Online Learning with Sublinear Best-Action Queries

NeurIPS 2024, Vancouver (Canada)

Matteo Russo
Sapienza University Rome

Andrea Celli
Bocconi University

Riccardo Colini-Baldeschi
Meta

Federico Fusco
Sapienza University Rome

Daniel Haimovich
Meta

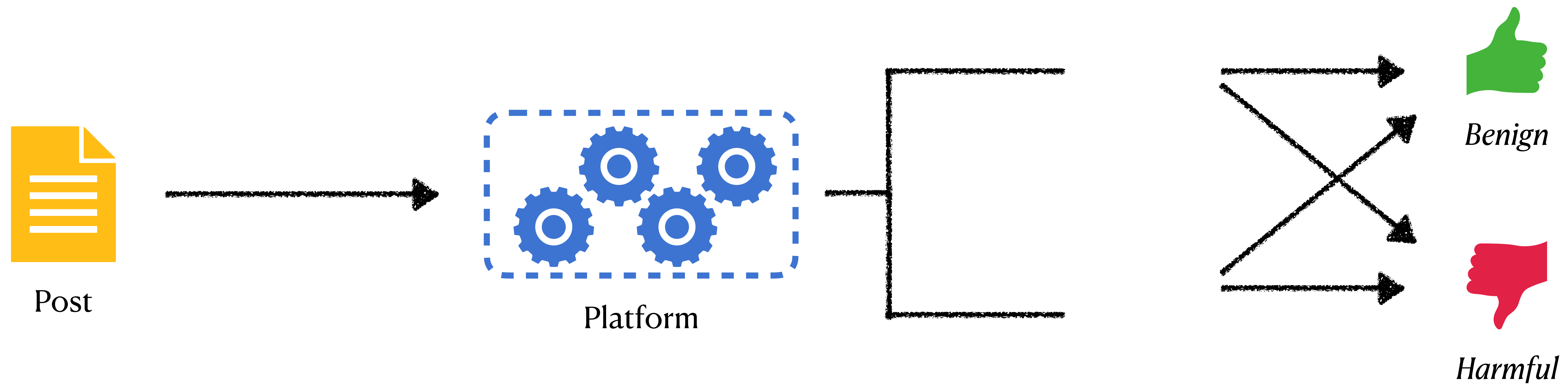
Dima Karamshuk
Meta

Stefano Leonardi
Sapienza University Rome

Niek Tax
Meta

Example: Online Platforms

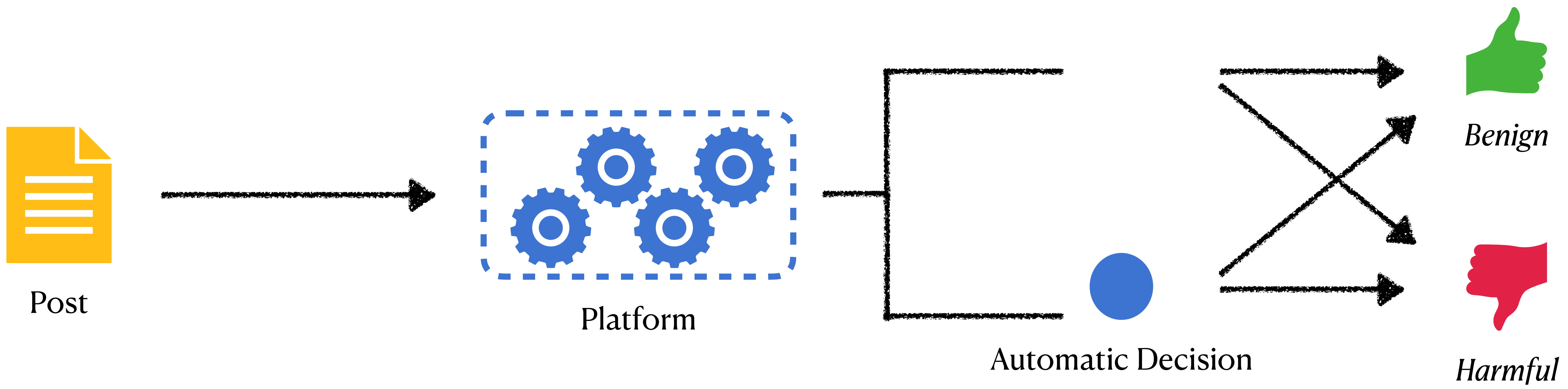
How to continuously moderate posted content



- Posts come one after the other and platform has to flag content as harmful or not

Example: Online Platforms

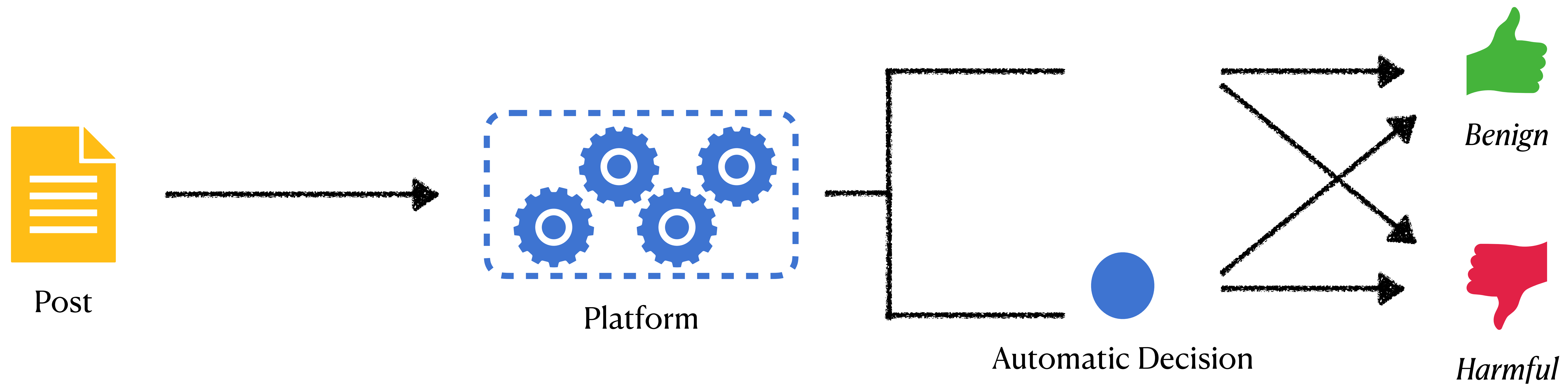
How to continuously moderate posted content



- Posts come one after the other and platform has to flag content as harmful or not

Example: Online Platforms

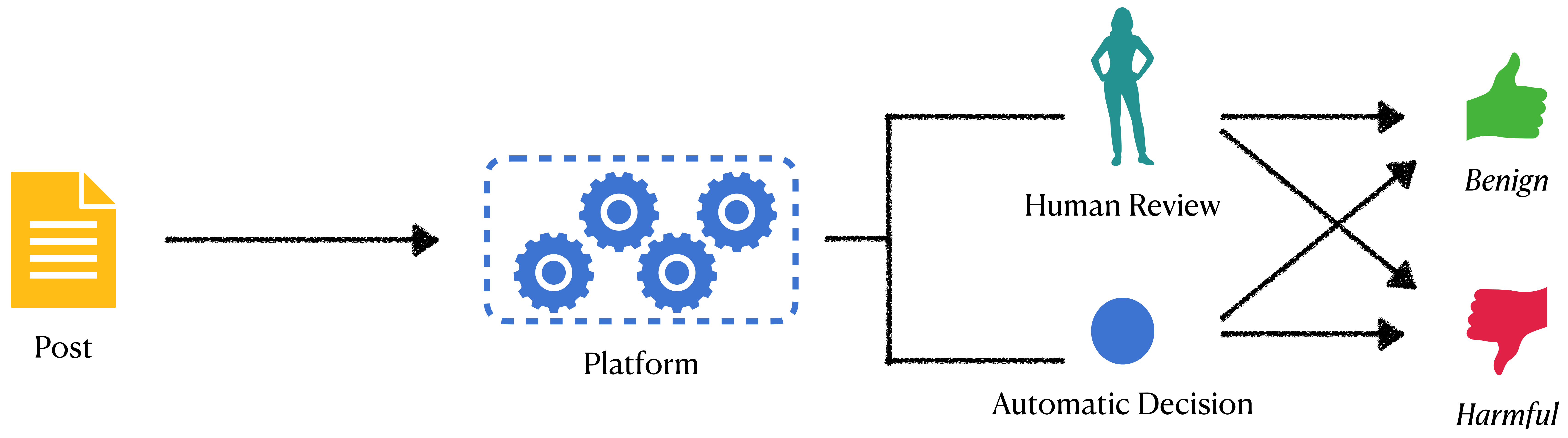
How to continuously moderate posted content



- Posts come one after the other and platform has to flag content as harmful or not
- Either by an automatic decision that can make mistakes

Example: Online Platforms

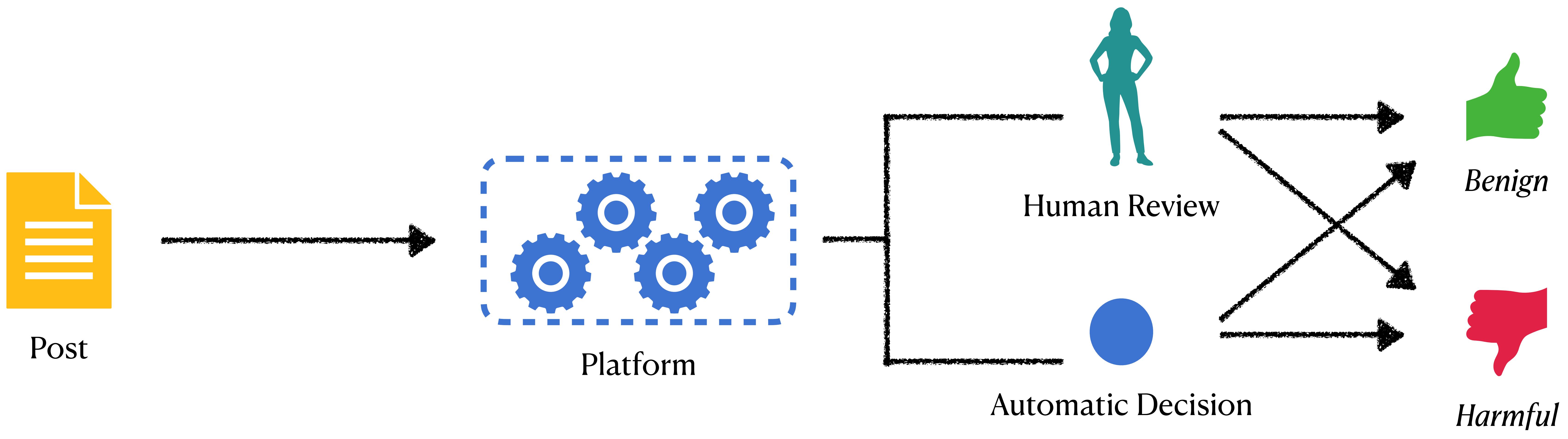
How to continuously moderate posted content



- Posts come one after the other and platform has to flag content as harmful or not
- Either by an automatic decision that can make mistakes

Example: Online Platforms

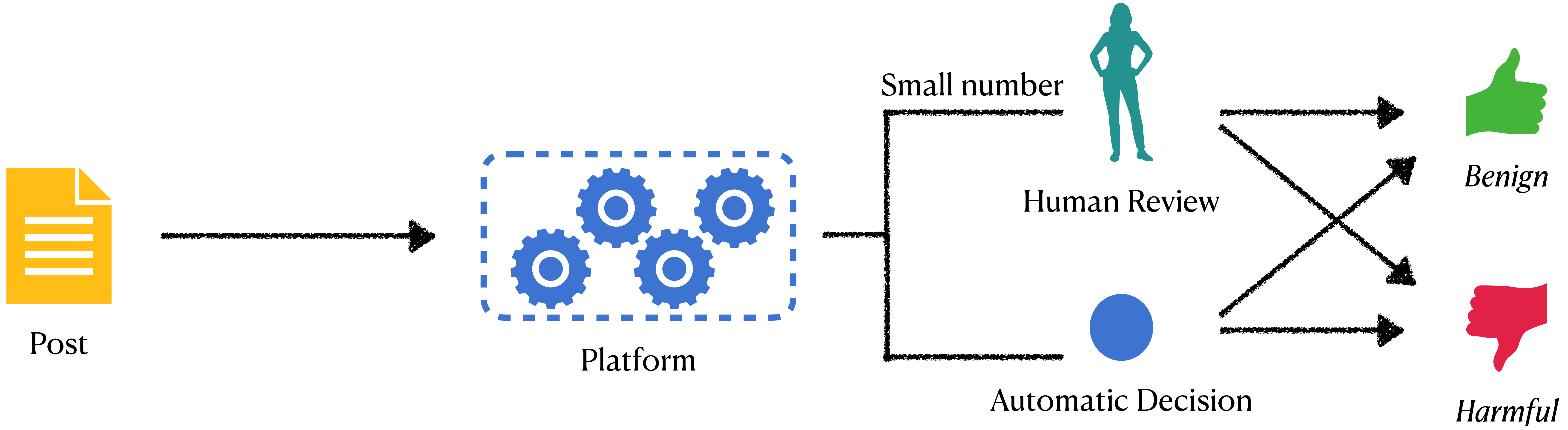
How to continuously moderate posted content



- Posts come one after the other and platform has to flag content as harmful or not
- Either by an automatic decision that can make mistakes
- Or by asking for an (expert) human review which we assume to be perfect

Example: Online Platforms

How to continuously moderate posted content



- Posts come one after the other and platform has to flag content as harmful or not
- Either by an automatic decision that can make mistakes
- Or by asking for an (expert) human review which we assume to be perfect

Learning Protocol

Online Learning with Best-Action Queries

Learning Protocol

Online Learning with Best-Action Queries

- **Setting**: n possible actions and k best-action queries available

Learning Protocol

Online Learning with Best-Action Queries

- **Setting**: n possible actions and k best-action queries available
- **For** time $t = 1, \ldots, T$:

1. A (hidden) loss $\ell_t(i)$ arrives for each action $i \in [n]$
2. The learner

   A. Either takes action i_t at time t

   B. Or is told the identity of the best action i_t^* at time t, and takes it

3. The learner incurs a (hidden) loss $\ell_t(i_t)$ or $\ell_t(i_t^*)$
4. A feedback z_t is revealed

Learning Protocol

Online Learning with Best-Action Queries

- **Setting**: n possible actions and k best-action queries available
- **For** time $t = 1, \ldots, T$:

1. A (hidden) loss $\ell_t(i)$ arrives for each action $i \in [n]$
2. The learner
   A. Either takes action i_t at time t
   B. Or is told the identity of the best action i_t^* at time t, and takes it
3. The learner incurs a (hidden) loss $\ell_t(i_t)$ or $\ell_t(i_t^*)$
4. A feedback z_t is revealed

Learning Protocol

Online Learning with Best-Action Queries

- **Setting**: n possible actions and k best-action queries available
- **For** time $t = 1, \ldots, T$:

1. A (hidden) loss $\ell_t(i)$ arrives for each action $i \in [n]$
2. The learner

   A. Either takes action i_t at time t

   $\leq k$ times

   B. Or is told the identity of the best action i_t^* at time t, and takes it
3. The learner incurs a (hidden) loss $\ell_t(i_t)$ or $\ell_t(i_t^*)$
4. A feedback z_t is revealed

The Model

Adversary, Queries & Feedback, Regret

The Model

Adversary, Queries & Feedback, Regret

- We assume losses to be generated by an **oblivious adversary**

The Model

Adversary, Queries & Feedback, Regret

- We assume losses to be generated by an **oblivious adversary**
- At time step t, before feedback is received, a **best-action query** reveals the *identity* of the best action at that time step, i.e., $i_t^* := \arg\min_{i \in [n]} \ell_t(i)$

The Model

Adversary, Queries & Feedback, Regret

- We assume losses to be generated by an **oblivious adversary**

$\leq k$ times

- At time step t, before feedback is received, a **best-action query** reveals the *identity* of the best action at that time step, i.e., $i_t^* := \arg\min_{i \in [n]} \ell_t(i)$

The Model

Adversary, Queries & Feedback, Regret

- We assume losses to be generated by an **oblivious adversary**
- At time step t, before feedback is received, a **best-action query** reveals the *identity* of the best action at that time step, i.e., $i_t^* := \arg\min_{i \in [n]} \ell_t(i)$

$\leq k$ times

- We study two types of feedback regimes

1. **Full feedback**: All losses revealed at all time steps, i.e., $z_t = (\ell_t(i))_{i \in [n]}$
2. **Label-efficient feedback**: All losses revealed *only after* a querying time step

The Model

Adversary, Queries & Feedback, Regret

$\leq k$ times

- We assume losses to be generated by an **oblivious adversary**
- At time step t, before feedback is received, a **best-action query** reveals the *identity* of the best action at that time step, i.e., $i_t^* := \arg\min_{i \in [n]} \ell_t(i)$
- We study two types of feedback regimes

  1. **Full feedback**: All losses revealed at all time steps, i.e., $z_t = (\ell_t(i))_{i \in [n]}$
  2. **Label-efficient feedback**: All losses revealed *only after* a querying time step

- We want to understand how the **regret** grows:

$$R_T := \sum_{t \in [T]} \mathbb{E}[\ell_t(i_t)] - \min_{i \in [n]} \sum_{t \in [T]} \ell_t(i)$$

Our Results

Upper and Lower Bounds

Regret	Classical No Query	Low Query	Sublinear Query
Full feedback	$k = 0$ $R_T \in \Theta\left(\sqrt{T}\right)$	$k \in O\left(\sqrt{T}\right)$ $R_T \in \Theta\left(\sqrt{T}\right)$	$k \in \Omega\left(\sqrt{T}\right)$ $R_T \in \Theta\left(\frac{T}{k}\right)$
Label-efficient feedback			

Our Results

Upper and Lower Bounds

Regret	Classical No Query	Low Query	Sublinear Query
Full feedback	$k = 0$ $R_T \in \Theta\left(\sqrt{T}\right)$	$k \in O\left(\sqrt{T}\right)$ $R_T \in \Theta\left(\sqrt{T}\right)$	$k \in \Omega\left(\sqrt{T}\right)$ $R_T \in \Theta\left(\frac{T}{k}\right)$
Label-efficient feedback	$k = 0$ $R_T \in \Theta\left(\frac{T}{\sqrt{k}}\right)$	$k \in O\left(T^{2/3}\right)$ $R_T \in \Theta\left(\frac{T}{\sqrt{k}}\right)$	$k \in \Omega\left(T^{2/3}\right)$ $R_T \in \Theta\left(\frac{T^2}{k^2}\right)$

Our Results

Upper and Lower Bounds

Regret	Classical No Query	Low Query	Sublinear Query
Full feedback	$k = 0$ $R_T \in \Theta\left(\sqrt{T}\right)$	$k \in O\left(\sqrt{T}\right)$ $R_T \in \Theta\left(\sqrt{T}\right)$	$k \in \Omega\left(\sqrt{T}\right)$ $R_T \in \Theta\left(\frac{T}{k}\right)$
Label-efficient feedback	$k = 0$ $R_T \in \Theta\left(\frac{T}{\sqrt{k}}\right)$	$k \in O\left(T^{2/3}\right)$ $R_T \in \Theta\left(\frac{T}{\sqrt{k}}\right)$	$k \in \Omega\left(T^{2/3}\right)$ $R_T \in \Theta\left(\frac{T^2}{k^2}\right)$

Upper Bound

- **Full feedback**: *Hedge* on *true* losses equipped with k uniform random queries across the time horizon (+ refined analysis)
- **Label-efficient feedback**: *Hedge* on *estimated* losses equipped with uniform probability querying until query budget exhaustion (+ refined analysis)

Our Results

Upper and Lower Bounds

Regret	Classical No Query	Low Query	Sublinear Query
Full feedback	$k = 0$ $R_T \in \Theta\left(\sqrt{T}\right)$	$k \in O\left(\sqrt{T}\right)$ $R_T \in \Theta\left(\sqrt{T}\right)$	$k \in \Omega\left(\sqrt{T}\right)$ $R_T \in \Theta\left(\frac{T}{k}\right)$
Label-efficient feedback	$k = 0$ $R_T \in \Theta\left(\frac{T}{\sqrt{k}}\right)$	$k \in O\left(T^{2/3}\right)$ $R_T \in \Theta\left(\frac{T}{\sqrt{k}}\right)$	$k \in \Omega\left(T^{2/3}\right)$ $R_T \in \Theta\left(\frac{T^2}{k^2}\right)$

Upper Bound

- **Full feedback**: *Hedge* on *true* losses equipped with k uniform random queries across the time horizon (+ refined analysis)
- **Label-efficient feedback**: *Hedge* on *estimated* losses equipped with uniform probability querying until query budget exhaustion (+ refined analysis)

Lower Bound

- **Full and label-efficient feedback**: Two actions where queries cannot help more than T/k and T^2/k^2

Our Results

Upper and Lower Bounds

Regret	Classical No Query	Low Query	Sublinear Query
Full feedback	$k = 0$ $R_T \in \Theta\left(\sqrt{T}\right)$	$k \in O\left(\sqrt{T}\right)$ $R_T \in \Theta\left(\sqrt{T}\right)$	$k \in \Omega\left(\sqrt{T}\right)$ $R_T \in \Theta\left(\frac{T}{k}\right)$
Label-efficient feedback	$k = 0$ $R_T \in \Theta\left(\frac{T}{\sqrt{k}}\right)$	$k \in O\left(T^{2/3}\right)$ $R_T \in \Theta\left(\frac{T}{\sqrt{k}}\right)$	$k \in \Omega\left(T^{2/3}\right)$ $R_T \in \Theta\left(\frac{T^2}{k^2}\right)$

Upper Bound

- **Full feedback**: *Hedge* on *true* losses equipped with k uniform random queries across the time horizon (+ refined analysis)
- **Label-efficient feedback**: *Hedge* on *estimated* losses equipped with uniform probability querying until query budget exhaustion (+ refined analysis)

Lower Bound

- **Full and label-efficient feedback**: Two actions where queries cannot help more than T/k and T^2/k^2

Future

- What about **bandit feedback, feedback graphs, partial monitoring feedback**?
- What if queries are **not perfect**?

Thank you!