15-442/15-642: Machine Learning Systems

Introduction

Spring 2025

Tianqi Chen
Carnegie Mellon University

Outline

Why study machine learning systems

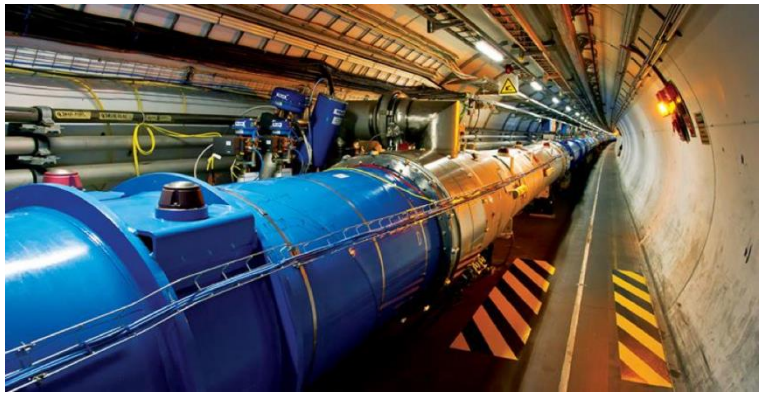
Course info and logistics

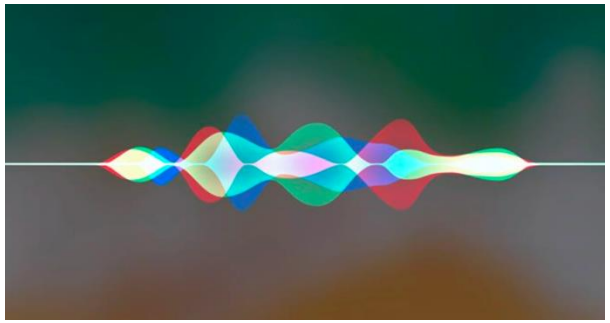
Outline

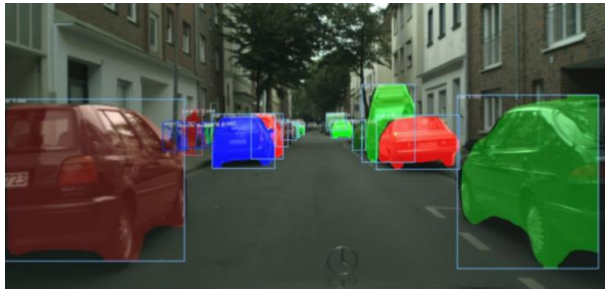
Why study machine learning systems

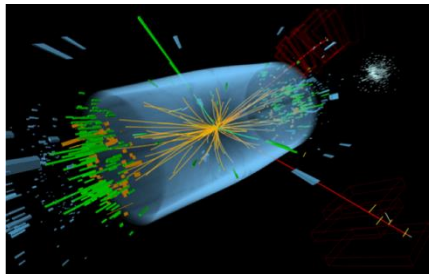
Course info and logistics

Successes of Machine Learning Today







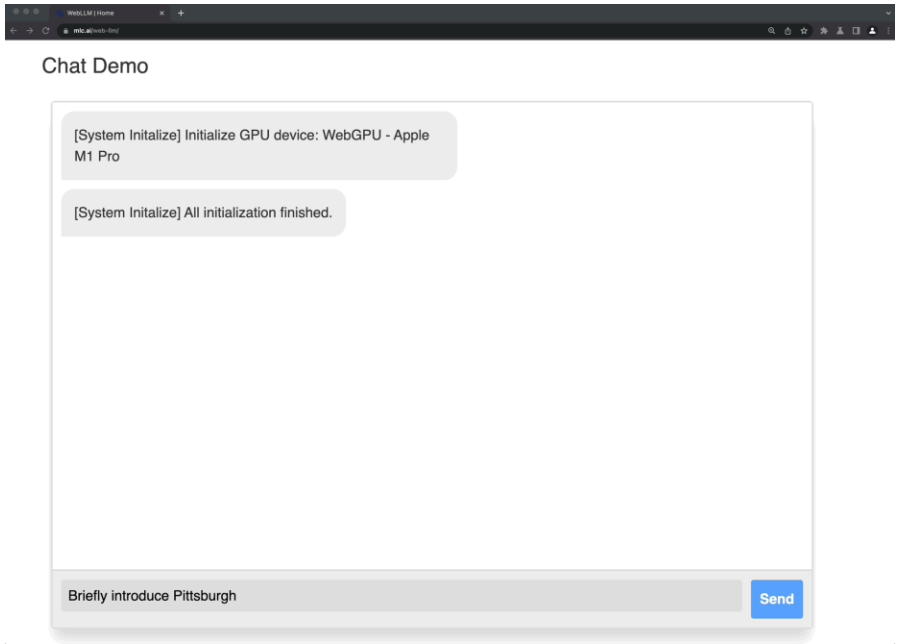


Big Bang of Generative AI

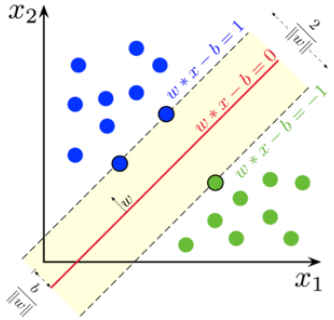
ImageGen models

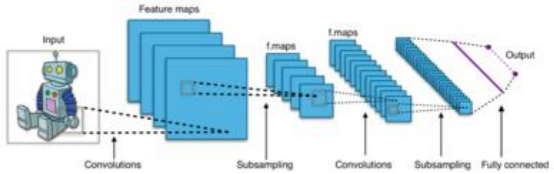
A photo of an astronaut riding a horse on mars

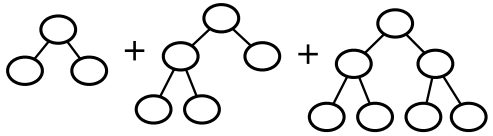
Large language model



1958 – 2000: Research







Perceptron Algorithm — 1958

Backprop — 1986

Support Vector Machine (SVM) — 1992

ConvNet — 1998

Gradient Boosting Machine (GBM) — 1999

Many algorithms we use today are **created before 2000**

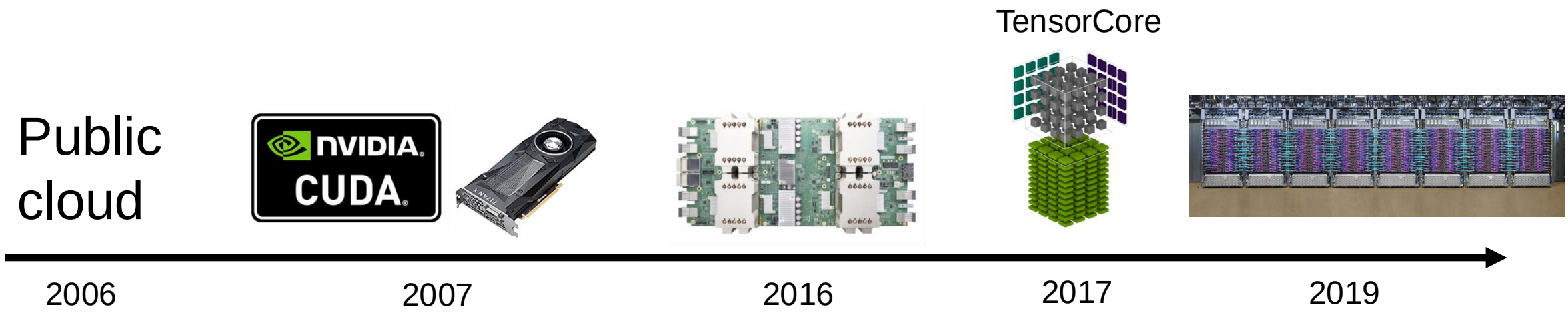
Based on personal view.
Source: Wikipedia

2000 – 2010: Arrival of Big Data

Data serves as fuel for machine learning models

Based on personal view.
Source: Wikipedia

2006 – Now: Compute and Scaling



Compute scaling

Based on personal view.
Source: Wikipedia, Nvidia, Google

Three Pillars of ML Applications

SVM ConvNet

Backprop GBM

ML Research

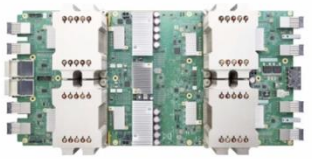
1958

IMAGENET

Data

2000

Public cloud



Compute

2007

Case Study: Ingredient of AlexNet

Year 2012

Methods

SGD
Dropout
ConvNet
Initialization

Data

1M labeled images

Compute

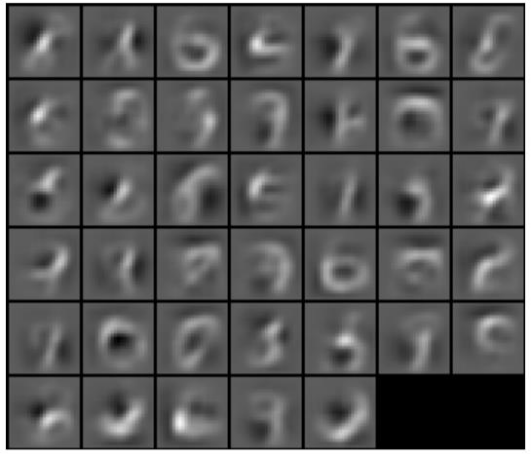
Two GTX 580

Six days

Krizhevsky et.al ImageNet Classification with Deep Convolutional Neural Networks.

Instructor's Story: First Deep Learning project

Year



Language	files	blank	comment	code
C	3	84	721	22755
C/C++ Header	43	1773	2616	12324
CUDA	21	1264	1042	7871
C++	17	268	343	1472
MATLAB	9	49	9	245
make	3	26	10	84
Python	2	12	0	42
SUM:	98	3476	4741	44793

One model variant

44k lines of code, including CUDA kernels for GTX 470

Six months of engineering effort

The project did not work out in the end.

Machine Learning Systems

Researcher

ResNet ….

Transformer

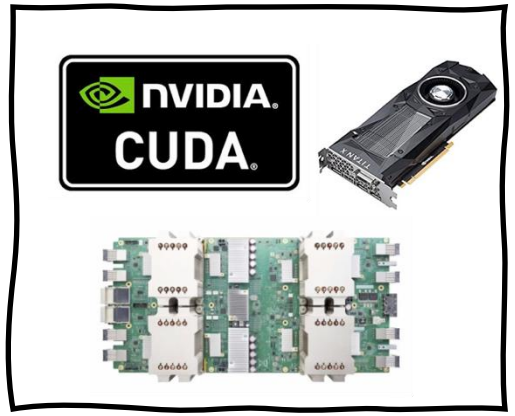
ML Research

44k lines of code

Six months

IMAGENET

Data



Compute

Machine Learning Systems

Researcher

ResNet ….

Transformer

ML Research

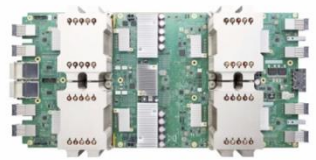
100 lines of python A few hours

System Abstractions

Systems (ML Frameworks)

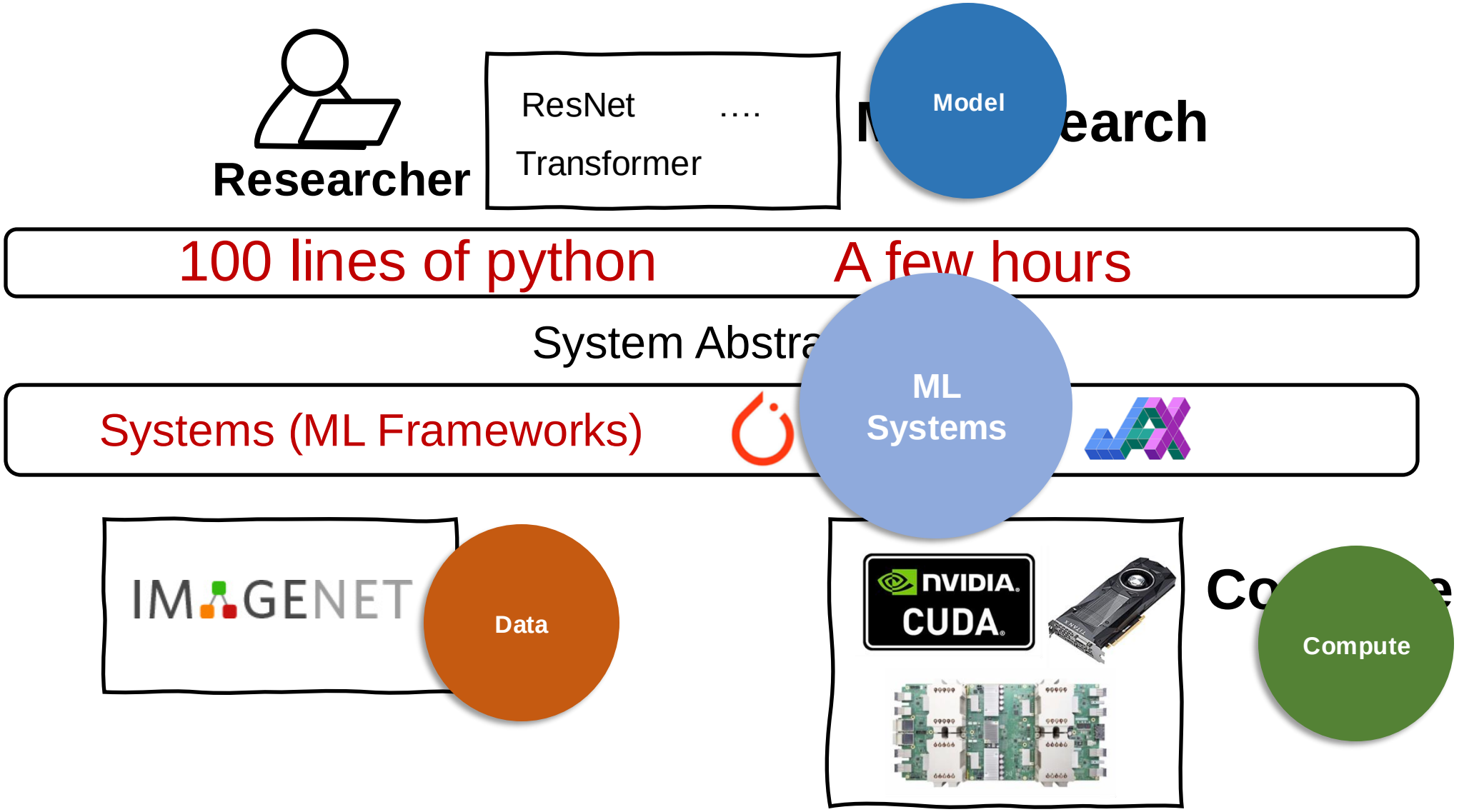
IMAGENET

Data

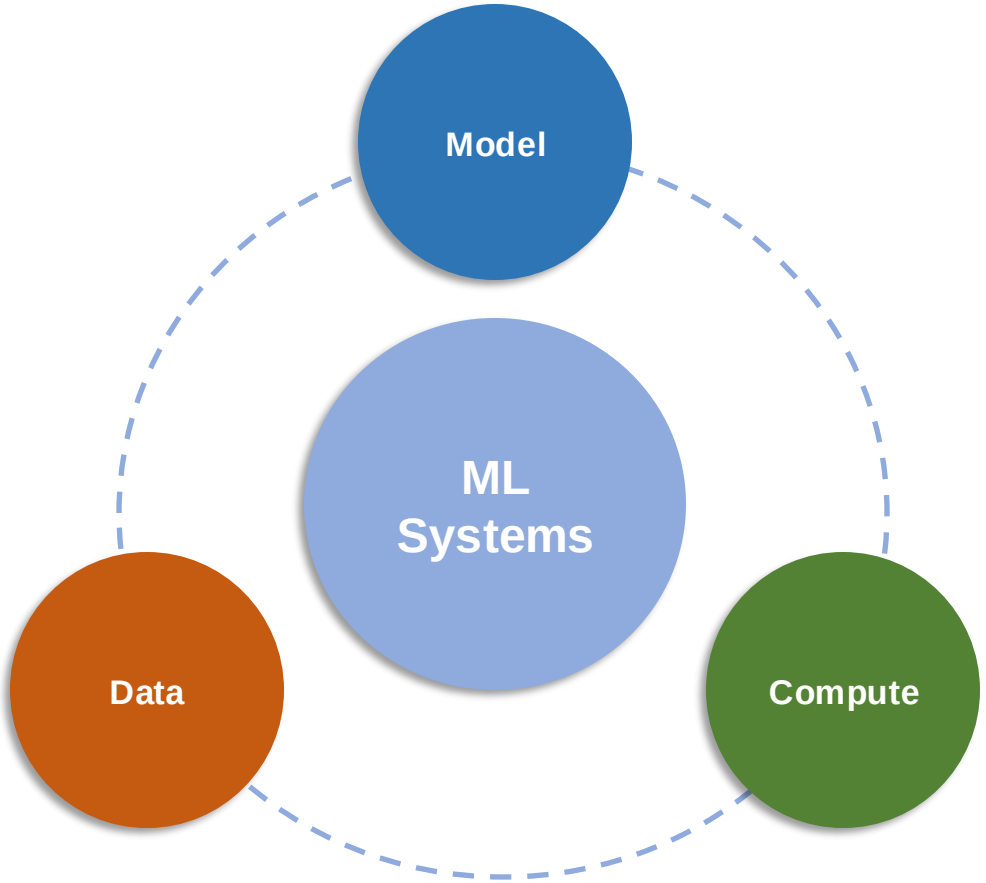


Compute

Machine Learning Systems



MLSys as a Research Field



Problems

A holistic approach (ML, Data, Systems, Hardware) to solve the problem of interest.

Question

Need to improve self-driving car's pedestrian detection to be X-percent accurate, at Y-ms latency budget

A Typical ML Approach

Need to improve self-driving car's pedestrian detection to be X-percent accurate, at Y-ms latency budget

Design a better model with smaller amount of compute via pruning, distillation

A Typical Systems Approach

Need to improve self-driving car's pedestrian detection to be X-percent accurate, at Y-ms latency budget

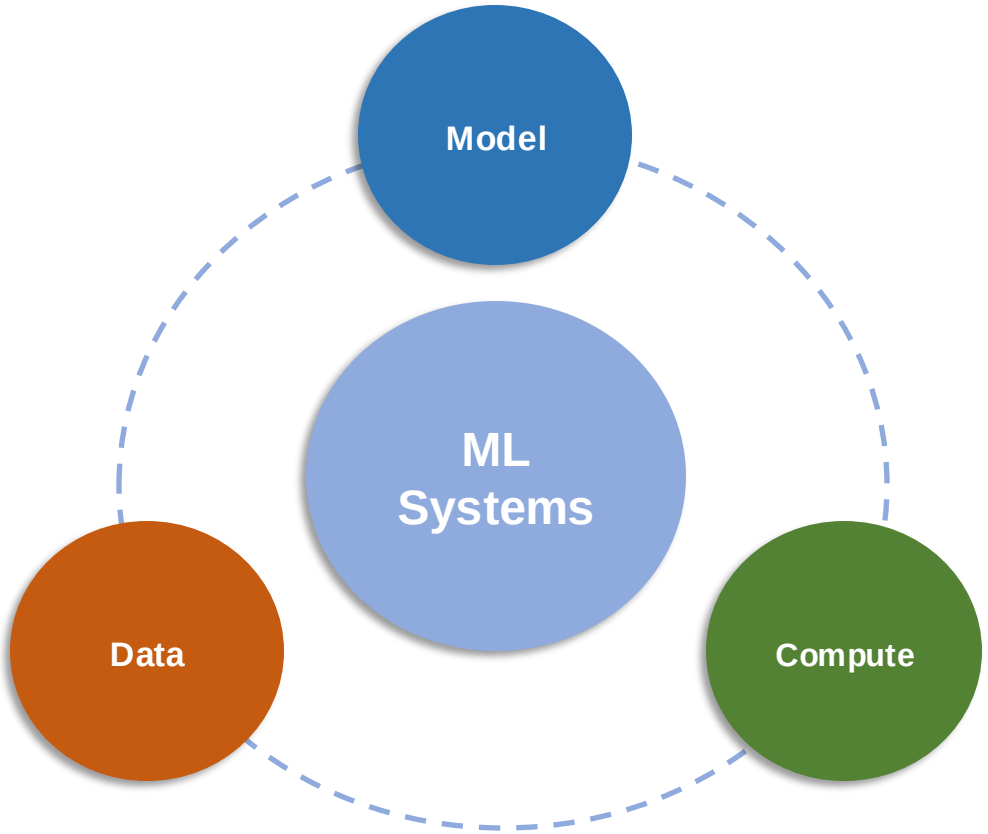
Build a better inference engine to reduce the latency and run more accurate models.

An Example MLSys Approach

Need to improve self-driving car's pedestrian detection to be X-percent accurate, at Y-ms latency budget

- Collect more data
- Incorporate specialized compute hardware
- Develop models that optimizes for the specific hardware
- Build end-to-end systems that makes use of the above points

MLSys as an Emerging Research Field



AI Systems Workshop at NeurIPS

MLSys tracks at Systems/DB conferences

Conference on Machine Learning and Systems (MLSys.org)

MLSys: The New Frontier of Machine Learning Systems

Why Study Machine Learning and Systems?

Reason #1 To push the frontier of modern AI applications, we need to have a holistic approach to the problem, understand and make use of existing systems more efficiently.

Reason #2 Prepare ourselves to build machine learning systems and work in the area of machine learning engineering.

Reason #3 Have fun building our own ML systems!

Outline

Why study machine learning systems

Course info and logistics

Course instructors + TAs

Tianqi Chen

Professor

catalyst

Carnegie Mellon University
School of Computer Science

Creator of Major Learning Systems

tvm

mxnet

dmlc XGBoost

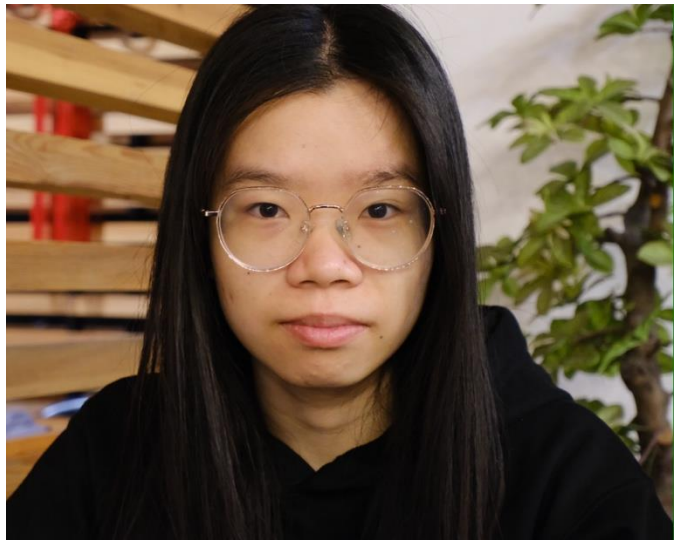
Cook and Foodie

Course instructors + TAs

Bohan Hou

Ruihang Lai

Zhihao Zhang



Mengdi Wu

Big bold disclaimer

This is the second time offering of this course. The material and outline will likely adjust throughout the semester. There will almost certainly be some bugs in the content or assignments. By signing up for this course, you're implicitly agreeing to bear the brunt of all these issues.

Learning Objects

By the end of this course, you will …

… understand the general components of modern machine learning frameworks, including concepts like automatic differentiation, hardware accelerations, parallelization and memory-saving techniques

… understand systems techniques for emerging generative AI applications

… implement your own machine learning systems project

Course Resources

- **Website**: https://mlsyscourse.org
  - Contains links to all resources

- **Piazza**: discussions and announcements
  - Links on the website, register today

- **Gradscope**: submit assignments, project proposals, final papers

Prerequisites

In order to take this course, you need to be proficient with

- Systems programming (e.g., 15-213)
- Linear algebra (e.g., 21-240 or 21-241)
- basic mathematical background (21-127 or 15-151)
- Python and C++ development
- Prior experience with machine learning or AI

This is a system-focused course, so make sure you are comfortable in system programming

Components of the Course

This course will consist of four main elements

1. Class lectures
2. Programming-based (individual) assignments
3. (Group) final project
4. Interaction/discussion in piazza

Grading breakdown: 45% assignments, 45% project, 10% class participation

Final Project

In addition to homeworks, there will also be a final project, done in groups of 2-3 students. The final project contains the following components

- Proposal: 5%
- Poster presentation: 20%
- Project report: 20%

Class Forum

Join the Piazza: Link in the course Website

At end of class, register the piazza ***using your Andrew email***

In order to receive a full participation score, you will need to be involved in at leave ***five*** online discussions during the course

Collaboration policy

All submitted content (code and prose for homeworks and final project) should be your own content, written yourself (or written by the group members, for projects)

However, you *may* (in fact are encouraged to) discuss the homework with others in the class and on the piazza

- Use best judgement, discuss as you see fit, but don't simply share answers

Generative AI “ChatGPT” Policy

You may use code from generative AI tools (e.g., ChatGPT or Co-pilot), no need to cite or specify it was from these tools.

You are ultimately responsible for anything the tools generate, including any flaws this code may contain. Content generated by LLMs that could be construed as plagiarism or scientific misconduct (e.g., fabrication of facts).

For our own information, we might conduct an (optional) poll on the extent to which students find such tools valuable for this course

Student Well-being

Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress. All of us benefit from support during times of struggle. You are not alone. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is often helpful.

In the Remaining Time…

- Sign up for the piazza ***using your @andrew.cmu.edu email***
  https://piazza.com/home/spring2025/15442

- Post a note in the "Say Hello" post
  1. Your name
  2. Your background and what you're interested in learning in this course
  3. Anything cool you've done at al relevant to machine learning systems

- Read other people's notes and respond if you feel like it