Data Re-identification

in the Age of AI

Cytopia Conference Cycle 1

May 12, 2026

Speaker: Chi Tran

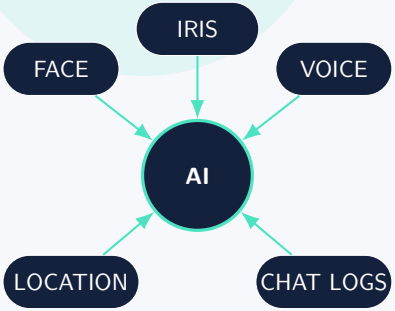
AI knows you. Even when it shouldn't.

AI is everywhere, and personal data can be collected, inferred, and re-identified faster than you think.

Real-world examples

Face recognition at airports, voice assistants at home, phone location, and chat logs: everyday AI can leave identity traces.

- AI does not just "see" you; it connects signals from multiple sources to infer identity.
- The risk is rarely one data point. It is the ability to link enough of them together.



Identity exposure grows

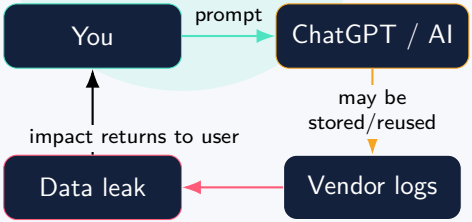
From Convenience To Leakage

Prompts and uploads can expose sensitive data.

A common workflow risk

Users paste text, upload files, and share logs for convenience, but those inputs may persist in systems outside their direct control.

- Samsung engineers leaked source code through ChatGPT (2023).
- CVs, emails, and personal details are frequently pasted into AI tools.
- Leaked prompts can become another source for linking people, behavior, and identity.



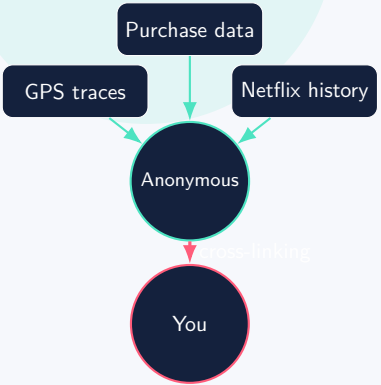
What Is Re-identification?

Anonymous-looking data plus cross-linking can reveal identity.

Short definition

Re-identification is the process of combining multiple harmless-looking data points to infer a specific person.

- A dataset may remove names, but timestamps, location, behavior, and metadata can still remain.
- Once combined with another source, the real identity can emerge.



Examples

GPS traces, purchases, and Netflix history can point back to you when combined.

Who Tries To Re-identify Data?

The same technique can be used for safety, profit, or harm.

Legitimate

Privacy teams test if a dataset is safe to share.

- Privacy testing
- Bias audit
- Disclosure risk

Gray area

Companies link app data, public records, and profiles.

- Advertising
- Surveillance
- Monetization

Malicious

Attackers use re-identified data to scam or harass people.

- Fraud
- Targeting
- Stalking

Different motives, same method: linking data back to real people

AI Amplifies The Risk

AI makes linking, inference, and scale much faster than humans can manage.

Memory	Linking	Probing
AI systems may reproduce pieces of data they have seen.	Public photos, posts, and profiles can be connected across platforms.	Repeated questions can reveal hidden data or sensitive patterns.

When identity is exposed, the harm becomes personal.

Real Damage, Real People

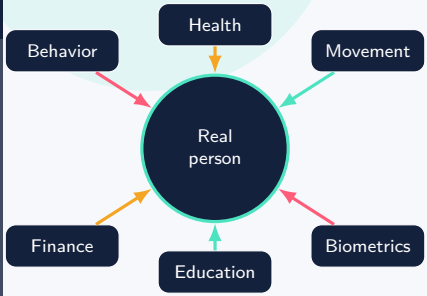
The consequences are not theoretical anymore.

Three memorable examples

- AOL search logs (2006): search history exposed a real person's identity.
- Netflix Prize (2008): anonymous movie ratings could be matched back to users.
- Location data: phone traces can reveal homes, workplaces, and routines.

What it means

Re-identification can lead to lost privacy, stalking, loss of trust, and legal risk for companies.



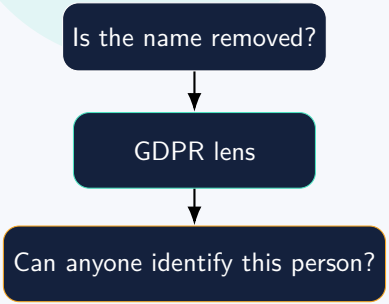
Different traces can point back to someone

GDPR: The Rules of The Game

The right question is not "did we remove the name?" but "can anyone re-identify this with reasonable effort?"

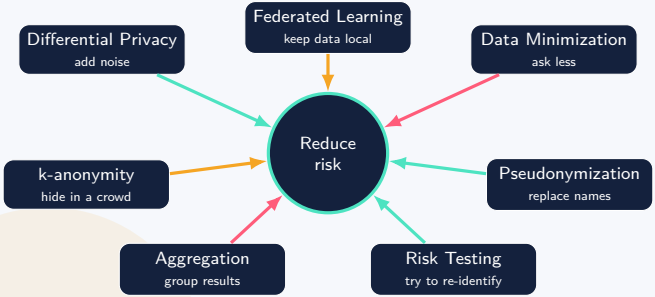
The GDPR lens

- If a person can still be identified, the data may still be personal data (Art. 4).
- Identifiability includes what others can reasonably link together (Recital 26).
- Privacy must be built in by design and by default (Art. 25).
- Serious violations can lead to major fines (Art. 83).



A Privacy Defense Toolbox

Some methods to reduce what can be linked back to you.



Q&A

AI is convenient, but privacy is never safe by default.

Thank you