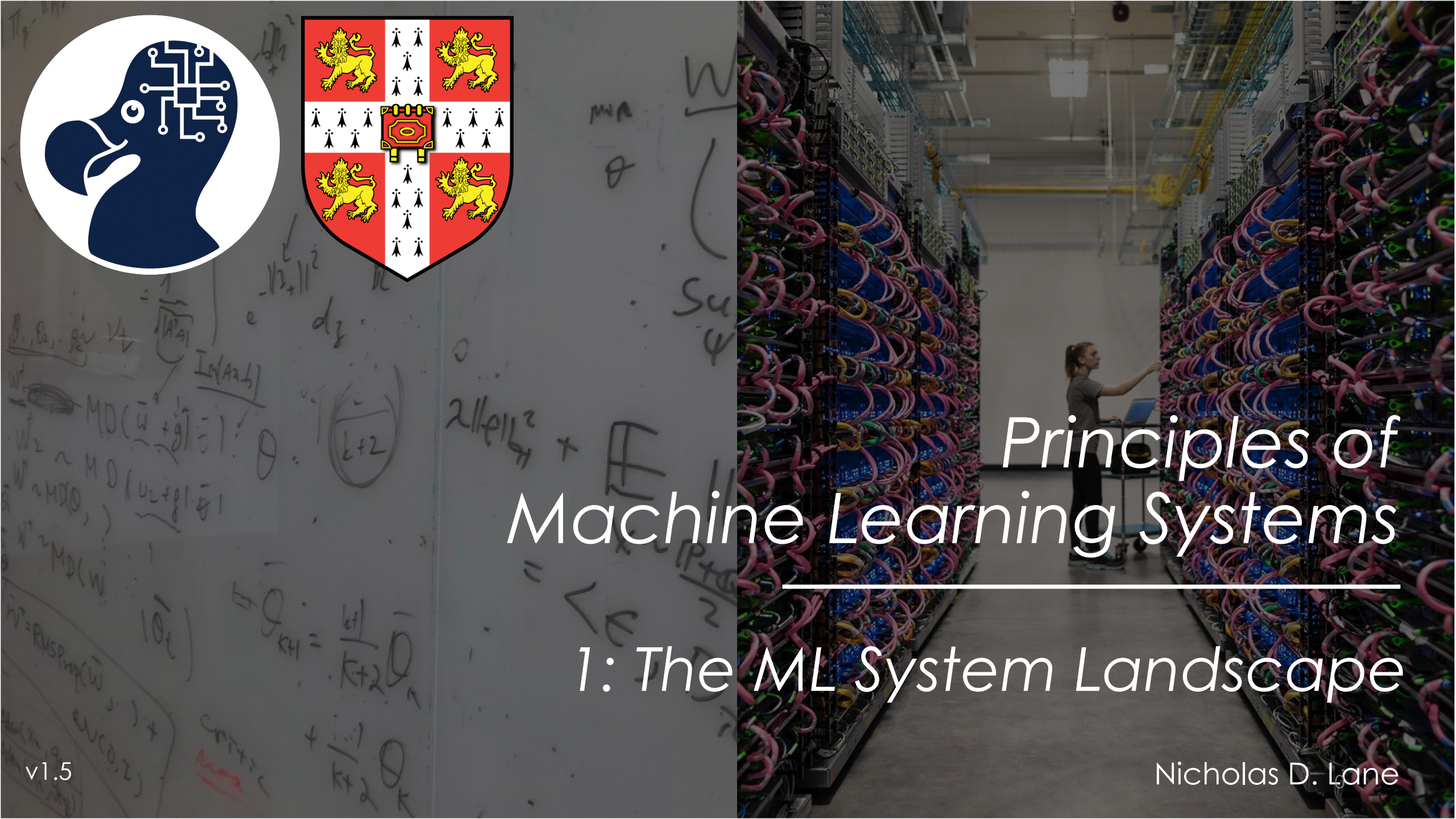


Principles of Machine Learning Systems

1: The ML System Landscape

v1.5

Nicholas D. Lane

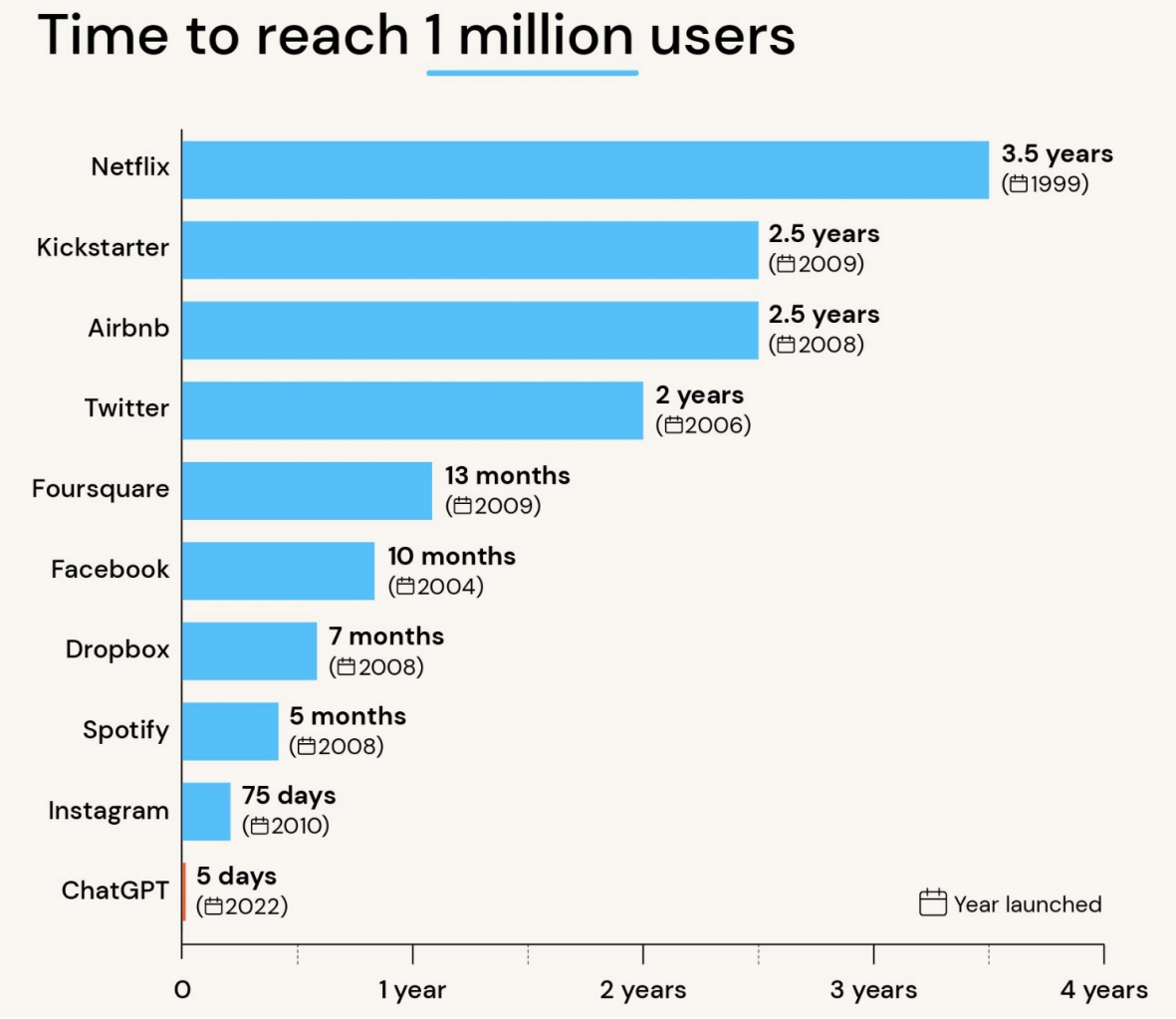
Roadmap for Today

1. Introduction
2. Illustrative Examples
3. Efficiency
4. Open Problems

Roadmap for Today

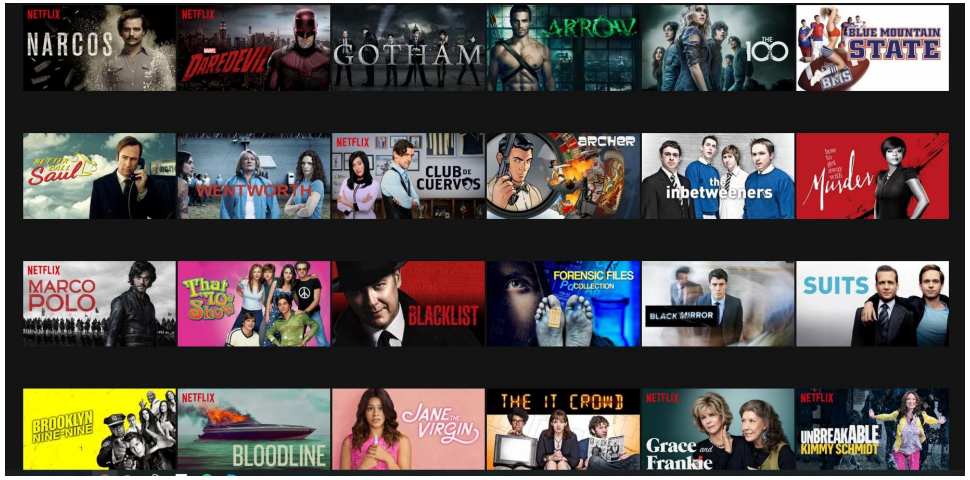
1. **Introduction**
2. Illustrative Examples
3. Efficiency
4. Open Problems

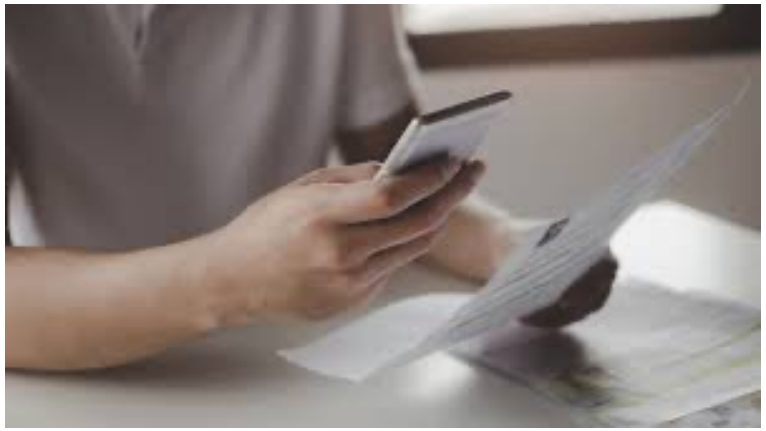
ML Systems in the Wild

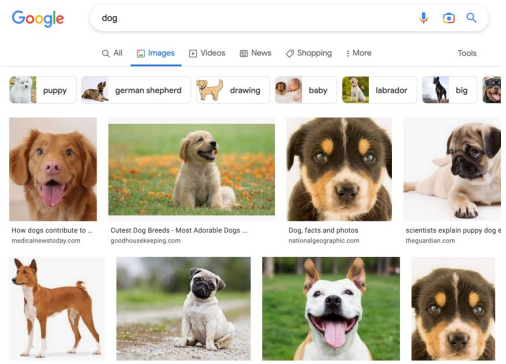


Source: Economist

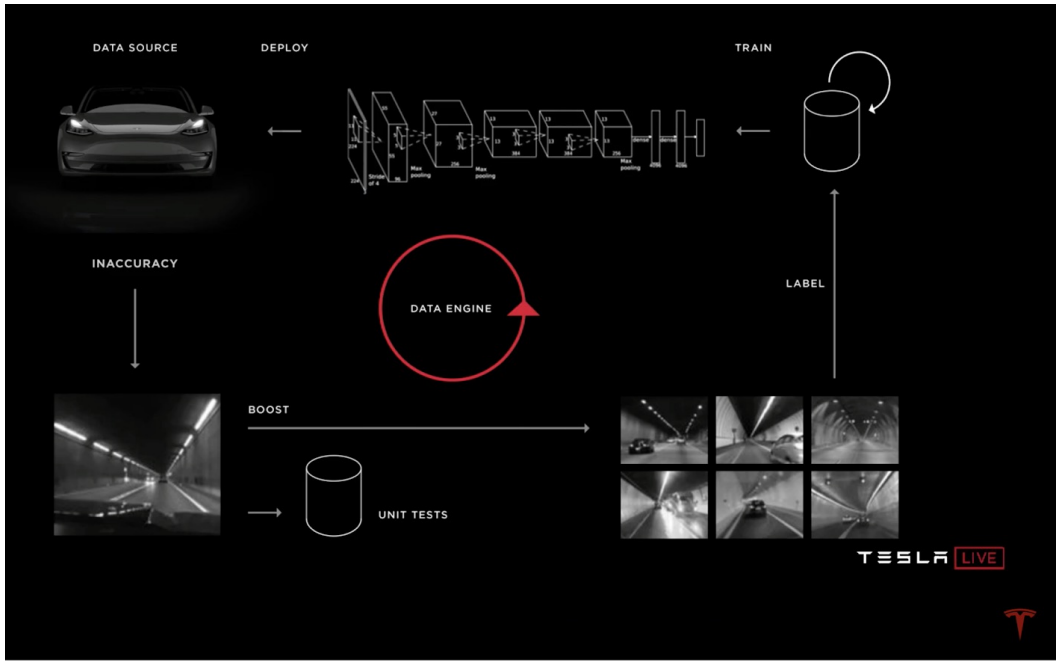
ML Systems in the Wild

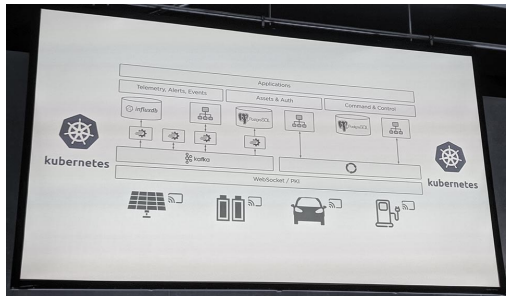


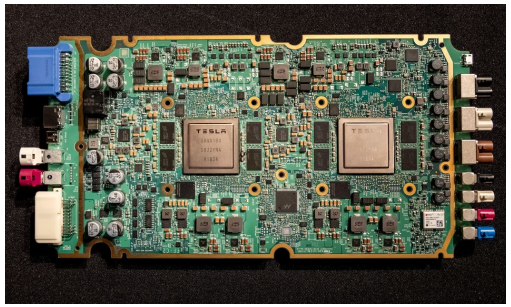


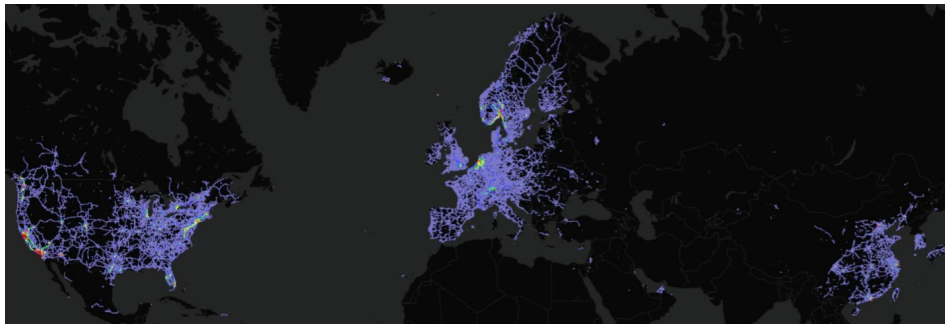


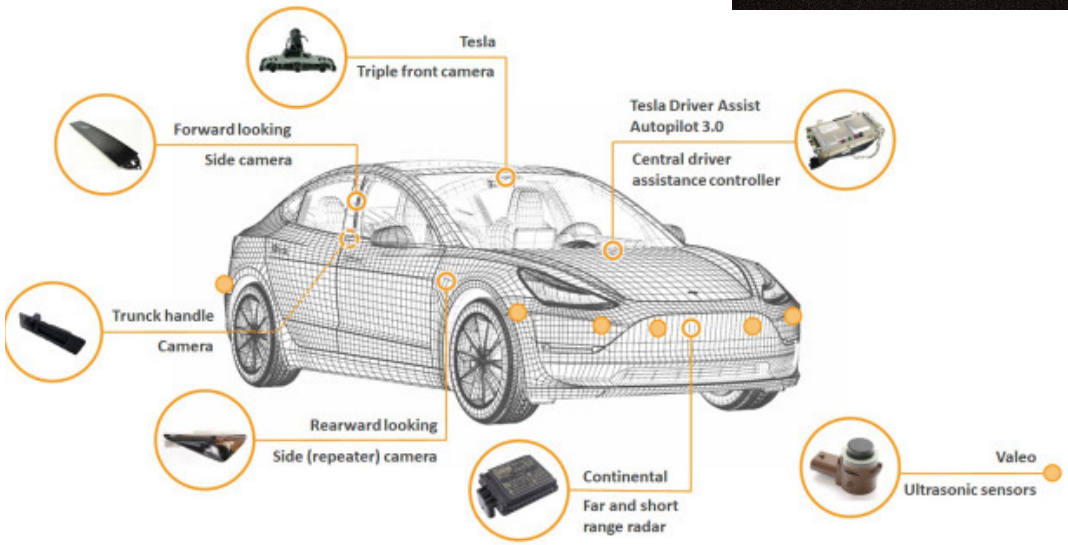
ML Systems in the Wild



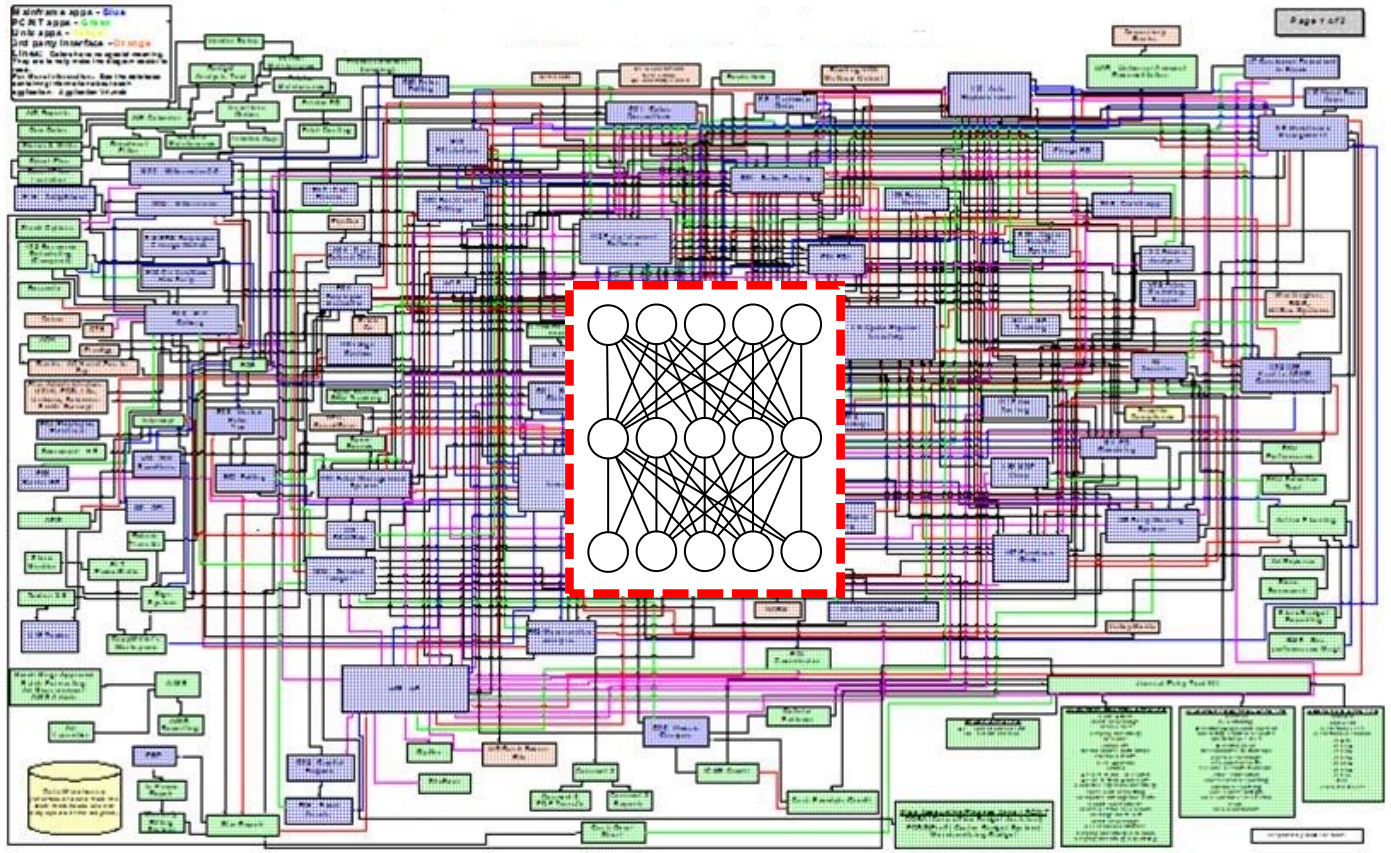








What is an ML System?



Varieties

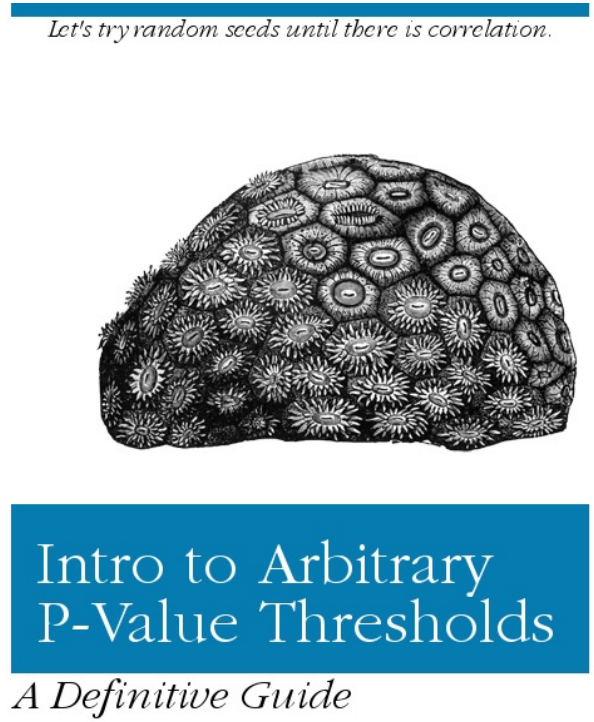
- Enabling an ML-powered service
  - e.g., ASR system
- An ML-enabler
  - e.g., NAS system

Computing system with ML at its core

Role of this course

Fundamental ML Course

Practical ML Course



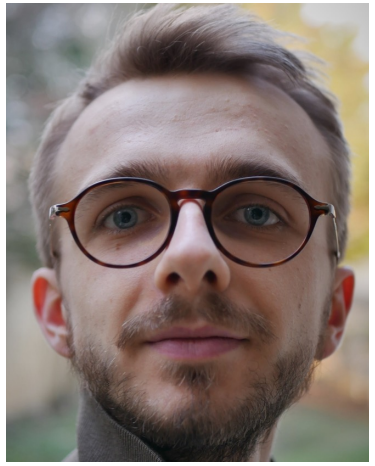
We are here

Principles of ML Systems

10 Lectures, covering:

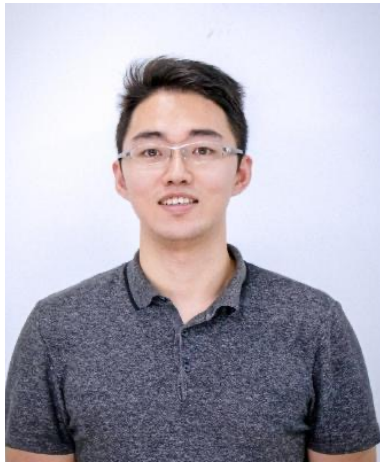
- ML Systems Landscape
- Mapping to Hardware
- Model Compression
- Accelerators: GPUs, NPUs
- Frameworks and Run-times
- Single/Multi GPU Training
- Scalable Inference Serving
- Deep Learning Compilers
- Automated ML
- Federated Learning
- Development Practices
- MLOps related

The Team



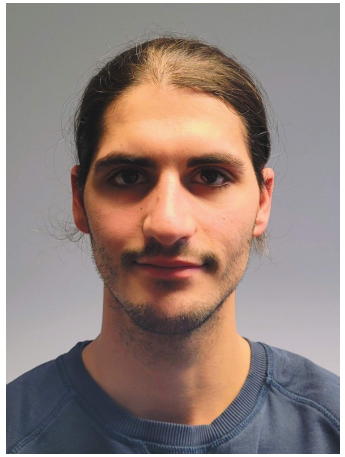
Titouan Parcollet

Filip Svoboda

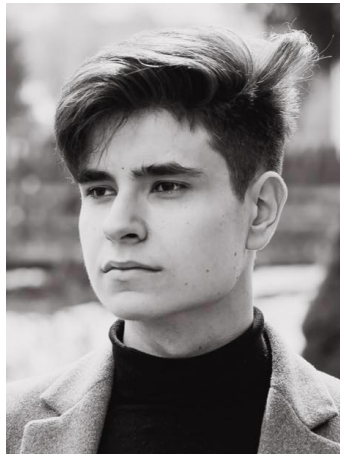


Hongxiang Fan

Nic Lane



Lorenzo Sani

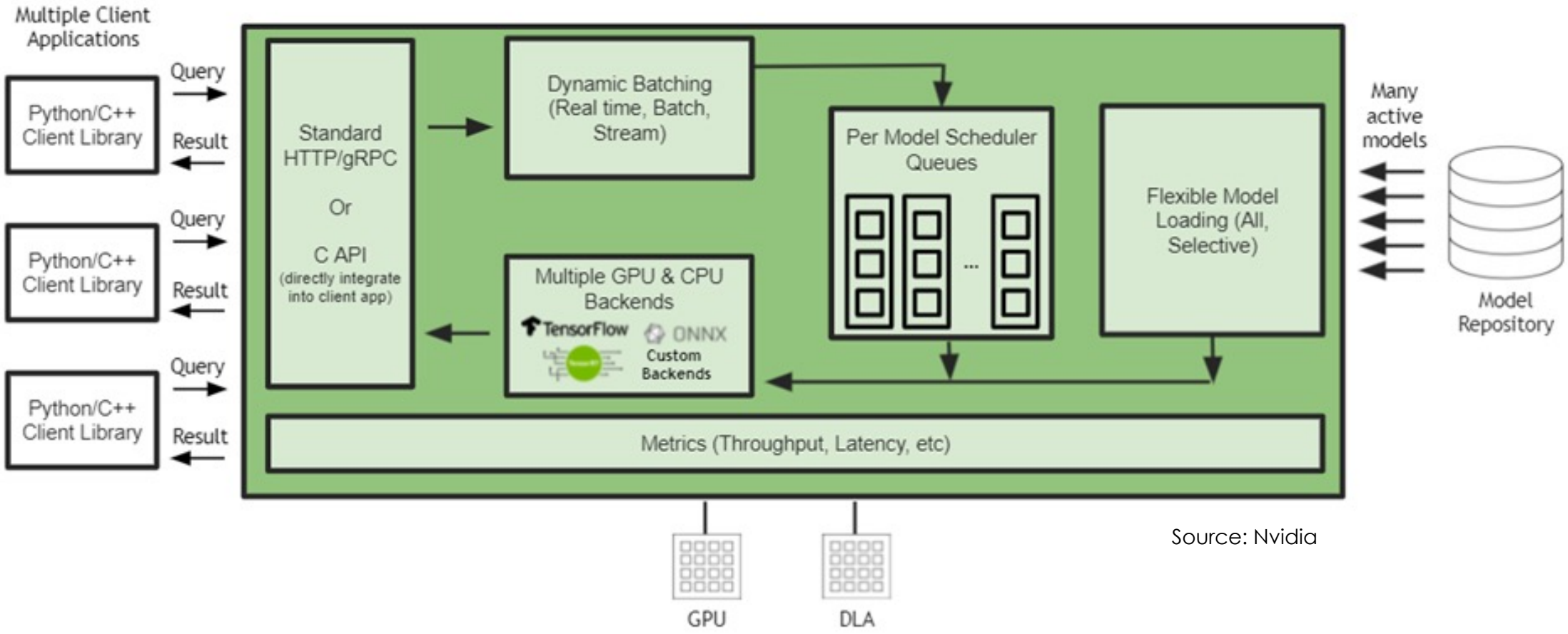


Alex Iacob

Roadmap for Today

1. Introduction
2. **Illustrative Examples**
3. Efficiency
4. Open Problems

Model Serving

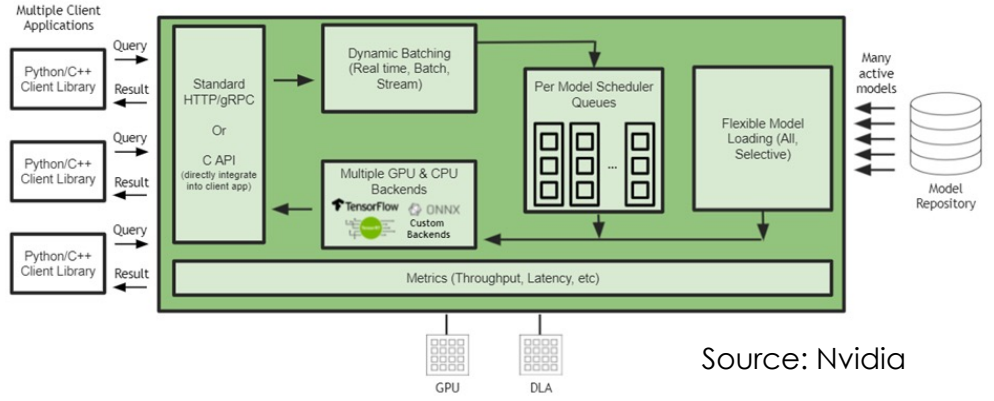


Source: Nvidia

Performance Measures

- **Accuracy** – determines if it can perform the task
  - In an ML class: often only metric a model is judged on
  - Others.. {uncertainty, generalization, noisy/clean conditions, robustness, variety of accuracy metrics}
- **Throughput** – determines if it can handle the data flow
  - Number of PEs utilized (not just peak performance)
  - Real-time performance 3
- **Latency** – operate in real time
- **Flexibility** – range of tasks…

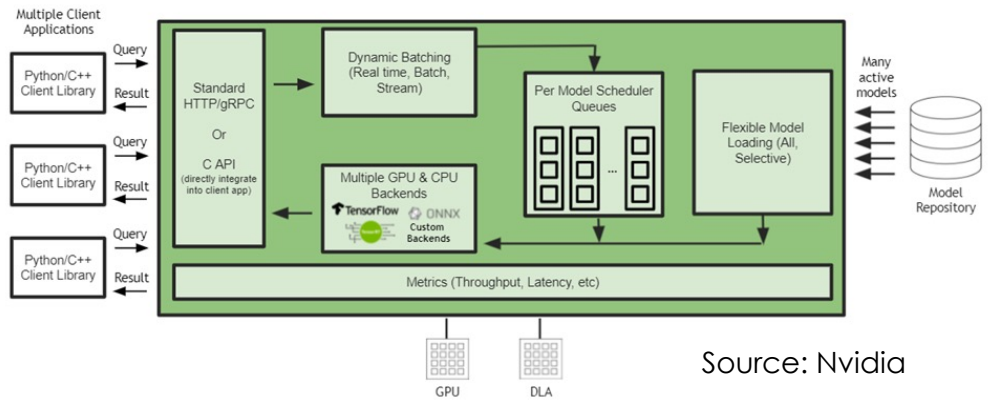
Source: V. Sze



Source: Nvidia

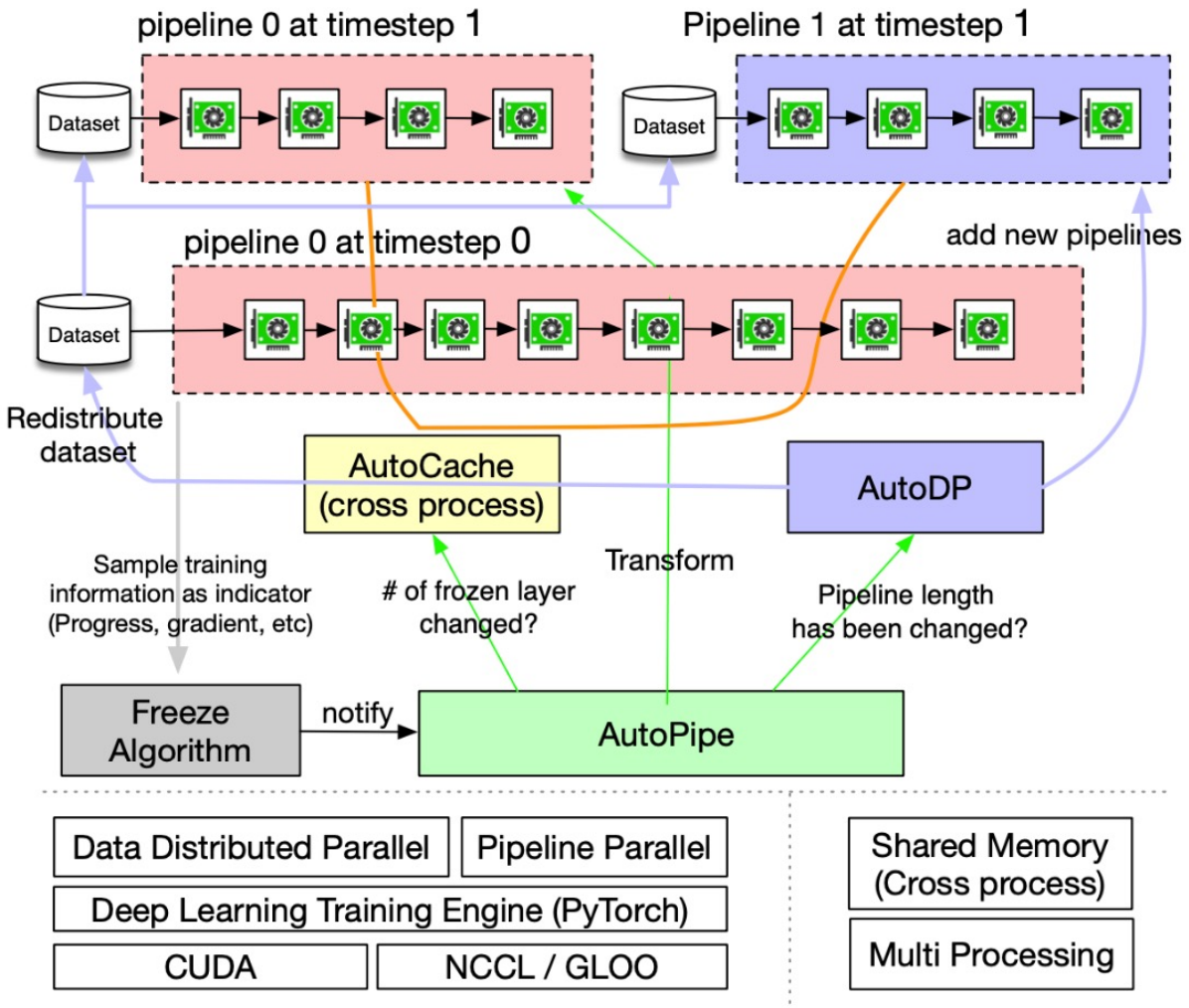
Design Considerations

- **Hardware** – GPU or CPU, maybe FPGA
- **Model Compression** – Accuracy vs. Model Size/Compute; Quantization
- **Remote vs. Local Computation**
- **Frequency of Model Updates** – Weeks or Months
- **Localization of Models** – e.g., languages
- **Auto-scaling** – Burst events



Source: Nvidia

Model Training



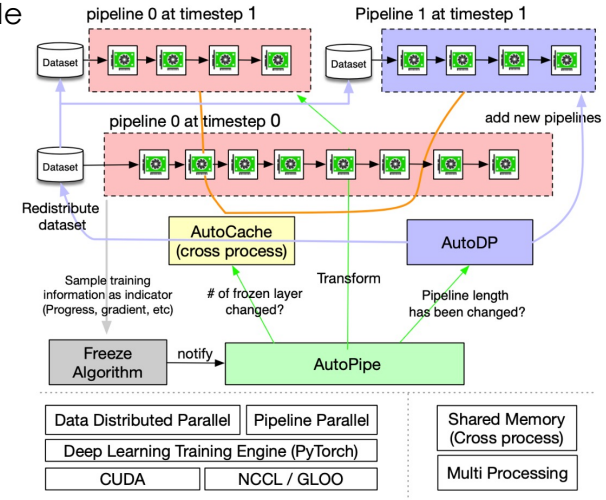
Source: He

Performance measures

- **Energy and Power**
  - Power consumption for running specific model architectures
  - Off-chip memory access (e.g., DRAM)
- **Cost and Training Time**
- **Scalability**
  - Data and model architecture scaling
  - How much room is there in design to grow
  - How to metrics change as scale increases (Does efficiency decline?)

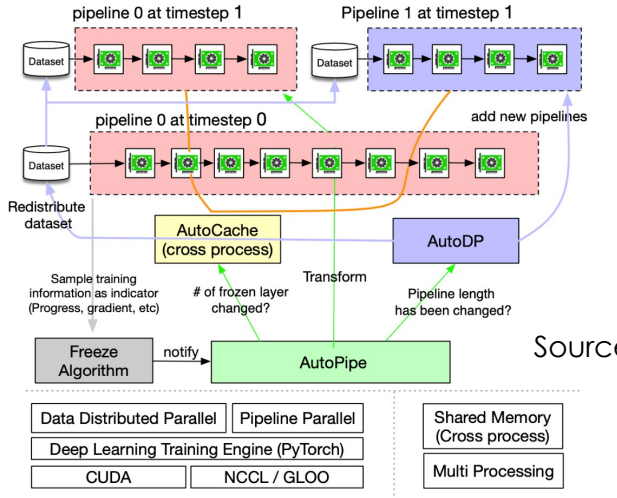
Source: V. Sze

Source: He



Design Considerations

- **Level of Specialization** – NPU, communication
- **Model Architectures and Size**
- **Data Storage** – Will data change frequently
- **Carbon Footprint**– Source of energy
- **Storage Hierarchy**– Speed and location



Source: He

Roadmap for Today

1. Introduction
2. Illustrative Examples
3. **Efficiency**
4. Open Problems

The Deep Learning Era

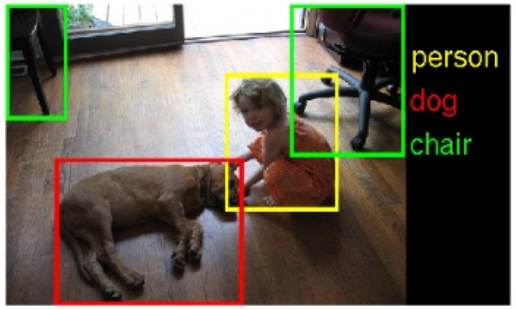
IMAGENET

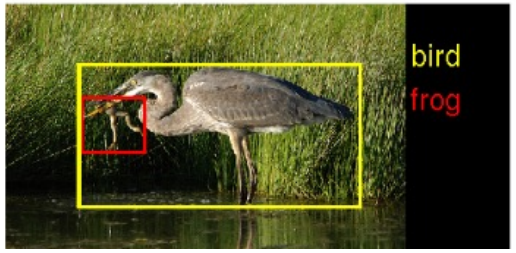
Image Classification Task:

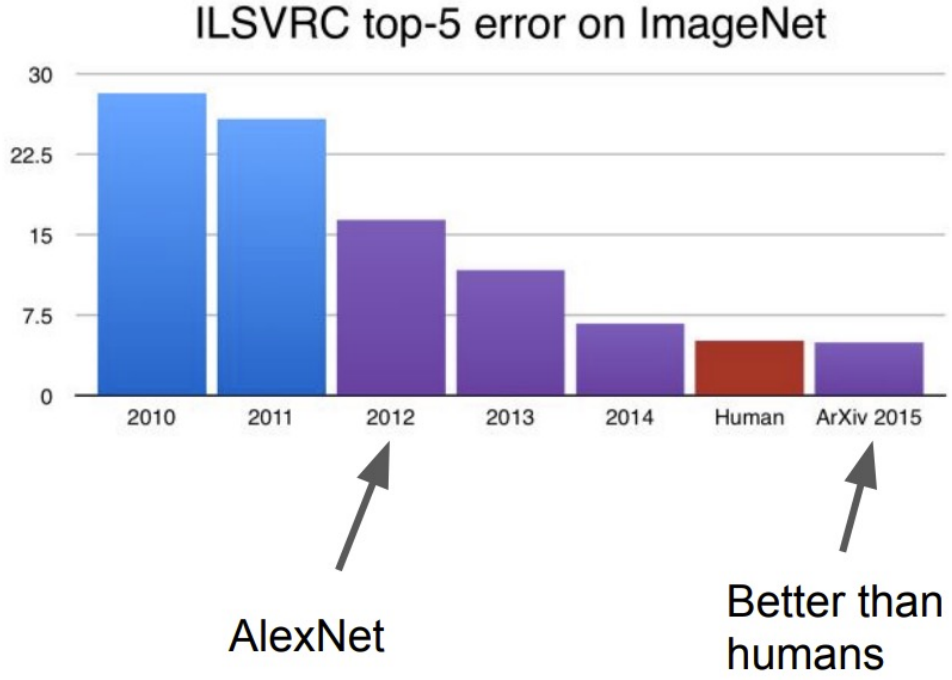
1.2M training images • 1000 object categories

Object Detection Task:

456k training images • *200 object categories*

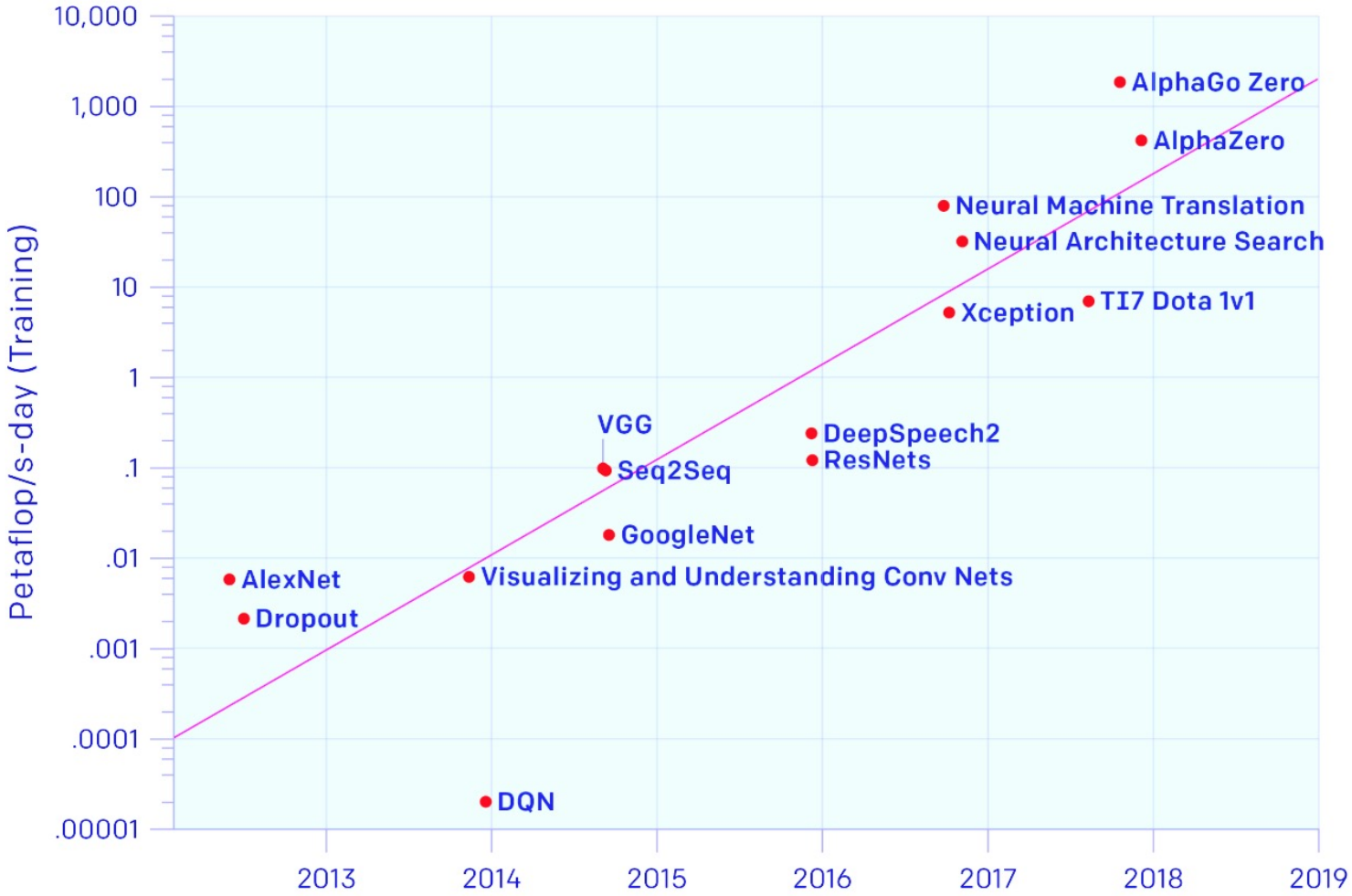






Krizhevsky, Sutskever, Hinton 2012

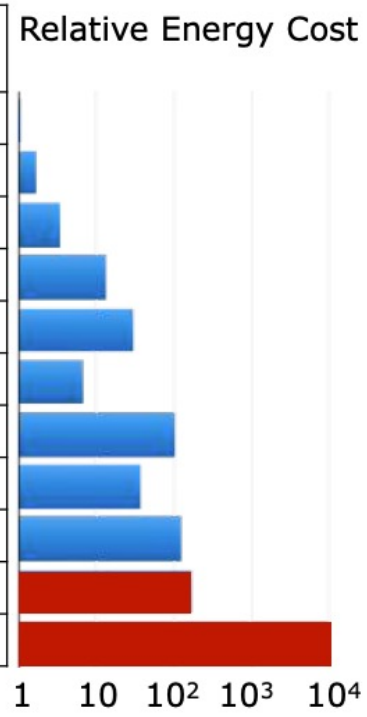
Deep Learning ❤ Resources



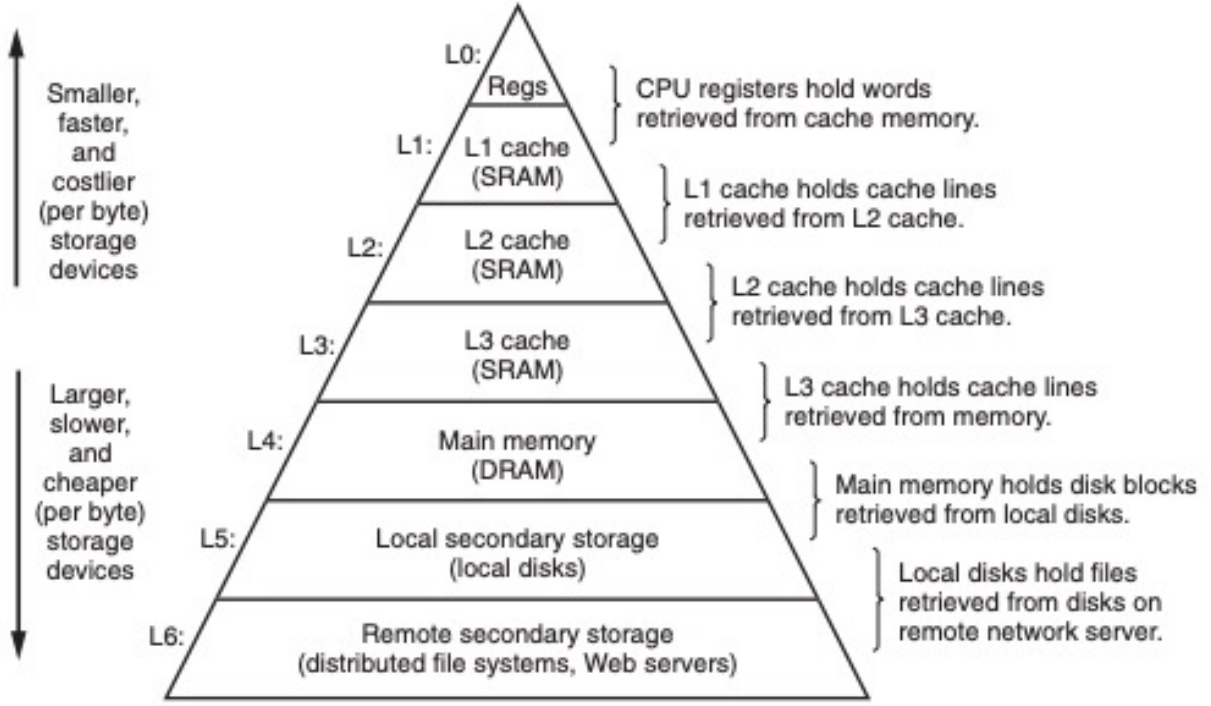
Source: Open.AI

Memory Hierarchy Pressure

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Multiply	0.2
32b Multiply	3.1
16b FP Multiply	1.1
32b FP Multiply	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

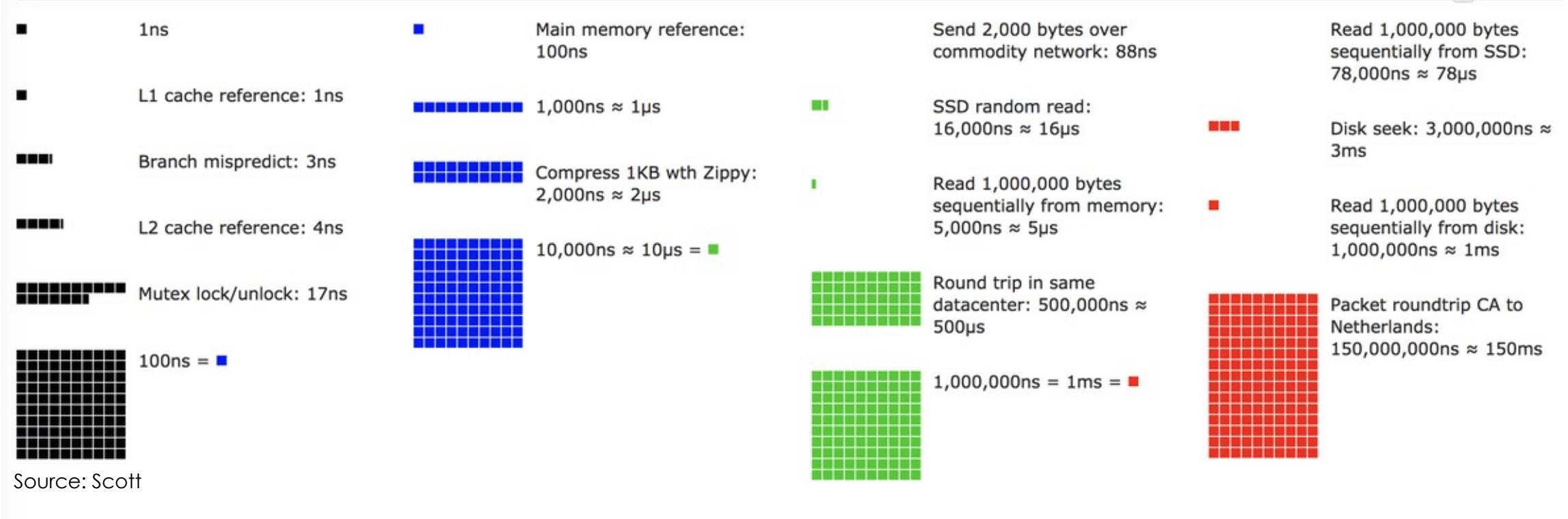


Source: Horowitz



Source: Bryant and O'Hallaron

Data Movement Overhead



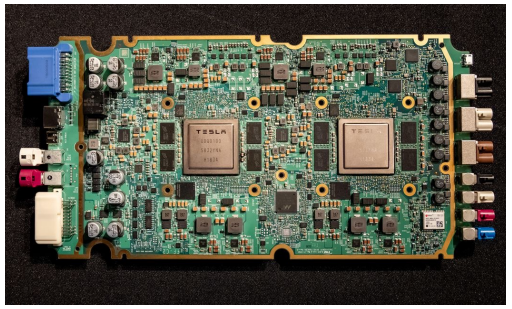
Source: Scott

Roadmap for Today

1. Introduction
2. Illustrative Examples
3. Efficiency
4. **Open Problems**

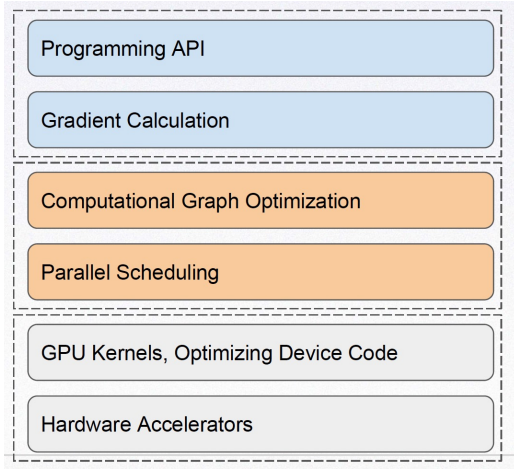
Open Problems

Hardware



Privacy

Tool and Software

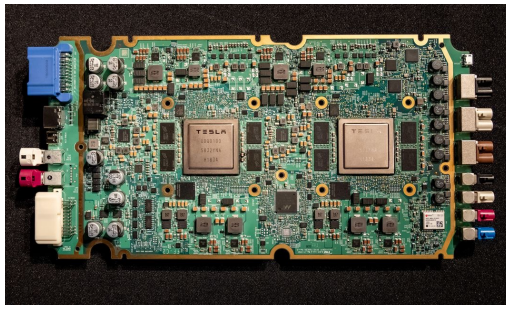


Scalability

Open Problems

Hardware

(Lecture 6)

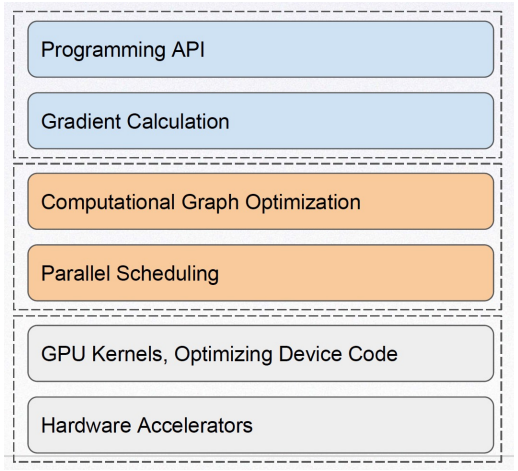


Privacy

(Lecture 8)

Tool and Software

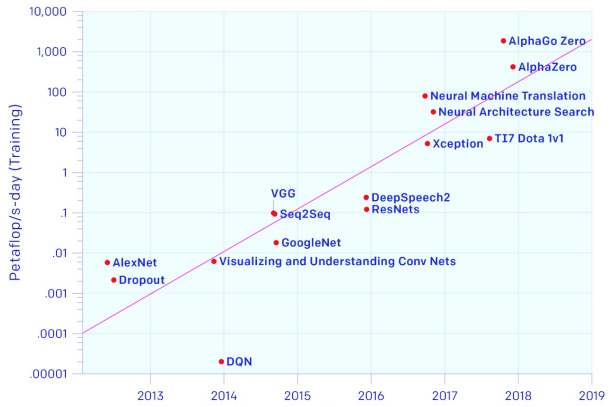
(Lecture 3)



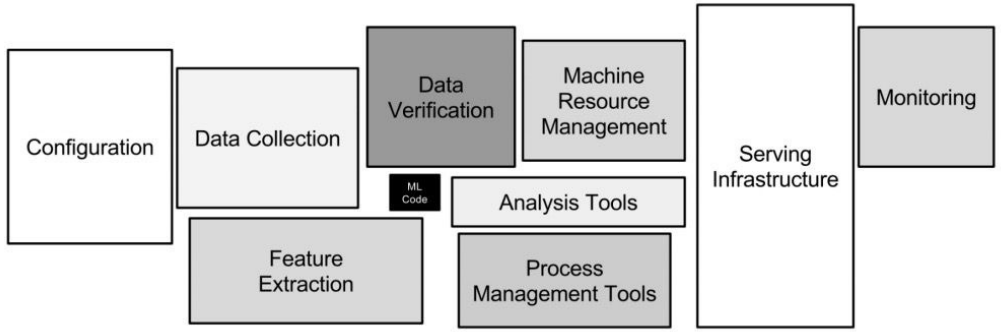
Scalability *(Lecture 7)*

Open Problems

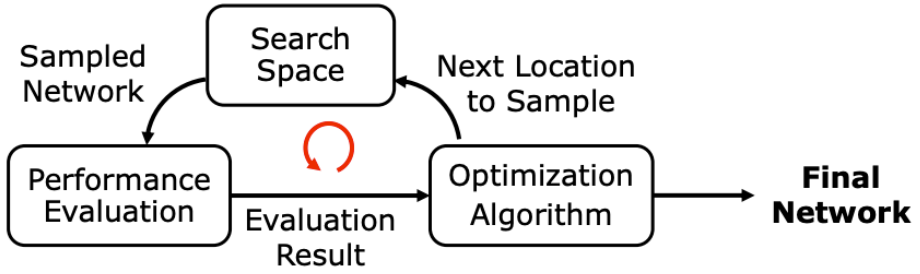
Metrics



System Resources



System Architecture

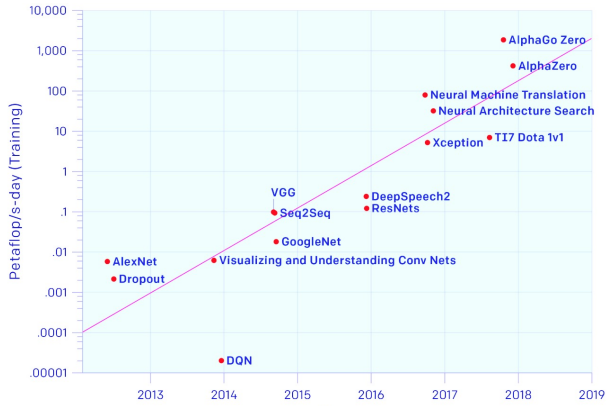


Automated Design

Open Problems

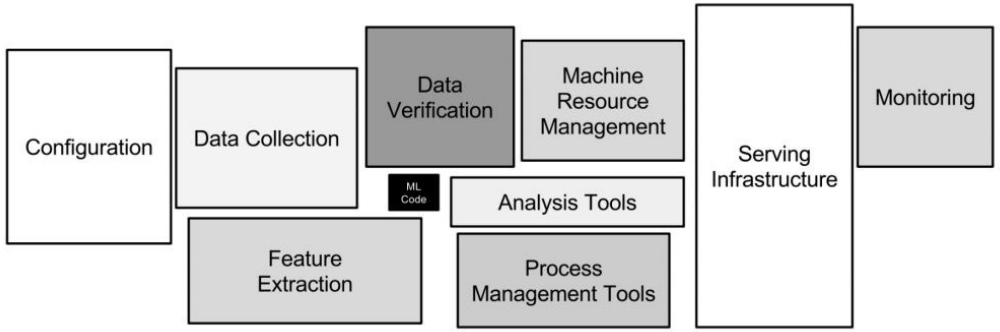
Metrics

(Lecture 1)

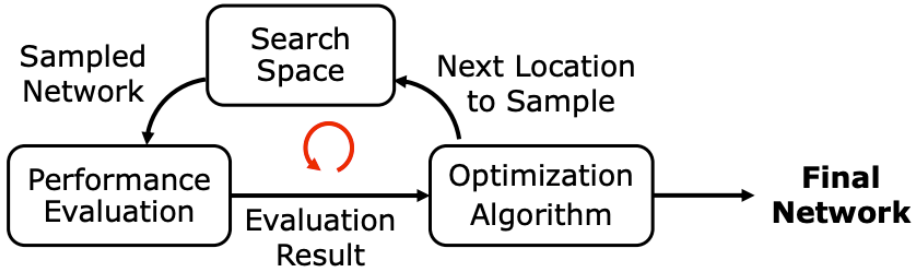


System Resources

(Lecture 2)



System Architecture *(Lecture 10)*



Automated Design

(Lecture 9)

Summary of the Day

1. Introduction
2. Illustrative Examples
3. Efficiency
4. Open Problems