DATA+AI SUMMIT 2022

Human-in-the-loop ML systems for platform integrity

Nihit Desai
Co-Founder & CTO, Refuel.AI

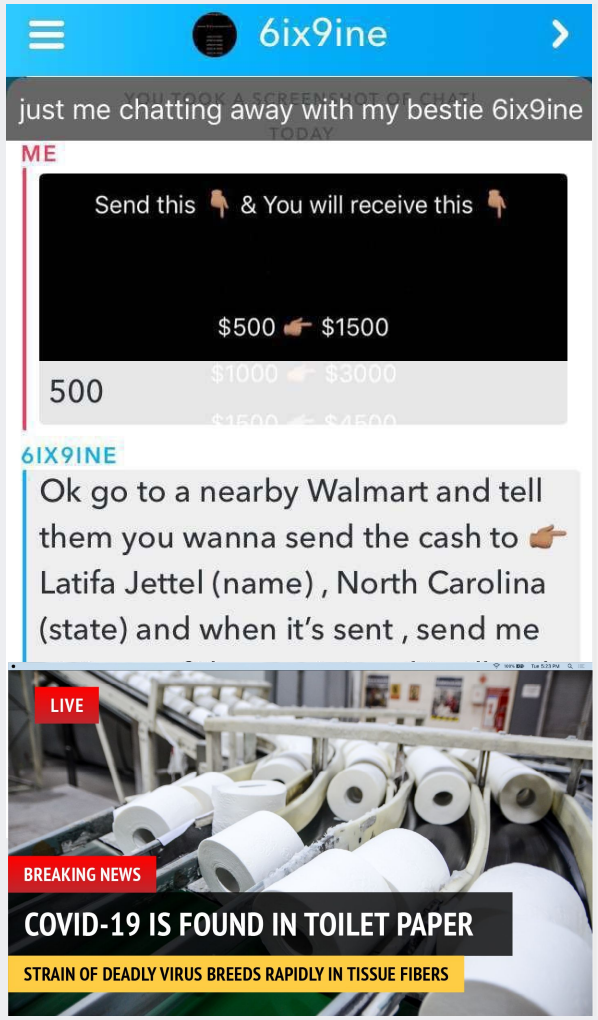
ORGANIZED BY databricks

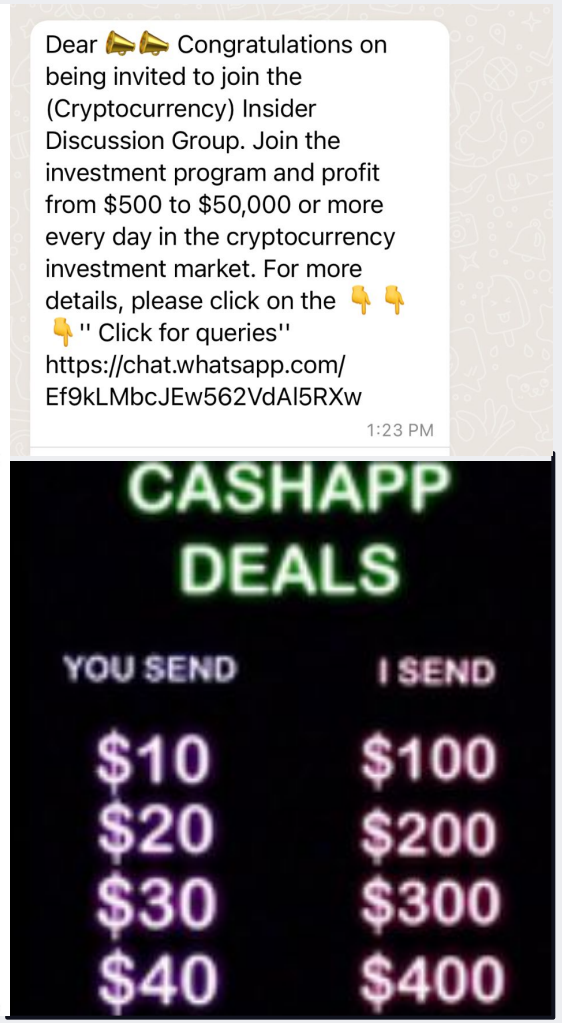
Agenda

- Problem overview
- Architecture overview
- Learnings in building integrity systems
  - Data Collection
  - Model Evaluation
  - Adaptiveness
- Questions

Problem Overview

With Scale comes responsibility



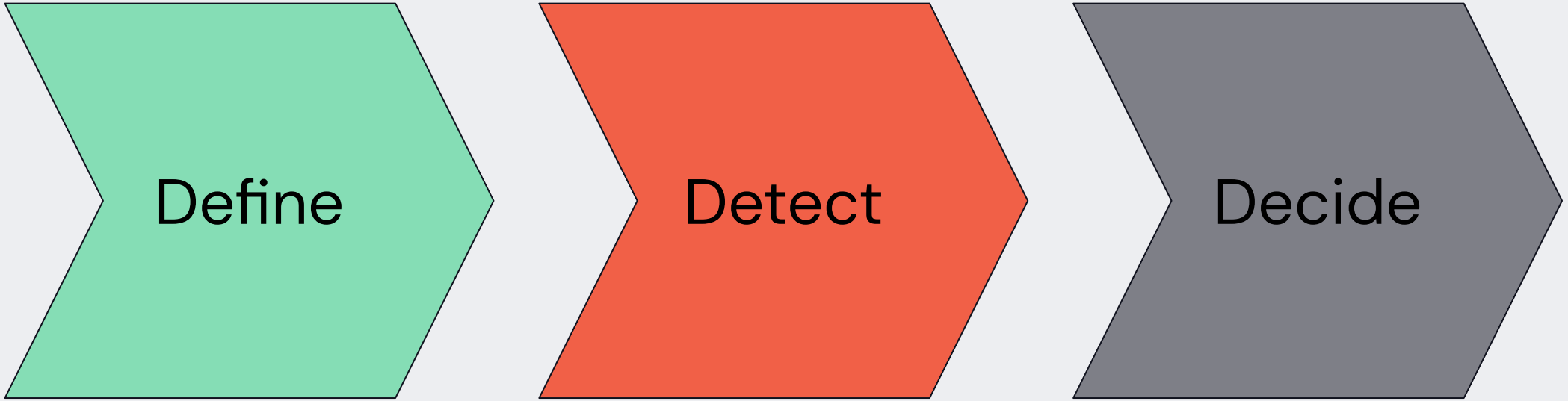


Mission statement

Reduce harm *to the user community,*

and ***increase trust*** *of interactions on the platform*

Three pillars of integrity systems



Define what is acceptable on the platform

⚖️

Detect content that violates defined policies

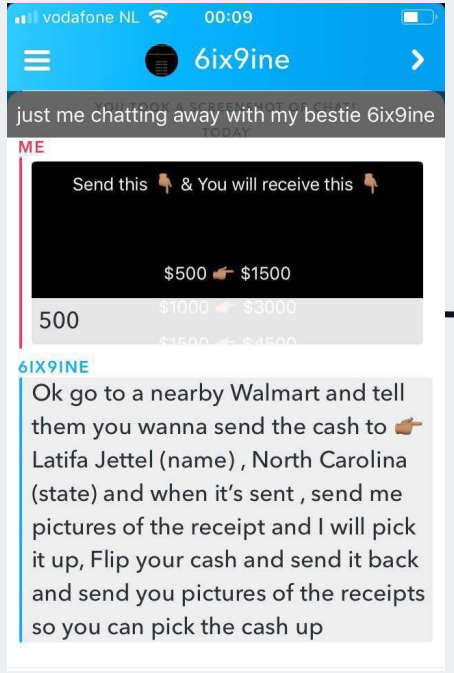
👮

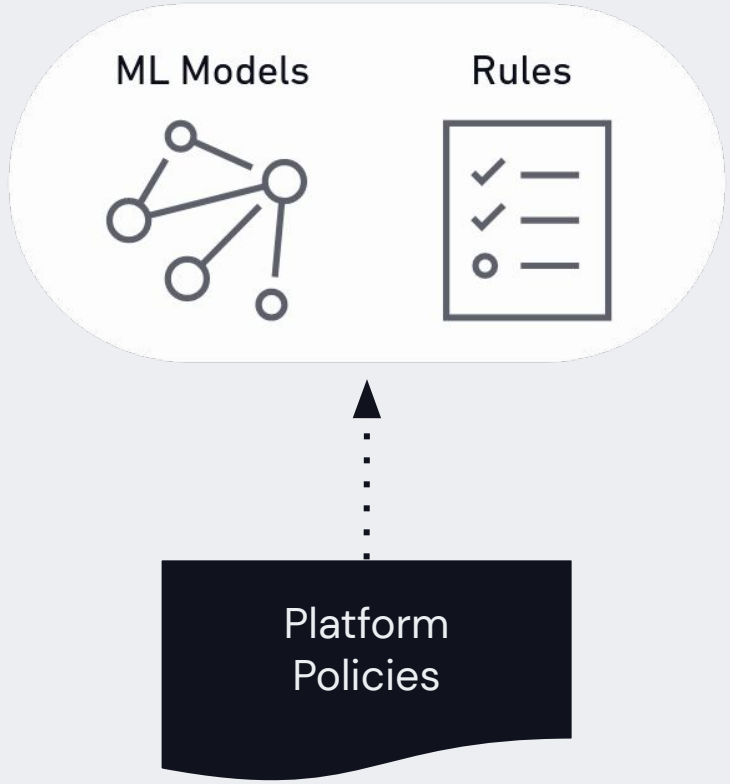
Take enforcement action on violating content/actors

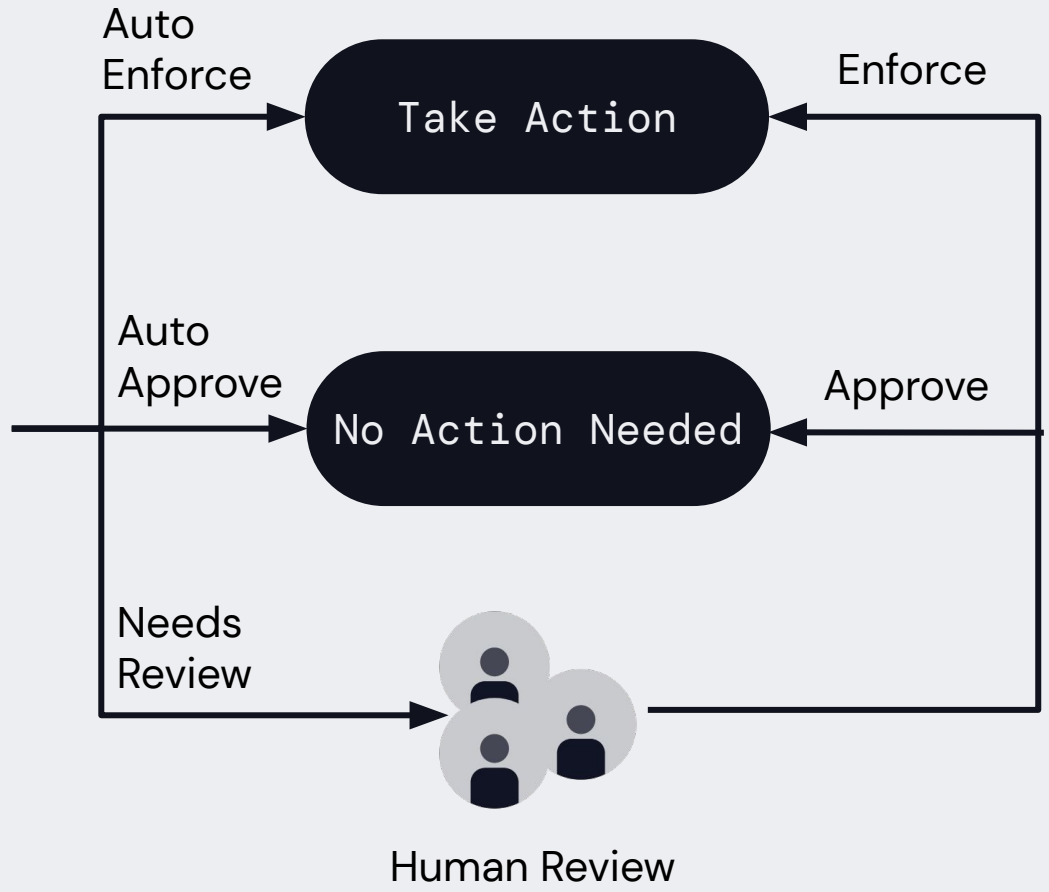
🏛️

Architecture Overview

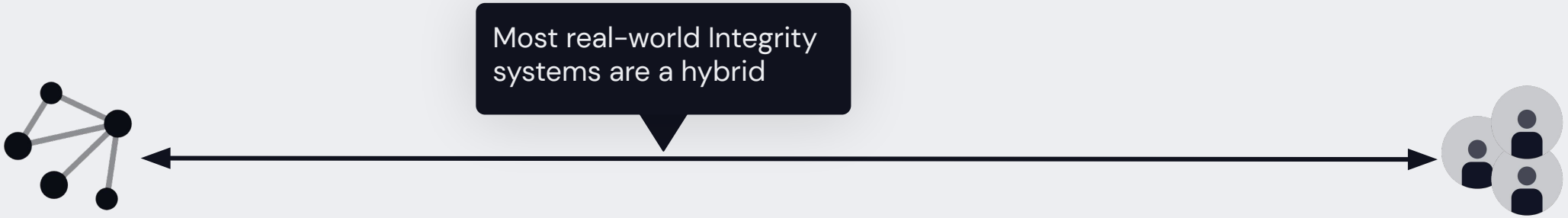
Architecture overview







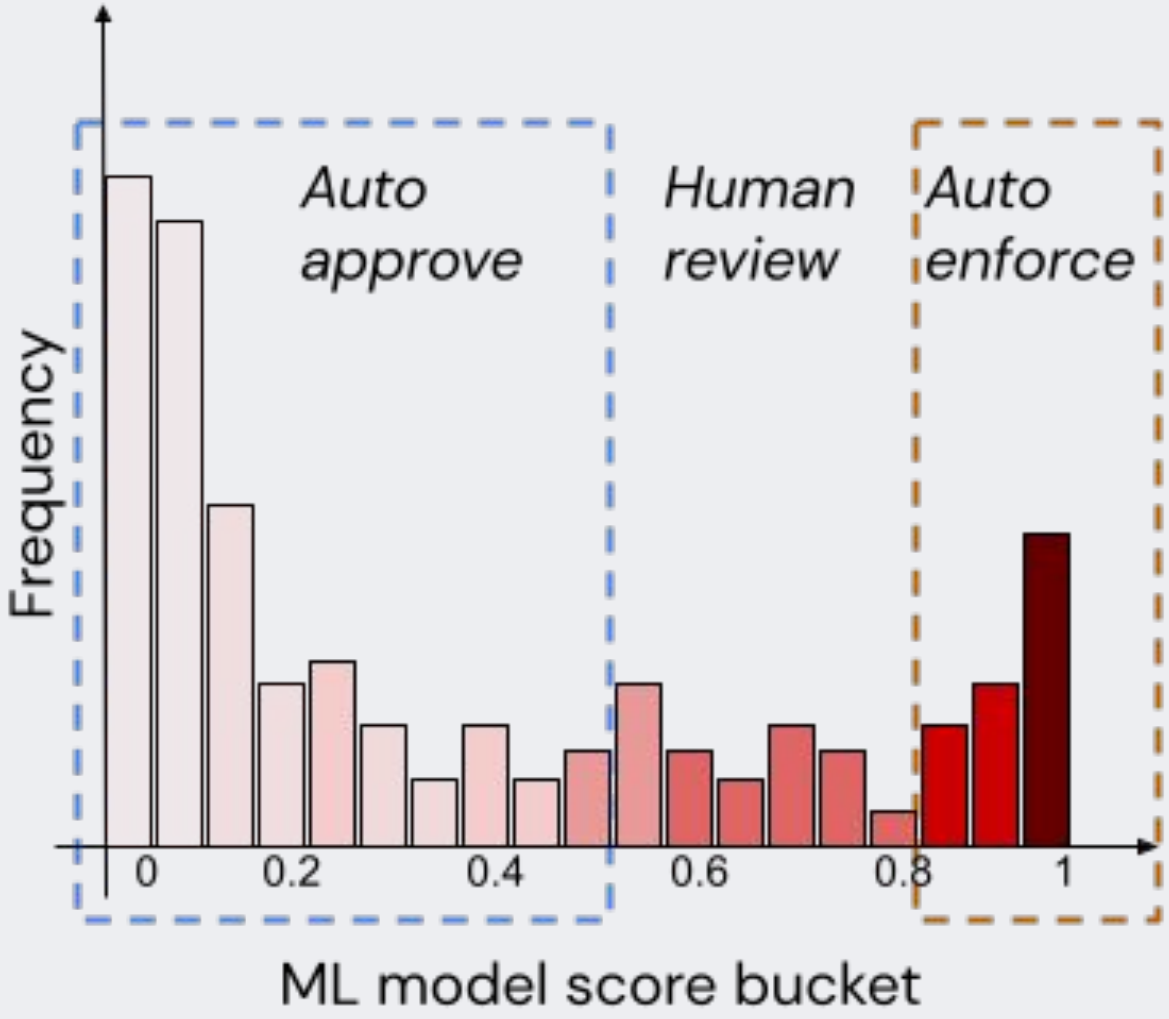
Human-in-the-loop ML: Best of both worlds

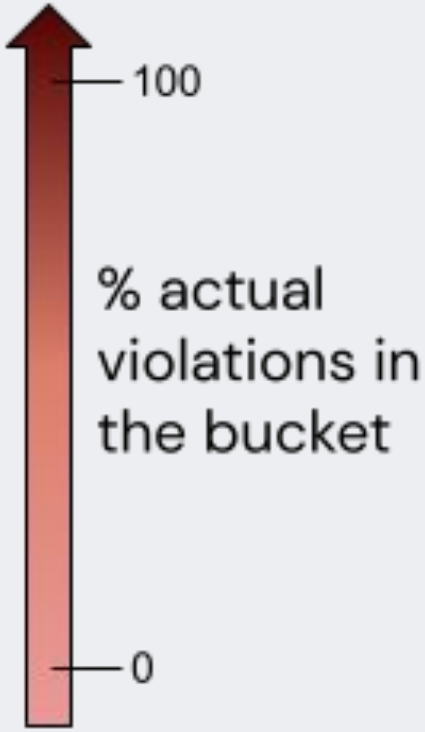


- ✓ O(seconds) latency
- ✓ Scale to O(billions)
- ✓ Compute is cheap(er)
- ✗ Performance degradation, lack of adaptiveness

- ✗ O(hours) latency
- ✗ Harder to scale
- ✗ Cost prohibitive at scale
- ✓ Adaptive, easy to incorporate external context

Human-in-the-loop ML: Best of both worlds





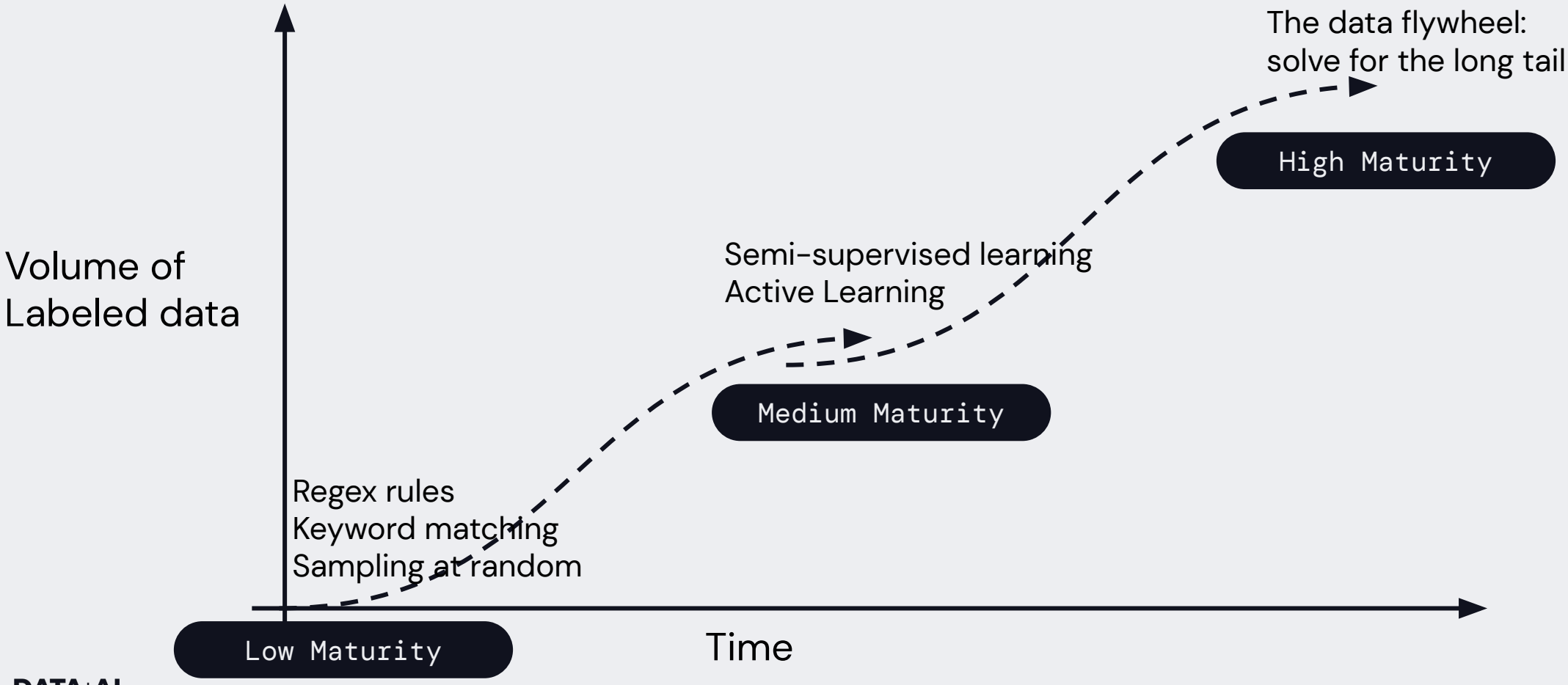
Measuring success

1. *Maximize **recall***
2. *Minimize **false positive** errors*
3. *Keep **operational expenses** within budget*

Data Collection

Data Collection

Strategies evolve as the problem domain matures over time



Active Learning

Intelligently prioritize what to label next

Active learning samples what to label, from the pool of unlabeled data. Strategies:

- **Uncertainty-sampling:** based on what the model is most uncertain about.
- **Diversity-sampling:** based on what is most "novel" compared to training data

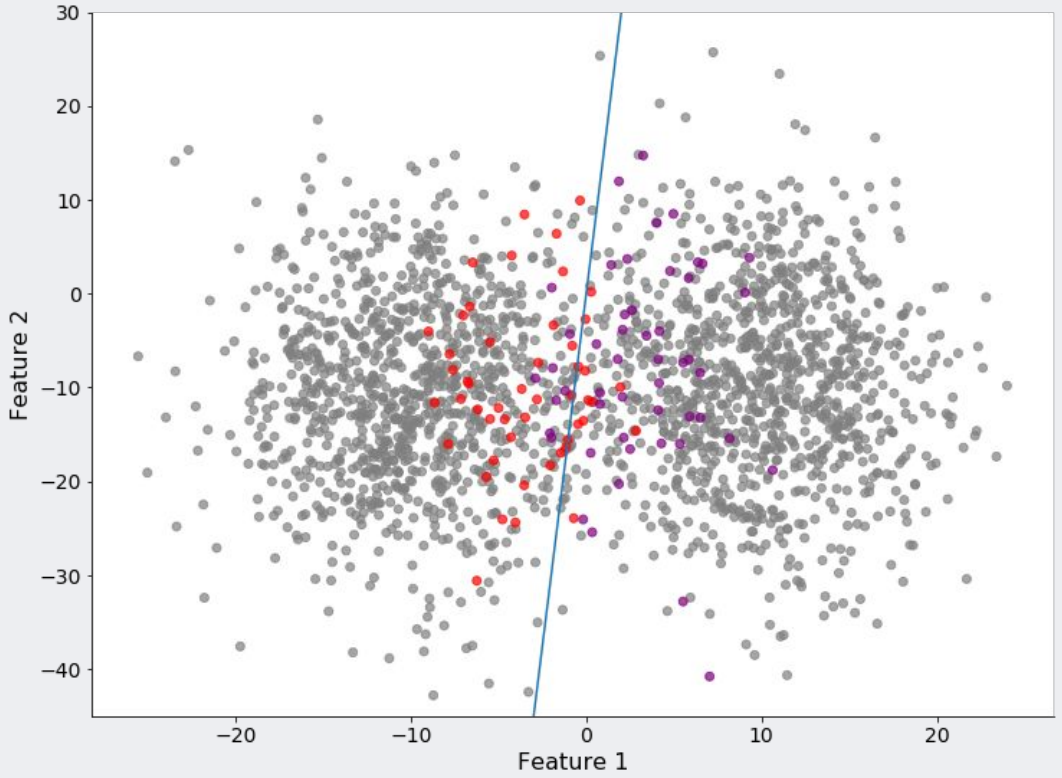


Image Credit: Towards Data Science

The Data Flywheel

Real-world is complex, evolving and long-tailed

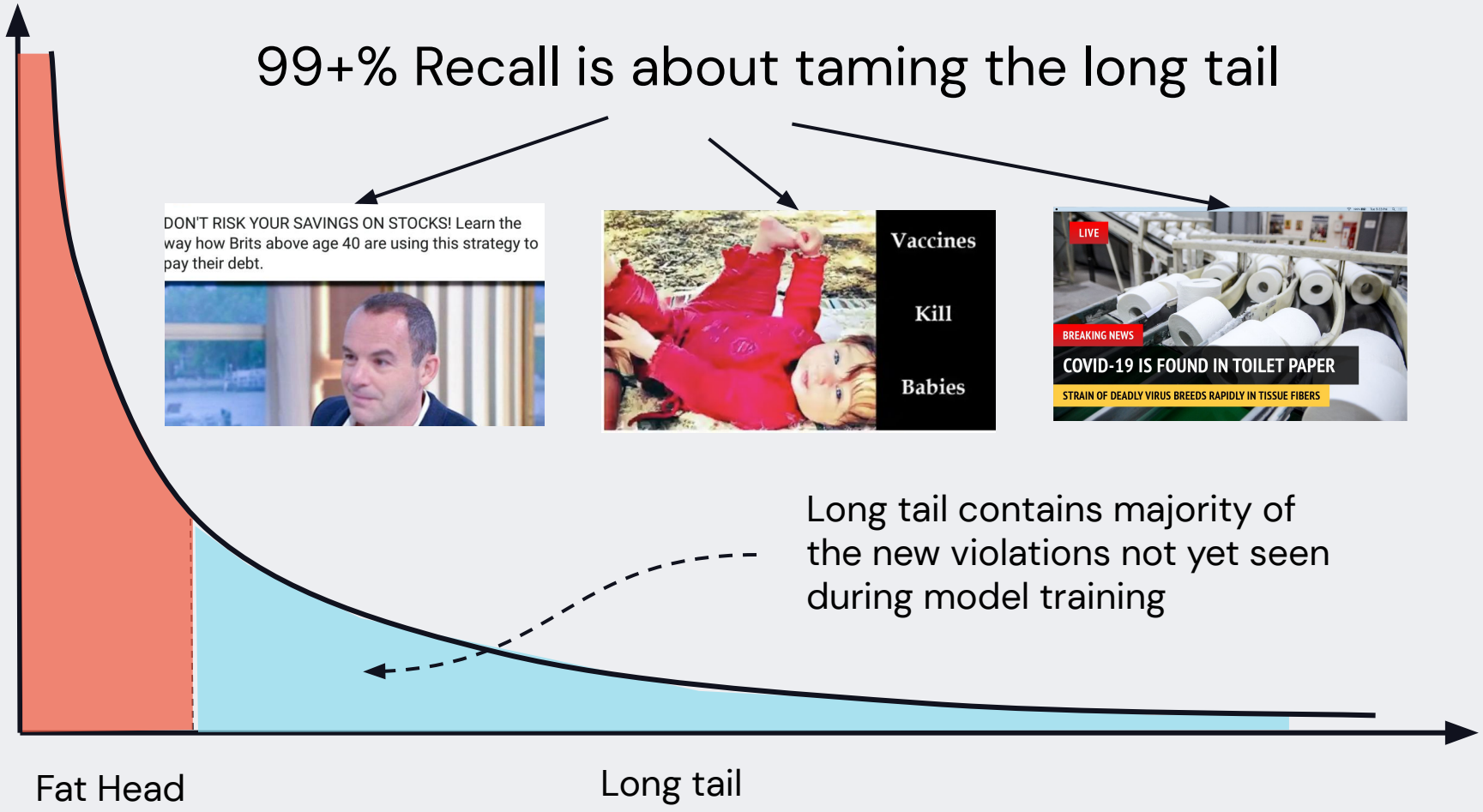


Image inspired by similar graphic from Tesla Autonomy Day 2019

Model Evaluation

Components of Model Evaluation

Evaluation setup should proxy real-world performance

Datasets

Should represent what the model will see after launch

- ✓ Dynamic benchmark sampled from real-world traffic
- ✗ Static datasets

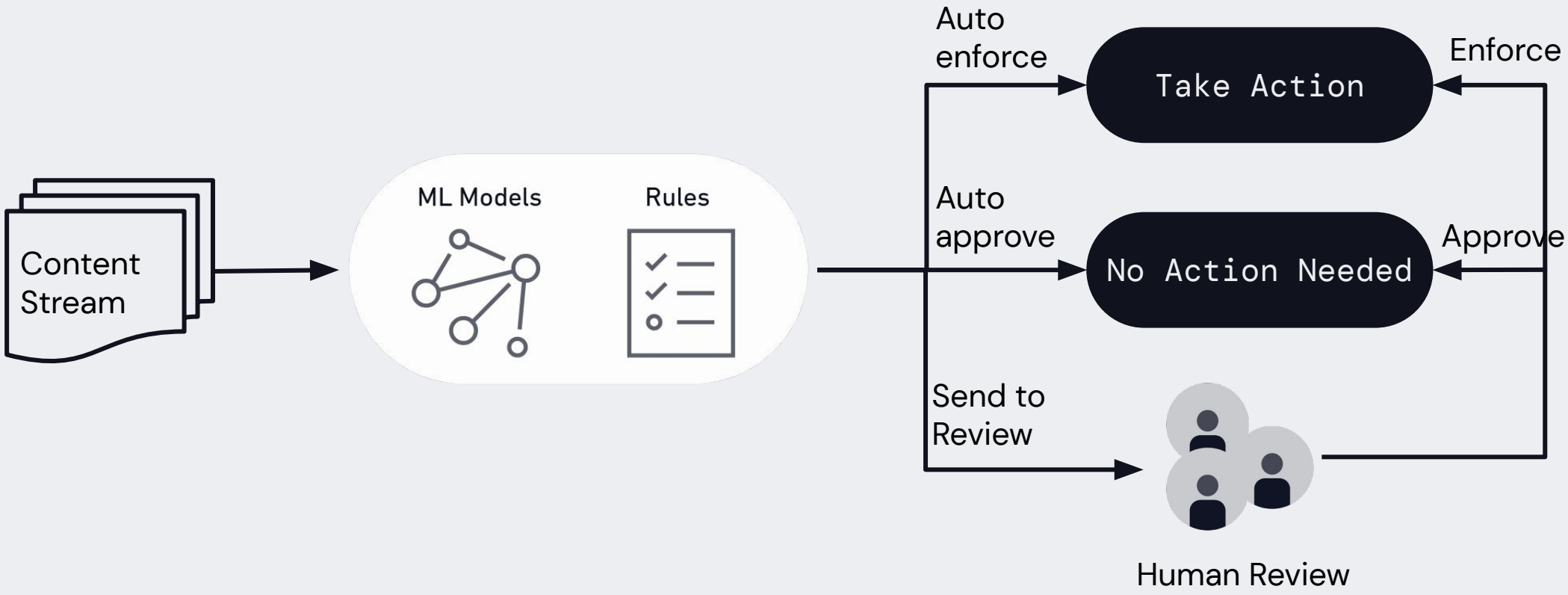
Metrics

Should align with the product and business objectives

- ✓ Recall @ target precision
- ✗ Accuracy; no measurement for dataset slices

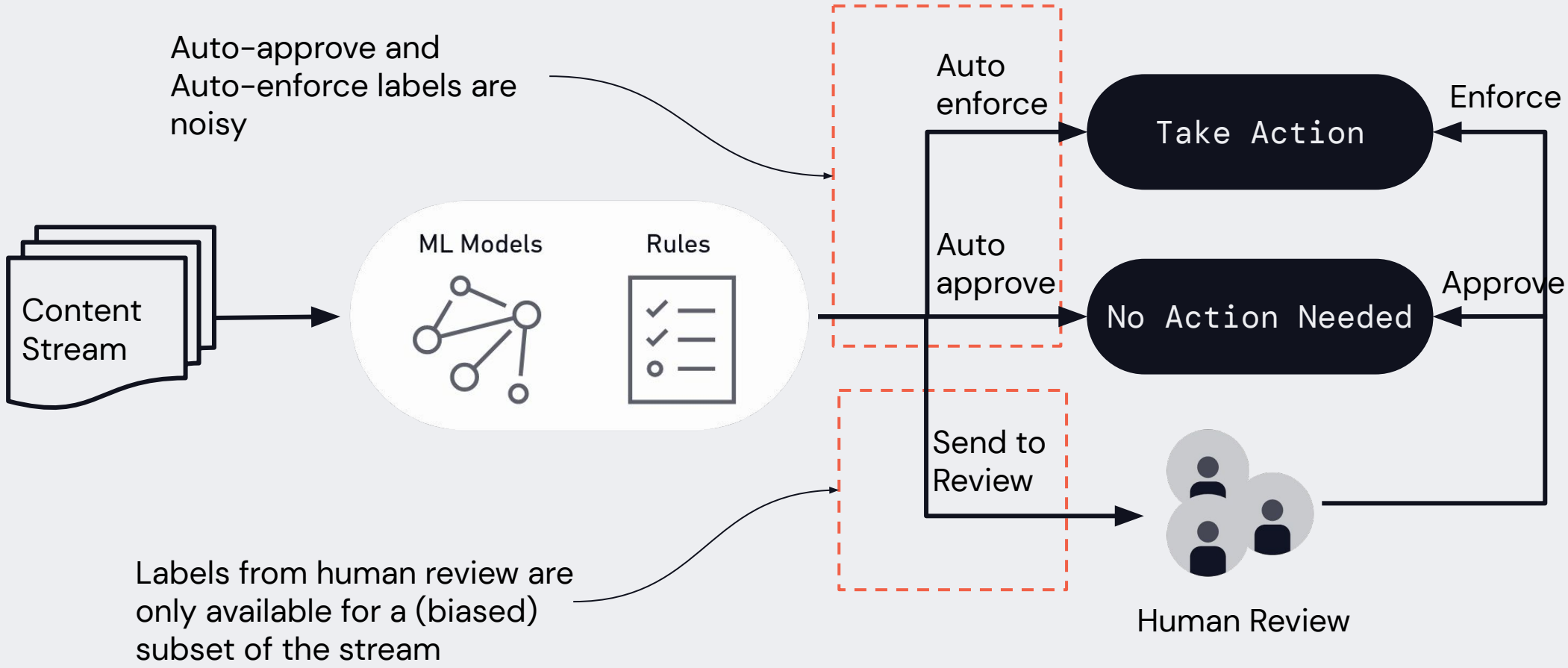
Model Evaluation: Datasets

Constructing unbiased evaluation datasets is hard



Model Evaluation: Datasets

Constructing unbiased evaluation datasets is hard



Model Evaluation: Metrics

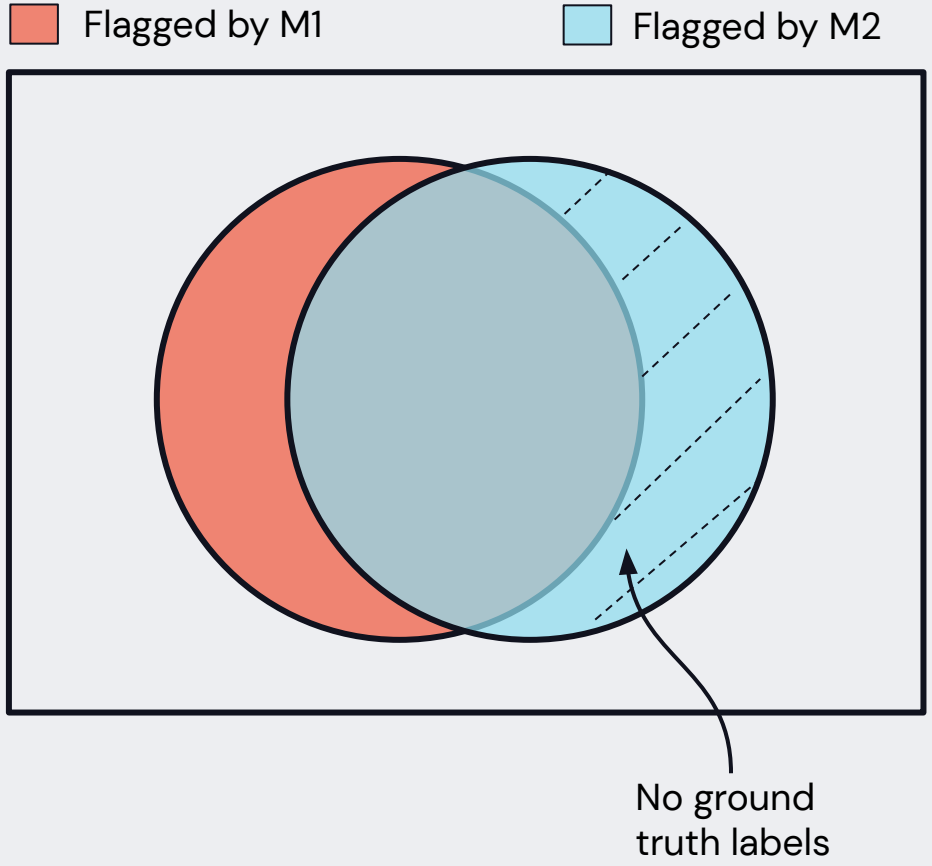
Unbiased evaluation needs continuous annotations

Common scenario:

- We have production model (M1).
- Want to evaluate new candidate (M2) to decide whether to ship

Challenge:

Subset of traffic with ground truth labels [■] is biased. Evaluating just on this dataset gives us no idea of recall improvement



Model Evaluation: Metrics

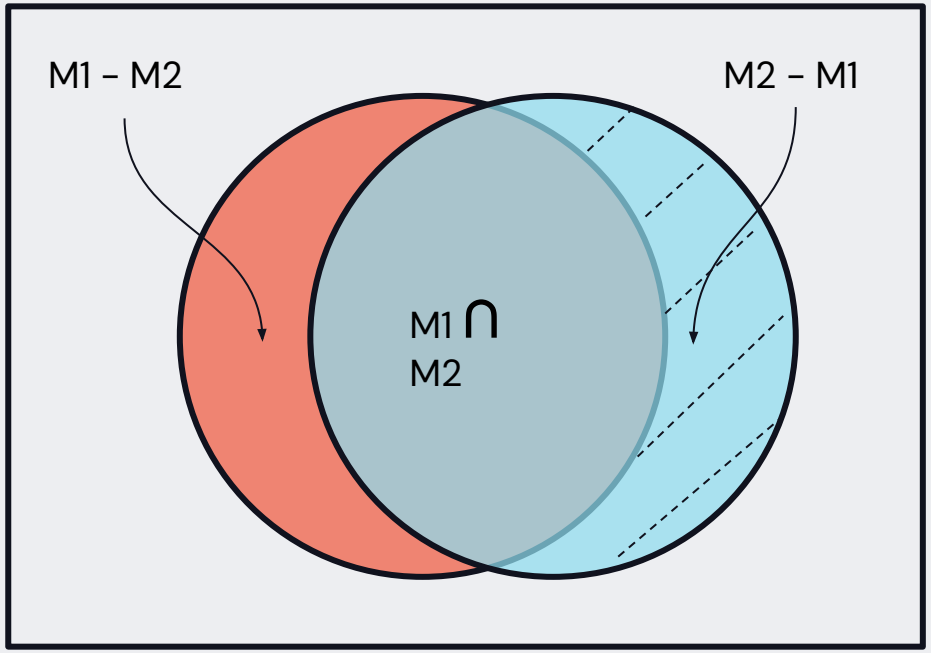
Unbiased evaluation needs continuous annotations

Relative recall: How many violations caught by baseline (M1) are also caught by M2 ?

= `Precision(M1 ∩ M2) * |M1 ∩ M2|`

Additive recall: How many new violations are caught uniquely by M2 (for the same budget)?

= `Precision(M2 - M1) * |M2 - M1|`



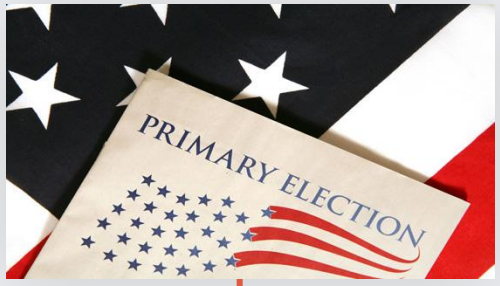
Estimate precision (density of TP) in each bucket through annotations

Adaptiveness

Real world evolves. Adaptiveness is key

Online discourse changes in response to real-world events

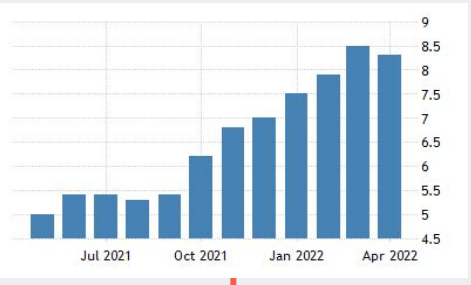
Primary Elections



Jan 6

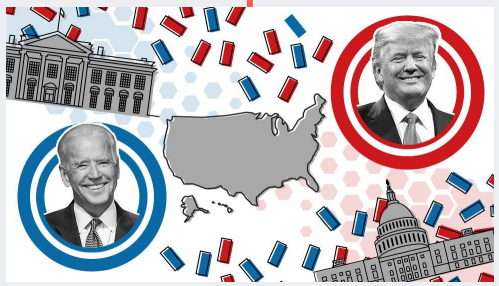
Infrastructure bill

Inflation & Economy

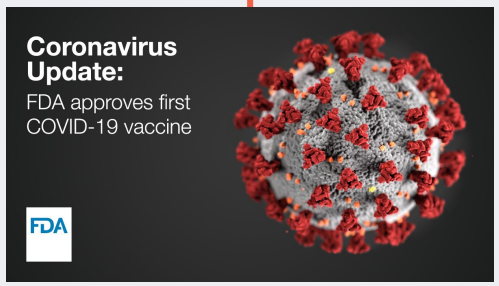


2020

2022



General Elections

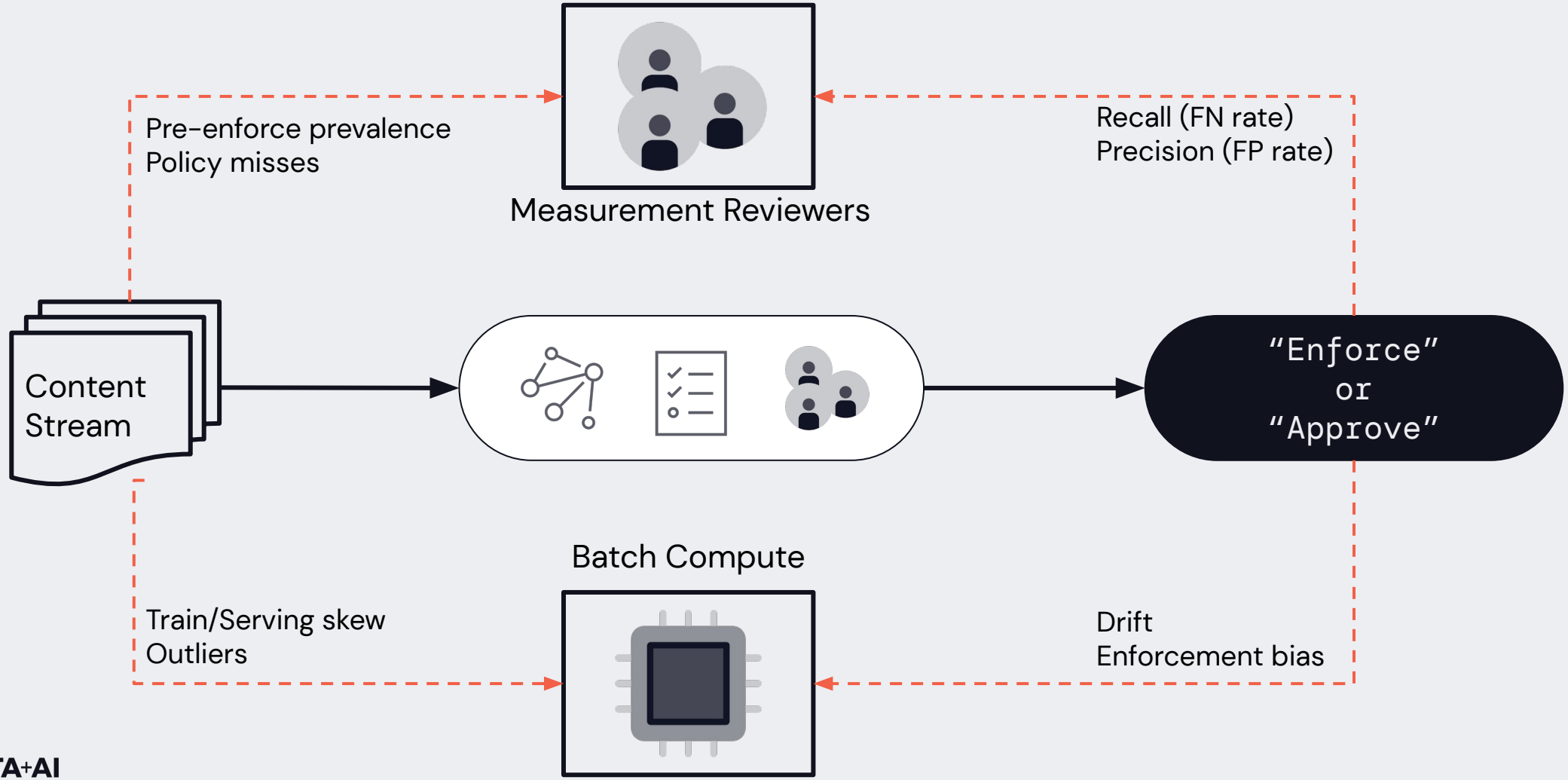


COVID vaccines

War in Ukraine

Measure Everything

Fine-grained, realtime and attributable measurement is key

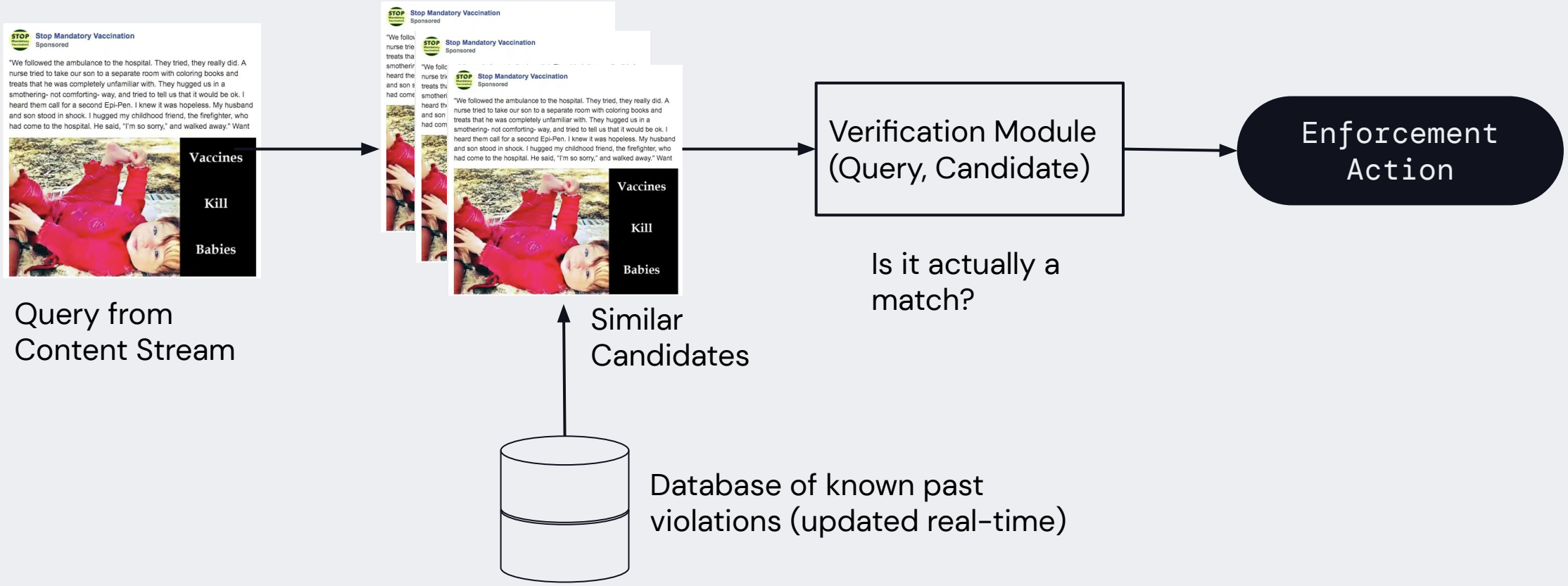


Leverage content similarity

Fanout decisions to near-duplicate content

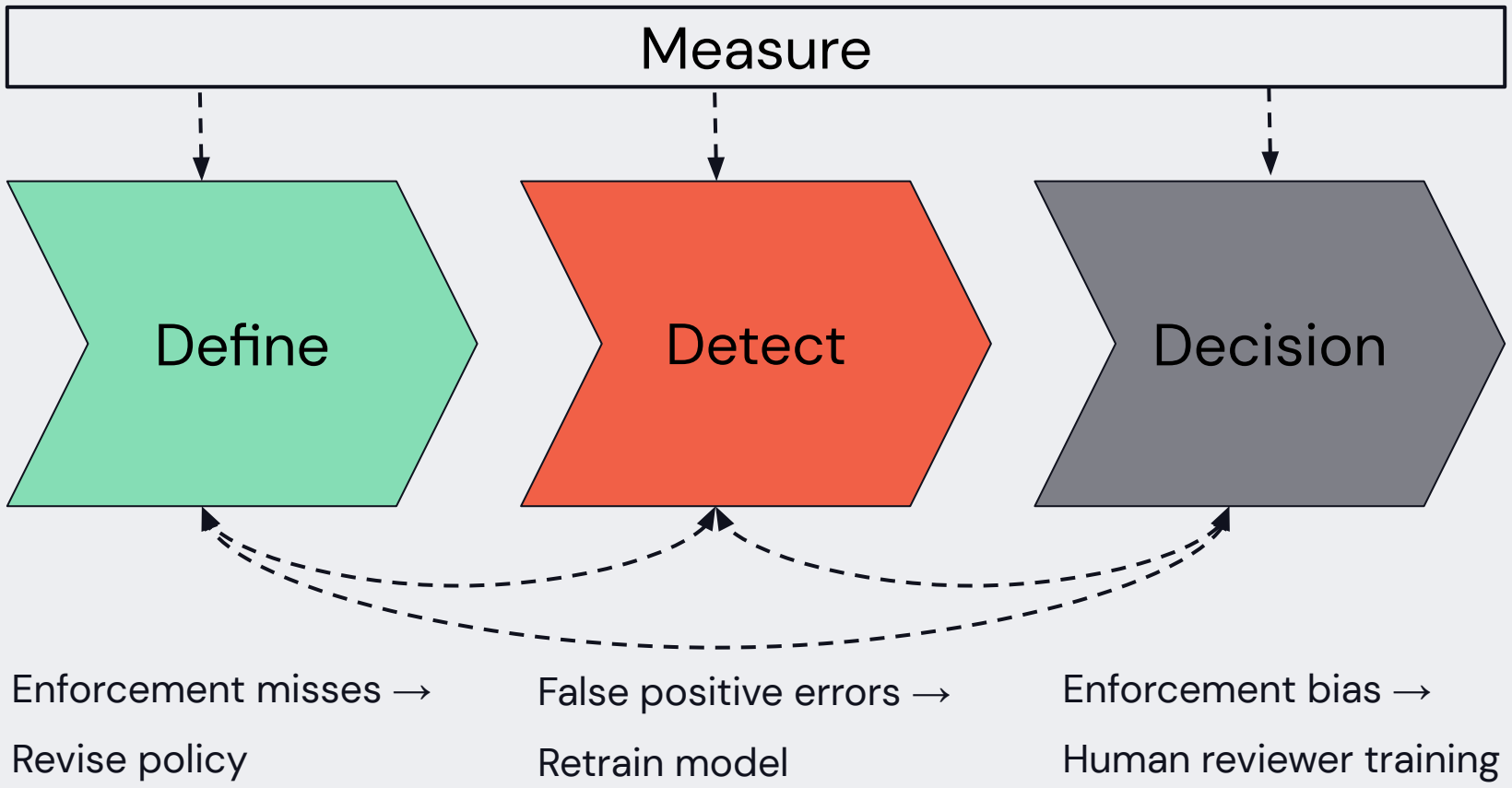
Leverage content similarity

Fanout decisions to near-duplicate content



Collaboration shortens feedback loops

Components need to talk to each other to adapt to changes



Key Takeaways

- **Goal:** Maximize recall, reduce false positives, keep opex within budget
- **Taming the tail**: Accurately enforcing on the long tail of content violations is necessary for high recall
- **Adaptiveness:** Build for adaptiveness rather than perfection

DATA+AI
SUMMIT 2022

Drop us a note: hello@refuel.ai

p.s. - We are hiring (a lot)

Nihit Desai
Co-founder & CTO, Refuel.AI

nihit@refuel.ai

@nihit_desai