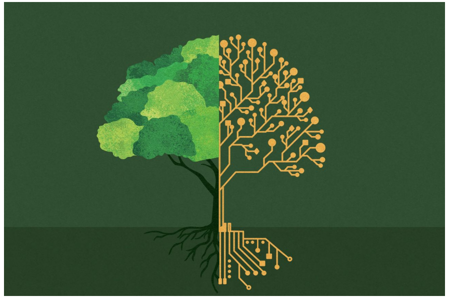
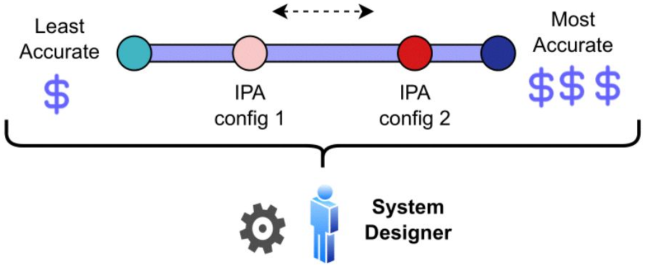
Considering Energy Consumption in IPA Towards Sustainable AI

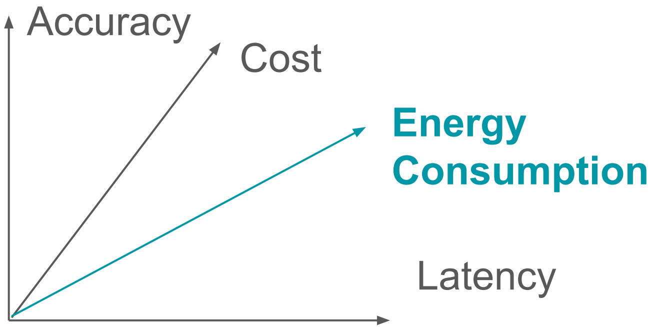
Regan Willis, Chase Bryson, Osasuyi Agho





Considering Energy Consumption in IPA

- Machine learning solutions require a significant amount of energy.
- IPA focuses on the trade-off space between accuracy, cost, and latency, but does not consider energy consumption.
- **<u>Question</u>: Does including energy consumption as a metric expand the trade-off space significantly?**



Objective Space

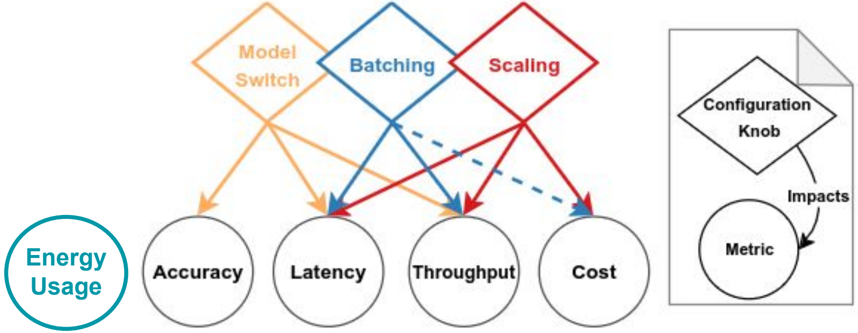


Figure from IPA showing the impact of configuration knobs on metrics with energy consumption added as another metric.

Determining Energy Consumption in IPA

- Design an experiment:
  - Implement IPA on an existing machine learning pipeline.
  - Use IPA to output the adaptation configuration space for a given input.
  - **For every adaptation configuration in the adaptation configuration space: measure the energy consumption.**
- Measuring energy consumption:
  - Energy cost can be difficult to measure, especially retroactively.
  - Chameleon provides tools to capture power measurements on CPUs.
  - NVIDIA Power Capture Analysis Tool helps collect and analyze power data on NVIDIA GPUs.
  - Open source energy consumption measurement tools.
  - NVIDIA GPU power capping.

Does Energy Consumption Expand the Trade-Off Space?

- **Determine if energy consumption is tightly coupled with any existing metric in IPA** (accuracy, latency, cost) by analyzing the results of all four objectives on every adaptation configuration in the adaptation configuration space.
  - **If not, energy consumption is a useful metric in IPA** and can be considered in the adaptation configuration search space.
    - Is our chosen energy consumption measurement generalizable?
  - **If so, energy consumption is not a relevant metric in the existing IPA system.**
    - Ask what changes to IPA may cause energy consumption to be a useful metric.
      - GPU time-slicing: shared access to a GPU with guaranteed equal share of time.