15-884: Machine Learning Systems

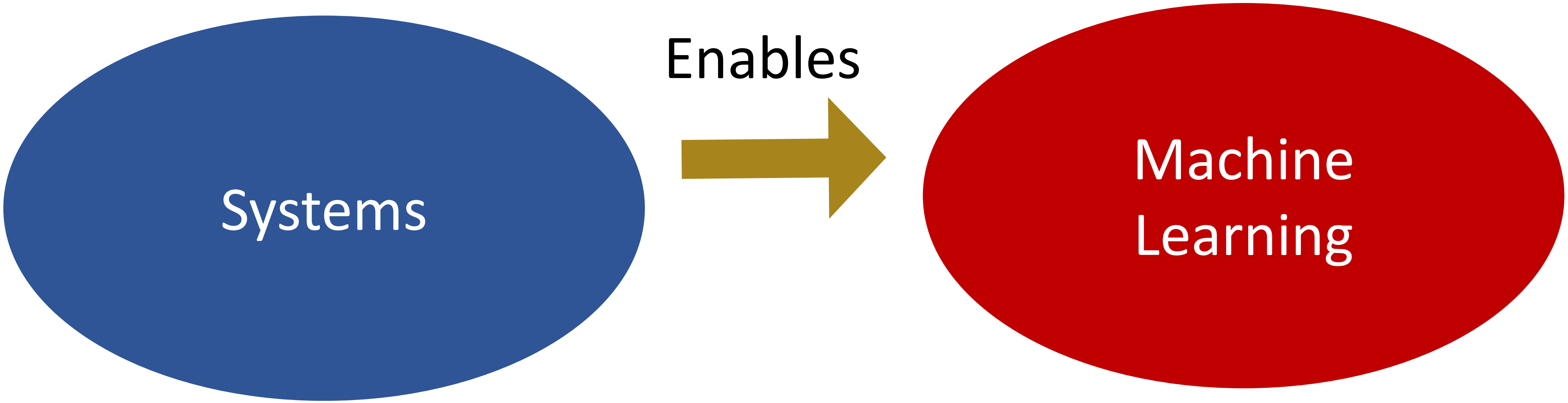
(Problems in) Machine Learning for Systems

Instructor: Tianqi Chen

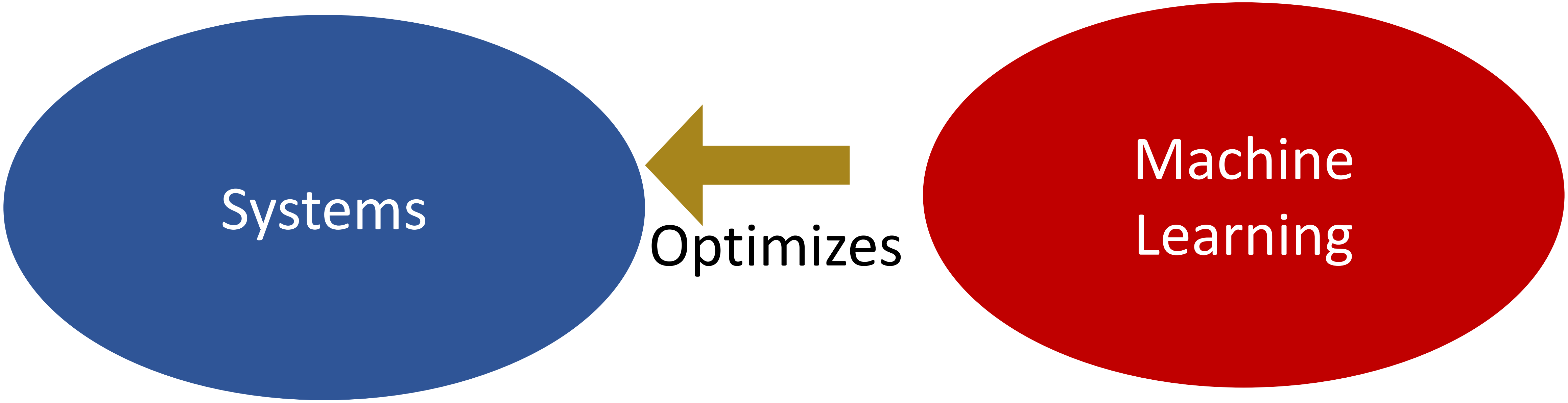
Carnegie Mellon University
School of Computer Science

catalyst

Lectures so Far



Today’s Topic



System Comes with Heuristics

- Compiler Optimizations
- Cache replacement policy
- Scheduling and resource allocation
- …

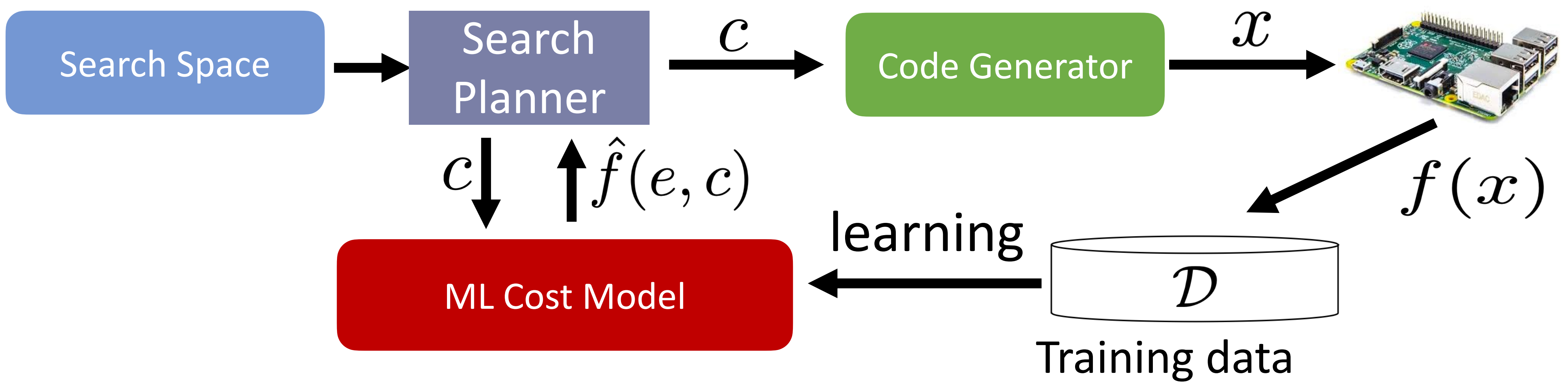
Discussion

What are good characteristics of ML for systems that we can leverage

How do they relates to techniques we apply

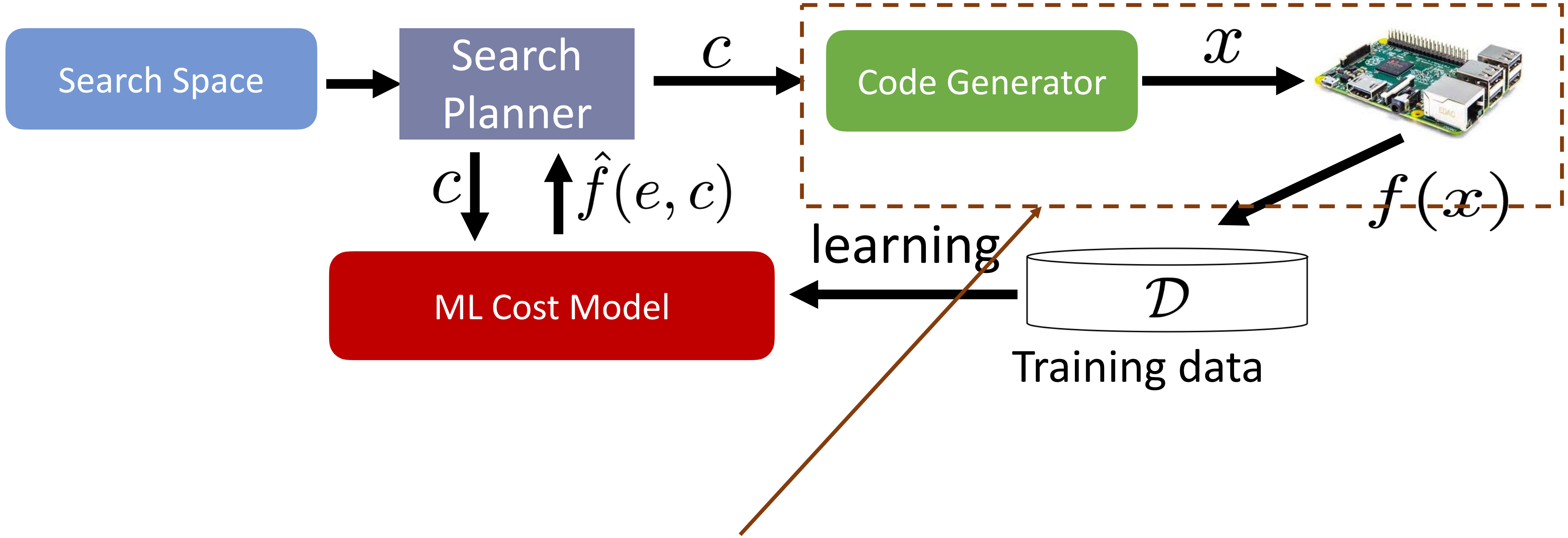
Problem 1: System Constraints

Case Study: Learning a Program Cost Predictor



Use ML based cost model to speedup program search

Case Study: Learning a Program Cost Predictor



Average latency cost: 1 second

System Constraints

- Cost model in program optimization: must run faster than benchmarking
- Cache replacement policy: latency constraint
- …

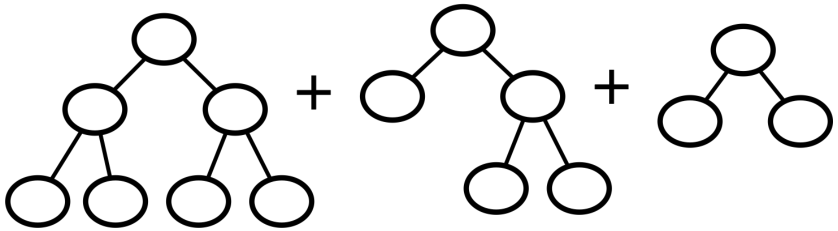
Discussion

What are other example systems constraints

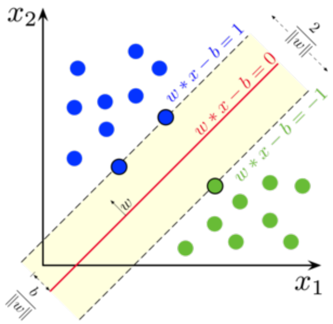
What are the implication for ML models

Common Models

Decision Trees (ensembles)
- Can compiles to actual code.
- Shallow trees interpretable as rules.



Linear model
- Can speedup with pruning (L1 regularization).
- Requires feature engineering.



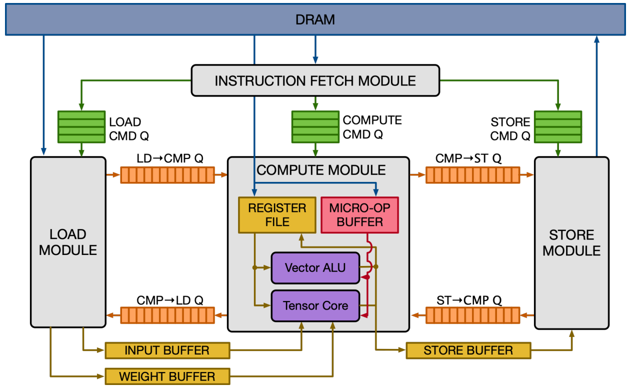
Problem 2: Domain Specific Modeling

Problems of Interest

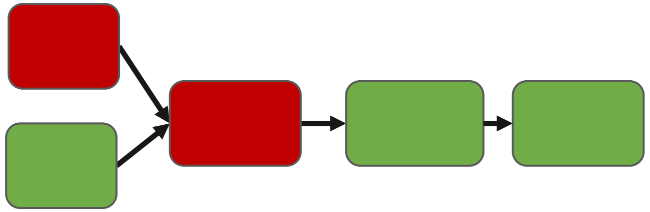
Program ASTs

```
for y in range(8):
  for x in range(8):
    C[y][x]=0
    for k in range(8):
      C[y][x]+=A[k][y]*B[k][x]
```

Hardware Architectures



Device Placement, Planning



Common theme: Structured Data

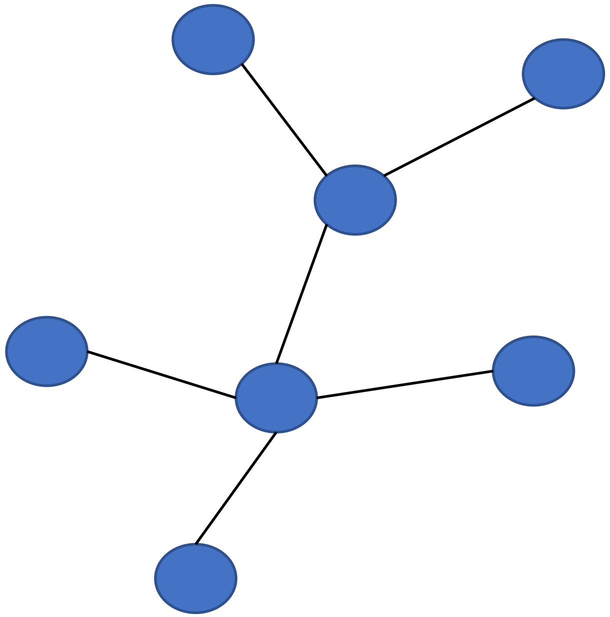
- Trees
- Graphs
- Hierarchical

Discussion

What are other example problem structures

How to build effective models for them

Graph Neural Networks



Embedding of each node

Message

$$m_{u \to v}^{(l)} = M^{(l)} \left(h_v^{(l-1)}, h_u^{(l-1)}, e_{u \to v}^{(l-1)} \right)$$

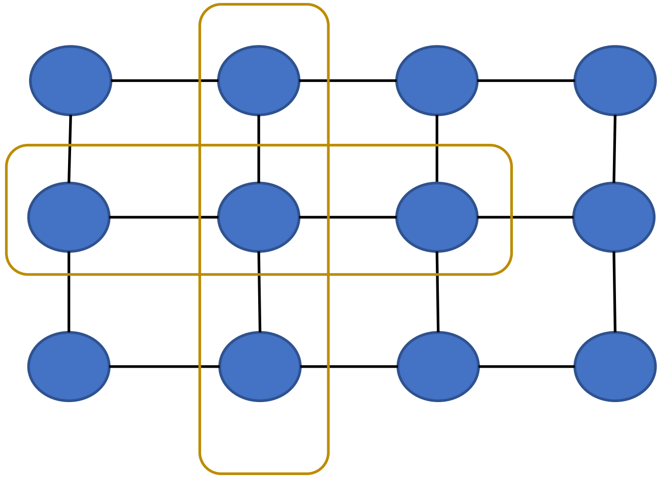
Aggregator

$$m_v^{(l)} = \sum_{u \in \mathcal{N}(v)} m_{u \to v}^{(l)}$$

$$h_v^{(l)} = U^{(l)} \left(h_v^{(l-1)}, m_v^{(l)} \right)$$

Update after aggregation

CNN as GNN



Graph serves a way to represent generic spatial locality

Graph Neural Networks and Structure Learning

- Represent structure as graphs
- Embed node information
- Run GNNs to get updated node state
- Decode per node, or globally

Other Models

- Tree-LSTM
- LSTM for instruction sequence
- Model Cascades(index)

Problem 3: Model Invariance and Generalization

Case Study: Loop Modeling

```
for y in range(8):
  for x in range(8):
    C[y][x]=0
    for k in range(8):
      C[y][x]+=A[k][y]*B[k][x]
```

touched memory

	C	A	B
y	64	64	64
x	8	8	64
k	1	8	8

outer loop length

y	1
x	8
k	64

Per loop level features

Only generalizable to a fixed loop nest structure.

Distribution Drift Problems: Scheduling

- Train on offline simulations of cluster workloads
- Deploy to new cluster workloads

Discussion

What are possible ways to combat distribution drift

Ways to Improve Generalization

- Collect more data
- Build more invariant models
- Disentangle factors (e.g. hardware setup and program)

Summary: Problems in ML for Systems

- System constraints
- Domain structure
- Generalization