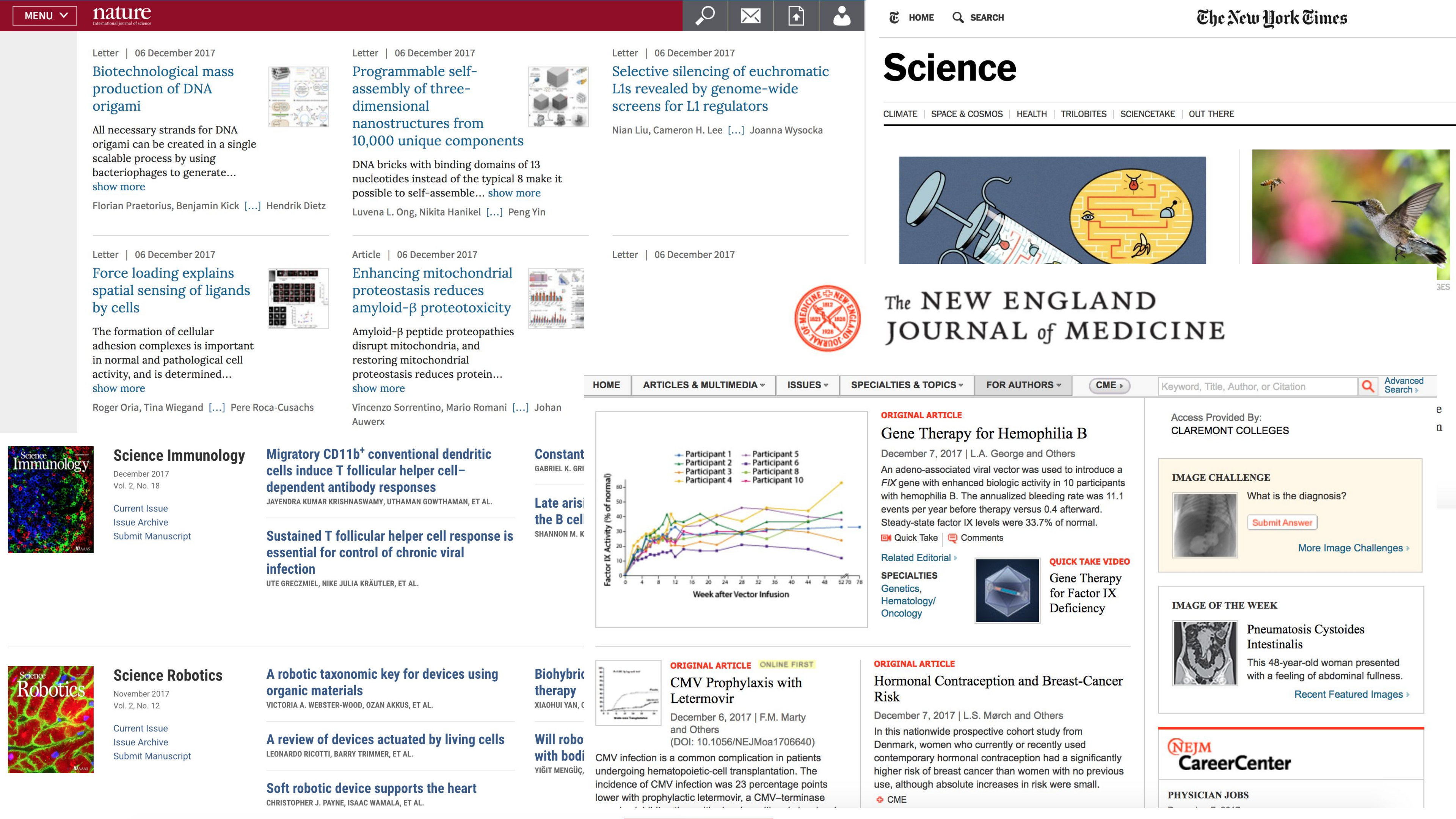
What makes a science article popular?

Sal Fu, Jerry Xuan, Brian Lorenz



MENU nature International journal of science

Letter | 06 December 2017

Biotechnological mass production of DNA origami

All necessary strands for DNA origami can be created in a single scalable process by using bacteriophages to generate… show more

Florian Praetorius, Benjamin Kick […] Hendrik Dietz

Letter | 06 December 2017

Programmable self-assembly of three-dimensional nanostructures from 10,000 unique components

DNA bricks with binding domains of 13 nucleotides instead of the typical 8 make it possible to self-assemble… show more

Luvena L. Ong, Nikita Hanikel […] Peng Yin

Letter | 06 December 2017

Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators

Nian Liu, Cameron H. Lee […] Joanna Wysocka

Letter | 06 December 2017

Force loading explains spatial sensing of ligands by cells

The formation of cellular adhesion complexes is important in normal and pathological cell activity, and is determined… show more

Roger Oria, Tina Wiegand […] Pere Roca-Cusachs

Article | 06 December 2017

Enhancing mitochondrial proteostasis reduces amyloid-β proteotoxicity

Amyloid-β peptide proteopathies disrupt mitochondria, and restoring mitochondrial proteostasis reduces protein… show more

Vincenzo Sorrentino, Mario Romani […] Johan Auwerx

Letter | 06 December 2017

HOME SEARCH The New York Times

Science

CLIMATE | SPACE & COSMOS | HEALTH | TRILOBITES | SCIENCETAKE | OUT THERE

GES

The NEW ENGLAND JOURNAL of MEDICINE

HOME | ARTICLES & MULTIMEDIA | ISSUES | SPECIALTIES & TOPICS | FOR AUTHORS | CME | Keyword, Title, Author, or Citation | Advanced Search

ORIGINAL ARTICLE

Gene Therapy for Hemophilia B

December 7, 2017 | L.A. George and Others

An adeno-associated viral vector was used to introduce a *FIX* gene with enhanced biologic activity in 10 participants with hemophilia B. The annualized bleeding rate was 11.1 events per year before therapy versus 0.4 afterward. Steady-state factor IX levels were 33.7% of normal.

Quick Take | Comments

Related Editorial

SPECIALTIES

Genetics, Hematology/Oncology

QUICK TAKE VIDEO

Gene Therapy for Factor IX Deficiency

Access Provided By:
CLAREMONT COLLEGES

IMAGE CHALLENGE

What is the diagnosis?

Submit Answer

More Image Challenges

IMAGE OF THE WEEK

Pneumatosis Cystoides Intestinalis

This 48-year-old woman presented with a feeling of abdominal fullness.

Recent Featured Images

NEJM CareerCenter

PHYSICIAN JOBS

ORIGINAL ARTICLE ONLINE FIRST

CMV Prophylaxis with Letermovir

December 6, 2017 | F.M. Marty and Others
(DOI: 10.1056/NEJMoa1706640)

CMV infection is a common complication in patients undergoing hematopoietic-cell transplantation. The incidence of CMV infection was 23 percentage points lower with prophylactic letermovir, a CMV–terminase

ORIGINAL ARTICLE

Hormonal Contraception and Breast-Cancer Risk

December 7, 2017 | L.S. Mørch and Others

In this nationwide prospective cohort study from Denmark, women who currently or recently used contemporary hormonal contraception had a significantly higher risk of breast cancer than women with no previous use, although absolute increases in risk were small.

CME

e
n

Science Immunology

December 2017
Vol. 2, No. 18

Current Issue
Issue Archive
Submit Manuscript

Migratory CD11b^{+} conventional dendritic cells induce T follicular helper cell–dependent antibody responses

JAYENDRA KUMAR KRISHNASWAMY, UTHAMAN GOWTHAMAN, ET AL.

Sustained T follicular helper cell response is essential for control of chronic viral infection

UTE GRECZMIEL, NIKE JULIA KRÄUTLER, ET AL.

Constant

GABRIEL K. GRI

Late arisi the B cel

SHANNON M. K

Science Robotics

November 2017
Vol. 2, No. 12

Current Issue
Issue Archive
Submit Manuscript

A robotic taxonomic key for devices using organic materials

VICTORIA A. WEBSTER-WOOD, OZAN AKKUS, ET AL.

A review of devices actuated by living cells

LEONARDO RICOTTI, BARRY TRIMMER, ET AL.

Soft robotic device supports the heart

CHRISTOPHER J. PAYNE, ISAAC WAMALA, ET AL.

Biohybric therapy

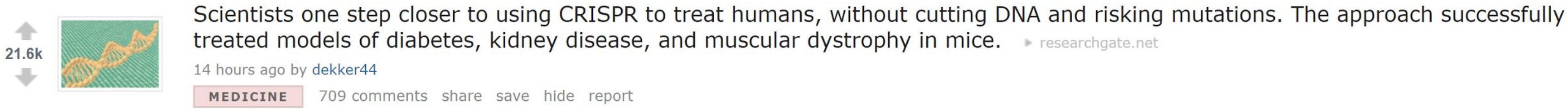
XIAOHUI YAN, C

Will robo with bodi

YIĞIT MENGÜÇ,

Consider Reddit

- Over 500 million users - 8th most visited website worldwide
- Subreddits: specific forums for topic areas
- Upvote system:
    - Users can post content and “upvote” ones they like
    - High upvotes makes an articles more visible (by appearing on the front page)
- Our work is restricted to /r/science
    - Only links to science articles are allowed



21.6k

Scientists one step closer to using CRISPR to treat humans, without cutting DNA and risking mutations. The approach successfully treated models of diabetes, kidney disease, and muscular dystrophy in mice. ▸ researchgate.net

14 hours ago by dekker44

MEDICINE 709 comments share save hide report

Project Overview

Question: What makes a science post popular on Reddit?

Possible explanatory variables: Scientific field of post, time post was created, content of the title, author of the post, etc...

Response variable: Number of upvotes

Method: build random forests!

- Also returns variable importance

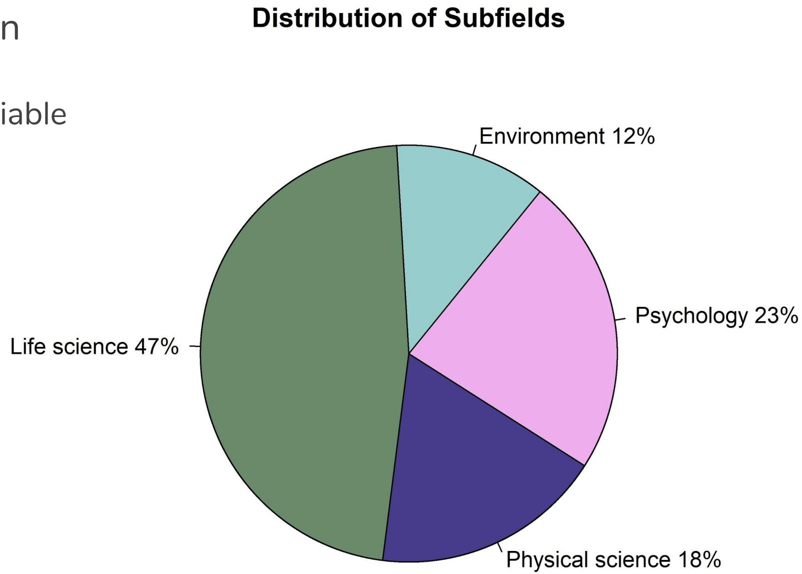
Data Acquisition: PRAW

praw-dev / praw

- "Python Reddit API Wrapper"
- Created Reddit account, authenticated, and connected to Reddit API with PRAW
- Data from November 2015 - November 2017
- 21,921 total posts to /r/science in that time

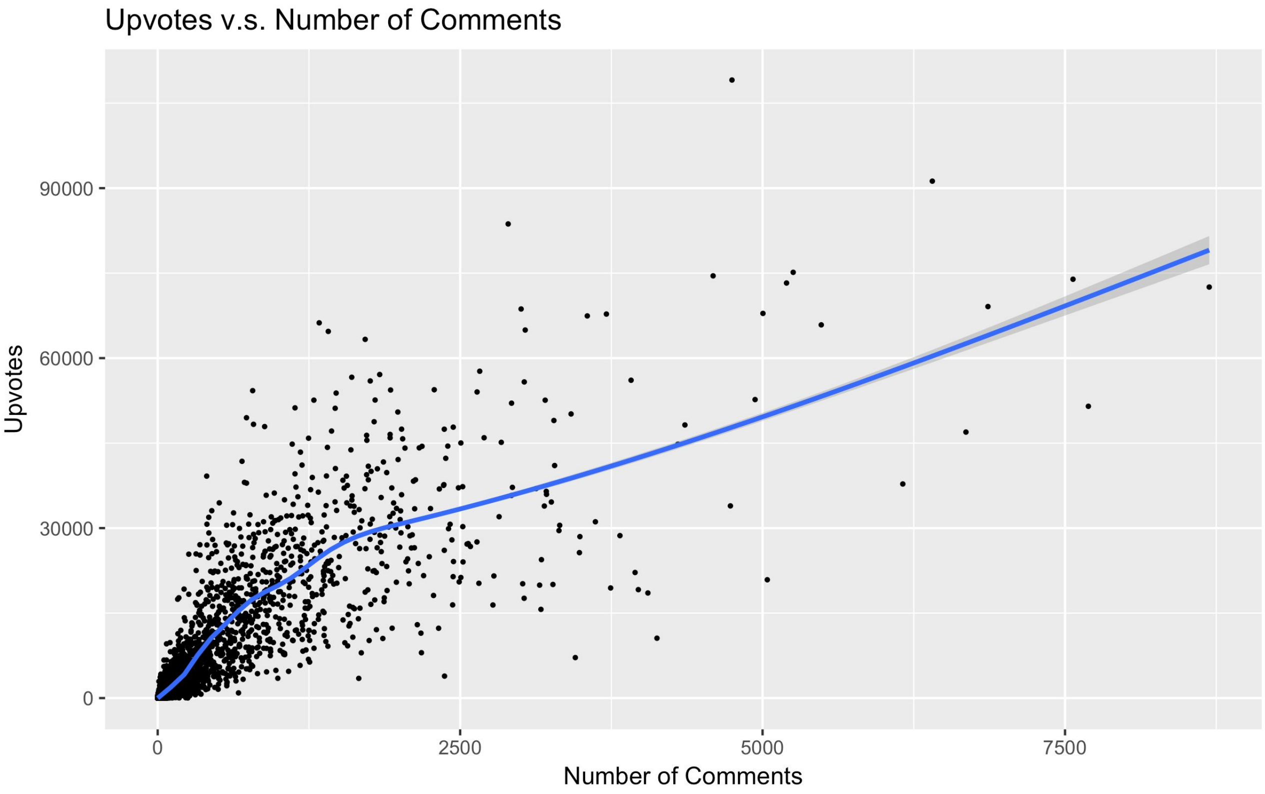
Data Characteristics

- Basic columns: post title, upvotes, time posted, subfield...
- Wrangled explanatory variables to extract information
  - Consolidated subfields, turned time into categorical variable
- Added journal *h*-index (impact factor)
  - Low: Journal *h* index: below 500 (SJR)[1]

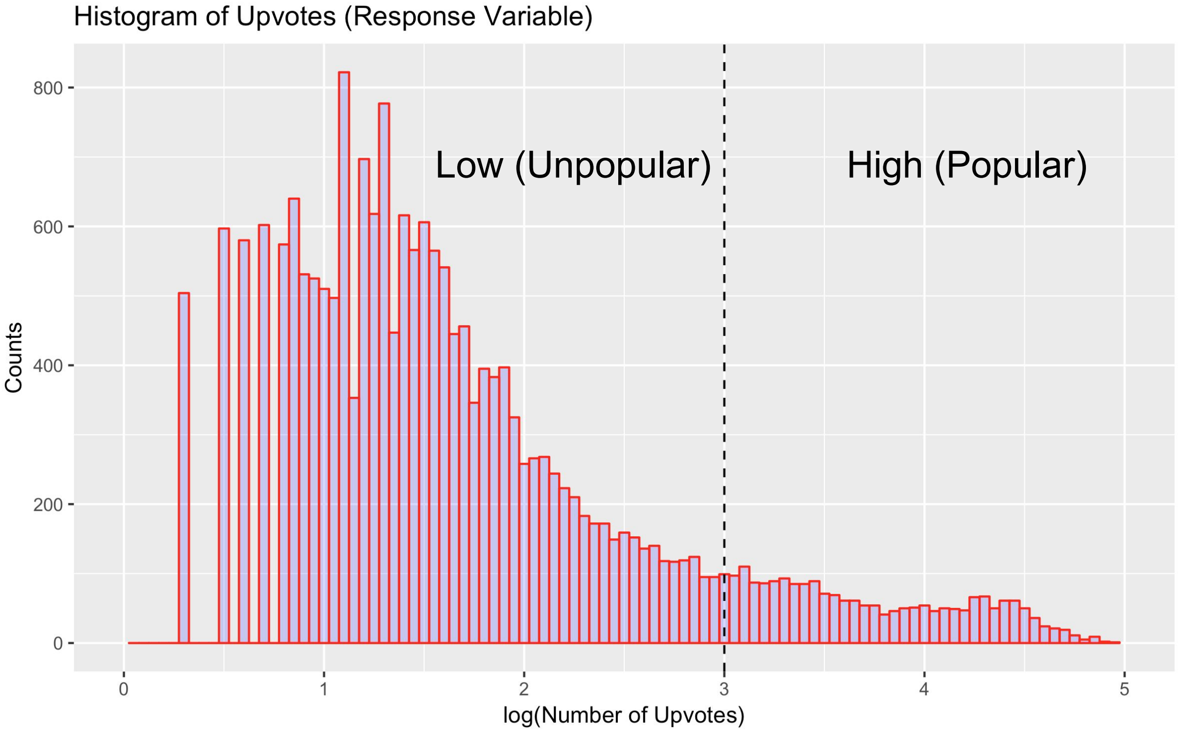


[1] *Scimago Journal & Country Rank*

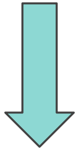
- Removed variables that are closely coupled with response
    - Author’s link karma, post’s number of comments



Response: Upvotes



Continuous Response



Binary Response

Title Analysis

- Computed mean and max word length
- Title Sentiment - SentiWordNet
  - Each word is given a score from 0 to 1 for “Positive” and “Negative”
  - Removed common “stopwords” (e.g. “which,” “the,” “and”)
  - Scaled to length of title

Highly Positive:

Highly Negative:

Title Analysis

- Computed mean and max word length
- Title Sentiment - SentiWordNet
  - Each word is given a score from 0 to 1 for “Positive” and “Negative”
  - Removed common “stopwords” (e.g. “which,” “the,” “and”)
  - Scaled to length of title

<u>Highly Positive:</u>

Reasonable?
“Babies know when they don’t know something”

<u>Highly Negative:</u>

Reasonable?
“Induced Earthquakes”

Title Analysis

- Computed mean and max word length
- Title Sentiment - SentiWordNet
  - Each word is given a score from 0 to 1 for “Positive” and “Negative”
  - Removed common “stopwords” (e.g. “which,” “the,” “and”)
  - Scaled to length of title

<u>Highly Positive:</u>

Reasonable?
“Babies know when they don’t know something”

Umm….
“Happiness doesn’t bring good health, study finds”

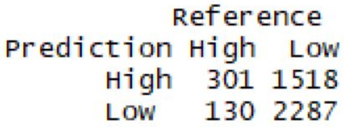
<u>Highly Negative:</u>

Reasonable?
“Induced Earthquakes”

Umm…
“Cats protect newborns against asthma”

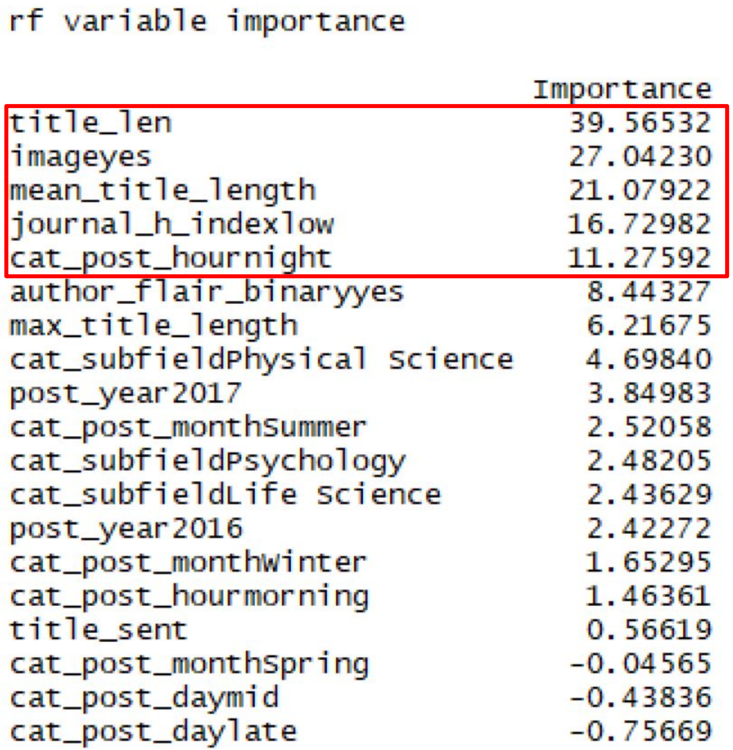
Building the Random Forest...

- Set aside test data (20%)
- Create training data with equal numbers of high/low response
  - Biased training set - vast majority are low upvotes
- Random forest with 1000 trees, tuning mtry
  - mtry = 3
  - OOB error rate: 35%
  - Test error rate: 39%
  - 95% CI for accuracy: 59.6%-62.6%



```
          Reference
Prediction High  Low
      High  301 1518
      Low   130 2287
```

Variable Importance Analysis



```
rf variable importance

                               Importance
title_len                        39.56532
imageyes                         27.04230
mean_title_length                21.07922
journal_h_indexlow               16.72982
cat_post_hournight               11.27592
author_flair_binaryyes            8.44327
max_title_length                  6.21675
cat_subfieldPhysical Science      4.69840
post_year2017                     3.84983
cat_post_monthSummer              2.52058
cat_subfieldPsychology            2.48205
cat_subfieldLife Science          2.43629
post_year2016                     2.42272
cat_post_monthWinter              1.65295
cat_post_hourmorning              1.46361
title_sent                        0.56619
cat_post_monthSpring             -0.04565
cat_post_daymid                  -0.43836
cat_post_daylate                 -0.75669
```

Most important variables!

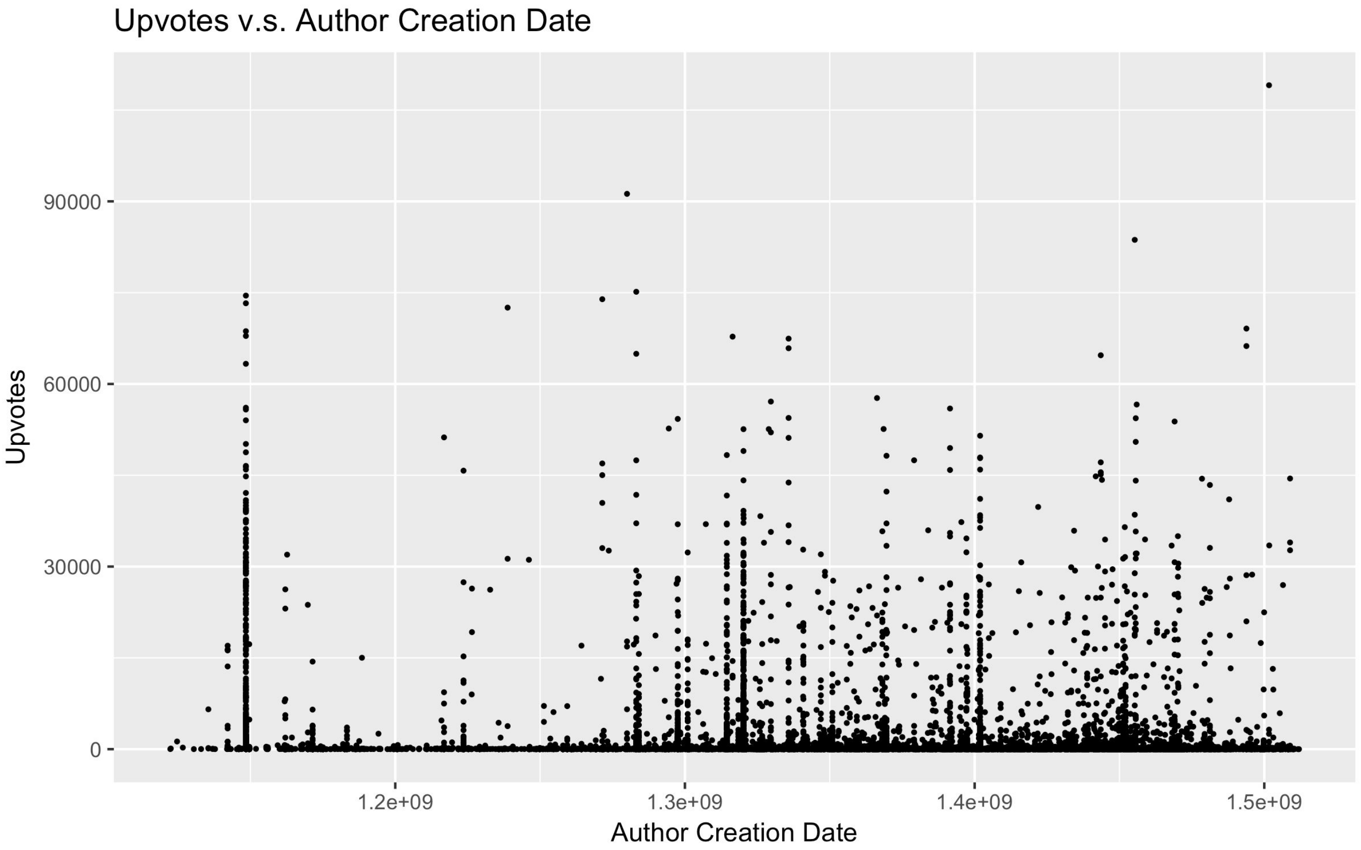
- Mean_title_length = mean length of words in title
- Nighttime defined as: 6PM-3AM PST

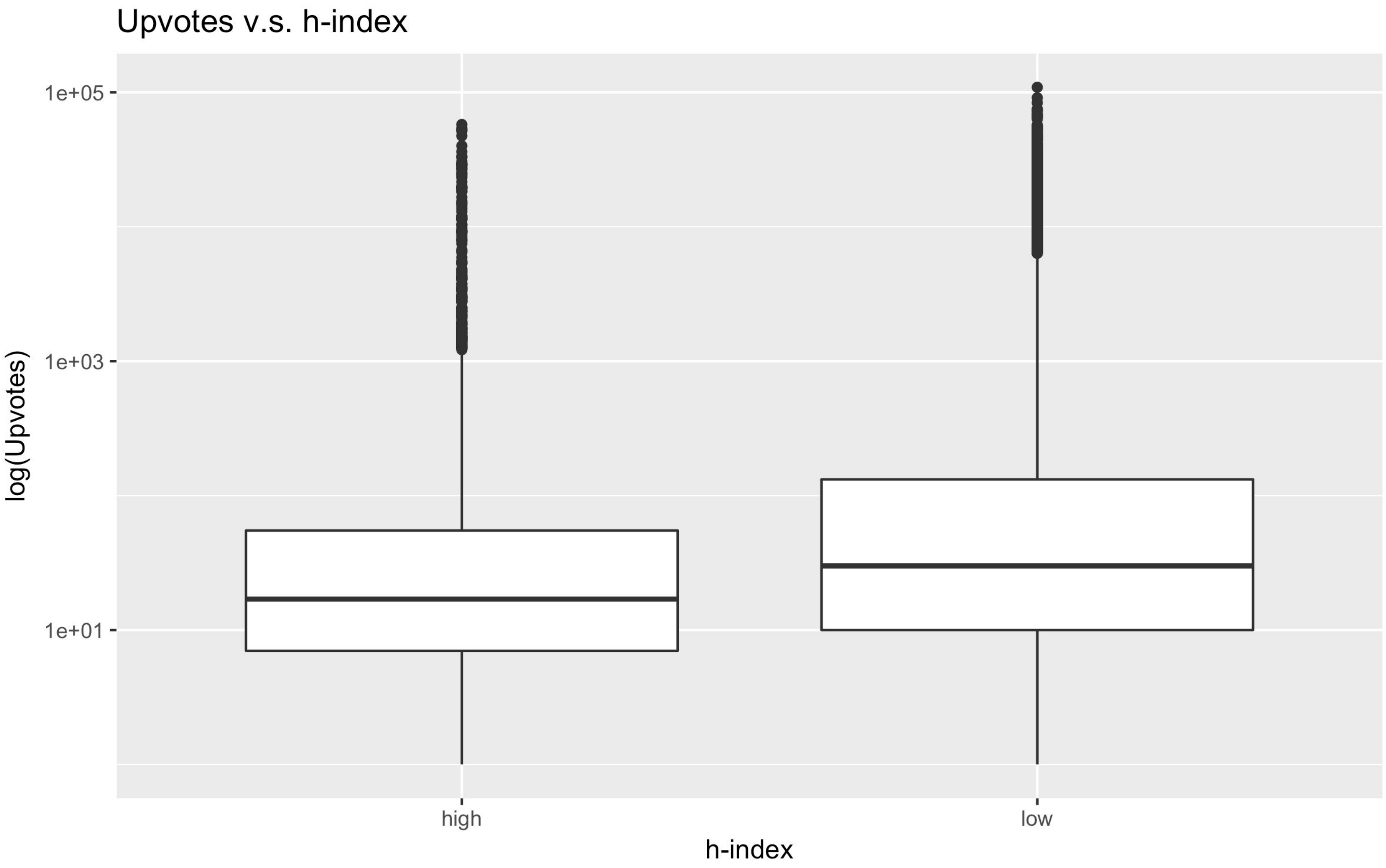
Conclusions

- Data represents those on Reddit who are interested in science
  - Demographic: primarily ages 18-25, primarily male[1]
  - NOT representative sample of overall population
- “Fast-food” Science Media Consumption
  - Popularity likely related to long post titles and low journal *h* index
- Effective (?) science communication: keep things colloquial!

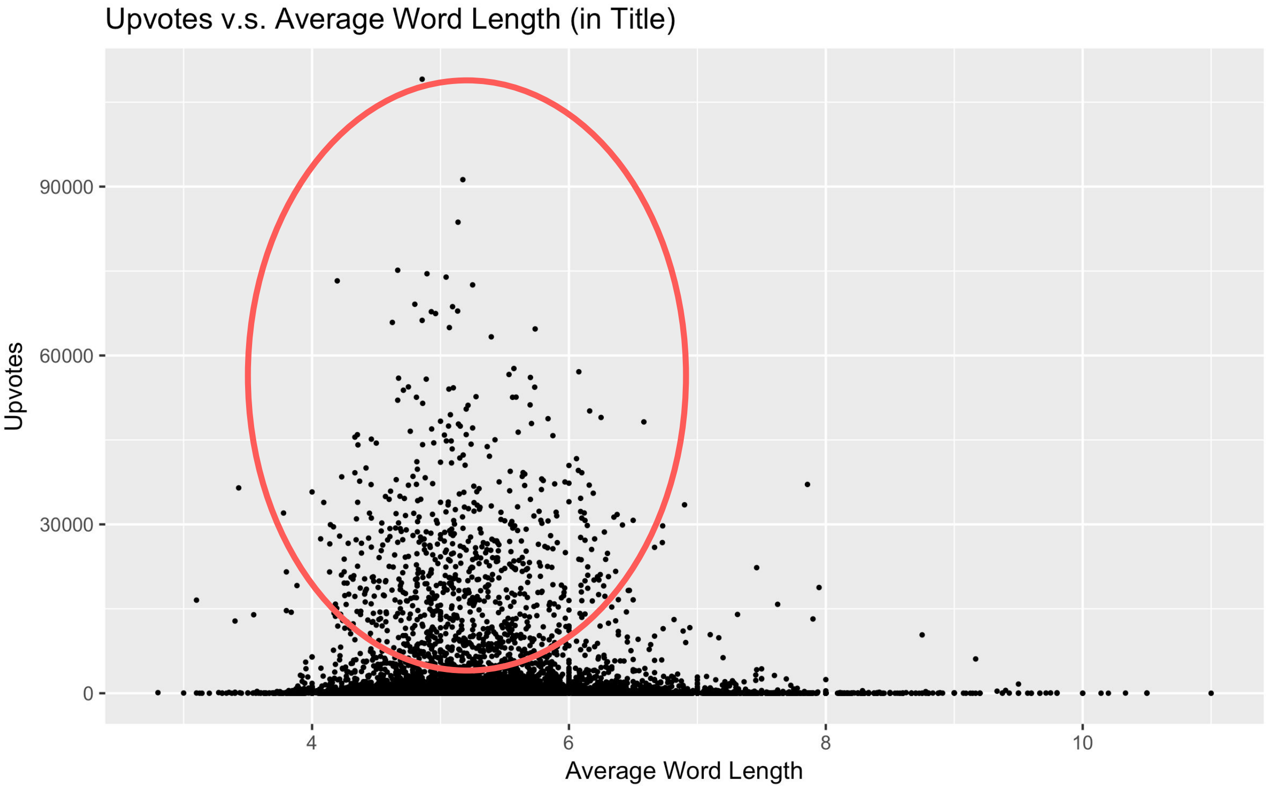
[1] *https://www.reddit.com/r/dataisbeautiful/comments/5700sj/octhe_results_of_the_reddit_demographics_survey/*

Extras


Upvotes v.s. Author Creation Date
Upvotes
90000
60000
30000
0
1.2e+09
1.3e+09
1.4e+09
1.5e+09
Author Creation Date


Upvotes v.s. h-index
1e+05
1e+03
1e+01
log(Upvotes)
high
low
h-index

The Unexpected Effect of Word Length


Upvotes v.s. Average Word Length (in Title)
Upvotes
Average Word Length