Efficient Machine Learning at the Edge in Parallel

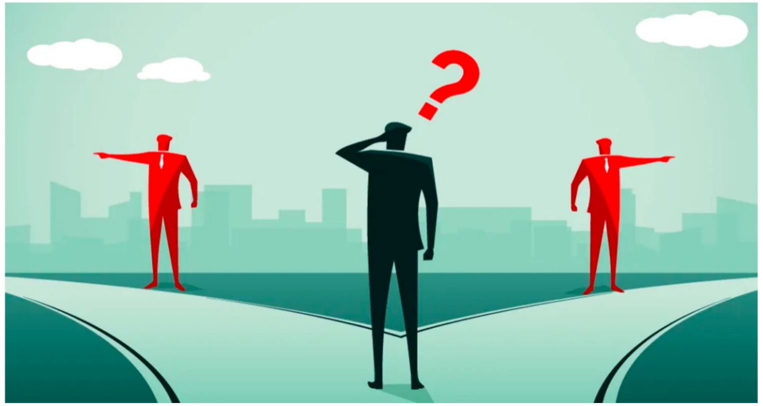
Furong Huang

University of Maryland

furongh@umd.edu

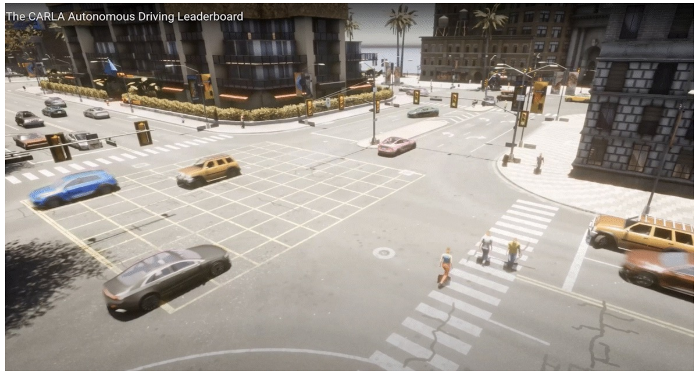
https://furong-huang.com

Real-time Machine-Augmented Intelligence



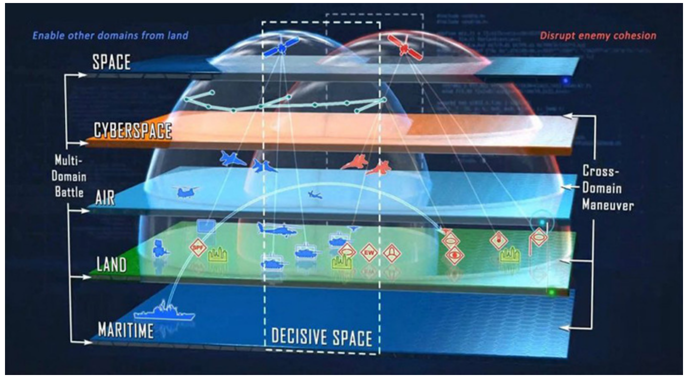
Multisource information solicitation

Highly dynamic and massive data streams



Driving at traffic junction

Healthcare & medicine



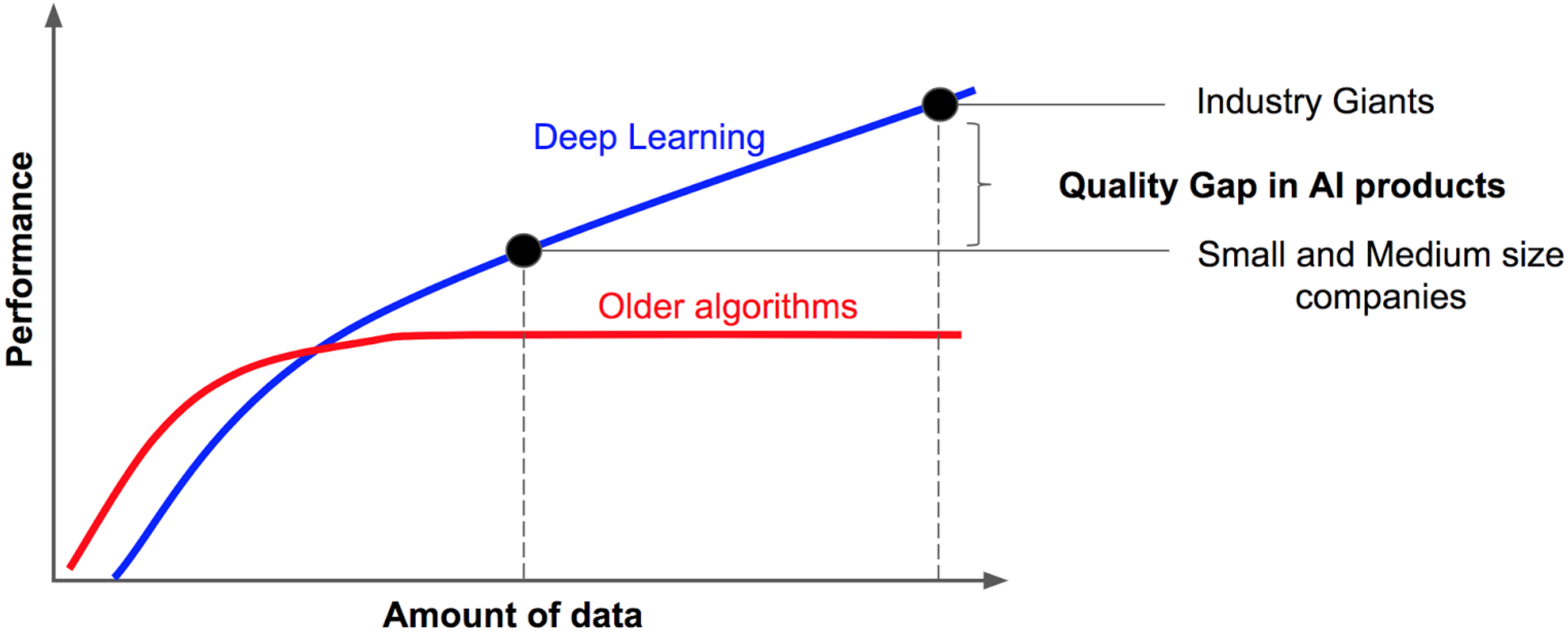
Command & control in battle space

Augment Intelligence for Efficient *Decision-Making*

Challenges in Decision-Making

- Low learning efficiency

  Training a decision-maker takes tons of samples and computations

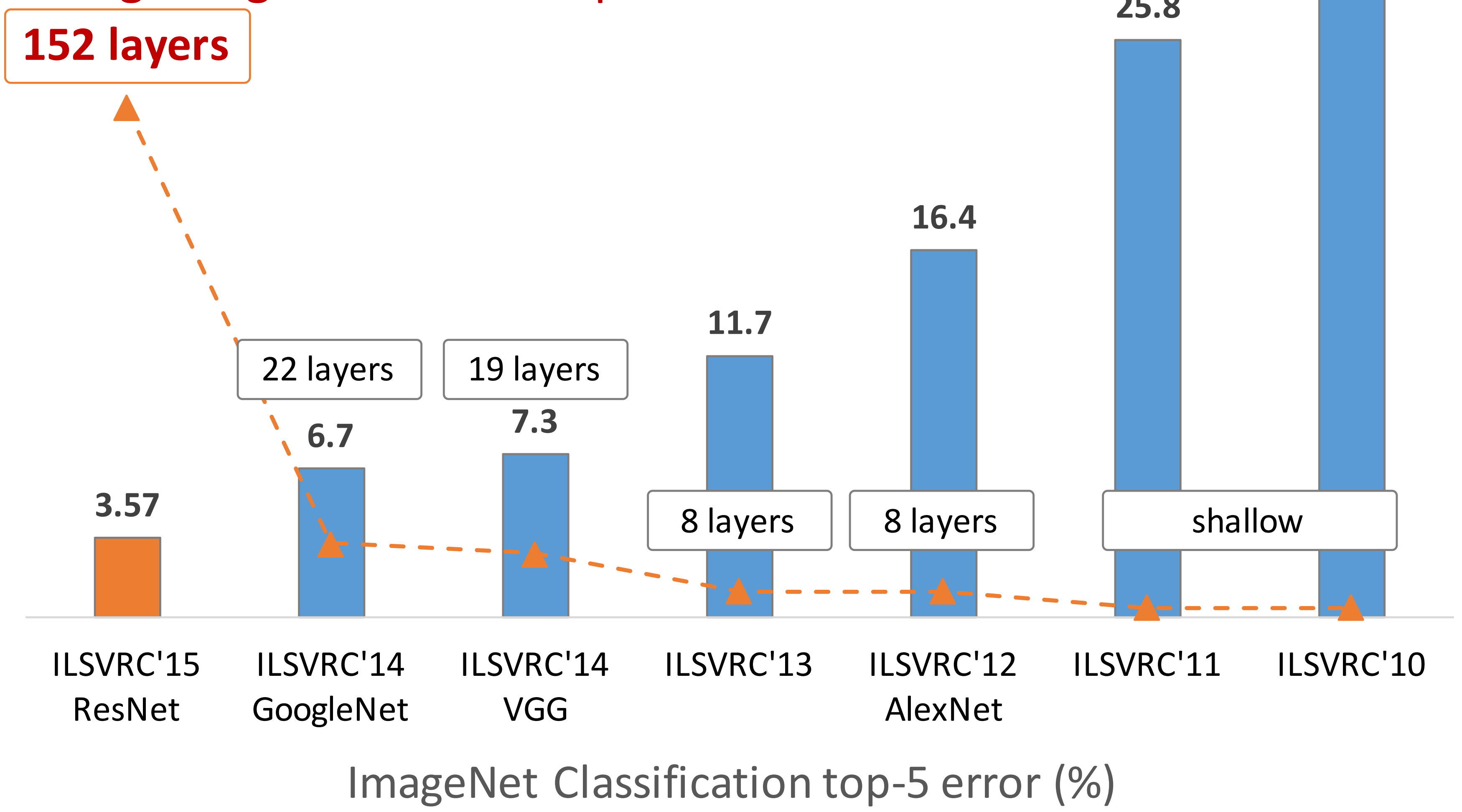


"Scale drives deep learning progress" by Andrew Ng

Challenges in Decision-Making

- Low model efficiency

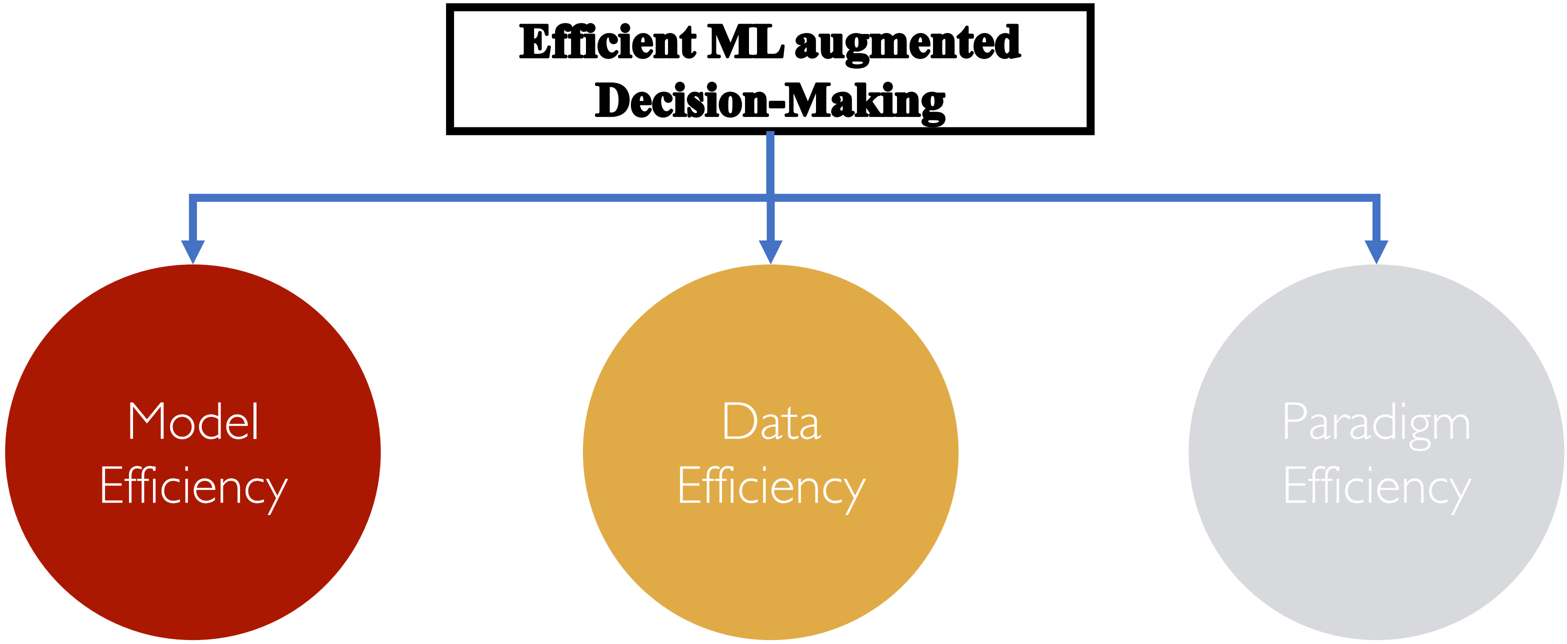
Models are getting wider and deeper



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016.

Learning to Scale

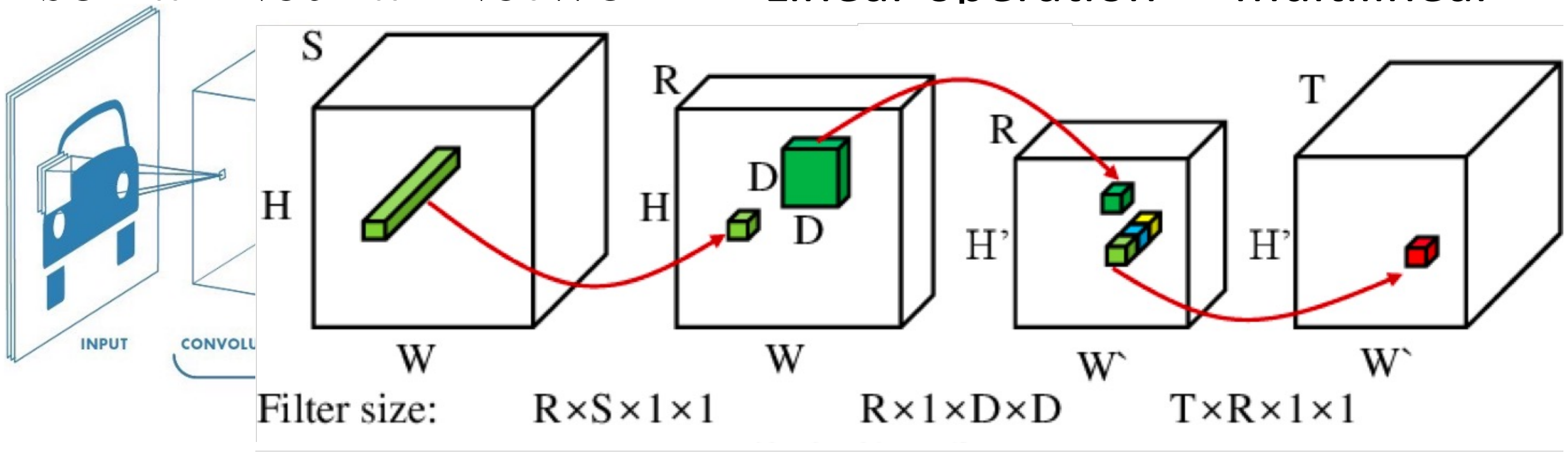
[DRAWH,NeurIPS’22]
[BSBEHGG, NeurIPS’22]
[ZYLTH, ICLR’22]
[LSH,ICLR’22]
[RFYRVH,AAAI’22]
[SLLRCTH,Frontiers’22]
[SZHLFGH, ICML’22]
[WWH,ICML’22]
[ZAH,NeurIPS’21]
[DKLZDHG,NeurIPS’21]
[SWH, NeurIPS’20]
[SBKHKA,NeurIPS’20]
[SCH,ICLR’20]
[LSSSH,AISTATS’20]
[LSLSH,ICML’19]
[HUPCSA,UAI’19]
[HALS,ICML’18]



Model Efficiency via Network Model Design and Interpretation

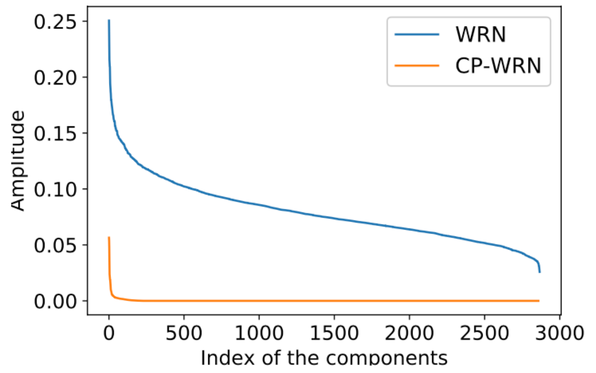
Tensorial Neural Network

Linear operation → multilinear



- Tensor factorized form inspired neural network
- Model compression via tensor representation

Generalization Improvement through the lens of Compression



CP Layer exhibits "Low Rankness"

Main Theorem: Generalization Error Bound

To achieve γ compression on sample S

$\tilde{O}\left(\sum_{k=1}^{n} \hat{R}^{(k)}\right)$ number of parameters is required to achieve γ compression on sample S

$$L_0(g) \leq \hat{L}_\gamma(f) + \tilde{O}\left(\sqrt{\frac{\sum_{k=1}^{n} \hat{R}^{(k)}}{m}}\right)$$

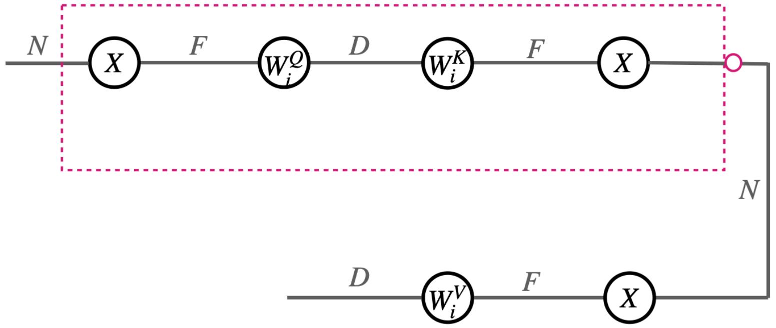
Personalized ML, Federated learning **in edge devices**

Su, Li, Liu, Ranadive, Coley, Tuan, H., "Compact Neural Architecture Designs by Tensor Representations", Frontiers 2022.
Li, Sun, Su, Suzuki, H., Understanding Generalization in Deep Learning via Tensor Methods. AISTATS 2020.

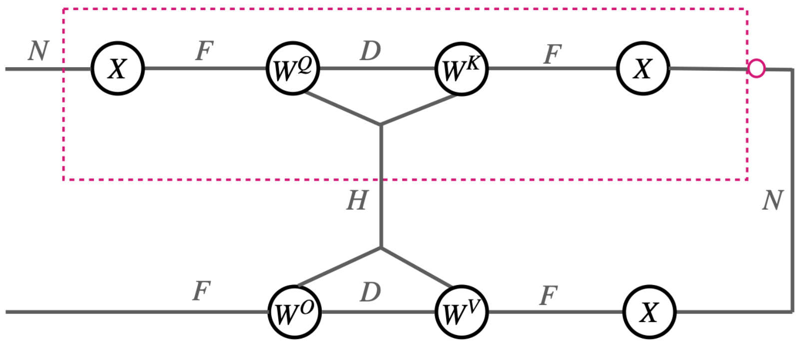
Model Efficiency via Network Model Design and Interpretation

Interpret & Improve Multi-Head Self-Attention in Transformers

A Rigorous Visual Interpretation of Self-attention



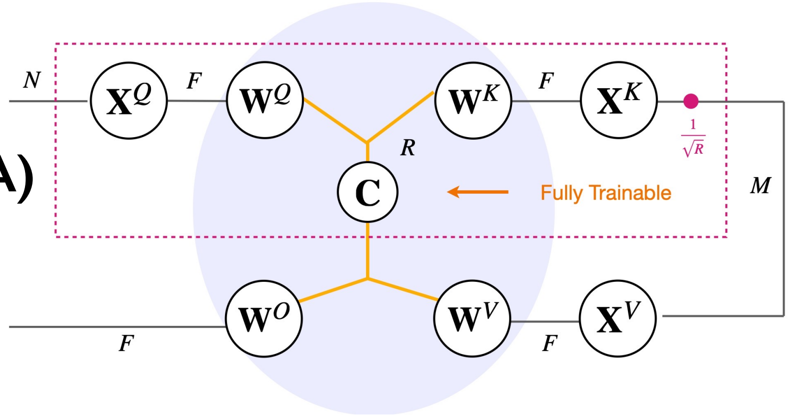
Single-head Tensor Diagram



Multi-head Tensor Diagram

New Architecture

Tunable-Head Self-Attention (THSA)

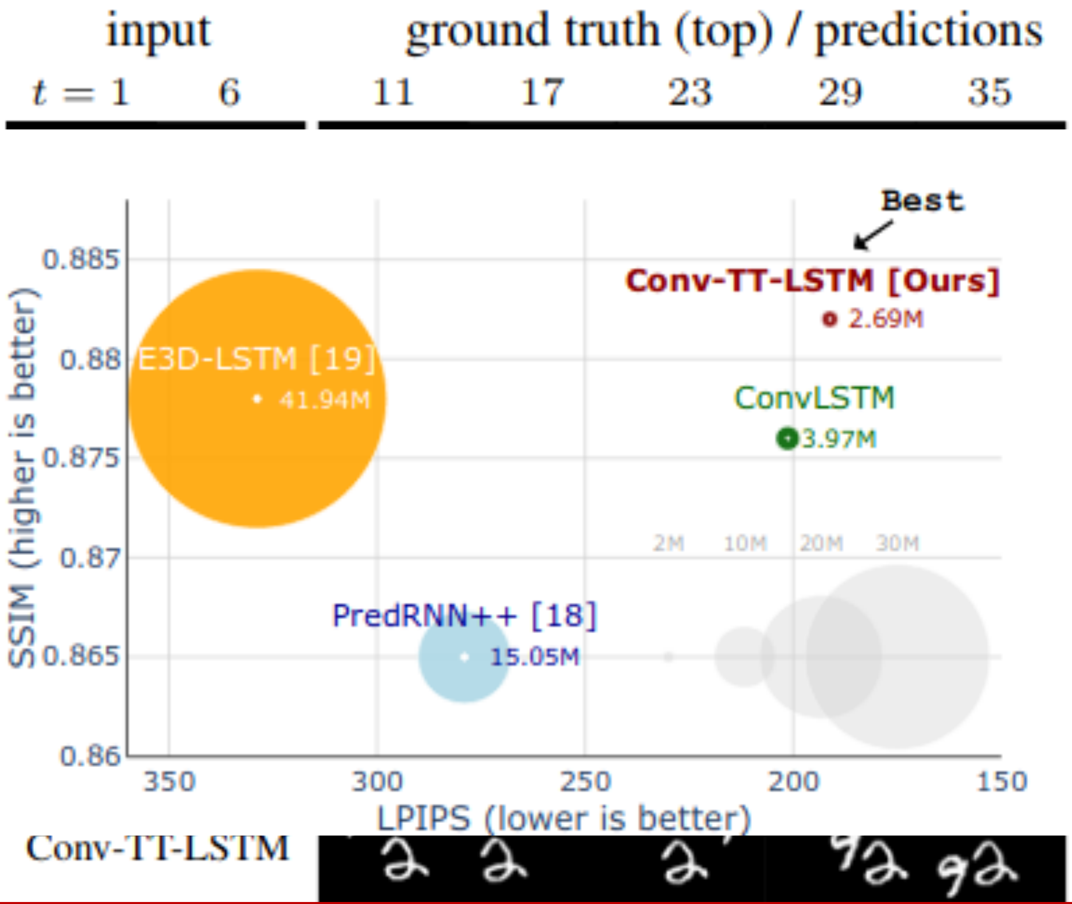


Provably Guaranteed Higher Expressive Power Under Same Size

Liu, Su, H., Tuformer: Data-driven Design of Transformers for Improved Generalization or Efficiency, ICLR 2022.

Model Efficiency via Network Model Design and Interpretation

Long-Term Video prediction (10 -> 30 frames):
predict the future based on spatiotemporal correlations.



Highest performance with fewest parameters.

Image Classification:
On CIFAR 10 Resnet-32 (460K parameters)

Compression Rate	Performance
Original	93.20%
10%	91.28%
5%	89.86%
2%	85.70%

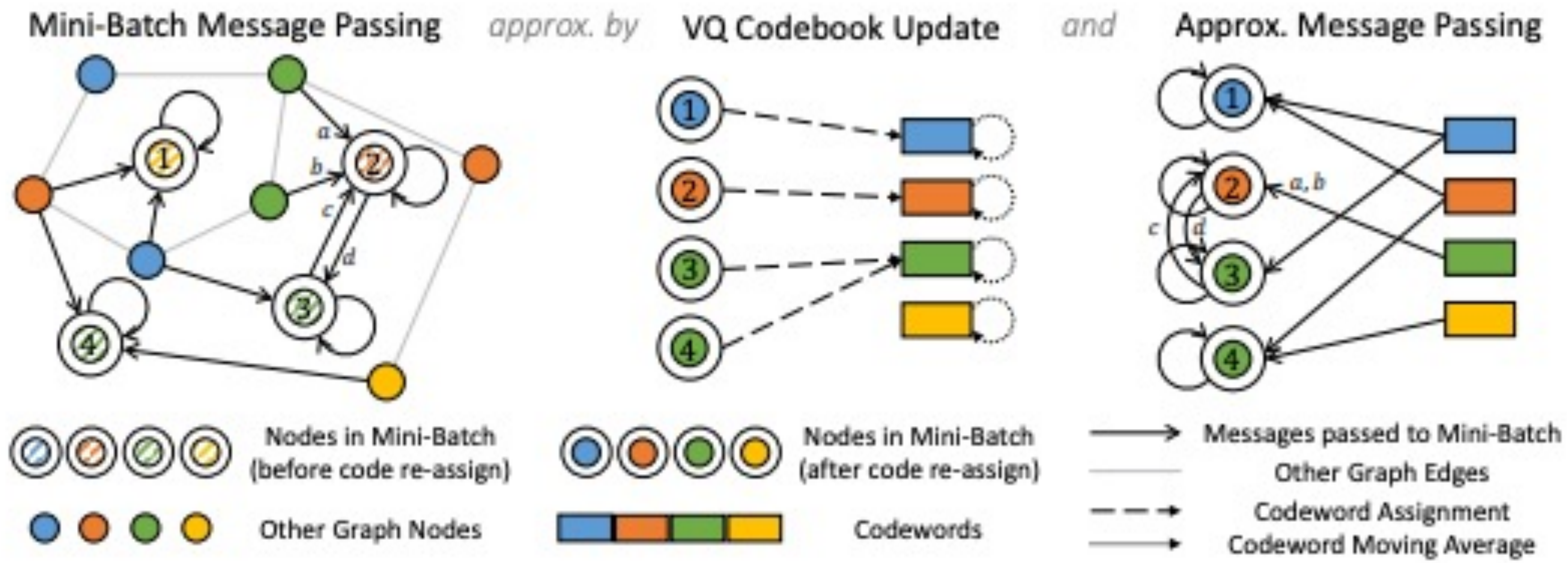
High performance small models

Su, Wang and H., ARMA Nets: Expanding Receptive Field for Dense Prediction, NeurIPS 2020.

Su, Byeon, Kossaifi, H., Kautz, Anandkumar, Convolutional Tensor-Train LSTM for Spatio-Temporal Learning, NeurIPS 2020.

Model Efficiency via Network Model Design and Interpretation

Scalable Graph Neural Networks



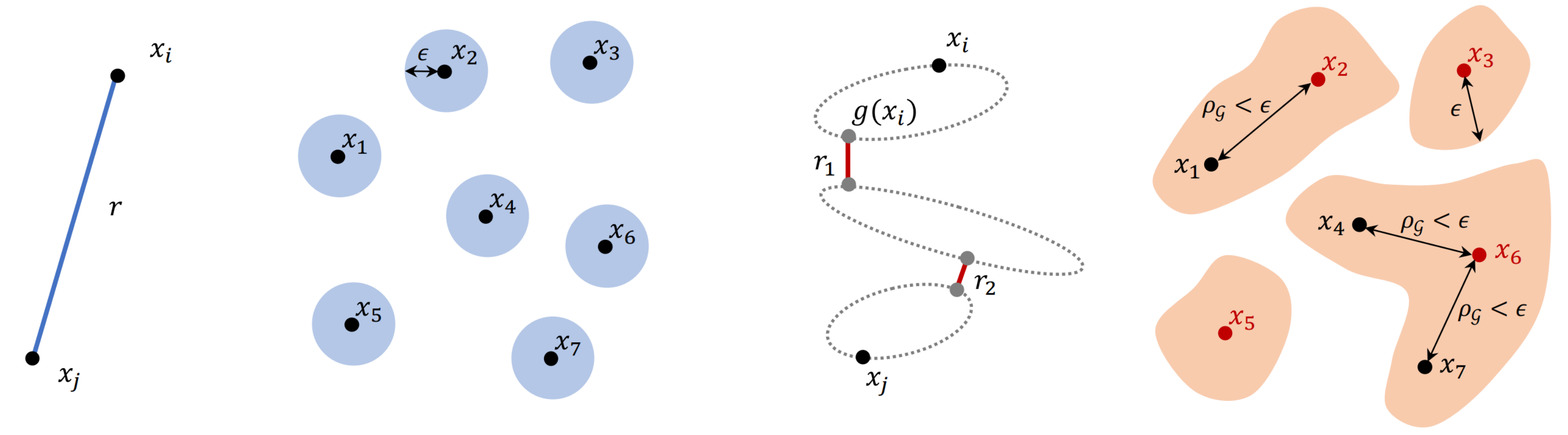
VQ-GNN, a universal framework to **scale up** any GNNs via **Vector Quantization** w/o compromising the performance

Sketch-GNN: a **sublinear complexity** training framework via **Polynomial Tensor-Sketch theory** for sketching non-linear activations and graph convolution matrices in GNNs

Ding, Kong, Li, Zhu, Dickerson, H., Goldstein, VQ-GNN: A Universal Frame- work to Scale up Graph Neural Networks using Vector Quantization, NeurIPS 2021.
Ding, Rabbani, An, Wang, H., Sketch-GNN: Scalable Graph Neural Networks with Sublinear Training Complexity, NeurIPS 2022.

Small Number of Effective Samples Covers

Theoretical Understanding of Model Invariance & Data Augmentations



$\rho_{\mathcal{G}}(x_i, x_j) = r$ sample cover = $\{x_1, x_2, \ldots, x_7\}$

(a) $\mathcal{G} = \emptyset$

$\rho_{\mathcal{G}}(x_i, x_j) = r_1 + r_2$ sample cover = $\{x_2, x_3, x_5, x_6\}$

(b) $\mathcal{G} = \{\text{"rotation"}\}$

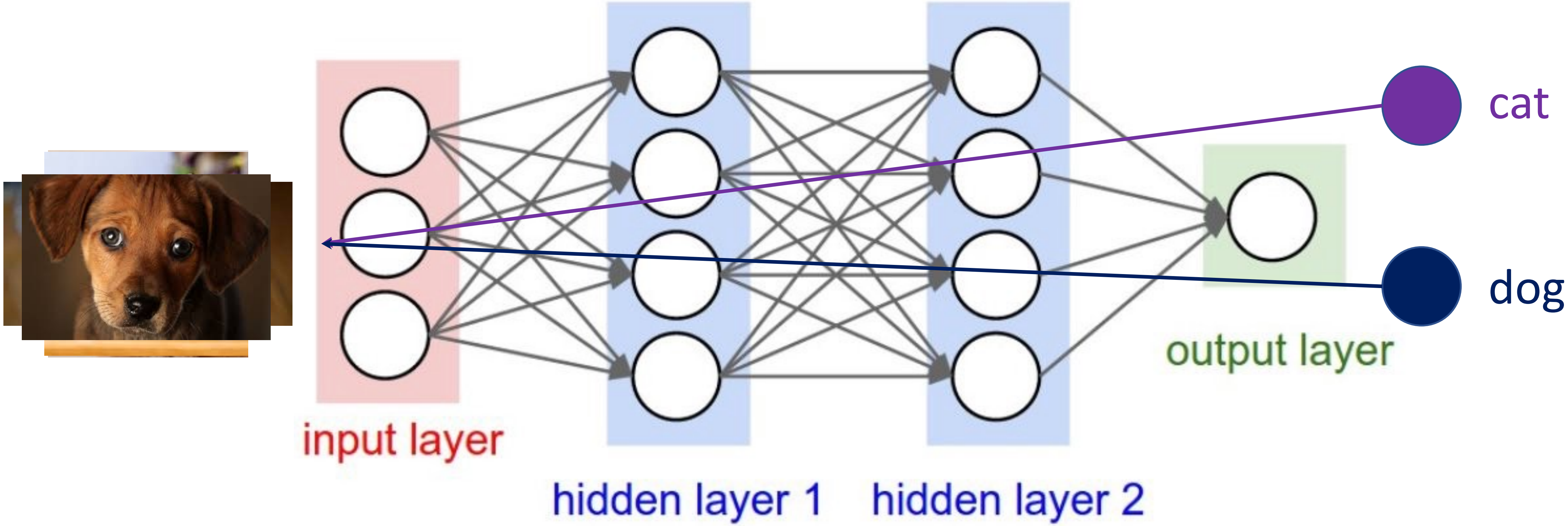
Study the **generalization benefit of model invariance** by introducing the sample cover induced by data transformations/augmentations

Zhu, An, H., Understanding the Generalization Benefit of Model Invariance from a Data Perspective, NeurIPS 2021.

Challenges in Decision-Making

- Inefficient learning paradigm

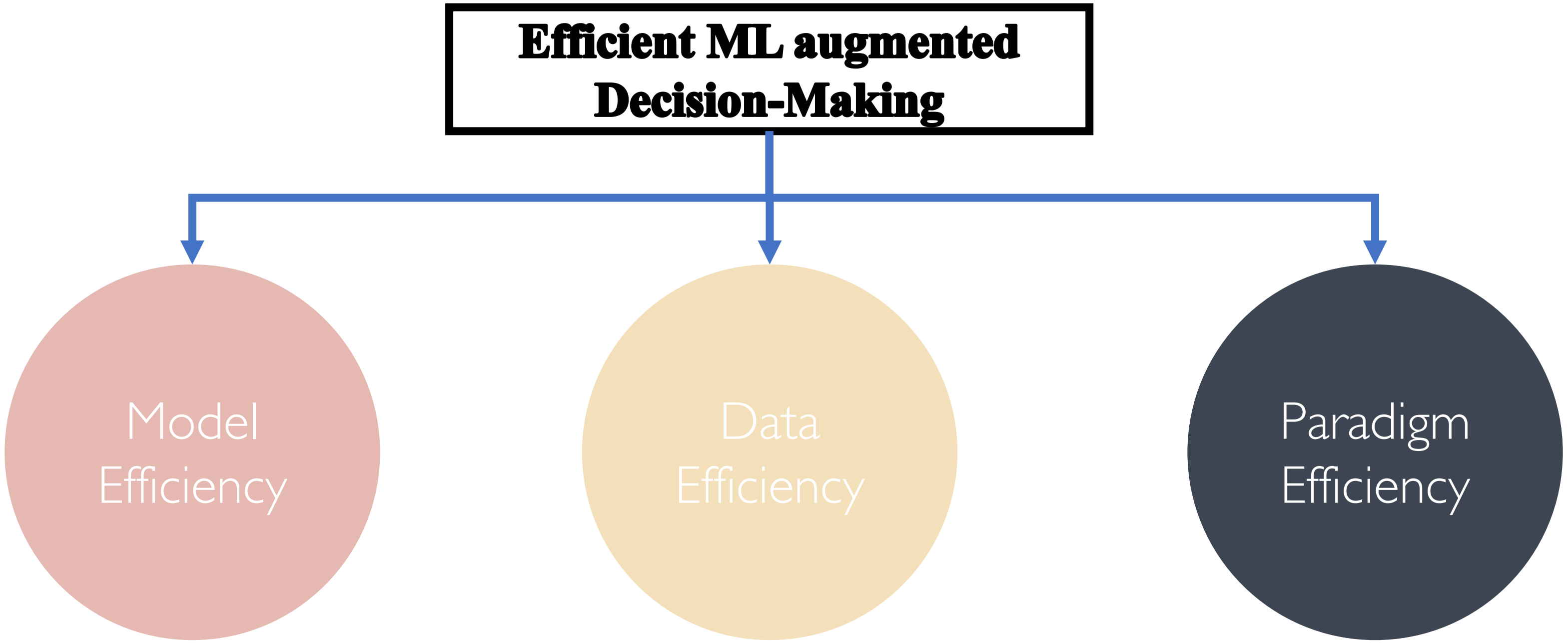
  Models are learned in a “center controller” sequentially



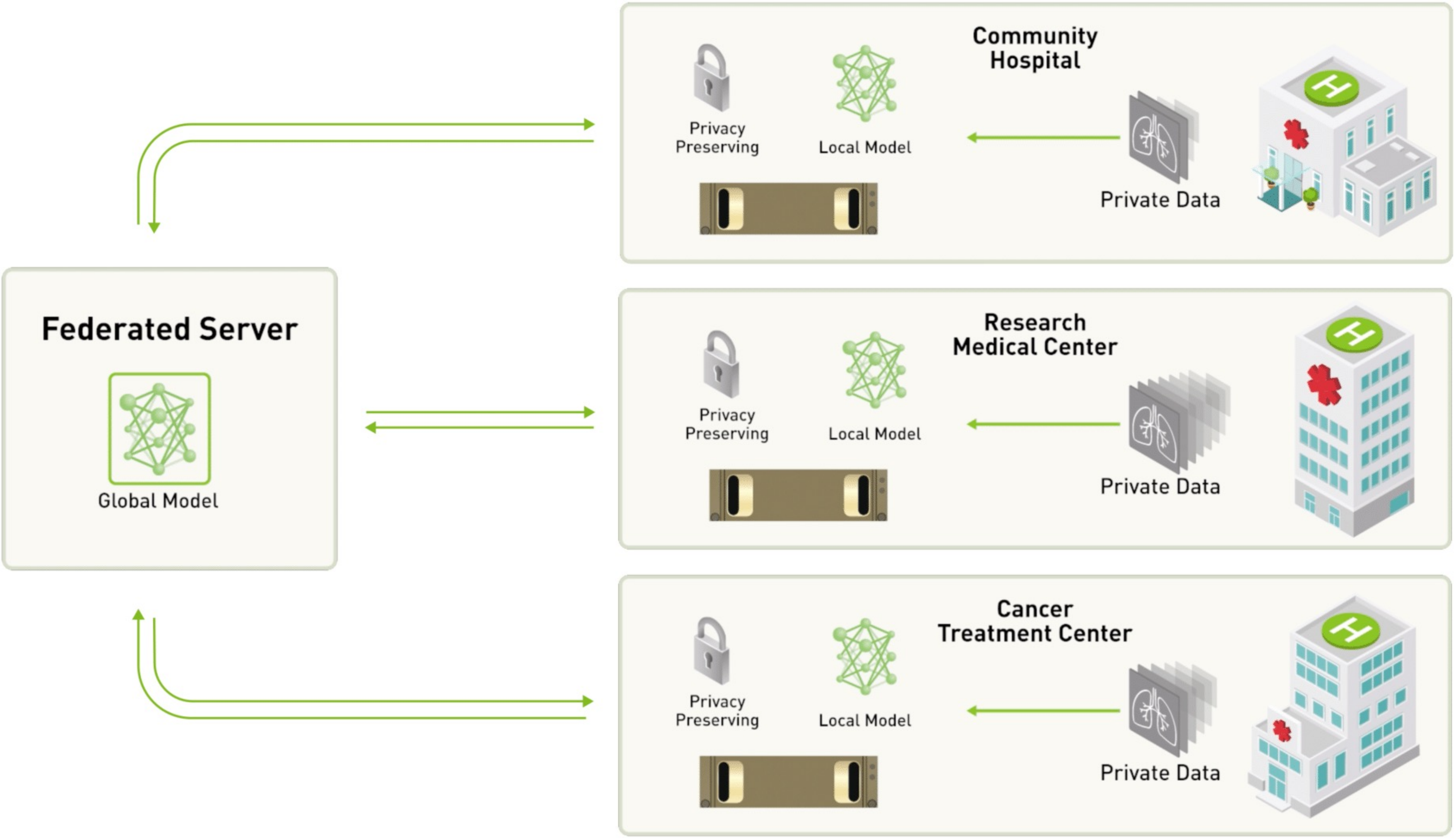
$$t \rightarrow t+1 \rightarrow t+2 \rightarrow t+3$$

Learning to Scale

[DRAWH,NeurIPS’22]
[BSBEHGG, NeurIPS’22]
[ZYLTH, ICLR’22]
[LSH,ICLR’22]
[RFYRVH,AAAI’22]
[SLLRCTH,Frontiers’22]
[SZHLFGH, ICML’22]
[WWH,ICML’22]
[ZAH,NeurIPS’21]
[DKLZDHG,NeurIPS’21]
[SWH, NeurIPS’20]
[SBKHKA,NeurIPS’20]
[SCH,ICLR’20]
[LSSSH,AISTATS’20]
[LSLSH,ICML’19]
[HUPCSA,UAI’19]
[HALS,ICML’18]



Centralized Federated Learning



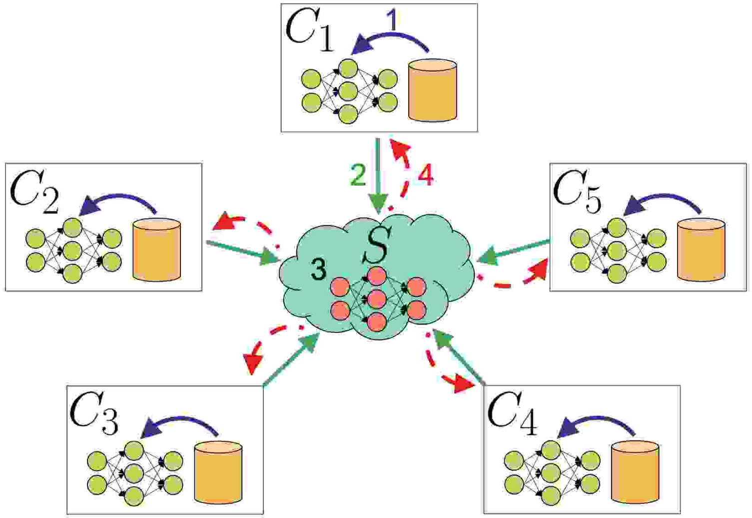
A centralized-server approach to federated learning.

Image credit: Nvidia

Challenges in **Centralized** Federated Learning

Limited Scalability

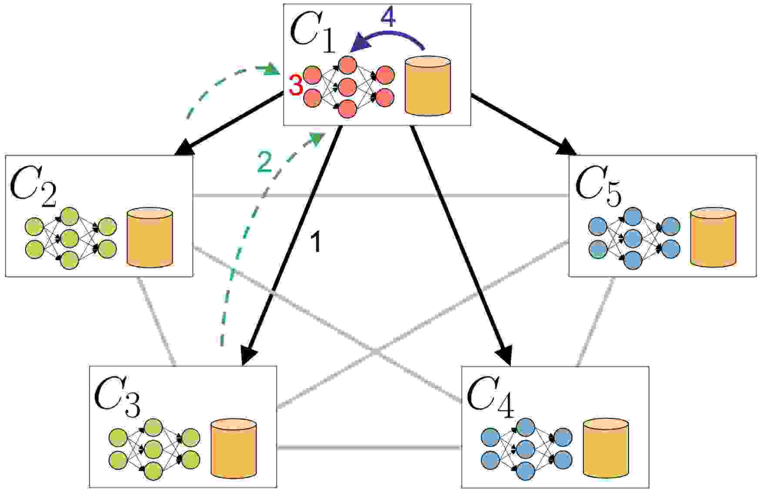
- ❑ Centralized host becomes a single point of failure
- ❑ Data-privacy breaches
- ❑ High communication latency



central host → peer-to-peer communication

Decentralized Federated Learning:

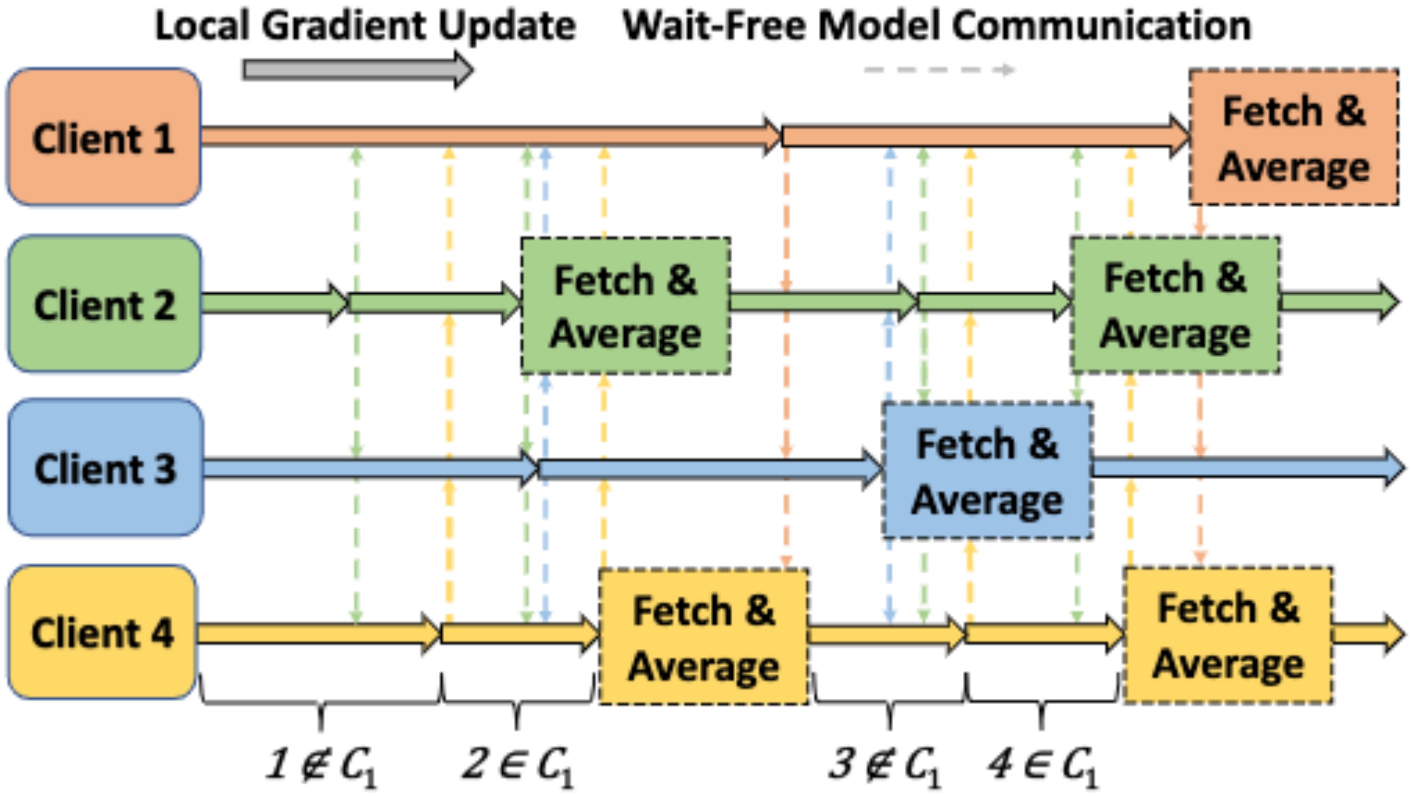
- ❑ Remove single point of failure
- ❑ Improve data privacy
- ❑ Lower communication latency?



Challenges in **Decentralized** Federated Learning

- ❑ Constructing **efficient communication protocols** amongst clients
- ❑ Ensuring the **convergence** of a global model under asynchronous updates
- ❑ Dealing with changing or sparse network topologies
- ❑ Being **robust** to deal with non-IID data between heterogeneous clients.

Shared WaIt-Free Transmission (SWIFT) Federated Learning

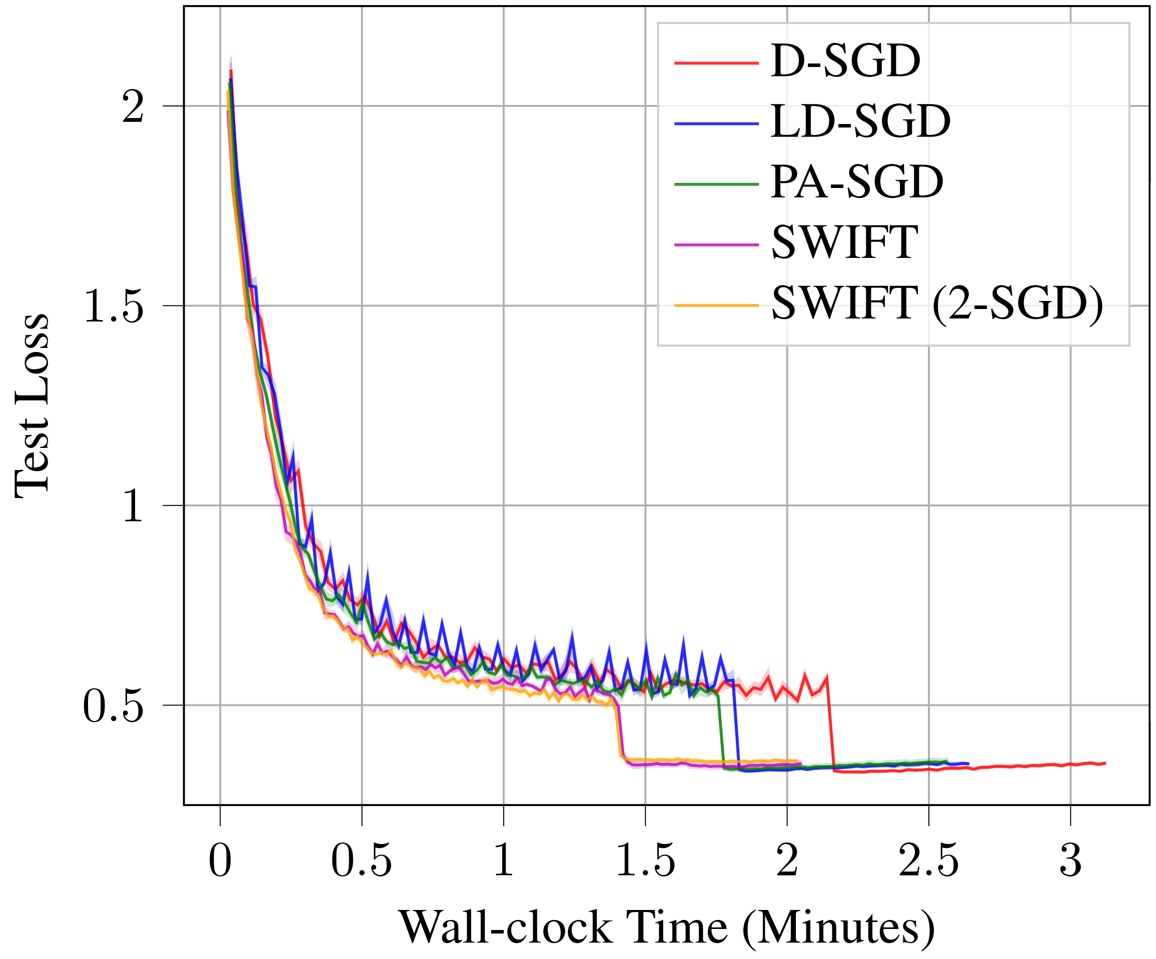


SWIFT schematic with clients communicate every 2 local updates

- ❑ Asynchronous and wait-free, SOTA communication-time complexity
- ❑ Does not require a bound on the speed of the slowest client in the network
- ❑ Golden-standard iteration convergence rate $O(1/\sqrt{T})$ of parallel SGD

Bornstein, Rabbani, Wang, Bedi, H., SWIFT: Rapid Decentralized Federated Learning via Wait-Free Model Communication

Evaluations on Real Data



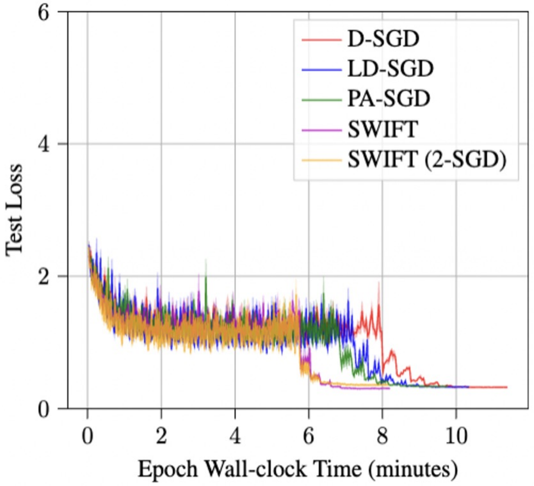
Decentralized FL Algorithms	16 Client Ring			
	Epoch (s)	% Change	Comm. (s)	% Change
SWIFT ($\mathcal{C}_0$)	**1.019**	-34.60	**0.086**	-86.28
D-SGD ($\mathcal{C}_0$)	1.558	—	0.627	—-
AD-PSGD* ($\mathcal{C}_0$)	—	-15.86	—	—
SWIFT ($\mathcal{C}_1$)	**1.016**	-34.79	**0.064**	-89.79
LD-SGD ($\mathcal{C}_1$)	1.320	-15.28	0.428	-31.74
PA-SGD ($\mathcal{C}_1$)	1.281	-17.78	0.358	-42.90

* AD-PSGD results come from Table 4 in (Lian et al., 2018).

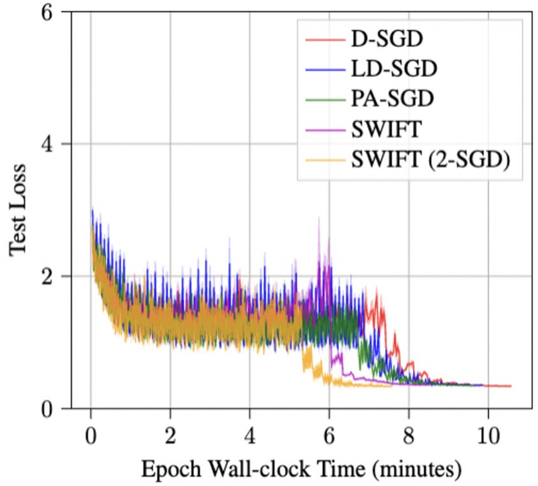
SOTA communication efficiency

SOTA convergence efficiency

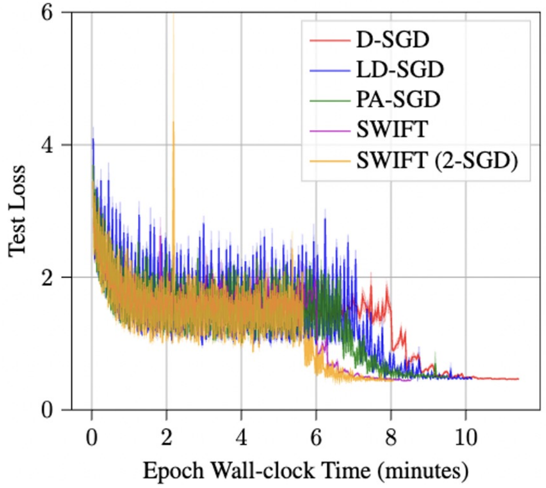
Evaluations on Real Data



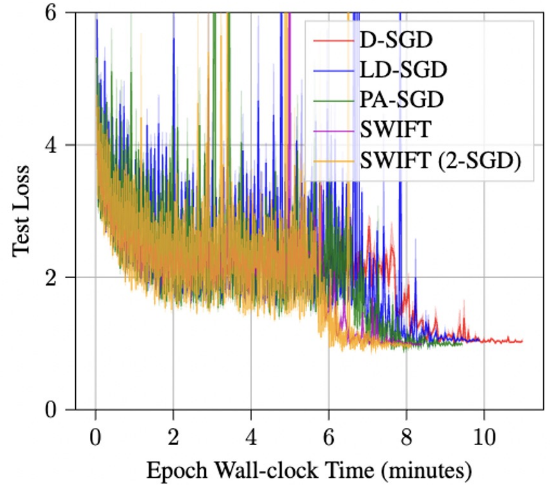
(a) 1/4 degree non-IID data.



(b) 1/2 degree non-IID data.



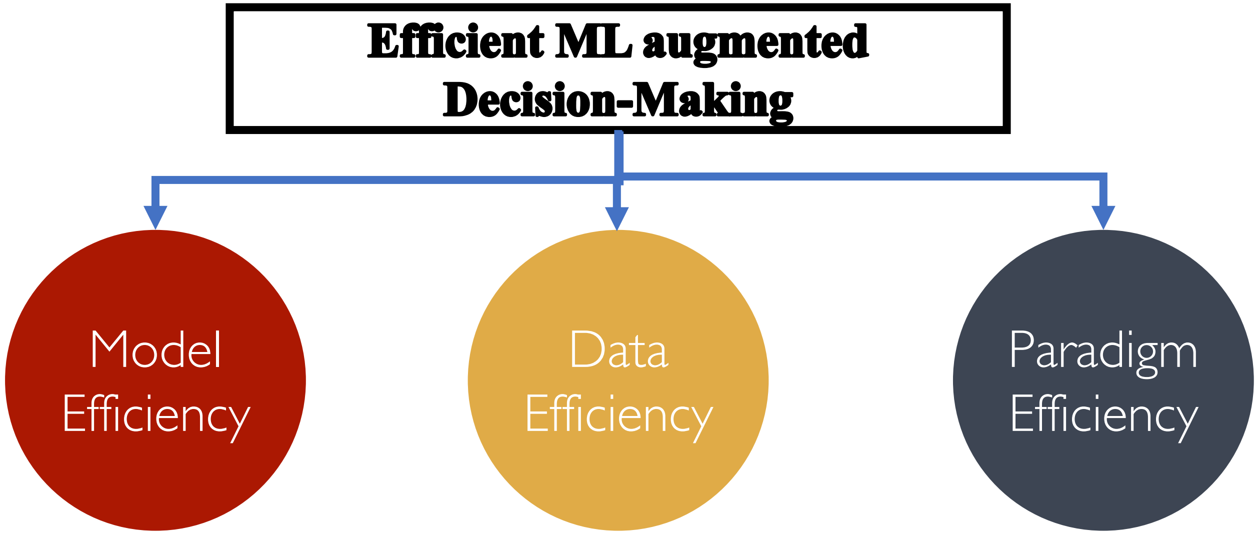
(c) 7/10 degree non-IID data.



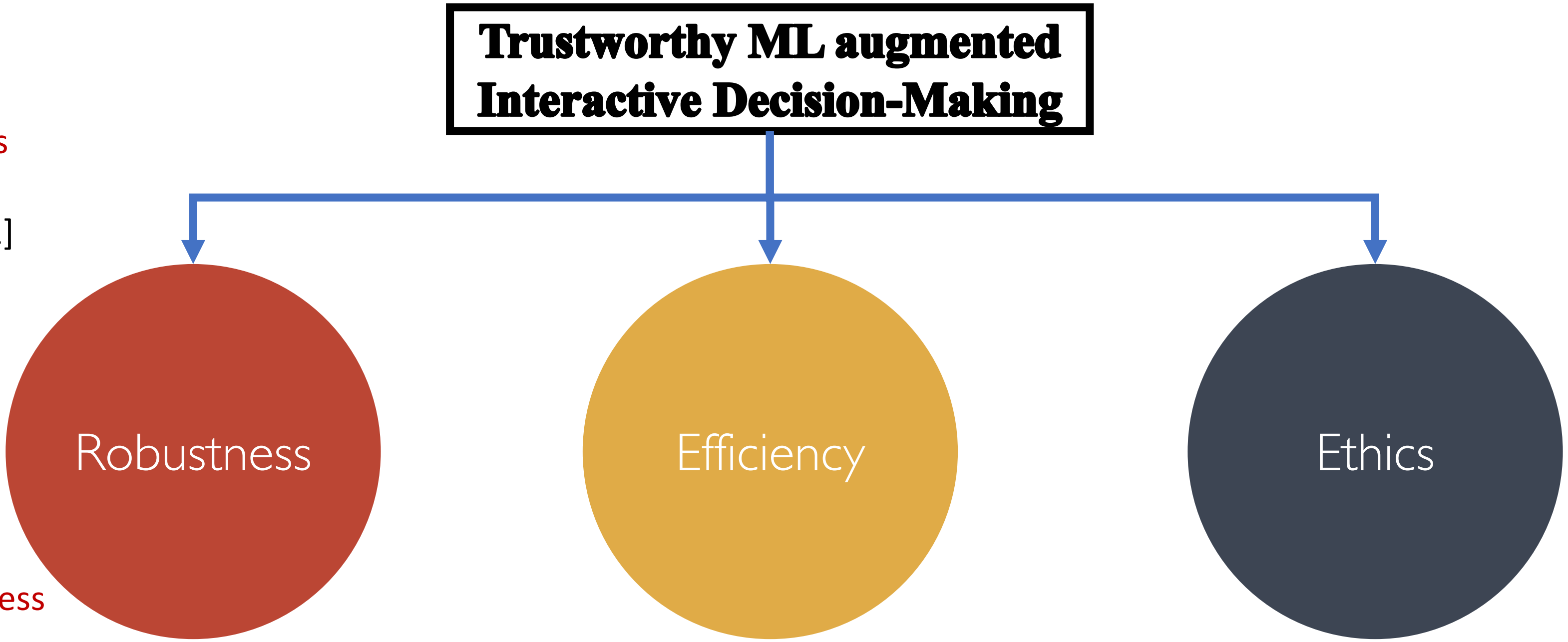
(d) 9/10 degree non-IID data.

SOTA adaptability to heterogeneous data across clients

Efficient Machine Learning in Parallel



Our Solutions to Trustworthy Decision-Making via Machine Learning



Adversarial Robustness

[LSZH,NeurIPS'22]
[YSSHTHZT, NeurIPS'22]
[LSBHGG, NeurIPS'22]
[SBH, ICML'22]
[SZHLFGH, ICML'22]
[XSH,ICML'22]
[SZLH,ICLR'22]
[SHH,ICLR'21]
[ZZGH, AAAI'21]

Distributional Robustness

[SZHLFGH, ICML'22]
[WWH,ICML'22]
[SZWCH,ICLR'22]
[SBGHVGG, NeurIPS'21]
[DKCKGWHG,NeurIPS'21]
[SYH, AAAI'21]
[SH,AAMAS'20]

Generalizability

[SZHLFGH, ICML'22]
[WWH,ICML'22]
[ZAH,NeurIPS'21]
[LSSSH,AISTATS'20]
[LSLSH,ICML'19]

[DRAWH,NeurIPS'22]
[BSBEHGG, NeurIPS'22]
[ZYLTH, ICLR'22]
[LSH,ICLR'22]
[RFYRVH,AAAI'22]
[SLLRCTH,Frontiers'22]
[DKLZDHG,NeurIPS'21]
[SWH, NeurIPS'20]
[SBKHKA,NeurIPS'20]
[SCH,ICLR'20]
[HUPCSA,UAI'19]
[HALS,ICML'18]

[ACDH,NeurIPS'22]
[DRELH,ICML'20]
[XZAAH,'22]
[HKAAH,'22]

Furong Huang
University of Maryland
furongh@umd.edu
https://furong-huang.com

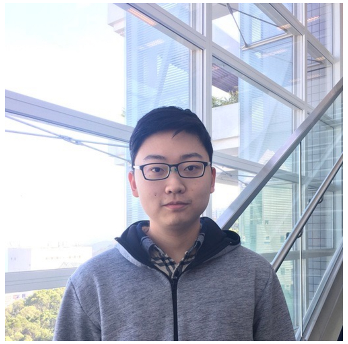
Dr Jiahao Su

Bang An

Marco Bornstein

Souradip Charkraborty

Chenghao Deng



Mucong Ding

Xiangyu Liu

Xiaoyu Liu

Tahseen Rabbani

Yanchao Sun

Xiyao Wang

Joy Wongkamjan

Yuancheng Xu

Sicheng Zhu

Frank Zheng

A Selected List of Related Work

- X. Liu, J. Su, F. Huang, "Tuformer: Data-driven Design of Transformers for Improved Generalization or Efficiency", ICLR 2022.
- J. Su, W. Byeon,, F. Huang, "Scaling-up Diverse Orthogonal Convolutional Networks with a Paraunitary Framework", ICML 2022.
- J. Su, J. Li, X. Liu, T. Ranadive, C. Coley, T.C. Tuan, F. Huang, "Compact Neural Architecture Designs by Tensor Representations", Frontiers 2022.
- S. Zhu, B. An, F. Huang, Understanding the Generalization Benefit of Model Invariance from a Data Perspective, NeurIPS 2021.
- J. Li, Y. Sun, J. Su, T. Suzuki, F. Huang, "Understanding Generalization in Deep Learning via Tensor Methods", AISTATS 2020.
- J. Su, S. Wang and F. Huang, "ARMA Nets: Expanding Receptive Field for Dense Prediction", NeurIPS 2020.
- J. Su, W. Byeon, J. Kossaifi, F. Huang, J. Kautz, A. Anandkumar, "Convolutional Tensor-Train LSTM for Spatio-Temporal Learning", NeurIPS 2020.
- A. Reustle and T. Rabbani and F. Huang, "Fast GPU Convolution for CP-Decomposed Tensorial Neural Networks", IntelliSys 2020.
- M. Ding, T. Rabbani, B. An, E. Wang, F. Huang, "Sketch-GNN: Scalable Graph Neural Networks with Sublinear Training Complexity", NeurIPS 2022.
- M. Ding, K. Kong, J. Li, C. Zhu, J. Dickerson, F. Huang, T. Goldstein, VQ-GNN: A Universal Frame- work to Scale up Graph Neural Networks using Vector Quantization, NeurIPS 2021.

An Incomplete List of Related Publications

Robust ML

- Yongyuan Liang*, Yanchao Sun*, Ruijie Zheng, Furong Huang. "Efficiently Improving the Robustness of RL Agents against Strongest Adversaries". NeurIPS 2022.

- Yanchao Sun, Ruijie Zheng, Yongyuan Liang, Furong Huang. "Who Is the Strongest Enemy? Towards Optimal and Efficient Evasion Attacks in Deep RL". NeurIPS 2021 Safe and Robust Control of Uncertain Systems Workshop (Oral, Best Paper Reward), ICLR 2022.

- Yanchao Sun, Ruijie Zheng, Xiyao Wang, Andrew Cohen, Furong Huang. "Transfer RL across Observation Feature Spaces via Model-Based Regularization". ICLR 2022.

- Zhi Zhang, Zhuoran Yang, Han Liu, Pratap Tokekar, Furong Huang. "Reinforcement Learning under a Multi-agent Predictive State Representation Model: Method and Theory". ICLR 2022.

- Yanchao Sun, Da Huo, Furong Huang."Vulnerability-Aware Poisoning Mechanism for Online RL with Unknown Dynamics". ICLR 2021.

- Yanchao Sun, Xiangyu Yin, Furong Huang. "TempLe: Learning Template of Transitions for Sample Efficient Multi-task RL". AAAI 2021.

- Huimin Zeng, Chen Zhu, Tom Goldstein, Furong Huang. "Are Adversarial Examples Created Equal? A Learnable Weighted Minimax Risk for Robustness under Non-Uniform Attacks", AAAI 2021.

- Yanchao Sun, Furong Huang. "Can Agents Learn by Analogy? An Inferable Model for PAC Reinforcement Learning". AAMAS 2020.