System for Artificial Intelligence

Introduction

Siyuan Feng
Shanghai Innovation Institute

OUTLINE

01 Why AI System

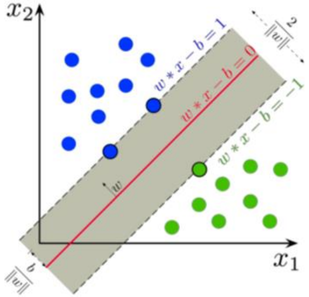
02 How to Learn

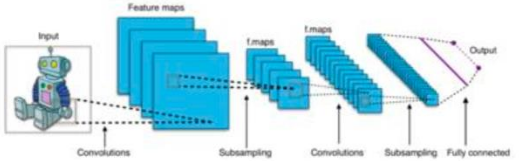
03 Course Logistics

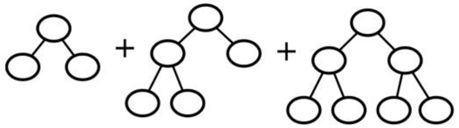
01

Why AI System

1958 – 2000: Research







Perceptron Algorithm

1958

Backprop

1986

Support Vector Machine (SVM)

1992

ConvNet

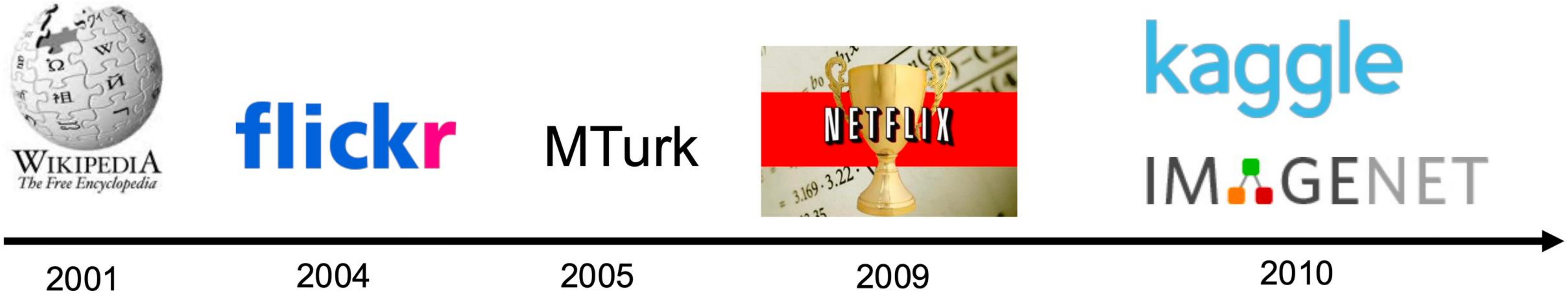
1998

Gradient Boosting Machine (GBM)

1999

Slides credits to Tianqi Chen

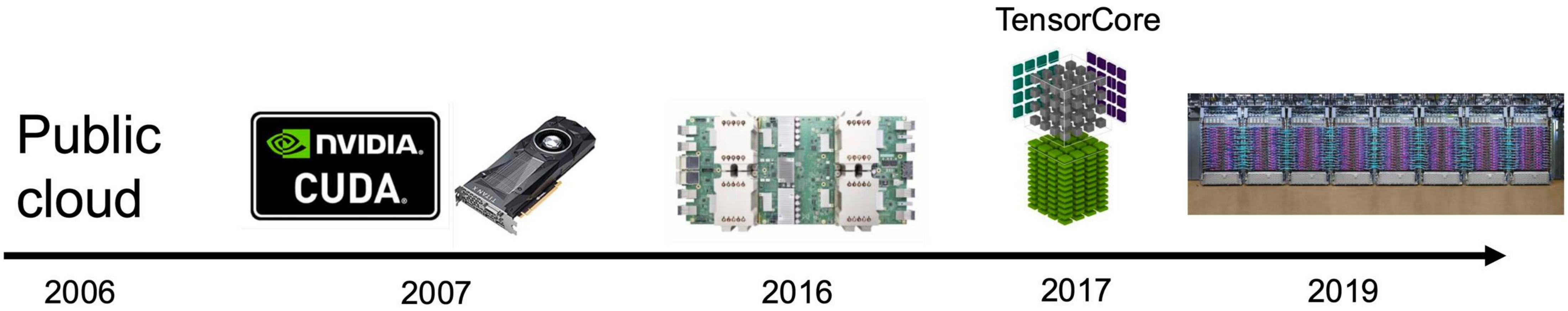
2000 – 2010: Arrival of Big Data



Data serves as fuel for machine learning models

Slides credits to Tianqi Chen

2006 – Now: Compute and Scaling



Public cloud

TensorCore

2006

2007

2016

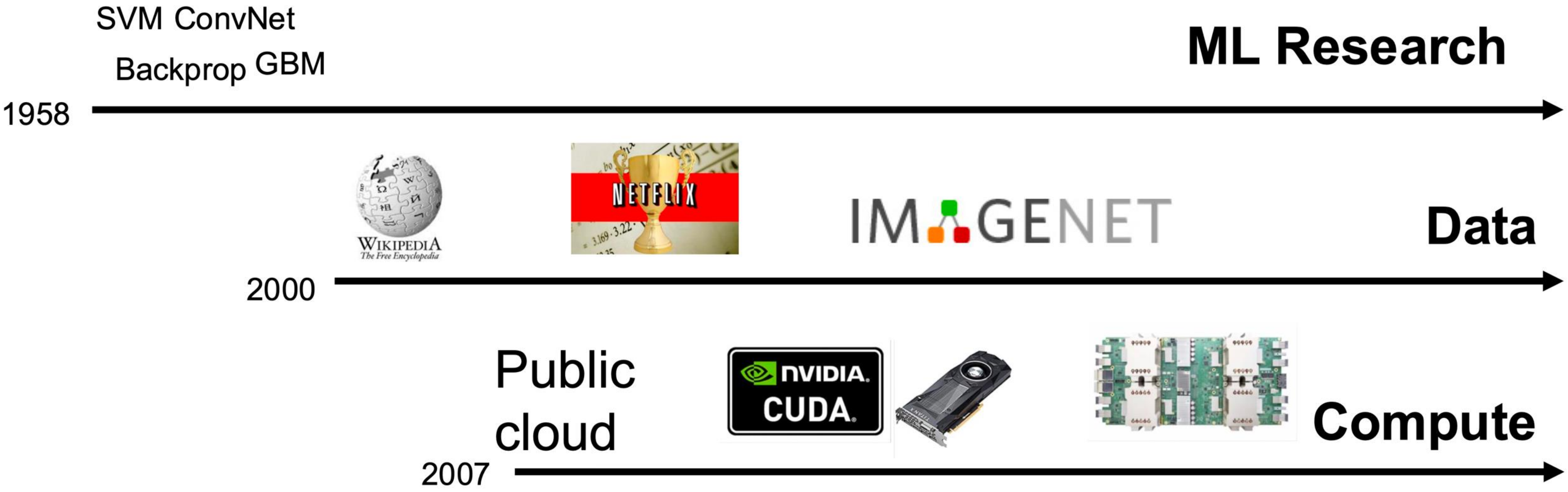
2017

2019

Compute Scaling

Slides credits to Tianqi Chen

Three Pillars of ML Applications



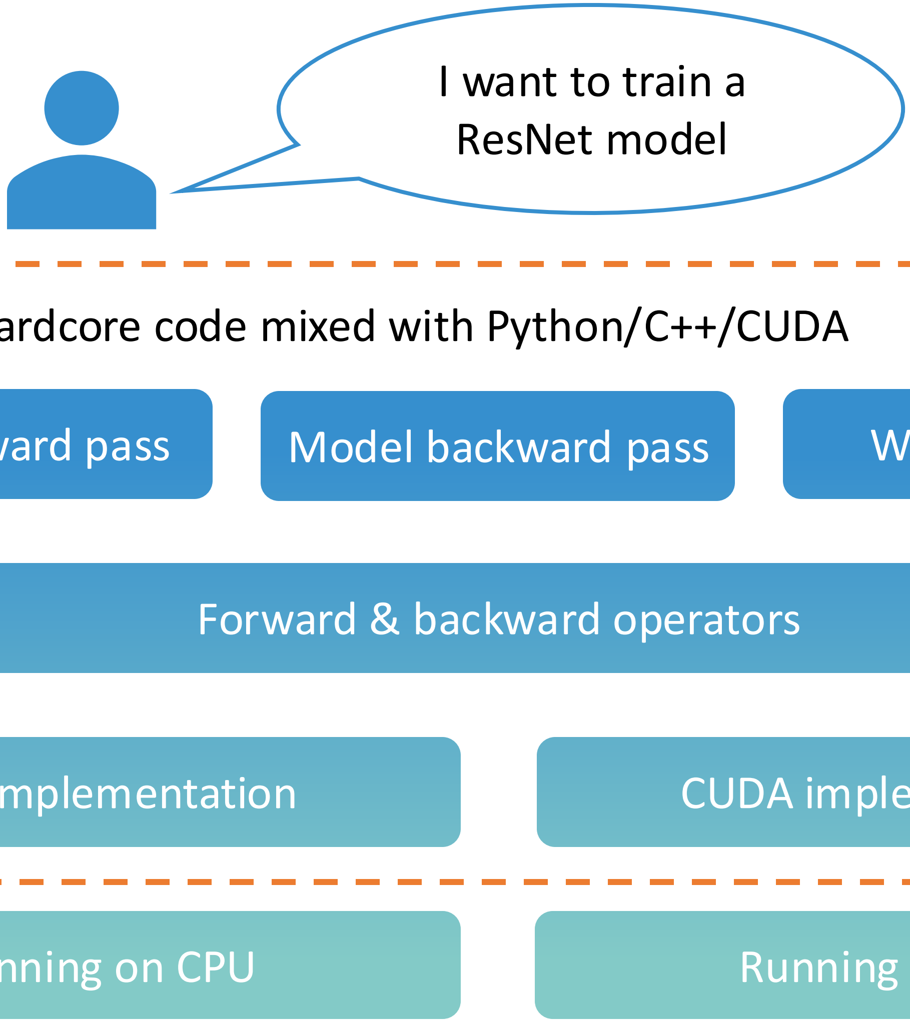
1958 — SVM ConvNet Backprop GBM — **ML Research**

2000 — **Data**

2007 — Public cloud — **Compute**

Slides credits to Tianqi Chen

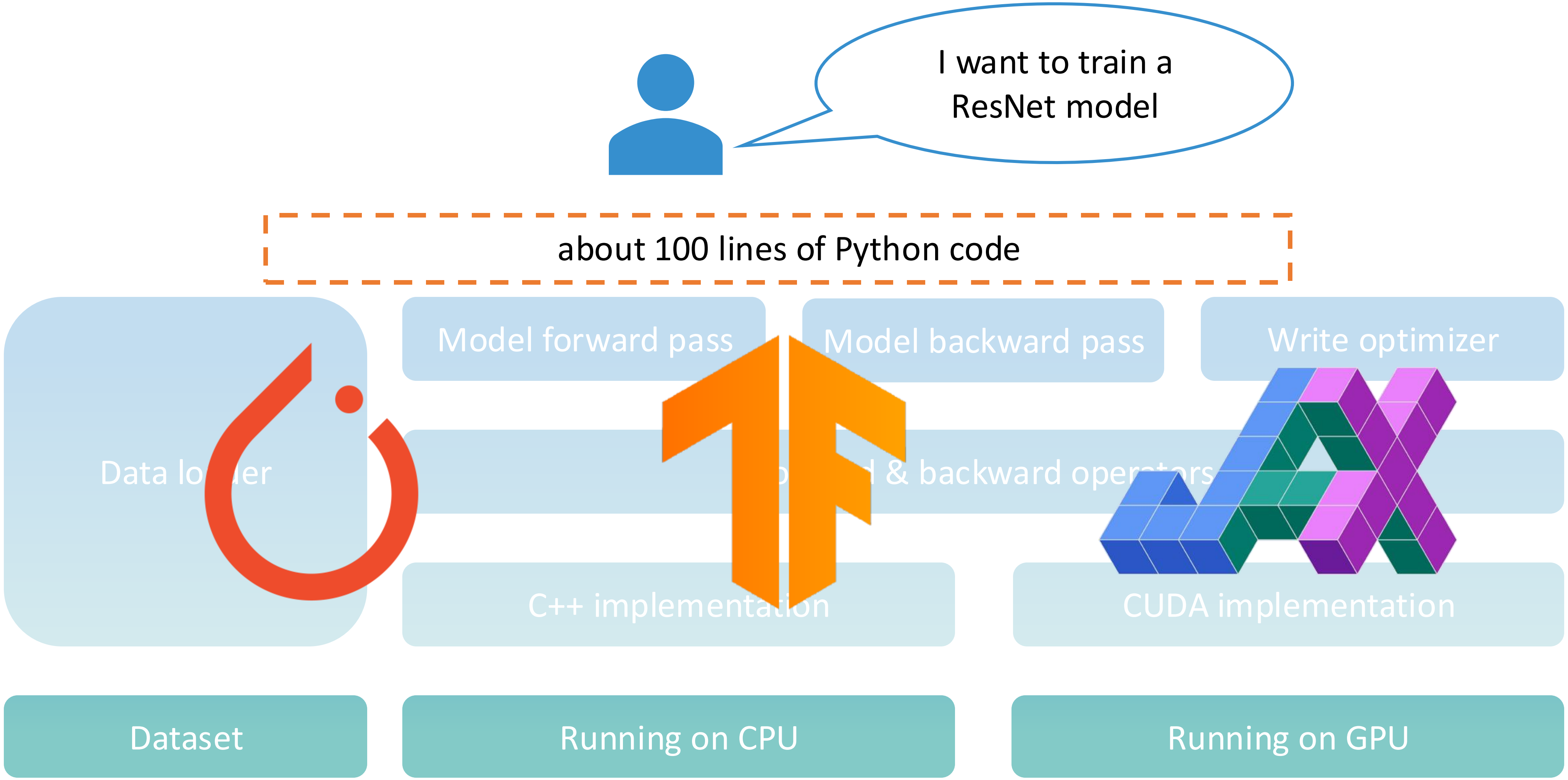
Research w/o ML System



I want to train a ResNet model

10k-100k lines of hardcore code mixed with Python/C++/CUDA

Research w/ ML System



I want to train a ResNet model

about 100 lines of Python code

Model forward pass

Model backward pass

Write optimizer

Data lo ler

o l & backward oper rs

C++ implementation

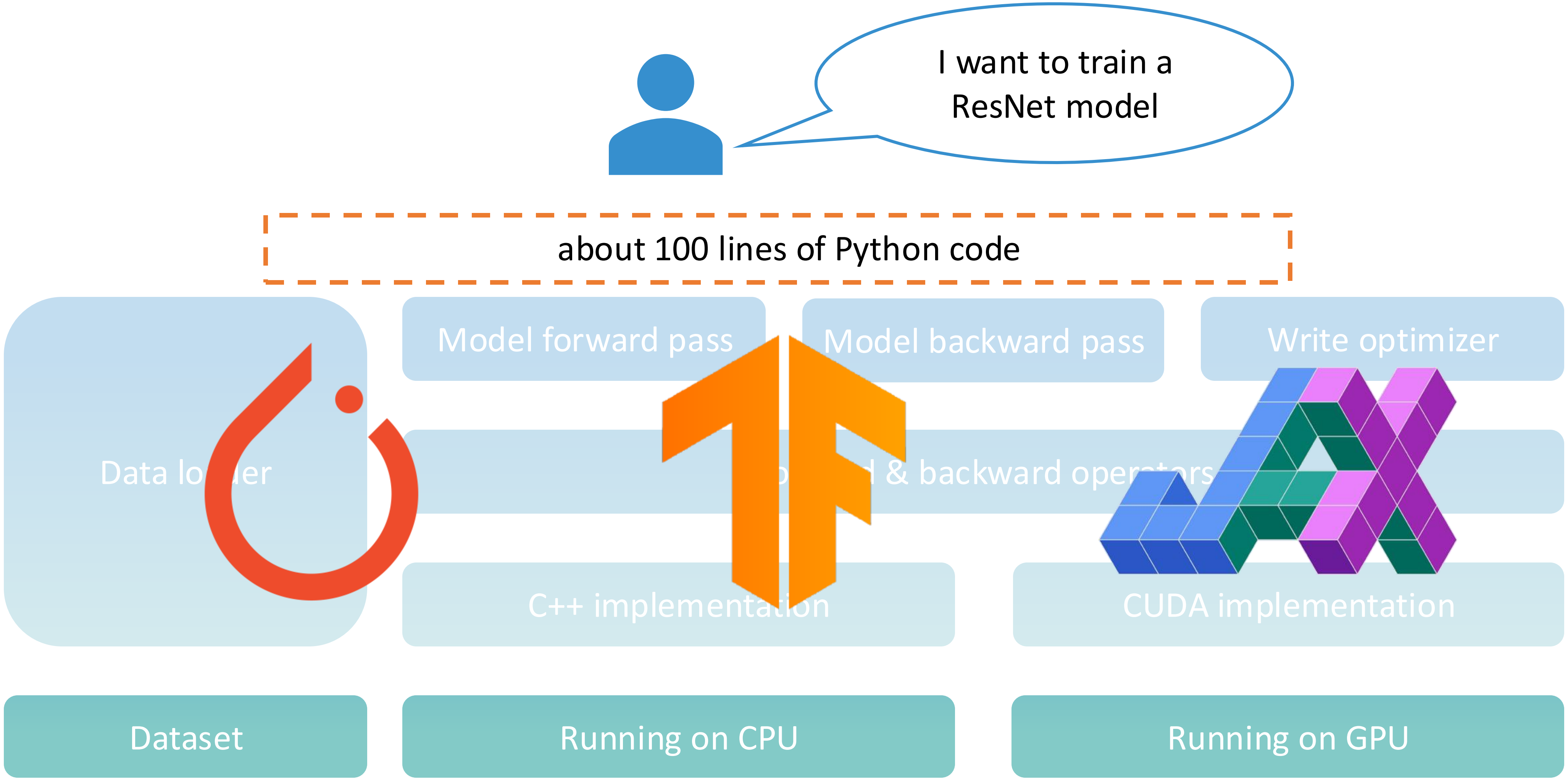
CUDA implementation

Dataset

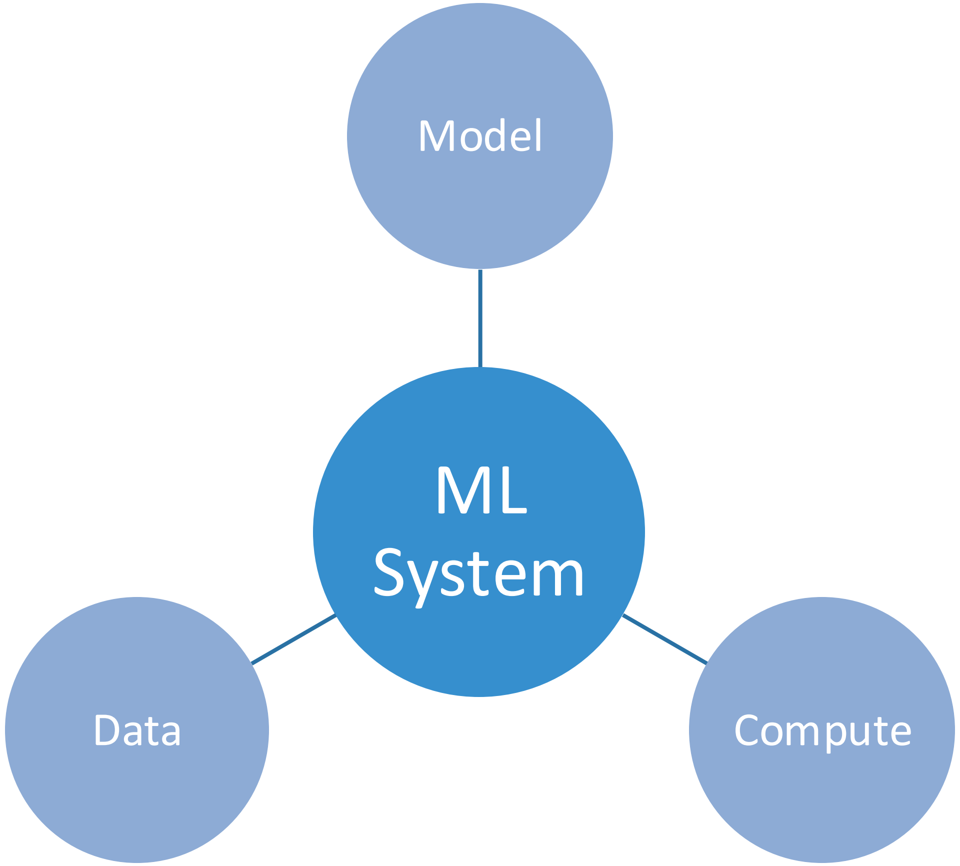
Running on CPU

Running on GPU

Research w/ ML System



ML system as a Bridge



ML system is **the bridge** across model, data, and compute hardware

ML system plays **a central role** during the whole ML research problem

Why Study AI System?

1. AI is revolutionizing everything, and systems are foundation.
2. Understanding the fundamental principles of how AI works facilitates better development of models and algorithms.
3. In the context of geopolitical rivalry, it is crucial to support our nation's own computing power.
4. Huge industry demand and high salaries.

02

How to Learn

Rapid Developing – When Popular Project Release

Project	Scenario	Release Date
TensorFlow	DL framework	Nov 2015
PyTorch	DL framework	Oct 2016
Transformers (Hugging Face)	Library	Nov 2018
Megatron-LM	LLM training	Sep 2019
DeepSpeed	LLM training	May 2020
FlashAttention	Kernel	May 2022
vLLM	LLM serving	Jun 2023
SGLang	LLM serving	Dec 2023
Mooncake	KVCache management	Jun 2024
verl	RL post-training framework	Oct 2024
DeepEP	Communication library	Feb 2025

Most of projects for LLMs started at about 2 years ago and are still under rapid developing

No Textbook but only Materials

- Machine Learning / Deep Learning
  - Dive into Deep Learning: https://d2l.ai
- Machine Learning Systems
  - Open-Sourced Book Machine Learning Systems: https://mlsysbook.ai
  - Machine Learning Compilation: https://mlc.ai
  - Microsoft AI-System Education Resource (Chinese): https://github.com/microsoft/AI-System

English materials are recommended

Ask AI when Possible

- API is not enough, **MAKE SURE** to enable online search
- Spend some time to learn how to write prompt

03

Course Logistics

Course Instructor

Siyuan FENG (冯思远)

Assistant professor at Shanghai Innovation Institue

- Ph.D. in computer science, SJTU
- B.Sc. in computer science, ACM class, SJTU
- Experiences in compiler, AI system, AI accelerator
- Office at 1203-1
- Email: syfeng@sii.edu.cn

Course Instructor

Chengcheng WAN (万成城)

Associate professor at East China Normal University / SII

- Postdoctoral in computer science, University of Chicago
- Ph.D. in computer science, University of Chicago
- B.Sc. in computer science, SJTU
- Experiences in machine learning software system
- Email: ccwan@sei.ecnu.edu.cn

Course TAs

Yubing GAO (高毓兵)

Zhixin WANG (王治鑫)

Tianyi ZHOU (周天怡)

上海创智学院
Shanghai Innovation Institute

Course Schedule

Date	Plan	Lab schedule	Lab Topic
Sep 18	Introduction to system for AI		
Sep 25	Automatic differentiation	Lab 1 release	autograd system
Oct 2	[SKIP] National Holiday		
Oct 9	Hardware acceleration		
Oct 16	GPU architecture and CUDA programming	Lab 1 due	
Oct 23	NPU architecture and Ascend C programming	Lab 2 release	cuda **AND** ascend operator implementation
Oct 30	Machine learning compilation		
Nov 6	Introduction to LLMs and optimizations		
Nov 13	Introduction to distributed computing	Lab 2 due / Lab 3 release	tensor parallel
Nov 20	LLMs: parallelization and training techniques I		
Nov 27	LLMs: parallelization and training techniques II		
Dec 4	LLMs: parallelization and training techniques III	Lab 3 due	
Dec 11	LLMs: serving techniques I	proposal due	
Dec 18	LLMs: serving techniques II		
Dec 25	LLMs: serving techniques III		
Jan 1	[SKIP] New Year's Day		
Jan 8	LLMs: post-training techniques		
Jan 15	Project presentation		

Prerequisites

- Basic mathematical background
- Basic linear algebra knowledge
- Strong Python programming skill
- (Optional) C++/CUDA programming
- (Optional) Computer architecture and network

- **Prompt engineering and vibe coding skill**

Grading

- [10%] Class Participation
  - 1st unexcused absence: Warning
  - 2nd unexcused absence: Attendance grade is halved.
  - 3rd unexcused absence: Attendance grade becomes 0.
- [45%] Assignments
  - Each of three assignments is 15%.
- [45%] Course Project (Groups of 2-3)
  - Proposal: 5%
  - Technical Report: 20%
  - Presentation: 20%
  - Topic: AI **SYSTEM** related.
- [10%] Extra Bonus
  - Each meaningful question during the class: +1
  - Each bug finding (except typo): +2

Course Exemption Policy

- Students who have been **leading an AI systems research project** and **achieved significant milestones** may be exempted from attendance and assignment requirements. Their grade for these components will be awarded based on a fixed score of **10+40/10+45**, but they are still required to complete the course project.
- Students who have **published a paper as the (co-)first author in a top-tier systems conference** or make outstanding contributions in **globally influential communities** within the **last two years** . may be granted a full exemption from the course and will receive a final grade of **90/100**.
- Sending email to **Siyuan FENG** if you'd like to apply an exemption before **Sep 25**

Disclaimers

- This is a first time offering of this course, may have bugs or errors in content or assignments
- Industry & open-source world evolving ultra fast.
- The material and outline will likely adjust throughout the semester.

Policy for the Use of AI Tools

The use of any form of Artificial Intelligence (AI) tools is **permitted** and **encouraged** in this course to support learning and research. Students are not required to declare or cite the use of AI tools in their submissions, including but not limited to assignments, projects, and reports.

Students **are held fully accountable** for all submitted materials, including but not limited to code, experimental data, and technical reports. Should the use of AI tools result in any adverse consequences—such as the submission of malicious or destructive code, data fabrication, or other forms of academic misconduct including plagiarism or excessive similarity to other works—the student who made the submission will bear sole responsibility.

Acknowledgement

The development of this course, including its structure, content, and accompanying presentation slides, has been significantly influenced and inspired by the excellent work of instructors and institutions who have shared their materials openly. We wish to extend our sincere acknowledgement and gratitude to the following courses, which served as invaluable references and a source of pedagogical inspiration:

- Machine Learning Systems[15-442/15-642], by **Tianqi Chen** and **Zhihao Jia** at **CMU**.
- Advanced Topics in Machine Learning (Systems)[CS6216], by **Yao Lu** at **NUS**

While these materials provided a foundational blueprint and a wealth of insightful examples, all content herein has been adapted, modified, and curated to meet the specific learning objectives of our curriculum. Any errors, omissions, or shortcomings found in these course materials are entirely our own responsibility. We are profoundly grateful for the contributions of the educators listed above, whose dedication to teaching and knowledge-sharing has made the creation of this course possible.

System for Artificial Intelligence

Thanks

Siyuan Feng
Shanghai Innovation Institute