AI-EDGE: NG Networks meet Distributed Intelligence

Eylem Ekici

EWD Co-Director, NSF AI Institute for Future Edge Networks and Distributed Intelligence

Professor of ECE

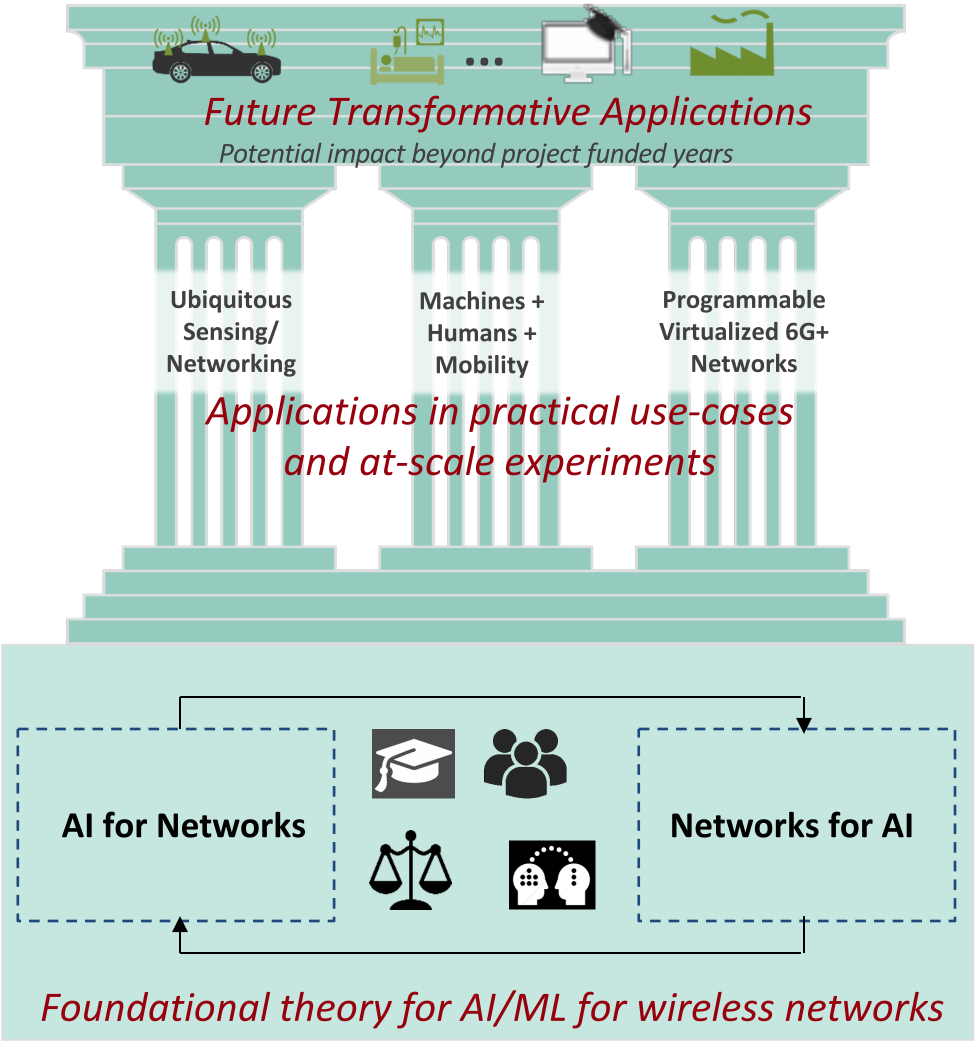
The Ohio State University

AI EDGE Institute

Outline

- **Institute Vision**
- Organization/Key Personnel
- Overview and Rationale
- Research Plan
- Overview of Thrusts
    - Example Research Challenges
- Synergies with Industry/DoD

Institute Vision



To create a research, education, knowledge transfer, and workforce development environment that will help develop research leadership in future generation edge networks (6G and beyond) and distributed AI for many decades to come

Organization and Key Personnel: Academia

PI: Ness Shroff

Expertise: Net. Theory, Bandits, RL, Optimization, Algorithms, MDP, Games

Co-PI: Elisa Bertino

Expertise: Information Security, Database, Privacy and Trust

Co-PI: Gauri Joshi

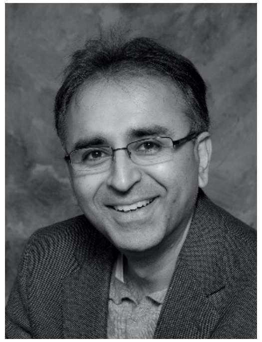
Expertise: Distributed Learning, Bandits, Bayesian Optimization

Co-PI: Jim Kurose

Expertise: Computer Network Arch. & Protocols, Network Measurements

Co-PI: Rob Nowak

Expertise: Machine Learning, Stat. Signal Processing, Statistics



SP: Anish Arora

Expertise: Network Systems Scalability & Dependability

SP: Kaushik Chowdhury

Expertise: Network Systems, 5G, Protocols, Experiments At-Scale

SP: Mingyan Liu

Expertise: Net. Resource Allocation, Sequential Decision Theory

SP: Sanjay Shakkottai

Expertise: Net. Optimization, Stat. Learning and Wireless Communication

SP: Ameet Talwalkar

Expertise: Stat. Learning, Democratize Machine Learning, Fed. Learning

SP: Raef Bassily

Expertise: Privacy-Preserving Data Analysis, ML, Optimization, Info. Theory

SP: Constantine Caramanis

Expertise: Decision Making in Complex Systems, High Dim. Statistics, Optimization

SP: Eylem Ekici

Expertise: Cognitive Radio, Vehicular Communication, Net. Resource Management

SP: Atilla Eryilmaz

Expertise: Stochastic Network Optimization, Bandits, Control

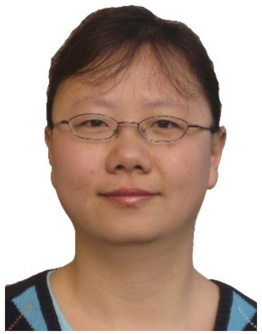
SP: Stratis Ioannidis

Expertise: Distributed Systems, Networking, Optimization, ML, Privacy

Organization and Key Personnel: Academia

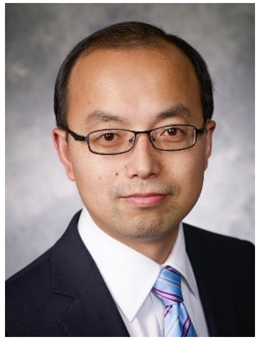
SP: Nan Jiang

Expertise: Reinforcement Learning, Online Learning



SP: Yingbin Liang

Expertise: Info. Theory, Wireless Communications, Optimization, Statistical SP



SP: Zhiqiang Lin

Expertise: Security, Trusted Computing, Program Analysis,

SP: Jia (Kevin) Liu

Expertise: ML, Distri. Optimization, Stochastic Network Optimization,

SP: Tommaso Melodia

Expertise: Wireless Networks, Cognitive Radio Experiments at Scale

SP: Aryan Mokhtari

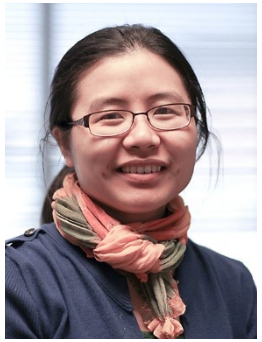
Expertise: Convex and Non-convex Optimization, Large-scale ML & Data Science

SP: Sewoong Oh

Expertise: Theoretical ML, Robust Statistics, Social Comp., Diff. Privacy

SP: Srini Parthasarathy

Expertise: Data Analytics, Graph Analytics, Network Science ML, Database Systems



SP: Chunyi Peng

Expertise: Mobile Networking Systems, Security, 5G, 6G Systems

SP: Hulya Seferoglu

Expertise: Coded Comp., IoT, Anomaly Detection in Video Streaming

SP: Kannan Srinivasan

Expertise: WirelessSys, Protocols, Measurements, Communication Security

SP: Aylin Yener

Expertise: Info. Theory, Cybersecurity, Wireless Comm., Optimization, Learning

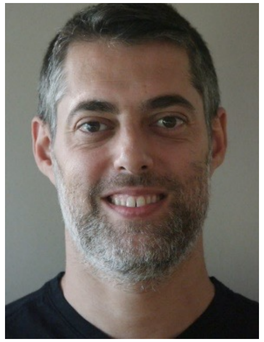
SP: Lei Ying

Expertise: Complex Stochastic Systems, Big Data, Graph Data Mining

Organization and Key Personnel: Industry/DoD

Microsoft: Victor Bahl

Expertise: Edge Comp., 5G, Mobile Computing, Wireless Sys., Cloud Comp.



IBM: Lior Horesh

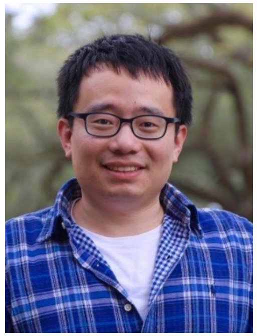
Expertise: Optimization, Applied Inverse Problems, Large-Scale Simulations, ML

NRL: Sastry Kompella

Expertise: Network Optimization, Scheduling, Cognitive Radio, ML, AoI

Qualcomm: Junyi Li

Expertise: Wireless Communication, Mobile Broadband, OFDMA



IBM: Songtao Lu

Expertise: ML, Distr. Optimization, Stat. Signal Processing, Networking

AT&T: Milap Majmundar

Expertise: Mobile Netw., Radio Access Network, Spectrum Strategy

AFRL: Chris Myers

Expertise: Computational Cognitive Models for Complex Tasks

AFRL: Lee Seversky

Expertise: Autonomy, Command & Control Systems

IBM: Mark Squillante

Expertise: Mathematical Foundation of Complex Sys. Modeling and Analysis

ARL: Anathram Swami

Expertise: Network Science, Signal Processing, Wireless Communications

Mitre: Venki Ramaswamy

Expertise: Cellular Nets, 5G Mobile, Blockchains, AI/ML

Organization and Key Personnel

NSF AI Institute Ecosystem

- Internship Opportunities
- Industry Supported Research

- Annual Retreat Site

AI Institute for Networks

- Government Supported Research
- Internship Opportunities

Use-Inspired Research Partners

- **Ubiquitous Sens. & Networking**
  - IBM
  - AT&T
  - Qualcomm
  - ARL
- **Machine + Human + Mobility**
  - AT&T
  - Microsoft Research
  - AFRL
  - ARL
- **Programmable / Virtualized 6G+**
  - Qualcomm
  - Microsoft Research
  - AFRL
  - NRL

Connections

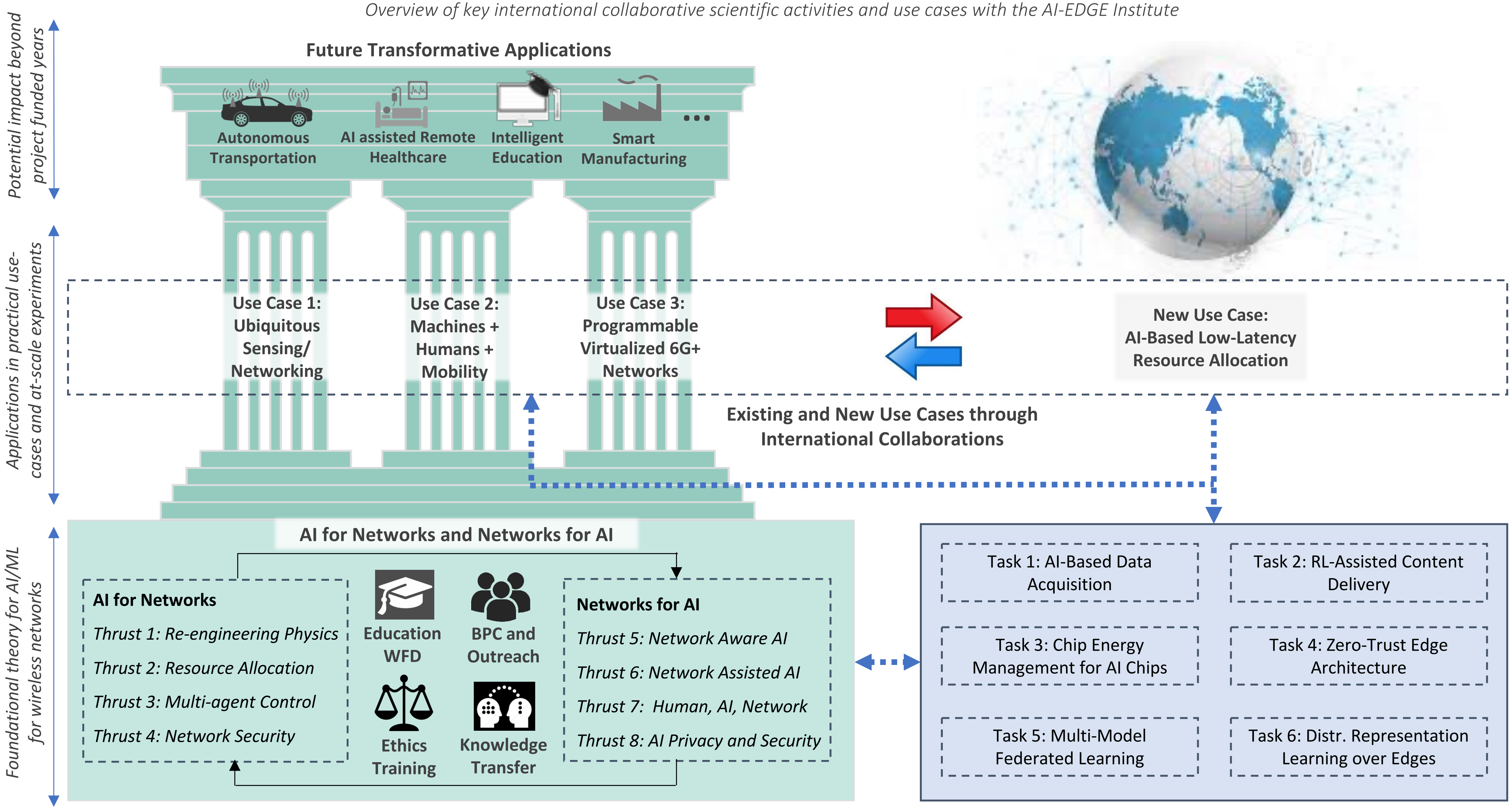
- Collaborative Research Projects
- Knowledge Transfer
- Education & Workforce Development
- Broadening Participation Efforts
- Knowledge Transfer Efforts

- Shared Staffing
- Institute Supported Research
- Good Systems-led Ethics Training

AI-EDGE goes Global

AI-EDGE Institute International Collaborations

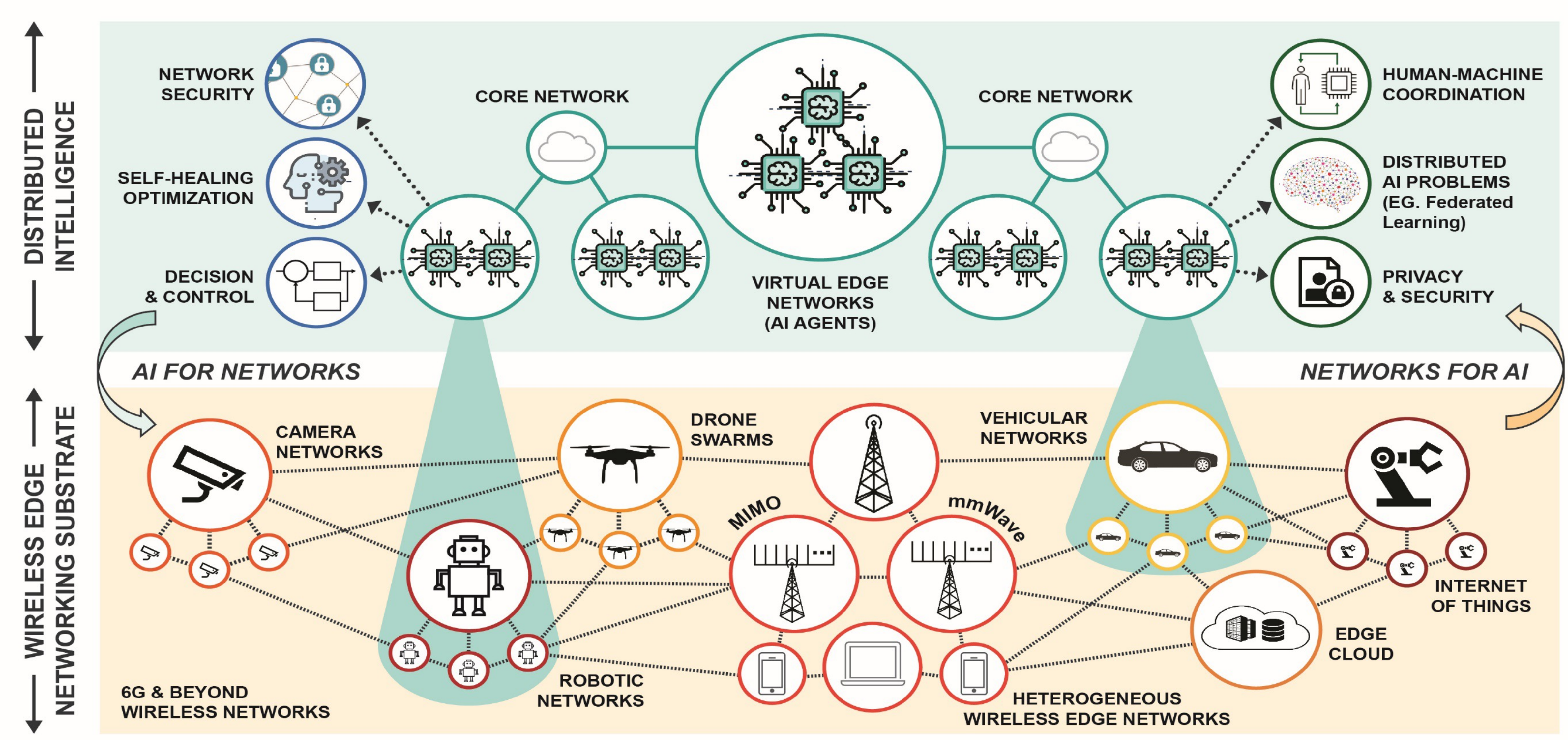
Overview of key international collaborative scientific activities and use cases with the AI-EDGE Institute



AI-EDGE's Current Research Thrusts

New Proposed Tasks through International Collaboration

Scope

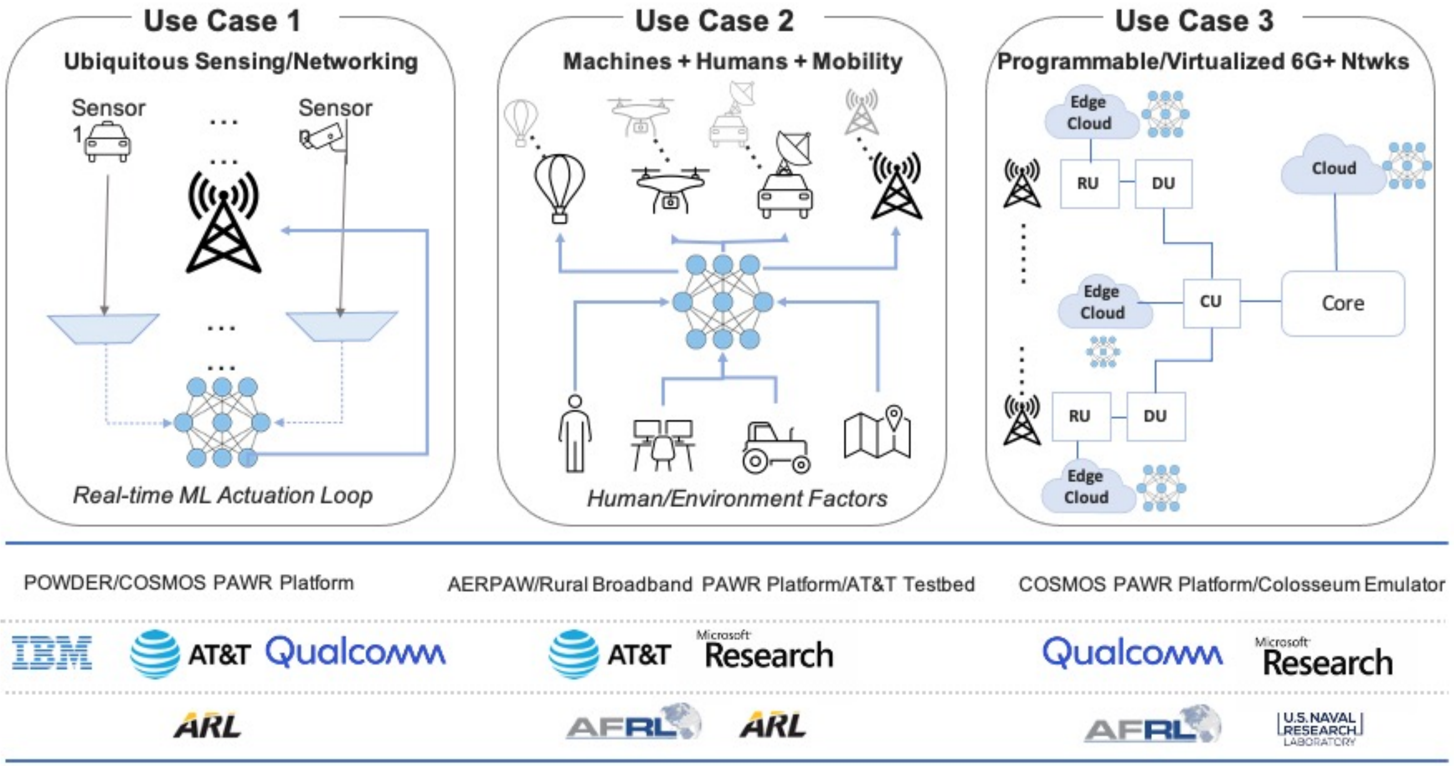

DISTRIBUTED INTELLIGENCE
WIRELESS EDGE NETWORKING SUBSTRATE
NETWORK SECURITY
SELF-HEALING OPTIMIZATION
DECISION & CONTROL
CORE NETWORK
CORE NETWORK
VIRTUAL EDGE NETWORKS (AI AGENTS)
HUMAN-MACHINE COORDINATION
DISTRIBUTED AI PROBLEMS (EG. Federated Learning)
PRIVACY & SECURITY
AI FOR NETWORKS
NETWORKS FOR AI
CAMERA NETWORKS
DRONE SWARMS
VEHICULAR NETWORKS
MIMO
mmWave
INTERNET OF THINGS
EDGE CLOUD
6G & BEYOND WIRELESS NETWORKS
ROBOTIC NETWORKS
HETEROGENEOUS WIRELESS EDGE NETWORKS

Research Plan

- Research Plan is organized across 8 thrusts that span two broad symbiotic research areas
  - **AI for Networks** (Thrusts 1-4): A new distributed intelligence plane will be developed to ensure that these networks are self-optimized, self-healing, secure, and resilient with little or no human intervention.
    - Require new AI tools and techniques that take into account network constraints and dynamics
    - Build upon decades long domain knowledge
  - **Networks for AI** (Thrust 5-8): Develop intelligent and adaptive networks that will interconnect different different AI agents, making AI more efficient, interactive, and privacy preserving.
    - Future of AI is distributed AI. The goal of this network will be to unleash the power of collaboration among different AI agents to help solve long-standing open challenges in distributed AI.
    - Require new network and AI integration that will make AI operations more network aware

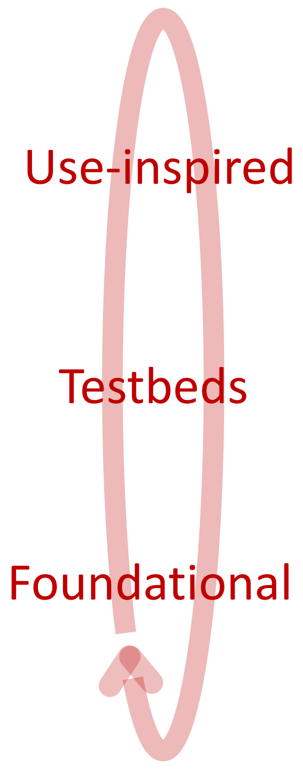
Research Plan (2)

- Research tasks will explore three important wireless edge network use cases in depth
  - Ubiquitous sensing/networking
  - Machines + humans + mobility
  - Programmable/Virtualized 6G+ networks
- Key issues
  - How to connect the key research thrusts to specific experimental platforms
  - How to translate them so they are adopted by our industry and DoD partners.



Key Differentiators of the Institute

- **New foundational AI:** Astonishing success of AI provides a unique opportunity to design distributed intelligent (efficient, self-healing, secure, adaptive) next generation edge networks (AI for Networks)
  - Simply applying known AI techniques is not good enough
  - Requires foundational AI advances that take into account network constraints, dynamics, and domain knowledge
- **AI-Aware Networks and Network Aware AI:** Intelligent and adaptive network edge will unleash power of collaboration to solve large-scale distributed AI problems – which is where most of the AI growth will take place (Networks for AI)
  - An intelligent and adaptive network is needed to coordinate the AI running on vehicles and edge devices
  - Distributed AI algorithms need to be aware of network constraints on determining how to share information
- **Virtuous cycle from theory to implementation to tech transfer:**
  - Shortening the time-scales of interaction between use case and foundational research across multiple disciplines and Industry/DoD ➔ cascading impact dramatically accelerating the time from research to tech transfer.

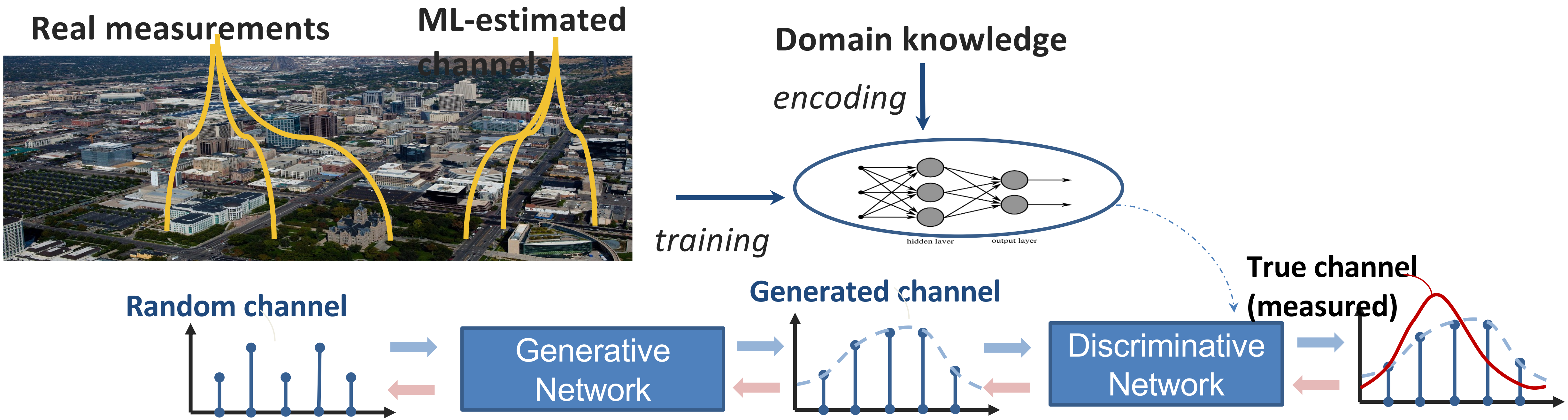


Brief Introduction to the Research Thrusts

T1: Reengineering the Physics/Constraints

Re-engineer the physical fabric for NG (6G+) wireless communications through AI, thus treating the fabric itself as a controllable entity.

- **Leverage Physical Knowledge**
- **Engineering the Physical Environment**
- **Deep-learning Facilitated Communication Algorithms**



T1: Example Research Challenges (1)

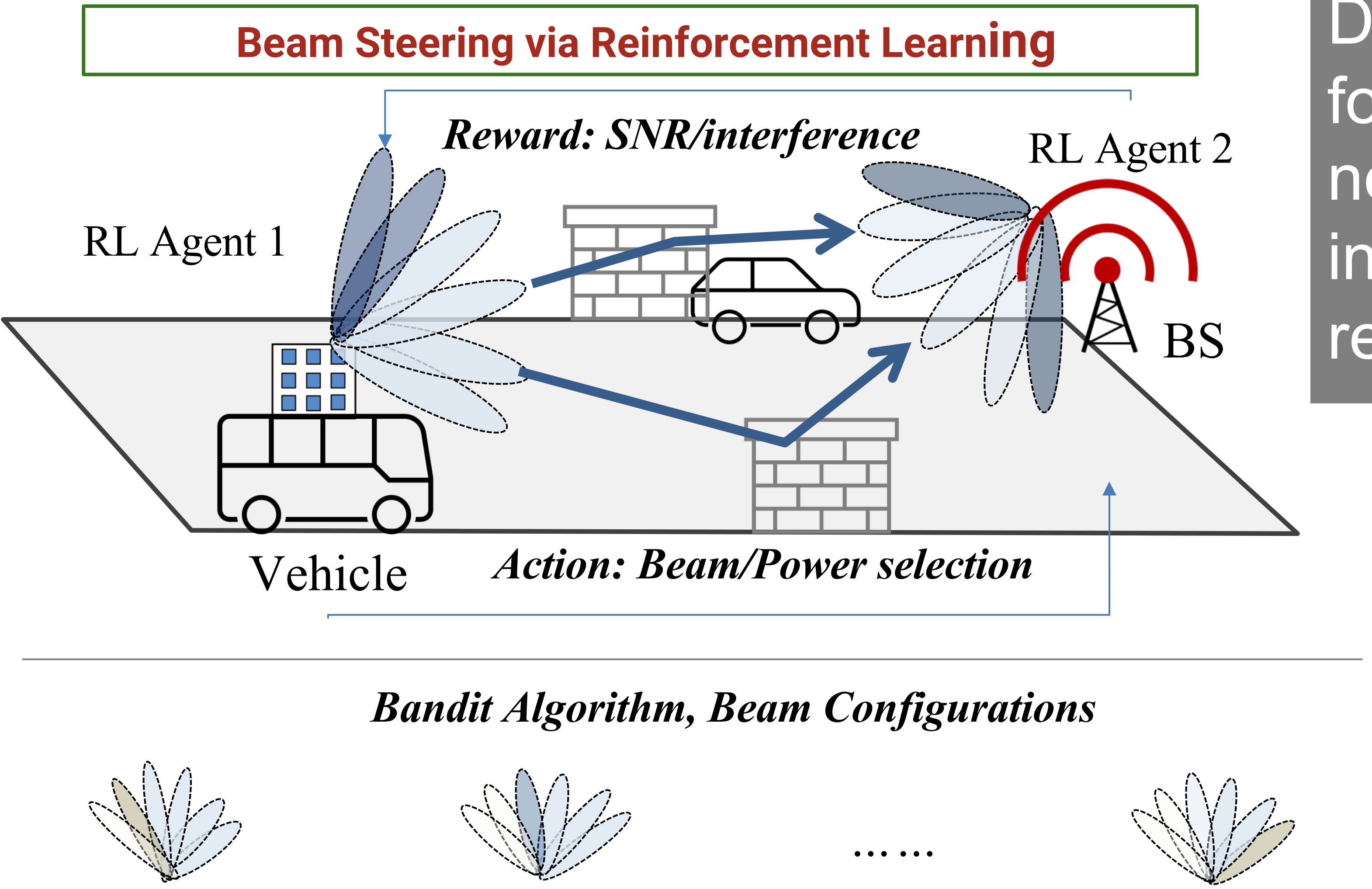
EDGE Institute

- **Leveraging Physical Knowledge:**
  - Physics based models have been used to solve complex problems (e.g., by leveraging meaningful physical models into a NN to speed up prediction).
  - Can such models be used to design ML tools to better estimate parameters (multi-path components, mobility, …), make initial guesses, provide boundary conditions, etc?
  - Can physics based models be integrated with ML tools to help make better control decisions (e.g., beam-searching, beam-forming, etc.), and substantially accelerate performance (convergence time, etc.).
    - Promising preliminary results [Srinivasan & Parthasarathy] on PHY guided ML to reduce beam-searching time by order of magnitude

T1: Example Research Challenges (2)

- **Discovering Efficient Codes in Communication Systems via ML**
  - Developing efficient codes is based on human ingenuity with sporadic breakthroughs on linear codes
  - Can ML be used to expedite discovery and DNN be used to expand the search space to nonlinear codes? Several key challenges:
    - Number of codewords is extremely large;
    - Training at one signal-to-noise ratio (SNR) and a small block length does not easily generalize to other SNRs and larger block lengths
  - To overcome these challenges, domain knowledge from communication, coding, and information theories is critical.
    - Promising preliminary results [Oh et. al] in designing low latency non-linear codes (important under extreme mobility)

T2: AI-Based Network Resource Allocation



Develop new AI techniques for the design and control of next-gen networks taking into account practical resource constraints.

- Low-complexity and Sample-efficient AI-network Algorithms
- Algorithms with Mis-specified Models
- Learning from Historical Data and Incomplete Network State

T2: Example Research Challenges (1)

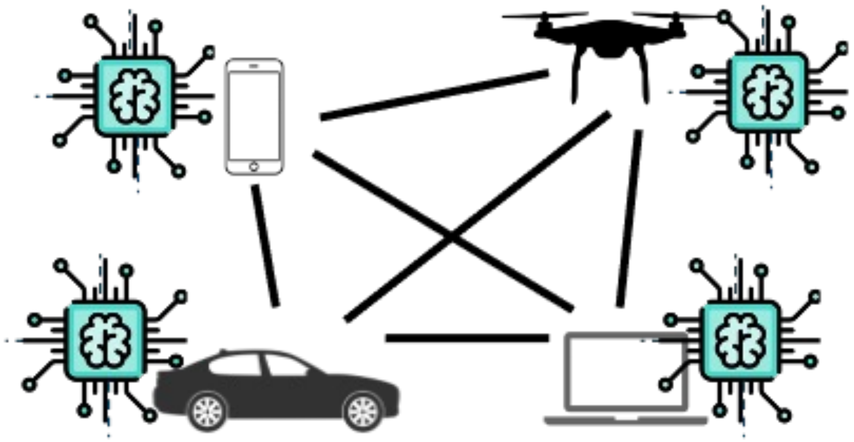
- **Low-complexity and Sample-efficient AI/ML Algorithms**
  – Need new techniques to rapidly learn and adapt to time-varying network conditions
    - To provide efficient and timely operation of critical network services
    - Identify errors and faults in an automated manner
    - Automated parameter tuning at scale (e.g., for cellular network configs) …

  – How to design ML tools for network control that have low sample, compute, and communication complexity, are safe, and can account for network environment/objectives
    - (i) non-stationary dynamics (ii) hard, soft, and safety constraints of network operation as well as user requirements; (iii) distributed/heterogeneous nature of large networked systems; (iv) multiple objectives…

T2: Example Research Challenges (2)

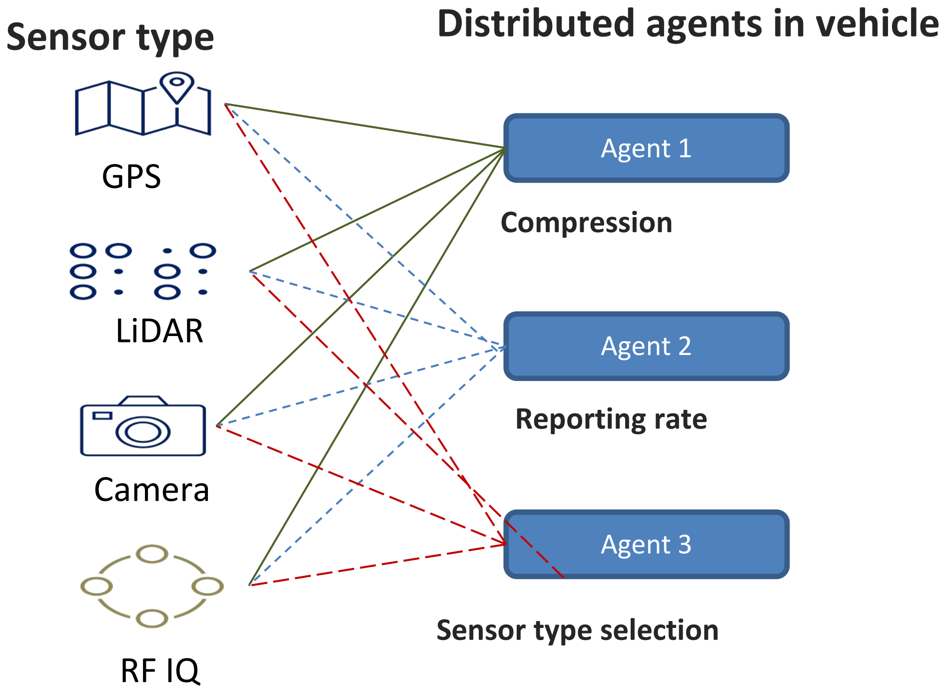
- **Incomplete Information:**
  - Practical challenge for network control is insufficient data at runtime
  - Many network control functions cannot wait for all data to arrive (fast decisions need to be made)
  - Need to develop multi-scale ML tools
    - Ex. One could use local data for real-time control, but use historical global data to determine policy.
    - Can such techniques provide near-optimal performance guarantees?
    - Can they be made scalable?

T3: Multi-Agent Network Control



Develop multi-agent AI techniques for distributed intelligence and control across possibly non-cooperative, network entities.

- Network as a Multi-Agent System
- Fair Network Operations Among (Non-Cooperating) Users
- Data Sharing and Augmented Learning for Distributed Network Operation and Resource Utilization
- Overcoming curse of dimensionality



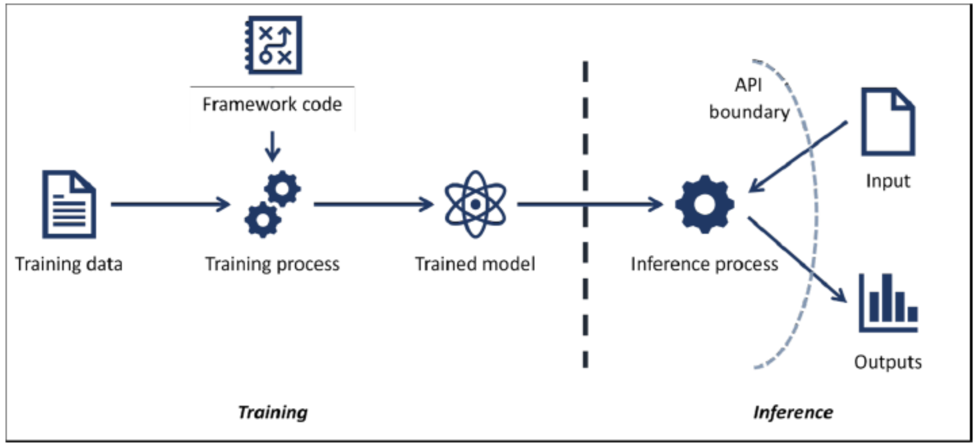
T4: AI-Powered Network Security

Develop new AI tools and techniques to guarantee the network is secure, intrusion free, and highly robust.

- **Creating Automated Tools to Ensure Security**
  - **Systematic Analysis of Network Protocol Specification**
  - **Systematic Analysis of Network Protocol Implementation**
  - **Network Anomaly Detection**
- **Overcoming Data Poisoning Attacks**
- **Data and Reasoning-driven Forensics**
- **Understanding Security/Performance Tradeoffs**

Training time

Gradient-based attacks
Clean label poisoning
Backdoor attack
GAN-based attack



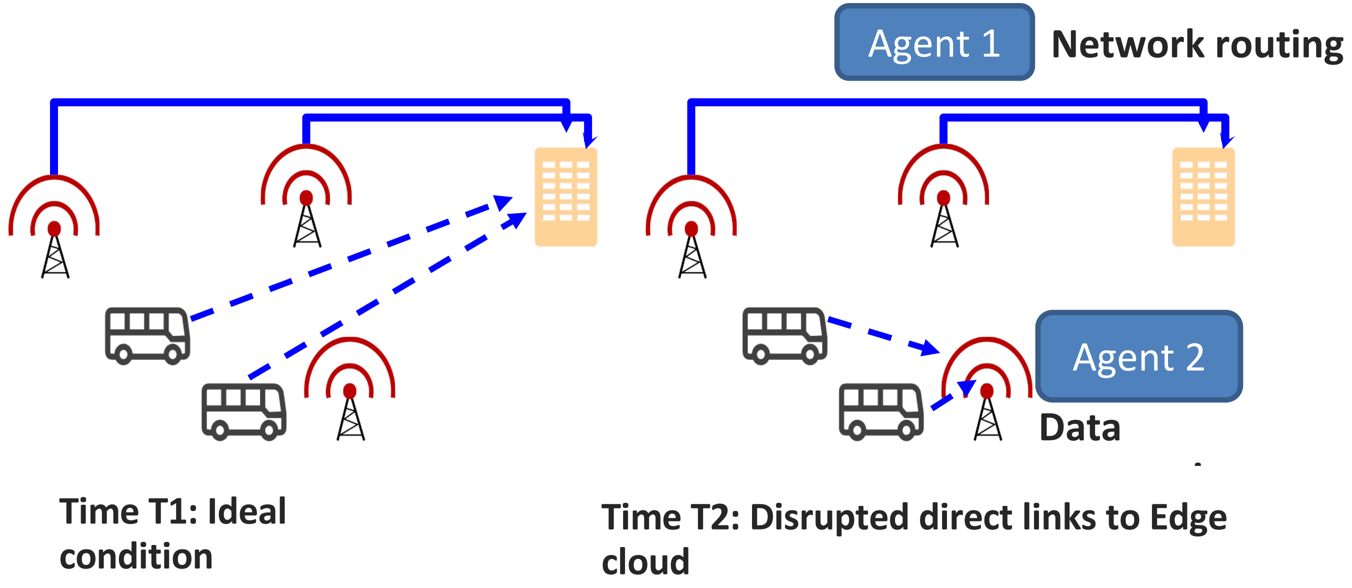
Defenses

Robust training
Feature squeezing
Data compression
GAN-based
Detector
Data randomization
Defensive distillation
Network verification
Gradient regularization
Deep compression network
Input reconstruction
Data sanitization
Certified defense

Inference time

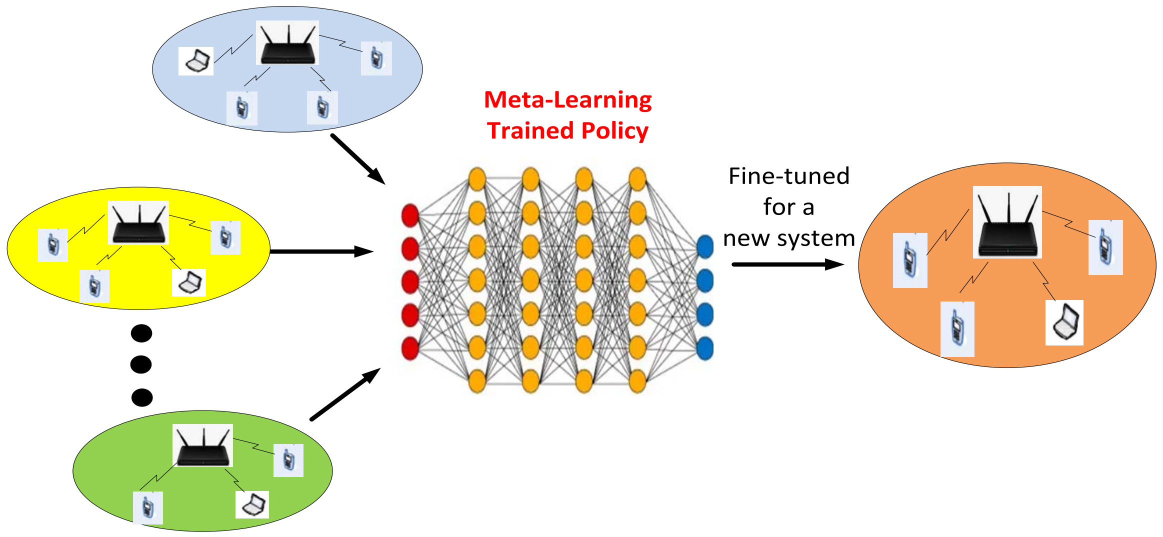
Model evasion (aka adversarial examples)
Functional extraction
Model inversion

T5: AI-Aware Network Operations



Develop distributed AI tools that will seamlessly adapt their operation by taking into account computation, communication and data constraints.

- Communication-Efficient and Network-Aware Distributed Optimization
- Scalable, Network-Aware Distributed Inference
- Meta-Learning, Hyperparameter Optimization, and Active Learning



T5: Example Research Challenges (1)

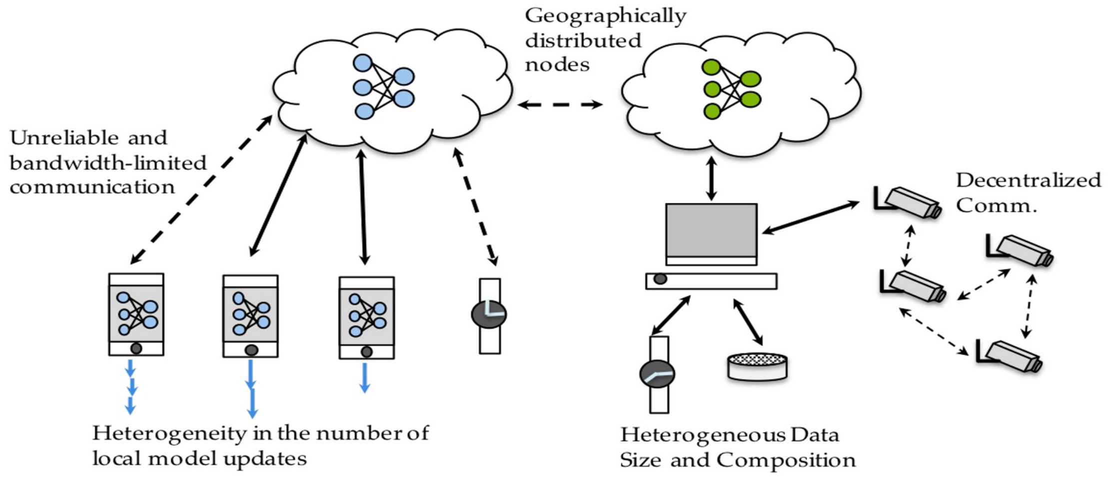
- **Network-Aware Distributed Optimization/ML**
  - Traditional distributed ML assumes reliable communication between a central aggregating server and worker nodes
  - In edge networks, workers are communication-constrained devices, such as mobile phones, IoT sensors, cameras, etc., with heterogeneous computing speeds and unreliable connectivity.
  - What are the fundamental tradeoffs between communication efficiency and ML convergence?
    - Preliminary works [Joshi et. al] for SGD show that communication gains can be substantial by transmitting infrequently when algorithms are far from convergence
    - Can we design appropriate ML tools explicitly taking network constraints (BW, delay) into account?
  - Can we further improve performance by overlapping communication, local computation, and compression/coding?

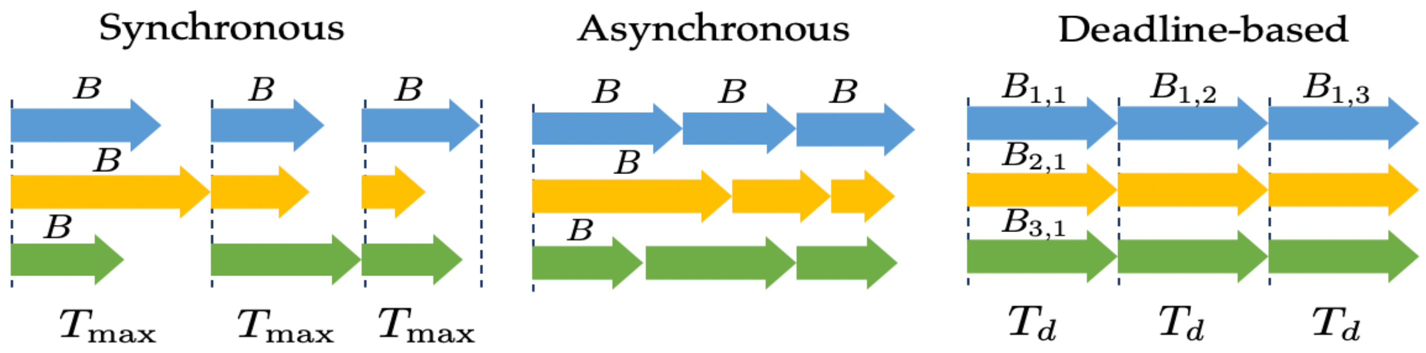
T5: Example Research Challenges (2)

- **Scalable, Network-Aware Distributed Inference**
  - Large scale ML problems can be distributed over many servers
    - Stragglers (slowest tasks due to low delay links, overloaded servers, etc.) substantially impact performance
    - Practical solutions (e.g., monitoring progress and replicating slower tasks) have significant limitations
    - Coded computation is promising and allows the final results to be recovered from a subset of completed tasks
    - Problem: Resolve straggler problem but may increase time because each parallel job now becomes more complex
  - Can coded computation take into account the structure of the ML problem (e.g., sparsity) and be adaptive to the heterogeneous edge?
    - Preliminary works on exploiting sparsity for code design [Shroff et. al]

T6: Network Operations for Distributed AI

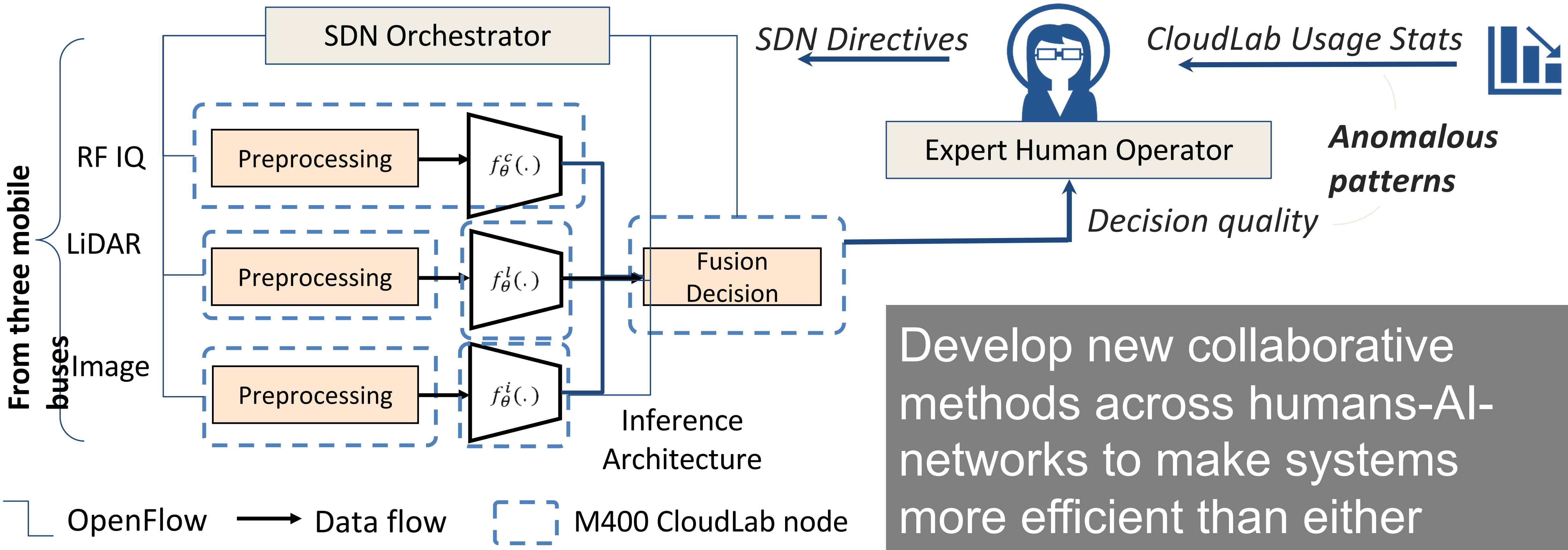
Re-engineer networks by adaptively allocating communication, computing, and storage resources for serving the needs of distributed AI applications.





- Network Operation for Managing AI-Side Uncertainty and Dynamics
- Network Operation for Managing Network-Side Uncertainty and Dynamics
- Unified, Distributed Network Operation for AI Applications

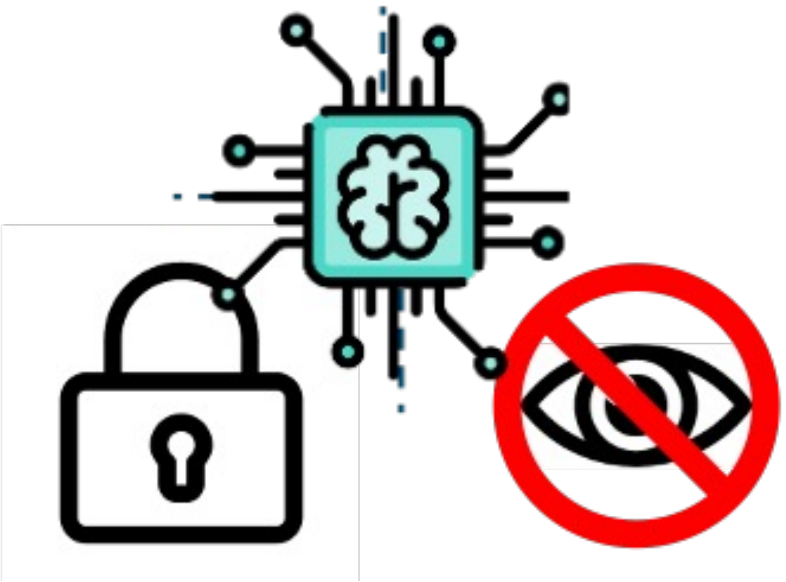
T7: Human, AI, & Network Research Interface



- ➢ **Human-AI Interface**
- ➢ **AI-Network Interface**
- ➢ **Human-Network Operations Interface**

Develop new collaborative methods across humans-AI-networks to make systems more efficient than either human or machines by themselves.

T8: Security and Privacy for Network Users

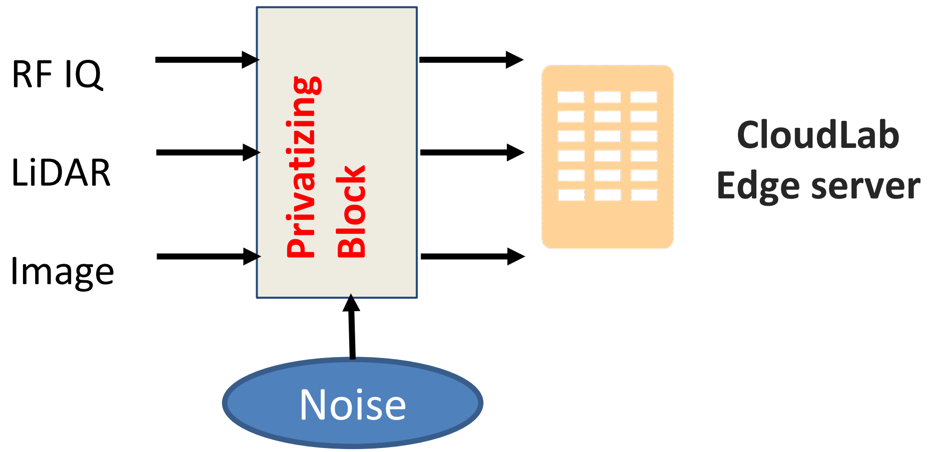


Design and control the networks such that they are privacy-aware and can be optimized to facilitate protection from information leakage and attacks.

Data/network/time-dependent privacy constraints

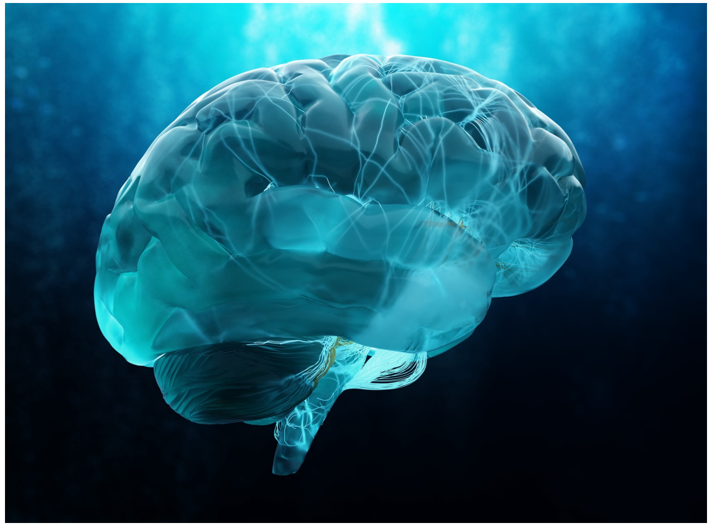
e.g., Local/centralized/shuffling-based differential privacy

- **Handling Heterogeneous and Dynamic Privacy Constraints and the Absence of Trusted Curators**
- **Protecting Data-in-Computing via Trusted Execution Environments (TEEs)**



Foundational AI-Advances

- **Various foundational AI advances are being made by the AI-EDGE Institute**
  - Reinforcement learning & bandits (constrained, online, offline, adversarial, deep…)
  - Federated learning
  - Meta learning
  - Transfer Learning
  - Continual Learning
  - Representation learning
  - Adversarial ML learning
  - Deep Neural Networks and Neural Tangent Kernels
  - Explainable AI…



➔ Allows for Synergies with various industry/DoD

Platforms Accessed by AI-EDGE Researchers

EDGE Institute

NSF PAWR Platforms

POWDER

Salt Lake City, UT

Bertino, Purdue, *5G security*

Chowdhury, NEU, *radar sensing*

COSMOS

West Harlem, NY

Chowdhury, NEU, *over-the-air ML models*

ARANET

Ames,IA

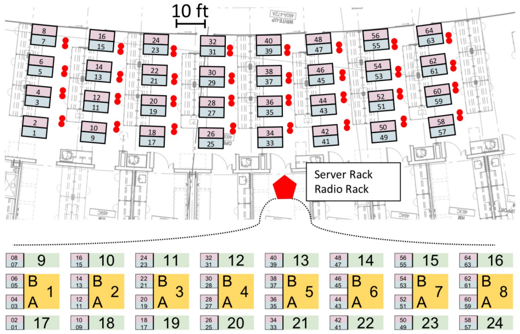
Arora, OSU, *creating rural broadband links*

RFDATA FACTORY

https://www.rfdatafactory.com

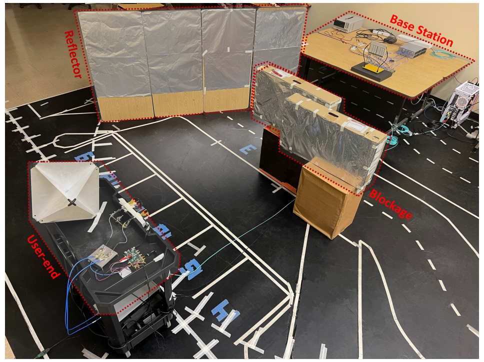
NSF Dataset and API repository

ARENA (indoor, SDR)



Melodia, NEU, *O-RAN test and measurement*

O-JRC: mmWave MIMO



Ekici, OSU, *mmWave beamforming*

Compute@IBM

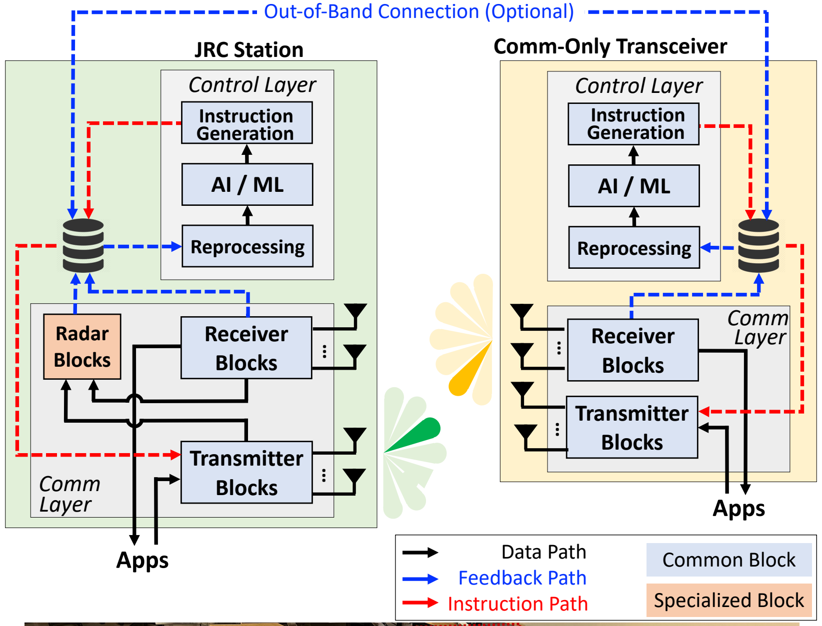
<u>Cognitive Computing Cluster (CCC):</u> *547 nodes with x86 processors and NVIDIA V100 GPUs*

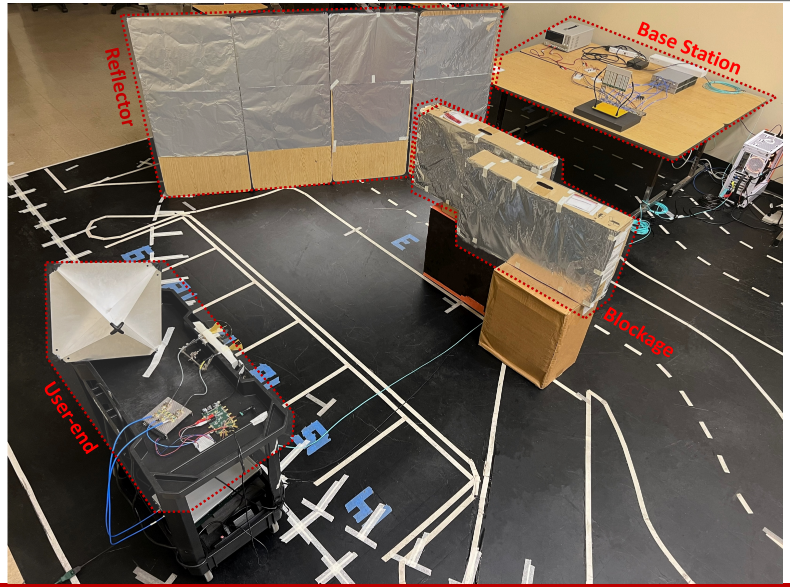
<u>AiMOS:</u> *268 nodes x86 processors and 1576 NVIDIA V100 GPUs*

Several AI-EDGE team members have access

O-JRC: An Open-Source Software Framework for mmWave MIMO Development and Experimentation

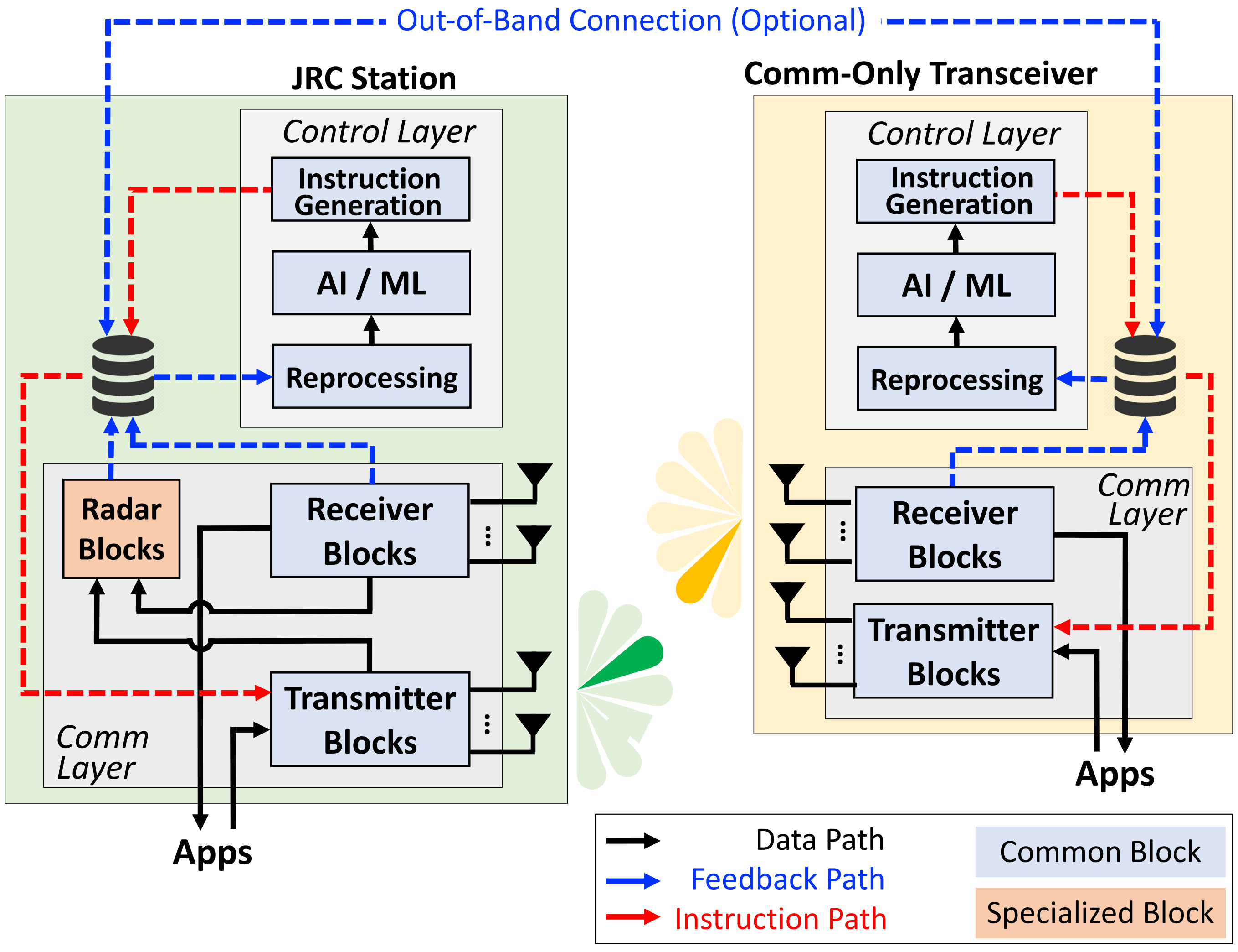
- **Key Properties**
  - **Layered and Modular Architecture**
    (Isolated Layer Communication, Flexible Module Replacement)
  - **Agile Development and Validation**
    (Intelligence, Agility, and Programmability)
- **Capabilities**
  - **Configurable System Structures**
    (Modulation Schemes, MIMO Setup, Beamforming Types)
  - **Online/Offline Model Training and Real-Time Control**
  - **Beam Training and Beam Tracking**
    (3 Control Algorithms verified for Single User)
  - **In Development:** Multi-User Detection,
    Blockage Prediction, RL-Driven Resource Allocation…
- **Testing Scenarios for Algorithm Validation**
  - **Static Scenarios:** Standard, Multipath, Obstruction
  - **Mobile Scenarios:** End User Mobility (indoors, up to 1 m/s), Multipath, Obstruction



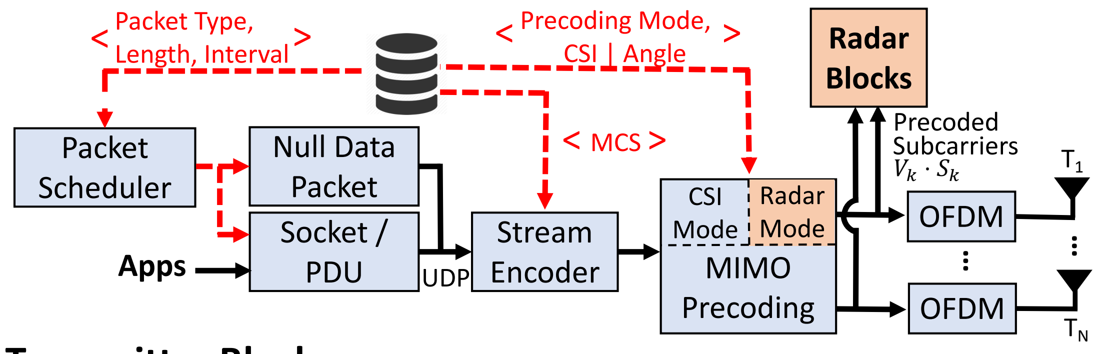


O-JRC: An Open-Source Software Framework for mmWave MIMO Development and Experimentation

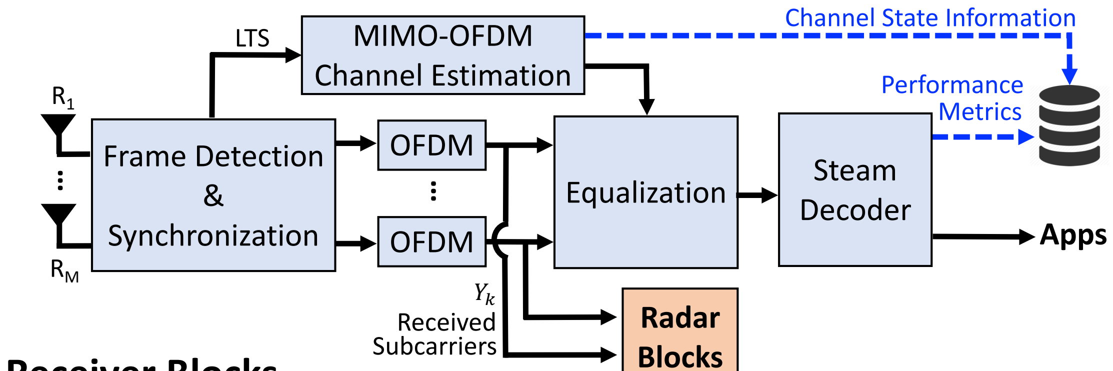
EDGE Institute



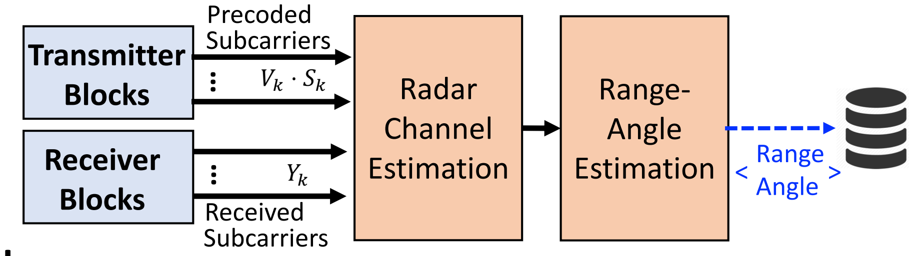
JRC Software Framework Overview



Transmitter Blocks



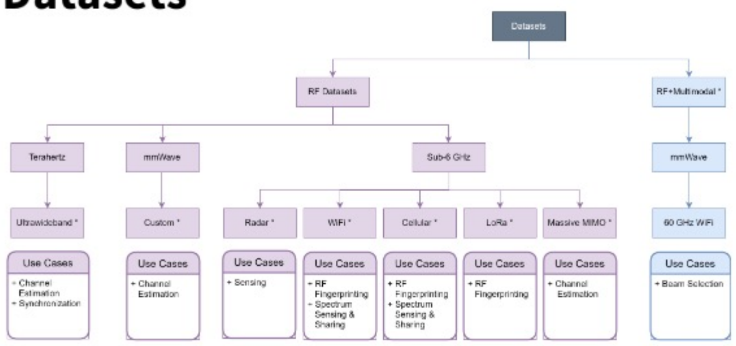
Receiver Blocks



Radar Blocks

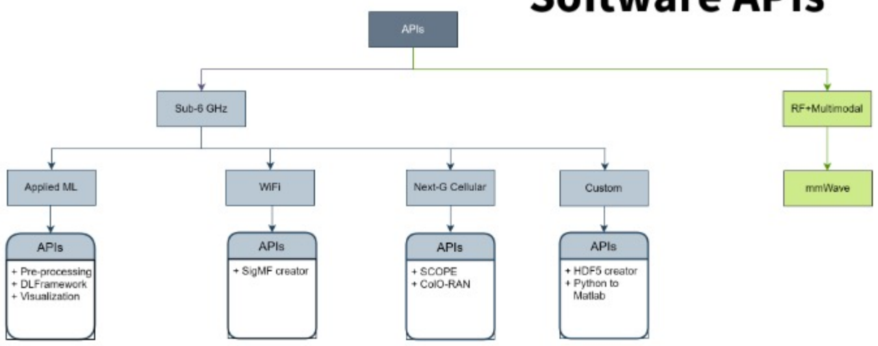
Datasets Generated by AI-Edge Faculty

Datasets

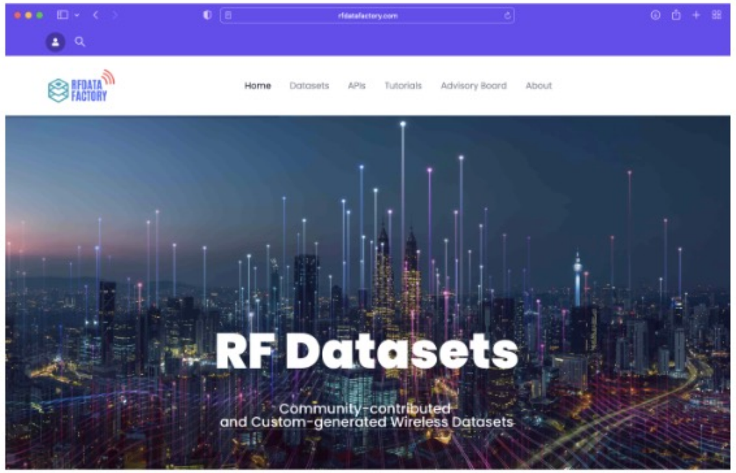


- Large number of data-sets/software APIs are being generated and collected
  - WiFi, LoRA, cellular LTE/5G signals, etc.
  - Applications in channel est., beamforming, modulation, etc.

Software APIs



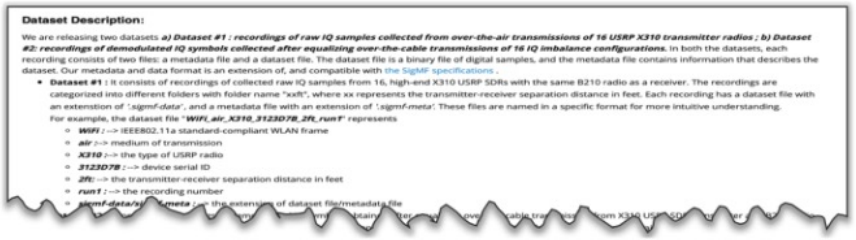
One-stop Resource for Datasets for the Wireless Community



Tutorials

https://www.rfdatafactory.com

SigMF Metadata



- Many of these are downloadable

Example Synergies with Industry/DoD (1)

- **Various foundational AI advances are being made by the AI-EDGE Institute**
  – Reinforcement learning and bandits (constrained, online, offline, adversarial…)
  – Federated learning
  – Meta learning
  – Transfer Learning
  – Representation learning
  – Adversarial ML learning
  – Deep Neural Networks and Neural Tangent Kernels
  – Explainable AI…

➔ Allows for Synergies with various industry/DoD

Example Synergies with Industry/DoD (2)

- **AI-Based Resource Allocation and Control of MANETs**
  - Congestion Control; Scheduling; Interference Management; Energy Management…
- **AI-Assisted Awareness**
  - Single-platform “sensor” fusion
  - Federated learning – perception from multiple sources
  - Fusion of dissimilar data sources (comm, on-board sensors, offline models)
- **AI-Driven Decision Making**
  - Coordinated vehicle routing (city-wide route optimization)
  - Fleet behavior
  - Multi-vehicle coordination for safety
  - Single platform collision avoidance
- **AI-Driven Safety Management**
  - Reinforcement learning (RL) with safety constraints, such as instantaneous safety requirements
  - Deep-learning based detection and collision avoidance
  - Imitation learning and reward-free RL for autonomous driving

Example Synergies with Industry/DoD (3)

- **Smart Manufacturing:**
  - Interconnected Robots
  - Automating manufacturing processes
- **AI based security/privacy**
  - Security detection; robustness to attacks; fast recovery
  - Differential privacy guarantees; prevention of information leakage, etc
- **Learning based optimization**
  - Faster/scalable chip designs
  - Energy efficient processors and IoT devices
  - More efficient data centers, networks, …
- **Human-Machine-AI Research**
  - Human-AI interface; AI-Human interface; Machine-AI operations interface
  - Improve performance over what humans and AI can deliver…

Data Intensive Sciences and AI-EDGE

- Large data volumes have historically been concentrated at the network core

- Million-Dollar Question:
  What happens when source and consumption shift to the EDGE?
  - Does it make sense to shift data to the core first?
  - Are challenges really still at the core?
  - Can we really plan ahead – What are the real-time dynamic requirements?
  - What is the role of ML/AI in moving and processing data?

- How can AI-EDGE be of service?

Data Intensive Sciences and AI-EDGE

- AI-EDGE can provide
  - Access to a wealth of testbeds, emulation platforms, development platforms
  - Customized joint Network/AI solutions specific to DIS
  - Collaborative innovation for JIT data processing
  - Capabilities to close the loop for real-time applications

We look forward to collaborating with the DIS community