Discovering Value from Community Activity on Focused Question Answering Sites:

A Case Study of Stack Overflow

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, Jure Leskovec

STANFORD UNIVERSITY

Cornell University

Intro + Motivation

Q&A sites have evolved:
from places to get one-off answers to questions
to **large repositories of long-lasting, valuable knowledge**

Intro + Motivation

In this work, we promote a systemic view of Q&A sites

Rather than focus on question-answer pairs,
we view a question together with its full set of answers

We show that this new approach can help solve
important problems in modern Q&A sites

- Early identification of pages with long-lasting value
- Finding questions with insufficient answers

Outline

1. Data
2. Introduce tasks
3. Empirical findings
4. Task performance

Outline

1. Data

2. Introduce tasks

3. Empirical findings

4. Task performance

Data

Large, focused programming-related Q&A site

Very well curated by the community

Users	440K
Questions	1M
Answers	2.8M (26% marked as accepted)
Votes	7.6M (93% positive)
Favorites	775K (on 318K questions)

Complete dataset

How to format a JSON date?

434

148

I'm taking my first crack at Ajax with jQuery. I'm getting my data onto my page, but I'm having some trouble with the JSON data that is returned for Date data types. Basically, I'm getting a string back that looks like this:

```
/Date(1224043200000)/
```

From someone totally new to JSON - How do I format this to a short date format? Should this be handled somewhere in the jQuery code? I've tried the `jQuery.UI.datepicker` plugin using `$.datepicker.formatDate()` without any success.

FYI: Here's the solution I came up with using a combination of the answers here:

This solution got my object from the callback method and displayed the dates on the page properly using the date format library.

jquery | asp.net | asp.net-mvc | ajax | json

link | edit | flag

edited **May 5 '11 at 16:09**

Peter Mortensen
4,761 ●4 ●25 ●57

asked **Oct 15 '08 at 20:43**

Mark Struzinski
5,884 ●11 ●49 ●81
88% accept rate

22 I'd upvote this twice if I could.... – Christian Payne Oct 2 '09 at 5:52

27 Answers

active | oldest | **votes**

491

Eval is not necessary. This will work fine:

```
var date = new Date(parseInt(jsonDate.substr(6)));
```

The substr function takes out the "\/Date(" part, and the parseInt function gets the integer and ignores the ")\/" at the end. The resulting number is passed into the Date constructor.

link | edit | flag

answered **Feb 23 '10 at 4:15**
Roy Tinker
5,273 •2 •7 •21

3 This worked for me, thanks ! – Jhonny D. Cano -Leftware- Mar 16 '10 at 21:20

21 +1 for not using eval, and for also working if a timezone offset is included in the string (which was the case for me). – Remi Despres-Smyth Apr 13 '10 at 17:34

12 +1 This answer should be marked the answer. It is the most universal. – Evildonald Aug 23 '10 at 19:20

72

Here's a good explanation of why Microsoft chose this format to represent JSON dates in ASP.NET:

http://weblogs.asp.net/bleroy/archive/2008/01/18/dates-and-json.aspx

link | edit | flag

answered **Apr 23 '09 at 14:14**
Chris Kentfield
1,192 •6 •8

1 Good link thx Chris – aromawebdesign.com Mar 6 '11 at 22:00

+1 This link actually contained just the info I needed. Thanks! – Gunder Aug 8 '11 at 8:52

1 exploiting loopholes in json is evil. I don't like it. – neoneye Nov 10 '11 at 10:58

49

You can use this to get a date from json:

```
var date = eval(jsonDate.replace(/\/Date\((\d+)\)\//gi, "new Date($1)"));
```

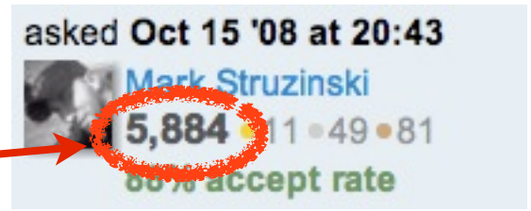
and then you can use JavaScript Date Format script (1.2 KB when minified and gzipped) to display it as you want.

link | edit | flag

answered **Oct 15 '08 at 21:14**
Panos
8,360 •1 •16 •36

Reputation

Stack Overflow is endowed with a highly respected reputation system



Action	Reputation Change
Q/A is upvoted	+5/+10
Q/A is downvoted	-2 (-1 to voter)
Answer is accepted	+15 (+2 to acceptor)
Answer wins bounty	+ bounty amount
Offer bounty	- bounty amount
Answer marked as spam	-100

Outline

1. Data

2. Introduce tasks

3. Empirical findings

4. Task performance

Tasks

Two questions from the Q&A site owner's perspective:

1. Predict long-term value of a question page

→ help guide consumers of information to high-quality content

2. Predict whether a question has been sufficiently answered

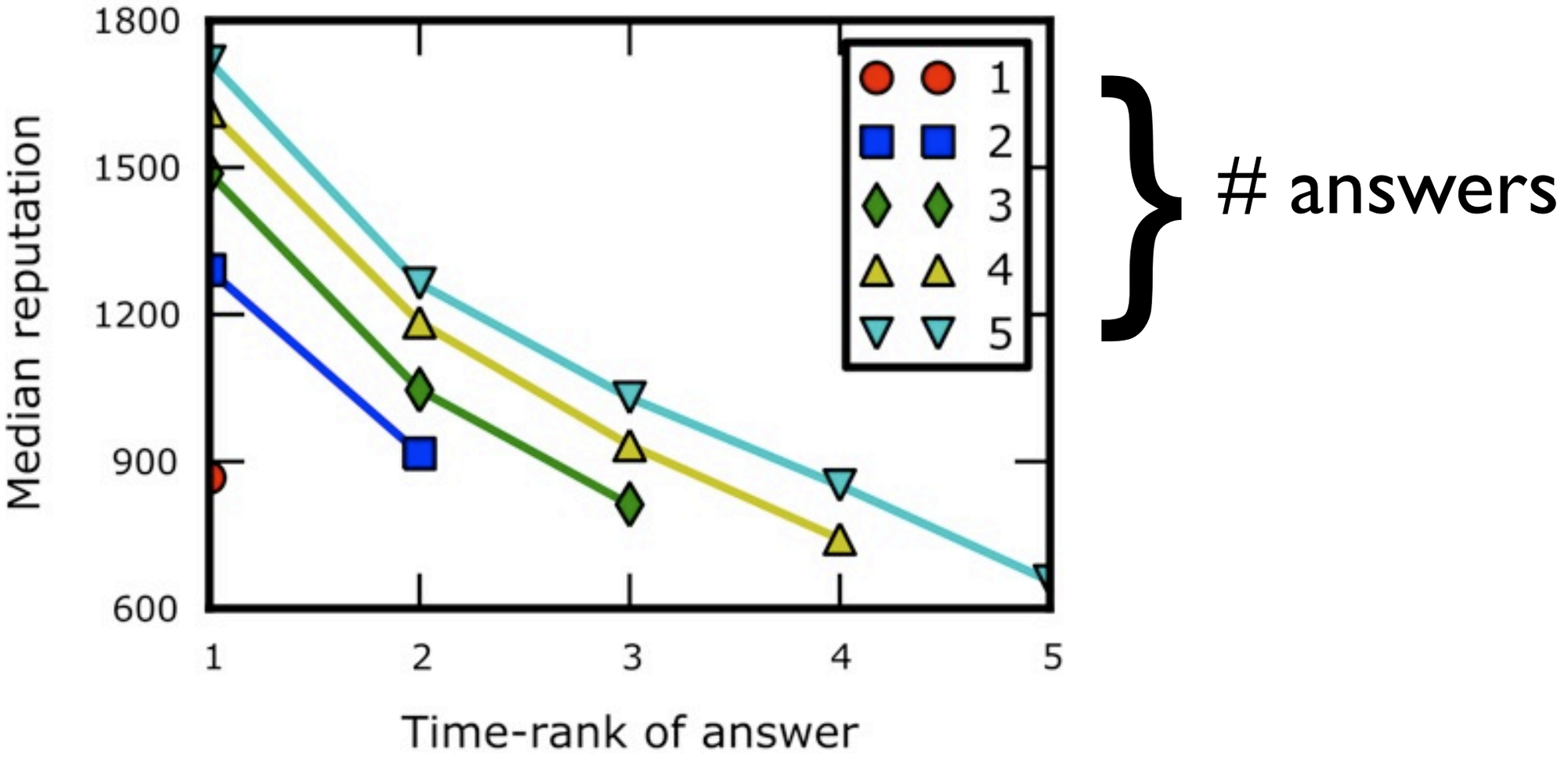
→ help direct producers of information to questions in need of expert attention

What features should we use to predict this?

Outline

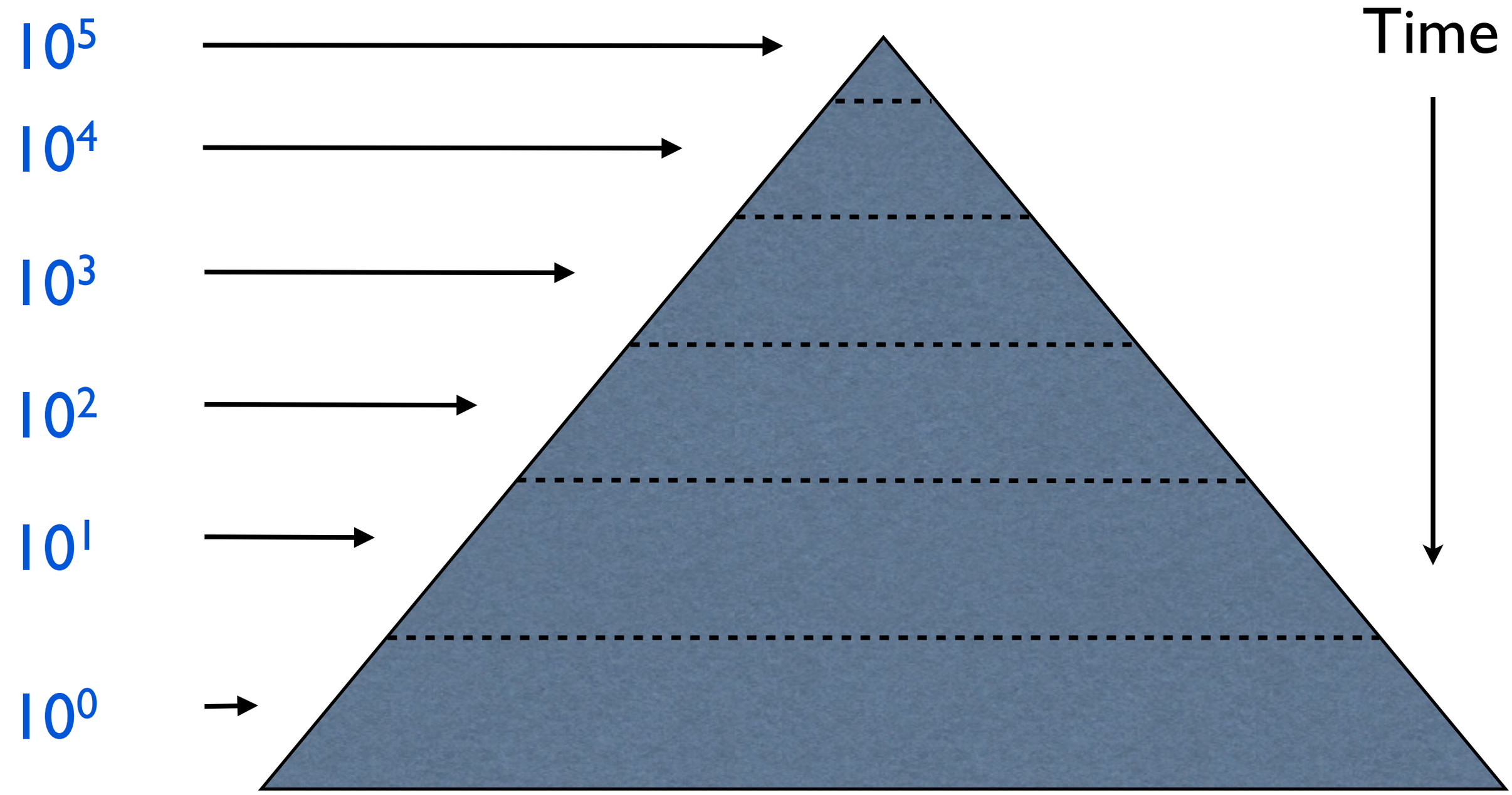
1. Data
2. Introduce tasks
3. Empirical findings
4. Task performance

Is there a relationship between the site-level reputation system and question-level dynamics?

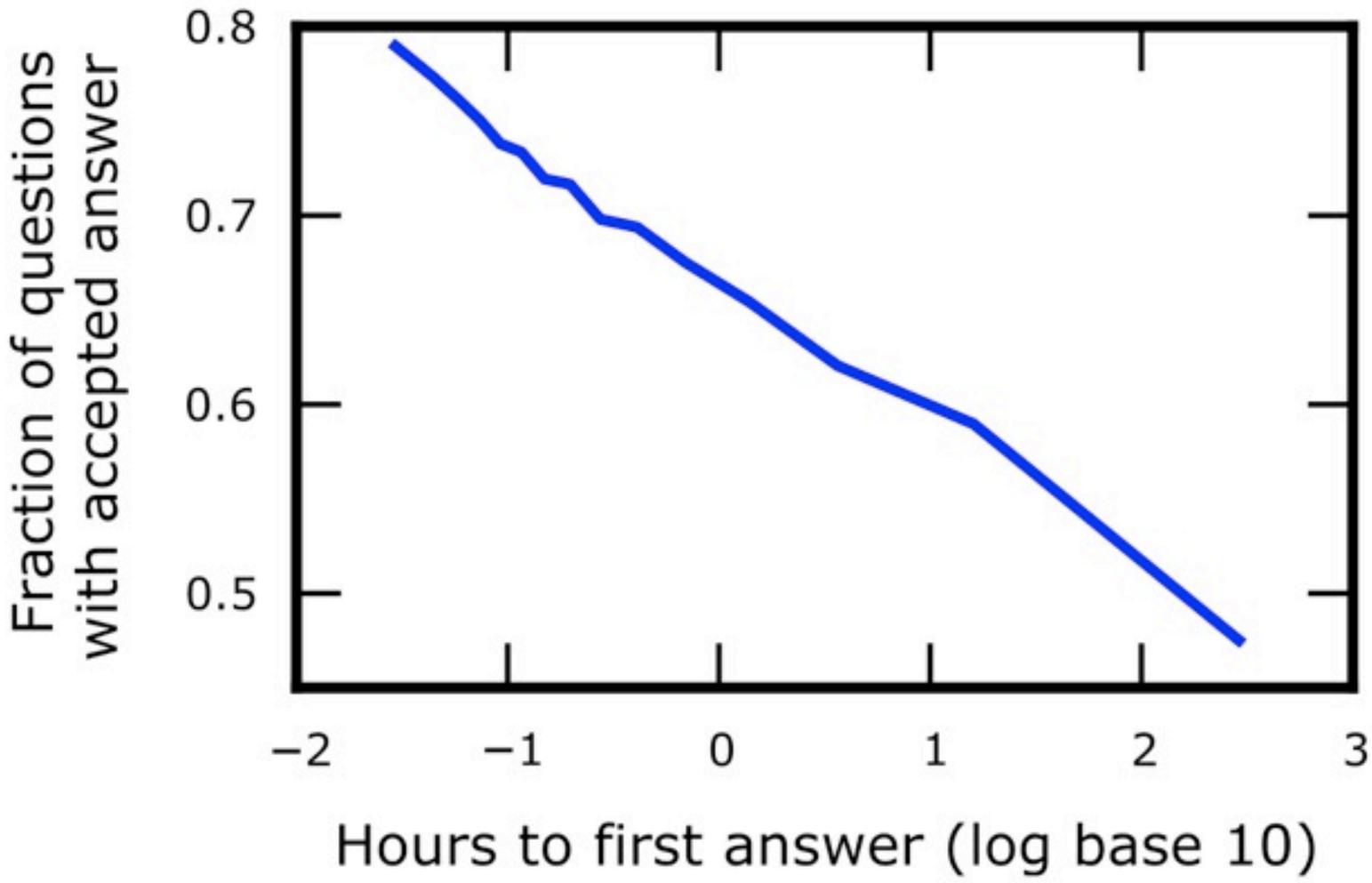


Higher-rep users arrive earlier

First principle: Reputation Pyramid



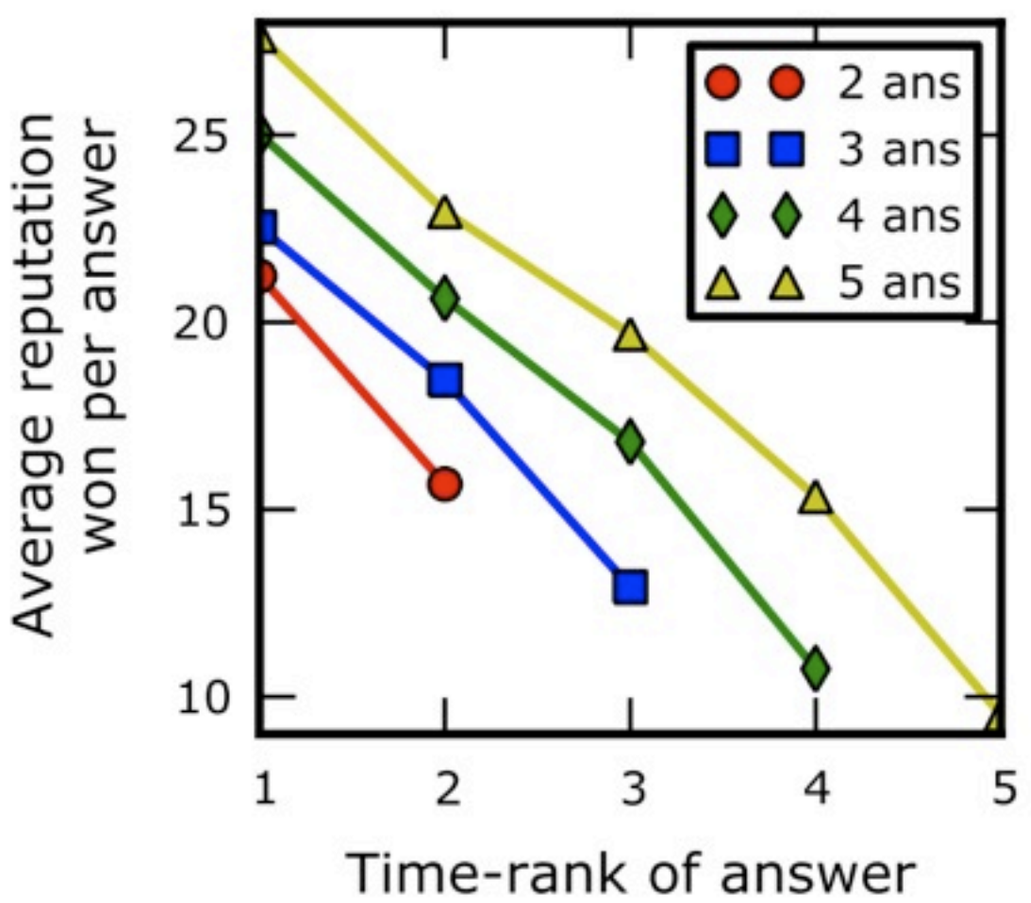
Mental model, not an explicit structure

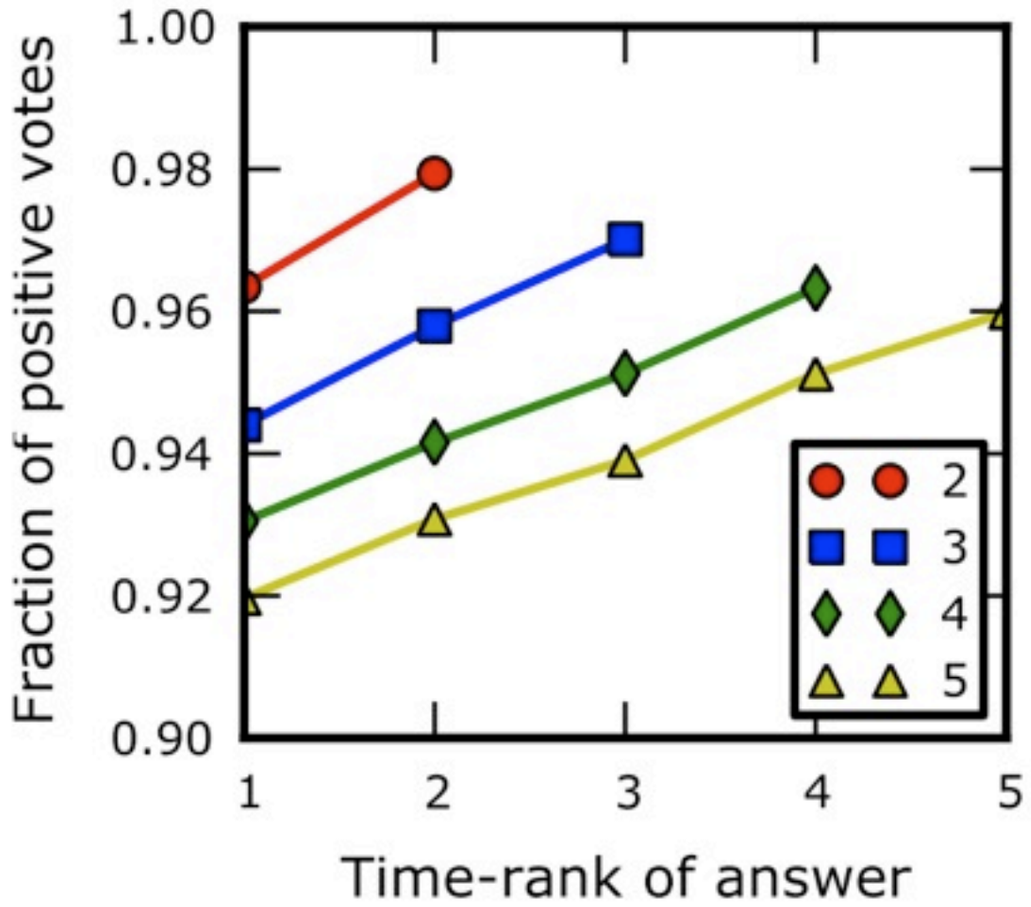


The longer it takes for the first answer to arrive, the less likely that any answer will be accepted

Consistent with reputation pyramid picture!

Two competing notions of answer quality:





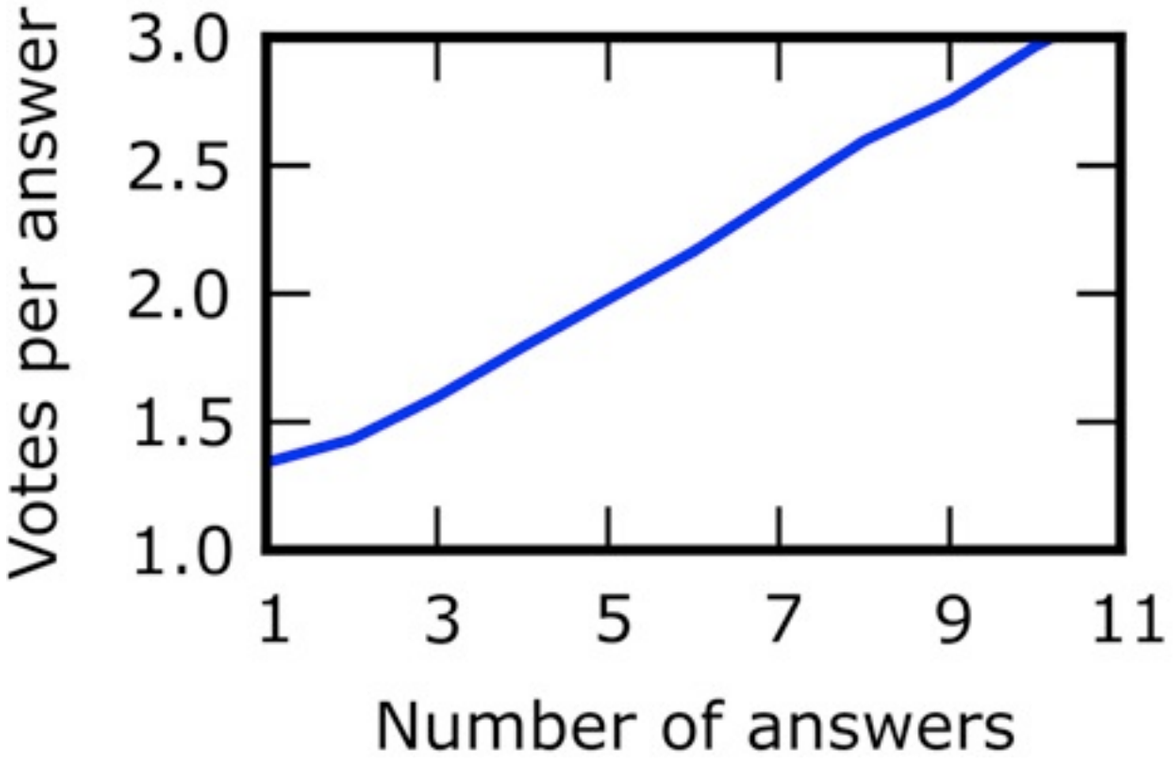
Earlier → More rep points

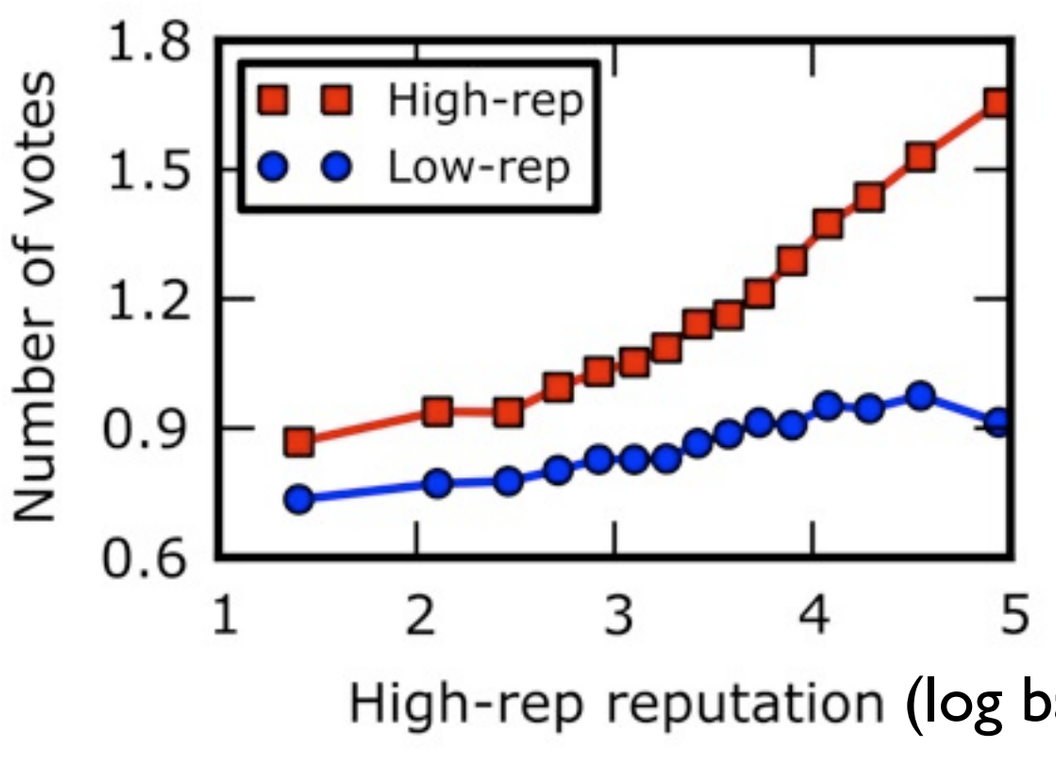
Later → Better vote score

Resolving these 2 notions is an open problem

Second Principle: "rising tide lifts all boats"

Is there competition between answers?





More activity ⟶ more votes for everybody

⟹ Supports our systemic view of Q pages

Outline

1. Data
2. Introduce tasks
3. Empirical findings
4. Task performance

Task 1: predict long-term value of a question page given how it looks a short time after it is created

Long-term value = Number of page-views one year after creation (in our data)

See one hour of data, predict views one year later

Set up as binary classification task: high/low page-views

We optimize for simplicity and interpretability

→ use logistic regression

Features

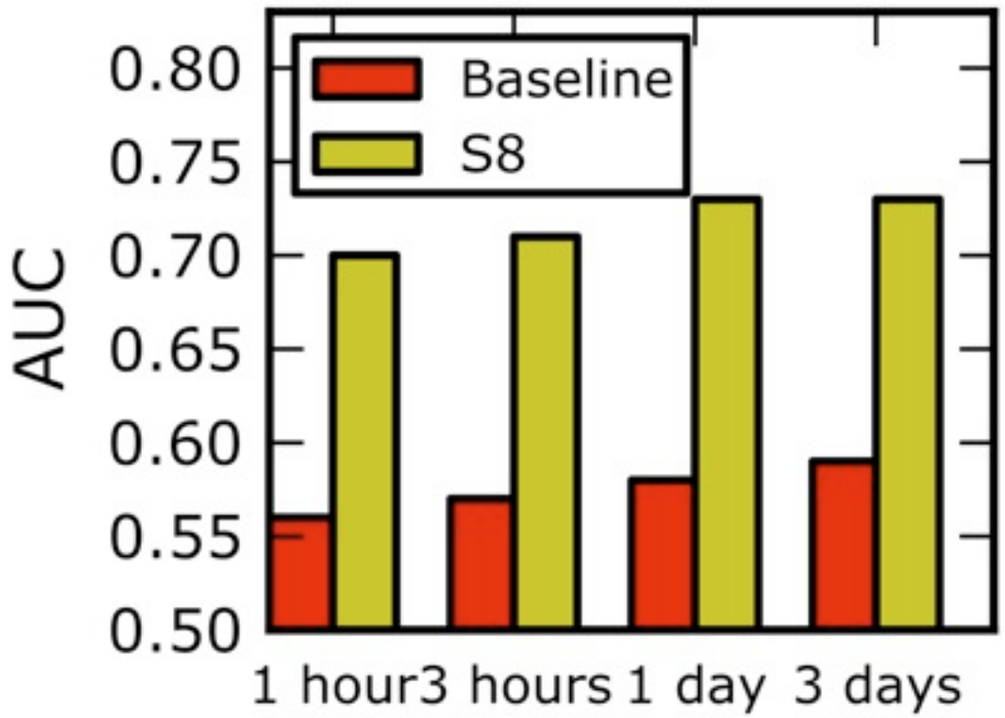
Set	Description (# feats)	Examples
A	Questioner features (4)	*reputation, number of previous Qs, ...*
B	Activity & Q/A quality (8)	*highest answer score, highest answerer rep, ...*
C	Community processes (8)	*average answerer reputation, # comments on answer by highest reputation answerer, ...*
D	Temporal processes (7)	*average time between answers, time for highest-scoring answer to arrive, ...*

Compare against "crowd-sourced" baseline: # favorites on question and question score (upvotes-downvotes) – 2 explicit mechanisms that measure value

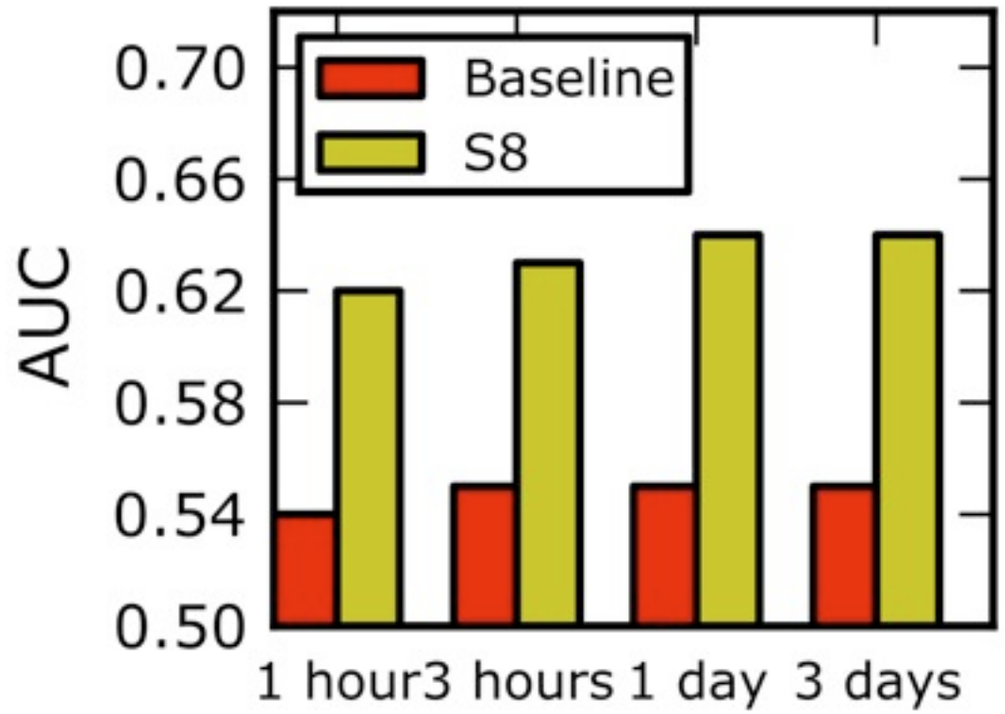
We perform feature selection and end up using 8 important features (S_8):

Feature	Coefficient
Number of answers	+0.61
Sum of answer scores	+0.47
# of questioner's questions (log scale)	-0.46
Length of highest-scoring answer	+0.38
Questioner's reputation (log scale)	+0.31
Time for highest-scoring answer to arrive	+0.22
# comments on highest-scoring answer	+0.19
# comments on highest-reputation answerer's answer	+0.17

Results



Top 25% vs. Bottom 25%



Top 50% vs. Bottom 50%

Features of the community processes that underlie the creation of the entire question page are useful for discovering long-term value at a very early stage

Task 2: Predict whether a question has been sufficiently answered

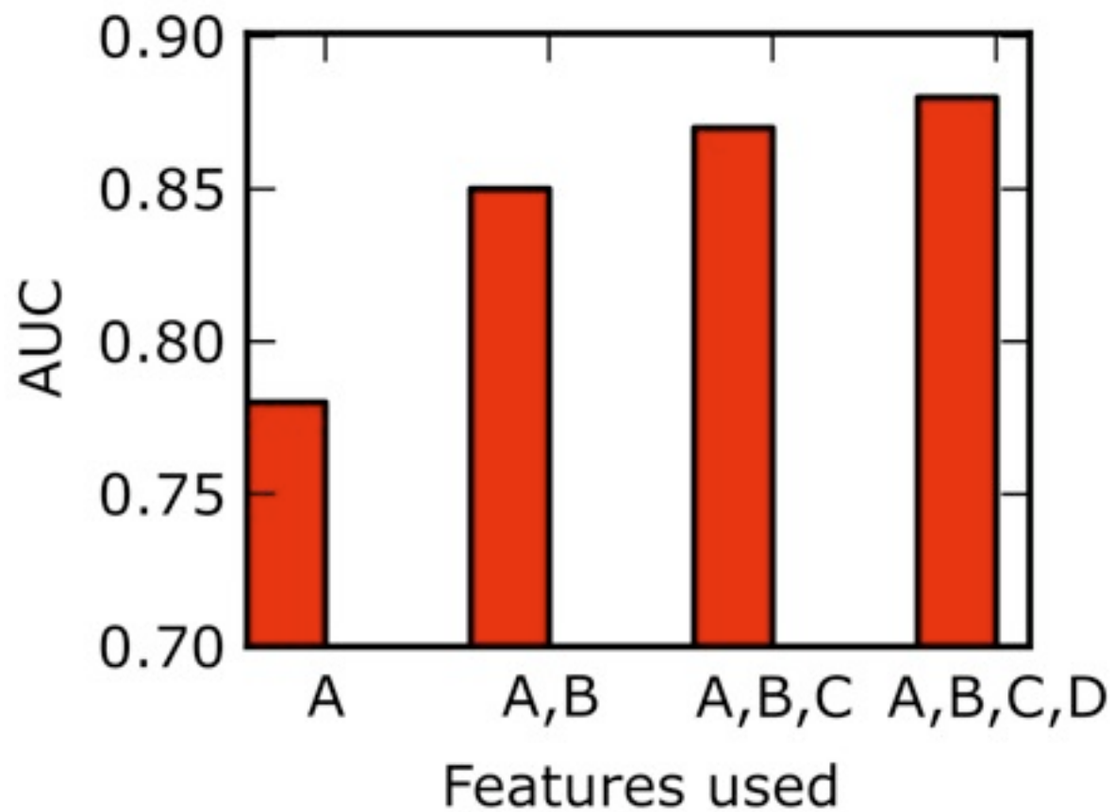
Setup: Given features of a question page, determine whether the question is about to accept one of the existing answers or offer a bounty

– Same logistic regression framework (with a balanced dataset)
– No natural baseline, so we compare our 4 classes of features
– Again perform feature selection, narrow down to set of 18 features

Results – Task 2

– Questioner features are powerful

– But adding features of community + temporal processes significantly boost performance



Features of the community processes underlying Q&A activity can provide important early indications

Conclusion

Q&A sites have evolved into focused communities

We suggest a shift in perspective from question-answer pairs to viewing questions together with their complete set of answers as one unit

There is useful information in the community and temporal processes for tasks like predicting long-term value and deciding if a question needs help

Thanks!