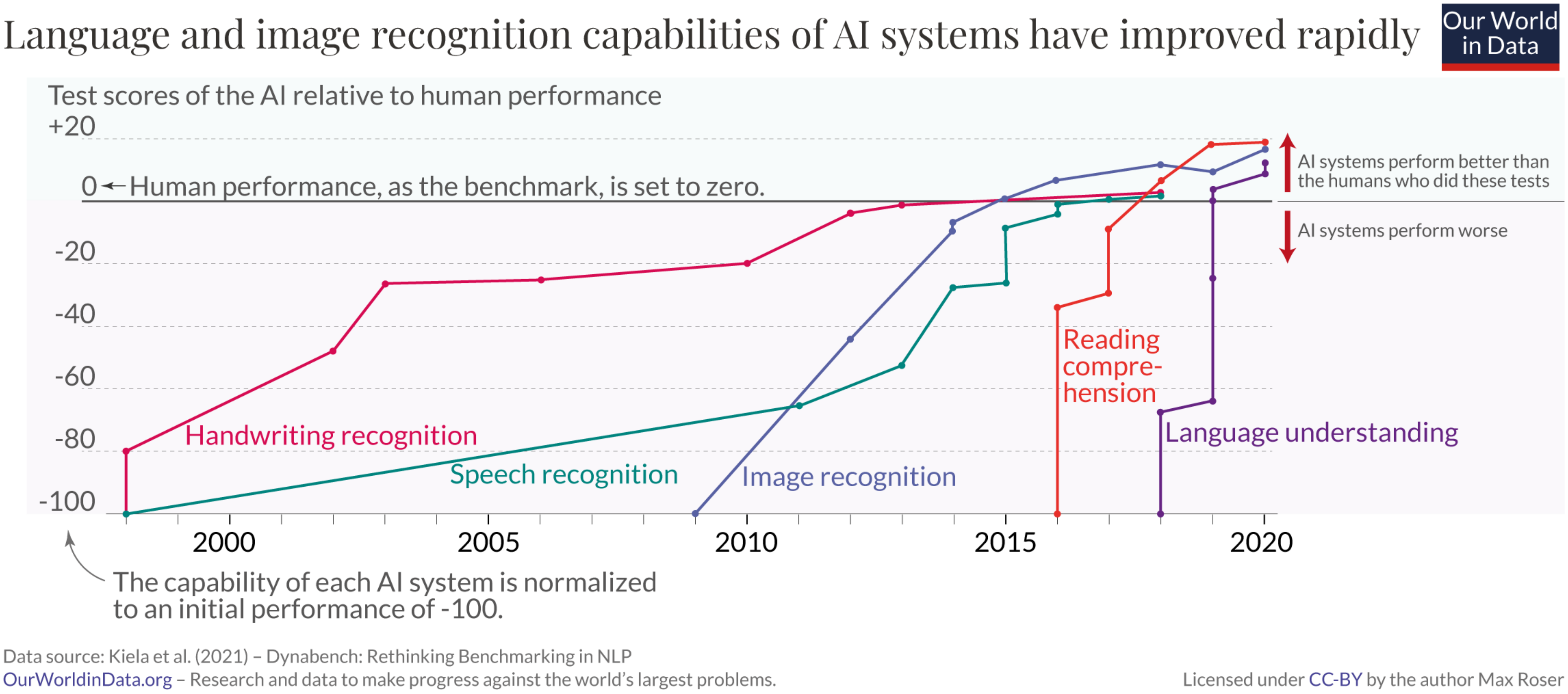
Semiconductor Research Corporation

CoCoSys
CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

CoCoSys Center Overview

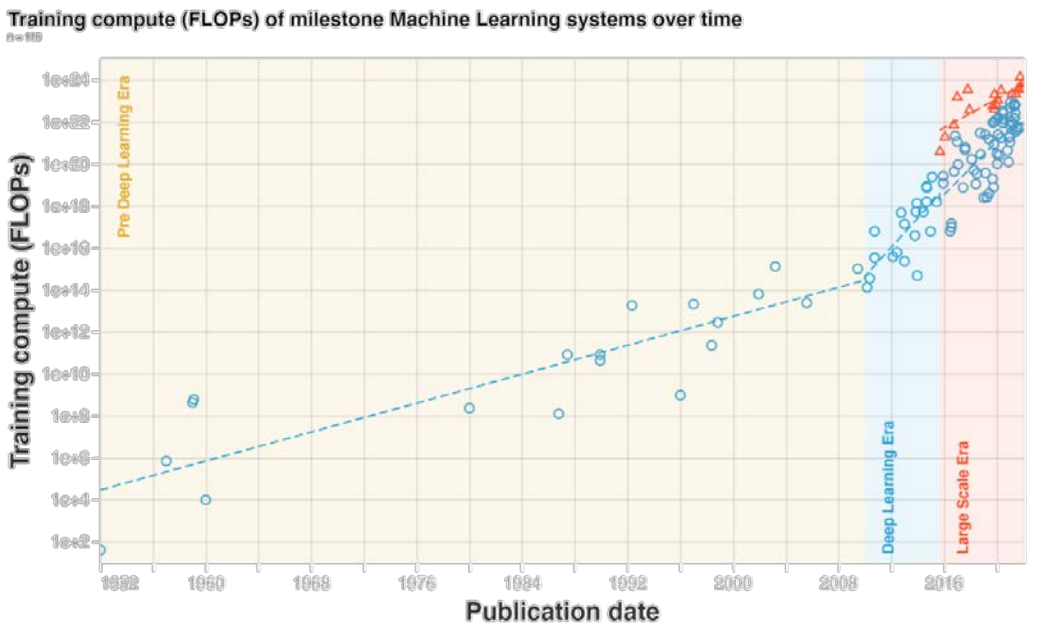
Arijit Raychowdhury
Anand Raghunathan
Anca Dragan
Azad Naeemi
Bruno Olshausen
Jae-sun Seo
James DiCarlo
Jan Rabaey
Josh Tenenbaum
Kaushik Roy
Larry Heck
Michael Carbin
Naresh Shanbhag
Priya Panda
Priyanka Raina
Sumeet Gupta
Tajana Rosing
Tushar Krishna
Vijay Raghunathan
Yingyan (Celine) Lin
Yu (Kevin) Cao

State of AI / Landscape



AI Challenges

- Unsustainable compute trajectory
- Lack of explainability and transparency
- Lack of robustness
- Narrow (specific to task or input modality)
- Algorithms driven by today's hardware (GPUs and digital accelerators)



Center Vision

Current AI Systems

- Black-box (not explainable or interpretable)
- Reliant on large datasets, networks and compute
- Mostly monolithic CMOS technology

CHALLENGES

- Can we stem the unsustainable trends in compute requirements for AI?
- Can a fusion of neural, symbolic and probabilistic methods lead to more scalable, robust and explainable AI?
- Can cognitive algorithms perform the entire gamut of tasks involved in collaborative AI systems (perception, reasoning and decision making)?
- Can cross-layer design of cognitive algorithms and hardware improve energy efficiency by over 100X?

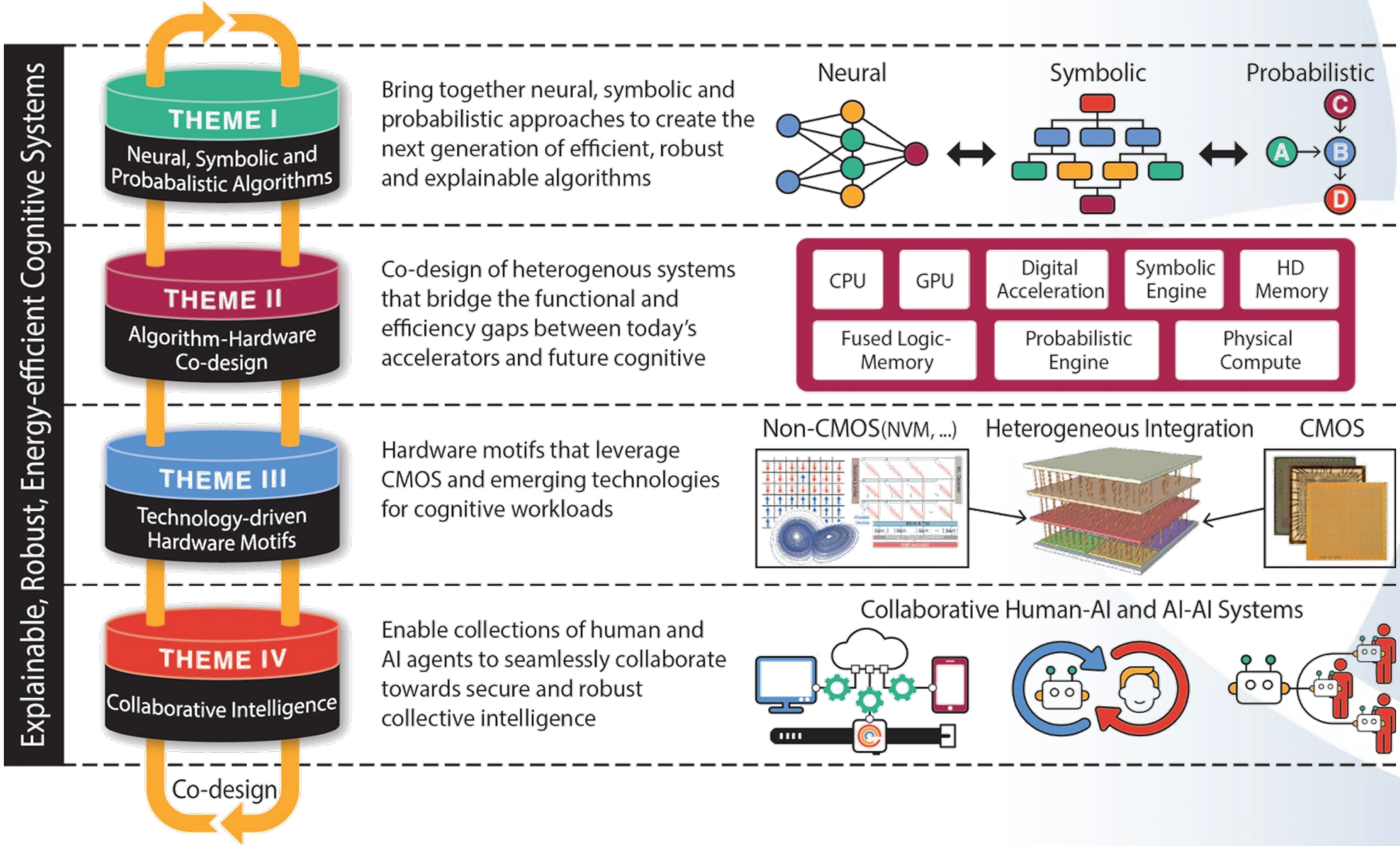
Future Collaborative AI Systems

- Seamless human-AI and AI-AI collaboration
- Explainable, robust and secure
- Hardware and algorithms co-designed to optimize energy efficiency, latency and throughput
- Leverage future logic, memory and integration technologies

SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS


CoCoSys
CENTER FOR THE
CO-DESIGN OF COGNITIVE SYSTEMS
James DiCarlo
MIT
Josh Tenenbaum
MIT
Michael Carbin
MIT
Anca Dragan
UC BERKELEY
Bruno Olshausen
UC BERKELEY
Jan Rabaey
UC BERKELEY
Anand Raghunathan
CO-DIRECTOR, PURDUE
Kaushik Roy
PURDUE
Sumeet Gupta
PURDUE
Yale
Priya Panda
YALE
Berkeley
UNIVERSITY OF CALIFORNIA
Stanford
Vijay Raghunathan
PURDUE
UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN
PURDUE
UNIVERSITY
Priyanka Raina
STANFORD
Arijit Raychowdhury
DIRECTOR, GEORGIA TECH
Jae-Sun Seo
ASU
Naresh Shanbhag
UIUC
UC San Diego
ASU
Azad Naeemi
GEORGIA TECH
Georgia
Tech
Yu (Kevin) Cao
ASU
Celine Lin
GEORGIA TECH
Tajana Rosing
UC SAN DIEGO
COCOSYS SUPPORT STAFF
Larry Heck
GEORGIA TECH
Emily Watson
PROGRAM & OPS MANAGER
Janna Young
FACULTY SUPPORT COORDINATOR
Tushar Krishna
GEORGIA TECH
Melissa Donahue
FINANCIAL ANALYST

Center Overview and Themes



Theme 1 - Neural, Symbolic and Probabilistic Algorithms: Vision

A principled approach to unified neuro-symbolic-probabilistic algorithms, supported by new information representations, computing models and analysis of fundamental limits, has the potential to radically advance the next generation of cognitive algorithms

State-of-the-Art Deep Learning Algorithms

- Require unsustainable growth in data sets, model sizes, and powerful compute
- Black-box models without theoretical underpinnings
- Lack of explainability, interpretability and robustness a fundamental limitation
- Deep learning, symbolic methods and neuromorphic approaches largely viewed as competing alternatives

CoCoSys Theme 1 →

Future Neuro-Symbolic-Probabilistic Algorithms

- Fusion of neural, symbolic and probabilistic approaches enables scalability without rapid increases in model size
- Interpretable, explainable and robust algorithms enable seamless human-AI collaboration
- Algorithms based on emergent and dynamical models expand scope of AI beyond traditional domains
- Theoretical analysis identifies fundamental limits of accuracy-robustness-efficiency tradeoffs

SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

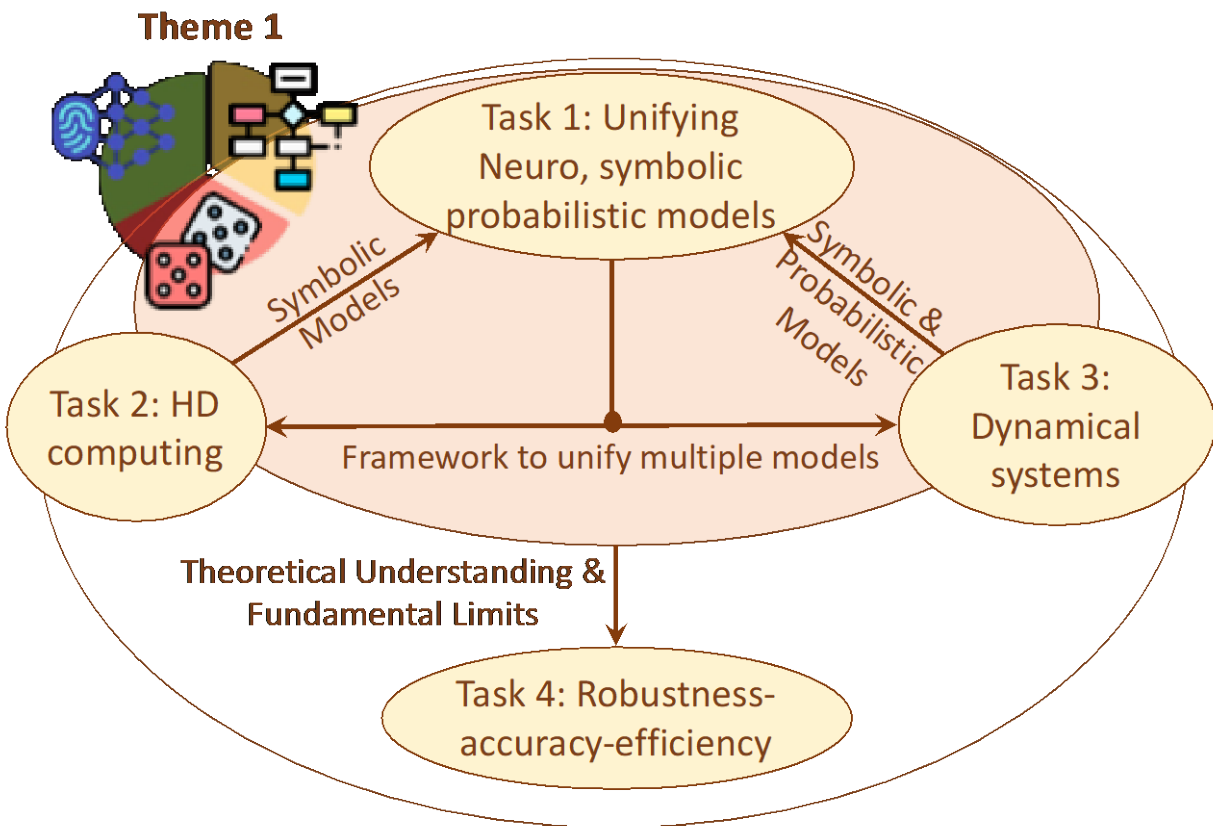
Theme 1 - Neural, Symbolic and Probabilistic Algorithms: Tasks

1.1 Unifying neural, symbolic, and probabilistic models

1.2 Hyper-dimensional (HD) information representations & processing

1.3 Computing with emergent and dynamical systems

1.4 Theoretical underpinnings of robustness-accuracy-efficiency tradeoffs



SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

Theme 2: Hardware-Algorithm Co-Design

The co-design of cognitive hardware driven by the evolution of cognitive workloads as well as the capabilities of future hardware technologies has the potential to unlock quantum improvements in processing efficiency

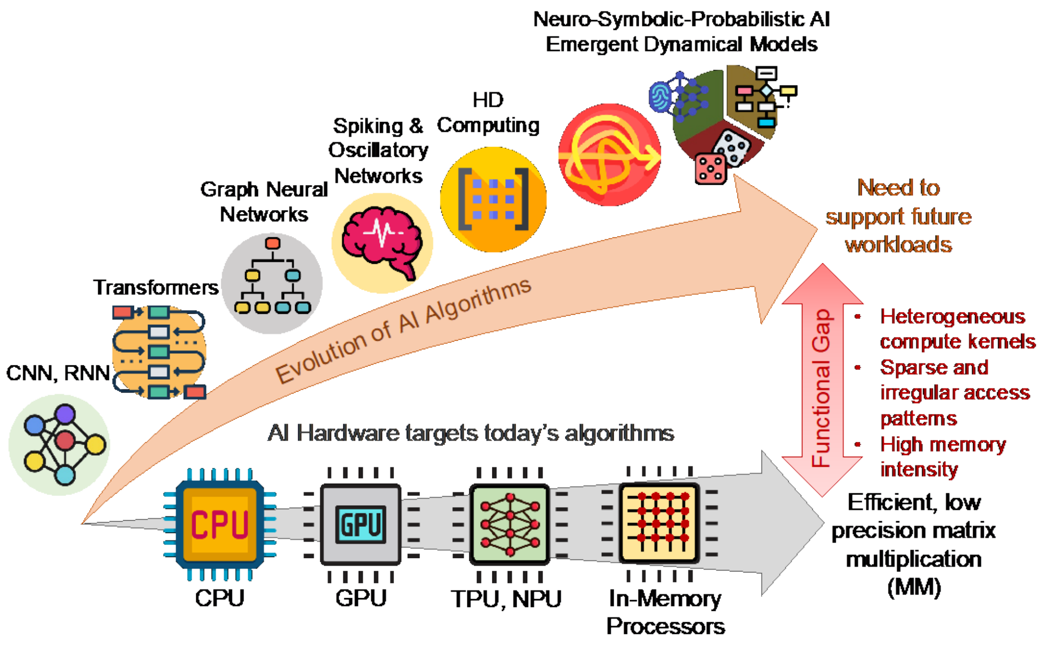
State-of-the-art AI Hardware

- Hardware architecture roadmap largely driven by deep neural networks
- Improvements in low-precision matrix multiplication are saturating
- Separate processing and memory leads to von Neumann bottleneck
- Require regular compute and access patterns for high efficiency
- In-memory compute fabrics limited by low-precision, low sensing margins, non-idealities, and limited array density

CoCoSys Theme 2

Future Cognitive Hardware-Algorithm Co-Design

- **Co-design of algorithms and software unlocks new sources of efficiency**
- **Architectures capture the computational kernels of future neuro-symbolic-probabilistic workloads**
- **Scalable and reliable compute on in-memory and physical compute fabrics**
- **Programming models and software frameworks enable seamless utilization of heterogeneous systems**
- **Re-configurable designs can be tuned for different scenarios across the compute spectrum**



SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

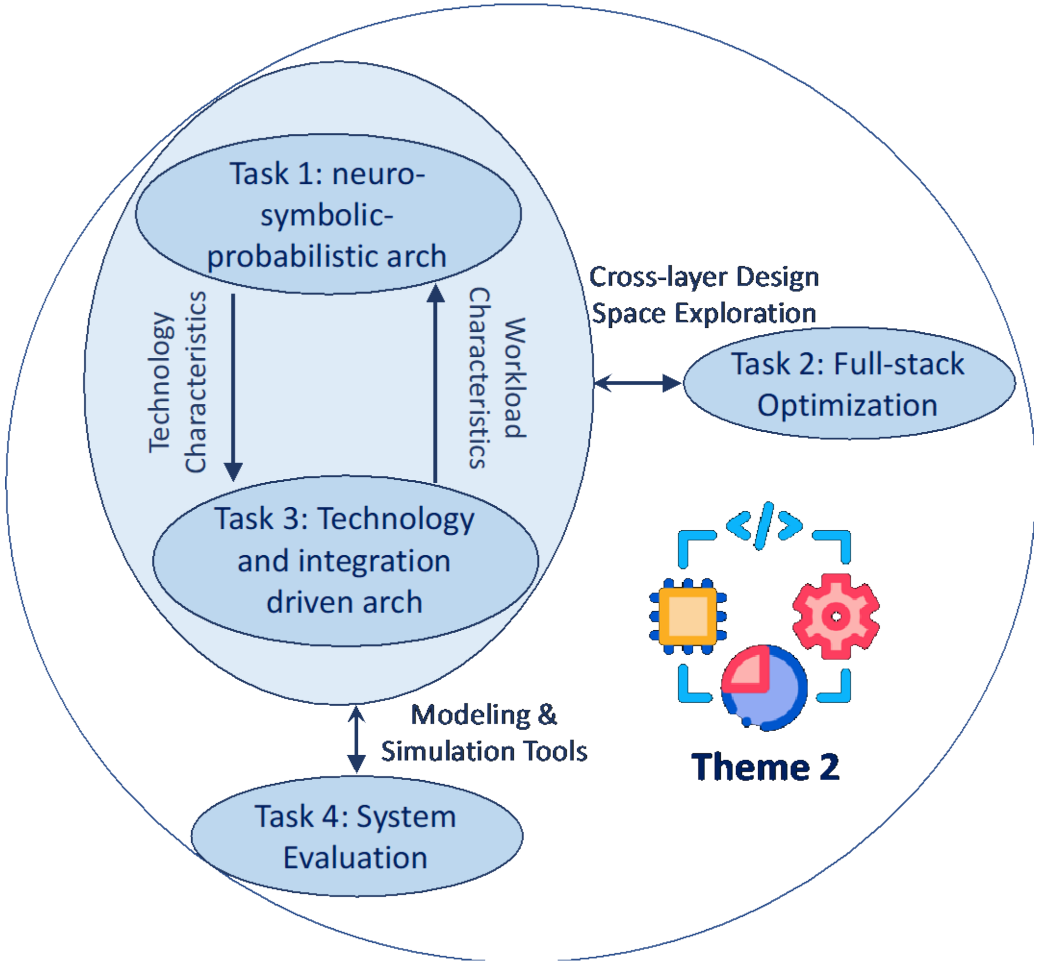
Theme 2 - Hardware-Algorithm Co-Design: Tasks

2.1 Architectures for neuro-symbolic-probabilistic workloads

2.2 Full-stack optimization and software frameworks for cognitive systems

2.3 Technology and integration-driven cognitive architectures

2.4 System evaluation and benchmarking



SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

Theme 3 : Vision

State-of-the-Art Circuit Fabrics for AI Hardware

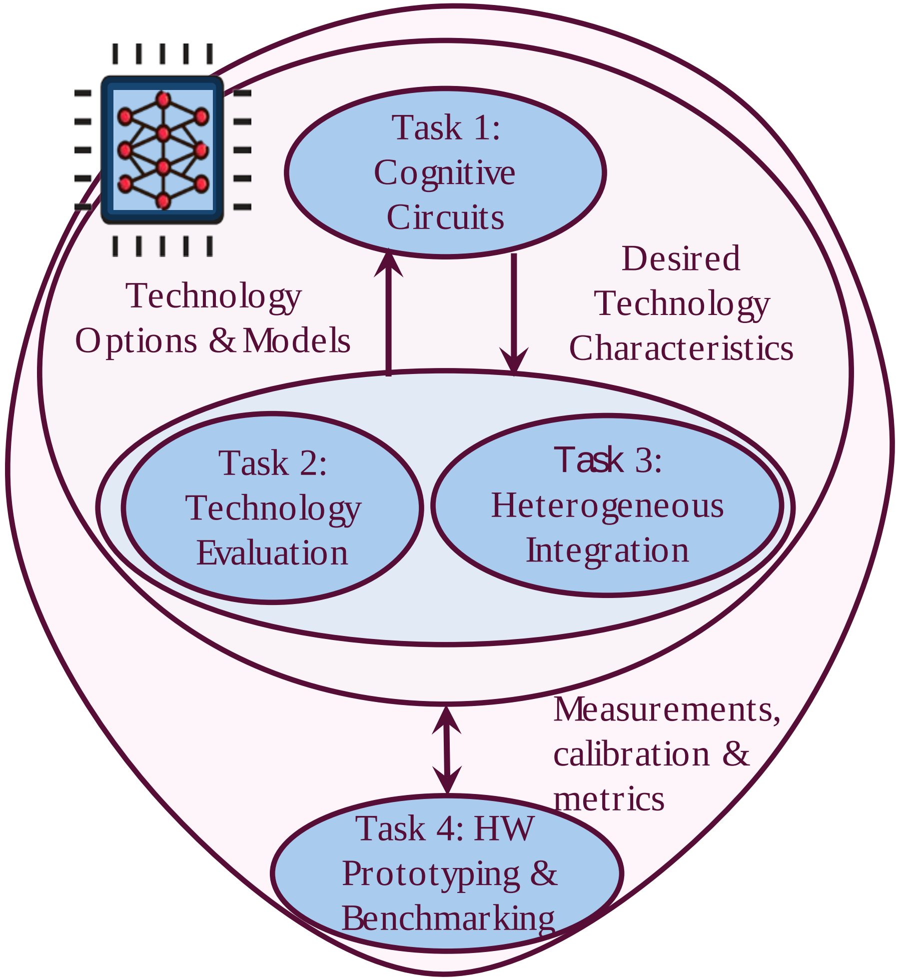
- Low-precision matrix multipliers do not adequately capture the computational kernels of future neuro-symbolic-probabilistic workloads
- Technology and voltage scaling limits sensing margins for in-memory compute
- Technology evaluation driven by general-purpose compute does not reflect needs of cognitive workloads
- Heterogeneous integration more of an afterthought

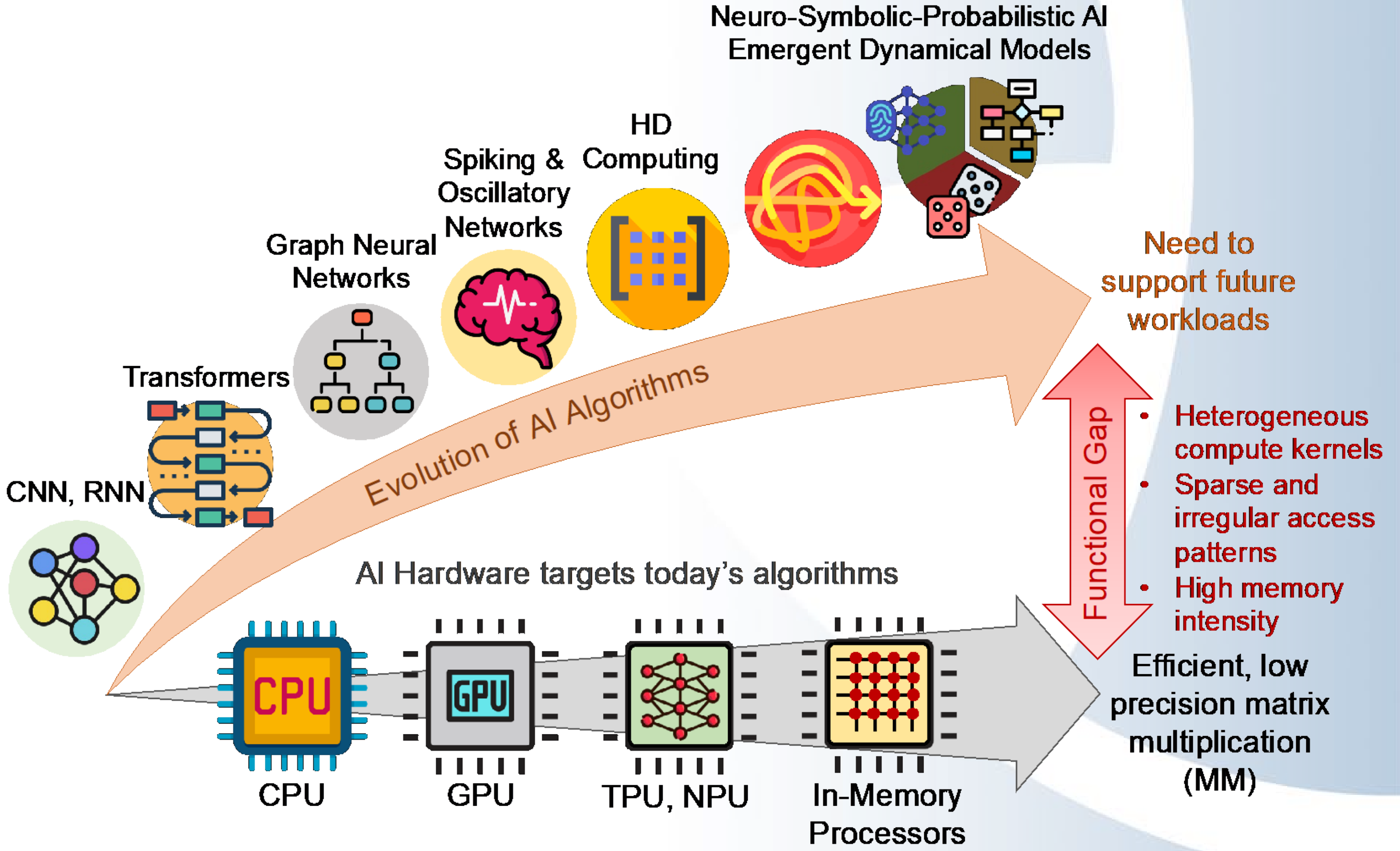
CoCoSys Theme 3

Future Digital, Mixed-signal and Physical Circuit Fabrics

- Hardware motifs for key compute kernels of neuro-symbolic-probabilistic workloads
- High-density, reliable and high-precision IMC fabrics go beyond matrix multiplication
- Composable digital, mixed-signal and physical compute fabrics
- Circuits that encode hard computational problems directly using emergent device dynamics
- Evaluation of deeply scaled CMOS and beyond CMOS technologies for cognitive workloads
- Heterogeneous integration driven circuit and system design

Theme 3 : Overview and Representative Tasks





SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

Theme 4 : Vision

State-of-the-Art in Collaborative AI

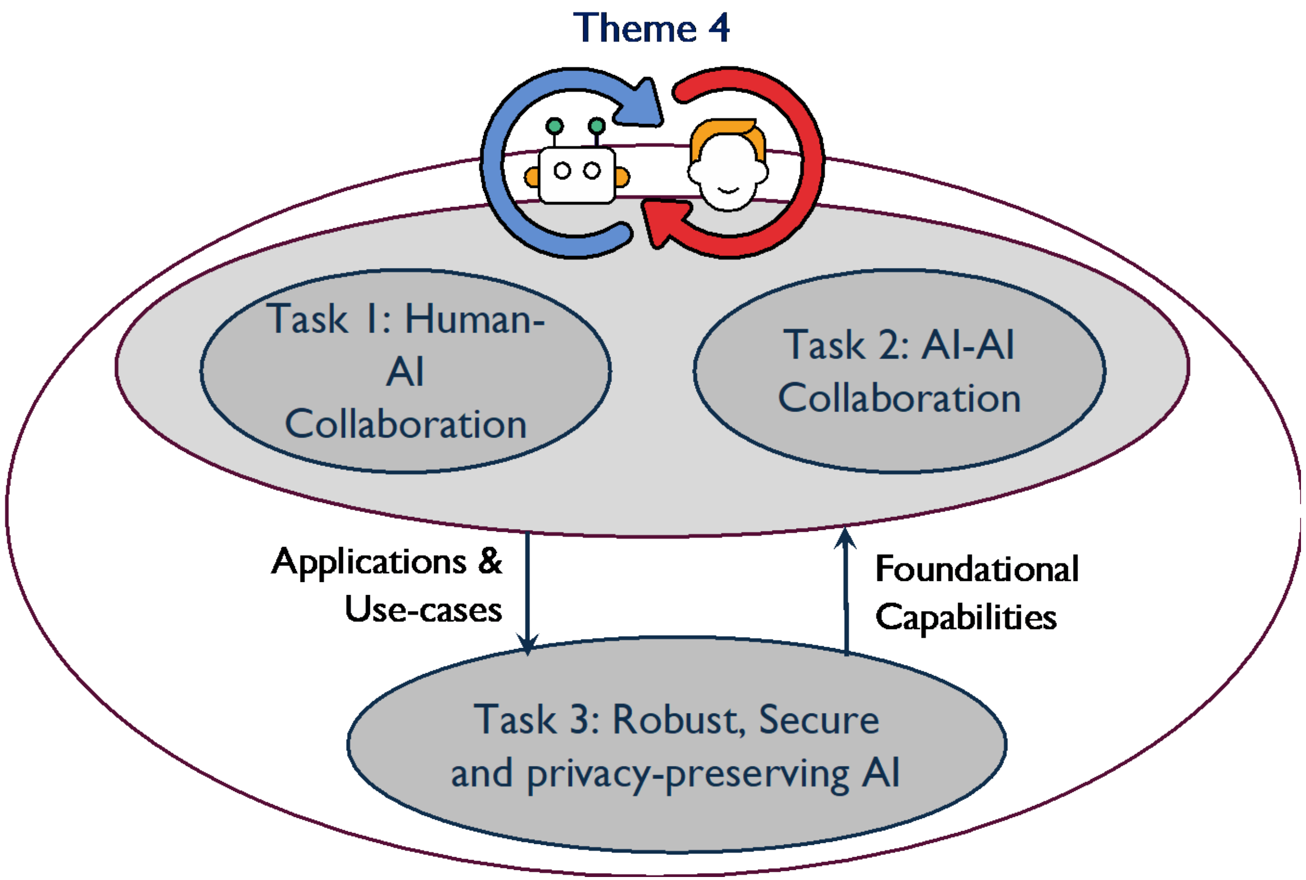
- Human-AI collaboration is highly limited in scope (one-on-one question answering) and learning capability (trained offline)
- AI-AI collaboration (multi-agent systems) mostly focused on homogeneous systems and sensing modalities
- Robustness, security and privacy concerns ranging from model divergence, adversarial attacks and data/model leakage

CoCoSys Theme 4

Future Digital, Mixed-signal and Physical Circuit Fabrics

- AI agents (co-bots and digital assistants) that seamlessly and continually learn from humans
- Neuro-symbolic-probabilistic algorithms enable explainable, interpretable and robust human-AI collaboration
- Multi-agent AI systems that are suitable for highly heterogeneous, unstructured and dynamic distributed computing substrates
- Robust operation in the presence of unreliable and untrusted human and AI agents and computing platforms

Theme 4 : Overview and Representative Tasks




Research


Teaching

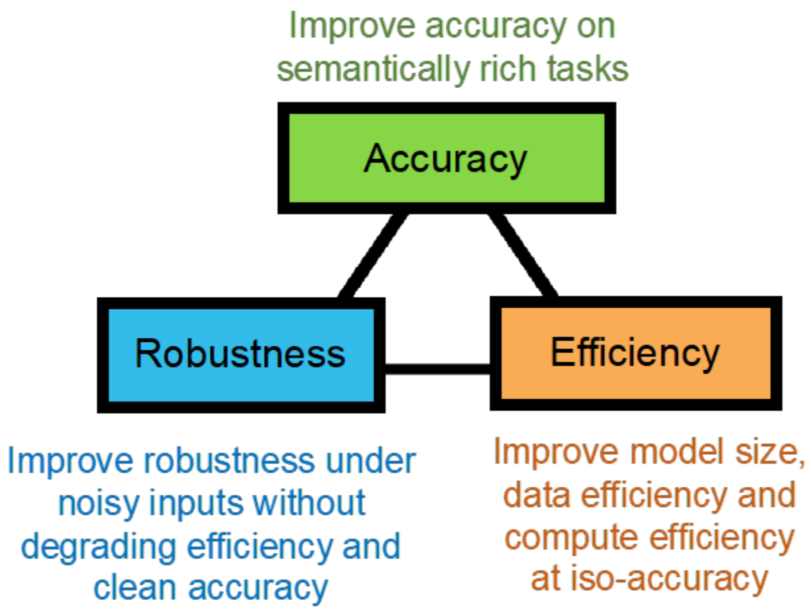

Creative

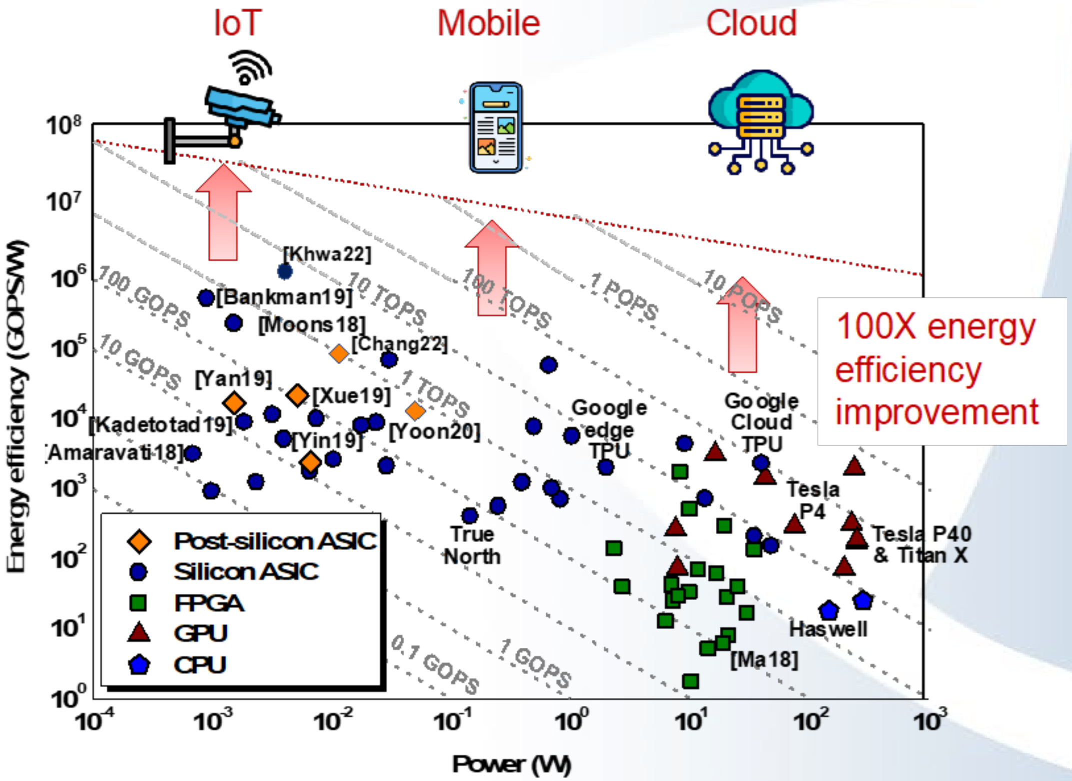

Medical

AI as a true partner for humans

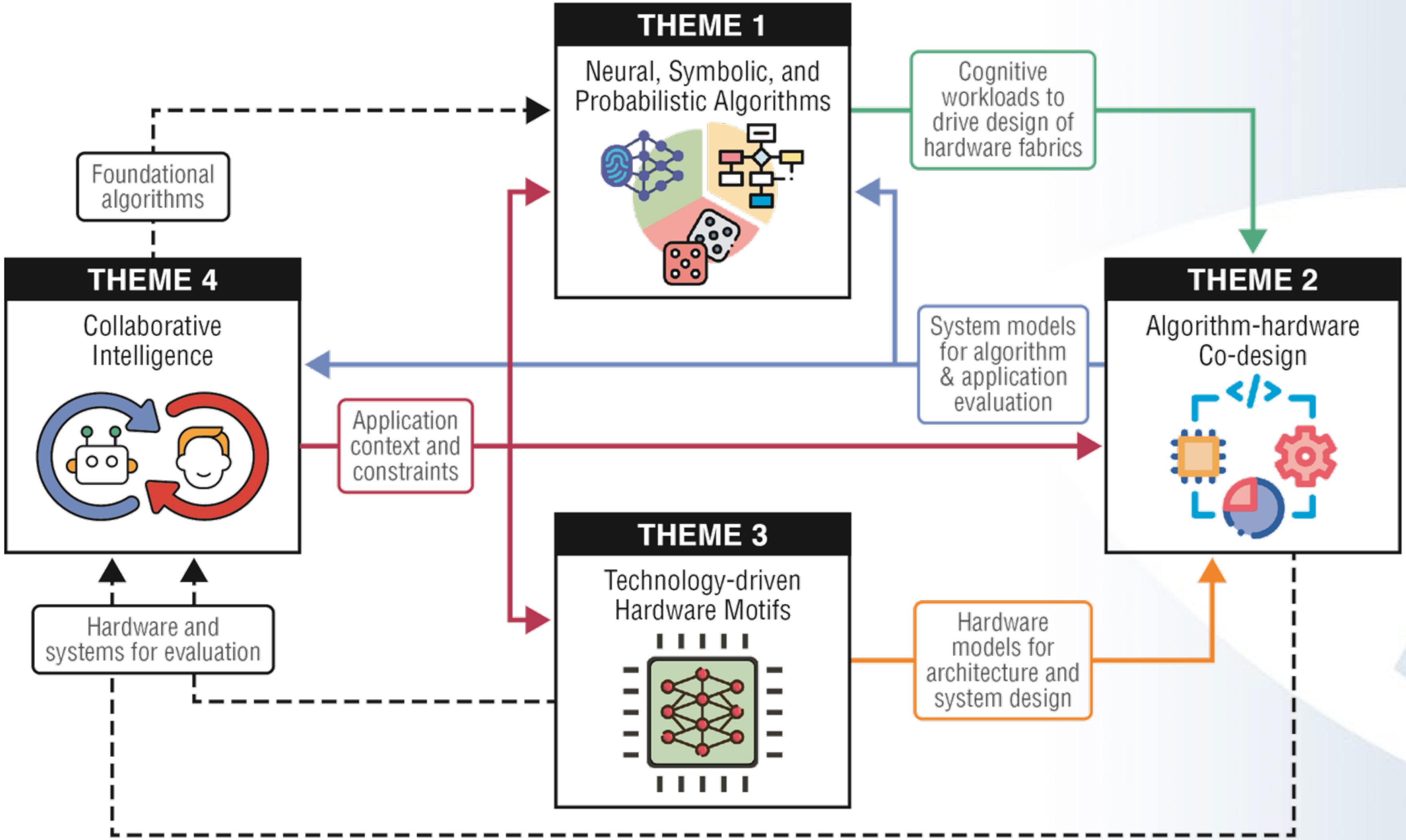
CoCoSys Grand Challenge

The grand challenge is to demonstrate end-to-end collaborative human-AI systems with quantum improvements in accuracy-robustness-efficiency metrics

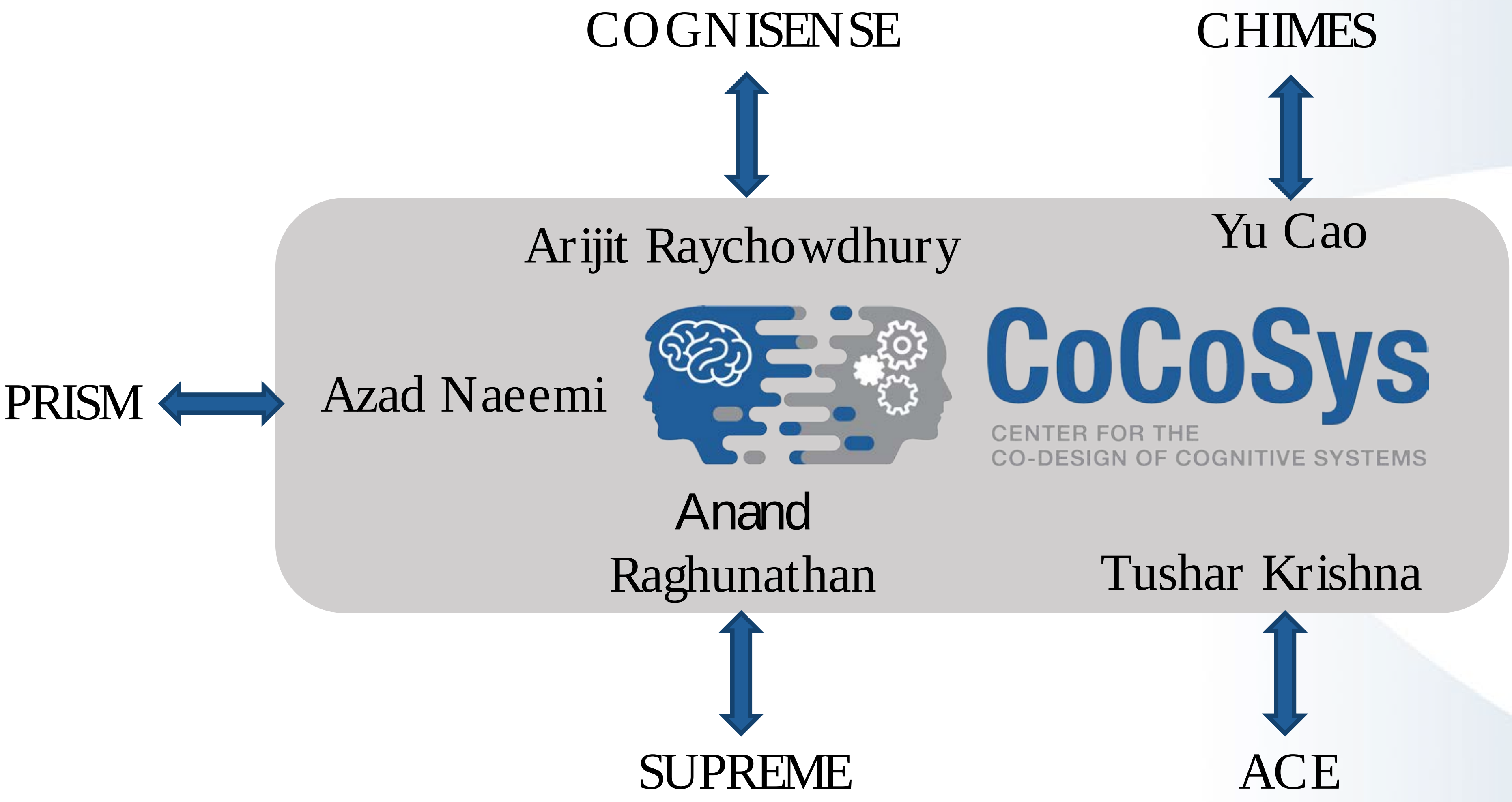




Fostering Intra-center Collaboration



Inter-center Collaboration



ACADEMIA • INDUSTRY • GOVERNMENT

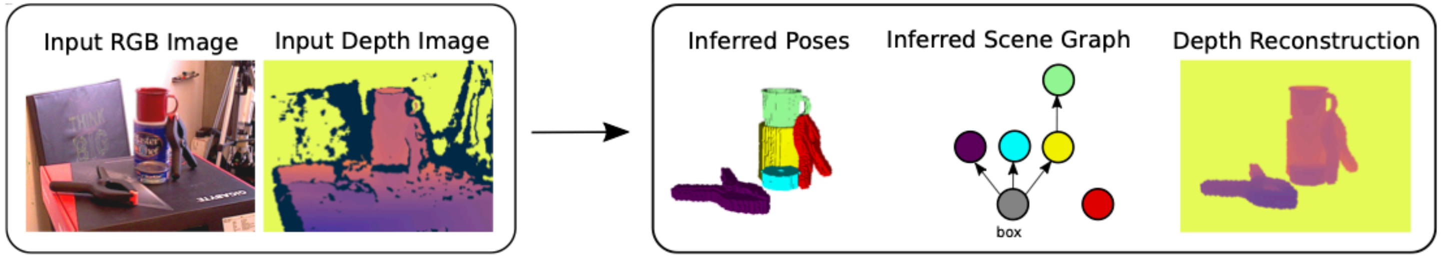
CoCoSys

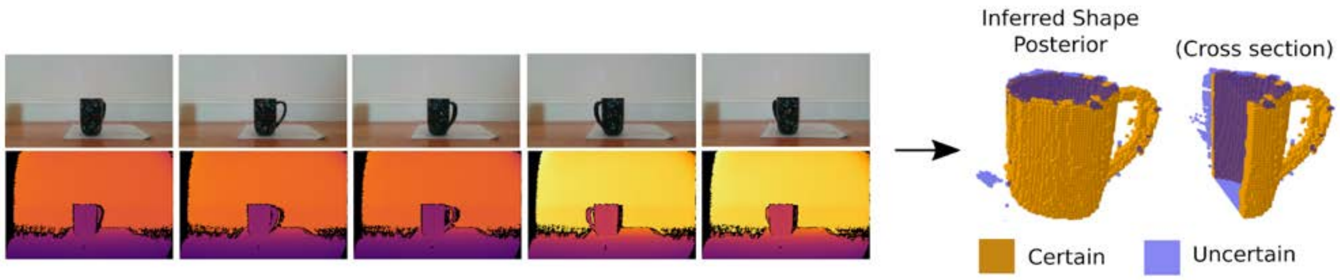
CENTER FOR THE
CO-DESIGN OF COGNITIVE SYSTEMS

Back-Up Slides

Theme 1: Neuro-Symbolic-Probabilistic Algorithms for 3D Scene Perception

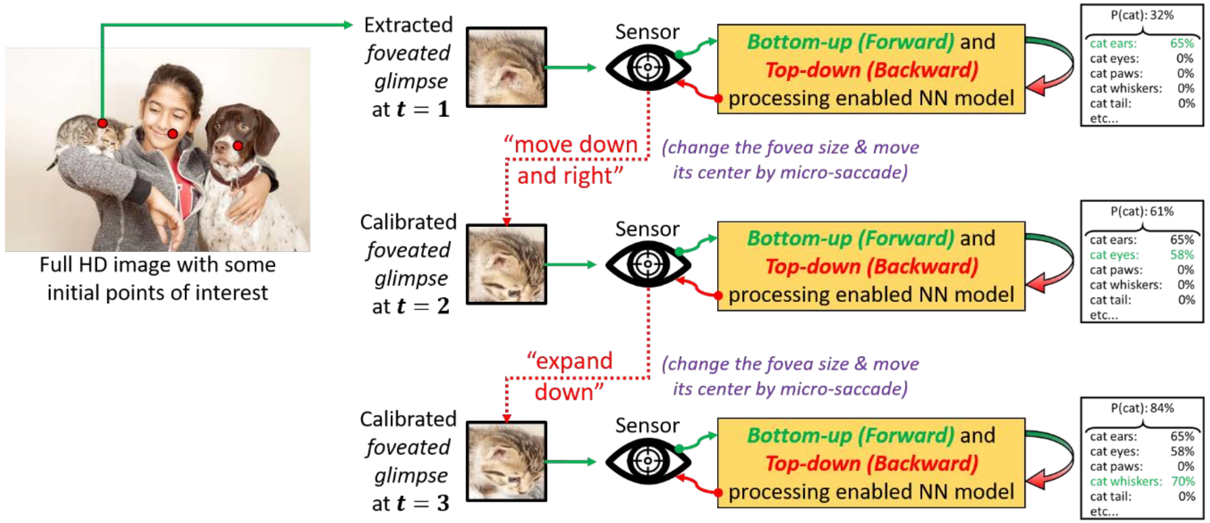
- Hybrids of neural, symbolic, and probabilistic models can simultaneously improve robustness and data efficiency in 3D scene perception
- Separation between "ventral" and "dorsal" functions (detection and location)

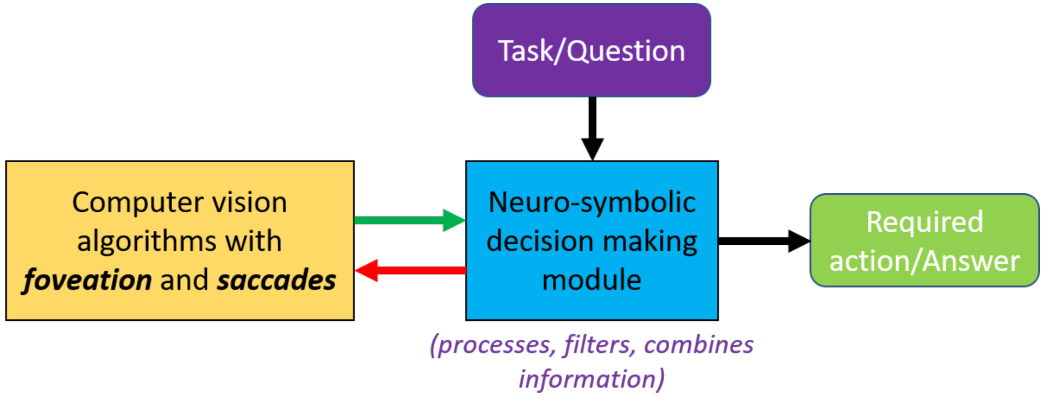




Theme 1: Top-Down and Bottom-Up Processing with Foveation based Learning

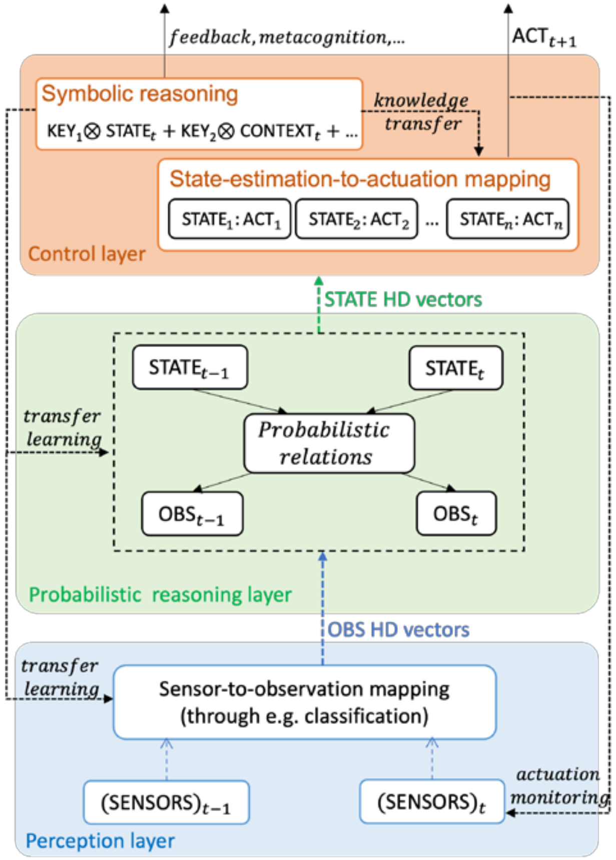
- Active vision system that employs an iterative method of processing relevant regions in the scene driven by top-down and bottom-up signals
  - Forward phase: Process selected patches of image and positional information to extract useful representations
  - Backward phase: Determine the location and size of next path (using RL)





Theme 1: Perception, Reasoning and Control with HD Computing

- Most prior efforts within HD computing focused on sensory and perceptual tasks
- Unified framework to design algorithms that can perform perception, reasoning and control with HD representations



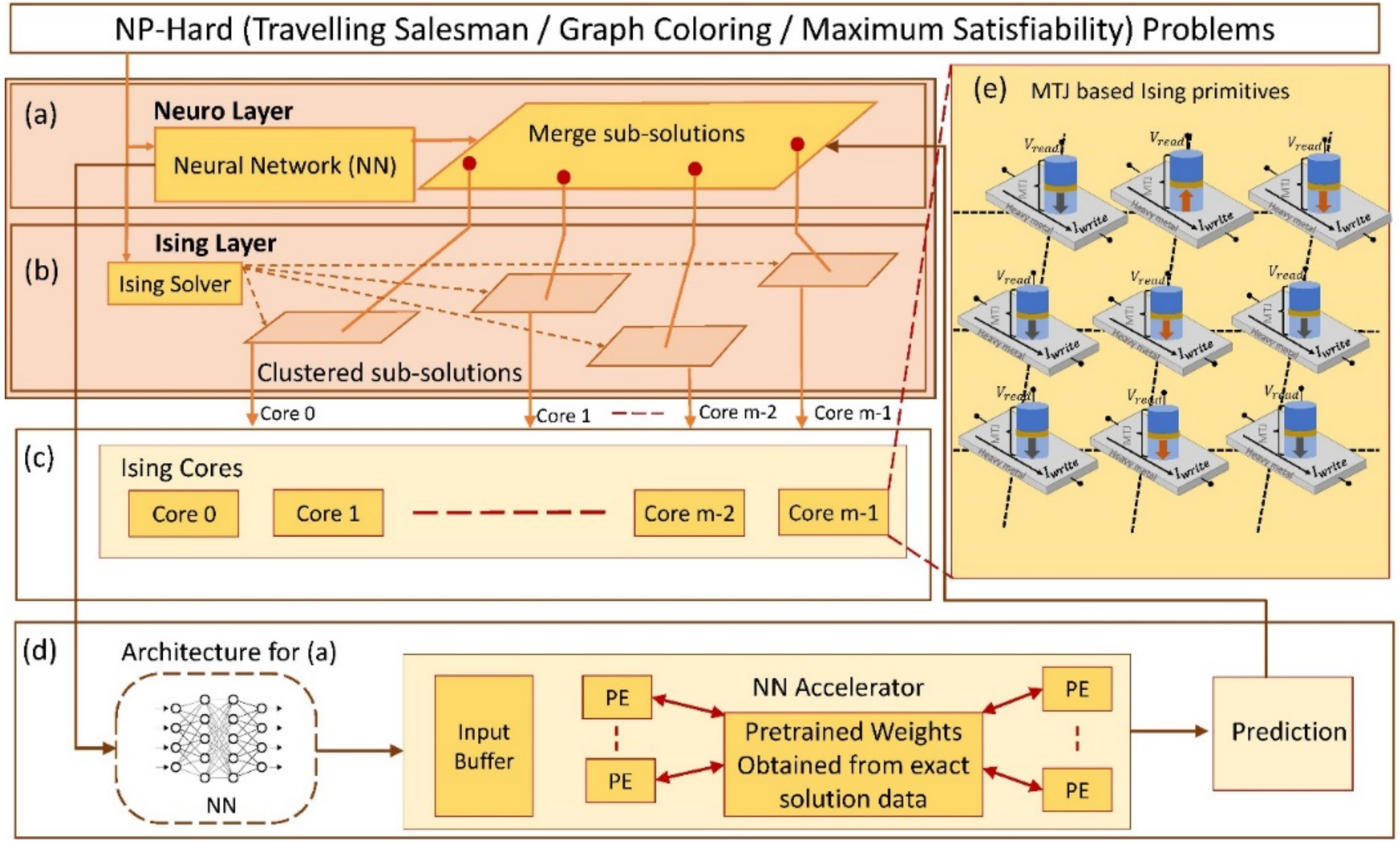
Control layer:
- Coordinate output actuation.
- (Symbolic) exogenous reasoning and feedback :
  - Reasoning across contextual changes (different location, new device, etc.)
  - Provide feedback to other layers.
- Reactive behavior recall, knowledge transfer through analogical reasoning, feedback, meta-cognition, etc.

Probabilistic reasoning layer:
- (Generative) model of the world.
- Reasoning through uncertain situations over time.
- Probabilistic Graphical Models, Probabilistic Circuits, probabilistic programs, etc.

Perception layer:
- Acquire and represent information (from e.g. sensors)
- Map to higher level representations.
- ML classification (e.g. DNNs), HD computing classification, etc.

Theme 1: Computing with Emergent and Dynamical Systems

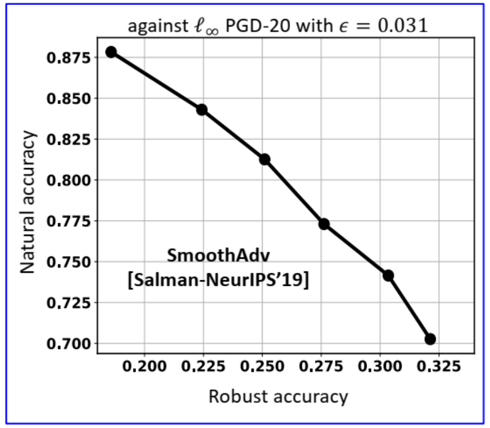
- A wide range of NP-hard problems can be mapped to Ising formulations
- Scalability is a key challenge
- Hybrid approach (Neuro+Ising) to find high-quality solutions by combining sub-problem results (Ising) using Neural methods



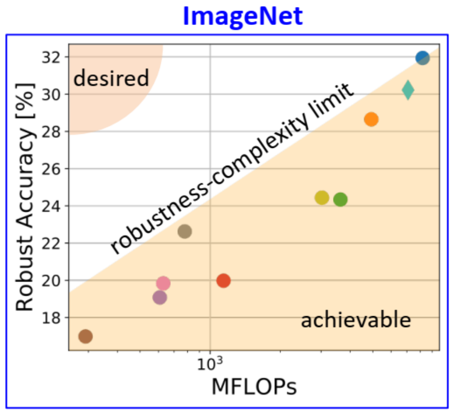
Theme 1: Theoretical Underpinnings of Accuracy-Robustness-Efficiency (ARE) Tradeoffs

- ARE tradeoffs are fundamental to cognitive systems but poorly understood
- Apply techniques from learning theory, signal processing and information theory to characterize the tradeoffs in cognitive systems

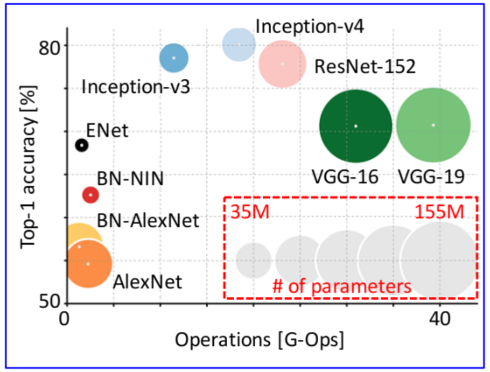
Accuracy vs. Robustness



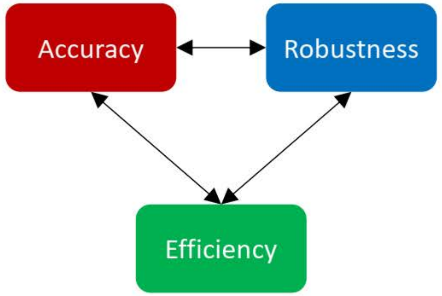
Robustness vs. Efficiency



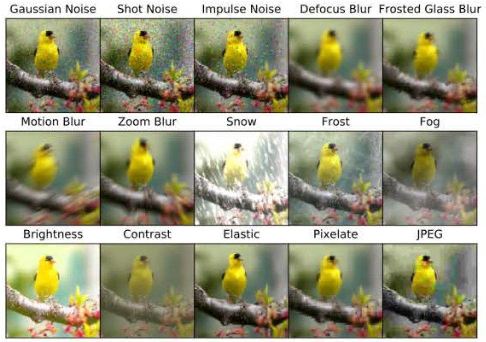
Accuracy vs. Efficiency



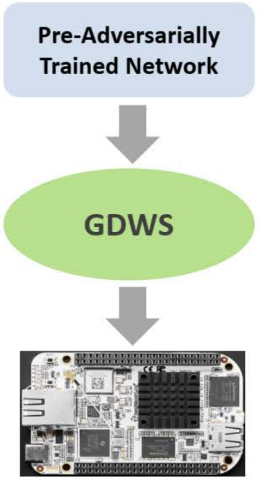
ARE trade-off



beyond adversarial vulnerabilities

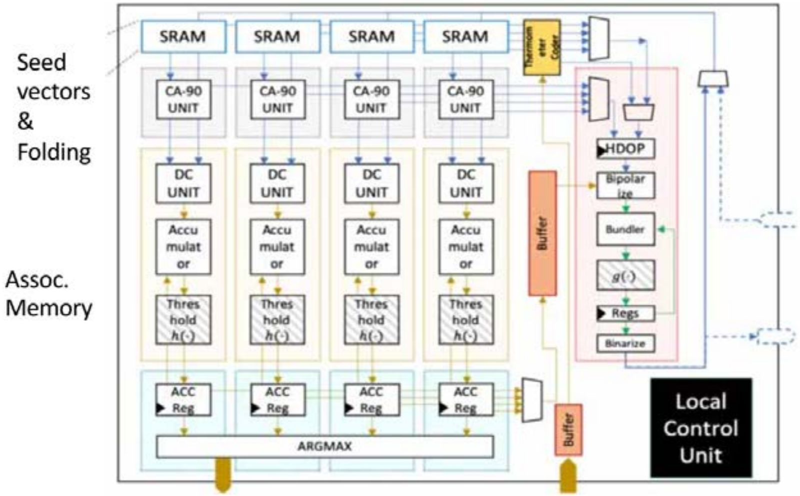


Real-life hardware

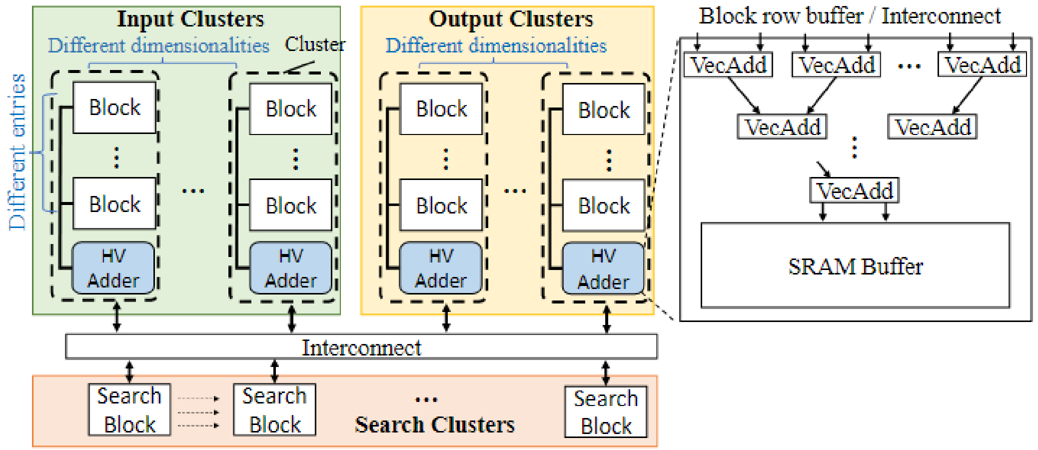


Theme 2: General-purpose HD Processors

- HD Computing enables low-complexity highly parallel and error-resilient hardware architectures
- CMOS and Beyond-CMOS processors for efficient HD computing and fusion with neural and probabilistic models



HPU: CMOS Processor for HD Computing

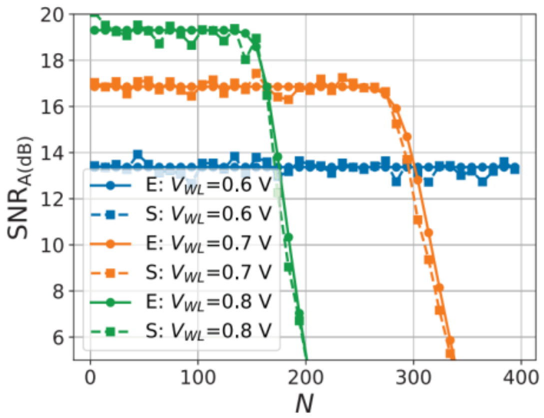


FeFET-based processor for HD computing on relational graphs

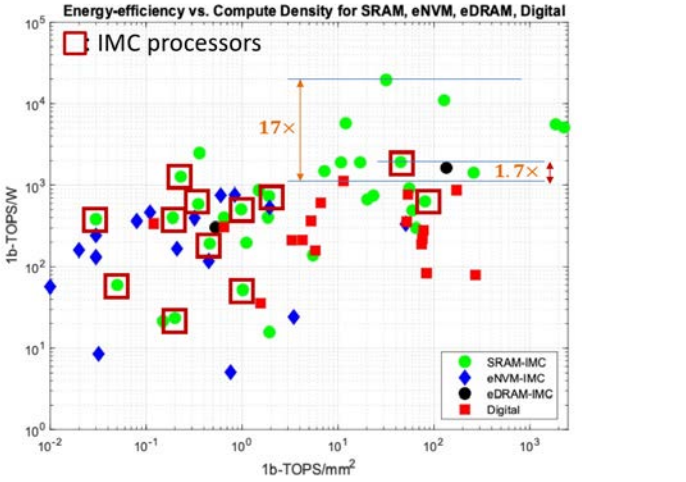
Theme 2: Limits of Latency-Energy-Accuracy for In-Memory Computing

- IMC has been a very active area of research over the past decade
- Analyze the fundamental limits on the latency-energy efficiency-accuracy (LEA) trade-off intrinsic to CMOS- and NVM-based CIMs
- Develop Shannon-inspired statistical error compensation methods to approach the LEA limits
- Determine metrics, benchmark published designs and compare them against the fundamental limits

limits

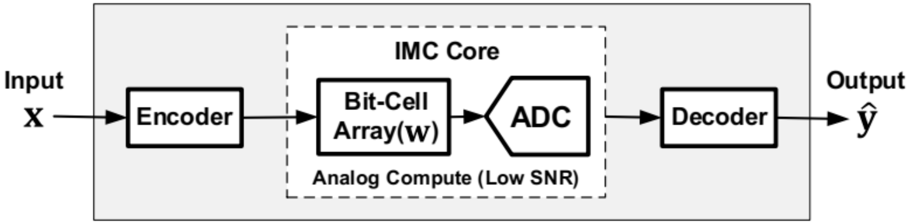


Benchmarking



https://github.com/naresh-shanbhag/UIUC-IMC-Benchmarking

SNR boost via SEC

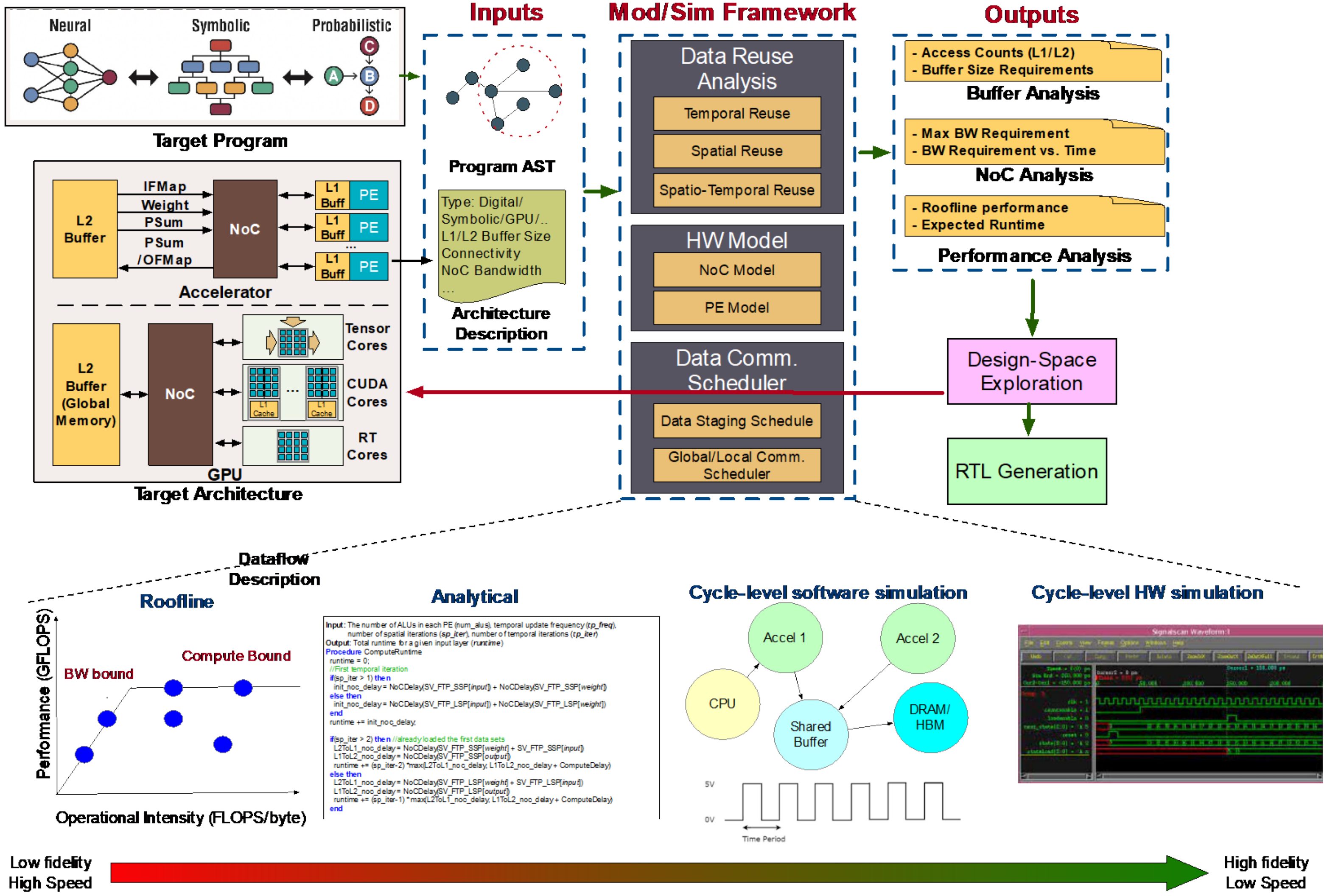


SRC

CoCoSys
CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

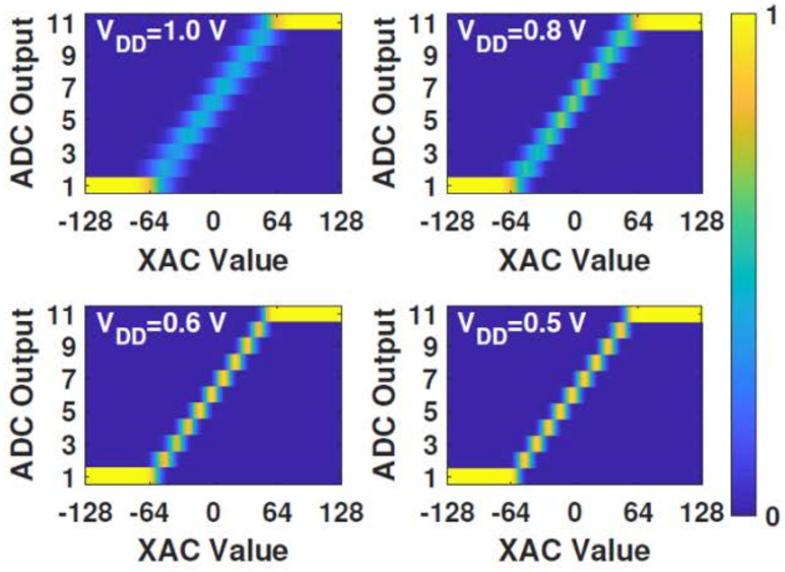
Theme 2: Modeling, Simulation and Exploration Framework

- Modeling and simulation framework to drive research within JUMP2.0 and the broader community towards the needs of future cognitive workloads

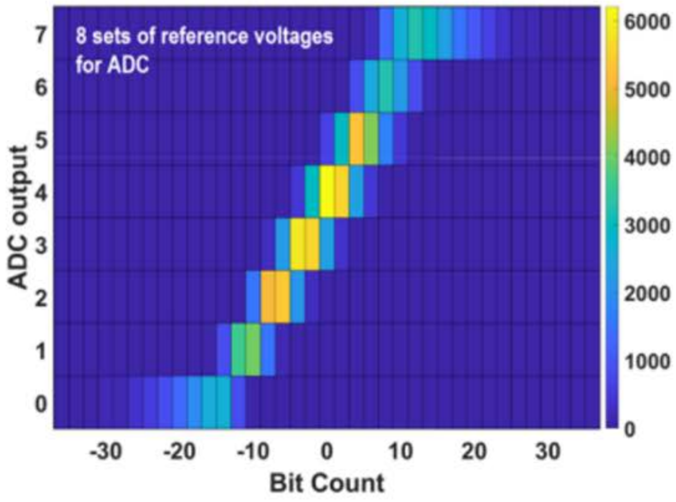


SRC
CoCoSys
CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

Theme 3 : Beyond Merged-Logic Memory Fabrics

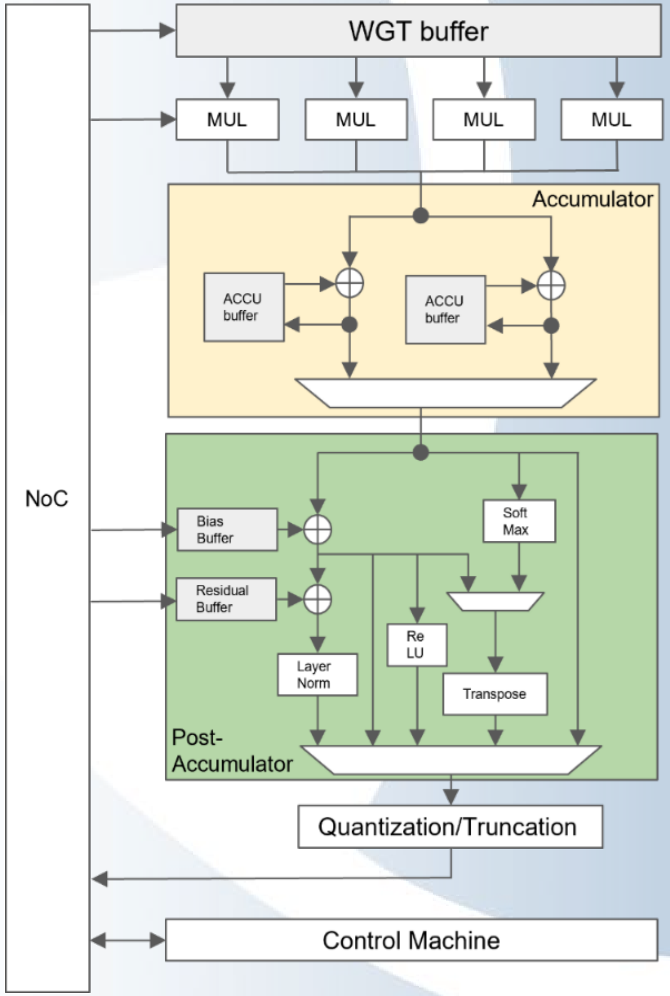


SRAM



RRAM

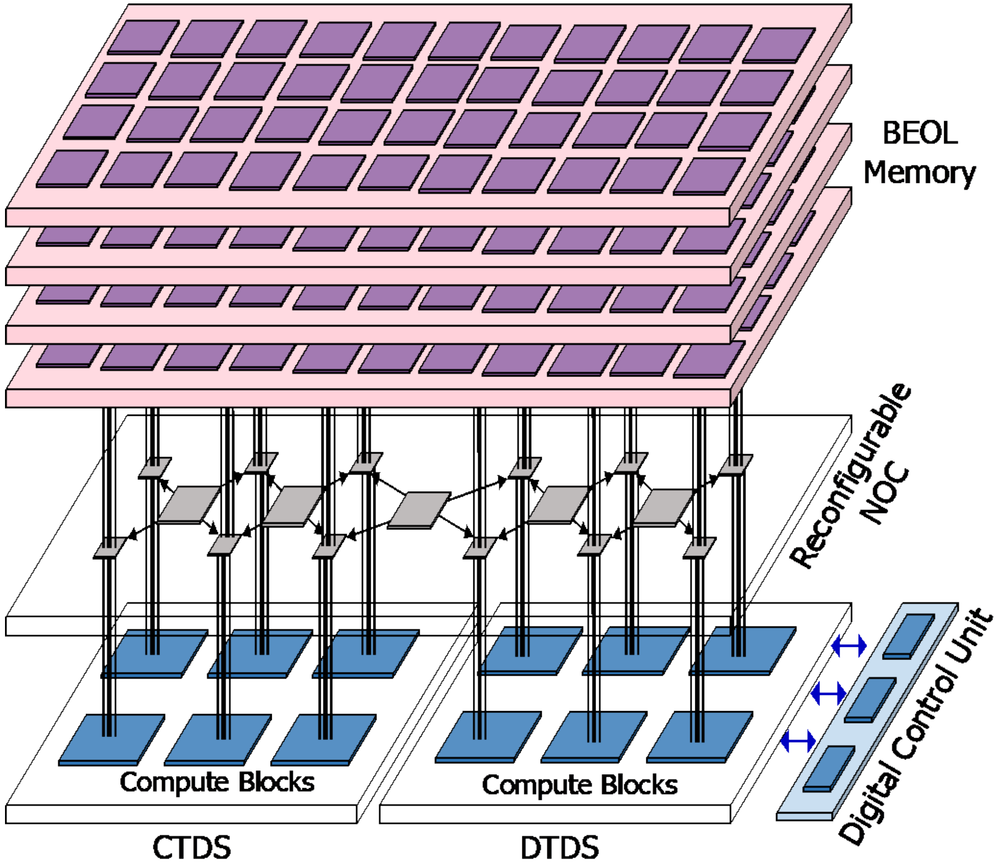
- Fundamental limitations of IMC in VMMs and beyond
- Exploiting noise for probabilistic inference
- Architectures to support emerging models

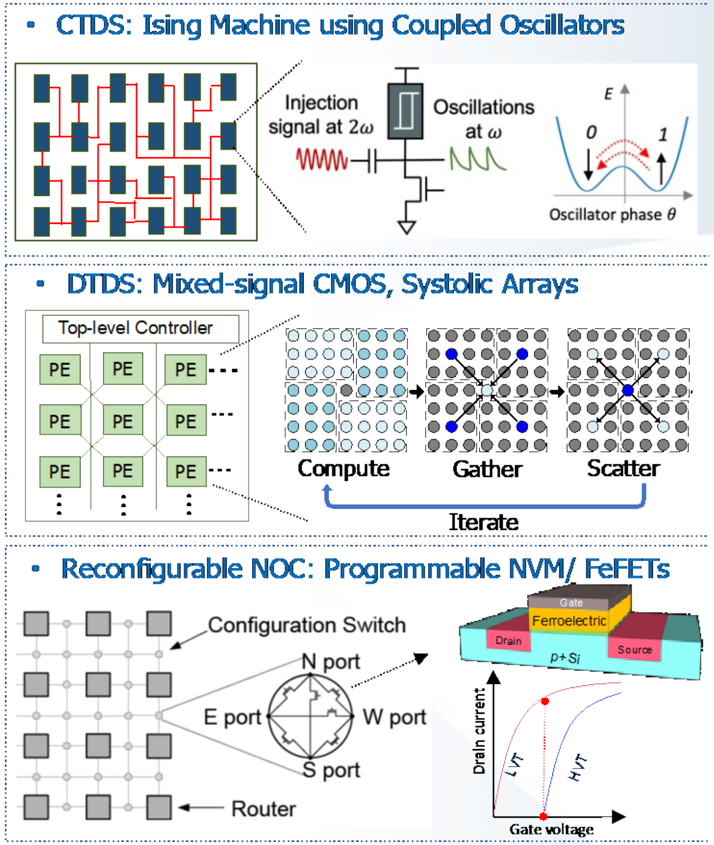


Transformer architecture exploiting column-wise sparsity

SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

Theme 3 : Continuous- & Discrete-Time Dynamics

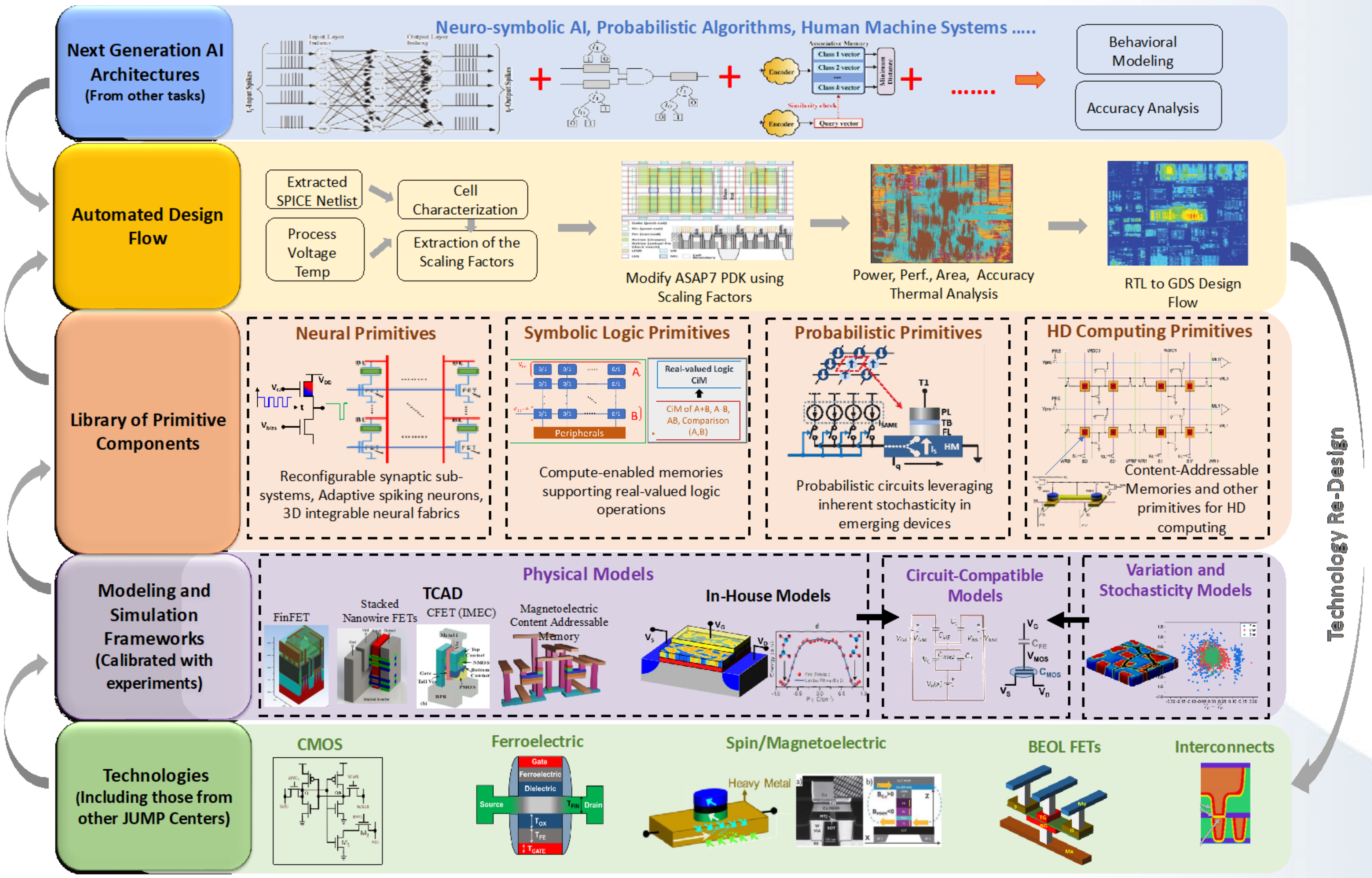




- Analog, mixed-signal, digital primitives in continuous- and discrete-time systems
- Vision of a heterogeneously integrated system that exploits spatio-temporal properties of algorithms

SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

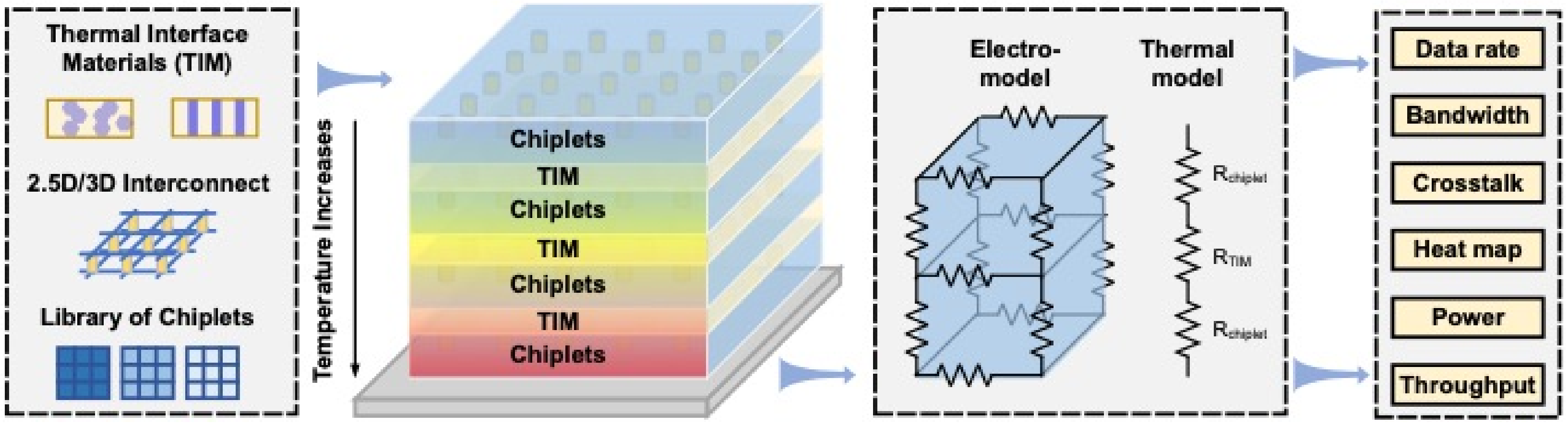
Theme 3 : Technology Evaluation



SRC

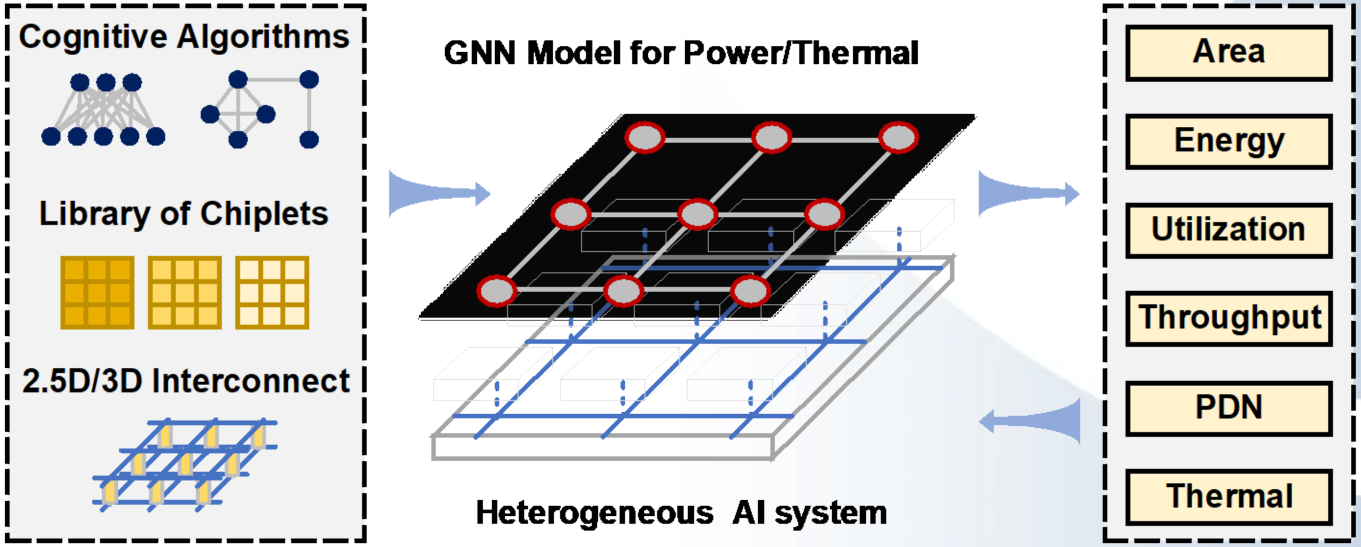
CoCoSys
CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

Theme 3 : HI and Packaging of AI hardware

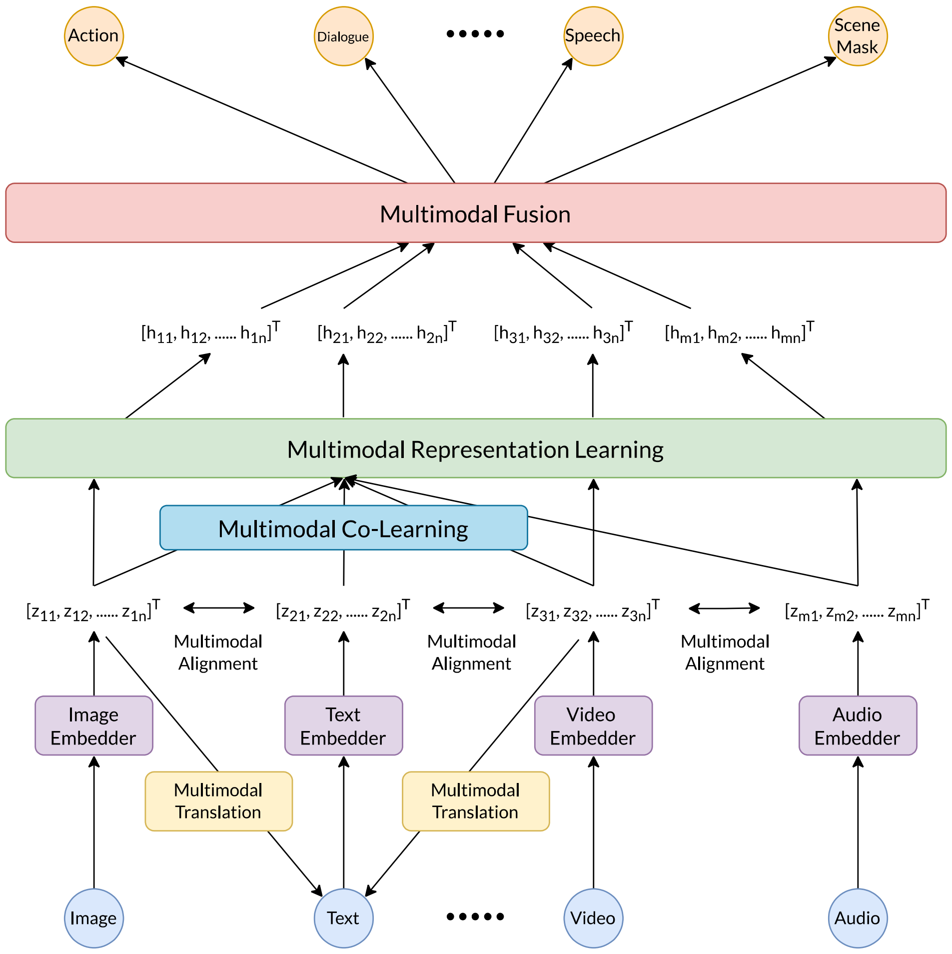


Electrothermal modeling of 2.5D and 3D architectures would enable accurate frameworks for PPA analysis

Closed loop architecture-HI exploration will provide optimization strategies for higher performance and energy-efficiency



Theme 4 : Always Helpful, Always Learning AI

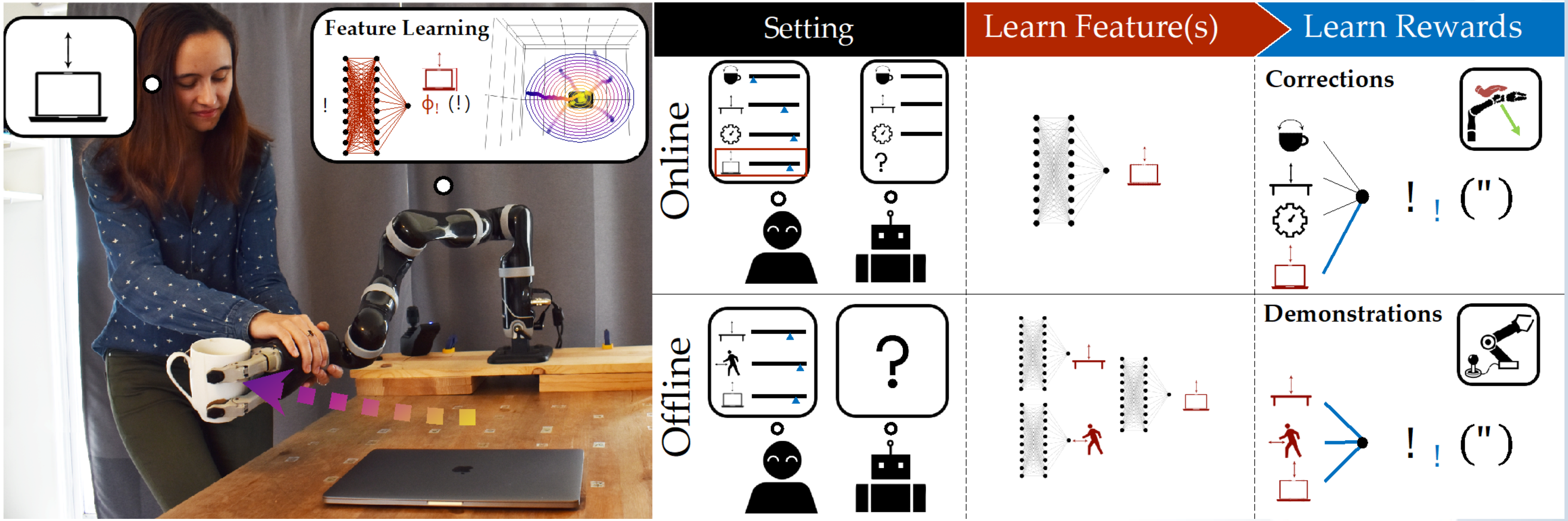


Conversational AI

Building a Digital Human – Open questions

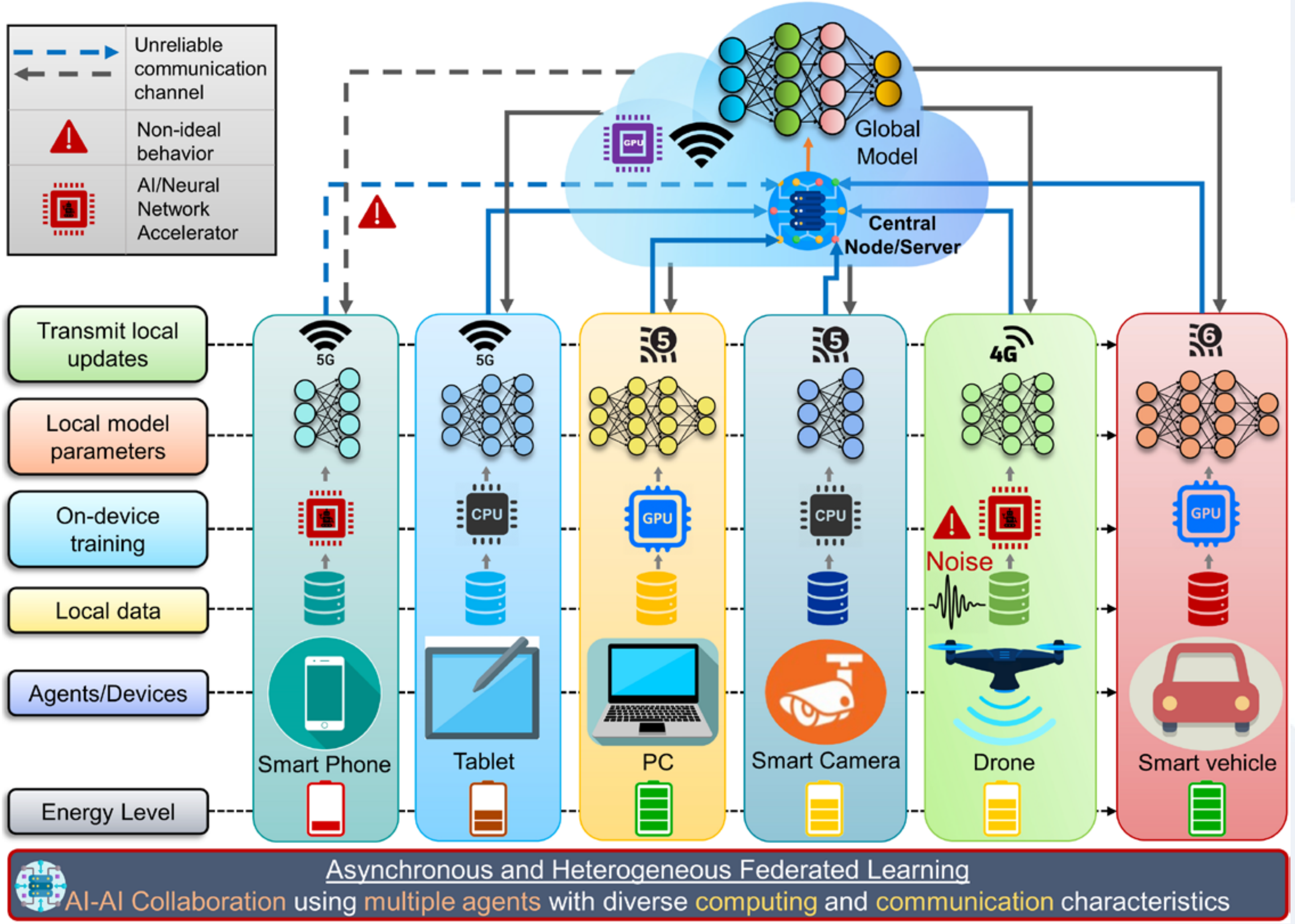
- How can we extend current SoTA end-to-end neural (deep learned) conversational response generation from language input/output to fully embodied digital humans?
- Can we achieve end-to-end learning of multimodal (language, vision, physical touch, gestures, facial expressions, audio, etc.) input/output?
- How do we make the digital human robust to failure? Will the introduction of the rich set of senses/expressiveness of digital humans change/amplify the effectiveness of multi-turn dialogue strategies to resolve ambiguities and clarify misunderstandings?

Theme 4 : Symbolic Models of Humans Objectives



- Unified Symbolic-Probabilistic Models of Humans Behavior
- Structure reward models by focusing specifically on learning causal representations on top of which to build rewards.
- Learning representations should not be done with the same general input meant to teach full reward functions - we should reinvent human input to be explicitly geared towards teaching the AI agent what matters causally

Theme 4 : Heterogeneous Federated Learning for AI-AI and Human-AI Collaboration



SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

CoCoSys events and engagement opportunities (2023)

- **CoCoSys Annual Review** on May 16-17 at Georgia Tech (register now on SRC's website)
- **Monthly Student Socials in Gather.Town** – we invite SRC leadership and industry liaisons to join our informal student gatherings.
- **Theme meetings** with CoCoSys task liaisons on Wednesdays at 11 AM ET

SRC CoCoSys CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

Beginning of a shared journey