

SNAC-Pack Tutorial

Surrogate **N**eural **A**rchitecture **C**odesign **P**ackage

Dmitri Demler¹, Jason Weitz¹, Ben Hawks², Nhan Tran^{2,3}, Javier Duarte¹

UC San Diego¹, Fermi National Accelerator Laboratory², Northwestern University³

Accessing the SNAC-Pack Tutorial

- Join the hls4ml-tutorial team on github
- <https://github.com/orgs/hls4ml-tutorial/teams/2025-snac-pack-fast-ml-tutorial>
 - If not already member: fill out this form: <https://forms.gle/4vk8khCgRBiNcEkK8>
- Once joined, go to <https://snac-tutorials.fastmachinelearning.org/>
 - Open nac-opt

@DimaPdemler has invited you to join the @hls4ml-tutorial organization on GitHub. Head over to <https://github.com/hls4ml-tutorial> to check out @hls4ml-tutorial's profile.

This invitation will expire in 7 days.

Join @hls4ml-tutorial

Motivation

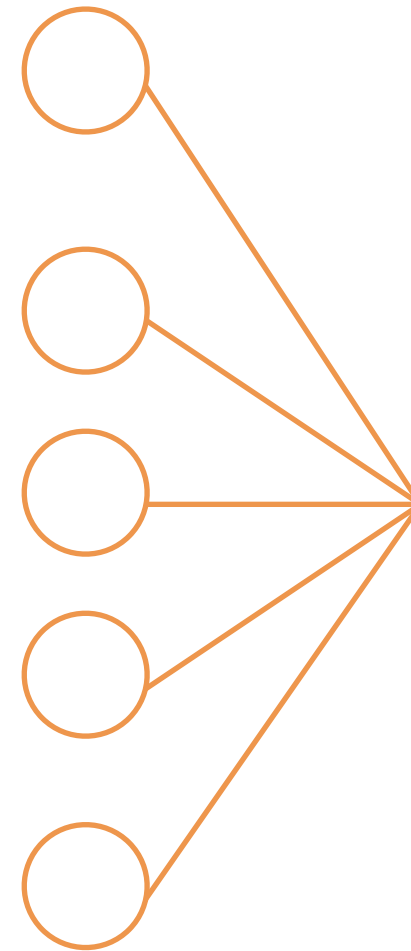
- Software tool to help anyone create cutting edge models for a given task
- User specifies input data, metrics, constraints
 - Output is a well performing, synthesized model

Input Data

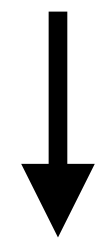
Metrics

Constraints

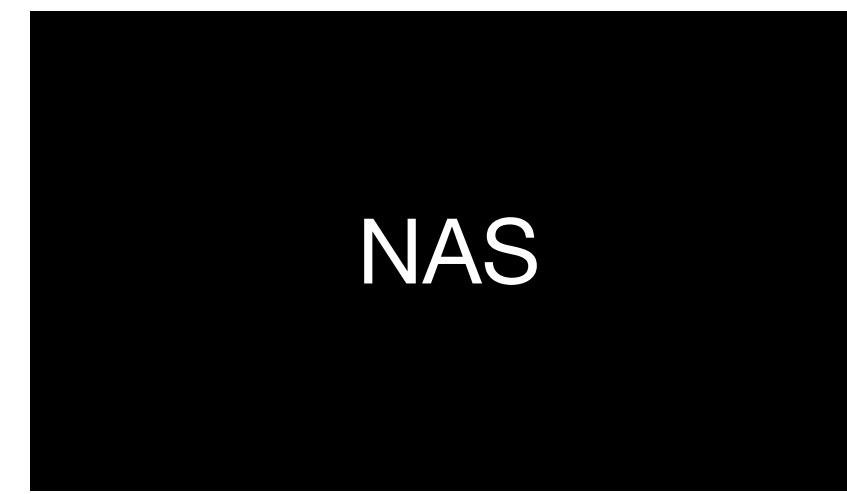
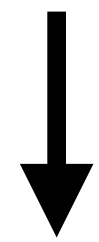
Input Data



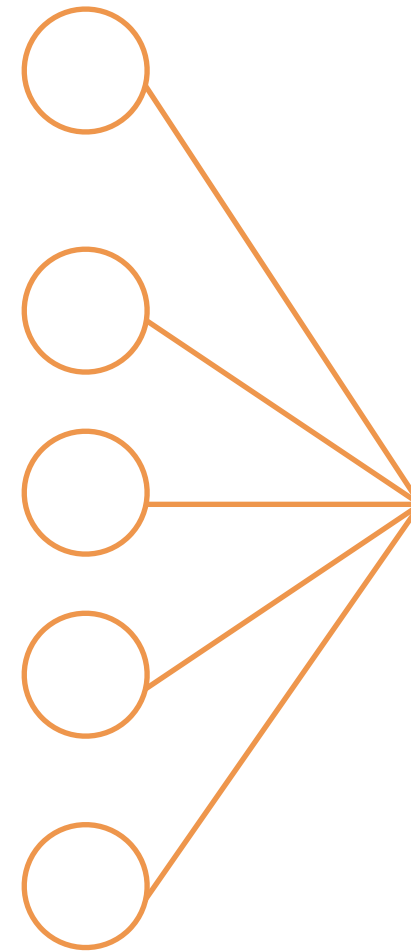
Metrics



Constraints



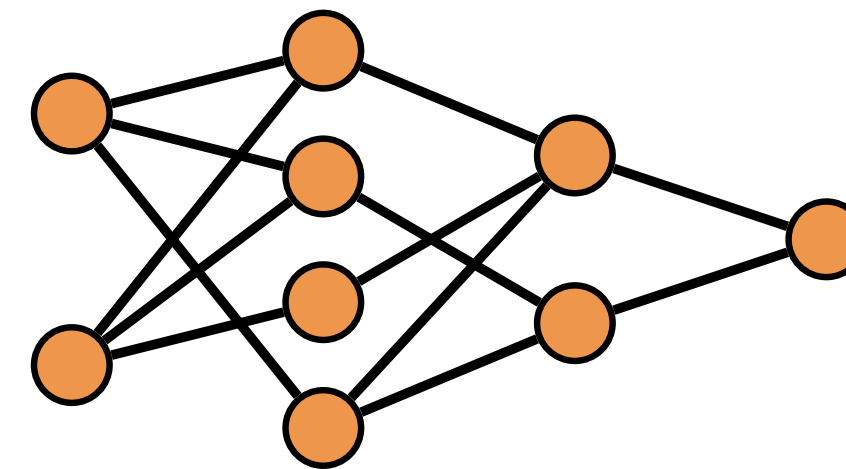
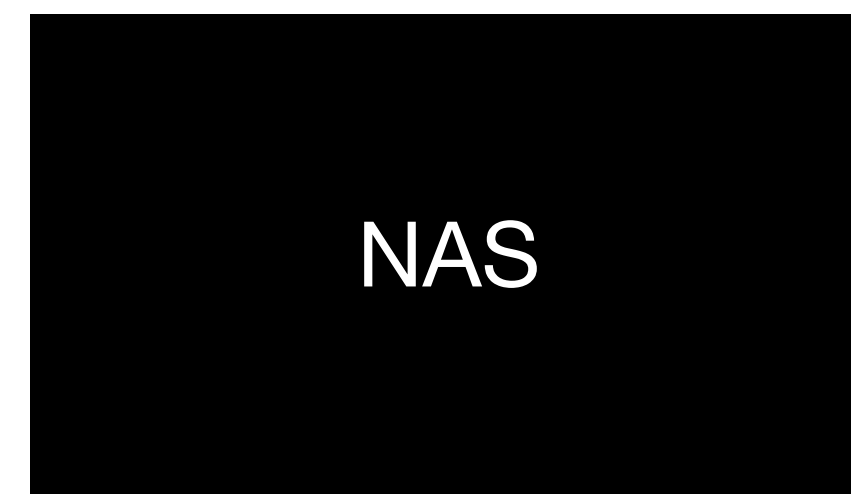
Input Data



Metrics

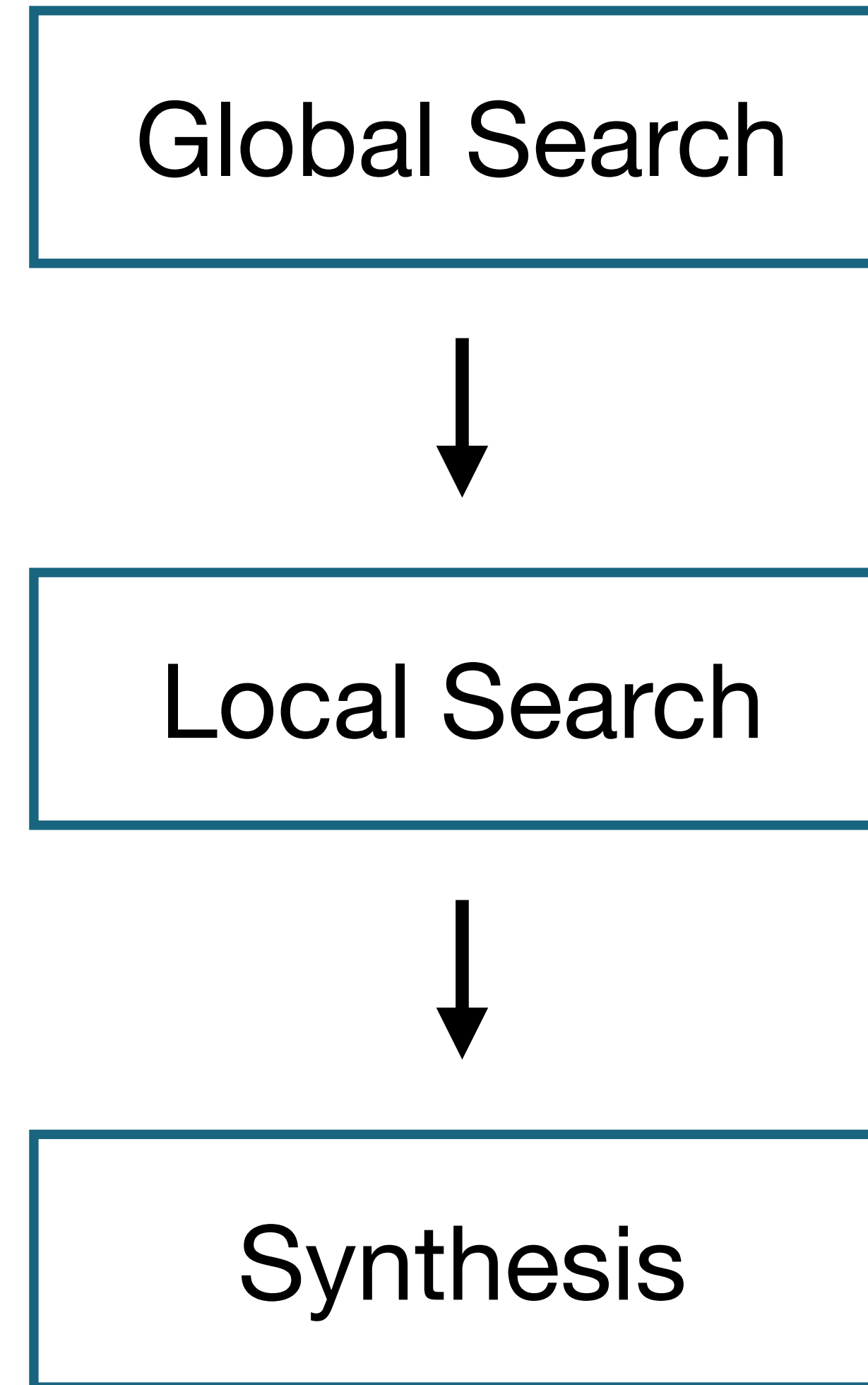


Constraints



NAS Outline

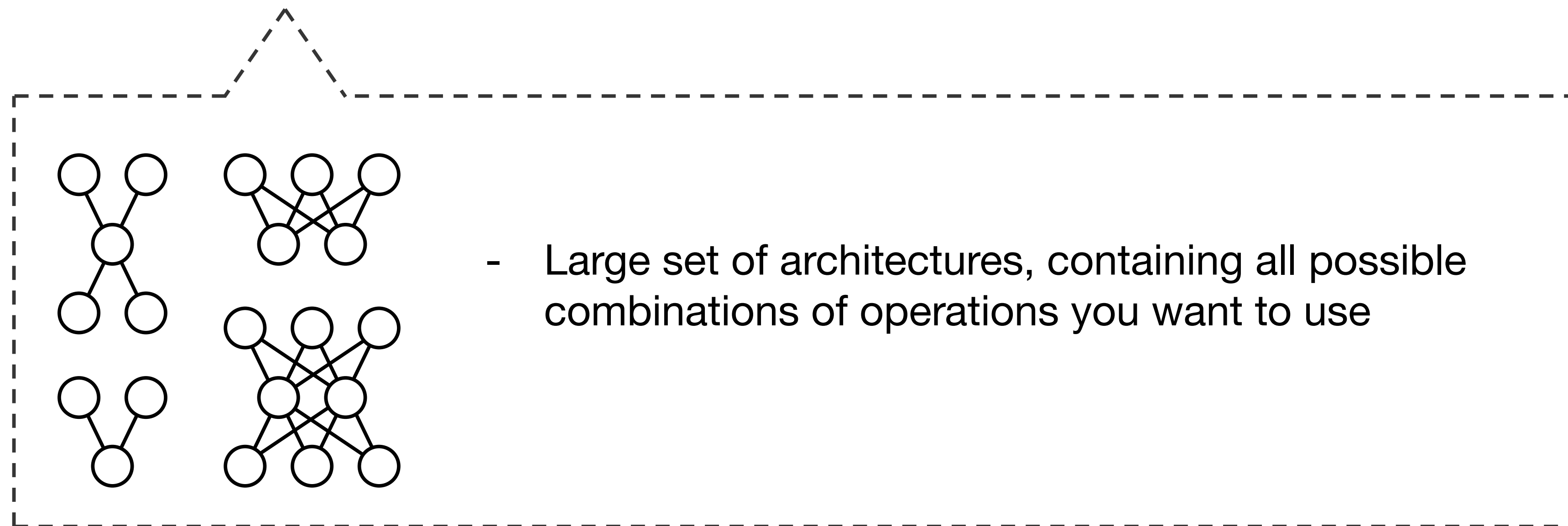
1. Global architecture search
2. Local architecture search
3. hls4ml synthesis



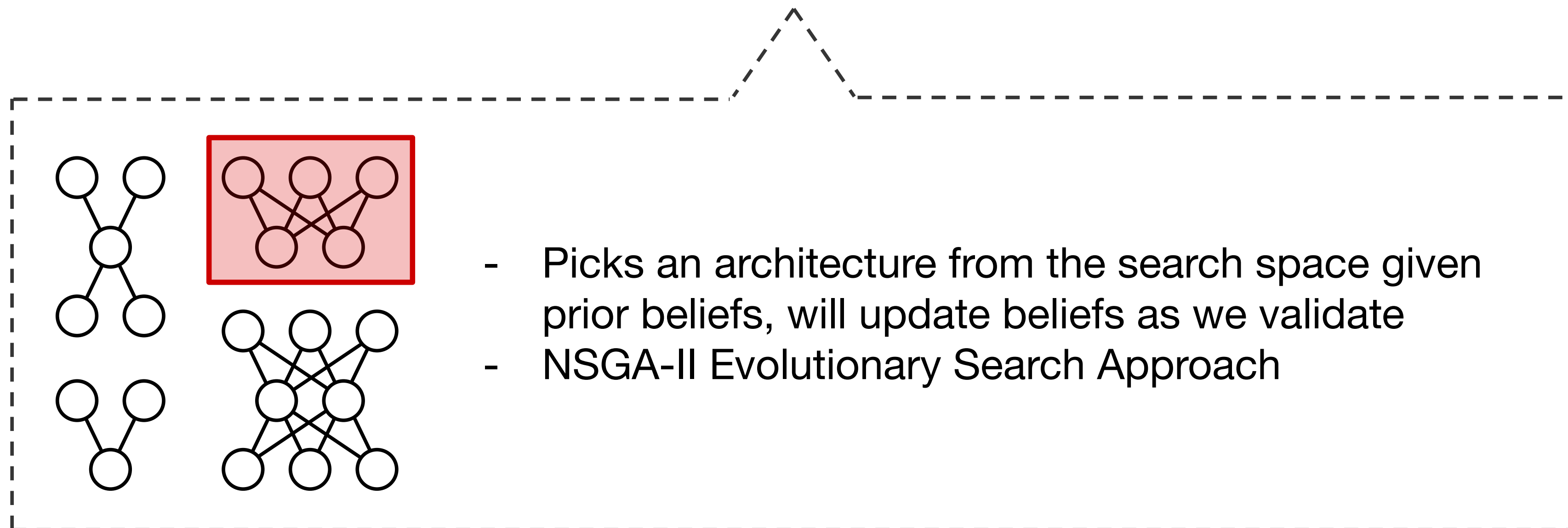
Global Search Outline



Search Space



Search Strategy

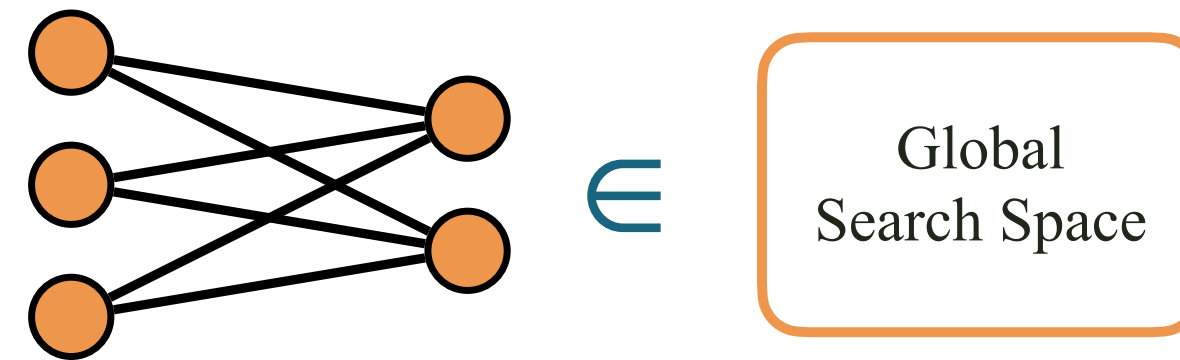


Evaluation

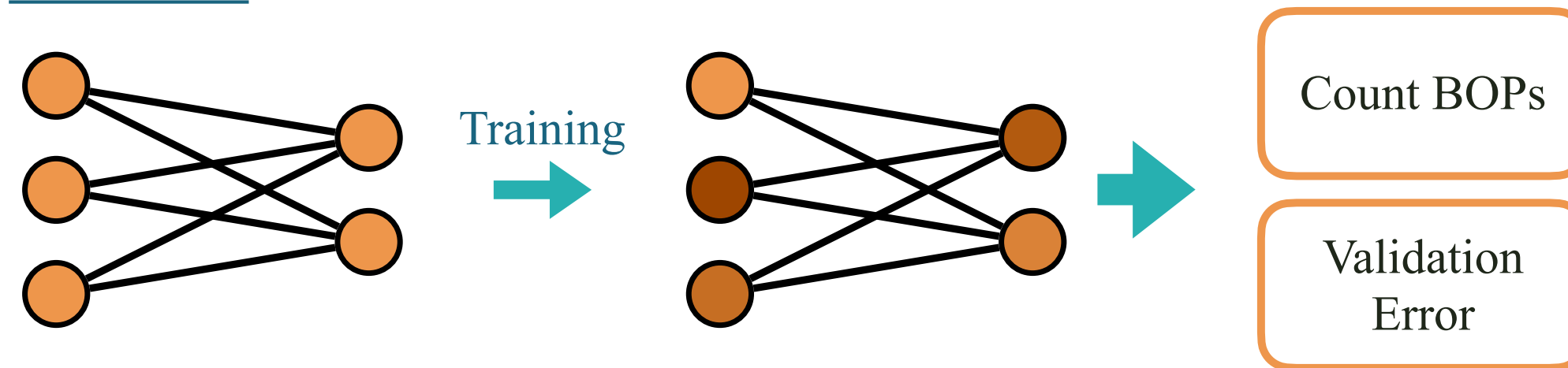


- Evaluates the architecture given set of metrics (parameter count, BOPs)

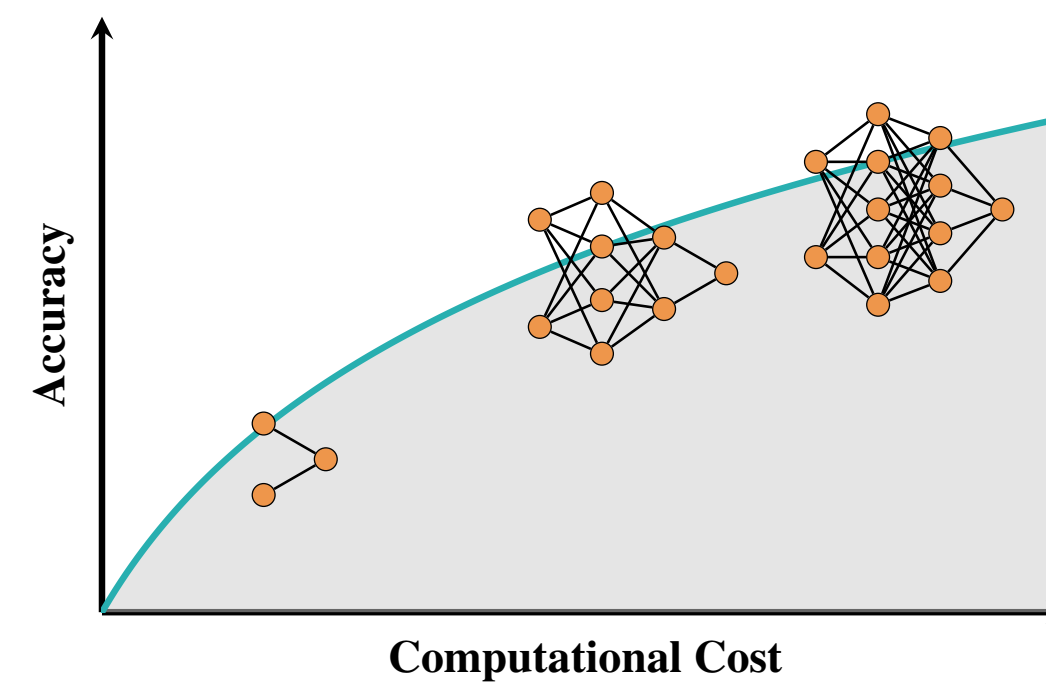
1) Sample



2) Evaluate



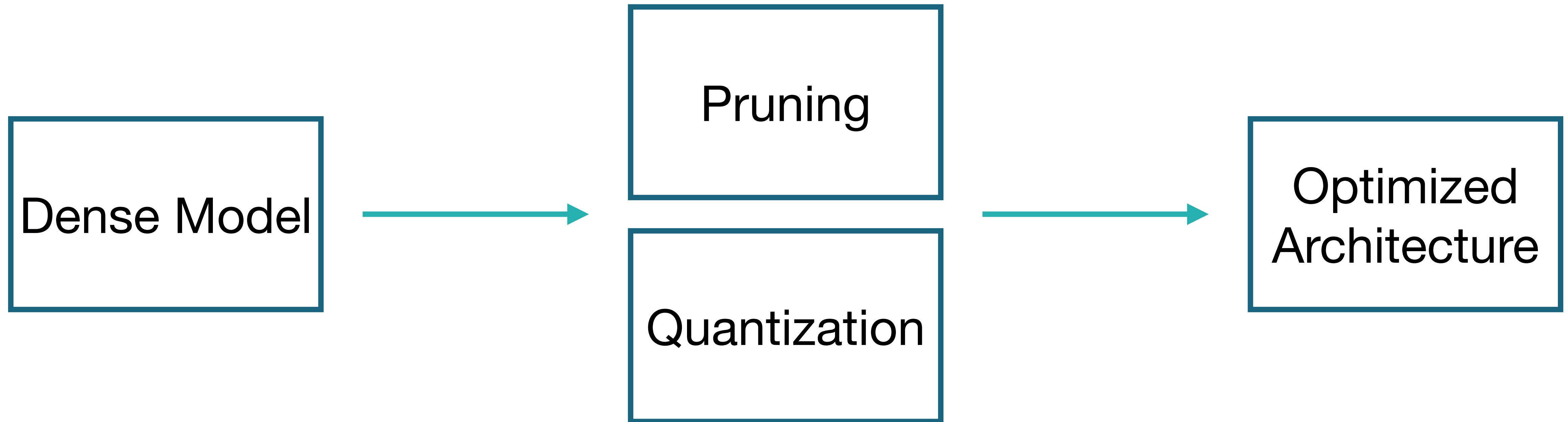
3) Iterate



Output
Pareto Front

Global Search

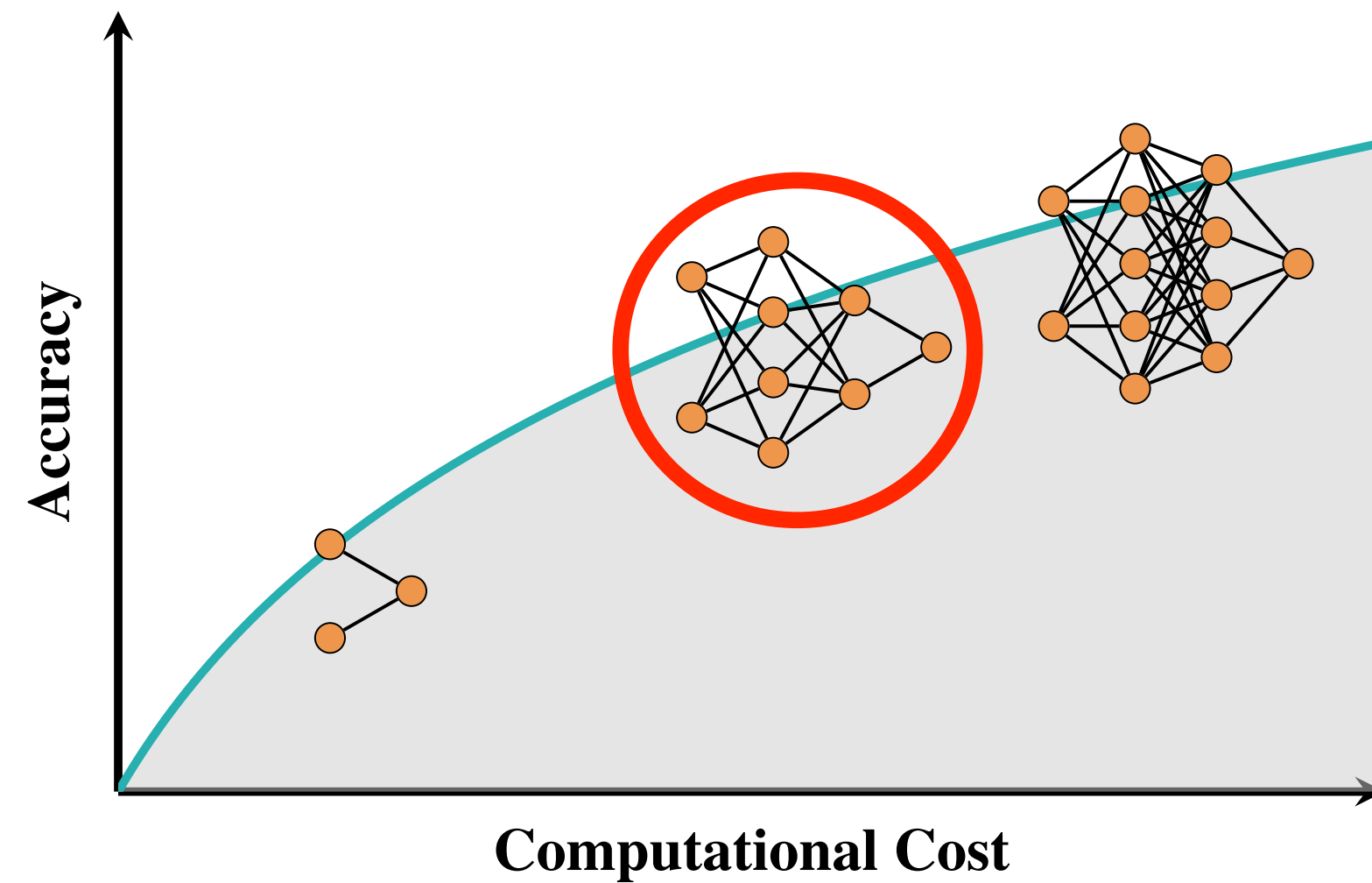
Local Search



Combination of techniques to optimize the dense model

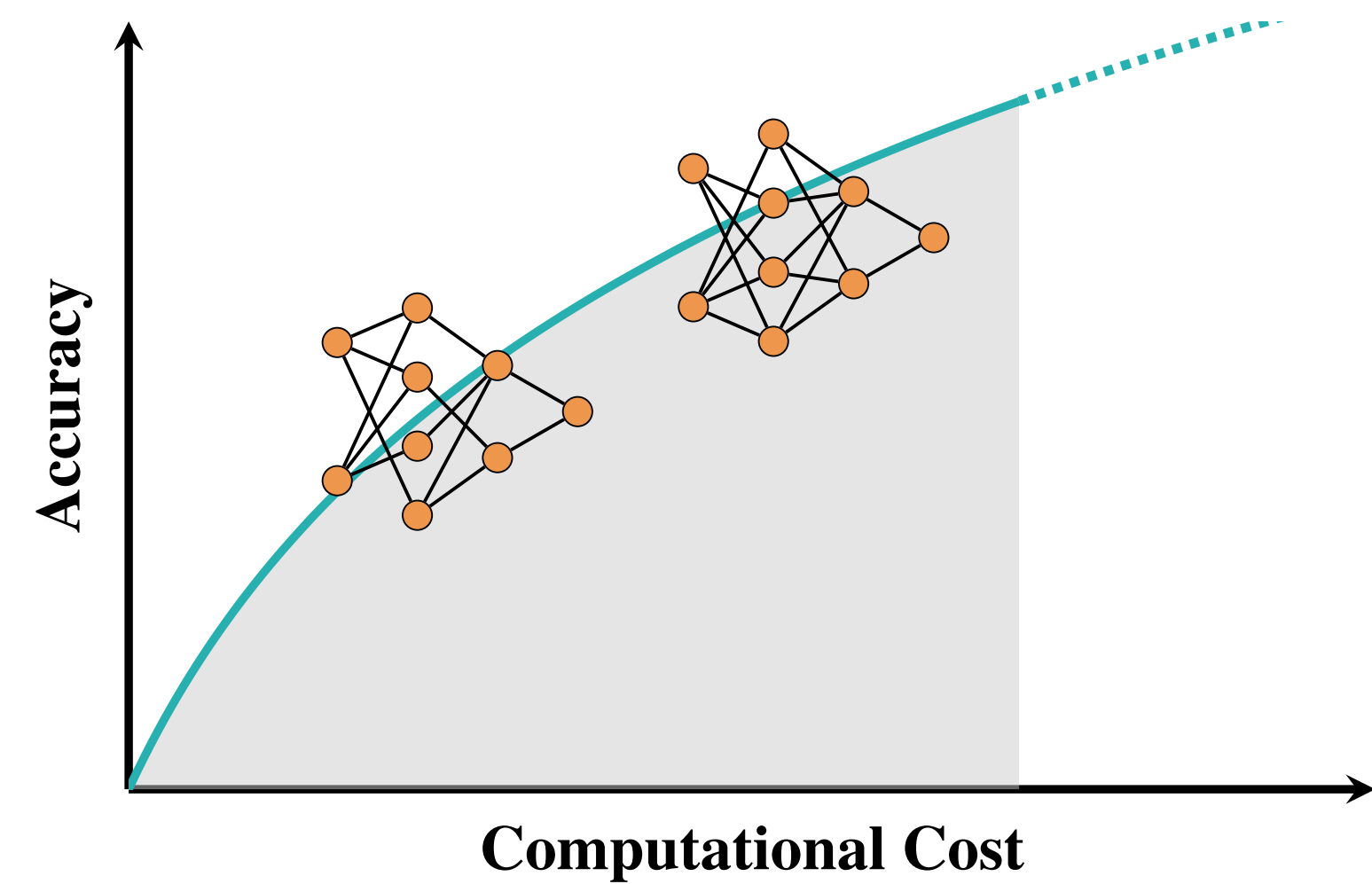
Local Search

3) Iterate



Global search output

4) QAT + Prune

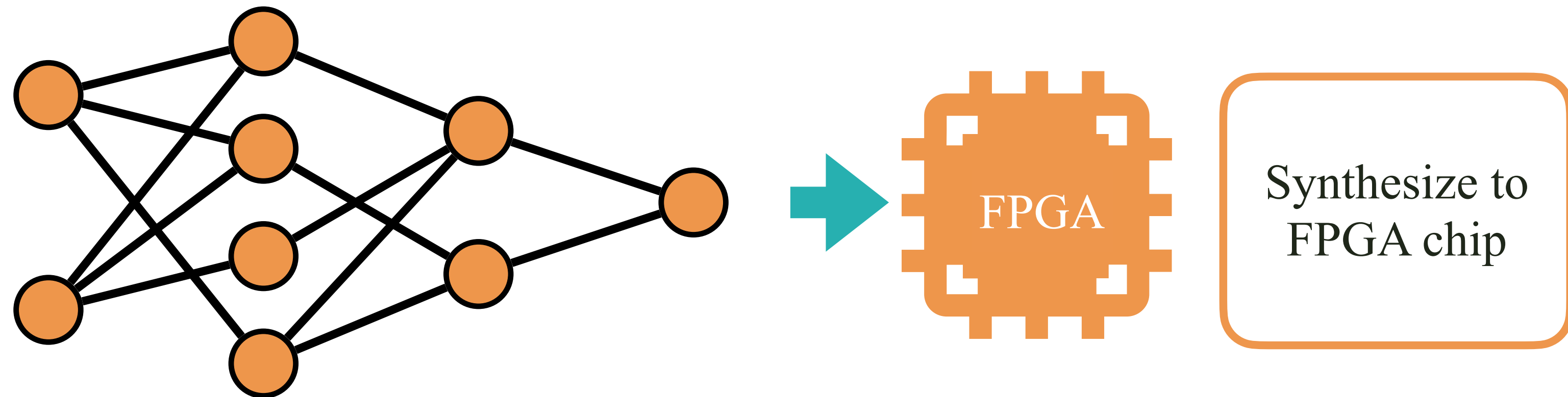


Local search output

Synthesis

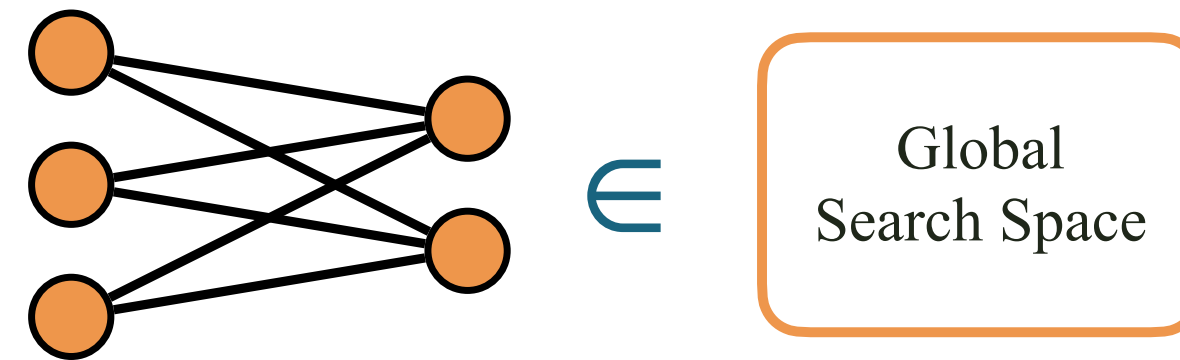
- Synthesis with hls4ml

5) Synthesis

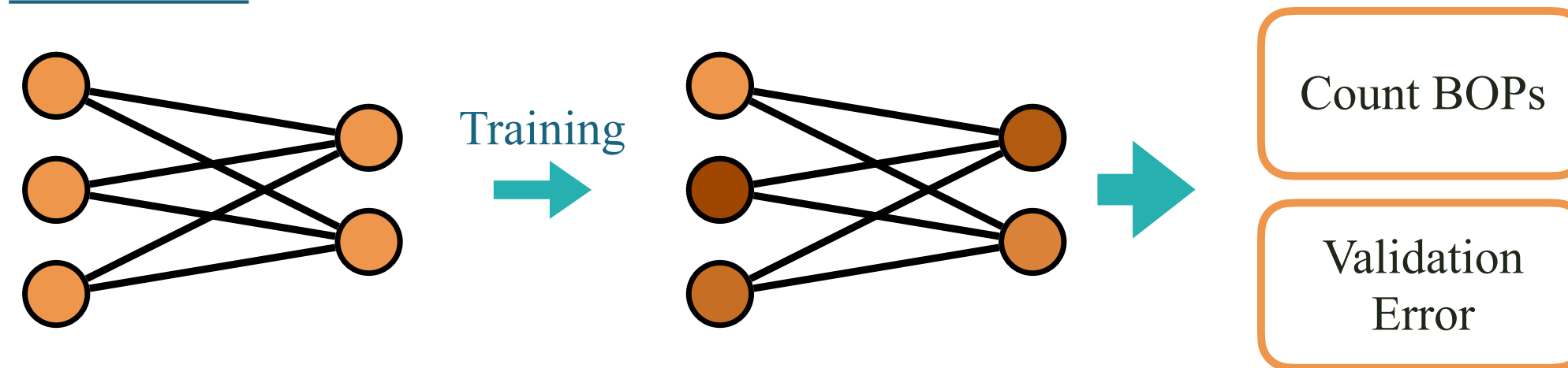


Resource Estimation

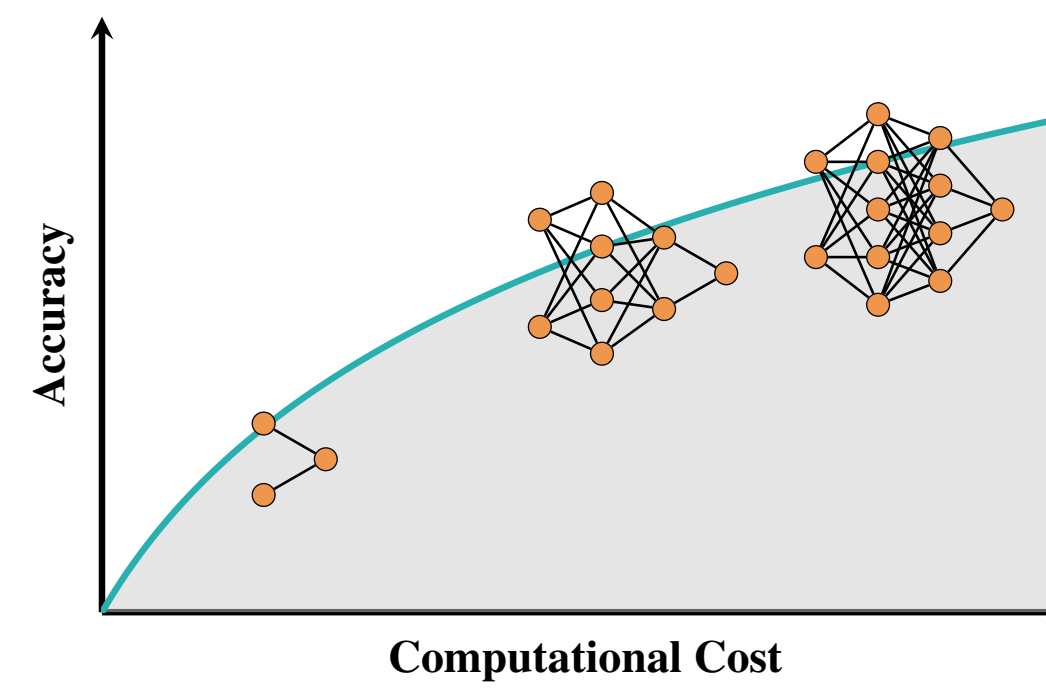
1) Sample



2) Evaluate



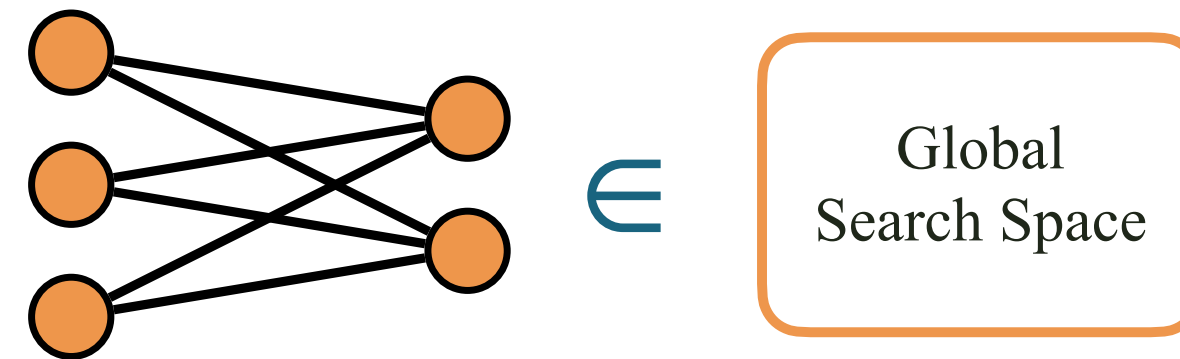
3) Iterate



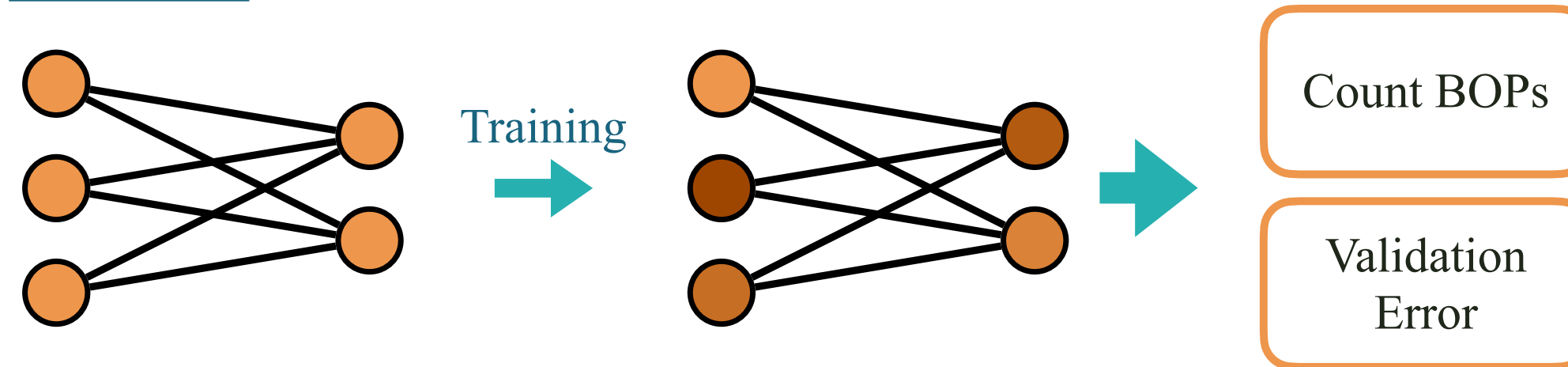
Output
Pareto Front

Global Search

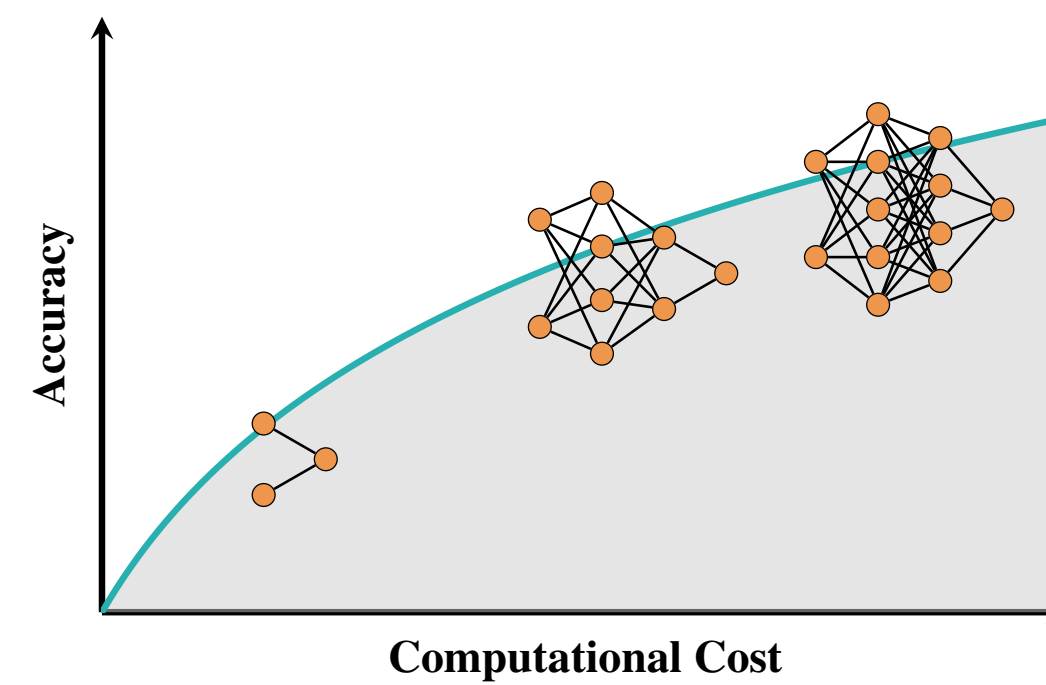
1) Sample



2) Evaluate



3) Iterate



Output
Pareto Front

Global Search

+ Resource utilization
and clock cycles

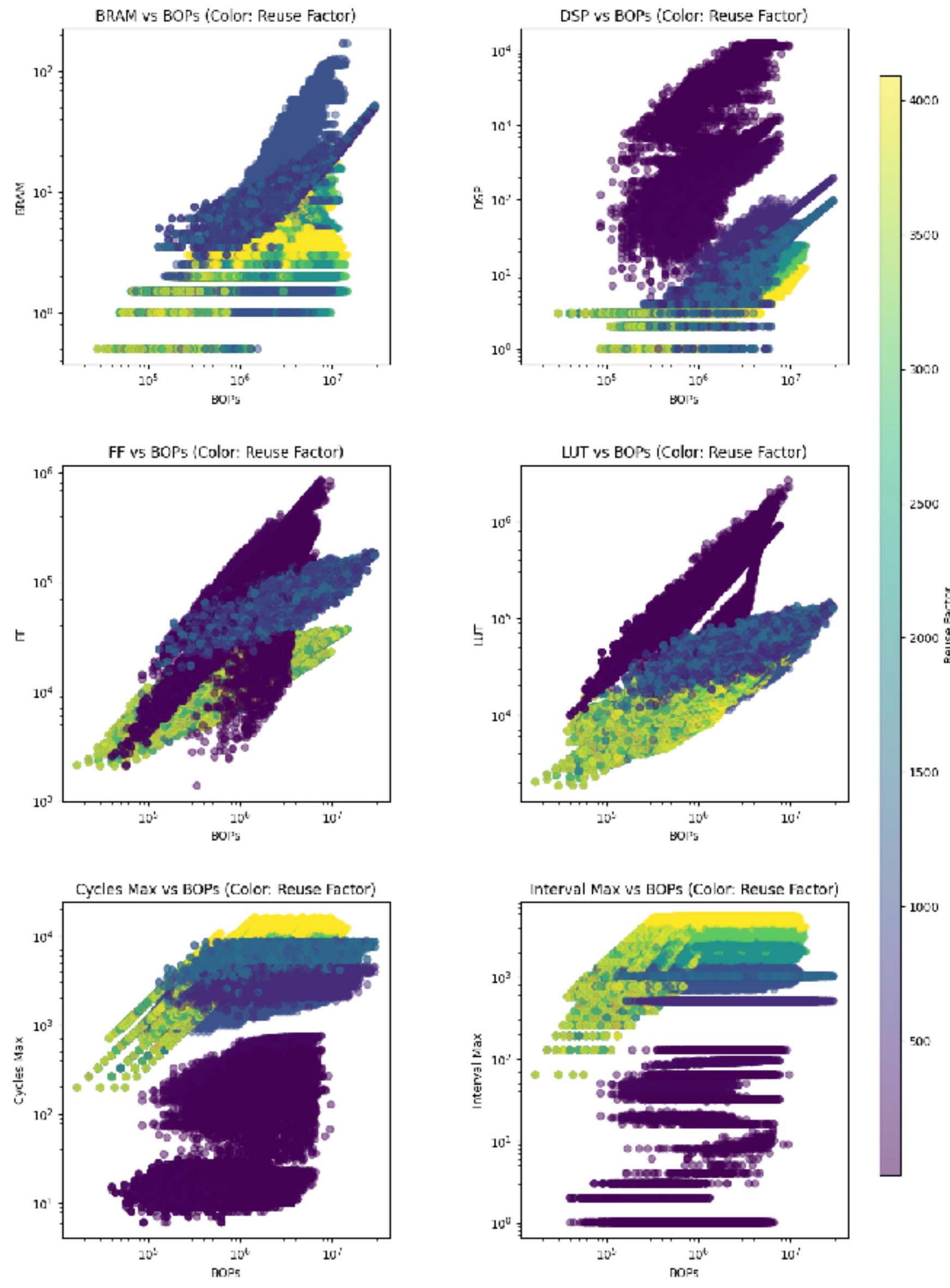
Resource Estimation

With wa-hls4ml and rule4ml

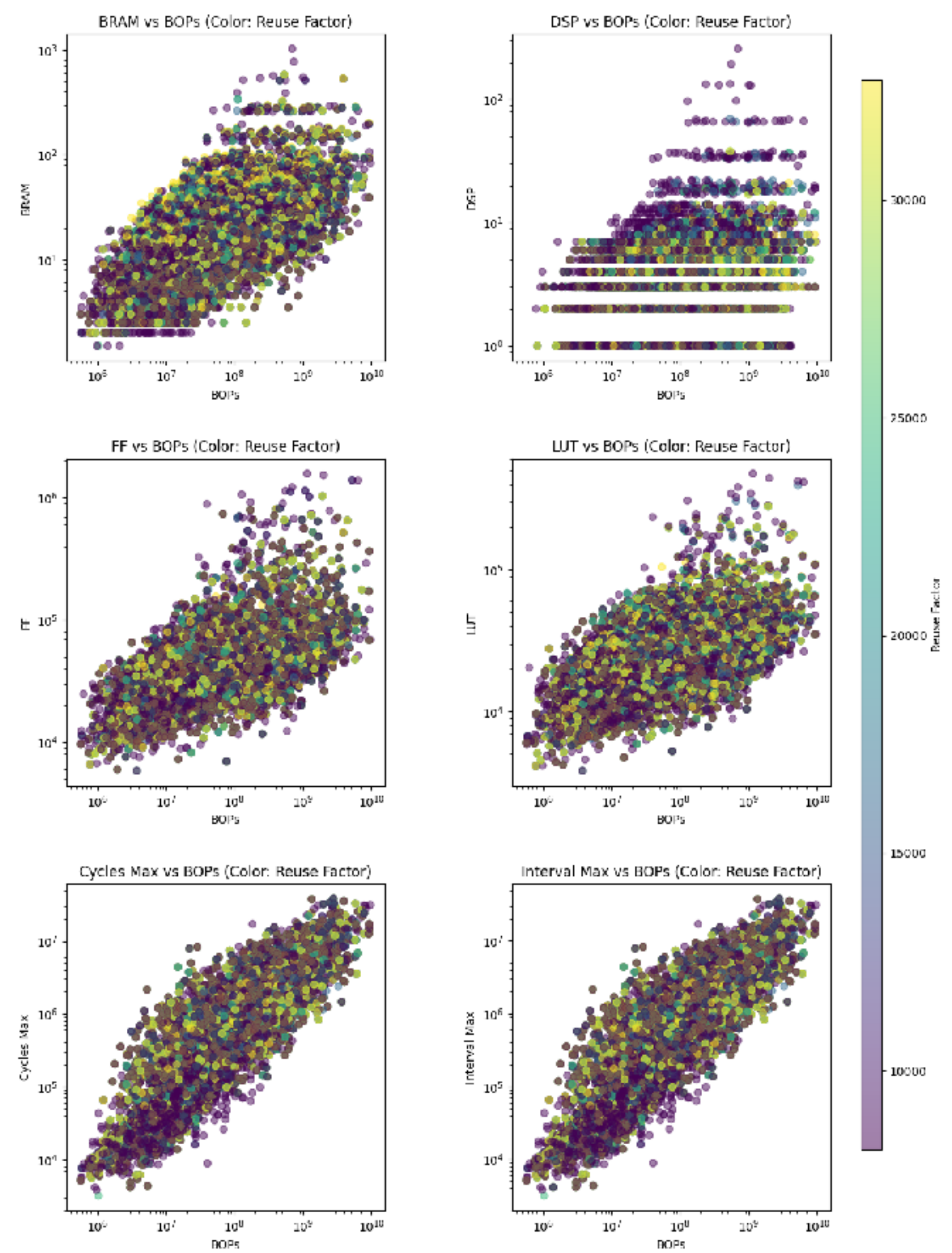
- Dataset

- 683,176 NNs
- Dense 2-20 layer, resource and latency, io_parallel
- Conv 3-7 layer, resource, io_stream
- Synthesized through AMD Vitis version 2023.2 and 2024.2, targeting the AMD Xilinx Alveo U250 FPGA board

Resources and Timing vs BOPs for Fully-Connected Models in Dataset



Resources and Timing vs BOPs for Convolutional Models in Dataset



SNAC-Pack Race

- On Fashion MNIST find the best global search model architecture
- Image will be shown every few seconds and contestants will take a step if they predict correctly
- Models with smaller BOPs will have 20% longer steps so accuracy isn't everything!
- Models are retrained for the race, the point is the architecture itself and not the weights. As a result there will be a lot of variability in the models however.
- Email me the global search created model architecture yaml and race on!

Thank you

