Architecture 2.0: Challenges and Opportunities with ML-Aided Design

Vijay Janapa Reddi | **Amir Yazdanbakhsh**
Harvard University | Google DeepMind

Acknowledgements

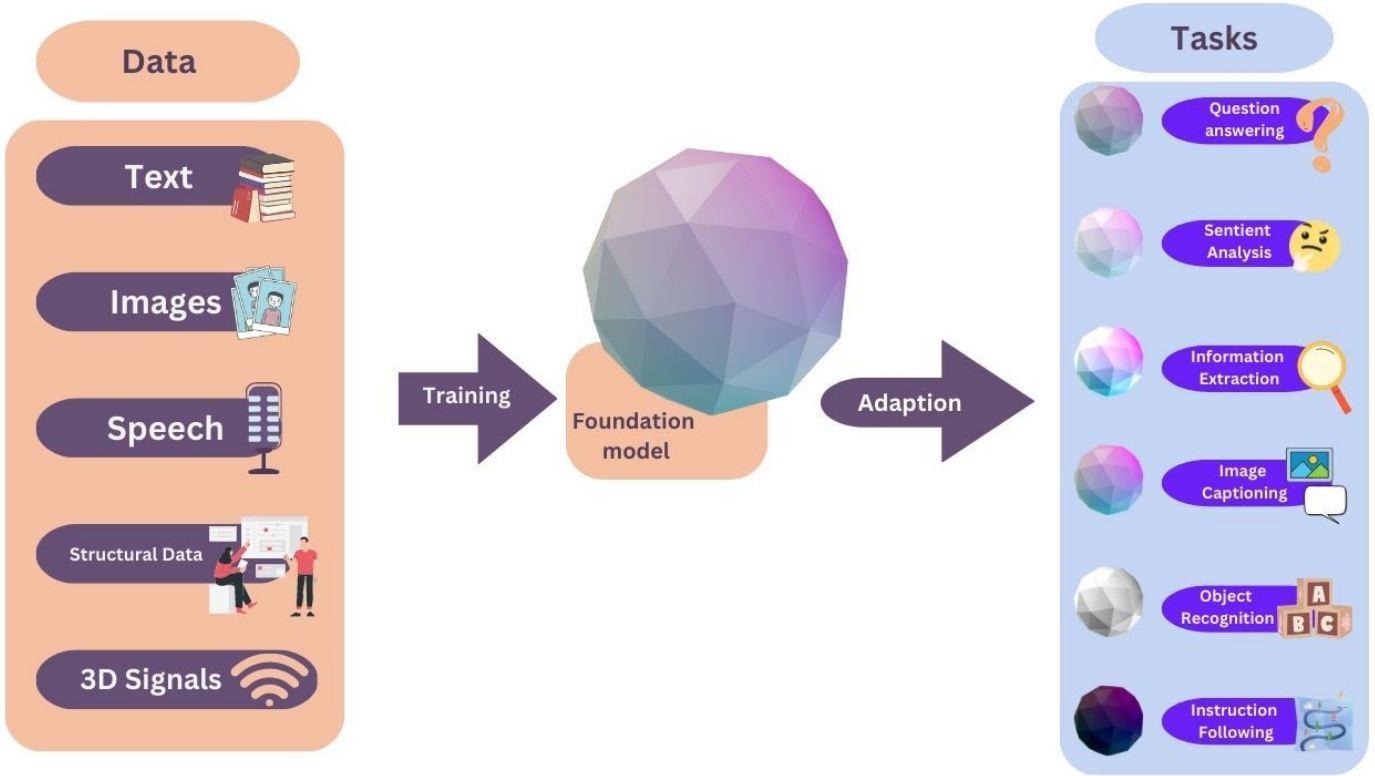
Architecture 2.0

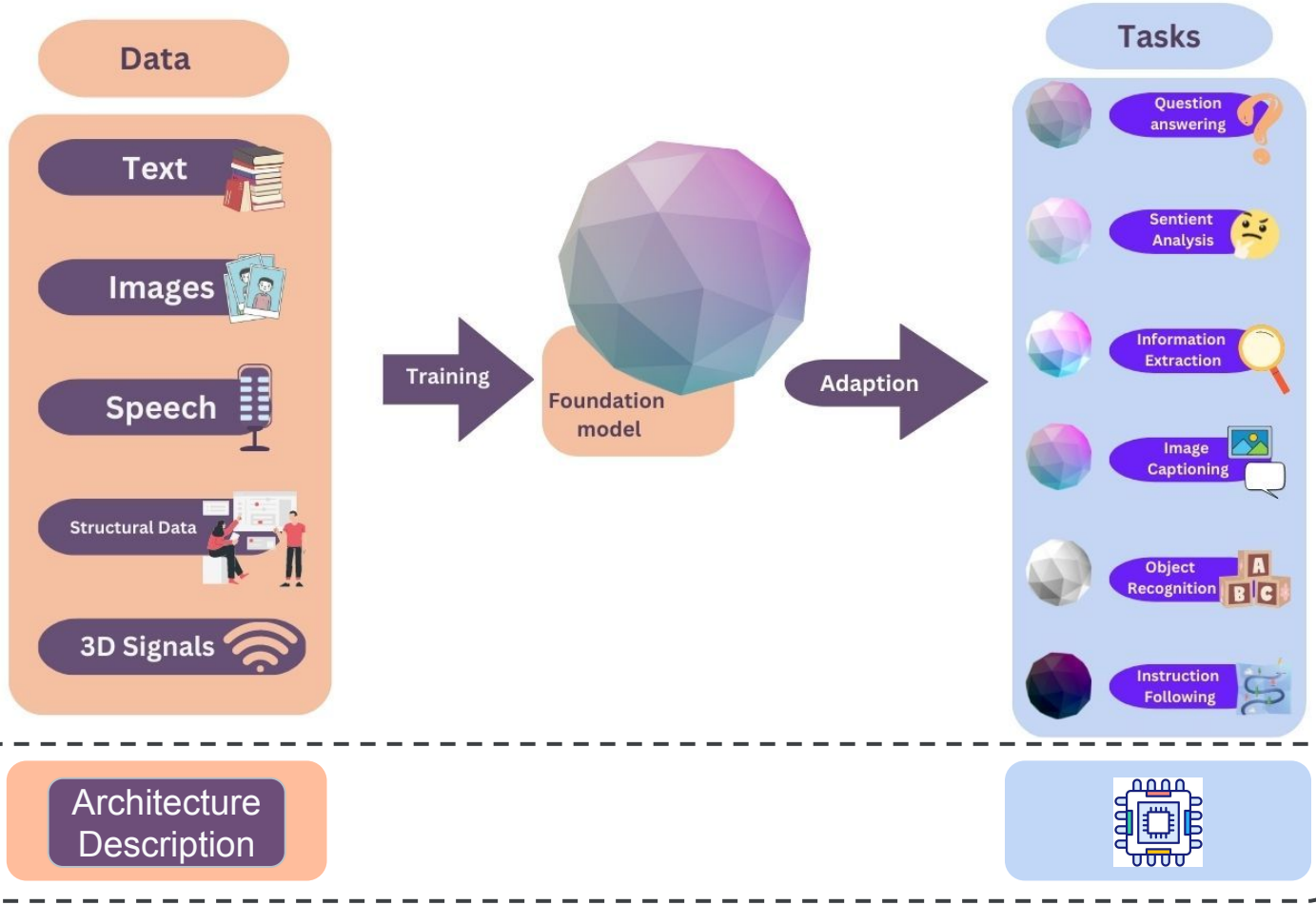
The era when we use AI/ML methods to

(1) minimize human intervention,

(2) build complex, efficient systems,

(3) in a shorter time frame.


Data
Text
Images
Speech
Structural Data
3D Signals
Training
Foundation model
Adaption
Tasks
Question answering
Sentient Analysis
Information Extraction
Image Captioning
Object Recognition
Instruction Following


Data
Text
Images
Speech
Structural Data
3D Signals
Training
Foundation model
Adaption
Tasks
Question answering
Sentient Analysis
Information Extraction
Image Captioning
Object Recognition
Instruction Following
Architecture Description

“Act like an architect — design me a custom 64-bit RISC-V processor with full vector extension support and optimize it for less than 3 Watt TDP in a 5 nm LP process node using the TSMC plugin library”

"... while you are at it add a few custom functional units that optimize the experience of XRBench [Hyoukjun et al. MLSys'23]"

"... and don't forget to generate all the unit test cases to verify the design and explain the design choices."

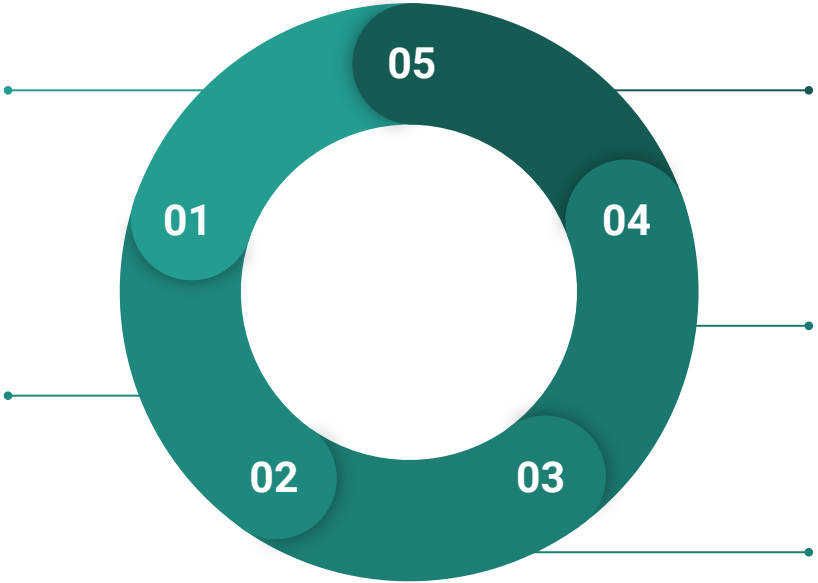
Challenges

Datasets

What datasets do we need? How we should collect these datasets for architecture research? What metadata should the datasets contain to enable broad usage? How do we create standard data formats from any ML algorithm?

ML Algorithms

How can we learn and apply new ML algorithms to effectively design high-performance/efficient systems? How do we make our community more accessible to ML researchers? How do we embrace ML algorithm design as part of architecture research?



Workforce & Training

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

Tools & Infrastructure

How do we reduce the sim2real gap? What instrumentation mechanisms do we need for creating the datasets? What gym environments do we need to enable data-centric AI? How do we define standard data formats for interoperability?

Best Practices

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

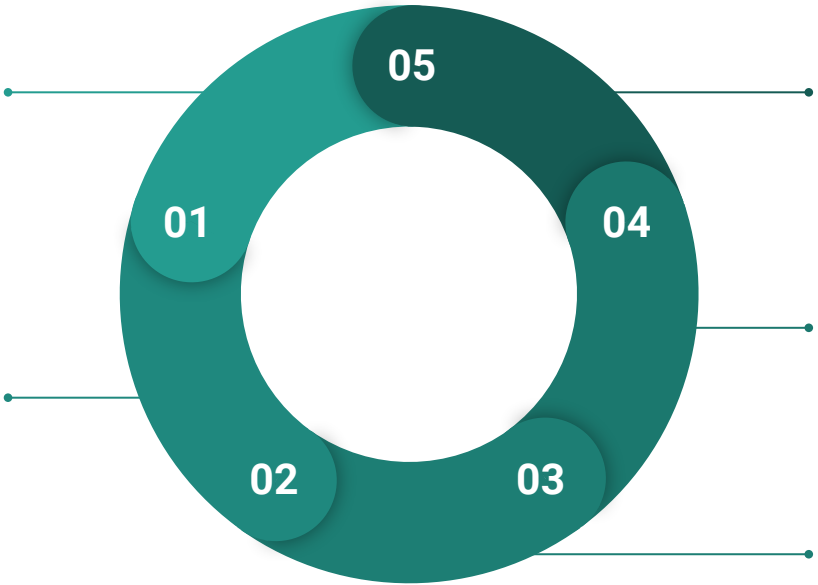
Challenges

Datasets

What datasets do we need? How we should collect these datasets for architecture research? What metadata should the datasets contain to enable broad usage? How do we create standard data formats from any ML algorithm?

ML Algorithms

How can we learn and apply new ML algorithms to effectively design high-performance/efficient systems? How do we make our community more accessible to ML researchers? How do we embrace ML algorithm design as part of architecture research?



Workforce & Training

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

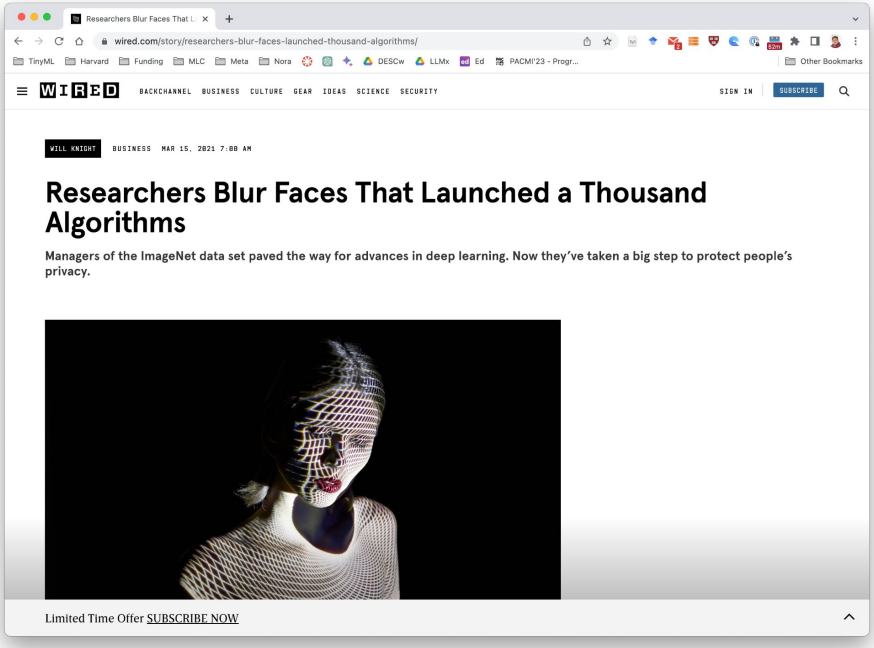
Tools & Infrastructure

How do we reduce the sim2real gap? What instrumentation mechanisms do we need for creating the datasets? What gym environments do we need to enable data-centric AI? How do we define standard data formats for interoperability?

Best Practices

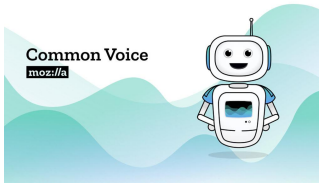
Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

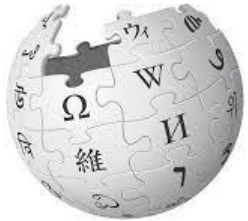
Lack of large, high-quality public datasets



- Need public data, but data needs to be held private
- Need to strike a safe balance

Inability to "scrape" the internet for creating public datasets

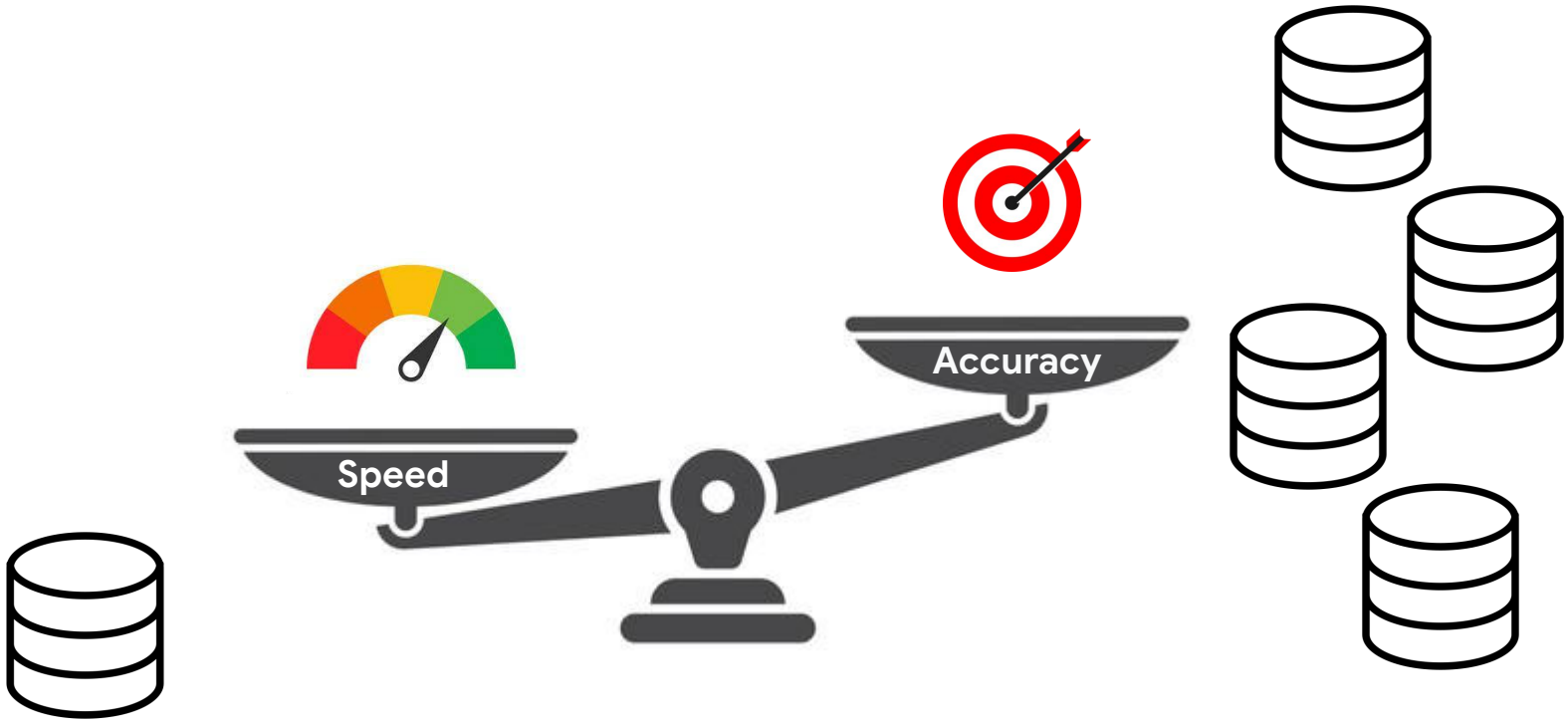




Common Crawl

GitHub

Data generation from cycle-level/accurate simulators is slow and difficult



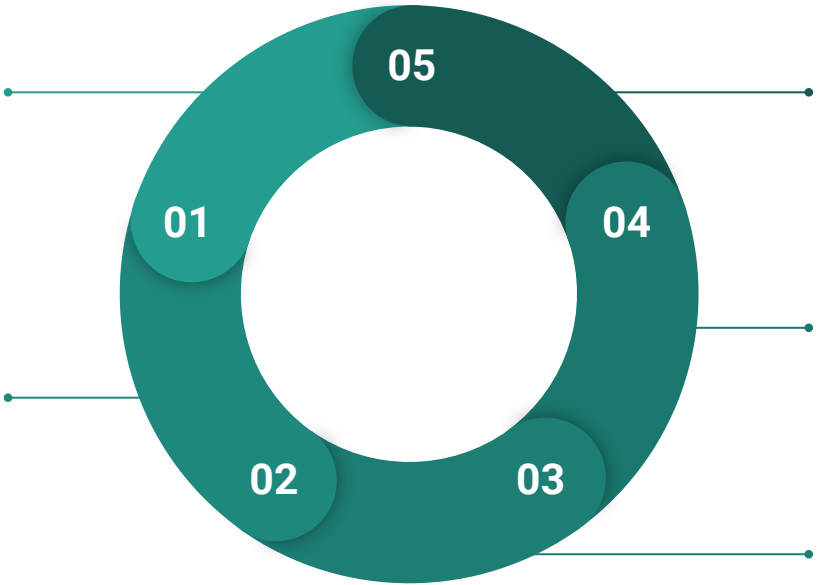
Challenges

Datasets

What datasets do we need? How we should collect these datasets for architecture research? What metadata should the datasets contain to enable broad usage? How do we create standard data formats from any ML algorithm?

ML Algorithms

How can we learn and apply new ML algorithms to effectively design high-performance/efficient systems? How do we make our community more accessible to ML researchers? How do we embrace ML algorithm design as part of architecture research?



Workforce & Training

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

Tools & Infrastructure

How do we reduce the sim2real gap? What instrumentation mechanisms do we need for creating the datasets? What gym environments do we need to enable data-centric AI? How do we define standard data formats for interoperability?

Best Practices

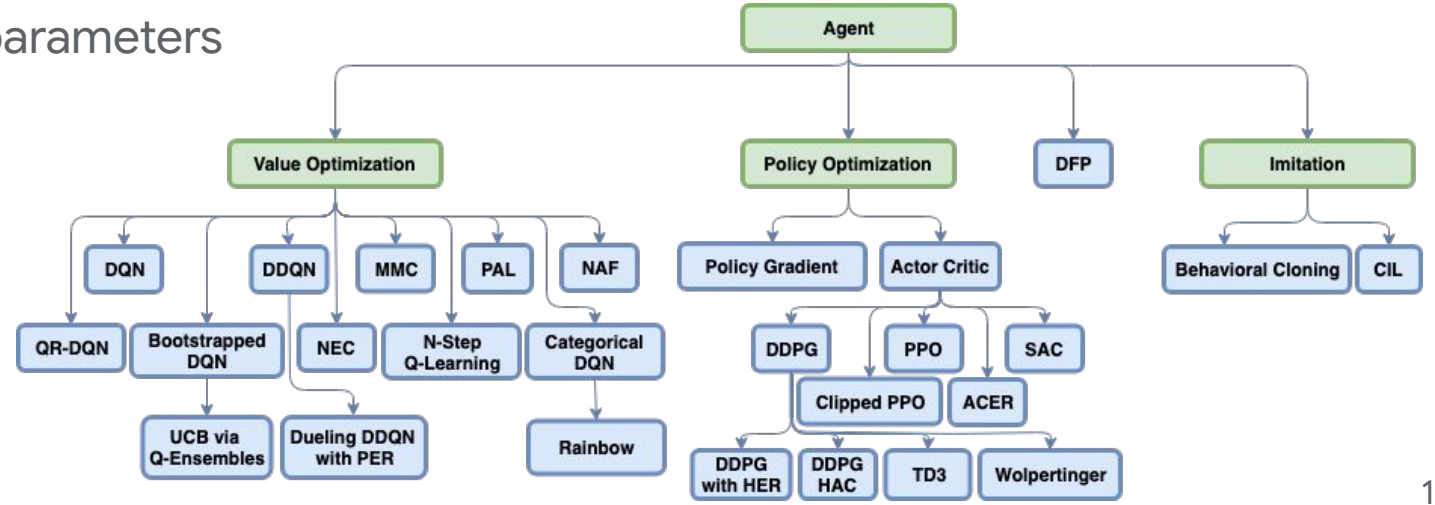
Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

Rapidly evolving ML algorithms landscape

- Many different algorithms out in the wild to choose from
- How do we know which algorithm is best suited for which architecture problem
- How do we compare these algorithms fairly against one another

Rapidly evolving ML algorithms landscape

- Take RL for example
  - Many different variants exist
  - New algorithms emerging
  - Hyperparameters



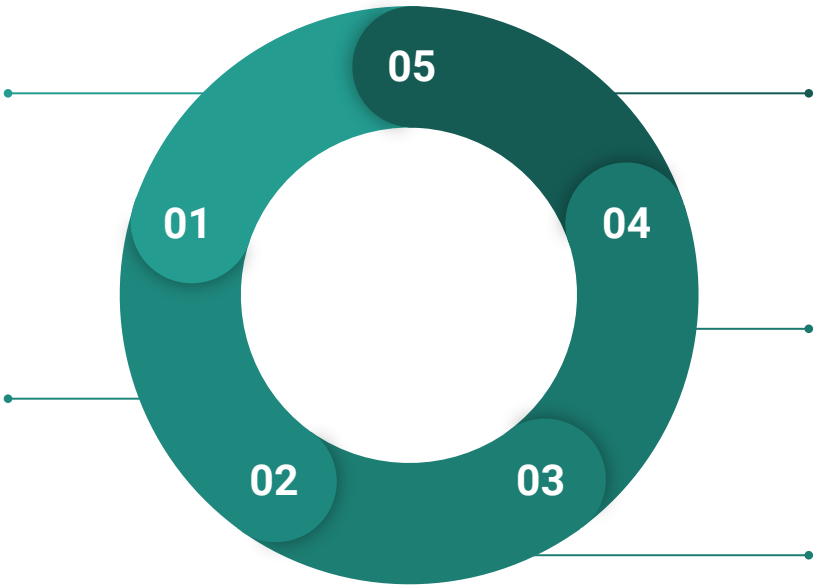
Challenges

Datasets

What datasets do we need? How we should collect these datasets for architecture research? What metadata should the datasets contain to enable broad usage? How do we create standard data formats from any ML algorithm?

ML Algorithms

How can we learn and apply new ML algorithms to effectively design high-performance/efficient systems? How do we make our community more accessible to ML researchers? How do we embrace ML algorithm design as part of architecture research?



Workforce & Training

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

Tools & Infrastructure

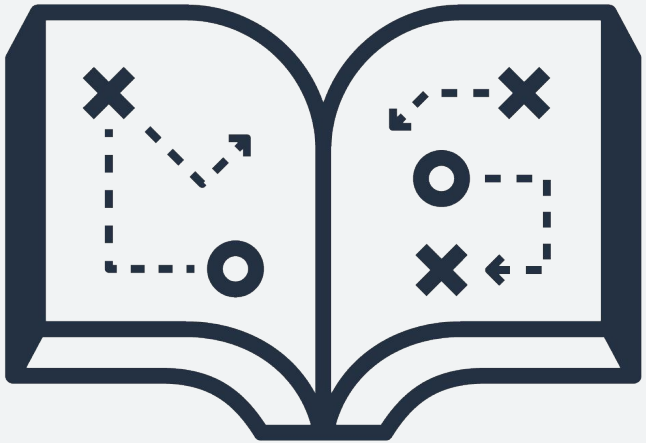
How do we reduce the sim2real gap? What instrumentation mechanisms do we need for creating the datasets? What gym environments do we need to enable data-centric AI? How do we define standard data formats for interoperability?

Best Practices

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

ML for Systems

- Problem suitability
  - High or low-dimensionality
- Deployment constraints
  - Latency
  - Space/time overheads
  - Hardware
  - Risk/robustness/interpretability
- Data availability
  - Privacy/security
  - Distribution shifts



Difficulty with verifying, validating, and interpreting ML algorithms

Task Performance	System Performance	Reliability	Generalization	Cost
*How well does the agent perform **the task it was trained for**?*	*What are the **compute requirements** needed to train and deploy the agent?*	*How **stable is the agent's performance** during training and inference?*	*How well does the agent perform on **outside tasks** of what it was trained on?*	*What are the **trade-offs** between using the various ML methods?*

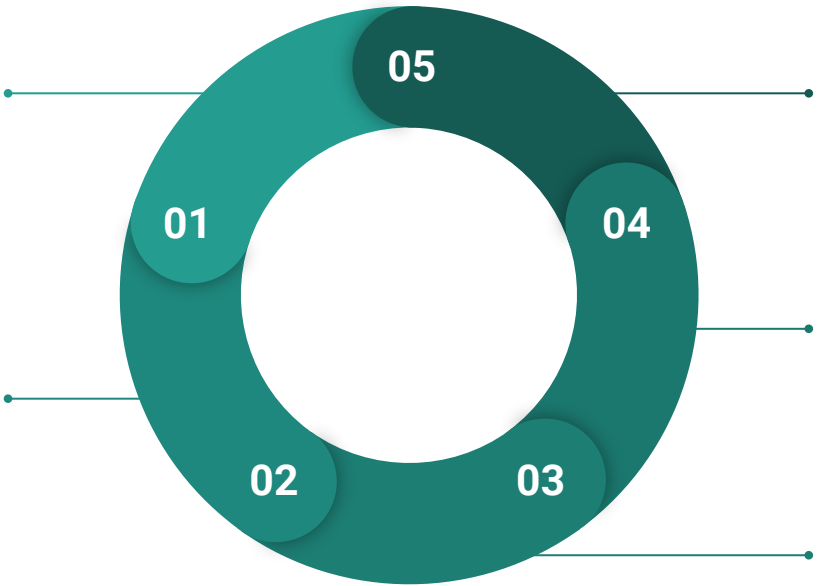
Challenges

Datasets

What datasets do we need? How we should collect these datasets for architecture research? What metadata should the datasets contain to enable broad usage? How do we create standard data formats from any ML algorithm?

ML Algorithms

How can we learn and apply new ML algorithms to effectively design high-performance/efficient systems? How do we make our community more accessible to ML researchers? How do we embrace ML algorithm design as part of architecture research?



Workforce & Training

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

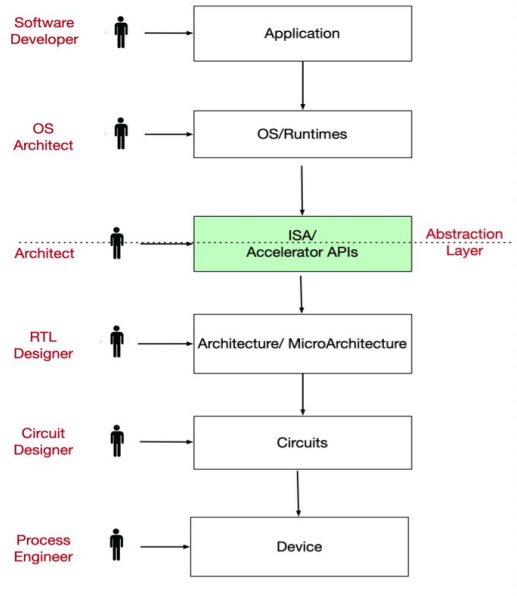
Tools & Infrastructure

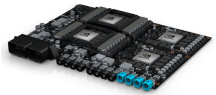
How do we reduce the sim2real gap? What instrumentation mechanisms do we need for creating the datasets? What gym environments do we need to enable data-centric AI? How do we define standard data formats for interoperability?

Best Practices

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

Full-stack Co-Design



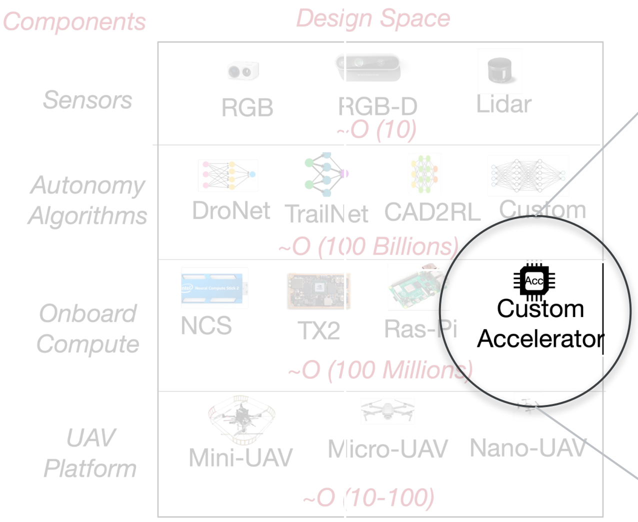


Full-stack Co-Design

Components	Design Space		
Sensors	RGB	RGB-D	Lidar
		~O (10)	
Autonomy Algorithms	DroNet	TrailNet	CAD2RL	Custom
		~O (100 Billions)	
Onboard Compute	NCS	TX2	Ras-Pi	Custom Accelerator
		~O (100 Millions)	
UAV Platform	Mini-UAV	Micro-UAV	Nano-UAV
		~O (10-100)	

Full-stack Co-Design

Large Design Space

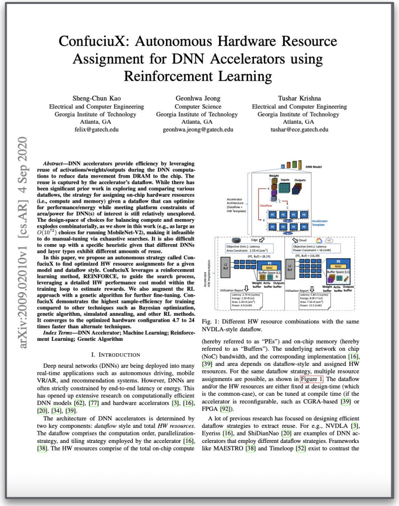


Parameters

~ 10^{14} to 10^{2300}

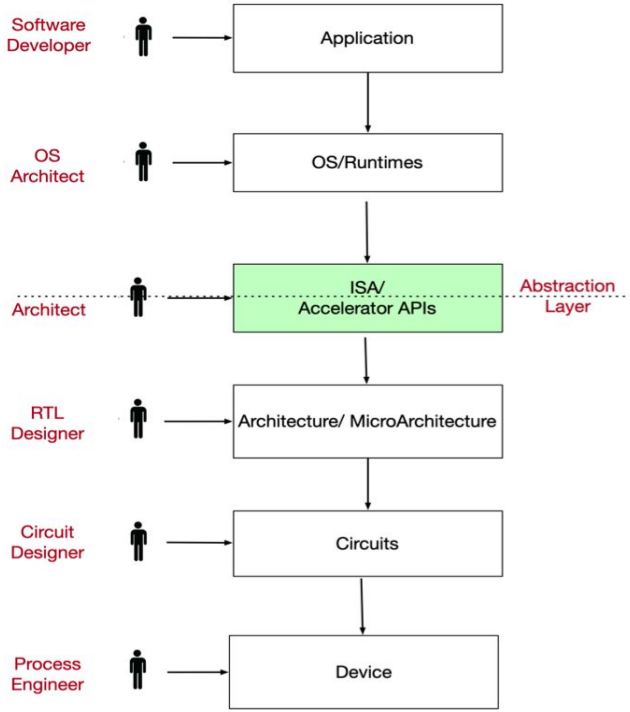

A Full-Stack Search Technique for Domain Optimized Deep Learning Accelerators

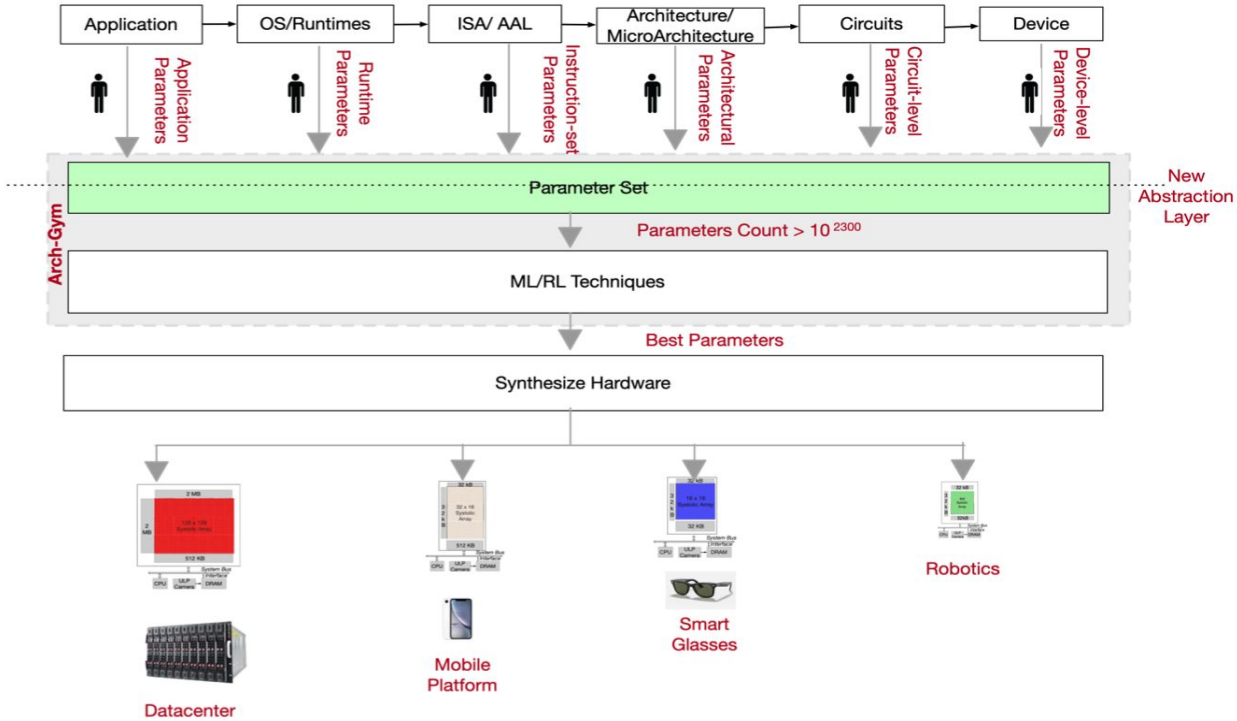
Dan Zhang, Safeen Huda, Ebrahim Songhori, Kartik Prabhu, Quoc Le, Anna Goldie, Azalia Mirhoseini


ConfuciuX: Autonomous Hardware Resource Assignment for DNN Accelerators using Reinforcement Learning

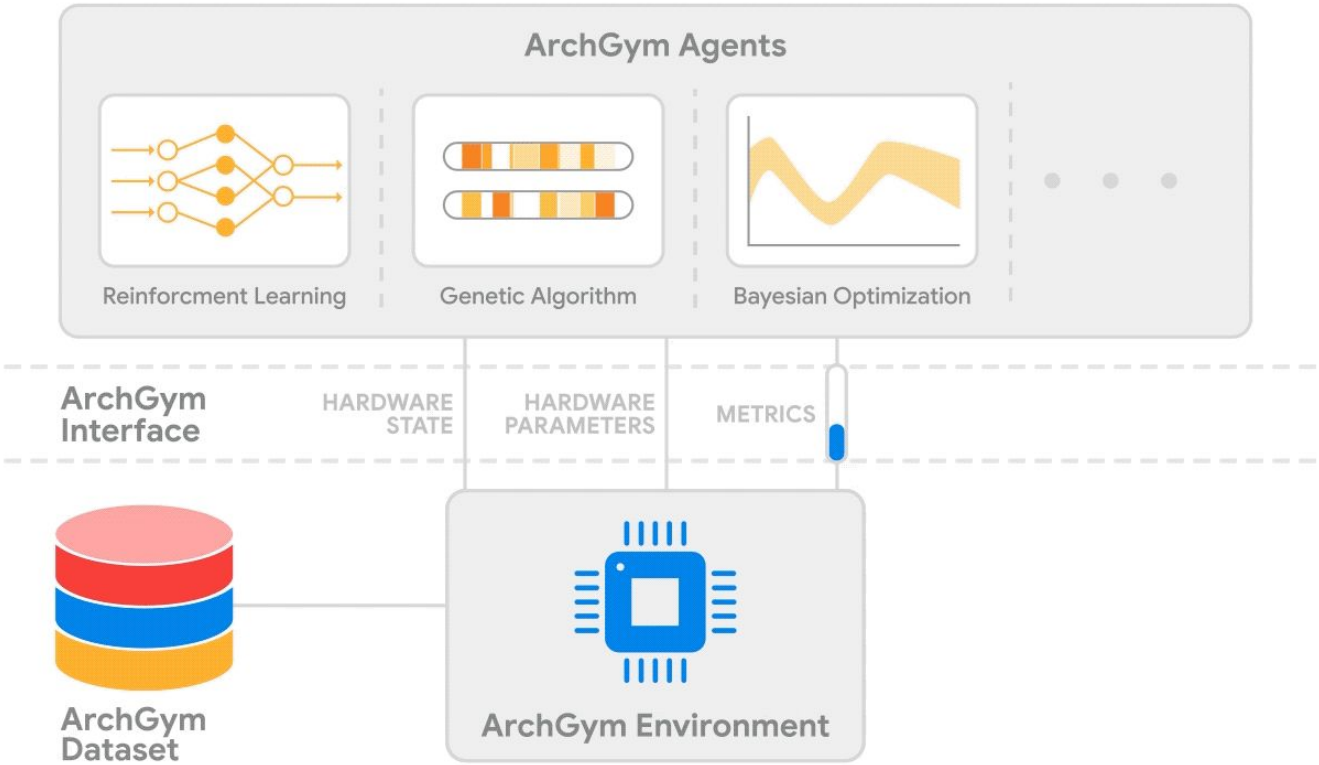
Sheng-Chun Kao, Geonhwa Jeong, Tushar Krishna

Full-stack Co-Design

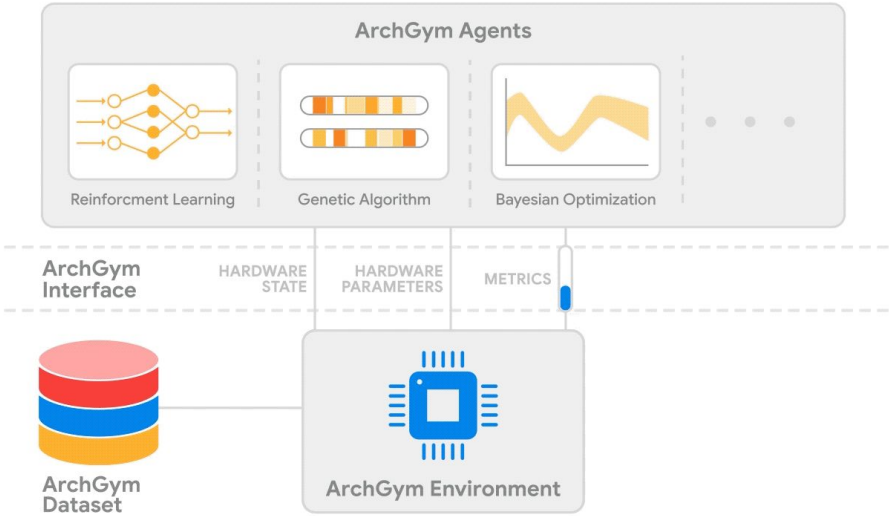




Tools & Infrastructure



Tools & Infrastructure



ArchGym: An Open-Source Gymnasium for Machine Learning Assisted Architecture Design

Srivatsan Krishnan
srivatsan@seas.harvard.edu
Harvard University
Cambridge, Massachusetts, USA

Amir Yazdanbaksh
ayazdan@google.com
Google Research, Brain Team
Mountain View, California, USA

Shvetank Prakash
sprakash@g.harvard.edu
Harvard University
Cambridge, Massachusetts, USA

Jason Jabbour
jasonjabbour@g.harvard.edu
Harvard University
Cambridge, Massachusetts, USA

Ikechukwu Uchendu
iuchendu@g.harvard.edu
Harvard University
Cambridge, Massachusetts, USA

Susobhan Ghosh
susobhan_ghosh@g.harvard.edu
Harvard University
Cambridge, Massachusetts, USA

Behzad Boroujerdian
behzadboro@utexas.edu
UT Austin/Harvard University
Cambridge, Massachusetts, USA

Daniel Richins
drichins@utexas.edu
UT Austin
Austin, Texas, USA

Devashree Tripathy
devashreetripathy@iitbbs.ac.in
IIT Bhubaneswar/Harvard University
Bhubaneswar, Odisha, India

Aleksandra Faust
faust@google.com
Google Research, Brain Team
Mountain View, California, USA

Vijay Janapa Reddi
vj@eecs.harvard.edu
Harvard University
Cambridge, Massachusetts, USA

ABSTRACT

Machine learning (ML) has become a prevalent approach to tame the complexity of design space exploration for domain-specific architectures. While appealing, using ML for design space exploration poses several challenges. First, it is not straightforward to identify the most suitable algorithm from an ever-increasing pool of ML methods. Second, assessing the trade-offs between performance and sample efficiency across these methods is inconclusive. Finally, the lack of a holistic framework for fair, reproducible, and objective comparison across these methods hinders the progress of adopting ML-aided architecture design space exploration and impedes creating repeatable artifacts. To mitigate these challenges, we introduce ArchGym, an open-source gymnasium and easy-to-extend framework that connects a diverse range of search algorithms to architecture simulators. To demonstrate its utility, we evaluate ArchGym across multiple vanilla and domain-specific search algorithms in the design of a custom memory controller, deep neural network accelerators, and a custom SoC for AR/VR workloads, collectively encompassing over 21K experiments. The results suggest that with an unlimited number of samples, ML algorithms are equally favorable to meet the user-defined target specification if its hyperparameters are tuned thoroughly; no one solution is necessarily better than another (e.g., reinforcement learning vs. Bayesian methods). We coin the term "*hyperparameter lottery*" to describe the relatively probable chance for a search algorithm to find an optimal design provided meticulously selected hyperparameters. Additionally, the ease of data collection and aggregation in ArchGym facilitates research in ML-aided architecture design space exploration. As a case study, we show this advantage by developing a proxy cost model with an RMSE of 0.61% that offers a 2,000-fold reduction in simulation time. Code and data for ArchGym is available at https://bit.ly/ArchGym.

CCS CONCEPTS

• **Computer systems organization → Architectures**; • **Computing methodologies → Reinforcement learning**; **Machine learning algorithms**; **Bio-inspired approaches**.

KEYWORDS

Machine learning, Machine Learning for Computer Architecture, Machine Learning for System, Reinforcement Learning, Bayesian Optimization, Open Source, Baselines, Reproducibility

ISCA '23, June 17–21, 2023, Orlando, FL, USA

ACM ISBN 979-8-4007-0095-8/23/06…$15.00
https://doi.org/10.1145/3579371.3589049

ACM Reference Format:
Srivatsan Krishnan, Amir Yazdanbaksh, Shvetank Prakash, Jason Jabbour, Ikechukwu Uchendu, Susobhan Ghosh, Behzad Boroujerdian, Daniel Richins, Devashree Tripathy, Aleksandra Faust, and Vijay Janapa Reddi. 2023. ArchGym: An Open-Source Gymnasium for Machine Learning Assisted Architecture Design. In *Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA '23), June 17–21, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3579371.3589049

[Krishnan et al. ISCA'23]

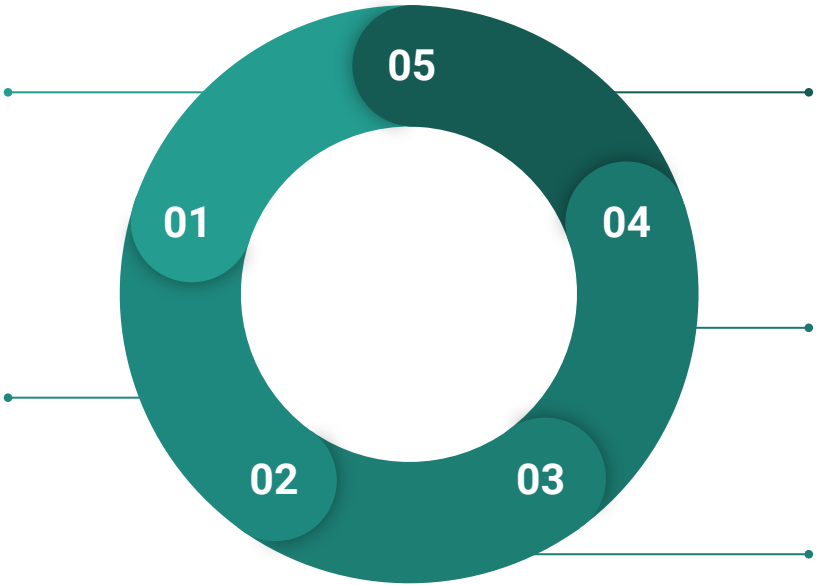
Challenges

Datasets

What datasets do we need? How we should collect these datasets for architecture research? What metadata should the datasets contain to enable broad usage? How do we create standard data formats from any ML algorithm?

ML Algorithms

How can we learn and apply new ML algorithms to effectively design high-performance/efficient systems? How do we make our community more accessible to ML researchers? How do we embrace ML algorithm design as part of architecture research?



Workforce & Training

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

Tools & Infrastructure

How do we reduce the sim2real gap? What instrumentation mechanisms do we need for creating the datasets? What gym environments do we need to enable data-centric AI? How do we define standard data formats for interoperability?

Best Practices

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

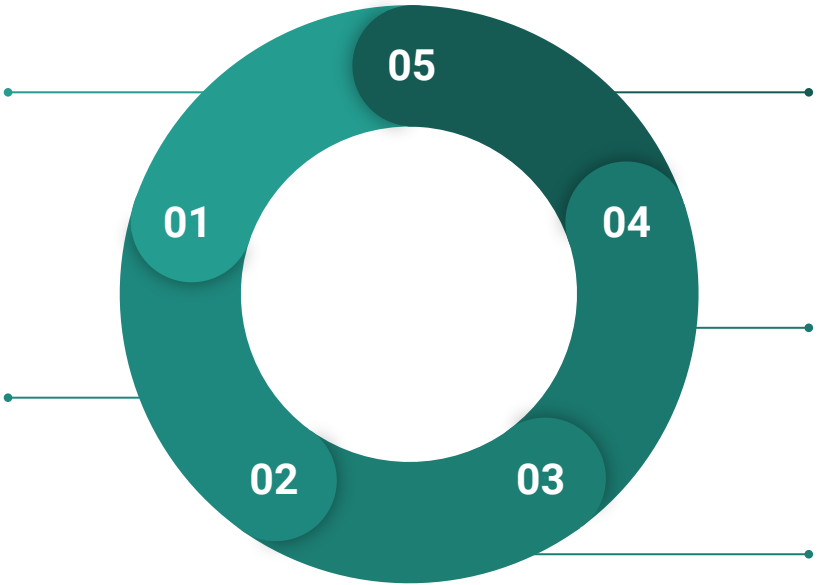
Challenges

Datasets

What datasets do we need? How we should collect these datasets for architecture research? What metadata should the datasets contain to enable broad usage? How do we create standard data formats from any ML algorithm?

ML Algorithms

How can we learn and apply new ML algorithms to effectively design high-performance/efficient systems? How do we make our community more accessible to ML researchers? How do we embrace ML algorithm design as part of architecture research?



Workforce & Training

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

Tools & Infrastructure

How do we reduce the sim2real gap? What instrumentation mechanisms do we need for creating the datasets? What gym environments do we need to enable data-centric AI? How do we define standard data formats for interoperability?

Best Practices

Can we create a systematic playbook for best known methods? How do we ensure strong baselines and reproducibility?

Solving hard problems needs a **community**

Foster a **collaborative community** with a shared vision of ML and systems researchers

Develop and **share curated datasets** that are representative of diverse workloads across the community

Encourage **data-driven AI research and innovation** for Architecture 2.0

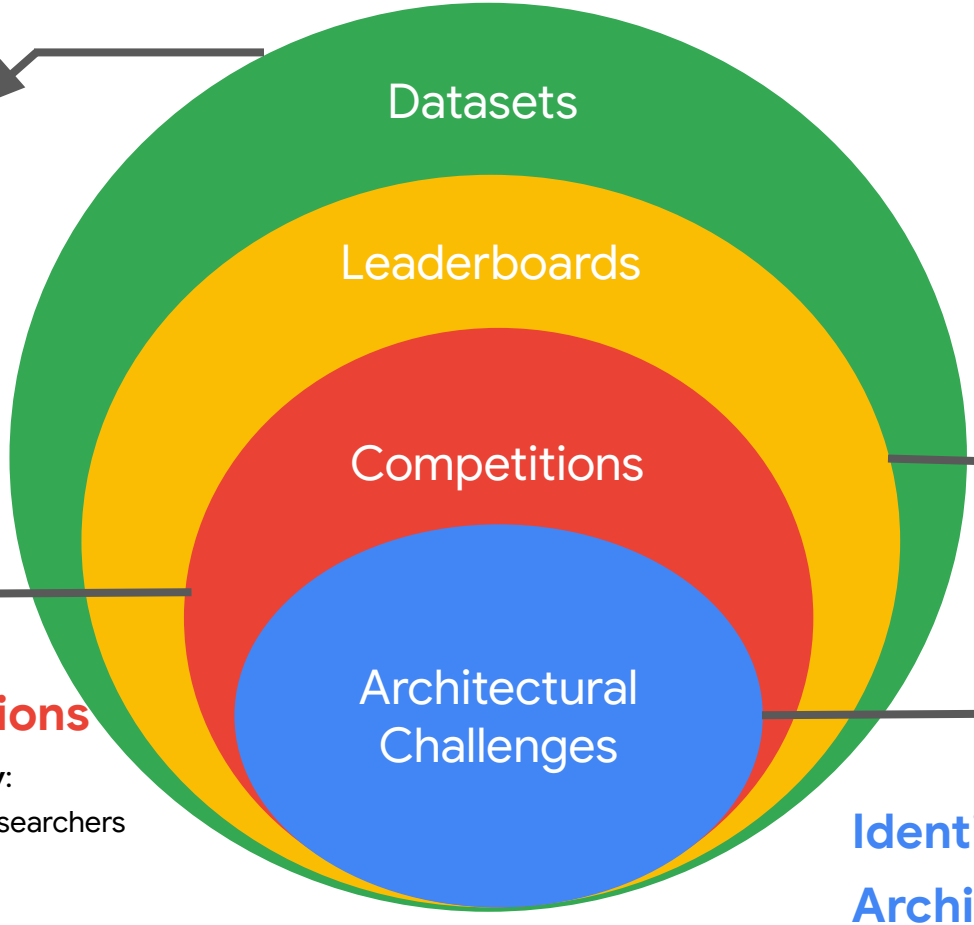
Promote **result replicability** to collectively measure progress & raise the SOTA

Ensure **equitable access** to ML hardware and cutting-edge software technologies

Curate

Datasets

Target Community: Architects, ML researchers, consortium



Maintain

Community-wide Leaderboards

Target Community: Academia + Industry, consortium

Hold **Competitions**

Target Community:
Computer Architects, ML researchers

Identify and **Curate Architectural Problems**

Target Community: Computer Architects

A Call to Action

- Join the activity to define the future of architecture 2.0
- Build a community around the fundamental challenges we have to collectively address
- **August 4th** virtual workshop https://sites.google.com/g.harvard.edu/arch2
- Kick-off a community project