#### Finding High-Quality Content in Social Media

Eugene Agichtein Emory University Atlanta, USA eugene@mathcs.emory.edu Carlos Castillo Yahoo! Research Barcelona, Spain chato@yahoo-inc.com Debora Donato Yahoo! Research Barcelona, Spain debora@yahoo-inc.com

Aristides Gionis Yahoo! Research Barcelona, Spain gionis@yahoo-inc.com Gilad Mishne Search and Advertising Sciences, Yahoo! gilad@yahoo-inc.com

#### ABSTRACT

The quality of user-generated content varies drastically from excellent to abuse and spam. As the availability of such content increases, the task of identifying high-quality content in sites based on user contributions—social media sites becomes increasingly important. Social media in general exhibit a rich variety of information sources: in addition to the content itself, there is a wide array of non-content information svallable, such as links between items and explicit quality ratings from members of the community. In this paper we investigate methods for exploiting such community feedback to autometically identify high quality content. As a test case, we focus on Yahoo! Answers, a large community question/answering portal that is particularly rich in the amount and types of content and social interactions swallable in it. We introduce a general classification framework for combining the evidence from different sources of information, that can be tuned automatically for a given social media type and quality definition. In particular, for the community question/answering domain, we show that our system is able to separate high-quality items from the rest with an accuracy close to that of humans.

#### Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing - indexing methods, linguistic processing; H.3.3 Information Search and Retrieval - information filtering, search process.

#### General Terms

Algorithms, Design, Experimentation.

#### Keywords

Social media, Community Question Answering, User Interactions.

Remittion to make digital or hard copies of all or part of this work for presental or characters are in granted without for provided that copies are not make or distributed for profet or commercial advantage and that copies have this notice and the full citation on the first page. To copy otherwise, to supublish, no post on servers or to redistribute to liter, apprise prior specific premission and/or a fee. WIDM'05 February 11–12, 2008, Palo Alto, California, USA.

WSDM'06; February 11–12, 2008, Palo Alto, California, USA Copyright 2008 ACM 978-1-59593-927-9708/0002...\$5.00.

#### 1. INTRODUCTION

Recent years have seen a transformation in the type of content scalishic on the web. During the first decade of the web's promisence—from the early 1900s onwards—most on-line content resembled traditional published material the majority of web users were consumers of content, created by a relatively small amount of publishers. From the early 2000s, user-generated content has become increasingly popular on the web: more and more users participate in content creation, rather than just consumption. Fopular user-generated content (or social nedia) domains include blogs and web forems, social bookmarking sites, photo and widoo sharing communities, as well as social networking platforms such as Facebook and MySpace, which offers a combination of all of these with an emphasis on the relationships among the users of the community.

Community-driven question/answering portals are a particular form of user-generated content that is gaining a large audience in recent years. These portals, in which users answer questions posed by other users, provide an abstractive channel for obtaining information on the web: rather than browsing results of search engines, users present detailed information assots—and get direct responses authored by humarkets, this information seeking behavior is dominating over traditional web search [28].

An important difference between user-generated content and traditional content that is particularly significant for knowledge-based media such as question/answering portals is the variance in the quality of the content. As Anderson [3] describes, in traditional publishing-mediated by a publisher-the typical range of quality is substantially norrower than is niche, unmediated markets. The mais challenge posed by content in social media sites is the fact that the distribution of quality has high variance: from very high-quality items to low-quality, sometimes abusive content. This makes the tasks of filtering and ranking in such systems more complex than in other domains. However, for information-retrieval tasks, social media systems present inherent advantages over traditional collections of documents: their rich structure offers more available data than in other domains. In addition to document content and link structure, social media exhibit a wide variety of user-to-document relation types, and user-to-user interactions.

In this paper we address the task of identifying highquality content in community-driven question/answering sites, exploring the benefits of having additional sources of infor-



## **Eugene Agichtein**

Emory University
Atlanta, USA

## Carlos Castillo Debora Donato Aris Gionis

Yahoo! Research Barcelona, Spain

### **Gilad Mishne**

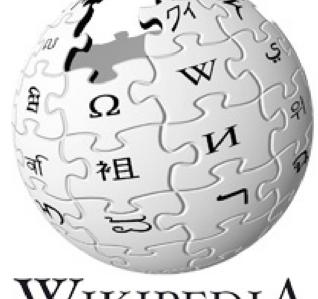
Yahoo! S&A Sciences Santa Clara, CA, USA

















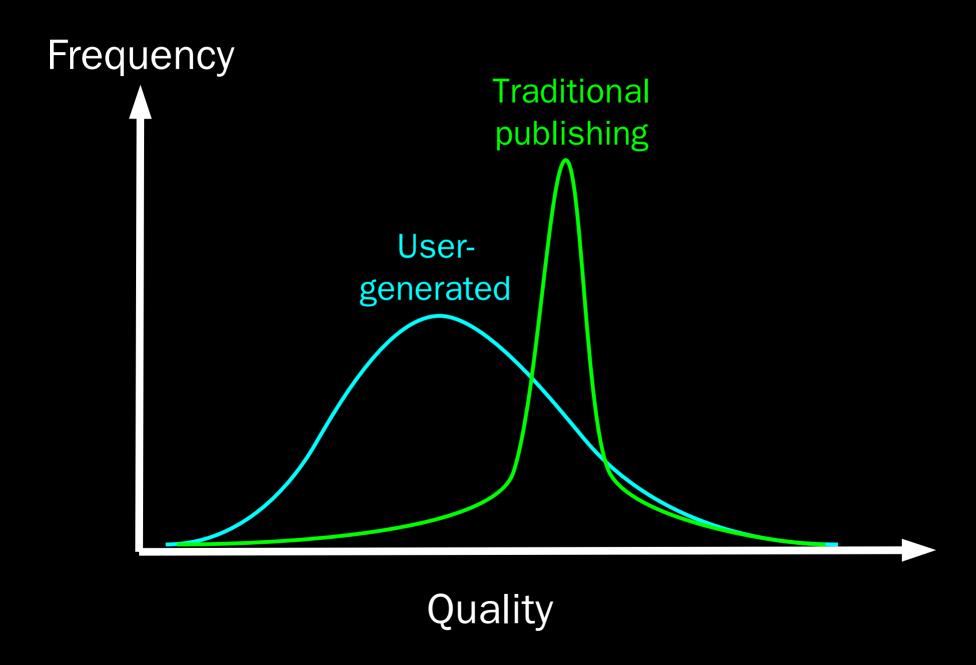
You Tube

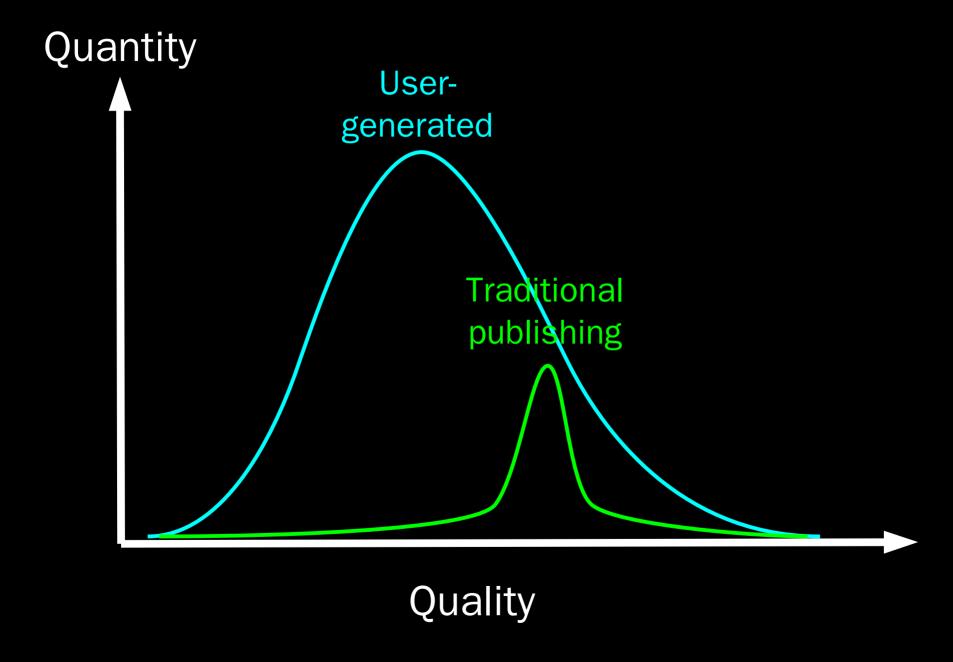
GARAGEBAND



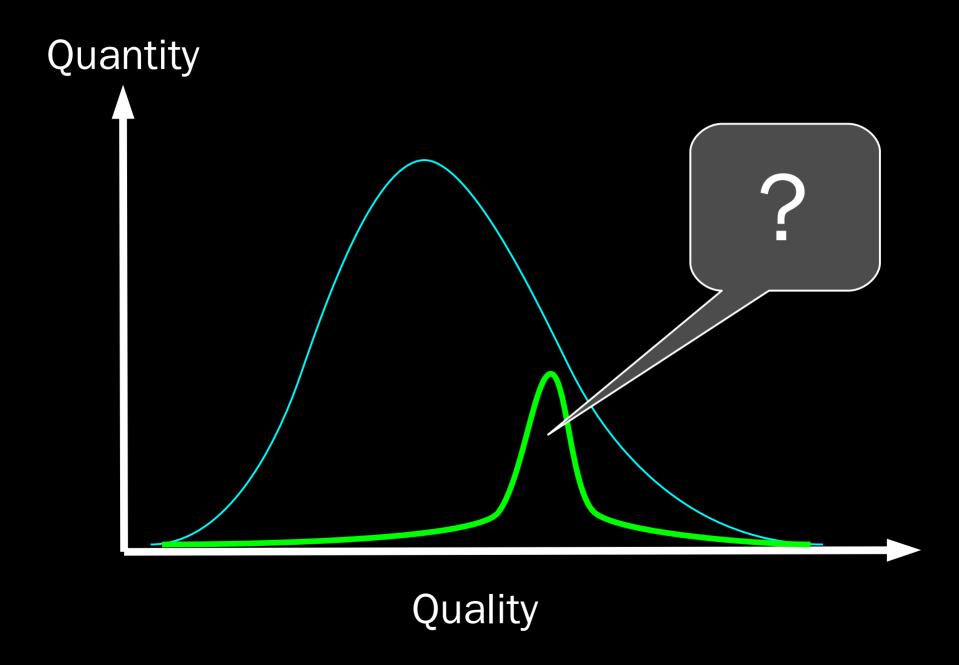
User-generated content  $\neq$  Traditional publishing



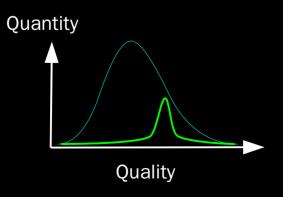








"We think it's all about quality over quantity now, because there's so much noise everywhere, there's no point in putting anything out unless it's fucking amazing."





## "There's still incredible merit in making albums."

How has new technology helped

I don't know if "democratizing" is the right word, but it's definitely levcling the playing field regarding who can hear what. Have you beard of this band? They have a song called "Cape Cod Kwassa Kwassa.

Vampire Weekend.

Yeah. What's amazing is anyone there in the world can hear three tracks within two minutes. s is to hear someone menig, and you can check it. coaning and grouning dustry, ultimately

noise everywhere, in putting anything or focking amazing. I still think incredible merit in making albupieces of work that are the same length as Beethoven's symphonies used to be. There's something pleasing about that amount of music from one artist.

What musicians do you ore leading us into the future?

Producers are like gold dust nowadays. Timbaland, Pharrell Williams and Dr. Dre, those are the people to watch.

Do you think music has a social aponsibility!

Our principal aim is to make that fifteen-year-old kid in Milwaukee who saved up to buy our record feel really great when he listens to it. If we manage to get Dick Chency impeached in the process, that's just a bonus.

What are the most important problems facing the world today?

Eventually, Curb Your Entirestarm. won't have any new seasons. I foresee. that being a problem.

Seriously!



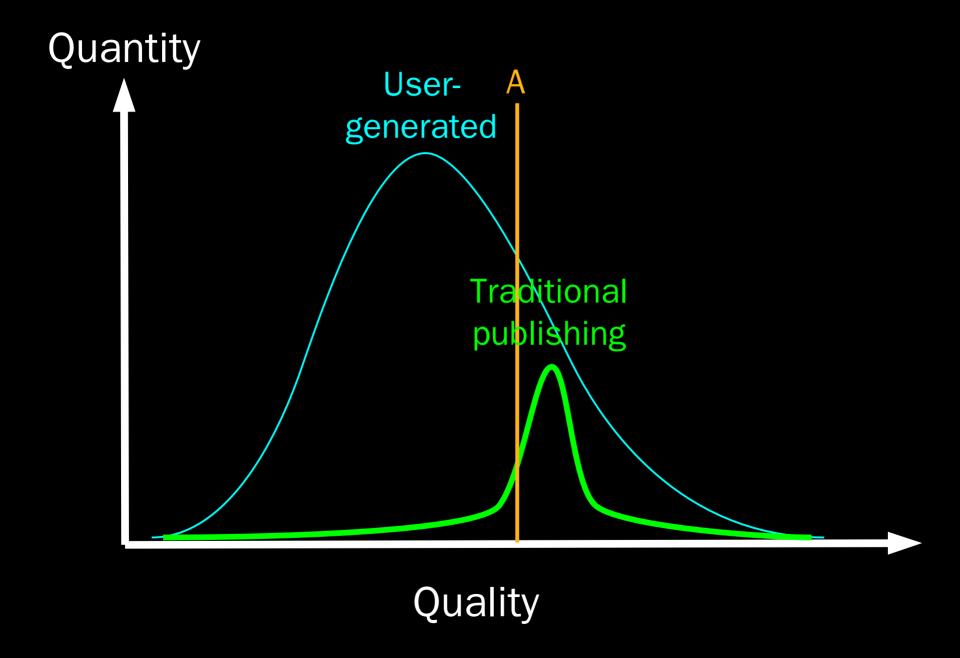
financial interests. That affects the environment, health, trade. It affects averything. Every night before bed, I have a little think, and I say, "Maybe tomorrow, someone's going to fix everything or something is going to happen that will make everything fixed. Maybe George and Karl will wake up tomorrow and completely change their minds." Maybe. Not implying that they sleep in the same bed.

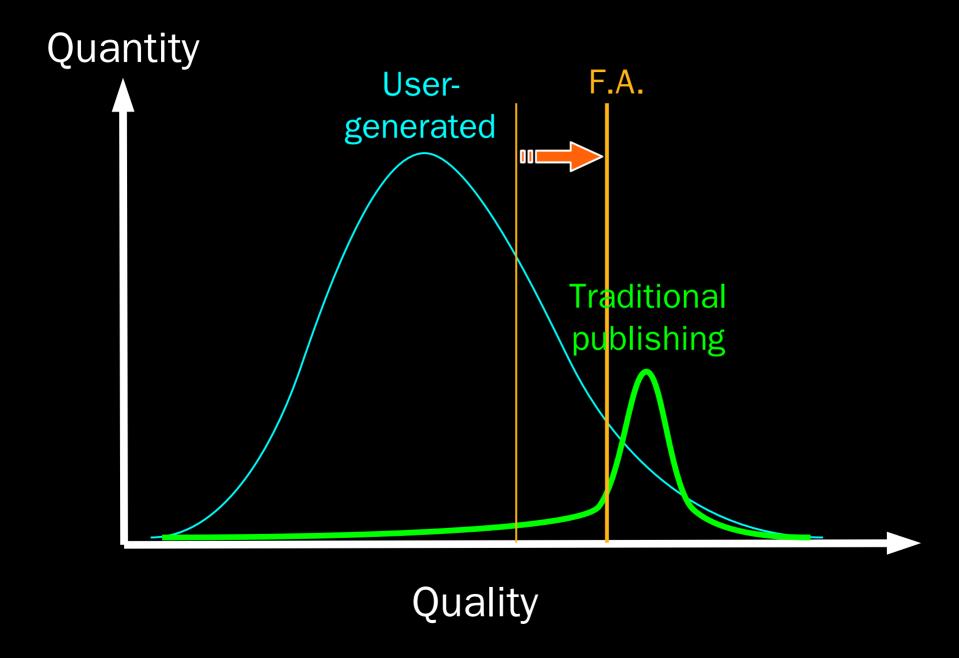
Who do you see having a positive impact on the future?

Bill Gates is a great example of the modern entrepreneur. He's made billions, then realized, "Oh, there's other things that are important too." That, to me, is very inspirational. Or Warren Buffett. I'm not enough of an authority on the ethics of his company, but I think it's very refreshing when you see the supersuccessful

Where do you see yourselves as a band in ten years, twenty years?

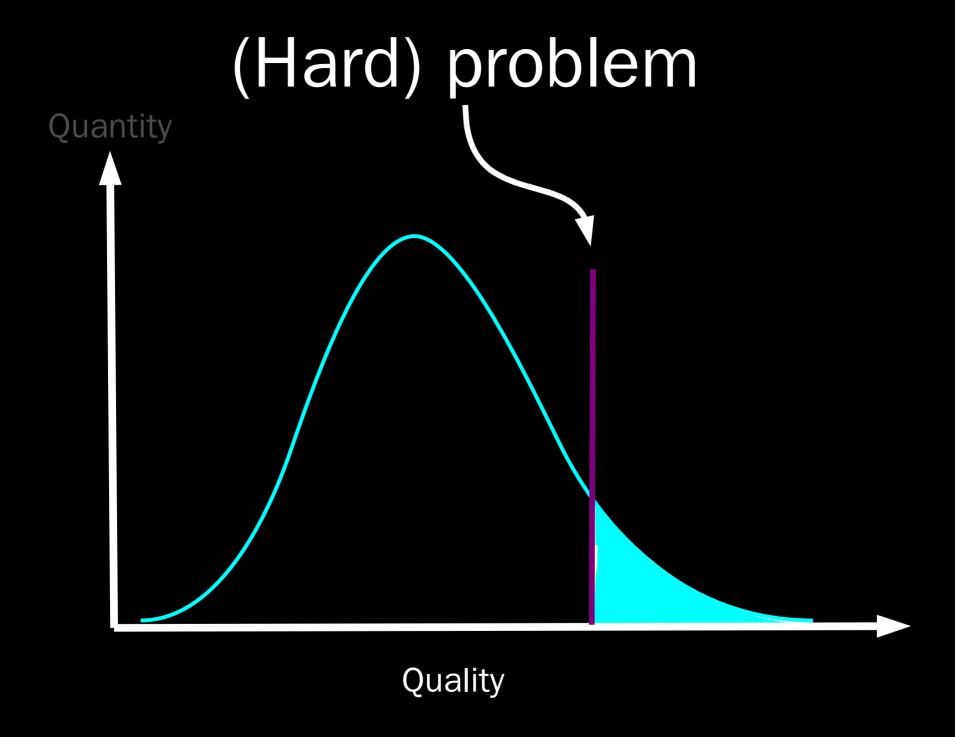
Won't it be the fiftieth anniversary of ROLLING STONE then? Hopefully I'll be talking to you, but we have a bell of a lot to do between now and then.















Welcome, chato [Sign Out, My Account]





Enter research question here:

What are the elements of social media that can be used to automatically discover high-quality content?

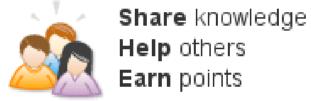
8 characters left

**Post Question** 

answer.



dis



What people think of Answers
How does it work?

Search for questions:

∽S∉arch

Α



Welcome, chato
[Sign Out, My Account]

Answers Home - Forum - Blog - Help

ask.



answer.



discover.

Search for questions:

Search

Advanced

My Profile

Home > Consumer Electronics > Land Phones > Resolved Question



ndvou

#### Resolved Question

Show me another »

## What's the best way to get telemarketers off my back?

i have caller id and usually don't answer. how can i get them to stop calling (i hear the donotcall registry doesn't work) and if i do pick up the phone aside from immediately hanging up what can i say to deter additional calls?

1 year ago

Report It



hrh grac...

#### Best Answer - Chosen by Asker

Register at the online do not call registry. Cell phones, business and home phones can be registered... You will still get some calls for about 30 days. Just tell anyone who calls in that time period that you are registered with the do not call registry and to please remove you from their calling list. If they give you any hassle advise them that you will file a report.

I had to do this too and every solicitor I spoke to was immediately ready to get off the phone and apologized quickly. Keep a log next to your phone for the first 30 days and file it with your phone bill after that. (You will then have a



Hello **ChaTo** Total Points 340 Level 2

#### Categories

- > All Categories
- Consumer Electronics
- Camcorders
- Cameras
- Cell Phones & Plans
- · Games & Gear
- · Home Theater

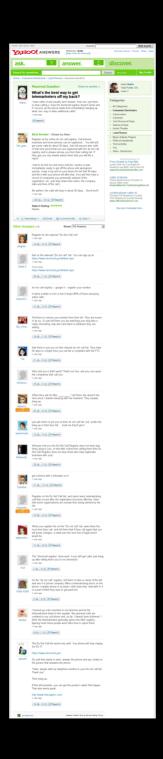
#### » Land Phones

- · Music & Music Players
- PDAs & Handhelds
- TiVO & DVRs
- TVs
- · Other Electronics

SPONSOR RESULTS

Free Grants to Pay Bills

Learn How You Can Apply for Grants to pay Bills. Get a Free Kit. www.thousanddollarprofits.com Best answer
Picked by votes
-orPicked by asker



## Question

+ "Stars"

## **All answers**

- + "Thumbs up"
- + "Thumbs down"

# 1/4 questions want an opinion: informal polls

3/4 questions seek for information or advice



#### Whats so great about life?

3 dr In Polls & Surveys - Asked by smashing pumpkins - 13 answers - 5 hours ago



#### Inviting guests + 1?



#### Doggie Clique???

n Dogs - Asked by Michelle - 7 answers - 5 hours ago



#### Where can I buy open toe panty hose?

In <u>Fashion & Accessories</u> - Asked by <u>THE CATWHISPERER</u> - 3 answers - 5 hours ago



#### Why is it People are Offended by the Silliest Things?

2 In General - Cultures & Groups - Asked by Californiagirl - 7 answers - 5 hours



#### How do i know if i am depressed or if i have an anxiety dissorder?

1 😭 In Psychology - Asked by Alejo - 5 answers - 5 hours ago



#### Can u think of a well paid job (not vet,doctor, or lawyer) thanks?

In <u>General - Careers & Employment</u> - Asked by <u>Verdun M</u> - 28 answers - 5 hours
ago



#### J.j. hardy...how will he do in 2008?

☆ In Baseball - Asked by Bailey - 3 answers - 5 hours ago





### **Resolved Question**

Show me another »

## Do girls like computer geeks / nerds?

kieran.b...

2 weeks ago





tabitha c

not really

2 weeks ago





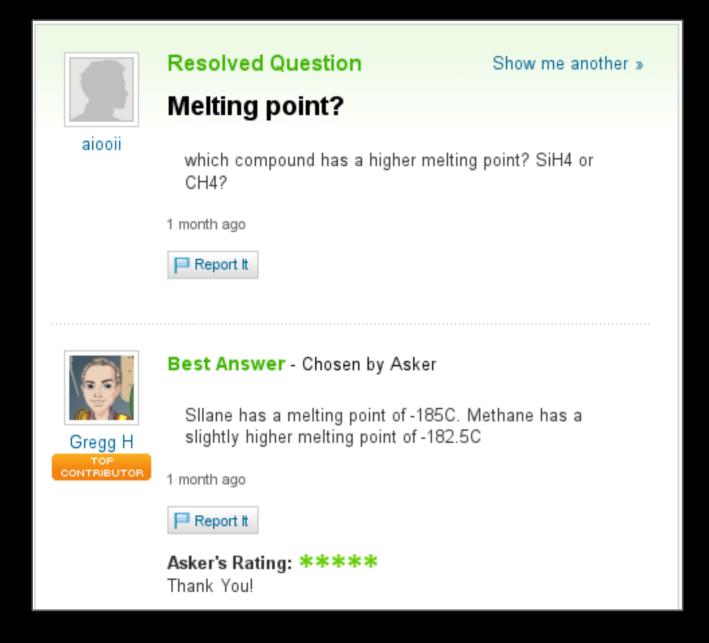
Ella G

a little geekiness is endearing, as long as they still have social skills and good personal hygiene!

2 weeks ago



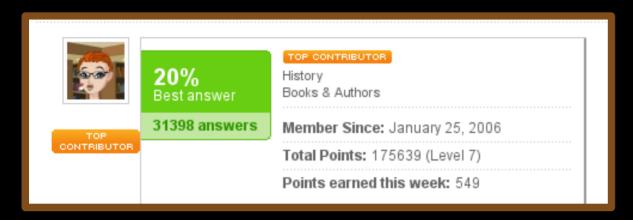
Q. Su, D. Pavlov, J.-H. Chow, W. C. Baker. "Internet-scale collection of human-reviewed data". WWW'07.



17%-45% of answers were correct

65%-90% of questions had at least one correct answer

## There are top contributors ...



## ... but they don't have all the answers



# Task: find high-quality items





# Existing tools

Link-based ranking methods

Propagation of trust/distrust

Automatic text analysis

Usage mining



## Sources of information

Content analysis

Usage data (clicks)

Community ratings



## Sources of information

Content analysis

(with errors)

Clicks

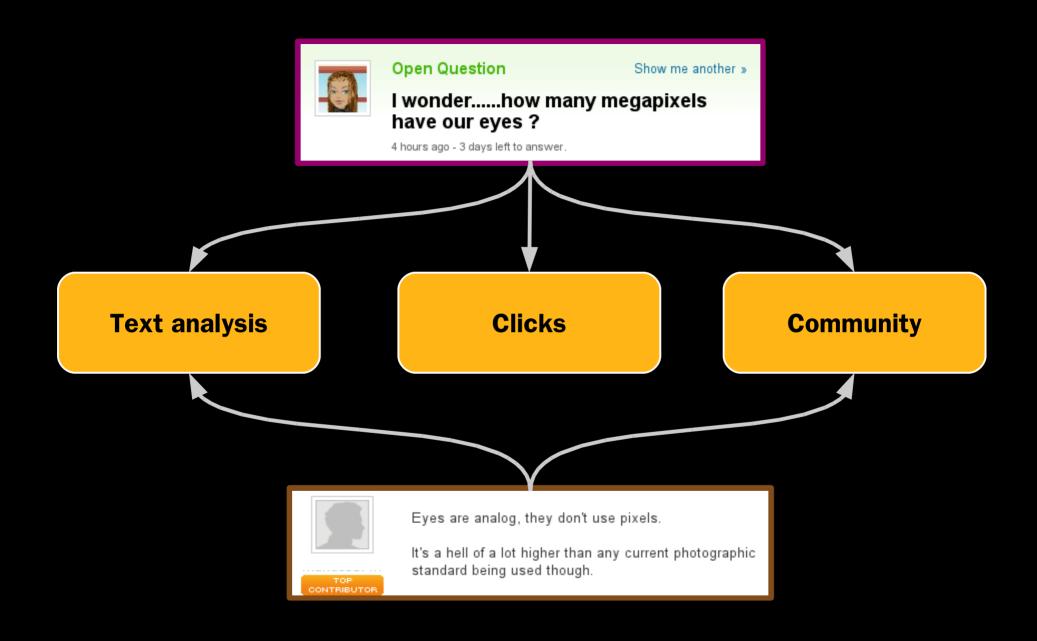
(with noise)

Community ratings

(sparse, with spam)





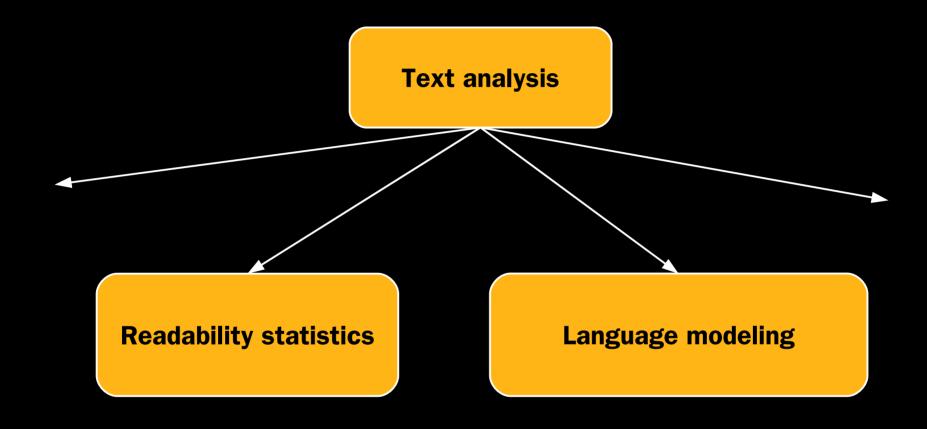


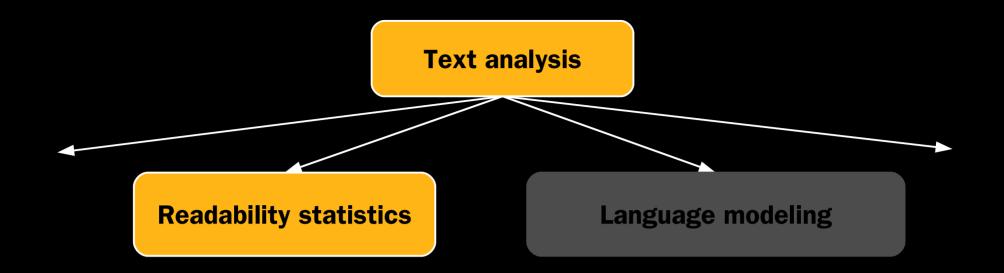
**Text analysis** 

Clicks

**Community** 







**Punctuation density** 



### Help! math! histogram! asap?

In Mathematics - Asked by Markyme 123 - 0 answers - 3 minutes ago

Capitalization errors



#### WHAT is heidi montag thinking WITH THIS MUSIC VIDEO?

😭 In <u>Celebrities</u> - Asked by <u>chrls\_bann88</u> - O answers - 3 minutes ago

Number of words

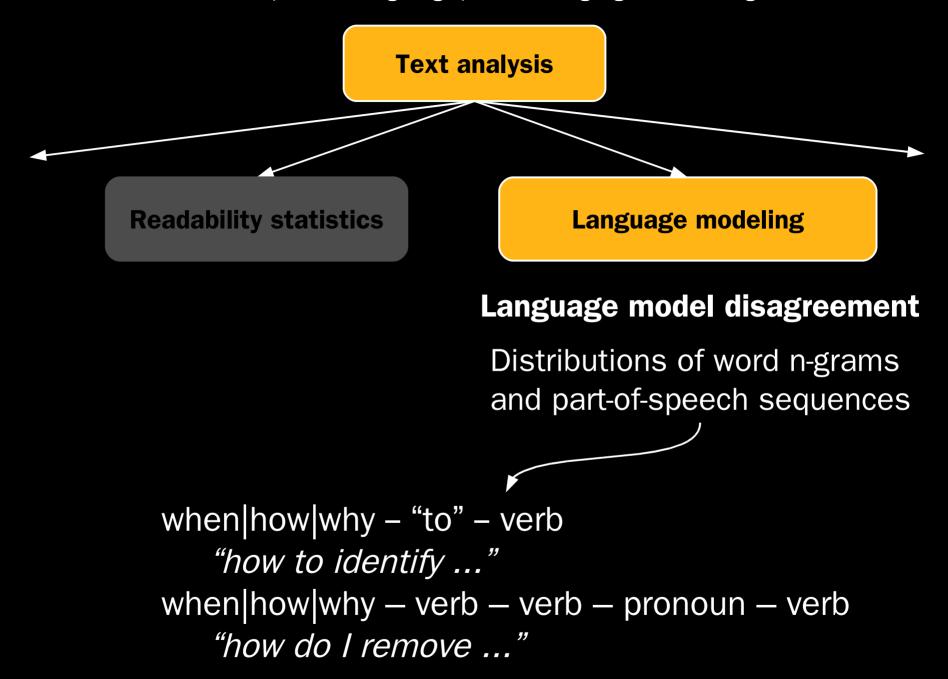


#### Help!!!!!!!!!!?

😭 In <u>General</u> - Asked by <u>\*So Confused\*</u> - 1 answer - 6 minutes ago

+ spacing density, sylablles per word,...





**Text analysis** 

**Clicks** 

**C**ommunity

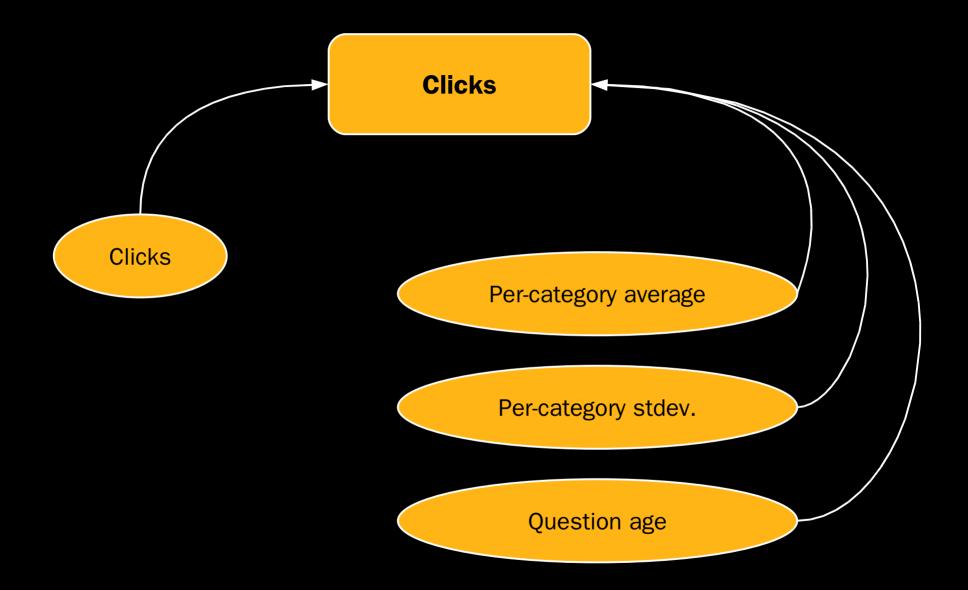


Clicks

If we know that a question is clicked **100** times, and another question is clicked **10,000** times ...

... we still know nothing





**Text analysis** 

Clicks

**Community** 



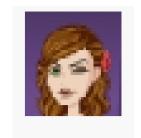


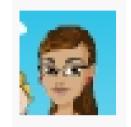


























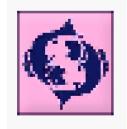


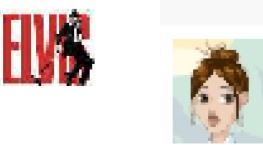


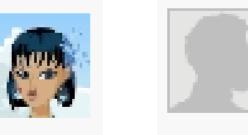












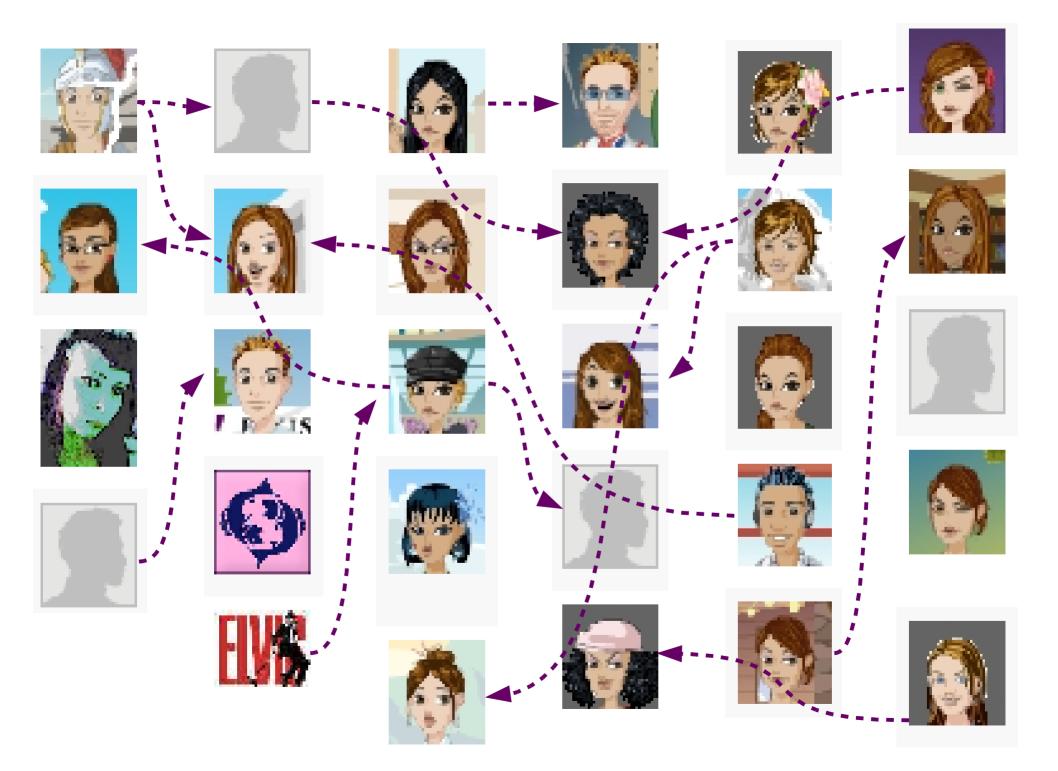


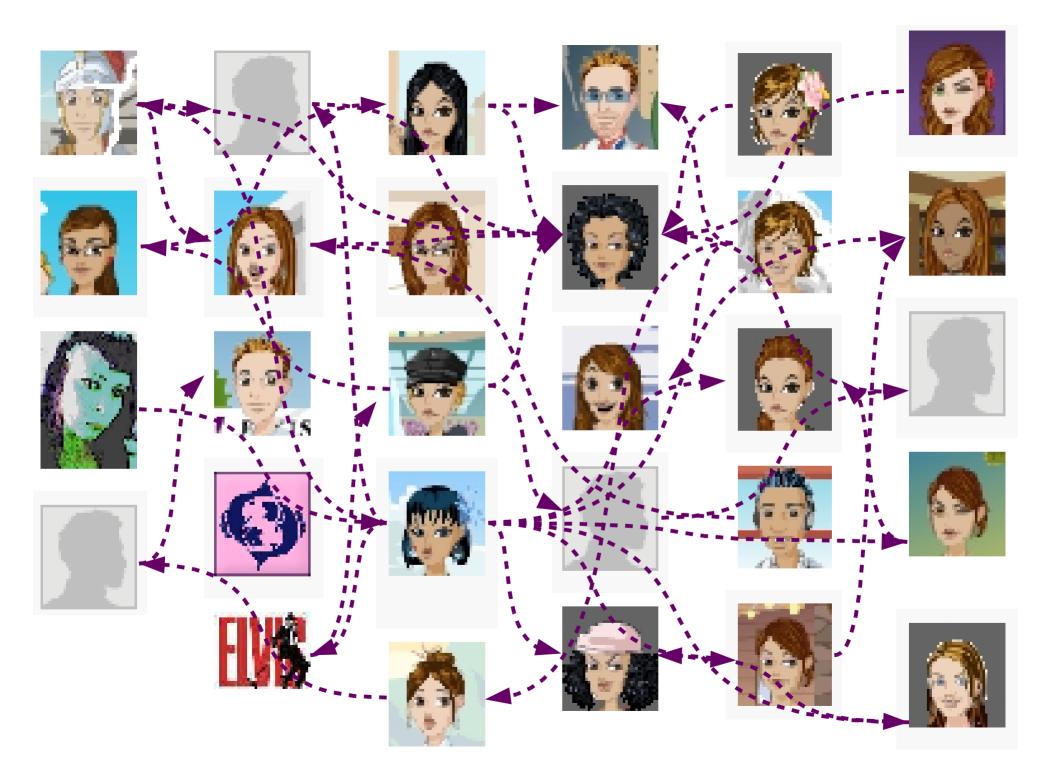




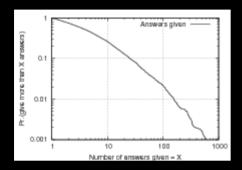


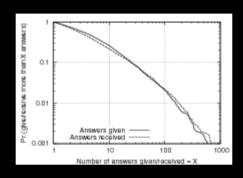


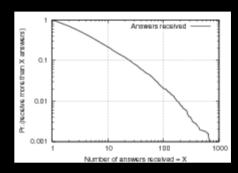


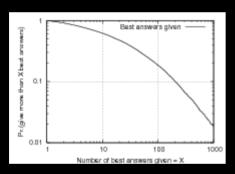


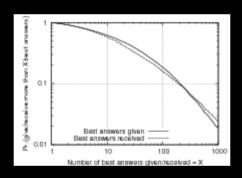
# Power laws

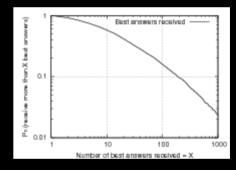


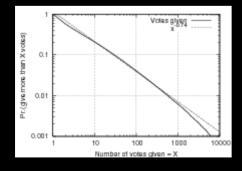


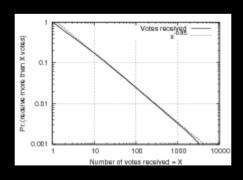


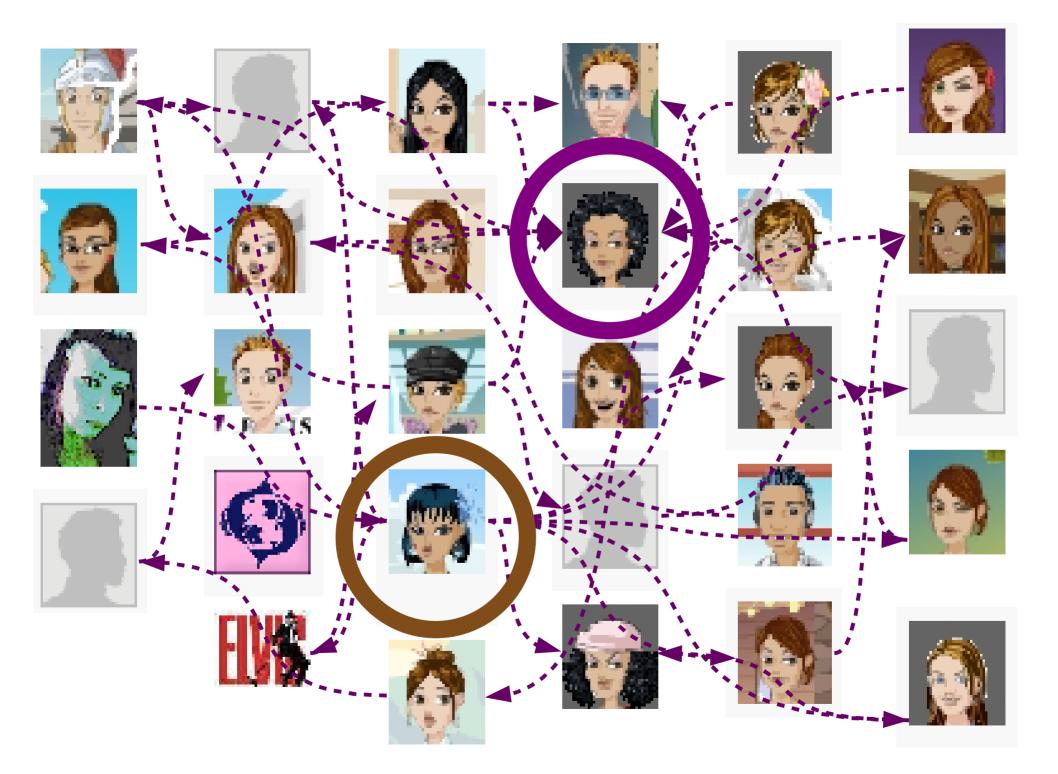




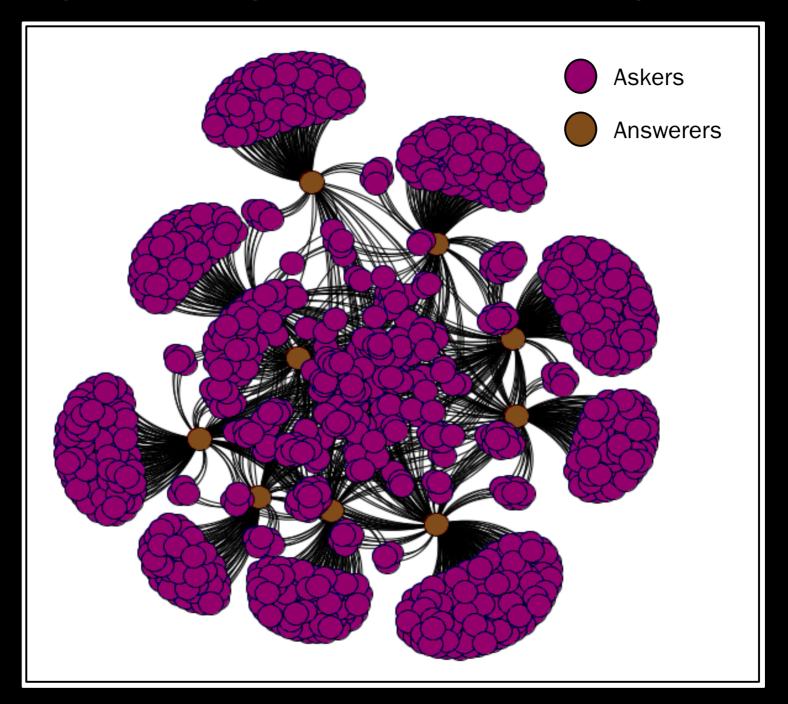




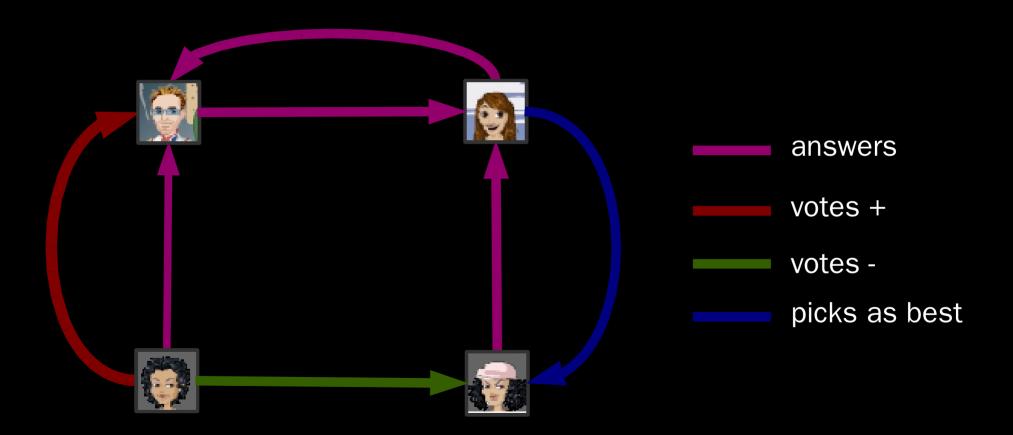




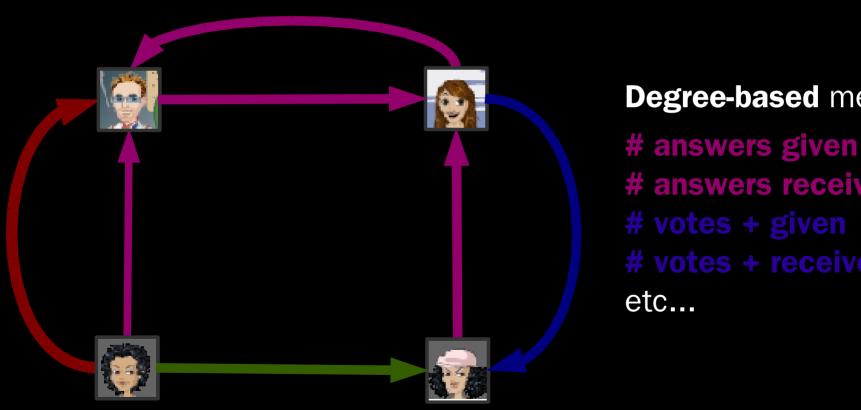
P. Jurczyk, E. Agichtein: "Discovering authorities in Q.A. communities by using link analysis" CIKM'07



**Community** 



#### **Community**

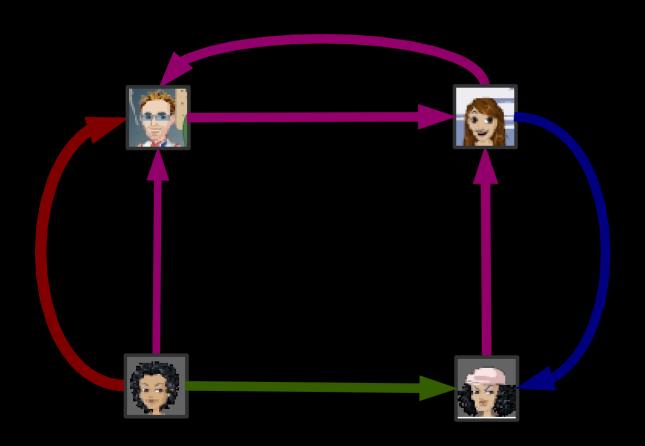


#### **Degree-based** metrics

```
# answers received
```



#### **Community**

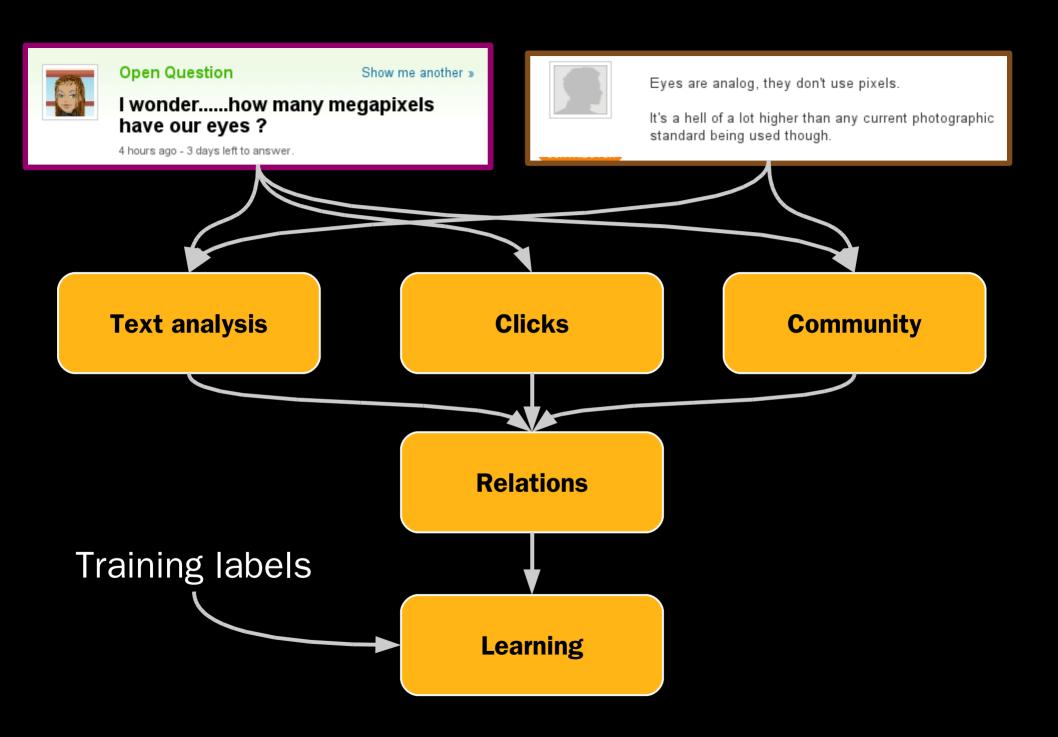


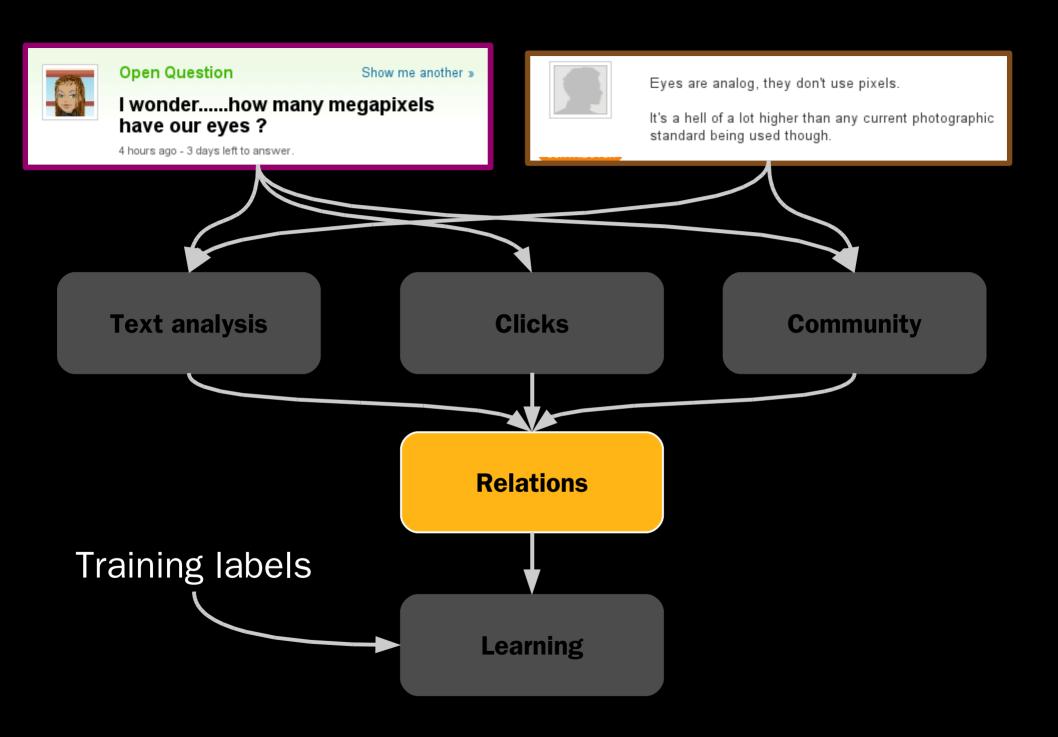
#### **Propagation-based** metrics

- 1. Pagerank score
- 2. HITS hub score
- 3. HITS authority score

Computed on each graph







**Answer** quality

	High	Medium	Low
High		<b>15</b> %	
Medium		<b>76</b> %	
Low		9%	
		100%	



Answer quality

	High	Medium	Low
High		15%	8%
Medium		76%	<b>74</b> %
Low		9%	18%
		100%	100%

Answer quality

	High	Medium	Low
High	<b>41</b> %	15%	8%
Medium	<b>53</b> %	76%	74%
Low	6%	9%	18%
	100%	100%	100%

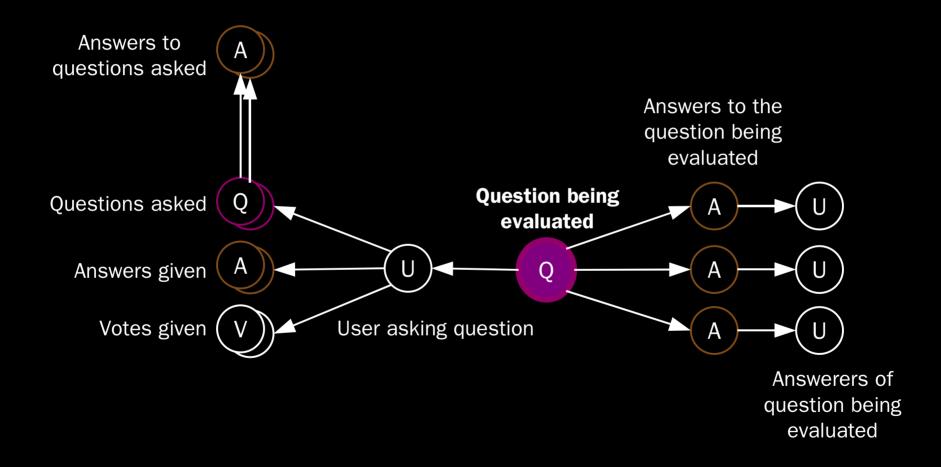
Answer quality

	High	Medium	Low
High	41%	15%	8%
Medium	53%	76%	74%
Low	6%	9%	18%
	100%	100%	100%

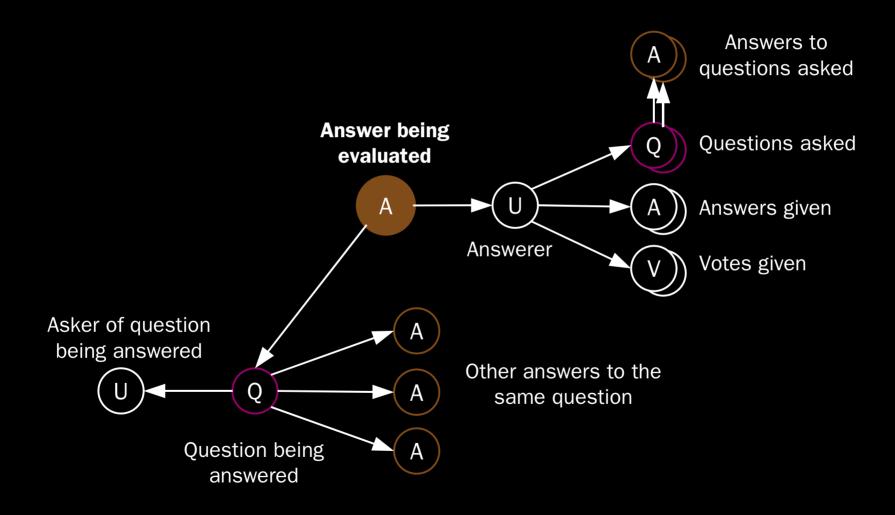
Question quality and answer quality are not independent

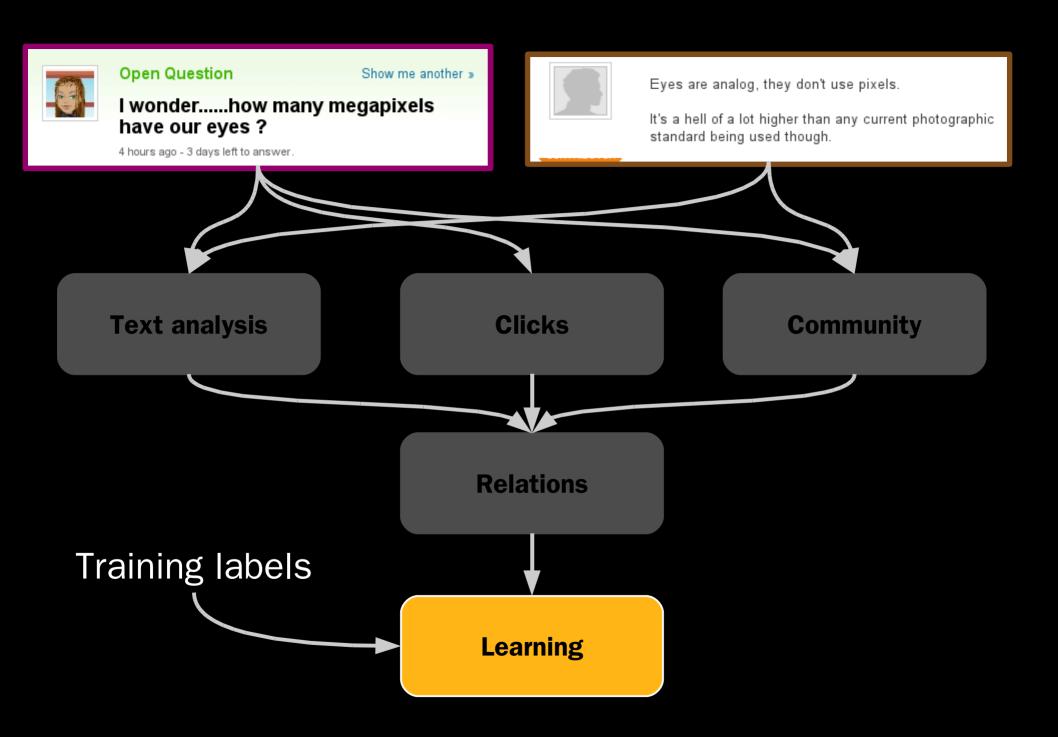


#### **Relations: questions**

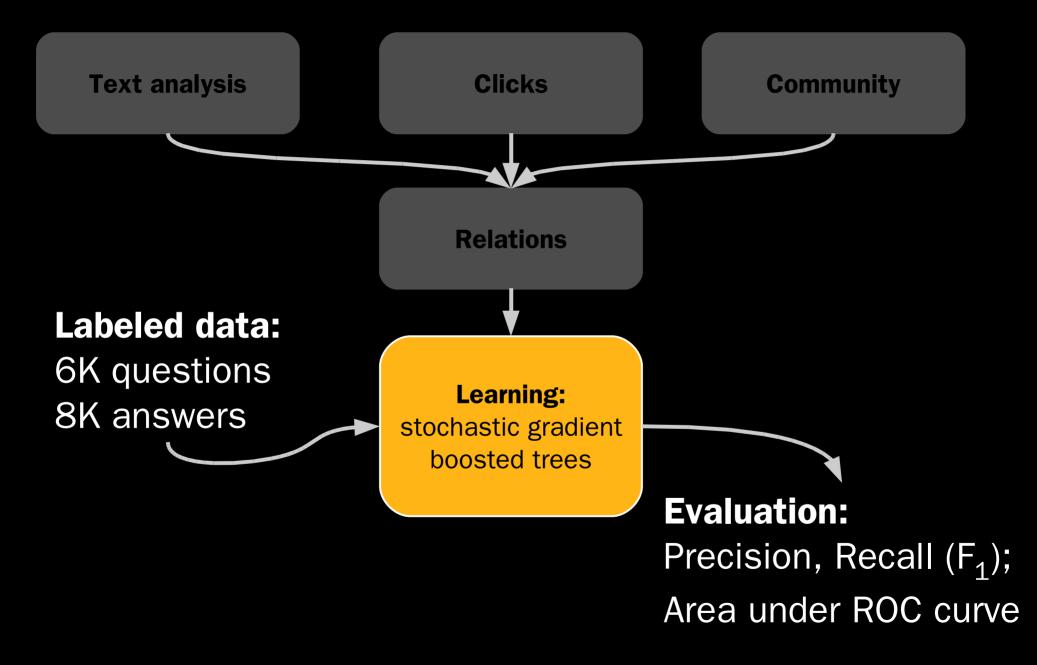


#### **Relations: answers**





J. H. Friedman: "Stochastic gradient boosting". Comp. Stat. Data. Anal., 38(4), 367-378, 2002.



# Task: high-quality questions

	Precision	Recall	AUC
N-grams (N)	65%	48%	0.52
N+ text analysis	76%	65%	0.65
N+ clicks	68%	57%	0.58
N+ relations	74%	65%	0.66
All	79%	<b>77</b> %	0.76



# Task: high-quality answers

	Precision	Recall	AUC
N-grams (N)	67%	86%	0.81
N + text analysis	<b>71</b> %	93%	0.88
N + clicks	-	-	-
N + relations	69%	85%	0.82
All	73%	91%	0.87



# In the paper ...

Framework for quality estimation in social media

Graph-based model of contributor relationships

Details on the relative importance of (sets of) features



## What did we learn?

Human assessments for this task

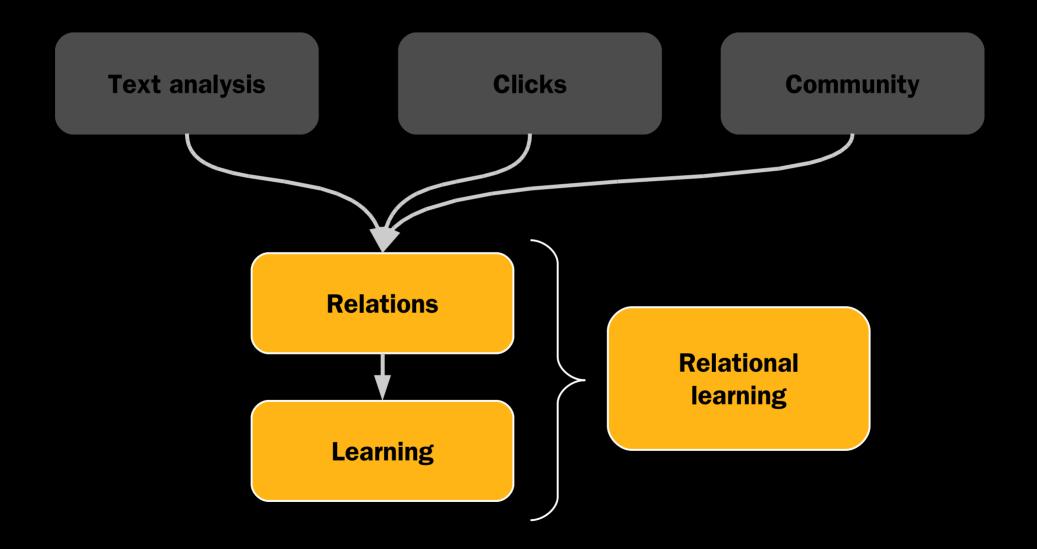
... have relatively low agreement Classifying questions/answers

... is substantially different from document classification

Look at orthogonal feature spaces



## Future work





#### Finding High-Quality Content in Social Media

The forest business for the business pearle viriantitud and Englishmen Additional 

Columbia and Balance Transference

CONTRACTOR DE CONTRACTOR CONTRACTOR CO.

and the state of the same Philadelphia and Alberta 

Filtrage State and State Services

Access resistance sides the state of the s College of the Colleg

1. Beckelennen an die ernem mit

E. C. Control of Control

Maria de Caracteria de Caracte Strong Strong

The state of the s

NRCPACTURE AND ADDRESS AND Man State Name Attended to

Submittee and deal below

E-Mindle - Tool Street

TO BE SEED OF THE PARTY OF THE

Approximation of the second All places. The March of Section with of the Mills and the Article of the Section of the Mills and Mills a

Same U. do at off taken Burk-flow-warm Lawrence and state of the

products and a point by an analysis of the

to an admitted of the published between the

man relative to the state of ----

Angelin harmanist drops

\$1400 beautiful and the second the filtre array of the other.

all rights and distributions by

Section of the second factor to the second factor of the second factor o

The effect of the term of the electric term of the effect of the effect

Margin Systems and Control

ld Greifenbalmintenret

Establishment and the According to the according to

the standard or harden of a gift a stayler to a

Automorphism and anti-attacks.





deltament of the

The state of the same of the same in

The delivery

Service description

byoth translation to the term

Annual Section 41 hard death and the state from

Tries and department



Barrier Springer

The parties are sense, at the parties of the partie

Marie Lande

and the state of t



See 1 See divine only to be taken

the special constraint of the first of the

Secretary and Company Management of the control of the con 222

The state of the s

Charles and Called Annual Control

#RESERVATION AND THE PARTY OF T

to the season production of the season of th transportation of the second equipment in the second

After Street, statute

Particular of the second of th

Indicated Tables on Manager about

Participants of the state of th Many 2003 adoption behalf of



and the state of the state of the The state of the state of

Amphibathan at the





Service State of parties of control of a

halither of a feel through the fit. Birth Court State and American

and the probability of the control o

COMMENSATION OF THE PARTY OF TH min facts and the state of the

El Gentles goding

the control of the best of the control of the contr



Note to the other transfer and the same

English and the base based at

A STATE OF THE PARTY OF THE PAR All Controls of the control of the c hope and an I have been

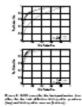
Andreas State Stat the state of the s Cities believe in County

(I have a county of the particular in the county of the coun The second secon

200 girl by weight by the day

Company of Street or Street Proceedings of the process of the contract of

AND THE PROPERTY AND THE



Portsout tracellos tab.

C 10 the C New State of State

Harris Charles Tel Charles Balance S. Bradelin, adjoint Englishment S. Bradelin adjoint Charles Balance Security of a Thomas of the Santon Security of Santon on State States

Character to be a second The part of the last the last

Continued in the continue of the late of the continue of the late of the continue of the late of the l

Today is not invited the

L. Charles of the Cha and the state of t

All the six and if the time, after the service community to a state for other time and the state of the service community to the state of the state

Control State Stat

The desired of the control of the test in the test in the second of the control o

### Thank you!







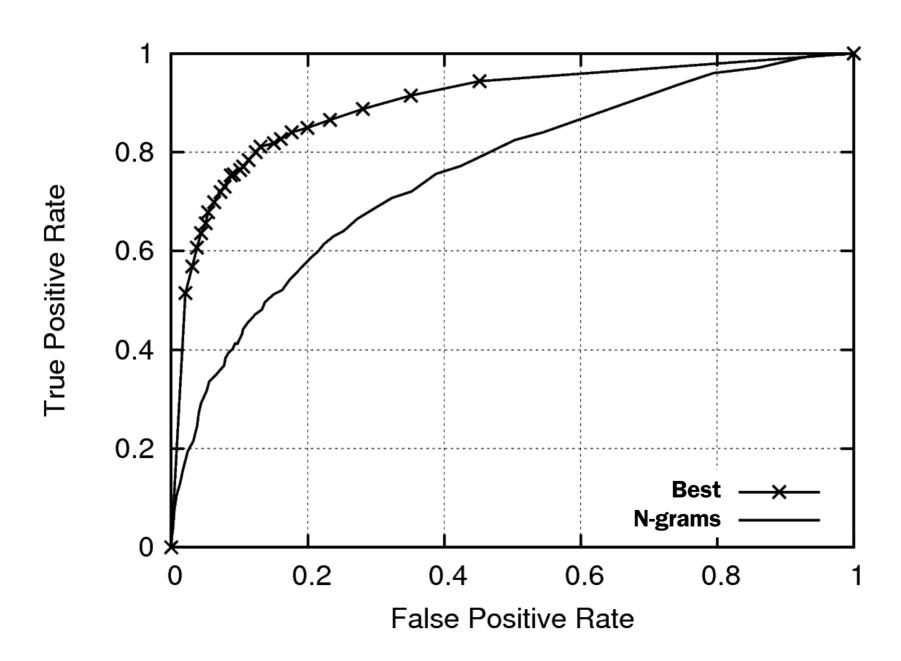
ask.

answer. discover.

YAHOO! ANSWERS



## ROC curve: high-quality questions



## ROC curve: high-quality answers

