A Roadmap for Real-time Embedded AI

Presented by: Tarek Abdelzaher, UIUC (PI)

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Real-time Research: A Time of Big (Collaborative) Growth!

Why is the recent AI/ML revolution a key opportunity for real-time computing?

- *We specialize in managing bottleneck computing resources*.
  - → AI/ML is creating the world's largest computing bottleneck!
- *We specialize in embedded computing*
  - → Embodied AI is embedded AI

AI + RT/Embedded collaborations could bring a wealth of new perspectives and applications

Acknowledgements

NEWS ▾ FEATURES ▾ OPINIONS ▾ SPORTS ▾ BUZZ ▾ CU FAITH & LIFE HOUSING PHOTO

2017

NEWS

Illinois chosen to lead $25 million research project

BY SAMANTHA BOYLE, STAFF WRITER

OCTOBER 21, 2017

Computers and other cyber technologies are playing
cyber threats, $25 million has been allocated to the U

Tarek Abdelzaher, academic leader of the Alliance for

Semiconductor Research Corporation

BOEING

IBM

NSF

Home > Computer Science News > Abdelzaher's IoBT REIGN Allian...

2022

Abdelzaher's IoBT REIGN Alliance Receives 5-year Extension Worth Up to $25.5M

University of Illinois and IBM to launch $200M Discovery Accelerator Institute

Tweet | Share Article | Print | Order Reprints

New Center Based at UIUC will Develop Distributed Computing Technology for 2030 and Beyond

Computer Science News

Overview

Donor Profiles

1/5/2023 8:04:41 AM

Funded by a $31.5 million grant from the Joint University Microelectronics Program 2.0 (JUMP 2.0), the University of

Challenge Set #1:
AI + Managing Bottleneck Resources

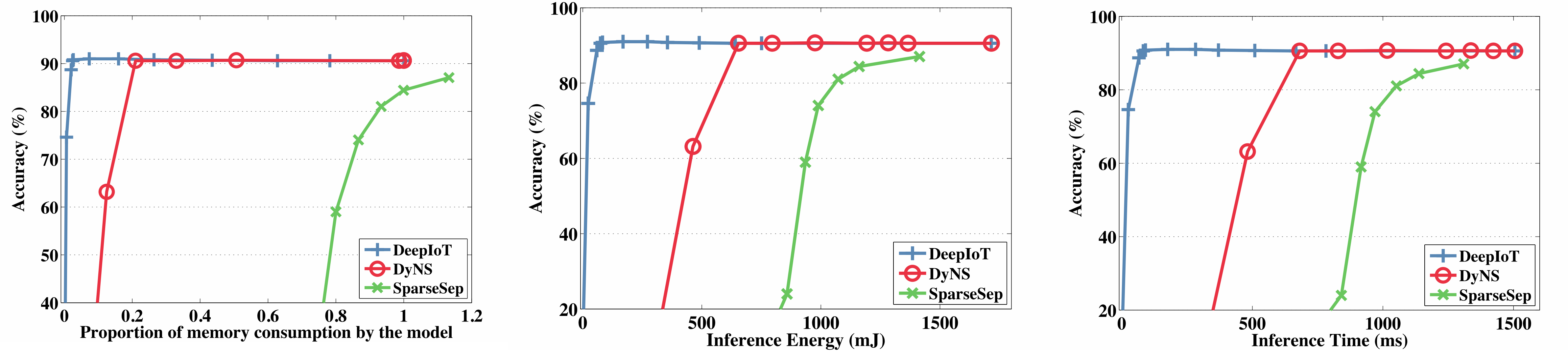
Challenge:

AI and Time Constraints

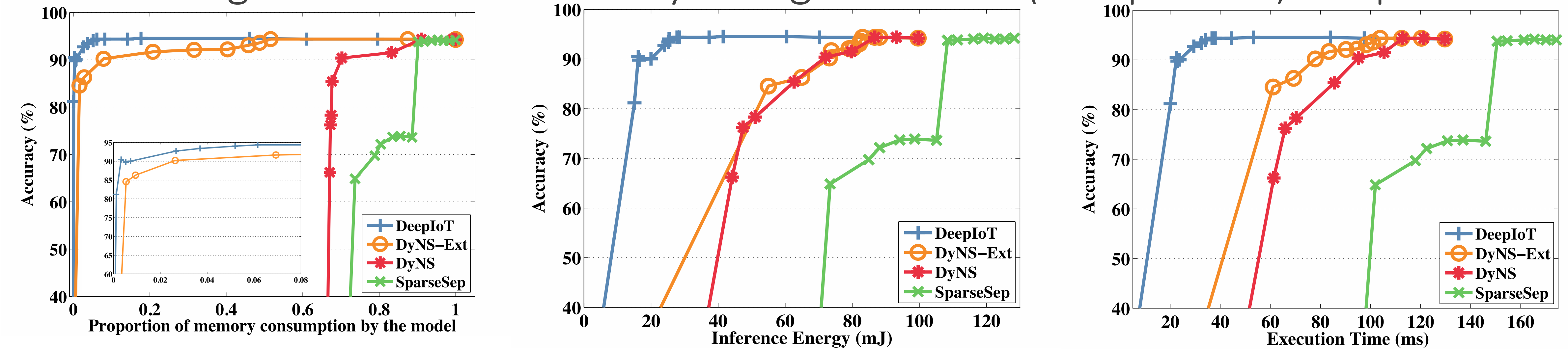
Exploit latency/quality trade-offs in AI to meet time constraints

[1] Shuochao Yao, Yiran Zhao, Aston Zhang, Lu Su, and Tarek Abdelzaher, "DeepIoT: Compressing Deep Neural Network Structures for Sensing Systems with a Compressor-Critic Framework," In Proc. *15th ACM Conference on Embedded Networked Sensor Systems (ACM SenSys)*, Delft, The Netherlands, November 2017.

Image Recognition with (Compressed) VGGNet

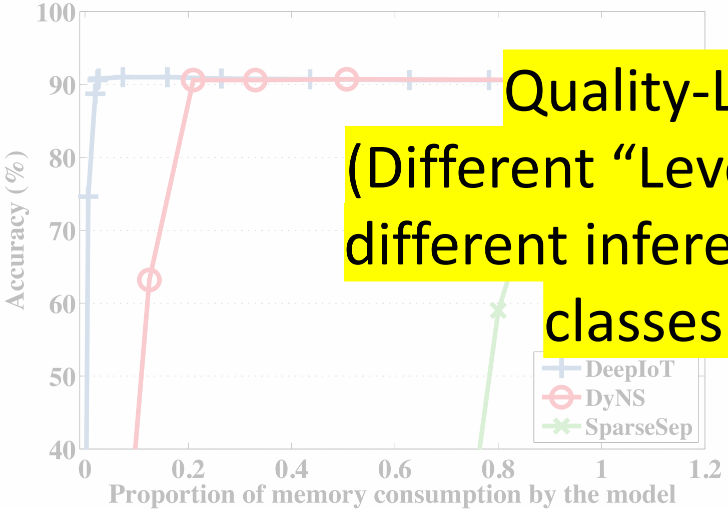


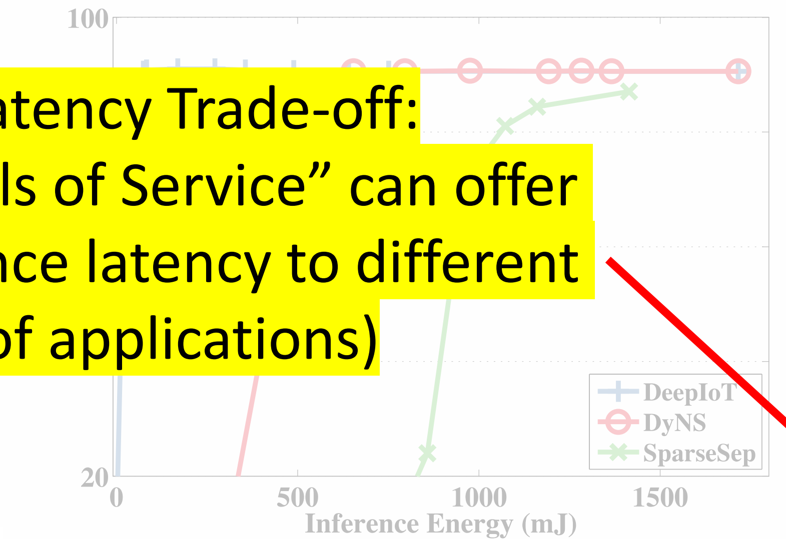
Heterogeneous Human Activity Recognition with (Compressed) DeepSense

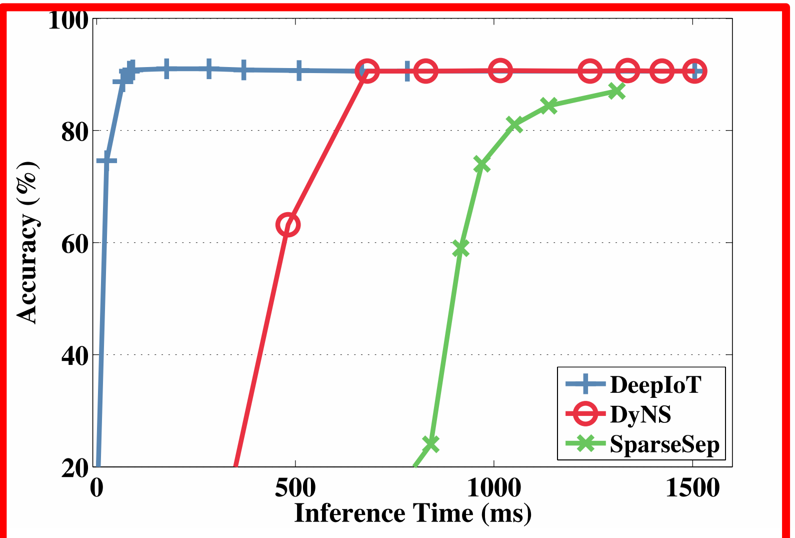


[1] Shuochao Yao, Yiran Zhao, Aston Zhang, Lu Su, and Tarek Abdelzaher, "DeepIoT: Compressing Deep Neural Network Structures for Sensing Systems with a Compressor-Critic Framework," In Proc. *15th ACM Conference on Embedded Networked Sensor Systems (ACM SenSys)*, Delft, The Netherlands, November 2017.

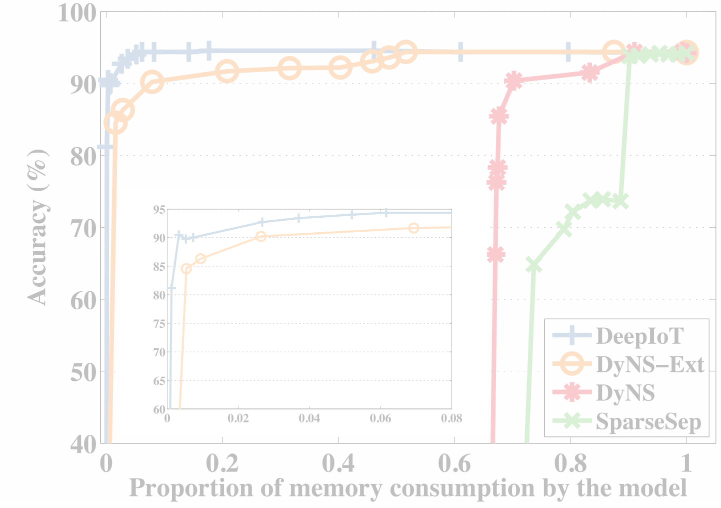
Quality-Latency Trade-off:
(Different "Levels of Service" can offer different inference latency to different classes of applications)

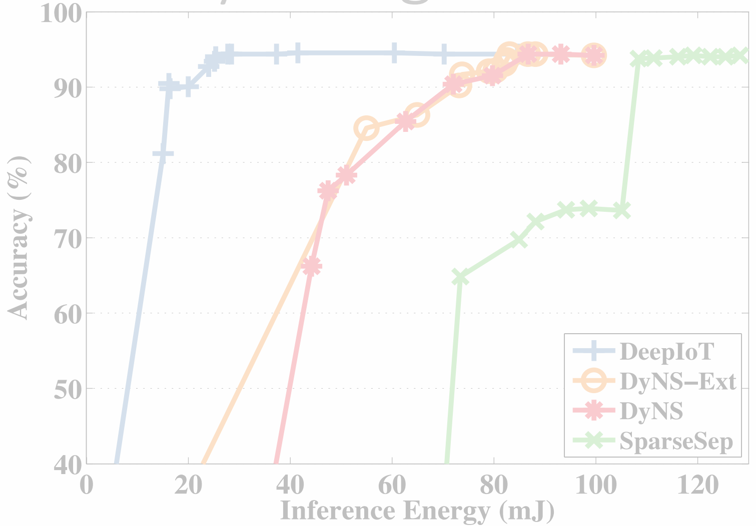


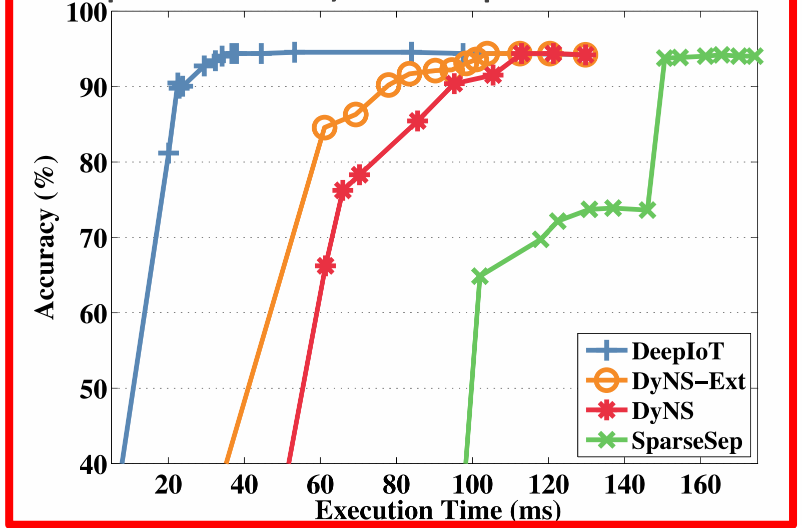




Heterogeneous Human Activity Recognition with (Compressed) DeepSense

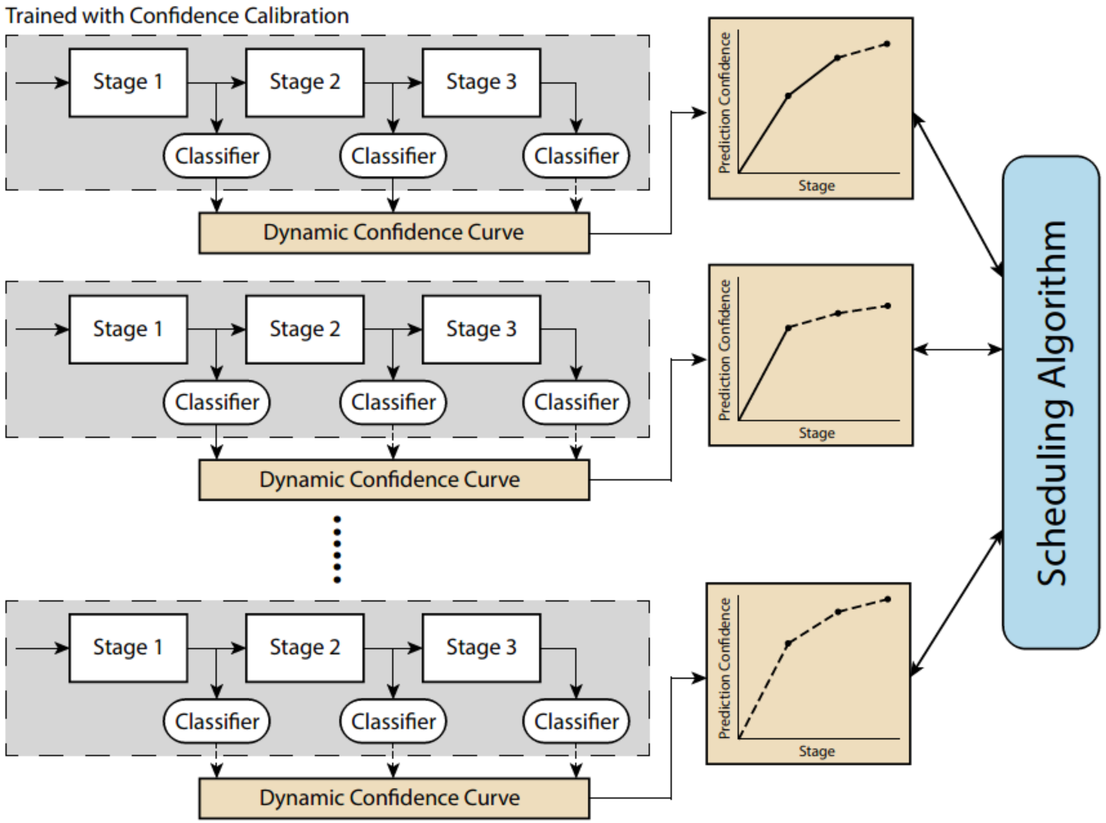






[2] Shuochao Yao, Yifan Hao, Yiran Zhao, Huajie Shao, Dongxin Liu, Shengzhong Liu, Tianshi Wang, Jinyang Li and Tarek Abdelzaher, "Scheduling Real-time Deep Learning Services as Imprecise Computations," In Proc. IEEE International Conference on Embedded and Real-time Computing Systems and Applications (RTCSA), South Korea, August 2020

Real-time Scheduling of Inference Tasks as "Imprecise Computations"

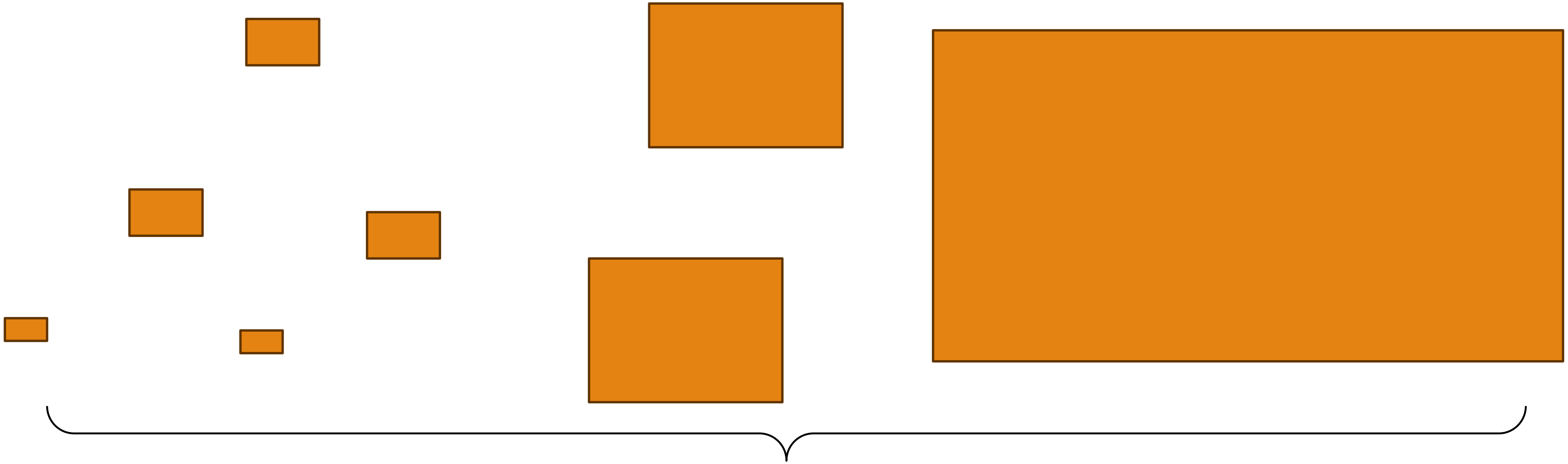


Challenge: Different data inputs offer different degrees of complexity. Some are easily recognizable patterns, but others are not.

Idea:

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

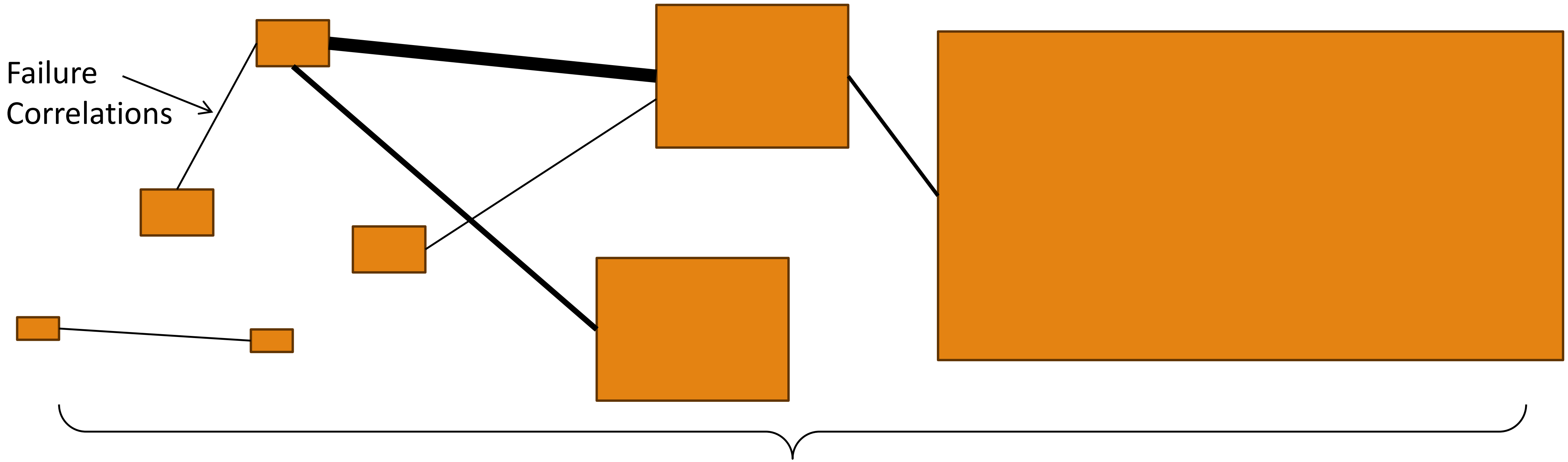
Real-time Model "Caching"
(An idea by Sanjoy Baruah, Alan Burns, and Rob Davis)



AI models of different quality and different computational complexity

What is the optimal sequence of models to try in order to minimize average latency to successful decision?

Real-time Model "Caching"
(An idea by Sanjoy Baruah, Alan Burns, and Rob Davis)



AI models of different quality and different computational complexity

What is the optimal sequence of models to try in order to minimize average latency to successful decision?

[3] Tarek Abdelzaher, Kunal Agrawal, Sanjoy Baruah, Alan Burns, Robert I. Davis, Zhishan Guo, Yigong Hu, “Scheduling IDK Classifiers with Arbitrary Dependences to Minimize the Expected Time to Successful Classification,” Journal of Real-time Systems, March 2023.

Multimodal Classifier Cascades and Execution Ordering

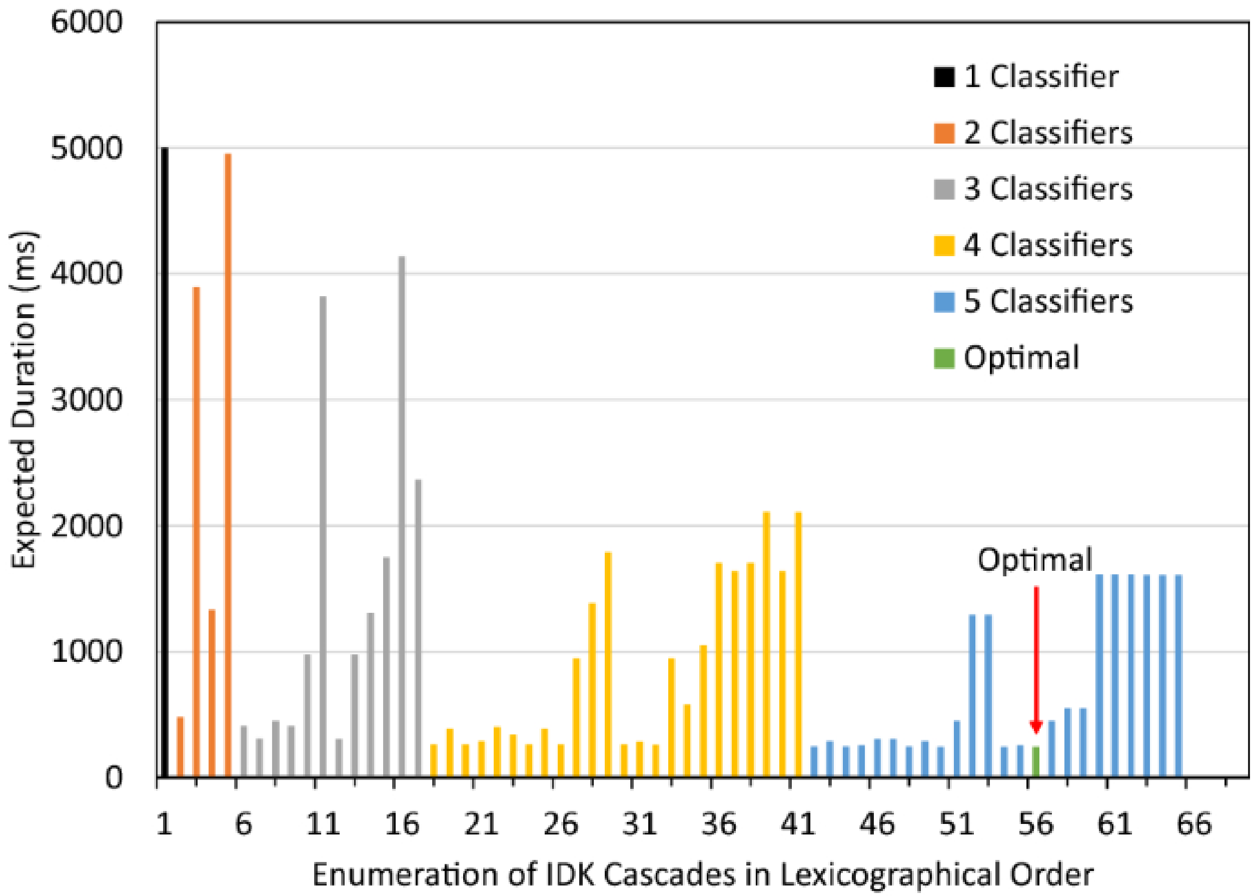


Figure shows expected durations of execution of classifier sequences made of acoustic, seismic, and camera-based object classifiers.

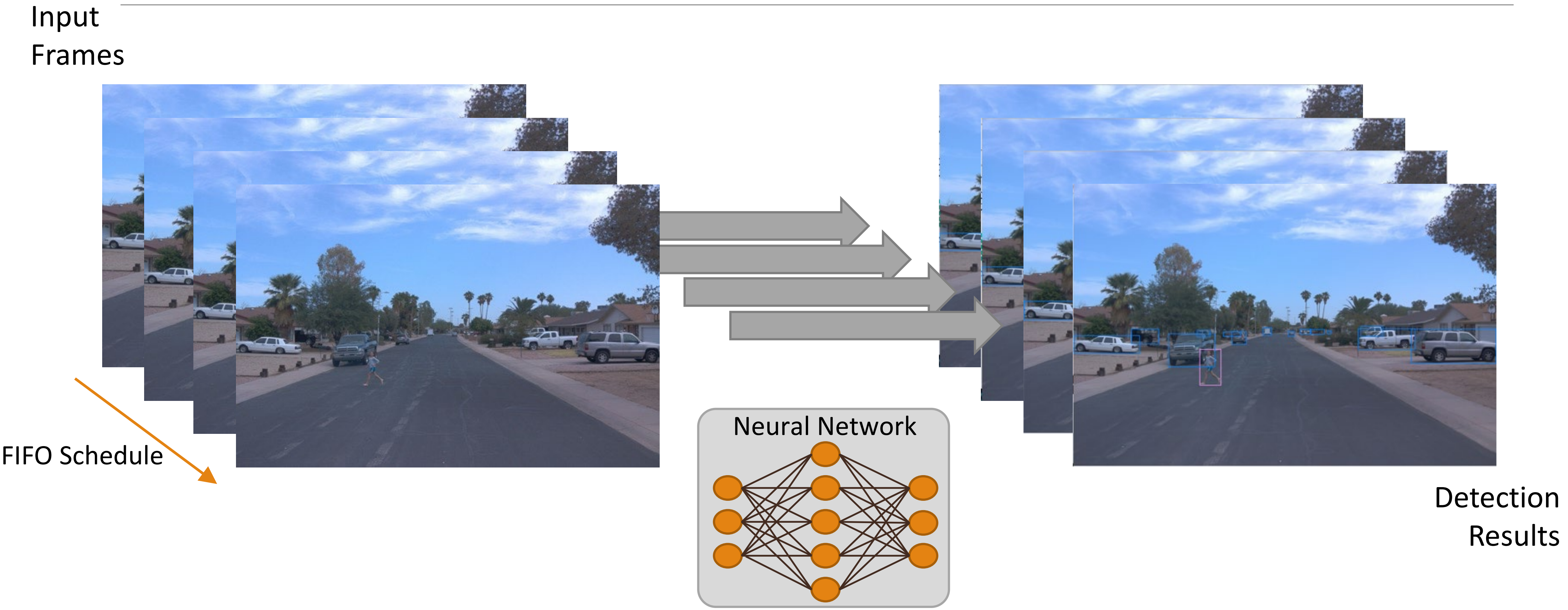
Significant average latency reductions are possible without jeopardizing expected accuracy by optimally ordering the execution sequence of different classifiers (where each escalates to the next when unsure)

Challenge:

Attention Management (Prioritization)

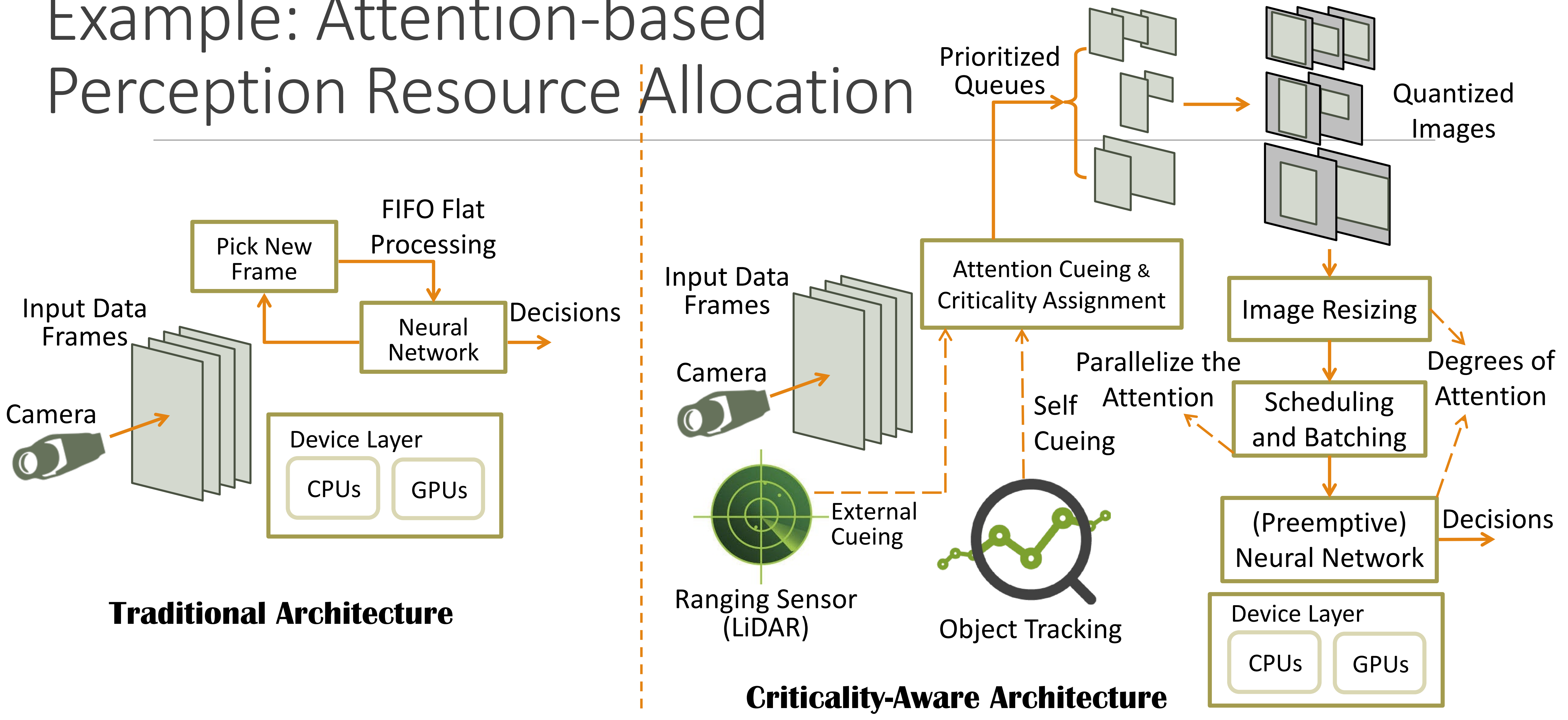
Attend to more relevant parts of the data first

Attention-based Resource Allocation at the Edge



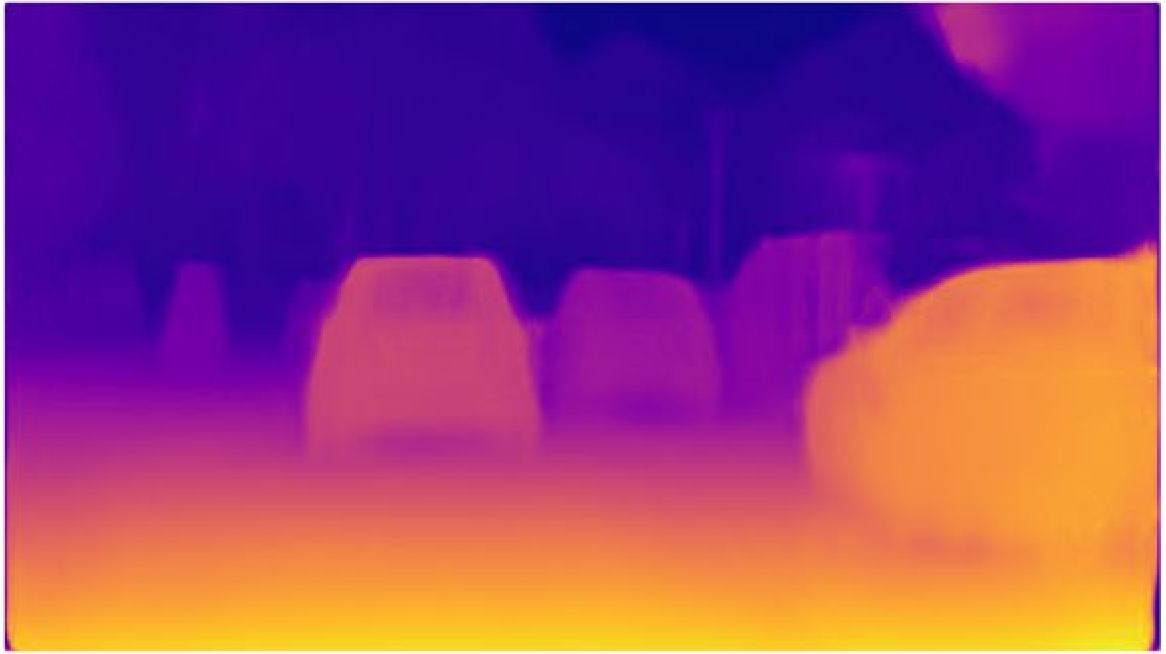
[4] Shengzhong Liu, Shuochao Yao, Xinzhe Fu, Rohan Tabish, Simon Yu, Ayoosh Bansal, Heechul Yun, Lui Sha and Tarek Abdelzaher, "On Removing Algorithmic Priority Inversion from Mission-critical Machine Inference Pipelines," In Proc. *IEEE Real-time Systems Symposium (RTSS)*, Houston, TX (Online), December 2020. **Best Paper Award**

Example: Attention-based Perception Resource Allocation



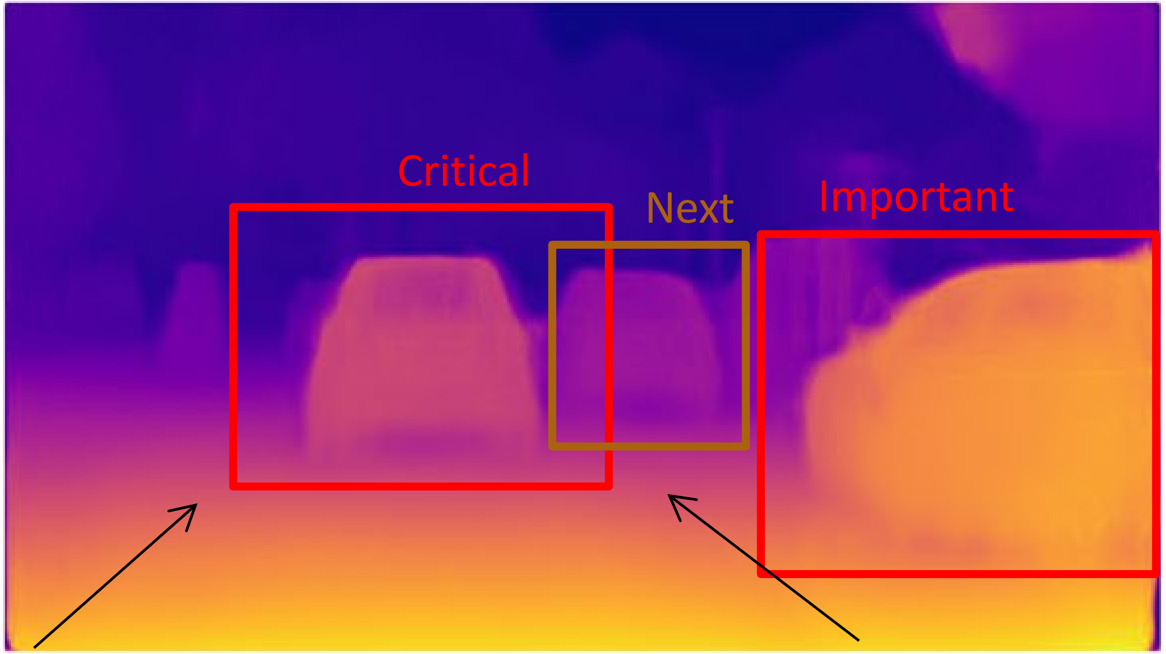
Attention Cueing: Decide What Data Are More Important

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



Attention Cueing: Decide What Data Are More Important

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



[5] Yigong Hu, Shengzhong Liu, Tarek Abdelzaher, Maggie Wigness, Philip David, "On Exploring Image Resizing for Optimizing Criticality-based Machine Perception," *Journal of Real-time Systems*, August 2022.

Give More Important Data Better Service (e.g., Differentiated Perception)

Idea: To save on less important segments, resize them and use a smaller neural network

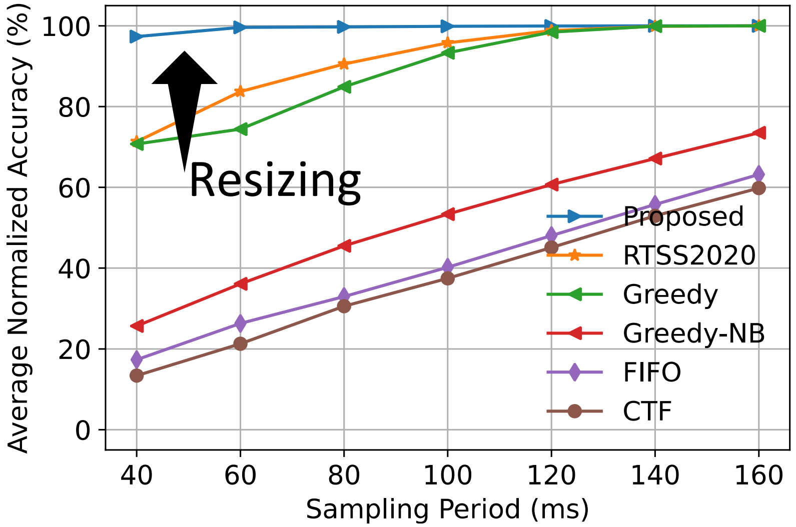
Observations:

Lowest deadline miss rate

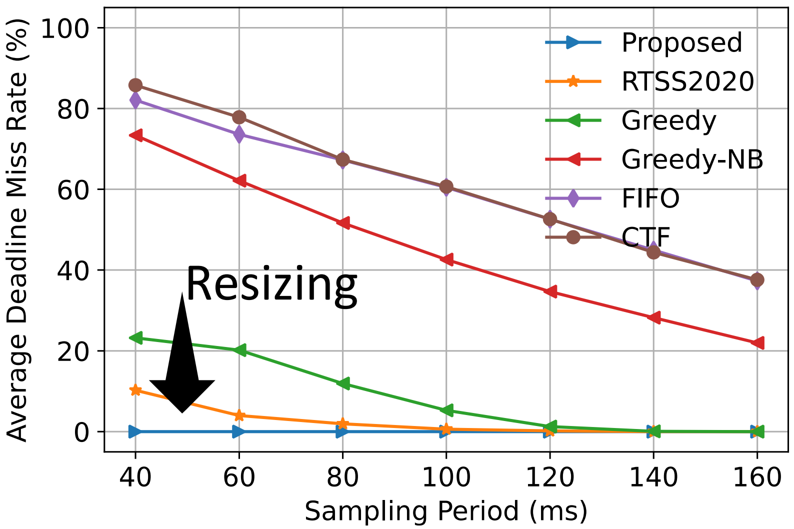
Highest accuracy

Lowest latency

Larger (better) batch size



(a) Normalized Accuracy



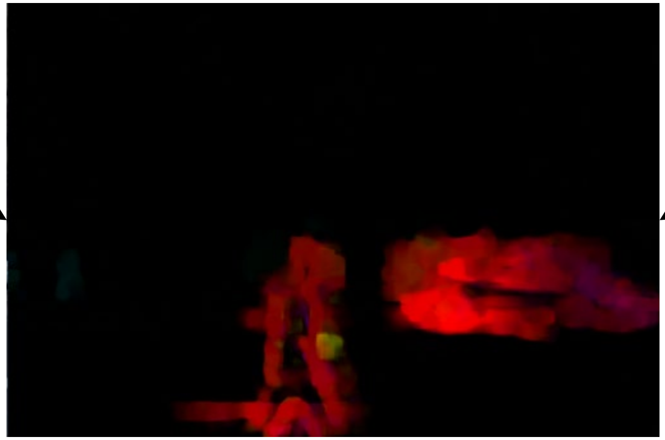
(b) Deadline Miss Rate

[6] Shengzhong Liu, Xinzhe Fu, Maggie Wigness, Philip David, Shuochao Yao, Lui Sha, Tarek Abdelzaher, "Self-Cueing Real-Time Attention Scheduling in Criticality-Driven Visual Machine Perception," In Proc. *28th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, Milano, Italy, May 2022. **Best Paper Award**

Attention (Self-)Cueing

- Optical flow: Pixel-level motion vectors between two frames, caused by the relative movement between objects and the observer.
- Cue attention to regions of larger change.

Previous Frame



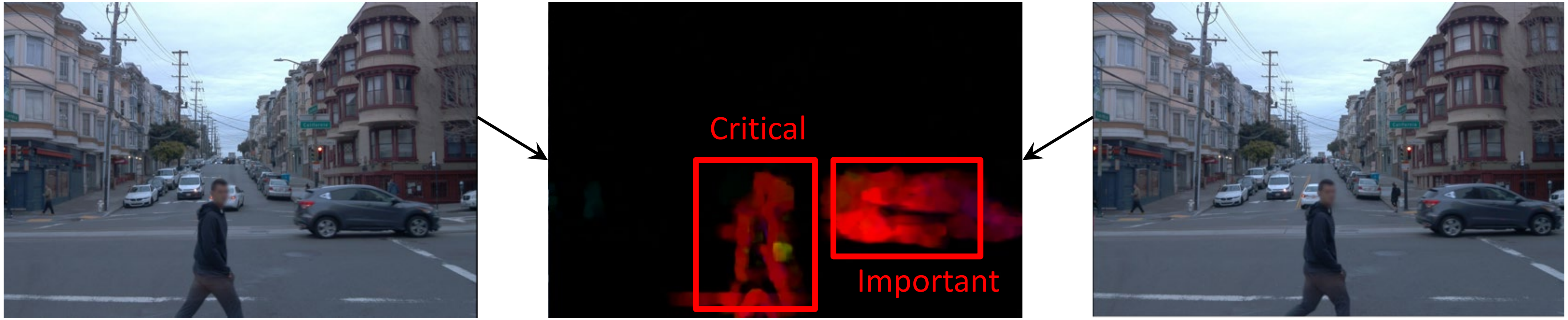
Optical Flow Map

New Frame

[6] Shengzhong Liu, Xinzhe Fu, Maggie Wigness, Philip David, Shuochao Yao, Lui Sha, Tarek Abdelzaher, “Self-Cueing Real-Time Attention Scheduling in Criticality-Driven Visual Machine Perception,” In Proc. *28th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, Milano, Italy, May 2022. **Best Paper Award**

Attention (Self-)Cueing

- Optical flow: Pixel-level motion vectors between two frames, caused by the relative movement between objects and the observer.
- Cue attention to regions of larger change.



Previous Frame | Optical Flow Map | New Frame


verizon

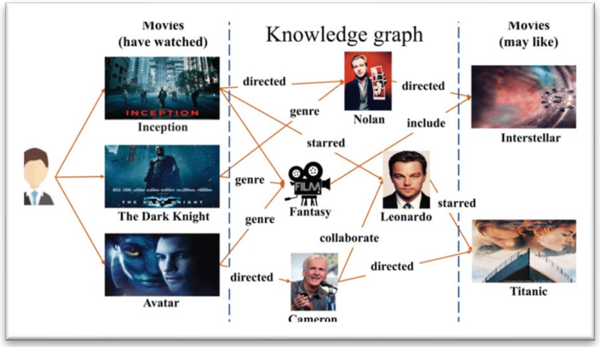
Attention Management Extends Beyond the Embedded Device!

Attention is a key concept in AI and a key bottleneck

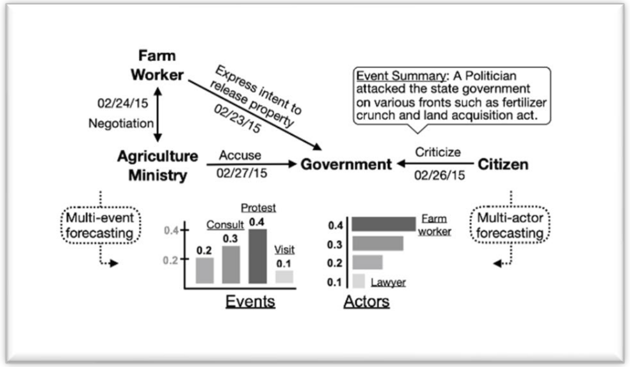
There is significant room for innovation in prioritizing attention to different data regions to meet deadlines and derive the corresponding (machine) *cognitive capacity*[1] constraints.

[1]As humans, we often do not fall "behind" real-time (in perception/reasoning) but rather limit our attention and ignore progressively more extraneous stimuli.

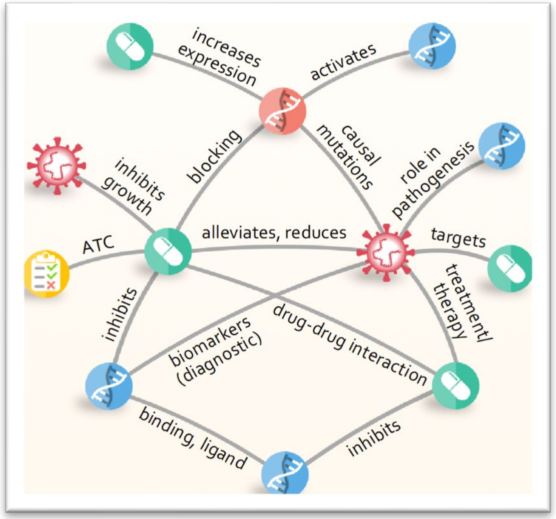
Example: Maintaining Temporal Knowledge Graphs



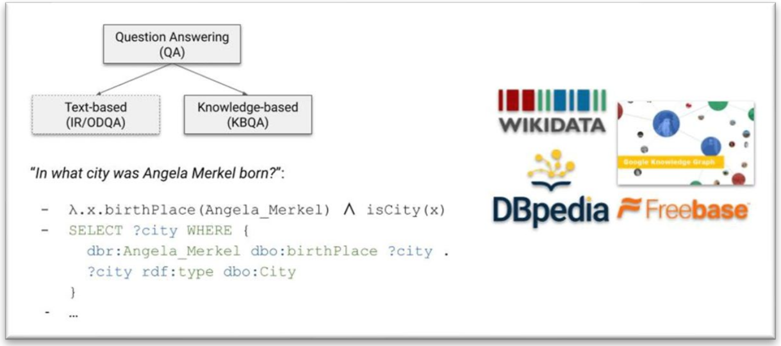
Recommendation[1]



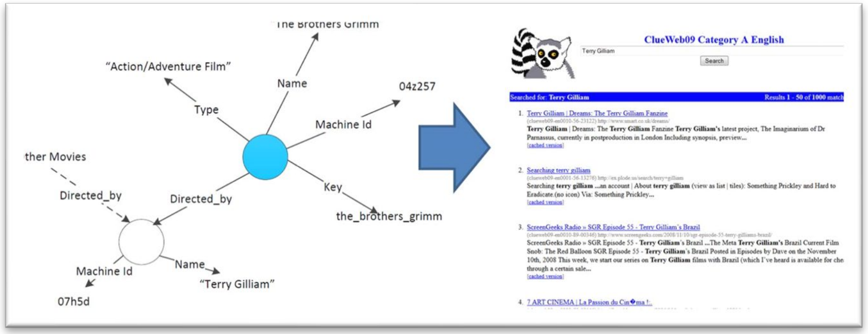
Social Event Forecasting[2]



Drug Analytics[3]



Question Answering[4]

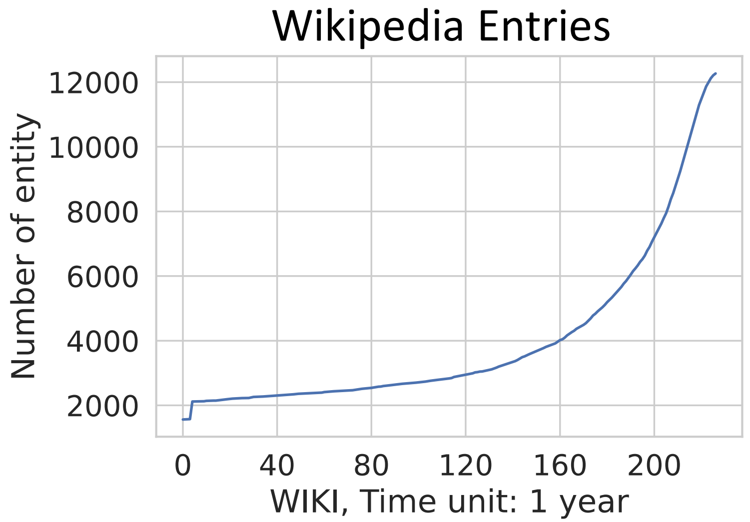


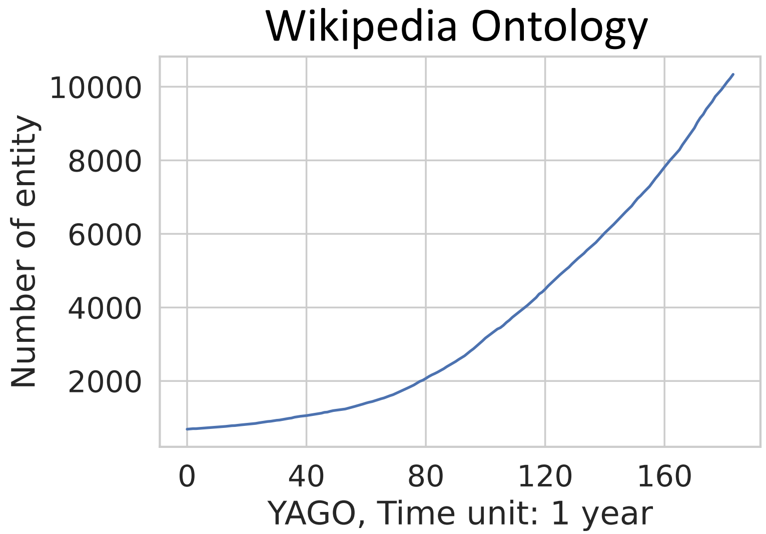
Information Retrieval[5]

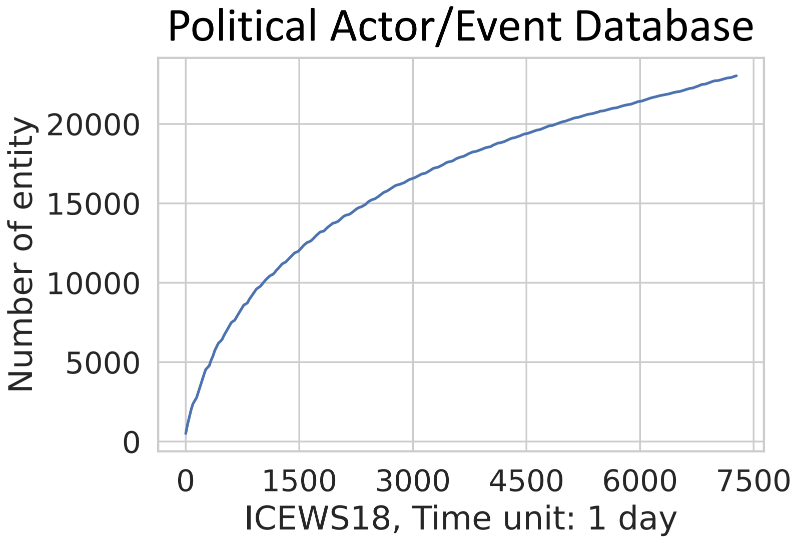
[1] Personalized recommendation system based on knowledge embedding and historical behavior; [2] Dynamic Knowledge Graph based Multi-Event Forecasting, in KDD 2020; [3] Xiangxiang Zeng et al. Repurpose open data to discover therapeutics for COVID-19 using deep learning. Journal of proteome research 2020; [4] https://towardsdatascience.com/the-new-benchmark-for-question-answering-over-knowledge-graphs-qald-9-plus-da37b227c995; [5] http://www.cs.cmu.edu/~callan/Projects/IIS-1422676/

New Entities Continuously Join Temporal Knowledge Graphs

- New entities continuously join graphs:





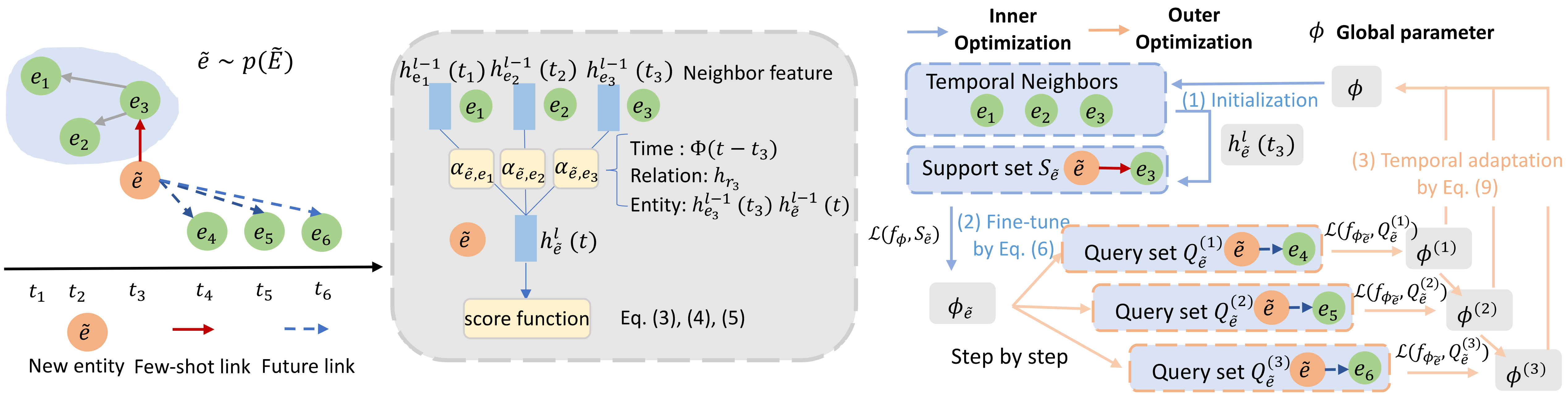


Attention Prioritization: What Data Are More Important (in Temporal Graph Learning)?

How best to Compute Embeddings of New Nodes and Update Old Nodes Given New Observations?

The solution learns the functions that compute/update the embeddings of new nodes given their (most important) interactions/relations with other nodes (neighbors).

The attention management contribution lies in a novel attention framework that *prioritizes the neighbors* to infer new embeddings from; updates are based on important neighbors only.



(a) Temporal KG with new entity ***(b) Temporal Encoder f_ϕ*** ***(c) Meta Temporal Reasoning***

Overall Performance

- **Observation:**
  - The approach improves accuracy given the same latency

Models	YAGO				WIKI				ICEWS18			
	1-shot		2-shot		1-shot		2-shot		1-shot		2-shot	
	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
TransE	0.183	0.268	0.193	0.304	0.144	0.186	0.146	0.213	0.049	0.077	0.058	0.086
TransR	0.189	0.270	0.198	0.312	0.160	0.183	0.160	0.225	0.050	0.080	0.060	0.090
RotatE	0.215	0.280	0.210	0.359	0.175	0.190	0.201	0.268	0.068	0.098	0.070	0.091
RE-NET	0.221	0.304	0.233	0.390	0.212	0.259	0.239	0.294	0.185	0.250	0.200	0.341
LAN	0.196	0.269	0.200	0.310	0.174	0.275	0.162	0.273	0.170	0.301	0.188	0.317
I-GEN	0.238	0.321	0.237	0.402	0.181	0.241	0.223	0.287	0.199	0.320	0.177	0.337
T-GEN	0.247	0.331	0.260	0.379	0.202	0.245	0.240	0.319	0.131	0.262	0.161	0.259
MetaDyGNN	0.269	0.396	0.316	0.496	0.241	0.371	0.271	0.390	0.249	0.420	0.269	0.441
MetaTKGR	**0.294***	**0.428***	**0.356***	**0.526***	**0.277***	**0.419***	**0.309***	**0.441***	**0.295***	**0.496***	**0.301***	**0.500***
Gains (%)	*9.43*	*8.04*	*12.69*	*6.14*	*14.64*	*12.93*	*14.04*	*13.20*	*18.45*	*17.87*	*11.47*	*13.39*

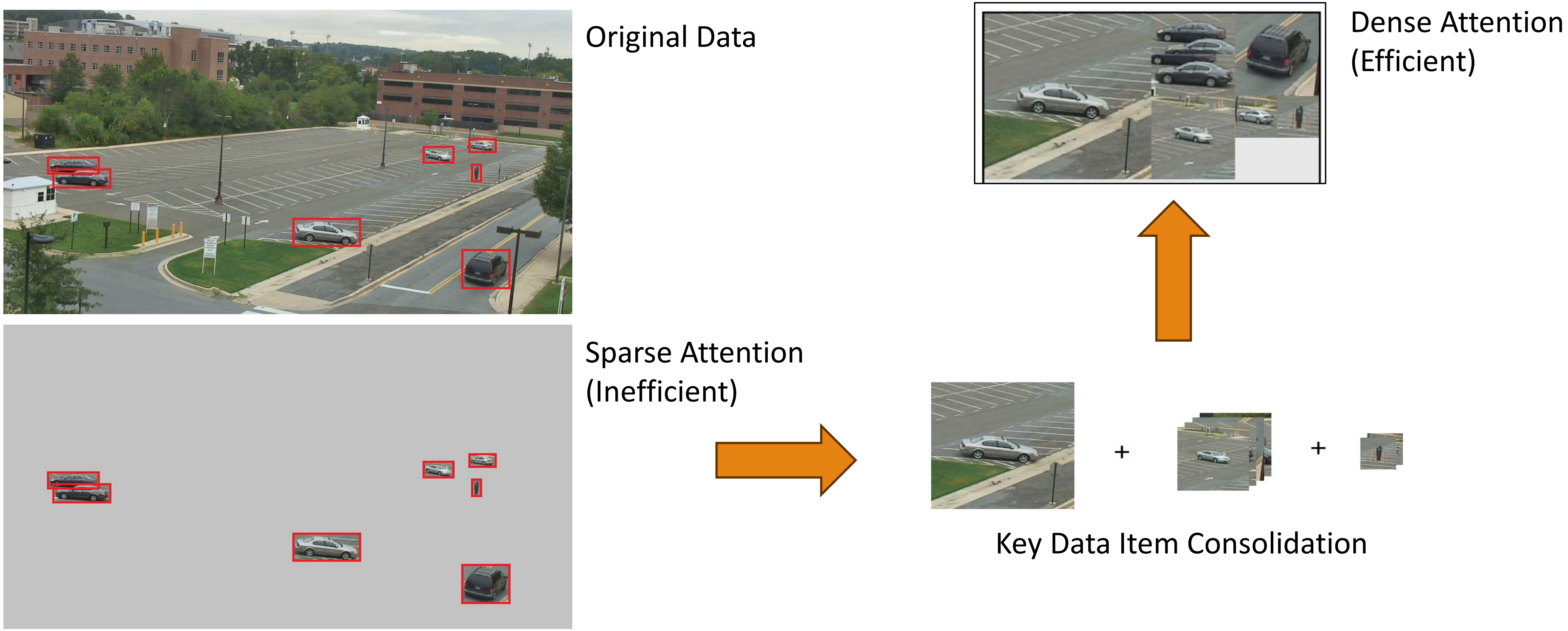
Challenge:

Attention Management (Scheduling)

Consolidate attention foci for efficient processing within time constraints

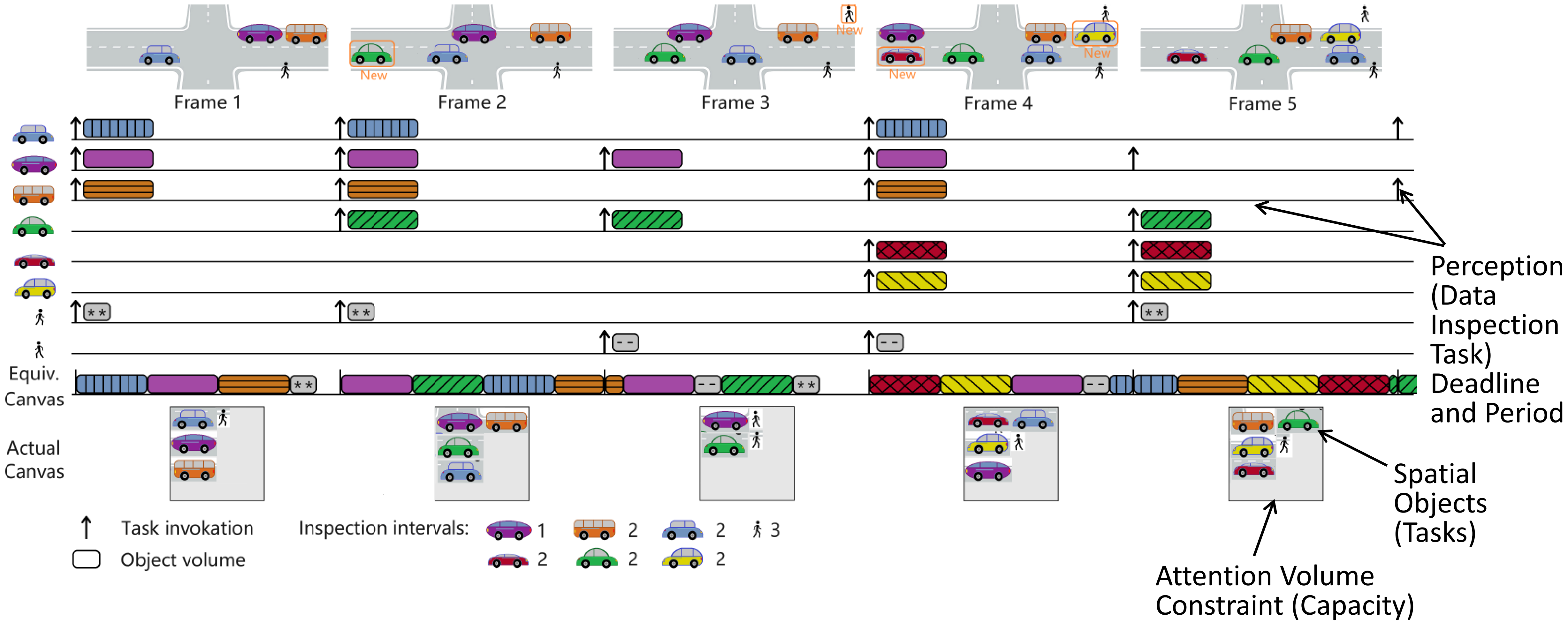
[8] Yigong Hu, Ila Gokarn, Shengzhong Liu, Archan Misra, Tarek Abdelzaher, "Algorithms for Canvas-based Attention Scheduling with Resizing," In Proc. *IEEE RTAS*, May 2024.

Consolidate the Most Pertinent Data for Efficient Downstream Processing



[8] Yigong Hu, Ila Gokarn, Shengzhong Liu, Archan Misra, Tarek Abdelzaher, “Algorithms for Canvas-based Attention Scheduling with Resizing,” In Proc. *IEEE RTAS*, May 2024.

A Spatial-temporal Scheduling Problem and Spatial-temporal “Perception Schedulability” Bound



[8] Yigong Hu, Ila Gokarn, Shengzhong Liu, Archan Misra, Tarek Abdelzaher, "Algorithms for Canvas-based Attention Scheduling with Resizing," In Proc. *IEEE RTAS*, May 2024.

A Spatial-temporal Scheduling Problem and Spatial-temporal "Perception Schedulability" Bound

Under EDF, a GPU that can process a volume of input data, V_{GPU}, per frame, will always meet all inspection deadlines if the sum of object volumes (each normalized by its relative inspection deadline, counted in the number of frame periods) does not exceed:

$$\frac{1}{2}V_{GPU} - v_{max},$$

where v_{max} is the largest object size.

$$\boxed{\sum_{o_i \in \mathcal{O}(k)} \frac{v_i}{D_i^{k_i}} \leq \frac{1}{2}V_{GPU} - v_{max}}$$

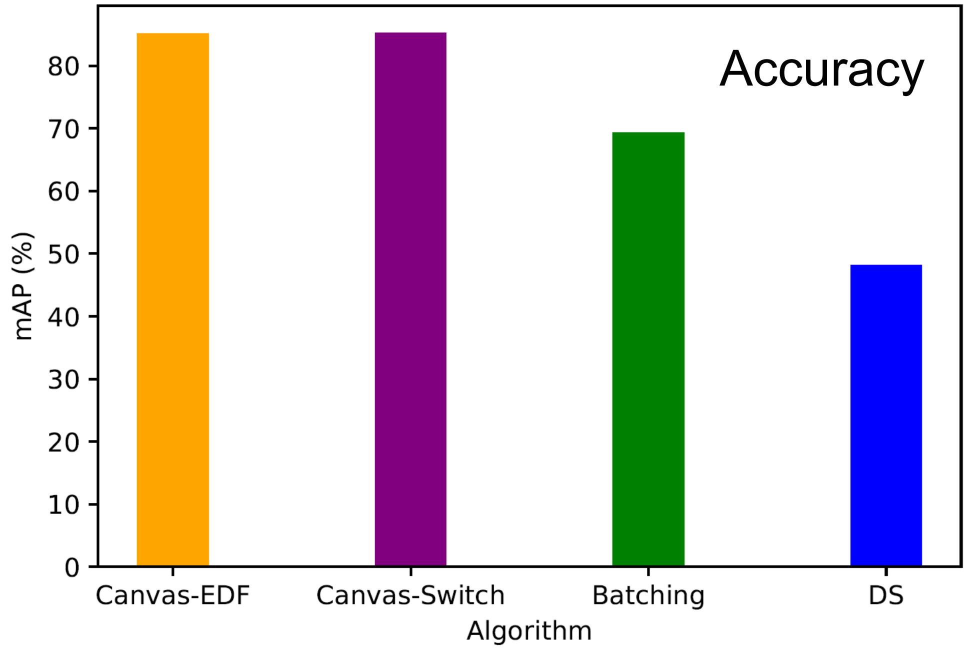
Evaluation Results

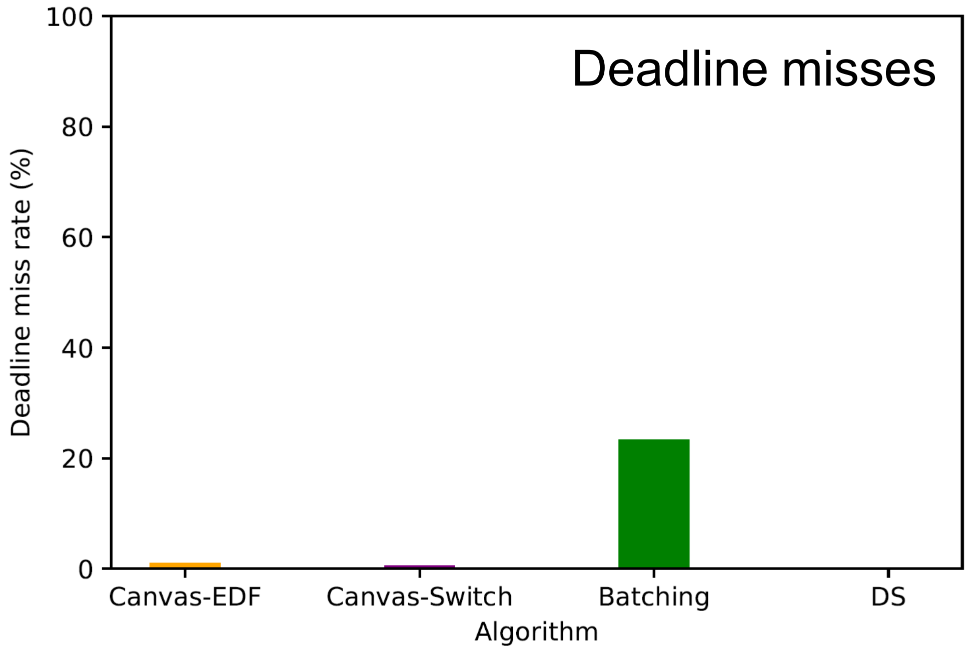
Canvas-EDF: canvas-based attention scheduling with EDF policy

Canvas-switch: canvas-based attention scheduling with task switching policy

Batching: attention scheduling with batching-based neural network execution

DS: downsize the entire frame to fit the frame rate





Challenge:

Latency/Quality Trade-offs in Data Communication (for Downstream AI)

Learn compressed data representations that improve latency/quality trade-offs

[9] Shengzhong Liu, Tianshi Wang, Jinyang Li, Dachun Sun, Mani Srivastava, and Tarek Abdelzaher, "AdaMask: Enabling Machine-Centric Video Streaming with Adaptive Frame Masking for DNN Inference Offloading," In Proc. *ACM Multimedia*, Lisbon, Portugal, October 2022.

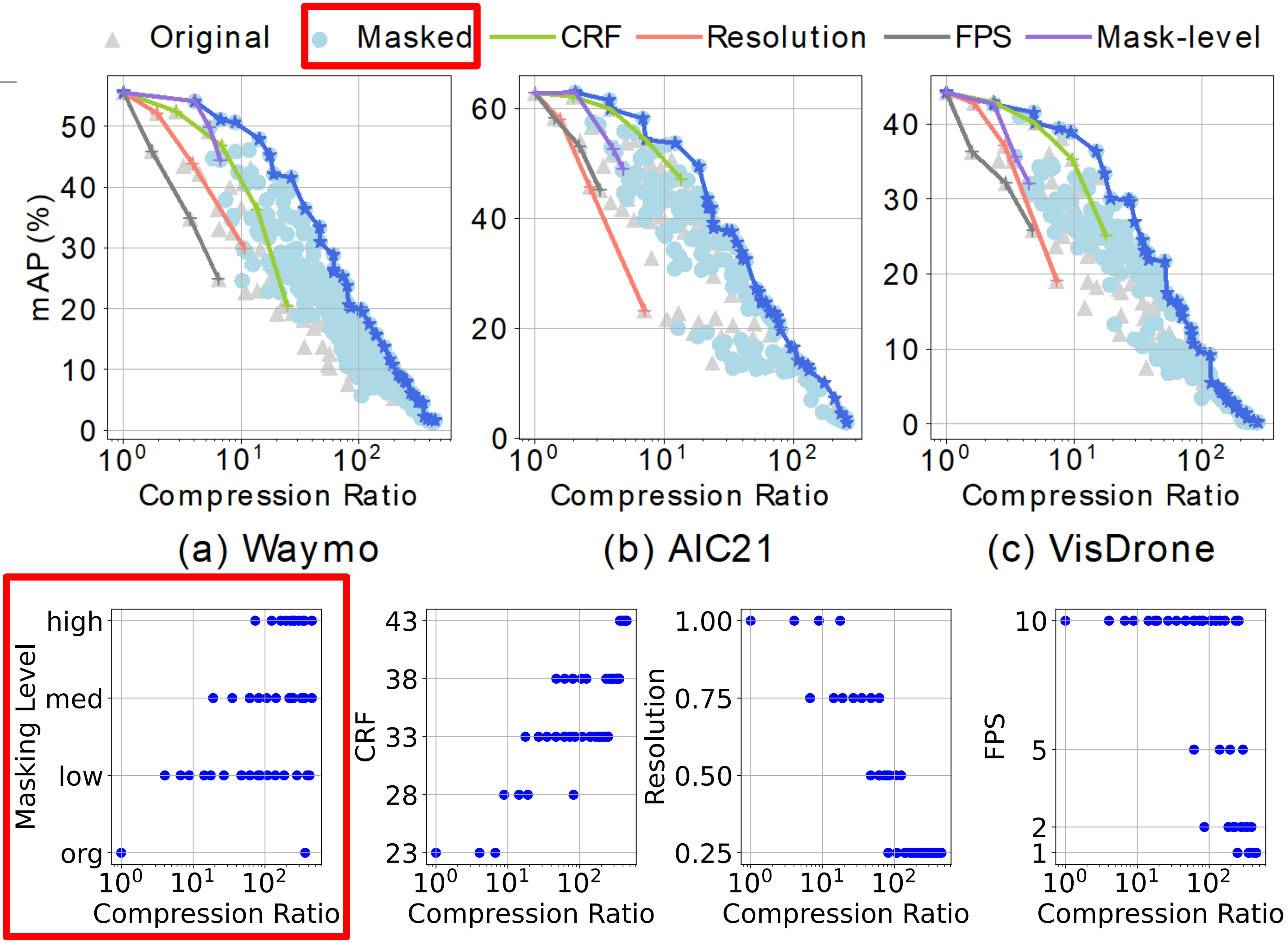
Pareto Boundaries and MPEG Encoding

Figures:

- The upper figure shows the accuracy-bandwidth tradeoff for different configurations. The Pareto boundary, along with the impact of individual knobs are highlighted with curves.
- The lower figure shows the value change of each control knob on the Pareto boundary.

Analysis:

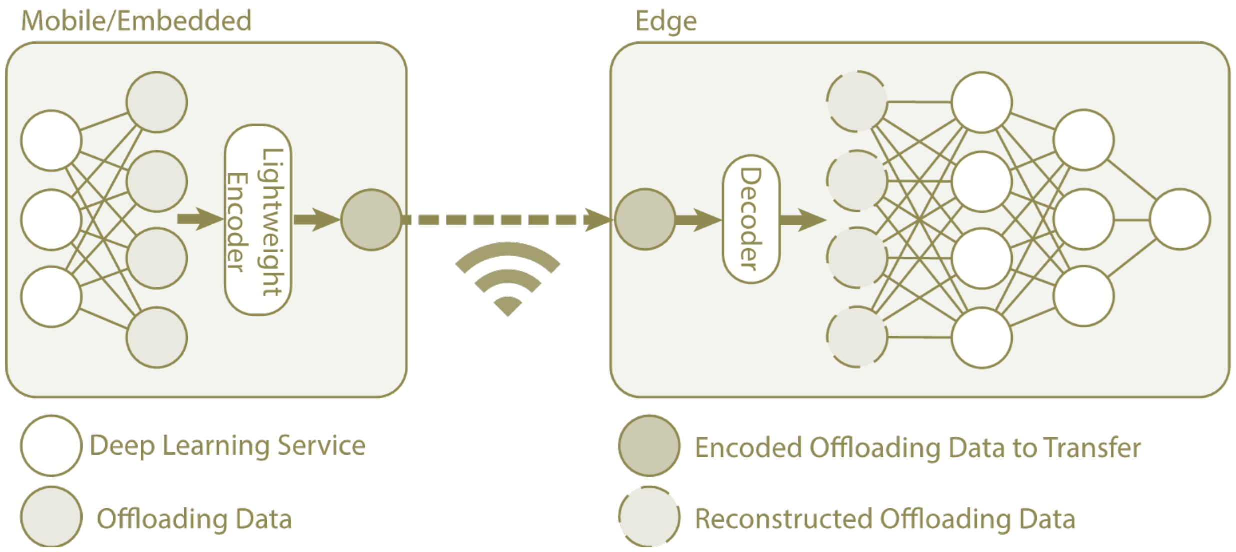
- Most points on the Pareto boundary use masked images (blue points).
- CRF (green curve) and masking level (purple curve) are better dimensions for trading less accuracy for higher compression ratios.

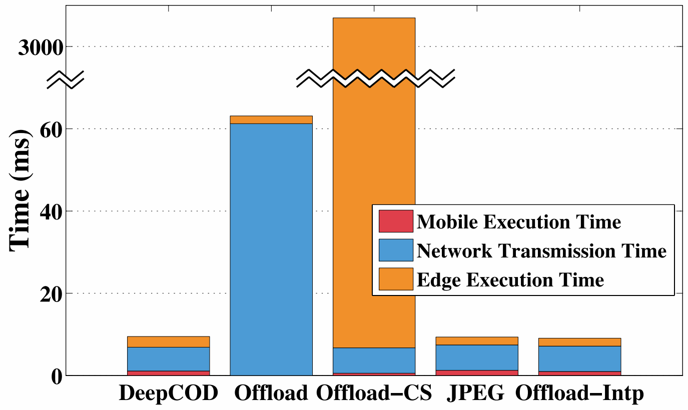


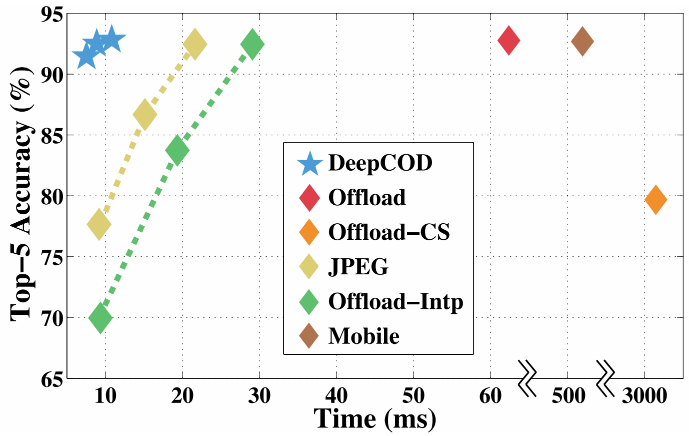
[10] Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Shengzhong Liu, Huajie Shao, Tarek Abdelzaher, "Deep Compressive Offloading: Speeding Up Neural Network Inference by Trading Edge Computation for Network Latency," In Proc. 18th ACM Conference on Embedded Networked Sensor Systems (SenSys), Japan (Online), November 2020.

Compressive Data Offloading

Contribution: Asymmetric auto-encoder (lighter on the client side)
Reduces network latency during offloading, while keeping accuracy







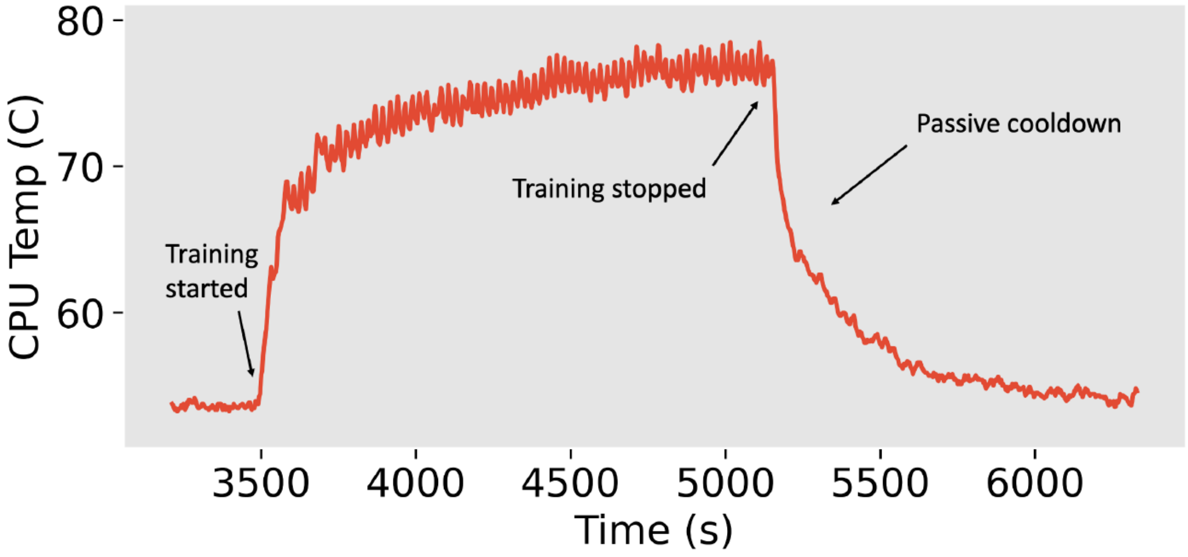
Challenge:

Embedded Real-time AI and Thermal Constraints

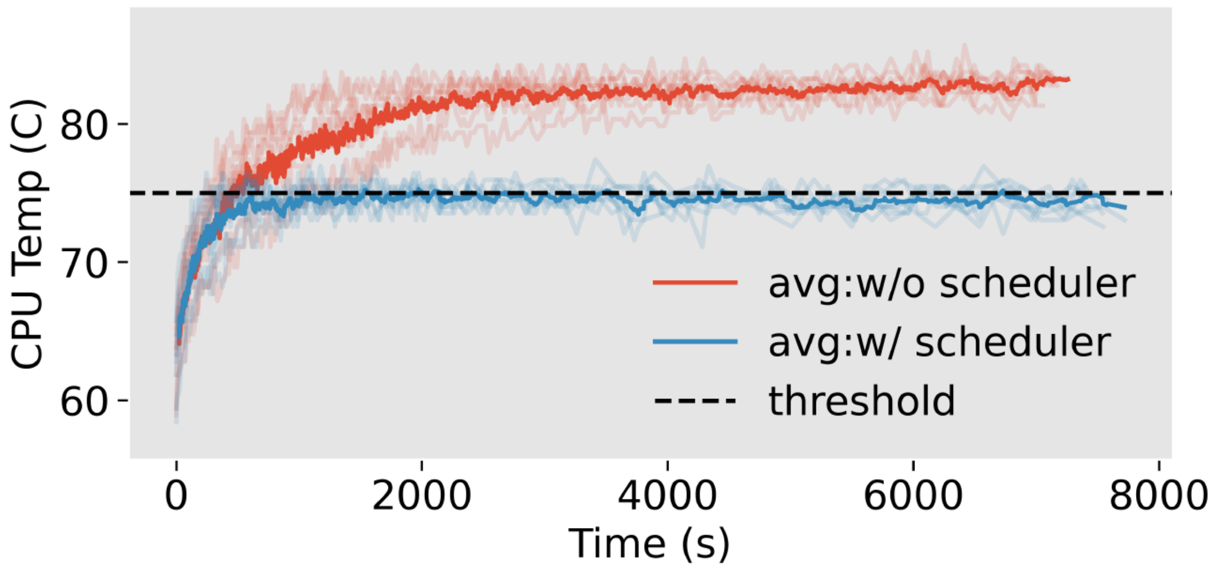
Latency/Quality Trade-offs and Temperature

Thermal Effects of an AI Module (on a Raspberry Pi)

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



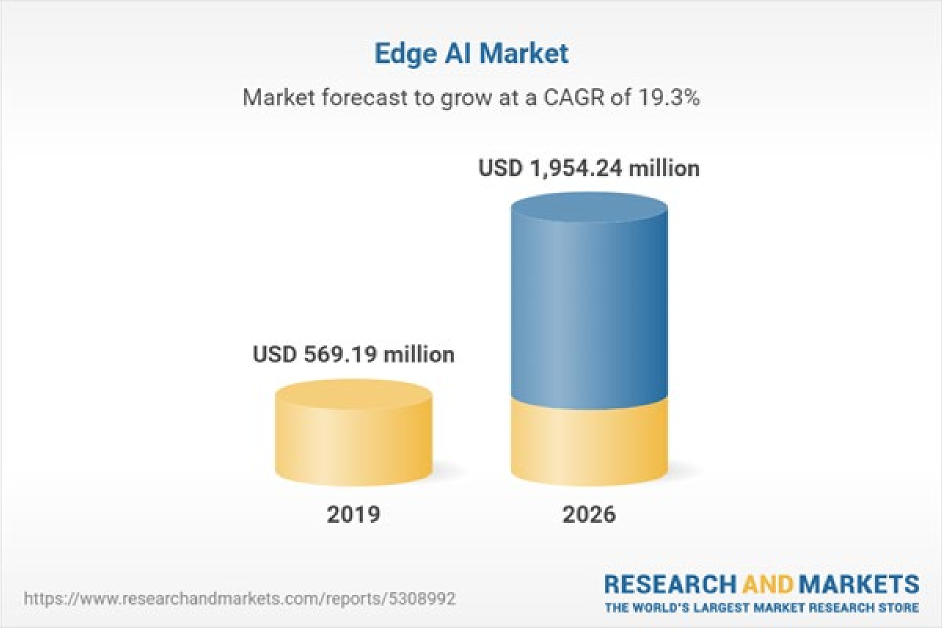
Overheating may trigger an emergency shutdown



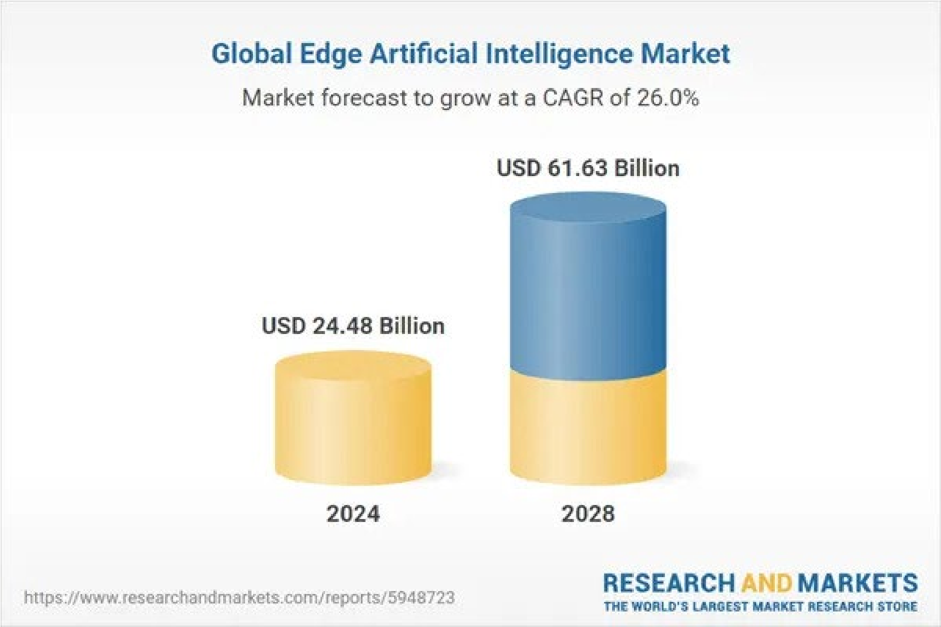
Temperature control prevents shutdown but increases latency, offering a novel trade-off space

Challenge Set #2:
AI + Embedded Computing

Intelligent Embedded Sensing (or “Edge AI”)
Growth Exceeding Expectations



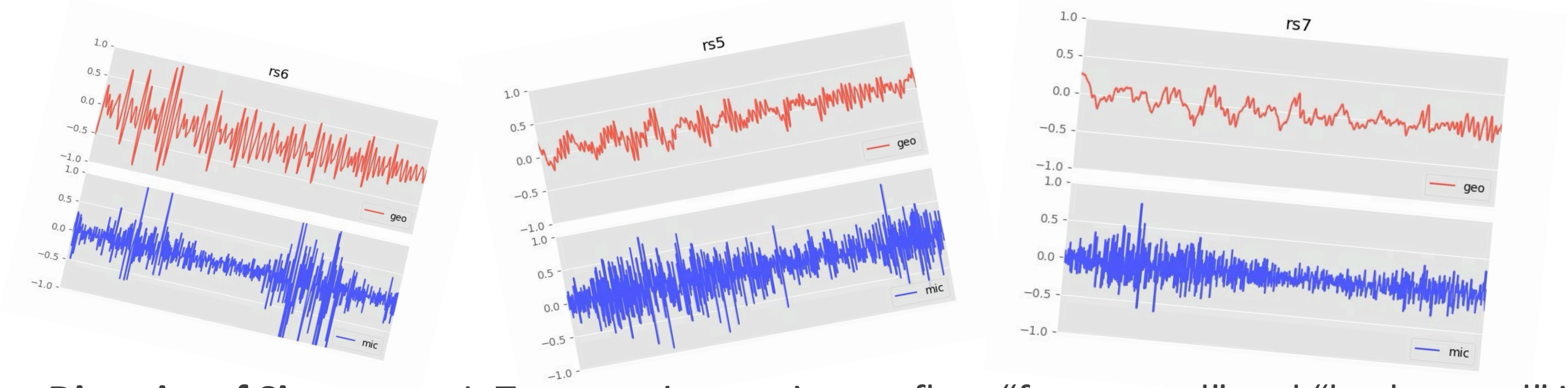
2021 Report: **$1.95B** by 2026



2024 Report: **$61.6B** by 2028

Challenge: Data Labeling for AI Training (to Support Embedded/IoT Applications)

- **Labeled Data Scarcity:** Difficulties finding sufficient labeled training data for IoT sensors
  - Can’t use standard (after-the-fact) labeling approaches due to lack of data interpretability

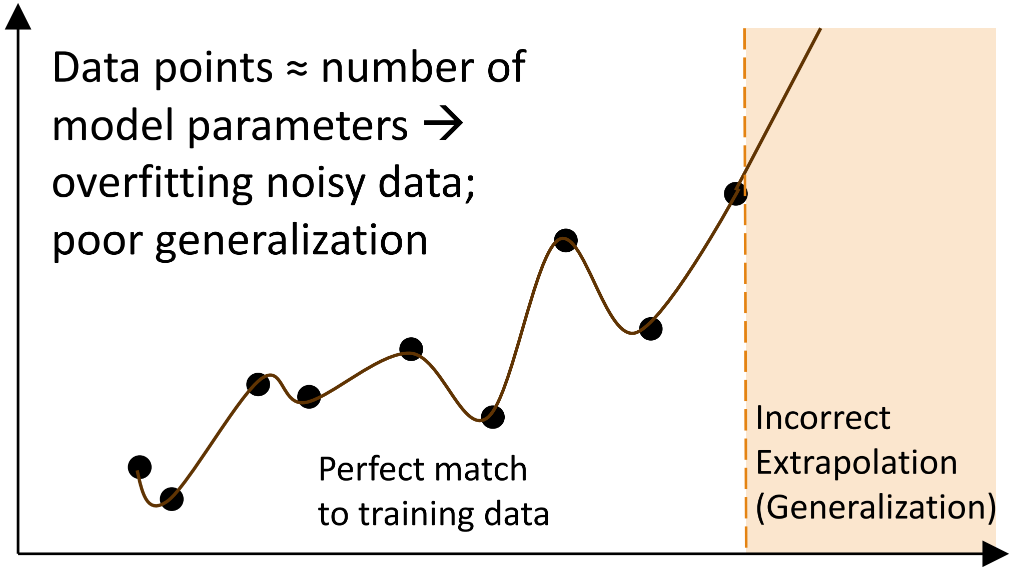


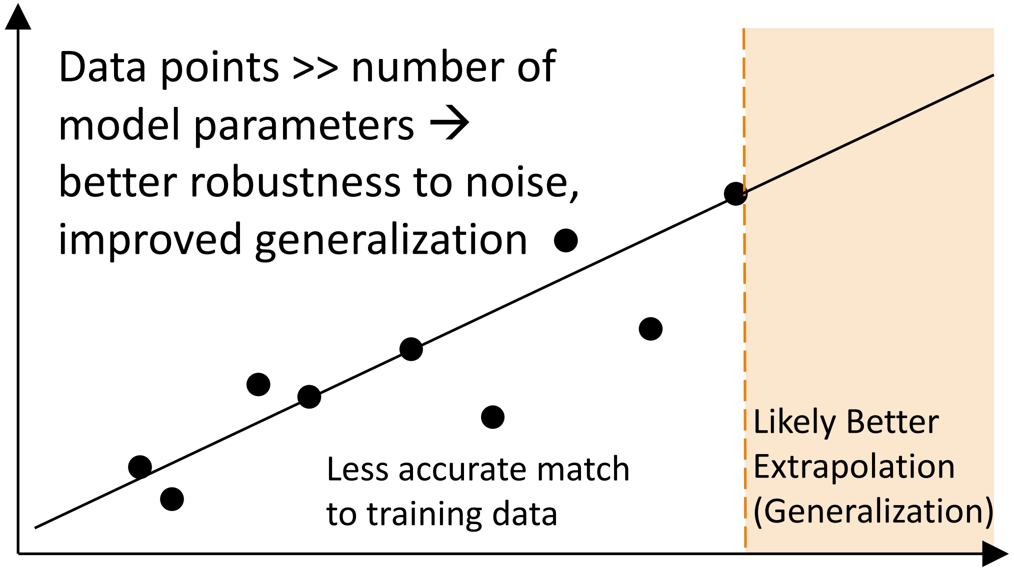
IoT time-series data are hard to interpret after the fact.

- **Diversity of Signatures:** IoT sensor time-series conflate “foreground” and “background” influences leading to an exponential explosion of different sensory signatures for the same phenomenon
  - Example: Acoustic and vibration sensors will be impacted by both the foreground activities and background noise (superimposed together), making it harder to isolate activity signature
  - Example: The sound of a moving car will depend not only on the car but also on the type of road/terrain, creating different signatures in different environments.

Implications of Labeled Data Scarcity and Diversity of Signatures: *Potential Overfitting!*

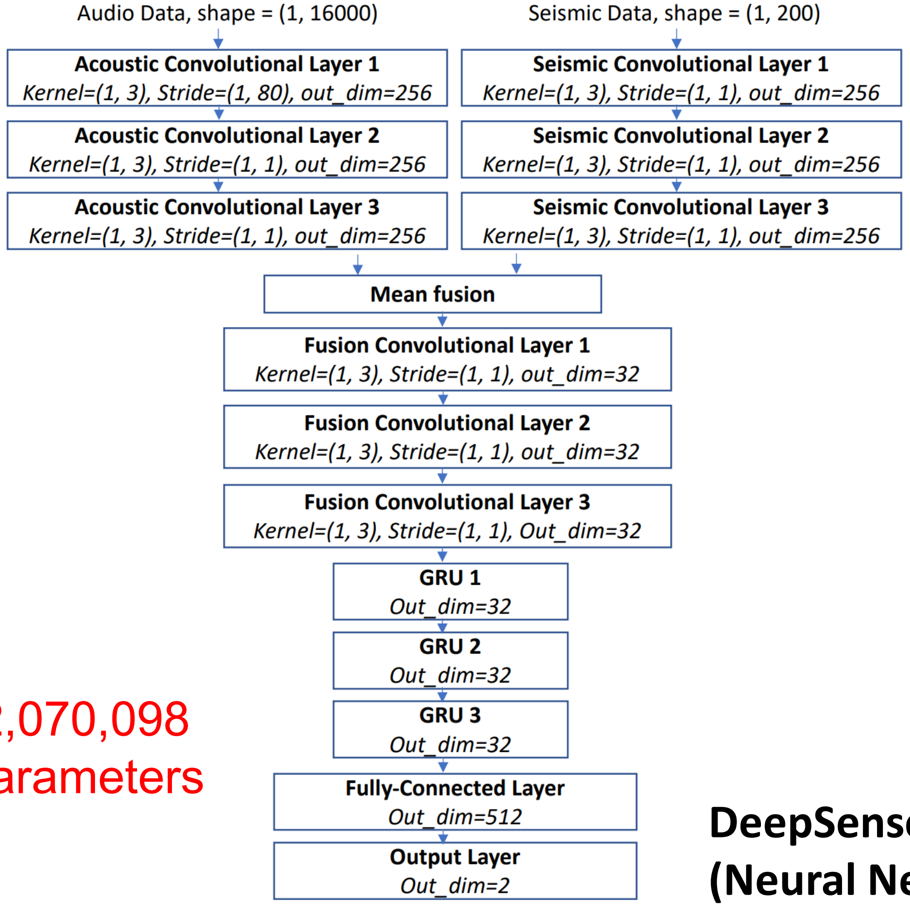
Lack of sufficient labeled training data prevents the use of modern AI models (they have too many parameters to train, thus requiring a lot of labeled samples)





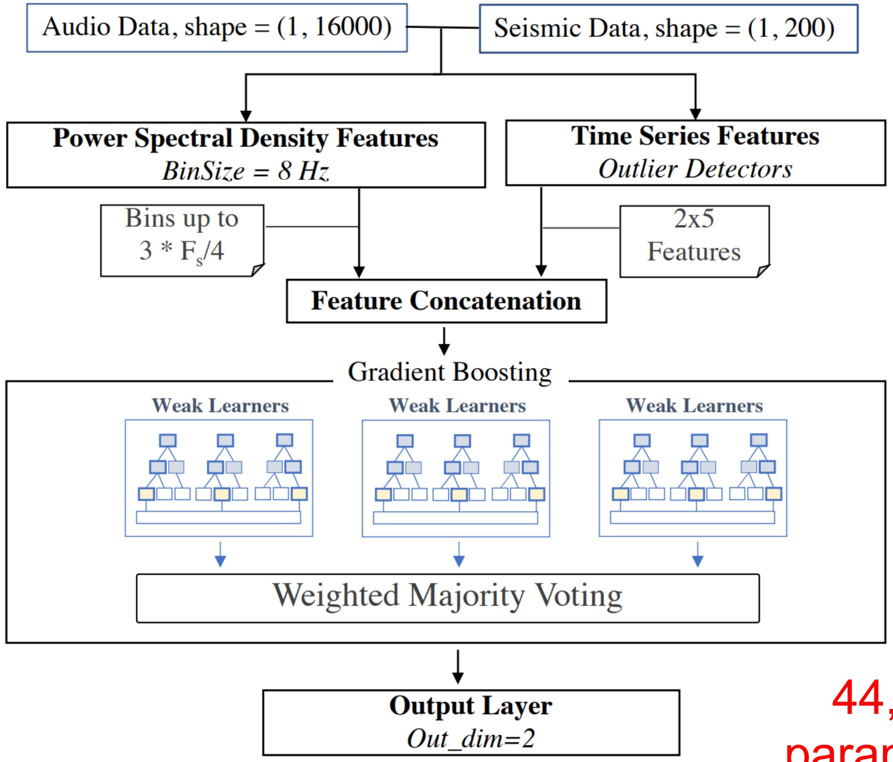
[11] Tianshi Wang, Denizhan Kara, Jinyang Li, Shengzhong Liu, Tarek Abdelzaher, Brian Jalaian, "The Methodological Pitfall of Dataset-Driven Research on Deep Learning: An IoT Example," In Proc. *Military Communications Conference (MILCOM), IoT-AE Workshop*, Rockville, MD, December 2022.

Overfitting Experiment: A Tale of Two Classifiers



2,070,098 parameters

DeepSense (Neural Network)



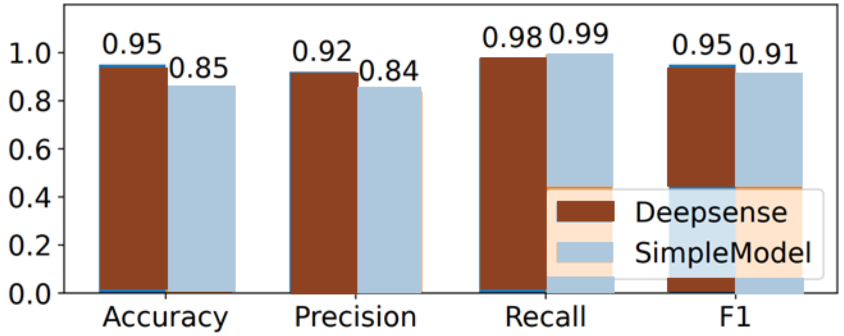
Simple Model: XGBoost (Decision-tree Classifier)

44,680 parameters

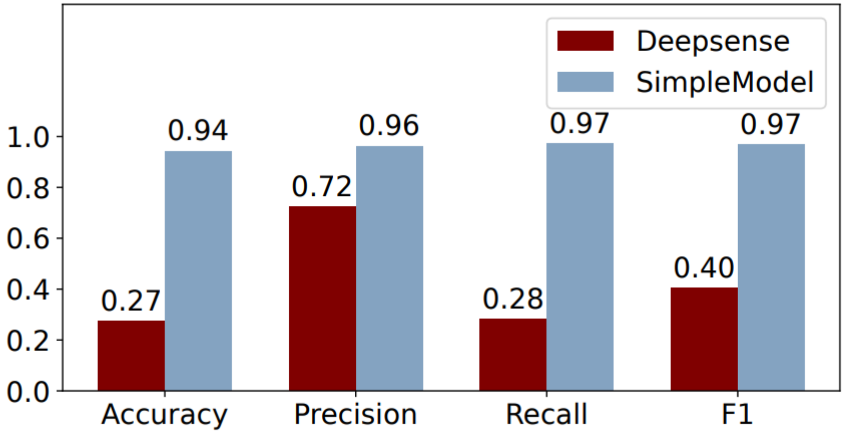
Overfitting Experiment: A Tale of Two Classifiers

- We collected seismic and acoustic data from multiple moving vehicles in multiple environments to train a classifier to determine vehicle type from its acoustic/seismic signature
- Separated the data into training, validation, and testing sets (80%, 10%, 10%).
- Trained the classifier to detect a specific type of vehicle; tuned hyper-parameters with validation set
- Testing results:
    - The larger classifier (DeepSense) is better in the absence of domain shift (on same roads, in the same environmental conditions)
    - Upon a small domain shift (testing in a new location not in training data), the smaller (simple) classifier is significantly better

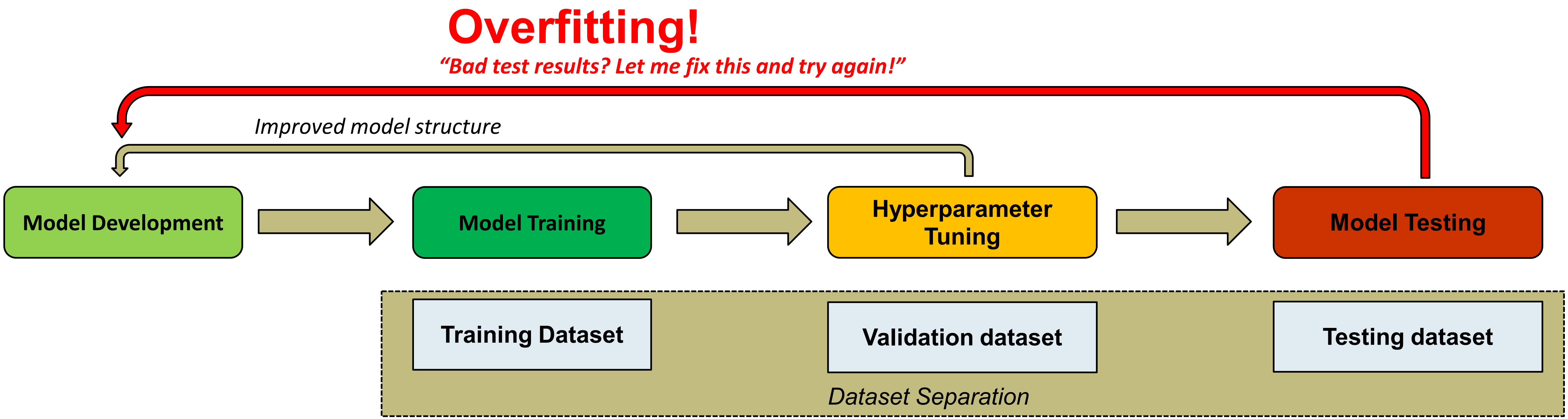
No Domain Shift



Small Domain Shift



Today's Academic Literature Greatly Underestimates the Brittleness of Embedded AI



See:
https://sigbed.org/2022/11/22/the-methodological-pitfall-of-dataset-driven-research-on-deep-learning-in-the-iot-space/

Solution

Can we use *unlabeled data* to train the AI instead of labeled data?

(Hint: the answer is yes)

Intuition: When we see a new type of object for the first time (e.g., a curved screen monitor), we are able to identify this type of objects thereafter without additional “labeled data”. Why?

Solution

Can we use *unlabeled data* to train the AI instead of labeled data?

(Hint: the answer is yes)

Intuition: When we see a new type of object for the first time (e.g., a curved screen monitor), we are able to identify this type of objects thereafter without additional “labeled data”. Why?

Because we learned to pay attention to “discriminative features” that help us distinguish different objects. These features can be learned *without knowing object labels*.

Supervised versus Self-Supervised Learning: A Difference in Objective

Supervised (Task-specific): The objective is to learn to associate data with particular object labels (specific to the classification task).

Self-supervised (Task-independent): The objective is to better represent notions of similarity in input data in order to help distinguish similar versus dissimilar objects (in multiple dimensions of similarity) and/or to predict "missing parts" of objects/contexts.

Foundation models: Self-supervised (task-independent) training at scale to extract representations of data that facilitate many downstream tasks

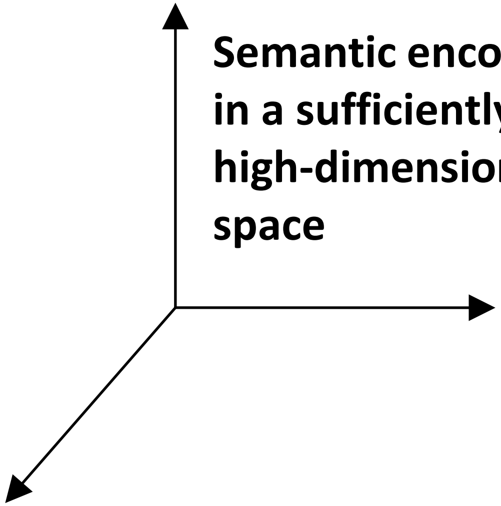
Challenge:

Foundation Models for Embedded Systems

Adapt self-supervised training to embedded application needs

Foundation Model Pre-training Encodes Inputs into a High-Dimensional Semantic Similarity Space; Fine-tuning Maps them to the Task

Pretraining



Semantic encoding in a sufficiently high-dimensional space

Downstream Tasks

Age classifier

Genre classifier

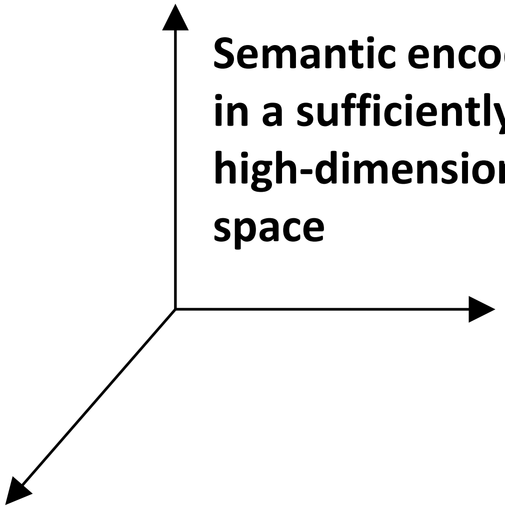
Language classifier

Finetuning

Towards Foundation Models for Embedded Systems: Design the “Right” Self-supervised Pretraining

Pretraining

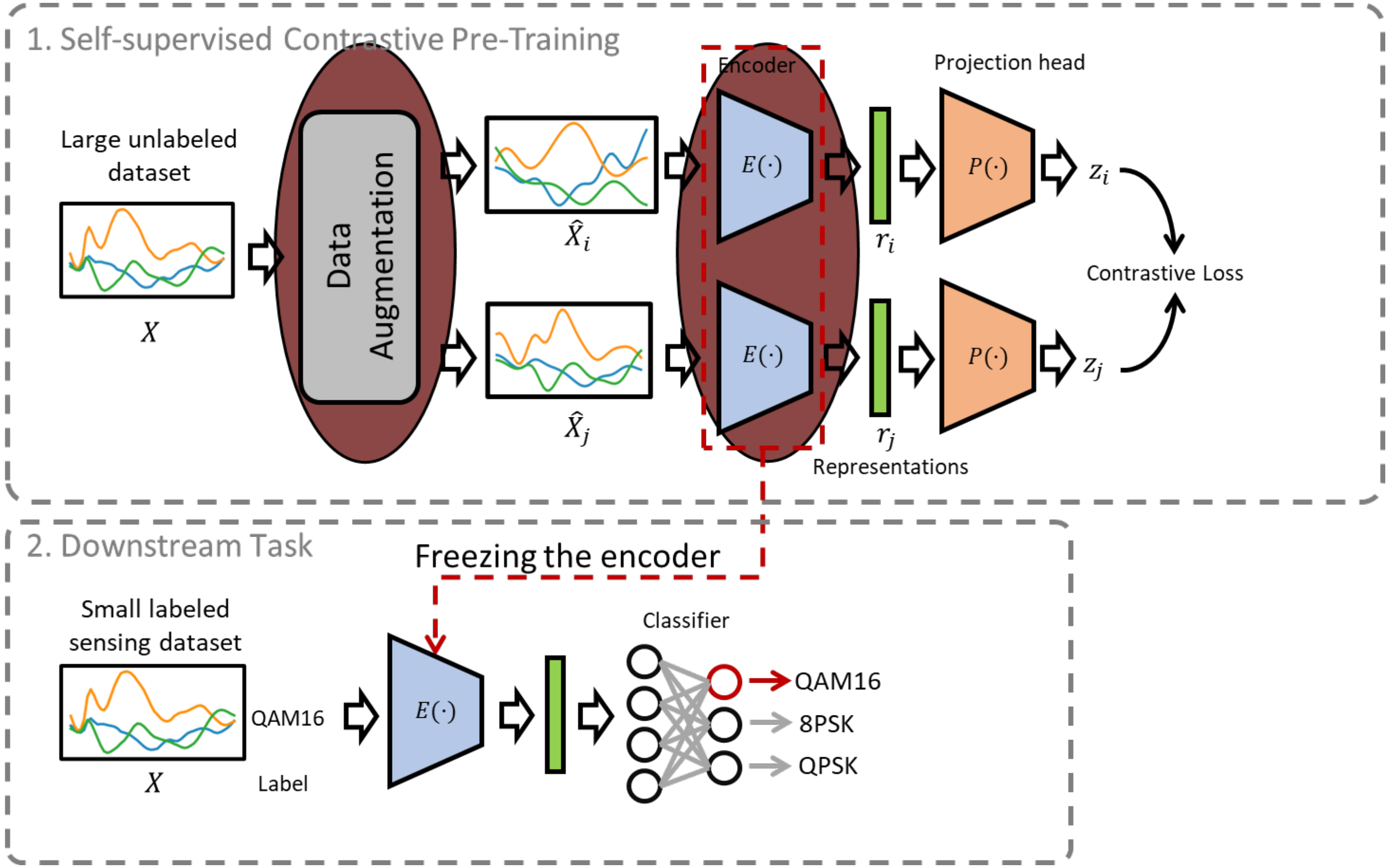
?

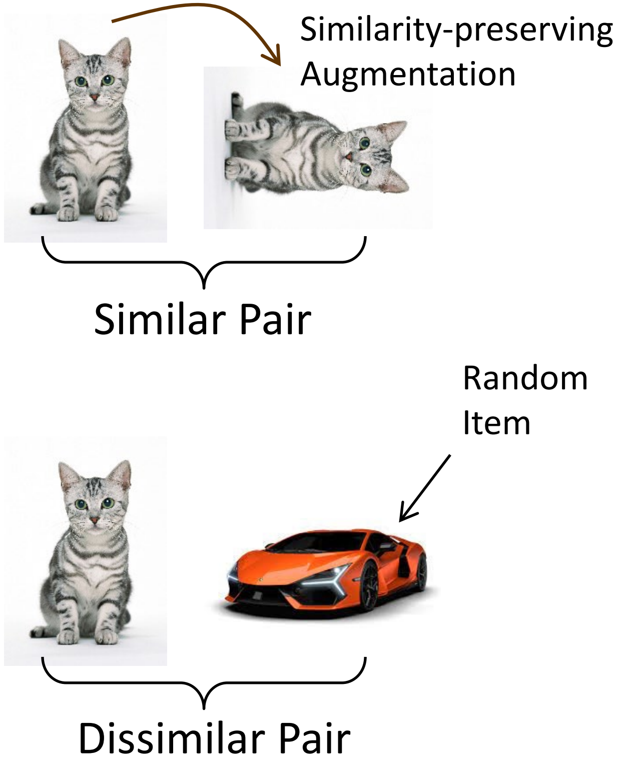


Semantic encoding in a sufficiently high-dimensional space

Common Self-Supervised Pretraining Approaches

1. Contrastive Learning: “Teach” Similarity

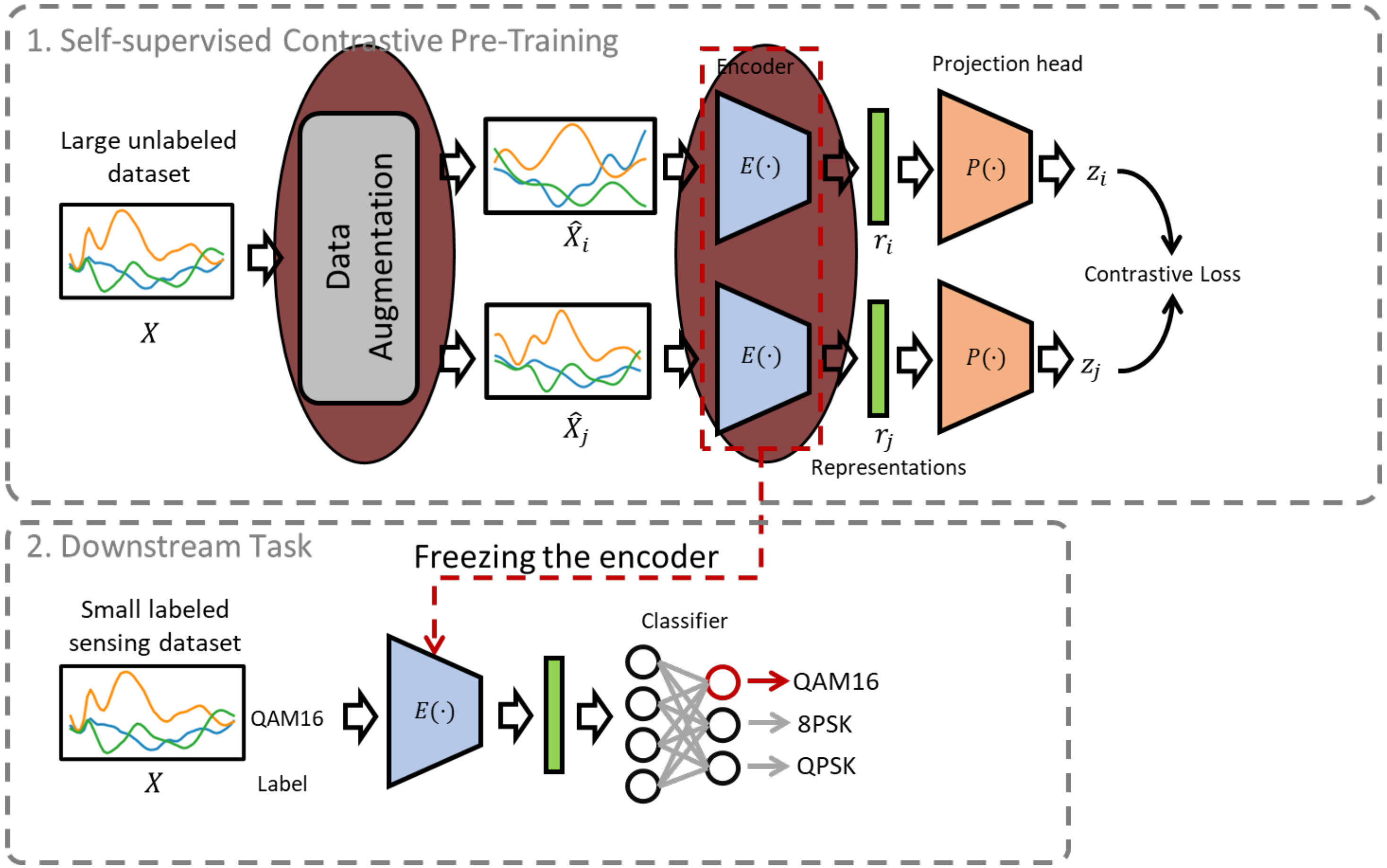


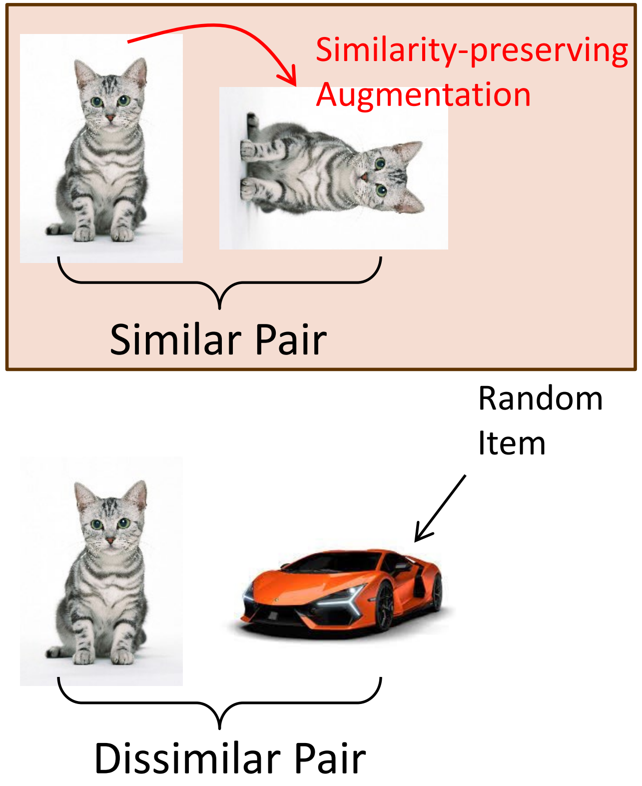


Common Self-Supervised Pretraining Approaches

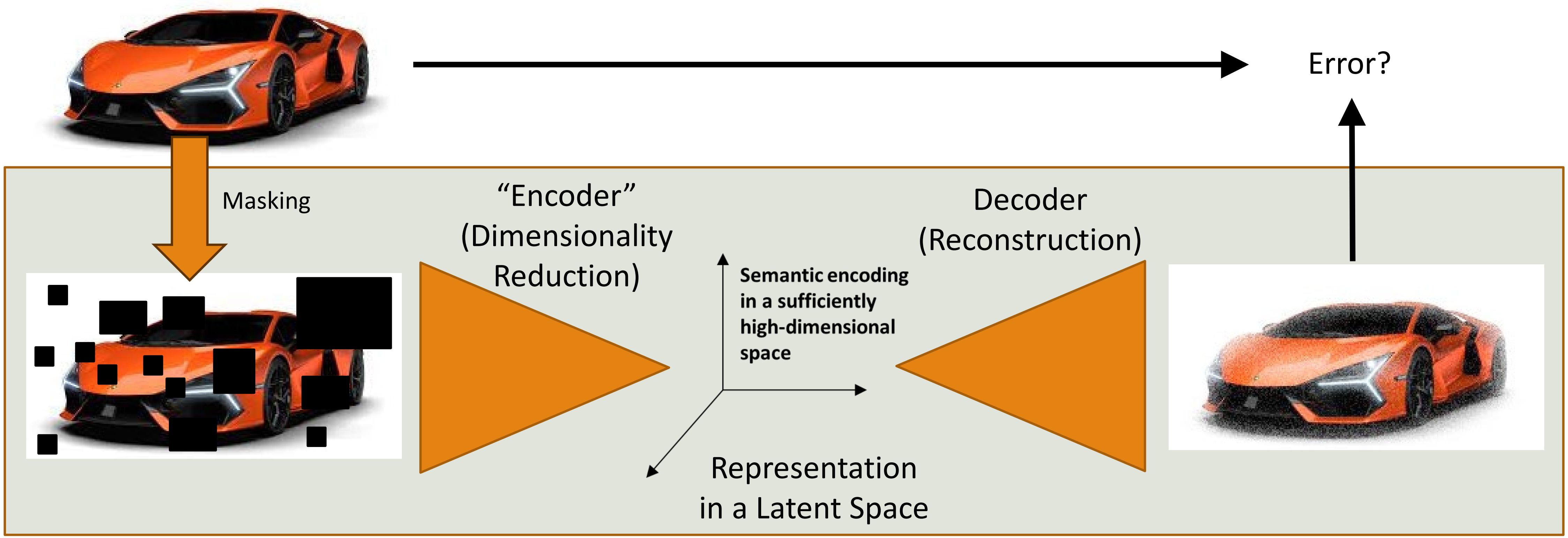
1. Contrastive Learning

Issue: Requires Augmentation Design; Introduces Inductive Bias





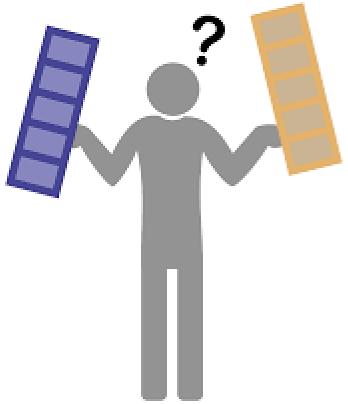
Common Self-Supervised Pretraining Approaches
2. Masked Autoencoders



Observation **Validation**

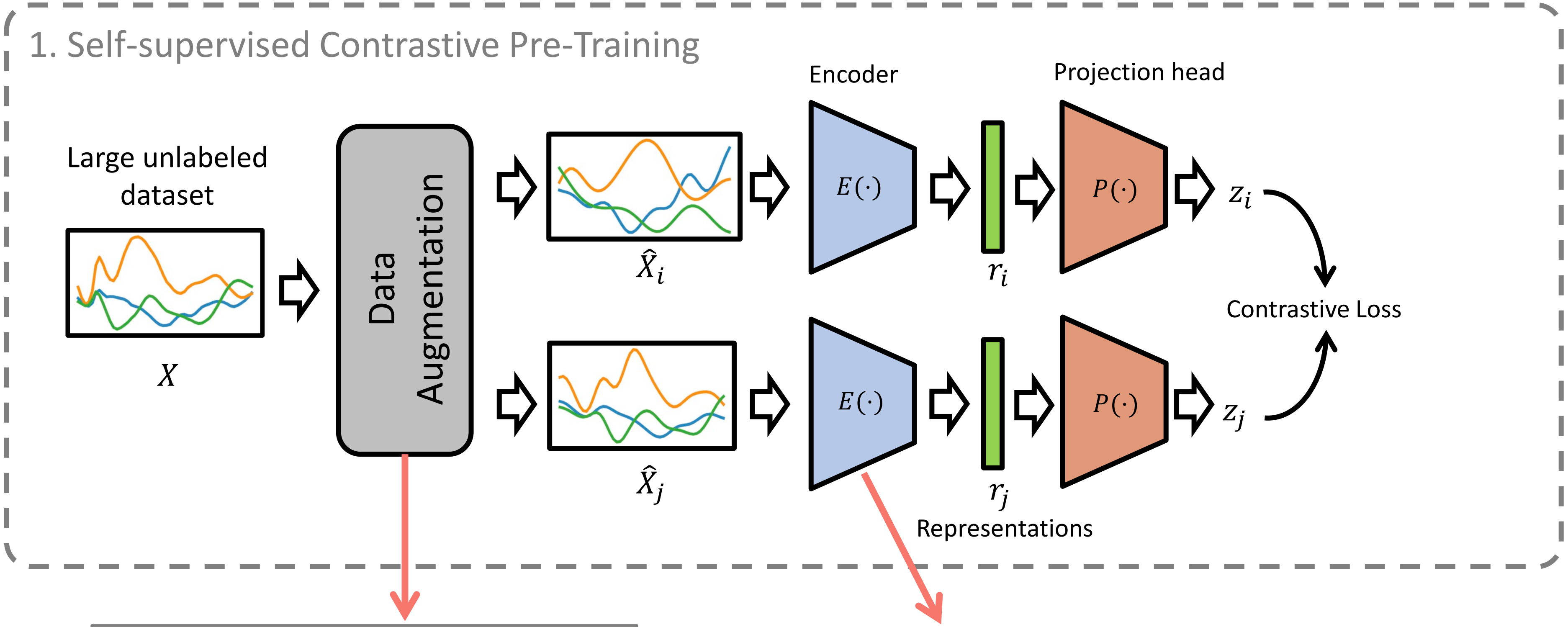
Challenge:

Contrastive Learning from Embedded Systems Data



[12] Dongxin Liu, Tianshi Wang, Shengzhong Liu, Ruijie Wang, Shuochao Yao, and Tarek Abdelzaher, "Contrastive Self-Supervised Representation Learning for Sensing Signals from the Time-Frequency Perspective," In Proc. *30th International Conference on Computer Communications and Networks (ICCCN)*, Athens, Greece, July 2021

Contrastive Learning from Embedded Sensing: Time versus Frequency Domain

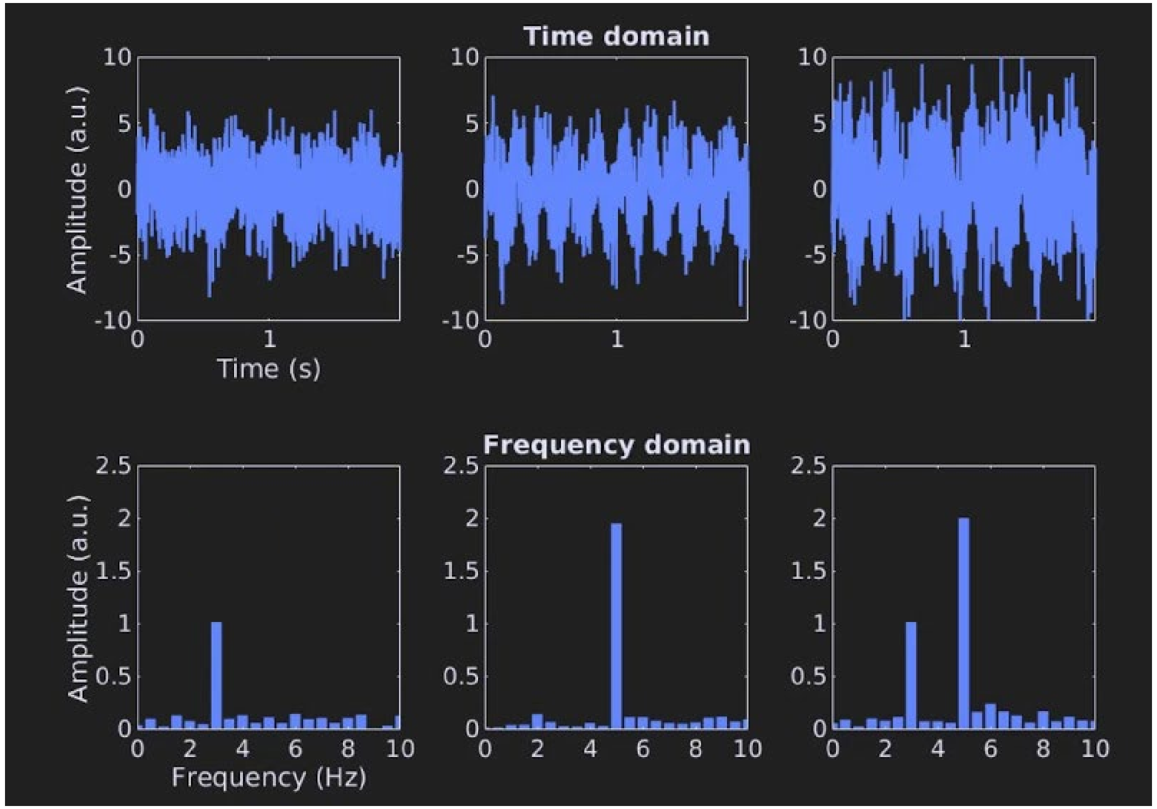


Data augmentation in time-domain or frequency domain?

Encoder for Frequency Domain Data?

Contrastive Learning from Embedded Sensing: Time versus Frequency Domain

- In IoT, sensing data measure physical phenomena
  - acceleration, vibration, or wireless signal propagation
- Underlying processes are fundamentally a function of signal frequencies
- IoT signals have a sparser and more compact representations in the **frequency domain**.

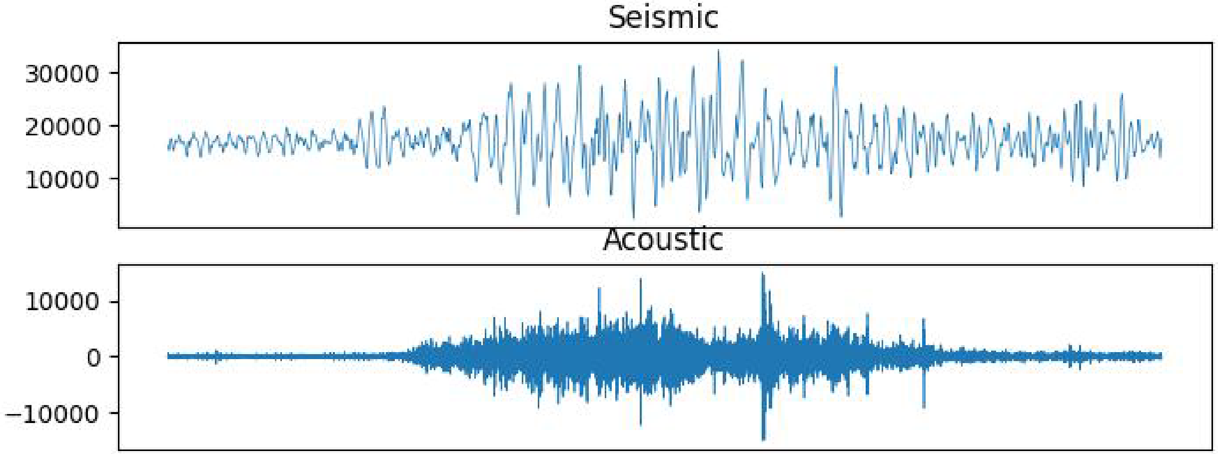


[13] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher, "FOCAL: Contrastive Learning for Multimodal Time-Series Sensing Signals in Factorized Orthogonal Latent Space," In Proc. *37th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, Louisiana, December 2023.

Contrastive Learning from Embedded Sensing: *Multimodal* Data

- Question #1: What is a notion of similarity between two different sensor time-series?

Physical Event/Activity

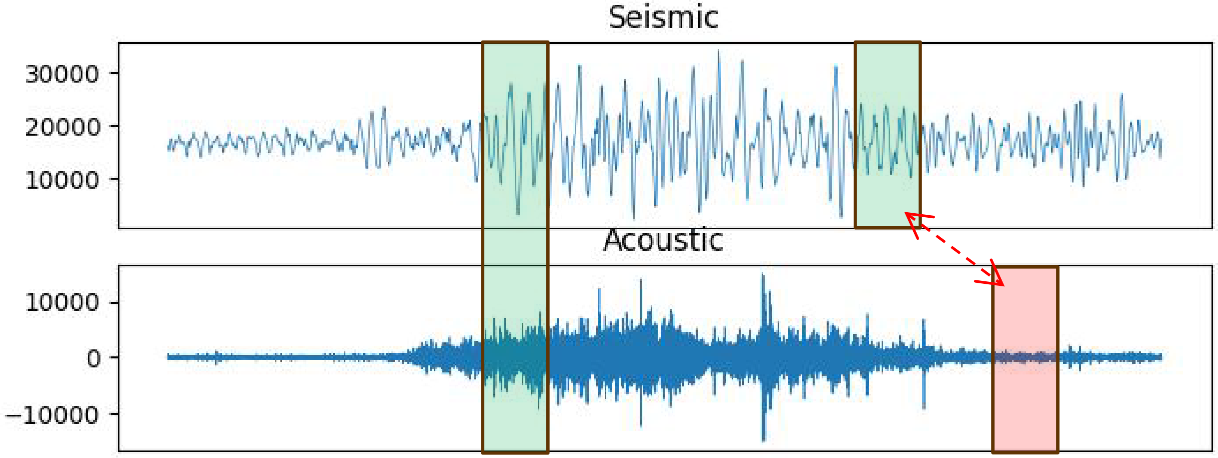


Multi-sensory Signature of Physical Event/Activity

Contrastive Learning from Embedded Sensing: *Multimodal* Data

- Suggestion: Similarity based on signature co-occurrence?

Physical Event/Activity



Multi-sensory Signature of Physical Event/Activity

Same time interval = similar

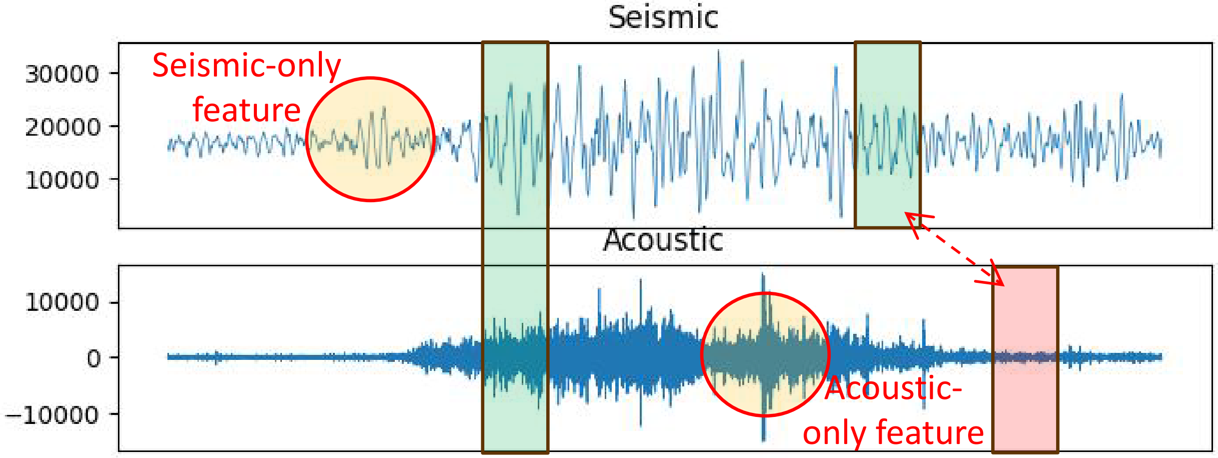
Different intervals = dissimilar

[13] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher, “FOCAL: Contrastive Learning for Multimodal Time-Series Sensing Signals in Factorized Orthogonal Latent Space,” In Proc. *37th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, Louisiana, December 2023.

Contrastive Learning from Embedded Sensing: *Multimodal* Data

- Question #2: How to capture the additional information visible to individual modalities only?

Physical Event/Activity



Multi-sensory Signature of Physical Event/Activity

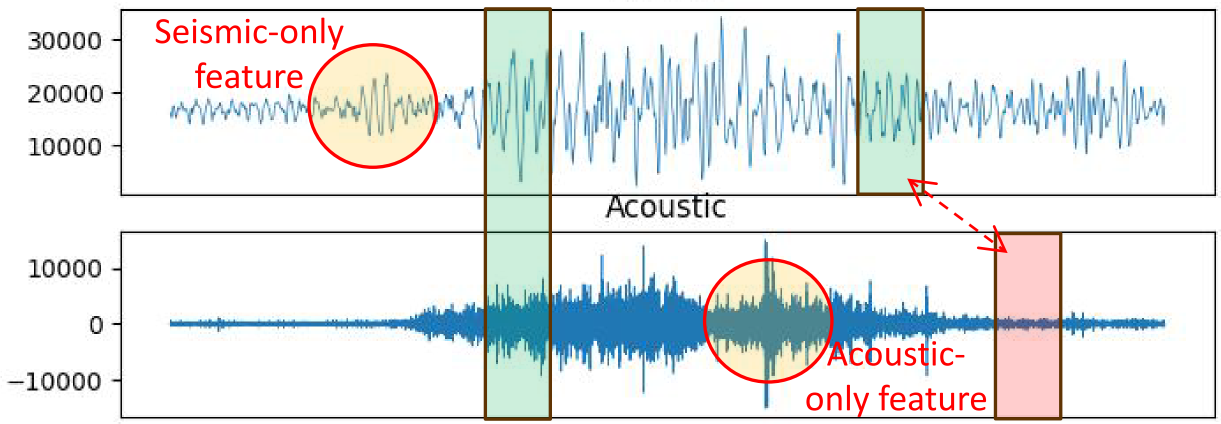
Same time interval = similar

Different intervals = dissimilar

[13] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher, "FOCAL: Contrastive Learning for Multimodal Time-Series Sensing Signals in Factorized Orthogonal Latent Space," In Proc. *37th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, Louisiana, December 2023.

Contrastive Learning from Embedded Sensing: *Multimodal* Data

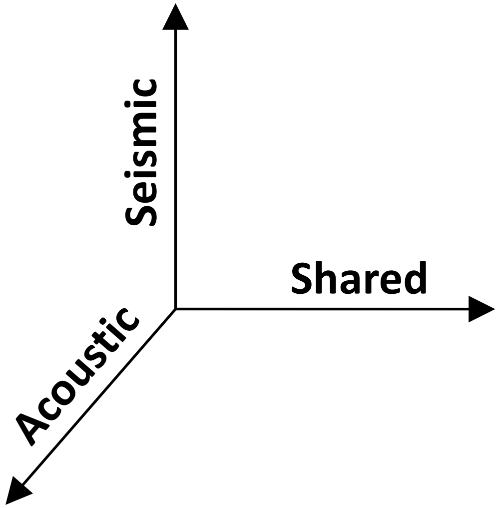
- Suggestion: Shared versus private latent subspaces



Multi-sensory Signature of Physical Event/Activity

Same time interval = similar

Different intervals = dissimilar

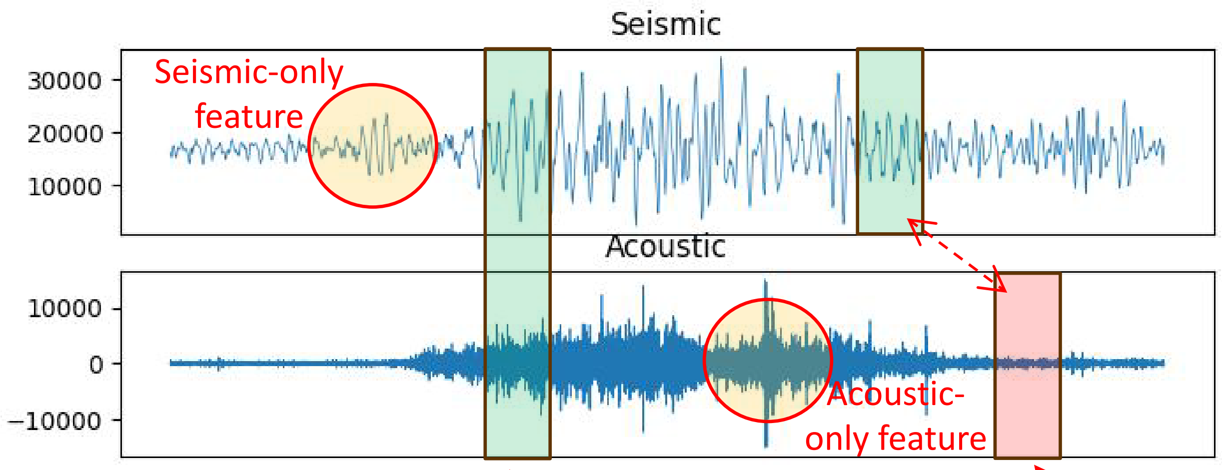


Latent Representation Space

[13] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher, "FOCAL: Contrastive Learning for Multimodal Time-Series Sensing Signals in Factorized Orthogonal Latent Space," In Proc. *37th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, Louisiana, December 2023.

Contrastive Learning from Embedded Sensing: *Multimodal* Data

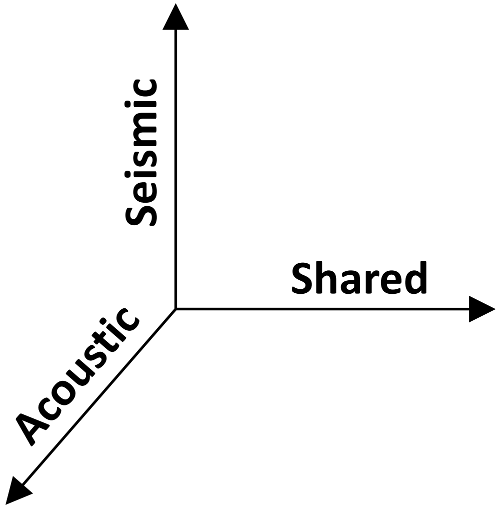
- Question #3: How to ensure a parsimonious (non-redundant) representation?



Multi-sensory Signature of Physical Event/Activity

Same time interval = similar

Different intervals = dissimilar

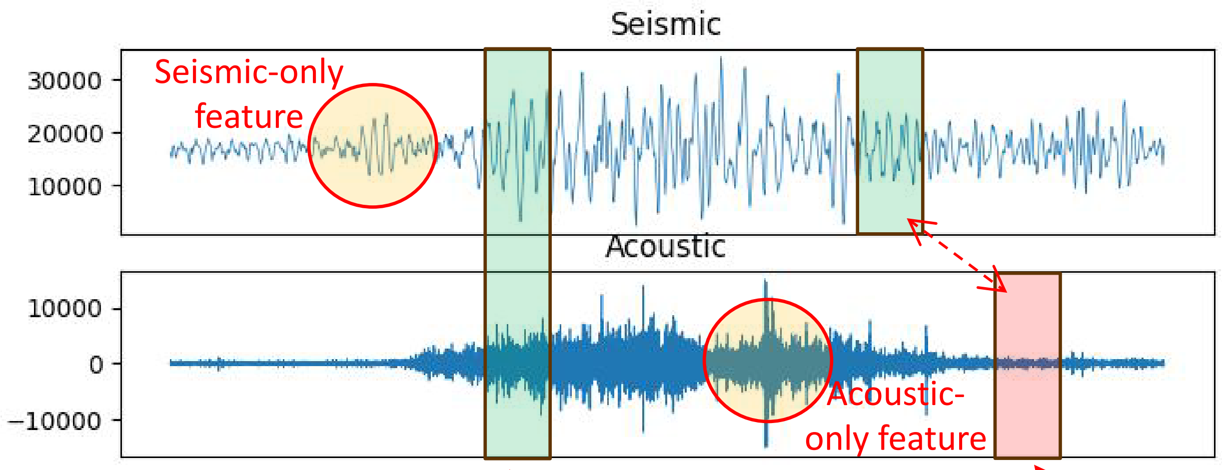


Latent Representation Space

[13] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher, "FOCAL: Contrastive Learning for Multimodal Time-Series Sensing Signals in Factorized Orthogonal Latent Space," In Proc. *37th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, Louisiana, December 2023.

Contrastive Learning from Embedded Sensing: *Multimodal* Data

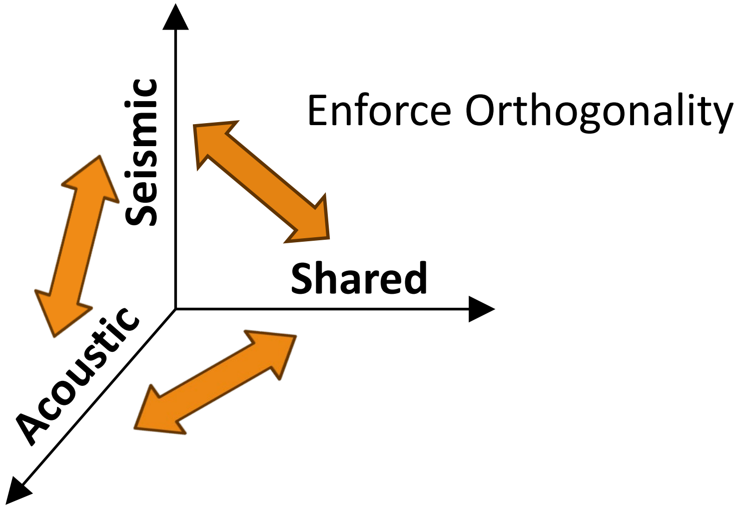
- Suggestion: Enforce orthogonality among shared and private latent subspaces



Multi-sensory Signature of Physical Event/Activity

Same time interval = similar

Different intervals = dissimilar

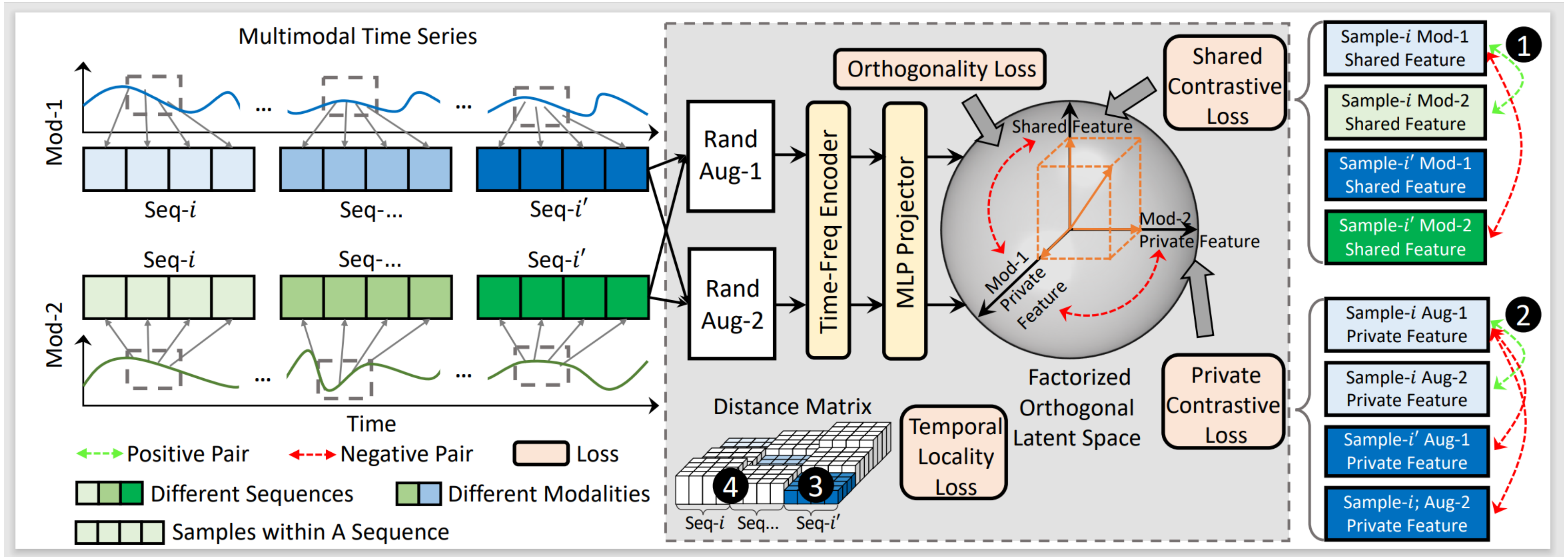


Latent Representation Space

[13] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher, "FOCAL: Contrastive Learning for Multimodal Time-Series Sensing Signals in Factorized Orthogonal Latent Space," In Proc. *37th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, Louisiana, December 2023.

FOCAL: A Miniature "Vibrometry" Foundation Model (Using *Multimodal* Contrastive Learning)

- Extract both ***shared*** and ***private*** information from multi-modal sensing signals in self-supervised manner.
- Appropriately address the information temporal locality within time series data.

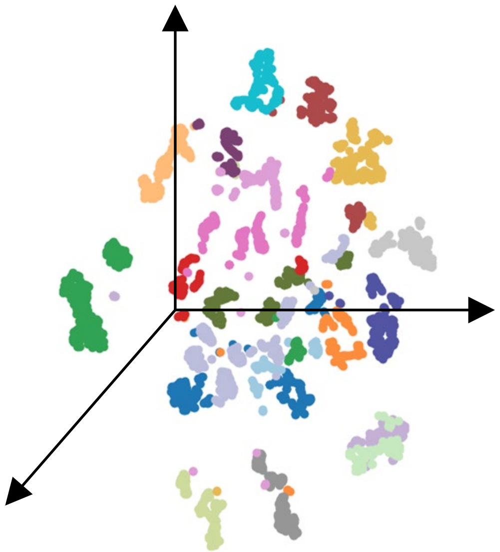


[13] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher, "FOCAL: Contrastive Learning for Multimodal Time-Series Sensing Signals in Factorized Orthogonal Latent Space," In Proc. *37th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, Louisiana, December 2023.

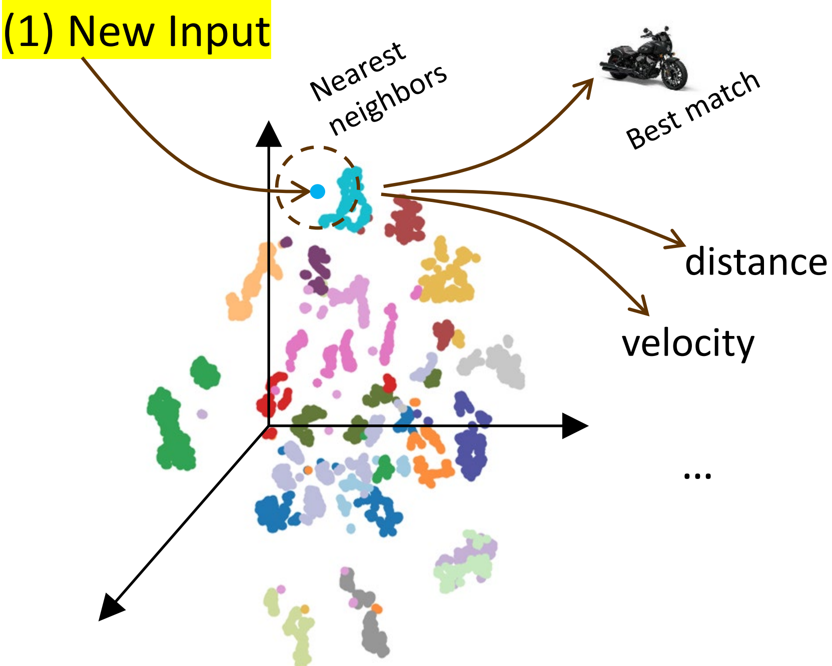
FOCAL: A Miniature "Vibrometry" Foundation Model (Using *Multimodal* Contrastive Learning)

Self-supervised Pretraining

?



Semantic encoding in a sufficiently high-dimensional space



Fine-tuning: insert a few labeled points to map space regions to labels

Evaluation

Downstream Performance with a Linear Classifier

Our method consistently outperforms SOTA time-series contrastive frameworks (TS2Vec, TNC, and GMC), visual contrastive frameworks (SimCLR, MoCo, CMC), and multi-modal contrastive frameworks (CMC, Cosmo, Cocoa, GMC).

Datasets

MOD: Self-collected data using seismic/acoustic signals to classify moving vehicle types.

ACIDS: Seismic/acoustic signals to classify military vehicle types.

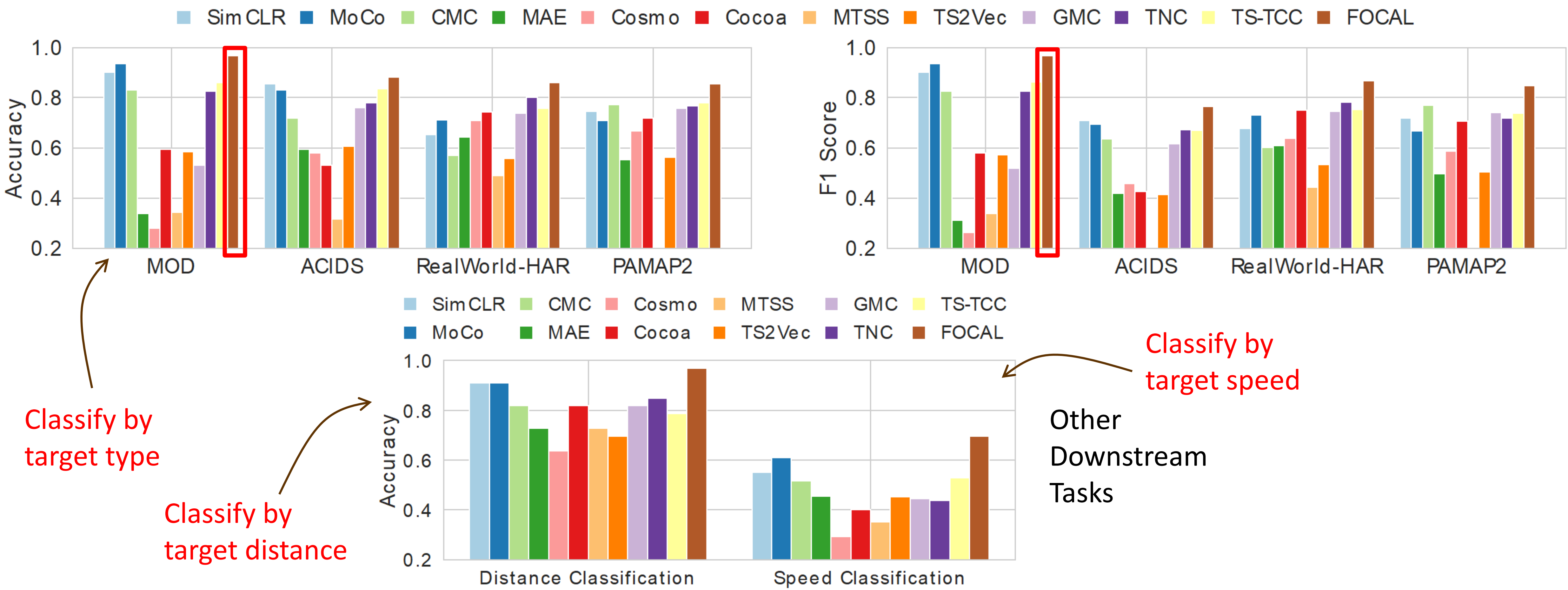
RealWorld-HAR: Use acc/gyro/mag/light signals to recognize human activities.

PAMAP2: Use acc/gyro/mag signals to recognize human activities.

Table 1: Finetune Results with Linear Classifier

Dataset		MOD		ACIDS		RealWorld-HAR		PAMAP2	
Encoder	Framework	Acc	F1	Acc	F1	Acc	F1	Acc	F1
DeepSense	Supervised	0.9404	0.9399	**0.9566**	0.8407	0.9348	**0.9388**	**0.8849**	**0.8761**
	SimCLR	0.8855	0.8855	0.7438	0.6101	0.7138	0.6841	0.6802	0.6583
	MoCo	0.8808	0.8812	0.7717	0.6205	0.7859	0.7708	0.7559	0.7387
	CMC	0.9196	0.9186	0.8443	0.7244	0.7975	0.8116	0.7906	0.7706
	MAE	0.5981	0.5993	0.6644	0.5618	0.7565	0.7515	0.7114	0.6158
	Cosmo	0.8989	0.8998	0.8511	0.6929	0.8956	0.8888	0.8356	0.8135
	Cocoa	0.8774	0.8764	0.6644	0.5359	0.8465	0.8488	0.7603	0.7187
	MTSS	0.4153	0.3582	0.4352	0.2441	0.2989	0.1405	0.3541	0.1795
	TS2Vec	0.7669	0.7648	0.5224	0.3587	0.6595	0.5984	0.5729	0.4715
	GMC	0.9257	0.9267	0.9096	0.7929	0.8869	0.8948	0.8119	0.7860
	TNC	0.9518	0.9528	0.8237	0.6936	0.8892	0.8971	0.8387	0.8143
	TS-TCC	0.8707	0.8735	0.7667	0.6164	0.8073	0.8010	0.7776	0.7250
	Our Method	**0.9732**	**0.9729**	0.9516	**0.8580**	**0.9382**	0.9290	0.8588	0.8463
SW-T Swin-TransformerV2	Supervised	0.8948	0.8931	0.9137	0.7770	0.9313	0.9278	**0.8612**	0.8384
	SimCLR	0.9250	0.9247	0.9128	0.8144	0.7046	0.7220	0.7705	0.7424
	MoCo	0.9390	0.9384	0.9174	0.8100	0.7813	0.8024	0.7717	0.7313
	CMC	0.9129	0.9105	0.8128	0.6857	0.8840	0.8955	0.8080	0.7901
	MAE	0.7803	0.7772	0.8516	0.7023	0.8829	0.8813	0.7910	0.7606
	Cosmo	0.3429	0.3378	0.7110	0.6086	0.8604	0.8169	0.7741	0.7366
	Cocoa	0.7040	0.7038	0.7096	0.5794	0.8892	0.8861	0.7689	0.7317
	MTSS	0.4206	0.4163	0.3429	0.2250	0.5136	0.4370	0.2847	0.1714
	TS2Vec	0.7254	0.7174	0.7183	0.5748	0.6151	0.5955	0.6195	0.5426
	GMC	0.8640	0.8611	0.9402	0.7766	0.9319	0.9379	0.8312	0.8083
	TNC	0.8533	0.8539	0.8352	0.7372	0.8817	0.8784	0.8013	0.7506
	TS-TCC	0.8734	0.8735	0.9041	0.7547	0.8731	0.8454	0.7997	0.7260
	Our Method	**0.9805**	**0.9800**	**0.9489**	**0.8262**	**0.9451**	**0.9503**	0.8580	**0.8401**

Results: Downstream Performance on Multiple Tasks with K-Nearest-Neighbor Classifier (K=5)



Classify by target type

Classify by target distance

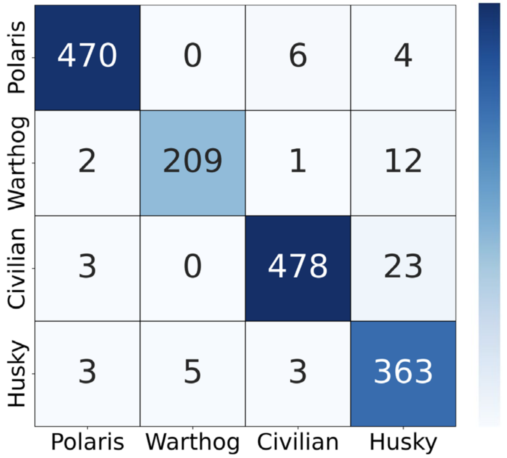
Classify by target speed

Other Downstream Tasks

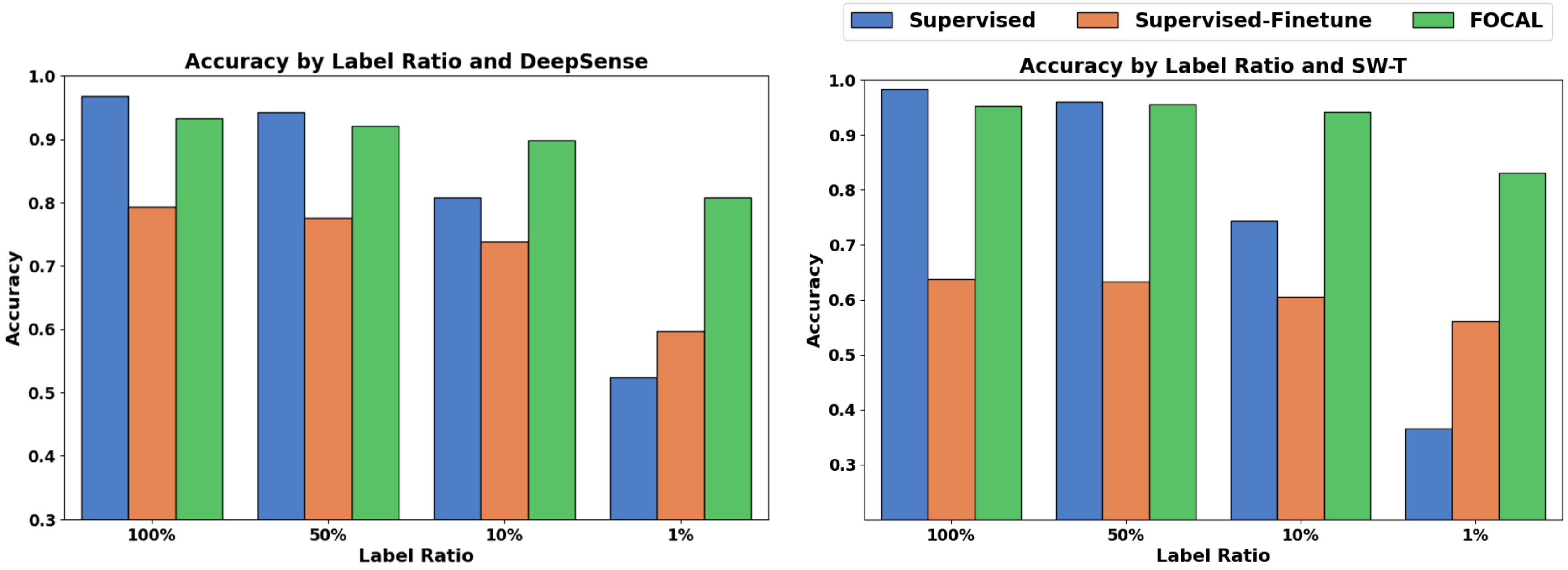
[14] Tomoyoshi Kimura, Jinyang Li, Tianshi Wang, Denizhan Kara, Yizhuo Chen, Yigong Hu, Ruijie Wang, Maggie Wigness, Shengzhong Liu, Mani Srivastava, Suhas Diggavi, Tarek Abdelzaher, "On the Efficiency and Robustness of Vibration-based Foundation Models for IoT Sensing: A Case Study," In Proc. *FM-Sys*, May 2024.

Evaluation of Robustness

How much fine-tuning (with labeled data) is needed to adapt a pre-trained model to a domain shift (new environment or new target)?



Test Confusion Matrix for Different Targets
(Husky not seen during pre-training)

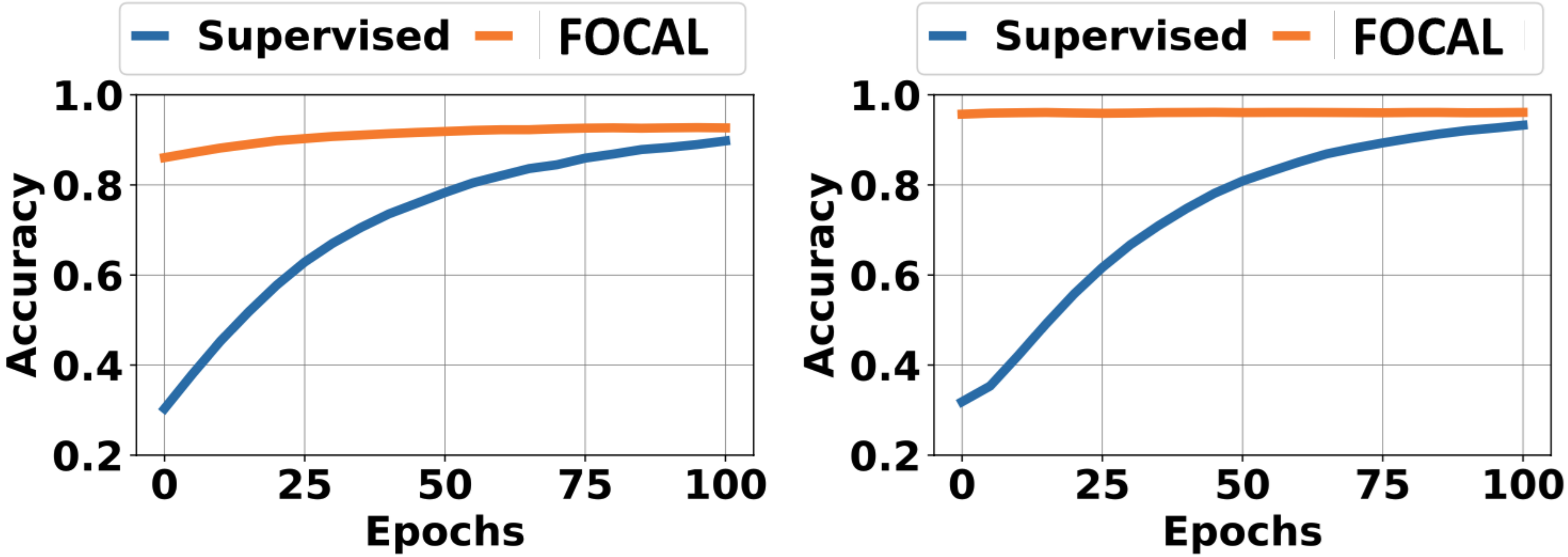


Fine-tuning performance at deployment for different labeled data sizes

[14] Tomoyoshi Kimura, Jinyang Li, Tianshi Wang, Denizhan Kara, Yizhuo Chen, Yigong Hu, Ruijie Wang, Maggie Wigness, Shengzhong Liu, Mani Srivastava, Suhas Diggavi, Tarek Abdelzaher, "On the Efficiency and Robustness of Vibration-based Foundation Models for IoT Sensing: A Case Study," In Proc. *FM-Sys*, May 2024.

Learning Speed

Accuracy curves of Supervised Training versus Fine-tuning (FOCAL)



(a) DeepSense (b) SW-T

[14] Tomoyoshi Kimura, Jinyang Li, Tianshi Wang, Denizhan Kara, Yizhuo Chen, Yigong Hu, Ruijie Wang, Maggie Wigness, Shengzhong Liu, Mani Srivastava, Suhas Diggavi, Tarek Abdelzaher, "On the Efficiency and Robustness of Vibration-based Foundation Models for IoT Sensing: A Case Study," In Proc. *FM-Sys*, May 2024.

Resource Overhead

Small enough to run fine-tuning and inference on Edge devices (e.g., Raspberry Shake)

Much faster than training a supervised model with the same amount of data

Inference Time (from 1 second of data) on Raspberry Pi Device

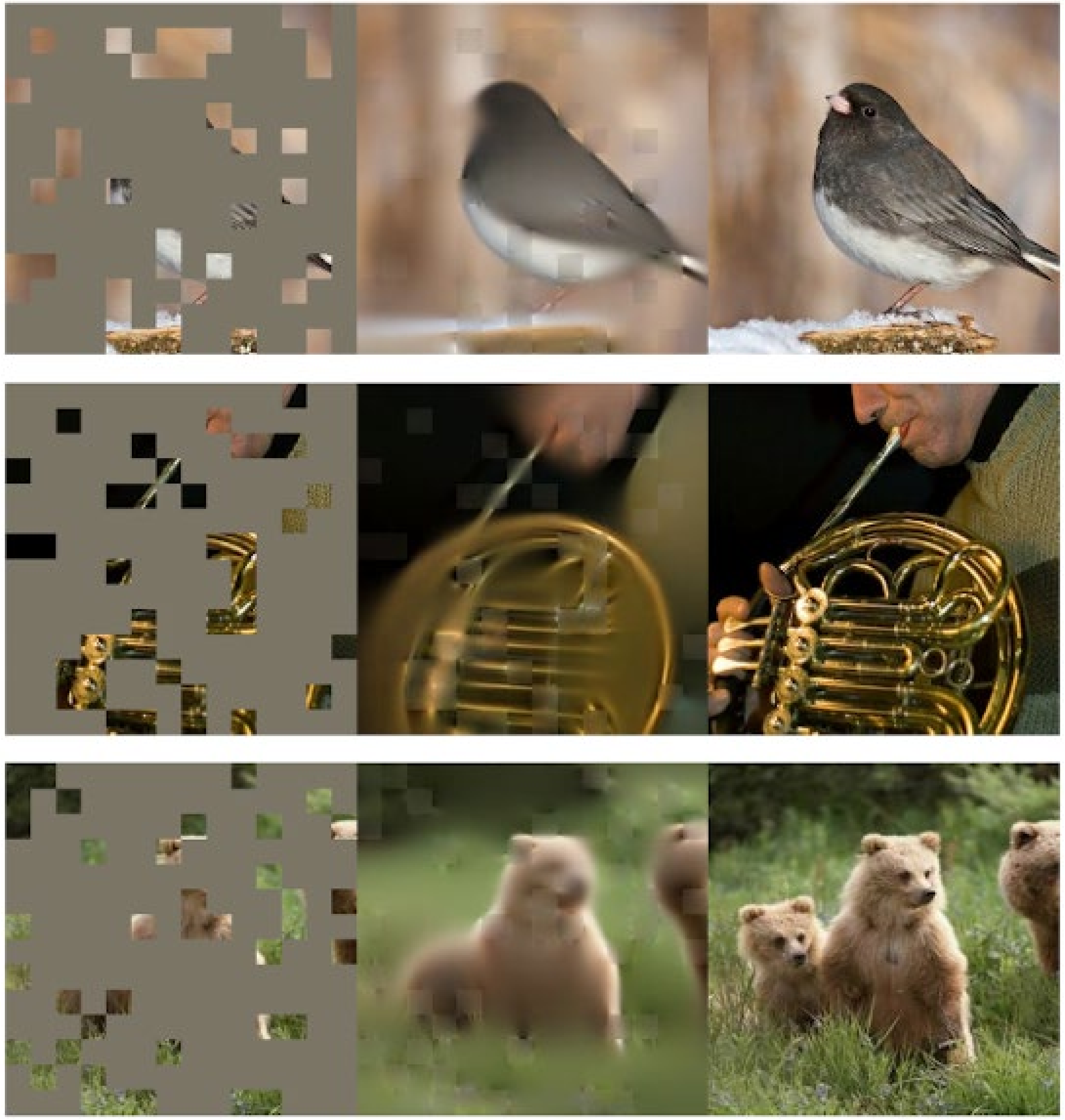
Encoder	Size (MB)	Parameters (M)	Infer Speed (s)
DeepSense	25.27	6.6220	0.1011
SWIN-Transformer	44.955	11.7725	0.1841
TSMixer	7.463	1.9523	0.0709

Model	Framework	B=1	B=2	B=4	B=8	B=16	B=32	B=64	B=128	Average Improvement
DeepSense	Supervised	0.6499	0.8850	1.2151	1.9488	3.4106	6.9621	13.3596	29.6567	88.49%
	FOCAL - Finetuned	0.1052	0.1374	0.1724	0.2452	0.3481	0.5901	1.0795	2.0166	
SW-T	Supervised	1.2364	1.5483	2.2258	3.5723	6.1268	11.2664	21.5920	42.6260	64.84%
	FOCAL - Finetuned	0.2639	0.4035	0.6932	1.2597	2.4553	4.5907	9.2683	18.6447	
TSMixer	Supervised	0.3526	0.5386	0.8981	1.5925	3.0116	5.8583	11.5925	24.8614	54.94%
	FOCAL - Finetuned	0.1215	0.2092	0.3825	0.7470	1.4690	2.8076	5.6527	12.9797	

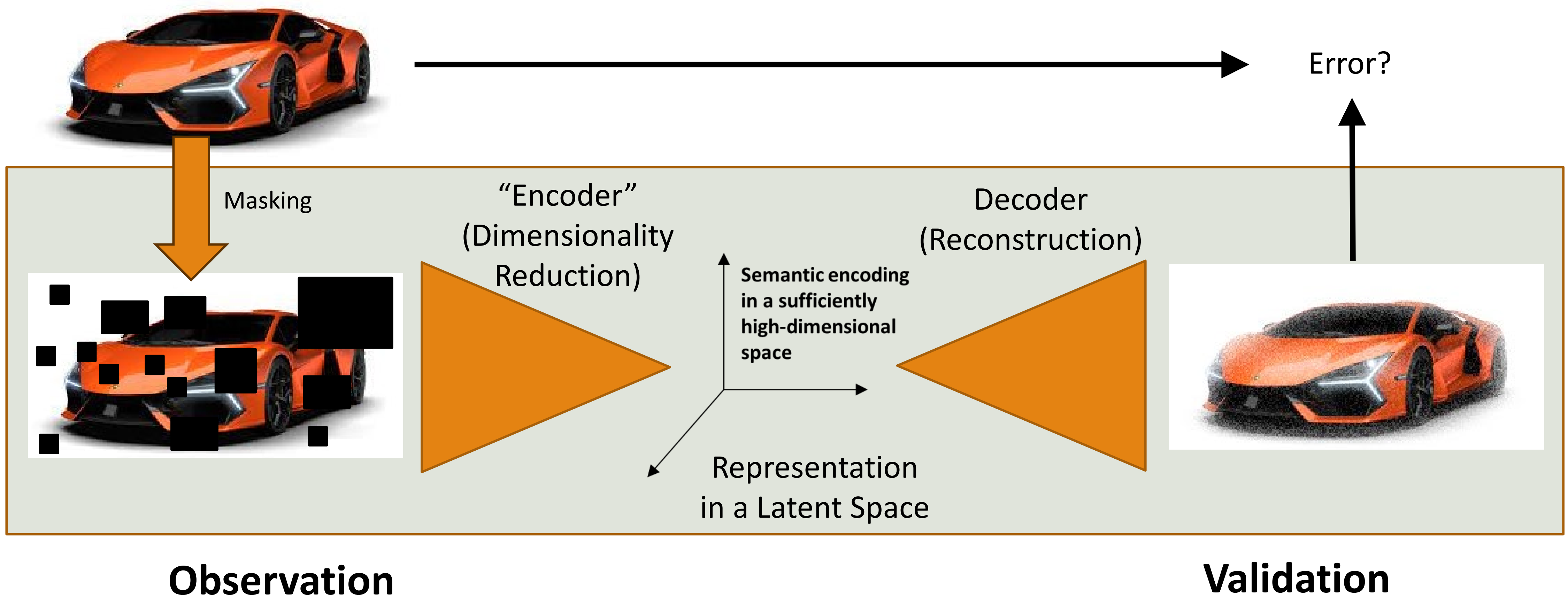
Training Time on Raspberry Pi Device

Challenge:

Masked Auto-Encoders for Embedded Computing



Masked Autoencoders



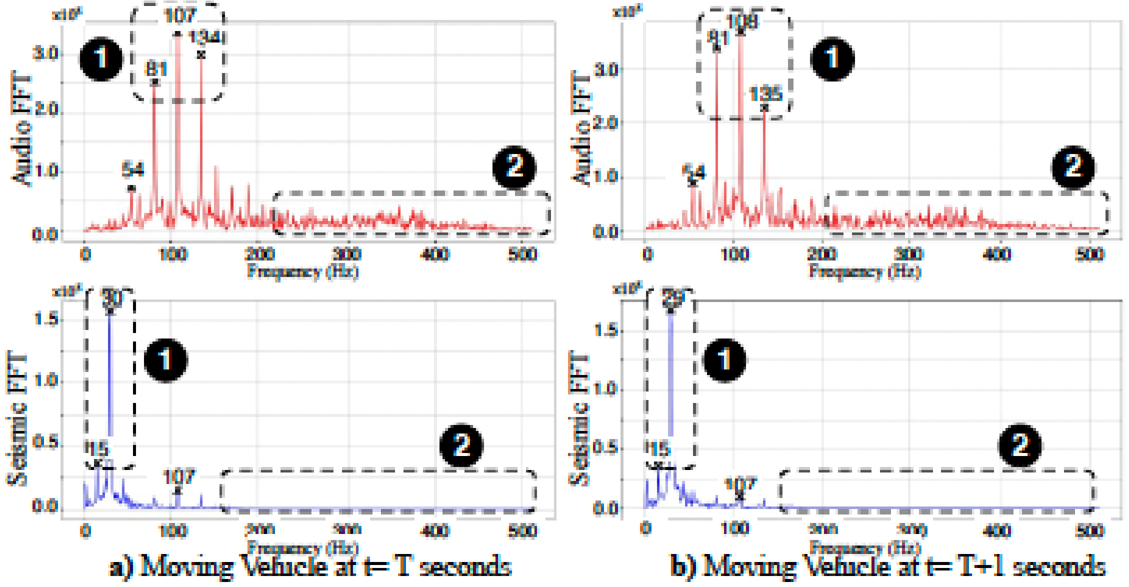
[15] Denizhan Kara, Shengzhong Liu, Jinyang Li, Dongxin Liu, Tianshi Wang, Ruijie Wang, Yizhuo Chen, Yigong Hu, Tarek Abdelzaher, "FreqMAE: Frequency-Aware Masked Autoencoder for Multi-Modal IoT Sensing," In Proc. *The Web Conference (WWW)*, May 2024.

Challenges

1. **No Scale and Shift Invariance**
   ❶ Position and scale shifts in spectrogram imply semantic differences.

2. **Multi-Modal Fusion is Essential**
   Each sensor modality offers unique insights, and their fusion leads to a richer understanding.

3. **Varied Information Density across Spectrum**
   ❷ Signal and noise have different densities in different parts of the spectrum.

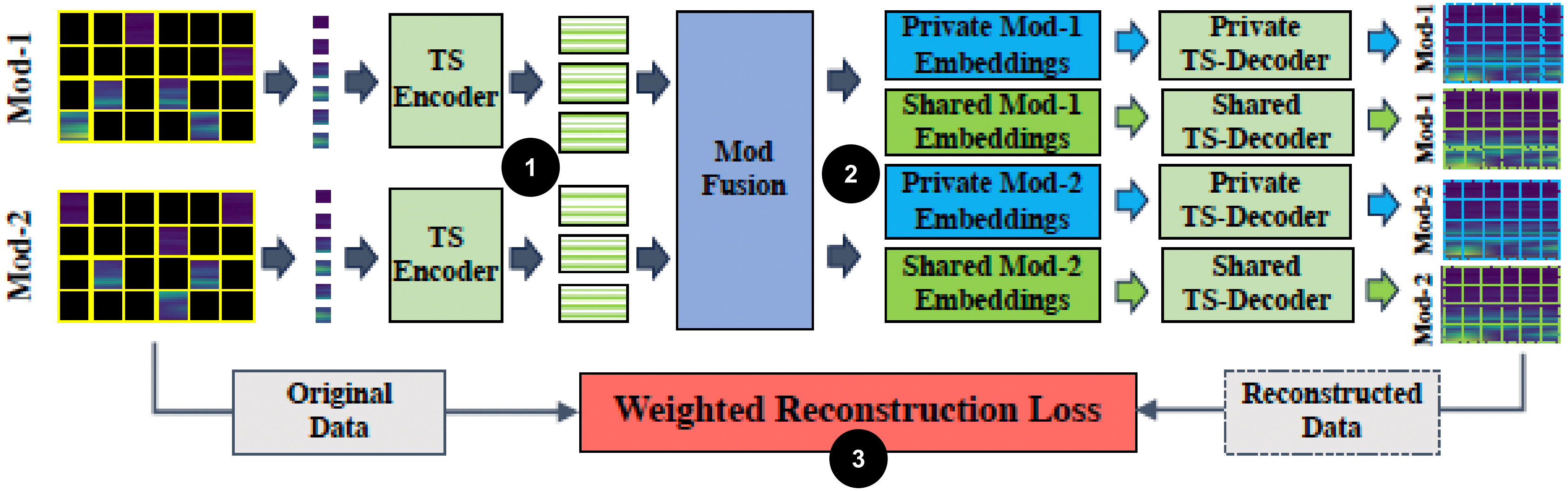


Audio FFT signatures for a moving vehicle. ❶ The presence of characteristic peaks in localized regions needs local harmonic associations and shift-sensitive representations. ❷ Higher frequency regions mostly contain noise.

[15] Denizhan Kara, Shengzhong Liu, Jinyang Li, Dongxin Liu, Tianshi Wang, Ruijie Wang, Yizhuo Chen, Yigong Hu, Tarek Abdelzaher, “FreqMAE: Frequency-Aware Masked Autoencoder for Multi-Modal IoT Sensing,” In Proc. *The Web Conference (WWW)*, May 2024.

FreqMAE

1. **Timeseries Spectrogram (TS) Transformer:** Transformer incorporating localized attention with a spectrogram-compatible shifting mechanism.
2. **Factorized Modality Fusion:** Learns private embeddings for modality-specific information and shared embeddings for cross-modal representations.
3. **Weighted Loss Function:** Emphasizes lower frequency within samples, and higher energy samples across datasets for efficient self-supervised pretraining.



[15] Denizhan Kara, Shengzhong Liu, Jinyang Li, Dongxin Liu, Tianshi Wang, Ruijie Wang, Yizhuo Chen, Yigong Hu, Tarek Abdelzaher, "FreqMAE: Frequency-Aware Masked Autoencoder for Multi-Modal IoT Sensing," In Proc. *The Web Conference (WWW)*, May 2024.

Evaluation

Datasets: Four different public datasets from two application domains

- **Vehicle Classification (VC):** ACIDS and MOD
- **Human Activity Recognition(HAR):** PAMAP2 and RealWorld-HAR

Preprocessing: Create spectrograms via FFT after splitting time-series to evenly sized sample windows

Training: Divide dataset runs to train-validation-test sets (roughly 8:1:1)

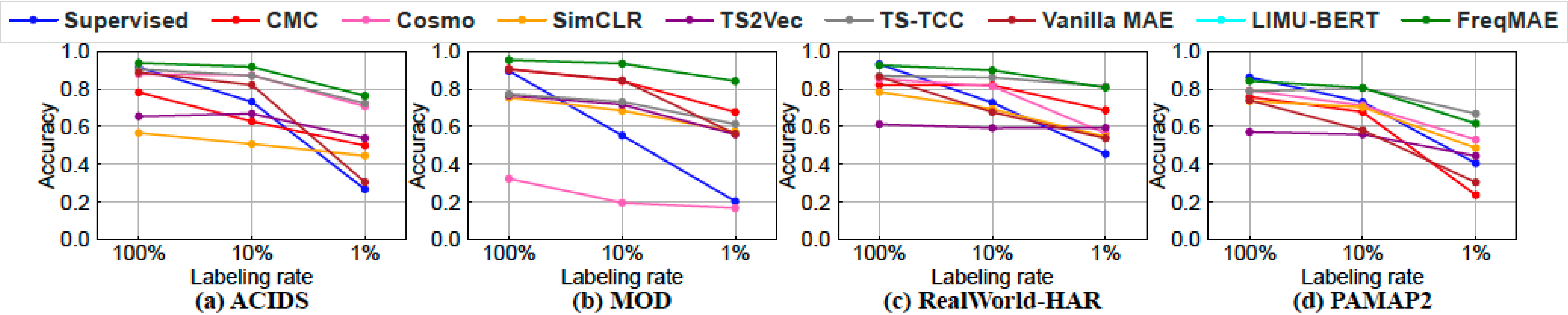
Table 1: Dataset Summary

Dataset	# Classes	Modalities[2]	# Samples	Application
MOD	7	MP, S	39,609	VC
ACIDS	9	MP, S	27,597	VC
RealWorld-HAR	8	A, G, M, L	12,887	HAR
PAMAP2	18	A, G, M	9,611	HAR

[15] Denizhan Kara, Shengzhong Liu, Jinyang Li, Dongxin Liu, Tianshi Wang, Ruijie Wang, Yizhuo Chen, Yigong Hu, Tarek Abdelzaher, “FreqMAE: Frequency-Aware Masked Autoencoder for Multi-Modal IoT Sensing,” In Proc. *The Web Conference (WWW)*, May 2024.

Evaluation

- Improved classification accuracy compared to other approaches (especially when the amount of labeled data (for training and/or fine-tuning) is low
- Reduced need for labeled samples



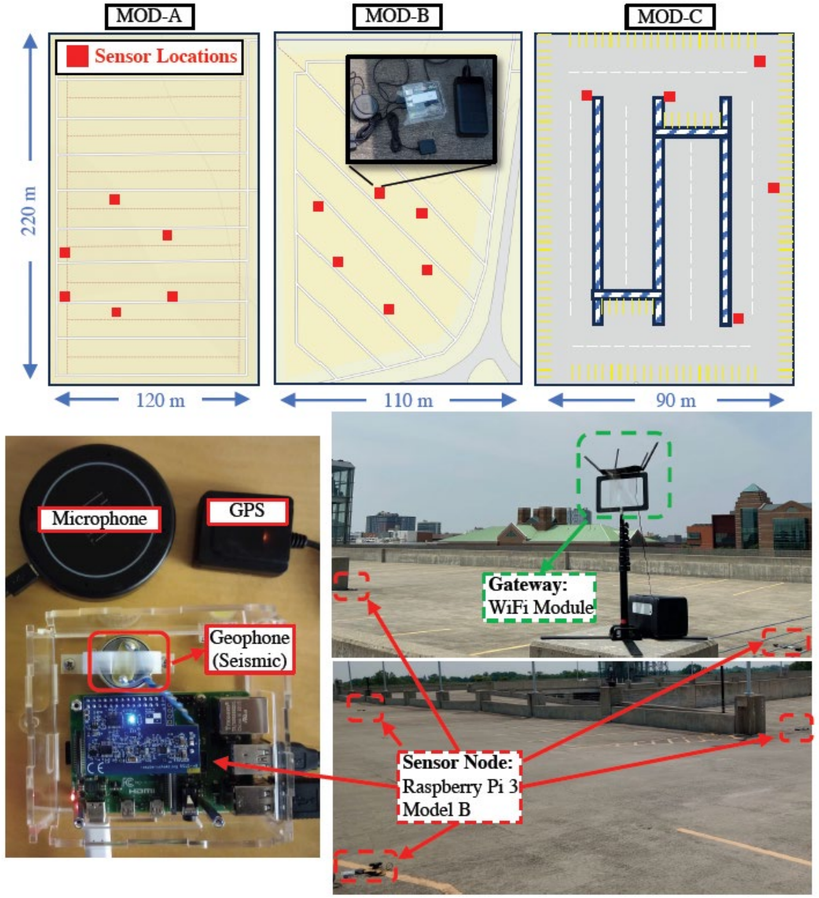
[15] Denizhan Kara, Shengzhong Liu, Jinyang Li, Dongxin Liu, Tianshi Wang, Ruijie Wang, Yizhuo Chen, Yigong Hu, Tarek Abdelzaher, "FreqMAE: Frequency-Aware Masked Autoencoder for Multi-Modal IoT Sensing," In Proc. *The Web Conference (WWW)*, May 2024.

Masking Strategies

	PAMAP		RWHAR		ACIDS	
Metric	Acc	F1	Acc	F1	Acc	F1
CMC	0.7571	0.7223	0.8211	0.8384	0.7836	0.6452
Cosmo	0.7910	0.7469	0.8529	0.7968	0.8776	0.7298
SimCLR	0.7346	0.6635	0.7830	0.7181	0.5658	0.4879
TS2Vec	0.5706	0.4942	0.6117	0.5002	0.6539	0.4913
TS-TCC	0.7871	0.7107	0.8684	0.8227	0.8758	0.7400
Vanilla MAE	0.7382	0.6999	0.8638	0.8700	0.8521	0.6908
LIMU-BERT	0.7847	0.7612	0.7946	0.7261	0.5023	0.3171
AudioMAE	0.7808	0.7478	0.8163	0.7437	0.7845	0.6120
PhyMask	**0.8056**	**0.7719**	**0.9059**	**0.9137**	**0.9265**	**0.8044**

Tables show improved performance with a new masking strategy (PhyMask) that prefers masking semantically significant regions

Deployment Experiments



	MOD-A		MOD-B		MOD-C	
Metric	Acc	F1	Acc	F1	Acc	F1
CMC	0.7415	0.7390	0.5760	0.4983	0.6412	0.5691
Cosmo	0.4205	0.3059	0.5816	0.5214	0.5496	0.2376
SimCLR	0.6733	0.6685	0.5377	0.3922	0.6107	0.3730
TS2Vec	0.6563	0.6439	0.5260	0.3521	0.5725	0.4487
TS-TCC	0.6051	0.5910	0.5012	0.1720	0.5802	0.4099
Vanilla MAE	0.8580	0.8602	0.6626	0.6347	0.6794	0.6326
LIMU-BERT	0.5000	0.1667	0.4233	0.1983	0.5649	0.2407
CAV-MAE	0.4801	0.4431	0.50309	0.21076	0.5419	0.3409
AudioMAE	0.5113	0.4981	0.4839	0.3475	0.4961	0.4571
FreqMAE	**0.8750**	**0.8766**	**0.6885**	**0.6622**	**0.7710**	**0.7340**

Testing in three locations: A, B, and C.

Conclusions

The recent AI/ML revolution is a key opportunity for real-time computing!

- *We specialize in managing bottleneck computing resources*.
  - → AI/ML is creating the world's largest computing bottleneck!
    - Exploit latency/quality tradeoffs in computing and communication
    - Prioritize data processing (i.e., attention scheduling) to meet latency constraints
    - Derive spatial-temporal real-time attention bounds
    - Explore the impact of thermal control, DVFS, etc.
- *We specialize in embedded computing*
  - → Embodied AI is embedded AI
    - Learning from Sensor Data (in frequency domain, multimodal, harmonic structure, …)

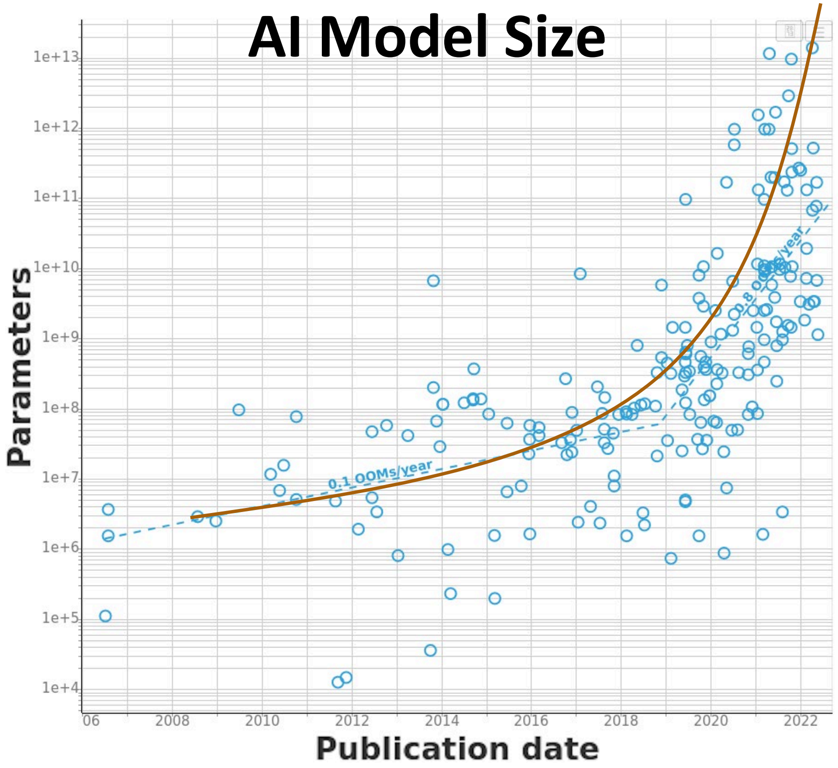
AI + RT/Embedded collaborations could bring a wealth of new perspectives and applications

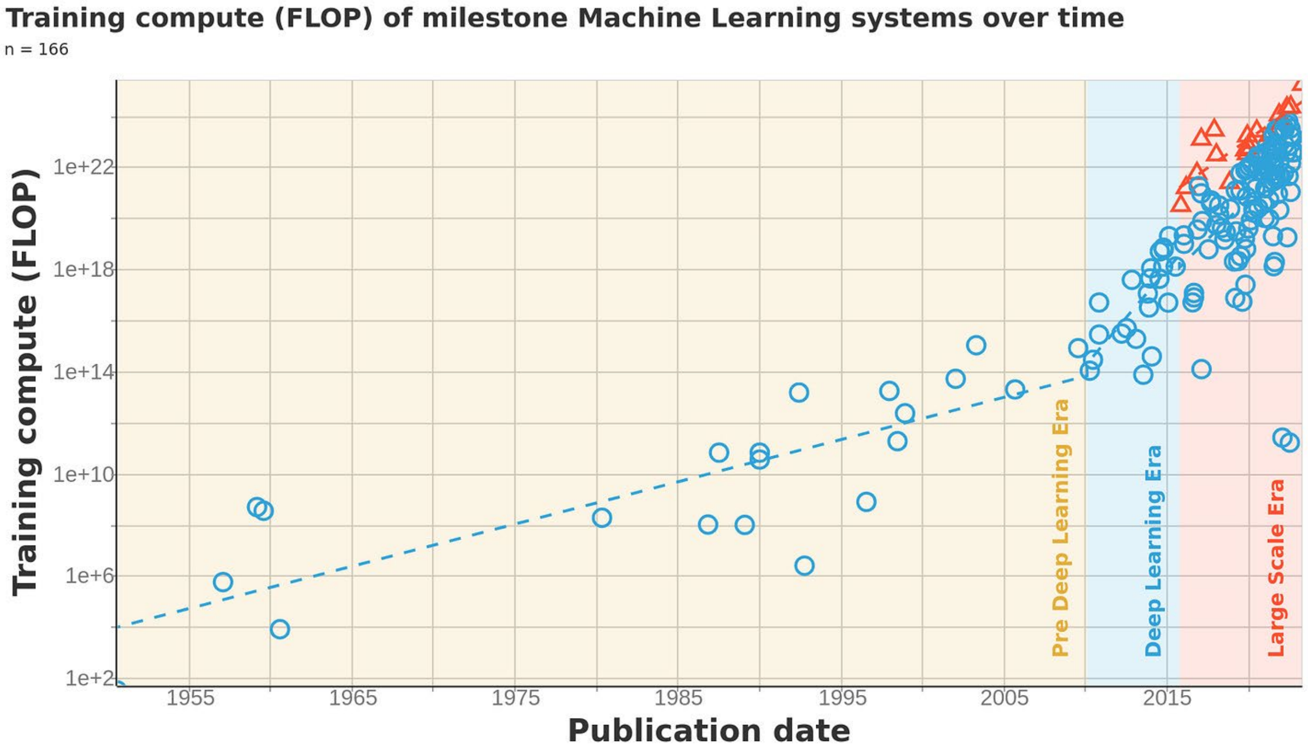
AI is Creating the World’s Largest Computing Bottleneck

Moore’s Law: Capacity doubles **every 18 months**.

AI model size doubles approximately **every 3.4 months**.

https://www.computerweekly.com/news/252475371/Stanford-University-finds-that-AI-is-outpacing-Moores-Law





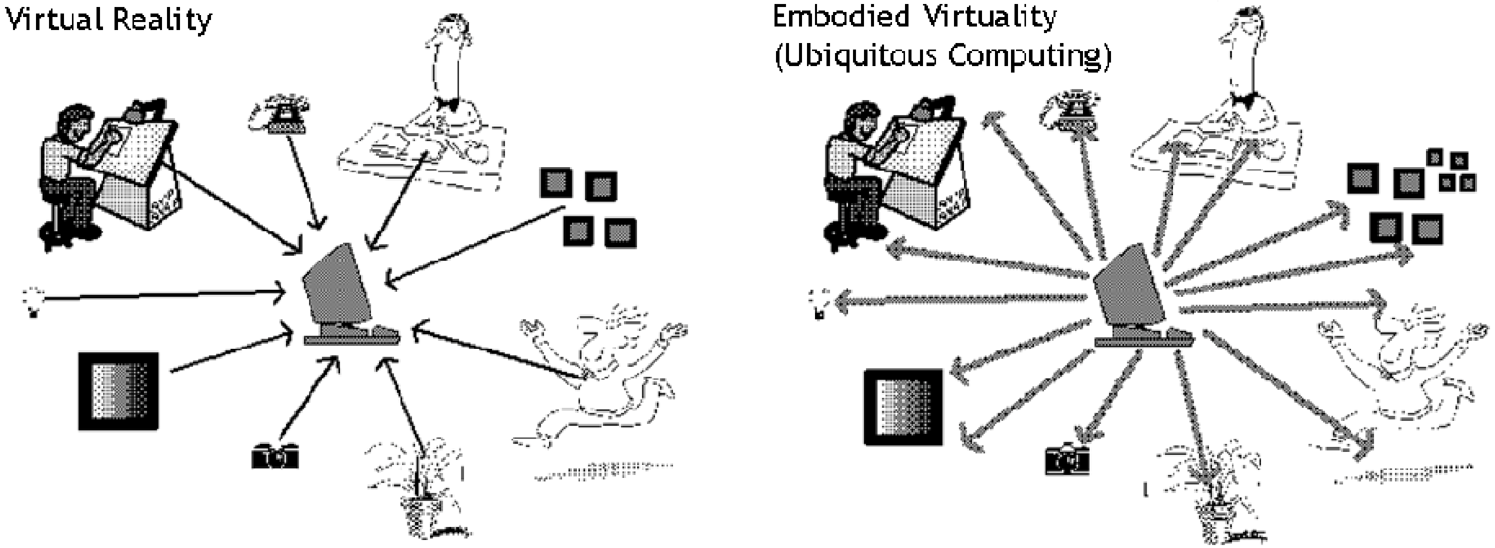
Emerging Applications in Human Interactions

Creating new interaction spaces (between humans and the environment), not natively supported by the underlying physical objects.

Virtual Reality: Manipulate human perception to create (virtual) spaces that allow novel computationally-enabled interactions

Ubiquitous Computing (IoT): Embed computation into the environment to create (smart) spaces that allow novel computationally enabled interactions

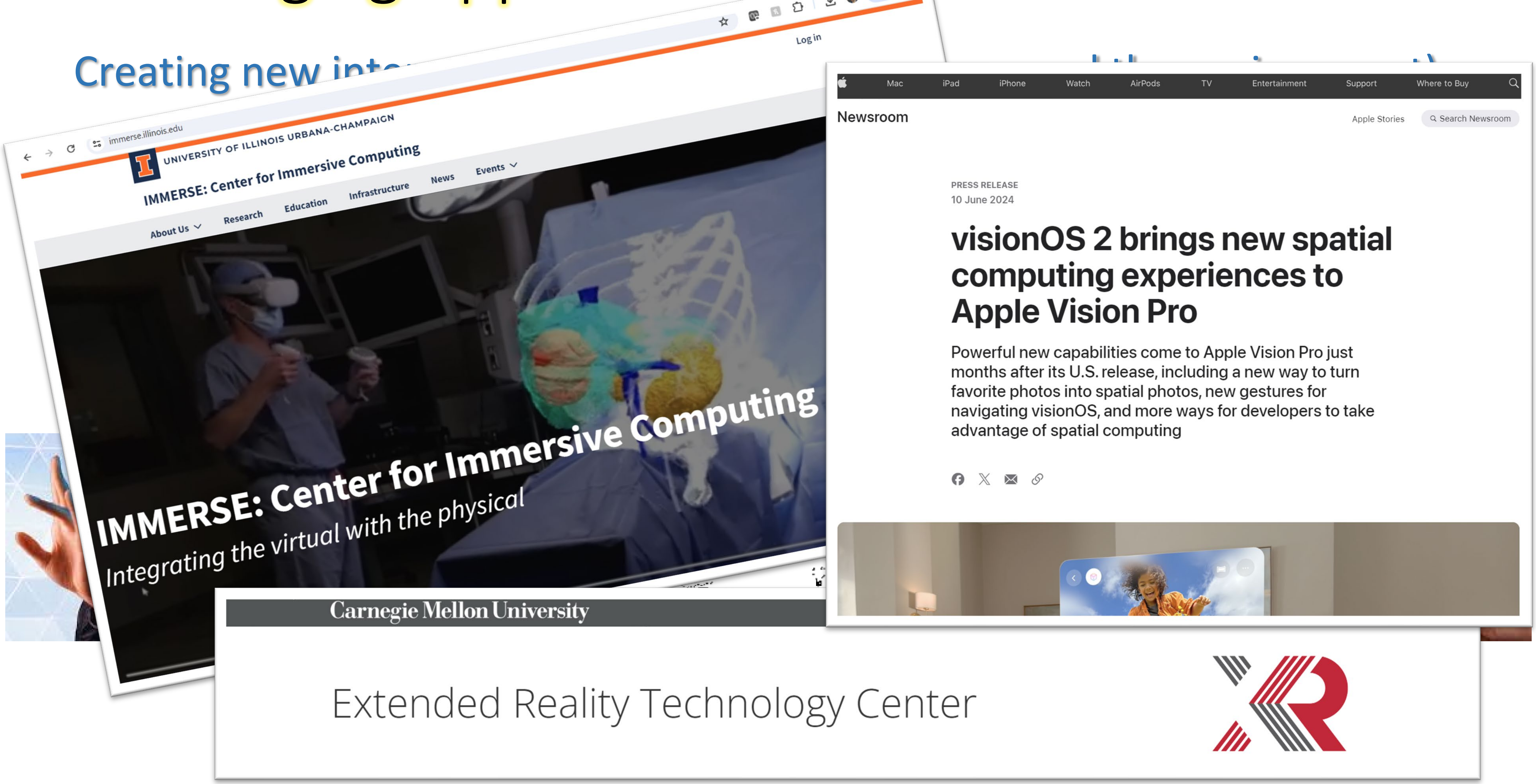
Virtual — Physical



Mark Weiser's cartoons about Ubiquitous Computing vs. Virtual Reality (late 80s)

Emerging Applications in Human Interactions

Creating new inte

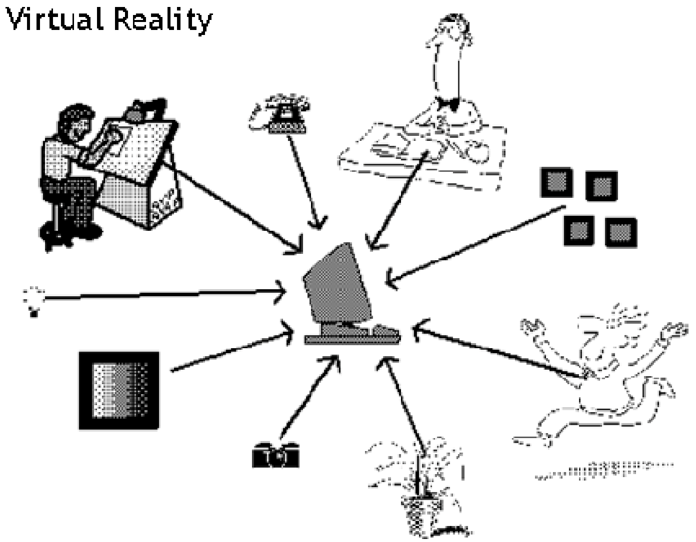


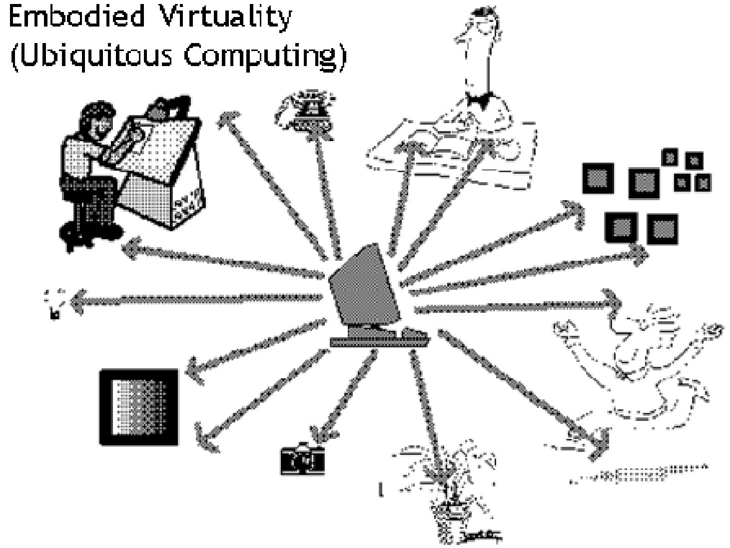
[16] Tarek Abdelzaher, Matthew Caesar, Charith Mendis, Klara Nahrstedt, Mani Srivastava and Minlan Yu, "Challenges in Metaverse Research: An Internet of Things Perspective," In Proc. *1st IEEE International Conference on Metaverse Computing, Networking and Applications (MetaCom)*, Kyoto, Japan, June 2023.

Why Now?

Virtual reality and ubiquitous computing visions have existed for over 40 years. Why an emerging application now?

Virtual | Physical





Mark Weiser's cartoons about Ubiquitous Computing vs. Virtual Reality (late 80s)

How Do Content-Centric Applications Rise?

How Do Content-Centric Applications Rise?

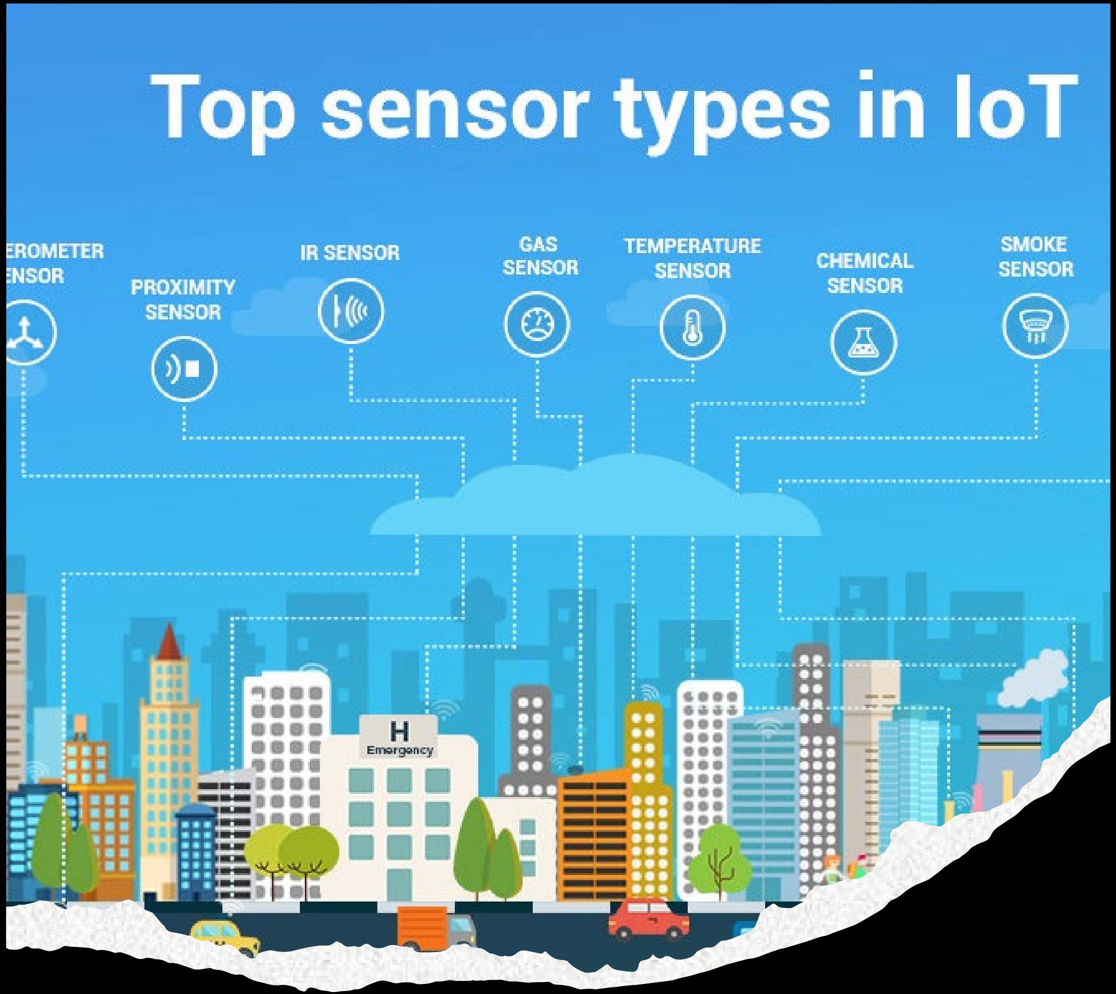
Hint: When the cost of content creation is lowered

- **YouTube? (2005)**
  - Promoted by the proliferation of camera phones

How Do Content-Centric Applications Rise?

Hint: When the cost of content creation is lowered

- **Instagram? (2010)**
  - Promoted by the proliferation of digital photography

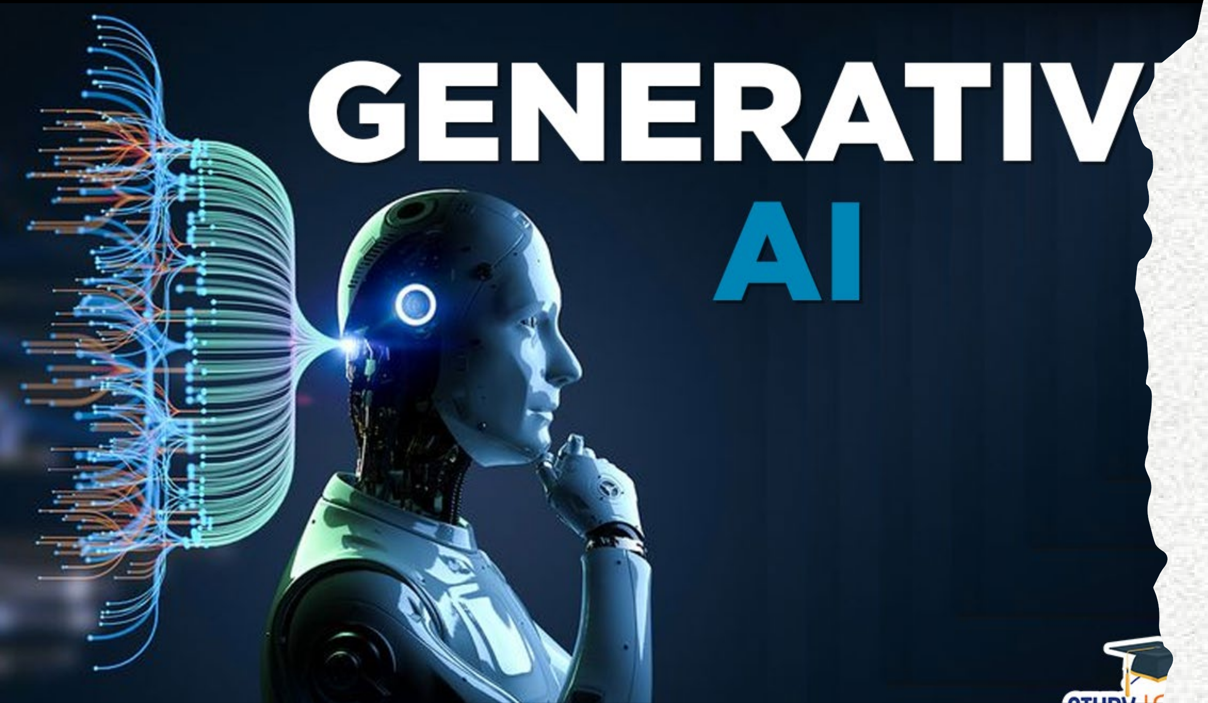


How Do Content-Centric Applications Rise?

Hint: When the cost of content creation is lowered

- **Internet of Things? (~ 2010)**
  - Promoted by the proliferation of cheap sensor data (and connectivity)

What about Immersive Computing?



What about Immersive Computing? (~ now)

Hint: When the cost of content creation is lowered

- 360 cameras (Content Capture)
- Generative AI (Creative Authorship)

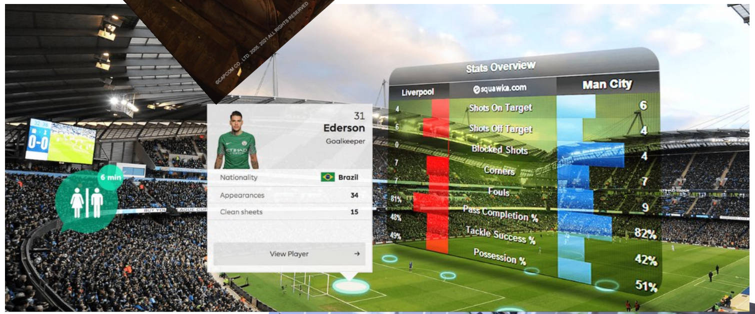
Cultural
Preservation

Art and Culture

Media and
Entertainment

VR Gaming

Applications



Live
Sports

Services
(Metaverse Seoul)

Training and Simulation

Digital Twins/Design

Teleconferencing/Workspace

Immersive Computing:
A Computing Services Perspective

Back-end Acceleration

Cloud and Data Center Networks, Storage, Reliability

Low-latency Networking, Verification, Digital Twins

360 Video Acceleration, Multimedia, Security, QoS

Wireless Sensing, Mobile Computing, IoT/CPS

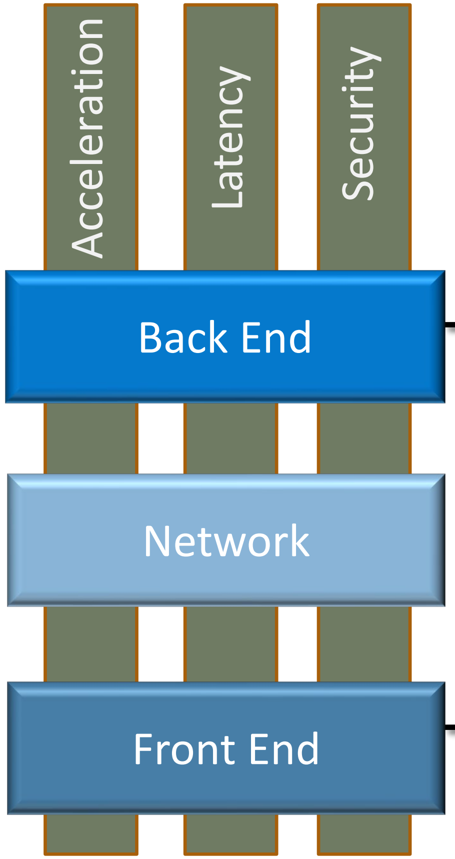


Image source: https://www.digitaljournal.com/

Application:
Observational Science at Scale
From Millions of Observations to Compact Models of Phenomena

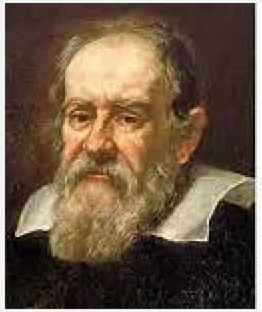
Observations

Laws of Gravity?

Validation

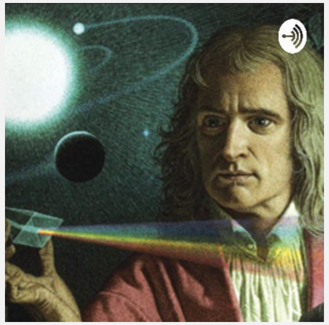
Towards a Science of *Observational Social-Information Dynamics*

In the 17th Century



A new **observational instrument** (Galileo's Telescope)

&

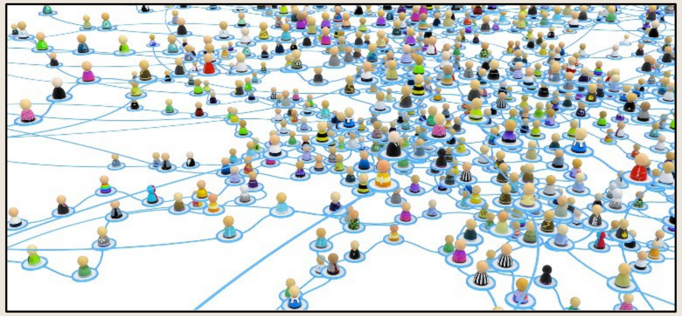


A new **latent state representation** (discovery of gravity)

A science of the motion of object positions in physical space

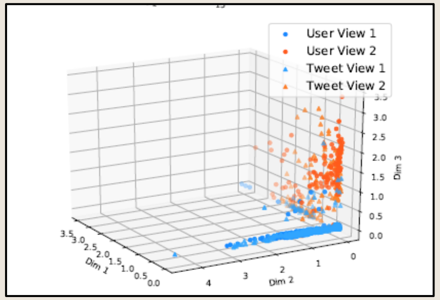
Newton's Laws of Mechanics

Today



A new **observational instrument** (Online Social Media)

&



A new **latent state representation** (network embedding)

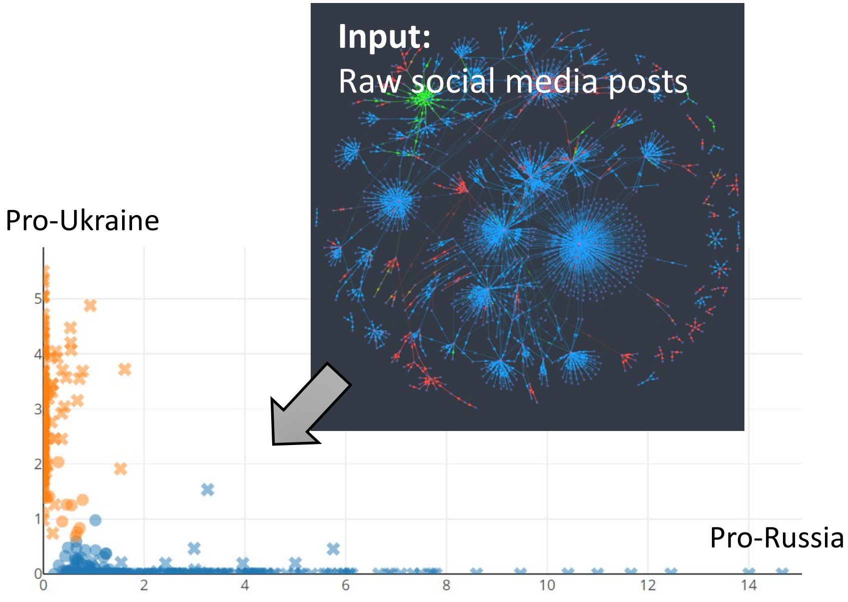
A science of the motion of human positions (beliefs) in ideological space

Observational Social-Information Dynamics

[17] Jinning Li, Huajie Shao, Dachun Sun, Ruijie Wang, Yuchen Yan, Jinyang Li, Shengzhong Liu, Hanghang Tong, Tarek Abdelzaher, “Unsupervised Belief Representation Learning with Information-Theoretic Variational Graph Auto-Encoders,” In Proc. *SIGIR*, Madrid, Spain, July 2022.

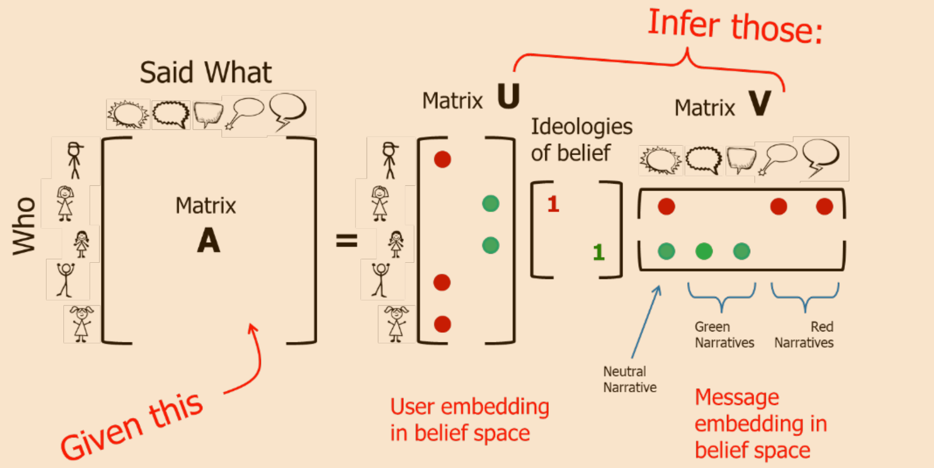
Ideological (Belief) Embedding

(i) Dimensions represent different views
(ii) Nodes move depending on their view adoption
(iii) The original is “neutral”

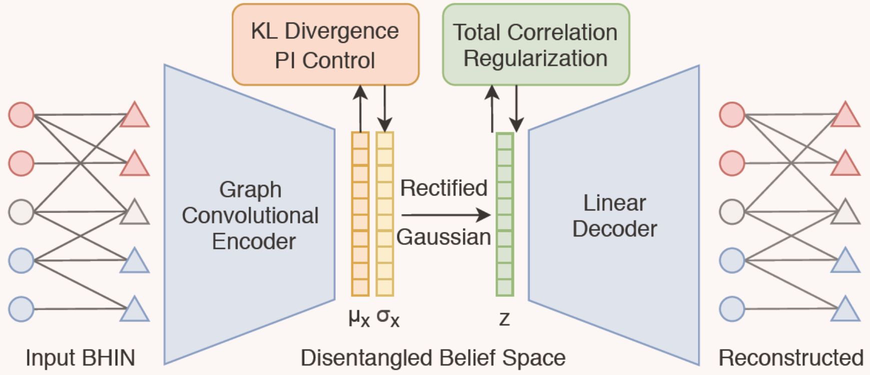


Example Output

Version 1.0: Non-negative matrix factorization. (Linear “encoding” and “decoding”.)



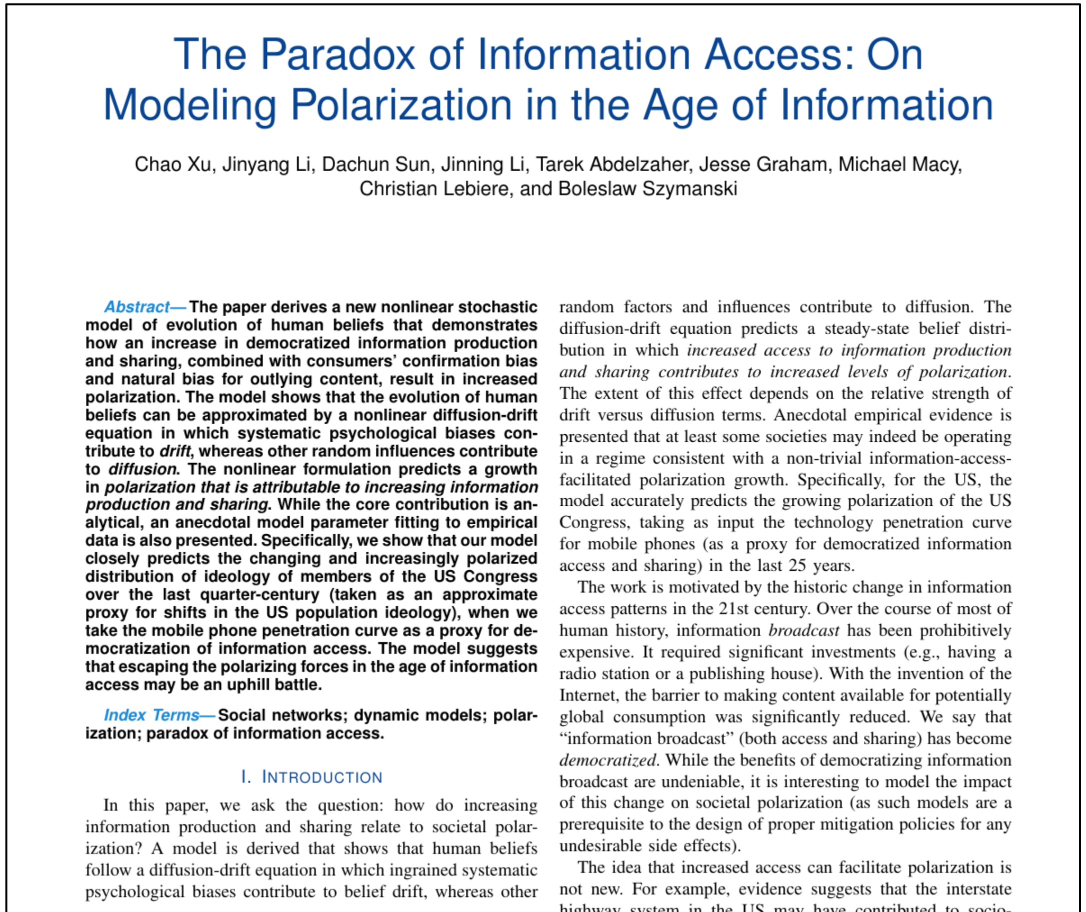
Version 2.0: Graph Auto-Encoder [1]. Non-linear (Graph Convolutional) “encoding” and linear “decoding” (taking both link and node attributed into account)



Application: Social Dynamics Forecasting
(Predict Escalation/Radicalization/Reconciliation)

[18] Chao Xu, Jinyang Li, Dachun Sun, Jinning Li, Tarek Abdelzaher, Jesse Graham, Michael Macy, Christian Lebiere, and Boleslaw Szymanski, "The Paradox of Information Access: On Modeling Polarization in the Age of Information," *IEEE Transactions on Control of Network Systems*, Accepted in 2023.

- The dynamic trajectories of beliefs predict future population opinion distribution
- Predicting (and defending against) the potential impact of adversarial manipulations in the information space



The Paradox of Information Access: On Modeling Polarization in the Age of Information

Chao Xu, Jinyang Li, Dachun Sun, Jinning Li, Tarek Abdelzaher, Jesse Graham, Michael Macy, Christian Lebiere, and Boleslaw Szymanski

***Abstract*—The paper derives a new nonlinear stochastic model of evolution of human beliefs that demonstrates how an increase in democratized information production and sharing, combined with consumers' confirmation bias and natural bias for outlying content, result in increased polarization. The model shows that the evolution of human beliefs can be approximated by a nonlinear diffusion-drift equation in which systematic psychological biases contribute to *drift*, whereas other random influences contribute to *diffusion*. The nonlinear formulation predicts a growth in *polarization that is attributable to increasing information production and sharing*. While the core contribution is analytical, an anecdotal model parameter fitting to empirical data is also presented. Specifically, we show that our model closely predicts the changing and increasingly polarized distribution of ideology of members of the US Congress over the last quarter-century (taken as an approximate proxy for shifts in the US population ideology), when we take the mobile phone penetration curve as a proxy for democratization of information access. The model suggests that escaping the polarizing forces in the age of information access may be an uphill battle.**

***Index Terms*—Social networks; dynamic models; polarization; paradox of information access.**

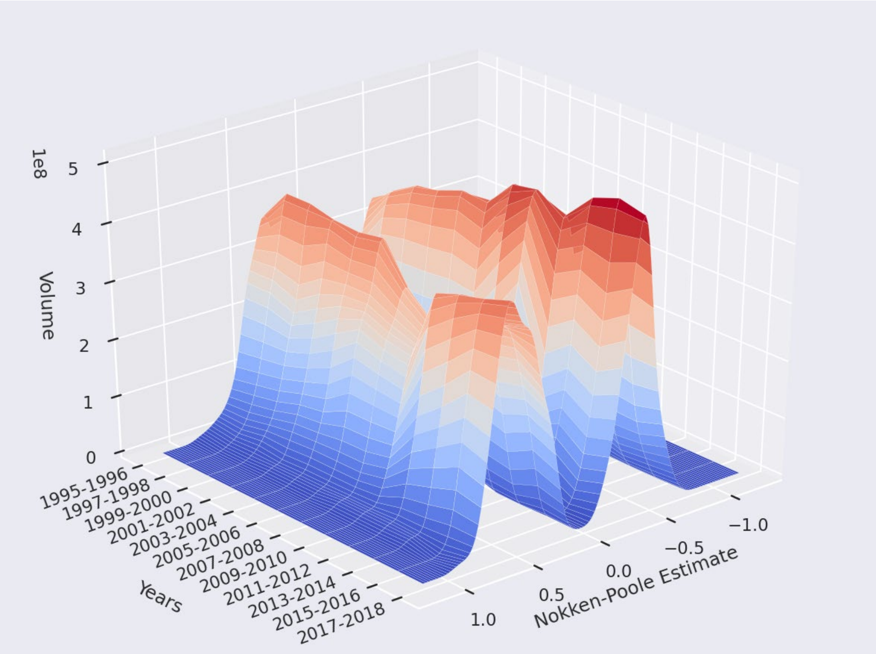
I. Introduction

In this paper, we ask the question: how do increasing information production and sharing relate to societal polarization? A model is derived that shows that human beliefs follow a diffusion-drift equation in which ingrained systematic psychological biases contribute to belief drift, whereas other random factors and influences contribute to diffusion. The diffusion-drift equation predicts a steady-state belief distribution in which *increased access to information production and sharing contributes to increased levels of polarization*. The extent of this effect depends on the relative strength of drift versus diffusion terms. Anecdotal empirical evidence is presented that at least some societies may indeed be operating in a regime consistent with a non-trivial information-access-facilitated polarization growth. Specifically, for the US, the model accurately predicts the growing polarization of the US Congress, taking as input the technology penetration curve for mobile phones (as a proxy for democratized information access and sharing) in the last 25 years.

The work is motivated by the historic change in information access patterns in the 21st century. Over the course of most of human history, information *broadcast* has been prohibitively expensive. It required significant investments (e.g., having a radio station or a publishing house). With the invention of the Internet, the barrier to making content available for potentially global consumption was significantly reduced. We say that "information broadcast" (both access and sharing) has become *democratized*. While the benefits of democratizing information broadcast are undeniable, it is interesting to model the impact of this change on societal polarization (as such models are a prerequisite to the design of proper mitigation policies for any undesirable side effects).

The idea that increased access can facilitate polarization is not new. For example, evidence suggests that the interstate highway system in the US may have contributed to socio-

Ideological polarization in the US congress



Understanding impact of messages on beliefs:

Message → Message Embedding

Message Embedding (of consumed messages)

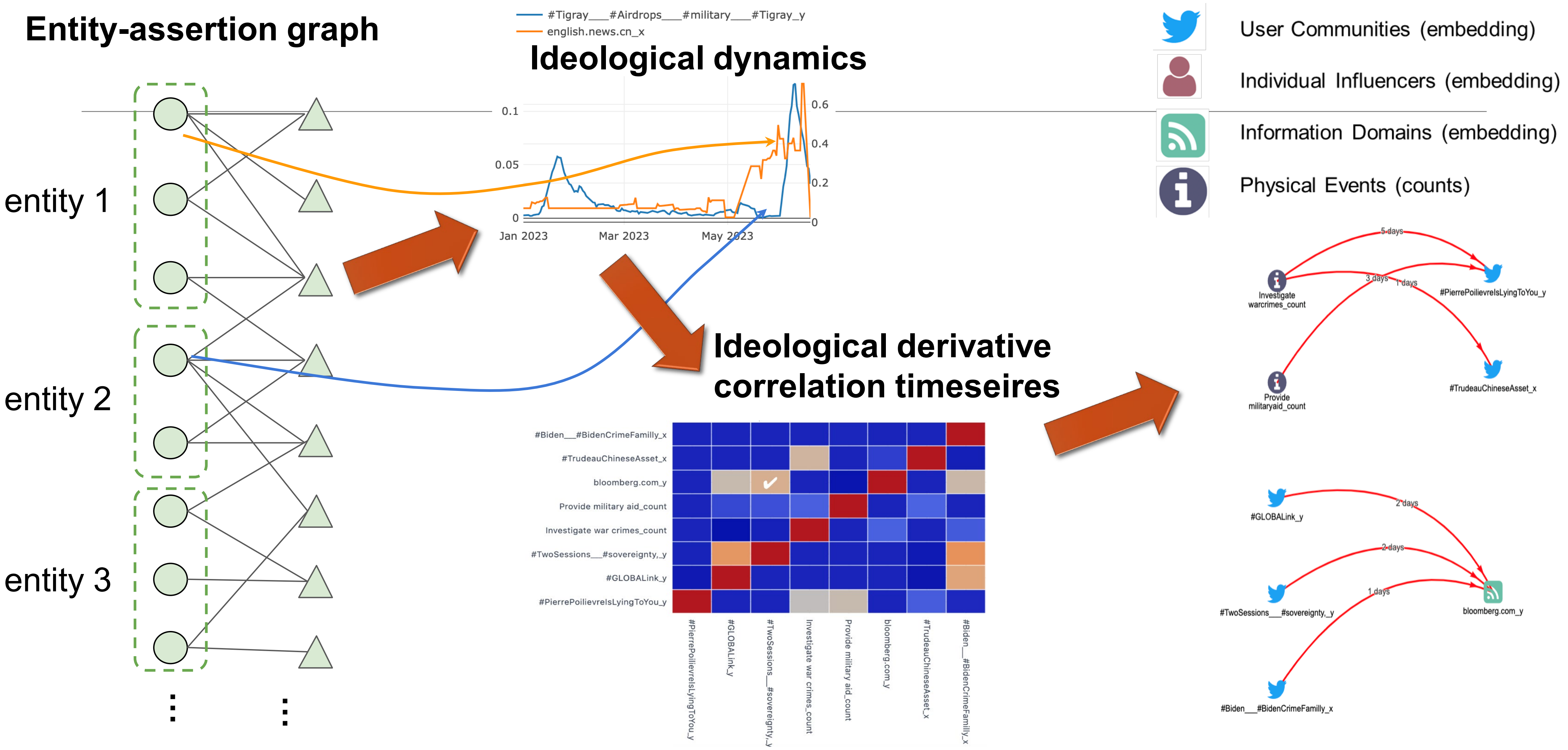
→ Actor Embedding

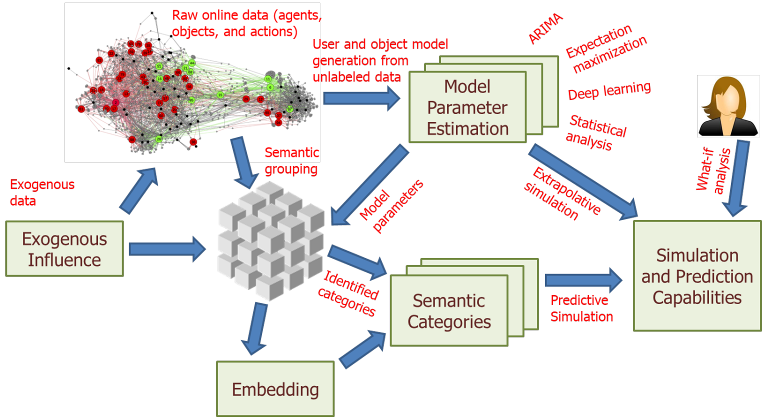
Actor Embedding (+ Interactions)

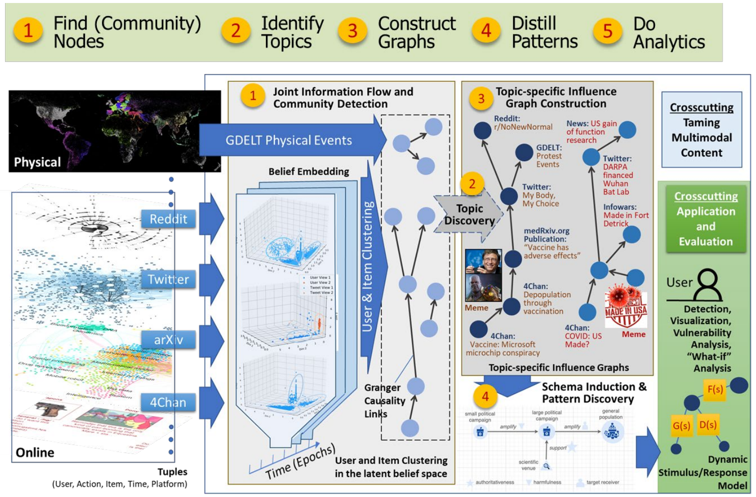
→ Next step Actor Embedding

Application: Influence Pathway Discovery

[19] Xinyi Liu, Ruijie Wang, Dachun Sun, Jinning Li, Christina Youn, You Lyu, Jianyuan Zhan, Dayou Wu, Xinhe Xu, Mingjun Liu, Xinshuo Lei, Zhihao Xu, Yutong Zhang, Zehao Li, Qikai Yang and Tarek Abdelzaher, “Influence Mapping on Social Media based on Interpretable Ideological Embedding,” In Proc. *9th International Conference on Collaboration and Internet Computing (IEEE CIC)*, Atlanta, GA, Nov 2023.







DARPA INCAS

- Characterize population response to information campaigns
- Segment populations by observed response to persuasion, and correlate persuasion tactics with population segment attributes

DARPA SocialSim

- Multiscale modeling and simulation techniques for online information propagation and belief dynamics
- Decoupling of macroscopic and microscopic models (e.g., detailed cascade models versus aggregate trends)

DARPA MIPs

- Develop a toolkit for the discovery, visualization, and analysis of influence pathways in the information space.
- Develop "what-if" capabilities for intervention modeling

Conclusions

The recent AI/ML revolution is a key opportunity for real-time computing!

- *We specialize in managing bottleneck computing resources*.
  - → AI/ML is creating the world's largest computing bottleneck!
    - Exploit latency/quality tradeoffs in computing and communication
    - Prioritize data processing (i.e., attention scheduling) to meet latency constraints
    - Derive spatial-temporal real-time attention bounds
    - Explore the impact of thermal control, DVFS, etc.
- *We specialize in embedded computing*
  - → Embodied AI is embedded AI
    - Learning from Sensor Data (in frequency domain, multimodal, harmonic structure, …)

AI + RT/Embedded collaborations could bring a wealth of new perspectives and applications