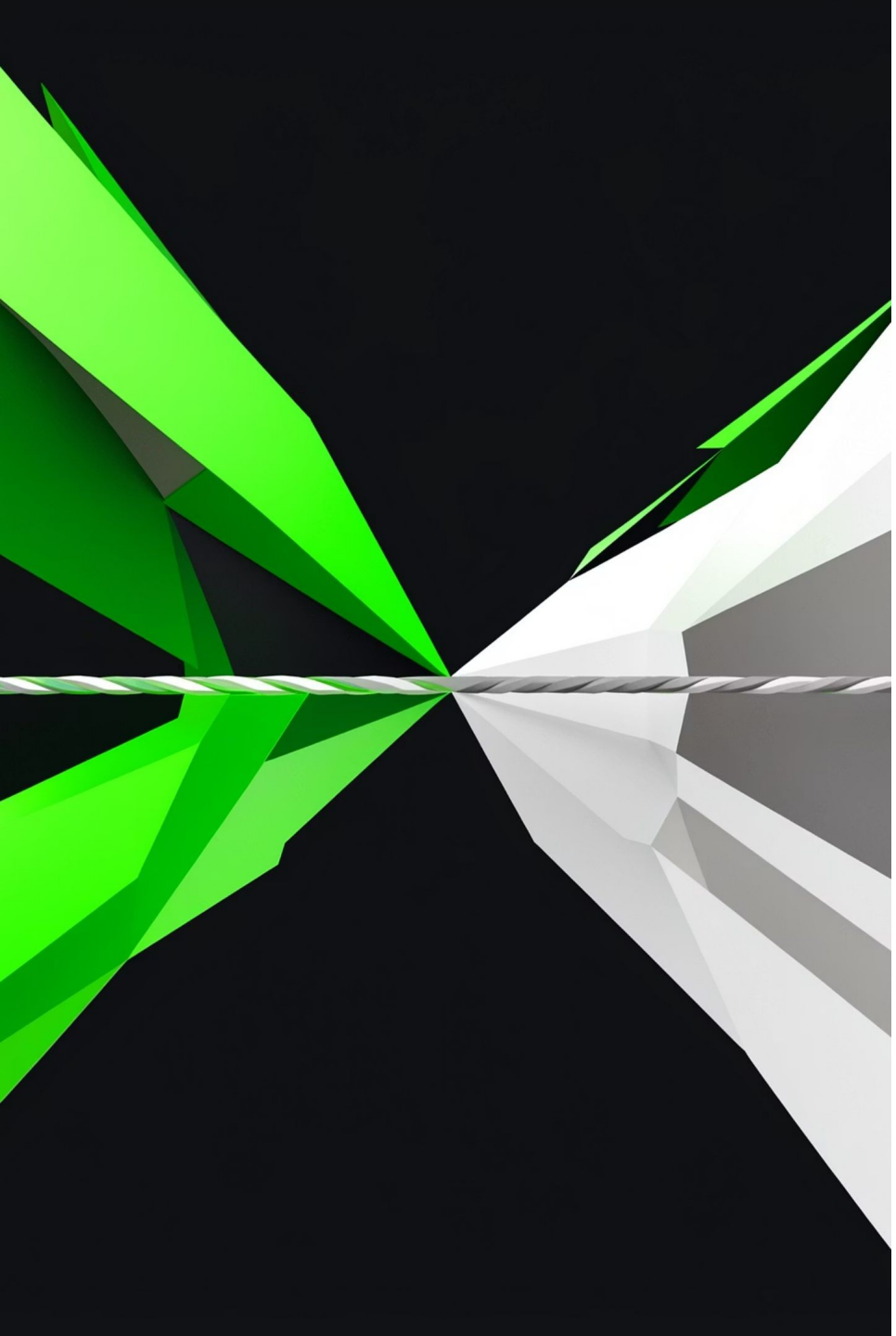


Real-Time Privacy-Preserving AI Infrastructure Using Distributed k-Anonymity Enforcement

Vamsi Viswanath Bethamsetty I University of Florida, Gainesville

RAPTORS CONFERENCE 2026



The Core Tension in Modern AI

AI platforms ingest continuous behavioral data at massive scale — the richer the data, the .better the model, but the greater the privacy risk

ML Performance

Needs high-resolution, granular event data

Privacy Compliance

Needs suppression and anonymization before processing

Regulatory Governance

Needs consistent, auditable enforcement

The Re-Identification Risk

No single field reveals identity — but a combination of common attributes can uniquely identify a person, even in “anonymized” data.

Anonymous Health Dataset

Names removed. Looks anonymous, right?

Age	ZIP Code	Disease
29	98105	Flu
34	98105	Diabetes
37	**98105**	**Rare Disease**
42	98107	Asthma
36	98109	Hypertension
…	…	…

+

Public Information

Available from news, social media, etc.

News Article

“A 37-year-old in ZIP 98105 was recently diagnosed with a rare disease.”

→

Person Identified

The combination uniquely identifies an individual.

The only 37-year-old in ZIP 98105 with this rare disease.

Sensitive health information exposed.

Key Takeaway

Even without a name, combinations of common attributes can reveal a person’s identity — leading to privacy violations, discrimination, or real-world harm.

Why Batch Anonymization Falls Short compared to streaming mode

.Batch privacy controls were built for bounded datasets, not continuous streams

Exposure Windows

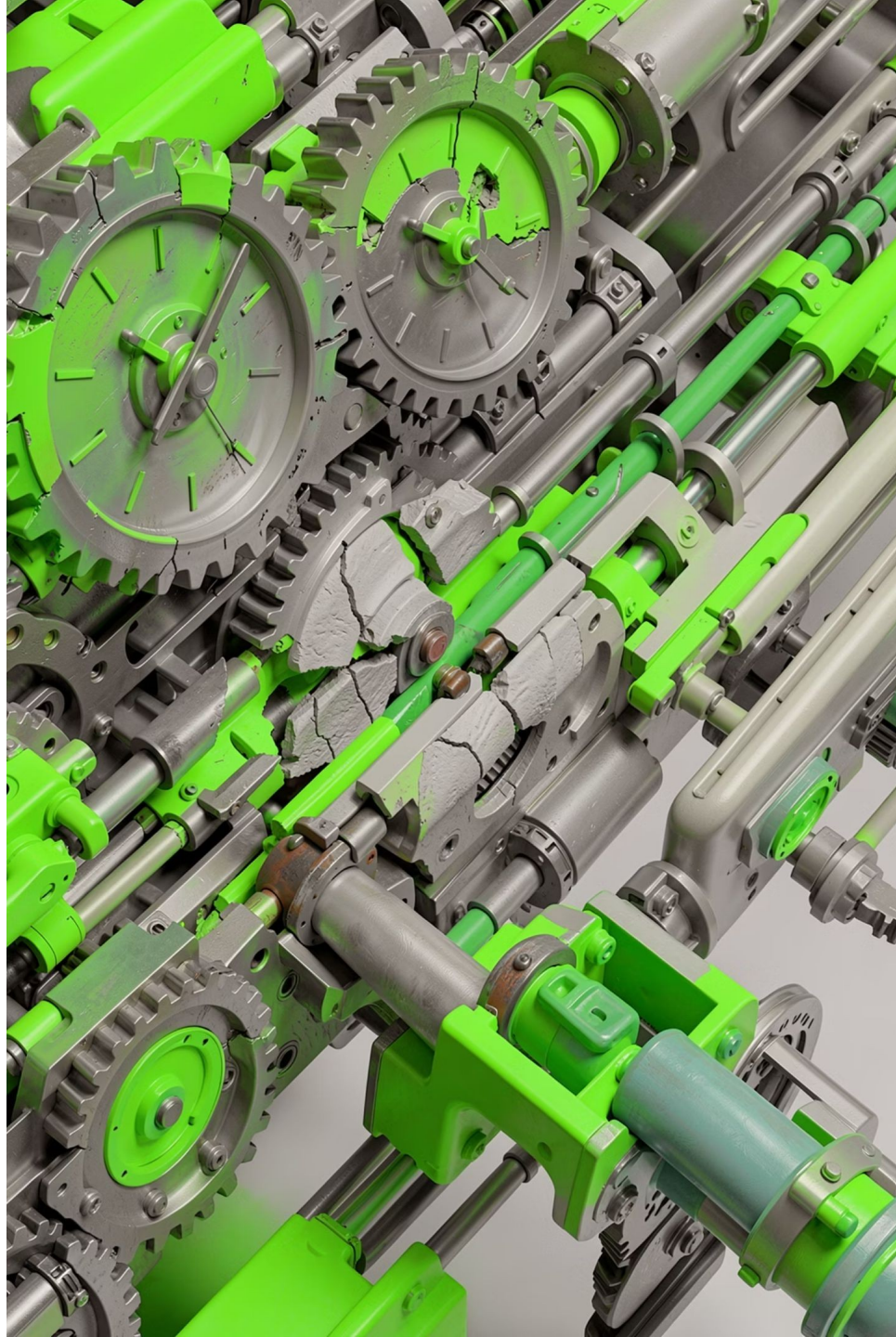
Raw data sits unprotected before batch processing runs

Backfill Gaps

Replay and regeneration often bypass anonymization

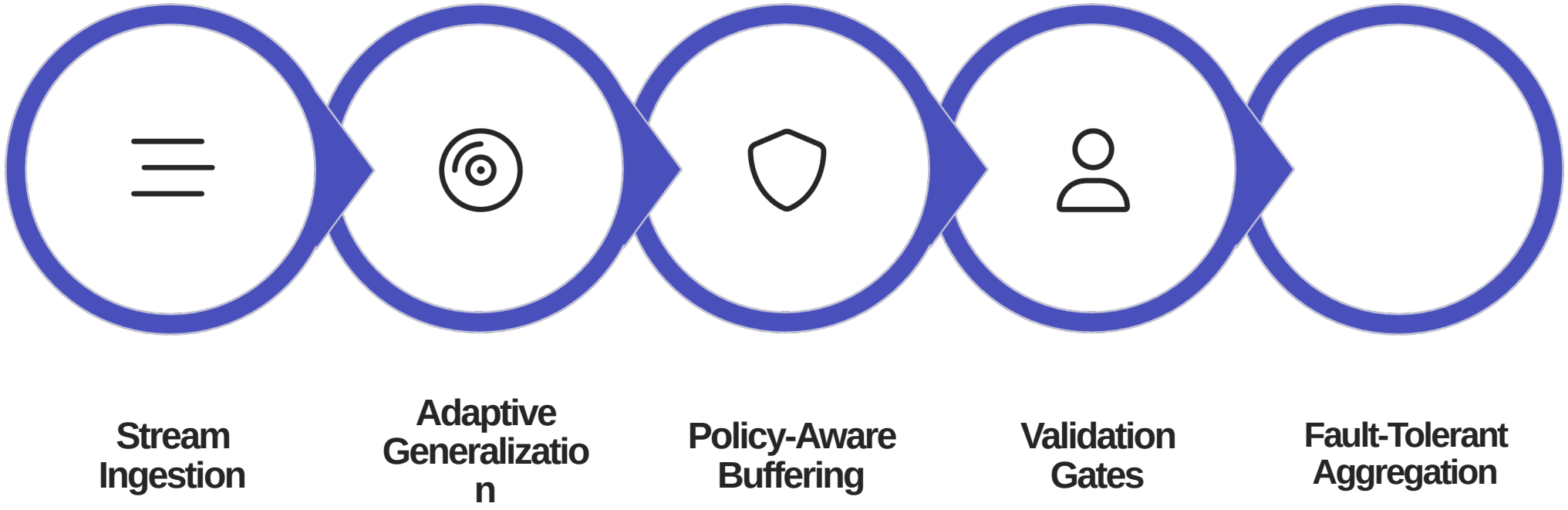
Inference Leakage

Real-time pipelines consume raw features directly

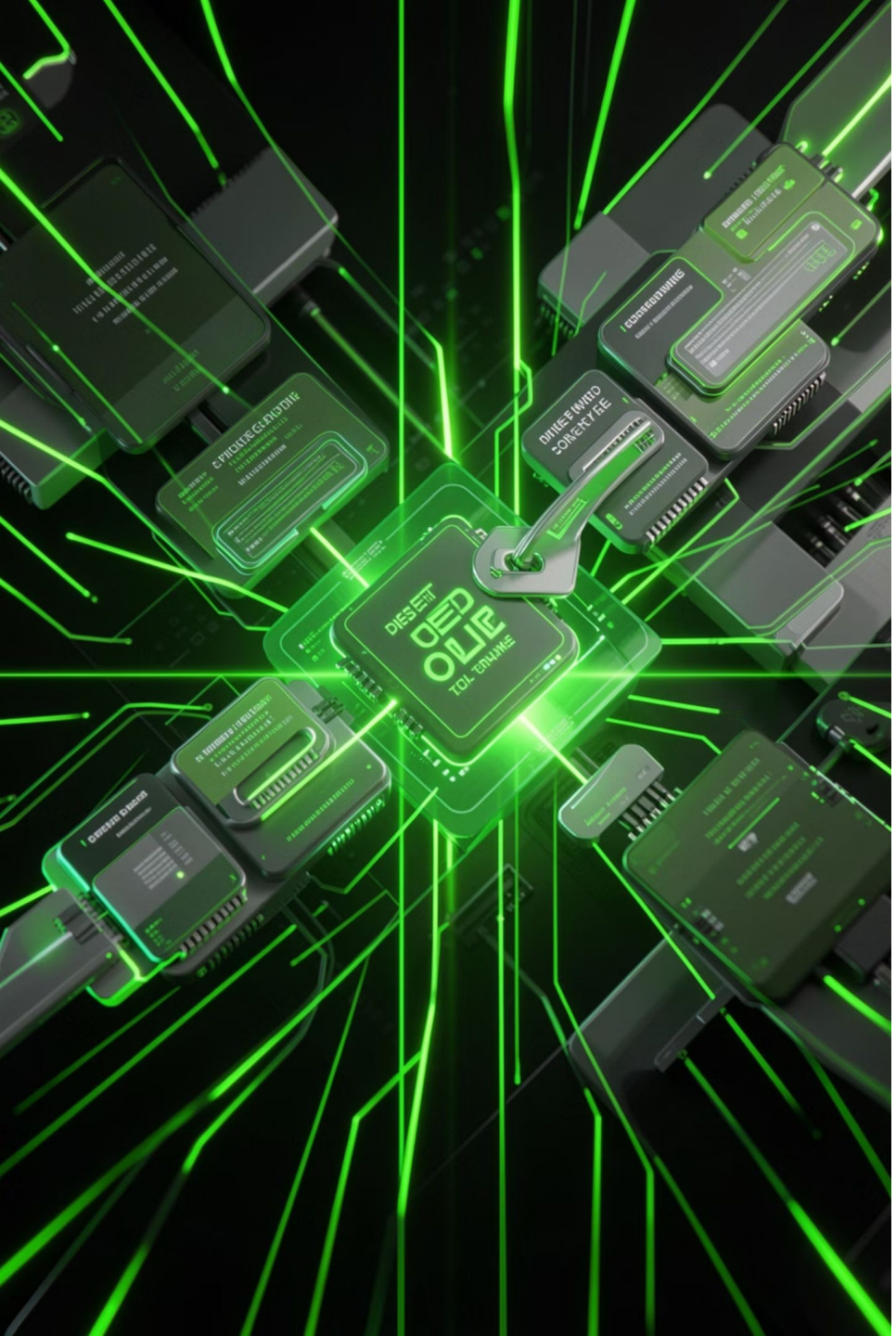


A Platform-Level Privacy Enforcement Framework

.Privacy must be a first-class capability of the infrastructure itself — not a downstream preprocessing step



.Every stage enforces privacy constraints before data reaches any downstream ML system



Ingestion & Generalization

.Events are intercepted at the point of entry, before storage or forwarding

01

Classify & Tag

Quasi-identifiers are classified and tagged on arrival

02

Generalize Values

Location, time, device, and behavior values are generalized into ranges

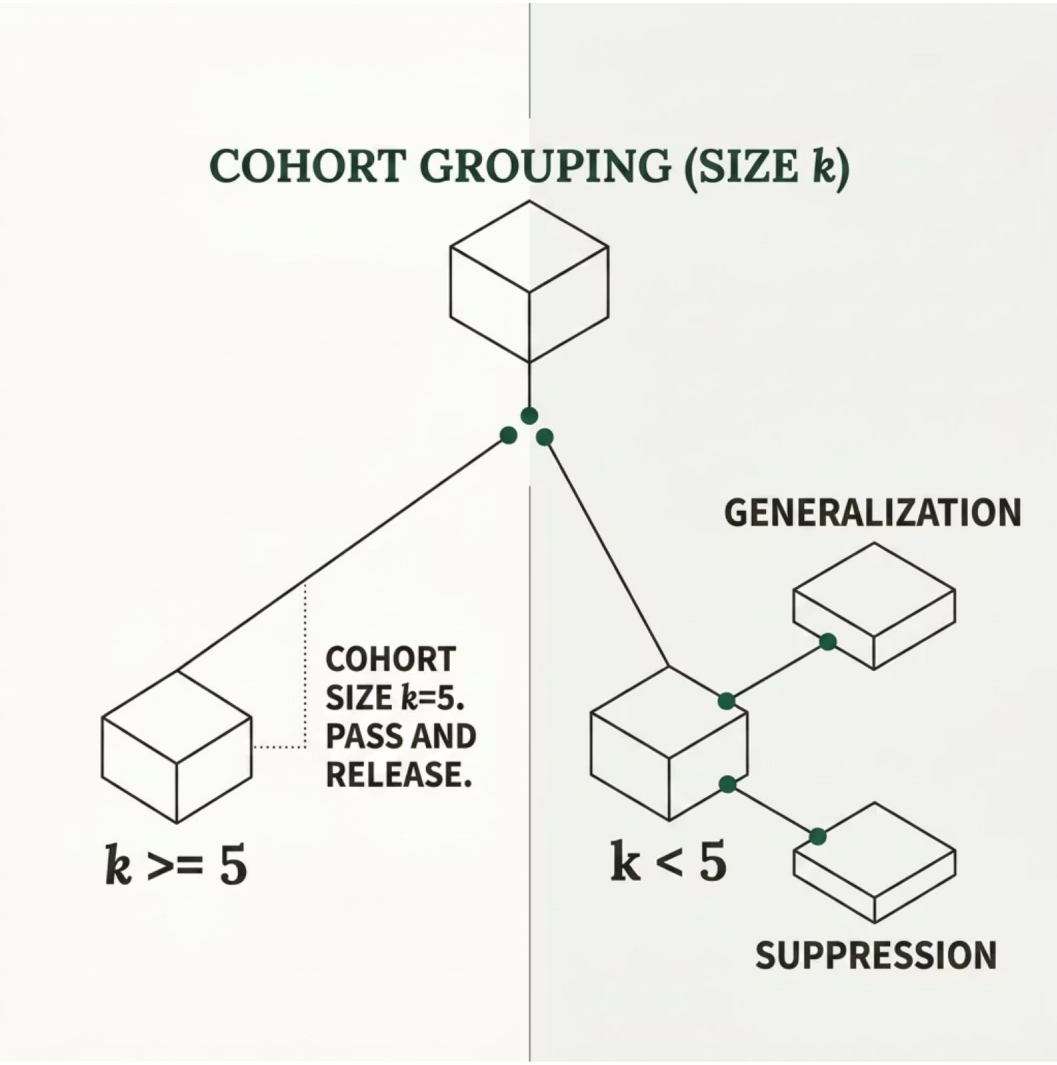
03

Apply Back-Pressure

Prevents bypass under high load

Building k-Anonymous Equivalence Classes

.Events sharing generalized values are grouped into cohorts — and only released once they're safe



Release Threshold

A cohort is released only when it reaches size **k**

Below Threshold

Generalize further, or suppress after timeout

Configurable k

Per data domain and sensitivity tier

Buffering & Edge Cases

Buffering holds events until a valid cohort forms — where latency and privacy are actively .traded off

Sparse Cohorts

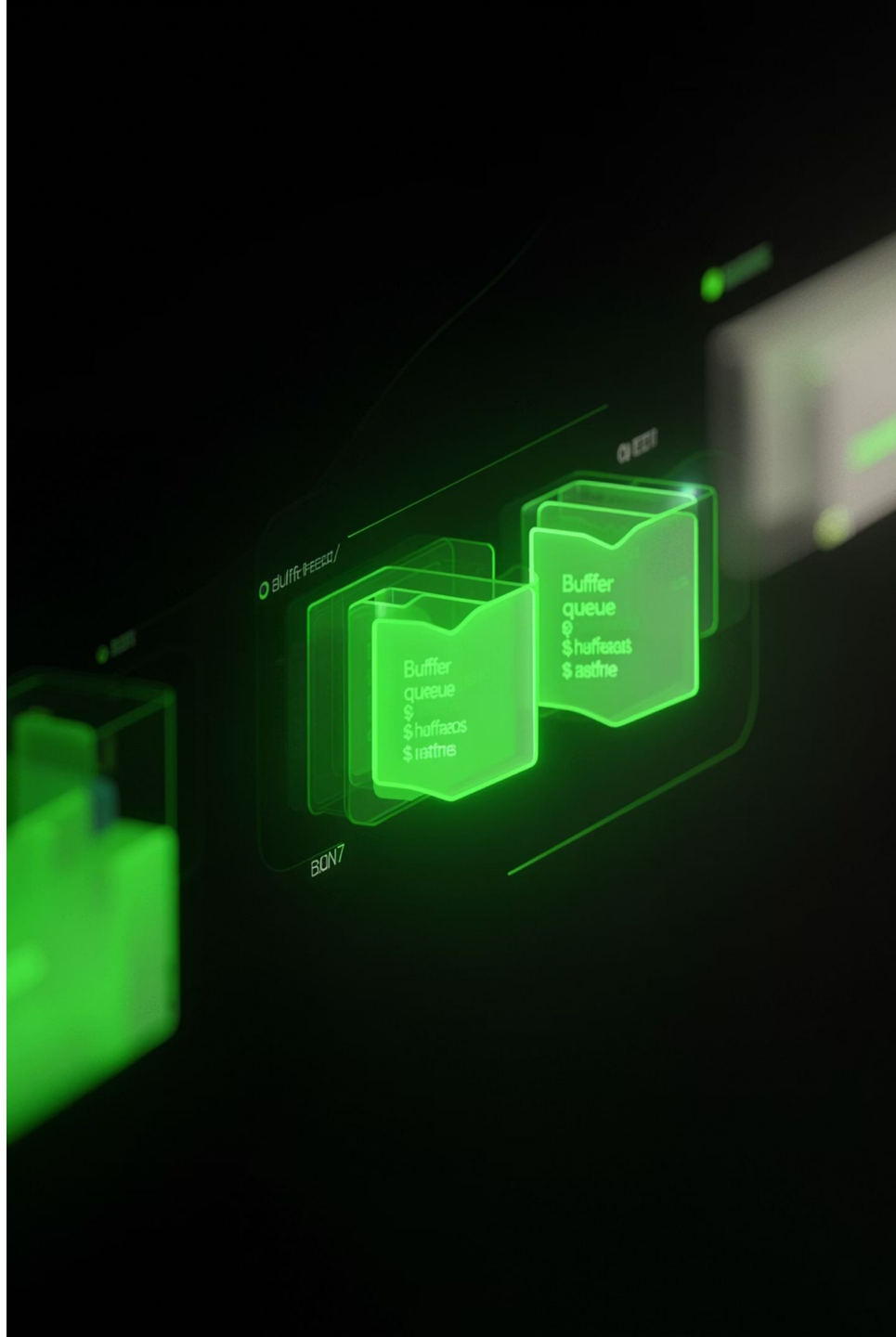
Escalate generalization or merge across domains

Late-Arriving Events

Watermark windows re-evaluate already-closed cohorts

Audit Logging

Suppression is always logged for audit





Validation Gates & Bypass Prevention

.No data reaches a downstream consumer without passing a mandatory checkpoint

Dual-Check Every Time

k-anonymity check + policy-version check on every event

Backfill as First-Class Consumer

Replay and backfill are never a shortcut — treated identically to live streams

Structurally Impossible Bypass

Bypass is architecturally prevented, not just policy-prohibited

Built to Survive Failure

.Privacy guarantees must hold through node failures and reprocessing, not just normal operation



Idempotent Processing

No duplicate cohort membership on retry



Distributed State

Survives worker failures without data loss



Exactly-Once Commits

Prevents partially applied policy



Versioned Regeneration

Historical data pinned to policy version at ingestion, with full audit trail

Integration with ML Platforms

.The anonymization layer sits transparently between raw streams and ML consumers — no changes to model training code required





Where This Applies

Healthcare

HIPAA-aligned de-identification for clinical AI

Financial Services

GLBA-compliant behavioral analytics

Public Sector

Accountable citizen-facing AI

Enterprise

GDPR/CCPA-compliant behavioral ML at scale

Key Takeaways for Practitioners

01

Platform-Layer Enforcement

Enforce privacy at the platform layer — not per-team

02

Design for Backfill

Design for backfill from day one

03

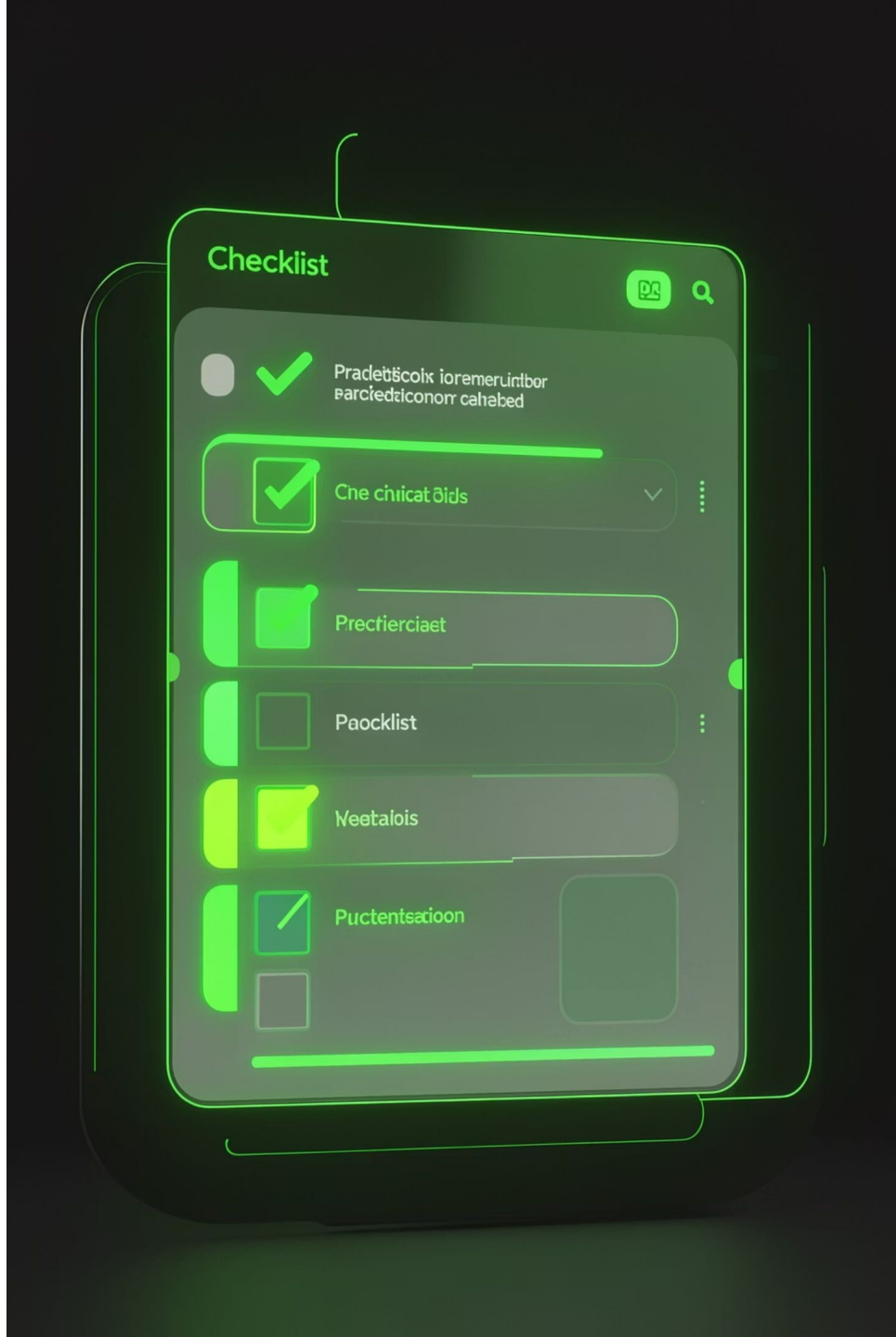
Version Everything

Policies, hierarchies, and k-thresholds must all be versioned

04

Monitor Continuously

Monitor data flow continuously — static gates aren't enough



Summary & Open Questions

This Work Delivers

A real-time architecture enforcing k-anonymity before data reaches ML systems, with governance for sparse cohorts, late arrivals, and bypass .risk

Open Directions

Streaming l-diversity and t-closeness
Combining with differential privacy
Automated generalization learning

.Thank you — questions welcome on architecture, latency trade-offs, and ML platform integration