CSC 116 Data Anonymization / Zero-Knowledge Proof

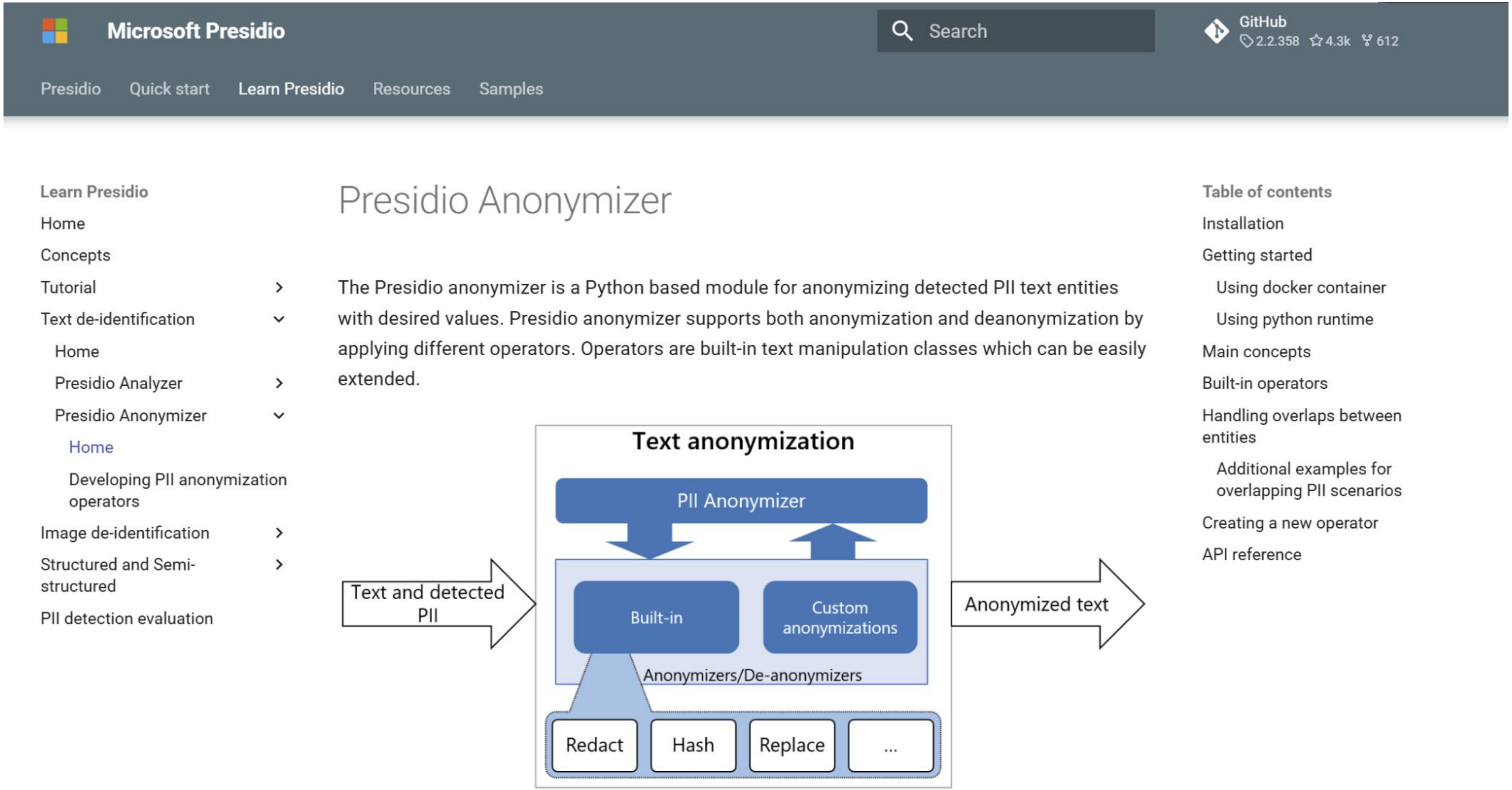
Using a small local LLM (llama 3.1 8B) to anonymize data for a remote big LLM (ChatGPT/Claude Sonnet)

Sacha Storz · Follow
5 min read · Aug 11, 2024

80

If you have data that includes sensitive information like names or other personal details, it's probably best not to send it to a remote LLM like ChatGPT or Claude Sonnet. With GDPR and other privacy regulations in play, it's better to be safe than sorry.

For instance, I often use these powerful LLMs to get well-structured summaries of interviews I conduct with clients. In the past, I would manually anonymize the interviews, replacing all names (whether of people or companies) with placeholders. Then, I would upload the anonymized content to ChatGPT, along with instructions on how to work with the material.



Microsoft Presidio

Search

GitHub
2.2.358 4.3k 612

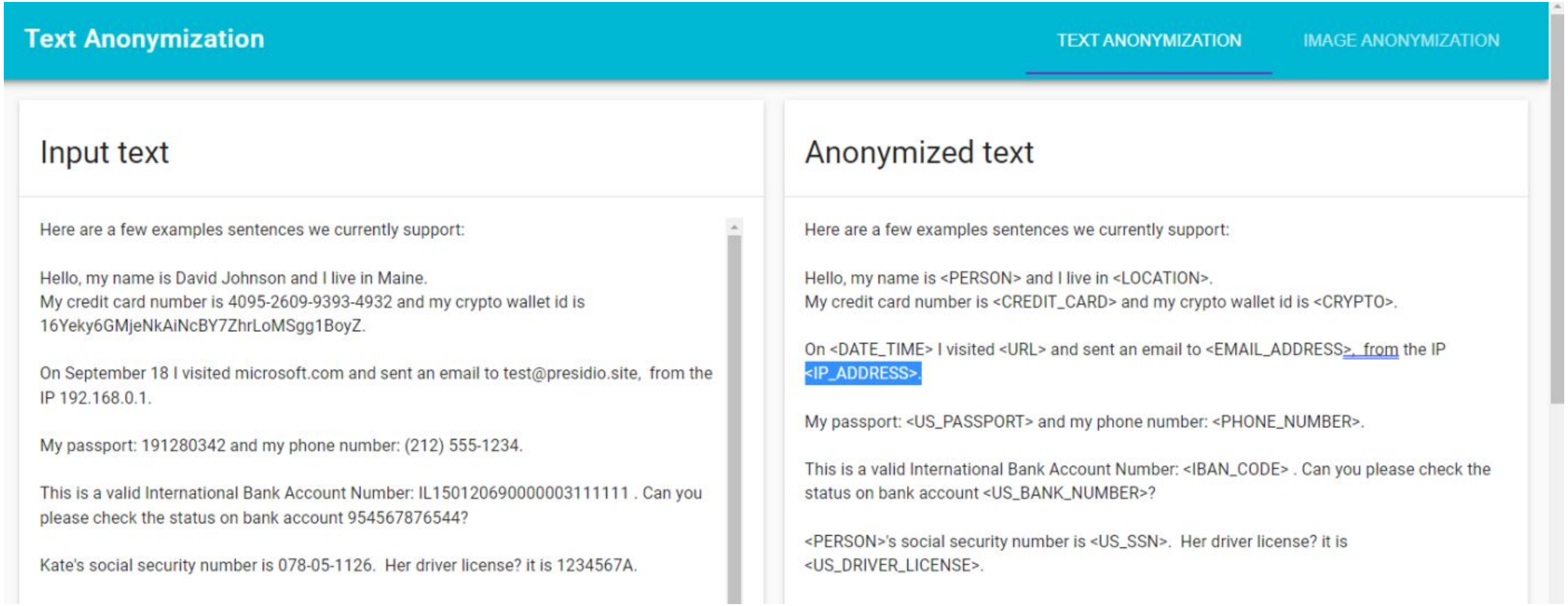
Presidio Quick start Learn Presidio Resources Samples

Learn Presidio
Home
Concepts
Tutorial
Text de-identification
Home
Presidio Analyzer
Presidio Anonymizer
Home
Developing PII anonymization operators
Image de-identification
Structured and Semi-structured
PII detection evaluation

Presidio Anonymizer

The Presidio anonymizer is a Python based module for anonymizing detected PII text entities with desired values. Presidio anonymizer supports both anonymization and deanonymization by applying different operators. Operators are built-in text manipulation classes which can be easily extended.

Table of contents



Input text

Here are a few examples sentences we currently support:

Hello, my name is David Johnson and I live in Maine.
My credit card number is 4095-2609-9393-4932 and my crypto wallet id is 16Yeky6GMjeNkAiNcBY7ZhrLoMSgg1BoyZ.

On September 18 I visited microsoft.com and sent an email to test@presidio.site, from the IP 192.168.0.1.

My passport: 191280342 and my phone number: (212) 555-1234.

This is a valid International Bank Account Number: IL150120690000003111111 . Can you please check the status on bank account 954567876544?

Kate's social security number is 078-05-1126. Her driver license? it is 1234567A.

Anonymized text

Here are a few examples sentences we currently support:

Hello, my name is <PERSON> and I live in <LOCATION>.
My credit card number is <CREDIT_CARD> and my crypto wallet id is <CRYPTO>.

On <DATE_TIME> I visited <URL> and sent an email to <EMAIL_ADDRESS>, from the IP <IP_ADDRESS>.

My passport: <US_PASSPORT> and my phone number: <PHONE_NUMBER>.

This is a valid International Bank Account Number: <IBAN_CODE> . Can you please check the status on bank account <US_BANK_NUMBER>?

<PERSON>'s social security number is <US_SSN>. Her driver license? it is <US_DRIVER_LICENSE>.

Zero-Knowledge Proof (ZKP)

A zero-knowledge proof (ZKP) is a cryptographic method that allows one party to convince another party that a statement is true, without revealing any information beyond the truth of the statement itself.

https://github.com/Xor0v0/awesome-zero-knowledge-proofs-security

Case 1

Someone says: **“I live in Tampa!”**

What can we infer?

- They are **likely from Florida** (Tampa is in Florida).
- They **may also be a U.S. citizen**, but we don’t know for sure.

The person hasn’t revealed **any document or ID** to **prove citizenship or state identity**.

Case 1

What if that person could prove they are from Florida, or even prove they are a U.S. citizen, **without** ever saying 'I live in Tampa' or showing their address or passport?

Zero Knowledge Proof

Case 2

A **patient visits a hospital for a medical procedure** that is only allowed for patients **above 18 years old**

Normally, the hospital would ask for:

- ID card
- Full date of birth
- Insurance details, etc.

⬅

This **reveals a lot of personal data** that isn’t necessary for just proving date birth.

Summary:

Zero-Knowledge Proof allows the users to prove themself without exposing private data.

It protects user privacy and reduces data leakage risks

Demo 1

Me:

A **ZKP for proving knowledge of a secret number x = 4**,

Random number r = 3 send to you

r = 3

You:
Generate a random value c = 5
Send to me!

Me:
s = r + c*x = 3 + 5 * 4 = 23
SEND 23 to you.

YOU:

$s=23$ total number

$x=4$ the secret

$r=3$ random number generated by me

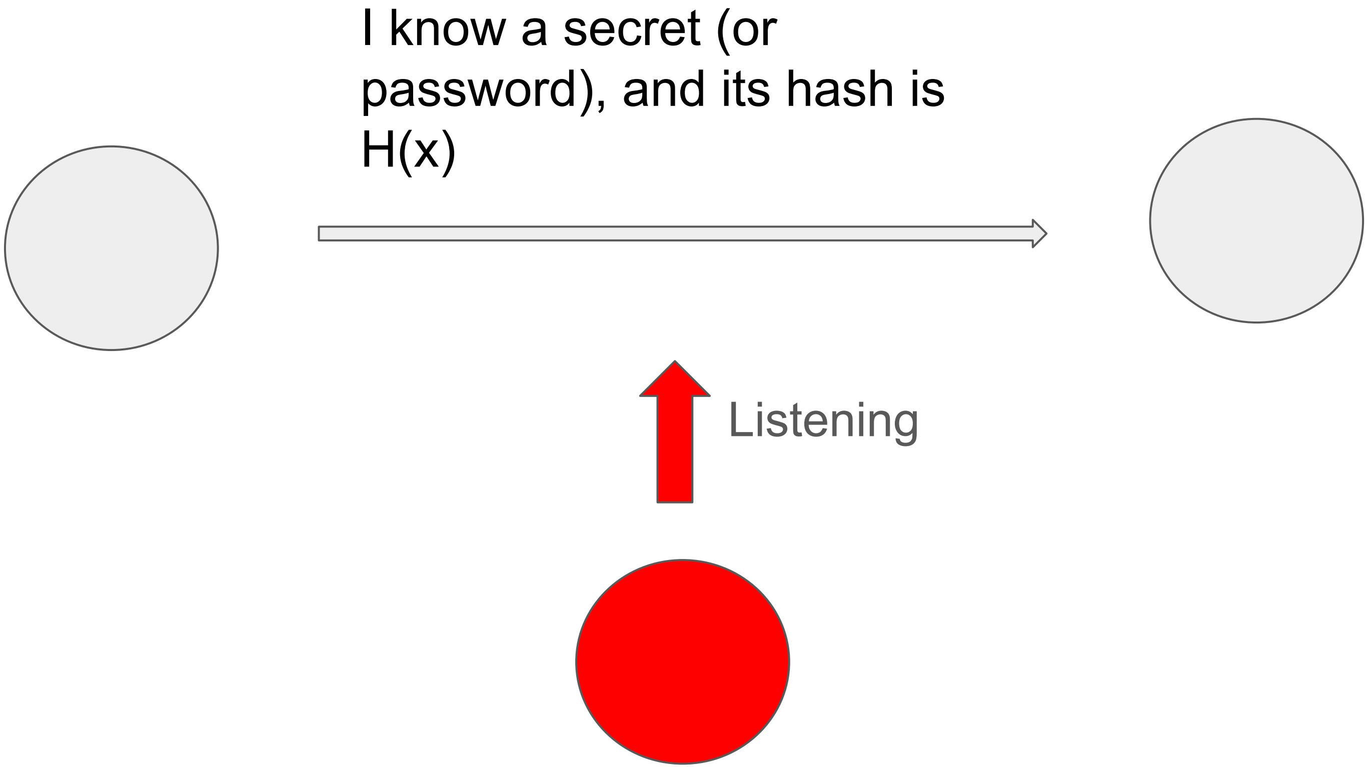
$c=5$ random number generated by you

$r+x*c=23$

So you understand me that I know it is 4.

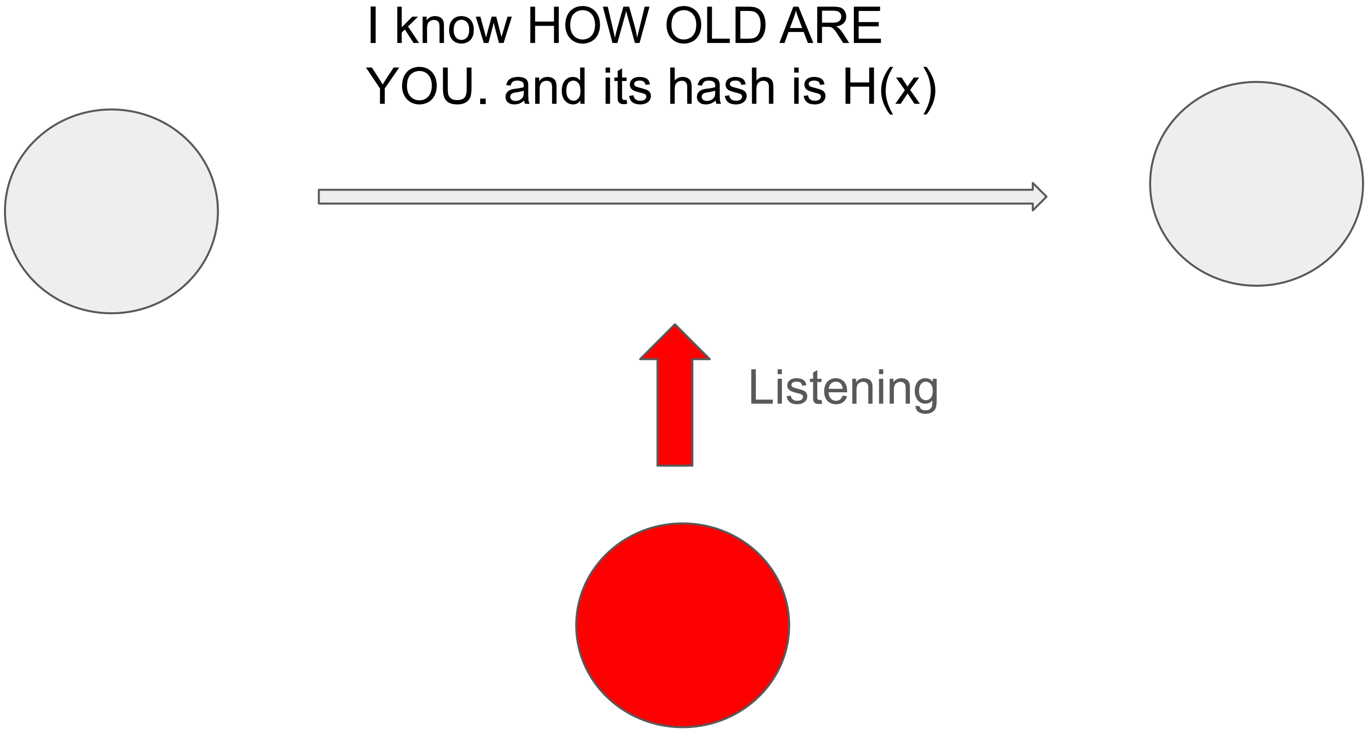
Demo 2： string

Question: If it is not numbers?



I know a secret (or password), and its hash is H(x)

Listening



I know HOW OLD ARE YOU. and its hash is H(x)

Listening

Question: Can we send this?

Applications of Zero-Knowledge Proof (ZKP) in Hospitals

Patient Identity Verification (Without Revealing Sensitive Information)

- Patients can prove they are **registered users**, **insured**, or **eligible for treatment**, without disclosing:
    - Name
    - ID number
    - Insurance details

✔ "I am a verified patient in the system."

Vaccination Status Proof (Without Revealing Medical History)

- Patients can prove they are **vaccinated**, without revealing:
  - When
  - Where
  - By whom
- Applicable in:
  - Admission protocols
  - Pre-surgery checks
  - Hospital staff onboarding

Privacy in Multi-Hospital Research Collaboration

- Hospitals collaborating in research can prove that:

  "A patient meets study criteria (e.g., has condition X),"
  without revealing detailed medical records.

- Enables privacy-preserving medical research.

Consent Proof for Data Access (Digital Authorization + ZKP)

- Patients can issue **digital consent proofs** saying:

  "Doctor A can access my data for the next 7 days," without disclosing the full content of the consent form.

- Helps protect patient rights while maintaining access control.

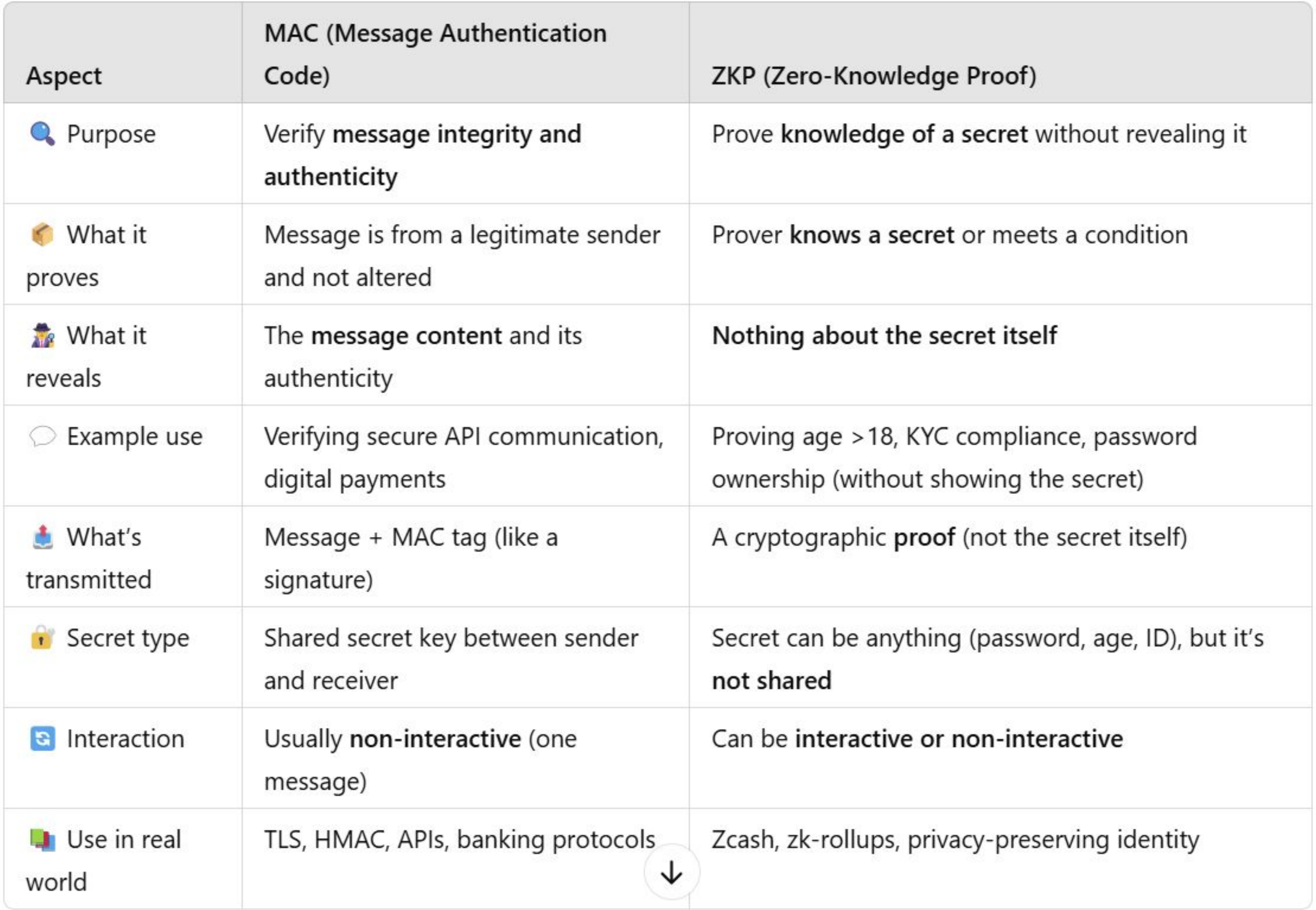
Applications of Zero-Knowledge Proof (ZKP) in finance

Private Transactions (Confidential Payments in Blockchain)

- ZKP enables **transactions to be verified without revealing amounts or parties involved**.
- Used in **privacy coins** like **Zcash**, where:
  The network verifies the transaction is valid, but no one knows how much was sent or who the sender/receiver is.

- Applicable in:
  - Private asset transfers

MAC and ZKP



Aspect	MAC (Message Authentication Code)	ZKP (Zero-Knowledge Proof)
🔍 Purpose	Verify **message integrity and authenticity**	Prove **knowledge of a secret** without revealing it
📦 What it proves	Message is from a legitimate sender and not altered	Prover **knows a secret** or meets a condition
🕵️ What it reveals	The **message content** and its authenticity	**Nothing about the secret itself**
💬 Example use	Verifying secure API communication, digital payments	Proving age >18, KYC compliance, password ownership (without showing the secret)
📤 What's transmitted	Message + MAC tag (like a signature)	A cryptographic **proof** (not the secret itself)
🔐 Secret type	Shared secret key between sender and receiver	Secret can be anything (password, age, ID), but it's **not shared**
🔄 Interaction	Usually **non-interactive** (one message)	Can be **interactive or non-interactive**
📚 Use in real world	TLS, HMAC, APIs, banking protocols	Zcash, zk-rollups, privacy-preserving identity

Property	**Meaning**
Completeness	If the prover is honest, the verifier will be convinced.
Soundness	If the prover is lying, they will be caught with high probability.
Zero-Knowledge	The verifier learns nothing about the secret, only that it’s correct.

Interactive vs Non-Interactive ZKP

Type	Explanation
Interactive ZKP	Prover and verifier communicate in several rounds.
Non-Interactive ZKP (NIZK)	Prover sends one proof. Verifier can check it anytime, no back-and-forth needed.

ZKP Limitations & Challenges

ZKP + Hash function is strong enough to solve some problems — especially when the secret is a fixed value, like a password, username, or ID number.

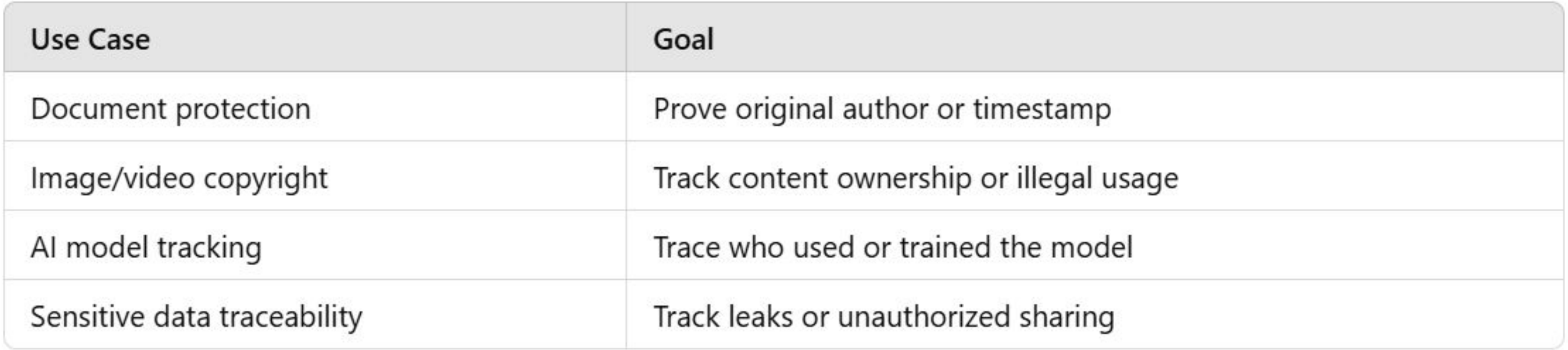
However, if the secret is a sentence or free-form text, it becomes harder to calculate or match the correct hash value — because you may not know the exact content or structure of the sentence. Even a small difference in the sentence will result in a completely different hash value.

For CS

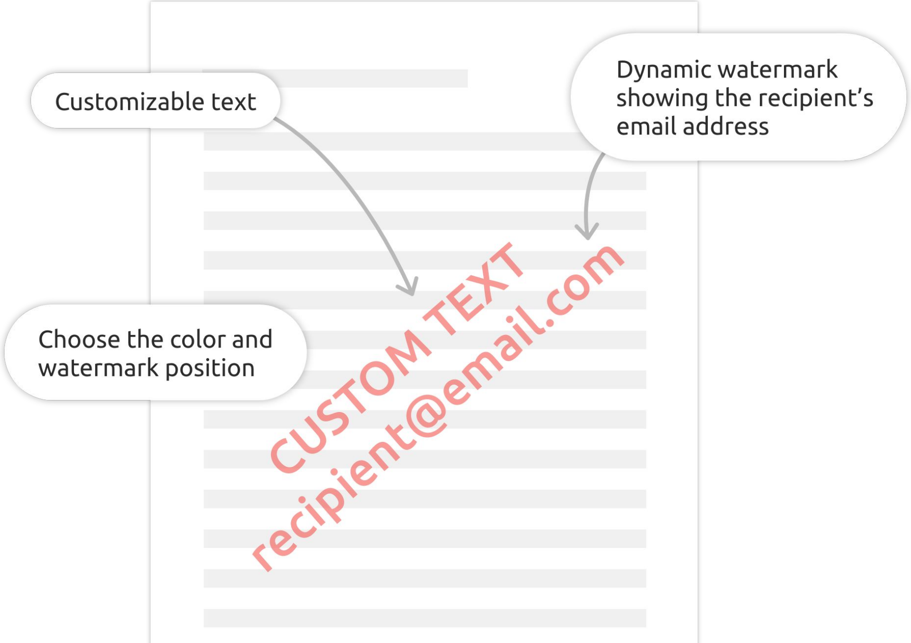
https://github.com/Xor0v0/awesome-zero-knowledge-proofs-security

Water marking solution

Watermarking is a technique used to **embed hidden information (a mark or ID) into data** — such as documents, images, videos, models, or even AI-generated content — to prove **authorship, integrity, or tracking usage**, without visibly altering the data.



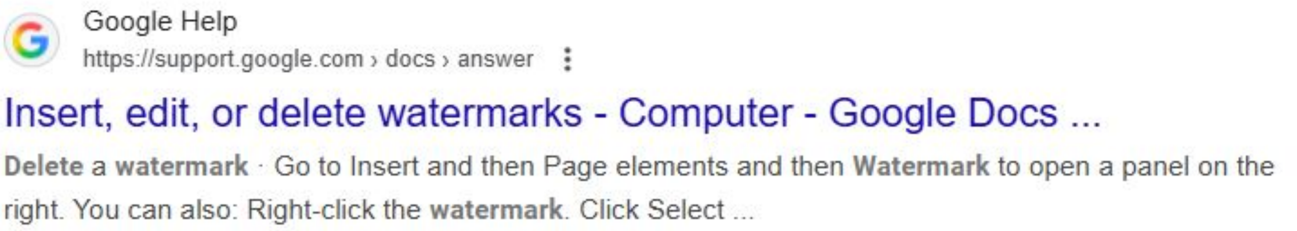
Use Case	Goal
Document protection	Prove original author or timestamp
Image/video copyright	Track content ownership or illegal usage
AI model tracking	Trace who used or trained the model
Sensitive data traceability	Track leaks or unauthorized sharing



Customizable text

Dynamic watermark showing the recipient’s email address

Choose the color and watermark position



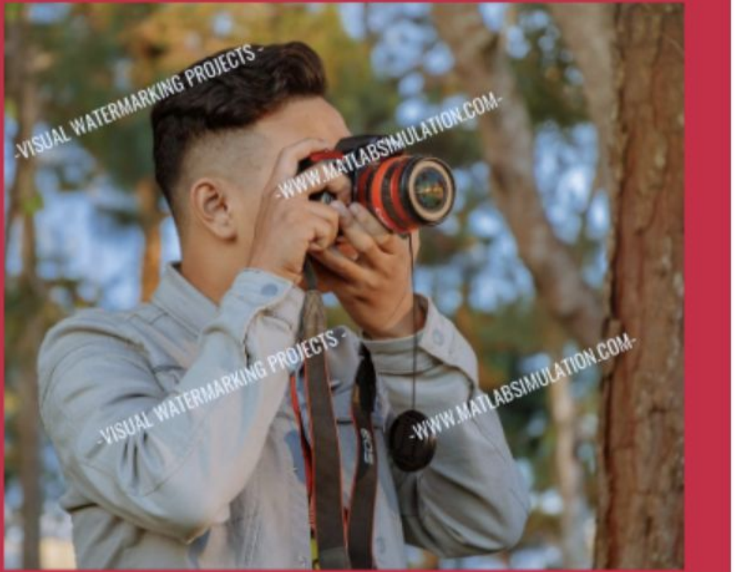
Google Help
https://support.google.com › docs › answer

Insert, edit, or delete watermarks - Computer - Google Docs ...

Delete a watermark · Go to Insert and then Page elements and then **Watermark** to open a panel on the right. You can also: Right-click the **watermark**. Click Select ...

NORMAL IMAGE

VISUAL WATERMARKING PROJECTS


-VISUAL WATERMARKING PROJECTS -
-WWW.MATLABSIMULATION.COM-
-VISUAL WATERMARKING PROJECTS -
-WWW.MATLABSIMULATION.COM-

Matlab Simulation
Technology with Trust

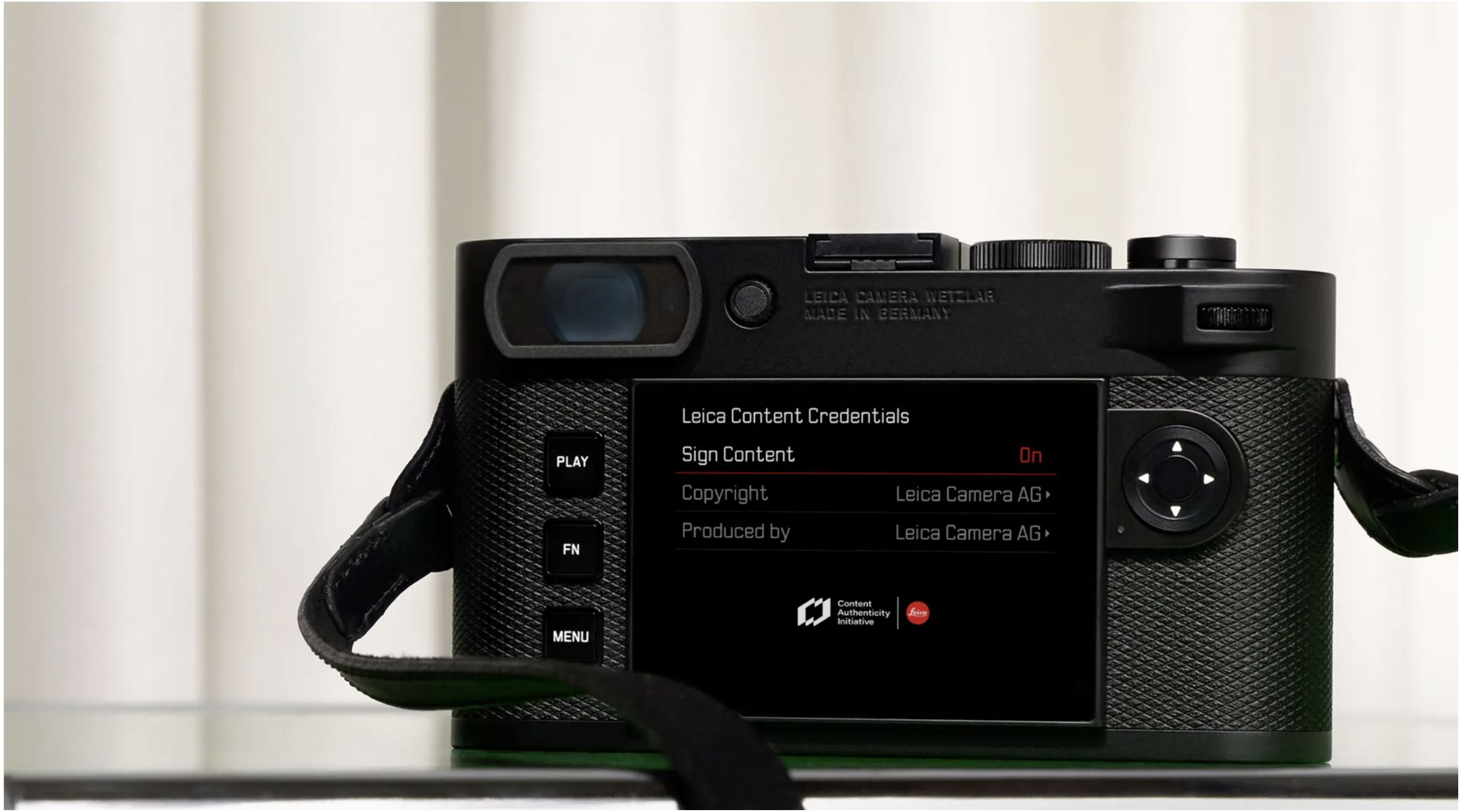
VISUAL WATERMARKING PROJECTS

WWW.MATLABSIMULATION.COM

Researchers are thinking about how to pretect the ownership of the photos

Products Experience Leica Company Service

Content Credentials



Leica Content Credentials

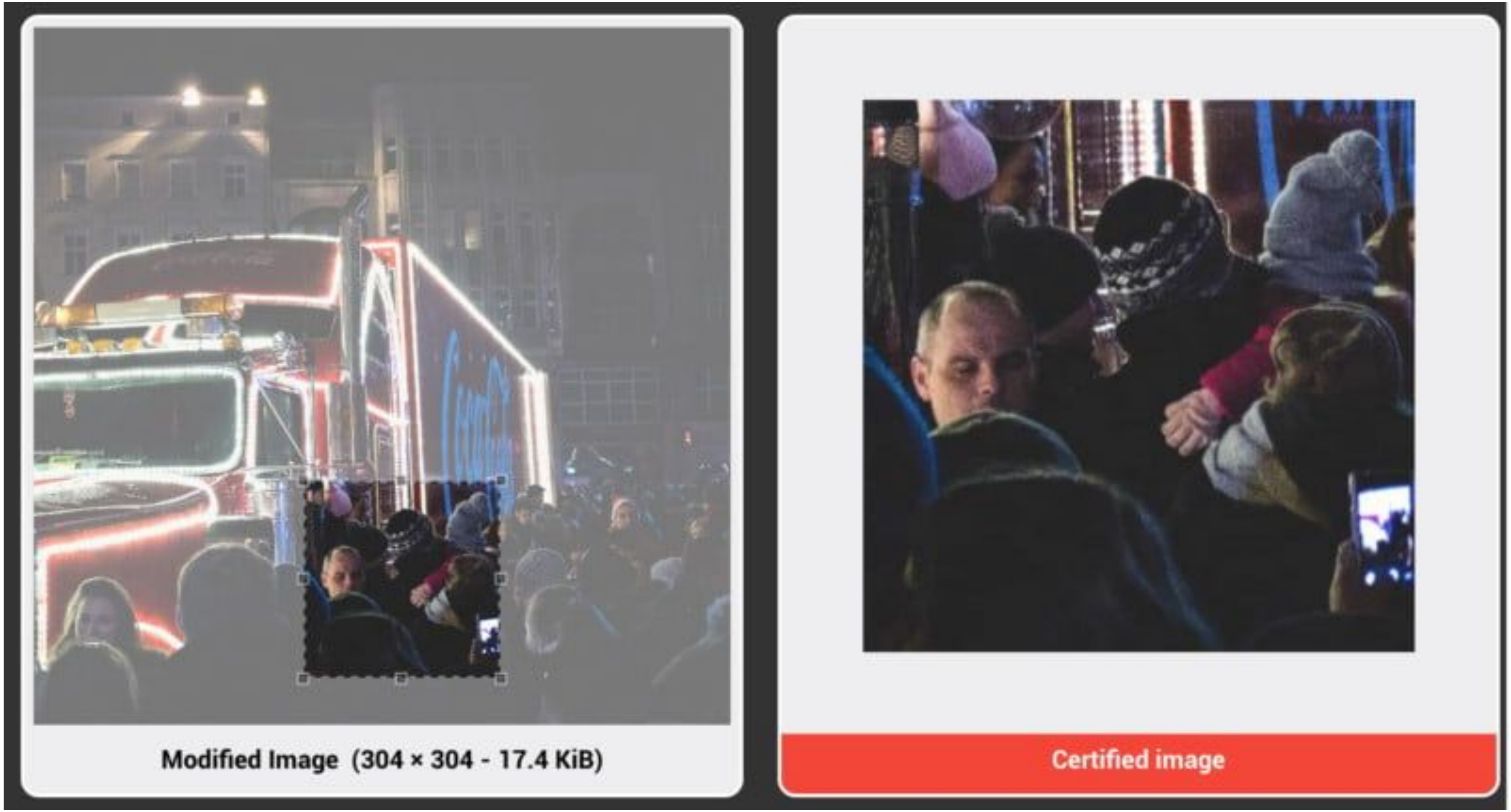
Sign Content On

Copyright Leica Camera AG

Produced by Leica Camera AG

Original Image

Watermarked Image


Modified Image (304 × 304 - 17.4 KiB)
Certified image

For CS

https://github.com/Stability-AI/invisible-watermark-gpu?utm_source=chatgpt.com

Feature	Digital Signature	Watermarking
Purpose	Prove sender identity (non-repudiation)	Embed ownership or tracking info inside content
Core Technology	Hash + private key encryption	Embedded signal/data in content (e.g., image, audio)
Attached or Embedded?	Attached (external tag)	**Embedded inside the content itself**
Visibility	Invisible	Can be visible or invisible
Used in	Legal document signing, blockchain transactions	Image/video copyright, medical records, AI models
Tamper Detection	Yes	Some resistance, but not cryptographic-grade
Can identify creator?	Yes (linked to private key owner)	Yes (tracks original owner or creator)

Applications of Watermarking in Hospitals

Watermarking in healthcare is about **embedding hidden, secure, or visible information** directly inside medical data (like images, reports, or documents), for the purposes of **provenance tracking, ownership, security, and data integrity**.

Medical Image Ownership & Copyright Protection

- Embed **hospital ID, radiologist ID, or department code** directly into medical images (e.g., X-rays, MRIs, CT scans).
- Helps prove **who generated the image**, especially when images are shared across systems or institutions.
- Prevents **unauthorized reuse or plagiarism** in research or publications.

Is a blockchain a form of watermarking?

What’s Similar?

They both can:

- Prove **ownership/authorship**
- Help with **data integrity**
- Enable **traceability**

But they **do it in different ways**.

Blockchain = "Record proof on-chain at this time"
Watermarking = "Embedding proof directly in the file or model"