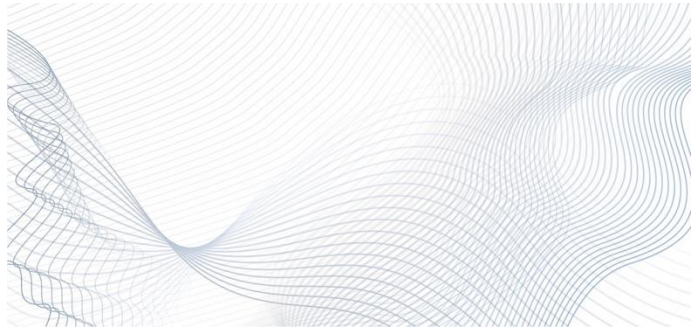
Toward Trustworthy Learning-Enabled Systems with Concept-Based Explanations

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

UC Irvine

KAIST

vmware by Broadcom



Learning-Enabled Systems are Transforming Networking

Congestion Control

Aurora (ICML ‘19)

Sage (Sigcomm ’23)

Video Conferencing

Tambur (NSDI ‘23)

Gemino (NSDI ‘24)

Cluster Scheduling

Decima (Sigcomm ‘19)

Pollux (OSDI ‘21)

Load Balancing

DeepRM (HotNets ‘16)

DRL for Network Slicing (IEEE Access ‘18)

Adaptive Bitrate Streaming

Pensieve (Sigcomm ‘17)

Fugu (NSDI ’20)

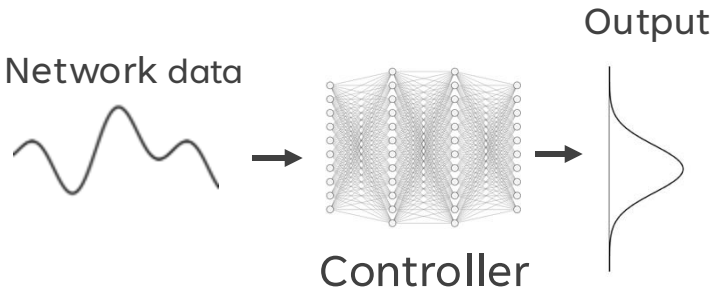
Microservice Management

Autothrottle (NSDI ‘24)

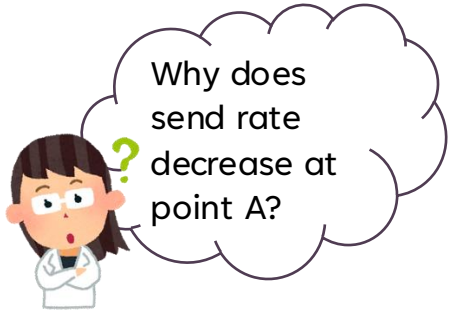
TopFull (Sigcomm ‘24)

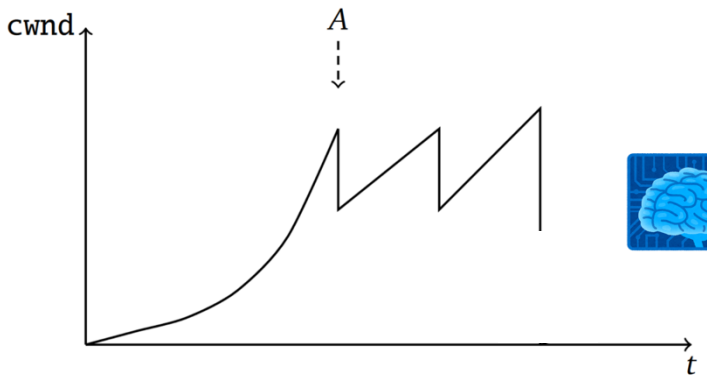
Learning Solutions Rely on Blackbox Neural Networks

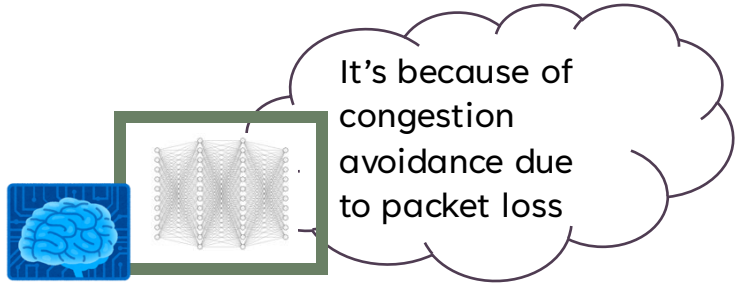
- Large amounts of network data
- Complex blackbox neural networks
  - Directly map input to output
  - Difficult to
    - Debug
    - Interpret

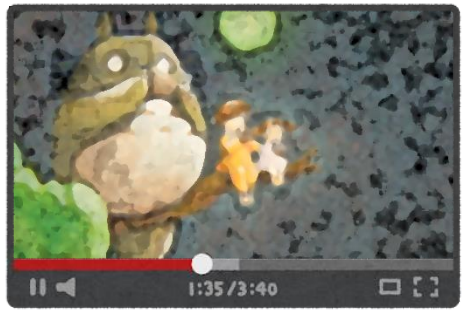


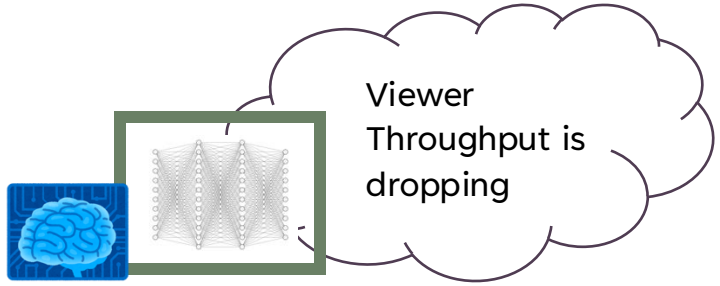
The Vision: Interactive Solution





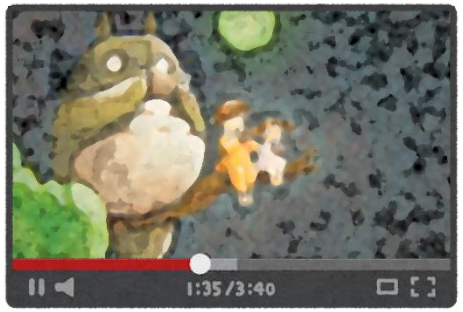


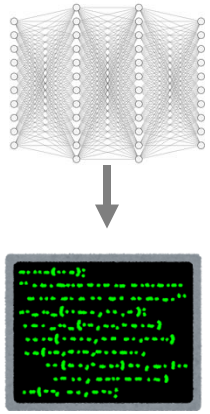


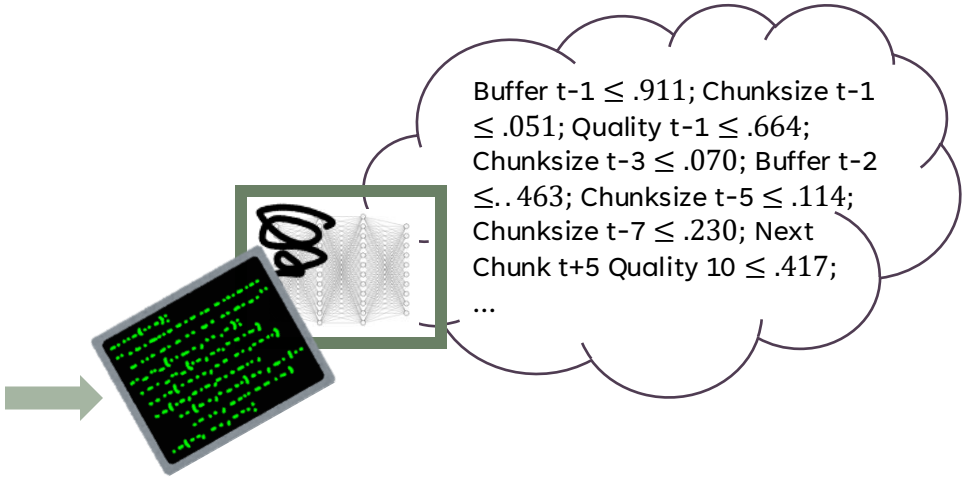


Current Techniques

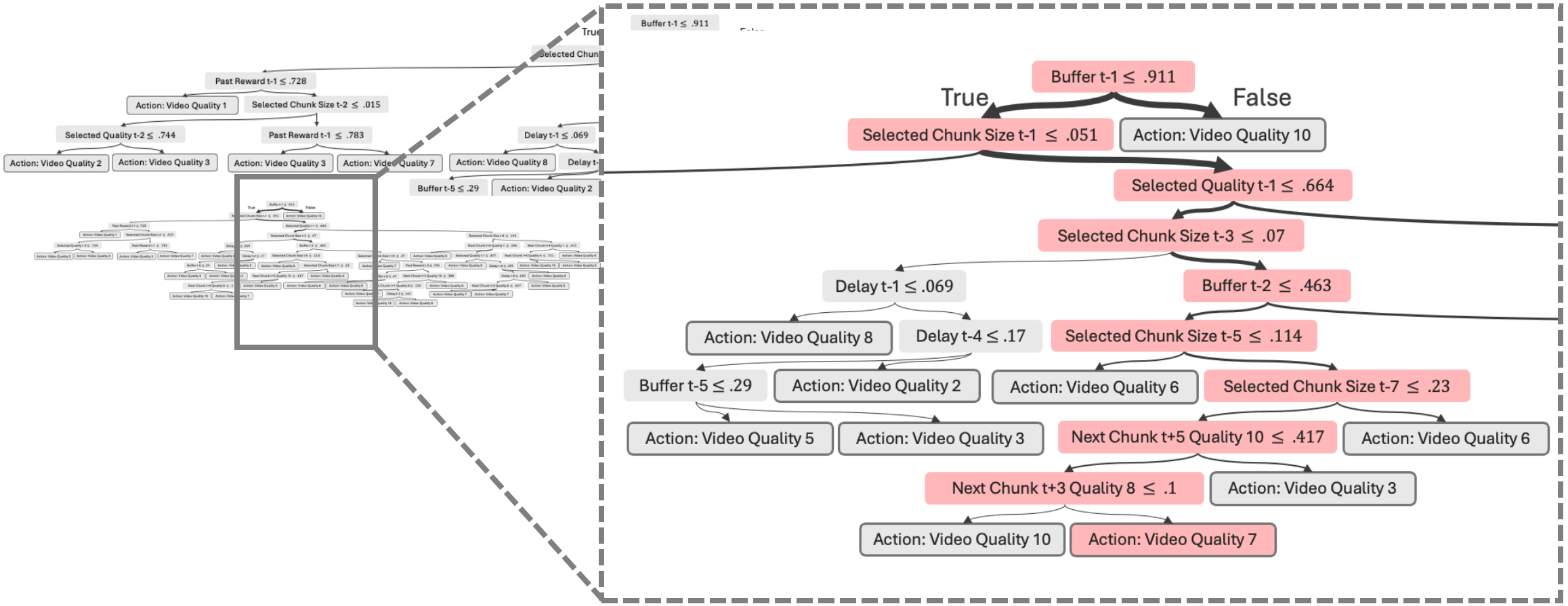
Current State-of-The-Art





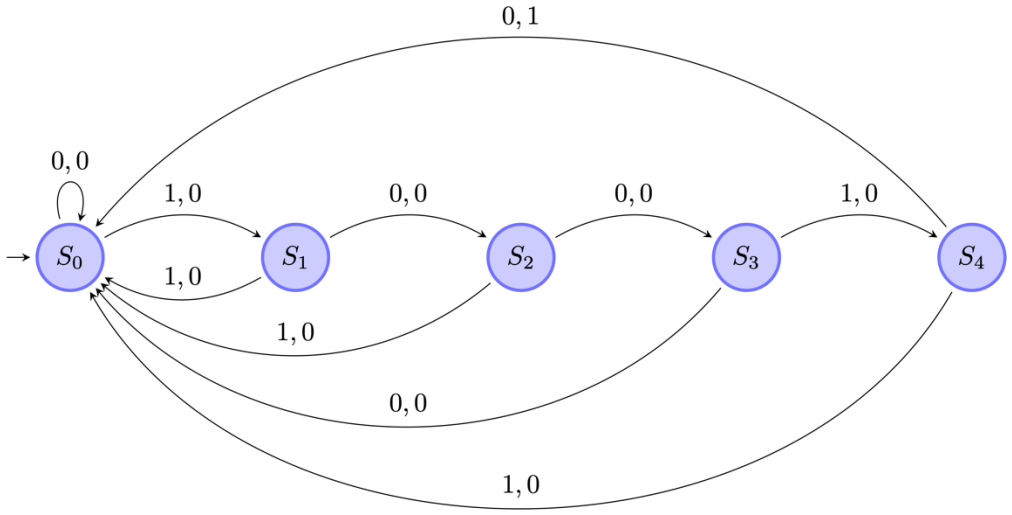


Decision Tree Explanations

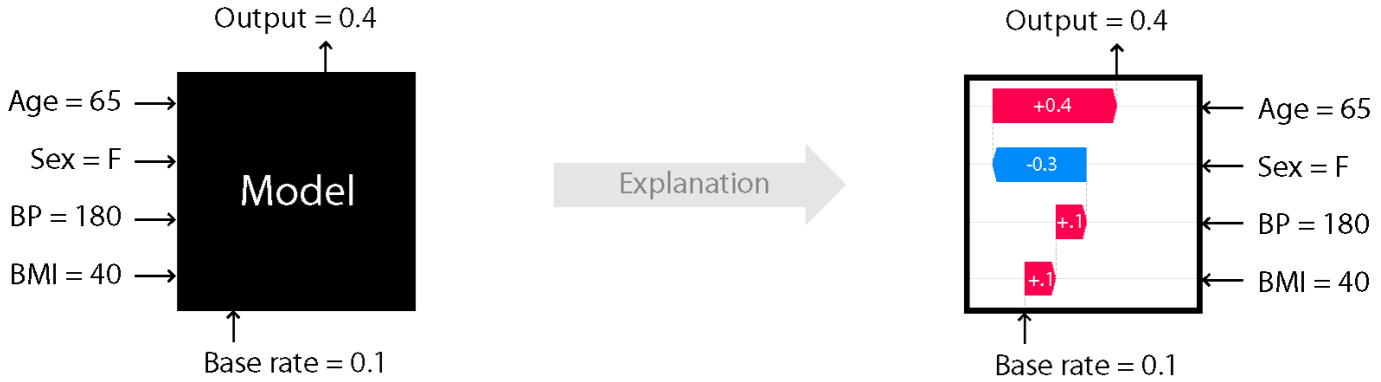


Other Techniques

- Distillation techniques
  - Graphs (Infocom ‘17)
  - Rule Sets (ICLR ‘17)
  - Finite-State Machines
    (IEEE Trans. Neural Netw. Learn. Syst. ‘20)
- Salient feature techniques
  - SHAP (NIPS ‘17)
  - LIME (SIGKDD ‘16)
  - ALE (J. R. Stat. Soc.‘20)



Finite State Machine Explanation

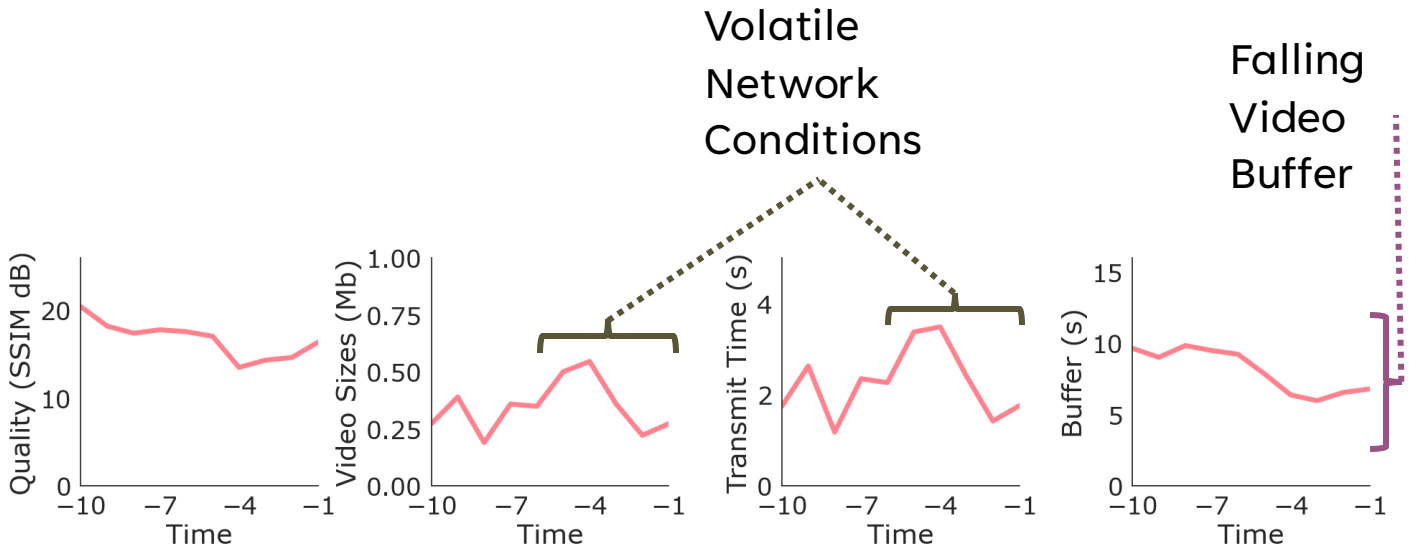


SHAP Explanation

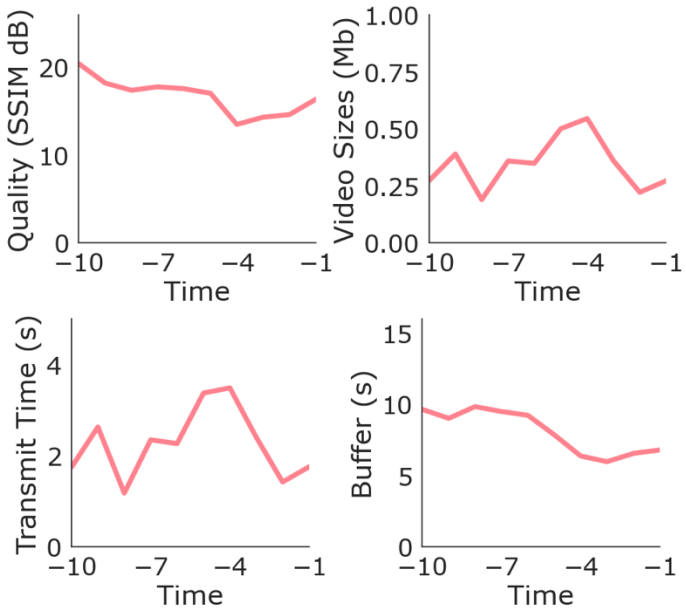
Introducing Concepts to Systems Explainability

Concepts in Systems

- Concepts
  - Human-understandable attributes that capture controller and environment characteristics
  - Can capture intricate patterns, trends, and behaviors in systems
  - Align with human intuition
- Involve multiple features
- Trends across time



Concept-Based Understanding



Controller Model

Blackbox

Bitrate
1440p

Concept-Based Model

Concepts

Weights

1. Volatile Network Conditions **0.98**
2. Falling Buffer 0.5

...

n. Medium Throughput 0.1

Linear Combination

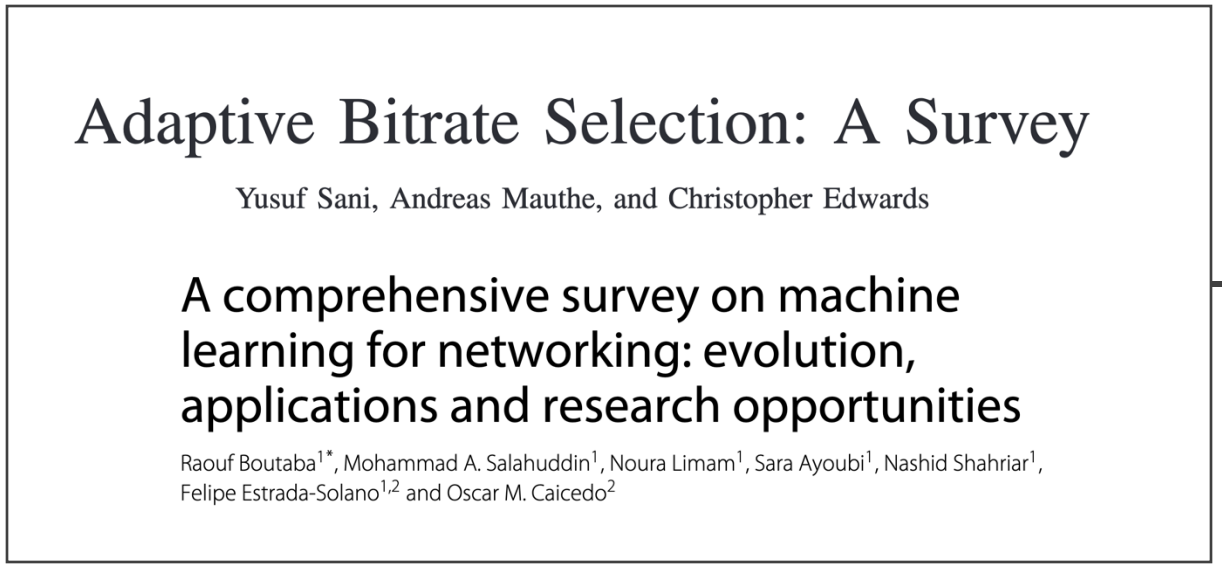
Bitrate
1440p

Training Pipeline: 1. Base Concept Generation

Prompt:
Summarize paper & identify concepts

Large Language Model

Base Concepts



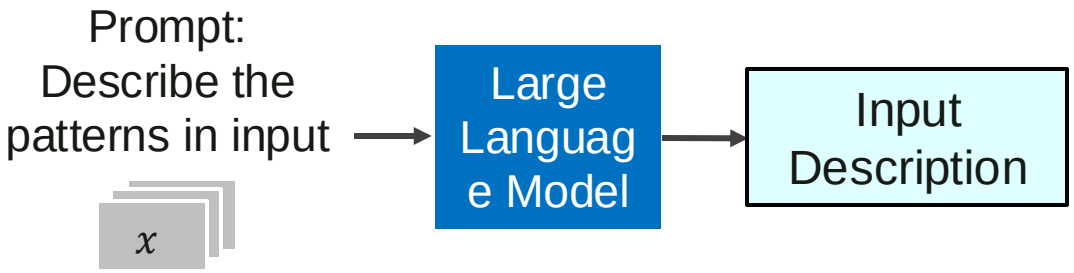
Concepts

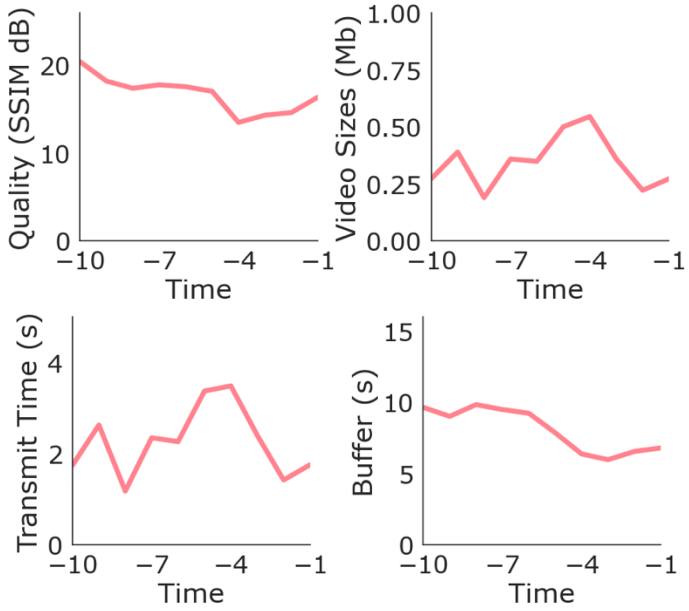
1. Volatile Network Conditions
2. Falling Buffer

...

n. Medium Throughput

2. Input Description Generation

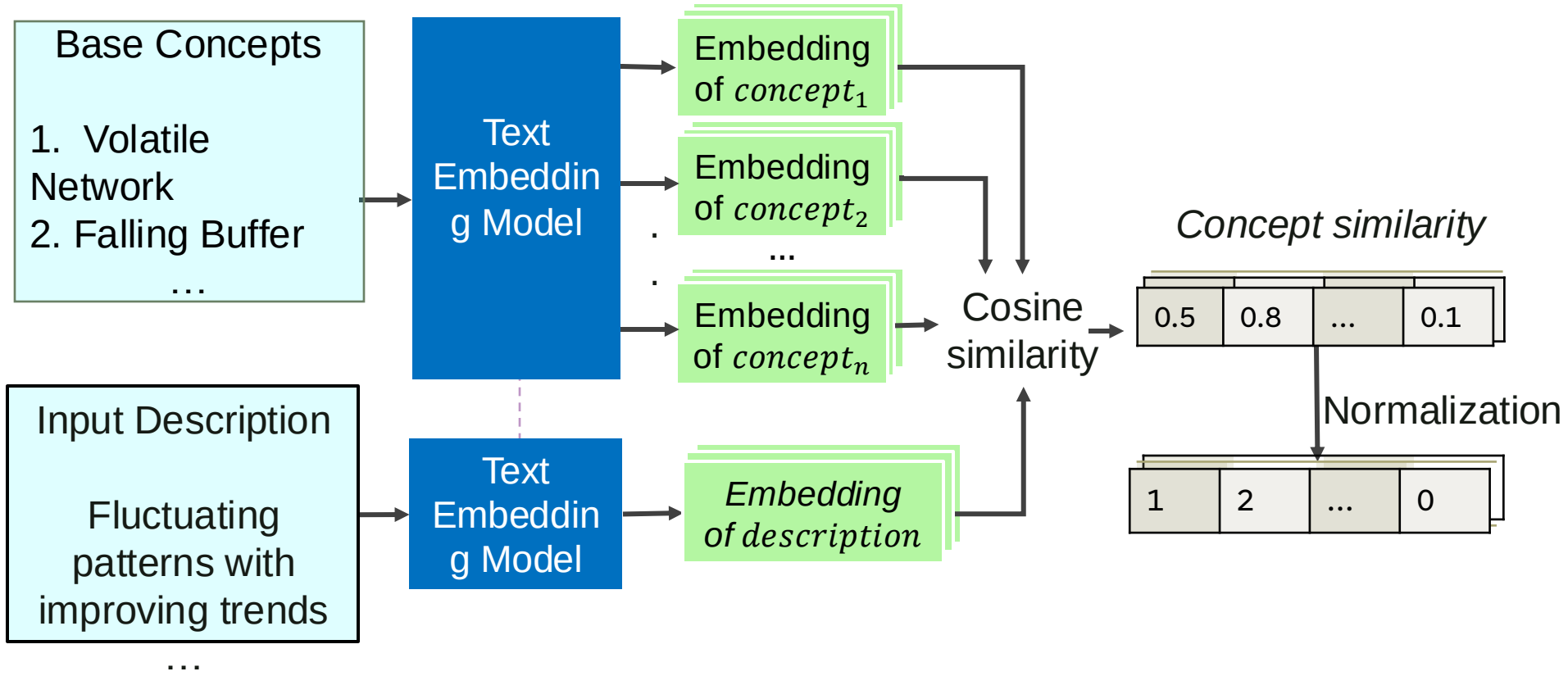




Network conditions:

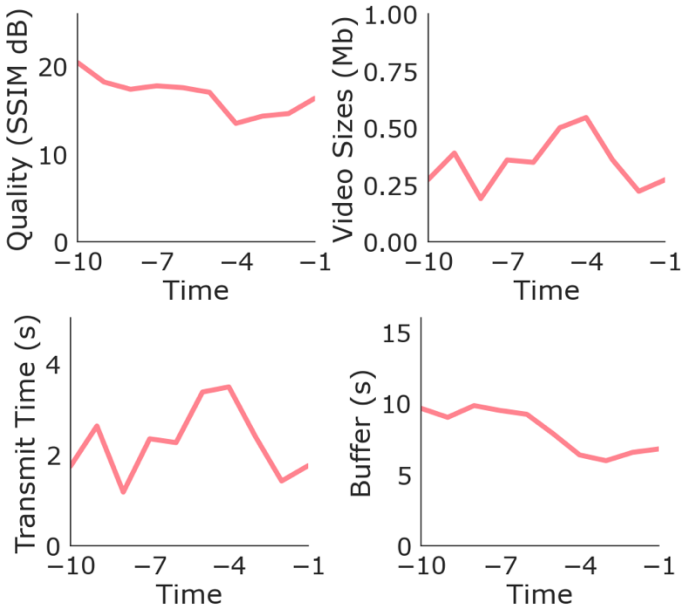
- Initially starts off with a fluctuating pattern, as observed from the features "Transmission Time of Chunk."
- In the middle...
- Overall, the trend is improving, indicating the presence of stable network conditions.

3. Input Concept Generation



4. Training Concept Mapping

Multi-dimensional classification



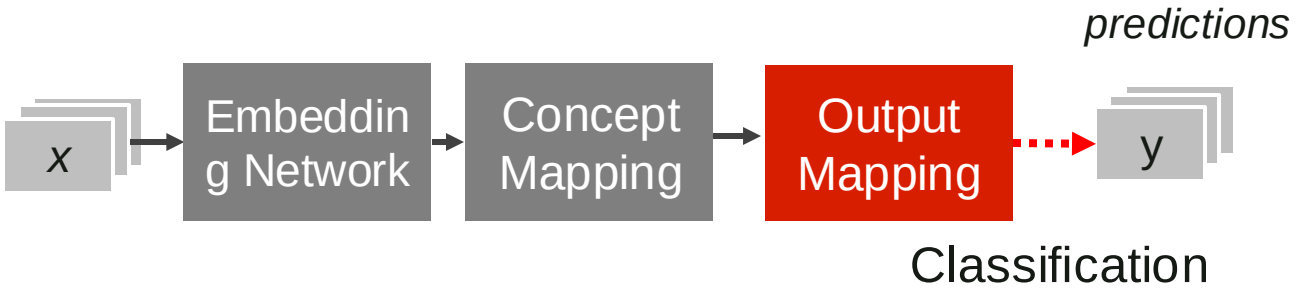
Concepts

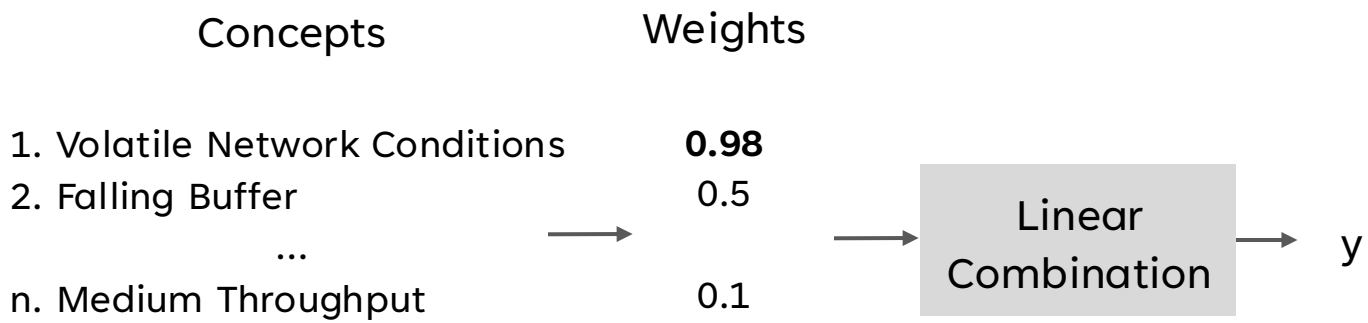
1. Volatile Network Conditions
2. Falling Buffer

...

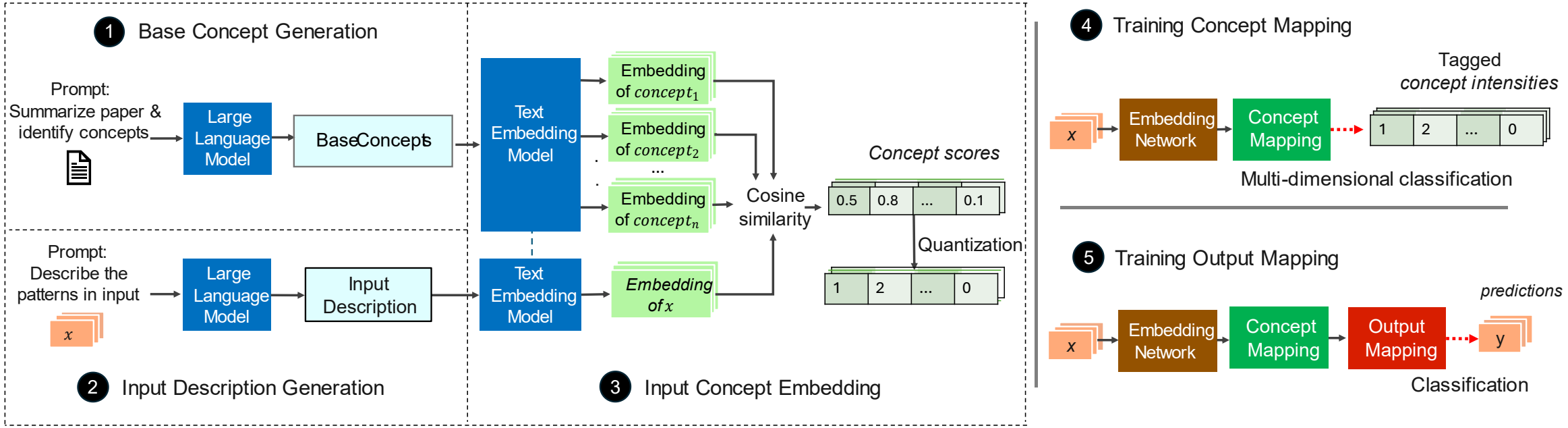
n. Medium Throughput

5. Training Output Mapping



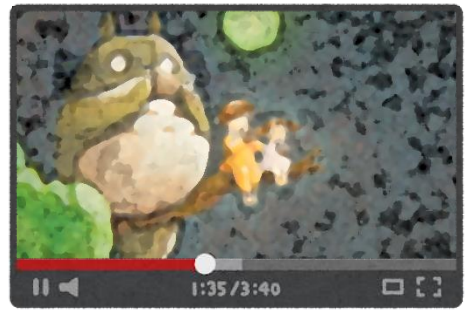


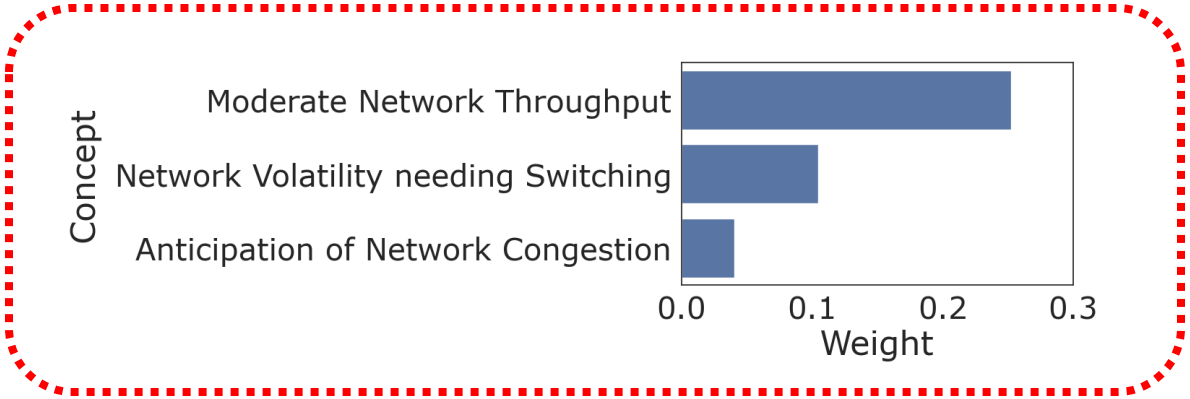
Training Pipeline



Using a Concept-Based Explainer: Unintended Behavior

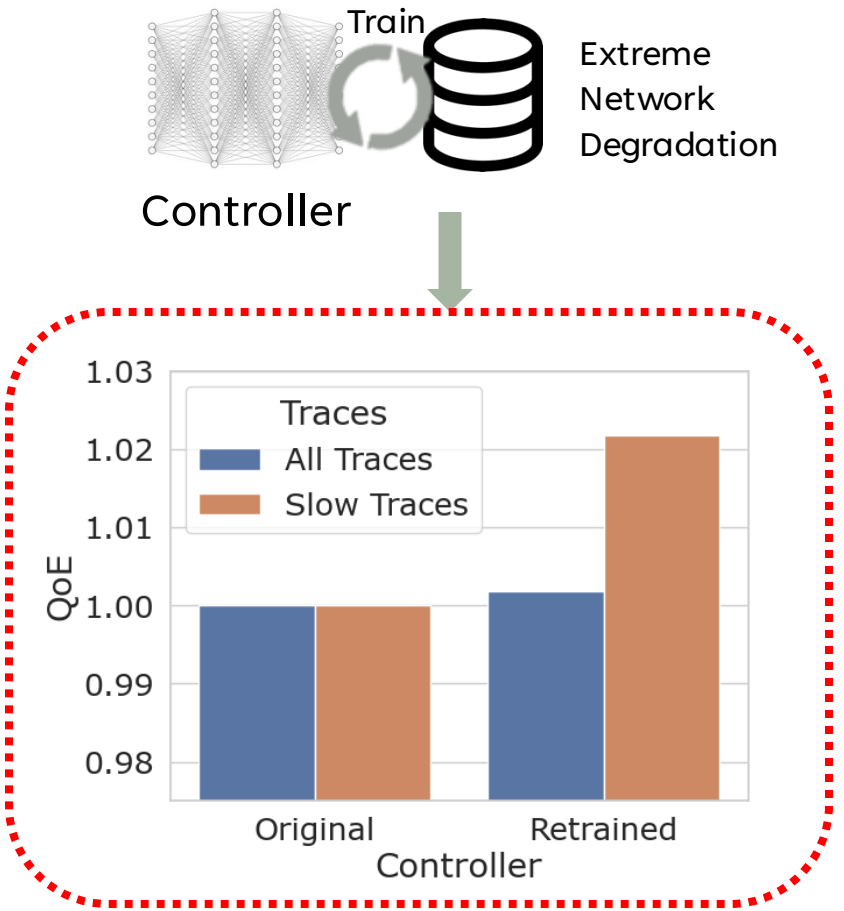
- Query for an explanation
  - Controller input
  - Chosen bitrate
- Understand the key high-level concepts



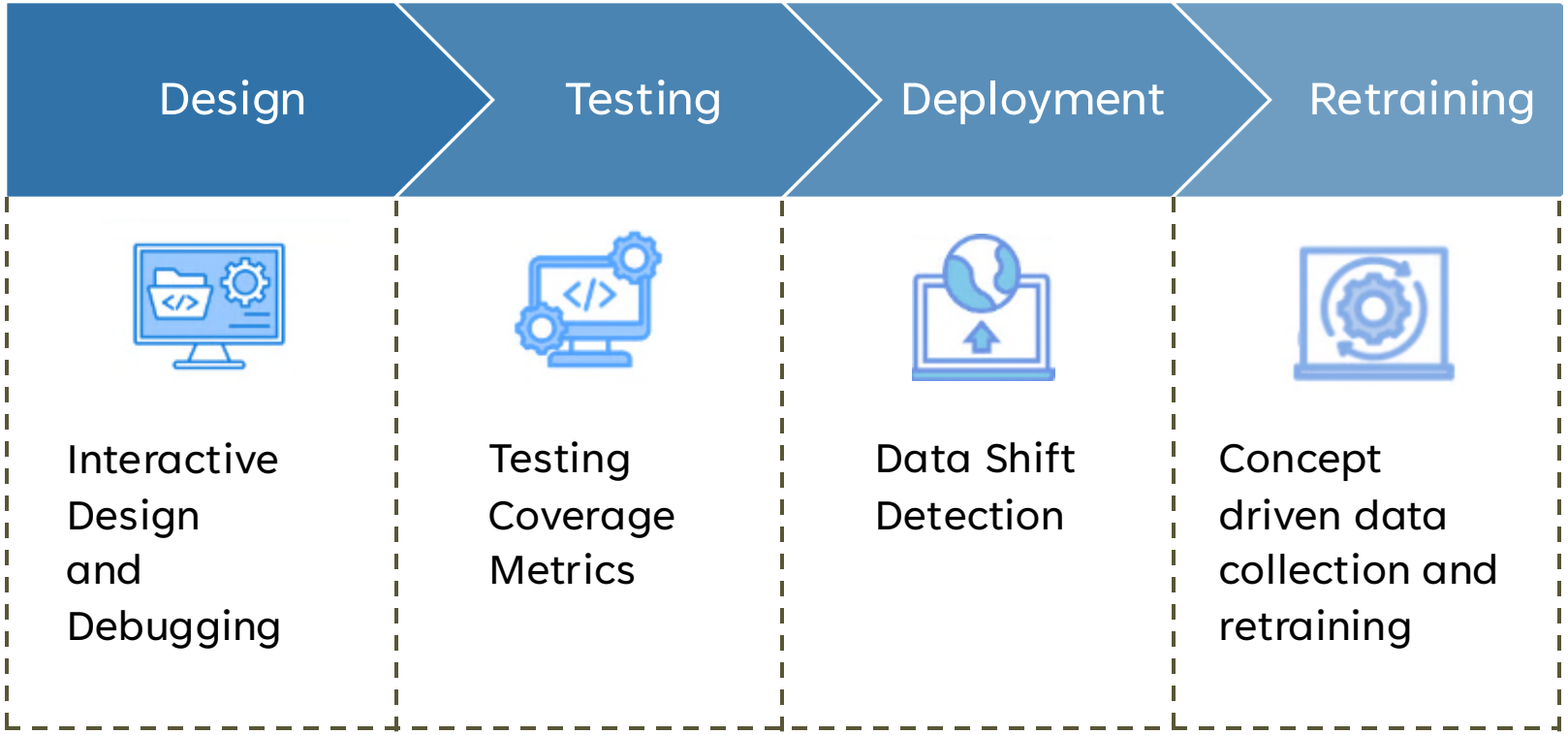


Using a Concept-Based Explainer: Debugging

- Define data-collection with concepts
- Target high-level areas
  - E.g. Network traces with
    "Extreme Network Degradation"
- Address concerns and attain higher performance



Transforming Controller Lifecycle with Concepts



Summary

- Concepts redefine explainability to a human-understandable level
- Align with human intuition
  - E.g. Volatile Network Conditions, Depleting Buffer
- Enable a fundamentally new approach to data-driven controllers