A Berkeley View of Systems Challenges for AI

Given the increasingly sophisticated application environments of the AI system and the limitation of future data technology, this paper proposes several open research directions in system, architectures, and security to address the challenges for greater AI system deployment.

It first summarizes the factors contributing to the success of AI, including massive data quantities, development of computing technology, and broad technical accessibility (open source and public infrastructures).

## Trends and challenges

Visioning the large-scale systems and ML frameworks, along with security and hardware architectures, playing a more important role in enabling the broad adoption of AI, this paper presents the future challenges

1. Mission-critical AI

AI systems are required to continually adapt and learn new skills as the environment changes

1. Personalized AI

AI systems are expected to consider user behavior and preferences, while protecting users’ privacy and security.

1. AI across organization

Training data owned by multiple organization without compromising their confidentiality

1. AI demands outpacing the Moore’s Law

With data continuing to grow exponentially, domain-specific architectures and hardware are required to address the performance needs of future AI applications.

## Research Opportunities

1. Dynamic environments
   1. Continual learning (Reinforcement Learning, but not yet widescale, and new system required) – building systems for RL that fully exploit parallelism
   2. Robust decisions (mission-critical applications)
   3. Explainable decisions (interactive diagnostic analysis with replay-able past execution)
2. Secure AI
   1. Secure enclaves
   2. Adversarial learning (learn from attack)
   3. Shared learning on confidential data (learn across multiple sources without leaking information from a data source during training or serving)
3. AI-specific architectures
   1. Domain specific hardware
   2. Composable AI systems (model composition in modular and flexible manner)
   3. Cloud-edge systems (design systems that are well suited to partitioned execution across the cloud and edge systems)