Building Trust at the Frontier:
Privacy, Fairness, and Security in an AI-Driven Society

Antigoni Polychroniadou

J.P. Morgan AI research
AlgoCRYPT CoE

AlgoCRYPT CoE

AI Research

J.P.Morgan

Challenges in Privacy and AI

AI is rapidly advancing

But..

What if sensitive information is stored across different silos and cannot be exchanged?

Challenges in Privacy and AI

- AI is rapidly advancing
- But..
- What if there no computational resources and sensitive data cannot be released?

Challenges in Privacy and AI

AI is rapidly advancing

But..

What if an AI model produces unfair outcomes—and we lack reliable methods to detect such bias?

Scenario 1

You have different labeled credit card transaction datasets in different regions. You want to train a fraud detection model jointly on these datasets.

How would you approach this problem?

Scenario 2

You have labeled brain cancer imaging datasets from various regions and want to jointly train a detection model to improve accuracy and generalize across diverse populations.

How would you approach this problem?

AlgoCRYPT CoE Research Areas

Information Sharing

1. Liberate Sensitive Data Safely

2. Ensure Client & Employee Data Protection and Privacy

Secure Collaborative Computation

3. Unlock the Power of AI on Private Distributed Data

4. Enable Secure, Intelligent and Safe Electronic Markets

5. Privacy Preservation to Prevent Financial Crime

Secure Platform

6. Supporting the Transition to the Cloud

7. Enable Secure Decentralized Finance Systems (DeFi)

8. Provide Security in a Quantum World

AlgoCRYPT CoE Research Pillars

- Secure Multiparty Computation
- Differential Privacy
- Privacy Preserving Machine Learning
- Privacy Preserving Federated Learning
- Fully Homomorhic Encryption
- Zero Knowledge Proofs
- Quantum Cryptography
- Privacy Preserving Blockchain

Agenda

Group Privacy

Privacy Preserving Federated Learning

Pairwise Privacy

Encrypted LLMs

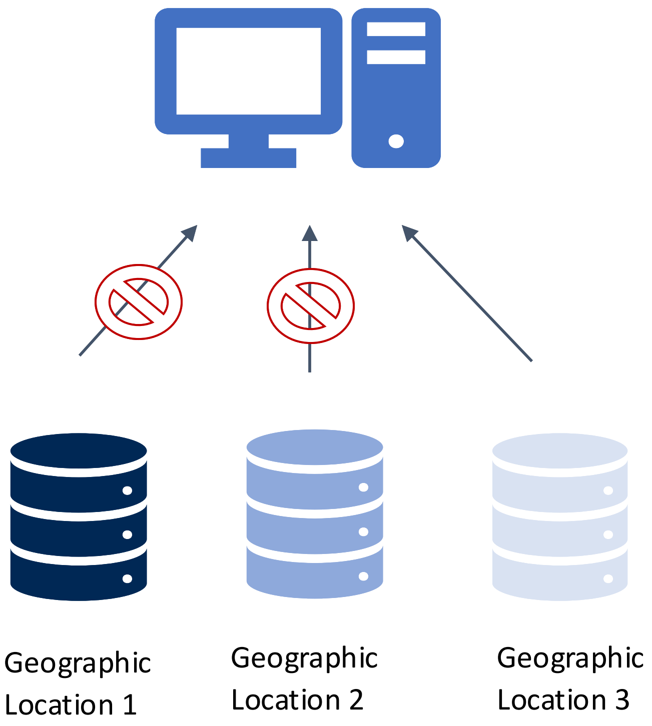
Checking AI Model Fairness

Individual Privacy

Biometric Authentication:

Issues with Centralized AI Model Development

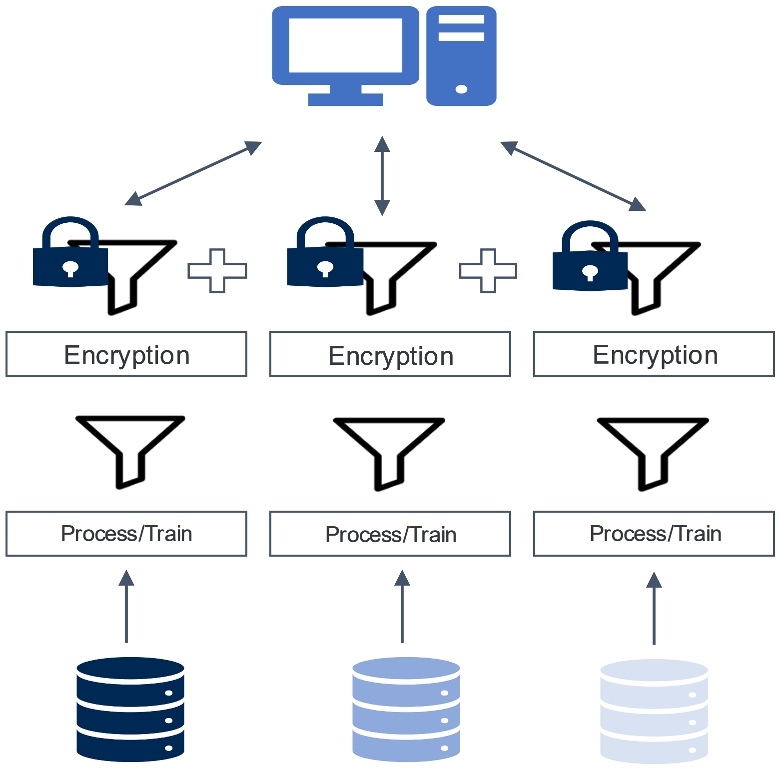
- Utility of Sensitive Data Limited by:
  - Data silos
  - Data controls
  - Regulatory Constraints



Example: Government Mandated Data Localization

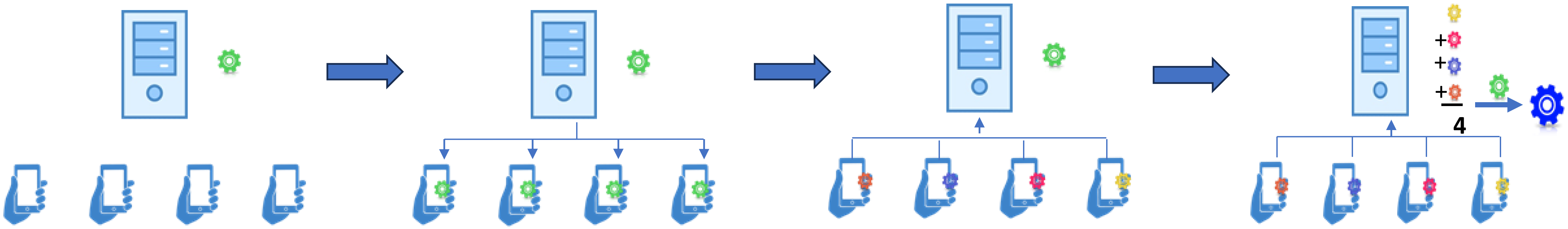
Federated Learning: Global Models, Local Data

- Data never leaves the original secure environment
- Strong cryptographic techniques used to protect sensitive data
- Only privacy-preserving aggregate models revealed



Federated Learning Process

- Users jointly learn shared ML model, managed by centralized server; data stays local



1. Server trains a global model (Optional)

2. Users deploy the global model
3. Each user locally trains

4. Users upload versions of their local trained model updates

5. Server aggregates users' encrypted updates and obtains decrypted, privacy-preserving next global model

Repeats steps 2 to 5, until the model converges

J.P.Morgan

AlgoCRYPT CoE

AI Research

Federated Learning: Bridging Data Silos in Healthcare

Use Case	What They Did	Outcomes
Mount Sinai & other hospitals (COVID-19 outcomes prediction)	Used EHR data from 5 hospitals to build models predicting COVID-19 progression, comparing federated vs local models.	Federated models outperformed or matched local models; showed better generalizability across sites.
Penn Medicine & 10 hospitals (Brain tumor imaging)	Trained models to distinguish cancerous vs non-cancerous MRI scans using federated learning without sharing raw images.	Performance was almost as good as centralized models (99% of centralized quality), with strong privacy preservation.
Oxford NHS partnership (COVID-19 screening in emergency depts)	Deployed a 'full-stack federated learning' setup using inexpensive hardware so hospitals could join easily; model trained across multiple NHS trusts.	Federated model performed significantly better (~27.6% improvement) compared with using each hospital’s data alone. Generalized well across sites.
FeTS / Penn + Intel (Brain tumor boundary detection, GBM patients)	Used data from 71 institutions across six continents, employing federated learning with privacy protections (SGX, etc.).	Improved tumor detection by ~33% over local models. Demonstrates that even very large, multi-institution setups can succeed.
Pediatric brain tumors (FL-PedBrain)	Classification & segmentation across 19 international sites, using federated learning.	Small drop compared to centralized training but much better generalization to out-of-network sites; big benefit for rare pediatric tumor domain.
MS lesion segmentation (3 hospitals, MRI)	Used federated learning and a U-Net model to segment lesions in Multiple Sclerosis; each hospital kept its data.	The federated model achieved acceptable performance (Dice ~0.66-0.80) on hold-out sets, showing feasibility in neuroimaging.

Differential Privacy (DP)

Main privacy metric for Federated Learning

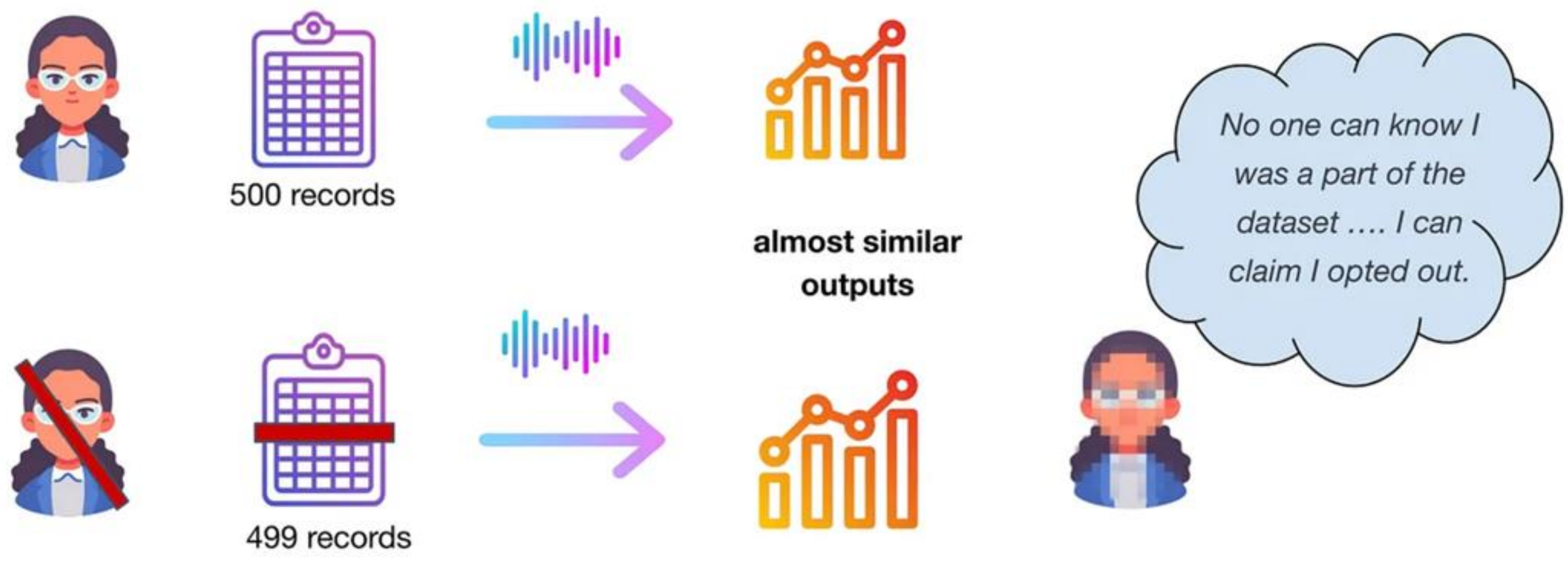
No privacy

Differential privacy

Encryption

Differential Privacy (DP)

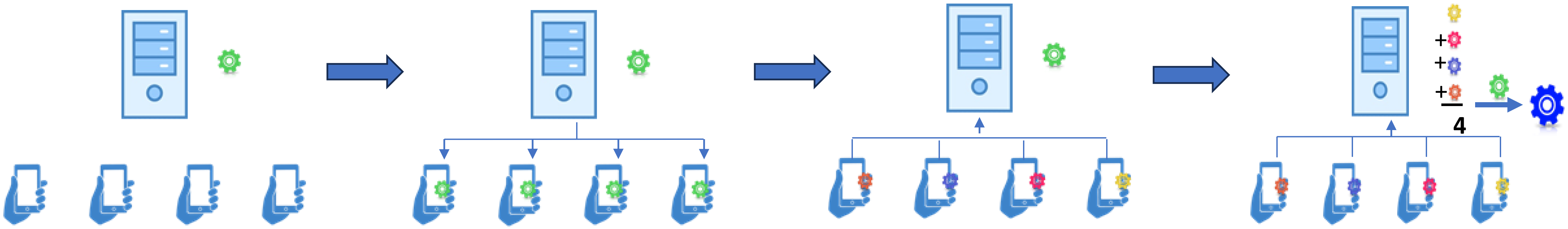
Main privacy metric for Federated Learning



Differential privacy adds a calculated amount of noise to hide each individual's contribution to data.

Federated Learning Process

- Users jointly learn shared ML model, managed by centralized server; data stays local



1. Server trains a global model (Optional)

2. Users deploy the global model
3. Each user locally trains

4. Users upload "DP" versions of their local trained model updates

5. Server aggregates users' encrypted updates and obtains decrypted, privacy-preserving next global model

Repeats steps 2 to 5, until the model converges

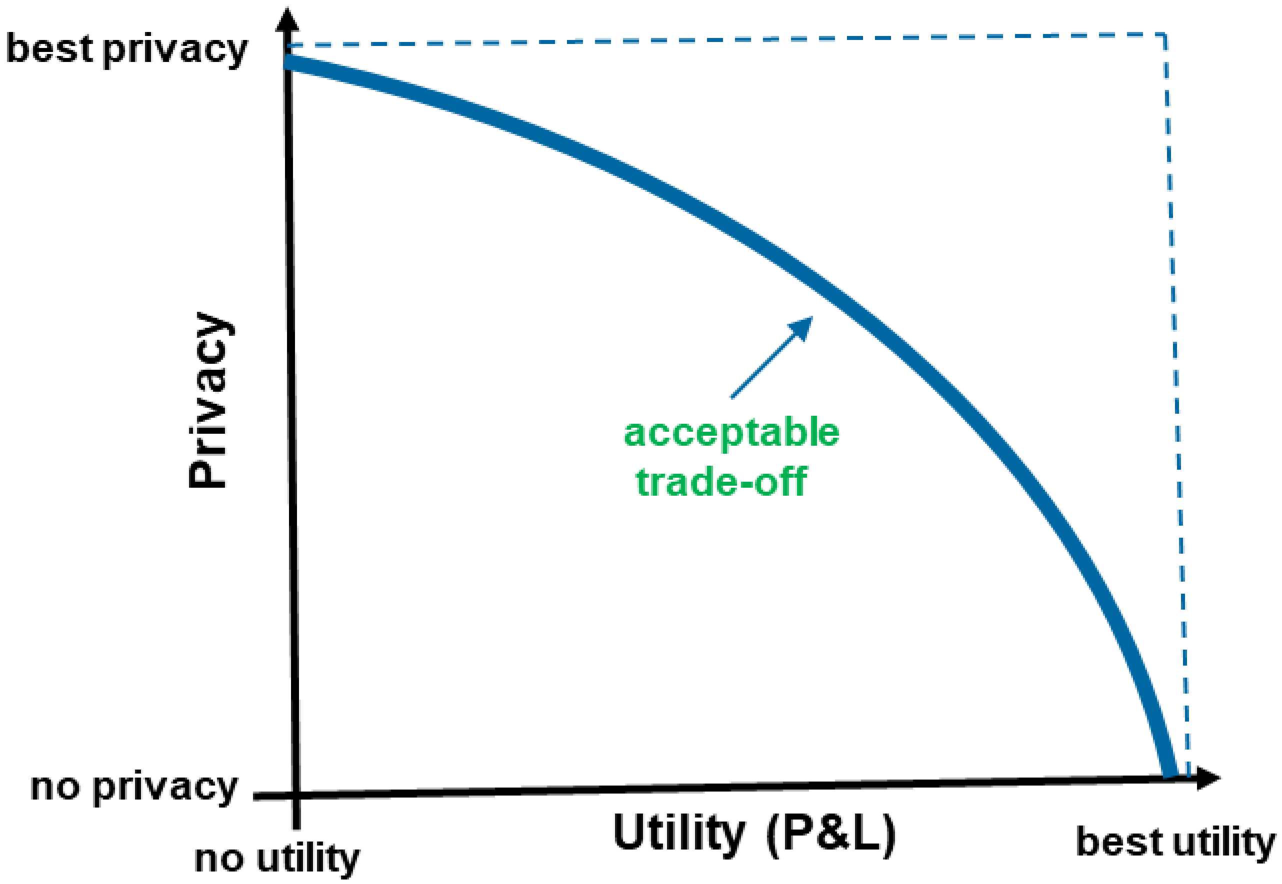
J.P.Morgan

AlgoCRYPT CoE

AI Research

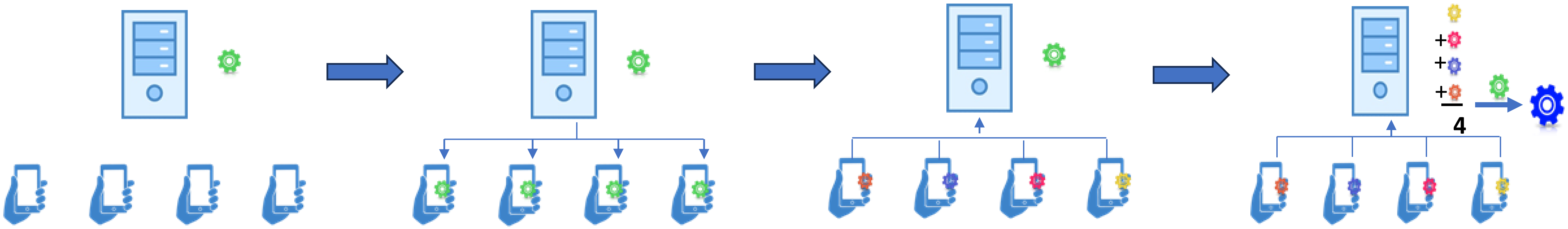
Privacy-Utility Tradeoff

- More noise → more privacy
- Less noise → more accuracy



Federated Learning Process

- Users jointly learn shared ML model, managed by centralized server; data stays local



1. Server trains a global model (Optional)

2. Users deploy the global model
3. Each user locally trains

4. Users upload "DP & Encrypted" versions of their local trained model updates

5. Server aggregates users' encrypted updates and obtains decrypted, privacy-preserving next global model

Repeats steps 2 to 5, until the model converges

Features of Federated Learning

Data Availability

- More data → better models
- Empirically shown for credit card fraud model

Privacy

- Strong security guarantees for sensitive data
- Data stays local

Dropout Resilience

- Ensures model integrity despite participant dropout.

Data Validation

- Data validation prevents poisoning attacks, ensuring integrity.

Quantum Protection

- Safeguards against quantum threat.

State-of-the-Art Protocols

- Outperform other systems (e.g., from Google)
- Papers at top conferences (e.g., ICML, CRYPTO, IEEE S&P; see Bibliography)

Protocol Applications

- **Our FL protocols are used by Amazon!**

CRYPTO 2025

Multiple papers on FL published at top tier conferences (MicroSecAgg, Flamingo, Lerna, Armandillo, DMM, OPA)

J.P.Morgan

Agenda

USENIX Security 2023

AAMAS 2024

Group Privacy

Privacy Preserving Federated Learning
(Products: Prime Match, Atlas-X)

Pairwise Privacy

Encrypted LLMs
Checking AI Model Fairness

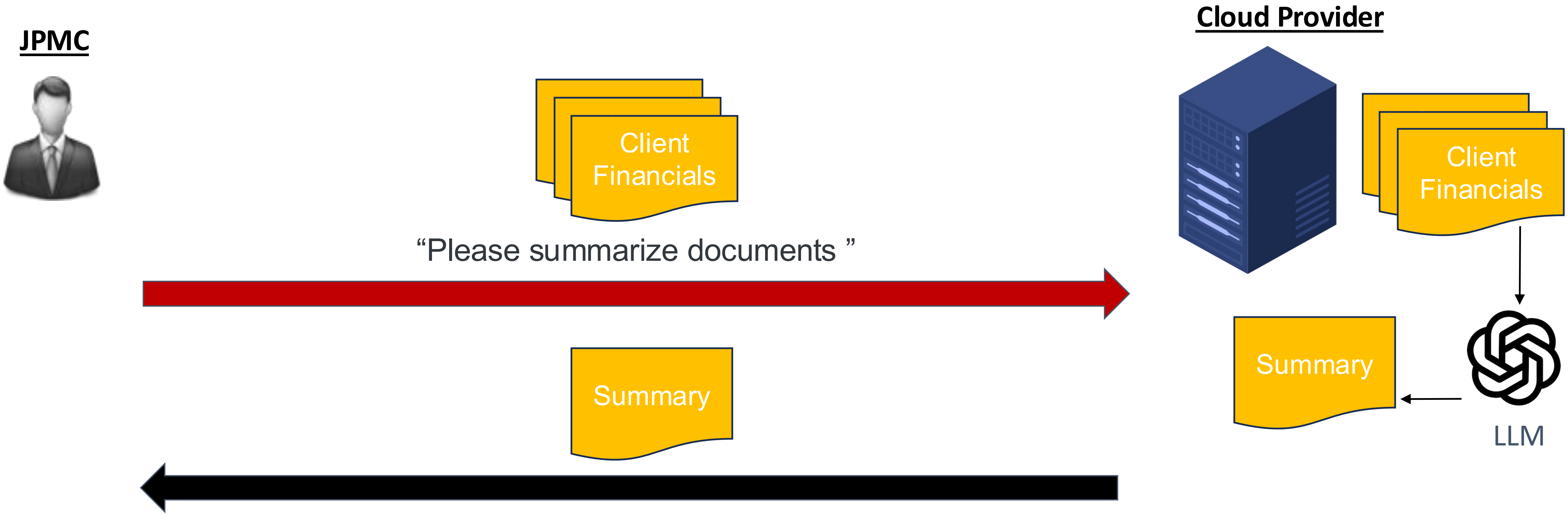
Individual Privacy

Biometric Authentication:

EncryptedLLM: LLM Evaluation on Encrypted Data

Examples
AWS
Azure
Google Cloud

Current LLM-based Applications: Send queries and data off-premises



Risk: Cloud provider could retain data, embed it into models, or fail to delete it **as promised**

Need: Remove the Need for Trust

J.P.Morgan

AlgoCRYPT CoE
AI Research

EncryptedLLM: LLM Evaluation on Encrypted Data

Examples
AWS
Azure
Google Cloud

Current LLM-based Applications: Send queries and data off-premises

1ST Attempt: Use traditional Encryption

Problem: LLMs do not understand traditional encryption

EncryptedLLM: LLM Evaluation on Encrypted Data

Examples
AWS
Azure
Google Cloud

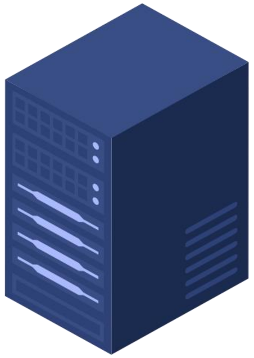
Current LLM-based Applications: Send queries and data off-premises

Cloud Provider

JPMC

Real example on GPT with an encrypted query:

DA WG48VeuHYOiE5kPuq6vDH8ymv7tUzZNy



DA WG48VeuHYOiE5kPuq6vDH8ymv7tUzZNy

It appears that you've entered a string of random characters: ~~WG48VeuHYOiE5kPuq6vDH8ymv7tUzZNy.~~ If you have a specific question or need assistance with something, please provide more context or information, and I'll be happy to help.

LLM

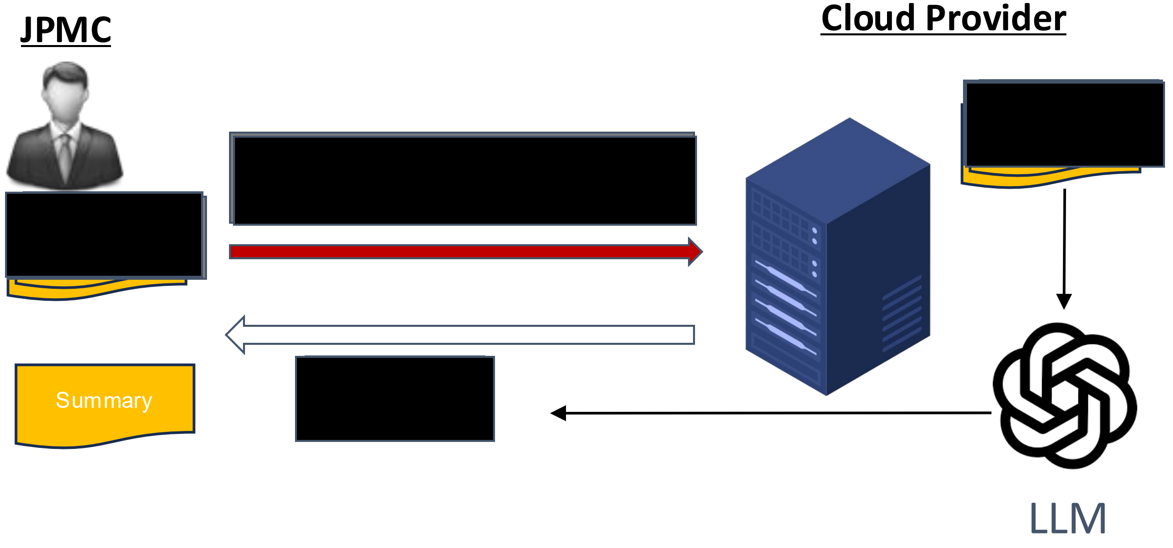
Problem: LLMs do not understand traditional encryption

J.P.Morgan

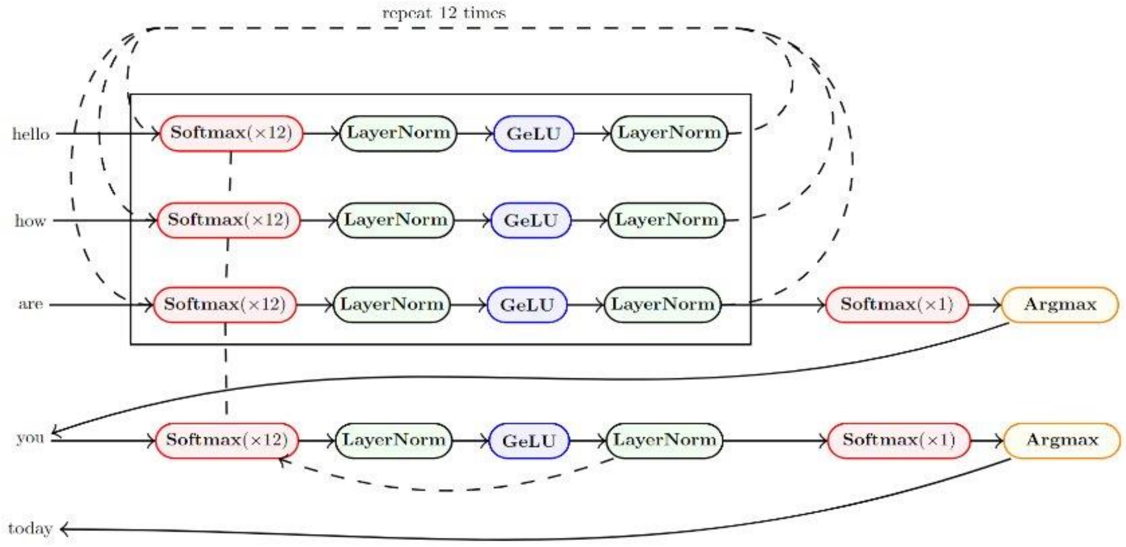
AlgoCRYPT CoE
AI Research

EncryptedLLM: LLM Evaluation on Encrypted Data

EncryptedLLM-based Applications: Send encrypted queries and data off-premises

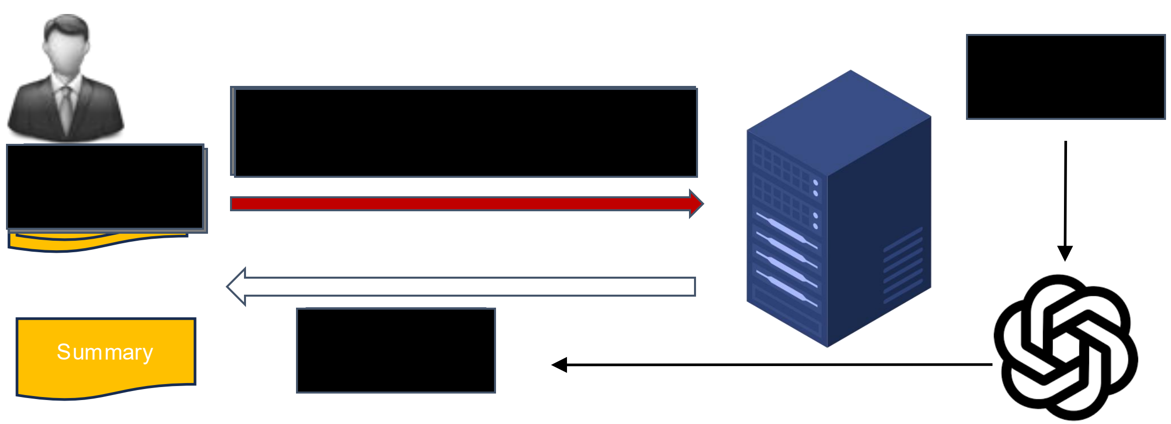


LLM complex computations:

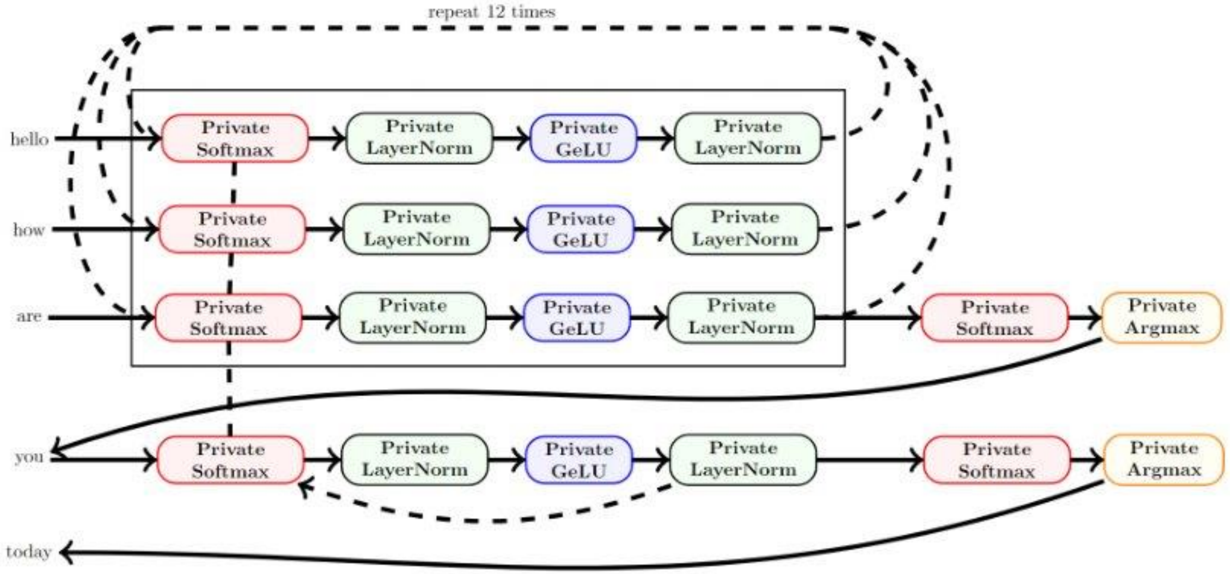


EncryptedLLM: LLM Evaluation on Encrypted Data

EncryptedLLM-based Applications: Send encrypte
queries and data off-premises



New LLM with new functions which operate on encrypted inputs:



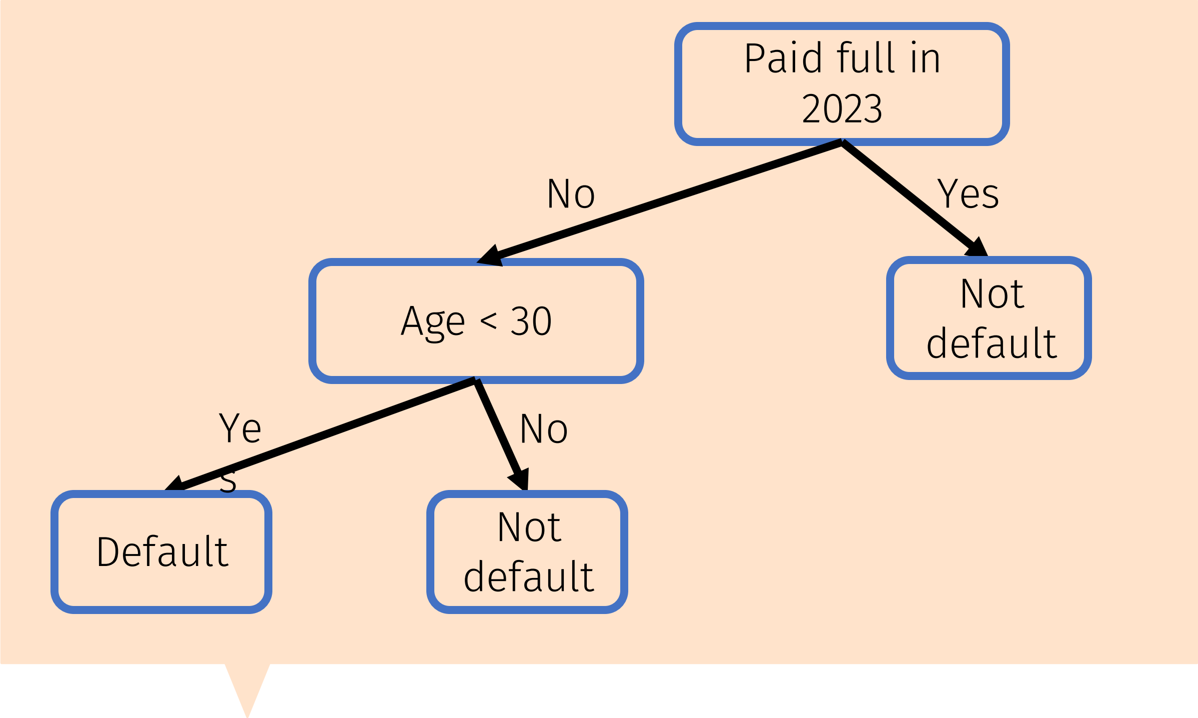
ICML 2025

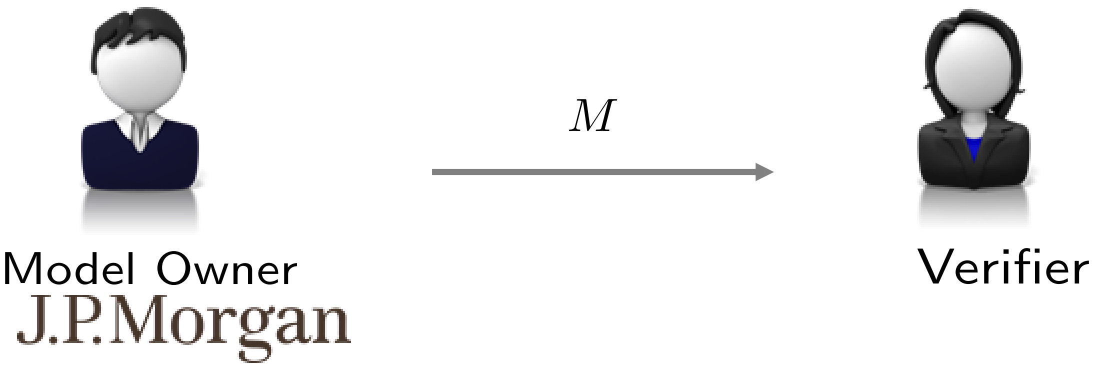
1st EncryptedLLM solution: **Invented a new LLM and a new advanced encryption** that can **support complex computations** over the data *while the data remains encrypted,* while also striving to support financial applications (such as summarization).

J.P.Morgan

AlgoCRYPT CoE
AI Research

Motivation: Checking Compliance of Proprietary Models

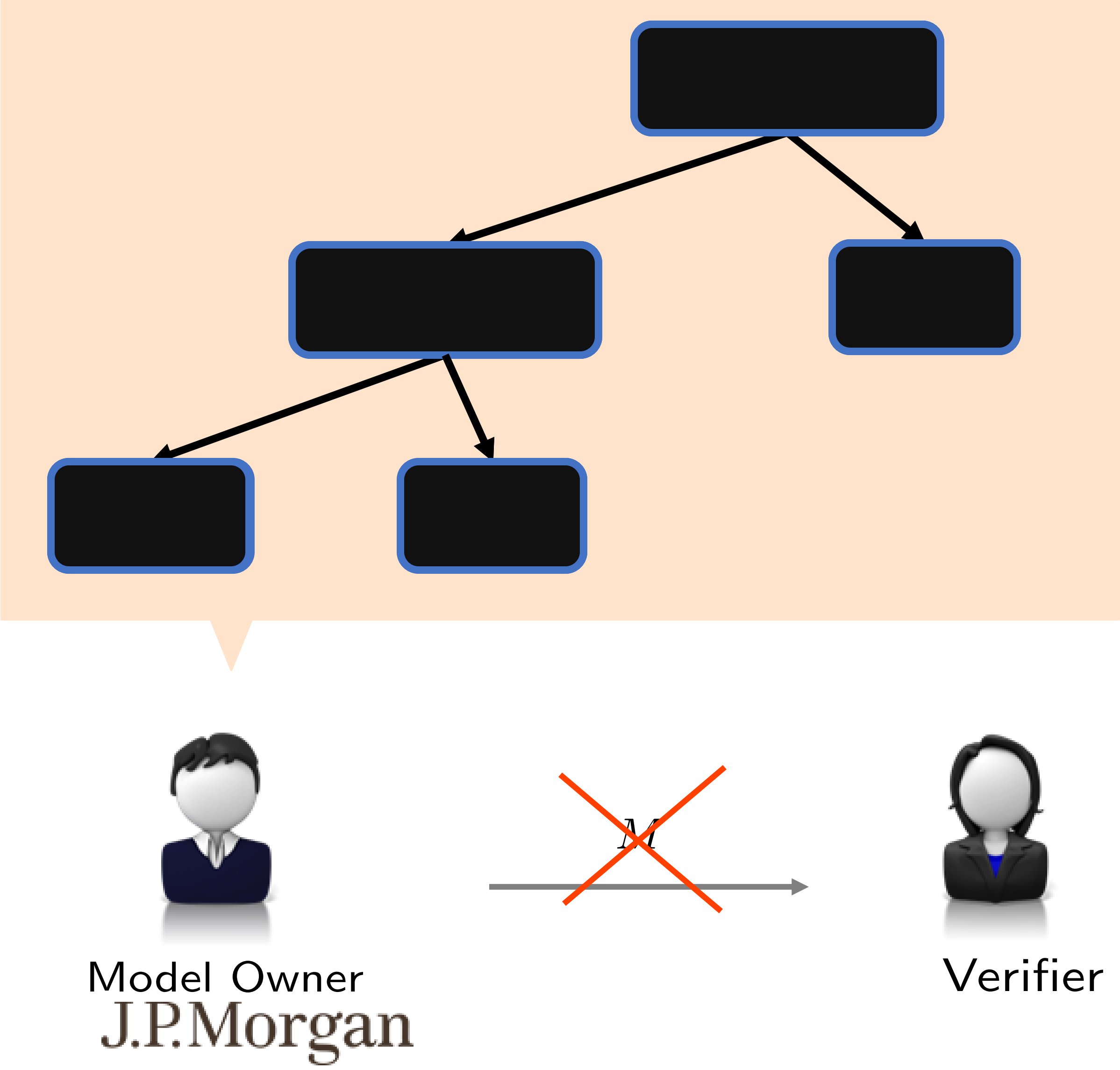




Naive Approach

- Office of fair lending statement prohibits discrimination in lending based on “red attributes” such as race, gender, age, religion, etc.
- Verifier examines the received model M in the clear
- **Problem**: If the model M is proprietary, Model Owner is not supposed to reveal M!

Motivation: Checking Compliance of Proprietary Models



Naive Approach

- Office of fair lending statement prohibits discrimination in lending based on “red attributes” such as race, gender, age, religion, etc.
- Verifier examines the received model M in the clear
- **Problem**: If the model M is proprietary, Model Owner is not supposed to reveal M!

Our Question

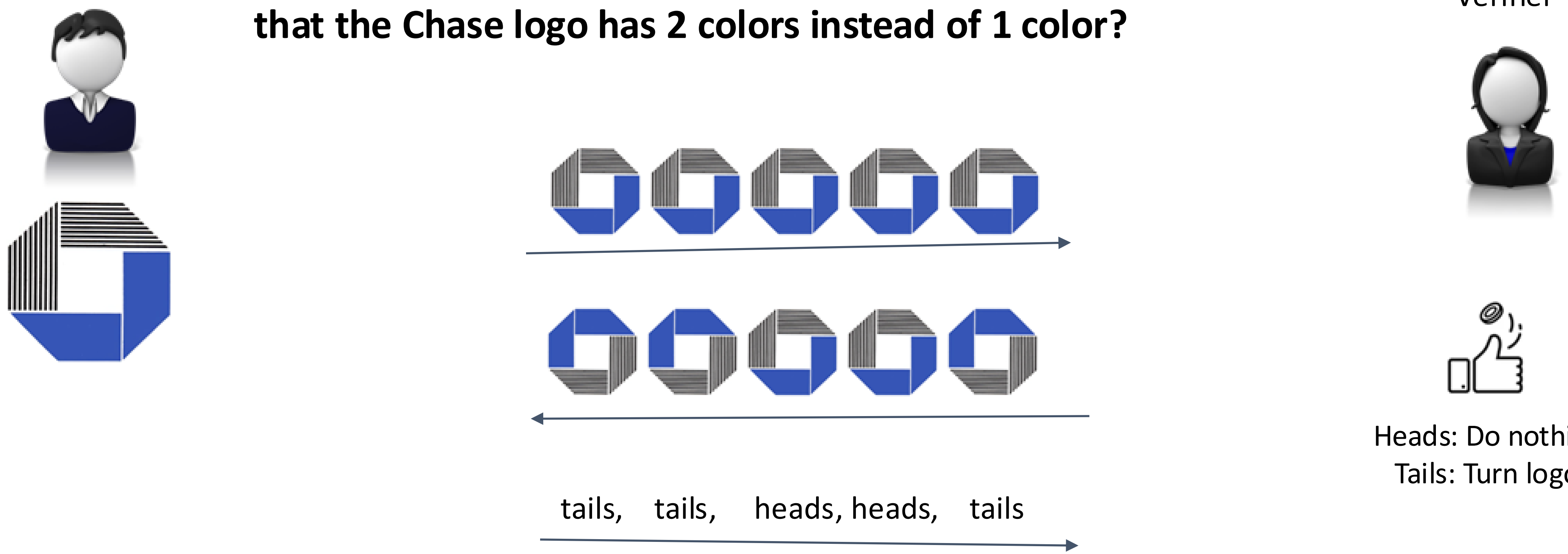
- Can we design a cryptographically secure protocol allowing Verifier to check that the proprietary model M is compliant with certain policies, while protecting confidentiality of M?

Zero Knowledge

Prover

Toy Example: Can the prover prove to colorblind Verifier that the Chase logo has 2 colors instead of 1 color?

Verifier

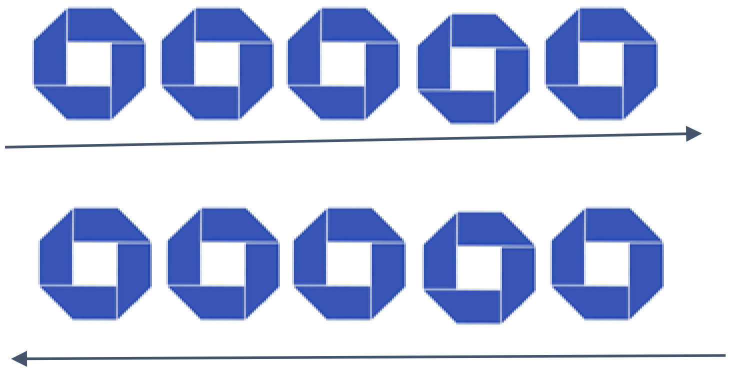


Heads: Do nothing
Tails: Turn logo 180 to the right

Zero Knowledge

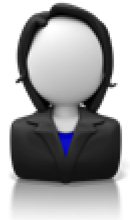
Prover

Toy Example: Can the prover prove to colorblind Verifier that the Chase logo has 2 colors instead of 1 color?



heads, tails, tails, heads, tails

Verifier



Heads: Do nothing
Tails: Turn logo 180 to the right

Agenda

Group Privacy

Privacy Preserving Federated Learning

Pairwise Privacy

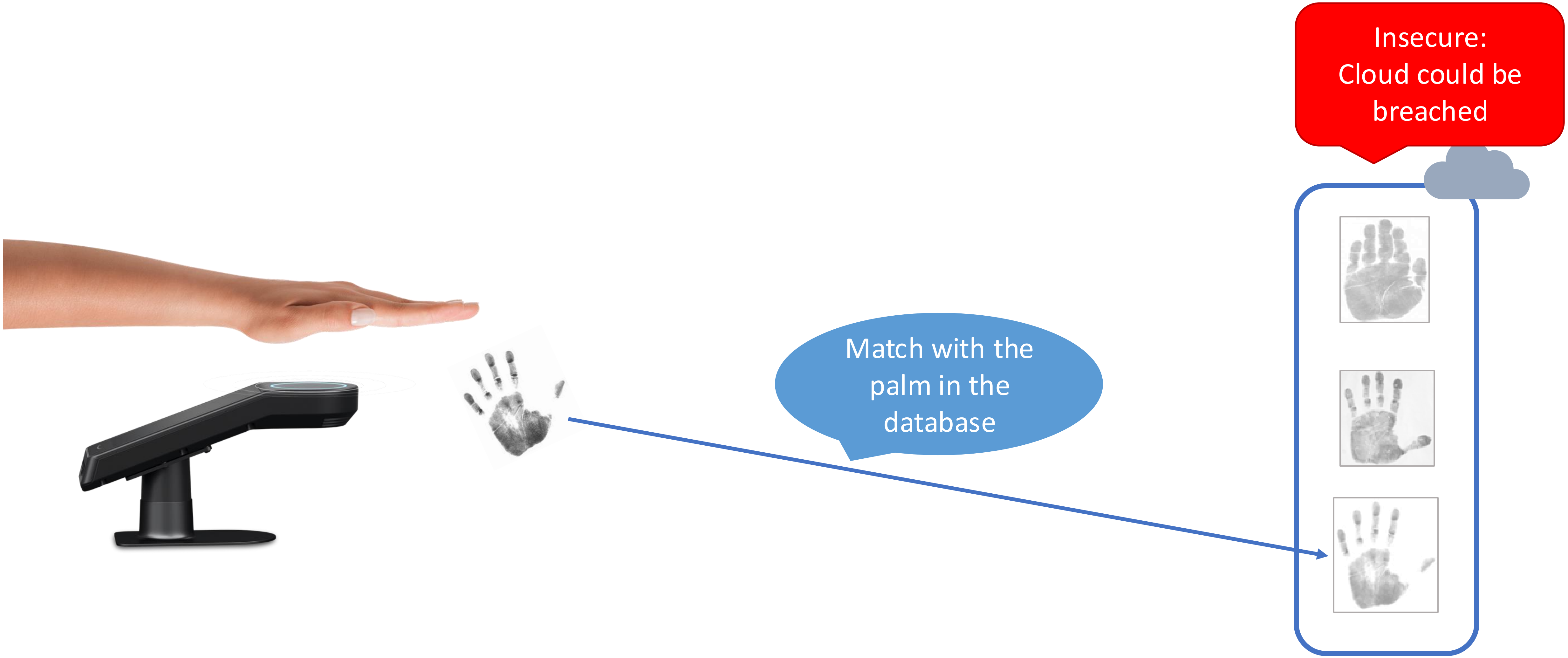
Encrypted LLMs

Checking AI Model Fairness

Individual Privacy

Biometric Authentication:

How Current *Insecure* Biometric Algorithms Work



Security Concerns of Existing Biometric Solutions

Vulnerability of State-of-the-Art Face Recognition Models to Template Inversion Attack

Publisher: IEEE Cite This PDF

23andMe Data Breach Settlement: $30M Deal Covers Millions Whose Info Was Stolen

Forbes

KEY FACTS

- The new app lets users sign up for Amazon One through their phones instead of having to visit a physical location to do so, requiring them to take photos of their palms for enrollment.
- Amazon One uses palms and their underlying vein structure to create a palm signature, which is created with the help of generative AI and verified by Amazon One scanners for things like retail purchases, age verification, entry and more.
- nners, once limited to Amazon stores, can now hundreds of Whole Foods locations, some ores and third-party locations including ts and fitness centers.
- Palm and vein images are encrypted and sent to the Amazon Web Service cloud, which Amazon says is "highly restricted to select AWS employees with specialized expertise."
- Albert Cahn, founder of the digital privacy advocacy Surveillance Technology Oversight Project, told Bloomberg he was skeptical of the trade-off between the convenience of biometric-based services and the user data required to run

Detour: Secure Password Authentication

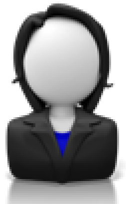
- First, how does password authentication work?

Account names	Passwords
Alice	“alice2003”
Bob	“jpmc@24”
Carol	“#000AI”

Server database

I don’t want to reveal my password to the server or have the server store it

“alice2003”



Alice

Detour: Secure Password Authentication

- First, how does password authentication work?

Use Chash!
Cryptographic
Hash Algorithm

Account names

Hash values

Alice	CHash(“alice2003”) =	47c2
Bob	CHash(“jpmc@24”) =	d604
Carol	CHash(“#000AI”) =	e921

Server database

I don’t want to reveal my password to the server or have the server store it

CHash(“alice2003”) = 47c2

Alice

Detour: Secure Password Authentication

- Why biometric setting creates a challenge? Biometric data are noisy

For a same person, *each scan is different* (even though they look similar)

Detour: Secure Password Authentication

- Why biometric setting creates a challenge? Biometric data are noisy

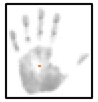
Account names

Hash values

Alice	CHash(“  ”) =	47c2
Bob	CHash(“  ”) =	d604
Carol	CHash(“  ”) =	e921

Server database

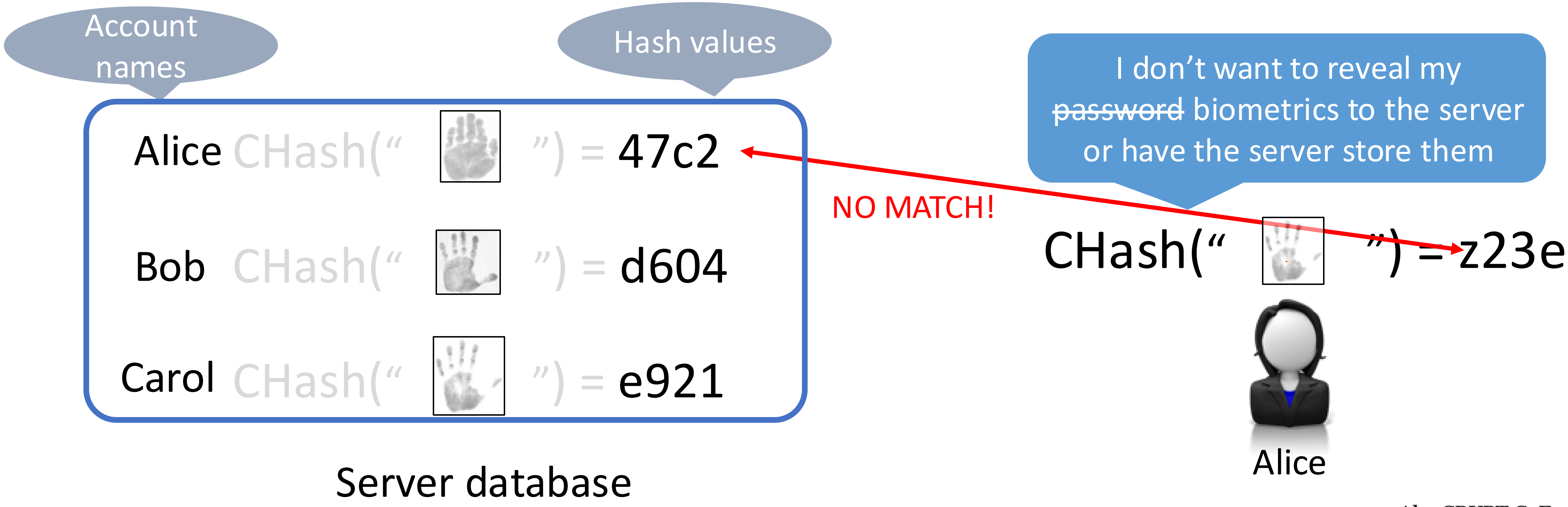
I don’t want to reveal my ~~password~~ biometrics to the server or have the server store them

CHash(“  ”) = z23e

Alice

Detour: Secure Password Authentication

- Why biometric setting creates a challenge? Biometric data are noisy



XBiometrics: *New Secure* Biometric Algorithm

XBiometrics →

- Unique and new algorithm
- Unmatched in both industry and academia
- Encrypt palms with standardized methods
- Enables authentication without decrypting (**ever**)

NIST Special Publication 800-63B

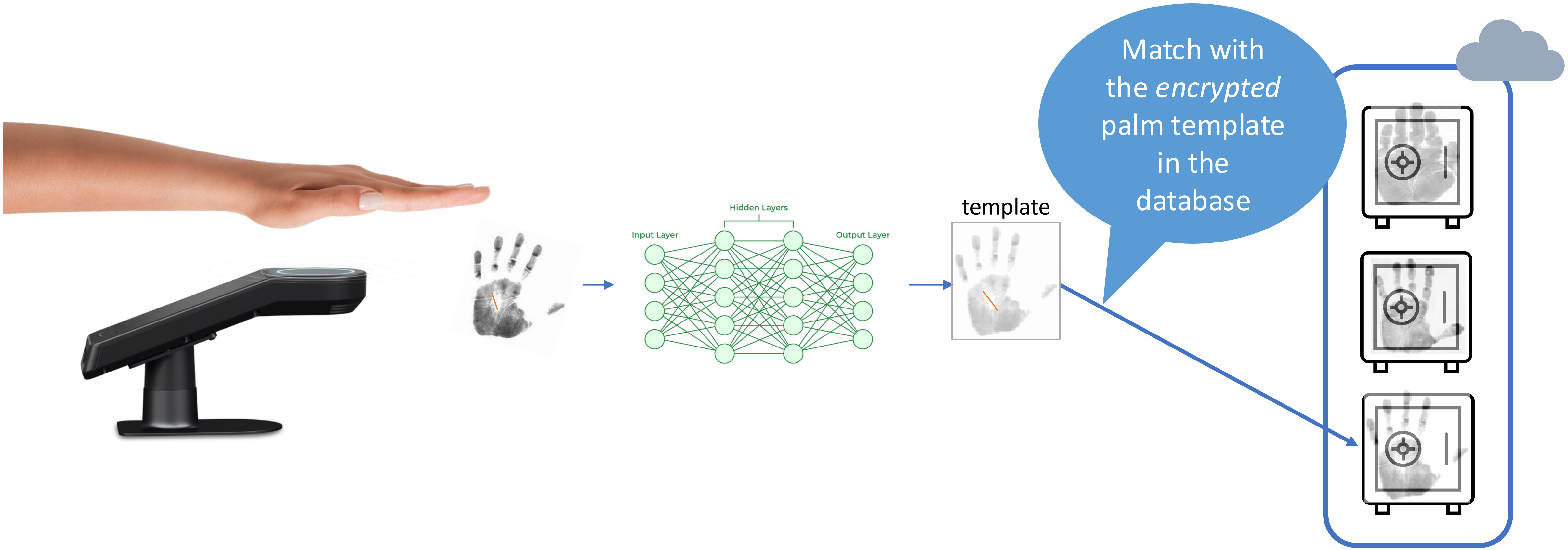
Digital Identity Guidelines

Authentication and Lifecycle Management

	Existing Algorithms	XBiometrics
Encryption in Transit	Yes	Yes
Encryption in Storage/Computation	No	Yes
Revocable	No	Yes
Unlinkable	No	Yes

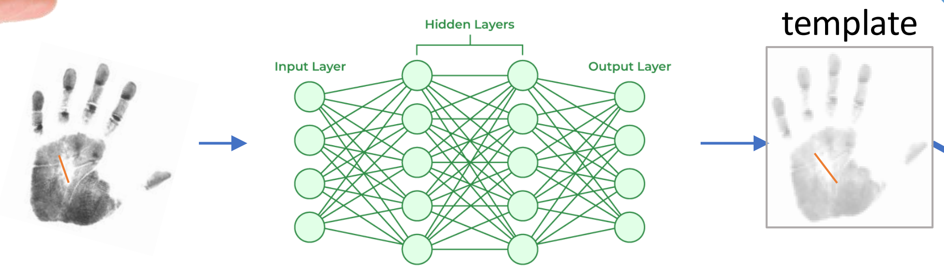
XBiometrics: *New Secure* Biometric Algorithm

Authentication

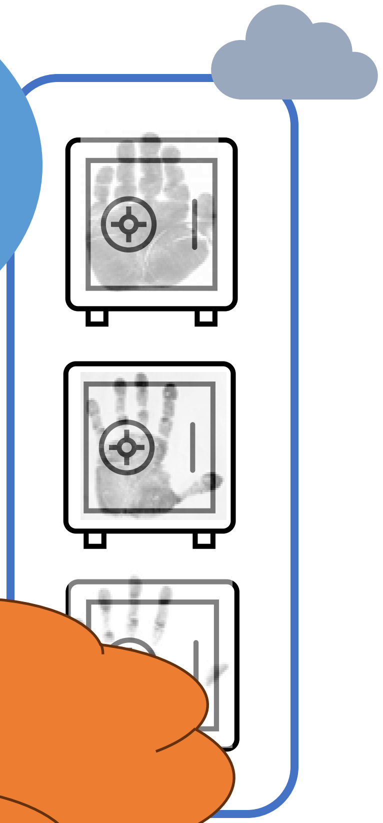


XBiometrics: *New Secure* Biometric Algorithm

Authentication



Match with the *encrypted* palm template in the database



Not as easy as it looks!
Our algorithm facilitates matching *encryptions* of *similar (but different)* palm templates from same person

XBiometrics uses cryptographic hashing in a novel way to match similar biometrics

J.P.Morgan

AlgoCRYPT CoE
AI Research

Conclusion

Group Privacy	**Privacy Preserving Federated Learning**
Pairwise Privacy	**Encrypted LLMs** **Checking AI Model Fairness**
Individual Privacy	**Biometric Authentication:**

Looking Ahead: Key Challenges

1. **Data Leakage**
•Generative AI models can inadvertently reproduce sensitive or private information from their training data, risking exposure of personal or confidential details.
2. **Deepfakes and Synthetic Media**
•Generative AI can create highly realistic fake images, videos, or audio (deepfakes), which can be used to impersonate individuals, spread misinformation, or violate privacy.
3. **Consent and Data Ownership**
•It is often unclear whether individuals have given informed consent for their data to be used in training generative models, raising ethical and legal concerns about data ownership and usage.
4. **Fairness and Bias**
•Generative AI models may reflect or amplify biases present in their training data, leading to unfair or discriminatory outputs that can impact individuals or groups and raise ethical concerns.