Foundations of Edge AI

Lecture01

Introduction and Course Overview

Lanyu (Lori) Xu

Email: lxu@oakland.edu
Homepage: https://lori930.github.io/
Office: EC 524

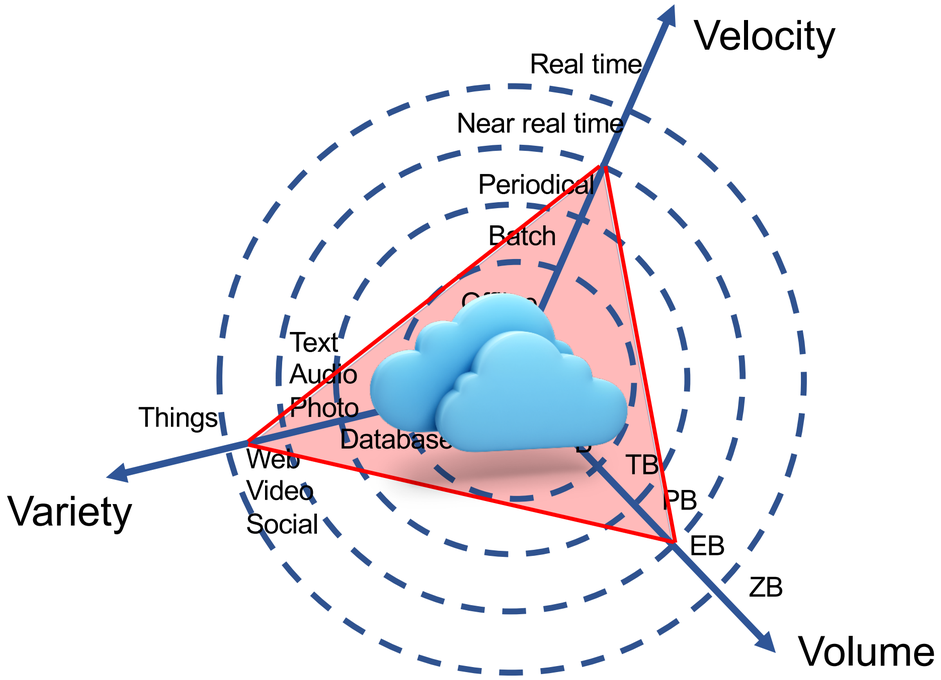
What is This Course About?

- **Edge AI = Edge Artificial Intelligence = Edge Computing + Artificial Intelligence**
- What is Edge Computing?
- What is the current landscape of Artificial Intelligence?
- Why Edge AI?
- How to achieve Edge AI?

Cloud Computing?

Digital Infrastructure 1.0 (2005 - 2015)

Personal Computer (PC) era -> “Cloud Computing” era



- Landmark event: IBM sold its PC business to Lenovo
- “The datacenter is the computer.”

TECHNICAL PERSPECTIVE:

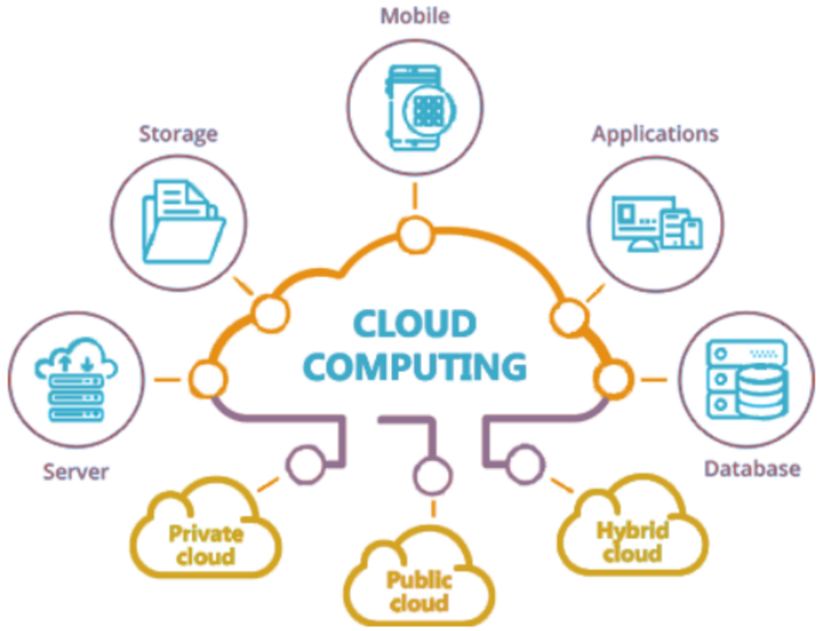
THE DATA CENTER IS THE COMPUTER

by David A. Patterson

nternet services are already significant forces in searching, retail purchases, music downloads, and auctions. One vision of 21st century IT is that most users will be accessing such services over a descendant of the cell phone rather than running shrink-wrapped software on a descendant of the PC.

Patterson, David A. "Technical perspective: the data center is the computer." Communications of the ACM 51.1 (2008): 105-105.

Cooking vs Cloud Computing



- Imagine a bustling kitchen where the world's **data flows in like ingredients** through a pipeline, ready to be transformed.
  - The **cloud** is a pot with **no limits** to storage and computing resources
  - The **data** (e.g., text, image, video, ...) is "stirred" and "cooked"
  - The **computation** is seasonings
  - The **results** are ready to be served to real-world applications (e.g., financial transactions, web services)

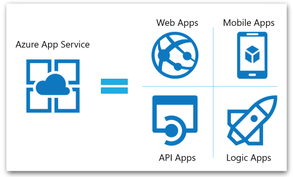
Cloud Service Providers

- Advanced infrastructure
- Transparent cost

👍👍👍

- Usability
- Scalability

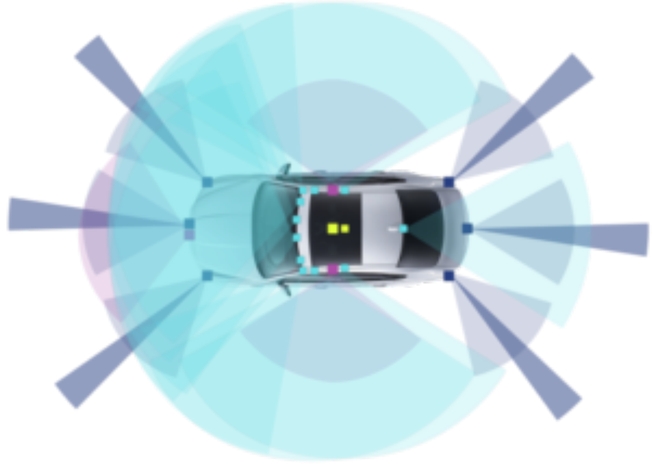
- **X as a Service (XaaS)**
  - Infrastructure as a Service (IaaS)
  - Platform as a Service (PaaS)
  - Software as a Service (SaaS)
  - Database as a Service (DBaaS)
  - Artificial Intelligence as a Service (AIaaS)
  - Function as a Service (FaaS)
  - ...



From Cloud Computing
To Edge Computing

Data Never Sleeps

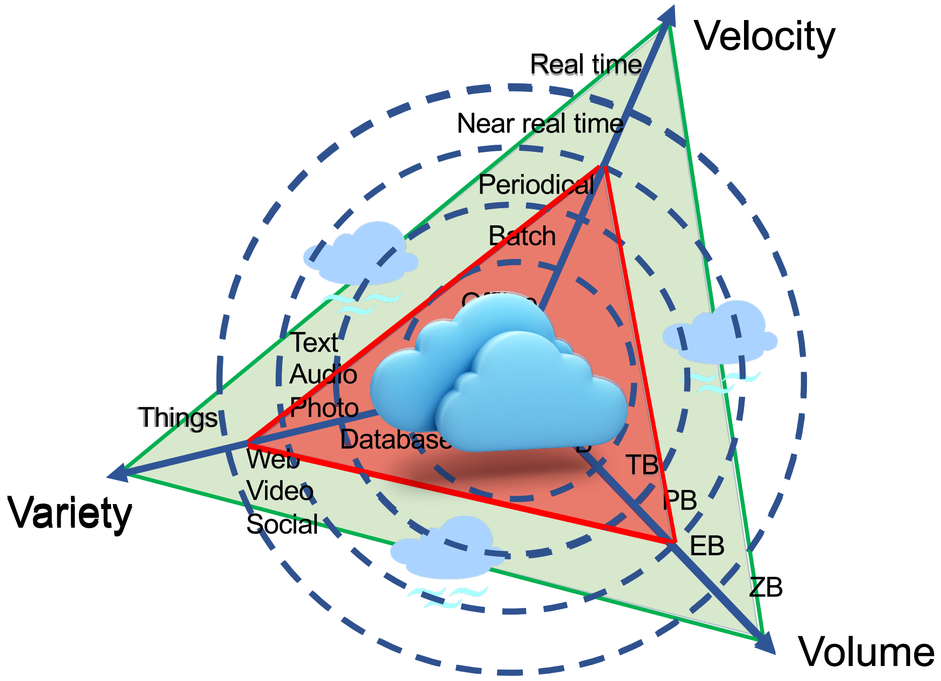
https://www.domo.com/learn/infographic/data-never-sleeps-11



- A single forward facing **radar** operating at 2800 MBits / s generates more than **1.26 TB of data per hour**, and a typical data collection **car** generates in excess of **30TB a day**.
- A two megapixel camera (24 bits per pixel) operating at 30 frames per second generates 1440 Mbits of data every second, so a **five camera setup** can generate in excess of **24 TB per day**.

Edge Computing (2015 -)

“Cloud Computing” era -> “Edge Computing” era



- **Push from cloud service providers**
  - **Reduce latency**
  - **Save bandwidth**
  - **Improve reliability**
- **Pull from the Internet of Things**
  - **Real-time context computing**
  - **Resource constraints**
  - **Security/privacy requirements**

Edge Computing is to Complement Cloud Computing

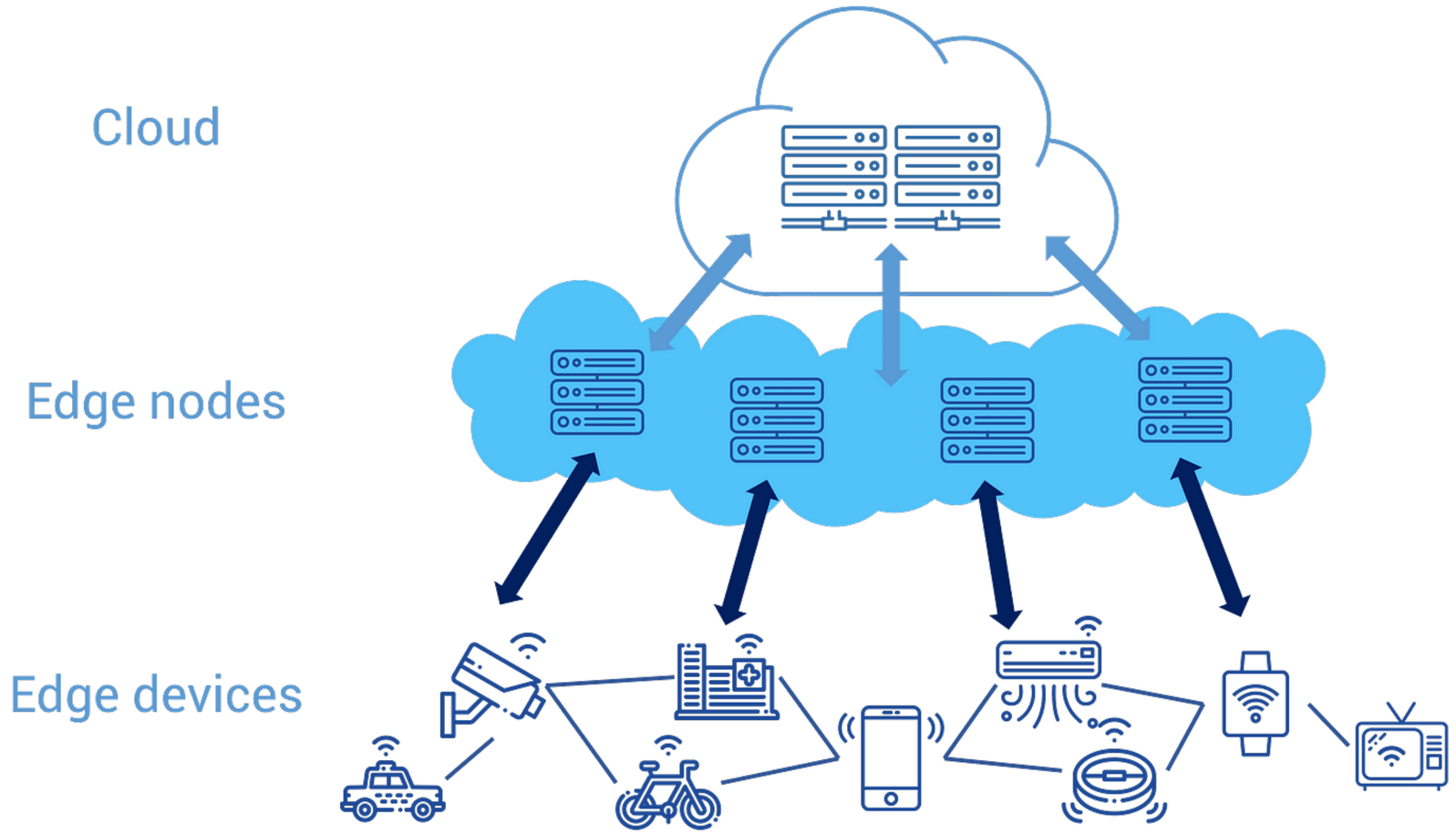
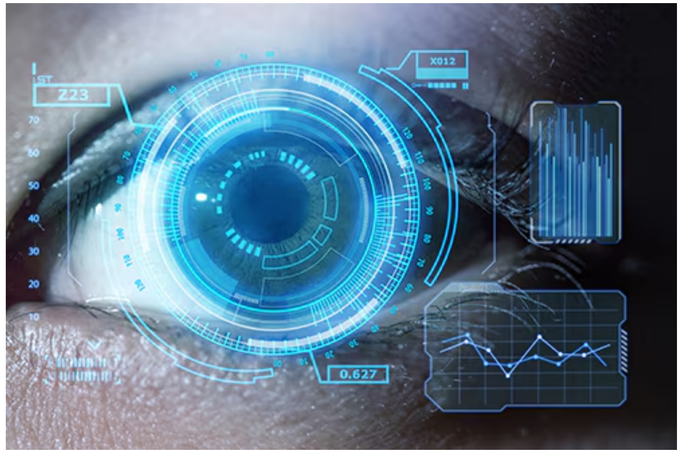


Image source: https://alibaba-cloud.medium.com/what-is-edge-computing-e2a00b713d92

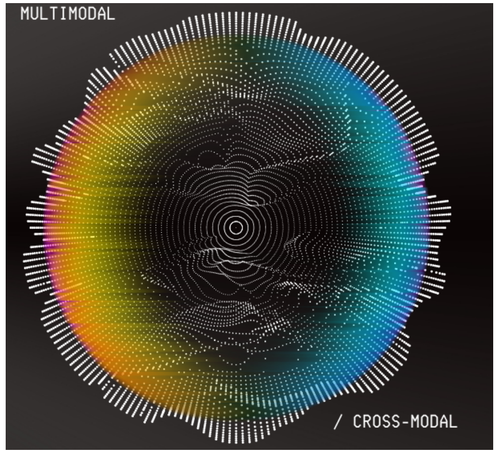
Current Landscape of AI

Deep Learning is Everywhere



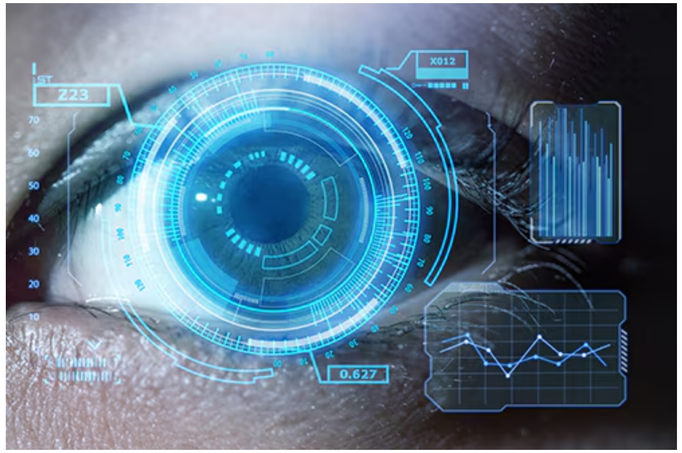
Vision

Language



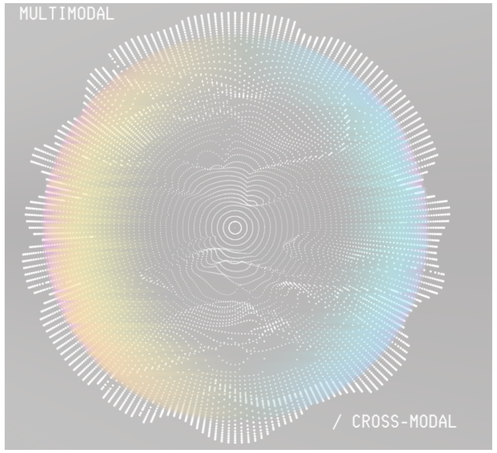
Multimodal

Deep Learning is Everywhere



Vision

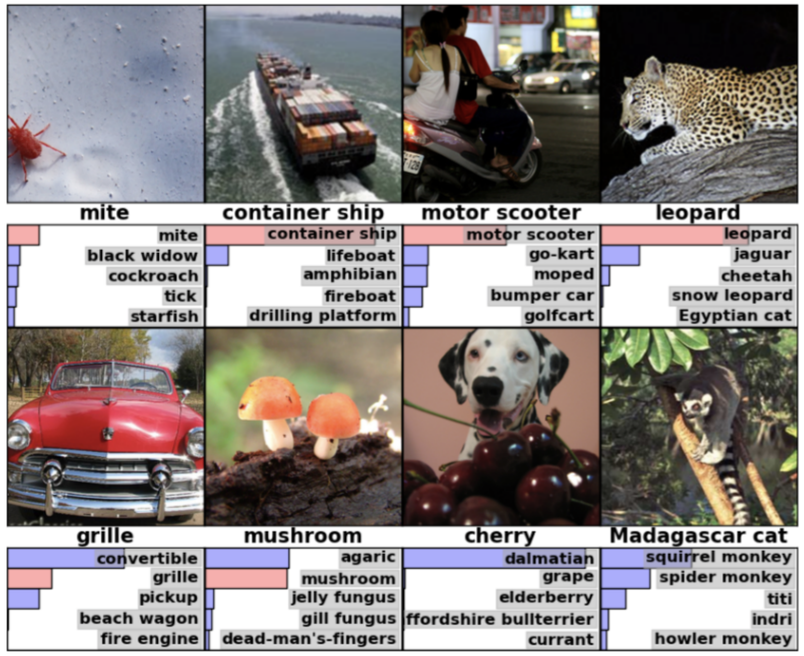
Language

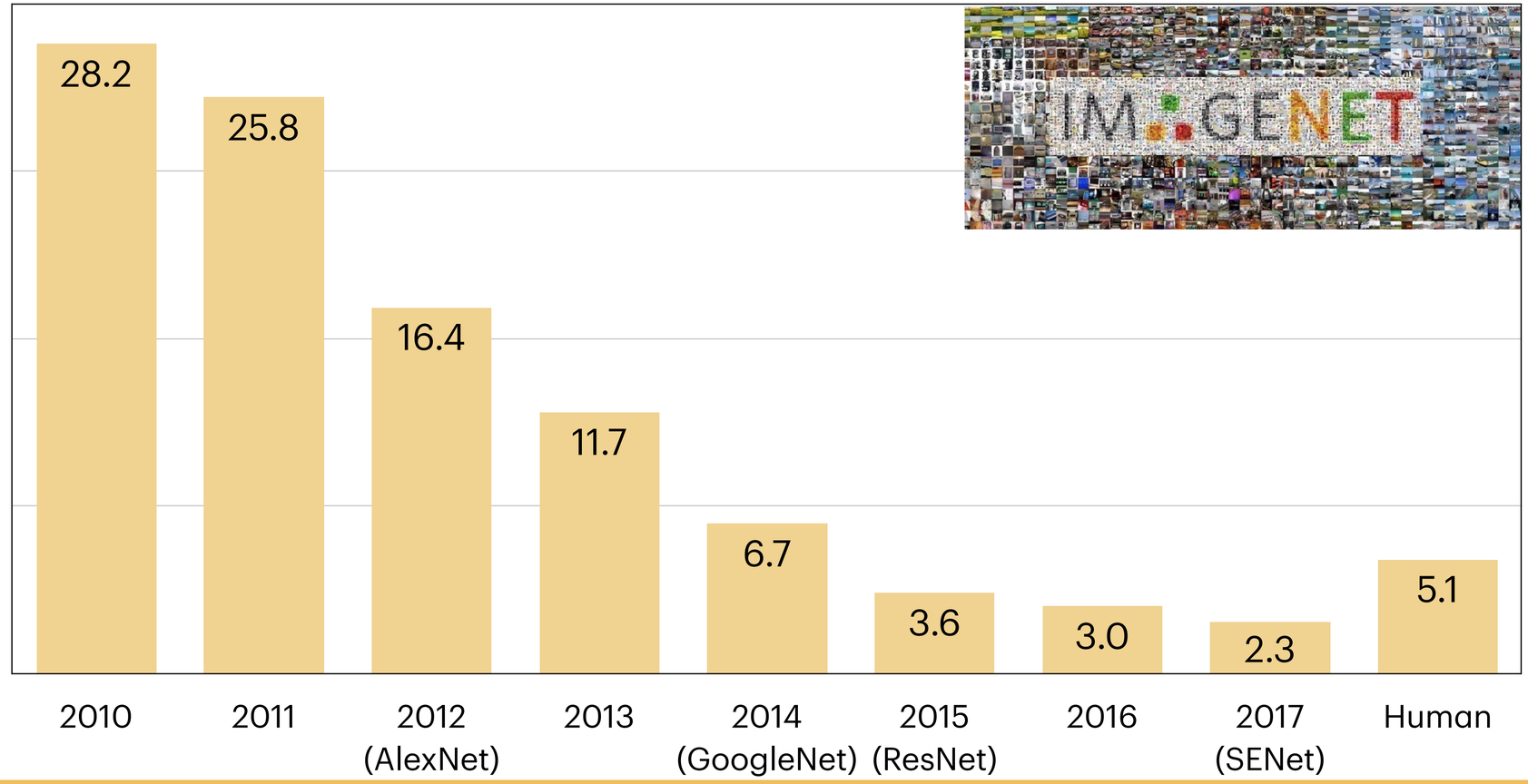


Multimodal

Deep Learning for Image Classificationœ

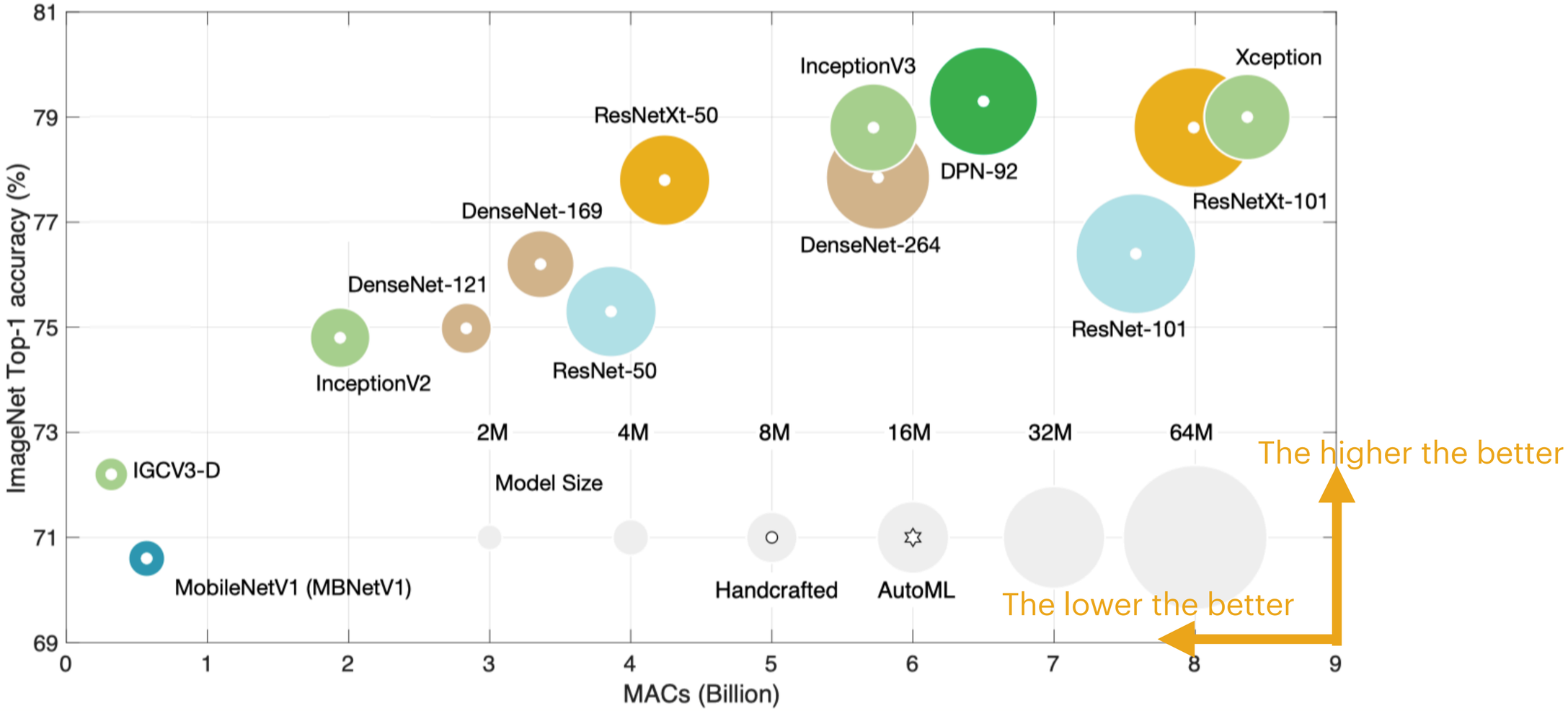
DNNs achieves super-human classification accuracy on ImageNet





Efficient Image Classification

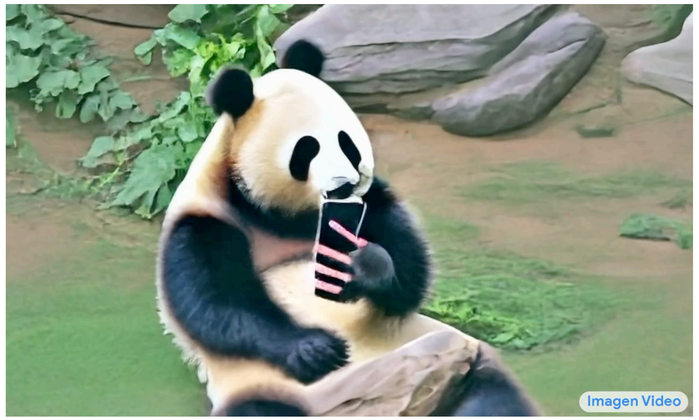
High accuracy comes at the cost of high computation



Model is Huge

Diffusion models create realistic images from a natural language description

An umbrella on top of a spoon.



A panda taking a selfie

A cat eating food out of a bowl, in the style of Van Gogh

- Video modeling is a harder task for which performance is not yet saturated at **5.6B model size**.

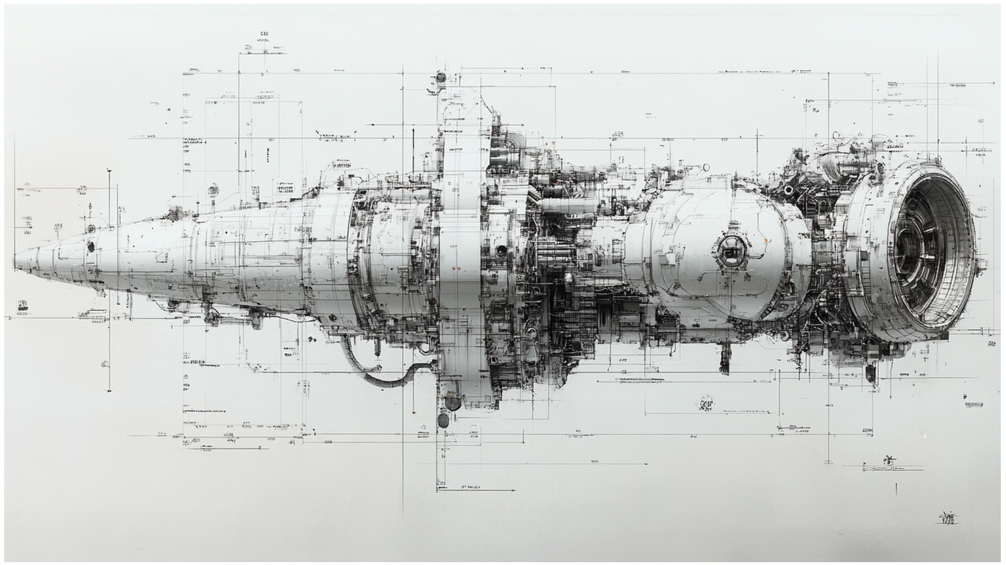
Imagen Video: https://imagen.research.google/video/

Training is Expensive

Diffusion models create realistic images from a natural language description

Healing Seven cats, stacked on top of each other, looking up, little a, childish and warm, Style cartoon, illustration, cute, healing

Exploded technical diagram of an enormous and mysterious futuristic machine. In the style of a blueprint.



Exploded technical diagram of an enormous and mysterious futuristic machine. In the style of a blueprint.

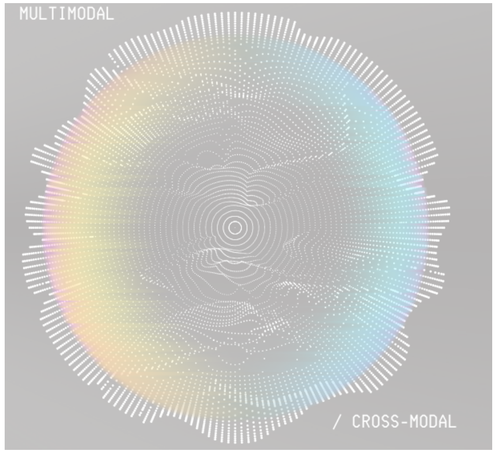
- Training stable diffusion costs **$600,000 (256 A100s, 150k hours)**

Midjourney: https://www.midjourney.com/

Deep Learning is Everywhere

Vision

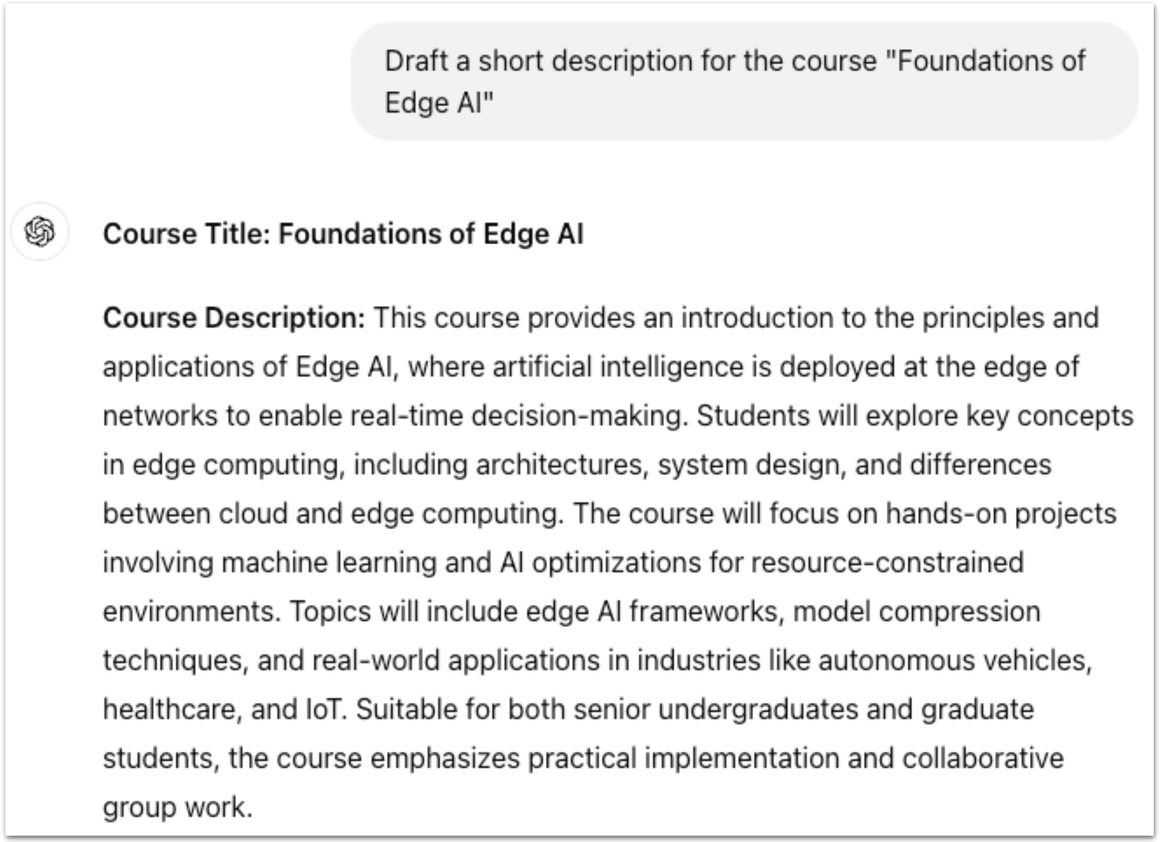
Language

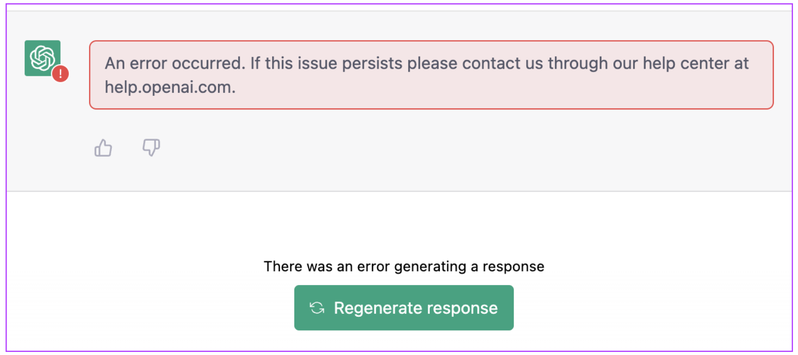


Multimodal

ChatGPT and Large Language Model

Large language models produce human-like text based on past conversations





- **Cloud-based** models are not always **stable**.

Code Generation

GitHub CoPilot can make meaningful coding suggestions based on context

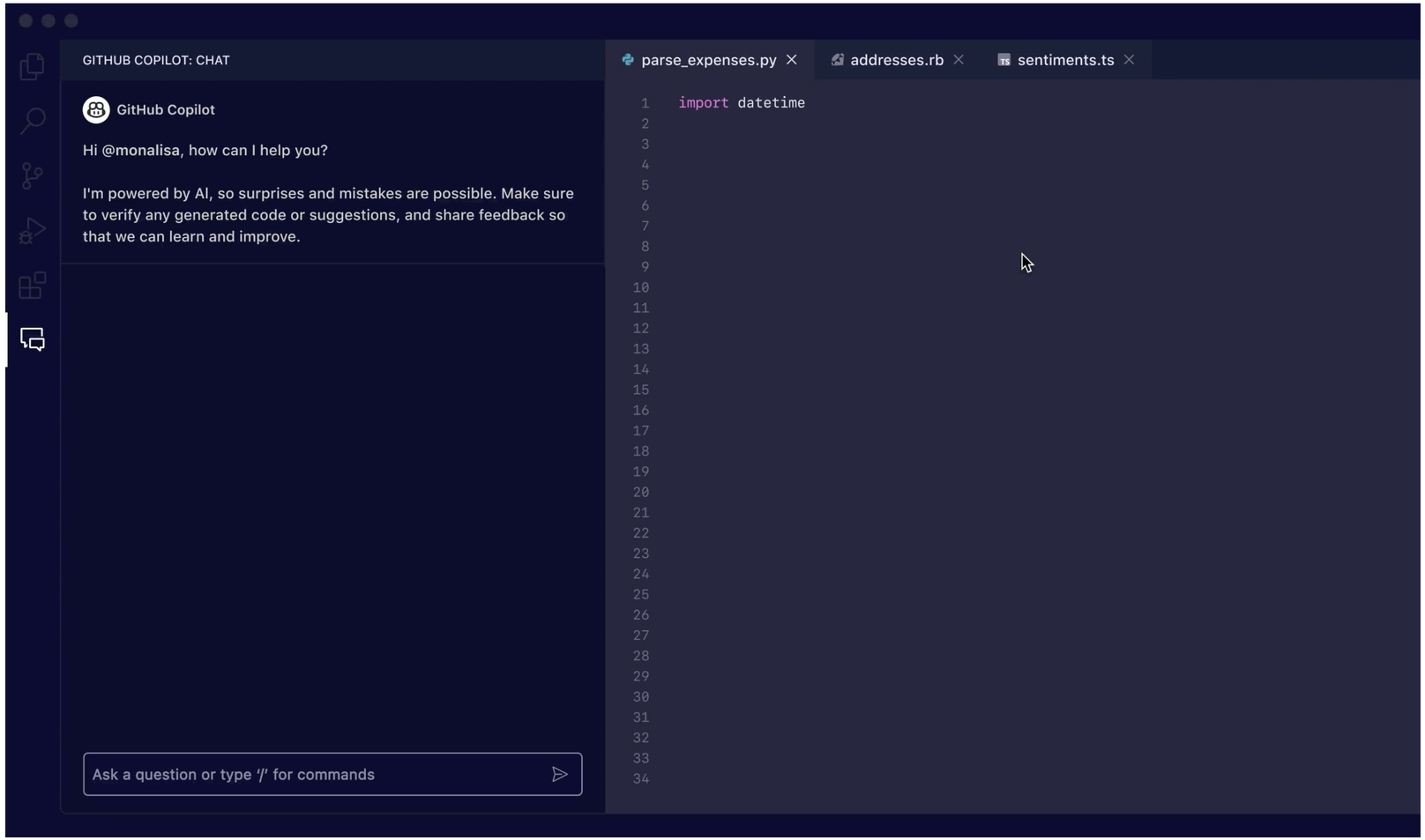
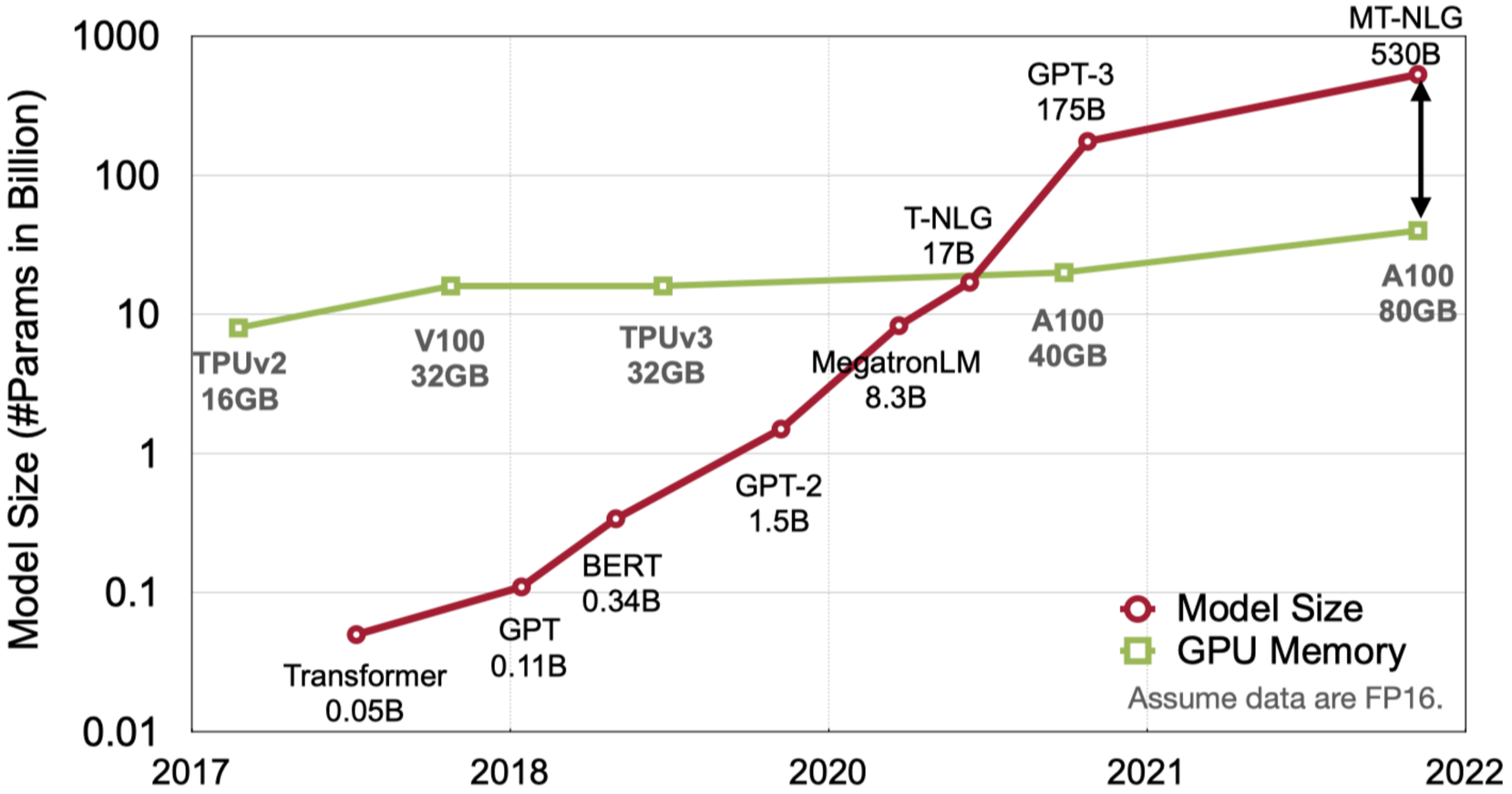
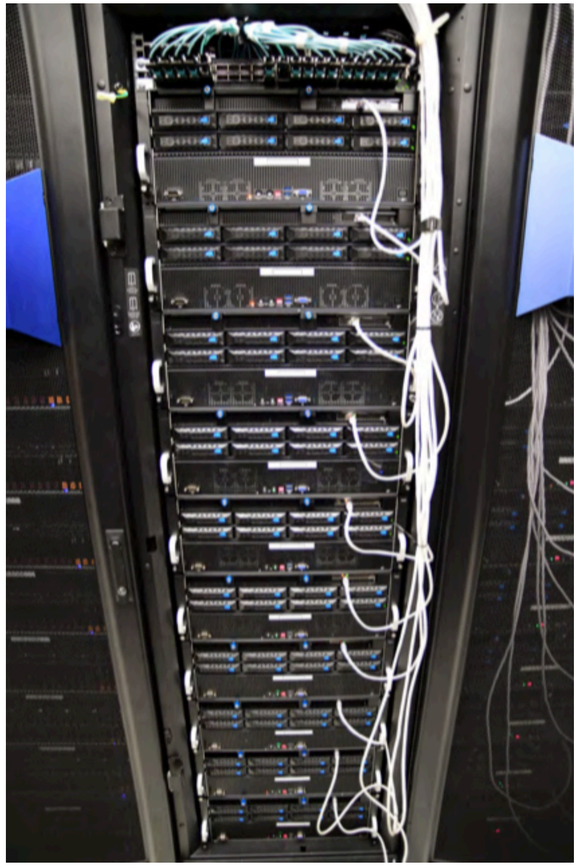


Image credit: https://github.com/features/copilot/

Large Language Models

Model size of language models is growing exponentially

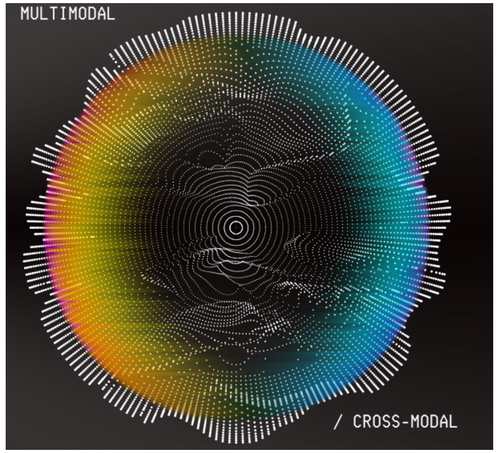




Deep Learning is Everywhere

Vision

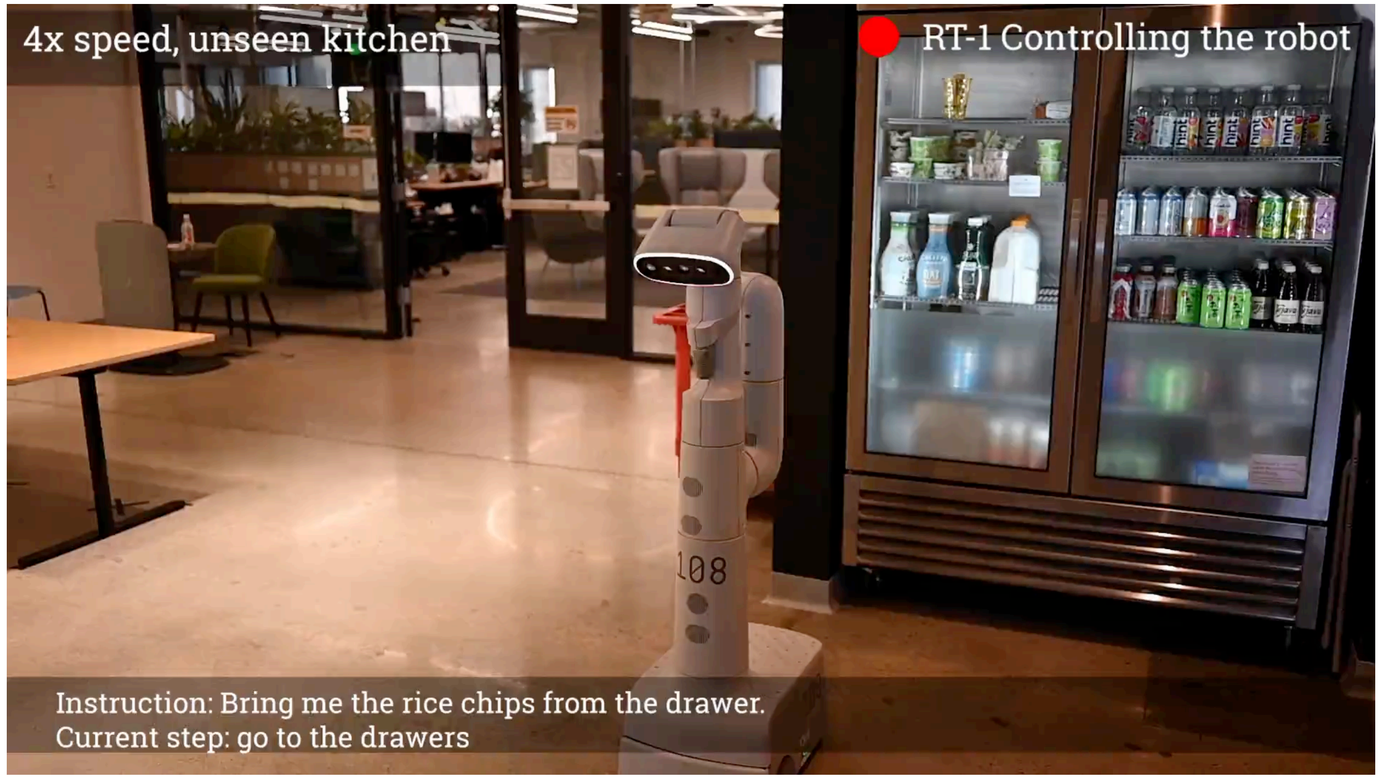
Language



Multimodal

Vision-Language-Action Models

Robotics transformers control robots based on language instructions

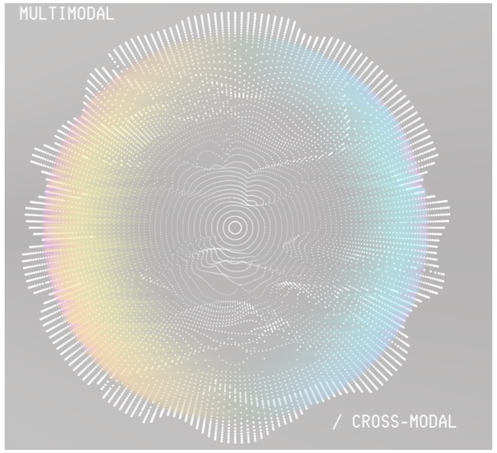


- Run at only **3Hz** due to the high computational cost and networking latency.

Deep Learning is Everywhere

Vision

Language

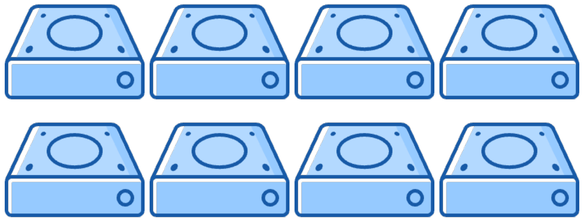


Multimodal

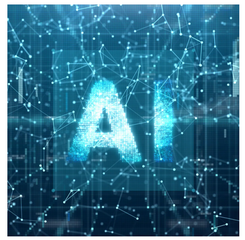
- **But they are computationally costly.**
- **How to make them light-weighted and fast?**

Current Landscape of AI

Big Computation, engineer and data



A lot of computation
A lot of carbon



Many engineers

A lot of data

Edge AI

Lightweight computation, engineer and data

Less computation
Less carbon

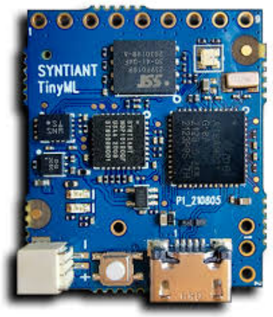
Edge AI

Fewer engineers

Less data

Edge AI Hardware

Edge AI devices still have a huge gap with cloud processors



	Cloud AI	Mobile AI	Tiny AI
Memory (Activation)	80GB	4GB	320KB
Storage (Weights)	~TB/PB	256GB	1MB

Course Overview

Course Overview

Edge Computing	Artificial Intelligence	Edge AI
Architectures	Basic Concepts	Efficient Inference
Applications	Popular Networks	Domain-Specific Optimization

Syllabus

Rules for this class

1. No mobile phone
2. No web surfing
3. No sitting in the last row
4. Email me if you cannot attend
5. You can leave early, and email me the reason later


ANY
QUESTIONS?