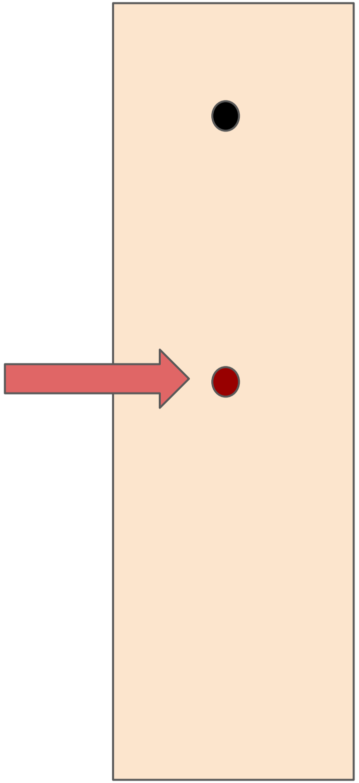
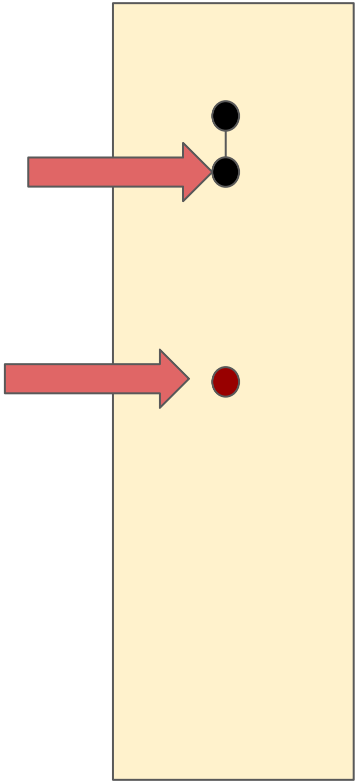
CSC 116 Federated Learning/Multi-modal Model

Gradient descent



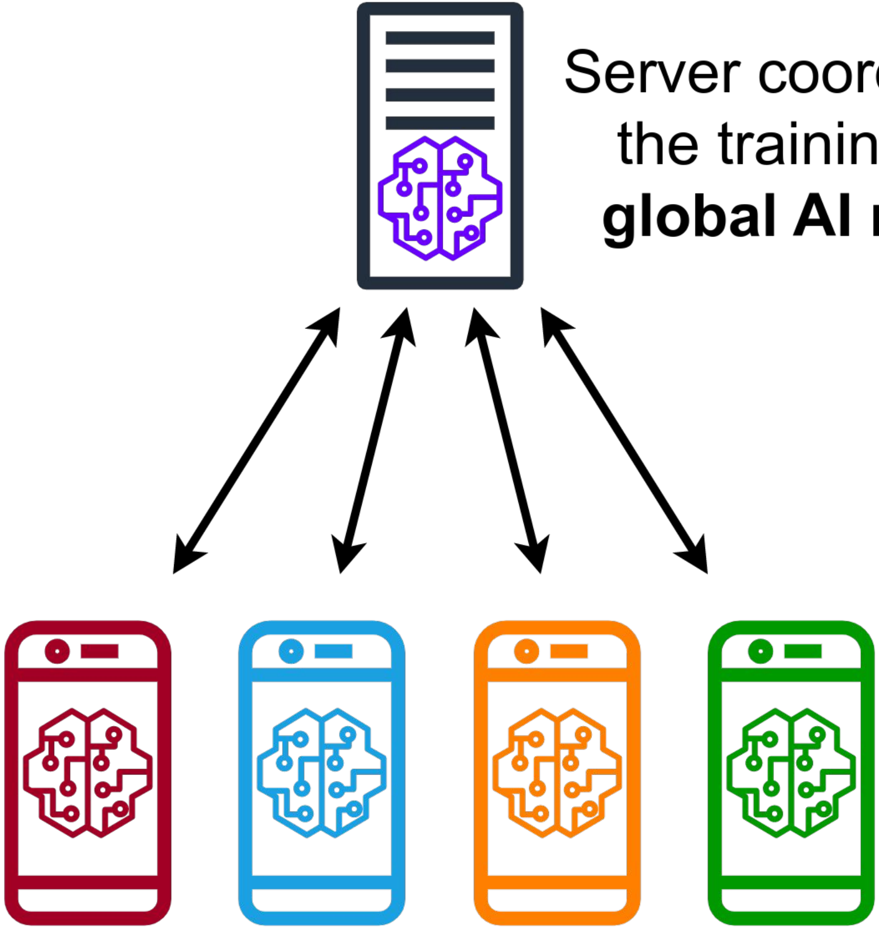


Move a little bit step

Introduction

What is “Federated Learning”?

- Also known as collaborative learning
- One method for large-scale distributed machine learning
- Allows distributed devices to collaboratively train a global ML model while maintaining distributed data privately.
  - A centralized parameter server
  - Nodes (workers or clients) participate



Server coordinating
the training of a
global AI model

Devices with
local AI models

Introduction

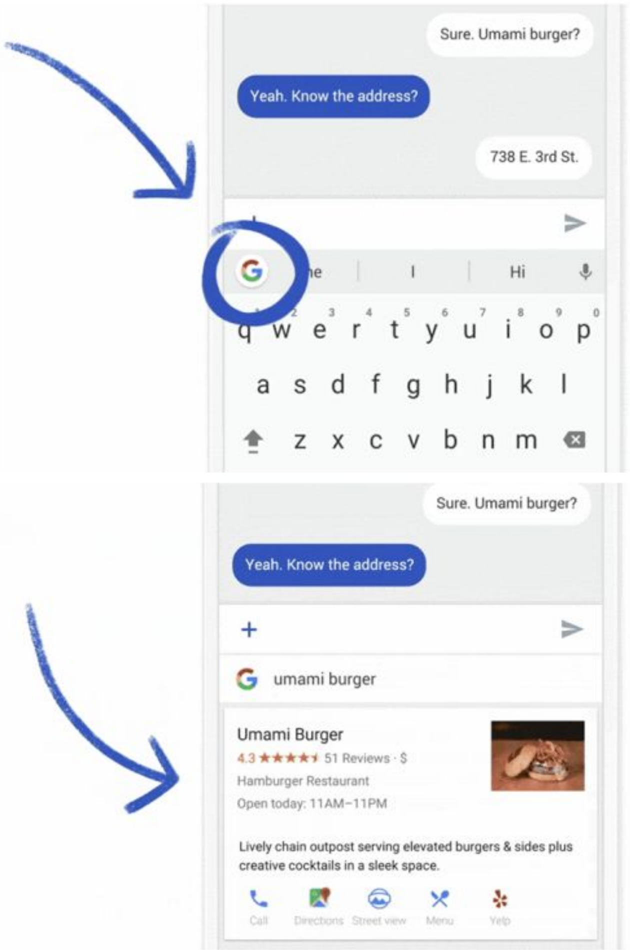
Typical federated training process:

- **Worker** preparation
- **Server** broadcast: workers download the current model weights and a training program.
- Worker computation: worker locally computes an update to the model.
- **Server aggregation**: server computes an aggregation of the worker updates.
- Model update: server locally updates the shared global model based on the aggregation. If training continued, server again broadcasts the global model and repeats the above steps.

Introduction

Federated Learning in Gboard on Android, the Google Keyboard.

- show suggested query based on your type style
- improve the next iteration of Gboard's query suggestion model
- rank photos that you prefer to look at, share, or delete



Sure. Umami burger?

Yeah. Know the address?

738 E. 3rd St.

Sure. Umami burger?

Yeah. Know the address?

umami burger

Umami Burger

4.3 ★★★★★ 51 Reviews · $

Hamburger Restaurant

Open today: 11AM–11PM

Lively chain outpost serving elevated burgers & sides plus creative cocktails in a sleek space.

Outline

- Introduction
  - Challenge 1 : Untrustworthy server
  - Challenge 2 : Untrustworthy clients

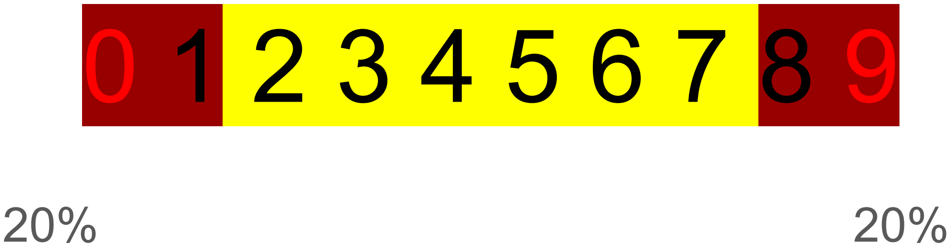
0 1 2 3 4 5 6 7 8 9

Average value = 45 / 10 = 4.5

0 1 2 3 4 5 6 7 8 9

Average value = 36 / 8= 4.5

Trim - Mean



0 1 2 3 4 5 6 7 8 9

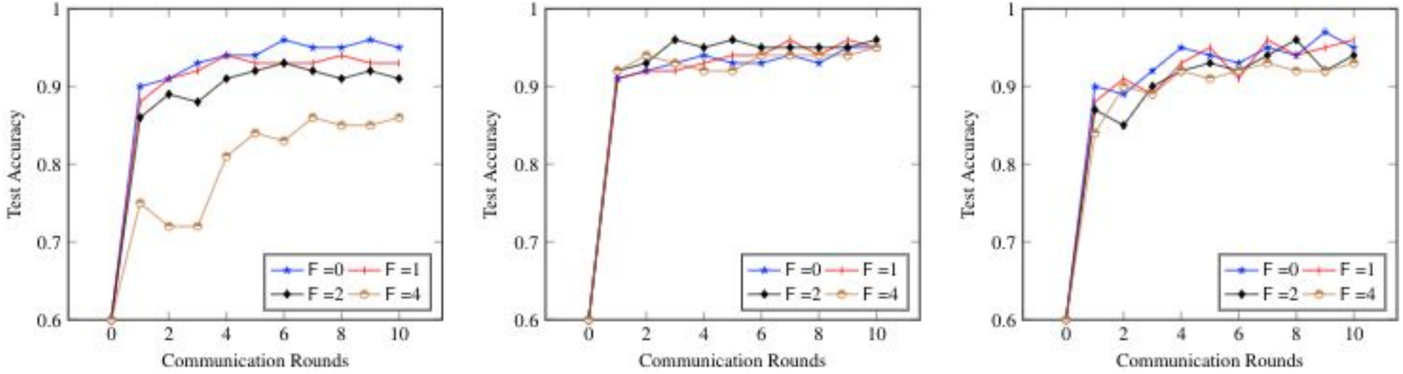
20% 20%

Reliable Federated Learning

Recap: Challenges of untrustworthy clients

- **Integrity of worker contributions** is hard to maintain. (worker bug, network packet loss, hardware crash, …)
  - The updated gradient is undesirable even lost.
- System may be injected with **poisoning contributions (gradients)** by malicious workers.
  - Such poisoning attacks mis-train the global model, reduce the global model’s accuracy, even “fully break” the training results.
  - The attack may appear in both sequential and asynchronous mode.

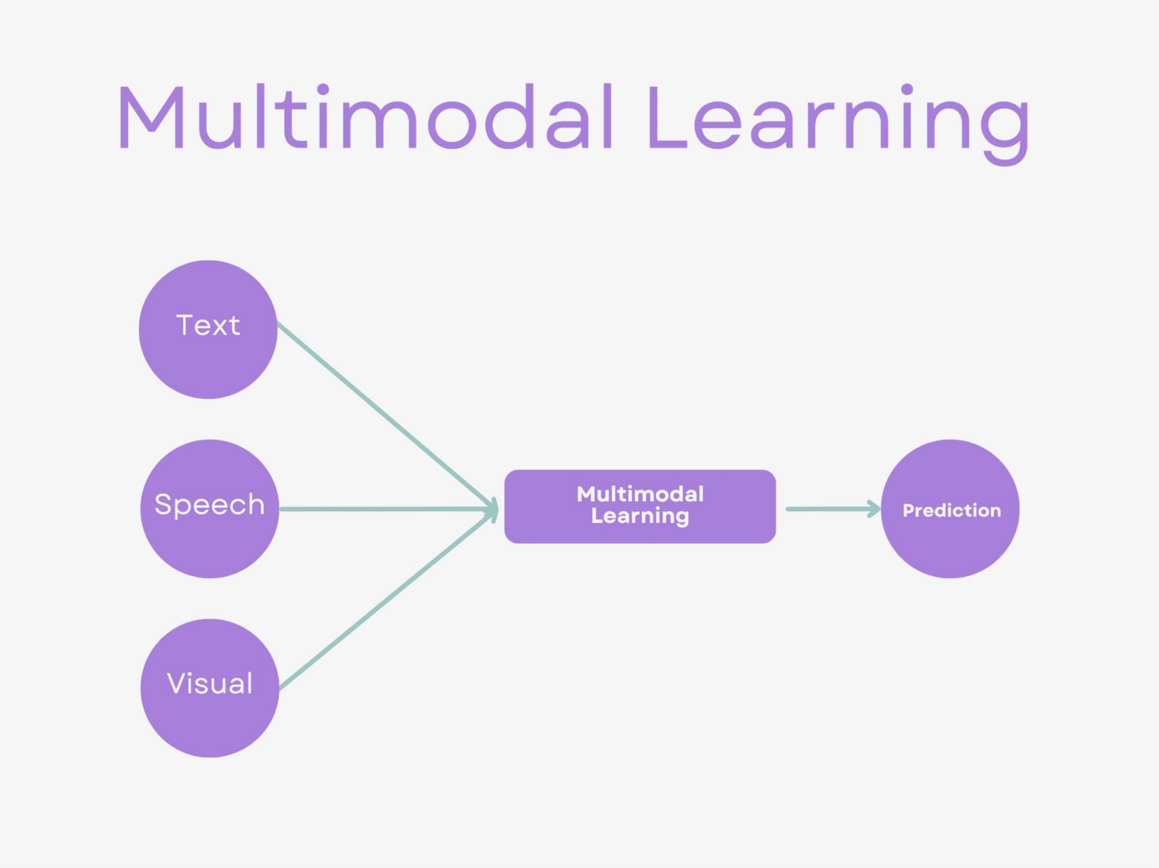
Evaluation: Poisoning Attack Defense



(a) FedL. (b) *SeqM*. (c) *AsyncM*.

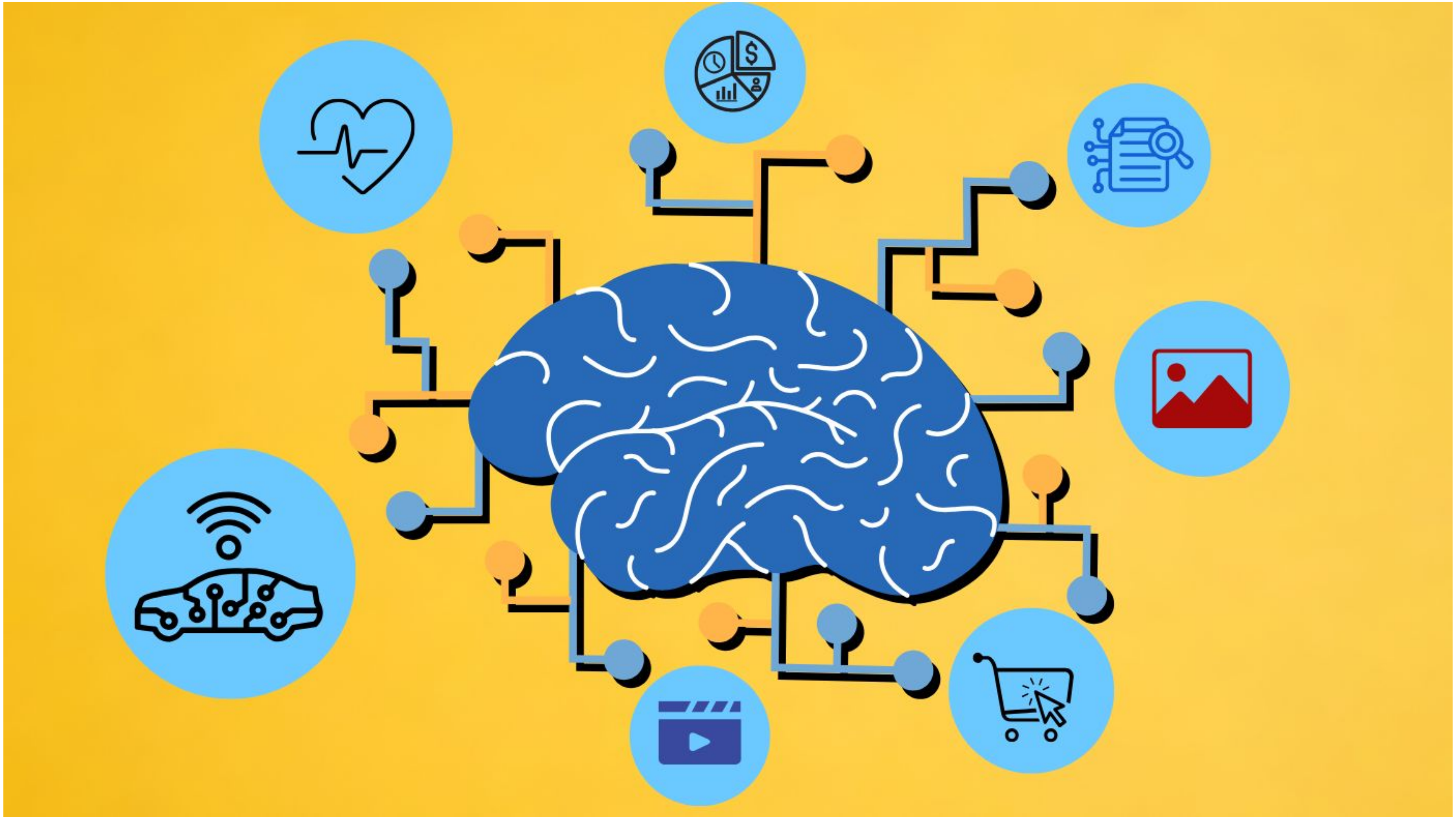
F: number of malicious workers (10 total workers)

Multi-Modal Model



GPT-4V, Google Gemini, Claude 3.5 Sonnet, and LLaVA

CLIP and Flamingo



Computer Architecture vs. Human Brain

The Human Brain: Parallel, Fuzzy, and Adaptive

The human brain isn’t really a “computer” — it’s a **dynamic pattern-recognition system**.

- **Massively parallel**: ~86 billion neurons, each connecting to thousands of others.
- **Fuzzy and probabilistic**: It doesn’t compute with exact numbers but with context, probability, and meaning.
- **Adaptive and self-modifying**: Neural connections change constantly (synaptic plasticity), so perception and consciousness *emerge* rather than being pre-programmed.

Perception, intuition, and semantic understanding.

Extracting meaning from incomplete or noisy data.

“Understanding” rather than “calculating.”

The Computer: Linear, Precise, and Closed

Linear — one instruction after another.

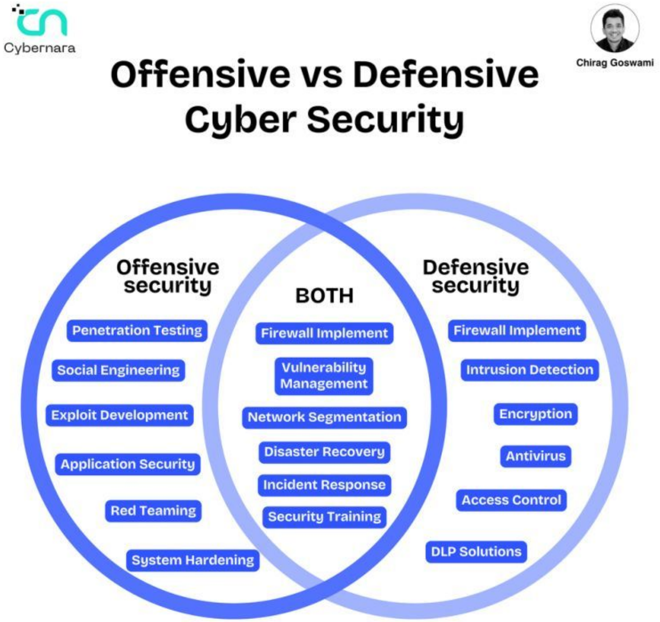
Deterministic — same inputs always yield same outputs.

Separated — data and computation are distinct.

1. Arithmetic precision
2. High-speed Computations

Offensive and defensive security are two sides of the same coin, with offensive security focused on proactively attacking and finding vulnerabilities, and defensive security concentrated on preventing, detecting, and responding to attacks.

Offensive security uses techniques like penetration testing and red teaming to simulate attacks and find weaknesses, while defensive security uses firewalls, antivirus software, and incident response plans to protect systems. The goal of both is the same—to improve an organization's overall security posture—but they use opposing mindsets and tactics to achieve it.



Why attack? Attacker motivations

Common motivations:

- Money (steal credit cards, ransom)
- Data theft (personal records, research)
- Disruption (take service offline)
- Steal Secrets
- Reputation or activism (political motives)
- Curiosity / challenge

Criminals want money; spies want secrets; some want to make a statement. Understanding motives helps defenders prioritize what attackers might try to get.

Attack methods

Recon / OSINT — gather public info

Phishing / Social engineering

Web application attacks (SQLi, XSS)

Credential attacks (weak passwords, brute force)

Misconfigurations & exposed services

Recon / OSINT — gather public info

What it is: collecting public info (websites, social media, DNS, job postings)

How attackers do it (steps):

1. **Search the target’s website and subdomains.**
2. **Look up employee names and emails on LinkedIn.**
3. **Find leaked credentials or code in public repos.**

   Map services by scanning public IPs (in lab/testing only).

Quick defense: reduce public exposure, use privacy settings, monitor for leaked data.

Phishing / Social Engineering

What it is: tricking people into revealing info or clicking malicious links

How attackers do it (steps):

1. **Use recon to find target names/emails.**
2. **Craft believable message (e.g., “IT: reset your password”).**
3. **Send email or message with link to fake login or an attachment.**
4. **Collect credentials or deliver malware if someone clicks.**

Quick defense: teach users, enable MFA, test with simulated phishing.

Web attacks (SQL Injection, XSS)

What they are:

- SQL Injection: give unexpected input to make the database run bad commands.
- XSS (Cross-Site Scripting): inject scripts that run in users’ browsers.

How attackers do it (steps):

1. **Find an input field (search box, form).**
2. **Send crafted input (like ' OR '1'='1 for SQLi).**
3. **Observe behavior — get database data or execute script in a victim’s browser.**

Credential attacks

What they are:

- **Guessing weak passwords, credential stuffing (re-using leaked passwords), and brute force.**

How attackers do it (steps):

1. **Use leaked credential lists or try common passwords.**
2. **Try many logins quickly or use automated tools**
3. **If successful, use account access to escalate.**

Quick defense: require strong passwords, rate-limit logins, enforce MFA, monitor for login anomalies.

Misconfigurations & exposed services

What it is: services left open or configured insecurely (open admin panels, S3 buckets, default passwords)

How attackers do it (steps):

1. **Scan for open ports and services.**
2. **Try default credentials or access public storage.**
3. **Use found access to move deeper.**

Quick defense: minimize exposed services, change defaults, review cloud storage permissions.

Thank you to everyone for completing
the Security and Privacy training.