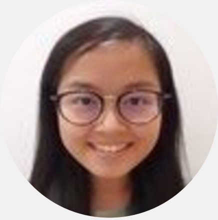
# Twitter Fake News Detection


**I like big BYTES and I cannot LIE**

*Eng Siang, Frances, Linh Chi, Naomi, Sophia, Stephen*

# Breakdown

# Who are they?

- Twitter Management Team
- Twitter Legal Team
- Twitter Operations Team


1208 Comments

TECH

Elon Musk completes $44B Twitter takeover, begins firing execs

By Thomas Barrabi and Theo Wayt

October 27, 2022 | 9:05pm | Updated

# Why should they care?

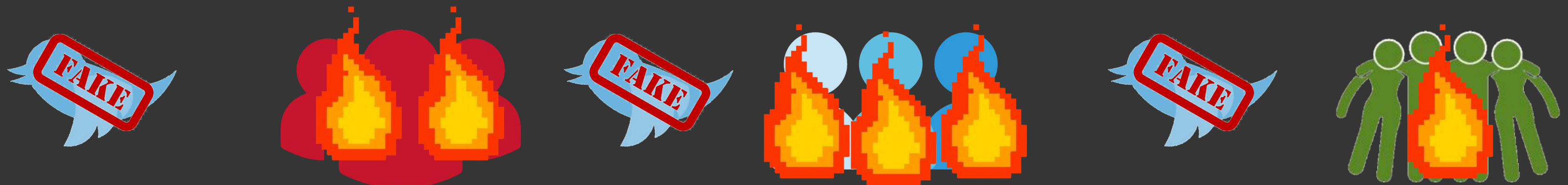
- Fake news spread at a faster rate as 70% are more likely to be retweeted than real news*'
    - Affects profits:
        - Twitter reliant on 89% of its profit
        - Fake news causes decrease in advertiser confidence
    - Increases expenses
        - Fake news can lead to legal lawsuits against twitter

Thus, there is a need to analyze the **trending topics** that could lead to fake news, and **how they have spread** via Twitter users and retweets.


*Source: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

# What is the extent of the influence of fake news on Twitter?

Big question to answer



# DEMO

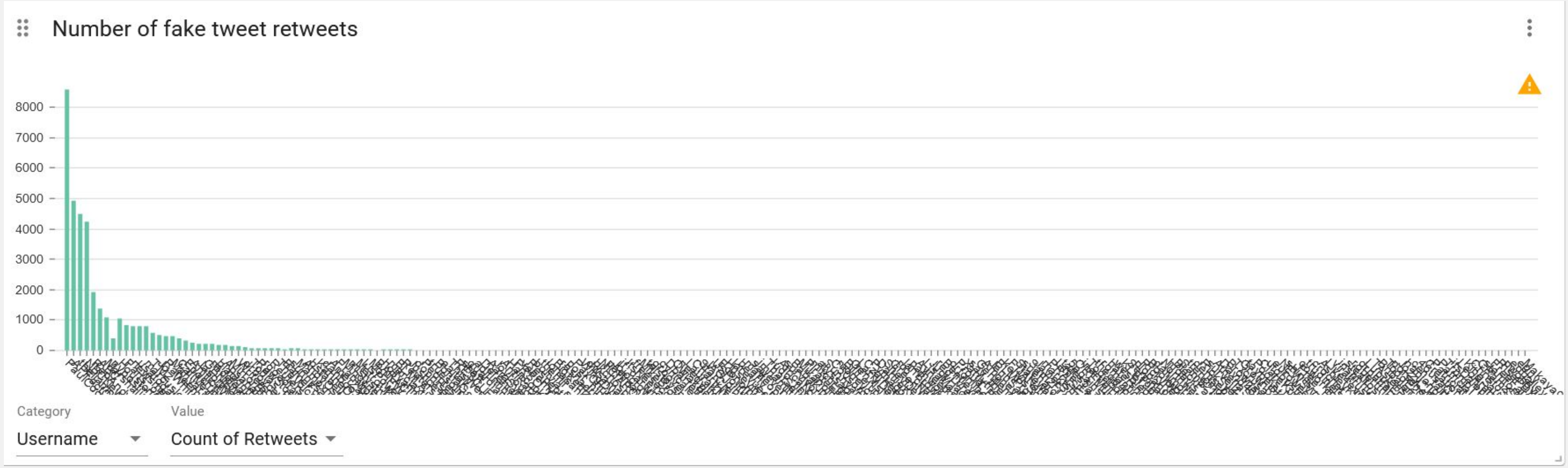
View our dashboard here

# 3% of scraped tweets are fake

# 77,042

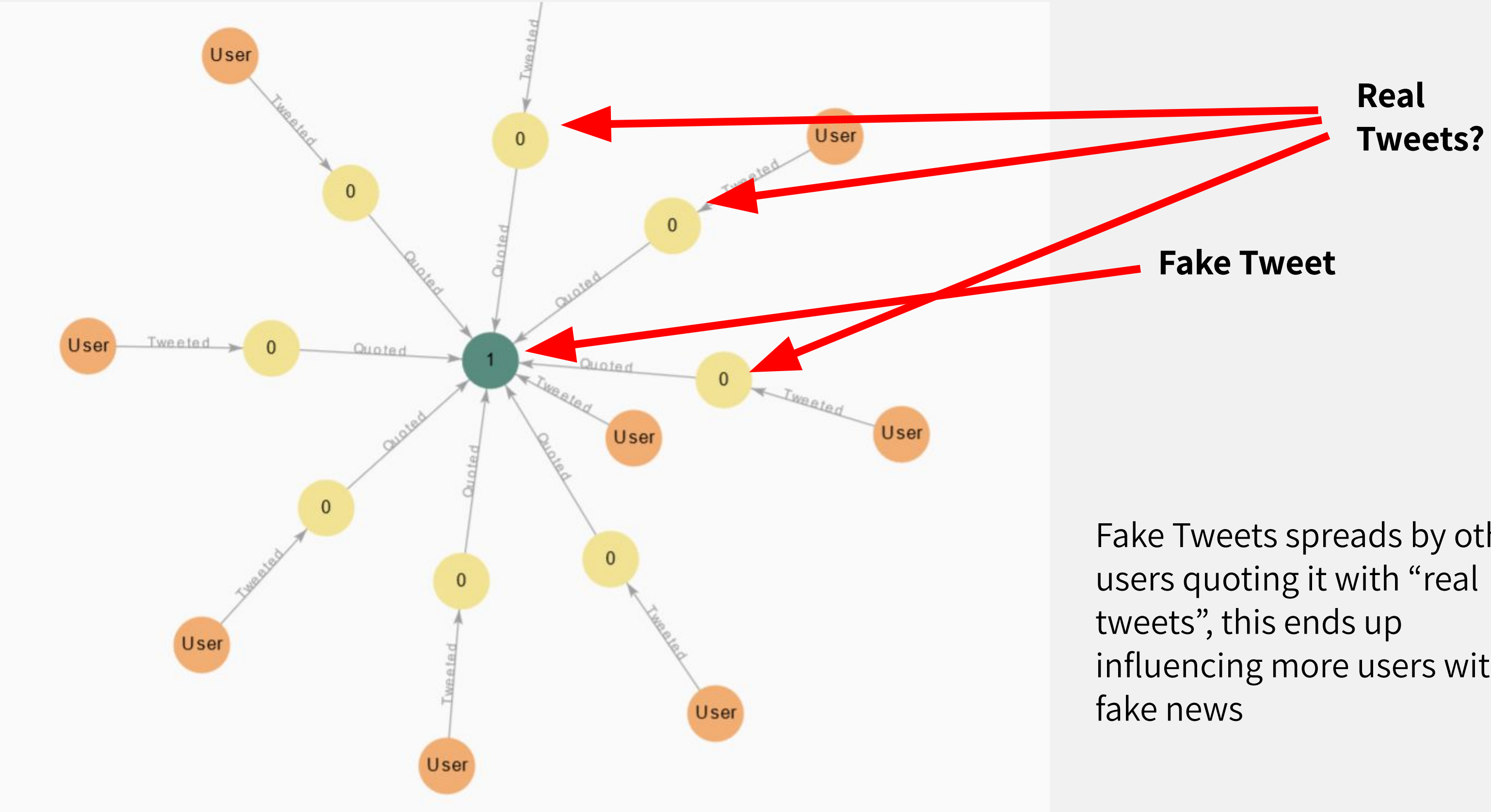
Fake Tweets retweeted

# Power Law Distributions



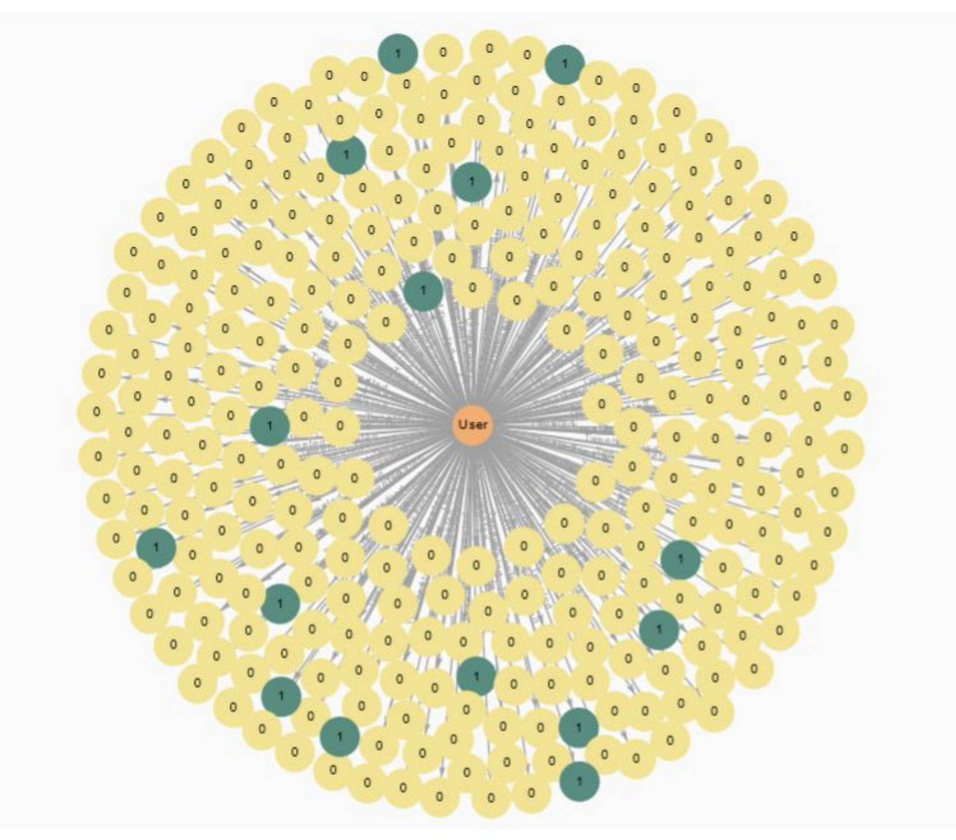

Only a handful of fake tweets are retweeted, however influence of quoted tweets are not accounted for

# Example




Fake Tweets spreads by other users quoting it with “real tweets”, this ends up influencing more users with fake news

# Possible False Negative



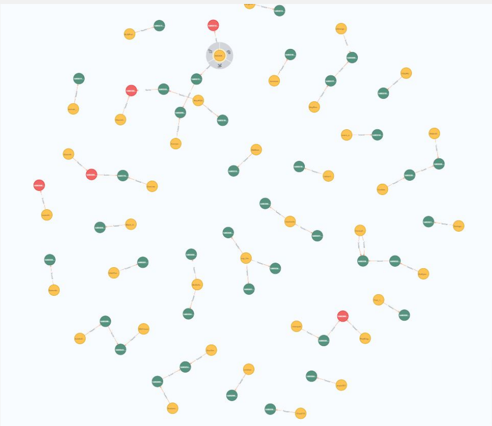

Example of a users with many "real" tweets and a handful of fake tweets.

Possible Explanations: Tweets contains words that are highly correlated with fake tweets

# How did we achieve such insights?

# Initial Scrape

- Scraping of surface level tweets
- Resulted in single nodes on neo4j dashboard
- Does not show any clusters



# Changing Scraping Landscape

- Changing Twitter leadership
- Scrape duration previously was 3 mins and now it runs for approximately 20 minutes as tested on EMR JupyterHub

# Process

## Model Training

- FakeNewsNet data to be processed
- Experimented with different models
- Best model pipeline trained and saved in a .pickle file.

## Scraping

- Using snscrape, we scrape information of tweets containing popular hashtag like #news
- Data includes datetime, id, content, username, url, quoted tweet, reply count, retweet count, like count, quote count
- Upload scraped tweets into S3's input folder

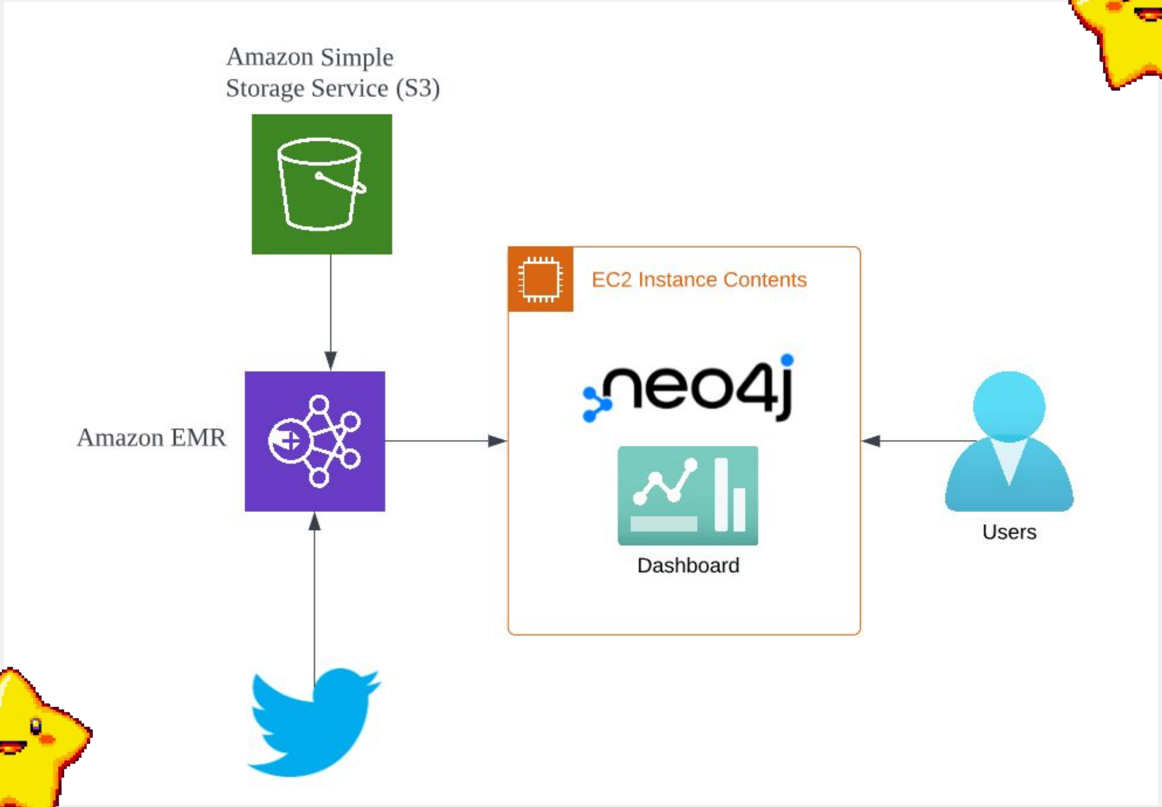
## Tagging

- Scraped data and trained pipeline stored in S3
- Notebook to pass in the scraped data into the pipeline, get label (fake/real)

## Visualisation

- Data loads into neo4j server on EC2 instance
- Experimented with different graphs and configurations to display on Neo4j Dashboard.

# AWS Architecture Diagram



# DEMO

# Tools Used

jupyter
pandas
NumPy
NLTK

neo4j
GitHub Pages

Amazon EC2
Amazon EMR
Amazon Simple Storage Service (S3)

neo4j aura™

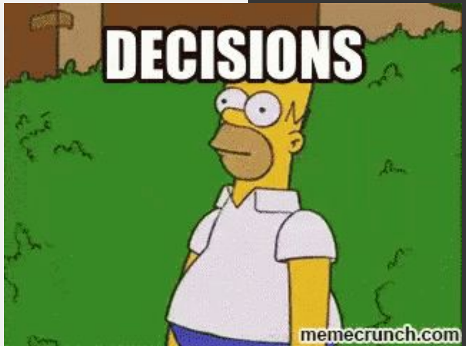
# Key Architectural Decisions #1

## Streaming ✗

- Twitter generates a lot of tweets on a daily basis
- Having the system running 24/7 will be costly

## Batch Processing ✓

- Event driven
- Not a critical function for Twitter
- Reduces overall of maintenance



# Key Architectural Decisions #2 - scraping

## Glue

- Runs on pyspark
- Does not support Snscrape library

## Lambda

- Original implementation works fine
- New scraping logic → Time out limitation

## EMR Notebook

- Runs only on Python 3.65
- PY version does not support Snscrape

## EMR App UI

- Able to import all needed libraries

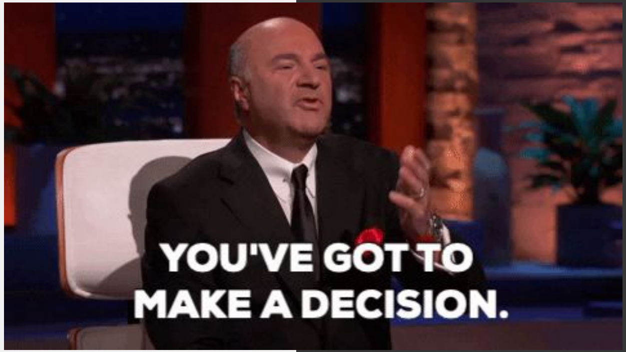
# Key Architectural Decisions #3 - load

## Glue

- Limited functionality
- Auto Scales but very expensive
- Cannot control the scale

## EMR

- Better suited for our use case
- Able to handle a greater load at a cheaper price



# $30.34

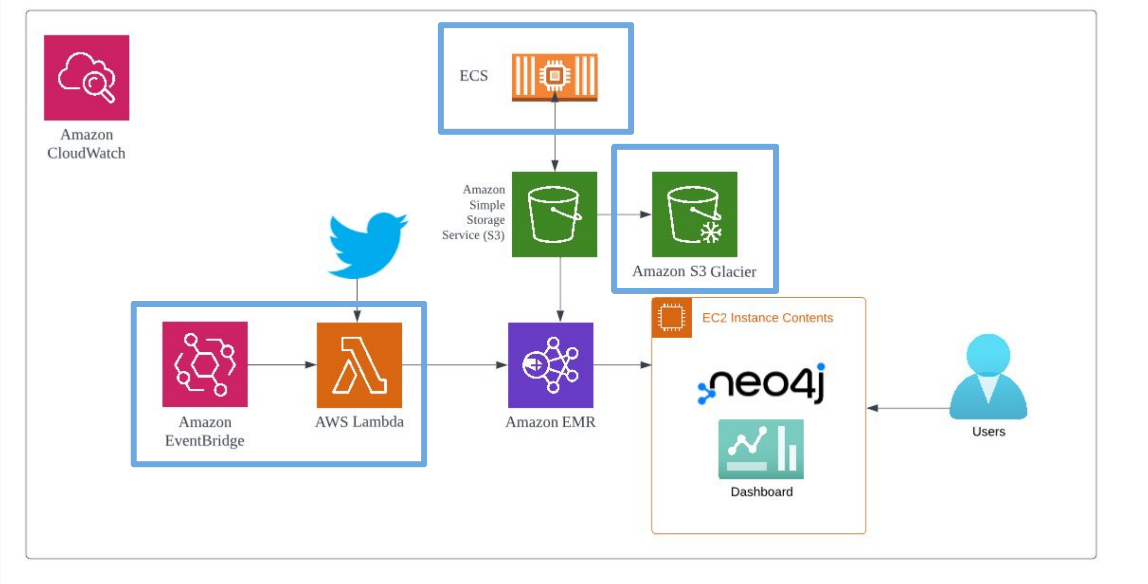
Monthly cost to run our system on AWS

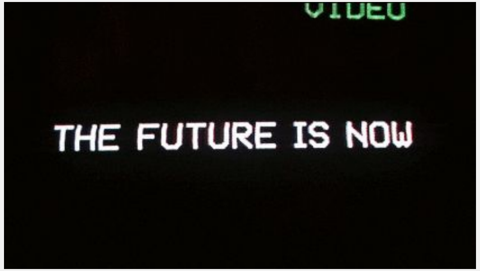
21.90

# Cost Breakdown

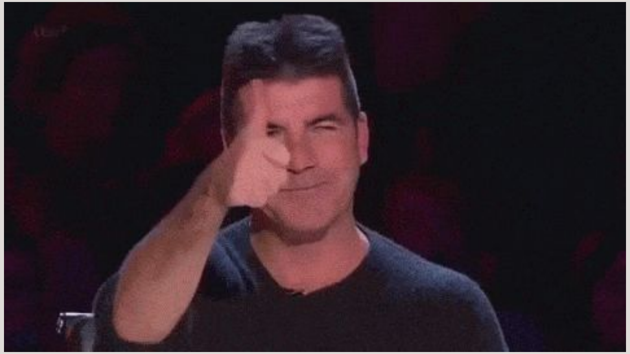

| AWS Service | Description | Cost (monthly) |
| --- | --- | --- |
| **Amazon EMR** | 3 m5.xlarge instances running 5 hours daily<br>*3 Hold instance(s) x 0.048 USD hourly x (5 / 24 hours in a day) x 730 hours in a month* | 21.90 |
| **Amazon S3** | 10gb monthly<br>*8 GB x 0.0230000000 USD = 0.18 USD*<br>*100 PUT requests for S3 Standard Storage x 0.000005 USD per request = 0.0005 USD (S3 Standard PUT requests cost)*<br>*100 GET requests in a month x 0.0000004 USD per request = 0.00 USD (S3 Standard GET requests cost)*<br>*0.184 USD + 0.0005 USD = 0.18 USD (Total S3 Standard Storage, data requests, S3 select cost)* | 0.18 |
| **Amazon EC2** | T2.micro<br>*1 instances x 0.0072 USD x 730 hours in month = 5.26 USD (monthly instance savings cost)*<br>*30 GB x 0.10 USD x 1 instances = 3.00 USD (EBS Storage Cost)* | 8.26 |
| | **Total** | **30.34 USD** |

# AWS Architecture Diagram (future plan)




- Automate the pipeline
- Archive past tweet inputs
- Make use of containers to store code for model retraining



# X Factors

# Neo4j Aura DB

- Fast and scalable graph platform on cloud
- Easily integrable with AWS Architecture
- Provides visualization

# Machine Learning

- Pre-trained model to label fake news
- Identify tweets that contain fake news

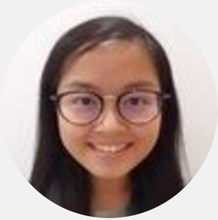
# Network analysis

- Tweets tagged to the users who posted and retweeted
- Better decision-making for Twitter


# Q&A

# Thanks!






# Business Problem Statement

- Fake news spread at a faster rate as 70% are more likely to be retweeted than real news*
- The effect of fake news on Twitter
  - Affects profits:
    - Twitter reliant on 89% of its profit
    - Fake news causes decrease in advertiser confidence
  - Increases expenses
    - Fake news can lead to legal lawsuits against twitter

Thus, there is a need to analyze the **trending topics** that could lead to fake news, and **how they have spread** via Twitter users and retweets.


*Source: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

# What is the extent of the influence of fake news on Twitter?

Big question to answer

# Data Sources

## FakeNewsNet data

- Data on fake and real political and gossip news were collected
- In each news, the information on the news articles and its relevant tweets, retweets, replies and likes are provided
- The text information on tweets can be trained and predicted to be fake or real.

## snscrape Twitter

A Python library that allows us to retrieve twitter posts based on a keyword or tag and user profile details of input username. Trending tags can be tracked using twitter-trends module in snscrape. Data collected are latest tweets, that are linked or related to the keyword or tag specified.

# Analysis

## Fake News Classification

Detect if trending tweets are real or fake using text mining on the text values of tweets

## Fake News Reach

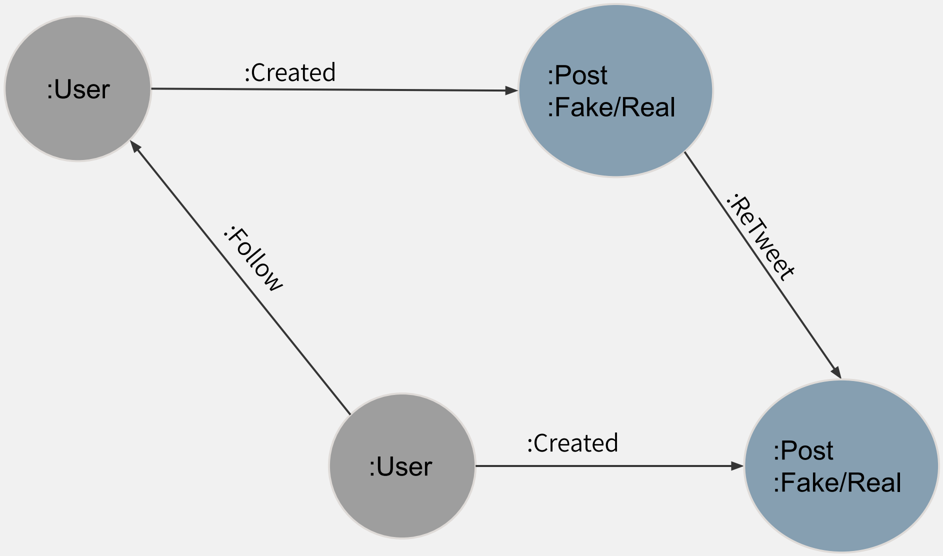
Visualisation of networks of tweets and users using graphs + Social Network Analysis



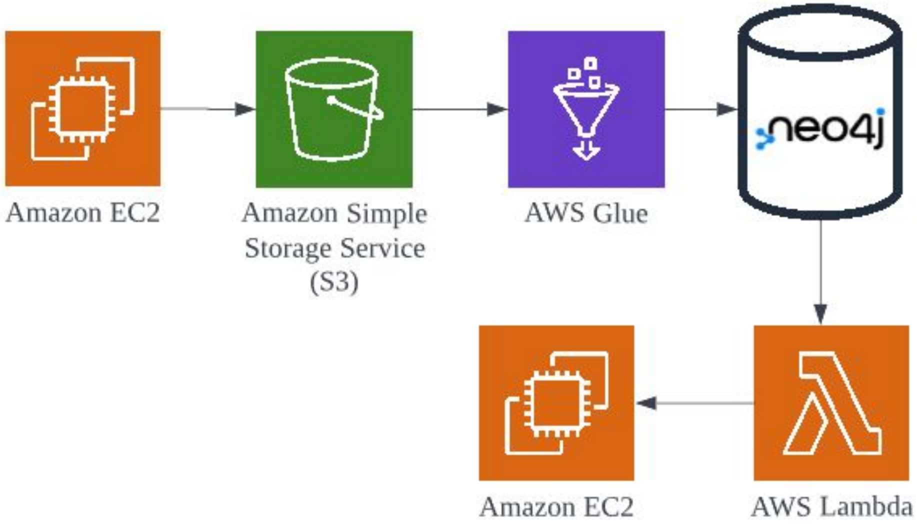
Real time fake news detection

Identify relationships between users and posts on Twitter

# Neo4j


:User
:Created
:Post
:Fake/Real
:Follow
:ReTweet
:User
:Created
:Post
:Fake/Real

# Big Data Architecture (Planned)




**Process**

- Data will be loaded through batch processing that will come from EC2.
- The AWS S3 will contain both the python pickle that holds the model and unprocessed raw data.
- AWS Glue will contduct the ETL process and load the the tagged tweets (contains/doesnt contain fake news and bot/human)
- The data will be queried using the lmabda and passed into ec2 instance for viauzliasation.

With estimated 5000 tweets a day for 30 days in a month the cost to maintain the architecture is $332.13 USD.

# X-factors

- **Neo4j Aura DB**
  - Neo4j Aura is a fast, scalable, always-on, fully automated graph platform offered as a cloud service. It is easily integratable with the AWS architecture and it provides visualization which we can pull to display making the build process faster.
- **Network/ relationship analysis**
  - Based on tagged tweets that contain fake news, nodes will take into account the relationships between the user, the tweet and its retweet. Through the identification of such relationships, Twitter is able to make better decisions to stop fake news from spreading on its platform to its core.
- **Machine Learning**
  - It is a pre-trained model that will label (has fake news or does not have fake news) the preprocessed tweet This serves as the brain for the project that would identify tweets that contain fake news.
- **Streaming in Data Pipeline**
  - Making use of AWS services that will allow the pipeline to withstand a constant incoming input of data. This allows the project to simulate Twitter in a much more realistic manner and allows the project to be a better mock up on what the actual Twitter environment is like