SONAR

Direct Architecture and System Optimization Search

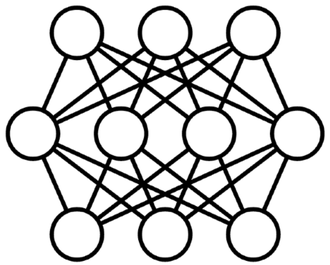
Elias Jääsaari, Michelle Ma, Ameet Talwalkar, Tianqi Chen

Carnegie Mellon University

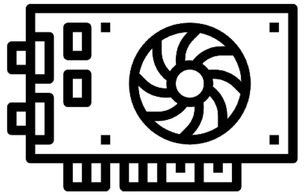
catalyst

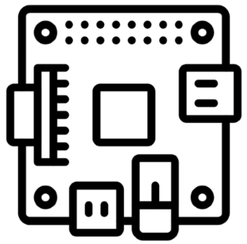
Constrained deployment

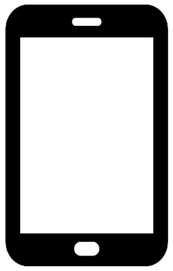
Model



Target



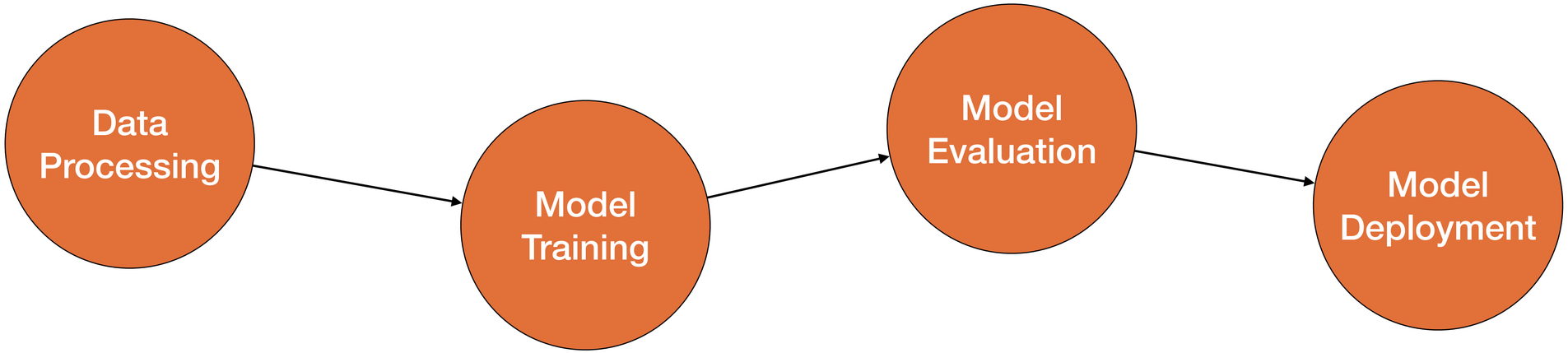




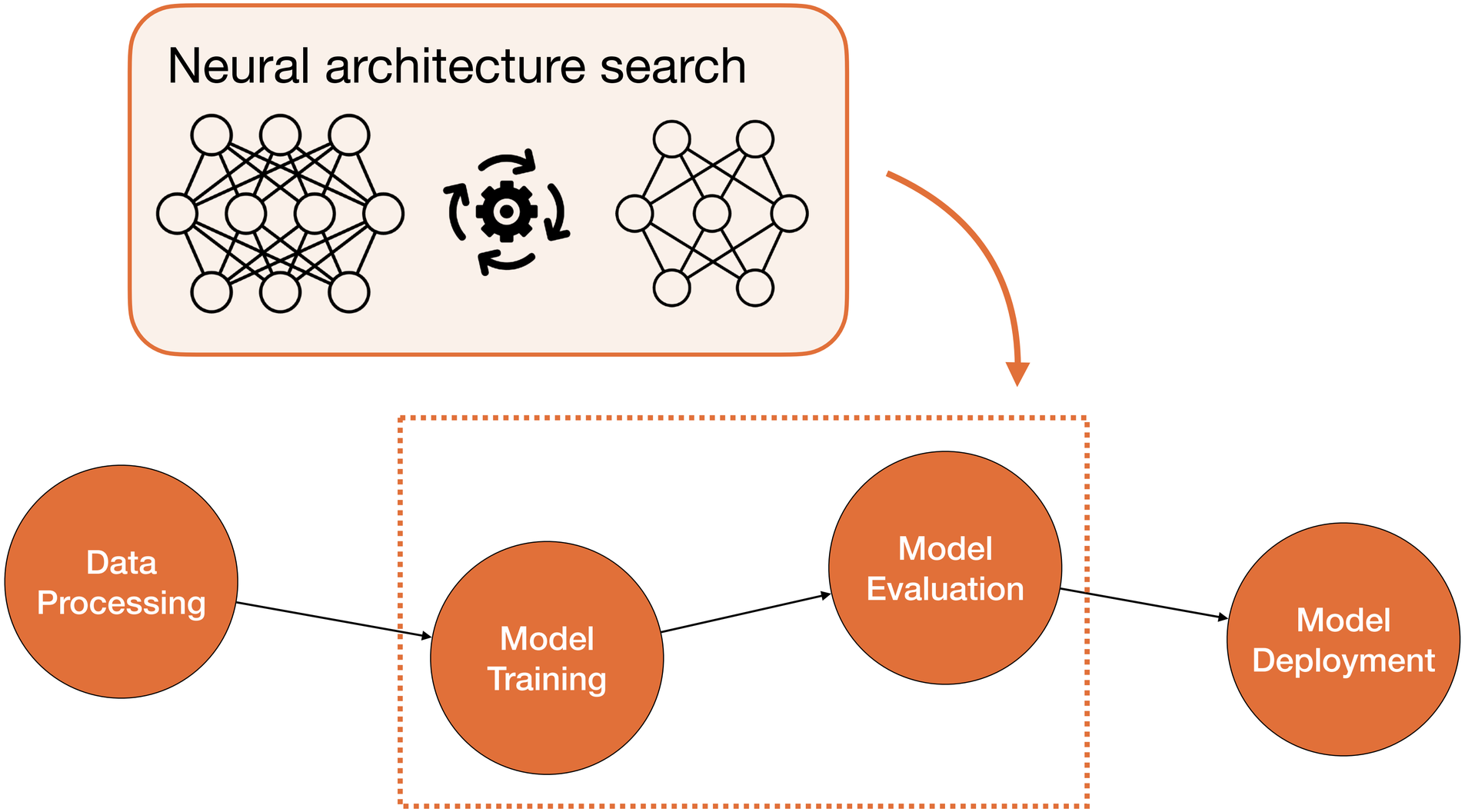
Objectives

- **Accuracy**
- **Latency**
- **Memory consumption**
- **Energy use**
- **Number of parameters**
- **FLOPS**
- **Fairness**
- **…**

The ML pipeline



The ML pipeline

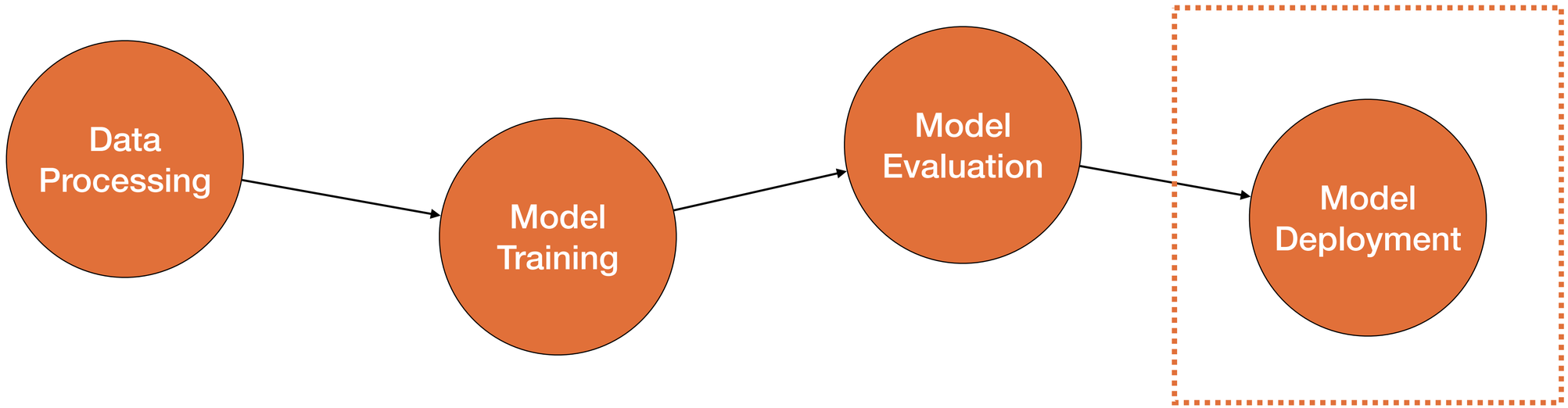


The ML pipeline

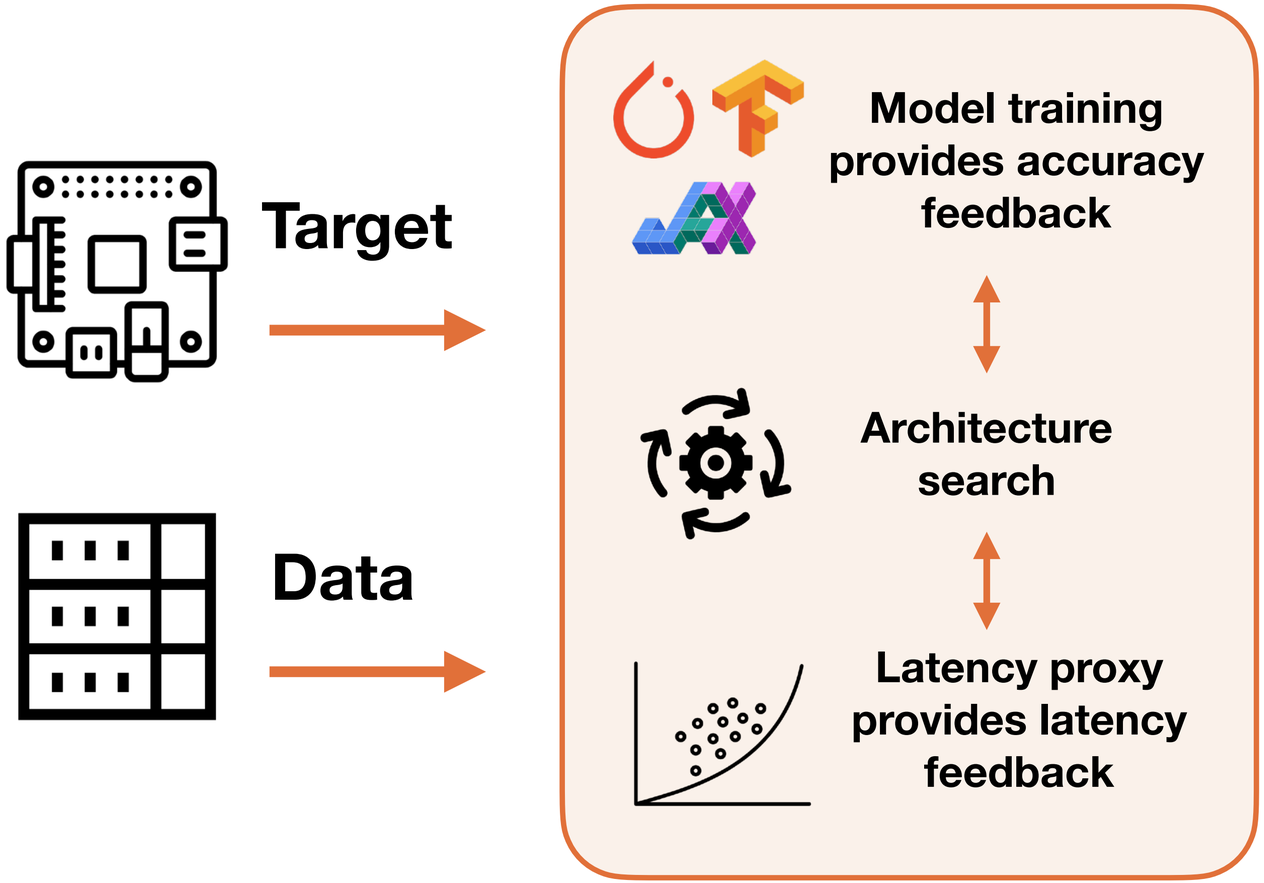
System optimization search

```
for yo in range(1024 / ty):
  for xo in range(1024 / tx):
    C[yo*ty:yo*ty+ty][xo*tx:xo*tx+x] = 0
      for k in range(1024):
        for yi in range(ty):
          for xi in range(tx):
            C[yo*ty+yi][xo*tx+xi] +=
              A[k][yo*ty+yi] * B[k][xo*tx+xi]
```

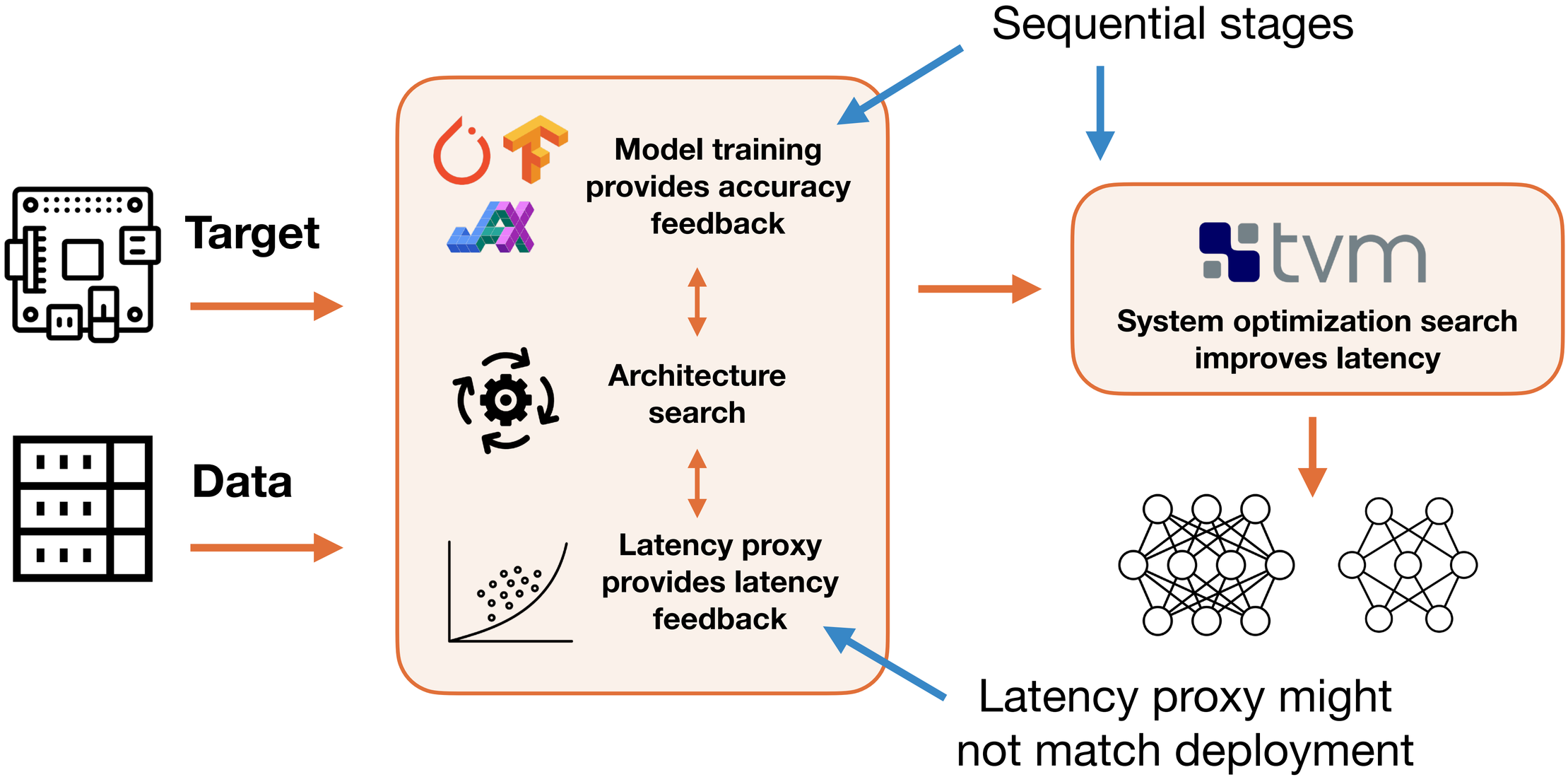
```
for yo in range(128):
  for xo in range(128):
    intrin.fill_zero(C[yo*8:yo*8+8][xo*8:xo*8+8])
    for ko in range(128):
      intrin.fused_gemm8x8_add(
        C[yo*8:yo*8+8][xo*8:xo*8+8],
        A[yo*8:yo*8+8][xo*8:xo*8+8],
        B[yo*8:yo*8+8][xo*8:xo*8+8])
```



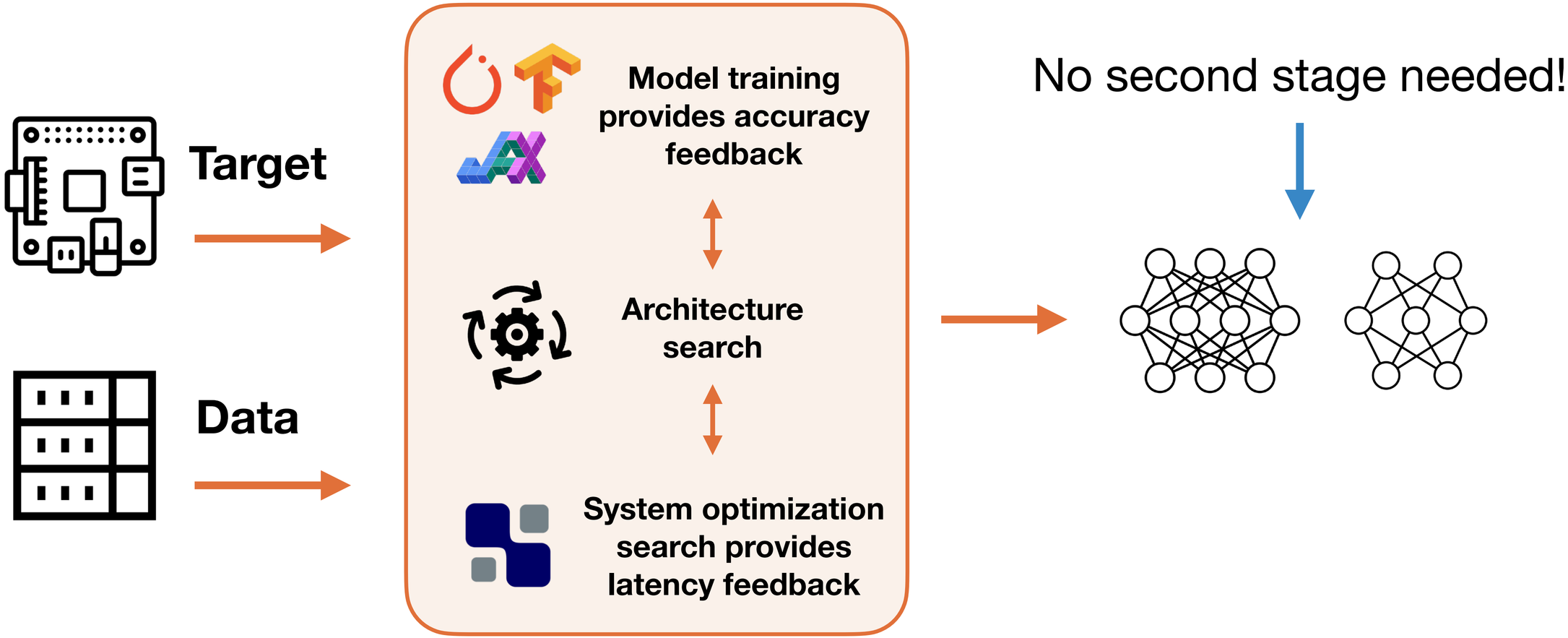
Indirect search



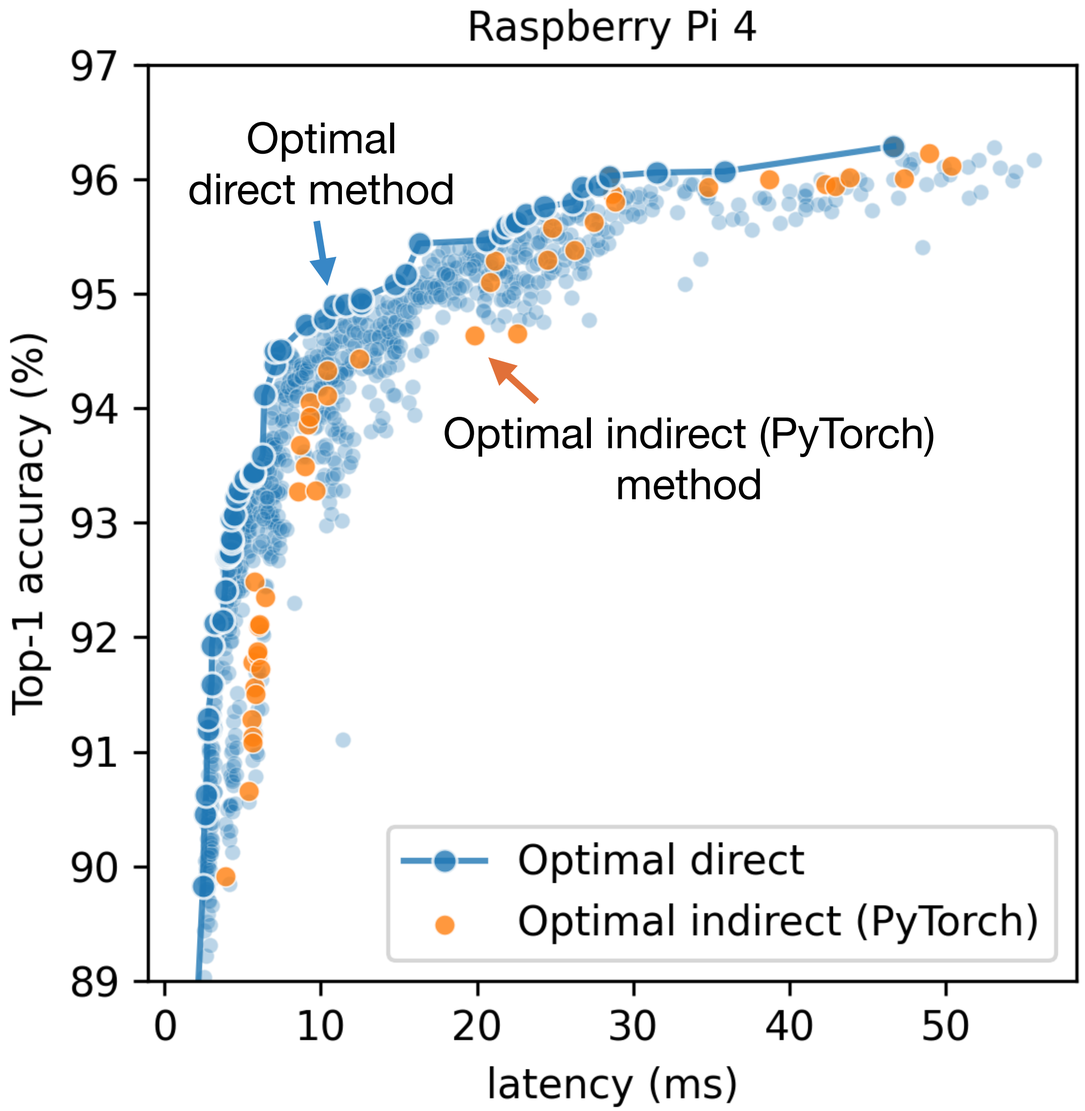
Indirect search



Direct search



It is better to be direct

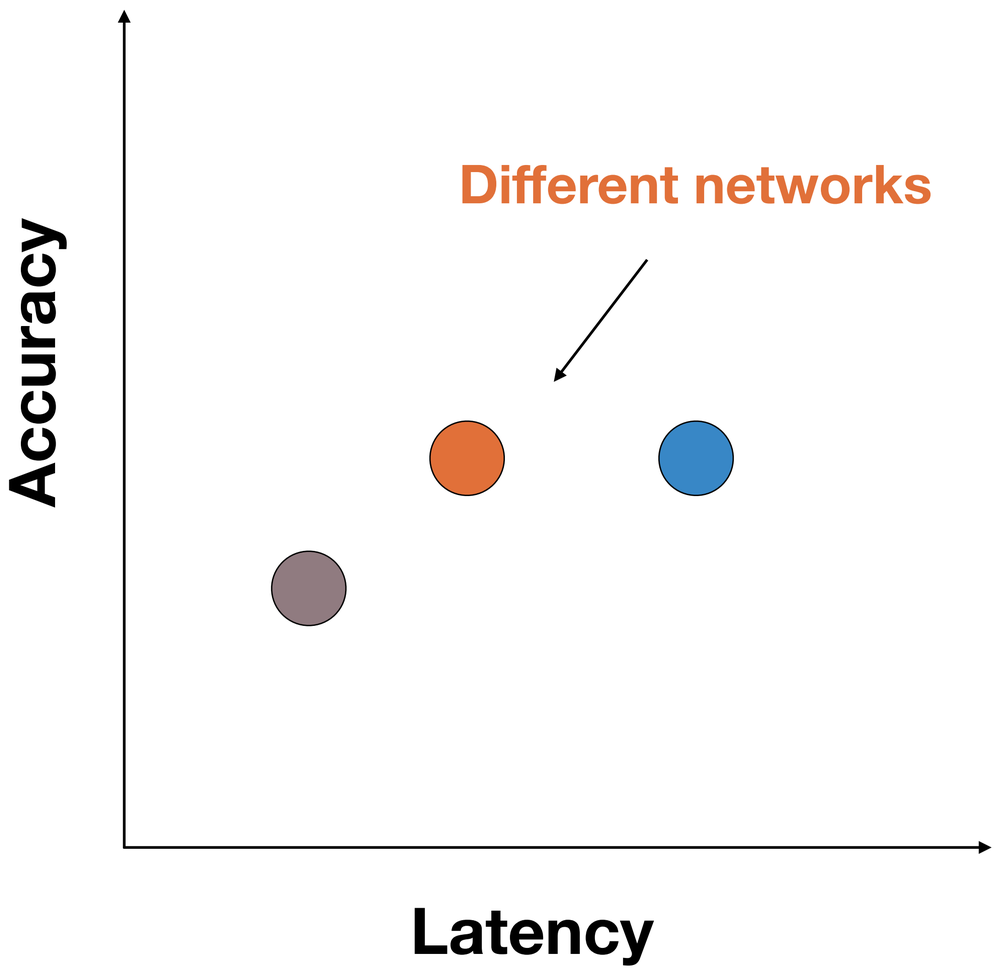


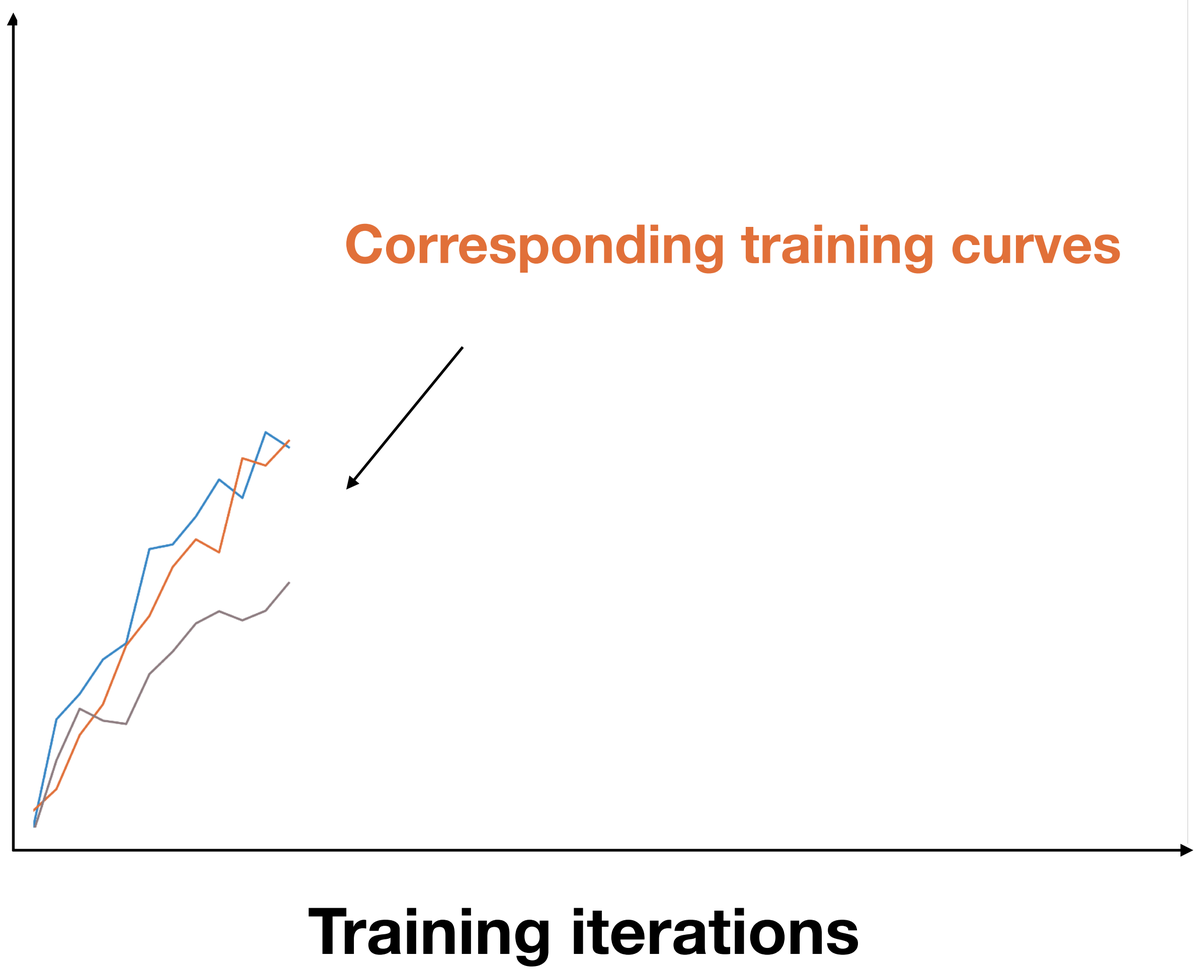
Efficient direct search

How to perform efficient direct search?

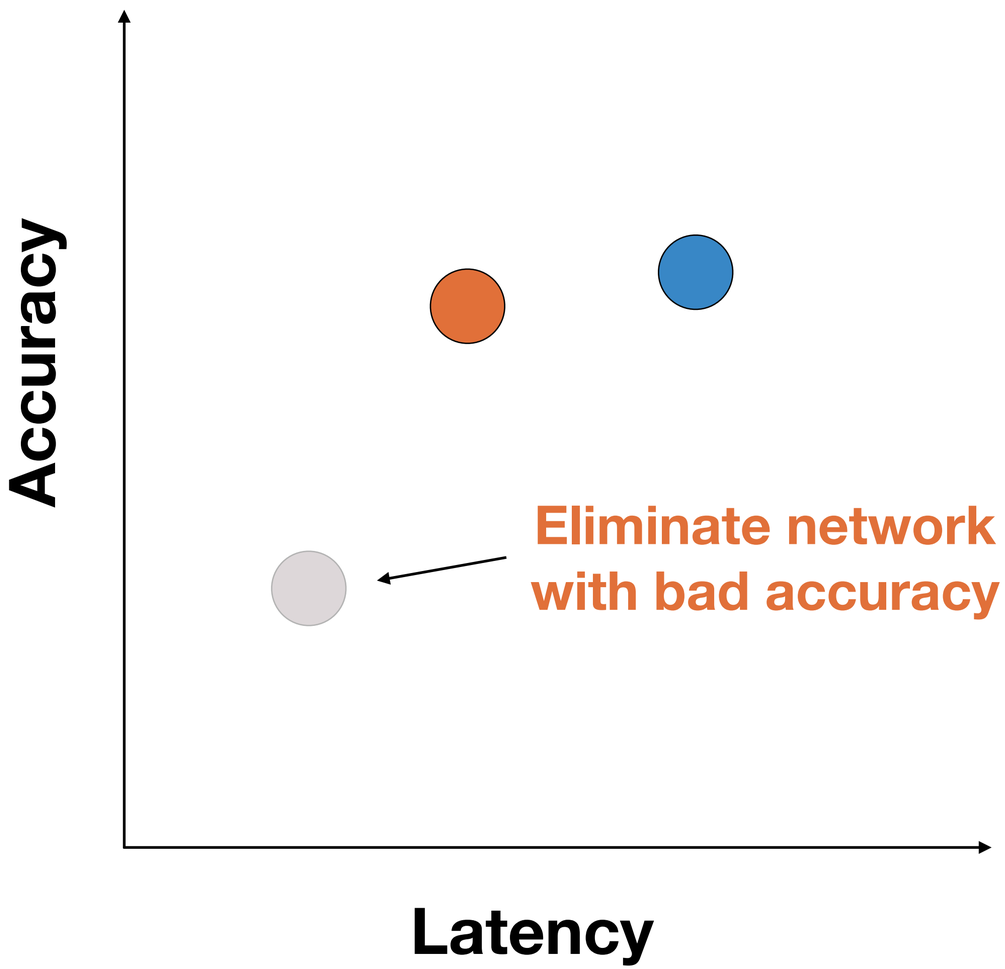
Use early stopping!

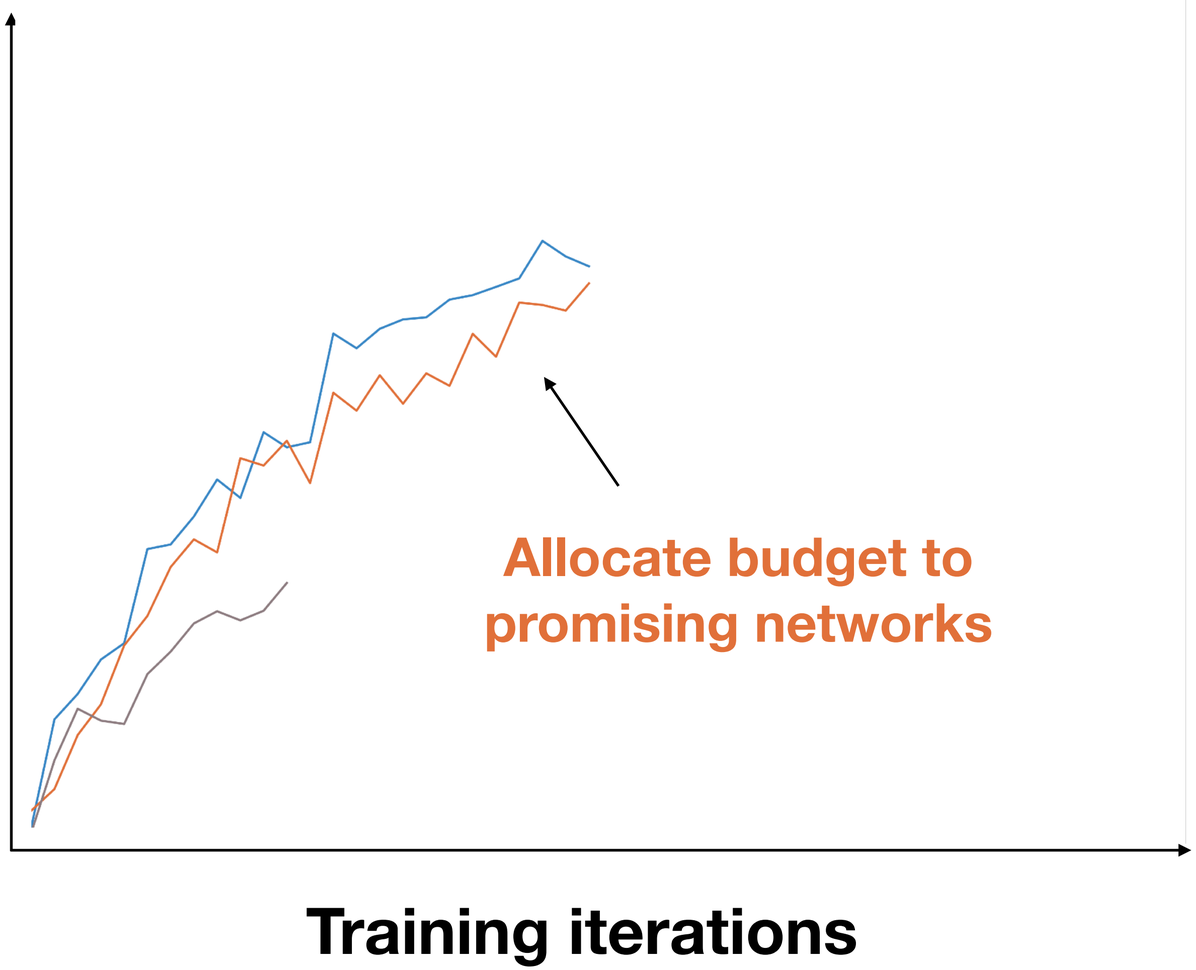
Early stopping for accuracy



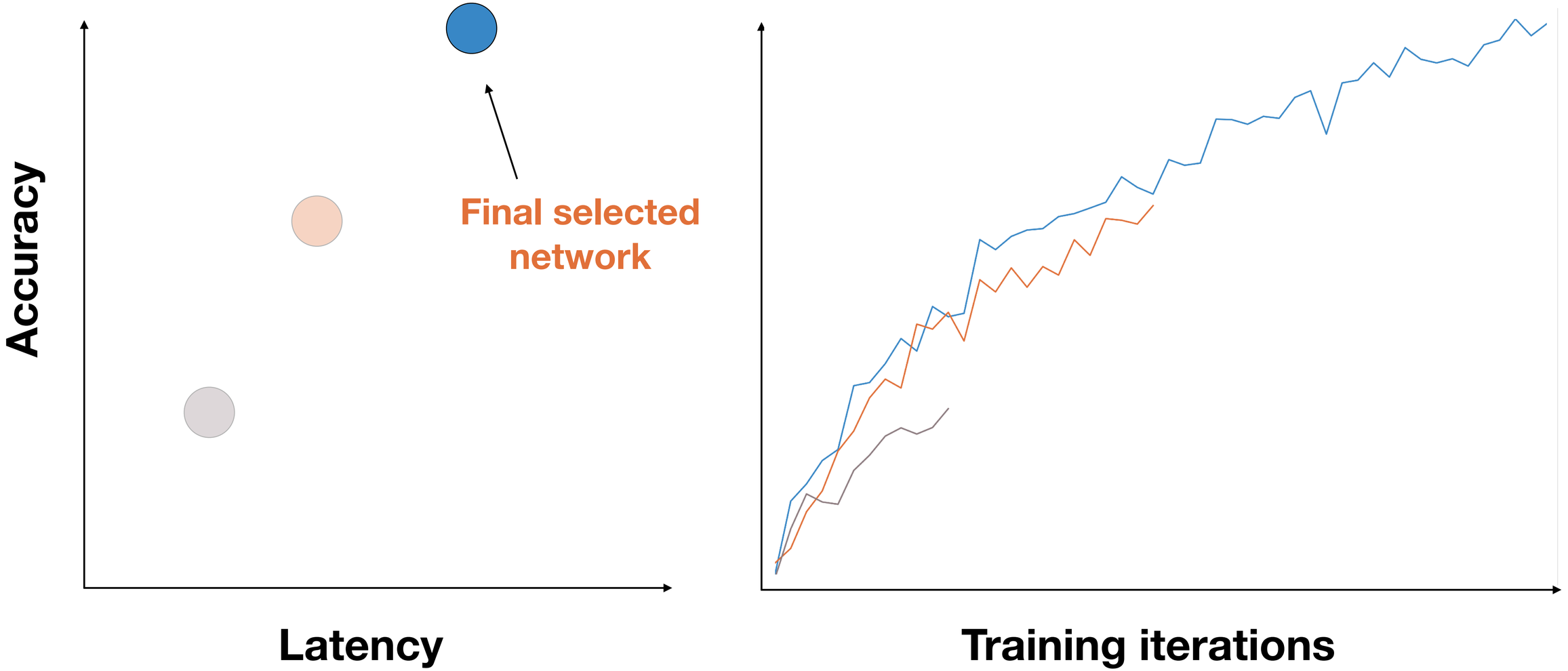


Early stopping for accuracy

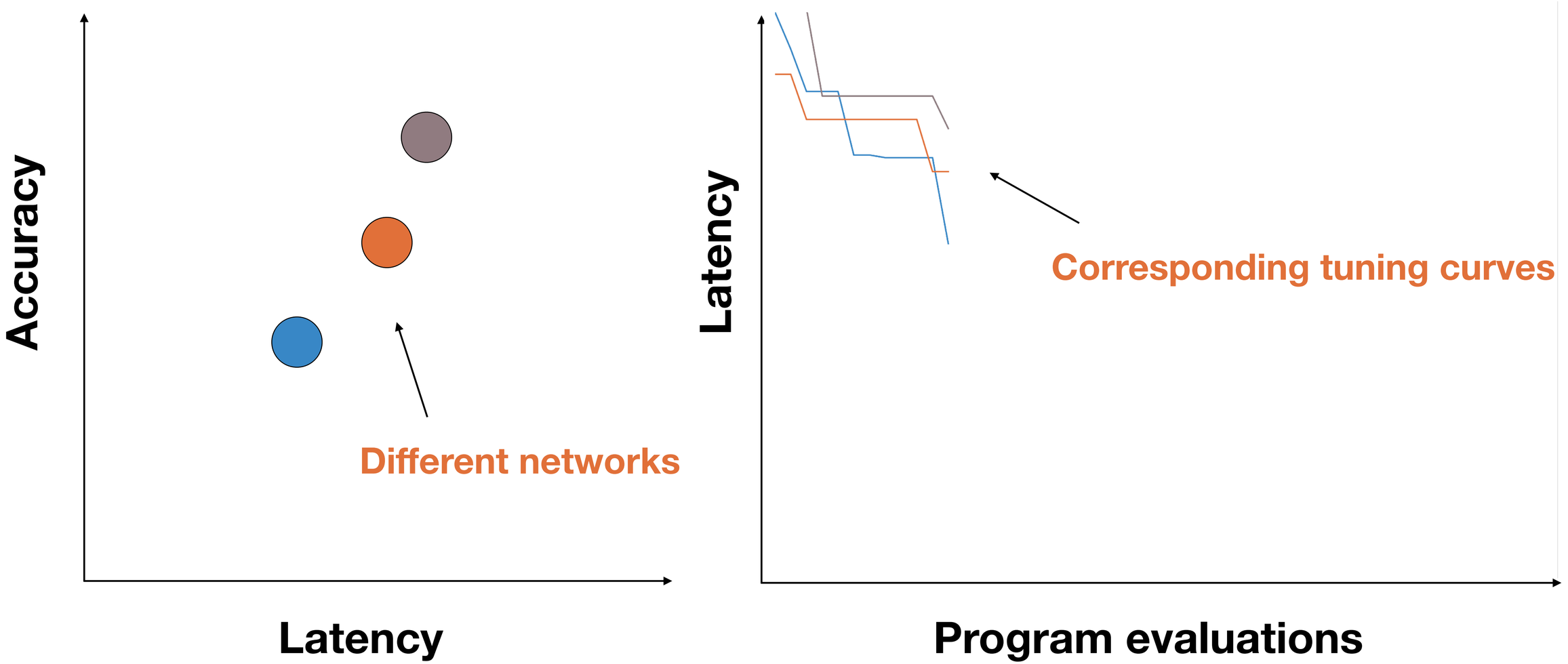




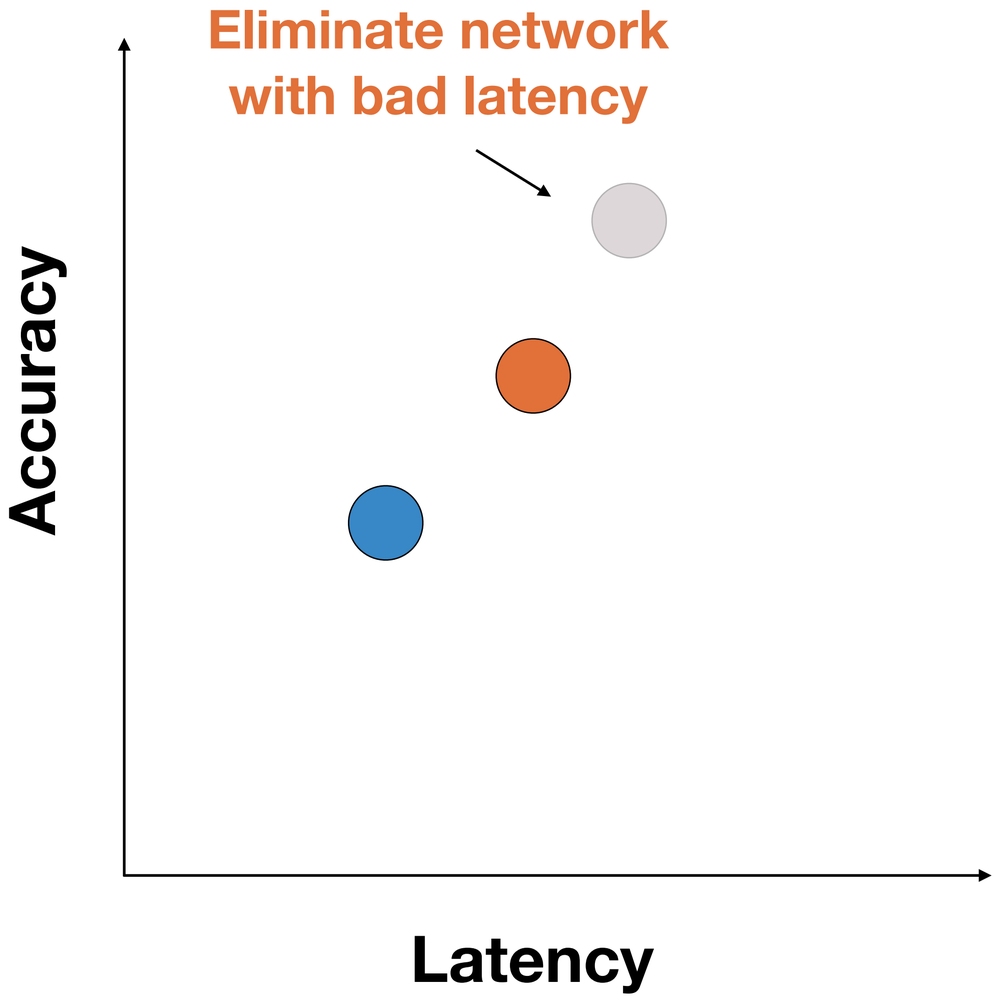
Early stopping for accuracy

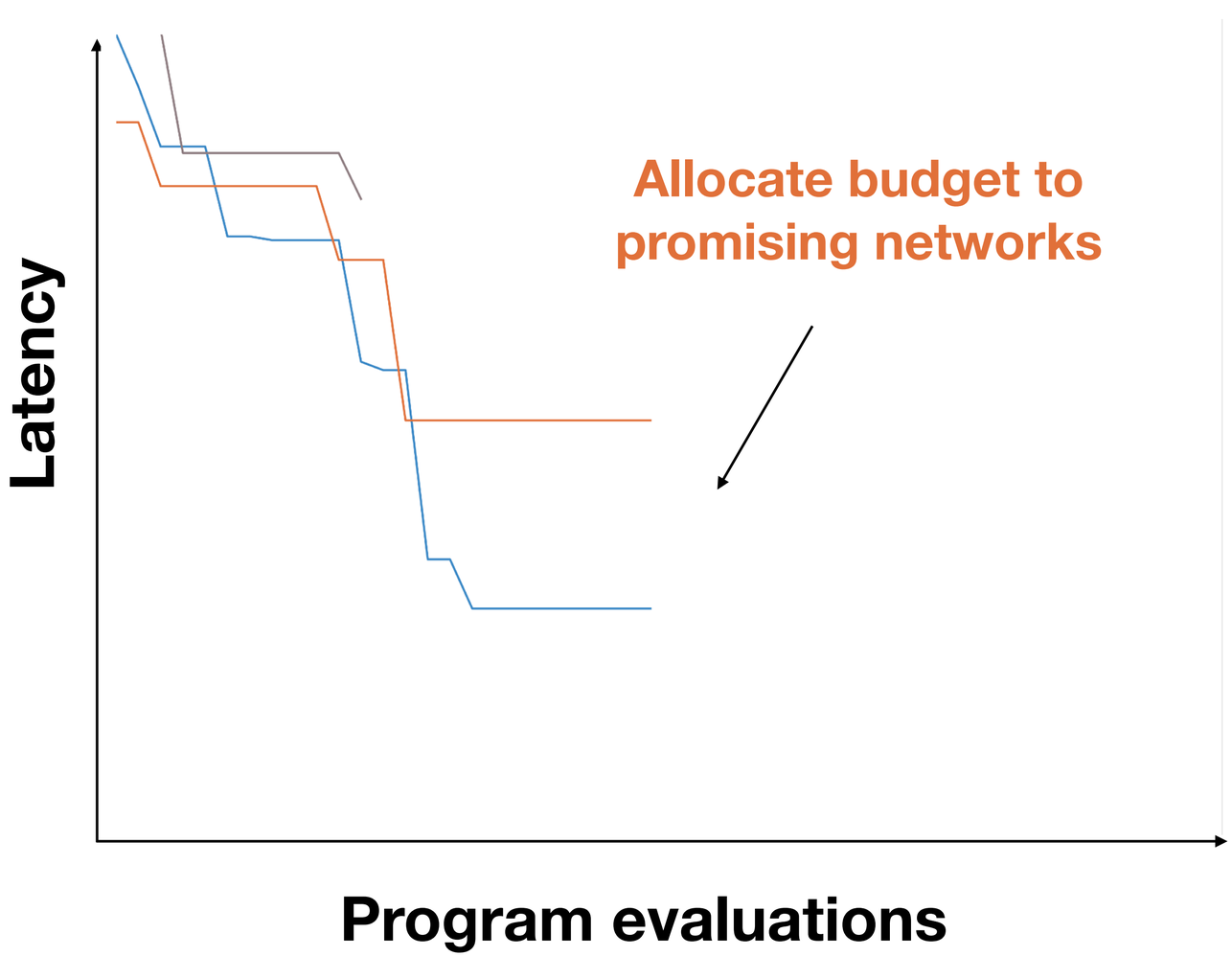


Early stopping for latency

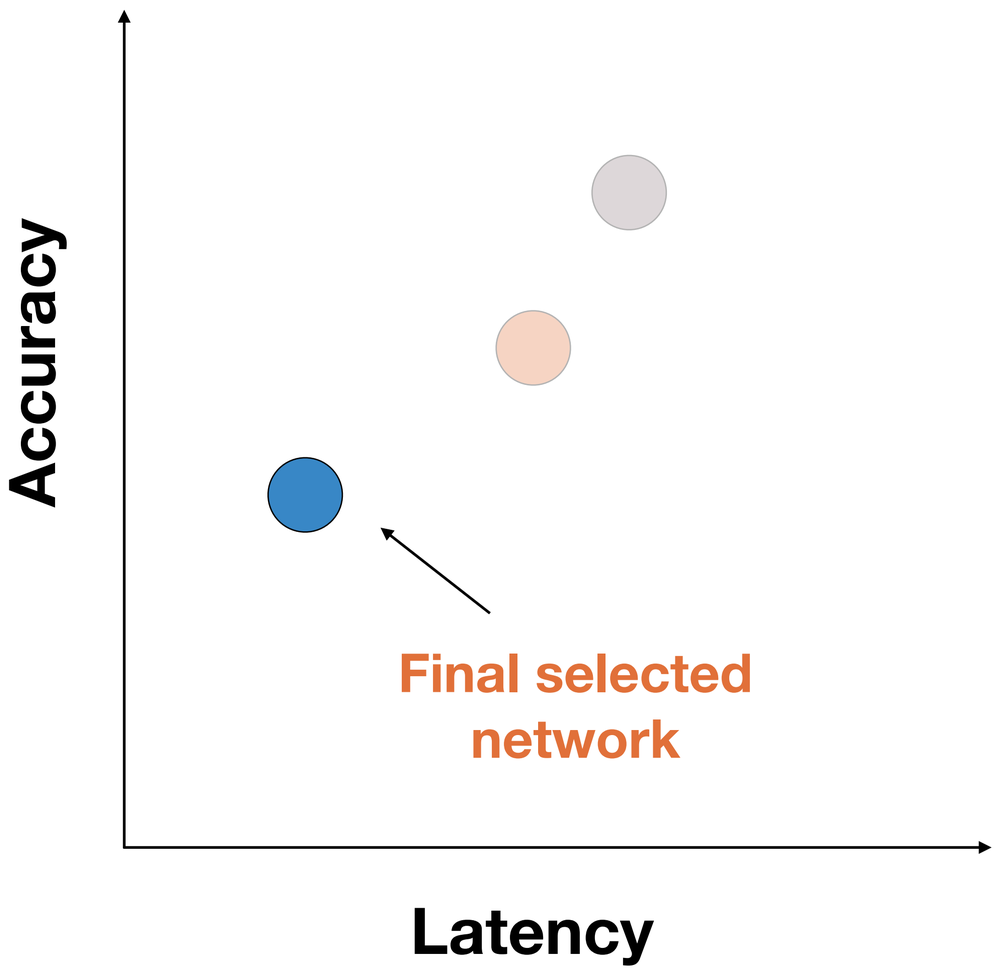


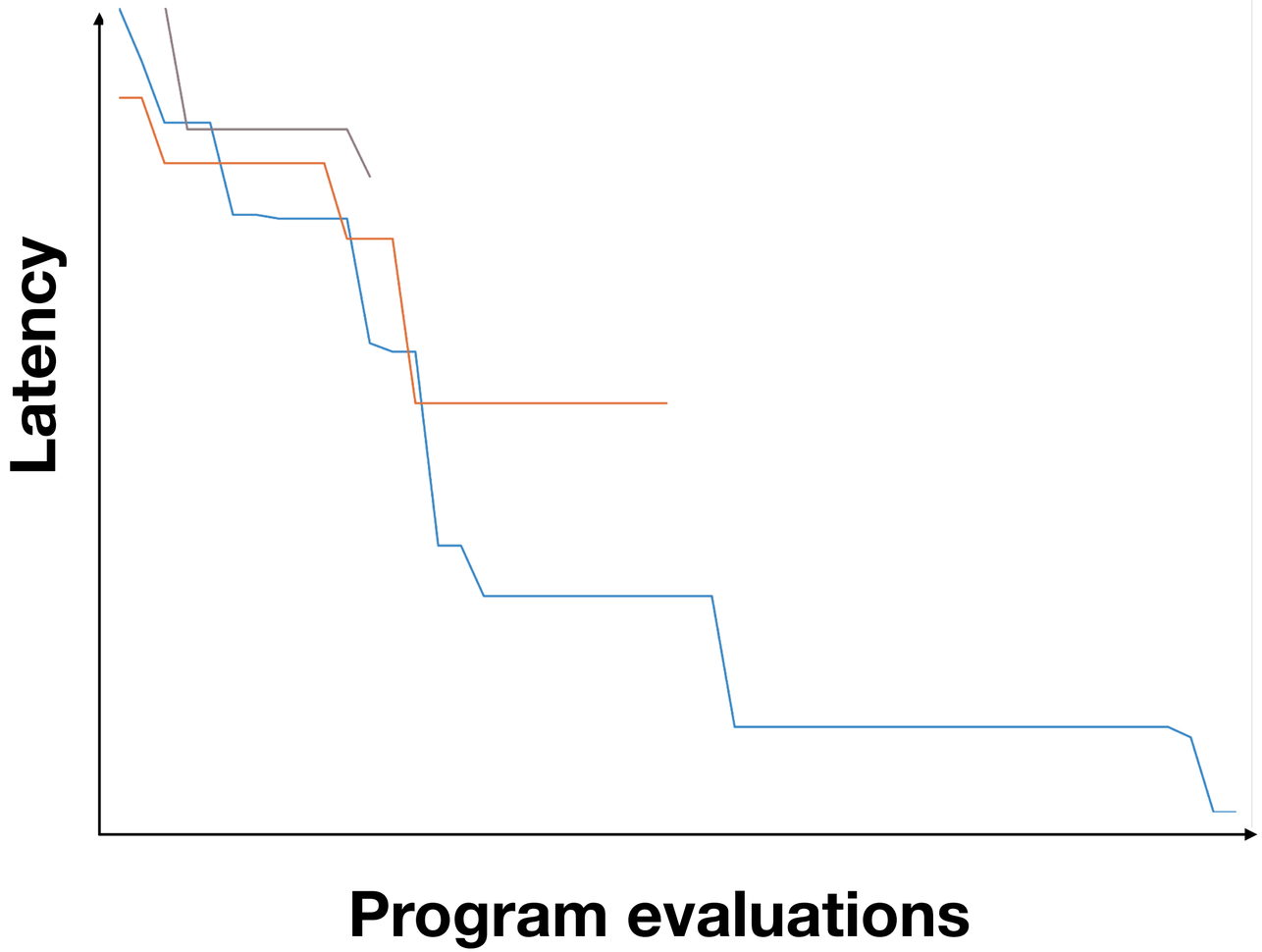
Early stopping for latency



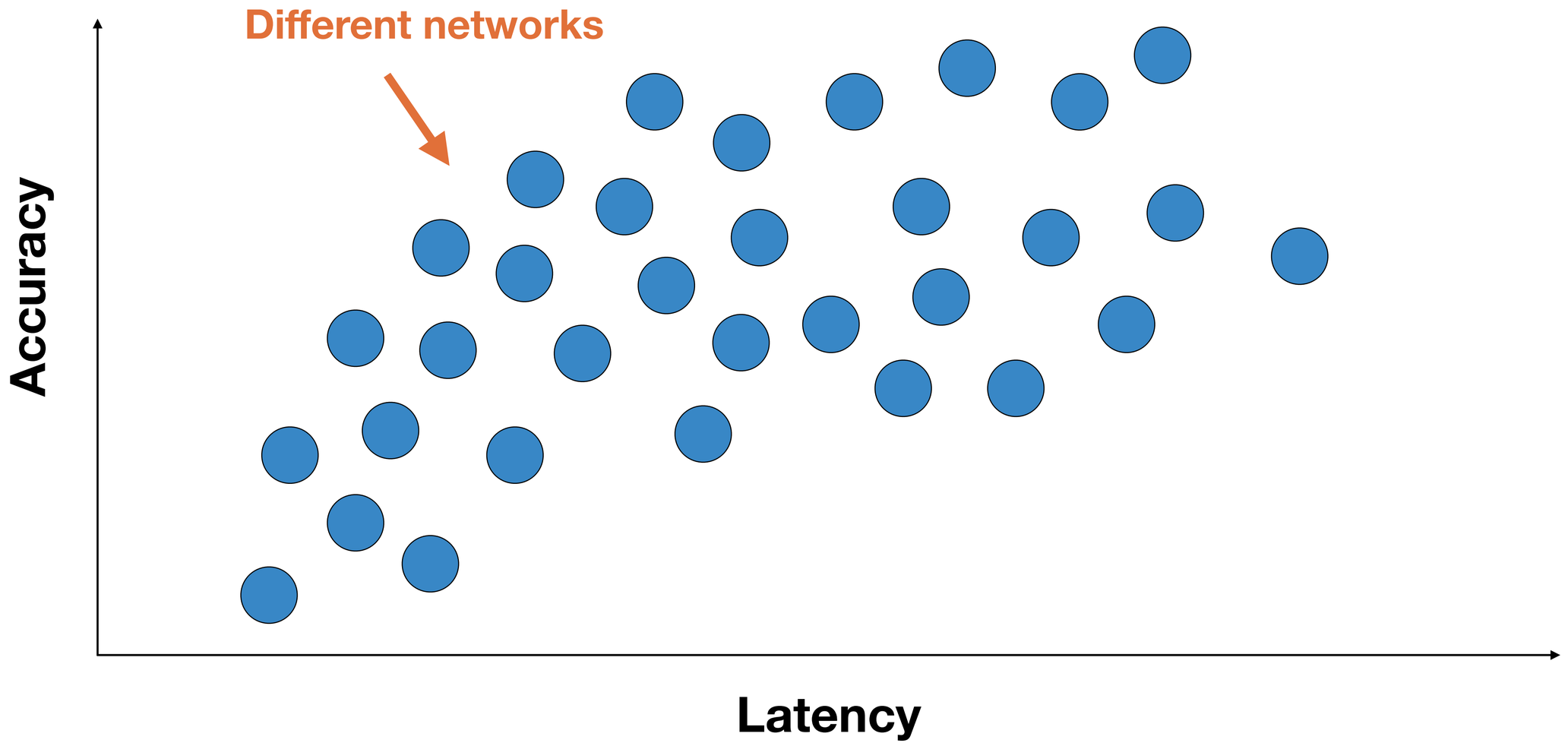


Early stopping for latency

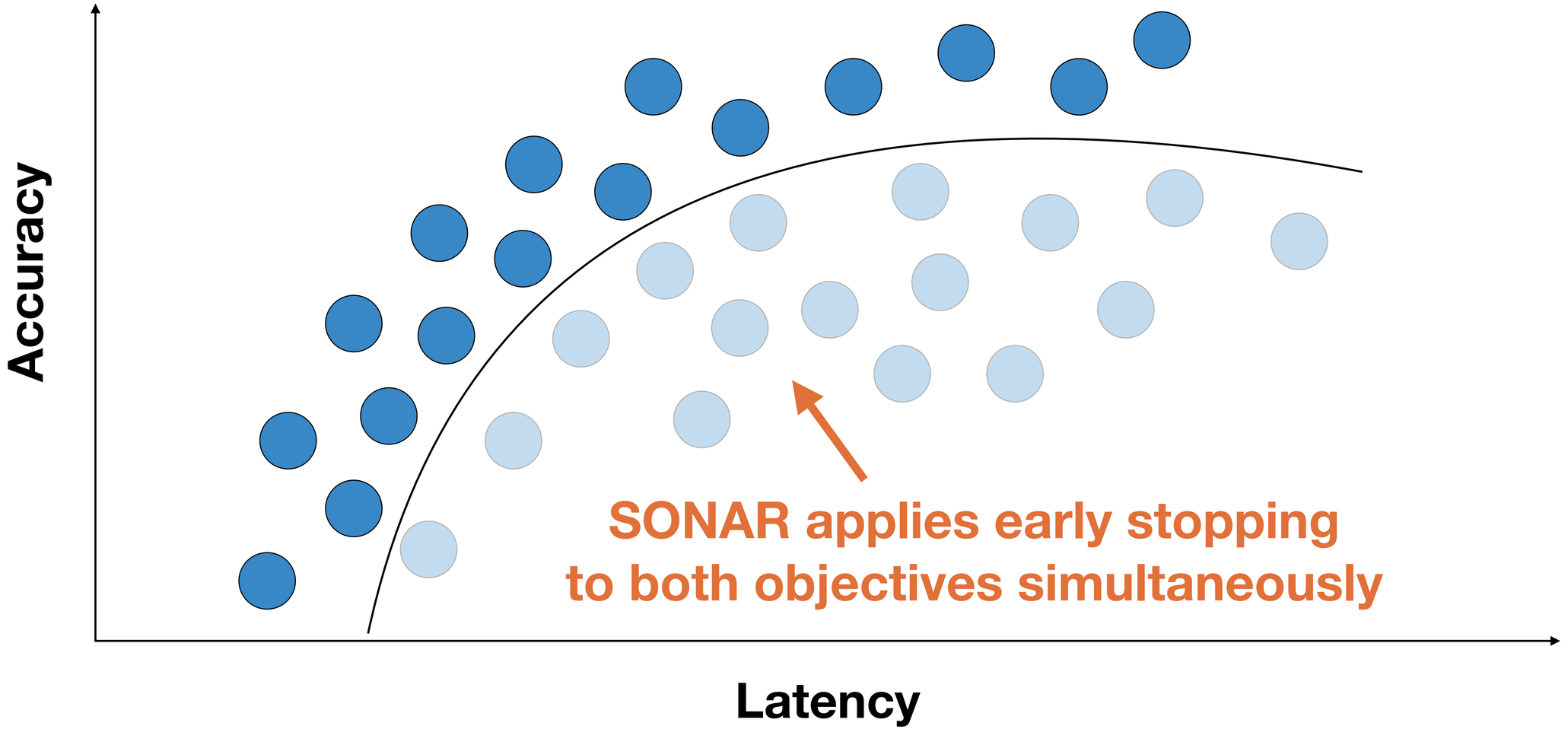




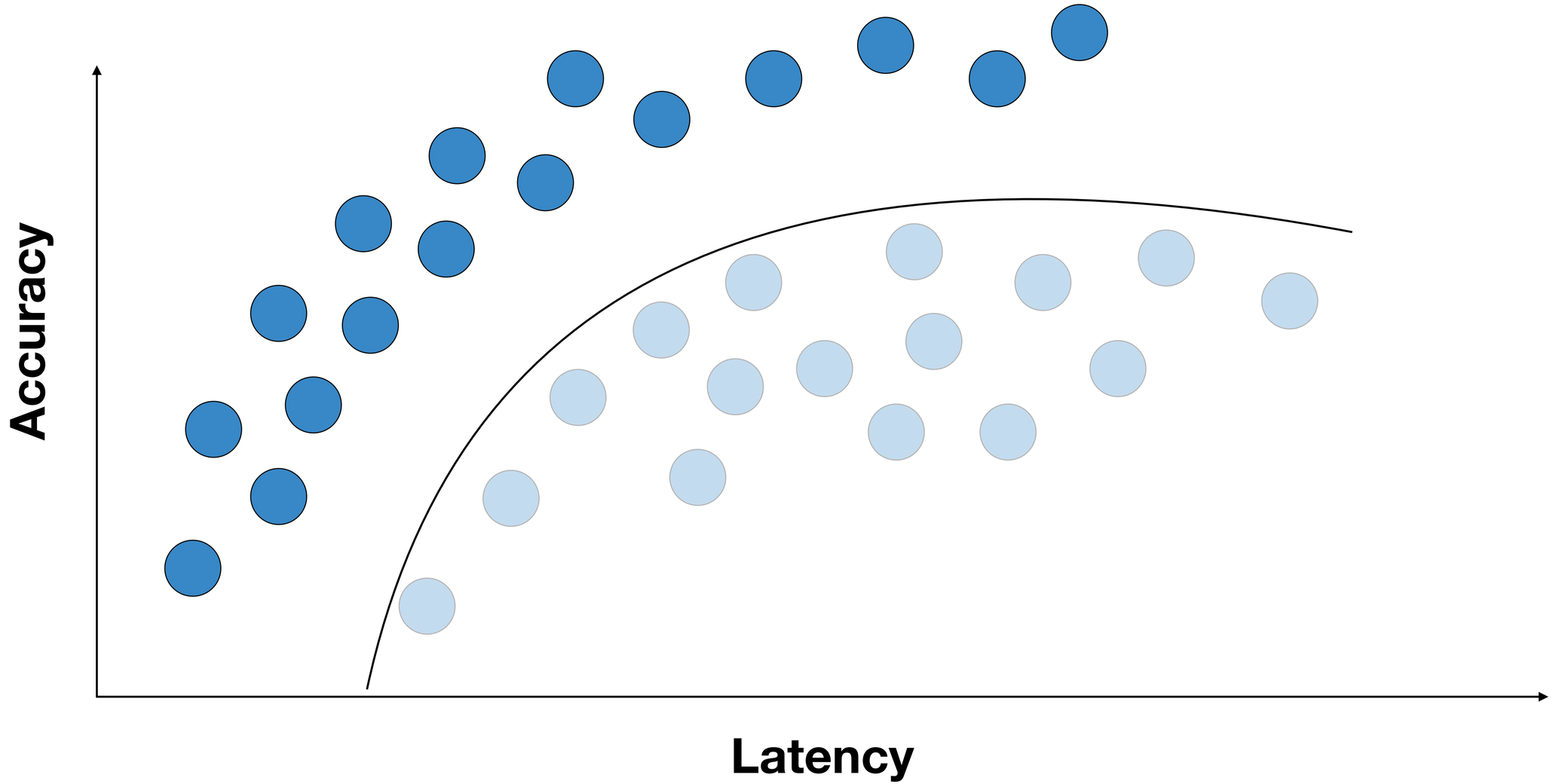
SONAR



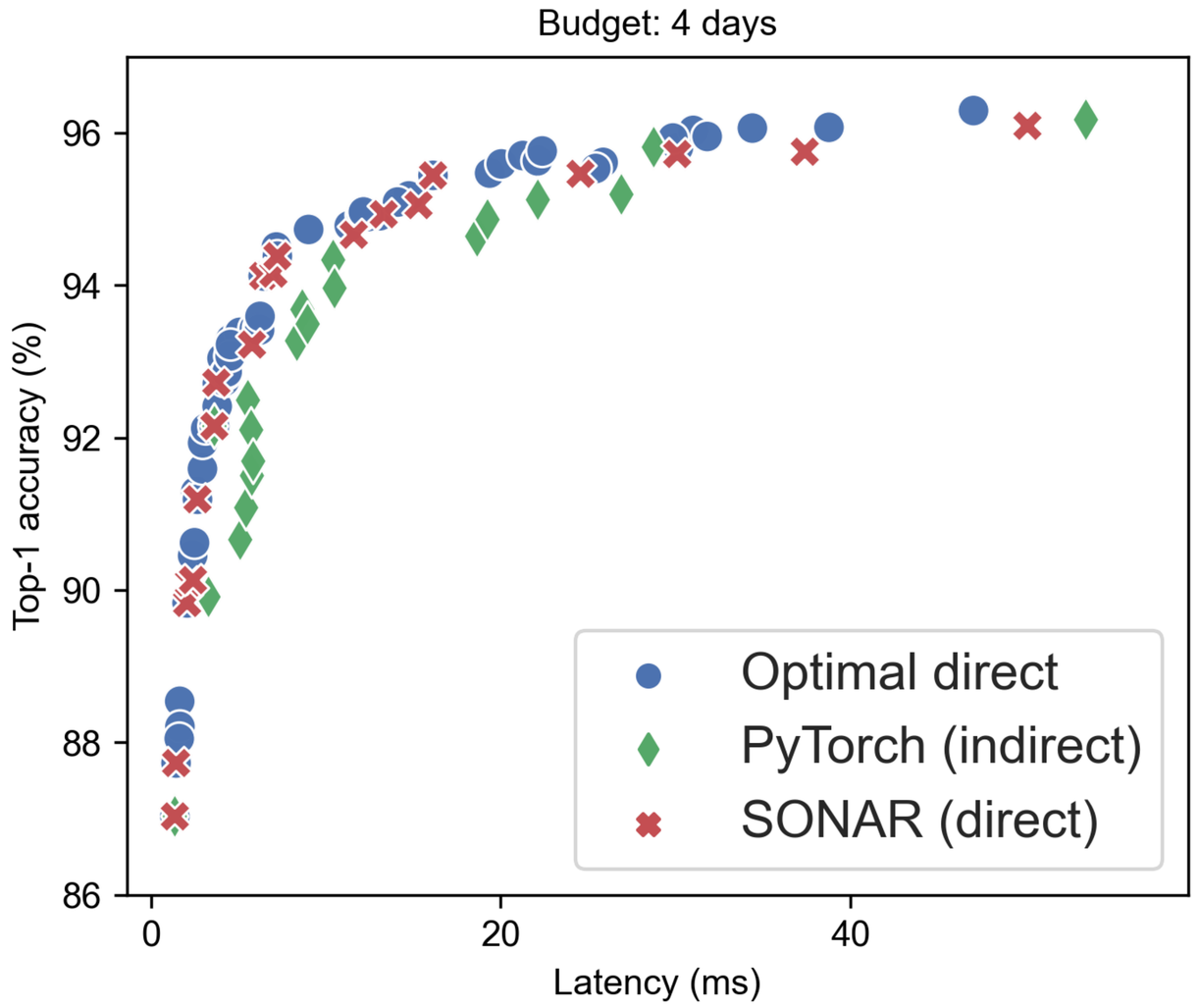
SONAR


Accuracy
Latency
SONAR applies early stopping
to both objectives simultaneously

SONAR



SONAR finds near optimal models



Thank you