PRIVACY & AI LAW — S06E02

PII vs Personal Data, Special Categories, Pseudonymisation & Retention

A Developer's Blueprint

General information, not legal advice – current as of November 2026.

Gemini Notebook

The Episode 1 Recap & The Intuition Bug

```
/* Recap Ep 1: We covered GDPR roles, 6 lawful bases (Art. 6), and 72-hour breach rules. */
```

Code Review: PII Sanitization Check

Left Side (The Pull Request)

```
{
  "user_uuid": "8f4...",
  "dynamic_ip": "192.0.2.1",
  "hashed_email": "e3b0c..."
}
```



Developer Comment

Stripped all names and SSNs. Zero PII here. Safe to log forever.

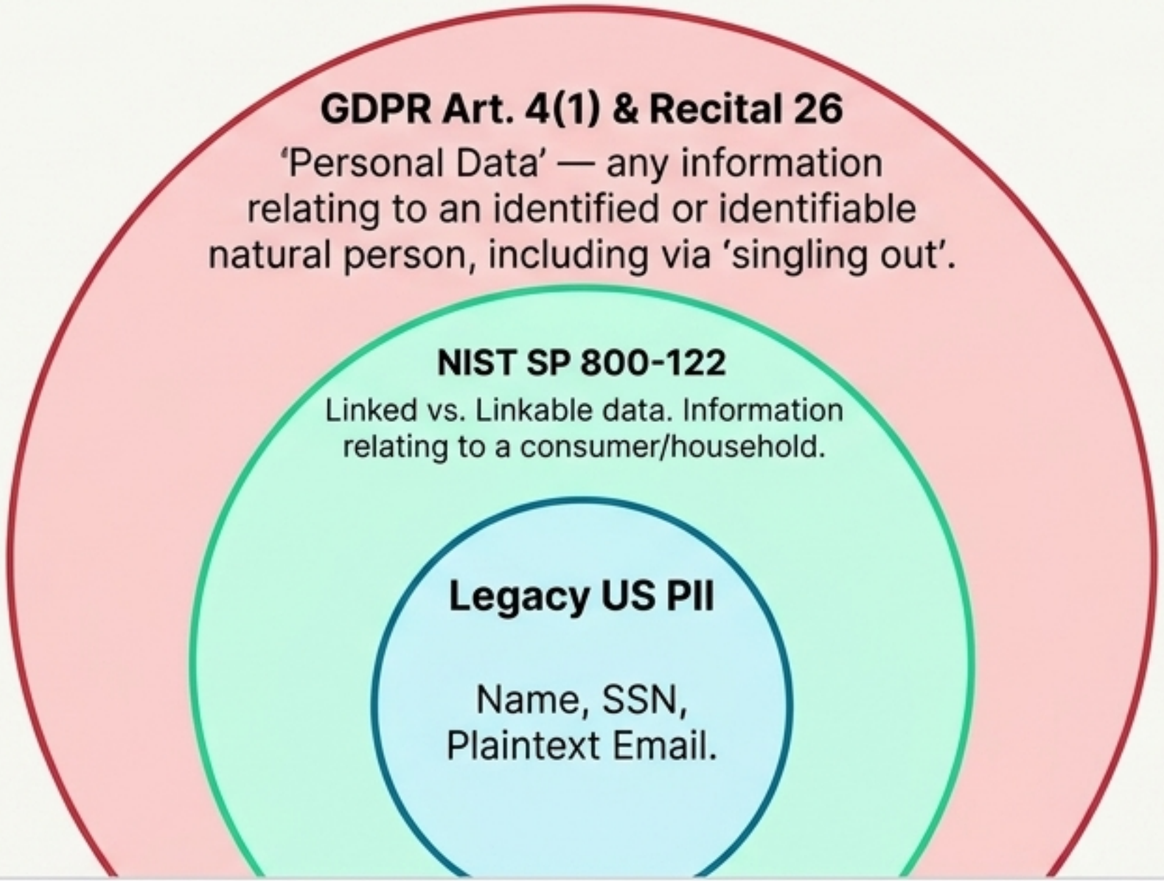
Right Side (The Syntax Error)

CRITICAL ERROR

The Bug: Relying on the legacy, narrow definition of 'PII' (Name + SSN + Email) in a modern privacy regime.
In the eyes of GDPR, this payload is fully regulated personal data is fully regulated personal data.

Gemini Notebook

Legacy PII vs. Modern Personal Data



⚠ The Takeaway: Stop searching for strings labeled 'Name'. **If a system can single a user out, track them, or link records to them** over time, it is **Personal Data.**

Gemini Notebook

The Expanding Radius of Identifiers (Art. 4(1))

```
// POST /api/v2/user/profile
{
  "user": {
    "direct_id": { full_name": "Alice Smith", "email": "alice.smith@example.com" },
    "telemetry_id": { "device_id": "d8c7a9…2f1", "ad_id": "IDFA_634b…." },
    "inferred_profile": { "segment": "high-income_urban_fitness", "score": 88 }
  }
}
```

Direct Identifiers

The obvious layer.

```
-- 'Direct' Tags:
[tag: Name]
[tag: government ID]
[tag: plaintext email]
```

Online / Telemetry Identifiers

(Recital 30)

```
-- 'Online' Tags:
[tag: Cookies]
[tag: device_id]
[tag: app_uuid]
[tag: advertising IDs (IDFA/AAID)]
[tag: RFID tags]
```

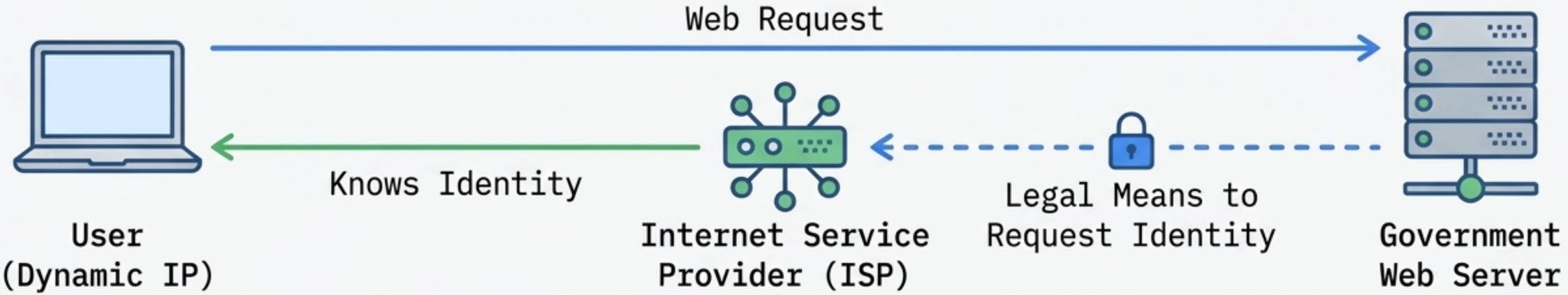
Indirect Identifiers & Inferences

“UK ICO guidelines: Inferred profiles (e.g., ‘high-income urban fitness enthusiast’) are personal data even if the underlying identity is completely hidden.”

```
-- 'Inference' Tags & Examples:
[tag: Inferred profiles]
[example: 'high-income urban fitness enthusiast']
[tag: Linkable data]
[tag: Behavioral score]
```

Gemini Notebook

The Breyer Precedent: Dynamic IPs as Personal Data



The Scenario (CJEU C-582/14)

A web server logs a user's dynamic IP address. The server operator has no idea who the user is.

The Ruling

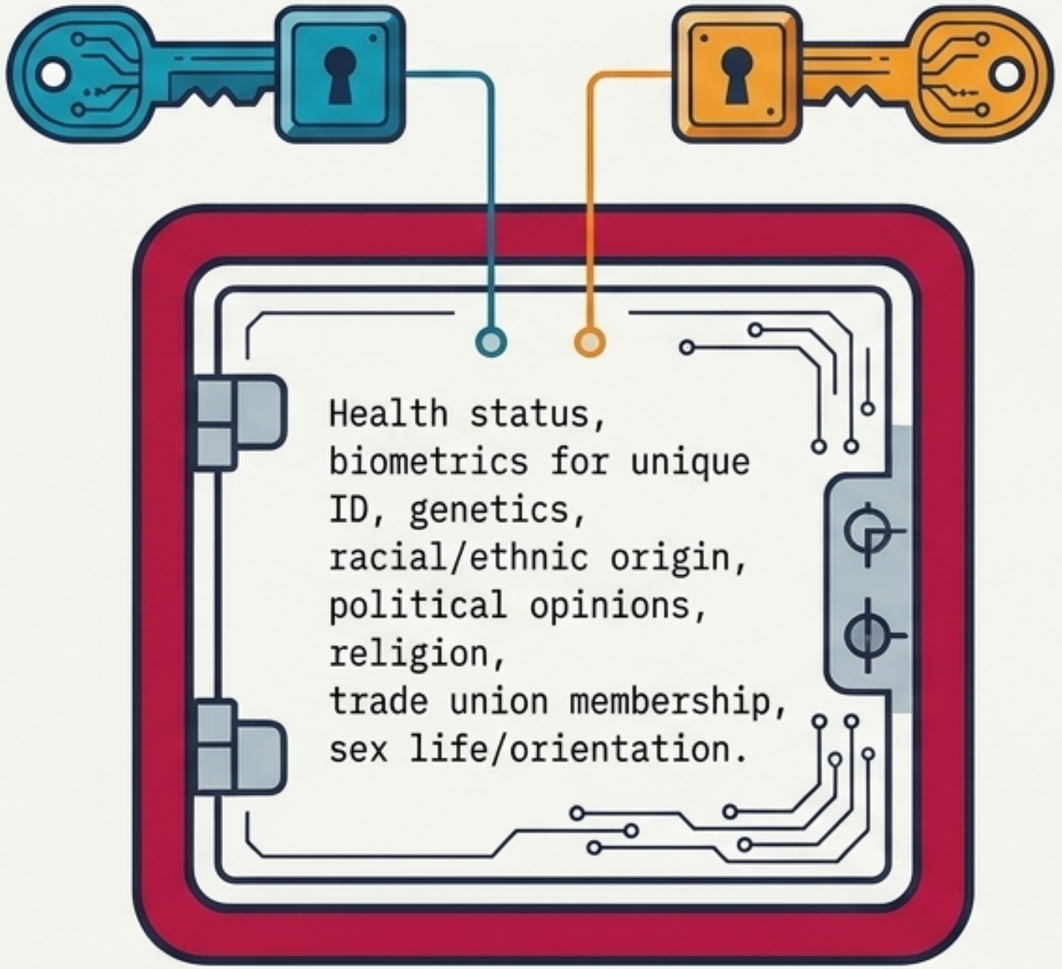
Because the ISP holds the linkage data, and the server operator has the legal means (e.g., cyberattack investigation) to ask the ISP to identify the user, the dynamic IP is personal data for the server operator operator.

Rule of Thumb: "Identifiable" includes indirect identification via third parties.

Gemini Notebook

High-Alert: Special Category Data (GDPR Art. 9) & US Parallels

Art. 6
Lawful Basis



Art. 9(2) Condition
(e.g., explicit consent, vital interests, employment law)

The Dual-Lock Rule

Standard data needs one key (Art. 6).

Art. 9 data requires **BOTH** an **Art. 6 basis** AND an **Art. 9(2) condition**.

US Parallels: Closely maps to California's 'Sensitive Personal Information' and triggers immediate regulatory scrutiny.

Gemini Notebook

Vulnerable Subjects: Children’s Data & Biometrics

Warning Modal



GDPR Art. 8 (Children)

Requires verifiable parental consent for information society services.

└ **Trap:** The age of digital consent varies by Member State (between 13 and 16 years old).

Warning Modal



US FTC Biometric Scrutiny

The FTC heavily polices facial recognition, voiceprints, and gait analysis.

└ Merely hashing biometric templates does not remove the compliance burden if unique identification remains possible.

Gemini Notebook

The Identifiability Spectrum

Direct Personal Data	Pseudonymised Data	US Deidentified Data	Irreversibly Anonymised
Law Art 4(1) Examples Names, emails. Status Fully regulated. Controls Encryption at Encryption at rest/transit.	Law Art 4(5) Examples Keyed HMACs, UUIDs. Status Fully regulated. Great security, but still personal data. Controls Isolate secrets, tokenization.	Law Cal. Civ. Code / NISTIR 8053 Status Often outside scope *if* strict technical/contractual non-reidentification limits are in place.	Law Recital 26 Examples Highly aggregated statistics. Status Outside GDPR entirely.

Gemini Notebook

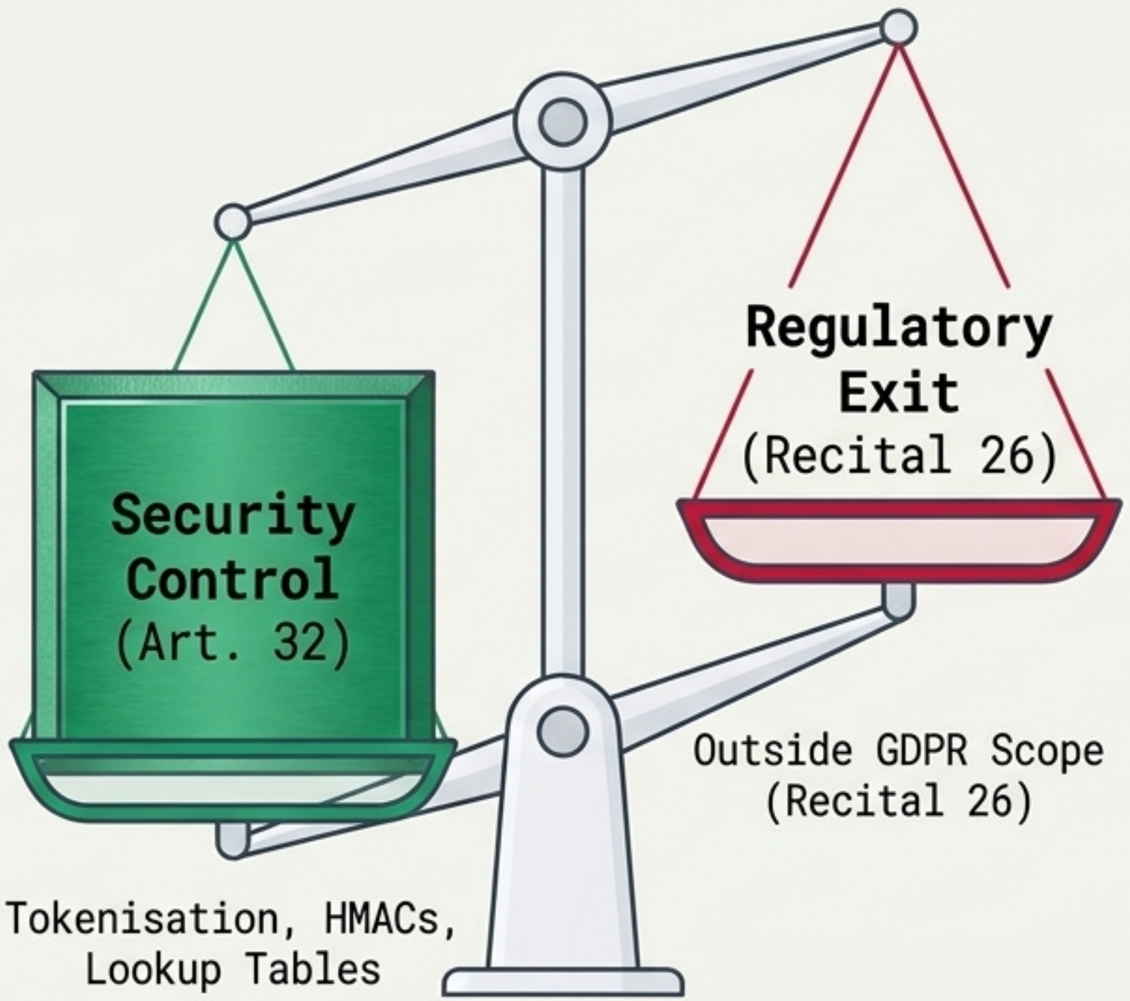
Pseudonymisation (Art. 4(5)) vs. True Anonymisation

The EDPB/ENISA Reality Check

EDPB Guidelines 01/2025 and ENISA are clear.

Tokenisation, isolated lookup tables, and keyed HMACs are brilliant security controls.

They reduce risk.



⚠ The Catch

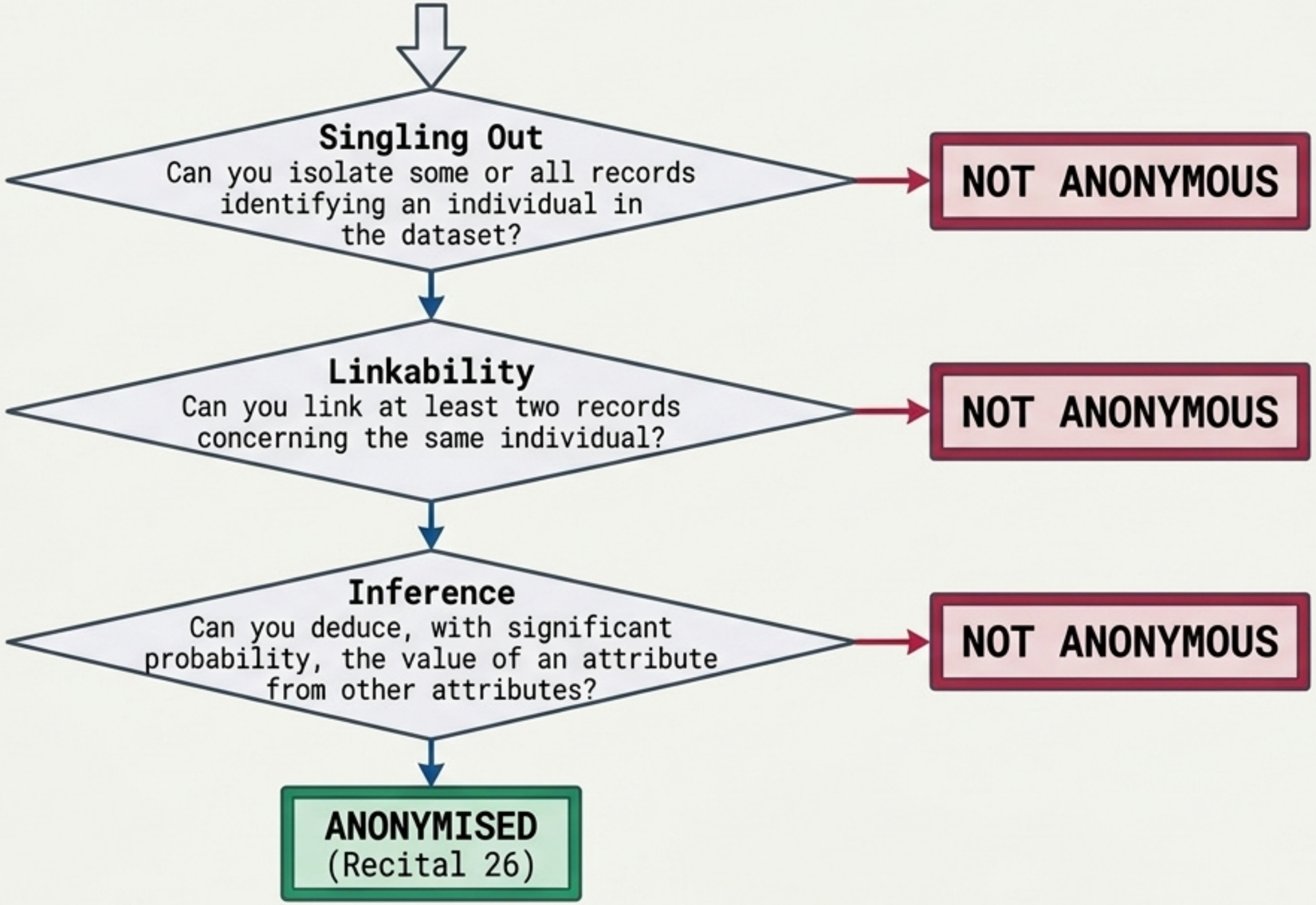
It remains regulated personal data.

If the lookup key exists anywhere in your organization, the pseudonymised rows are still in scope for GDPR DSARs, breaches, and deletion requests.

Gemini Notebook

The Anonymisation Threshold: 3 Tests

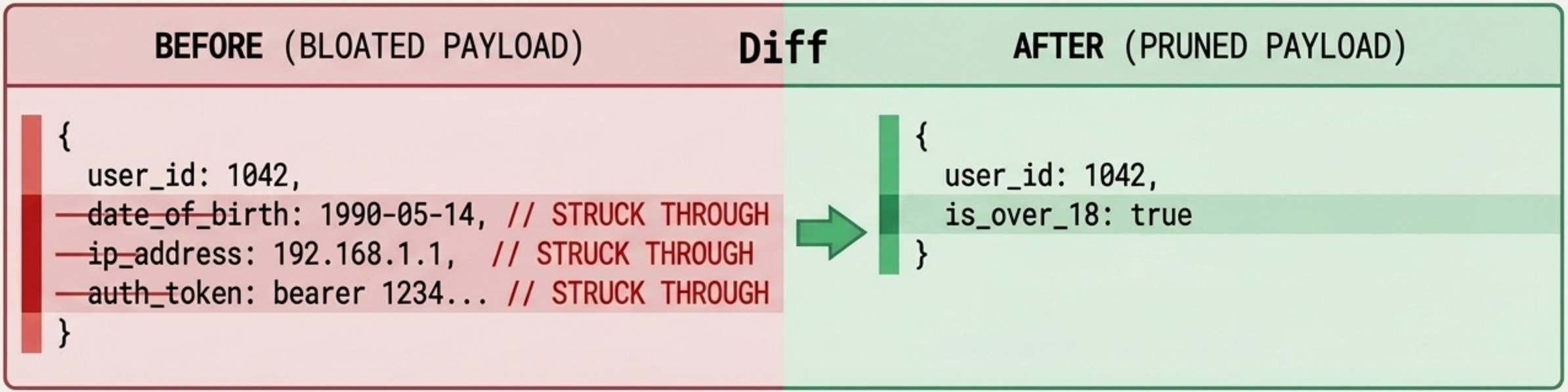
Threat Model: UK ICO ‘Motivated Intruder’ Test (WP216)



Gemini Notebook

Engineering Art. 5: Purpose & Data Minimisation

BEFORE (BLOATED PAYLOAD) **Diff** **AFTER** (PRUNED PAYLOAD)



The Code (EDPB 4/2019 & CPPA 2024-01)

Prune form fields aggressively. Do not collect date_of_birth if you only need a boolean age check.

Log Hygiene

Strip IP addresses and auth tokens from application logs before they hit your centralized logging systems (Datadog, Splunk).

Gemini Notebook

Engineering Art. 5: Storage Limitation & Cryptographic Erasure



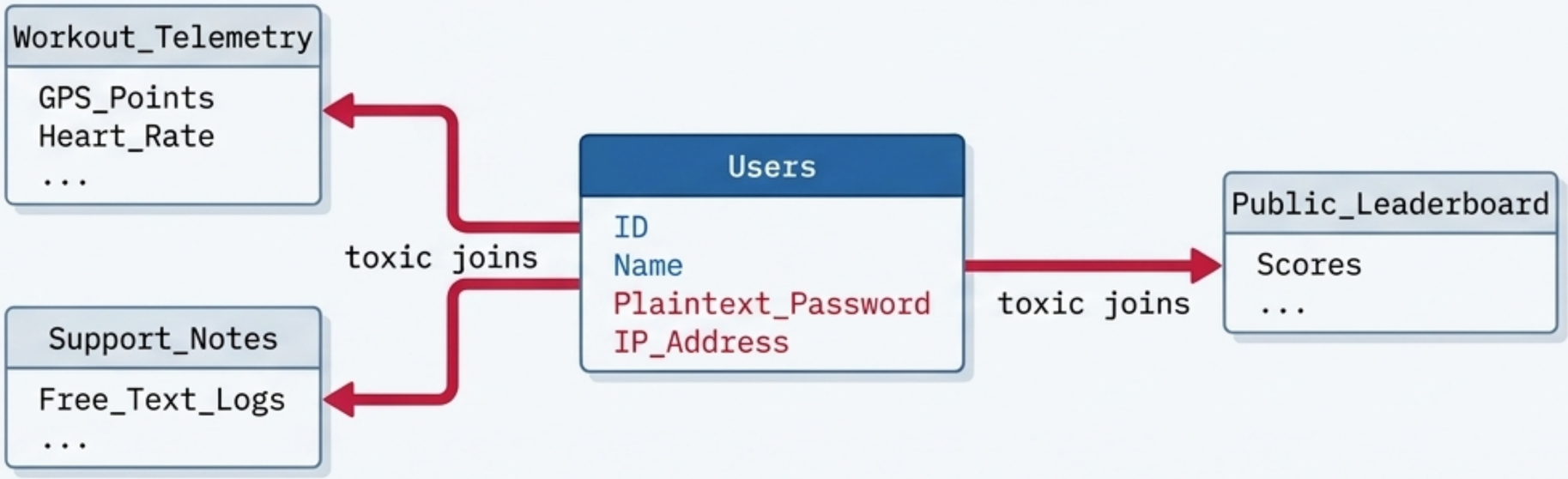
Collection → Storage → TTL (Time-To-Live) Trigger → Cryptographic Erasure

The Law	The Execution (CNIL / OWASP)
Art. 5(1)(e) requires keeping data no longer than necessary.	1. Implement automated **TTL** (Time-To-Live) cron jobs on database rows. 2. **NIST SP 800-88 Rev. 2**: Use **Cryptographic Erasure**. Encrypt a tenant's data with a unique **KMS key**. When they delete their account, delete the key. The data is instantly rendered irrecoverable, bypassing the need to hunt down individual **database rows** across backups.

Gemini Notebook

Architecture Audit: The Fitness App Scenario (Before)

Anti-Pattern: The Monolith



⚠ The Risk

Account authentication fields are stored in the same space as high-frequency telemetry and free-text notes. A single SQL injection or unauthorized dashboard access exposes highly sensitive, linked personal data.

Gemini Notebook

Isolating Identity Vaults from Event Tables



The Fix

Separate the strict Identity Vault from the event store. Connect them only via a pseudo-random, rotating app_uuid. The analytics team querying the telemetry table sees only the UUID, protecting the identity vault. (Compliant with EDPB 01/2025).

Gemini Notebook

Scrubbing Accidental Inferences & Support Logs

Support Log Parser - Regex Checker (Light Mode)

Ticket #9942: App keeps crashing when I log my rehab for recovering from knee surgery.

Art. 9 Health Data Detected

⚠ The Trap:

Free-text fields. User support logs or notes frequently contain accidental Art. 9 health data (e.g., injuries, pregnancies, medications).

🩹 The Fix:

- 📝 Treat all free-text fields as toxic.
- 🛠️ Apply aggressive TTLs (delete support logs after 30 days).
- 🔒 Restrict access via strict RBAC (Role-Based Access Control).
- 📱 Never feed raw customer support text into unsanitized LLM training pipelines.

Gemini Notebook

Architecture Audit: The Fitness App Scenario (After)

Auth Vault

Isolated, encrypted at rest, strict IAM limits.

Identity Data (PII)
Name
Email
Billing

Telemetry

TTL=90d

Pseudonymised via tokens, protected by Art. 32 security.

Telemetry Data
GPS points
Heart rate
Pseudonym (Rotating app_uuid)

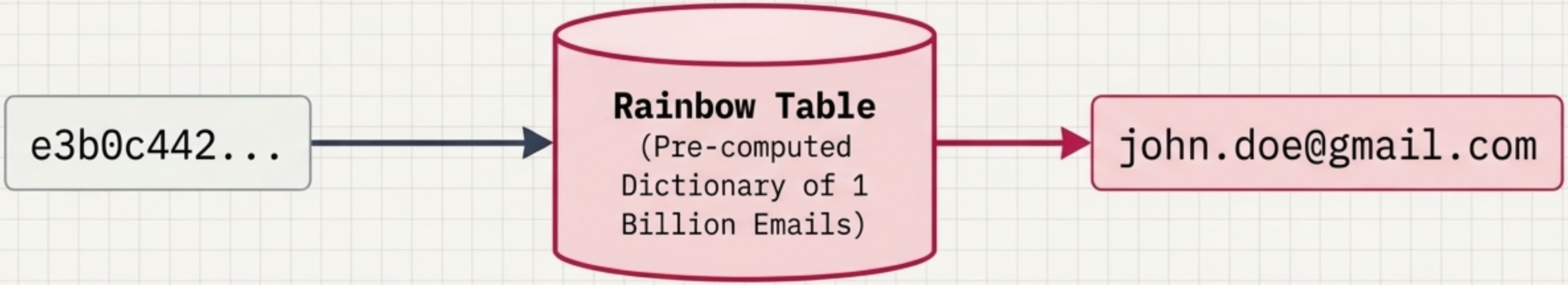
Public Leaderboards

Aggregated to pass the WP216 Motivated Intruder tests, achieving true Recital 26 anonymisation for public release.

Aggregated Data
Top Global Rankings
Weekly Steps Averages
City-level Activity Scores

Gemini Notebook

Tech Trap 1: Why Hashing Emails is NOT Anonymisation



The ENISA Warning

Using unsalted SHA-256 on structured, predictable fields (emails, phones) is mathematically broken.

The Reality

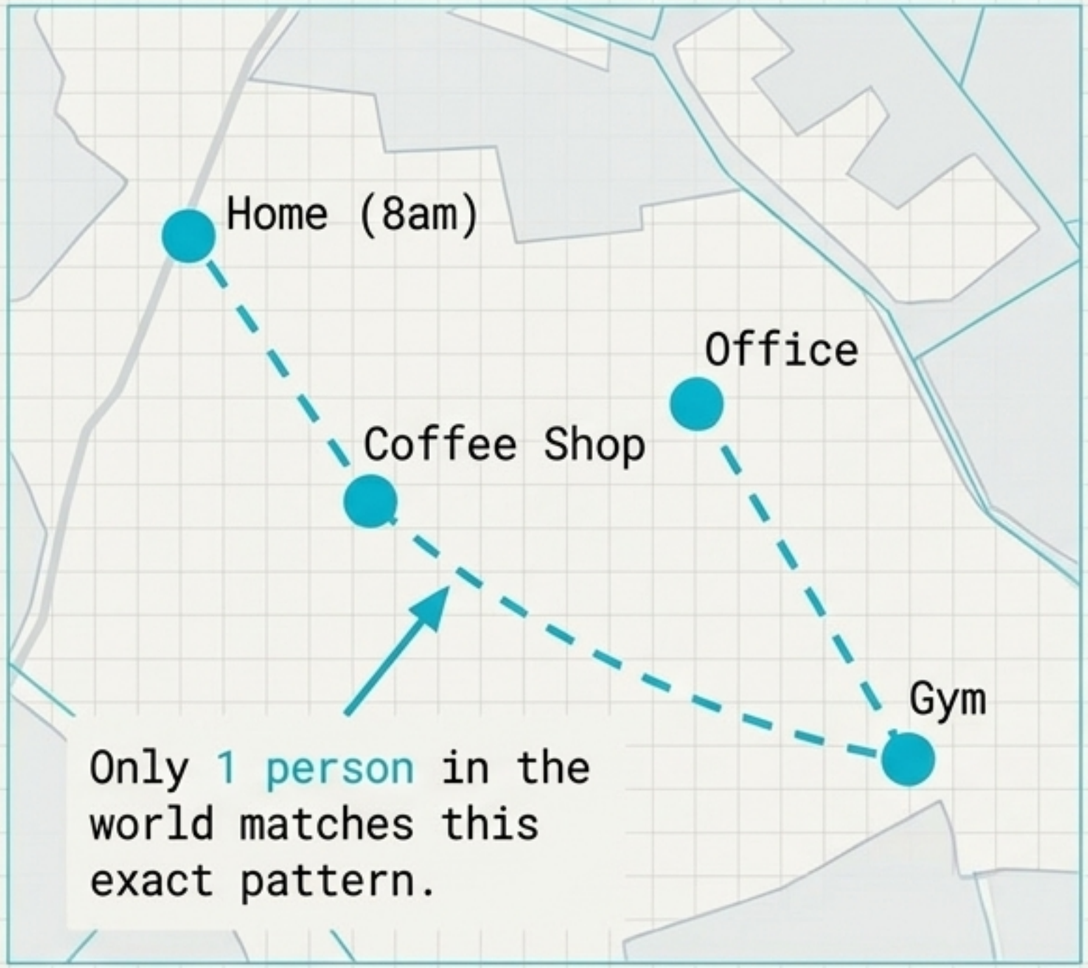
Because the input space of emails is small and predictable, attackers use rainbow tables to reverse the hash instantly.

The Rule

Hashed user rows are STILL personal data. If you must hash for lookup, use a strong salt (HMAC) and isolate the secret key.

Gemini Notebook

Tech Trap 2: Re-identifying Sparse Datasets



The Quasi-Identifier Trap

Datasets with high-dimensional event logs (time, location, unique behaviors) are nearly impossible to truly anonymise.

WP216 Reality Check

`4 spatial-temporal points` are enough to uniquely identify **95%** of individuals.

The Trade-off

You must choose between `k-anonymity` / `differential privacy` (which destroys data utility for precise analytics) or treating the dataset as heavily regulated `pseudonymised data`.

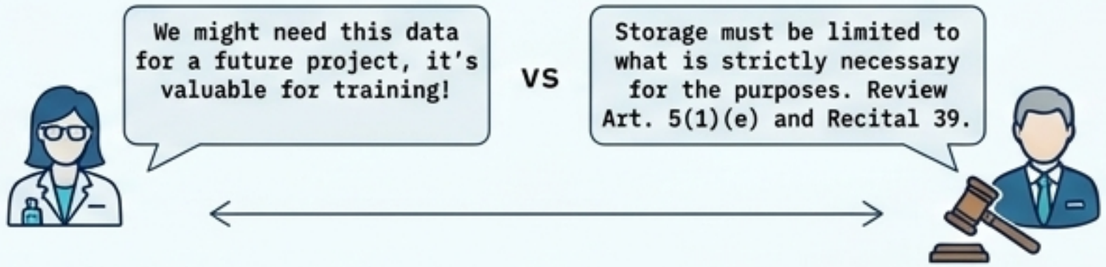
Gemini Notebook

Tech Trap 3: The Infinite S3 Hoarding Instinct



The Culture Clash

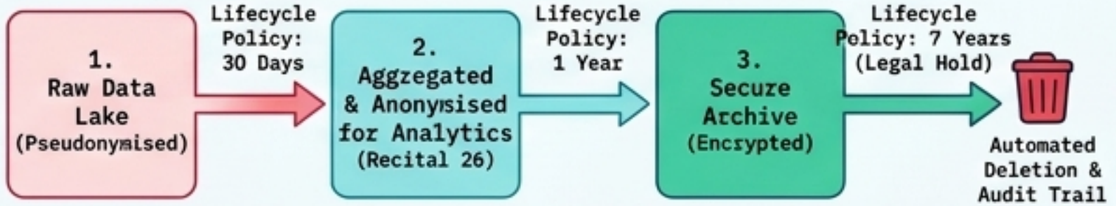
Data Science ('Store everything forever, just in case we need to train a model later') vs. European DPAs (Art. 5(1)(e) Storage Limitation).



The Enforcement Reality

DPAs aggressively police the line between strong pseudonymisation and Recital 26 anonymisation. If data in your data lake is not irreversibly aggregated, keeping it 'forever' violates GDPR.

Attach strict lifecycle policies to cloud storage buckets.



Gemini Notebook

The Pull-Request Schema Review Checklist

Reviewers | Conversation | Files changed | 100% resolved

- [x] **TTL Attached:** Do all new tables have an automated deletion or archiving policy?
- [x] **Identity Isolated:** Are direct identifiers separated from event/behavioral logs?
- [x] **Log Sanitisation:** Are API logs stripped of IPs, passwords, and session tokens?
- [x] **Art. 9 Tagging:** Are columns potentially containing health/biometric data explicitly tagged for strict RBAC?
- [x] **Cryptographic Erasure:** Is user data encrypted with a per-tenant key to allow instant deletion?

Gemini Notebook

Counsel Red Lines: When to Stop Coding and Call Legal

Trigger 1: Art. 9 Data

Designing a system that actively collects health, biometric, or genetic data. (Requires DPO / Legal to establish Art. 9(2) conditions and DPIAs).

Trigger 2: Children's Data

Building flows targeting users under 16 (or 13 in the US/certain EU states).

Trigger 3: 'Anonymised' Public Release

Planning to release a dataset publicly, claiming it passes WP216 tests for true, irreversible anonymisation.

Gemini Notebook

```
$ End of S06E02.

$ Next Episode:
$ > ./load_episode_3 --topic 'Cookies and Consent'
```

General information, not legal advice — current as of November 2026.

Gemini Notebook