Introduction

DS 5110: Big Data Systems

Spring 2025

Lecture 1

Yue Cheng

Introduction – Yue Cheng

- On the faculty of Data Science & Computer Science
  - Web: https://tddg.github.io
  - Email: mrz7dp@virginia.edu

- Current research: Designing **better data systems**
  - Serverless computing systems
  - Storage systems
  - Systems **X** + AI/ML

Course staff and getting help

- Instructor: Yue Cheng
  - Office hours: Thursday 3:15pm – 4:15pm, DS building, Room 435

- GTA:
  - Lehan Yang
  - Email: cjy6qy@virginia.edu
  - Office hours: Monday 2pm – 6pm, Zoom

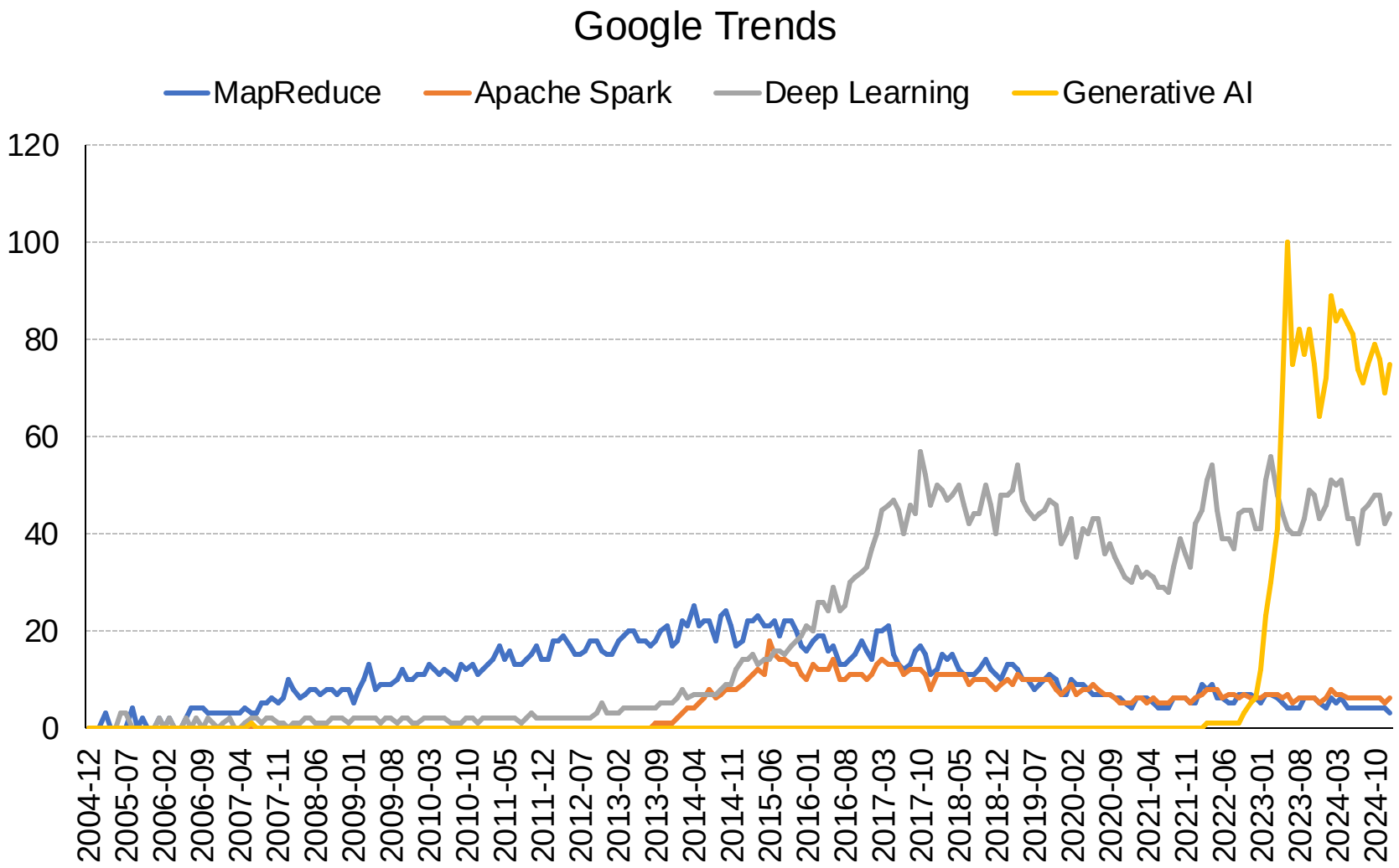
Course staff and getting help

- Offline discussion, questions: Ed
  - https://edstem.org/us/dashboard
  - Alternative place to ask questions about assignments, materials, and ideas
  - No anonymous posts or questions
  - Can use private posts to instructor/GTA
  - We are monitoring Ed several times a day
  - We will respond to questions in a batch manner

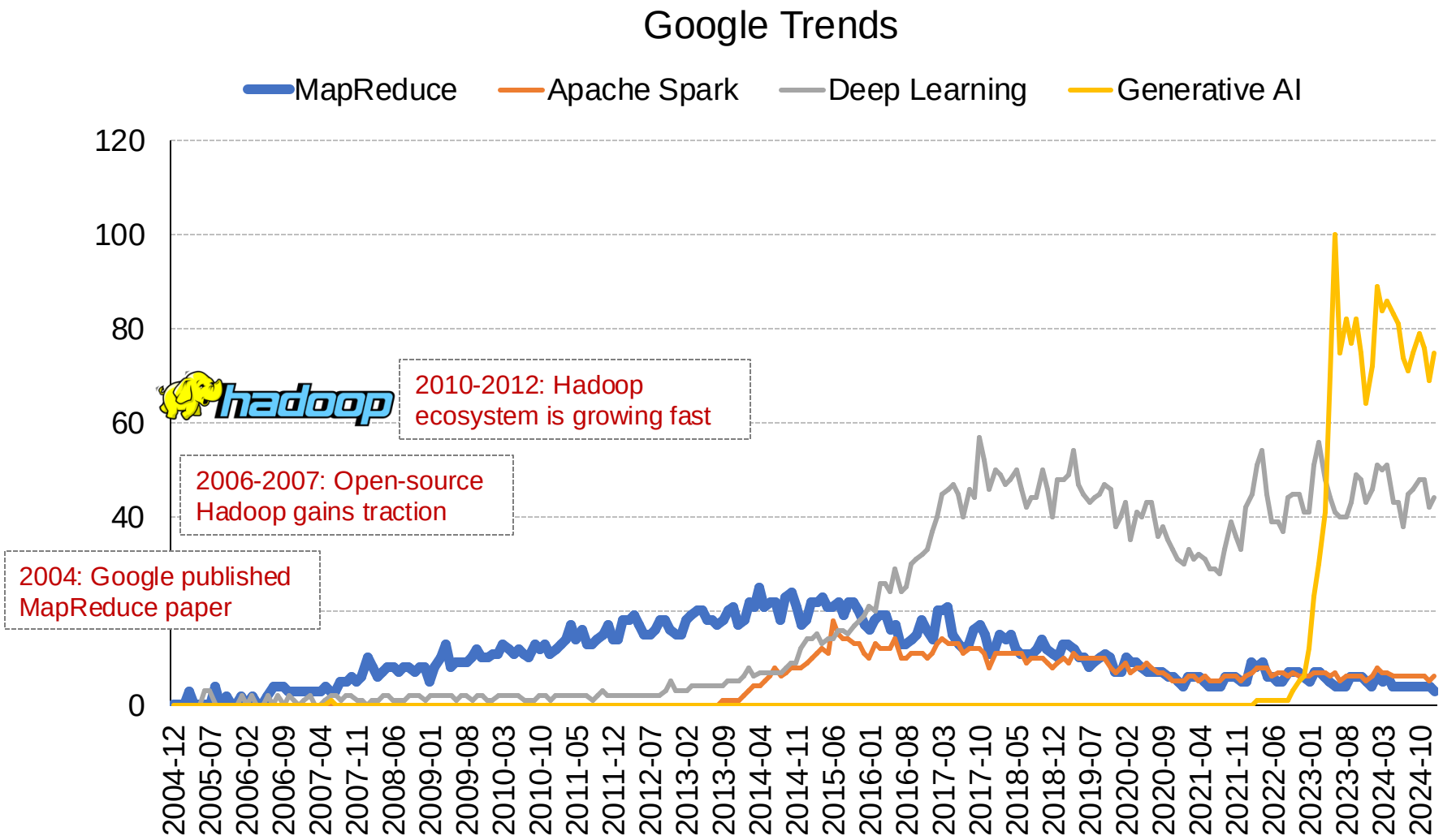
Today's agenda

- What is this course about?
- What will you do in this course?

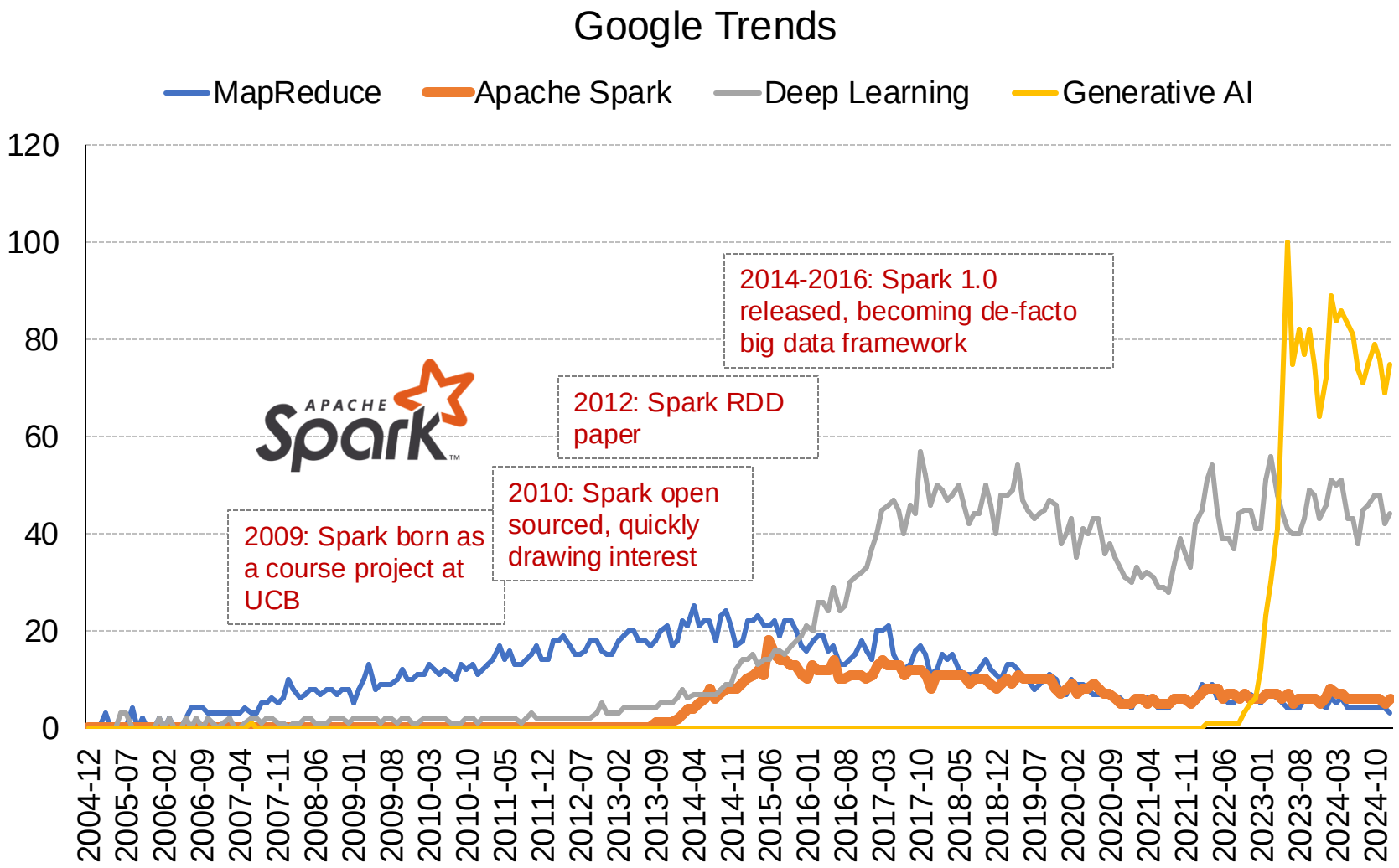
A brief history about Big Data



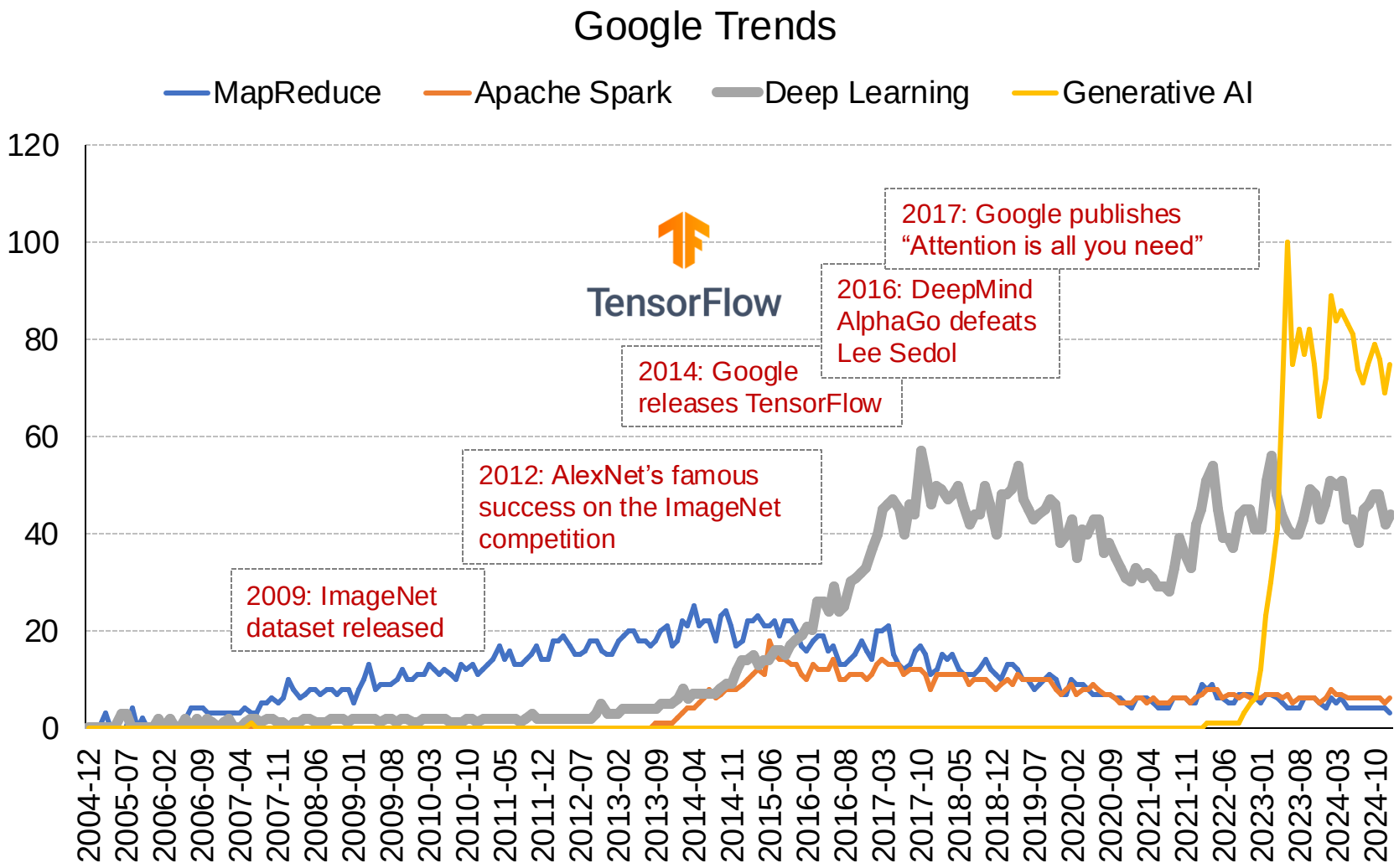
A brief history about Big Data



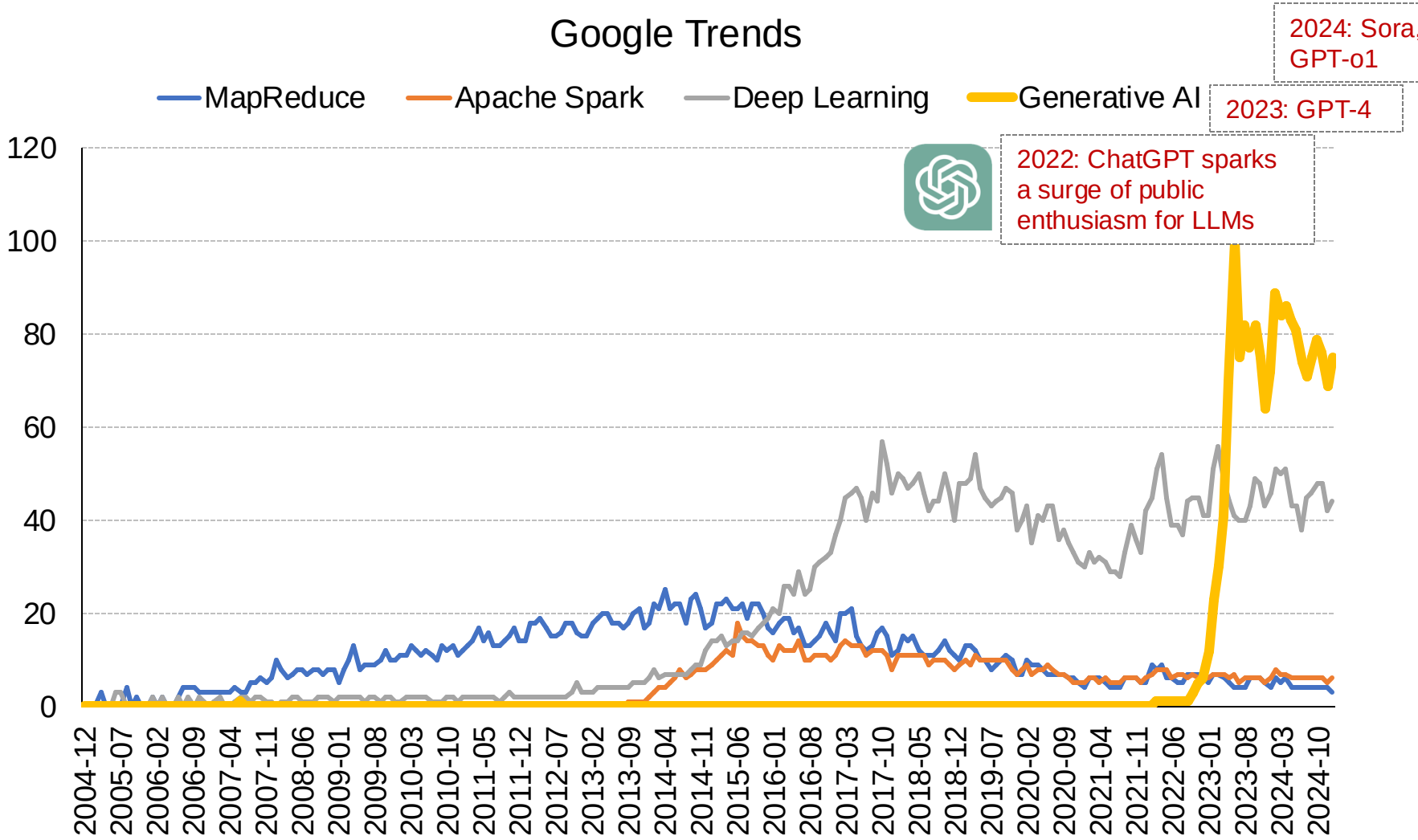
A brief history about Big Data



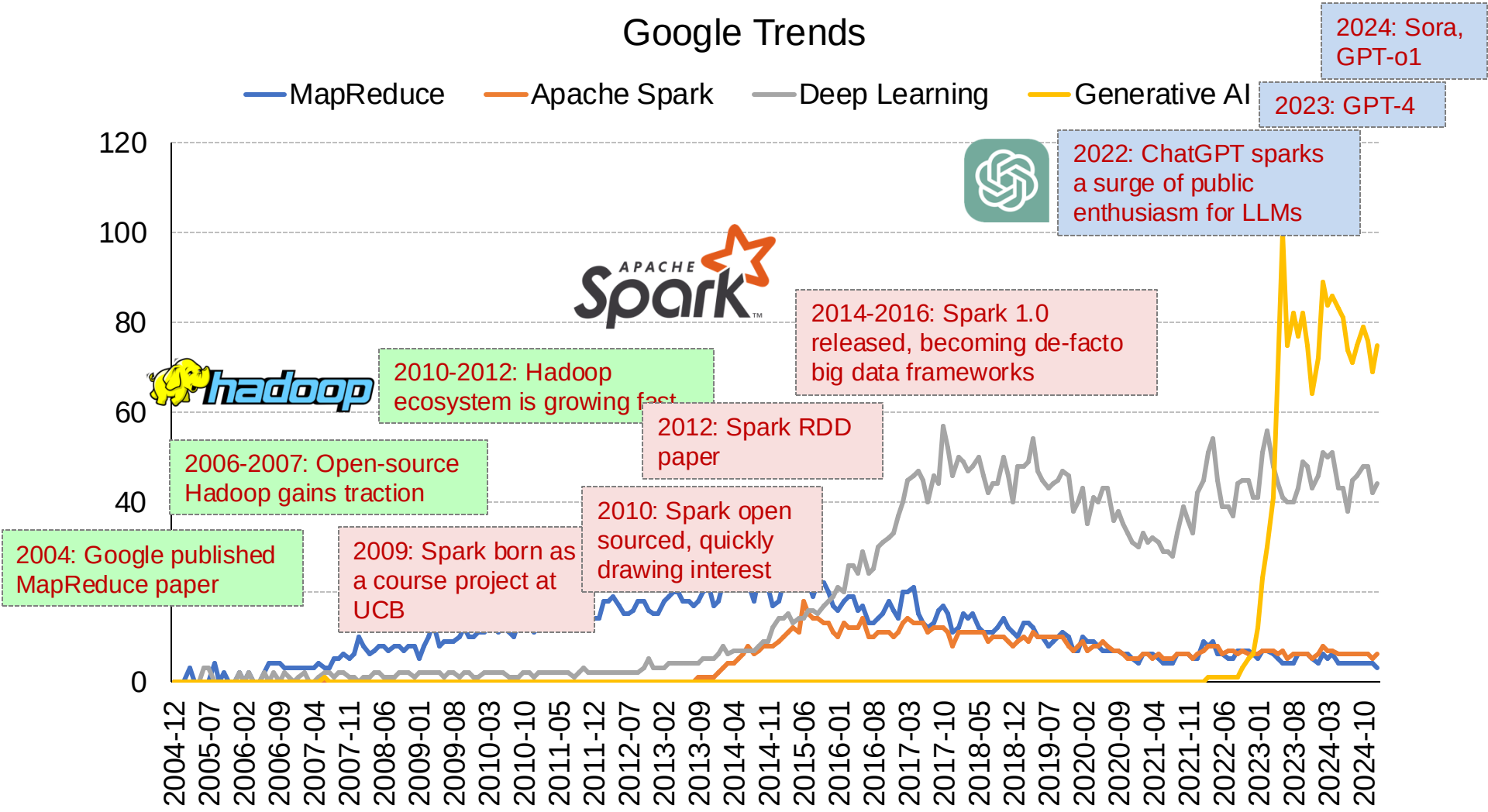
A brief history about Big Data



A brief history about Big Data



DS5110 narrative: A time machine



Google circa 1997

Everything is about data

"… Storage space must be used efficiently to store indices and, optionally, the documents themselves. The indexing system must process hundreds of gigabytes of data efficiently…"

"The system… downloading the last 11 million pages in just 63 hours… The sorter can be run completely in parallel; using four machines, the whole process of sorting takes about 24 hours…"

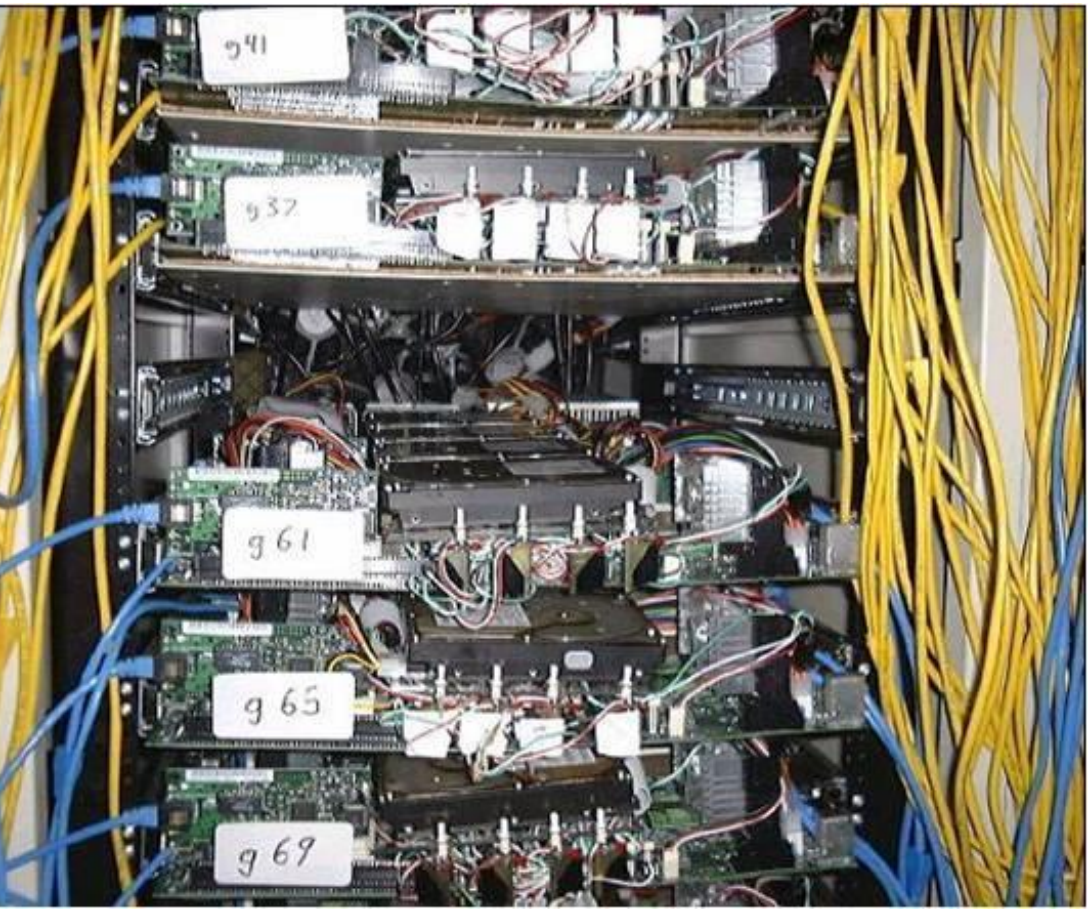
The anatomy of a large-scale hypertextual Web search engine [1]

Sergey Brin [2], Lawrence Page [*,2]

Computer Science Department, Stanford University, Stanford, CA 94305, USA

Abstract

Google circa 2000

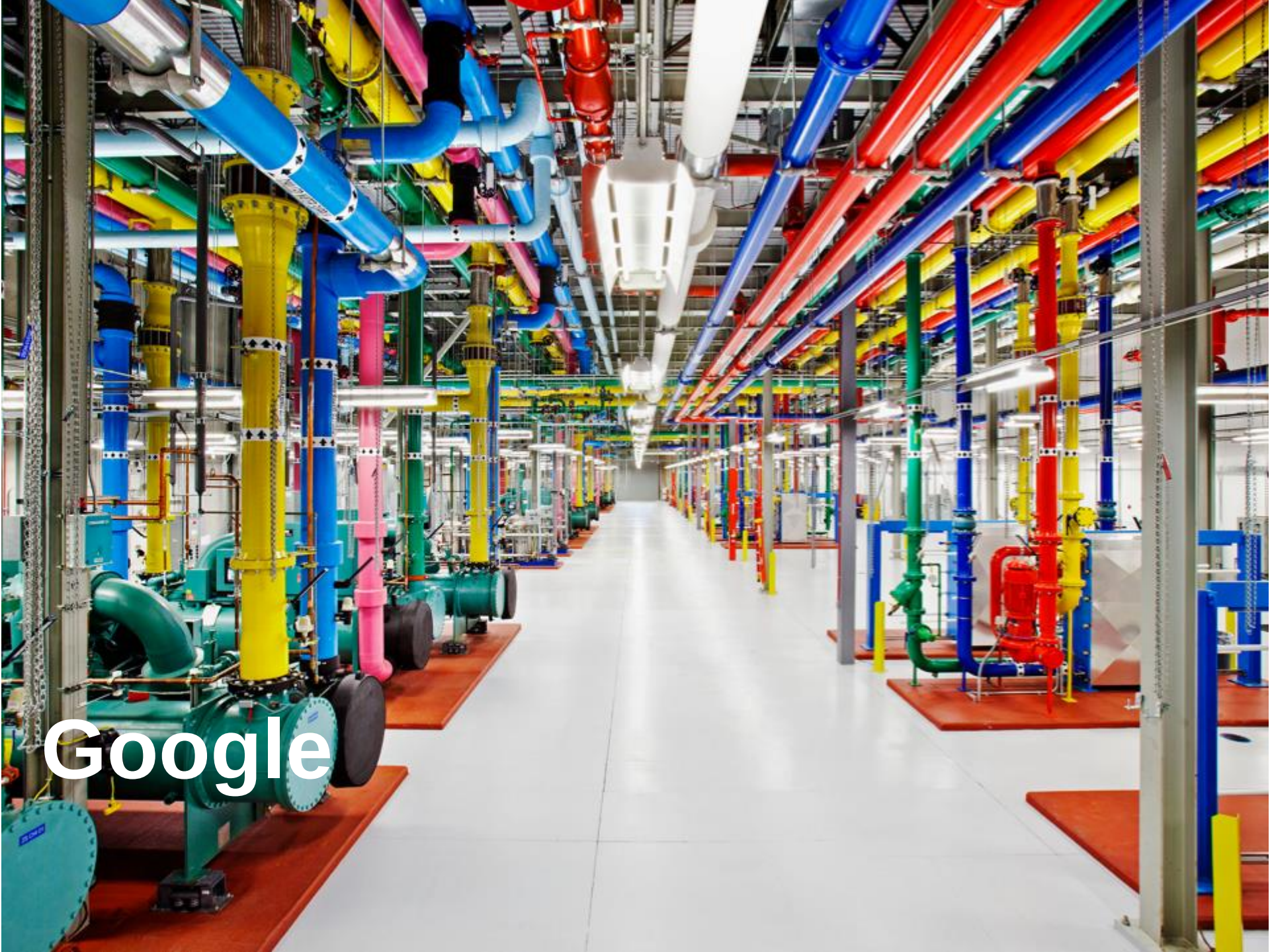


Commodity CPUs

Lots of disks

Low bandwidth network

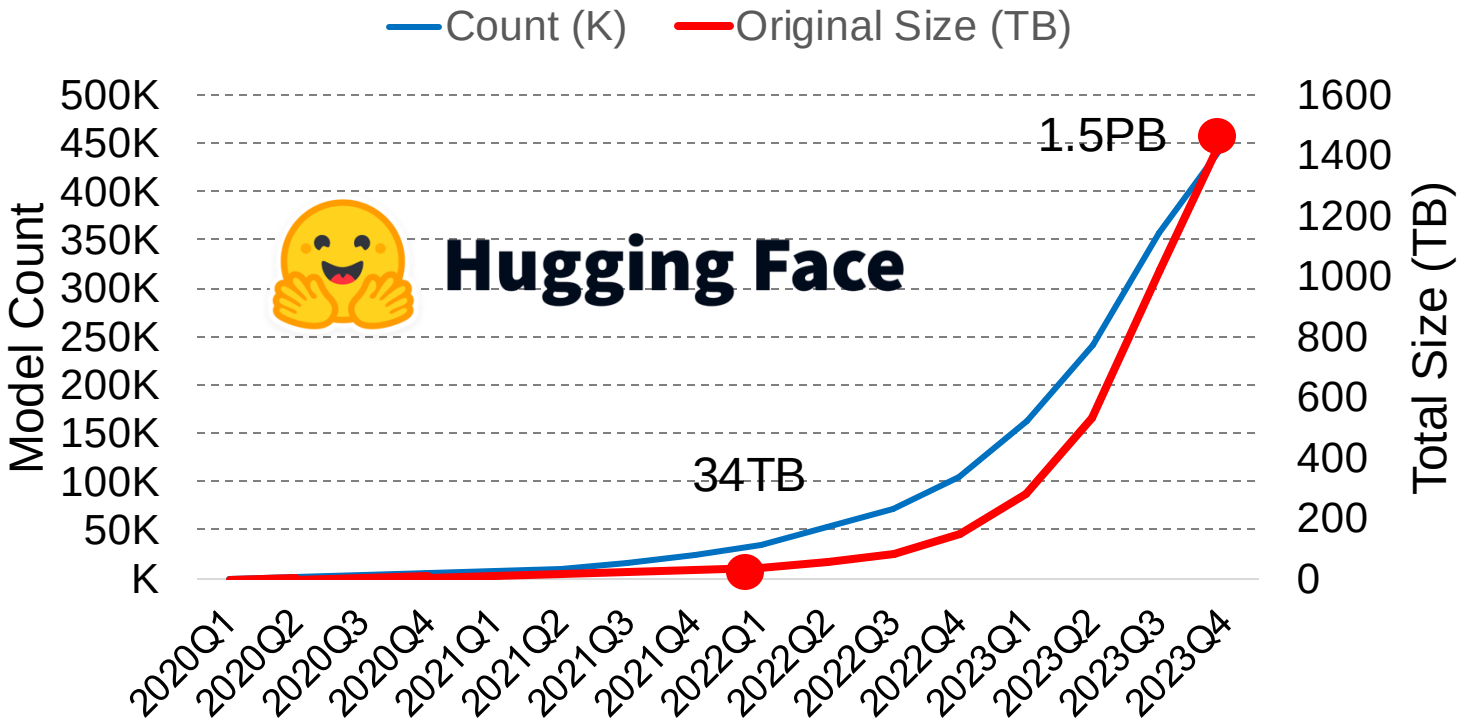
Cheap!


Google

A Google Datacenter in Hamina, Finland

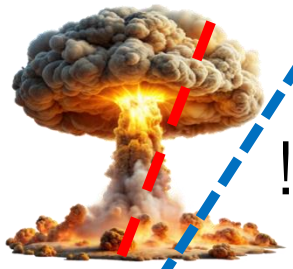
Data explosion in the GenAI era

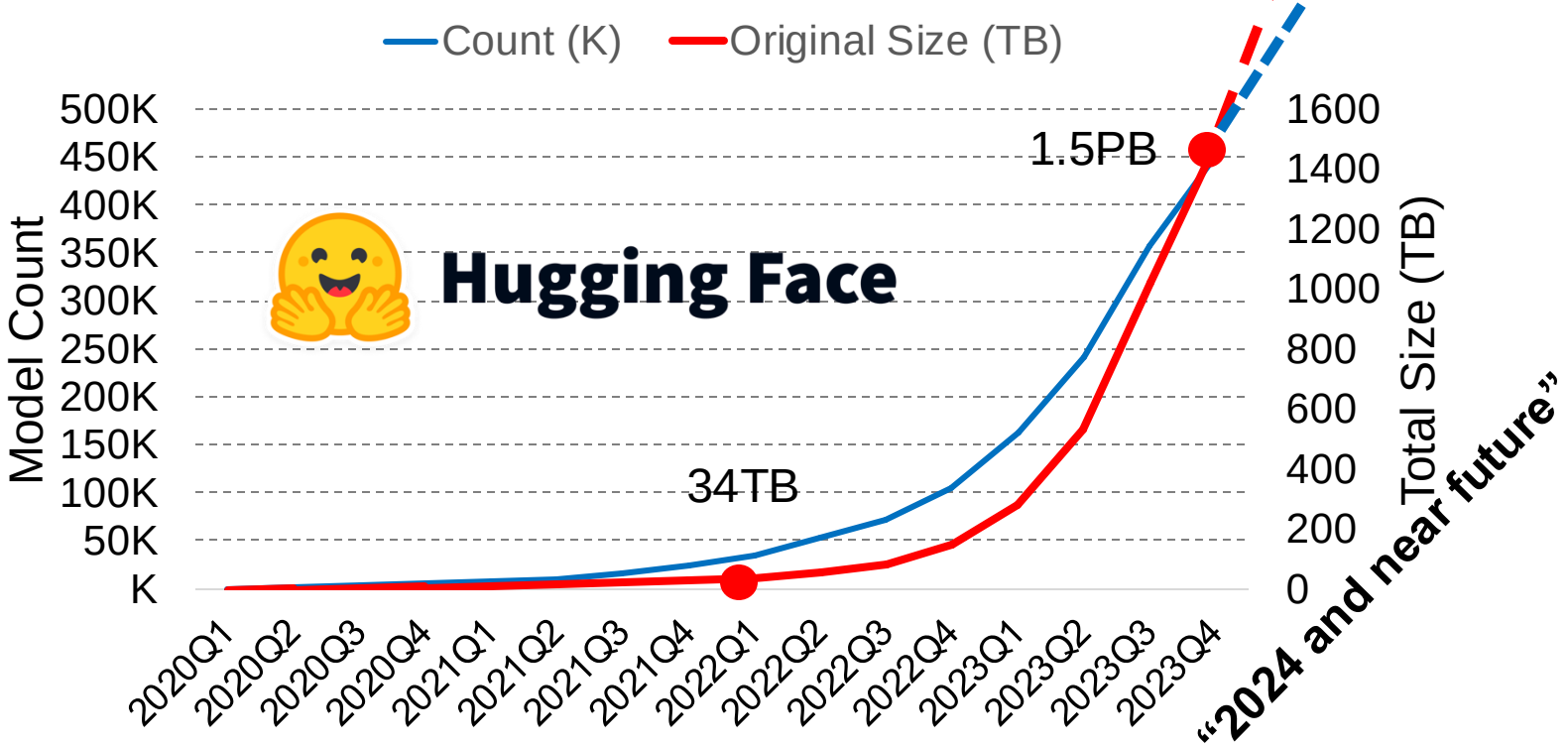
- Hugging Face: ML model storage is  !



Data explosion in the GenAI era

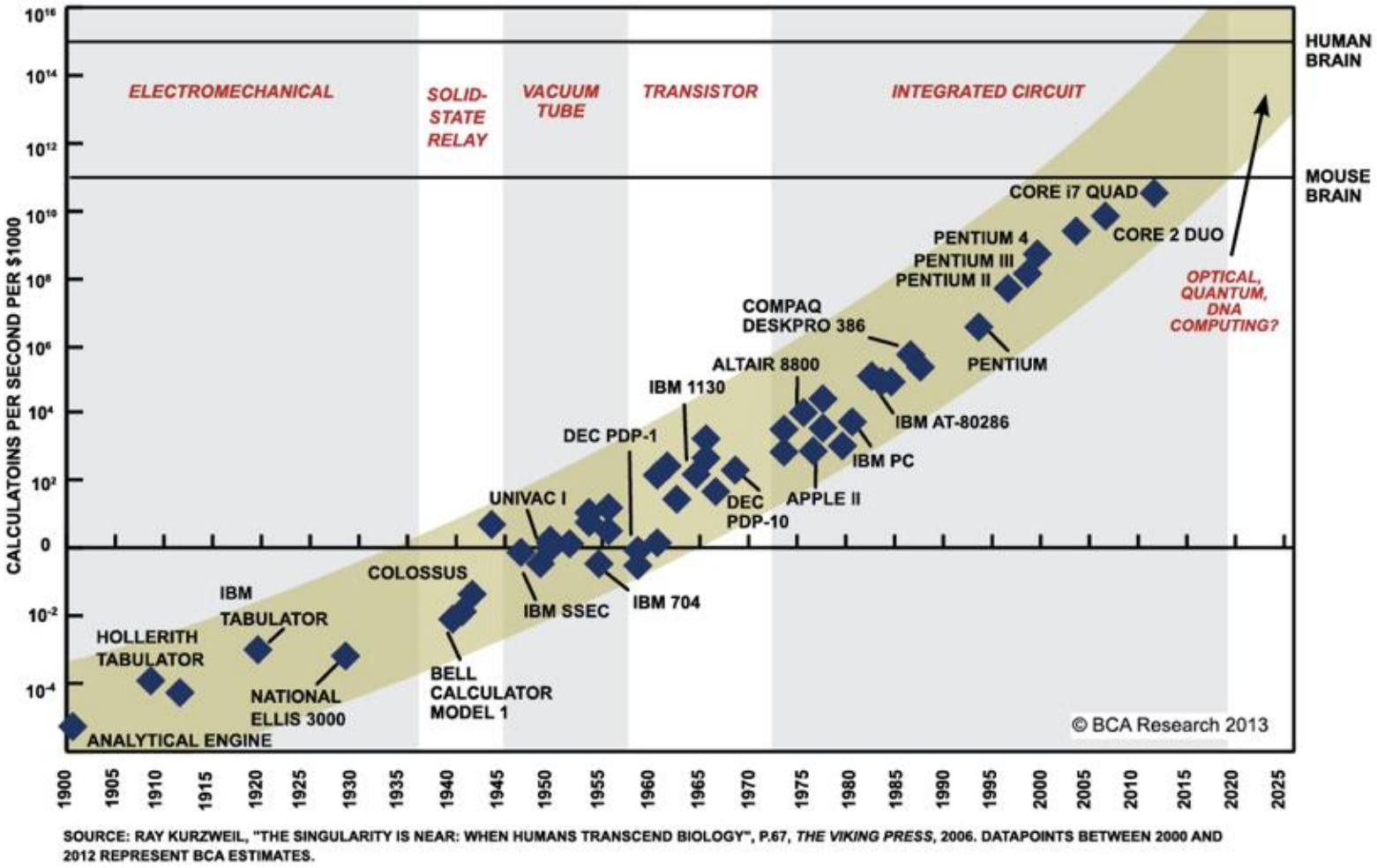
- Hugging Face: ML model storage is !





HuggingFace's AI/ML models are growing exponentially!

Moore's law is ending



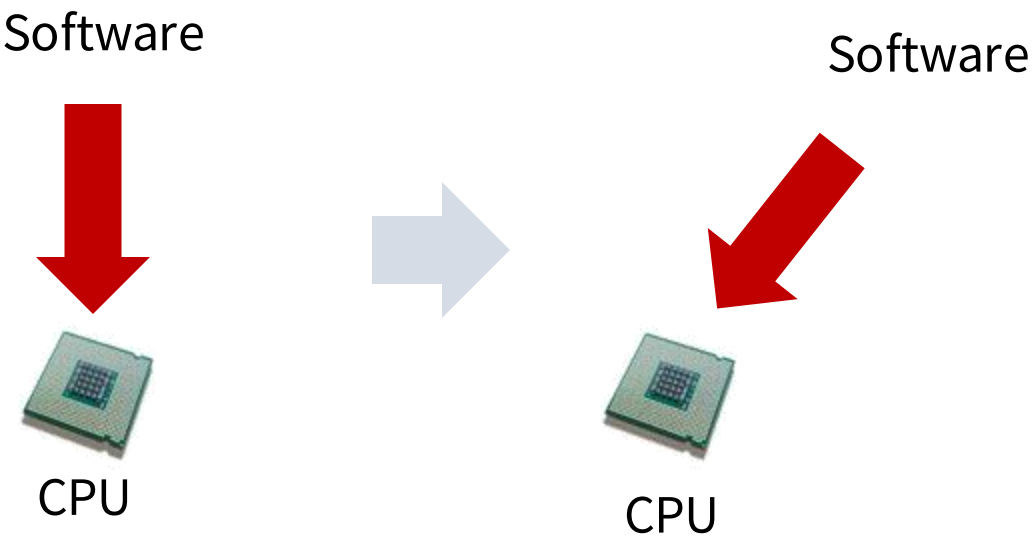
SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

Increased complexity – Computation

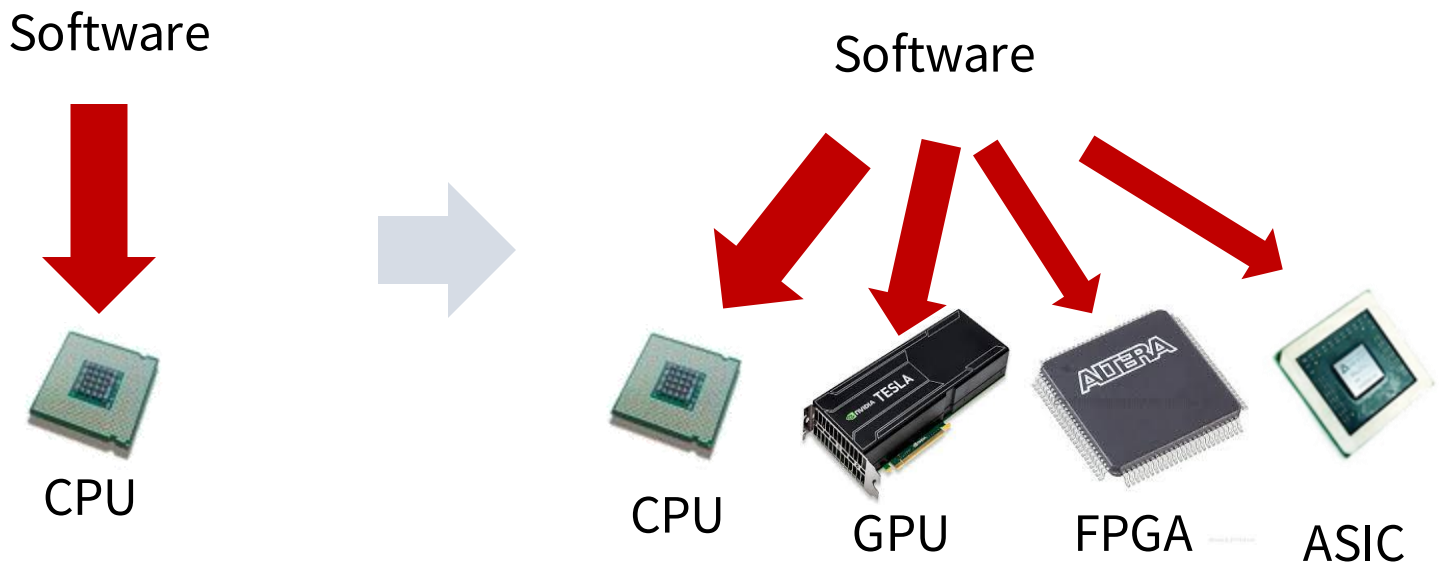
Software

CPU

Increased complexity – Computation



Increased complexity – Computation



Increased complexity – More and more choices in clouds

Basic tier: A0, A1, A2, A3, A4
Optimized Compute : D1, D2, D3, D4, D11, D12, D13
D1v2, D2v2, D3v2, D11v2,…
Latest CPUs: G1, G2, G3, …
Network Optimized: A8, A9
Compute Intensive: A10, A11,…

Microsoft Azure

t2.nano, t2.micro, t2.small
m4.large, m4.xlarge, m4.2xlarge, m4.4xlarge, m3.medium, c4.large, c4.xlarge, c4.2xlarge, c3.large, c3.xlarge, c3.4xlarge, r3.large, r3.xlarge, r3.4xlarge, i2.2xlarge, i2.4xlarge, d2.xlarge d2.2xlarge, d2.4xlarge,…

Amazon EC2

n1-standard-1, ns1-standard-2, ns1-standard-4, ns1-standard-8, ns1-standard-16, ns1highmem-2, ns1-highmem-4, ns1-highmem-8, n1-highcpu-2, n1-highcpu-4, n1-highcpu-8, n1-highcpu-16, n1-highcpu-32, f1-micro, g1-small…

Google Cloud

But how do we program this to tackle the challenges of big data?

Course syllabus

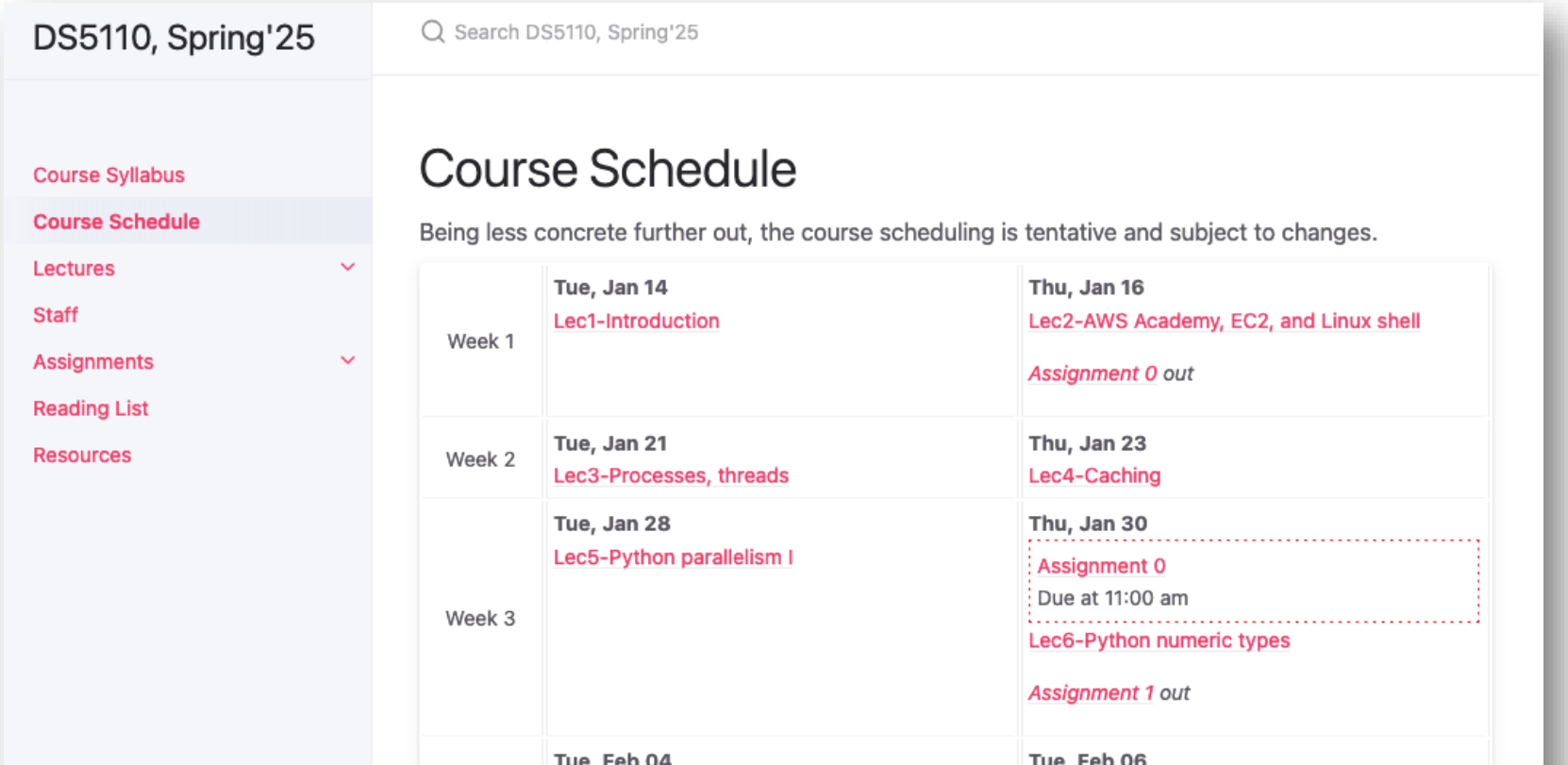
Big picture course goals

- Learn about some of the most influential works in (big) data systems
- Explain the design and architecture
- Read and evaluate some seminal papers
- Develop and deploy applications on open-source data systems (Spark, HDFS, Dask, Ray, PyTorch) and public cloud (AWS)
- Design and report some data systems ideas

Schedule (tentative)

- Readings, assignments, due dates
- Less concrete further out; don't get too far ahead

https://tddg.github.io/ds5110-spring25/



Lectures (tentative schedule)

- Lecture (+ discussion + demos)
  - Slides available on course website (night before or morning on the same day)
- First 3 weeks: Basics of computer systems
  - Mostly from textbook
- Week 4-6: Python analytics, MapReduce, Spark
- Week 7: Midterm exam
- Week 8-10: Ray (Week 9 is Spring break)
- Week 10-11: LLM systems
- Week 12-15: Cloud, serverless, cloud AI infra
- Guest speakers invited from industry (AWS, Hugging Face, Microsoft, etc.)

Readings

- Goal: read and help with better understanding

- Reading questions will be posted on Ed few days before lecture
  - Reading questions will cover required reading with a strong focus on stimulating a fruitful discussion
  - You don't need to fill out them (given excessive use of LLM tools like ChatGPT)
  - We will have **in-class discussions** (cold call if no one volunteers)
    - **Asking questions is highly encouraged!**

Textbooks?

- Papers, documentations, blog articles (required or optional) serve as reference for many topics that aren't directly covered by a text

- Slides/lecture notes

- Three optional textbooks (first two are free)
  - "**Operating Systems: Three Easy Pieces (OSTEP)**" by Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau
  - "**Distributed Systems (3rd edition)**" by van Steen and Tenenbaum will supply optional alternate explanations
  - "**Designing Data-Intensive Applications (1st edition)**" by Martin Kleppmann (can be accessed via UVA library)

Assignments

- Five programming assignments, in **Python**, on **AWS**
  - **Assignment 0:** Using AWS Academy, EC2, and Linux shell
  - **Assignment 1:** Parallelizing Python processing with Dask
  - **Assignment 2:** A tour of Apache HDFS and Spark
  - **Assignment 3:** A deeper dive with Ray
  - **Assignment 4:** (Small) LLM pipeline

- All assignments are individual

- Short coding-based assignments
  - Gain hands-on experience with AWS and popular open-source data systems and cloud tools
  - Get exposed to data processing & MLOps pipeline

Grading

- Assignments (65% total)
  - Assignment 0 (5%)
  - Assignment 1 (10%)
  - Assignment 2 (15%)
  - Assignment 3 (15%)
  - Assignment 4 (20%)
- Quizzes (5%)
- Two exams (30%): open-book, open-note, on gradescope
  - Midterm exam (15%)
  - Final exam (15%)
- Participation: in-class Q&A (5% extra credit as an incentive)

Time to introduce yourself