CS6216
Advanced Topics in Machine Learning (Systems)

Yao LU

2024 Semester 1

National University of Singapore
School of Computing

Course instructor

Yao LU, assistant professor in CS

- PhD in CS, University of Washington, 2018
- Researcher, Microsoft Research Redmond, 2018-2023
- Experiences in AI, databases, cloud systems, ML systems

Outline

- Why machine learning systems
- Some recent topics in ML systems research & production
- Logistics

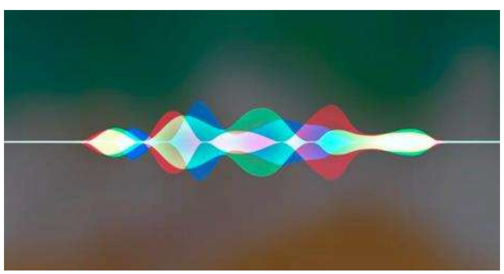
Successes of AI / ML Today

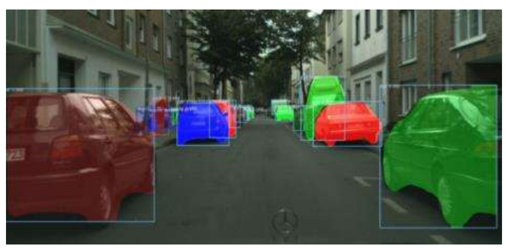
Personal Assistants

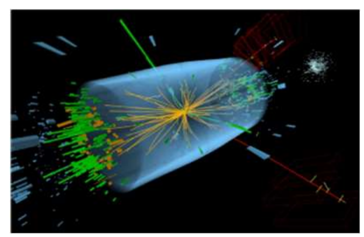
Robotics / Auto Driving

AI for Science

Search / Recom. / Ads





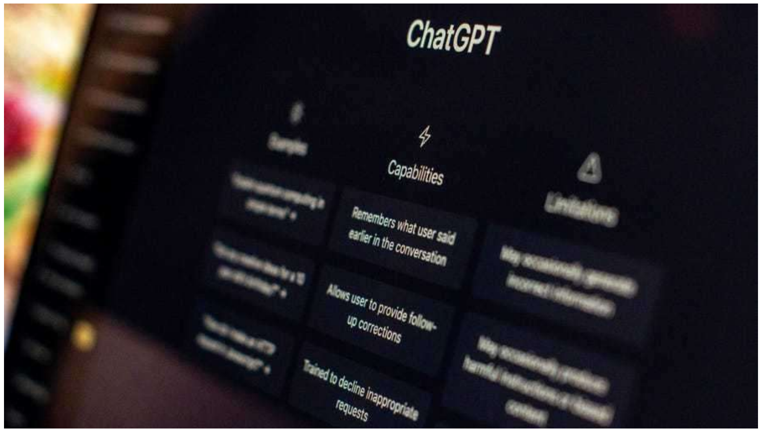


Slides credit to Tianqi Chen and Zhihao Jia

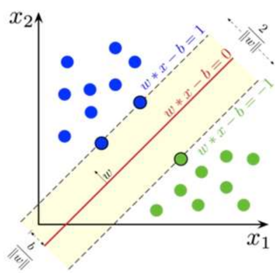
Big Bang of Generative AI

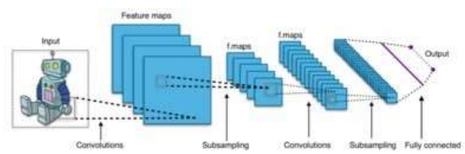
Colorful applications

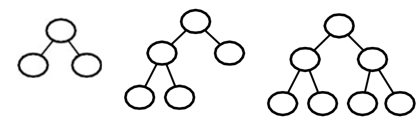
Large language models and ChatGPT



1958 – 2000: ML Research







Perceptron Algorithm	Backprop	Support Vector Machine (SVM)	ConvNet	Gradient Boosting Machine (GBM)
1958	1986	1992	1998	1999

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

Based on personal view.
Source: Wikipedia

2000 – 2010: Arrival of Big Data

Wikipedia — The Free Encyclopedia | flickr | MTurk | 

| kaggle, IMAGENET

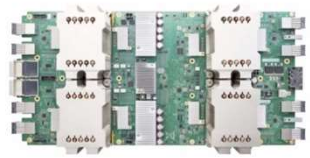
2001 | 2004 | 2005 | 2009 | 2010

Data serves as fuel for machine learning models

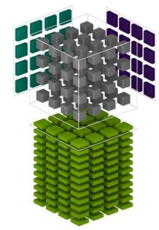
Based on personal view.
Source: Wikipedia

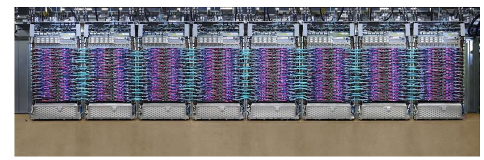
2006 – Now: Compute and Scaling

Public cloud



TensorCore





2006 2007 2016 2017 2019

Compute scaling

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

Three Pillars of ML Applications

SVM ConvNet

Backprop GBM

ML Research

1958

WIKIPEDIA
The Free Encyclopedia

IMAGENET

Data

2000

Public cloud

NVIDIA
CUDA

Compute

2007

AlexNet

Year 2012

ML Research

SGD
Dropout
ConvNet
Pooling

Data

IMAGENET

1M labeled images

Compute

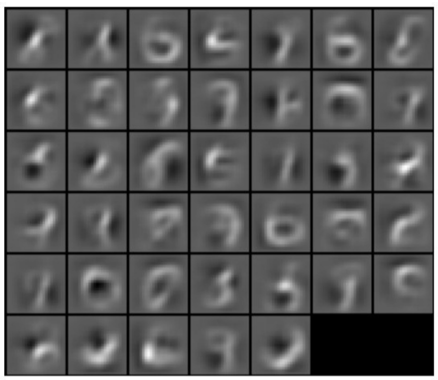
Two Nvidia GTX 580 GPUs

Six days

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

Tianqi Chen's First Deep Learning project

Year 2010



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

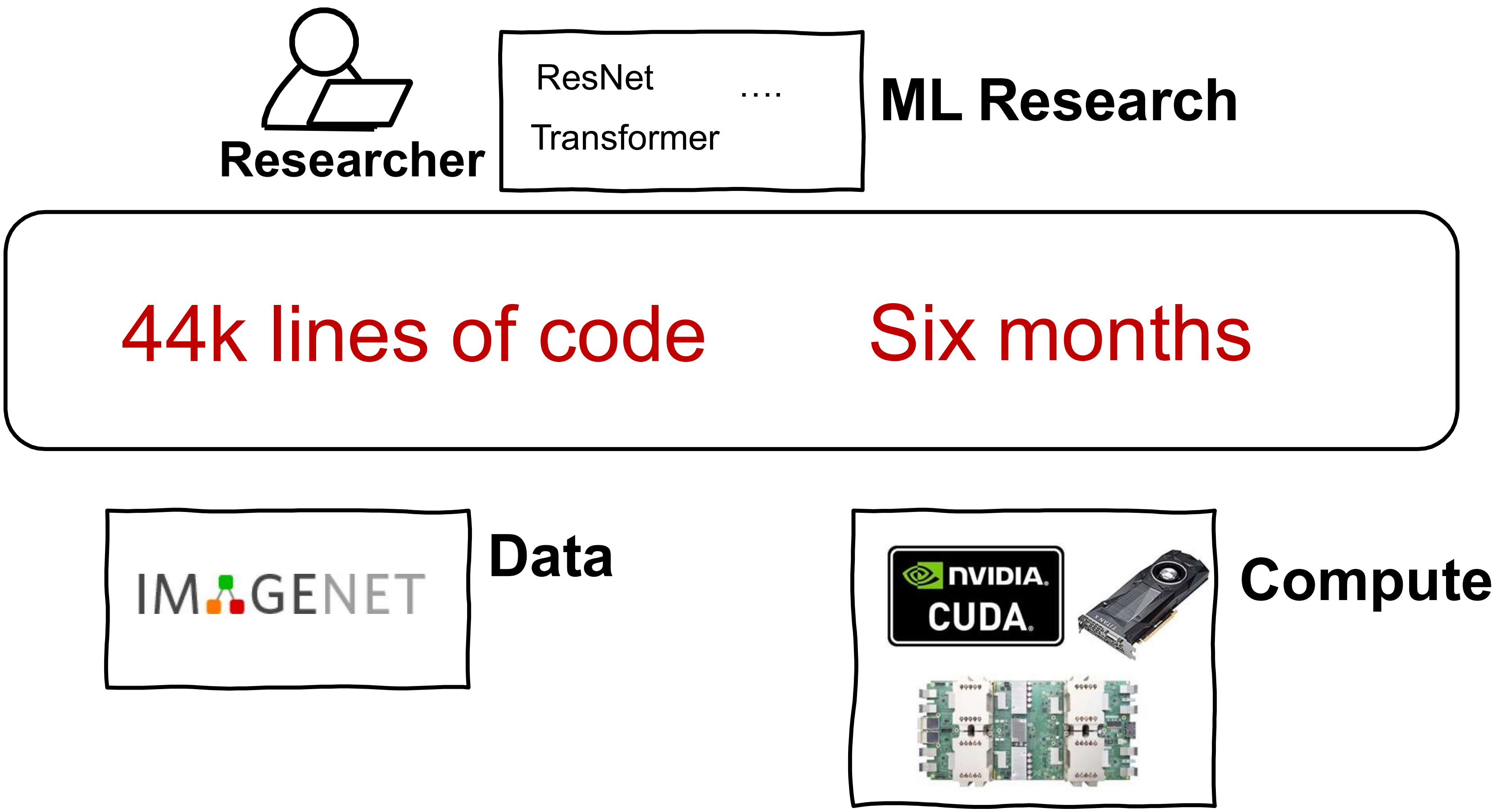
One model variant

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

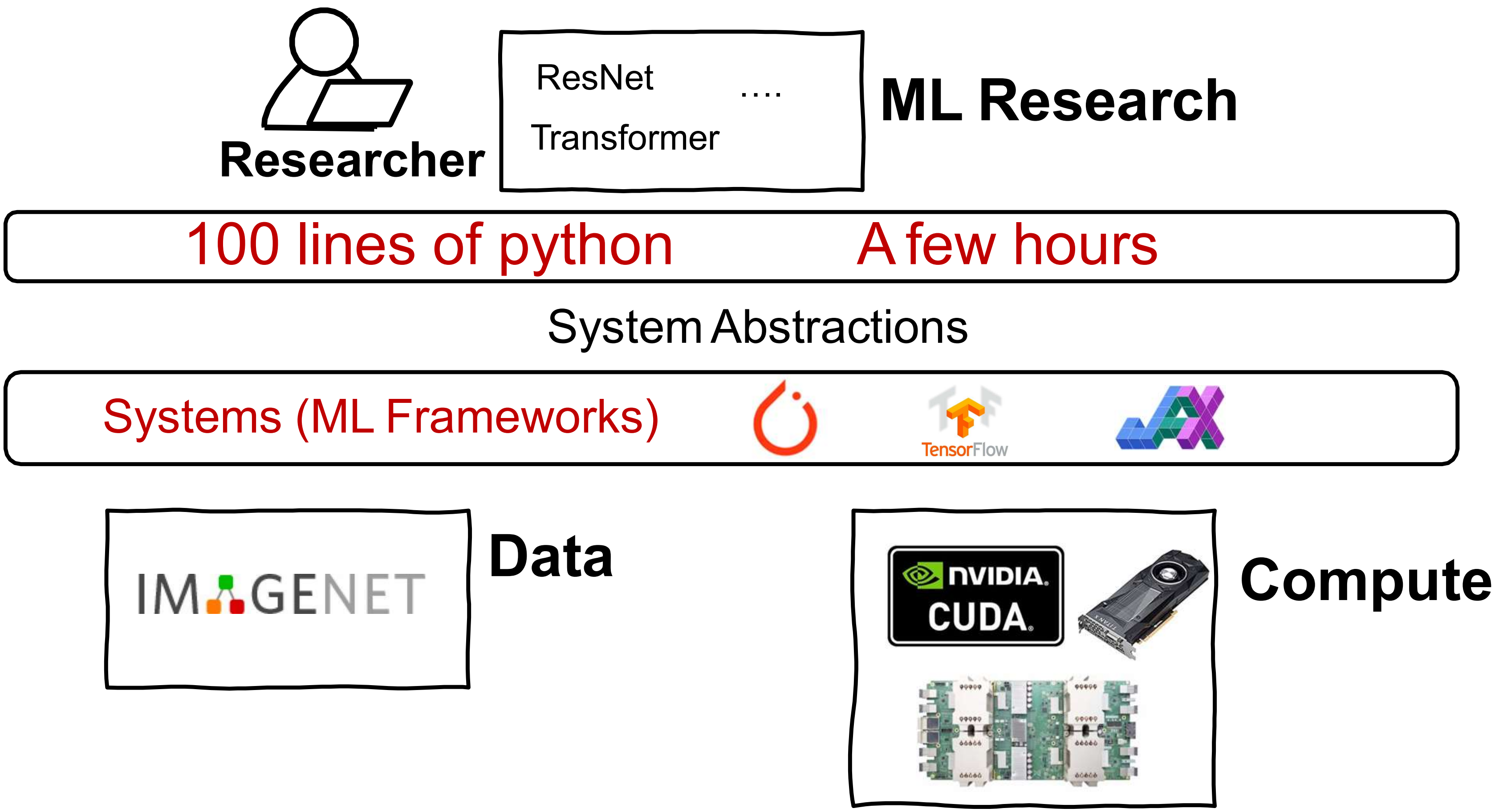
Six months of engineering effort

The project did not work out in the end.

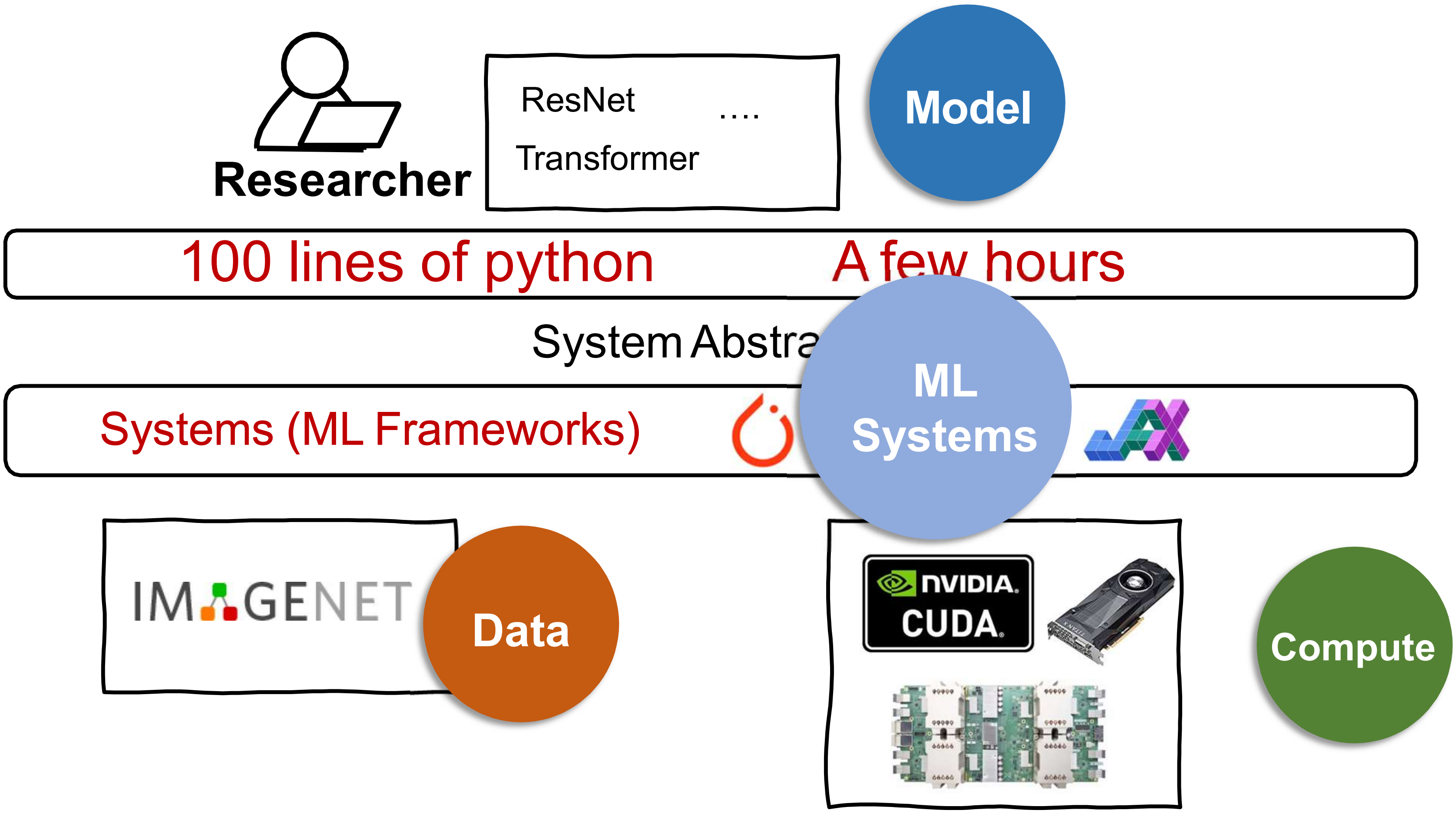
Machine Learning Systems



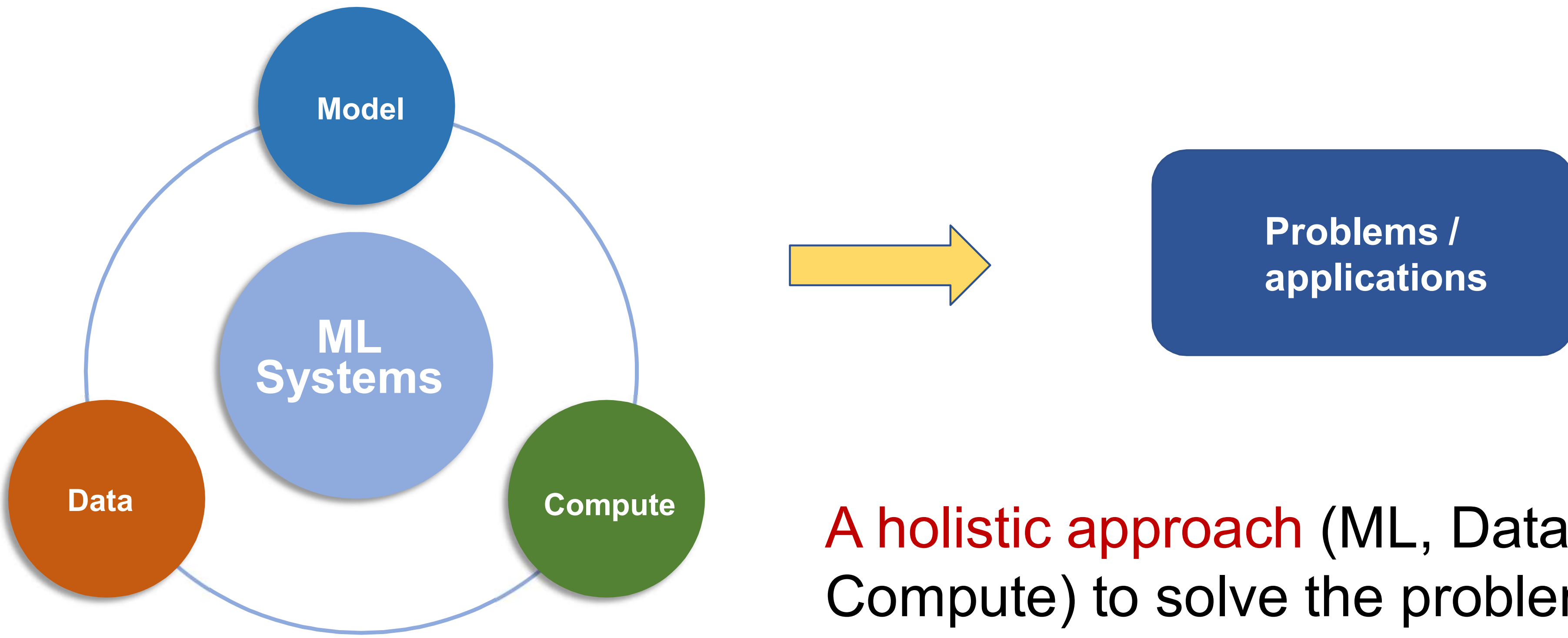
Machine Learning Systems



Machine Learning Systems



MLSys as a Research Field



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

A practical problem

To improve pedestrian detection to be X-percent accurate, at Y-ms latency budget with Z-watt hardware

A Typical ML Approach

To improve pedestrian detection to be X-percent accurate, at Y-ms latency budget with Z-watt hardware

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

A Typical Systems Approach

To improve pedestrian detection to be X-percent accurate, at Y-ms latency budget with Z-watt hardware

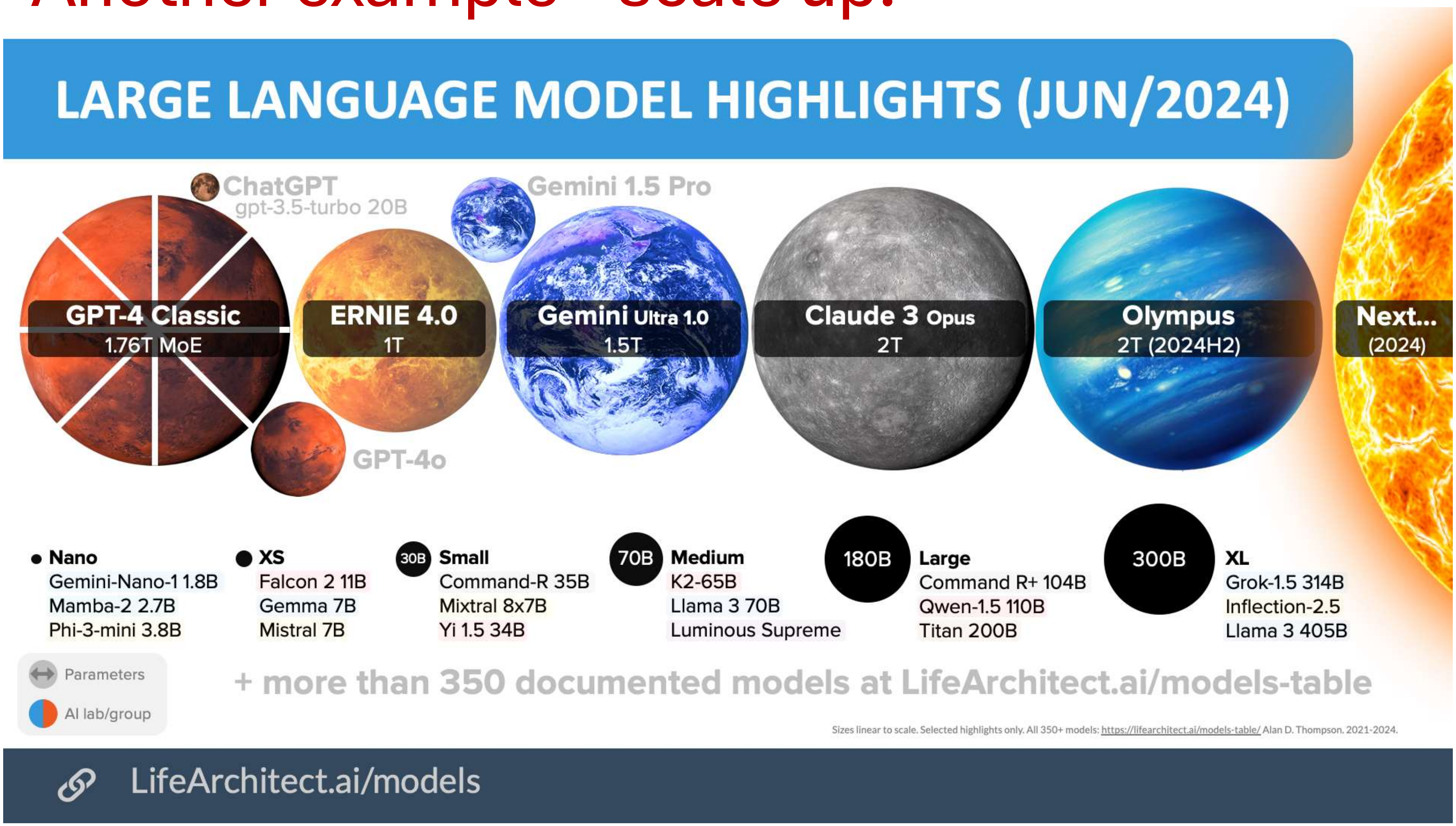
Build a better inference engine to reduce the latency and run more accurate models.

An MLSys Approach

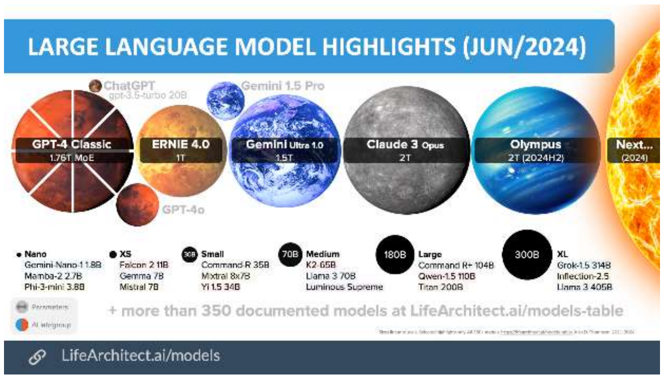
To improve pedestrian detection to be X-percent accurate, at Y-ms latency budget with Z-watt hardware

- **Data**: acquire more sensor data and preprocess them
- **Model**: Develop models that fit the accuracy & latency budget
- **Compute**: Build end-to-end systems for the specific hardware
  - Edge devices & accelerators, sensor data pipelines, decision making

Another example - scale up!



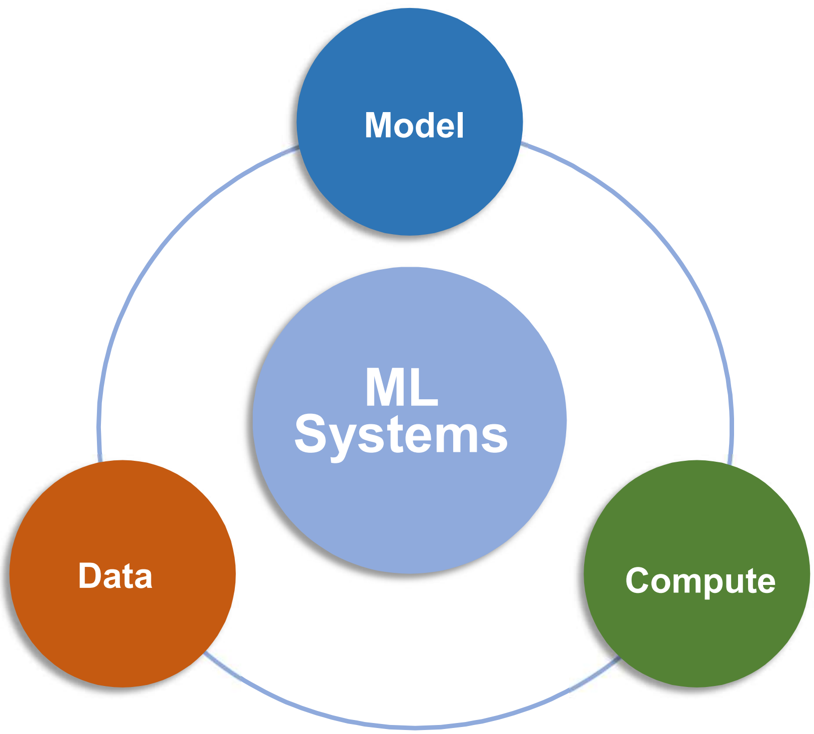
An MLSys Approach



Train an LLM with 1T parameters and maximize model quality

- **Data**: acquire more data and preprocess them
- **Model**: Design models that optimize for the specific model size
- **Systems**: Build end-to-end systems that enable training on a distributed cluster
  - Networking, storage, scheduling, failure recovery etc.

MLSys as an Emerging Research Field



AI Systems Workshop at NeurIPS

MLSys tracks at Systems/DB/Networking conferences

Conference on Machine Learning and Systems (MLSys.org)

MLSys as a Startup Arena

anyscale

LangChain

together.ai

mosaicML

Weaviate

Hugging Face

Lambda

LlamaIndex

Pinecone

Why Study Machine Learning and Systems?

Reason #1 AI is the future. Systems for AI is the foundation.

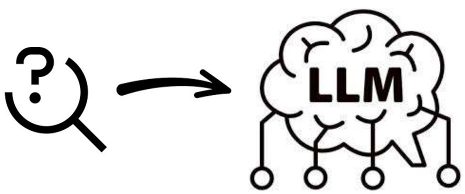
Reason #2 A full-stack and holistic approach to push the frontier of AI research and production.

Reason #3 Industry: high demand, low supply → high $$$

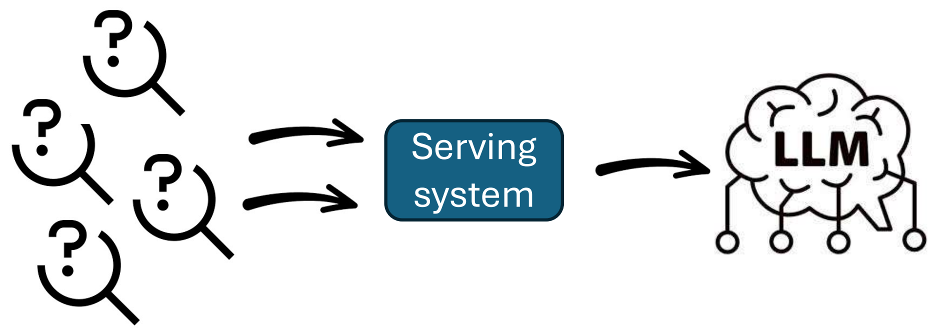
Outline

- Why machine learning systems
- Some recent topics in ML systems research & production
- Logistics

Serving systems



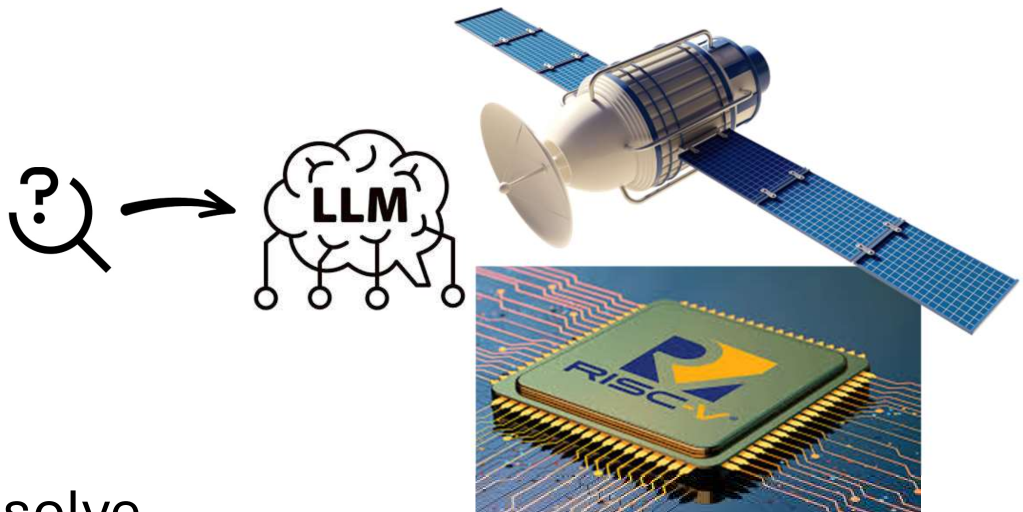
Model inference



Model serving

- Problems to solve
  - Batching, queueing, quota control
  - Improving latency and throughput

New hardware, edge / IoT devices



- Problems to solve
  - Improve accuracy / performance & reduce costs
  - Software-hardware co-design

Data infrastructures

Data management for AI :

acquisition, cleaning, structurization,

transformation, annotation, visualization

AI-aided data management & curation

Data systems for AI :

embedding, storage, indexing,

retrieval, query processing

Data quality is the key to high quality AI

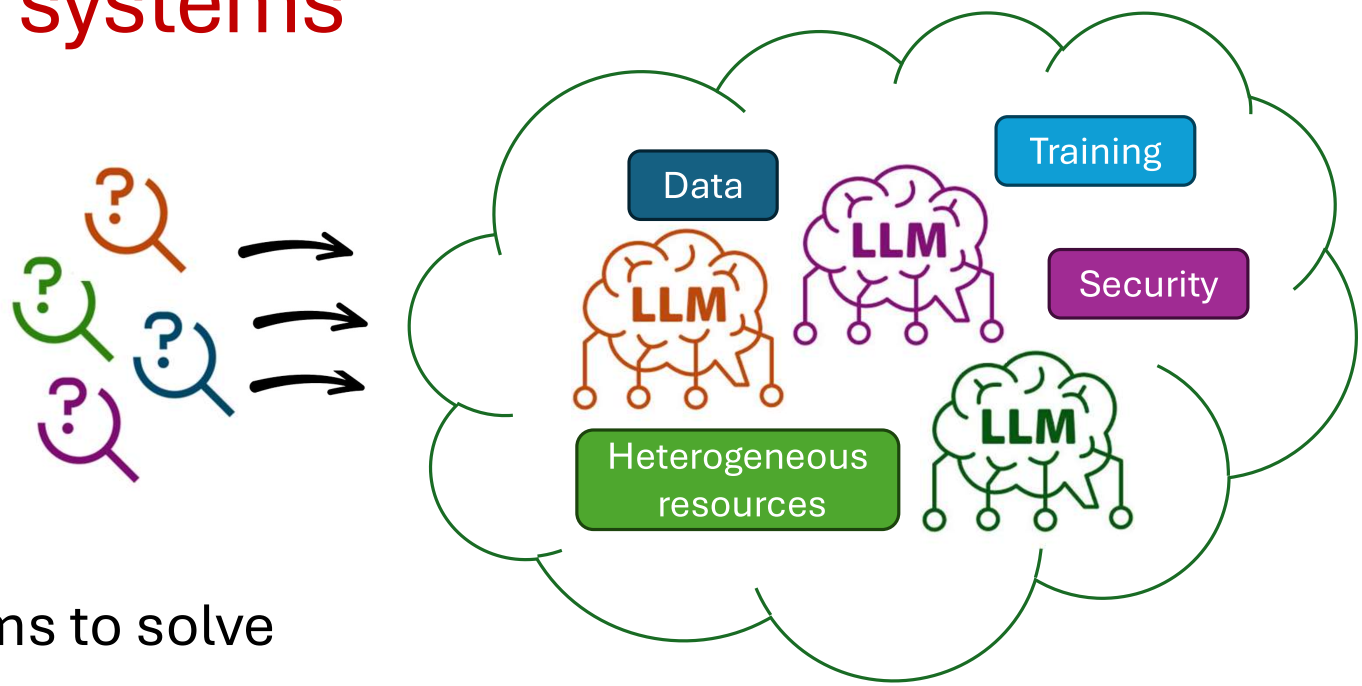
TABLE 41

YUBA RIVER DIVERSIONS–1938

Water User	*Mile and Bank	Number and Size of Pump	Monthly Diversions in Acre-feet: Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Total Diversion: March to October Acre-feet	Acreage Irrigated: General	Rice
— SEVENTH STREET BRIDGE - MILE 0.9 —													
California Lands, Inc.	0.9 L	1-5"			NO	DIVERSION							
Davis Brothers	1.6 L	1-12"				70	90	97	113		(1) 370	120	
Charles Shinkle (Harrington)	1.8 R	1-5"				6	14	14			34	12	
G. E. Edwards	1.9 L	1-6"			PLANT	REMOVED							
Davis Brothers (2)	3.0 L	1-10"				53	98	97	71		(3) 319	135	
Yuba River Farms (Higgins) (4)	3.0 R	1-6"				7	36	27	6		76	30	
G. F. Sherbourne (5)	4.1 L	1-8"				18	74	63	54		209	(6)46	
James Traynor (Covert)	4.2 R	1-3"				3	6	9			18	5	
S. J. Monaco	4.3 R	1-4"				8					8	5	
C. R. Perkins (Cunningham) (7)	(8)4.70L	1-6"					10	9			19	15	
Earl Fruit Company and Dinsmore	4.75L	1-6"				87	33				120	56	
Dantoni Orchards (Earl Fruit Co.)	5.3 L	1-8"				98					98	75	
Marysville River Farms Company	5.9 L	1-10"			NO	DIVERSION							
Marysville River Farms Company (Nagler and Pearson)	6.35L	1-10"			NO	DIVERSION							
Marysville River Farms Co. (Plantz)	6.35L				PLANT	REMOVED							
Hallwood Irrigation Company (9)	(9)11.0 R	Gravity		360	3284	6378	6419	6207	5844	1020	29512	4581	470
Cordua Irrigation District (9)	11.0 R (9)	Gravity			1523	2643	3202	2910	2196		12474	692 (10)	1135 (10)
Yuba Consolidated Gold Field Co.	14.5 L	Gravity			NO	DIVERSION							
Totals			0	360	4807	9371	9982	9433	8284	1020	43257	5772	1605

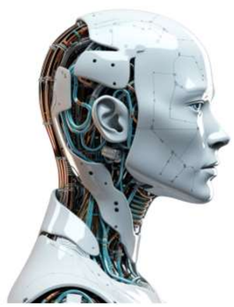
* Approximate mileage along river above highway crossing at Marysville.
(1) The diversion at this point is combined with that at Mile 3.0 Left.
(2) Formerly Davis and Cox.
(3) See plant at Mile 1.6 Left.
(4) Formerly Ward Hughins.
(5) Formerly E. O. Rubke.
(6) Includes 20 acres on Rubke land.
(7) Formerly J. S. Johnson.
(8) Former mileage of 4.8 in error.
(9) Hallwood Irrigation Company and Cordua Irrigation District have a common point of diversion and common canal for about one-half mile.
(10) Includes 285 acres rice and 35 acres general outside of District.

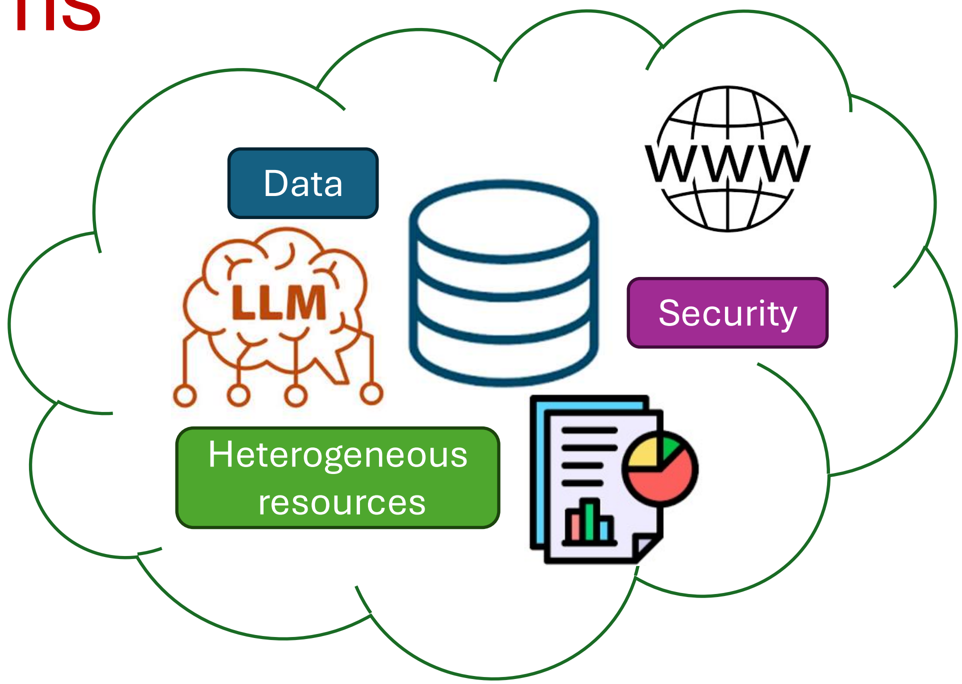
Cloud systems



- Problems to solve
  - Various data/ML/application pipelines
  - Scheduling the heterogeneous resources
  - Improve efficiency for individual users & operator

AI / ML for systems





- Problems to solve
  - Use AI to monitor & operate the cloud
  - Use AI/ML to improve individual cloud components
  - Reduce operating costs

Outline

- Why machine learning systems
- Some recent topics in ML systems research & production
- Logistics

Pre-requisitions

- UG machine learning or equivalent
- UG operating systems or equivalent
- Strong Python programming
- (Optional) C/C++/Rust programming
- This is a system-focused course, not intended for only LLM algorithms / modeling

Course TAs

Shenggan Cheng

PhD student @ HPC-AI Lab

Xuanlei Zhao

PhD student @ HPC-AI Lab

Course schedule (subject to change)

Week	Date	Lecture	Lecturer, if not LU Yao	HW schedule	HW Topic
1	08-14	Intro		HW1 out	ML and systems basics
2	08-21	ML sys foundations			
3	08-28	AI framework and autograd		HW1 due (more time)	
4	09-04	Transformers, attention and optimizations		HW2 out	AI framework + autograd
5	09-11	Hardware acceleration			
6	09-18	Training technologies		HW2 due	
		Recess			
7	10-02	Fine tuning technologies		HW3 out	LLM inference
8	10-09	Serving LLMs 1			
9	10-16	ML for systems	Guest lecturer	HW3 due	
10	10-23	Serving LLMs2		HW4 out	LLM serving
11	10-30	ML compilers (TBD)	Guest lecturer		
12	11-06	Cloud systems for ML		HW4 due	
13	11-13	Poster session			

Assignments and grading

- **Paper reading and discussion**
  - Mandatory, each weak 20%

- **Coding/Written assignments & course projects**
  - HW1 mandatory 20%
  - HW2-4 can be substituted partly or entirely by course projects 60% combined
    e.g.: all HW2-4 and no project , all project, no HW2-4, HW1 + project

- **Course projects** (normalized to 100%)
  - Group of 2-3 people
  - The fewer HW1-3 you take / the more people, the higher expectation
  - Choice & proposal by Week 3. (10%)
  - Mid-term report by the end of Recess week. (20%)
  - Final report by the end of Week 13. (40%)
  - Poster presentations in Week 13. (30%)
  - Topics: ML **systems** related. Pure ML/AI/CV/NLP projects are not acceptable.

- **Resources**
  - HW0: no GPU is needed. HW1-3 GPU programming as bonus
  - GPU clusters at SOC

Communications

- Office hour:

  Mondays 10AM-11AM, COM2-2-60

- Instructor email: luyao@comp.nus.edu.sg
- TA email: shenggan@comp.nus.edu.sg

  xuanlei@comp.nus.edu.sg

- Project discussion by appointment

- Canvas
  - Paper reading & discussions
  - Notifications
  - 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

Disclaimers

- This is a first time offering of this course. There are even not many similar offerings around the world.
- Industry & open-source world evolving ultra fast.
- The material and outline will likely adjust throughout the semester.
- There will be bugs in the content or assignments.