REDDIT.COM

HOW TO BREAK ONTO THE FRONT PAGE OF THE INTERNET

WHAT MAKES A POST ON REDDIT.COM ENGAGING?

- What features can increase engagement?
  - Thumbnails, Image, or Video
  - Questions
- Where do we post?
  - Subreddits
- What's in a title?
  - Natural Language Processing

DATA COLLECTION: WEB SCRAPING METHOD

- Selenium - mimics browser used by human
- Default home page (‘hot’ in r/popular)
- 500 pages every hour
- Periodically over a week
- Always reached ‘end’ - no more pages
- 50,000 unique posts

DATA COLLECTION: WEB SCRAPING

- Title
- Domain
- Age of post
- Thumbnail
- Poster (user)
- Crossposts

- Subreddit (where it as posted)
- How many subscribers to subreddit
- Number of comments
- Score
- Rank

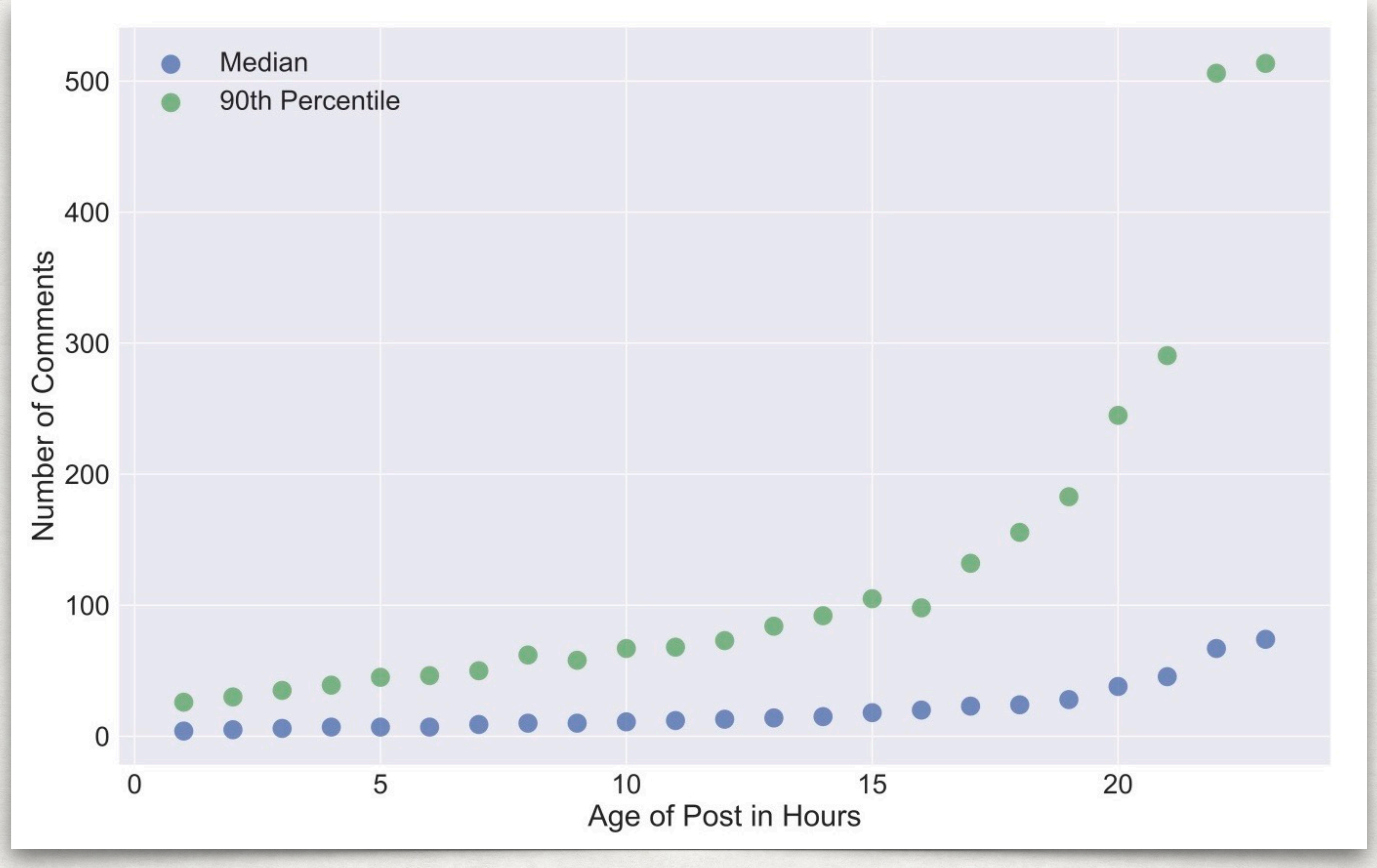
WHAT IS AN ENGAGING POST?

Comments!

Median overall: 12 comments

90th Percentile: 95 comments

BOTH MEDIAN AND 90TH PERCENTILE INCREASE OVER TIME



EVALUATING OUR MODELS

- Accuracy: How many posts are correctly predicted?
- Precision: Of the posts we thought would be above the 90th percentile, how many were actually?
- Sensitivity: Of the posts that were above the 90th percentile, how many did we guess?

RANDOM FOREST - PREDICTABILITY

- ExtraTreesClassifier - Uses many decision trees to collectively come to the best answer, uses random features, and tries random amounts
- Overall accuracy of 81% and Sensitivity of 73%
- Can boost that to 75% with XGB Classifier

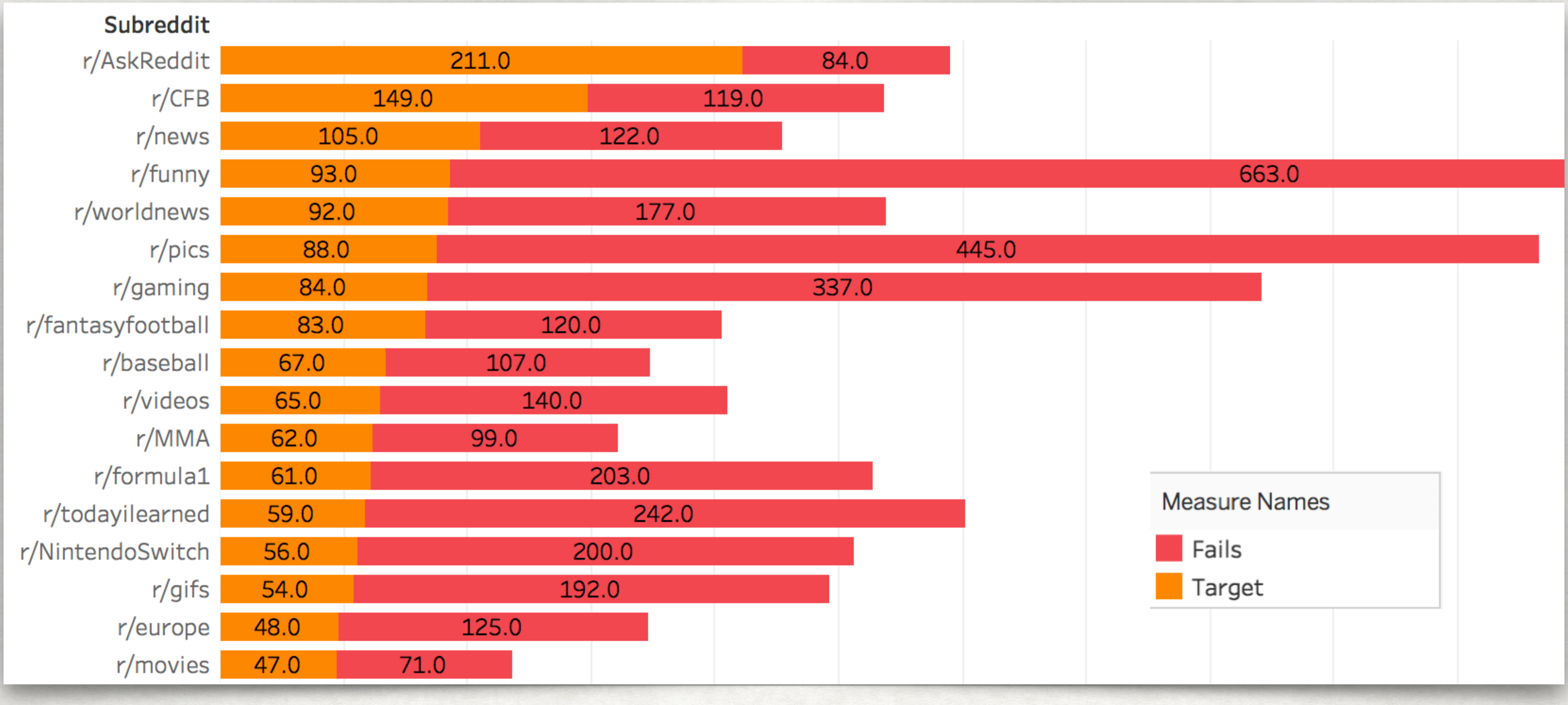
LOGISTIC MODEL - INTERPRETABILITY

- Question is 1.77 times as likely to generate comments in the target zone.
- Thumbnail is 0.55 times as likely
- Picture/GIF is 0.14 times as likely
- Video is 0.63 times as likely
- Game Thread is 26.9 times as likely

WHERE DO WE POST?

- Out of 50,000 posts, 3,700 subreddits were represented
- Subreddits are different forums dedicated to specific topics
  - Sports
  - Gaming
  - Movies etc.

BOTH MEDIAN AND 90TH PERCENTILE INCREASE OVER TIME



NATURAL LANGUAGE PROCESSING

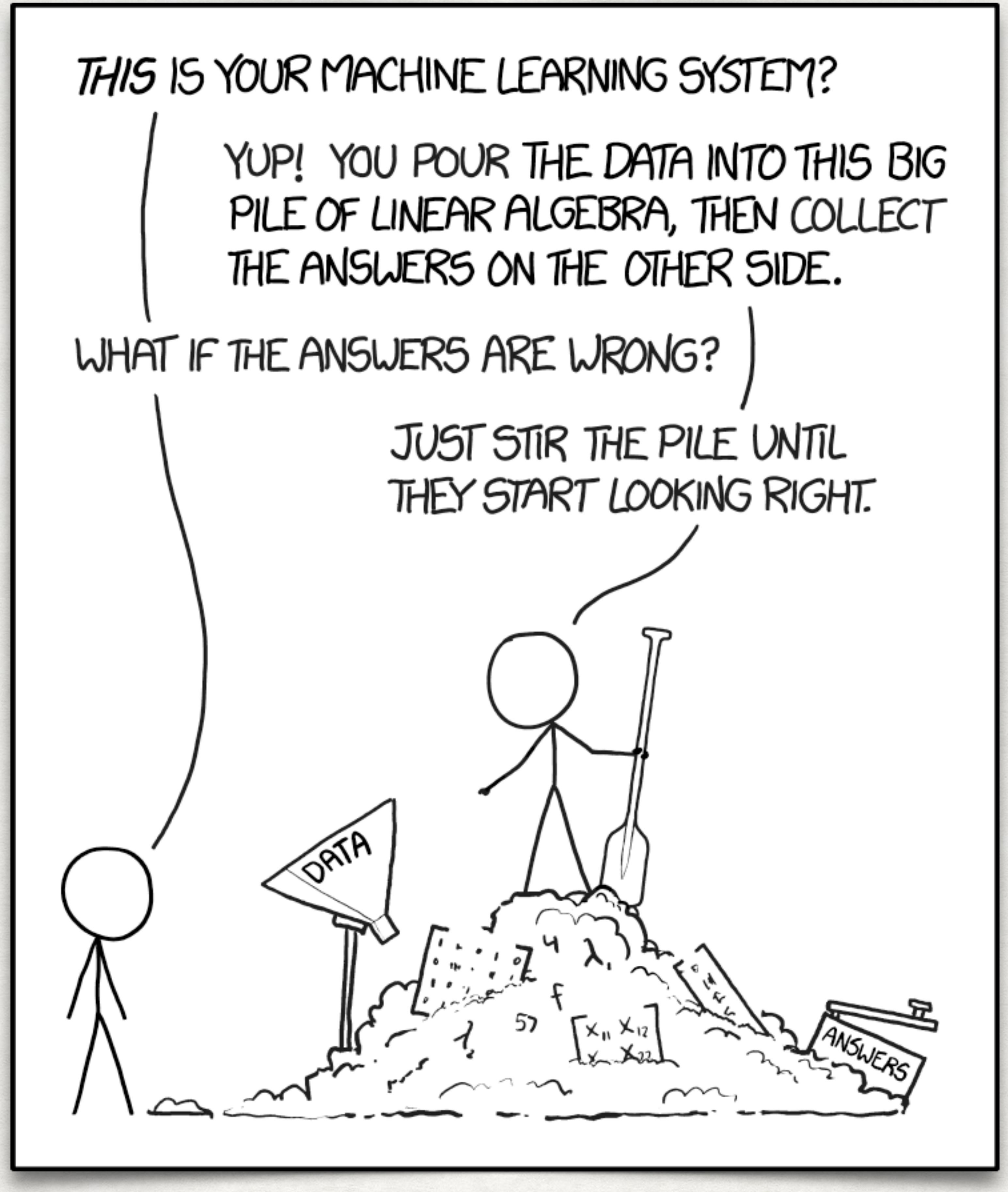
TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY

- What words were above the top 90th percentile, but not below?
    - Game
    - Thread
    - Like
    - New
    - State
    - World
    - Year
    - Time
    - Day
    - 2017
    - Week
    - Halloween
    - People
    - Post
    - Say
    - Make
    - Discussion
    - Think

RECAP

- Logistic Regression: Features such as forming title as a question, images can be a detriment, and current events improve post engagement
- Random Forests: Prediction based on title text, subreddits, and other features
- Natural Language Processing can help us find difference in successful posts

THANK YOU!


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Credit to: https://xkcd.com/1838/