Min(e)d Your Tags: Analysis of Question Response Time in StackOverflow

Vasudev Bhat
Stony Brook University

Adheesh Gokhale
Stony Brook University

Ravi Jadhav
Stony Brook University

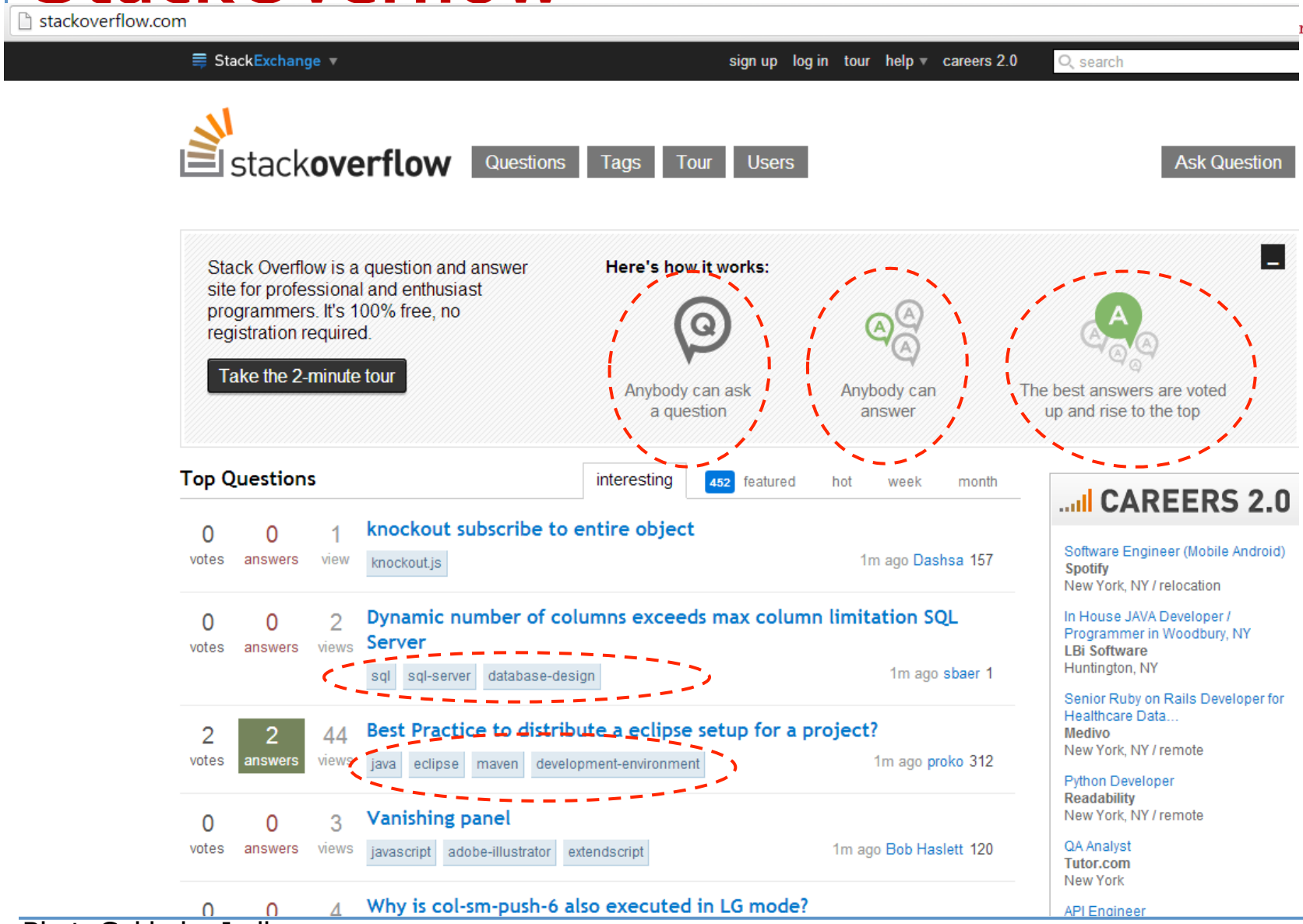
Jagat Pudipeddi
Stony Brook University

Leman Akoglu
Stony Brook University

ASONAM
Beijing, China
Aug 17-20, 2014

Stony Brook University
Computer Science

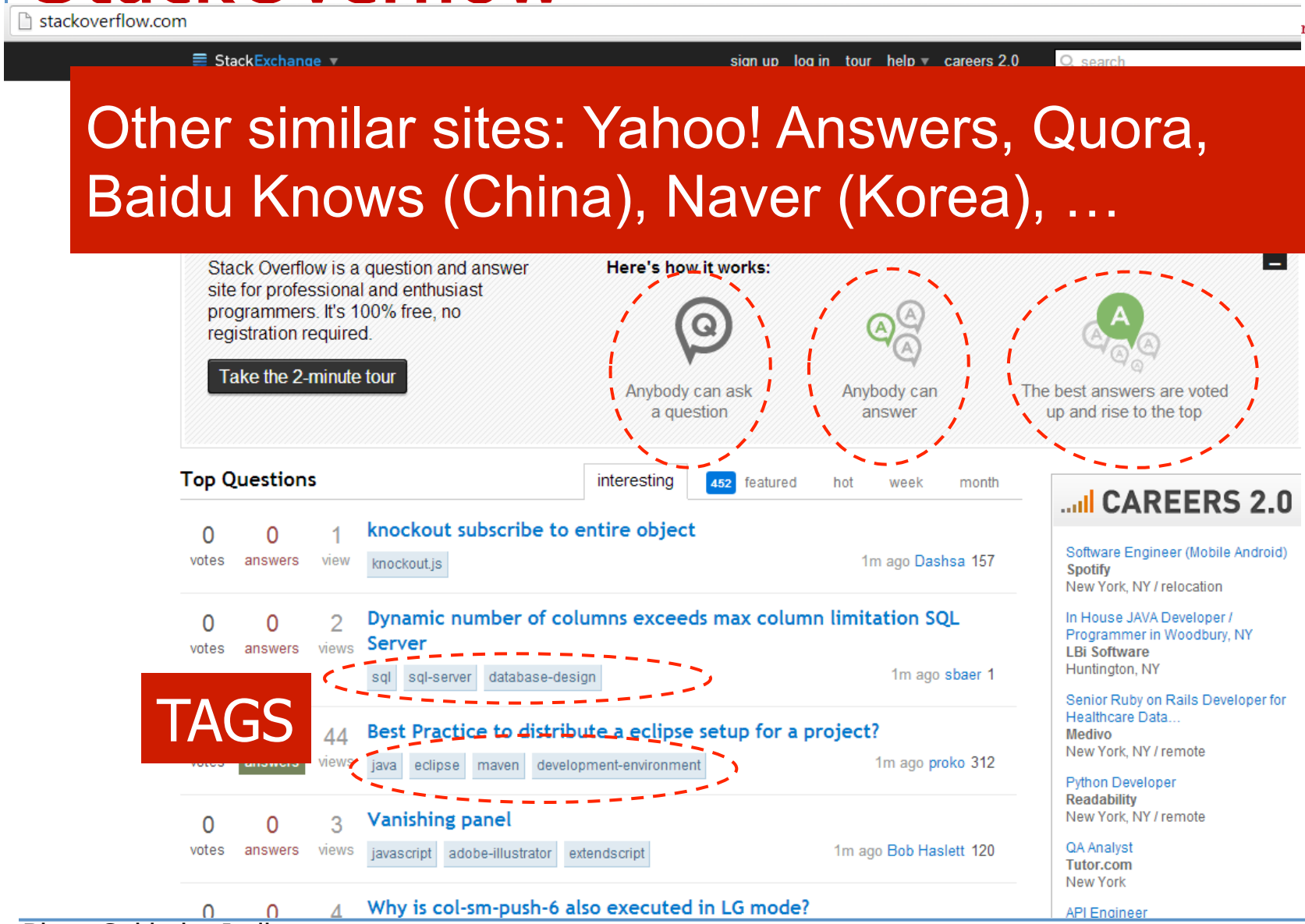
StackOverflow



StackOverflow

Stony Brook University

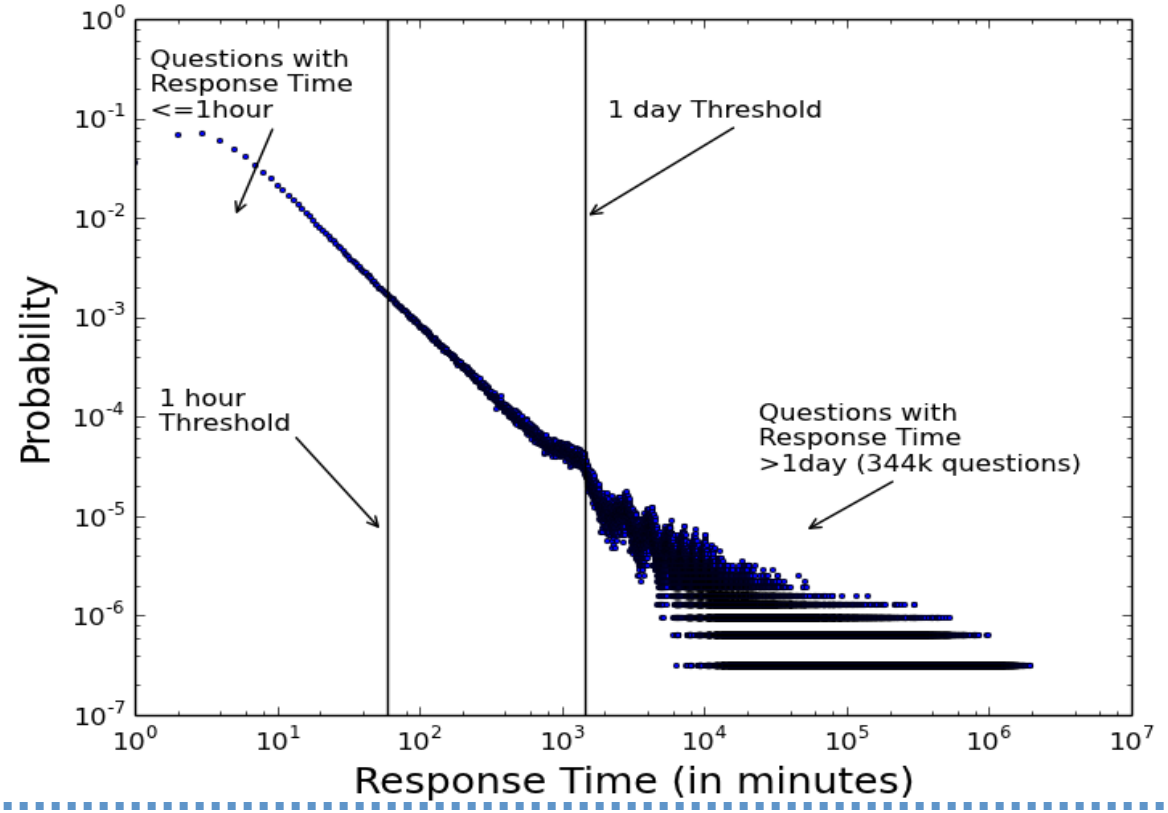
r Science



Response Time Distribution

Stony Brook University
Computer Science

30% of questions not answered within an hour have a response time of more than a day



Research Questions

Stony Brook University
Computer Science

- What are the intrinsic factors and signals that are likely to influence a question's response time?
- What site-level information is available that shows significant correlation with response time.
- How do tag-related factors relate to response time?
- Can we predict question response times using the evidential features available on the site? How effective are the tag-based features?

StackOverflow Data Statistics

Stony Brook University
Computer Science

- Users: 1.3 million
- Questions: 3.4 million, Answers: 6.8 million
- Questions answered: 91.3%
- Median time to receive an answer: 16 minutes
- Questions answered in ≤ 1hr: 63.5%
- Questions answered in > 1 day: 9.98%
- Expected number of tags a question has: 2.935

2 Prediction Tasks

Stony Brook University
Computer Science

Task 1.

Given a question (its tags, body, title, post date),
Predict if it will be answered in $\leq$ 16 minutes (median response time) or not.

Task 2.

Given a question (its tags, body, title, post date),
Predict if it will be answered in $\leq$ 1 hr or $>$ 1 day

Feature Extraction

- We find and organize 2 groups of features likely associated with response time

1

Tag based Question Features
tag_popularity: Average frequency of tags
num_pop_tags: Number of popular tags
tag_specificity: Average co-occurrence rate of tags
num_subs_ans: Number of active subscribers
percent_subs_ans: % of active subscribers
num_subs_t: Number of responsive subscribers
percent_subs_t: % of responsive subscribers

Feature Extraction

Stony Brook University
Computer Science

- We find and organize 2 groups of features likely associated with response time

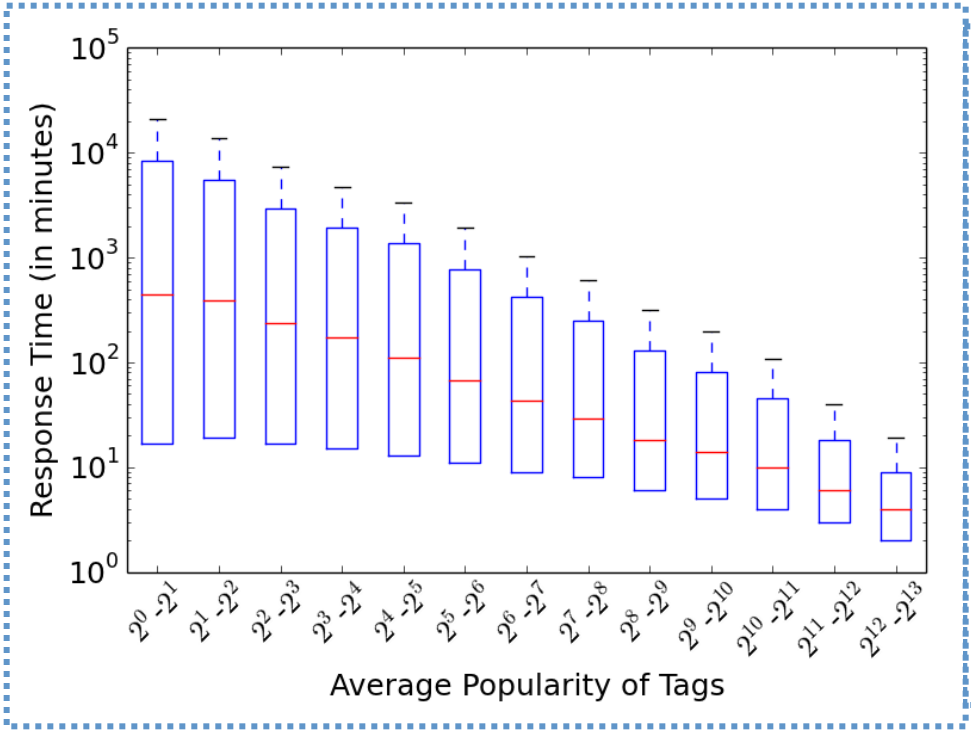
2

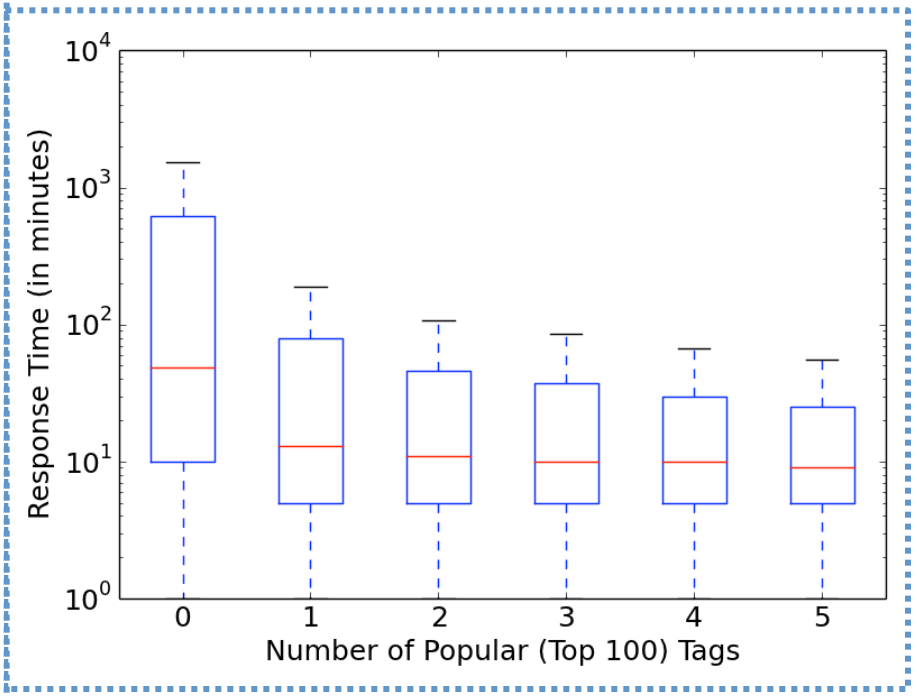
Non-tag based Question Features
num_code_snippet: Number of code segments
code_len: Total code length (in chars)
num_image: Number of images
body_len: Total body length (in chars)
title_len: Title length (in chars)
end_que_mark: Whether title ends with question mark
begin_que_word: Whether title starts with 'wh' word
is_weekend: Whether question posted on weekend
num_active_verb: Number of verbs that indicate action
num_selfref: Number of self references of the asker

Feature Analysis: Tag-Based

Stony Brook University
Computer Science

- Tag Popularity





Questions with popular tags receive an early response.

Feature Analysis: Tag-Based

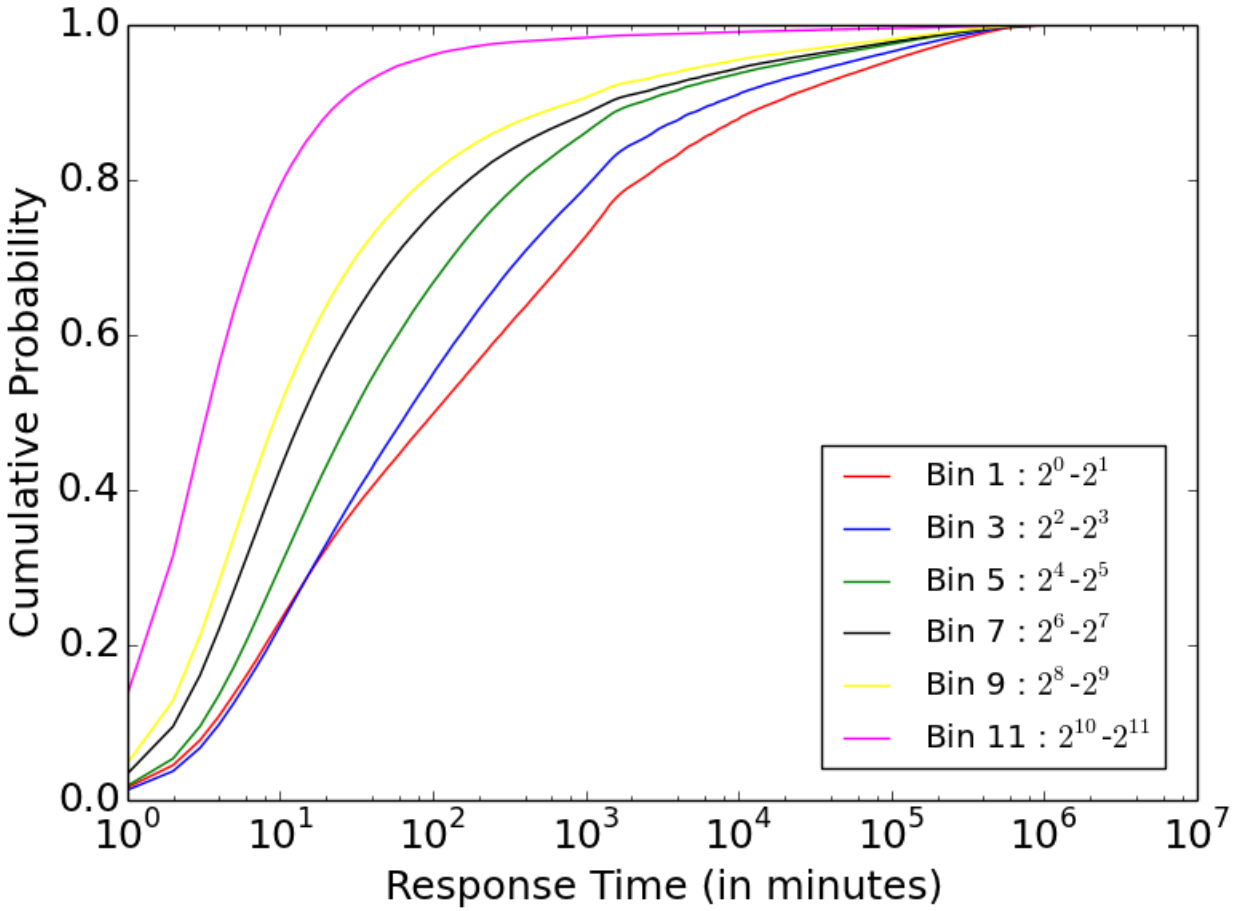
Stony Brook University
Computer Science

- Subscribers: 2 Types
  - Active Subscriber:
    - Activeness is associated with the amount of questions with a certain tag that a user is capable of answering.
    - A user who has posted "greater than a threshold δ(ans) number of answers" in the past predefined "number of months δ(mo)".
  - Responsive Subscriber:
    - Responsiveness is associated with the speed with which the user answers questions containing a certain tag.
    - User is a "responsive subscriber" of a tag t if his average response time for questions containing t and posted in "recent past" is less than a threshold δ(t).

Feature Analysis: Tag Based

Stony Brook University
Computer Science

- Active Subscribers

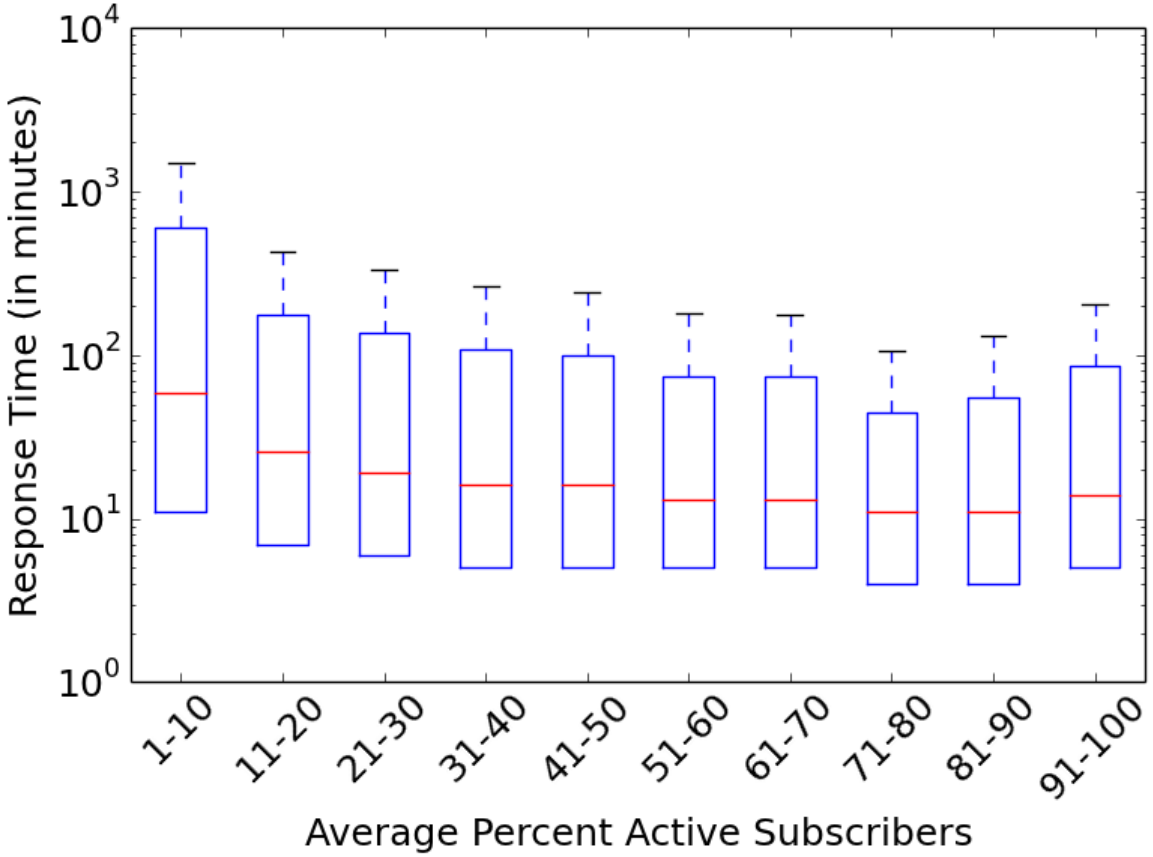


Higher bin numbers correspond to larger subscriber count and lower response time.

Feature Analysis: Tag Based

Stony Brook University
Computer Science

- Percent Active Subscribers

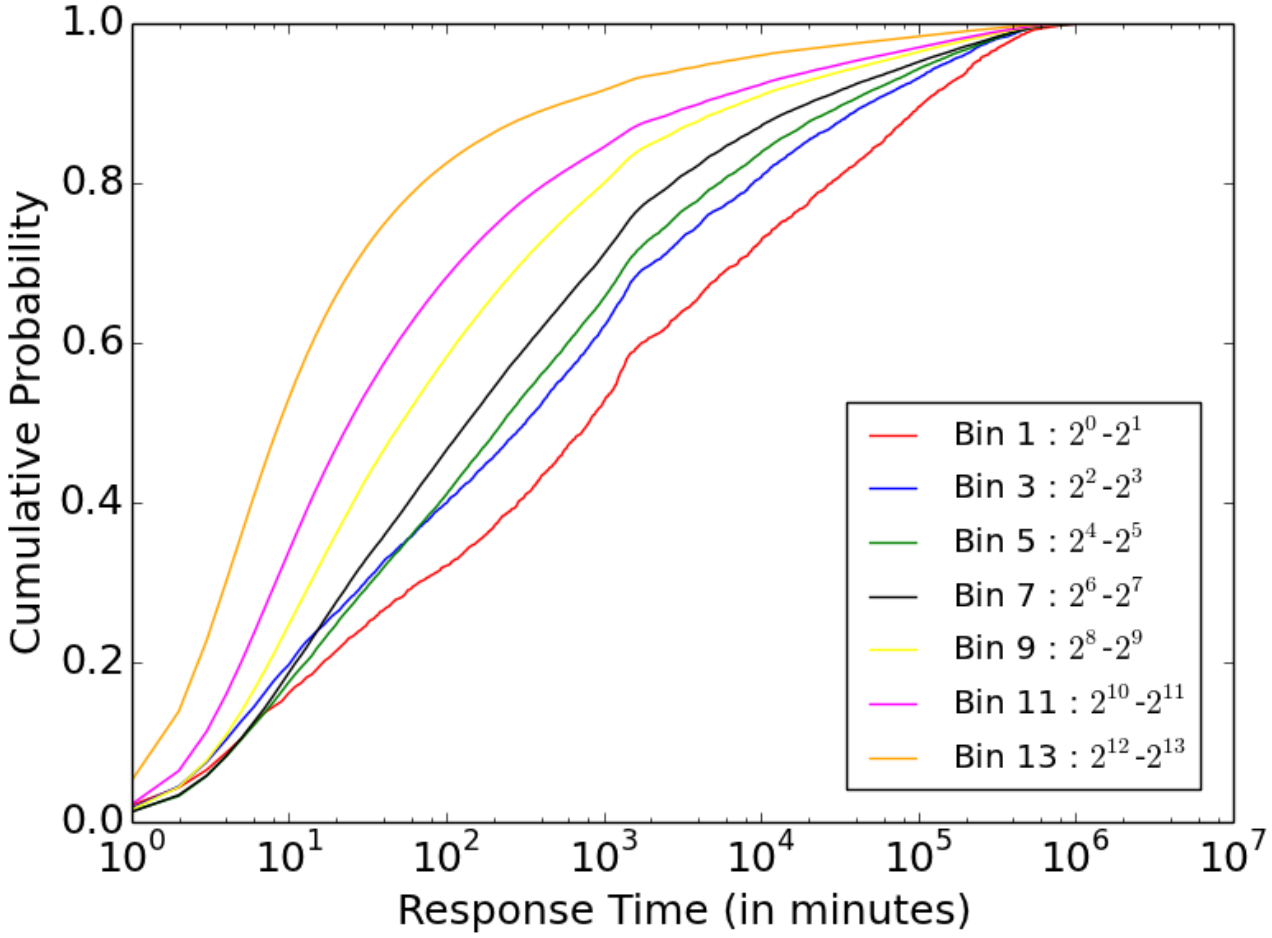


Greater the percentage of active subscribers, lesser is the response time.

Feature Analysis: Tag Based

Stony Brook University
Computer Science

- Responsive Subscribers

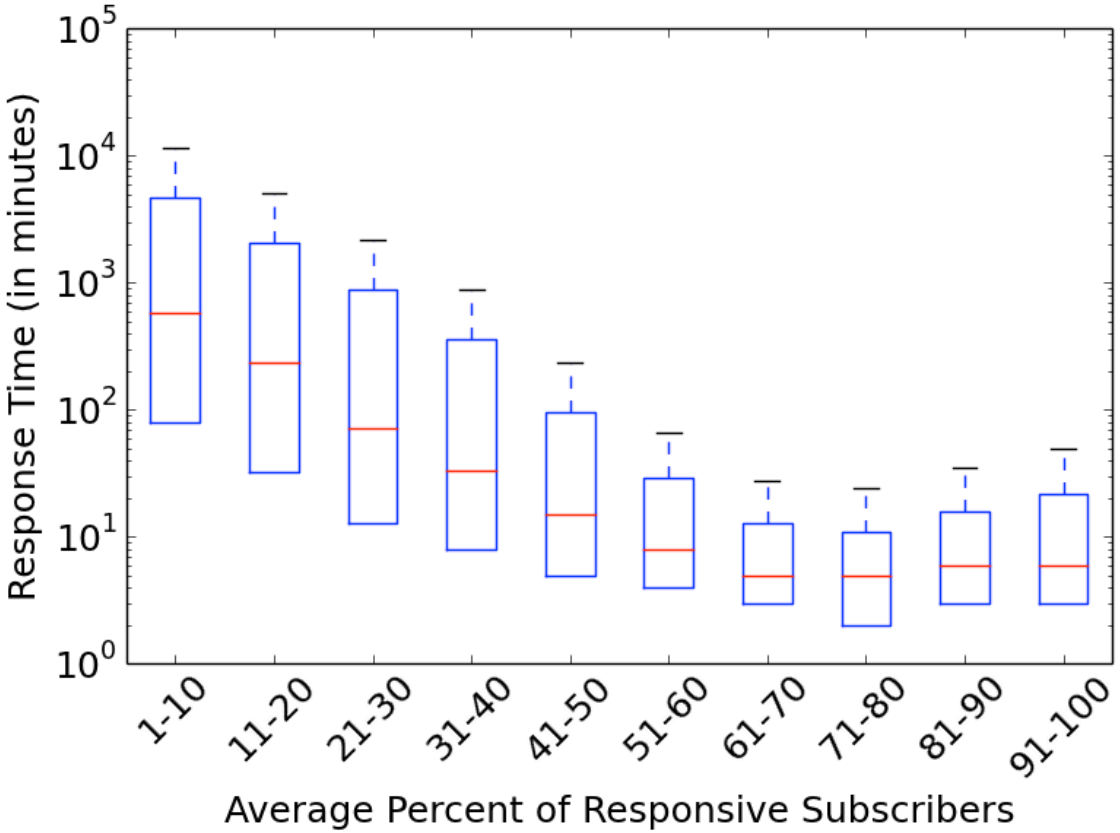


Higher bin numbers correspond to larger subscriber count and lower response time.

Feature Analysis: Tag Based

Stony Brook University
Computer Science

- Percent Responsive Subscribers

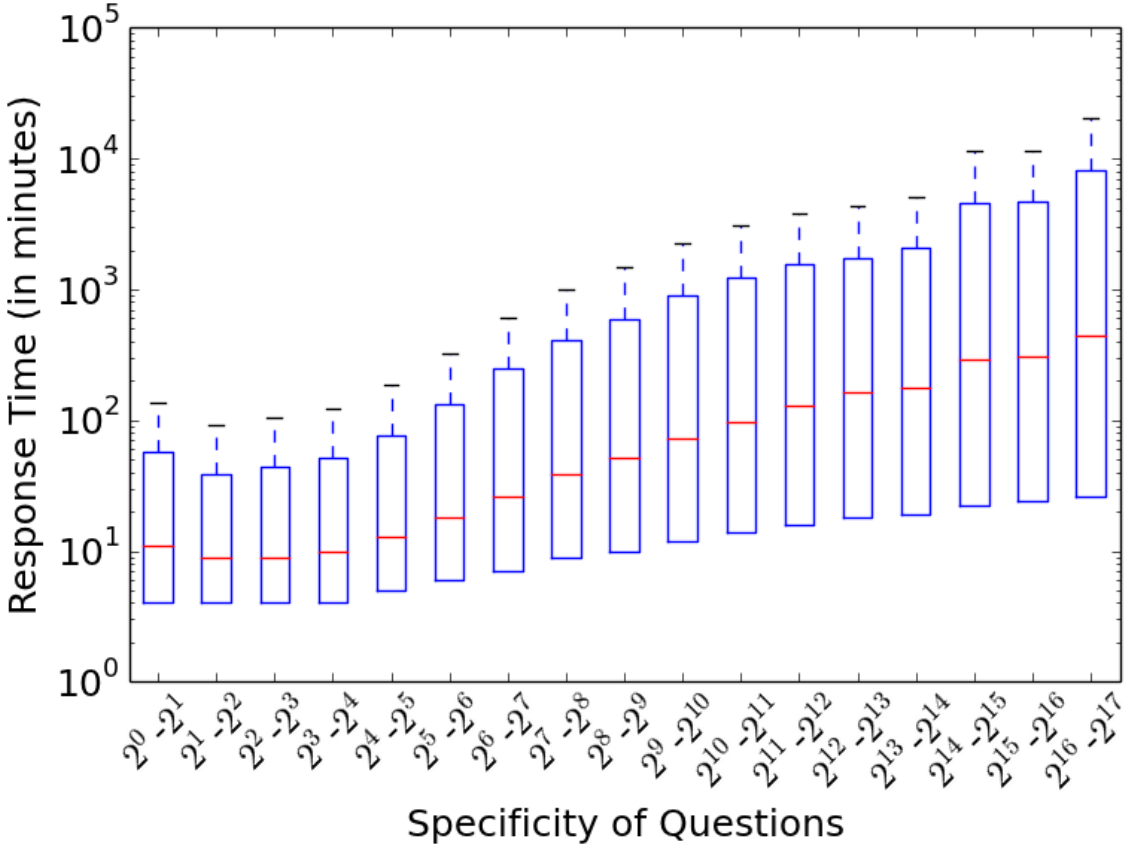


Greater the percentage of responsive subscribers, lesser is the response time.

Feature Analysis: Tag Based

Stony Brook University
Computer Science

- Specificity: Pairwise togetherness of tags in a question

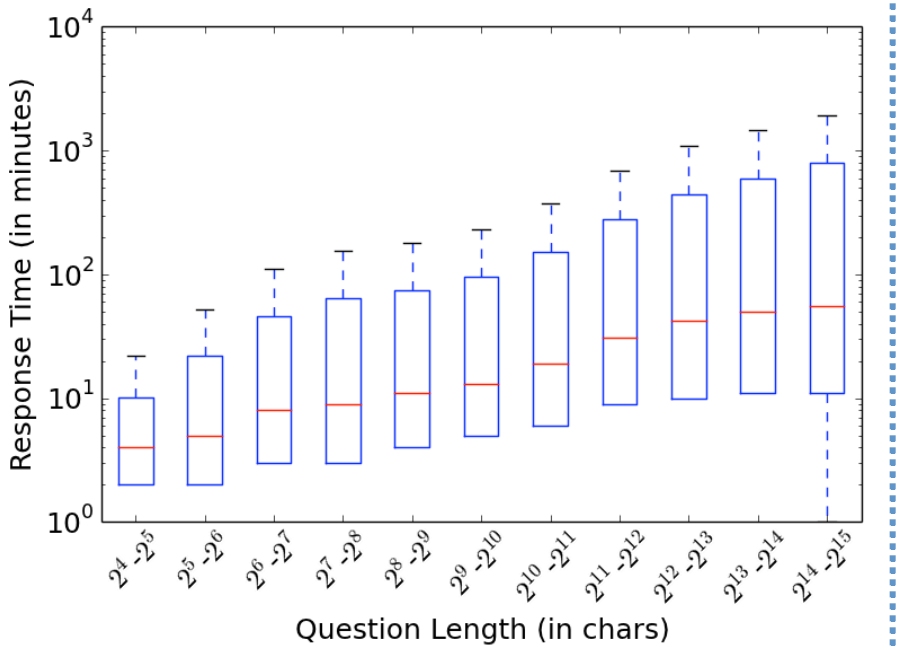


More the Specificity, greater is the response time

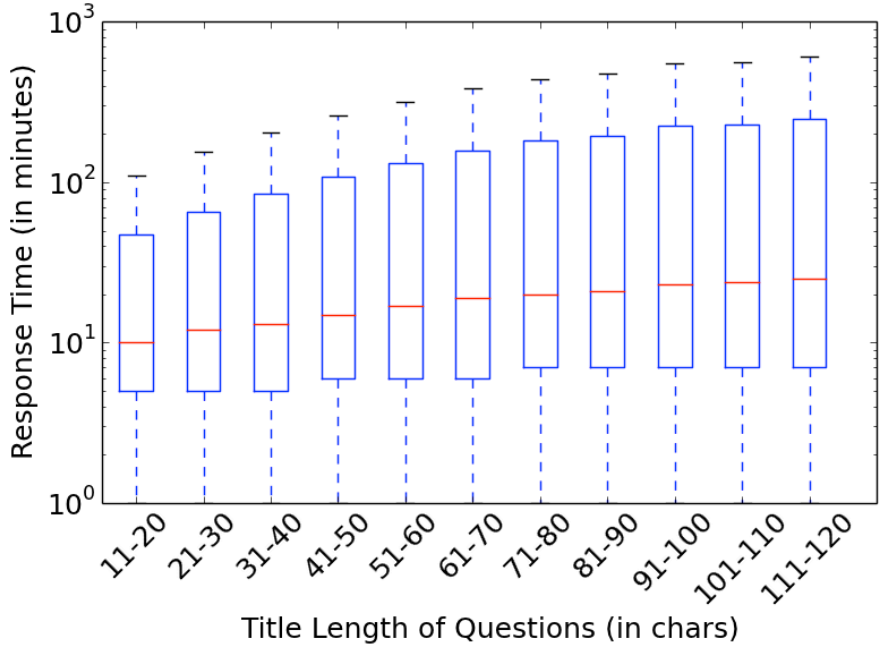
Feature Analysis: Non-Tag Based

Stony Brook University
Computer Science

- Body Length



- Title Length

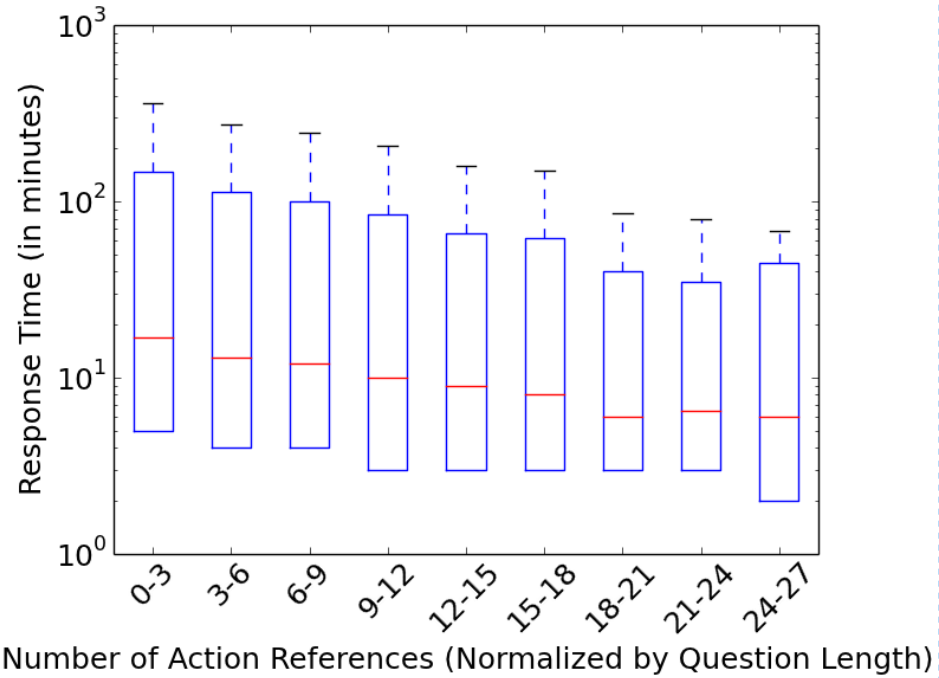


Questions with smaller length (body/title) receive quicker response.

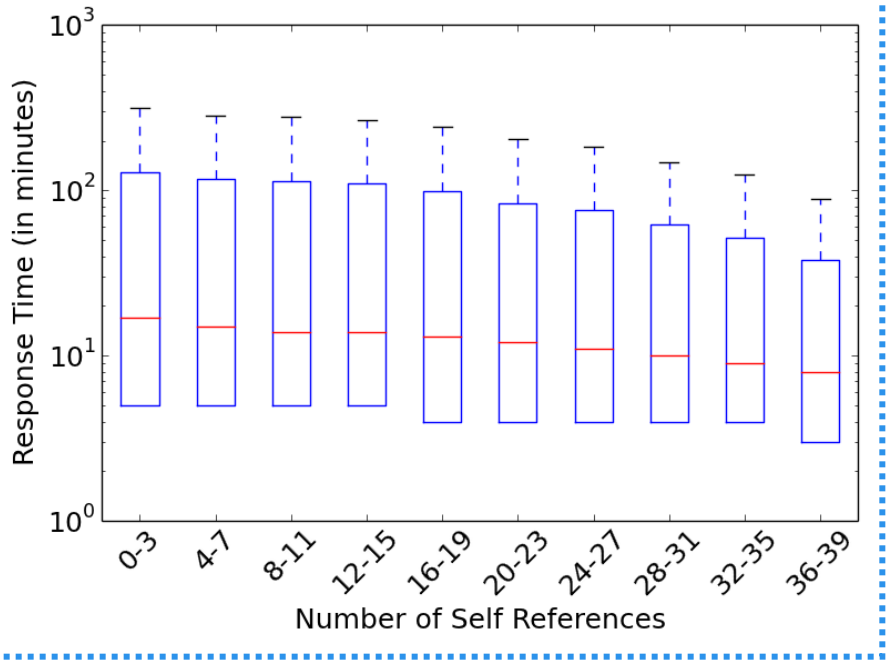
Feature Analysis: Non-Tag Based

Stony Brook University
Computer Science

- # Action Verbs



- # Self References



Feature Analysis: Summary

Stony Brook University
Computer Science

- Positive Correlations

Given a question, higher its

- body length
- title length
- code length
- tag specificity

the higher its response time tends to be.

Feature Analysis: Summary

Stony Brook University
Computer Science

- Negative Correlations

Given a question, the larger its

- tag popularity, number of popular tags
- active subscribers, responsive subscribers associated with its tags
- number of active verbs and self-referencing words

the lower its response time tends to be

Types of Experiments

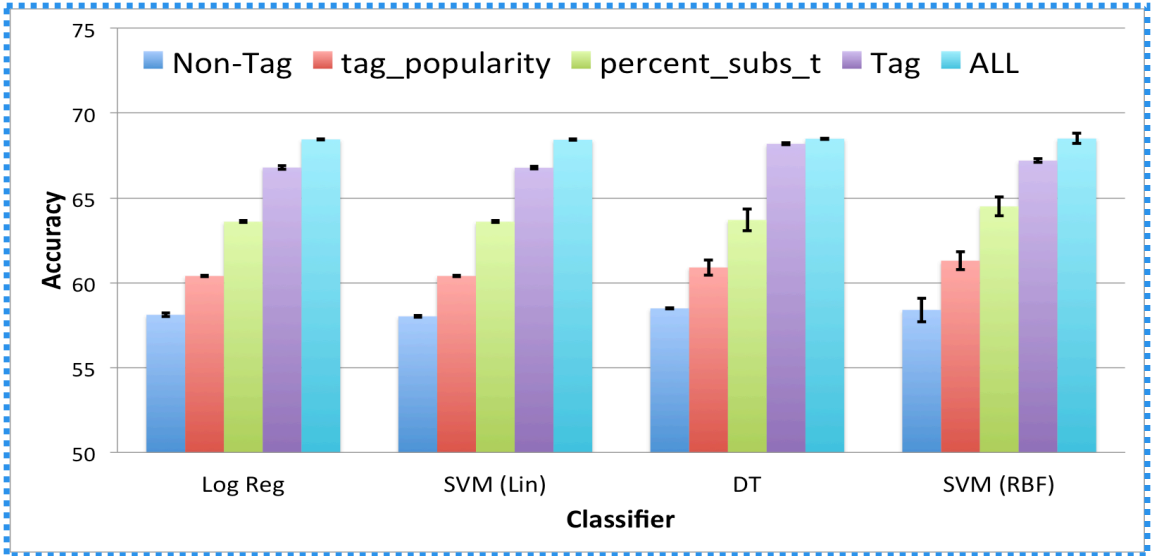
Stony Brook University
Computer Science

- Type 1: All but only non-tag based (10) features.
- Type 2: Only tag popularity (1).
- Type 3: Only percent responsive subscribers(1).
- Type 4: All but only tag-based (9) features.
- Type 5: All (19) features.

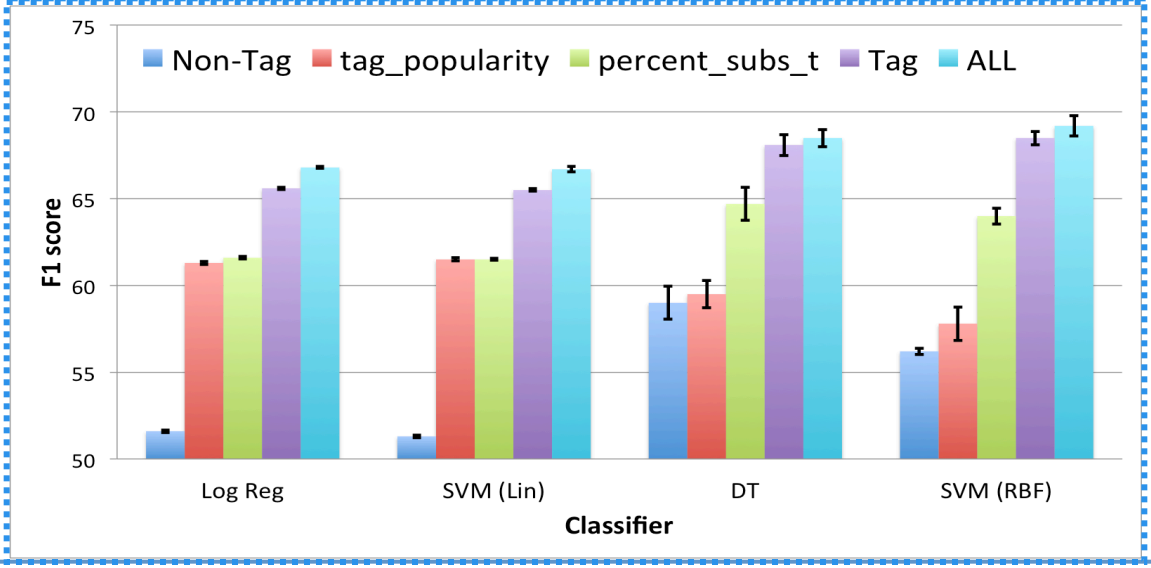
Prediction results: Task 1

Stony Brook University
Computer Science

Accuracy



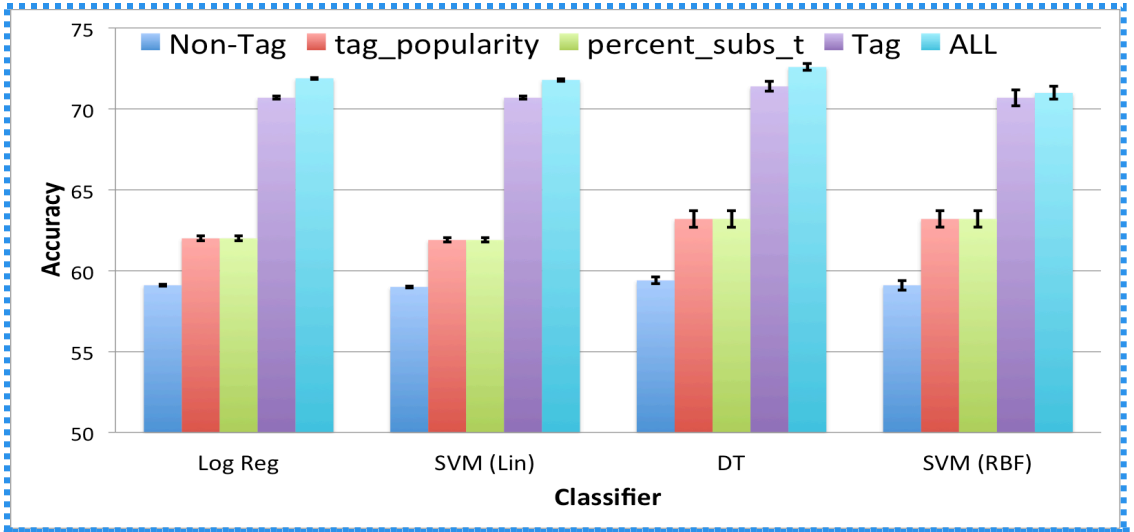
F1 Score



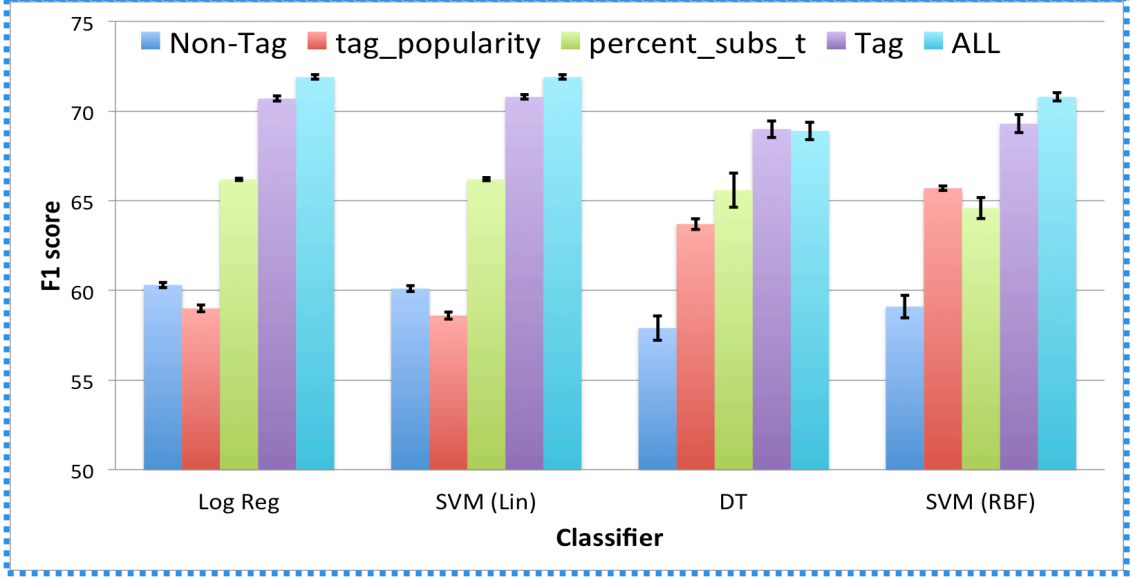
Prediction results: Task 2

Stony Brook University
Computer Science

Accuracy



F1 Score



Prediction analysis

Stony Brook University
Computer Science

- Top 3 features: Tag based

Task 1		Task 2	
* percent_subs_t	0.440	* percent_subs_t	0.506
* tag_popularity	0.173	* percent_subs_ans	0.266
* num_subs_t	0.130	* tag_popularity	0.085
body_len	0.123	body_len	0.057
* percent_subs_ans	0.033	* num_subs_ans	0.030
* num_subs_ans	0.026	end_que_mark	0.013
* tag_specificity	0.025	title_len	0.010
end_que_mark	0.013	* num_subs_t	0.010
title_len	0.013	code_len	0.009
code_len	0.012	* tag_specificity	0.007

Summary

- Studied a large set of factors likely to be associated with question response time.
- Analyzed tag-based features and illustrated the strong correlation between question tags and response time.
- Exploited tag-based features to estimate the response time.

Thank You

NORTHROP GRUMMAN

Stony Brook University