With the astounding success of machine learning in general, many researchers and practitioners have turned their attention to the edge. Bare-metal devices are commonly found on the edge, because they are cheap and energy-efficient.

**The Move Towards the Edge**

**The Problem**

Programming these devices is extremely difficult due to:

- Resource constraints
- Lack of compute power
- Lack of on-device memory management
- Restricted language and runtime support
- Tedium debugging

**μTVM's Approach**

- Plug directly into TVM as a backend
- Use the compiler to emit code for the device
- Gain access to TVM models and optimization

**Execution Model**

The current execution model is slow, but it allows us to use TVM's automatic tensor program optimizer.

**AutoTVM Compatibility**

**The End Goal**

**Contact and Acknowledgements**

This work is supported by the Semiconductor Research Corporation (SRC) and DARPA.

Logan Weber  Pratyush Patel  Tianqi Chen
{weberlo, patelp1, tqchen}@cs.uw.edu