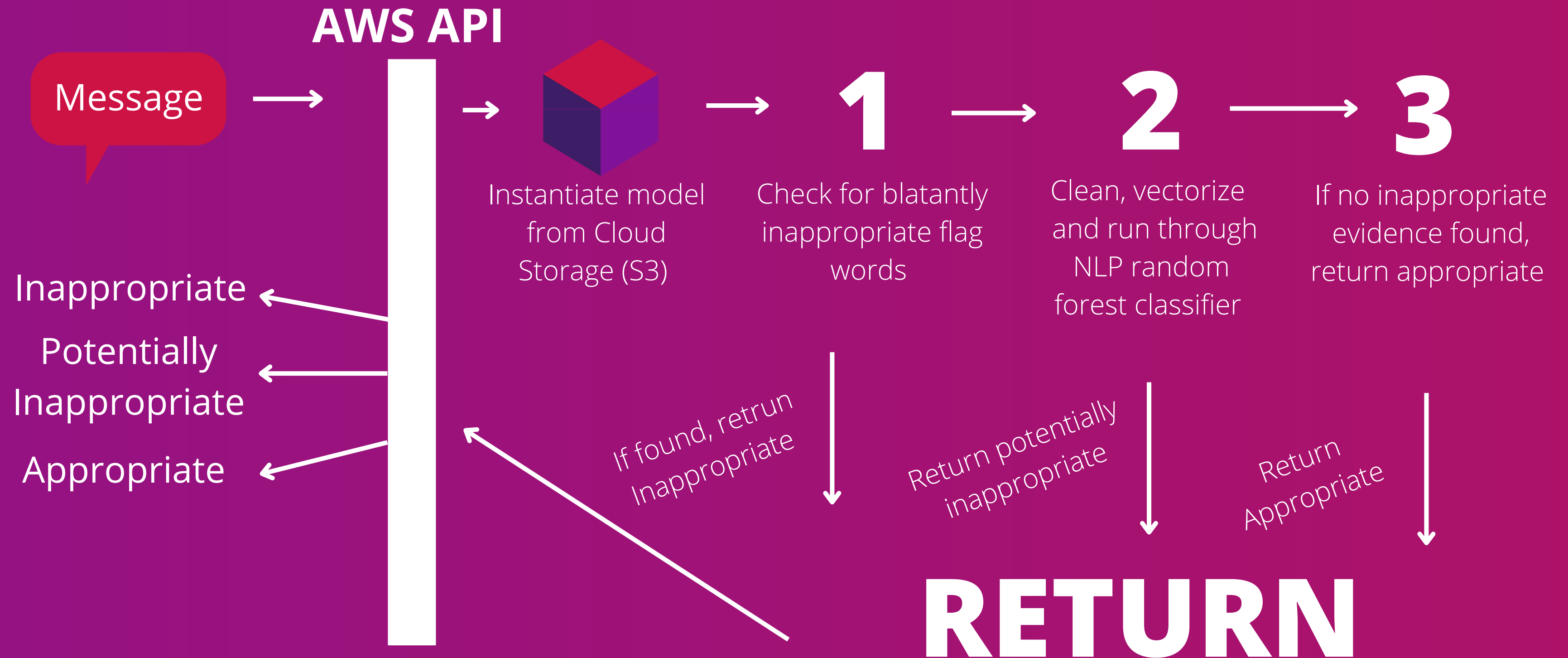# MUTUAL
ABUSE DETECTOR

TREVOR ANDRUS
LEXI DELOREY
LANCE PARRISH

OBJECTIVES

- **Identify** and flag potentially inappropriate or abusive messages sent by users
- **Maximize** Accuracy
- **Minimize** Latency
- **Create** an API to receive messages and return responses
- **Integrate** with existing Mutual servers

MODEL

AWS API

Message →

→

Instantiate model from Cloud Storage (S3)

→ **1**

Check for blatantly inappropriate flag words

→ **2**

Clean, vectorize and run through NLP random forest classifier

→ **3**

If no inappropriate evidence found, return appropriate

Inappropriate ←

Potentially Inappropriate ←

Appropriate ←

If found, retrun Inappropriate

Return potentially inappropriate

Return Appropriate

RETURN

RESULTS

Accuracy

Random Forest in conjunction with flag word dictionary has a prediction accuracy of 82%

Latency

All models, vectorizers and lexicons stored on the cloud (S3) for fast retrieval

After initial cold start, responses take less than 200ms on average

Integration

All code stored on AWS Lambda, and called by AWS API Gateway

API Gateway can integrate with other AWS services, linking with Mutual's front end to show user messages

EXAMPLE

```json
1  {"message": "I could sure go for a ncmo right now"}
```

Body Cookies Headers (5) Test Results

Status: **200 OK** Time: **107 ms** Size: **279 B** **Save Response** ∨

Pretty Raw Preview Visualize JSON ∨

```json
1  {
2      "Code": 402,
3      "Message": "I could sure go for a ncmo right now",
4      "Value": "Inappropriate",
5      "Flag_word": "ncmo"
6  }
```