Apple’s Deployment of Homomorphic Encryption at Scale

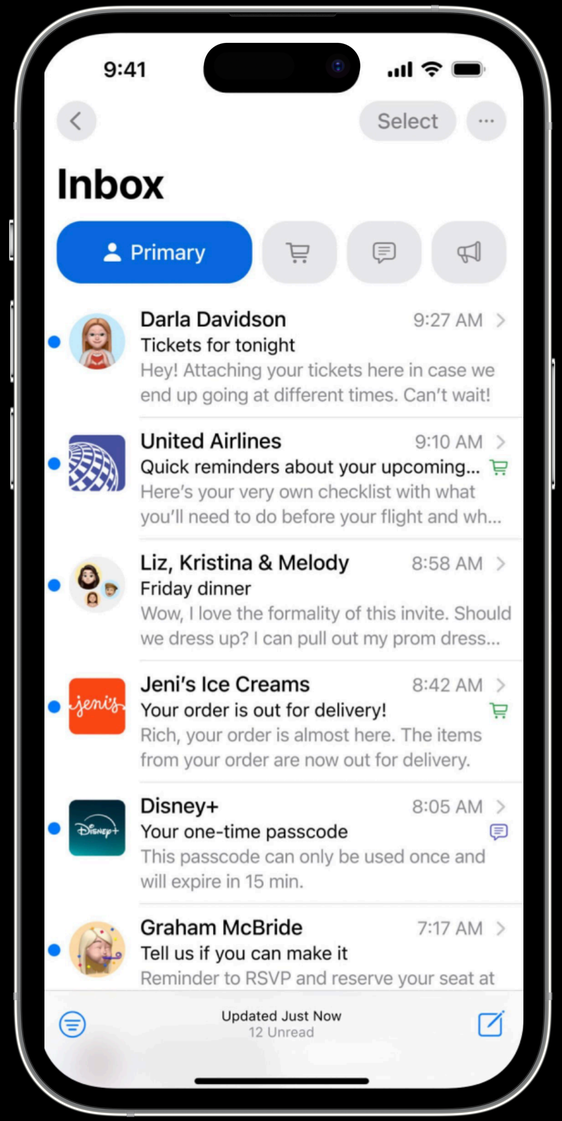
Rehan Rishi, Haris Mughees

RWC 2025 | Apple | 03/26/2025

Homomorphic Encryption across Apple features

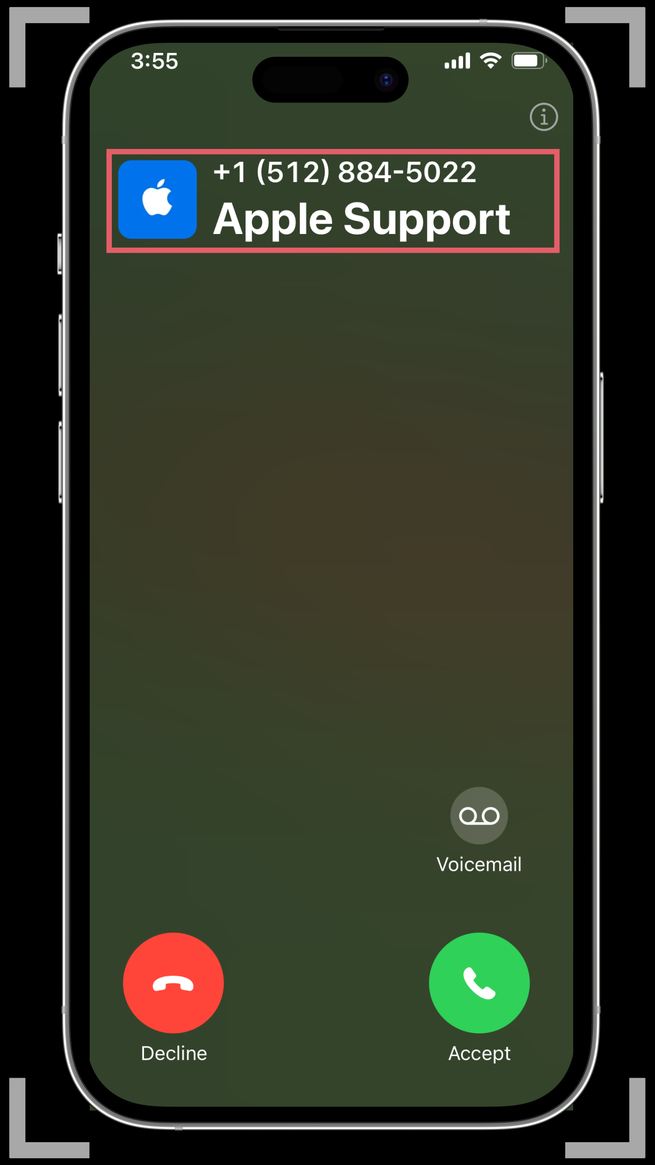
Caller ID

Enhanced Visual Search



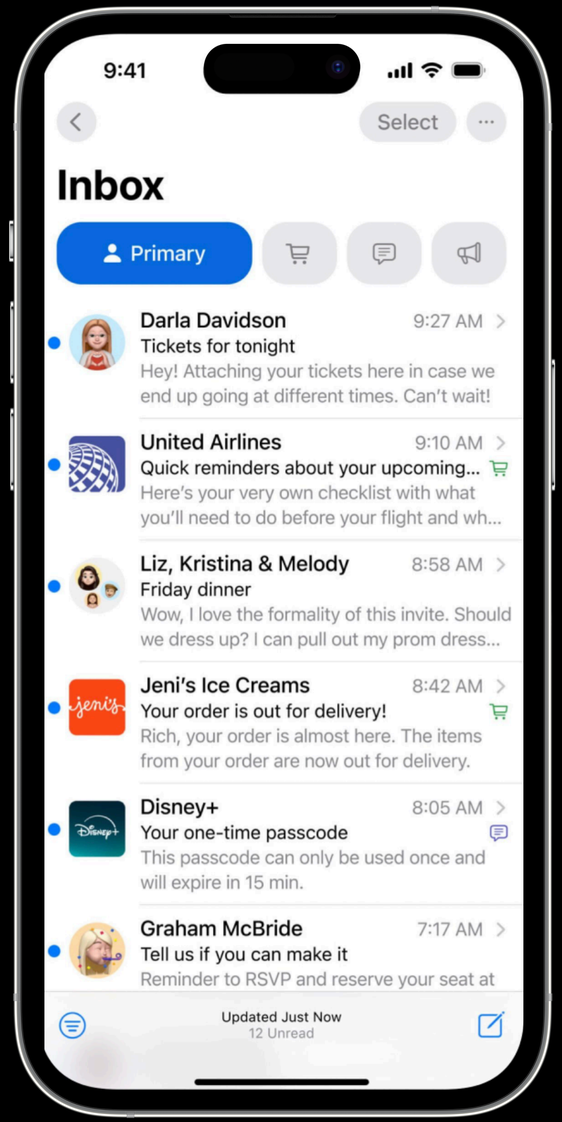
Business Services in Mail

Homomorphic Encryption across Apple features



Caller ID

Enhanced Visual Search

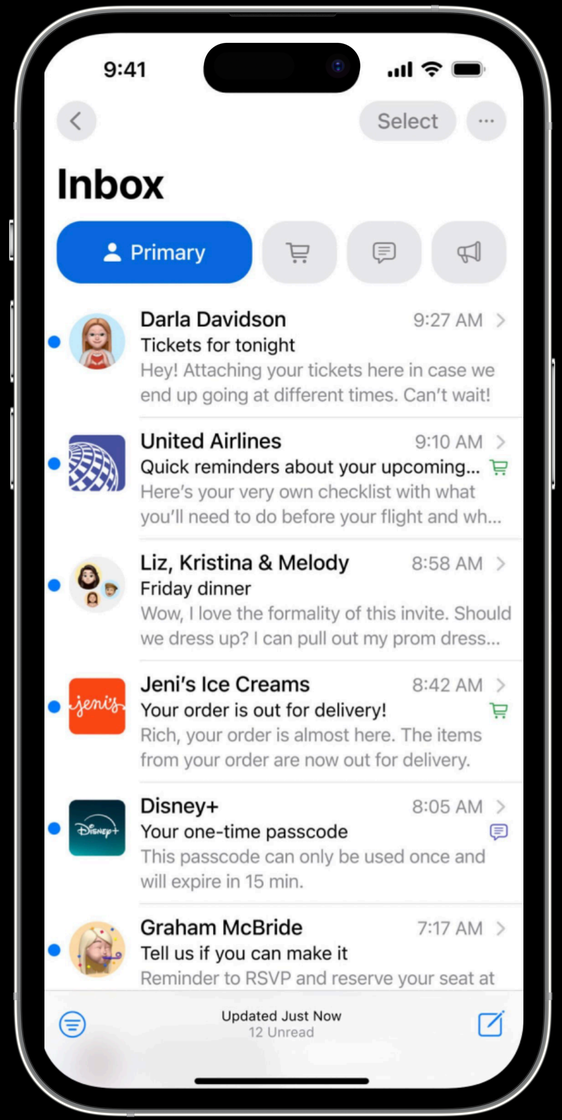


Business Services in Mail

Homomorphic Encryption across Apple features

Caller ID

Enhanced Visual Search

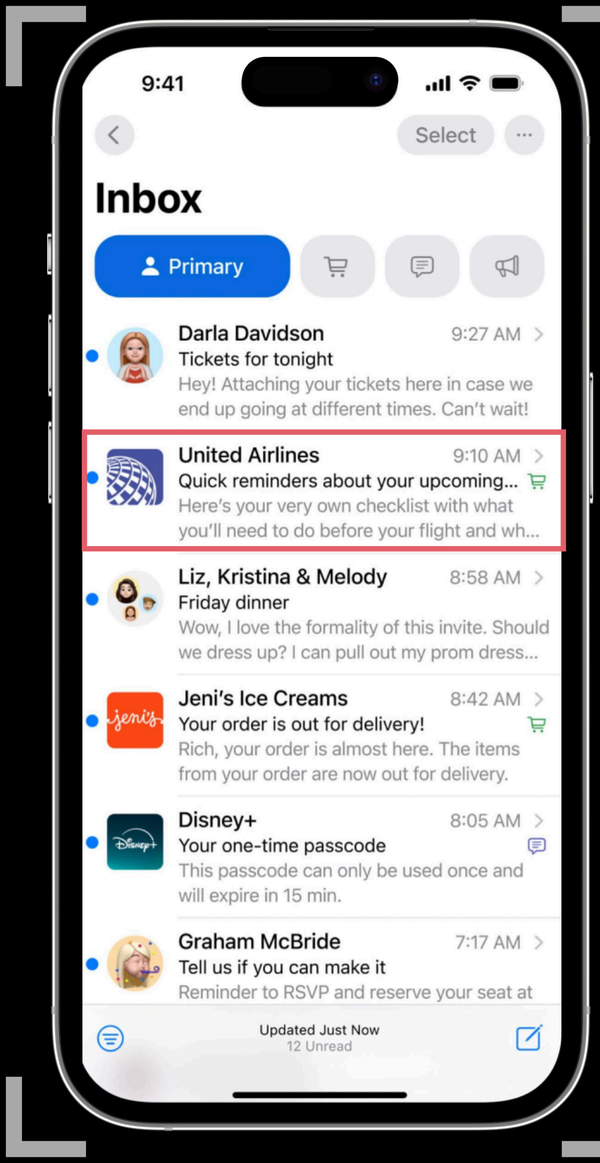


Business Services in Mail

Homomorphic Encryption across Apple features

Caller ID

Enhanced Visual Search



Business Services in Mail

Apple’s privacy commitment

Client queries remain private, even from Apple

Practical system requirements

Meets client constraints

- Small storage *A few MegaBytes*
- High quality experience

Meets server constraints

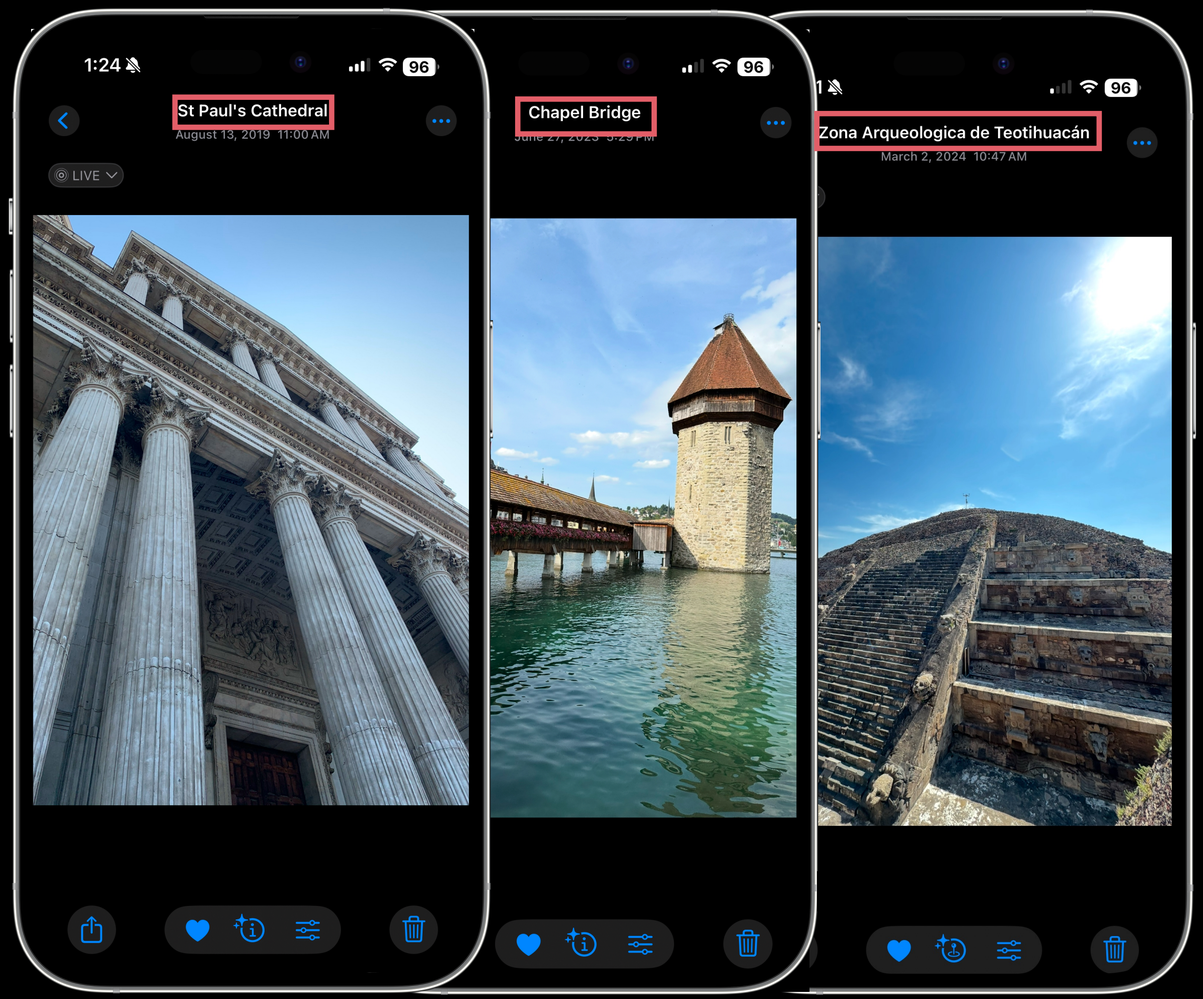
- High queries per second *Tens of thousands*
- Small communication *< 1 MegaByte*

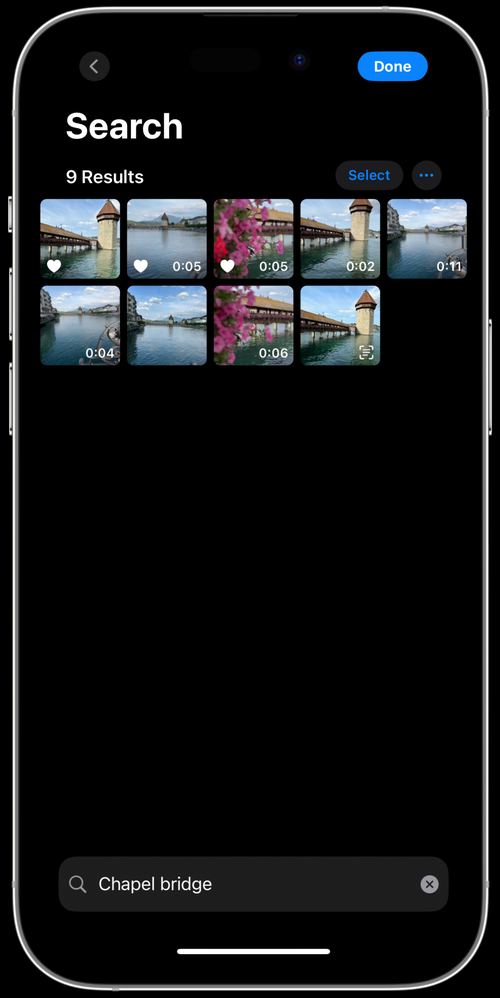
Feature Walkthrough: Enhanced Visual Search

Improving on-device search with privately tagging landmarks and places of interest

Enhanced Visual Search

Improving On-Device Search with Private Tagging





Enhanced Visual Search

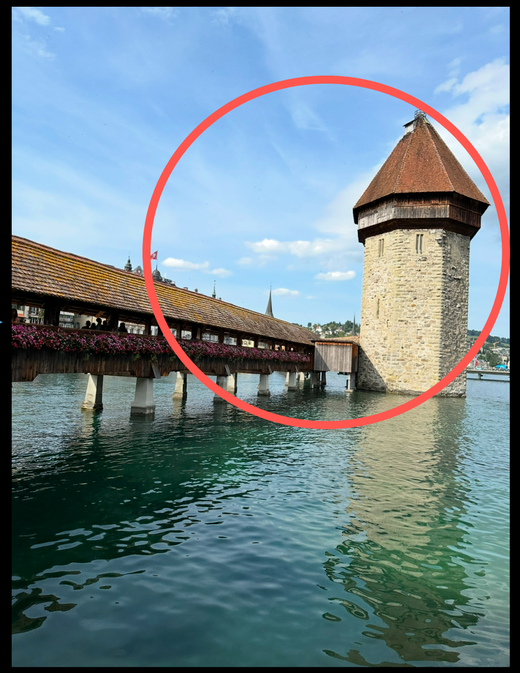
Improving On-Device Search with Private Tagging

Enhanced Visual Search

Improving On-Device Search with Private Tagging

Region-of-interest

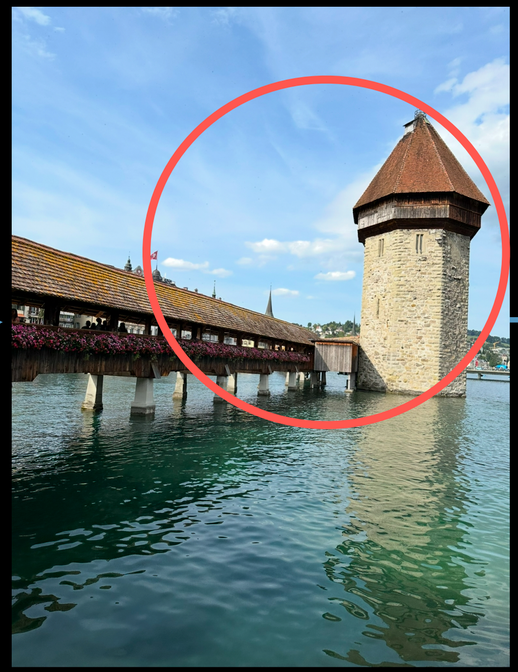
Detector



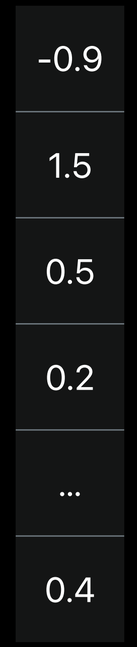
Enhanced Visual Search

Improving On-Device Search with Private Tagging

Region-of-interest Detector

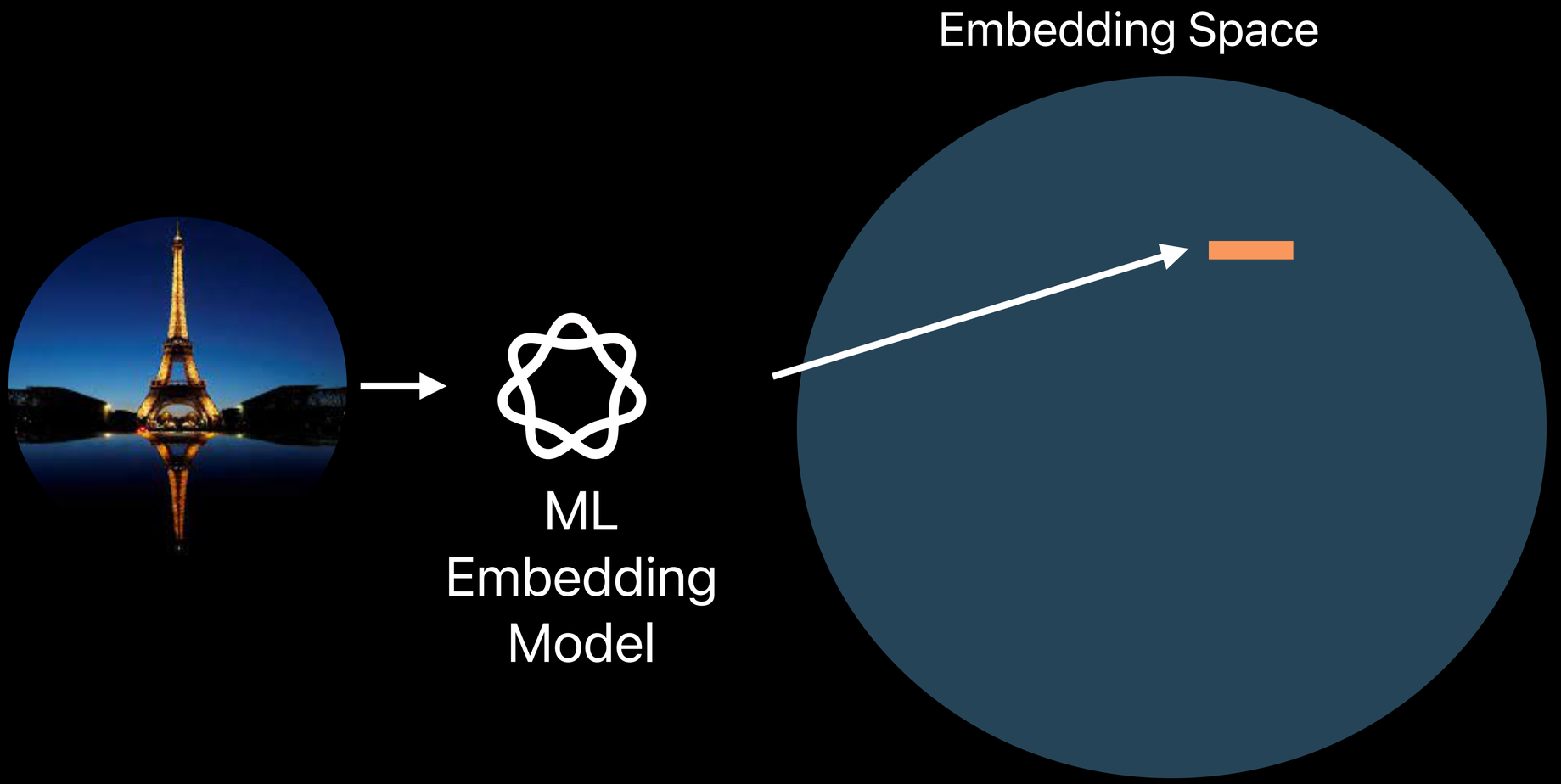


ML Embedding



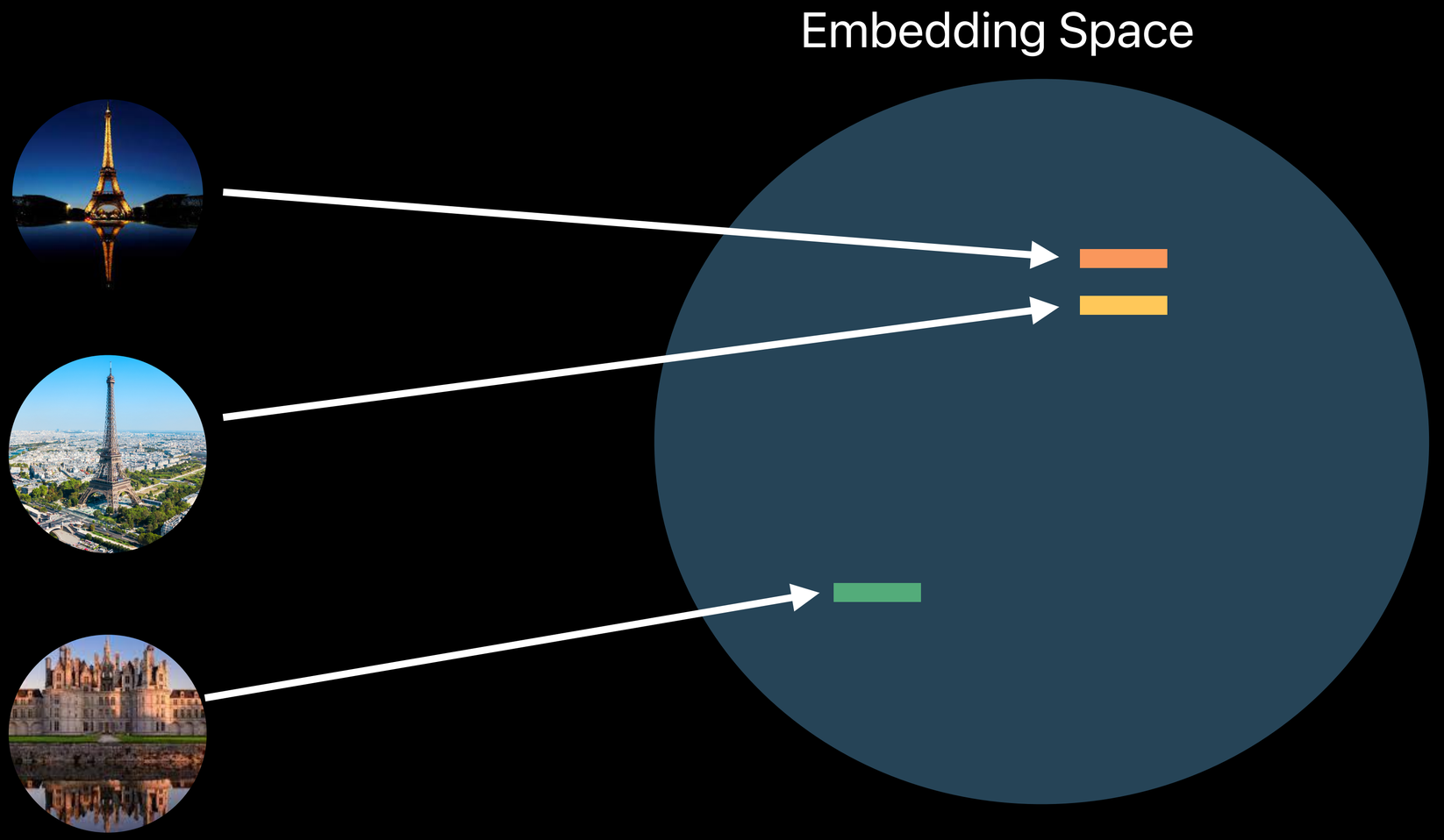
Landmark Region Of Interest in an image is converted into embeddings

ML Embeddings



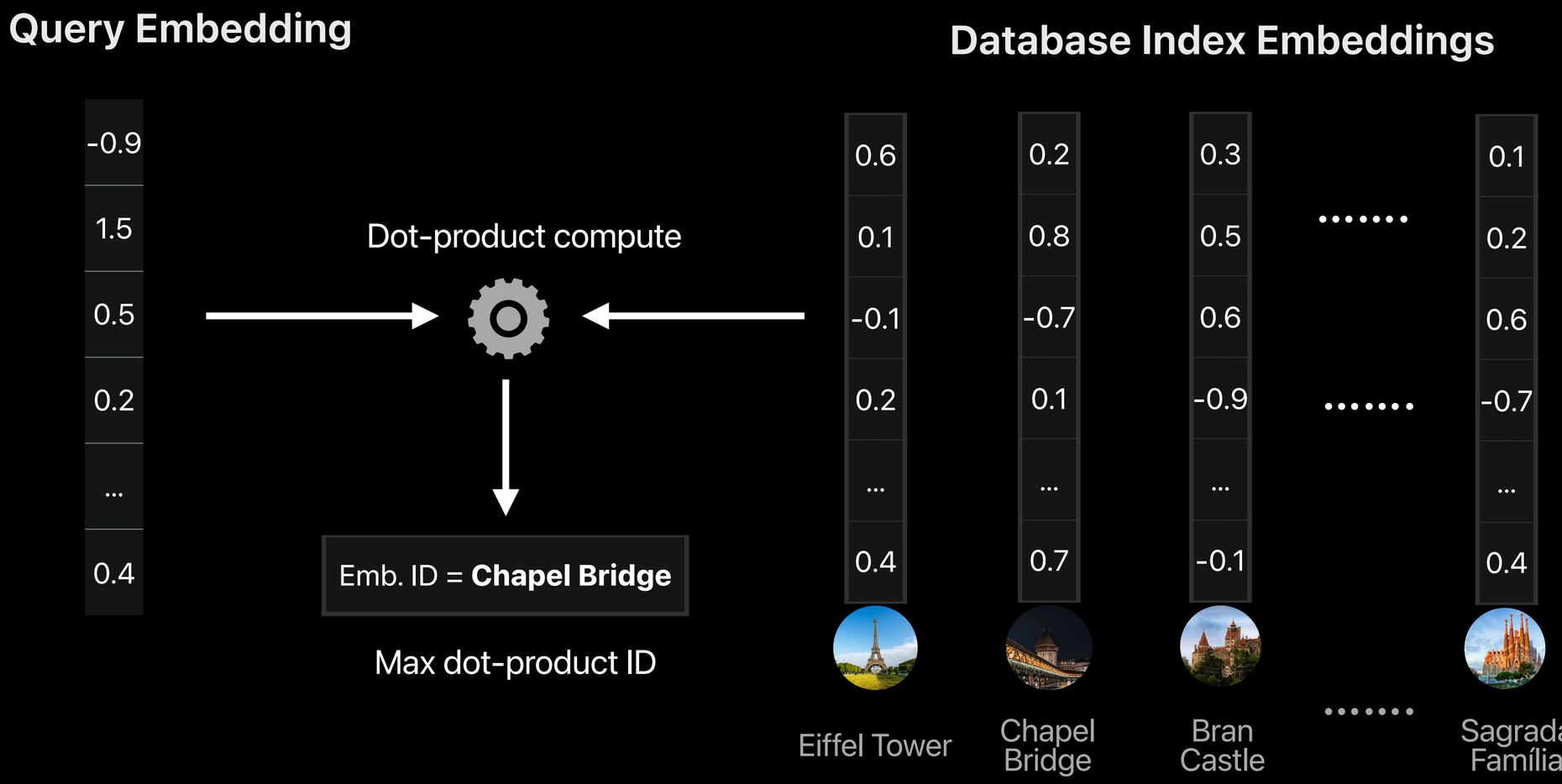
Embedding: A vector of floats of a fixed dimension that retains semantic meaning about the input

ML Embeddings



Similar images are closer in the embedding space

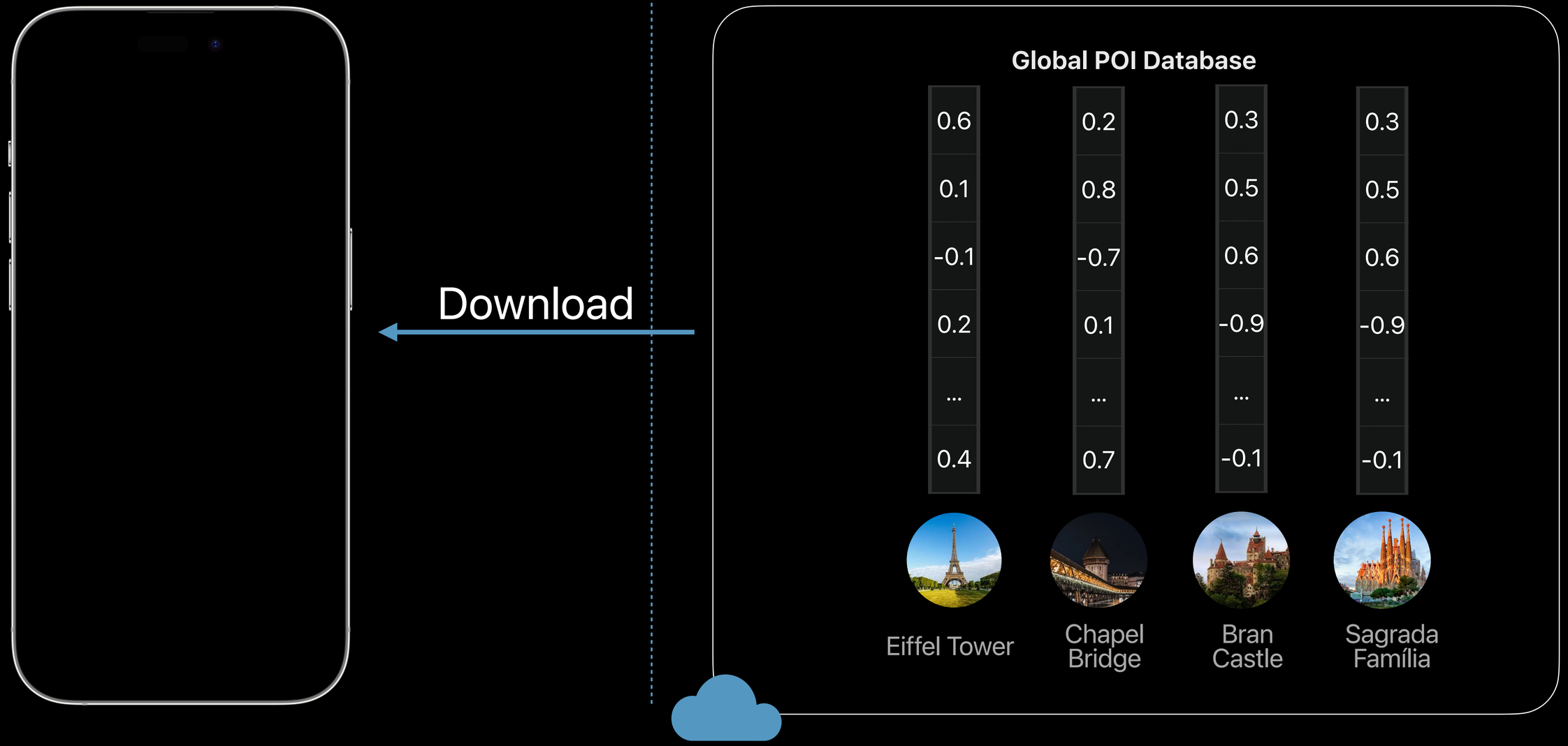
Nearest Neighbor Search with Plaintext Embeddings



Similarity metric computed using dot product

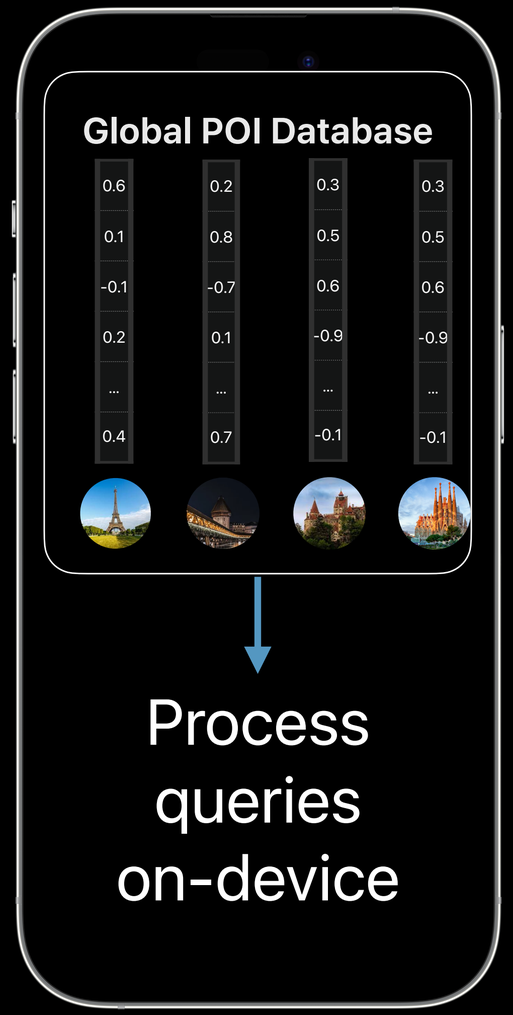
Trivial Solution

Download embeddings to the device



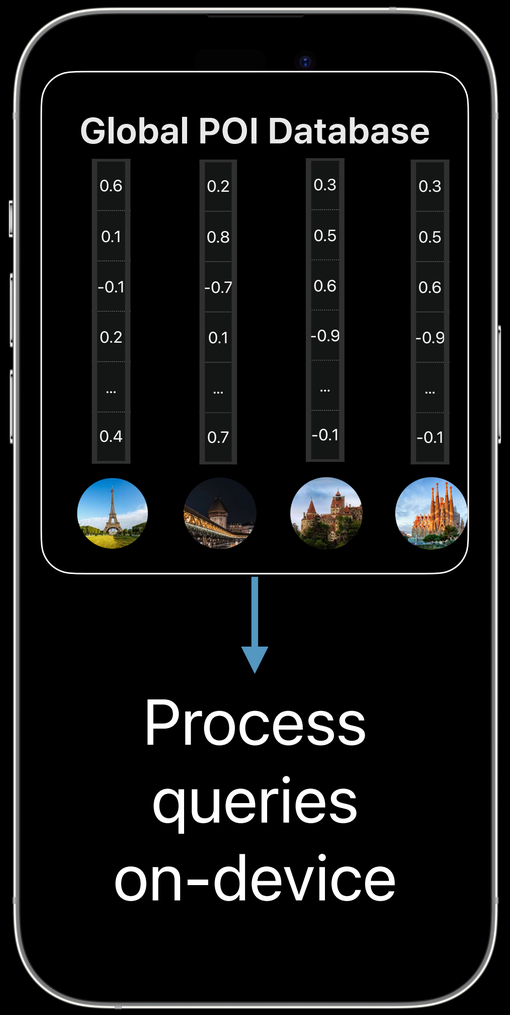
Trivial Solution

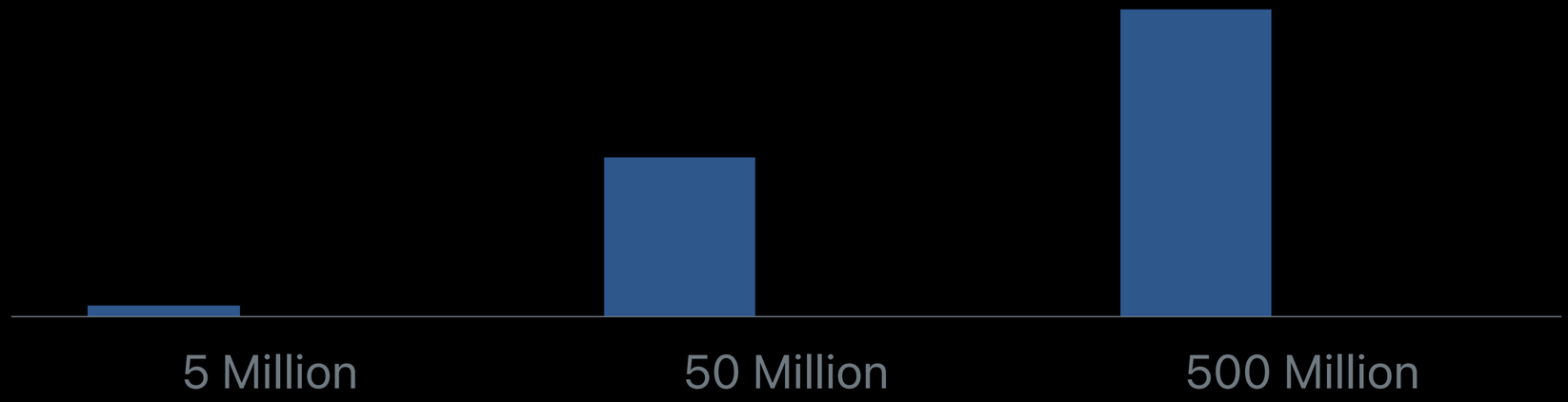
Download embeddings to the device



Trivial Solution

Download embeddings to the device





Devices are storage constrained

Trivial Solution

Download embeddings to the device

Apple’s privacy commitment

Client queries remain private, even from Apple ✅

Practical system requirements

Meets client constraints

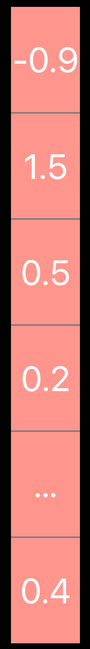
- Small storage *A few MegaBytes* ❌
- High quality experience

Meets server constraints

- High queries per second *Tens of thousands* ✅
- Small communication *< 1 MegaByte*

Background

Server assisted Nearest Neighbor Search without Query Privacy

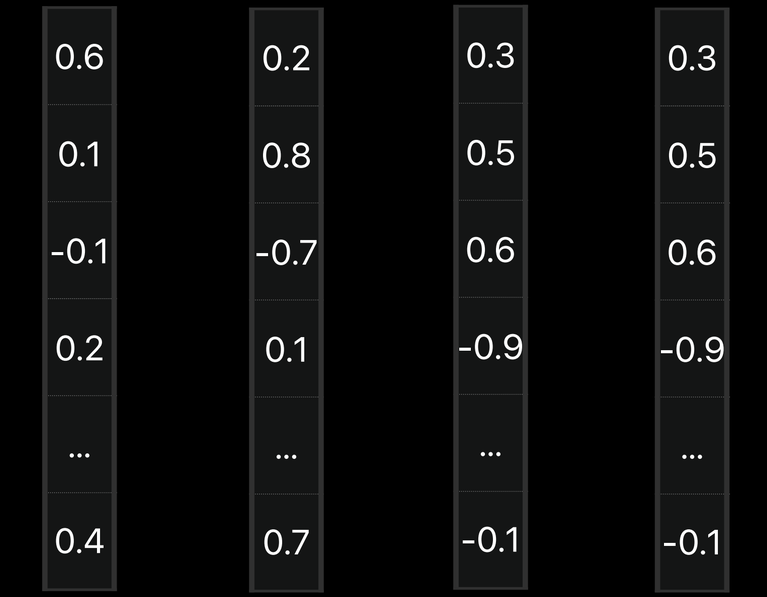


Dot-product compute

Max dot product ID

Emb. ID = **Chapel Bridge**

Global POI Index



Eiffel Tower | Chapel Bridge | Bran Castle | Sagrada Família

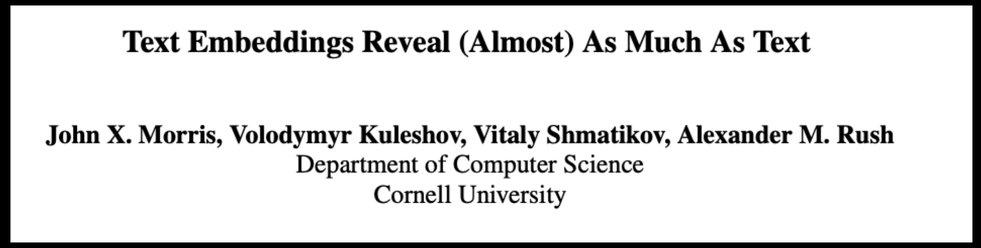
Embeddings Leak Original Data

Session 2A: ML and Information Leakage CCS '20, November 9–13, 2020, Virtual Event, USA

Information Leakage in Embedding Models

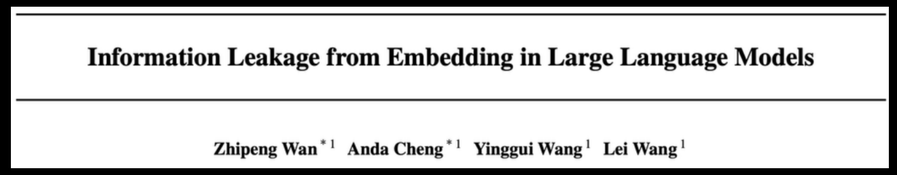
Congzheng Song
Cornell University & Google Brain
cs2296@cornell.edu

Ananth Raghunathan
Facebook & Google Brain
ananthr@cs.stanford.edu



Text Embeddings Reveal (Almost) As Much As Text

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, Alexander M. Rush
Department of Computer Science
Cornell University



Information Leakage from Embedding in Large Language Models

Zhipeng Wan [*1] **Anda Cheng** [*1] **Yinggui Wang** [1] **Lei Wang** [1]

Sentence Embedding Leaks More Information than You Expect: Generative Embedding Inversion Attack to Recover the Whole Sentence

Haoran Li, Mingshi Xu, Yangqiu Song
Dept. of CSE, Hong Kong University of Science and Technology
{hlibt, mxuax}@connect.ust.hk, yqsong@cse.ust.hk

Deep Private-Feature Extraction

Seyed Ali Osia, Ali Taheri, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, Hamid R. Rabiee

Quantifying Privacy Leakage in Graph Embedding

Vasisht Duddu*
Univ Lyon, INSA Lyon, Inria, CITI
vasisht.duddu@uwaterloo.ca

Antoine Boutet
Univ Lyon, INSA Lyon, Inria, CITI
antoine.boutet@insa-lyon.fr

Virat Shejwalkar
University of Massachusetts Amherst
vshejwalkar@cs.umass.edu

Analyzing Sensitive Information Leakage in Trajectory Embedding Models

Jiaxin Ding*
jiaxinding@sjtu.edu.cn
Shanghai Jiao Tong University

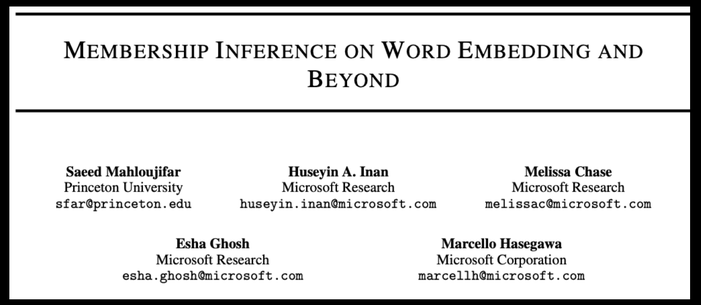
Shichuan Xi*
Xi_Shichuan@sjtu.edu.cn
Shanghai Jiao Tong University

Kailong Wu
1473686097@sjtu.edu.cn
Shanghai Jiao Tong University

Pan Liu
wslp1999@sjtu.edu.cn
Shanghai Jiao Tong University

Xinbing Wang
xwang8@sjtu.edu.cn
Shanghai Jiao Tong University

Chenghu Zhou
zhouch@lreis.ac.cn
Institute of Geographical Science and Natural Resources Research, Chinese Academy of Sciences



MEMBERSHIP INFERENCE ON WORD EMBEDDING AND BEYOND

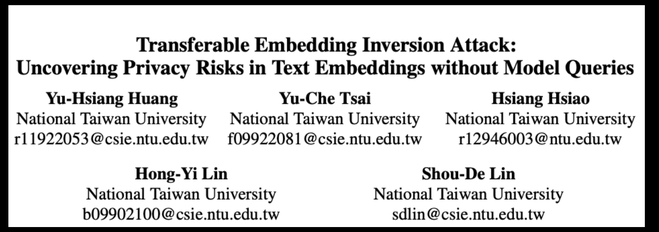
Saeed Mahloujifar
Princeton University
sfar@princeton.edu

Huseyin A. Inan
Microsoft Research
huseyin.inan@microsoft.com

Melissa Chase
Microsoft Research
melissac@microsoft.com

Esha Ghosh
Microsoft Research
esha.ghosh@microsoft.com

Marcello Hasegawa
Microsoft Corporation
marcellh@microsoft.com



Transferable Embedding Inversion Attack: Uncovering Privacy Risks in Text Embeddings without Model Queries

Yu-Hsiang Huang
National Taiwan University
r11922053@csie.ntu.edu.tw

Yu-Che Tsai
National Taiwan University
f09922081@csie.ntu.edu.tw

Hsiang Hsiao
National Taiwan University
r12946003@ntu.edu.tw

Hong-Yi Lin
National Taiwan University
b09902100@csie.ntu.edu.tw

Shou-De Lin
National Taiwan University
sdlin@csie.ntu.edu.tw

Candidate Secure Solution

Using homomorphic encryption only

Encrypted embedding

Encrypted dot-product compute

Global POI Database

Eiffel Tower	Chapel Bridge	Bran Castle	Sagrada Família
0.6	0.2	0.3	0.3
0.1	0.8	0.5	0.5
-0.1	-0.7	0.6	0.6
0.2	0.1	-0.9	-0.9
...	...	...	...
0.4	0.7	-0.1	-0.1

HE: Operations under encryption

Candidate Secure Solution

Using homomorphic encryption only

	Technique	Queries per second	Communication per query
Tiptoe [1]	Additive HE	**909**	**~17.4 MB**
Requirements		*Tens of thousands*	*< 1 MegaByte*

For 3.2 million MSMARCO dataset , assuming 10-thousand cores

[1]- Henzinger, Alexandra, et al. "Private web search with Tiptoe." Proceedings of the 29th symposium on operating systems principles. 2023

Candidate Secure Solution

Using homomorphic encryption only

Apple's privacy commitment

Client queries remain private, even from Apple ✅

Practical system requirements

Meets client constraints

- Small storage *A few MegaBytes* ✅
- High quality experience

Meets server constraints

- High queries per second *Tens of thousands* ❌
- Small communication *< 1 MegaByte*

Problem: High computation + communication

server performs HE operation for each entry in the database

Our solution

- **Use Differential Privacy to reduce server computation**
- **Use efficient HE to reduce communication**

Problem: **High computation + communication**

server performs HE operation for each entry in the database

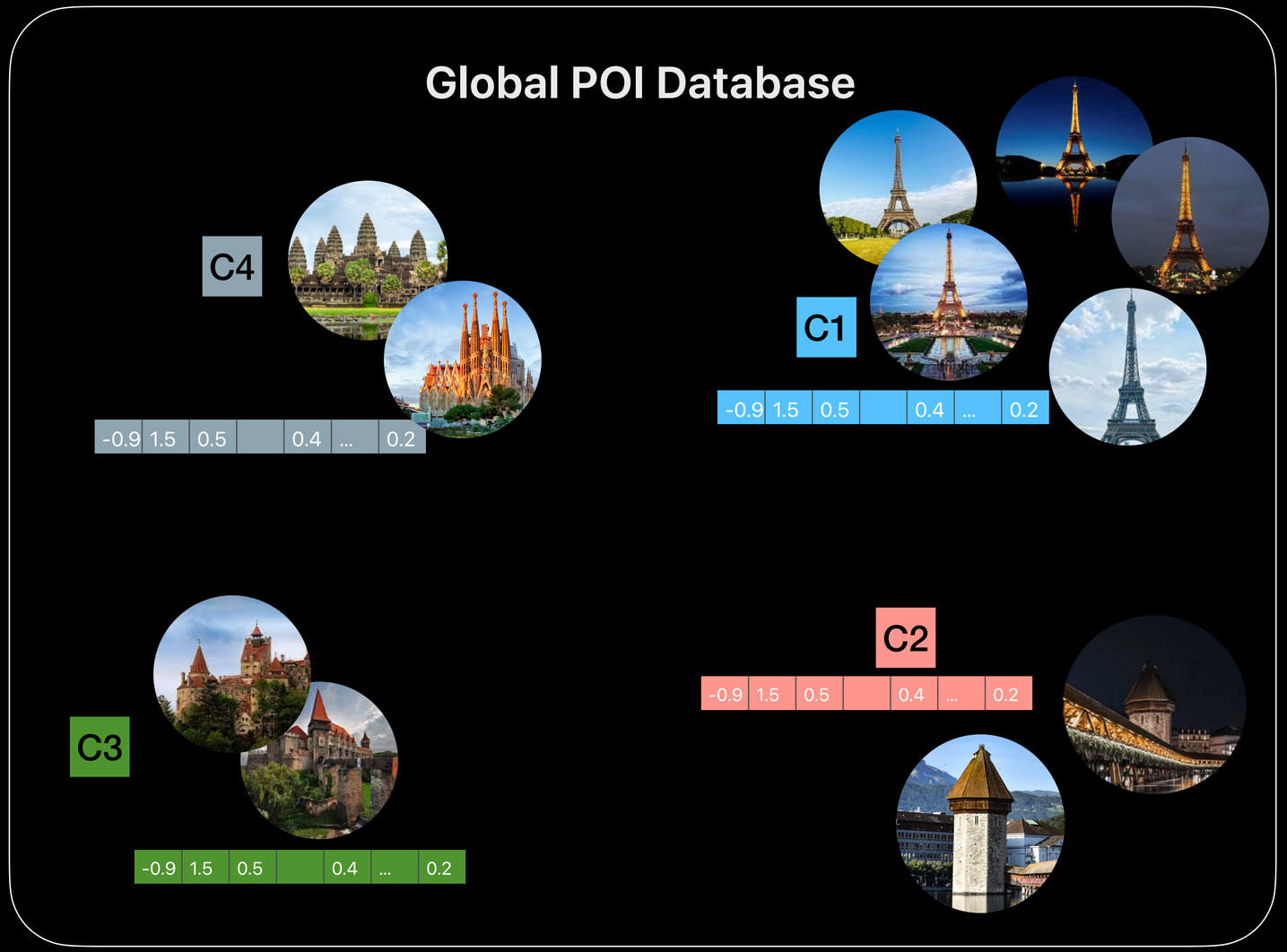
Our solution

- **Use Differential Privacy to reduce server computation**
- **Use efficient HE to reduce communication**

Our Solution

Reduce server computation with DP

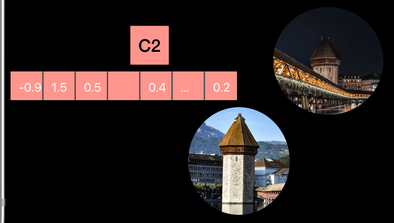
Clustering: standard ML technique that's a hard requirement for an efficient nearest neighbor search

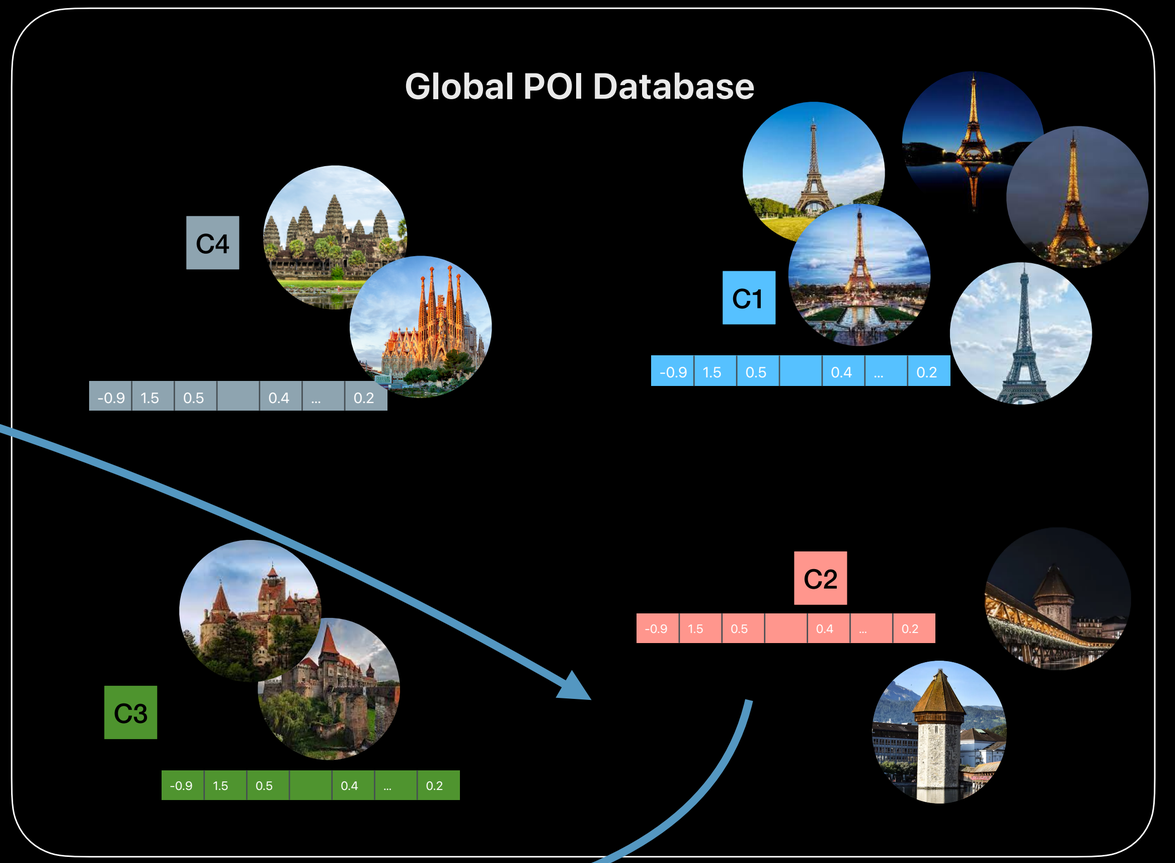


Our Solution

Reduce server computation with DP

C2





Our Solution

Reduce server computation with DP

Problem: **Revealing cluster might reveal semantic information about the client photo**

Observation: **Many users at any given time. We can hide a user query in batch of queries.**

Our Solution

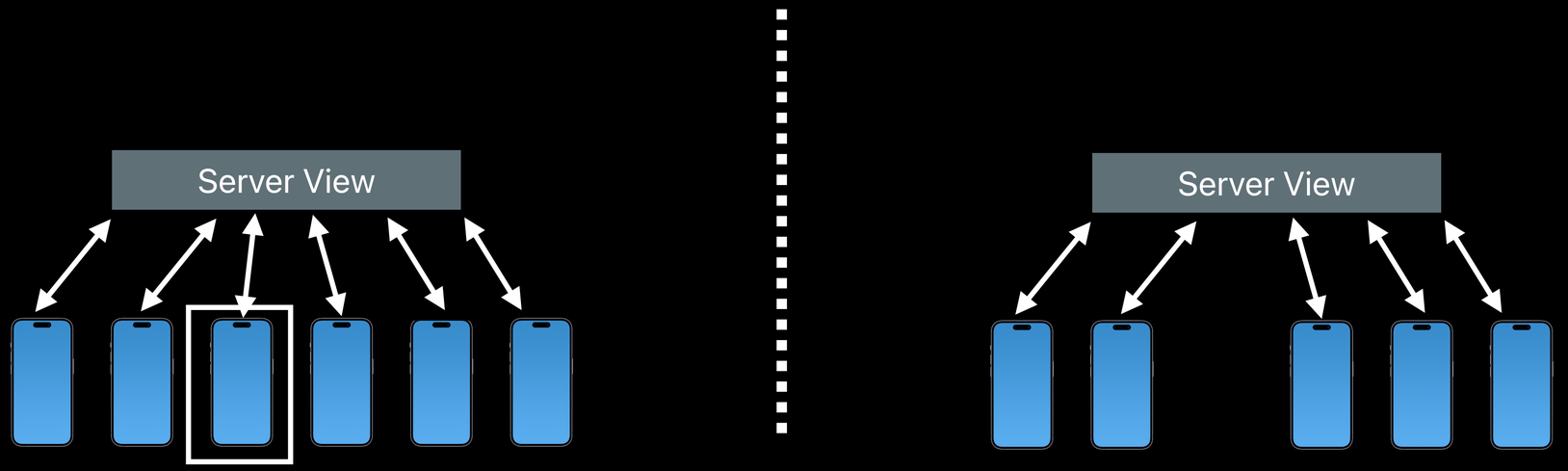
Reduce server computation with DP

- Provides trade-off between efficiency and privacy
- Formally bounds worst privacy leakage through clusters
- Guarantees (ϵ,δ)-differential privacy at the user level w.r.t user's photo library

Our Solution

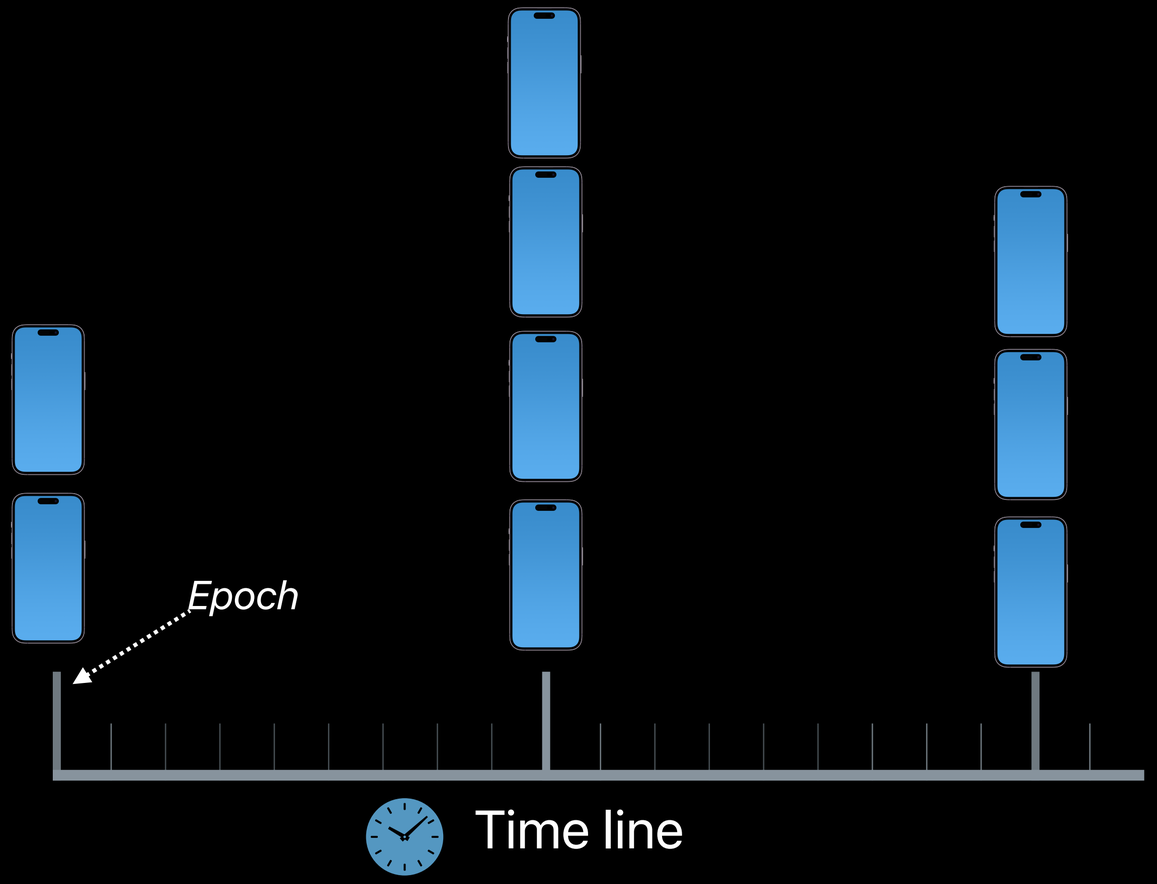
Reduce server computation with DP

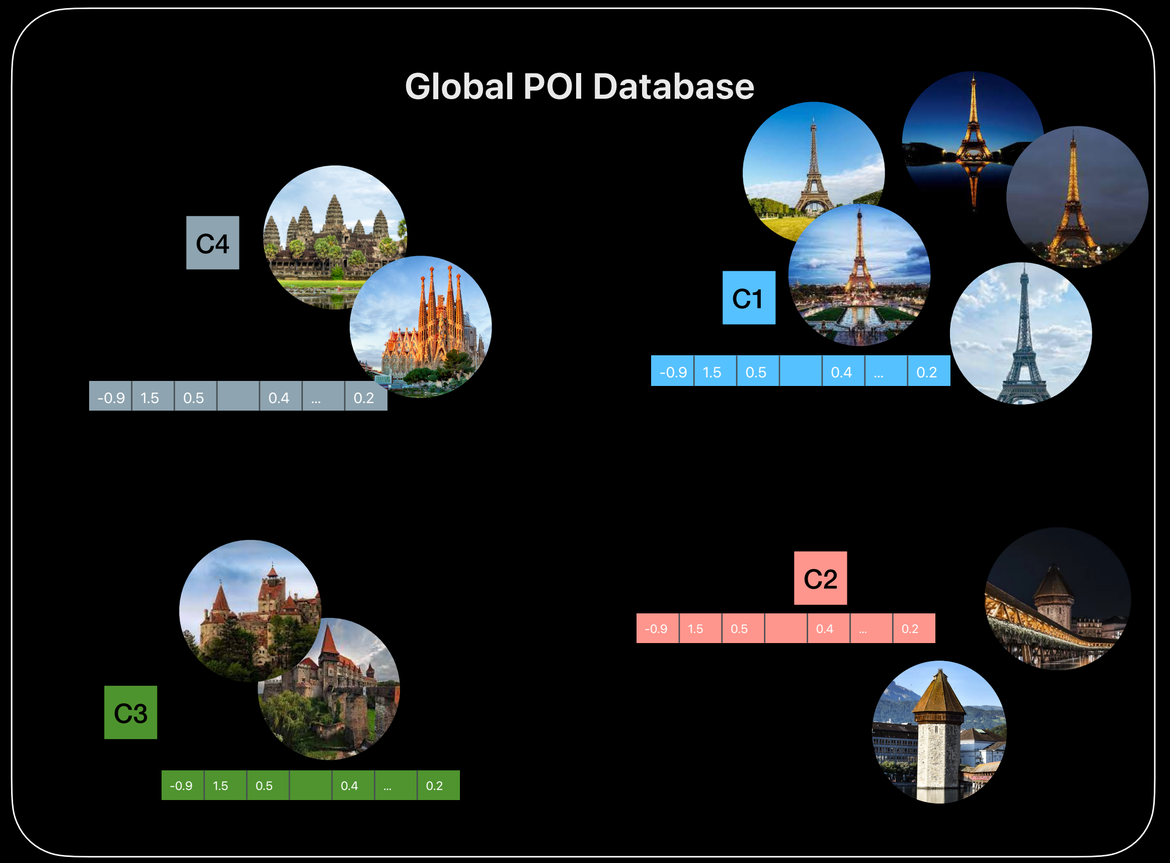
- **Intuition of our privacy guarantee**



We selected $\epsilon = 0.8, \delta = 10^{-9}$

Achieving (ϵ, δ)**-DP:** First Step





Works in epochs, with many clients

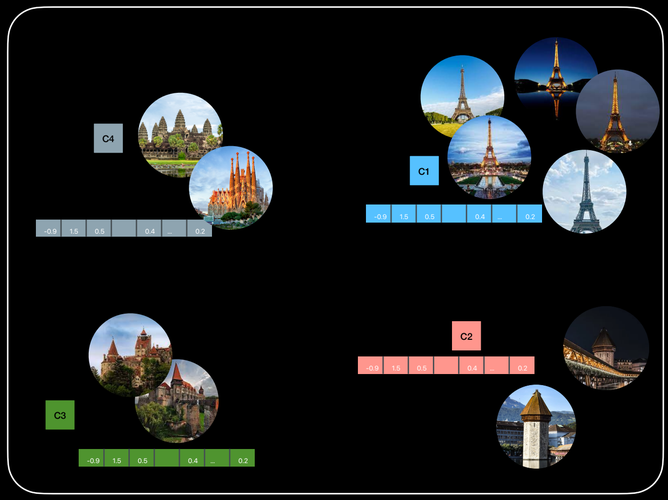
Achieving (ϵ, δ)**-DP:** Second Step

Request

User IP:
123.123.123.123

Anonymization Network

Request



Oblivious HTTP

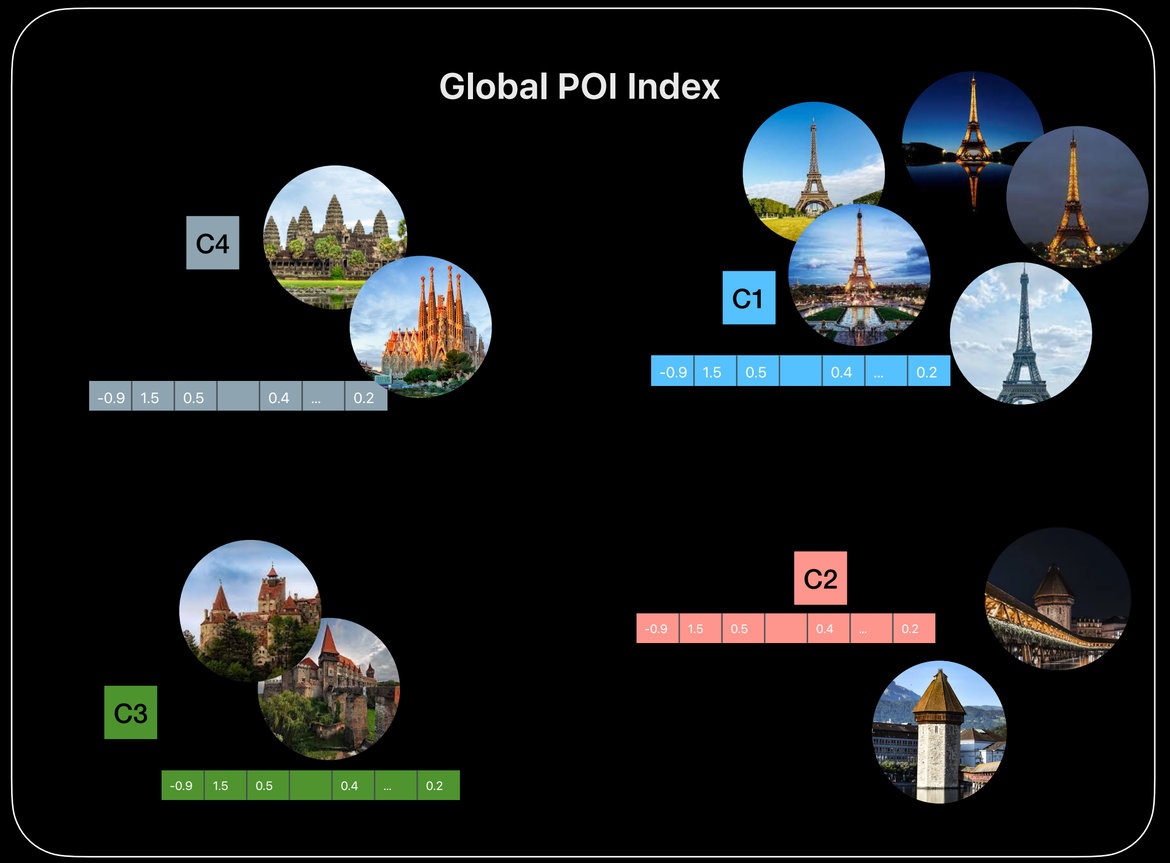
Achieving (ϵ, δ)-DP: Third Step

Real query

C2

Fake query

C4



Fake queries

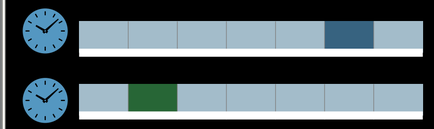
Achieving (ϵ, δ)-DP: Fourth Step

Real query

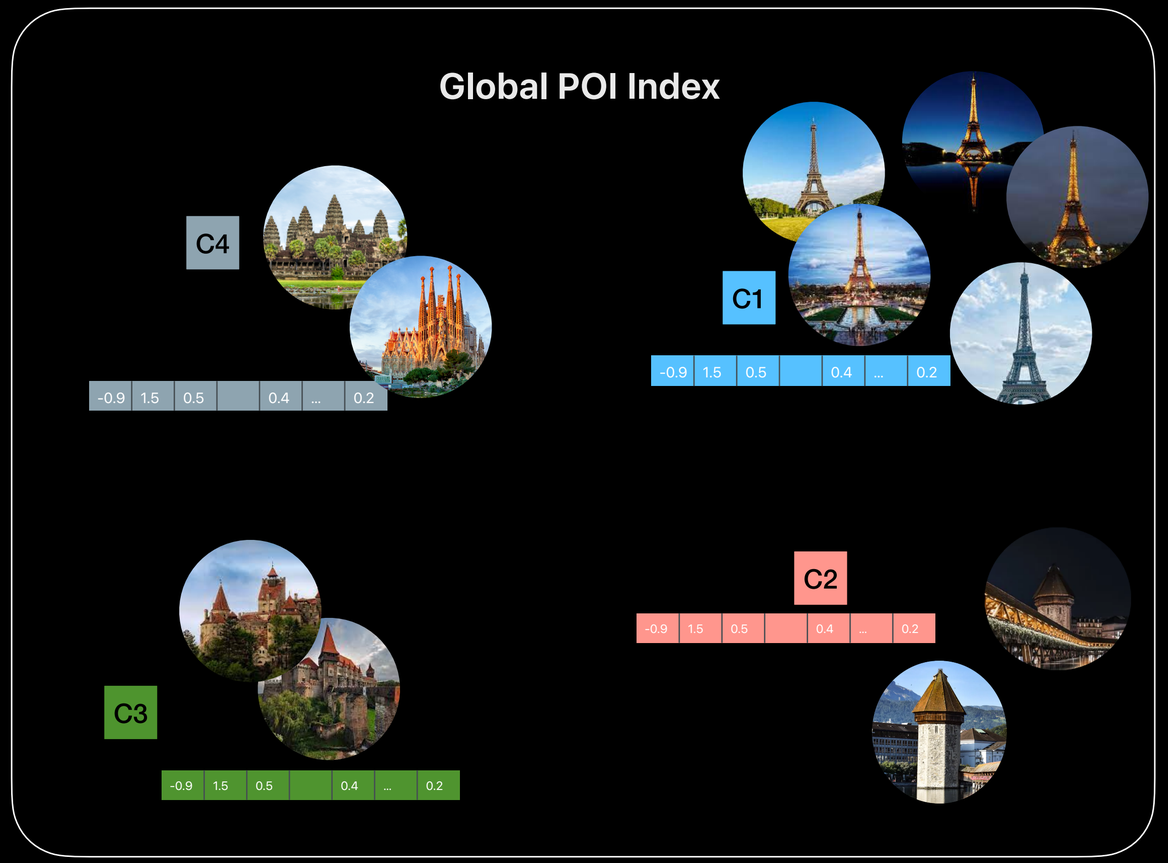
Fake query

C2

C4



Epoch length



Random query schedule

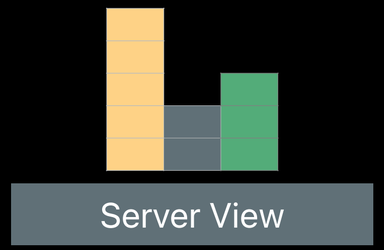
Our Solution

Reduce server computation with DP

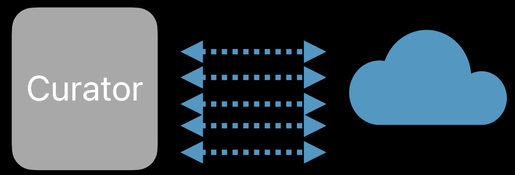
Proof details: *Scalable Private Search with Wally (https://arxiv.org/abs/2406.06761)*

Proof Intuition:

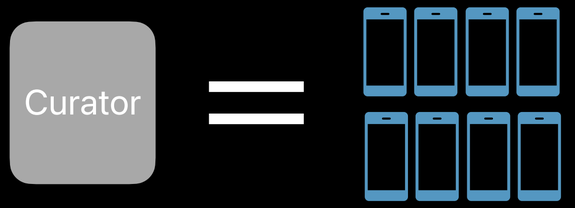
1. We show that server view is a noisy histogram of clusters



2. Prove this noisy histogram is (ϵ, δ)-DP in central model



3. The server gains no extra advantage in distributed model

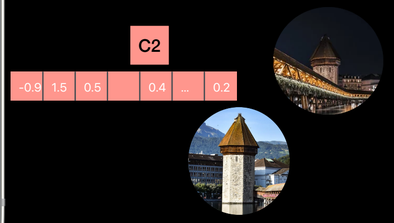


Our Solution

Reduce server computation with DP

C2

-0.9 1.5 0.5 0.4 ... 0.2



Global POI Index

C4

-0.9 1.5 0.5 0.4 ... 0.2

C1

-0.9 1.5 0.5 0.4 ... 0.2

C3

-0.9 1.5 0.5 0.4 ... 0.2

C2

-0.9 1.5 0.5 0.4 ... 0.2

Problem: **High computation + communication**

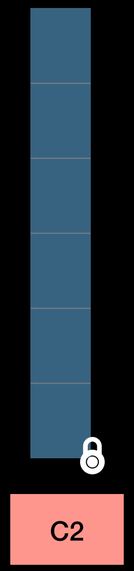
server performs HE operation for each entry in the database

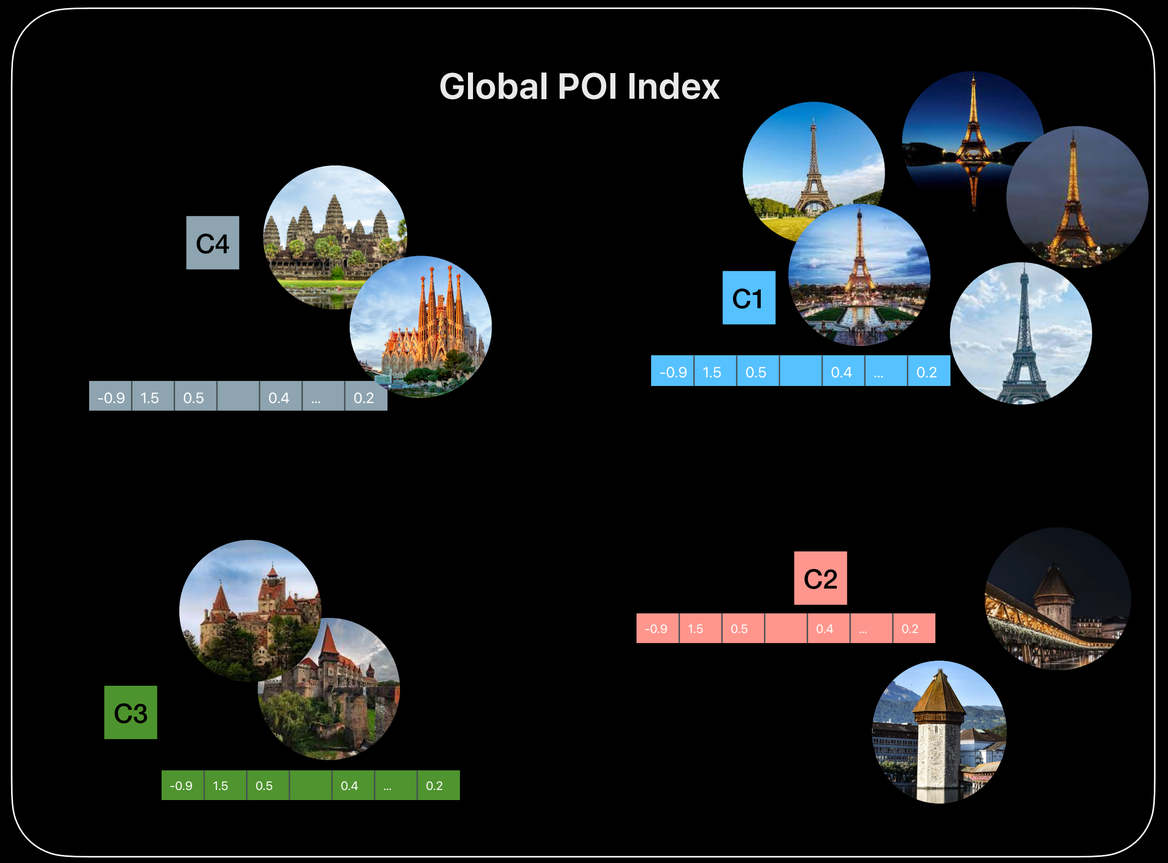
Our solution

- **Use Differential Privacy to reduce server computation**
- **Use efficient HE to reduce communication**

Our Solution
Reduce communication with HE

BFV HE [1]

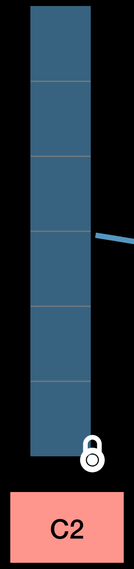


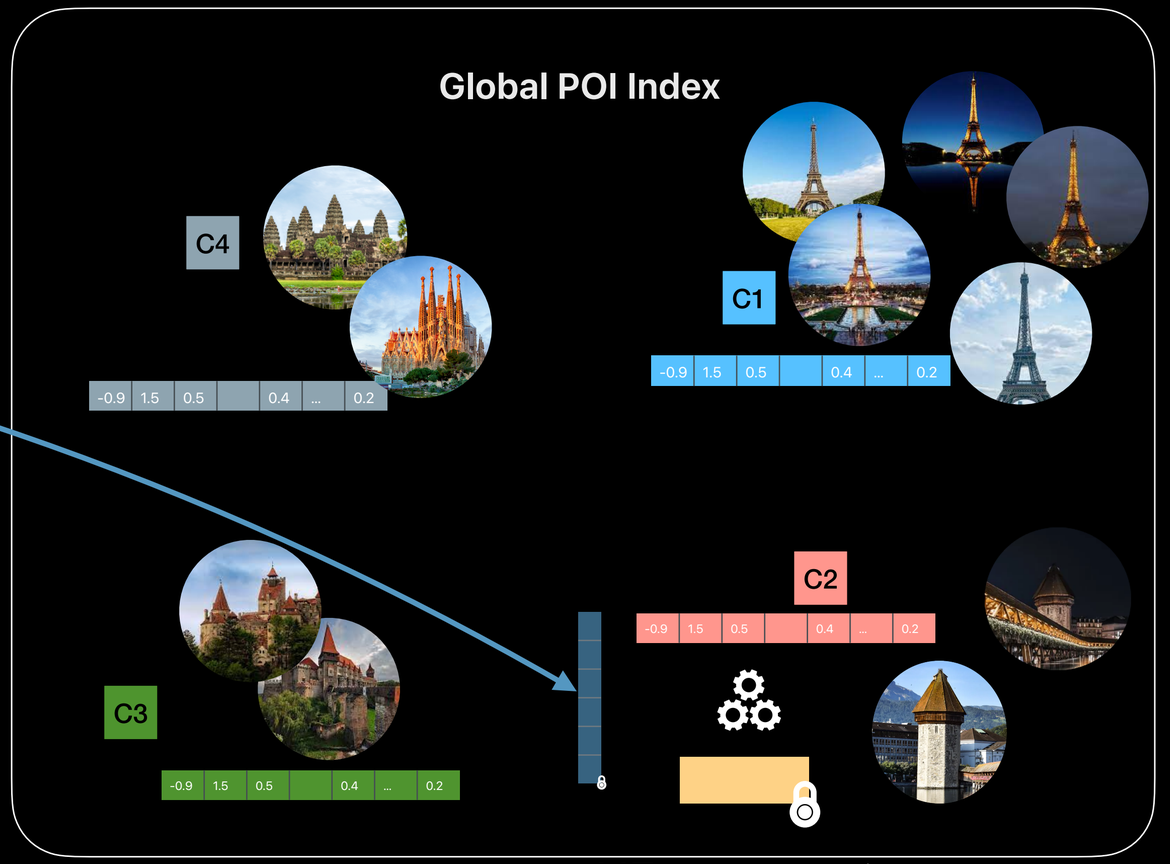


[1] Jean-Claude Bajard, et al., "A Full RNS Variant of FV-like Somewhat Homomorphic Encryption Schemes," International Conference on Selected Areas in Cryptography, 2016

Our Solution

Reduce communication with HE

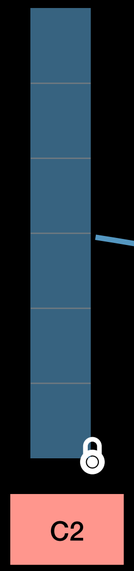




HE computation

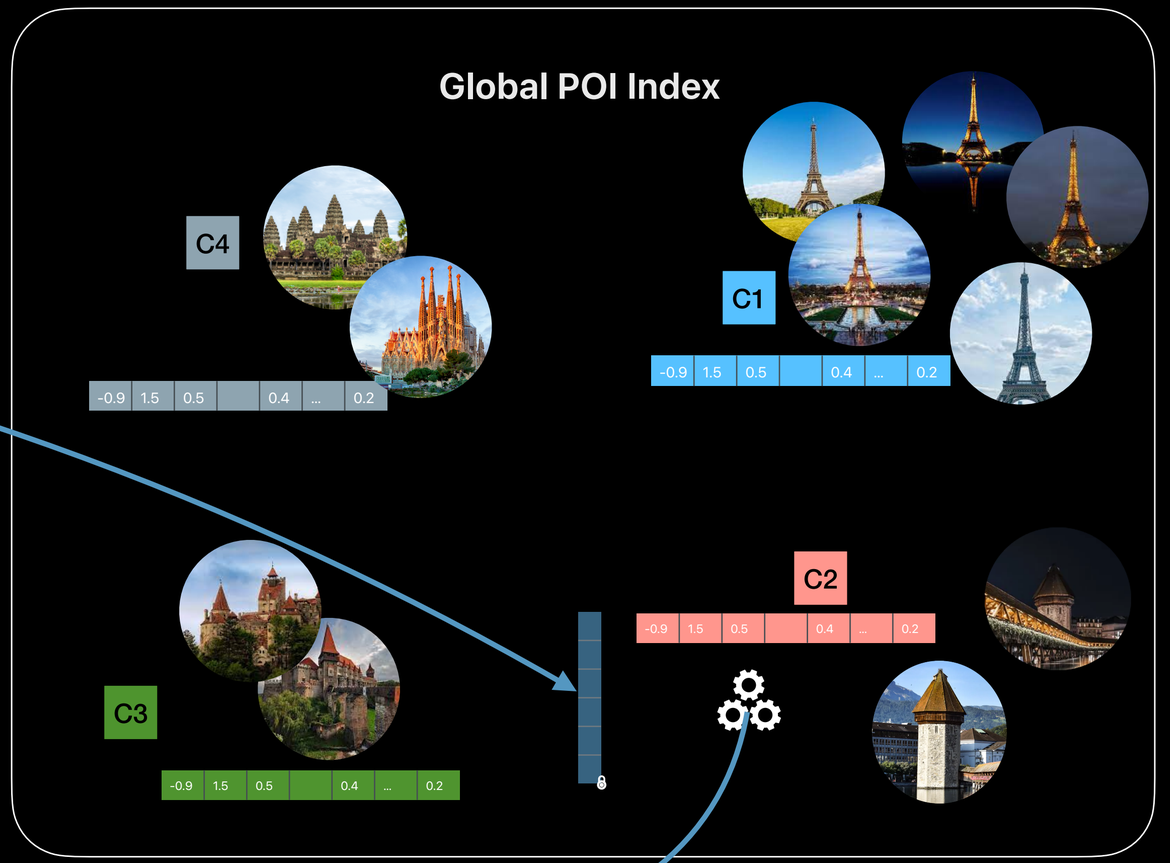
Our Solution

Reduce communication with HE



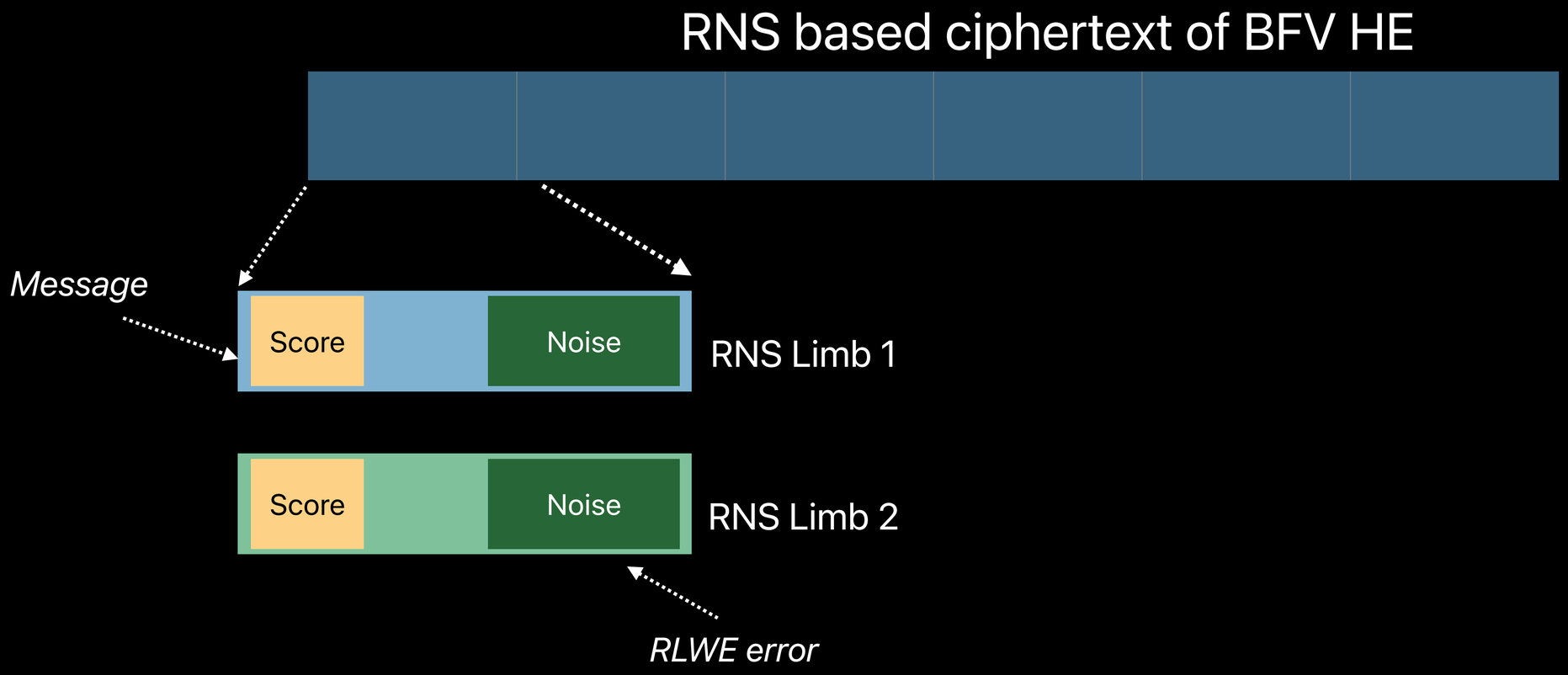
C2

Single response ciphertext



Our Solution

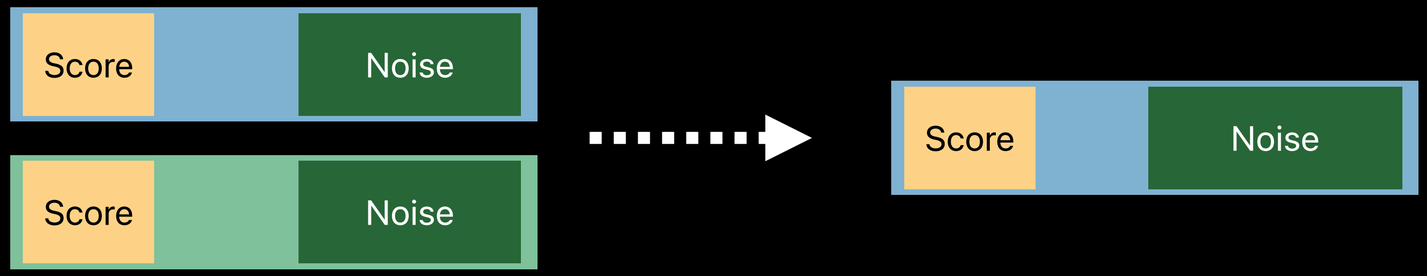
Reduce communication with HE



Message is encoded in higher order bits

Our Solution
Reduce communication with HE

- **Modulus switching [1]:** Keep single RNS limb



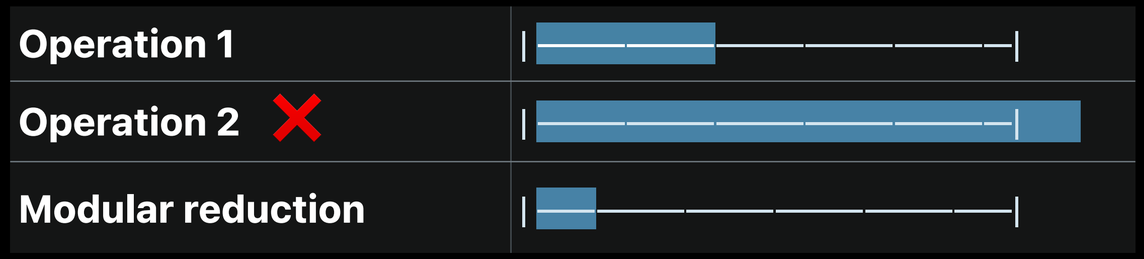
- **Dropping LSB [2]:** Further drop least significant bits from the remaining limb

[1] Zvika Brakerski and Vinod Vaikuntanathan, "Efficient Fully Homomorphic Encryption from (Standard) LWE," IEEE Symposium on Foundations of Computer Science, 2011
[2] Zhenyu Huang, et al., "Cheetah: Lean and Fast Secure Two-Party Deep Neural Network Inference," USENIX Security Symposium, 2022

Our Solution

Other Optimizations

- **Delayed modular reduction to reduce server compute**



Operation = arithmetic operations in field

- **Plaintext RNS to reduce evaluation key size**
  - Evaluation key dominates request size to maintain anonymity

Enhanced Visual Search
Results

	Technique	Queries per second	Communication per query
Our Results*	HE+DP+ Anonymization Network	**>25,000**	**0.56 MB**
Tiptoe [1]	Additive HE	**909**	**~17.4 MB**

**Includes overhead due to fake queries*

For 3.2 million MSMARCO dataset , assuming 10-thousand cores

[1] Alexandra Henzinger, et al., "Private Web Search with Tiptoe," Symposium on Operating Systems Principles, 2023

Our Solution

Apple’s privacy commitment

Client queries remain private, even from Apple ✅

Practical system requirements

Meets client constraints

- Small storage
- High quality experience

✅

Meets server constraints

- High queries per second
- Small communication

✅

Open source Server HE implementation

Server Side: Apple Swift Homomorphic Encryption

Auditable ✅

Novel optimizations ✅

Feedback welcome via Github ✅

https://github.com/apple/swift-homomorphic-encryption

Device Side HE implementation: Corecrypto

Auditable ✅

Novel optimizations ✅

https://developer.apple.com/security/#corecrypto

Summary of Apple's Deployment of Homomorphic Encryption at Scale

- Enhancing on-device experiences with information from the server while maintaining one of the strongest notions of privacy
- Efficient Homomorphic Encryption for several features running on over a billion devices
- Uniquely combined with other state-of-the-art privacy technologies:
  - Differential privacy
  - Anonymization network
  - Privacy pass

Public documentation and resources

apple / swift-homomorphic-encryption

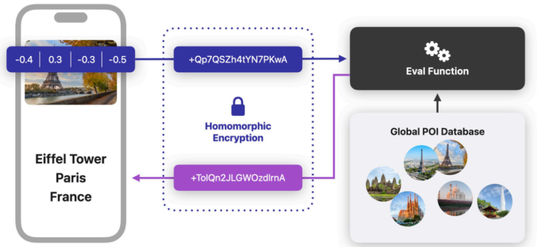
https://github.com/apple/swift-homomorphic-encryption

Highlight | October 24, 2024

Privacy

Combining Machine Learning and Homomorphic Encryption in the Apple Ecosystem

Using Private Nearest Neighbor Search for Enhanced Visual Search for photos.



https://machinelearning.apple.com/research/homomorphic-encryption

SMS and Call Reporting / Getting up-to-date calling and blocking information for your app

Article

Getting up-to-date calling and blocking information for your app

Implement the Live Caller ID Lookup app extension to provide call-blocking and identity services.

https://developer.apple.com/documentation/identitylookup/getting-up-to-date-calling-and-blocking-information-for-your-app

Scalable Private Search with Wally

Hilal Asi, Fabian Boemer, Nicholas Genise, Muhammad Haris Mughees, Tabitha Ogilvie, Rehan Rishi, Guy N. Rothblum, Kunal Talwar, Karl Tarbe, Ruiyu Zhu, Marco Zuliani

https://machinelearning.apple.com/research/wally-search