

facebook

The Abuse Uncertainty Principle

And Other Lessons Learned from Measuring Abuse on the Internet

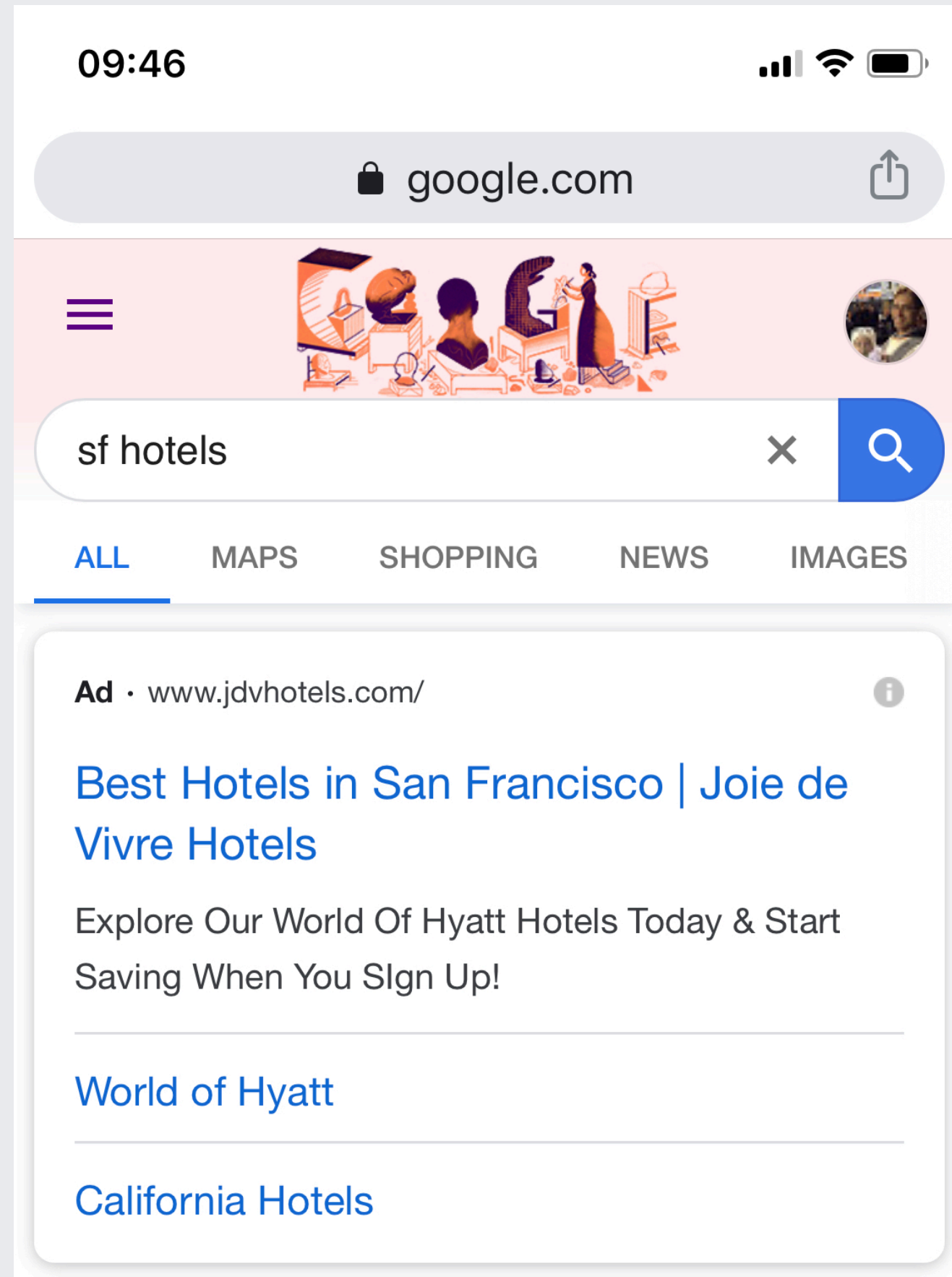
David Freeman

Research Scientist/Engineer, Facebook

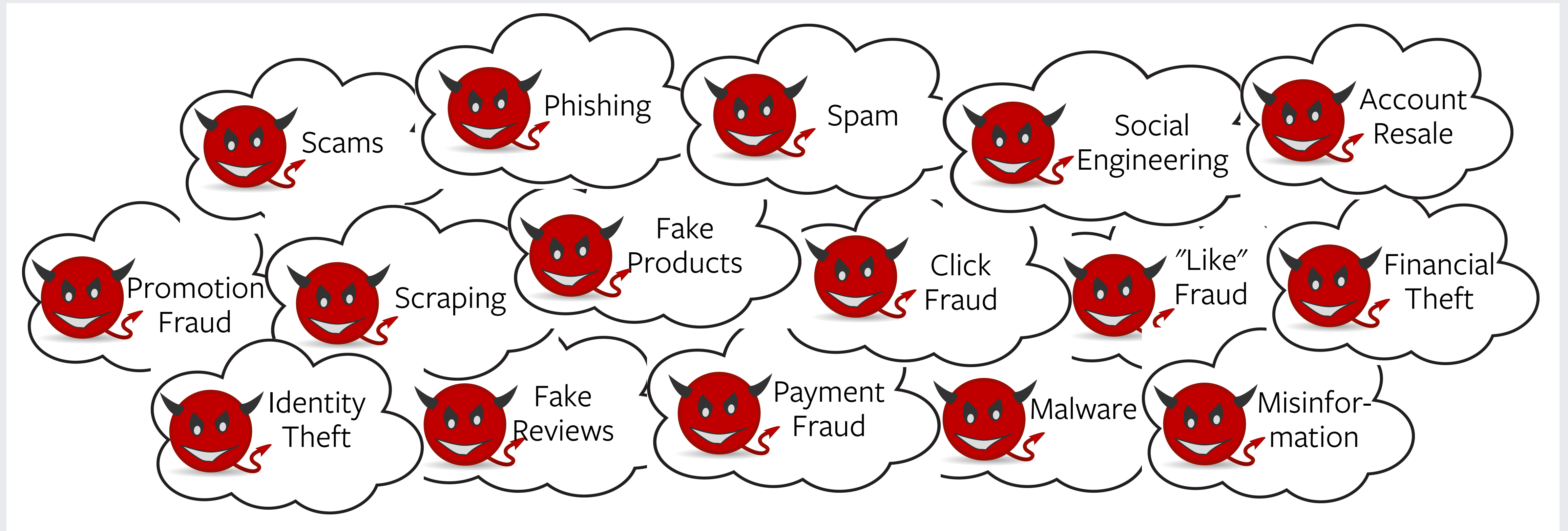
USENIX Enigma 2020

San Francisco, CA, 29 January 2020

Sometimes I wish I worked on ad ranking...

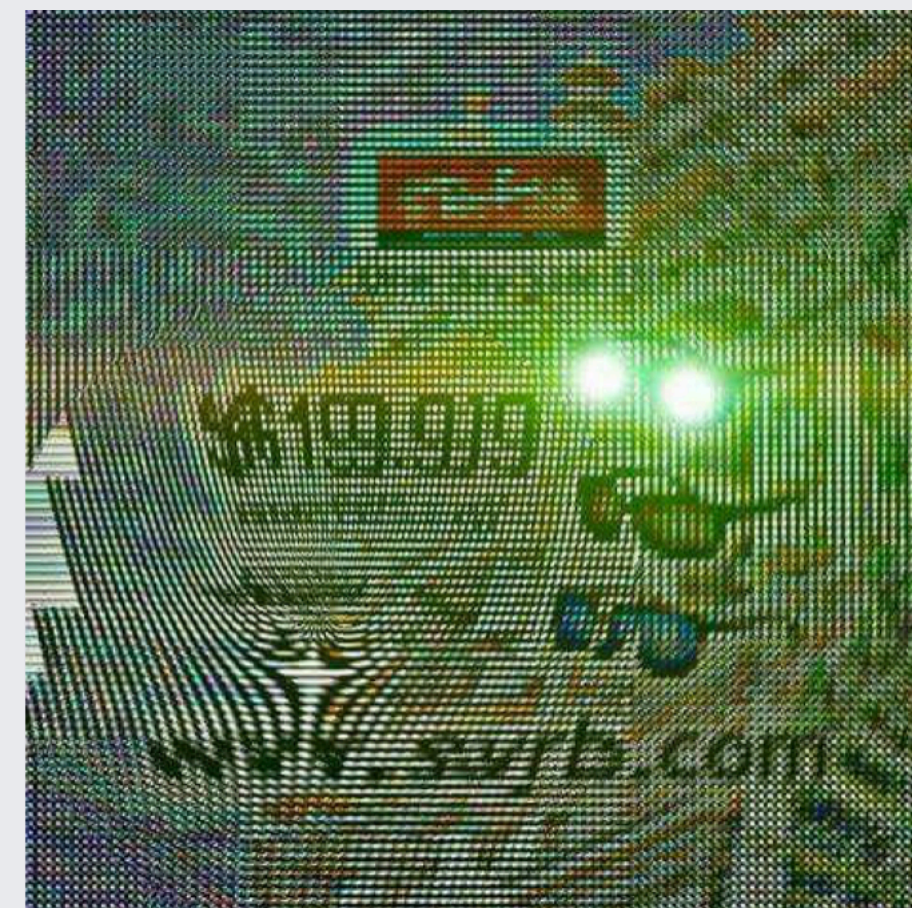


...but I don't!



- How do we figure out what to work on?
- How do we measure the impact of our work?

Abusive content on Facebook



facebook

WARNING !! Your Account Has Violated Terms on Facebook.

Warning: Your account will be disabled !!

Your Facebook account is Troubled. Your account has violated the provisions on Facebook. Security Systems has received reports from other users you violate the rules on Facebook which resulted in your account will be permanently disabled.

- » Post a rough profile or photos,
- » Insulting and threatening others (users)
- » Using facebook account just for promotion

Please confirm your account by clicking the link below:

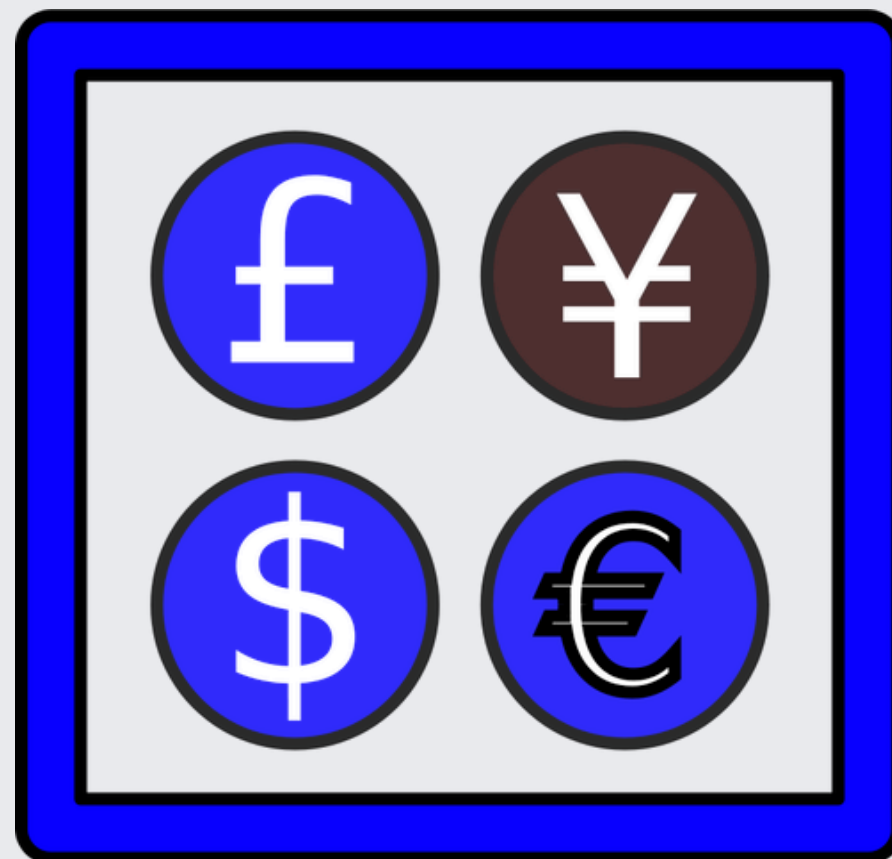
Confirm My Account

Attention:

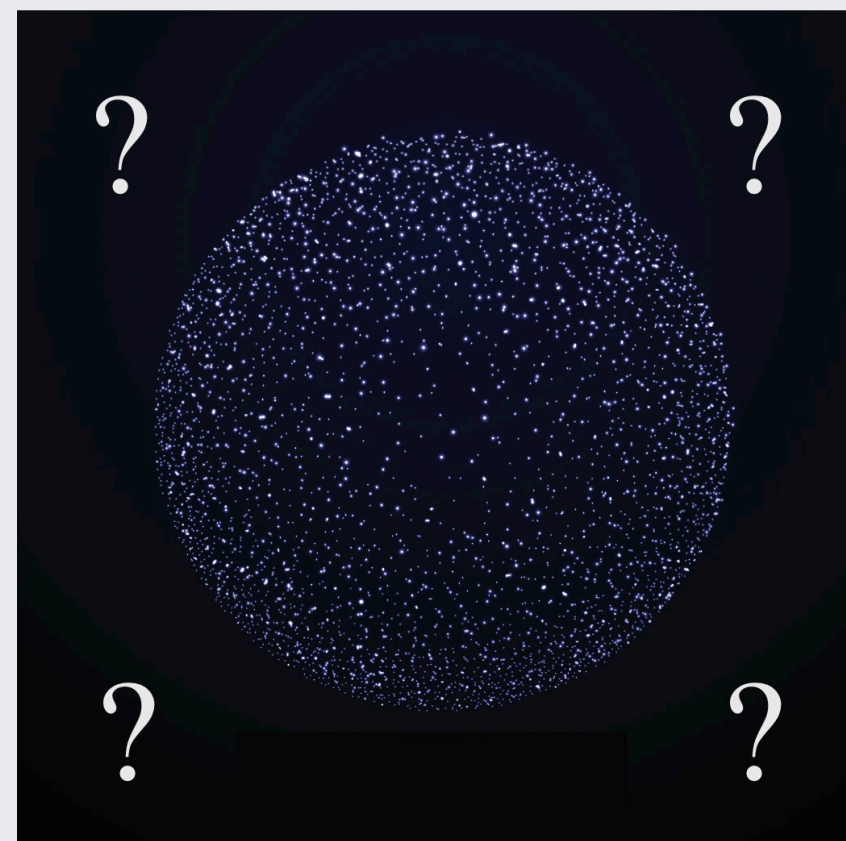
All accounts that are not verified within 24 hours

Measuring abuse is hard!

No common units



Unknown unknowns



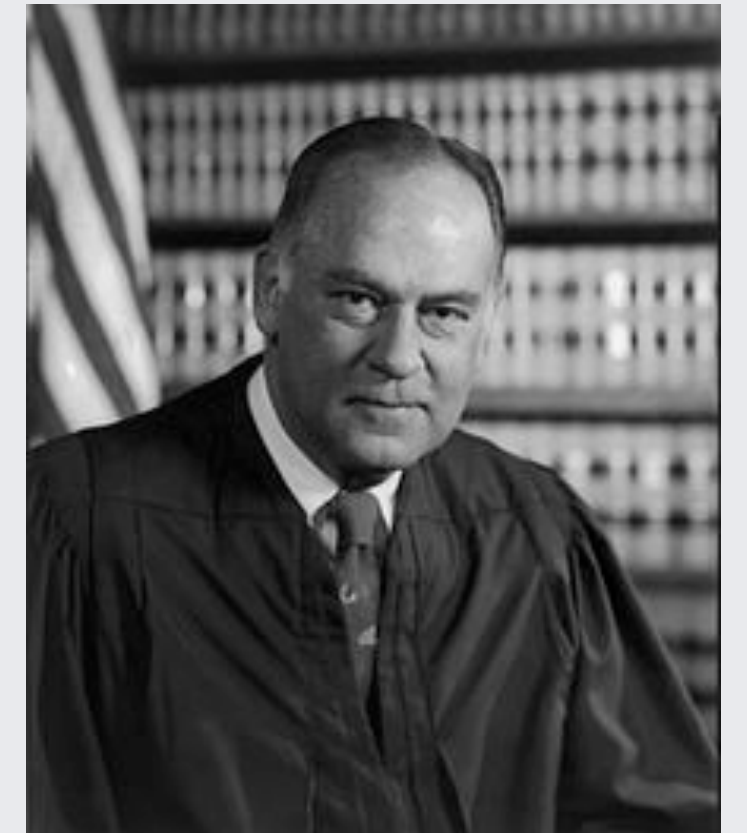
Adversarial response



Category imbalance

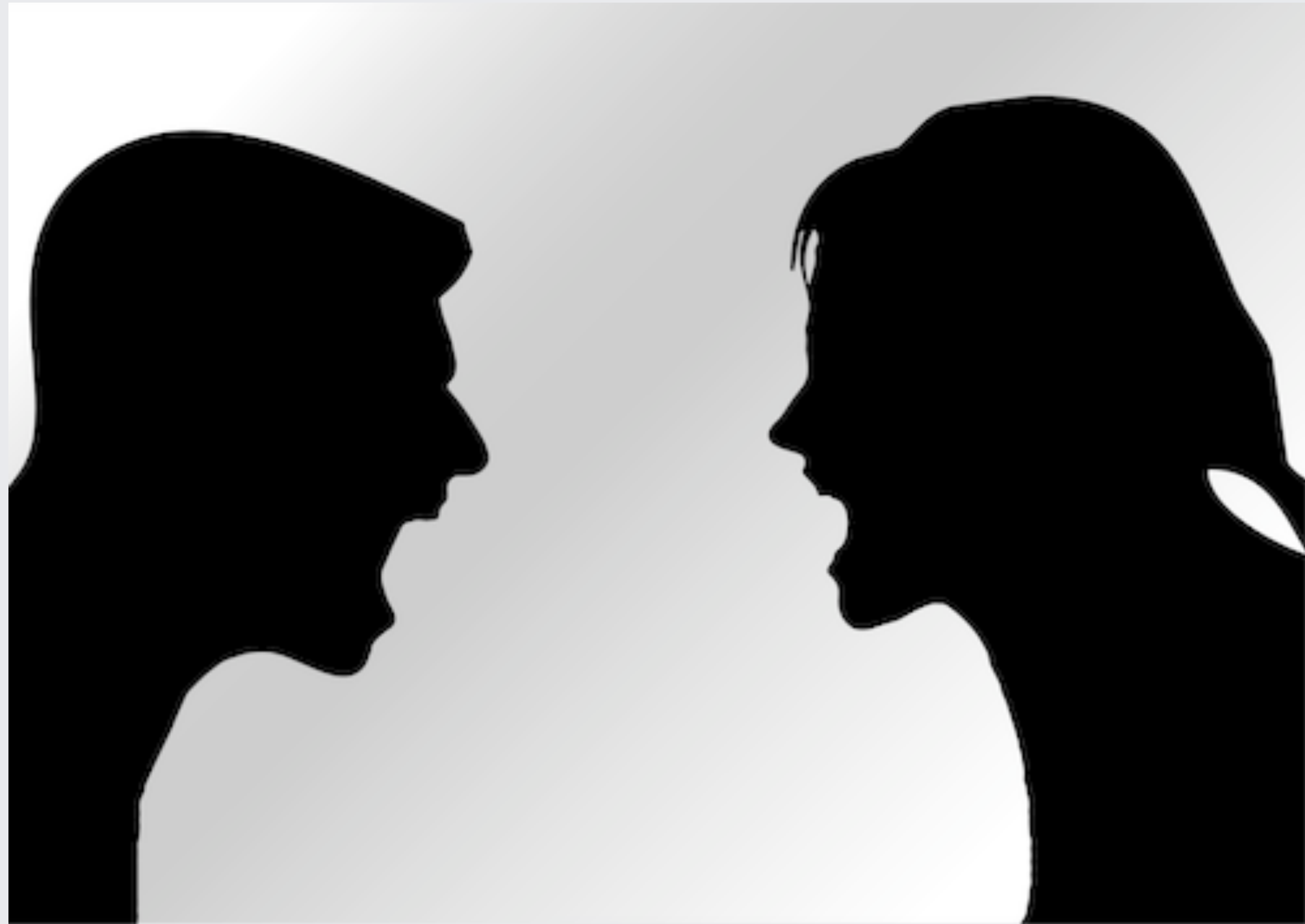


What is ground truth?



Measuring abuse: The Dark Ages

Prioritization

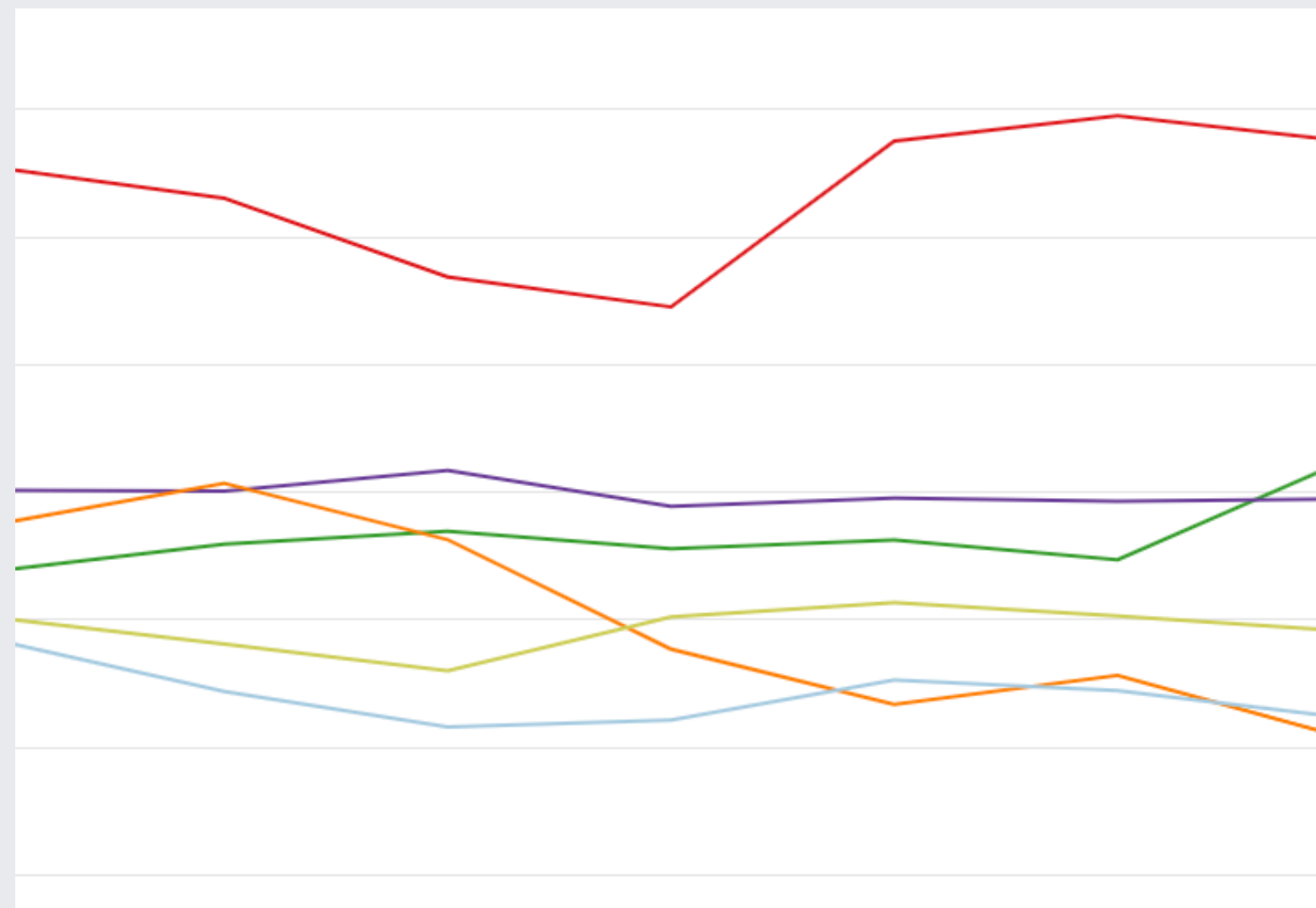


Impact



Measuring Abuse: The Renaissance

Prioritization



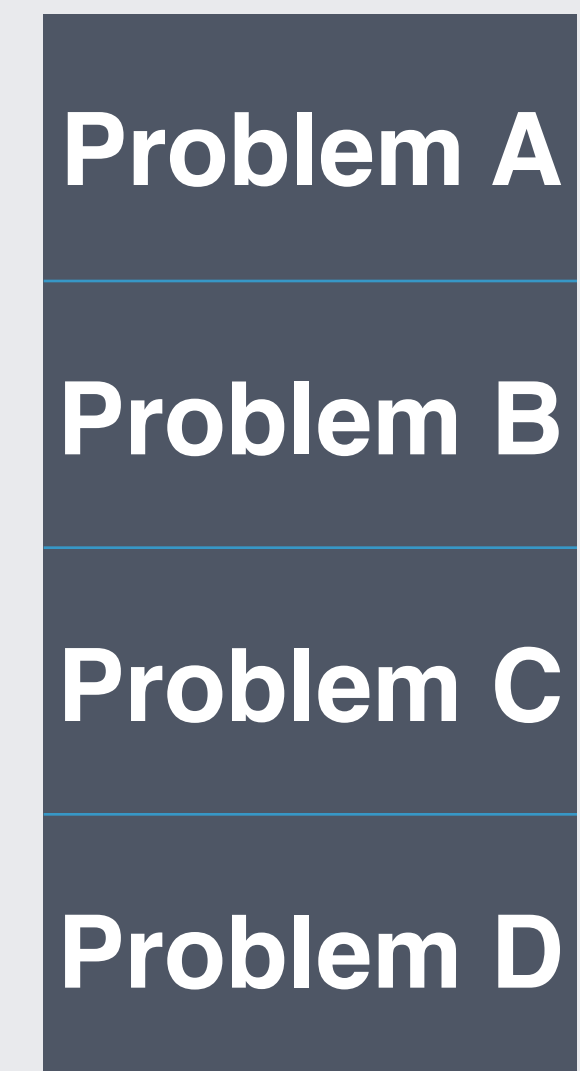
Volume

×

Problem	Harm Level
Spam	Yellow
Account Takeover	Orange
Identity Theft	Red

Harm
(per instance)

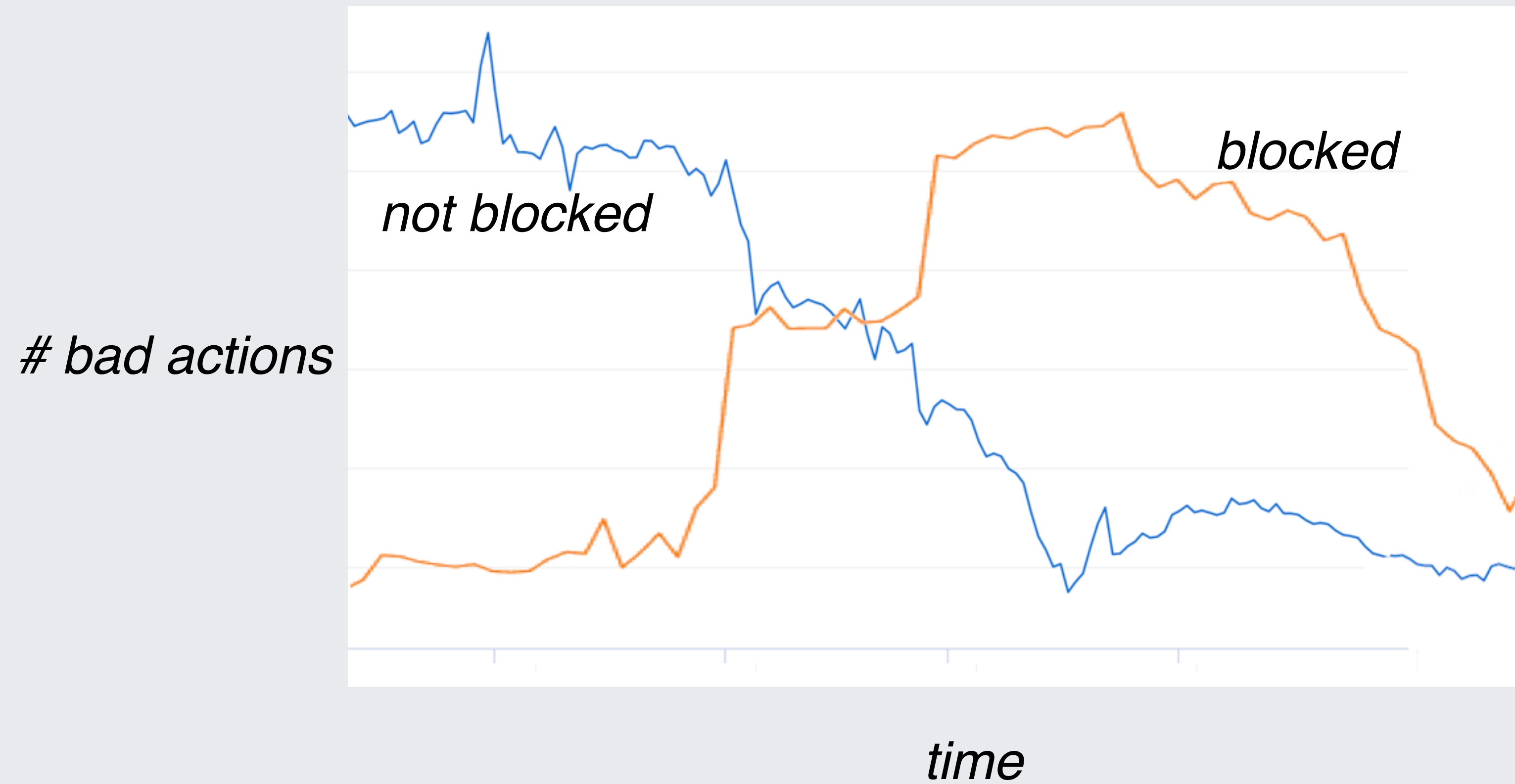
=



Stack Rank

Measuring Abuse: The Renaissance

Impact



Benefits of Measurement

(in addition to a quieter office)

- Track progress (and regressions) over time
- Provide data for ML training
- Enable deep dives
- Allow retroactive takedown*

How do we get there?



How do we get there?

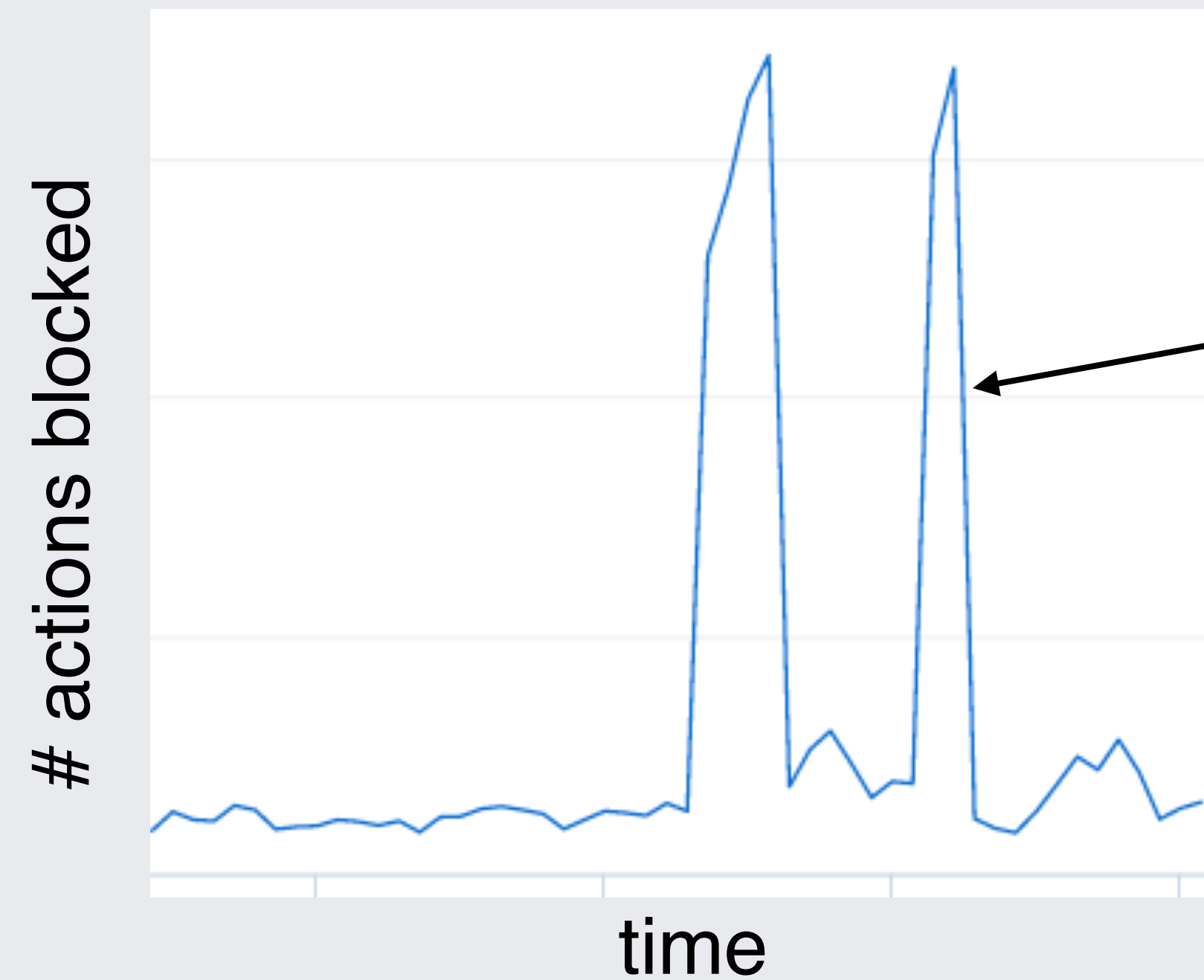


The “measurement printing press” is *Labeling*.

How do we label abuse?

Approach 1: Count what you blocked

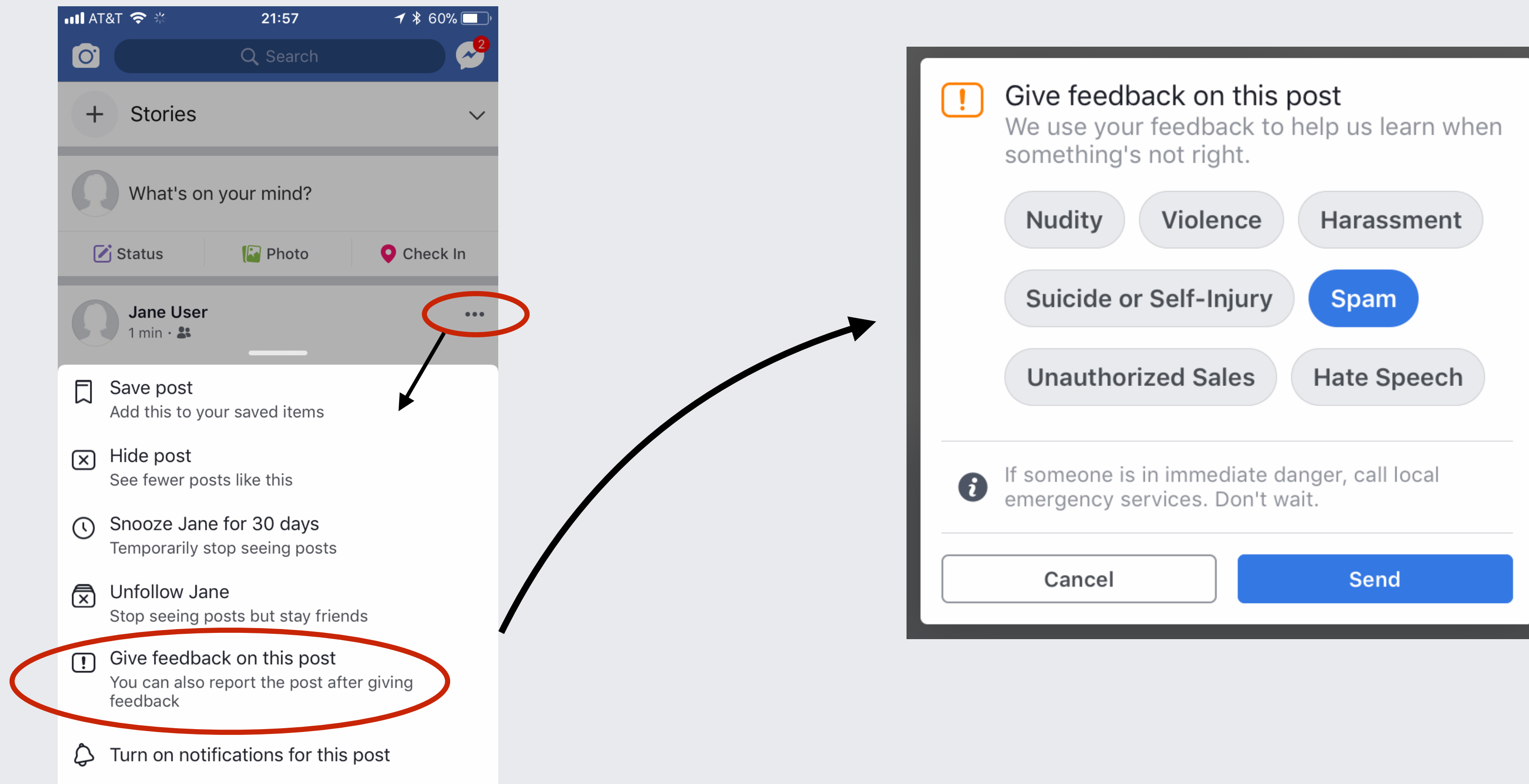
These things have to be bad, right?



is this spike
good or bad?

Approach 2: Have users do the labeling

Crowdsource the work via reporting/appeals flows.



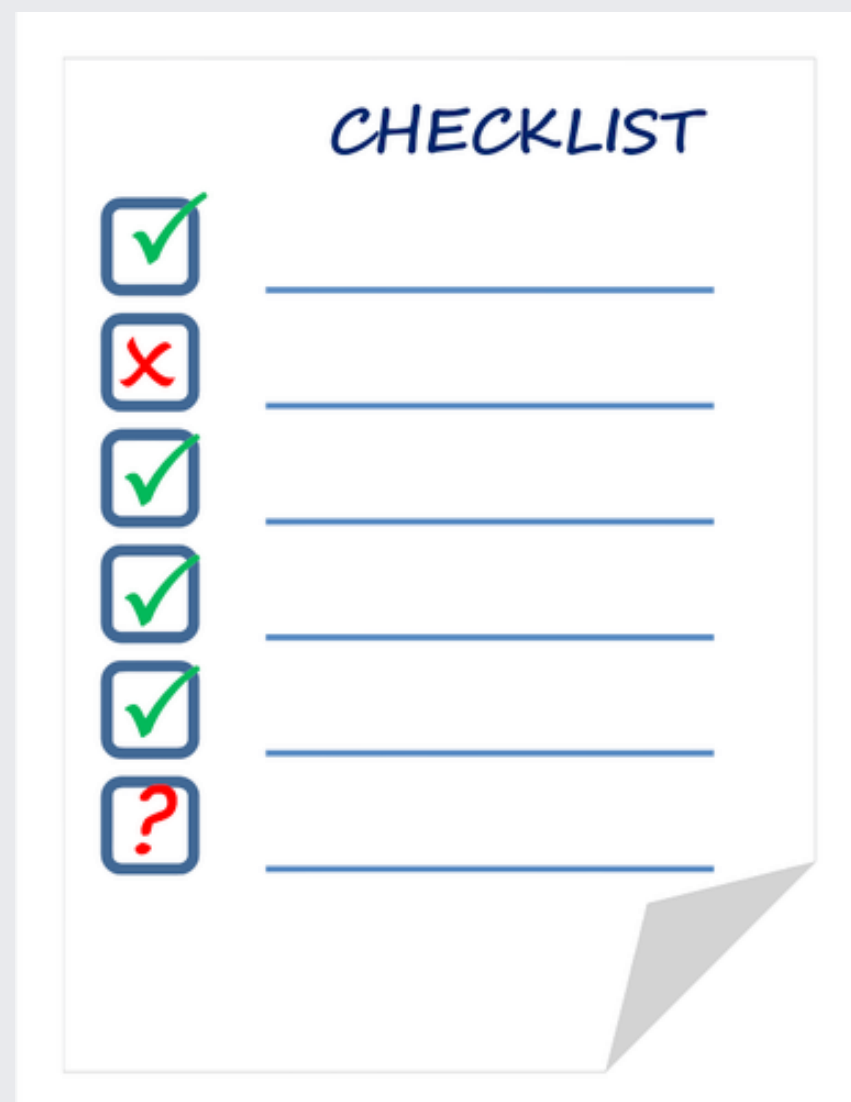
Approach 3: Human labeling

Get experts to decide for you.

i. Rubric

ii. Sampling

iii. Labeling

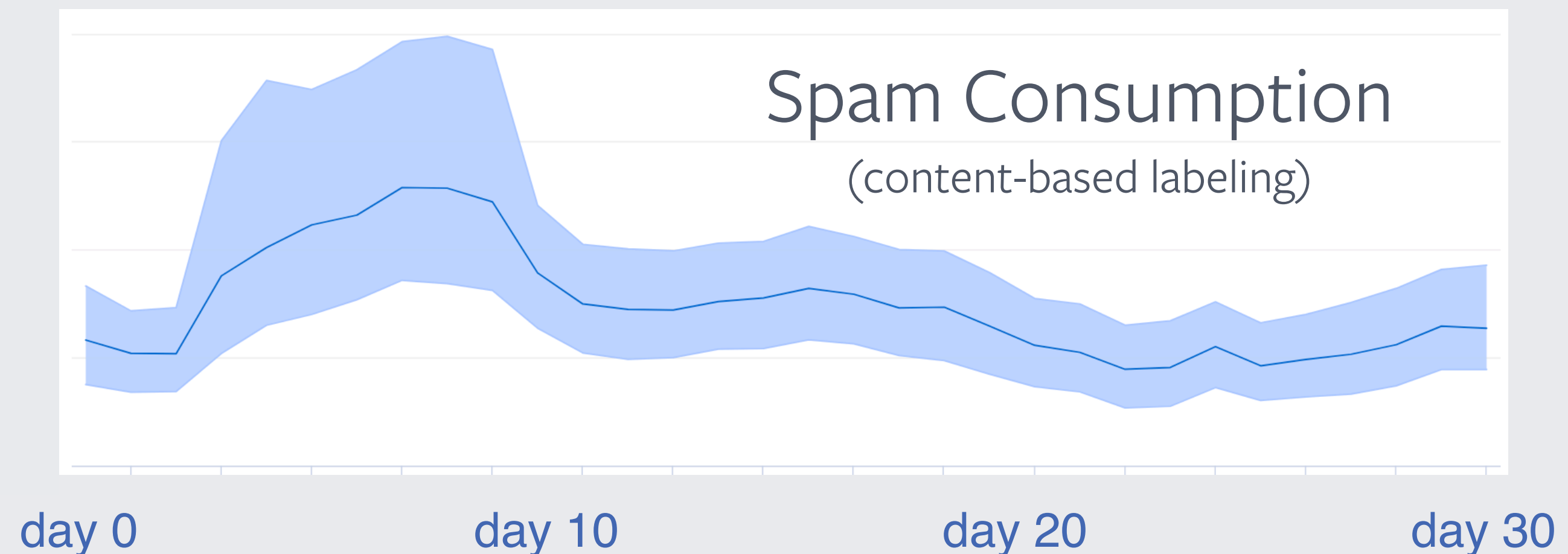
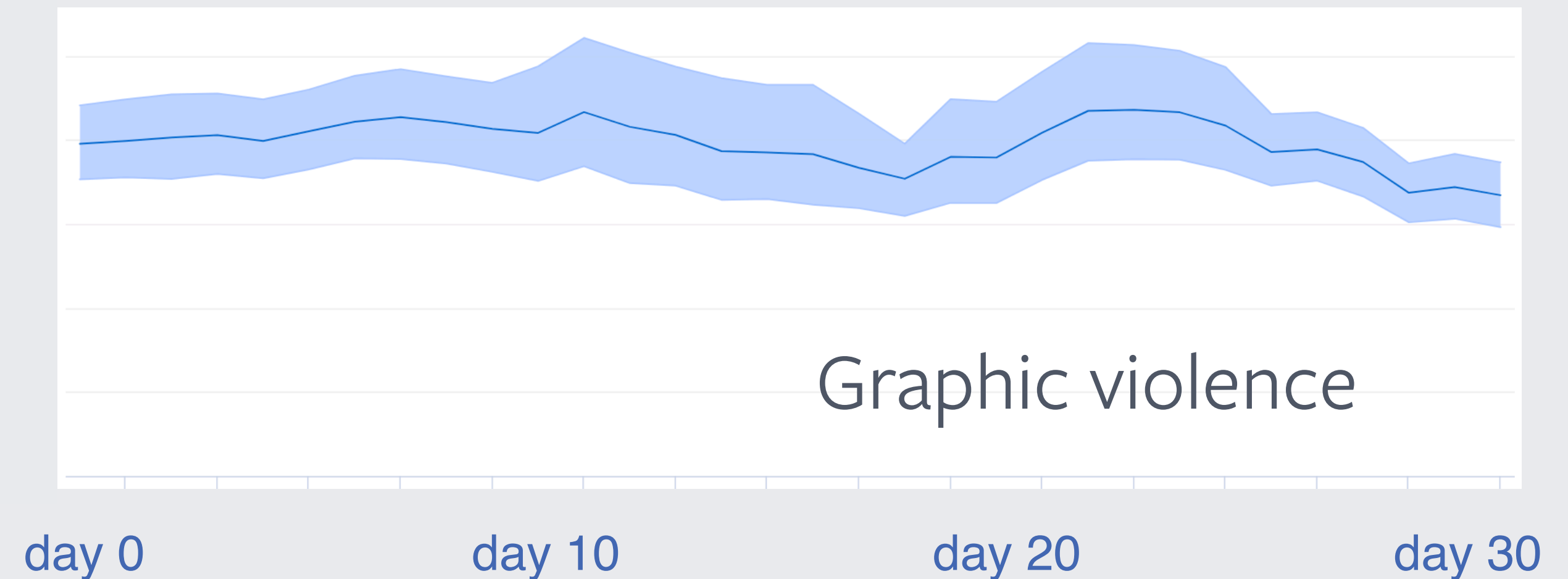


Key insight: use **stratified sampling** to oversample suspicious activity.

Deployment at Facebook

Used for internal direction and/or external reporting of

- Spam
- Fake/abusive accounts
- Nudity & pornography
- Graphic violence
- Hate speech
- Drug sales
- Ad farms
- Clickbait



Approach 4: Automated labeling

Expertise has its limits

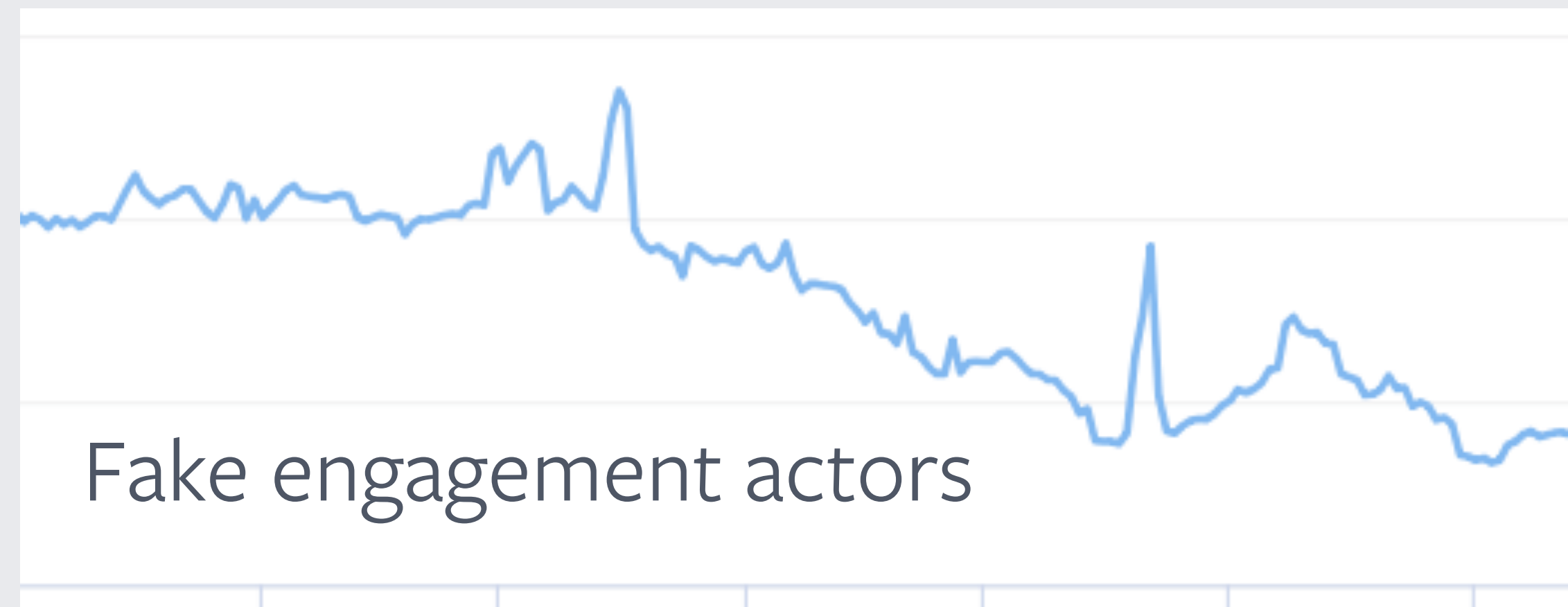
- Human labeling at scale requires standardized rubrics and infrastructure (e.g. content queue).
- Some problems don't fit!
 - e.g. fake likes, compromised accounts

Key idea: use **clustering & anomaly detection** to find examples of (scripted) abuse at scale.

Deployment at Facebook

Used for internal direction on

- Fake engagement
- Account compromise
- Scraping



Comparison of approaches

**Captures
all false
negatives**

**Points in the
correct
direction**







**Avoids
feedback
loops**

**Resists
adversarial
manipulation**













**Informs
detection**

**Scales to
new
problems**



















Comparison of approaches

	Captures all false negatives	Points in the correct direction	Avoids feedback loops	Resists adversarial manipulation	Informs detection	Scales to new problems
Count blocks						

























Comparison of approaches

	Captures all false negatives	Points in the correct direction	Avoids feedback loops	Resists adversarial manipulation	Informs detection	Scales to new problems
Count blocks						
User reporting						

Comparison of approaches

	Captures all false negatives	Points in the correct direction	Avoids feedback loops	Resists adversarial manipulation	Informs detection	Scales to new problems
Count blocks						
User reporting						
Human labeling						

Comparison of approaches

	Captures all false negatives	Points in the correct direction	Avoids feedback loops	Resists adversarial manipulation	Informs detection	Scales to new problems
Count blocks						
User reporting						
Human labeling						
Automated labeling						

Comparison of approaches

	Captures all false negatives	Points in the correct direction	Avoids feedback loops	Resists adversarial manipulation	Informs detection	Scales to new problems
Count blocks						
User reporting						
Human labeling						
Automated labeling						

Successful approaches ***amplify expert opinion.***

Labeling is only the first step

Goodhart's Law



Source: Wikipedia user Jamesfranklingresham

“When a measure becomes a target, it ceases to be a good measure.”

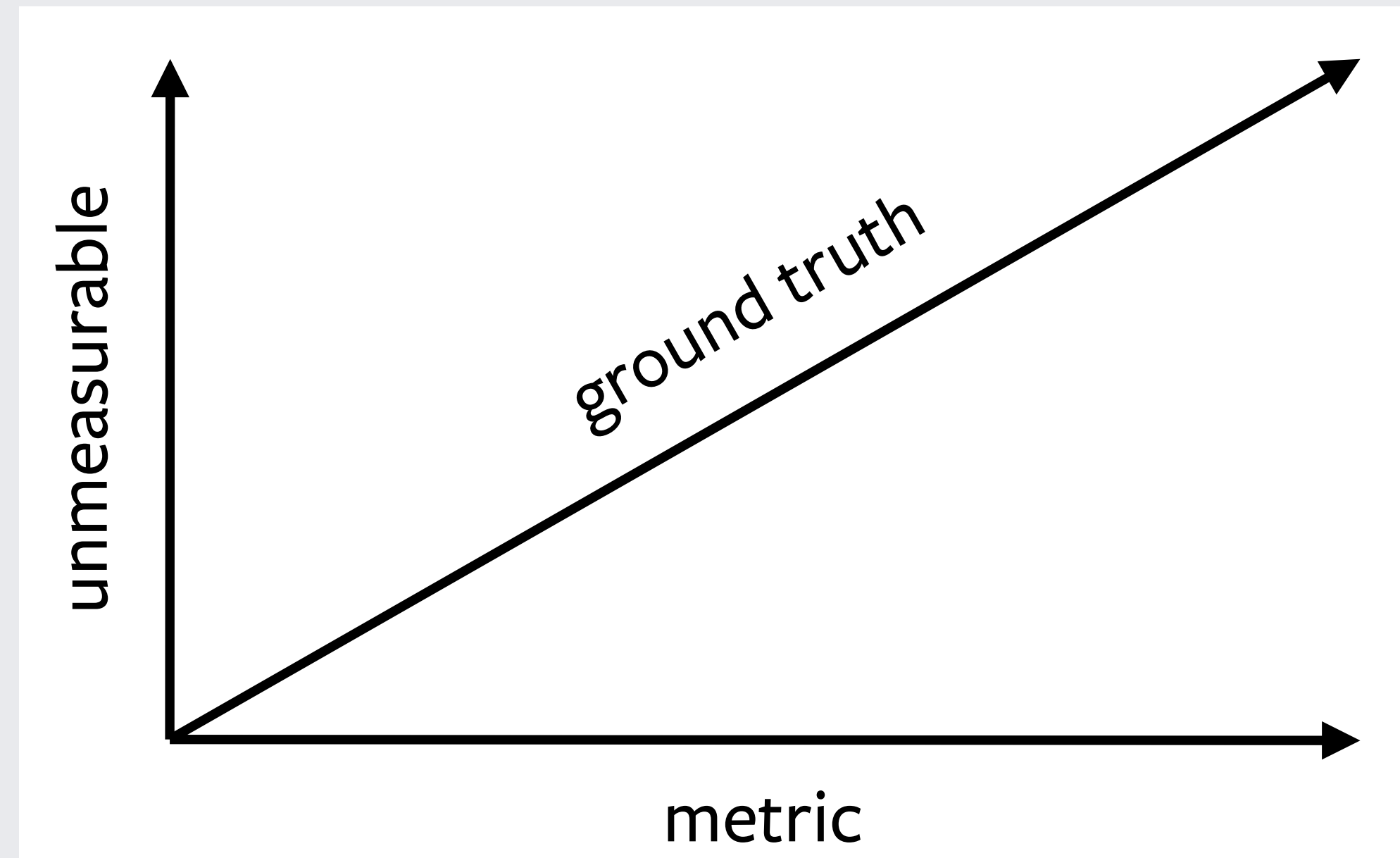
Example: Rate Limiting

Count users posting more than 100x/hour



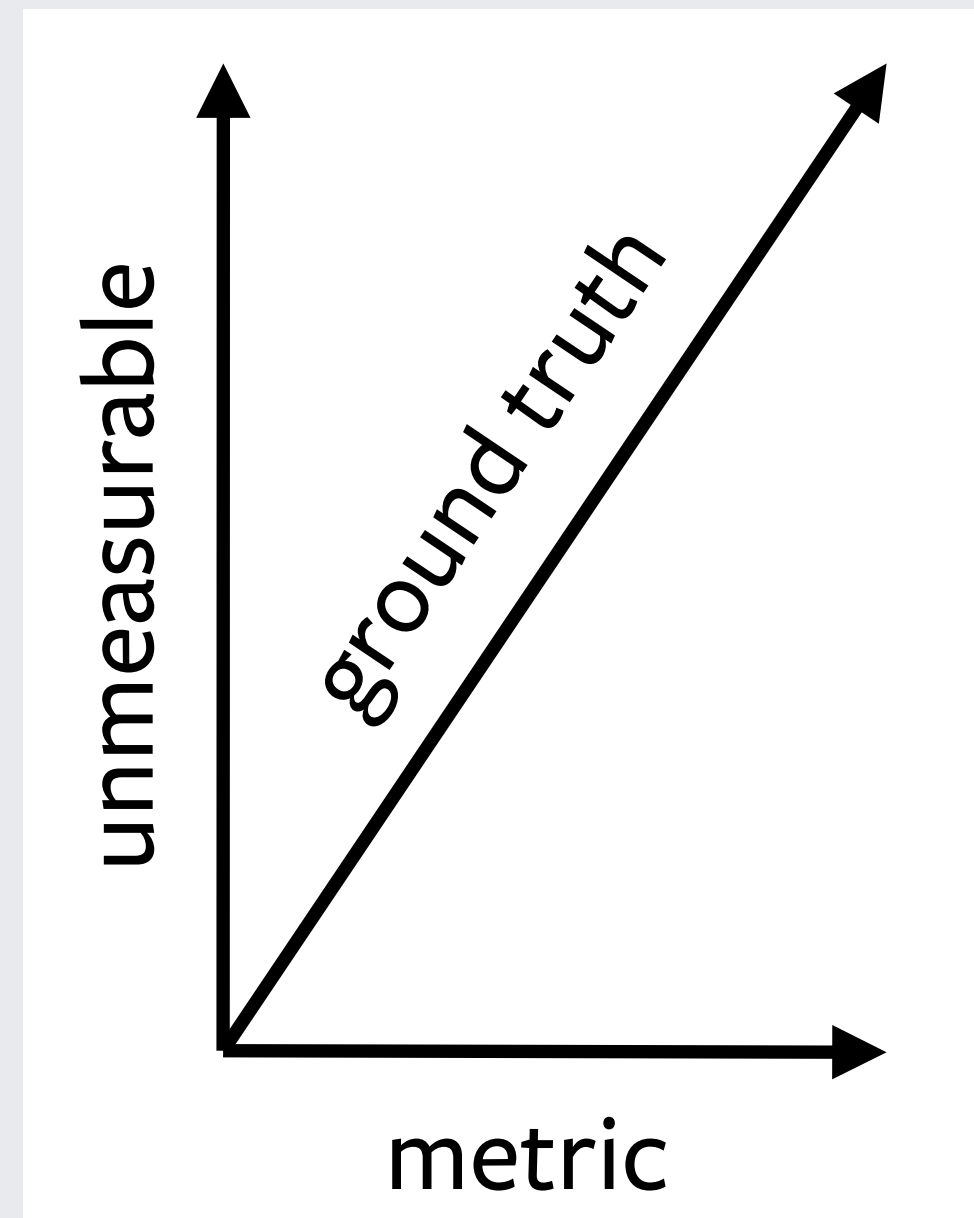
“Proof” of Goodhart’s Law

At metric creation time



“Proof” of Goodhart’s Law

After chasing the metric



Corollary: “Abuse Uncertainty Principle”



“Any abuse signal can be used for measurement or enforcement, but not both.”

Living with the Uncertainty Principle

For spam detection at Facebook, we split signals into two classes:

Measurement	Enforcement
Network connection	Counts and rates
HTTP request	Graph relations
User-generated content	Activity sequence

Living with the Uncertainty Principle

For spam detection at Facebook, we split signals into two classes:

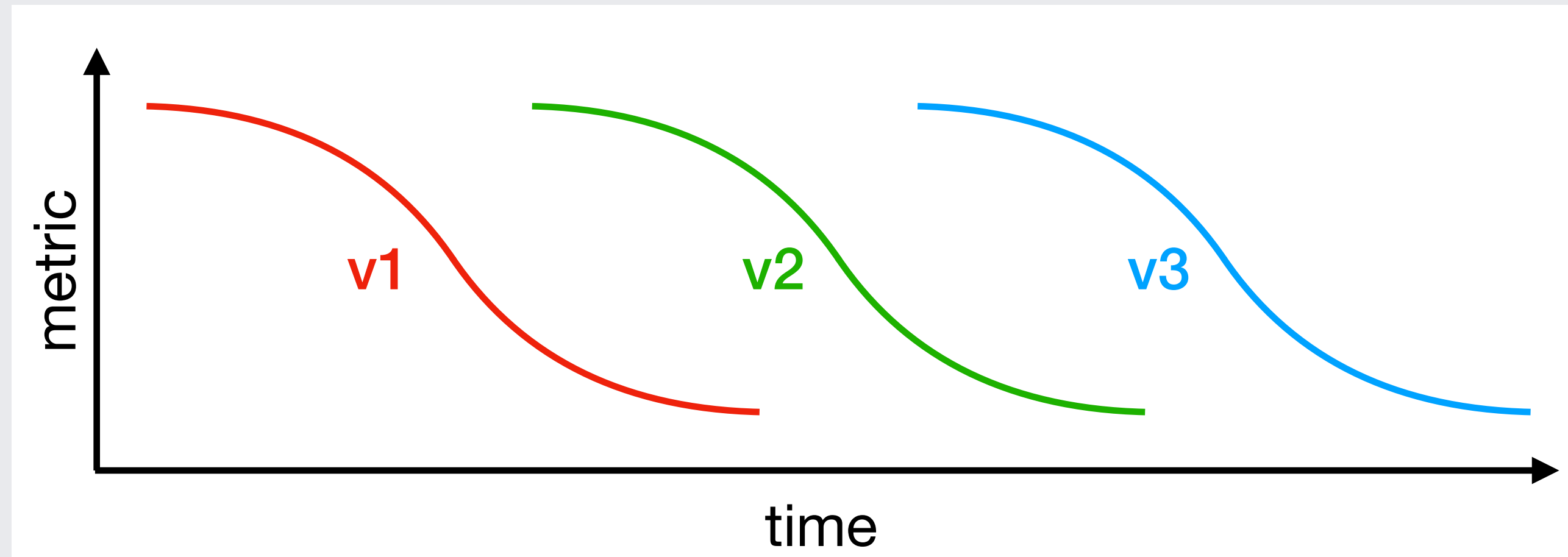
Measurement	Enforcement
Network connection	Counts and rates
HTTP request	Graph relations
User-generated content	Activity sequence

“We don’t want to be the ones solving the CAPTCHAs”



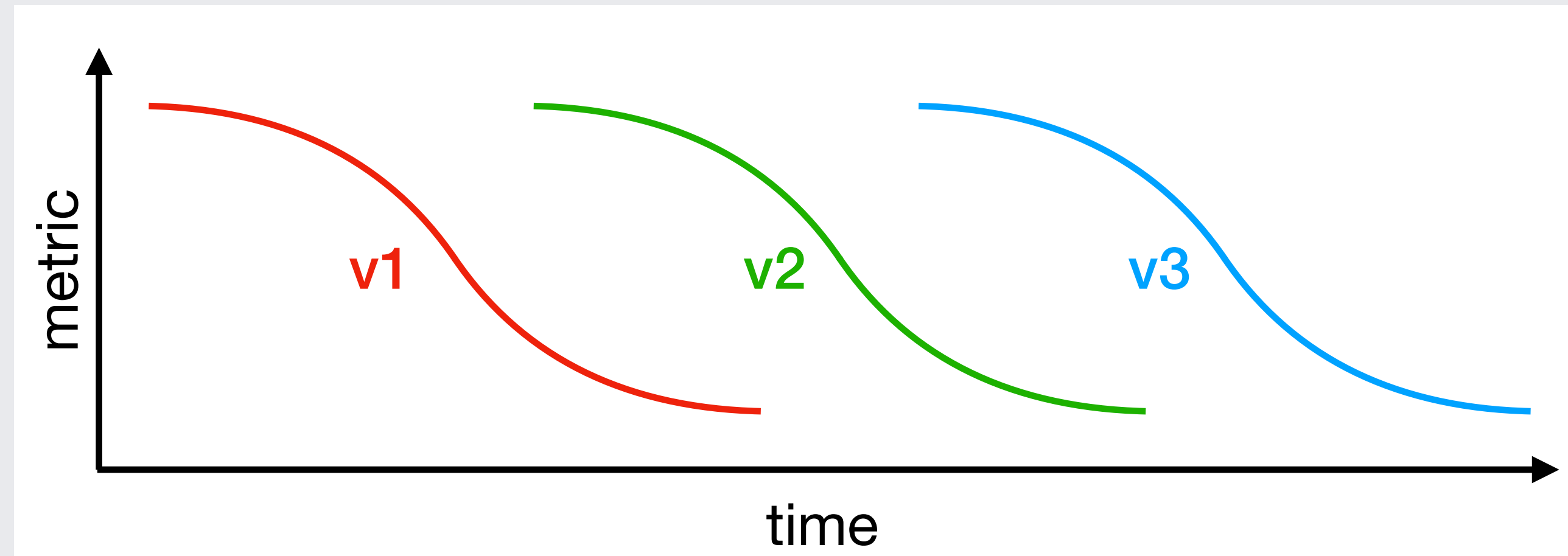
Other coping mechanisms

Iterate





Other coping mechanisms

Iterate



Validate

 Give feedback on this post
We use your feedback to help us learn when something's not right.



Now I don't wish I worked in Ads!

- Use labeling to build measurement.
- Use measurement to prioritize work & determine impact.
- But...everything is still in different units!

Challenge: Build an Abuse Adapter!



Open questions

- How can we obtain a reliable signal from user reporting?
- How can we combine different approaches into a unified system?



Thank you!

facebook

dfreeman@fb.com