UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Project Ideas

Presented by: Tarek Abdelzaher

Part I

Past Project Examples (with modern angles)

Outline

- Part I: Past Project Examples
- Part II: AI for IoT Challenges
- Part III: Systems and Applications

Example 1 (Past IoT Project): GreenGPS

Improve fuel-efficiency of transportation via "green" navigation

- Measure fuel-efficiency of vehicles
- Model fuel-consumption as a function of driver characteristics, road characteristics (average speed, speed variability, waiting time, slope, etc), and vehicle characteristics
- Compute least-energy routes for a given vehicle and driver

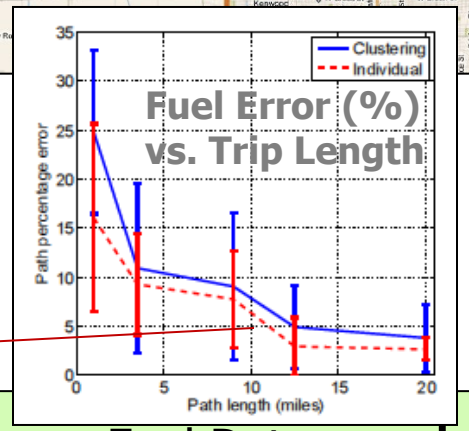
Green GPS

Saves 6% over shortest path and 13% over fastest path

Shortest and fastest

Green GPS

Most fuel-efficient



Subscribers:
Premium service
High savings

Subscribers

OBDII-WiFi Adaptor ($50)

+

GPS Phone

Fuel Data + Physical Models

Server

$$F_{engine} = \frac{\Gamma(\omega) G g_k}{r}$$

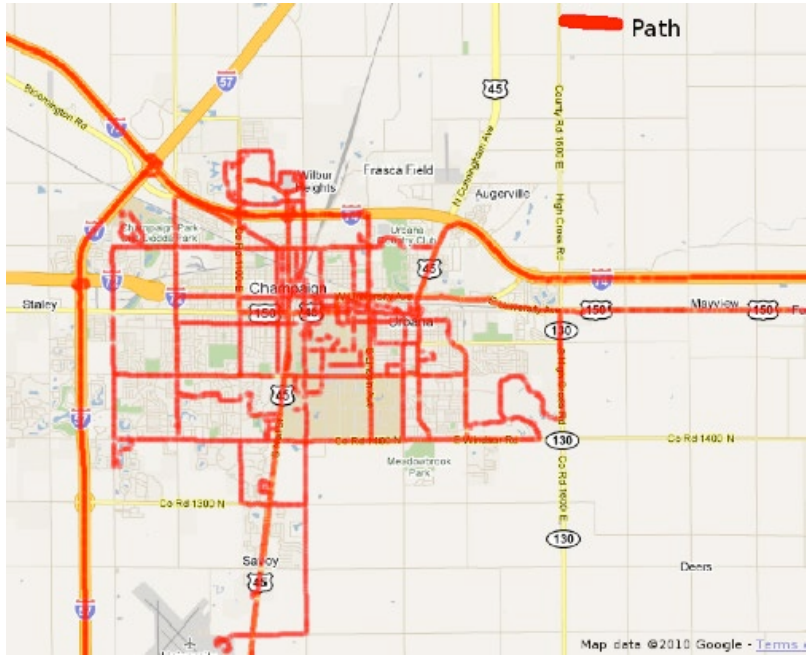
$$F_{air} = \frac{1}{2} c_d A \rho v^2$$

$$F_{friction} = c_{rr} mg\cos(\theta)$$

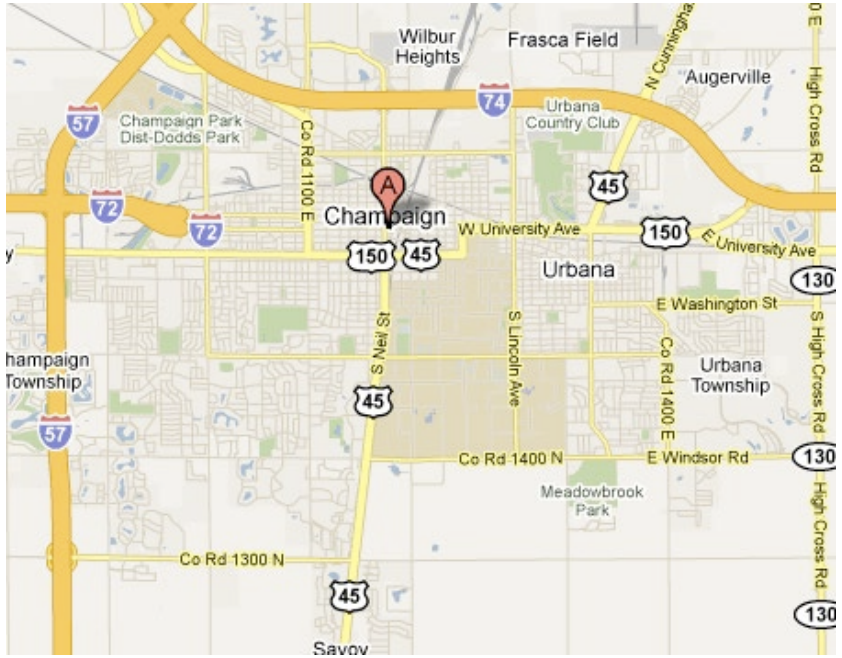
$$F_g^g = mg\sin(\theta)$$

$$F_{car} = F_{engine} - F_{friction} - F_{air} - F_g$$

A Modeling Challenge



Fuel consumption of a few cars driven on a few roads by a few driver



Predict fuel consumption of any car on any road by any driver

Fuel Savings Evaluation

How efficient is the fuel-efficient route?

Car Details%	Landmarks	Route	Savings
Honda Accord 2001	H1 to Mall	Shortest	31.4%
	H1 to Gym	Shortest	19.7%
Ford Taurus 2001	H2 to Restaurant	Shortest	26%
Toyota Celica 2001	H2 to Work	Fastest	10.1%
Nissan Sentra 2009	H3 to CUPHD	Fastest	8.4%
Honda Civic 2002	Grad to Work	Fastest	18.7%

Average fuel savings across 5 cars

A Modern Angle to GreenGPS

Instead of using an analytic model, train a foundation model for “Green Navigation”

- Learns to estimate fuel consumption on different streets as a function of car and street parameters
- Applies least cost routing to find the minimum fuel route between source and destination points

Example 2 (Past IoT Project): Semantic Disruption-tolerant Networks

Imagine: a big disaster strikes a city…

Images are collected from the Internet

Hurricane Katrina 2005

Nepal earthquake 2015

Thailand flood 2011

Storm Helene 2024

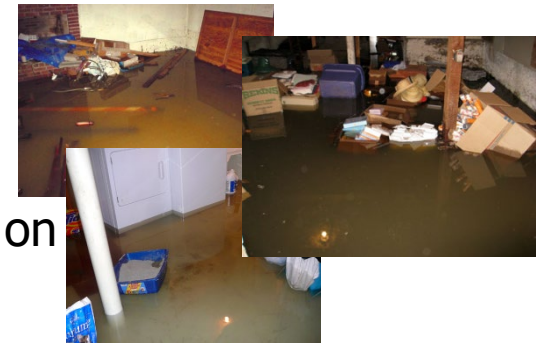
- Drones scout the area, capture pictures/videos, then form a network (IoT) to send these pictures to a rescue center
- Drone (IoT) network constraints prevent sending all pictures
- The network must prioritize the data to send representative samples

Challenge: Data Selection to Maximize Coverage

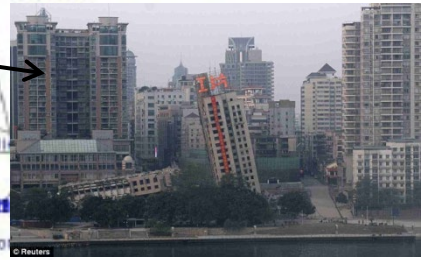
Fire on 6th and Main.

Collapse on Park Ave.

Flooding on State St.



Structural damage on Pier Square



Example of Bad Coverage

Fire on 6th and Main.

Collapse on Park Ave.

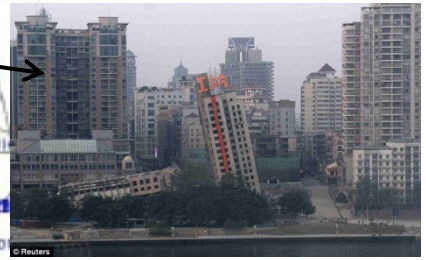
Example of Good Coverage

Fire on 6th and Main.

Collapse on Park Ave.

Flooding on State St.

Structural damage on Pier Square



A Scheduling Approach: Coverage-maximizing Priorities

Implement coverage-maximizing in-network prioritization for forwarding and storage

- Objects are forwarded/dropped in a priority order aimed to maximize coverage of delivered content
  - Objects similar to previously forwarded ones get lower priority
- Challenge: Forwarding and dropping must be made aware of the degree of semantic redundancy (i.e., similarity) between objects

A Scheduling Approach: Coverage-maximizing Priorities

Implement coverage-maximizing in-network prioritization for forwarding and storage

- Objects are forwarded/dropped in a priority order aimed to maximize coverage of delivered content
  - Objects similar to previously forwarded ones get lower priority
- Challenge: Forwarding and dropping must be made aware of the degree of semantic redundancy (i.e., similarity) between objects
  - Modern angle: Use representation learning to convert images/videos into a latent space that exposes semantic redundancy for prioritization purposes.

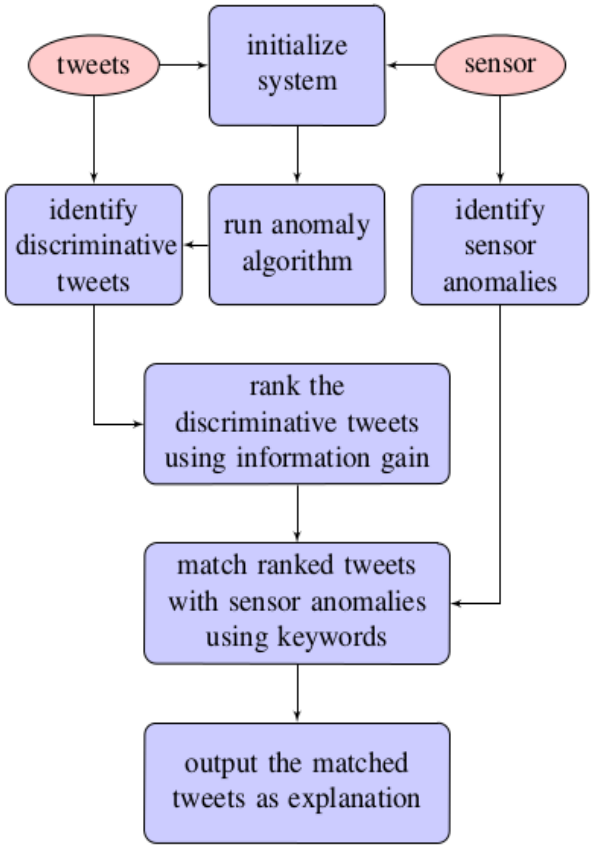
Example 3 (Past IoT Project): Anomaly Explanation by Combining Sensors and Social Media

Goal: Explain anomalous traffic conditions on freeways in three California cities: Los Angeles, San Francisco, San Diego

- Sensor data: Department of Transportation publishes freeway speed data (from thousands of sensors) once every 30 seconds
- Twitter data: Collected California Twitter (now X) data with keyword “Traffic”

Input: Location of sensor experiencing anomaly

Output: Ranked list of tweets (or other real-time social media posts) that comprise possible explanations.



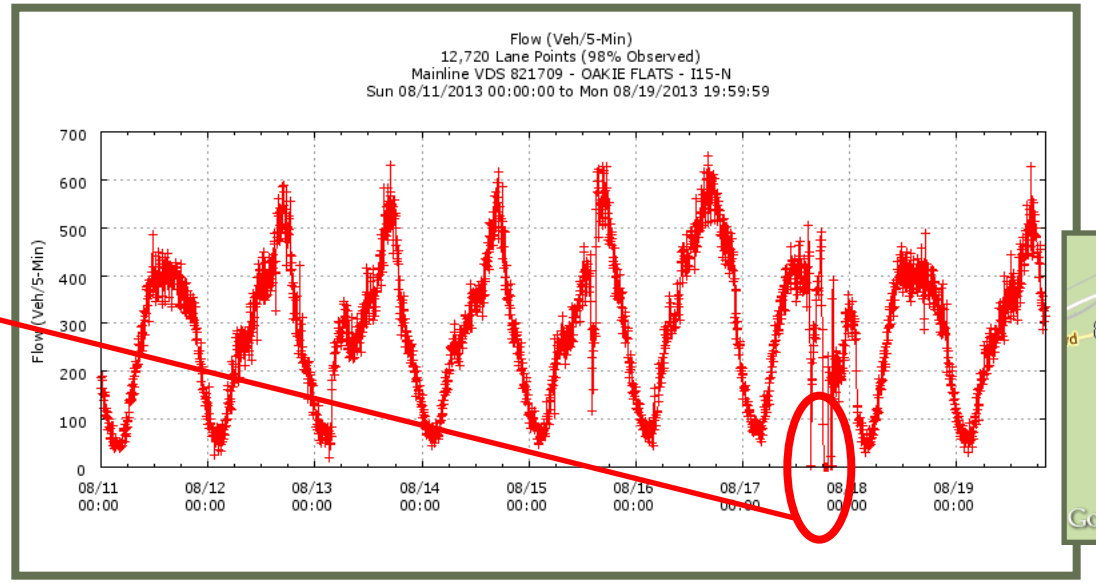
Detected Anomaly 1: Forest Fire

Time, Location:
6pm, Aug 17th, I-15 N, LA

Event Signature:
"Cajon Pass"

Possible Explanation (from Twitter):
Cleghorn Fire in **Cajon Pass** Snarls Traffic on I-15: (KTLA) One northbound lane on the 15 Freeway was reopened... http:\\t.co\nieqh4nsMX

URL:
http:\\t.co\nieqh4nsMX



One northbound lane on the 15 Freeway was reopened Saturday evening in the Cajon Pass as firefighters continued to battle the so-called Cleghorn fire, authorities said. All southbound lanes on Interstate 15 were open, according to the U.S. Forest Service. State Road 138 remained closed from the 15 Freeway to Summit Valley Road.

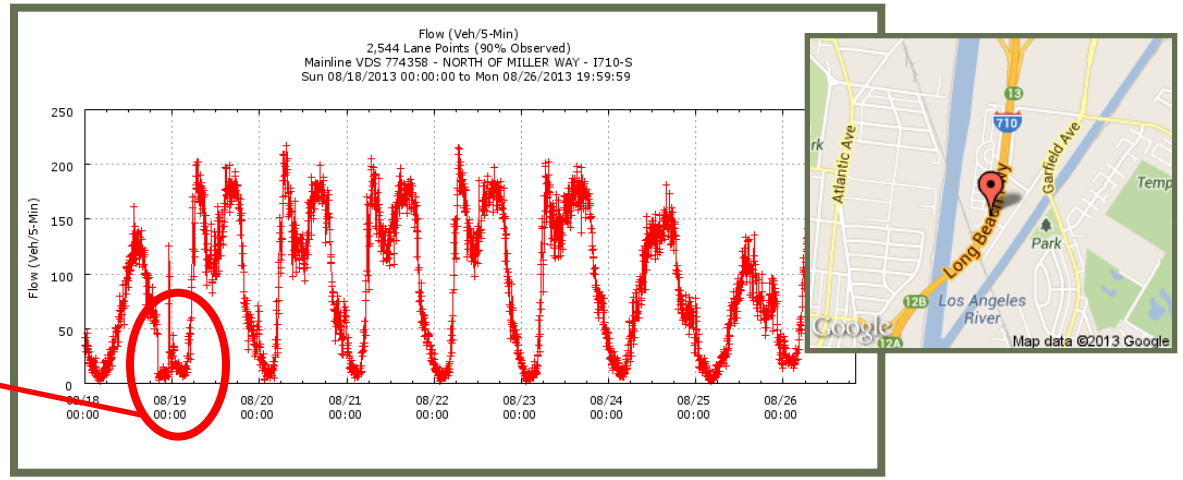
Detected Anomaly 2: Pedestrian Death

Time, Location:
8pm, Aug 18th, SB I-710, LA

Event Signature:
"Drunk Kills"

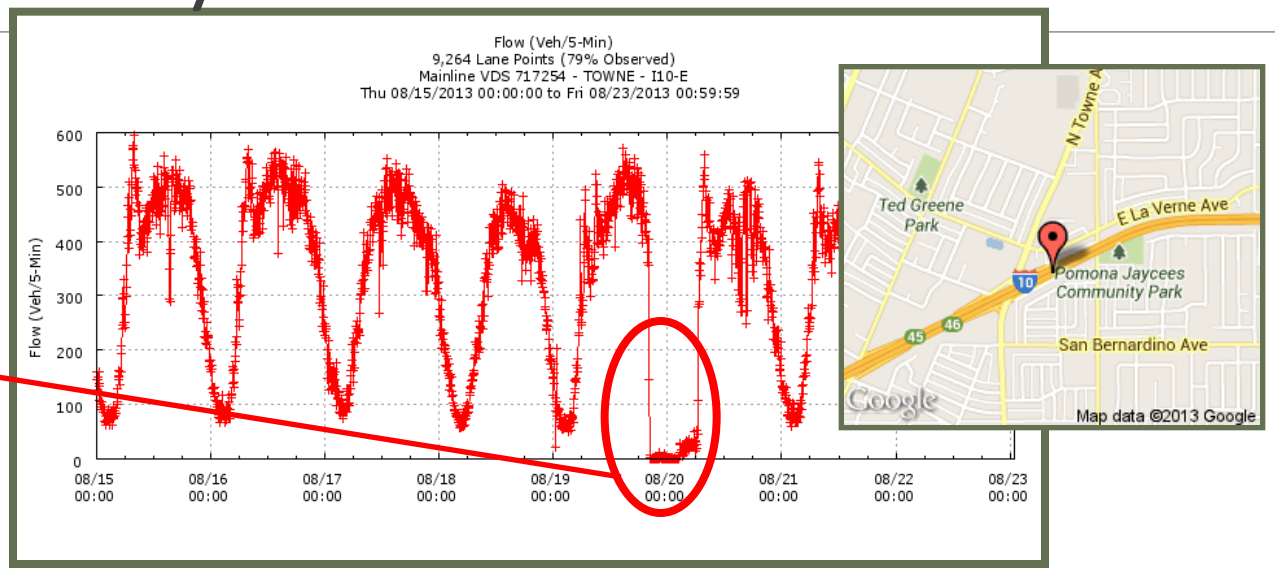
Possible Explanation (from Twitter):
27-Year-Old **Drunk** Driver Hits, **Kills** Man Trying To Stop Traffic On 710 http:\\t.co\rMoI7DxFH4

URL:
http:\\t.co\rMoI7DxFH4



A 27-year-old Long Beach woman faces a possible felony charge after fatally hitting a pedestrian on the 710 Freeway while she drove intoxicated.
Melanie Gosch struck a man in his 20s at about 8 p.m. on Sunday near Imperial Highway on the southbound 710 in South Gate, according to City News Service. The man, whose name has been withheld, was allegedly trying to stop traffic in the number three freeway lane.
Police responded to the scene after a caller reported a "long-haired man in dark clothing" on the freeway. Within a minute, the California Highway Patrol received a call that a pedestrian was lying in the freeway. The man was pronounced dead at the scene.
Gosch stopped her 2007 Nissan Sentra after striking the man, and she was arrested and booked on suspicion of causing injury or death while driving under the influence of alcohol or drugs.

Detected Anomaly 3: Vehicle Crash



Time, Location:
8pm, Aug 19th, I-10E, LA

Event Signature:
"Crashes Divider"

Possible Explanation (from Twitter):
1 dead, 9 hurt in fiery **crashes** on 10 Fwy in Pomona; EB 10 closed, traffic allowed to pass along center **divider**
http:\\t.co\cCu8xLzx1U

URL:
http:\\t.co\cCu8xLzx1U

POMONA, Calif. (KTLA) — The investigation continued Tuesday into a pair of chain-reaction crashes on the 10 Freeway in Pomona that left one person dead and eight others injured. One killed, eight hurt in pair of chain-reaction crashes on 10 Freeway in Pomona. It all happened around 8 p.m. on Monday on the eastbound 10 Freeway near Towne Avenue, according to the California Highway Patrol. The first crash involved four cars and created a traffic back-up, CHP officials said. That's when a second crash occurred involving a big rig with a full tank of diesel fuel and three other vehicles. The fuel tank of the big rig ruptured, causing it to burst into flames, authorities said. The driver of the semi was able to get out safely. However, the driver of a red BMW that became trapped under the big rig was not able to escape. That person, who was not immediately identified, died at the scene.

A Modern Angle

Embed sensor data/metadata and social media posts into the same space, allowing automated retrieval of posted explanations correlated with sensing anomalies.

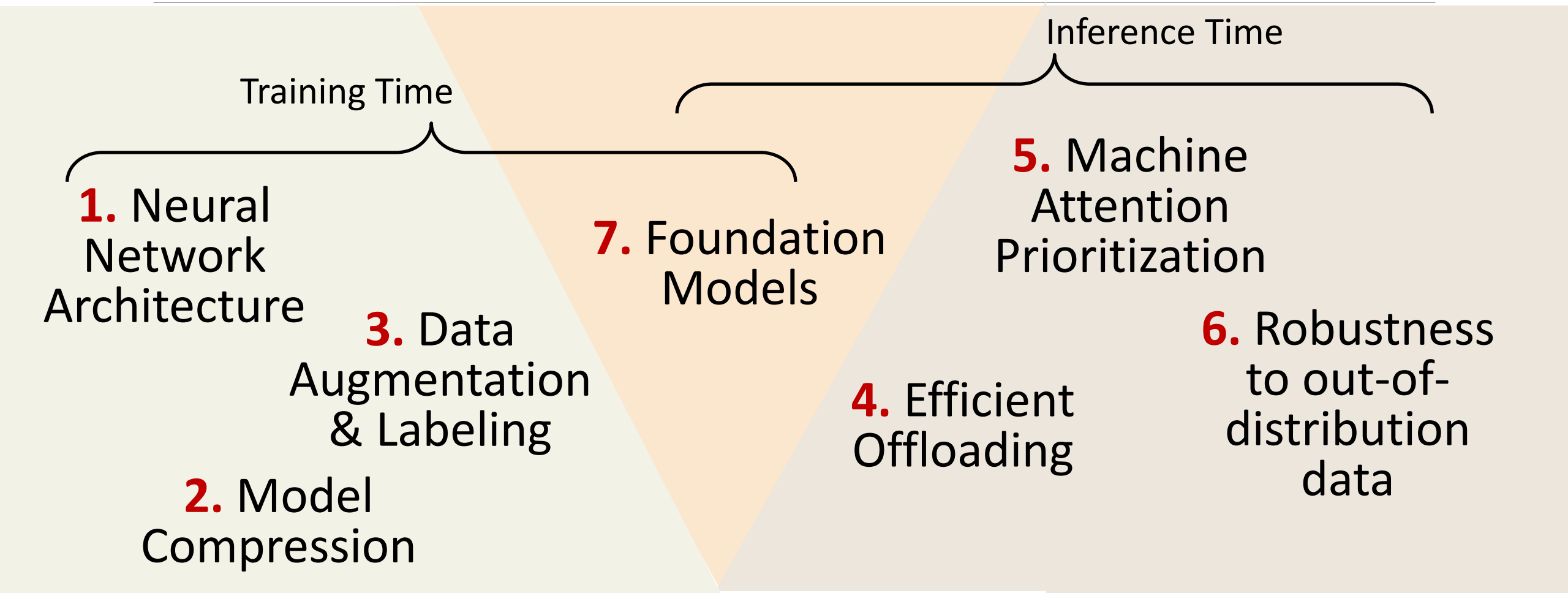
Part II

AI for IoT Challenges

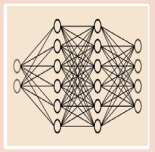
Outline

- Part I: Past Project Examples
- Part II: AI for IoT Challenges
- Part III: Systems and Applications

Projects by Challenge in ML for IoT

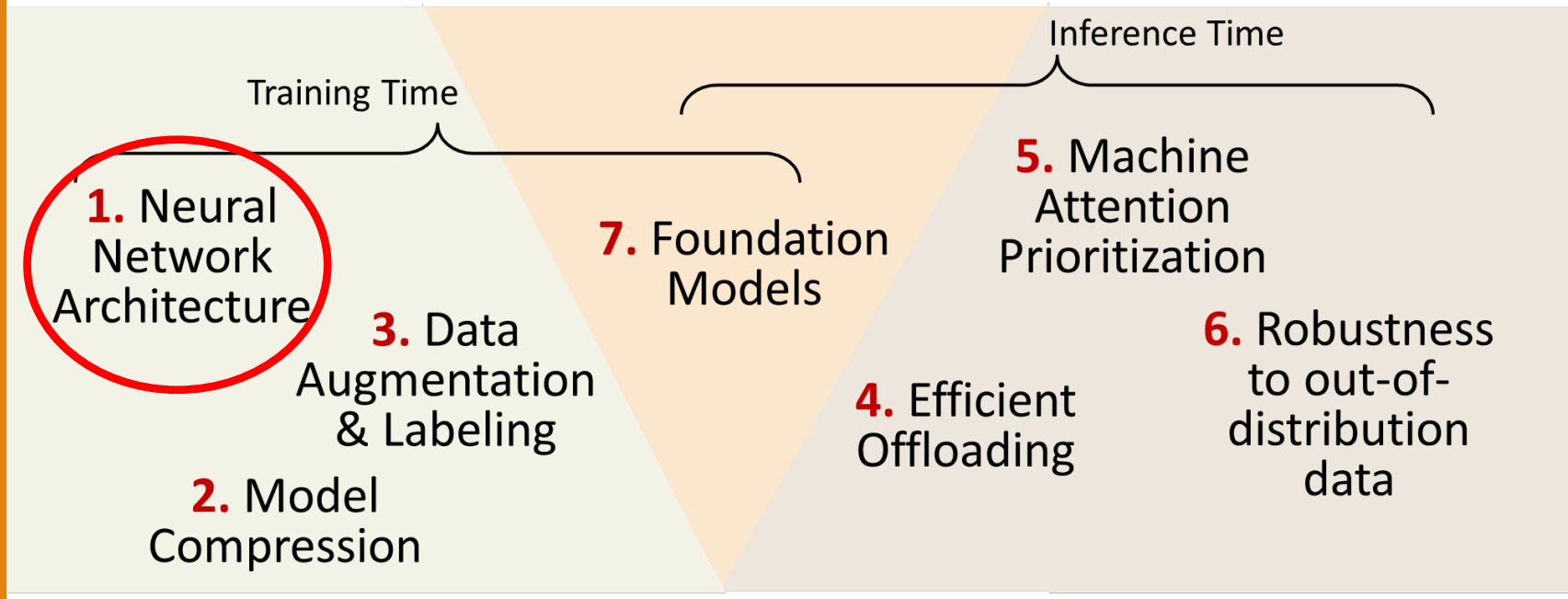


Challenge 1: Neural Network Architecture

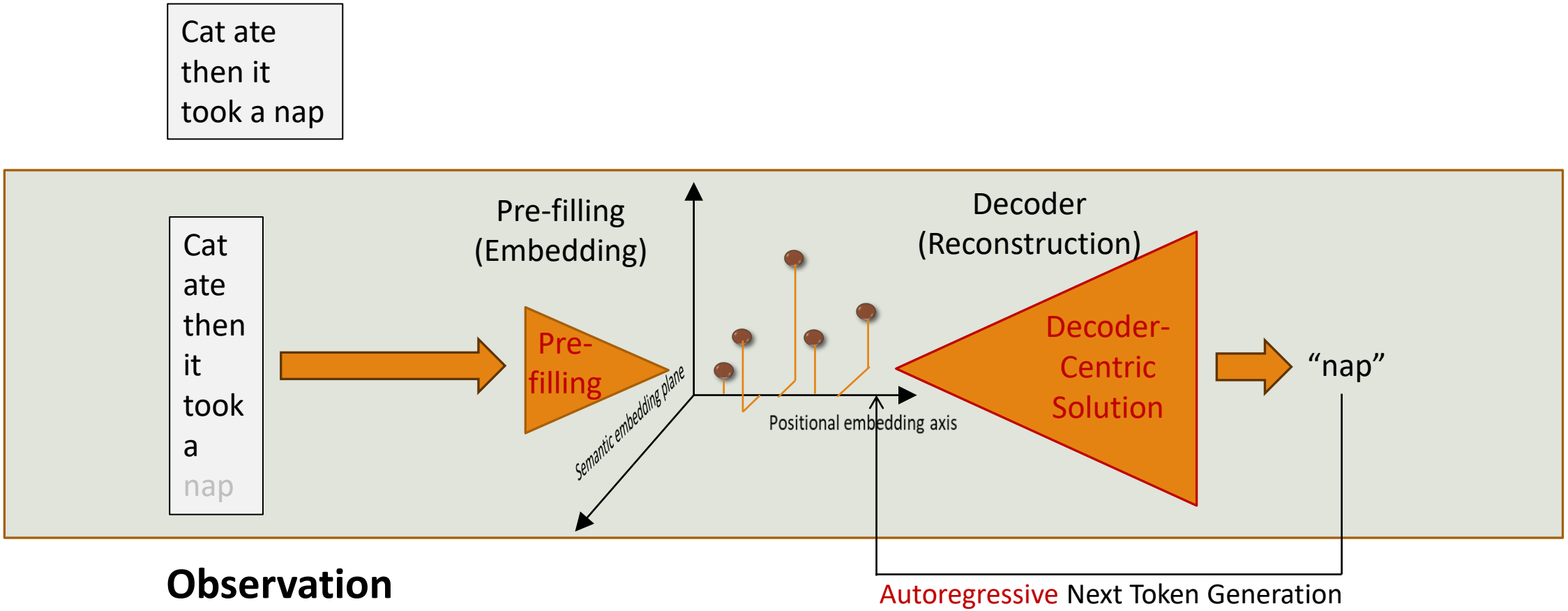


What neural network architecture is best suited for sensor data?

What new architectural components are useful?



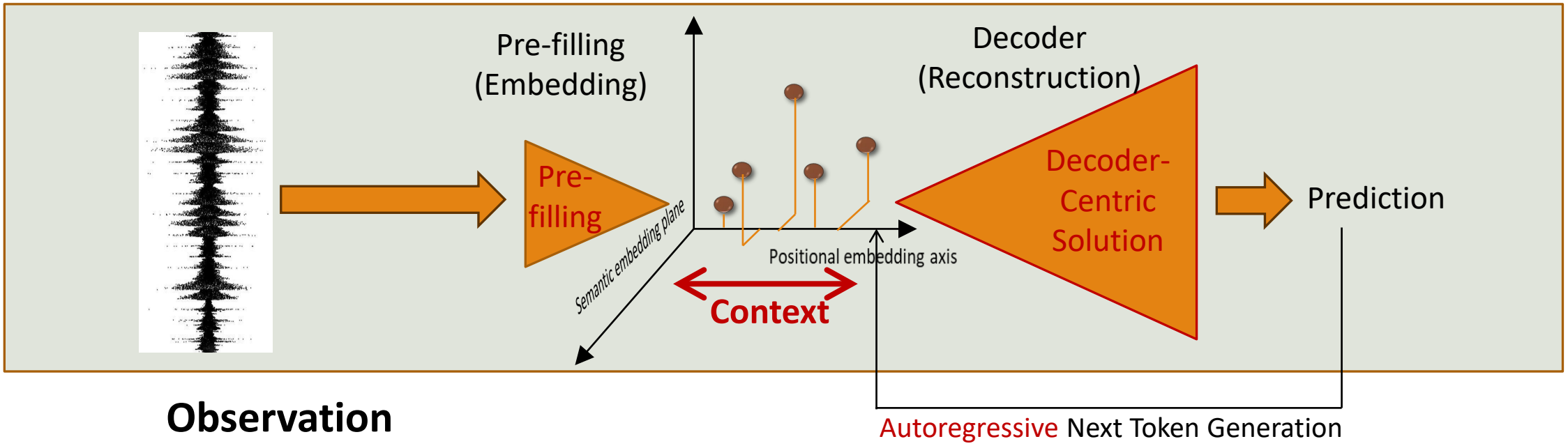
Project Idea: IoT-Centric Neural Network Architecture for Self-Supervised Learning



Project Idea: IoT-Centric Neural Network Architecture for Self-Supervised Learning

Problems: Architectures for better representing sensor data

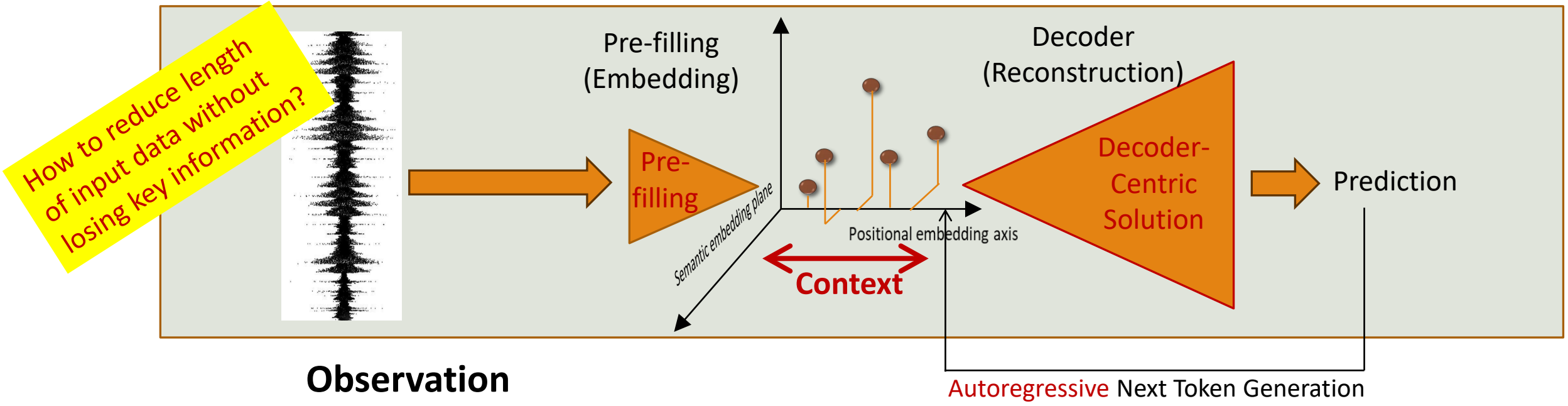
- Sensor can create a very long history very quickly (e.g., sound sampled at 10 KHz creates millions of samples in minutes. We cannot "prompt" IoT-compatible foundations models/LLMs with such a long context. How to summarize the data in a more manageable format without missing key events?



Project Idea: IoT-Centric Neural Network Architecture for Self-Supervised Learning

Problems: Architectures for better representing sensor data

- Sensor can create a very long history very quickly (e.g., sound sampled at 10 KHz creates millions of samples in minutes. We cannot "prompt" IoT-compatible foundations models/LLMs with such a long context. How to summarize the data in a more manageable format without missing key events?



Project Idea: IoT-Centric Neural Network Architecture for Self-Supervised Learning

Problems: Architectures for better representing sensor data

- Sensor can create a very long history very quickly (e.g., sound sampled at 10 KHz creates millions of samples in minutes. We cannot “prompt” IoT-compatible foundations models/LLMs with such a long context. How to summarize the data in a more manageable format without missing key events?

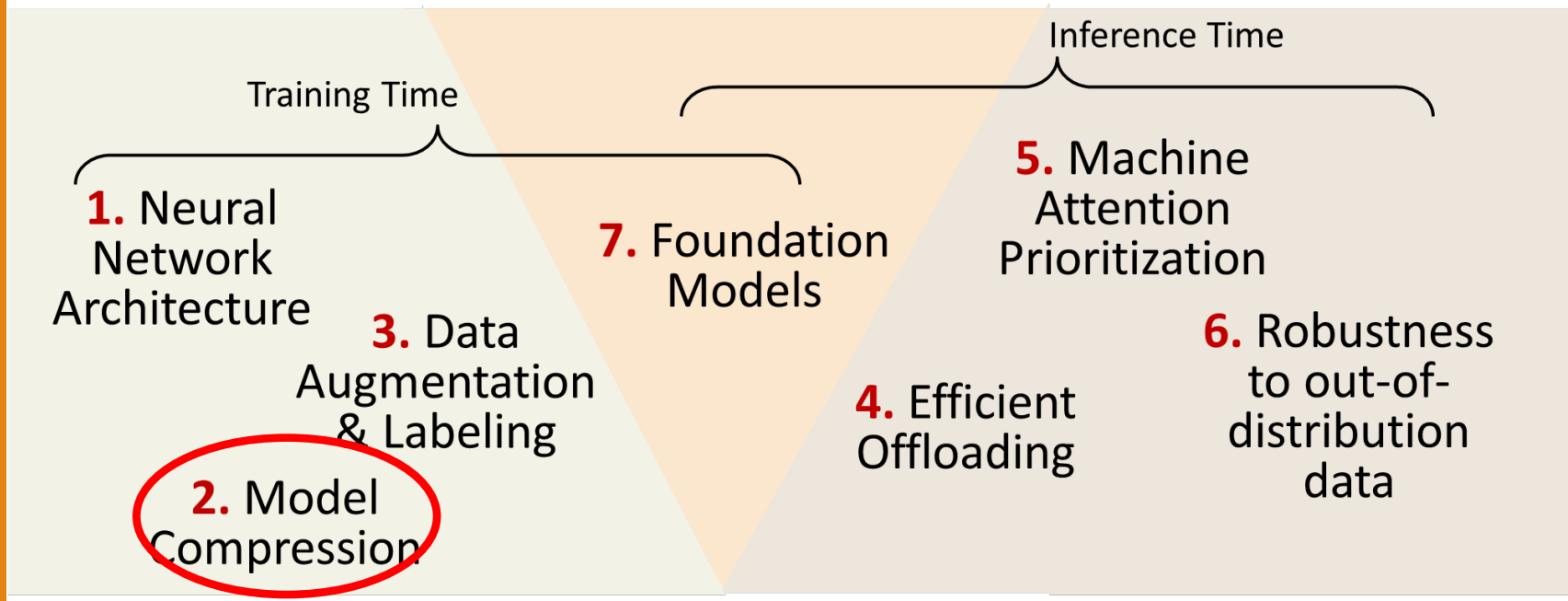
Directions

- Build an event-driven neural architecture.
- Tweak recent state-space models and neural ODEs.
- Develop heuristics from dropping “unimportant” data (e.g., silence).

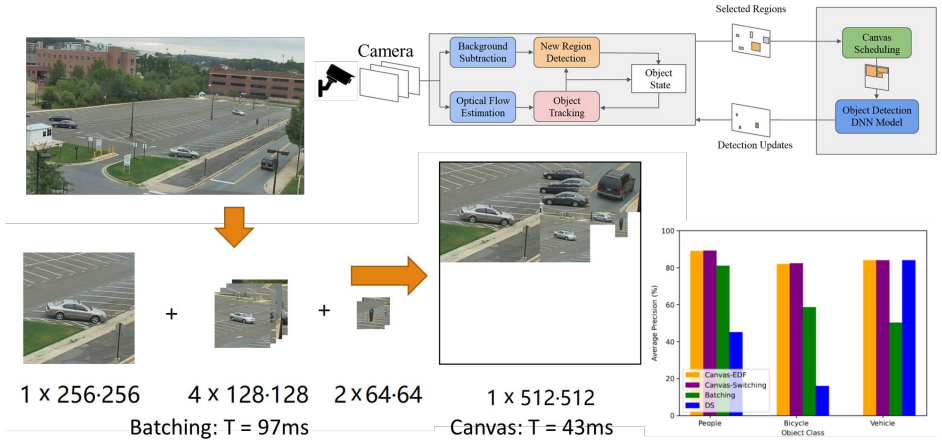
Challenge 2: **Model Compression**

How effective is neural network model compression?

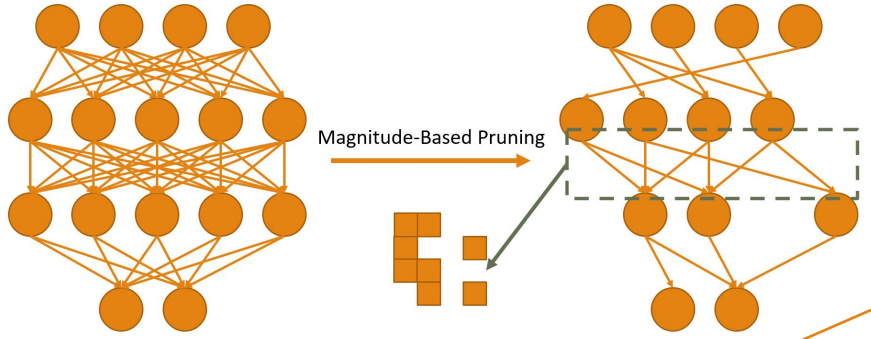
How to improve the latency/accuracy trade-offs?



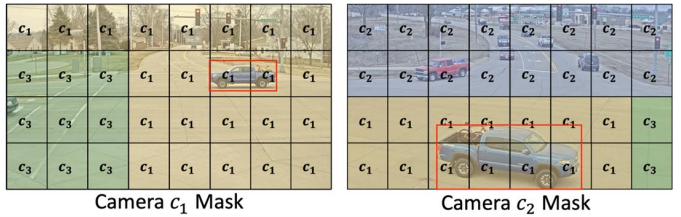
Project Idea: Reduce Computational and/or Memory Footprint of Edge AI

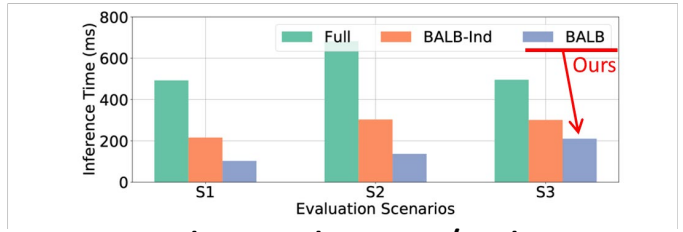


Attention management

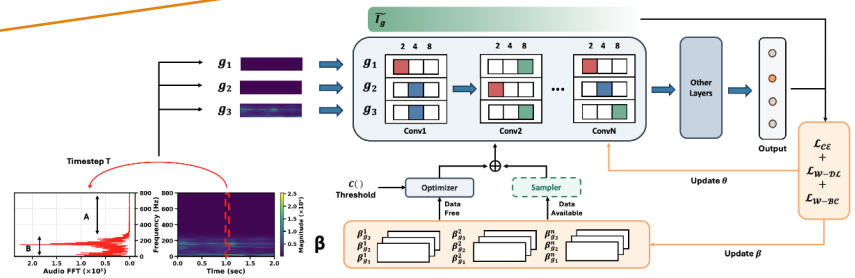


Neural Network Pruning



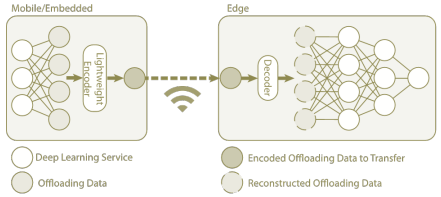


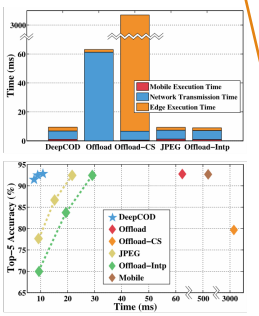
Load Distribution/Balancing



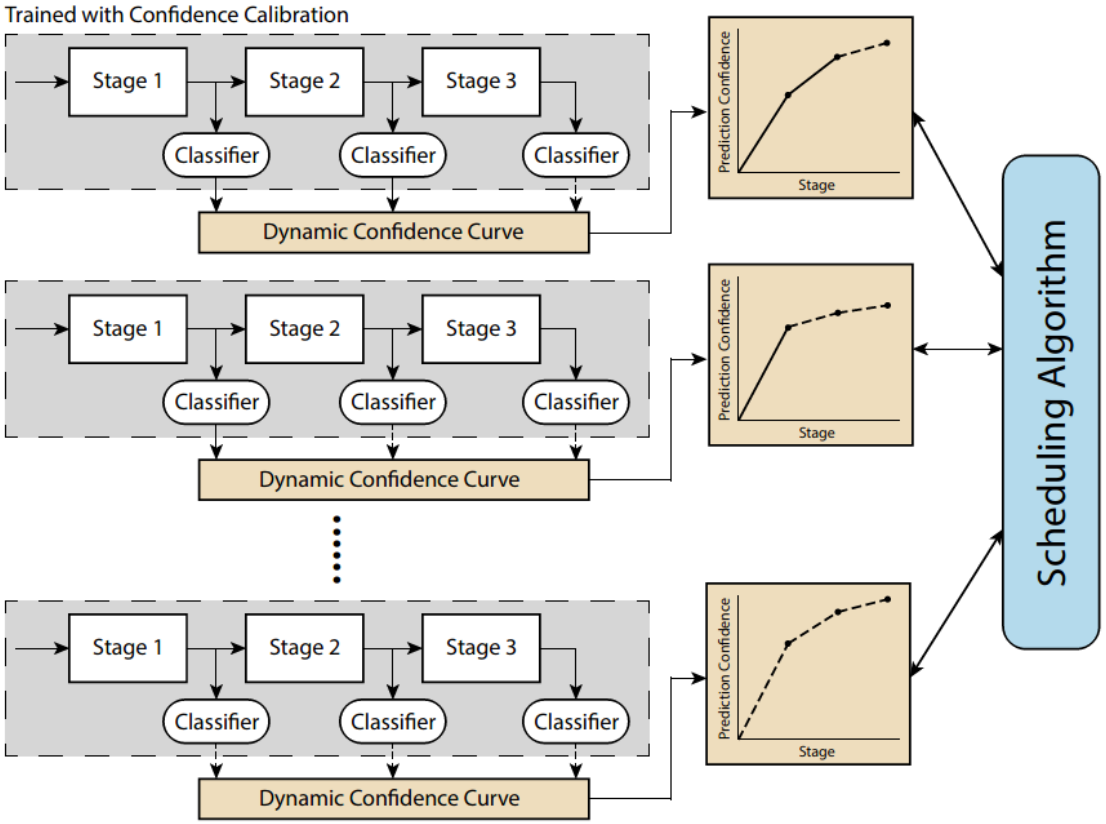
Quantization

Compressive Offloading





Example 1: Early Exit Networks



Idea:

- Break execution into stages
- Use confidence estimates to predict utility from executing the next stage of each AI task
- Scheduler executes the task (stage) with the highest marginal utility next

Example 2:
Real-Time Mixture of Experts Cascades

Key Idea: Disaggregate the AI model → Instead of one big model, use many small specialized case model called “experts”.
Problem: Given a choice of experts, in what order to try them to minimize time to successful inference (say, classification)?

Non-trivial trade-offs exist that are similar to optimizing a memory hierarchy:

- Start with a very simple classifier first:
  - *Pro:* Takes less time to execute
  - *Con:* More likely to fail to classify the target (cache miss), calling for the invocation of another classifier next.
- Start with a more mature classifier first:
  - *Pro:* More likely to succeed at classification
  - *Con:* Takes a long time to execute adding possibly needless latency to the workflow

Other factors affecting the optimal order (workflow) in which classifiers should be consulted:

- Hierarchical relations among classes:
  - Start with “high-level” classifiers (e.g., sedan versus SUV): Can classify more targets approximately (to a broader class)
  - Start with “specialist” classifiers: Can correctly identify the class of only a smaller subset of targets (e.g., type of sedan)
- Correlations among classifiers: Failures of certain classifiers earlier in the workflow help predict performance of later ones

Example 2:
Real-Time Mixture of Experts Cascades

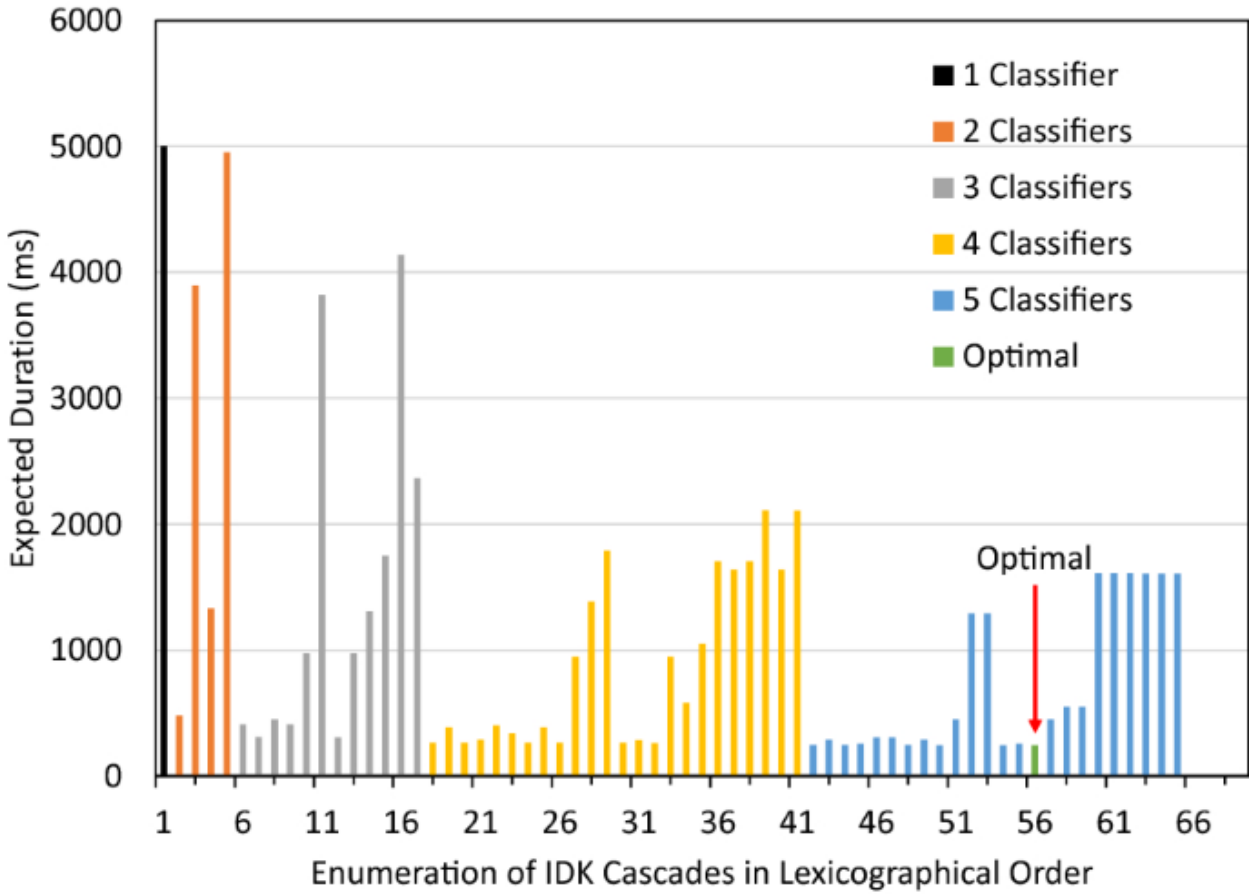
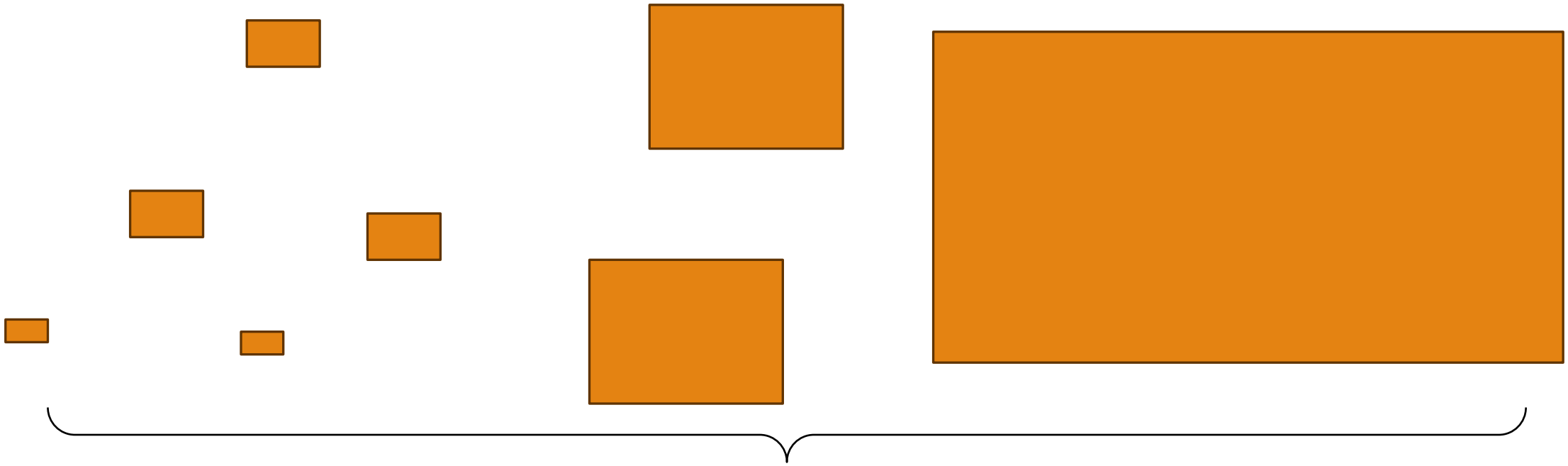


Figure shows expected durations of execution of classifier sequences made of acoustic, seismic, and camera-based object classifiers.

Significant average latency reductions are possible without jeopardizing expected accuracy by optimally ordering the execution sequence of different classifiers (where each escalates to the next when unsure)

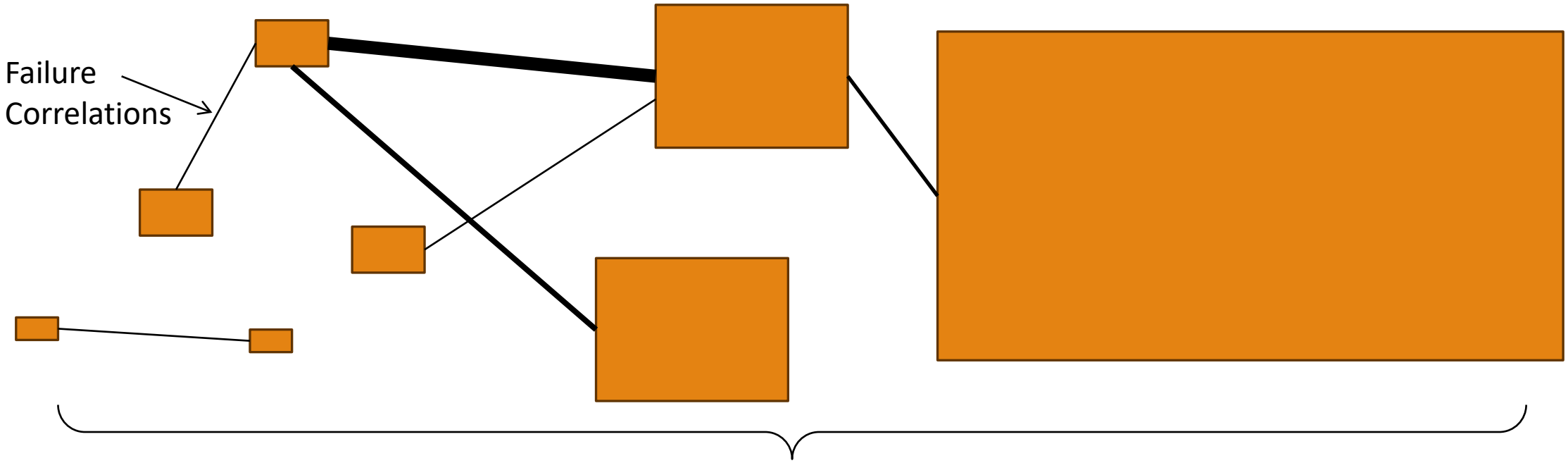
Example 3: Real-time Model/Expert “Caching”



AI models of different quality and different computational complexity

Which models to pre-load in GPU memory for optimal latency?

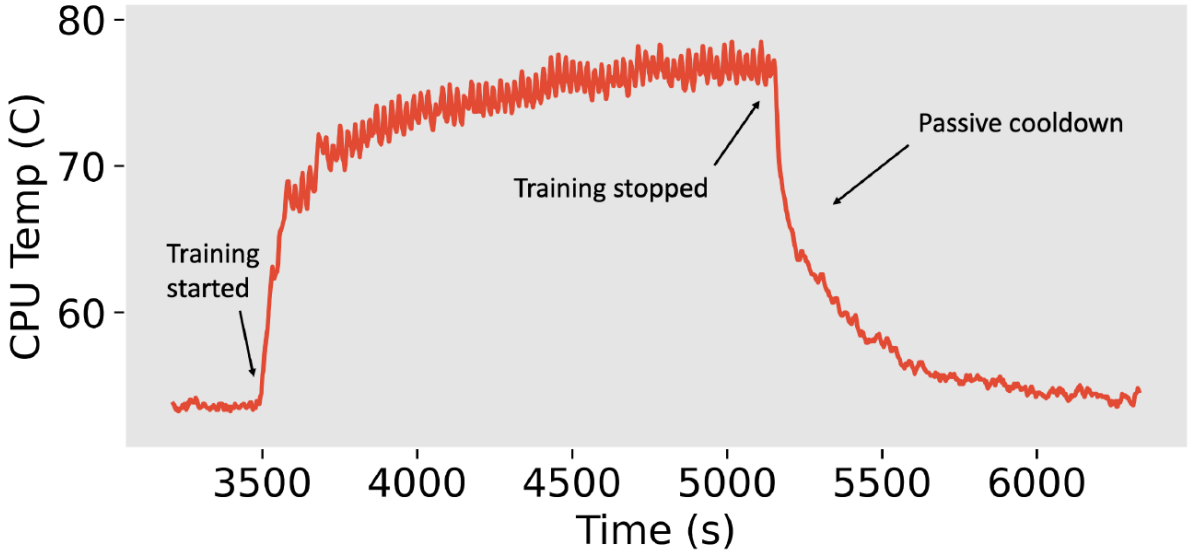
Example 3: Real-time Model/Expert “Caching”



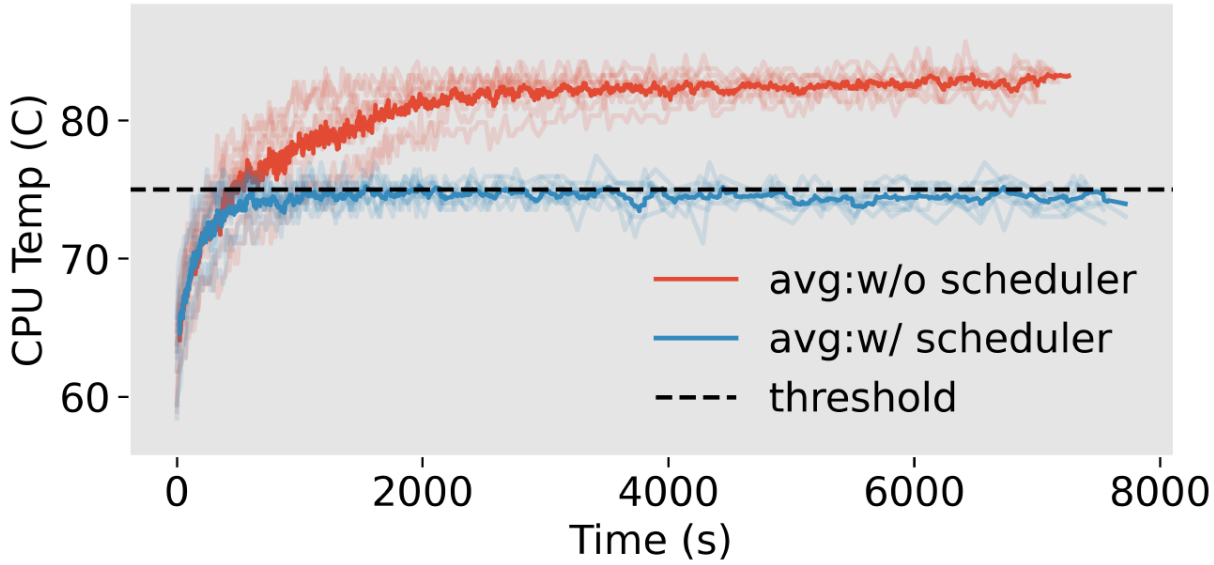
Which models to pre-load in GPU memory for optimal latency?

Example 4: Thermal Scheduling of an AI Module

The need to perform DVFS on the board creates latency/quality/temperature tradeoffs



Overheating may trigger an emergency shutdown

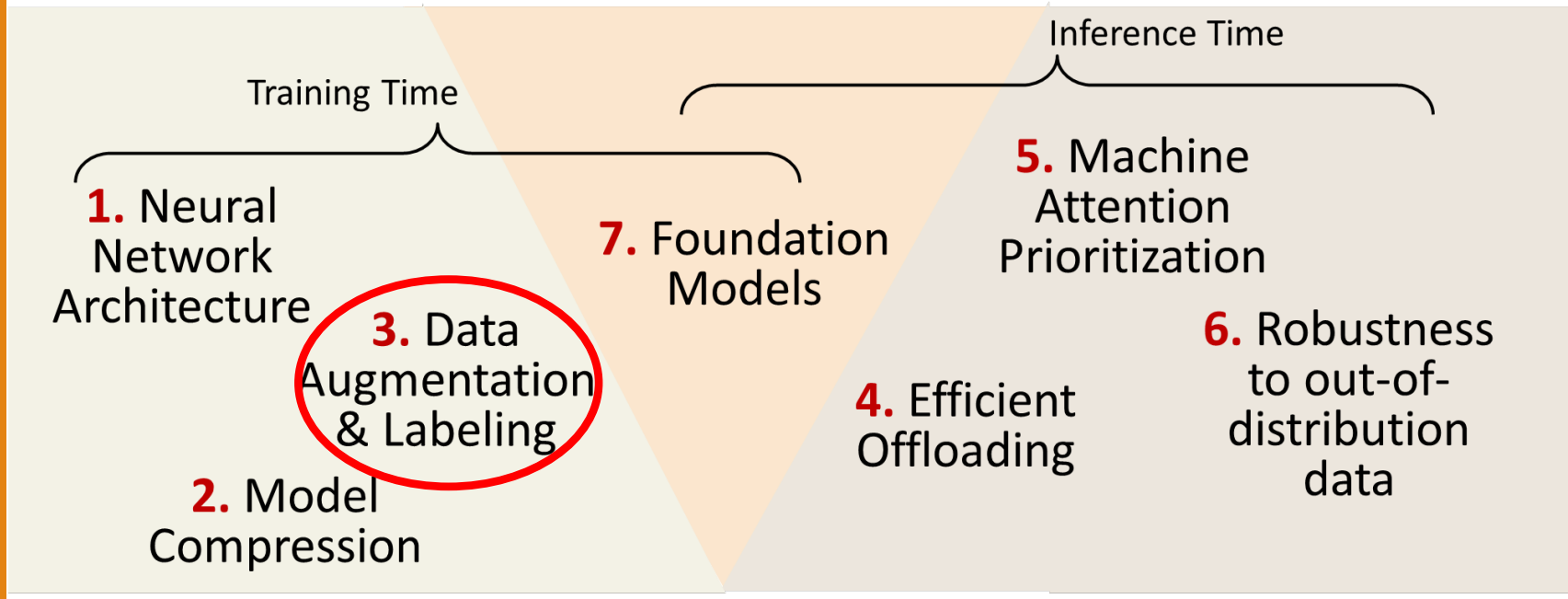


Temperature control prevents shutdown but increases latency, offering a novel trade-off space

Challenge 3: **Data Augmentation and Labeling**

How to generate realistic data for training the AI?

How to automate data labeling?



Project Idea: Data Augmentation by Extrapolating in the Condition Space

Condition 1: Type of Target





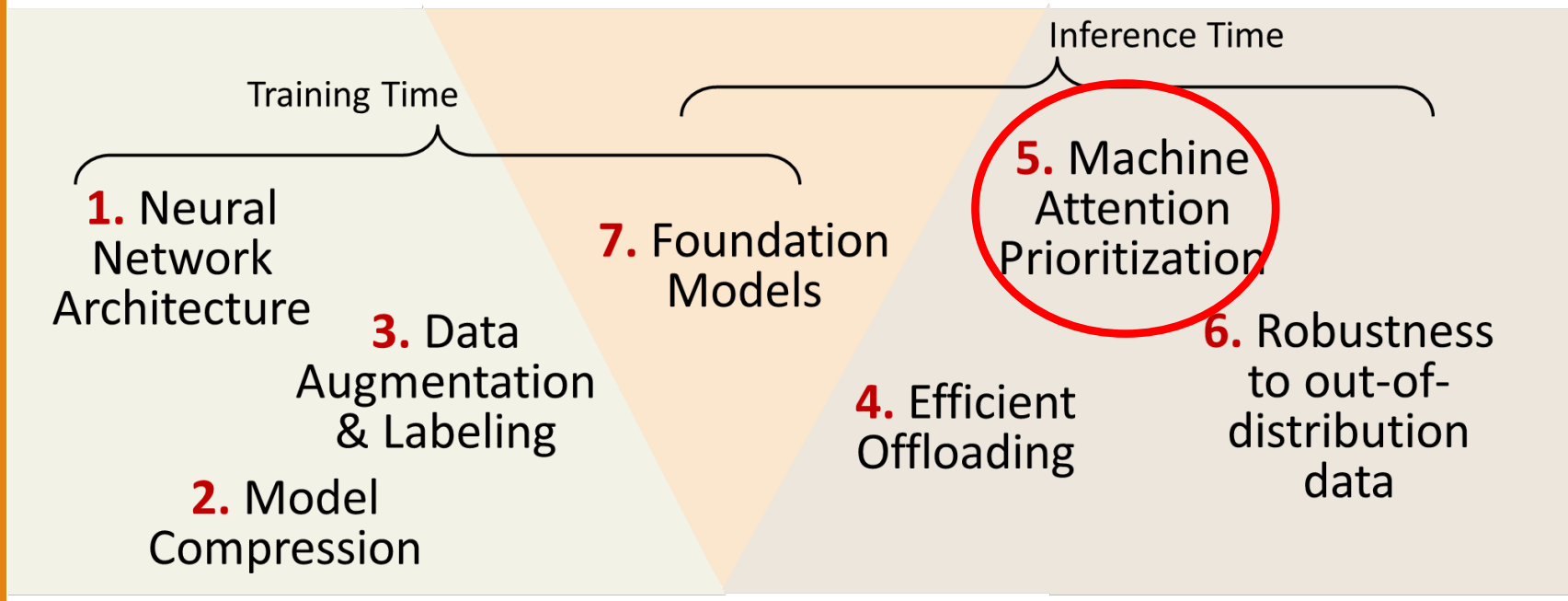
Condition 2: Type of Terrain

Dirt road			??			??
Freeway		??				
Snow					??	
Wet gravel				??		

Challenge 4: Attention Prioritization

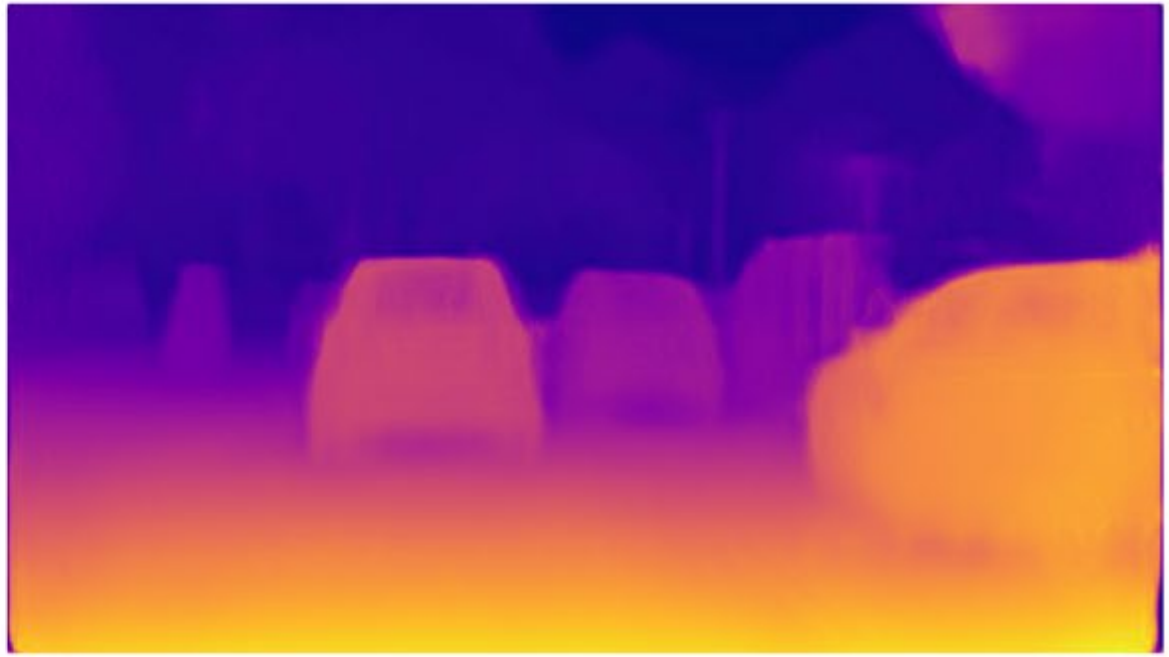
How to prioritize data processing?

How to cue machine attention?



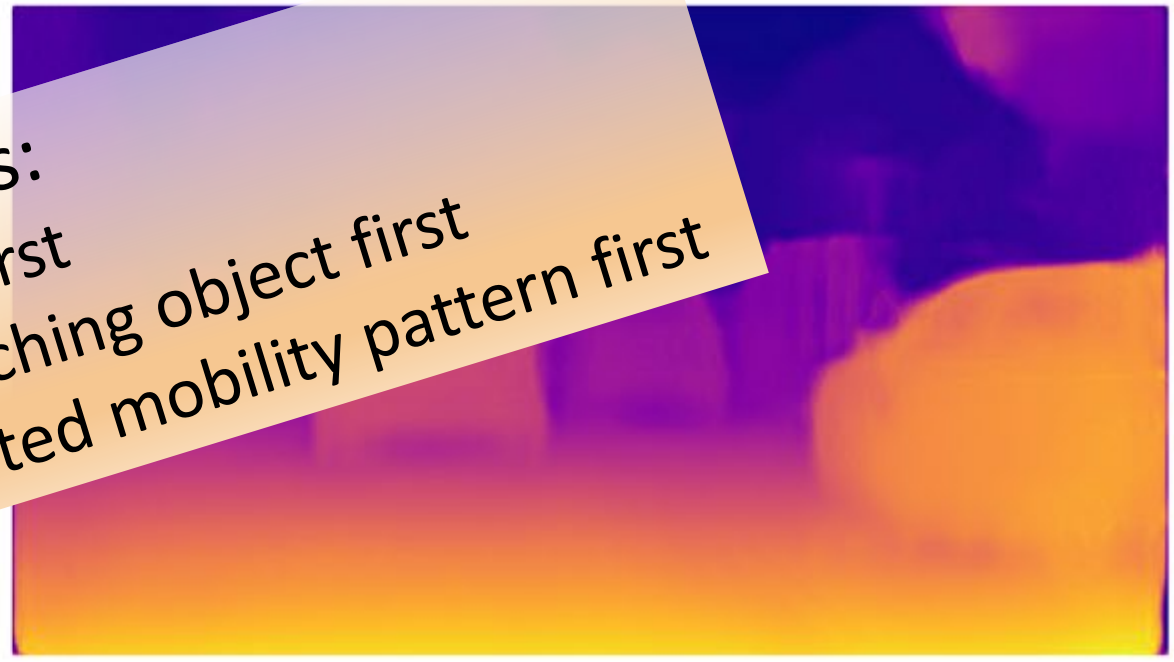
Project Idea: An Algorithm for Machine Attention Prioritization

- Purpose of prioritization:
  - Decide where to look (i.e., where to allocate computational attention)
  - Decide on (scene segment) prioritization and processing quality



Project Idea: An Algorithm for Machine Attention Prioritization

- Purpose of prioritization:
  - Decide where to look (i.e., where to allocate computational attention)
  - Decide on (scene segment) prioritization and processing quality

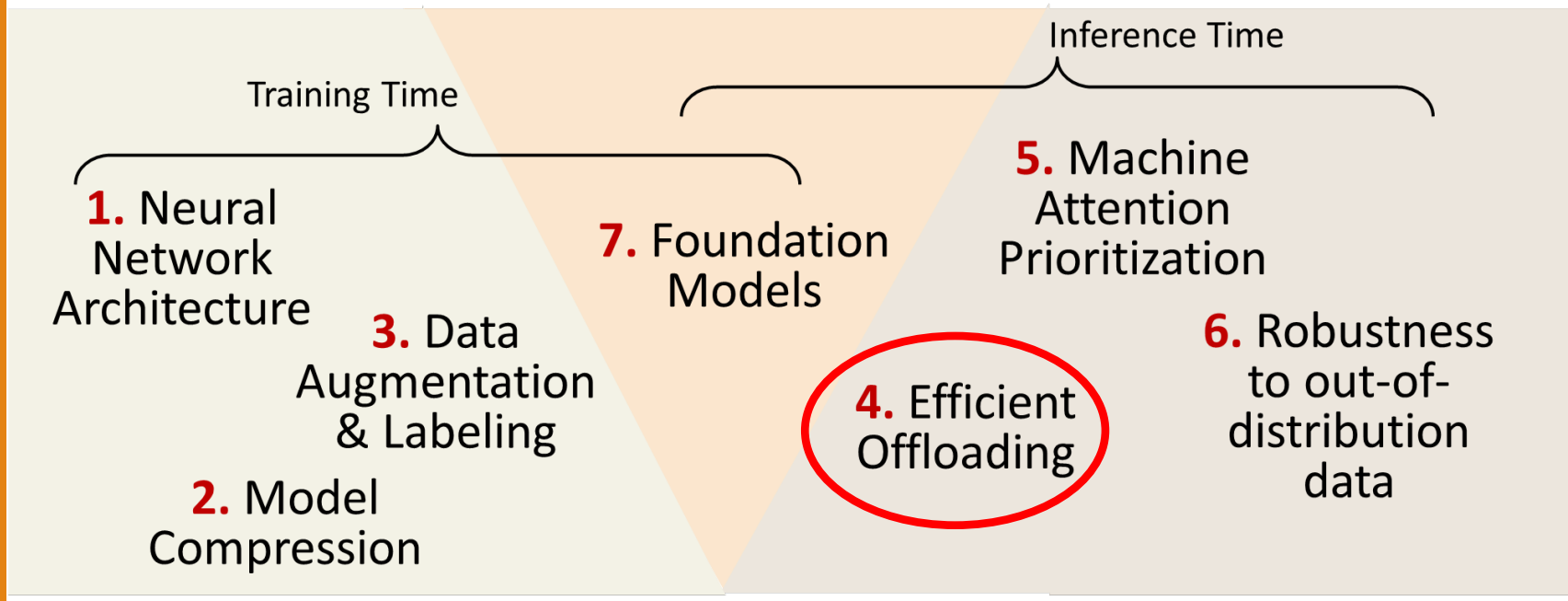


Examples policies:
1. Closest object first
2. Fastest approaching object first
3. Most unexpected mobility pattern first

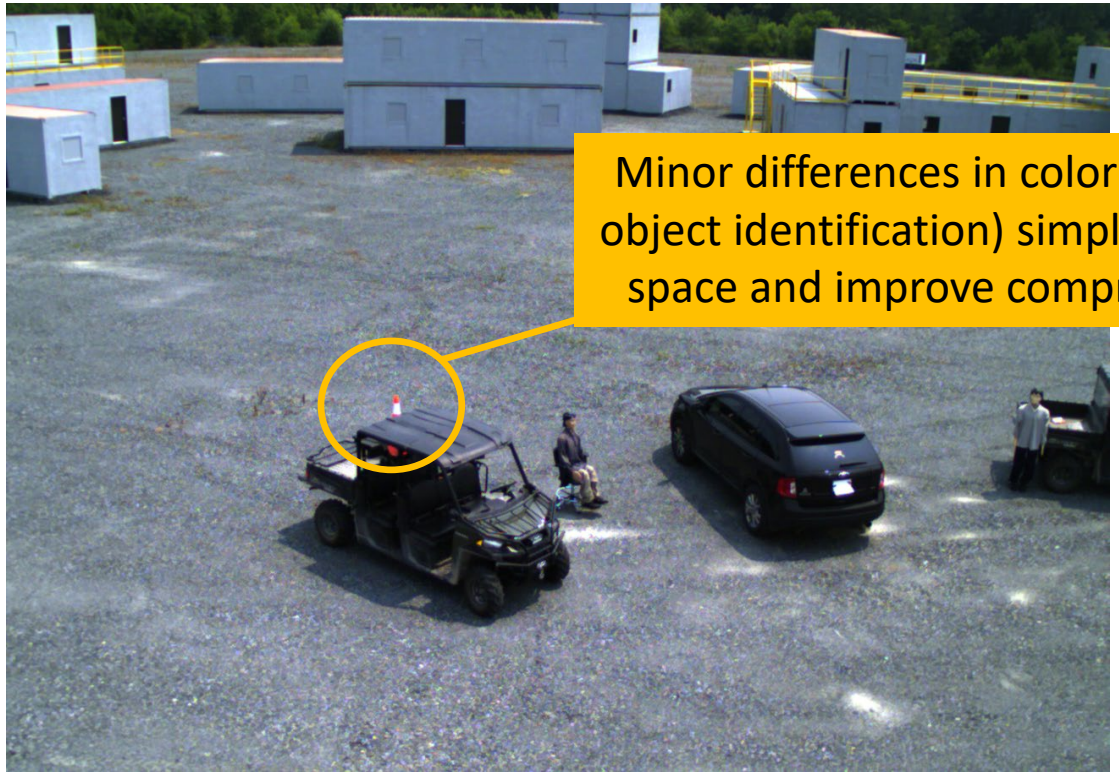
Challenge 5: **Efficient Data Offloading**

How to efficiently offload neural network processing?

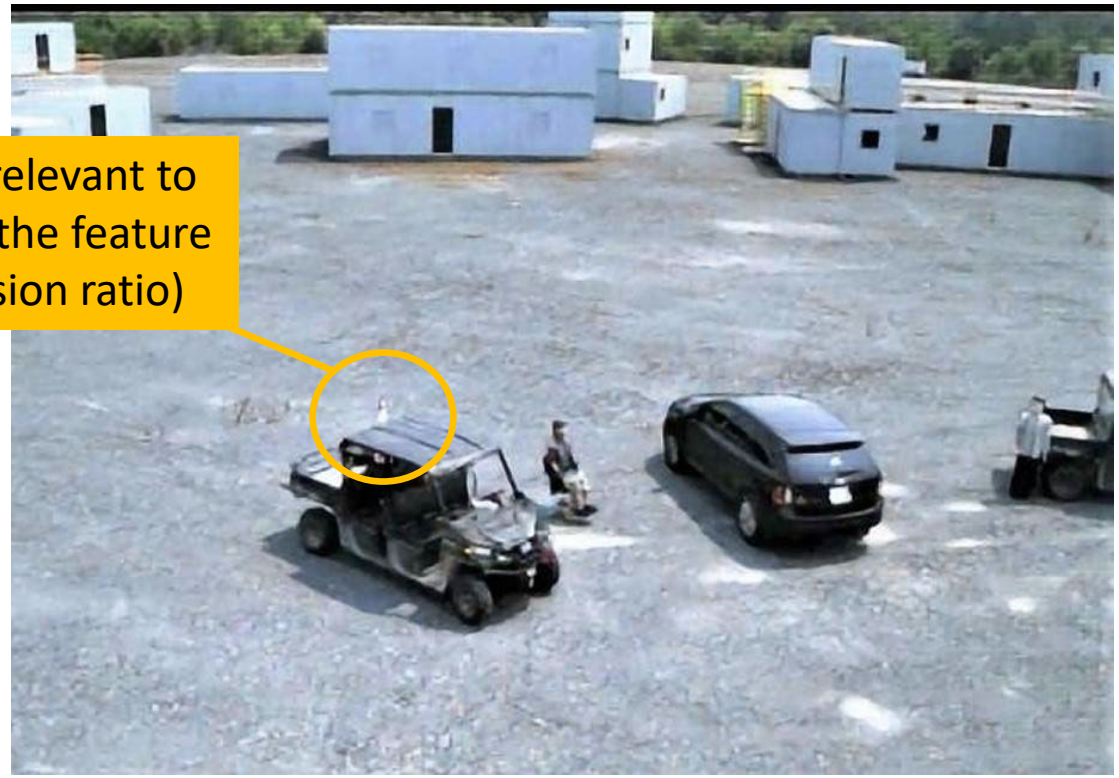
How to compress data with downstream analytics in mind?



Project Idea: AI-Aware Data Compression



Minor differences in color (irrelevant to object identification) simplify the feature space and improve compression ratio)



Original Image

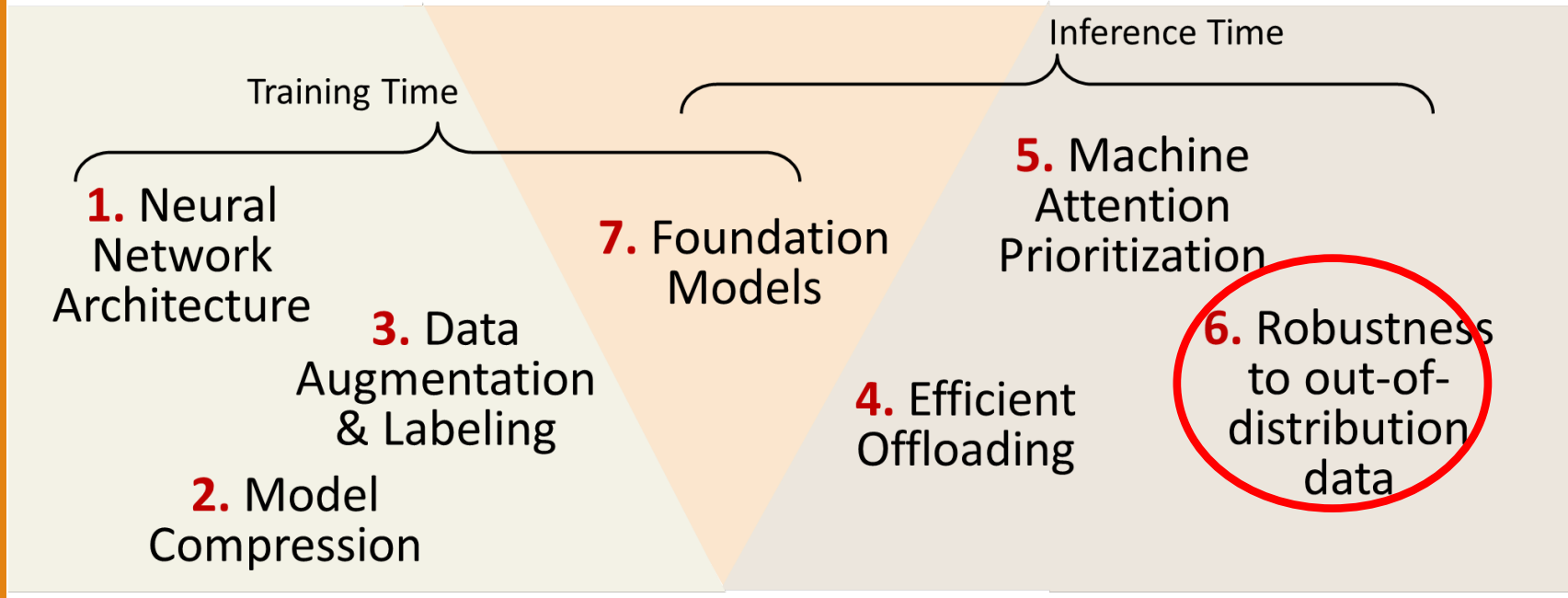
Reconstructed Image

Roughly 4-times improvement over JPEG compression

Challenge 6: **Robustness to Out-of-Distribution Data**

How to ensure robustness to out-of-distribution data in the field?

How to prevent overfitting?



Project Idea: Estimating Confidence in Inference Results

How can a classifier compute confidence in its own outputs? In general, if the environment is close to its training data, the classifier's output should be more reliable, but if the environment is different, the classifier's output may be wrong. How to estimate shift in the current environment (with respect to training) in order to assess output reliability?

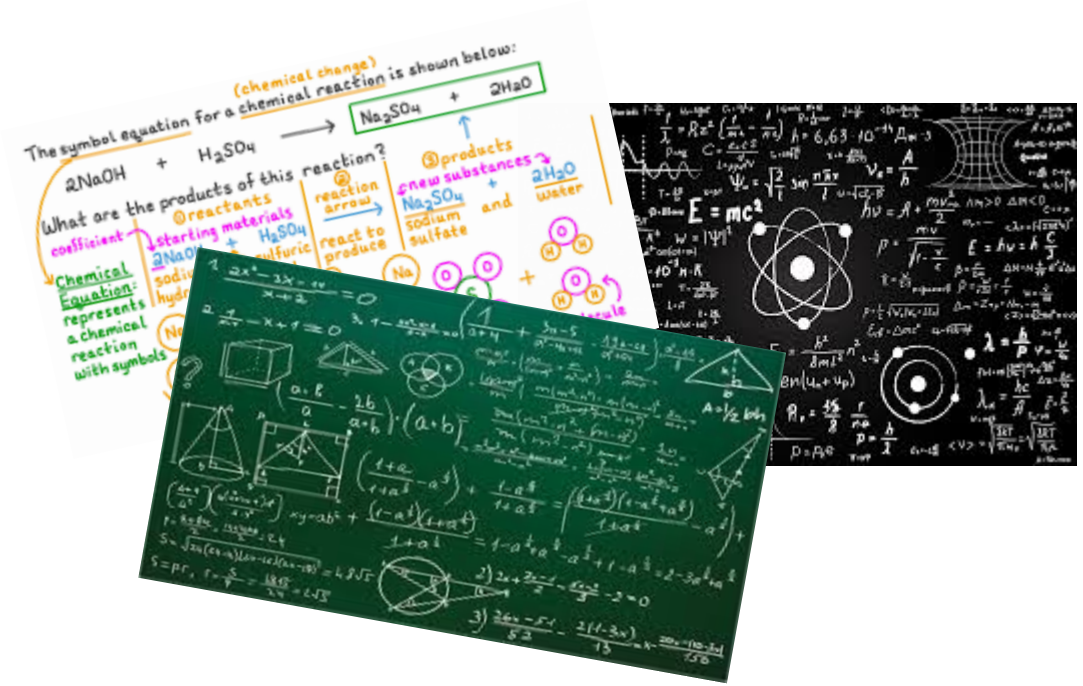
Part III

Systems and Applications

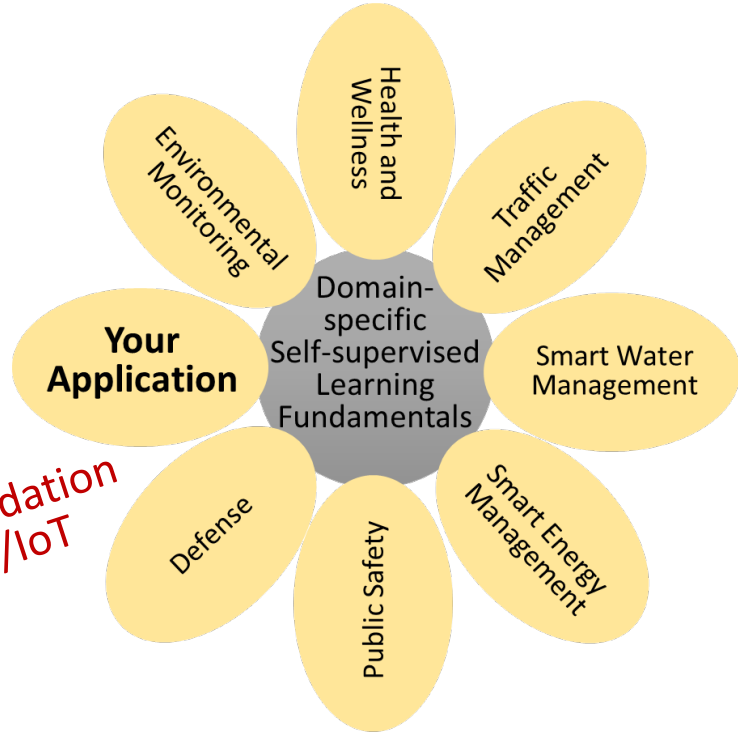
Outline

- Part I: Past Project Examples
- Part II: AI for IoT Challenges
- Part III: Systems and Applications

Project Idea:
Design a Foundation Model for an IoT Application



Today's human representation of knowledge



Train a novel foundation model for an CPS/IoT application

Tomorrow's foundation models

Example Applications

- Drone detection/classification (from acoustic data)
  - Based on analysis of correlations/covariances across a number of distributed microphones
- Vehicle tracking (from acoustic, seismic, mmRadar, LIDAR, neuromorphic camera, or other sensors)
- Human activity recognition (from wearables or ambient signals such as WiFi and RF sensors)
- Human emotion recognition (from wearables, microphones, or other sensors)
- Appliance detection from smart meter data
- Scene feature recognition from encrypted video

Project Idea: Improved Representation Learning for IoT Data

- Pick a modern representation learning approach
- Identify a deficiency when it comes to representing IoT data (e.g., deficiency representing frequency domain signals, spatial observations, sensor time-series, important events, etc)
- Redo the approach to eliminate the deficiency

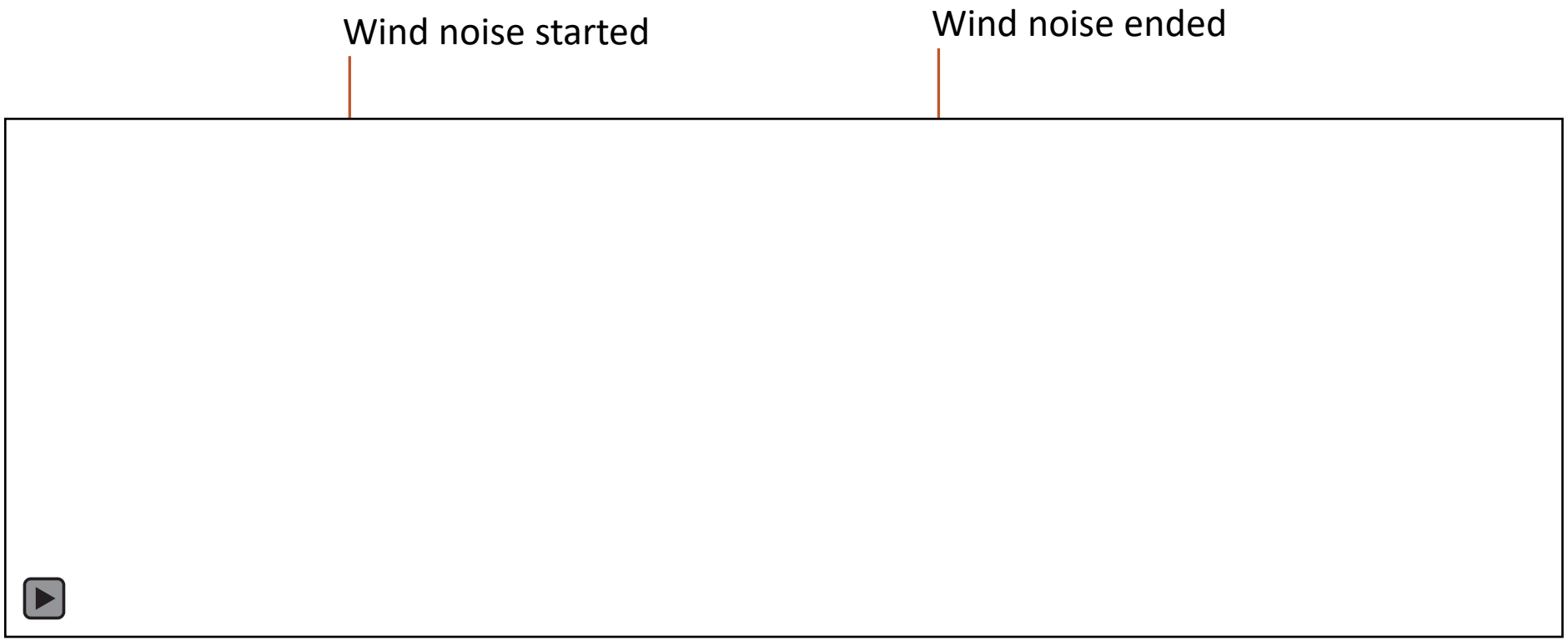
Project Idea: Benchmark LLM Deficiencies in Reasoning from IoT Data

- Generate a benchmark to test LLMs' ability to reason about IoT data. For example:
  - Benchmarks for spatial-temporal reasoning
  - Benchmarks for anomaly detection
  - Benchmarks for frequency domain analysis
  - …

Project Idea: Middleware for Edge AI

- Transparently optimizes communication
- Transparently saves computation (by caching, scheduling, approximation, etc)
- Transparently supports dynamic model selection
- Transparently supports fault tolerance
- Interfaces AI agents and sensor/actuator resources

Dynamic Model Selection

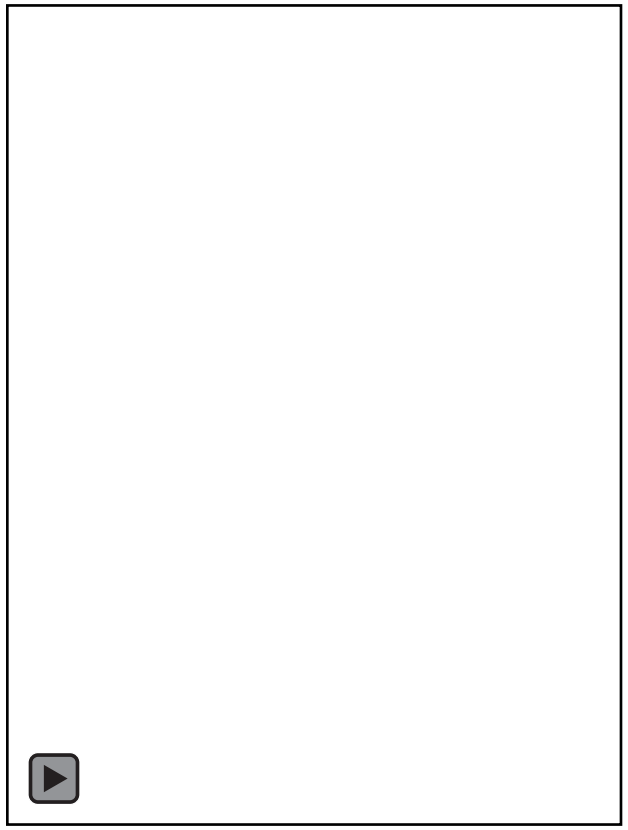


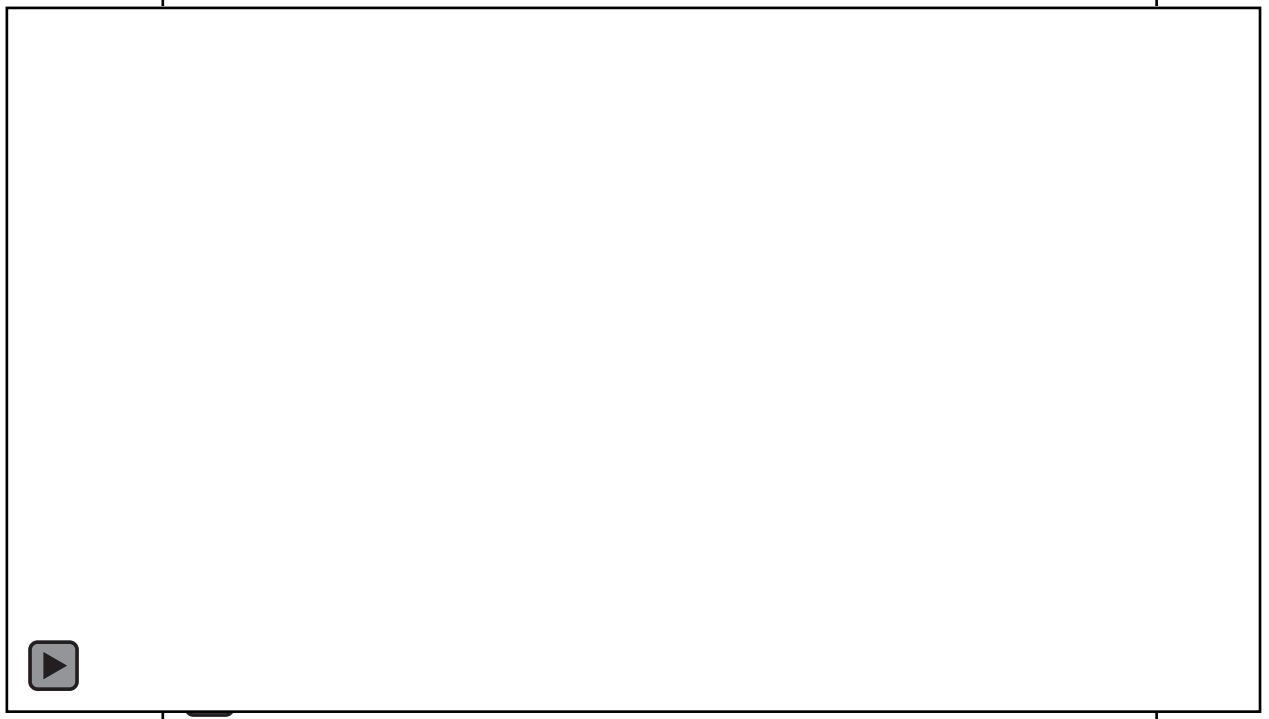
Service Failure

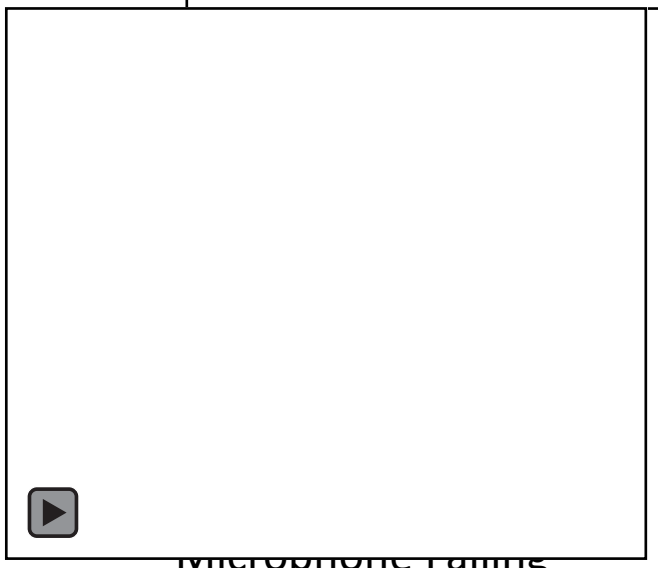
Activated

Standby

Failed

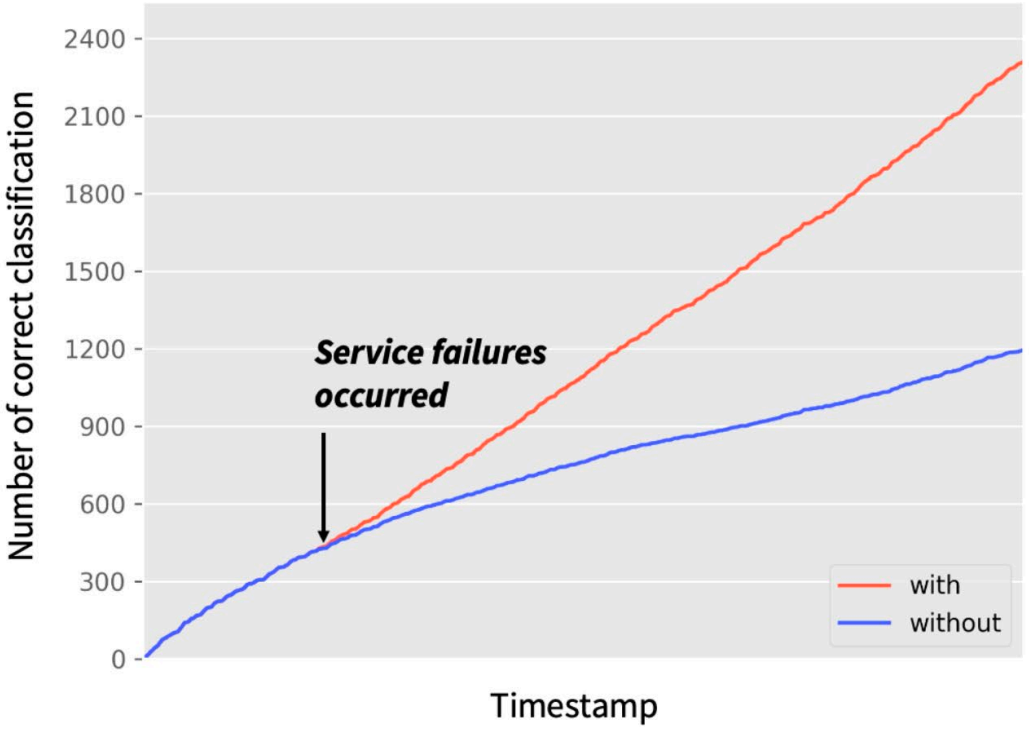


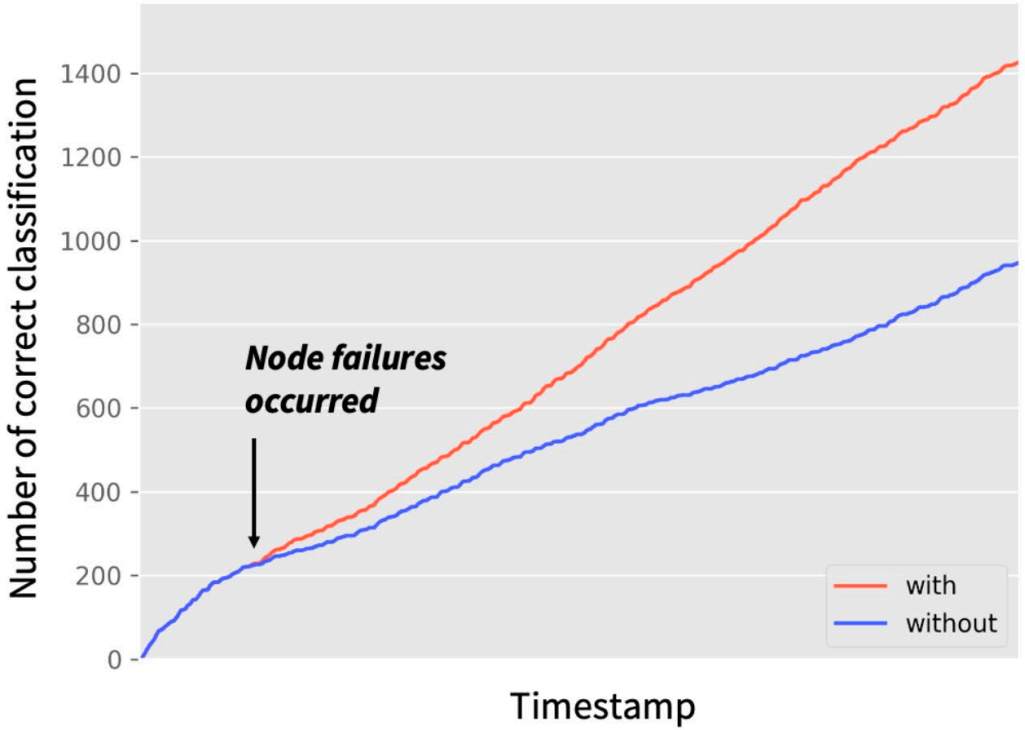




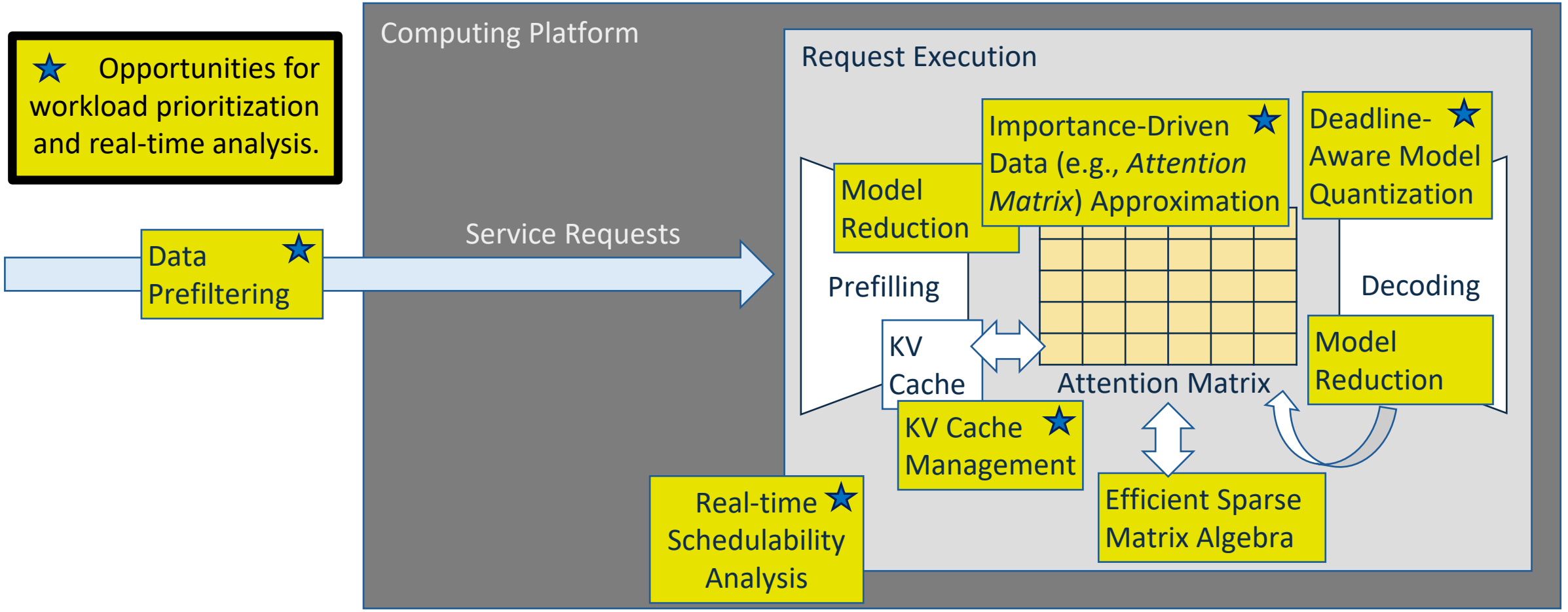
Microphone Failing

Number of Correct Classifications w/ and w/o Failover

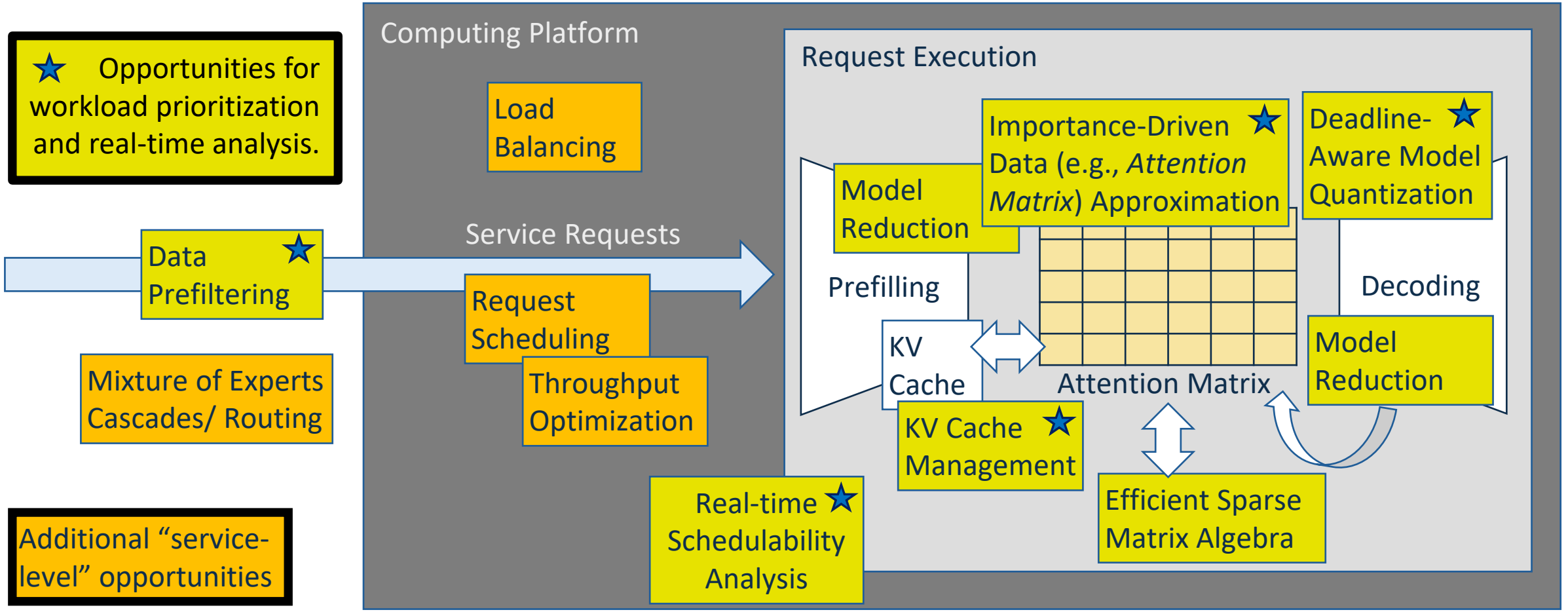




Project Idea: Optimize LLM Model Serving



Project Idea: Optimize LLM Model Serving



Hardware Examples

Seismic sensor

NVIDIA Jetson Nano

USB Wifi Antenna

USB microphone

Raspberry Pi

USB camera

Arduino board

USB 4G Dongle

Dataset Examples

- Acoustic drone signature data set (available on demand)
- Vehicular acoustic/seismic signature data set (available on demand)
- Human activity recognition data sets
  - Example: https://www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset
- Human emotion detection data sets
  - Example: https://www.kaggle.com/datasets/ananthu017/emotion-detection-fer
- Smart meter data sets for non-intrusive load monitoring (appliance detection)
  - Example: https://zenodo.org/records/10875988
- Autonomous car data sets
  - Example: https://waymo.com/open/
- Vehicle engine (OBD II) data sets
  - Example: https://adelaide.figshare.com/articles/dataset/CANdid/29068553

Reminders

- All project groups have been assigned
  - Same group is for the project and presentations
- Projects should be 2-3 people. If your project partners dropped the class and you are alone, please let me know. I will assign you a new group.
- Book (at least) a bi-weekly project zoom meeting with me asap. A list of bi-weekly slots starting next week or the week after is available for booking at the link below. Book one for a bi-weekly meeting or two in consecutive weeks for a weekly meeting. Please indicate your group number when booking.
  - https://doodle.com/meeting/participate/id/bqplRv7a
- Your project title and abstract are due Feb 10th. (Feel free to discuss the topic with me ahead of time on your group channel on Piazza, especially if you anticipate needing hardware.)
  - I will read your abstract and give you feedback on the project
  - You will generate a 2-page project description/plan by Feb 26th. It should include (i) the key topic (what are you going to do), (ii) what's innovative/different about it compared to the state of the art, (iii) why should people care, and (iv) an approximate execution timeline.
- Student-led talks:
  - There are 12 topics to choose from (see schedule). By Thursday 2/5, please bid on topics you'd like to cover by sending a post on your group channel on Piazza identifying your first, second, and third topic choices.
  - I will assign 1-2 topics per group. If your group is assigned 1 topic (those are mostly before Spring break), please expect to summarize 6 key papers in your talk. If your group is assigned 2 topics (those are mostly after Spring break), please expect to summarize 3 papers per topic.
  - Student-led talk assignments will be announced this Friday (together with recommended papers per topic).