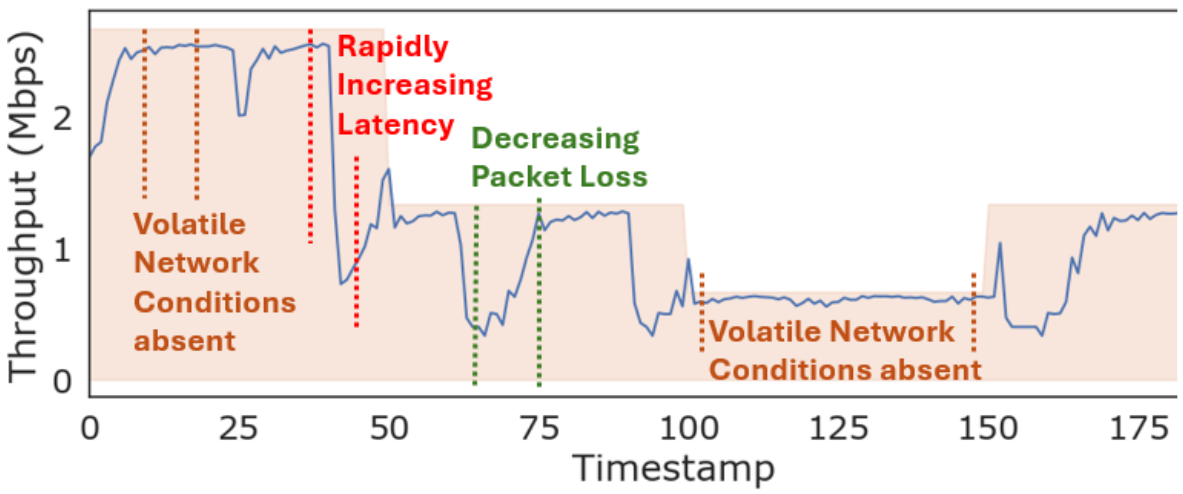
Application	Trustee		Agua	
	Full	Pruned	Llama 3.3 70B	GPT 4o
Adaptive Bitrate Streaming	0.946	0.949	0.982	**0.983**
Congestion Control	0.215	0.235	0.932	**0.936**
DDoS Detection	0.991	0.977	0.996	**1.000**

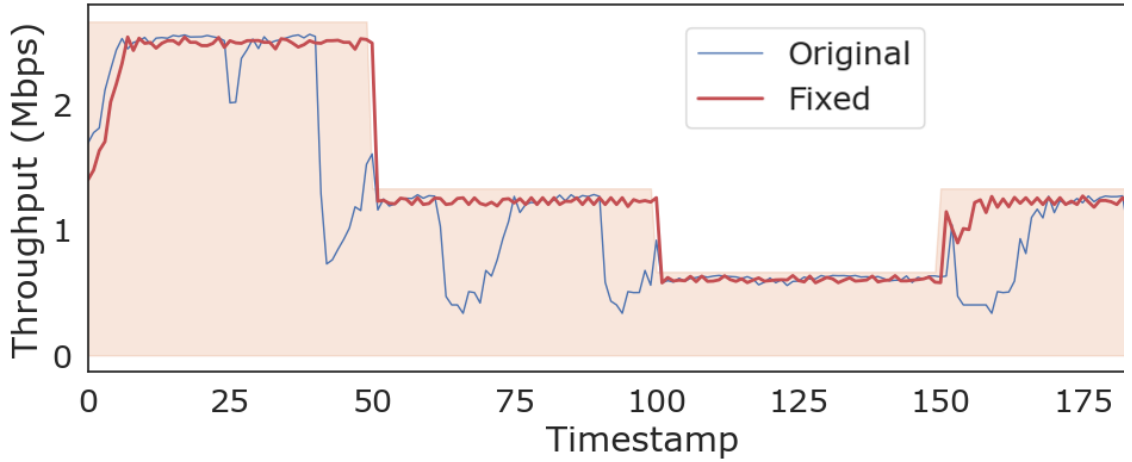
Use Cases

1. Debugging
2. Data shift detection
3. Concept-level retraining
4. Dataset expansion

Use Cases: Debugging

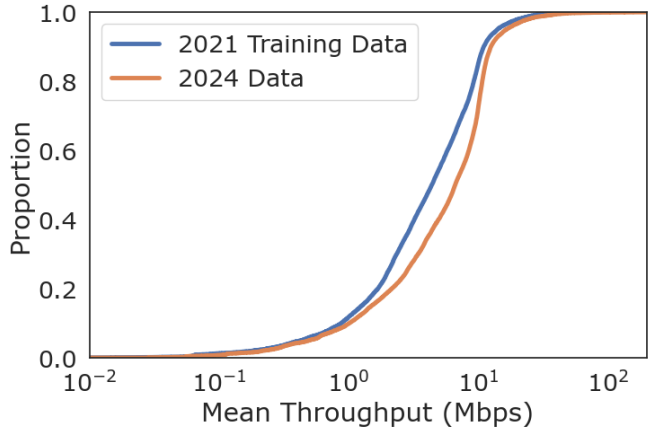
Aurora Throughput

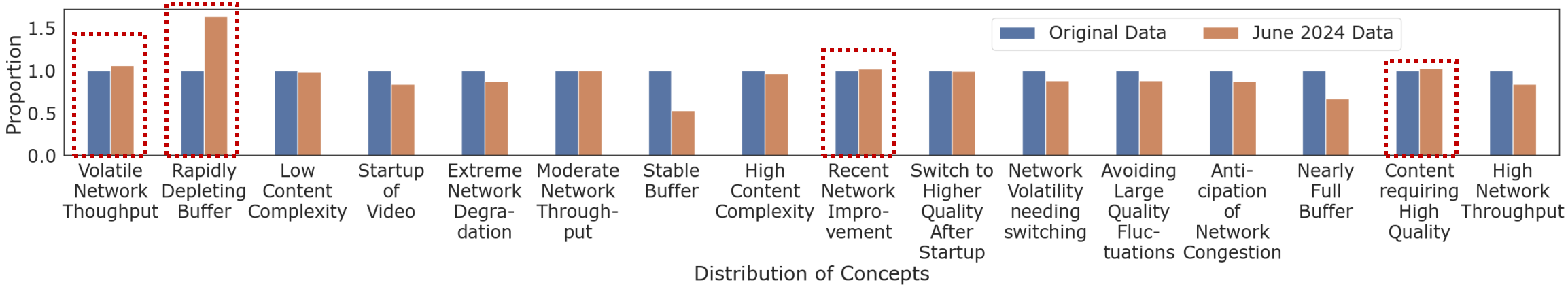




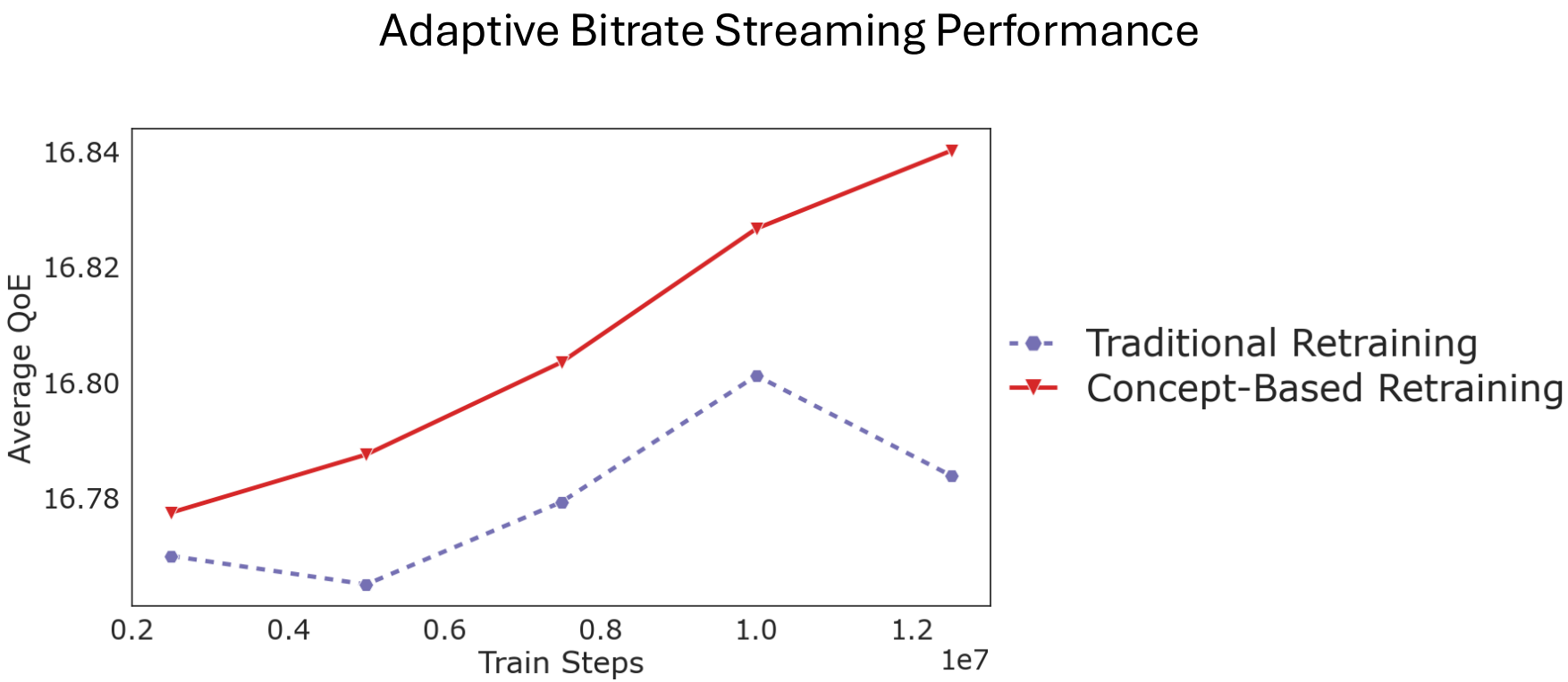
Use Cases: Distribution Shift Analysis

Gelato Datasets





Use Cases: Concept-Aware Retraining



Summary

- With Agua, we propose to move beyond feature-level explainability and towards an operator-aligned concept-level view
- Agua builds a surrogate concept-based model, attaining high fidelity and interpretability across
  - Congestion control, DDoS detection, adaptive video streaming
- Agua works with neural network controllers using supervised, unsupervised, and reinforcement learning
- We demonstrate Agua's ability to intuitively enable
  - Debugging, data shift detection, retraining, dataset expansion
- Code: https://github.com/NetSAIL-UCI/agua