Cross Validated

166,874 REPUTATION

🟨 22 ▢ 338 🟧 637

What Makes A Hit Stack Exchange Question?

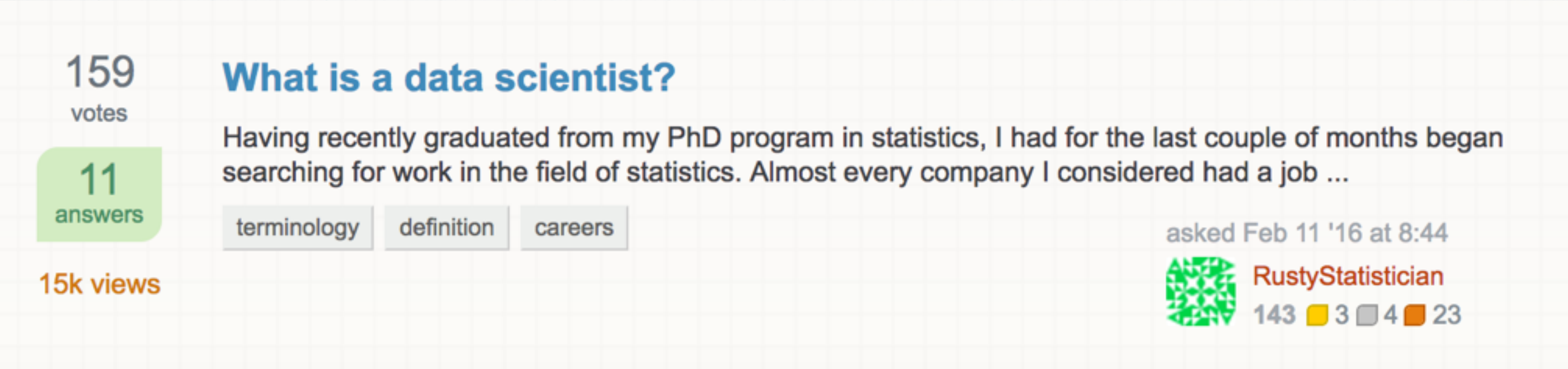
Predicting statistics stack exchange question views using metadata and text features

Introduction

- **Data:** 100,000+ questions scraped from SE; when, reputation, topics, text

- **Goal:** create regression model to predict total views

- **Application:** advertising, content selection, editing



159
votes

11
answers

15k views

What is a data scientist?

Having recently graduated from my PhD program in statistics, I had for the last couple of months began searching for work in the field of statistics. Almost every company I considered had a job ...

terminology definition careers

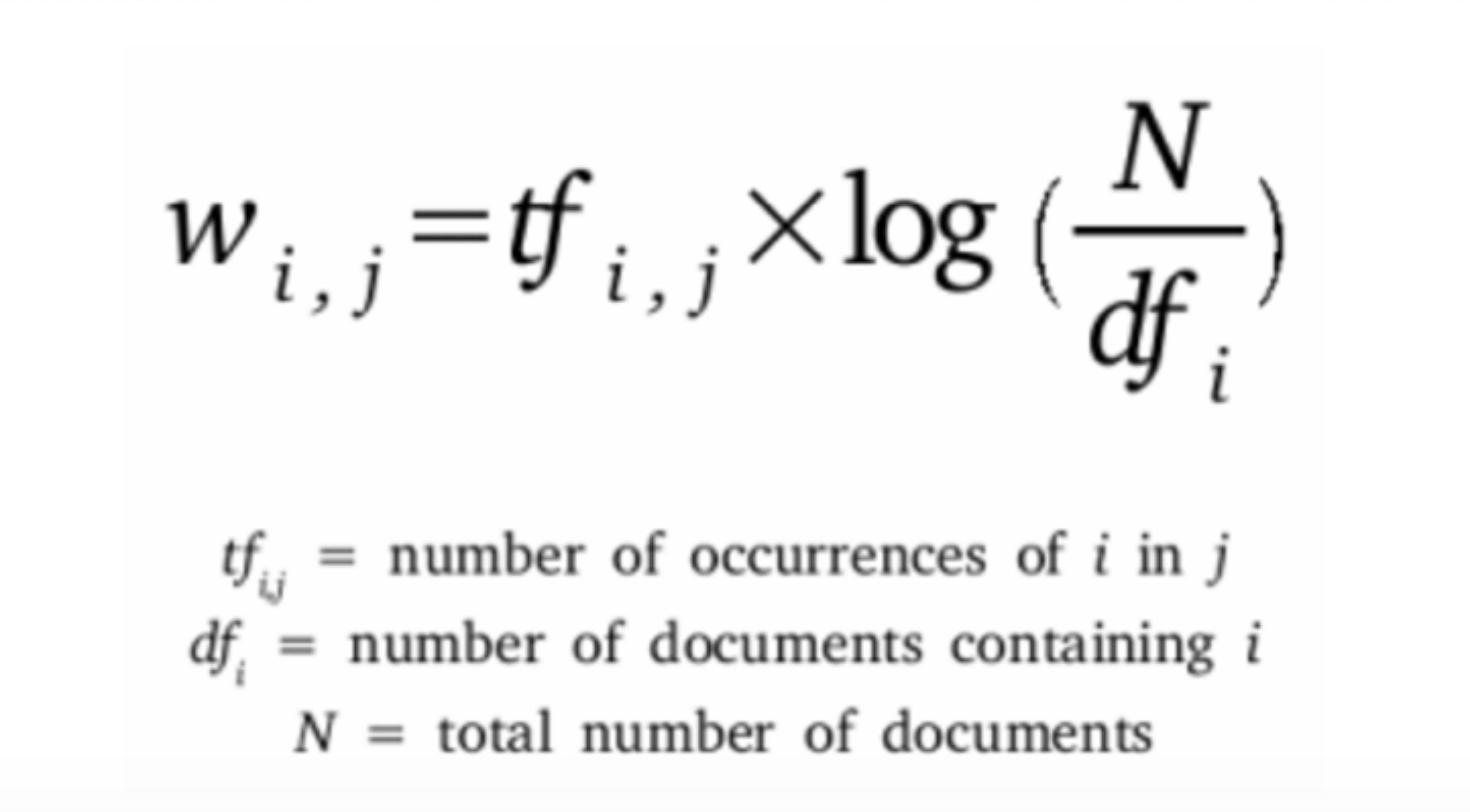
asked Feb 11 '16 at 8:44

RustyStatistician

143 🟨 3 ⬜ 4 🟧 23

Text features: tf-idf Primer

- Term frequency inverse document frequency
- Word importance in given question vs. entire collection



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

Source: DL4J

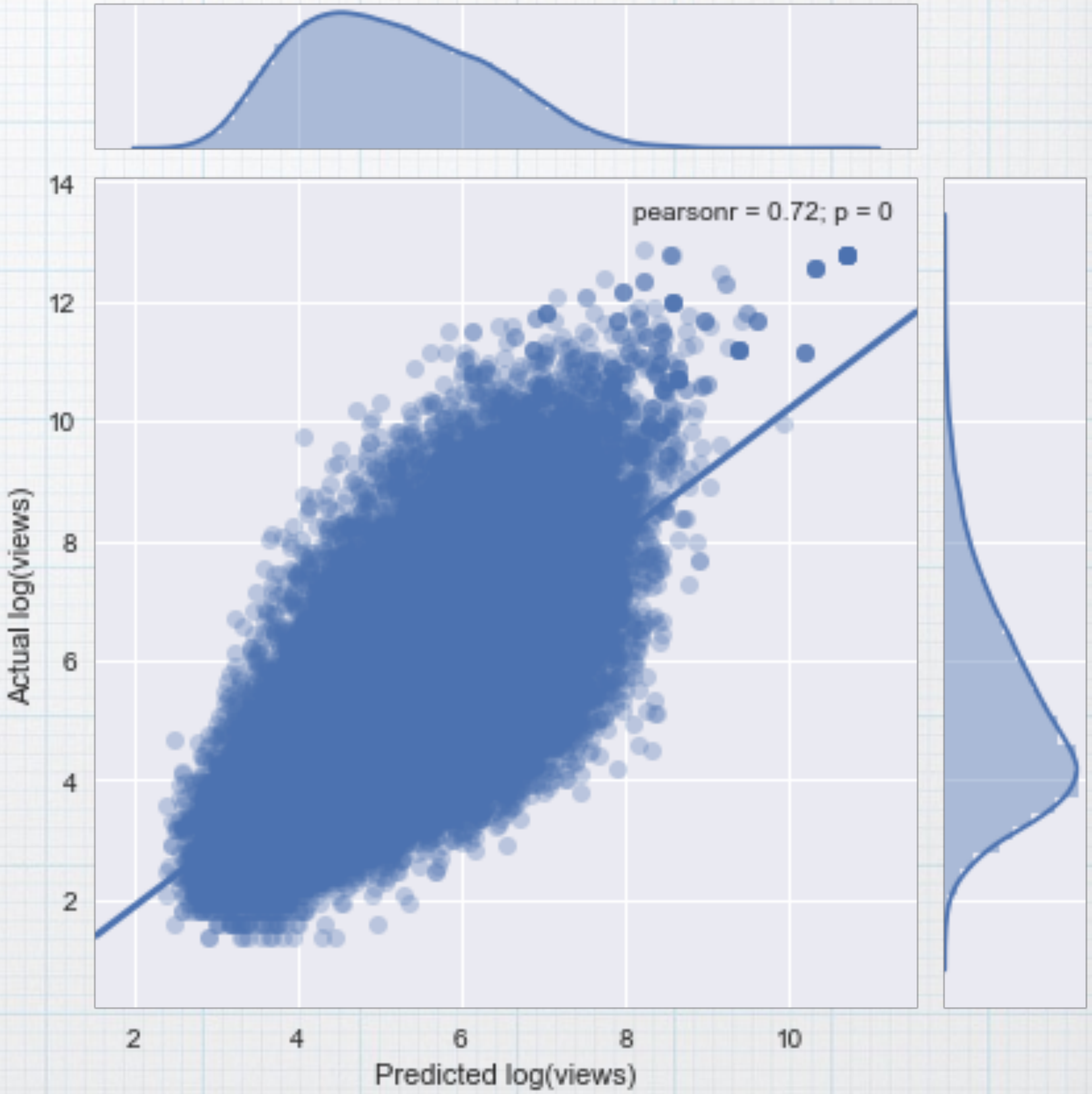
Model & Key Findings

- **Modest predictive accuracy:** 48% of variance explained (test R^2)

  - Ridge regularization

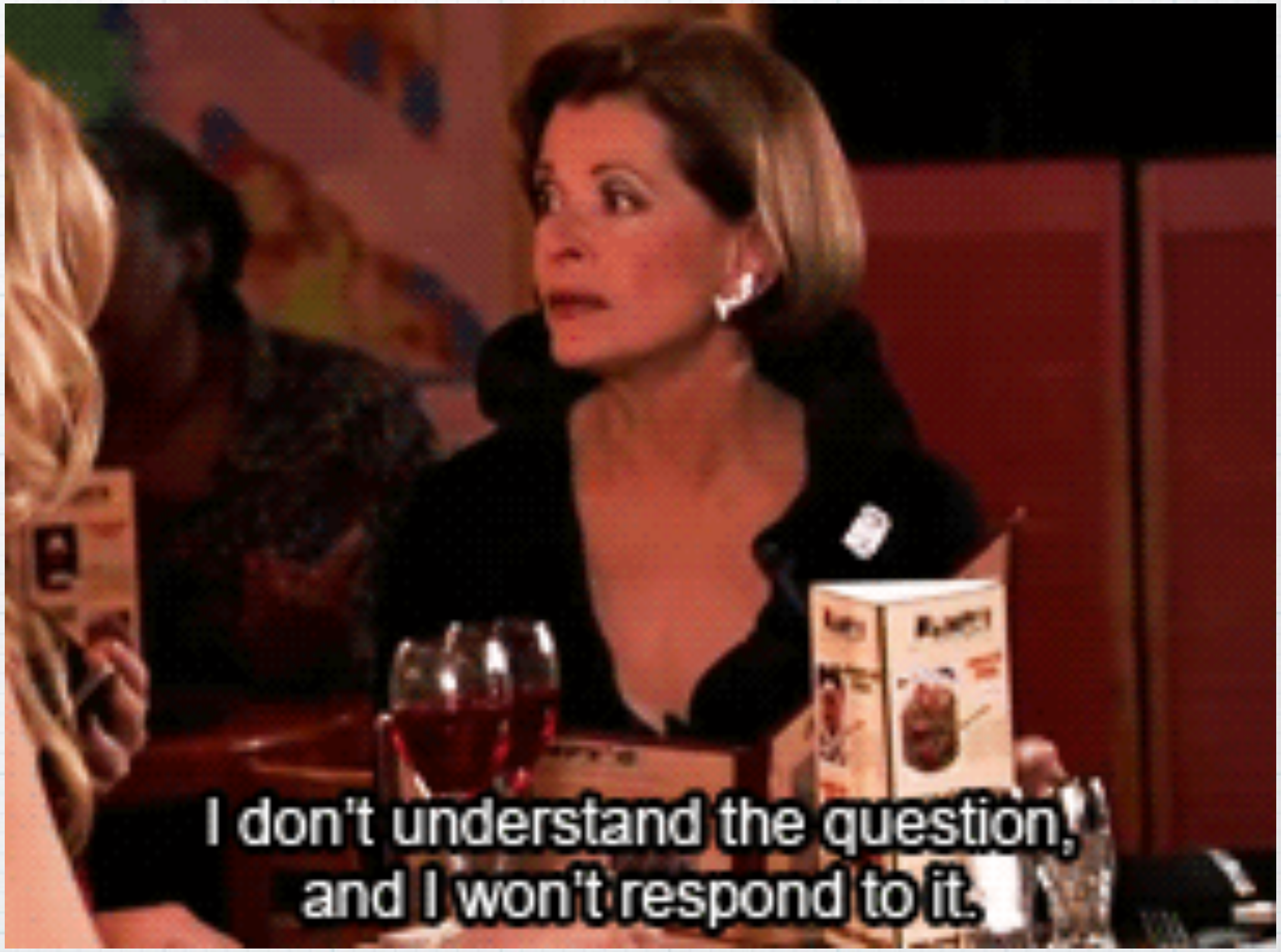
- **Word features:** understanding+, wanting-

- **Topic features:** careers, paradox, intuition+



Positive Word Features: Understanding

- **Strong positive** regression coefficient:
  - Edit, understand, explain, read, interpret, confused, wikipedia
  - Also, ARIMA! (autoregressive integrated moving average)



I don't understand the question, and I won't respond to it.

Negative Word Features: Wanting/Problem Solving

- Strong negative regression coefficient:
  - Want, certain, like, thanks, problem, idea, suppose.
  - Also, database!

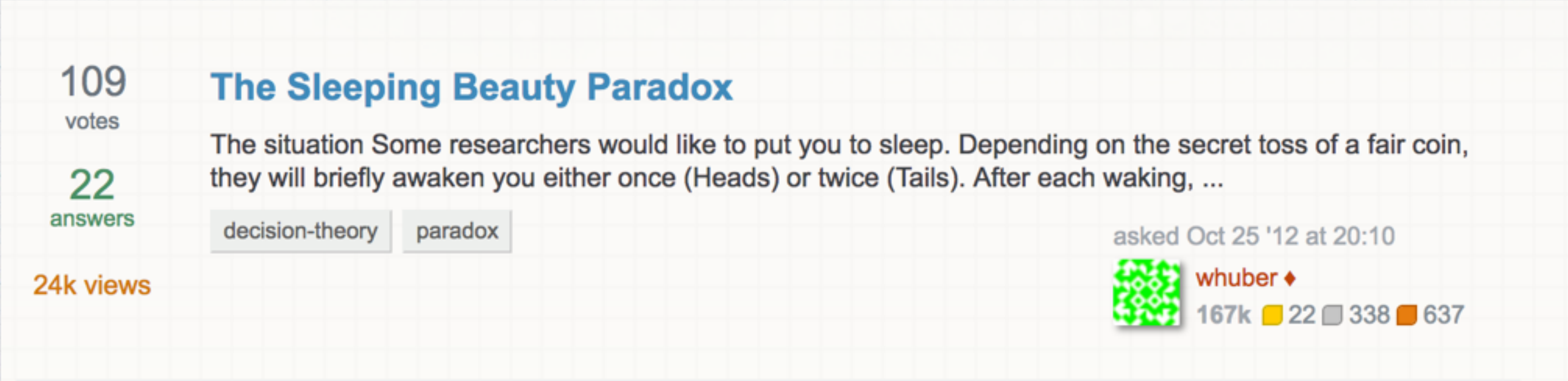
Topics Tag Features*

Most positive coefs:

- Careers
- Paradox
- Intuition
- Correlation Matrix
- Regression Strategies
- LSTM (Long Short Term Memory Network)

Most negative coefs:

- Phylogeny
- Active Learning
- Machine Translation
- Life Expectancy
- Structured Prediction
- Ranks



109
votes

22
answers

24k views

The Sleeping Beauty Paradox

The situation Some researchers would like to put you to sleep. Depending on the secret toss of a fair coin, they will briefly awaken you either once (Heads) or twice (Tails). After each waking, ...

decision-theory paradox

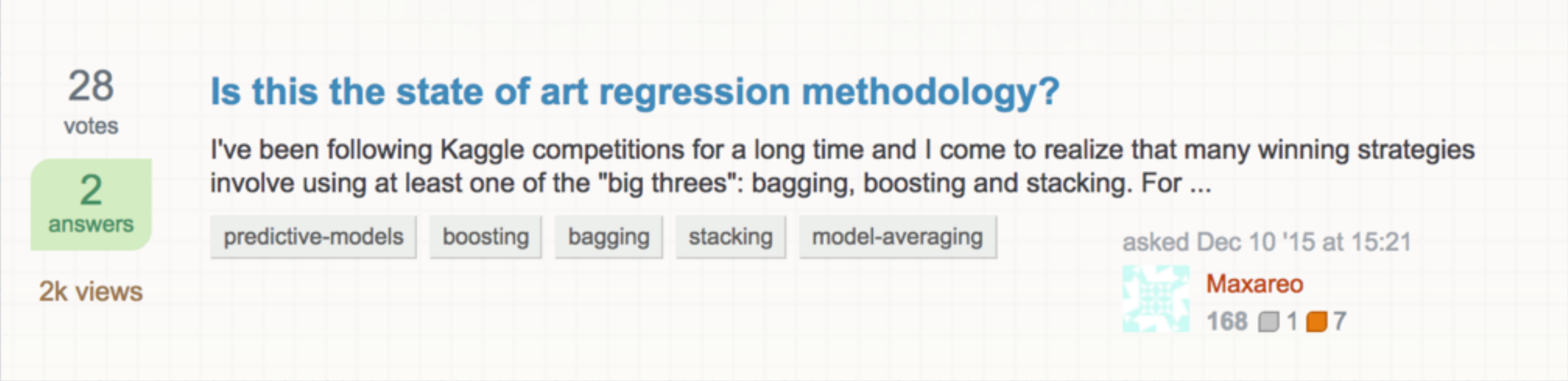
asked Oct 25 '12 at 20:10

whuber ♦
167k 🟨 22 ⬜ 338 🟧 637

*Take these with a grain of salt: small sample sizes for many of these tags

Next Steps?

- **Fancier Model:** the "big three": bagging, boosting, stacking
- **Better Text Features:** optimize tf-idf metrics (n-grams), convolutional neural network for feature extraction
- **Online learning:** model on the fly as questions come in



28 votes

2 answers

2k views

Is this the state of art regression methodology?

I've been following Kaggle competitions for a long time and I come to realize that many winning strategies involve using at least one of the "big threes": bagging, boosting and stacking. For ...

predictive-models boosting bagging stacking model-averaging

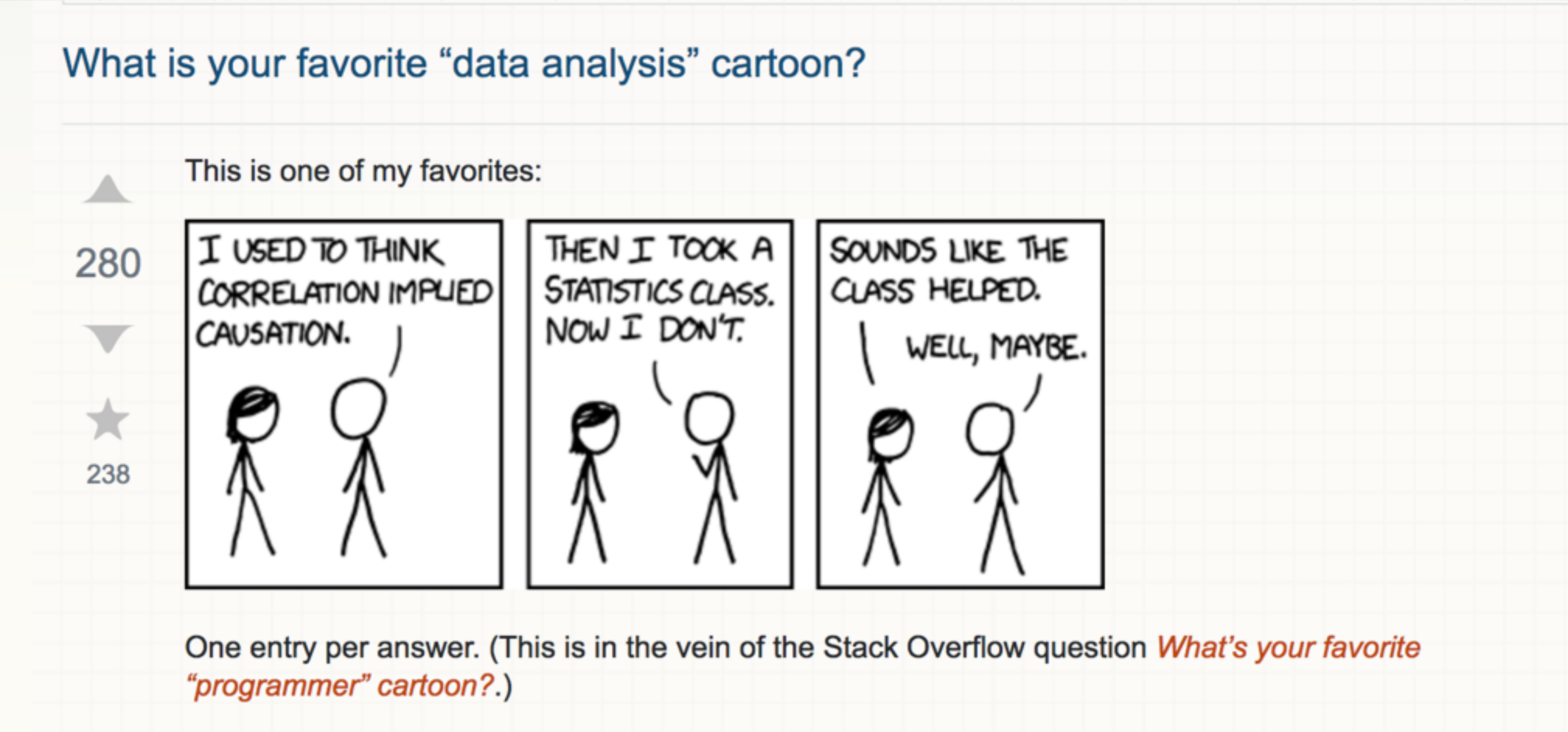
asked Dec 10 '15 at 15:21

Maxareo

168 □1 🟧7

Thank you!!!

Github: JEddy92



What is your favorite “data analysis” cartoon?

This is one of my favorites:

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

One entry per answer. (This is in the vein of the Stack Overflow question *What’s your favorite “programmer” cartoon?*.)

Source: xkcd