

Fighting Harmful Content with AI: Scalable Moderation for Real- World Safety

Georgina Tryfou

Machine Learning Engineer, Gcore

Conf42 AI & ML 2025



Why this matters?



Explosive video growth

User generated video
rising exponentially

Harmful content risks

CSAM and other
dangerous content
increasing rapidly

Manual limits

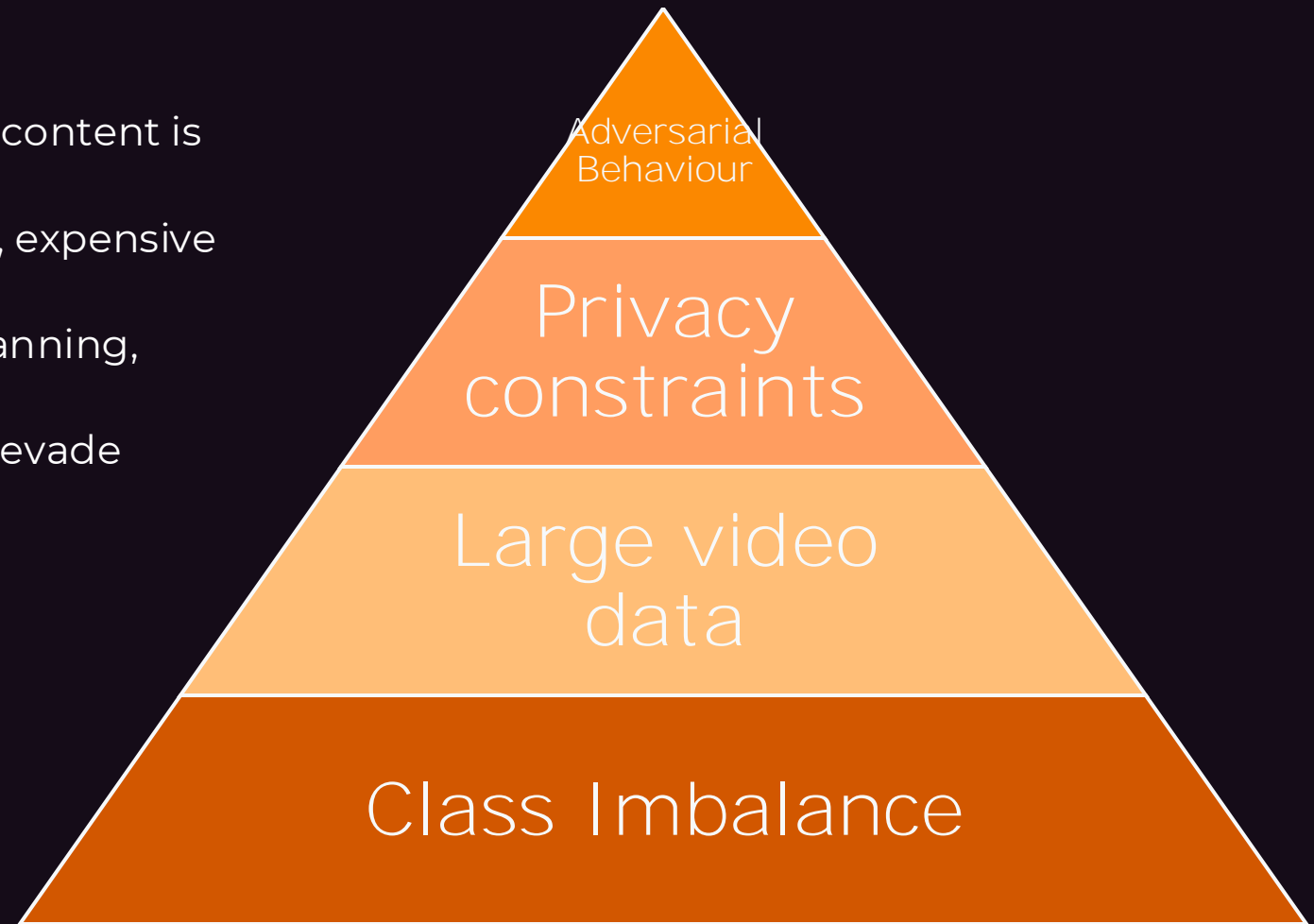
Human moderation
cannot keep pace

AI requirements

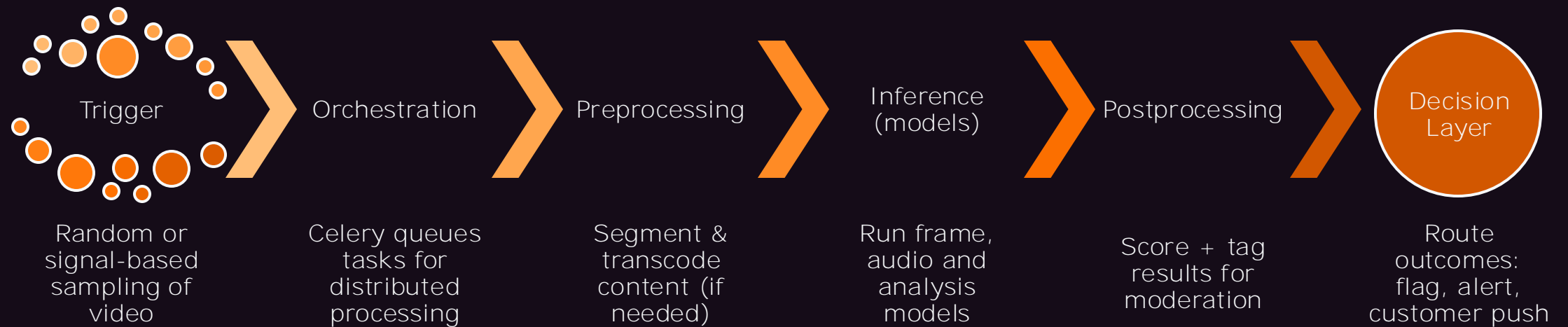
Solutions must be
scalable, private, real-time

The real-world complexity of harmful content detection

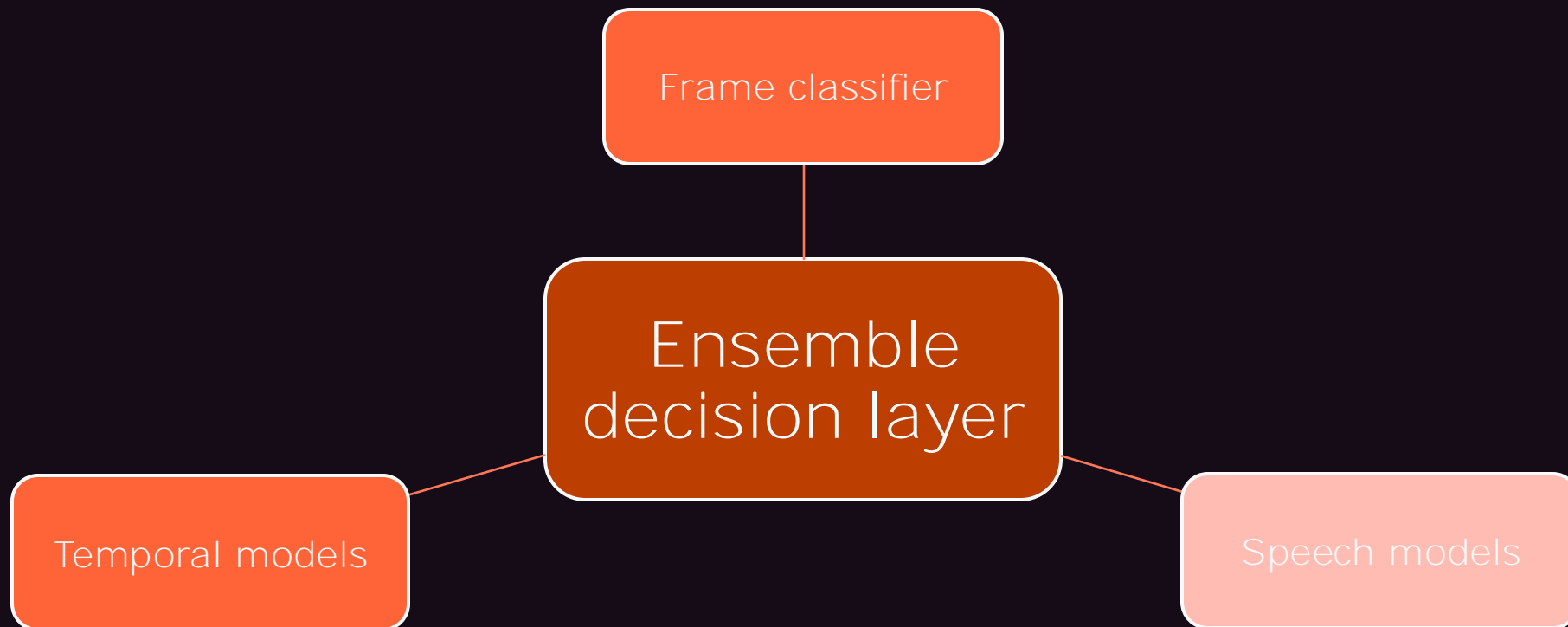
- **Extreme class imbalance** — harmful content is $<0.001\%$ of the data
- **High data volume** — videos are large, expensive to scan
- **Privacy constraints** — no invasive scanning, GDPR-compliant
- **Adversarial content** — attackers actively evade detection







Our scalable moderation architecture



Multi-modal AI for content detection



Customer use map

	Use Case	Description
	Video Uploads	Scan at ingestion to prevent publishing harmful content
	Live Streaming	Segment analysis in near-real-time
	File Hosting	Background moderation of stored files
	API Integration	Simple REST APIs & webhooks for decision routing

Privacy first, ethically designed moderation



Privacy by design

No media stored or reused



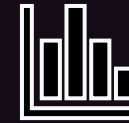
Random sampling

Avoid surveillance, enable fairness



Legal compliance

Follows GDPR, CCPA, local regulations



Bias & audit checks

Regular model validation & trace logs

Scaling and vision



Lessons Learned



Orchestration is everything

Pipelines make the model useful



Class imbalance

Impacts product behavior



Infra is the core of reliability

System health > model speed



Explainability builds trust

Metadata and logs matter



Thank you!

gcore.com

© 2025 Gcore