

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 available, 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 automatically 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 available 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WSDM'08, February 11–12, 2008, Palo Alto, California, USA.
Copyright 2008 ACM 978-1-59593-927-9/08/0002...\$5.00.

1. INTRODUCTION

Recent years have seen a transformation in the type of content available on the web. During the first decade of the web's prominence—from the early 1990s 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. Popular user-generated content (or social media) domains include blogs and web forums, social bookmarking sites, photo and video 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 alternative channel for obtaining information on the web: rather than browsing results of search engines, users present detailed information needs—and get direct responses authored by humans. In some markets, 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 narrower than in niche, unmediated markets. The main 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 high-quality 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*



GARAGEBAND

blinkx

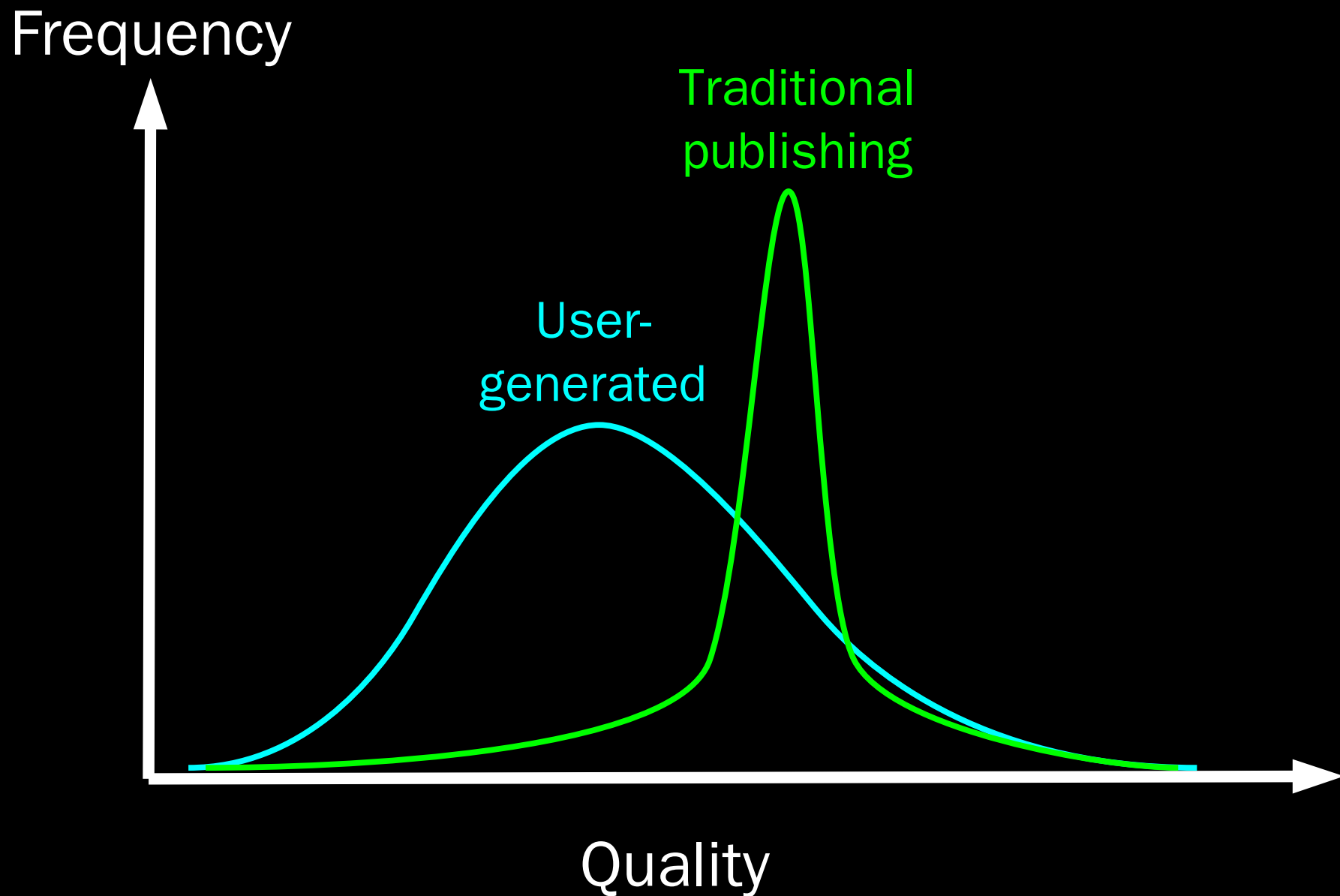


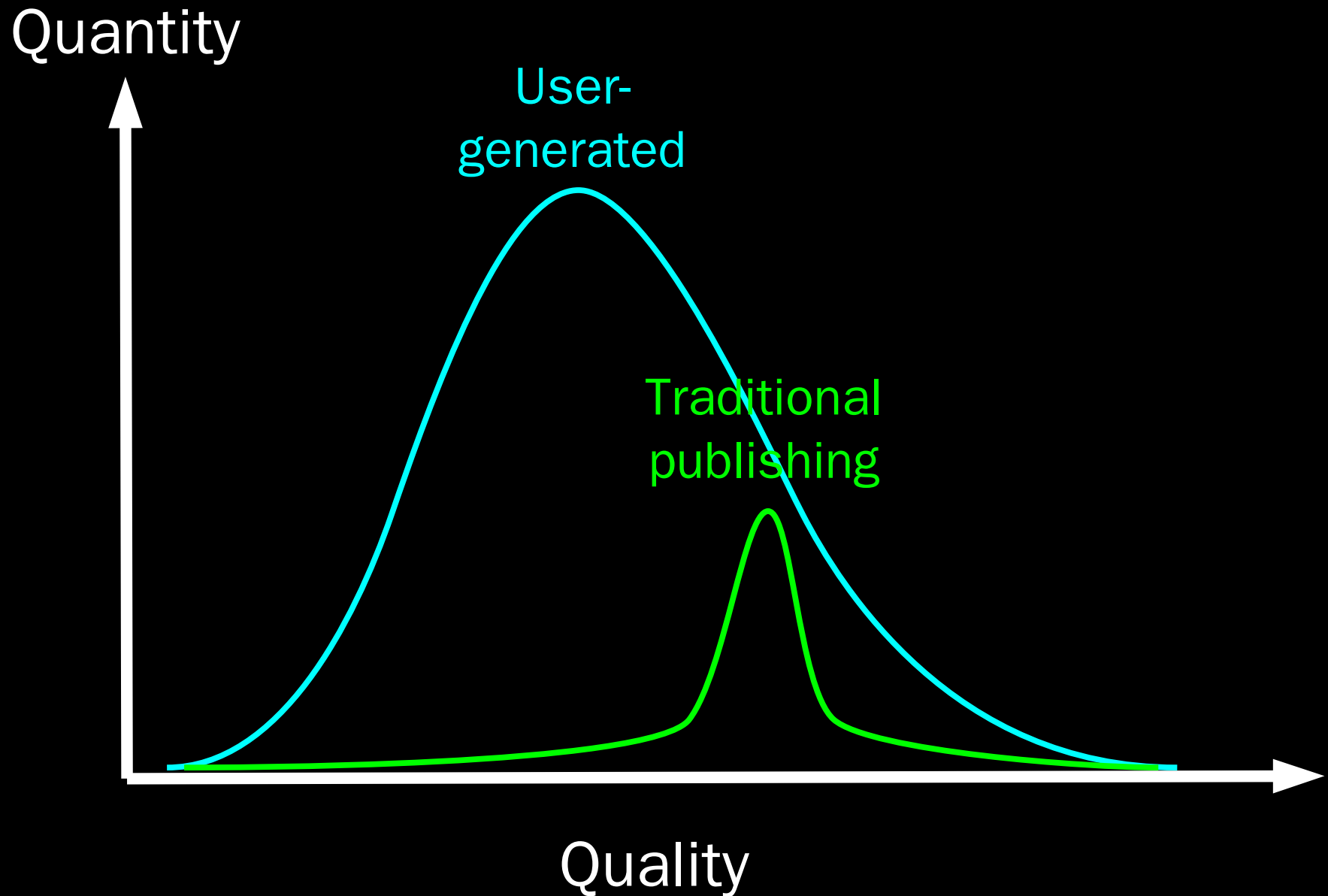
flickr BETA TM

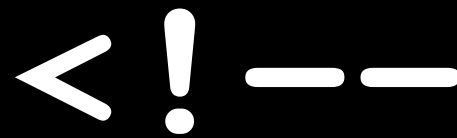
WIKIPEDIA



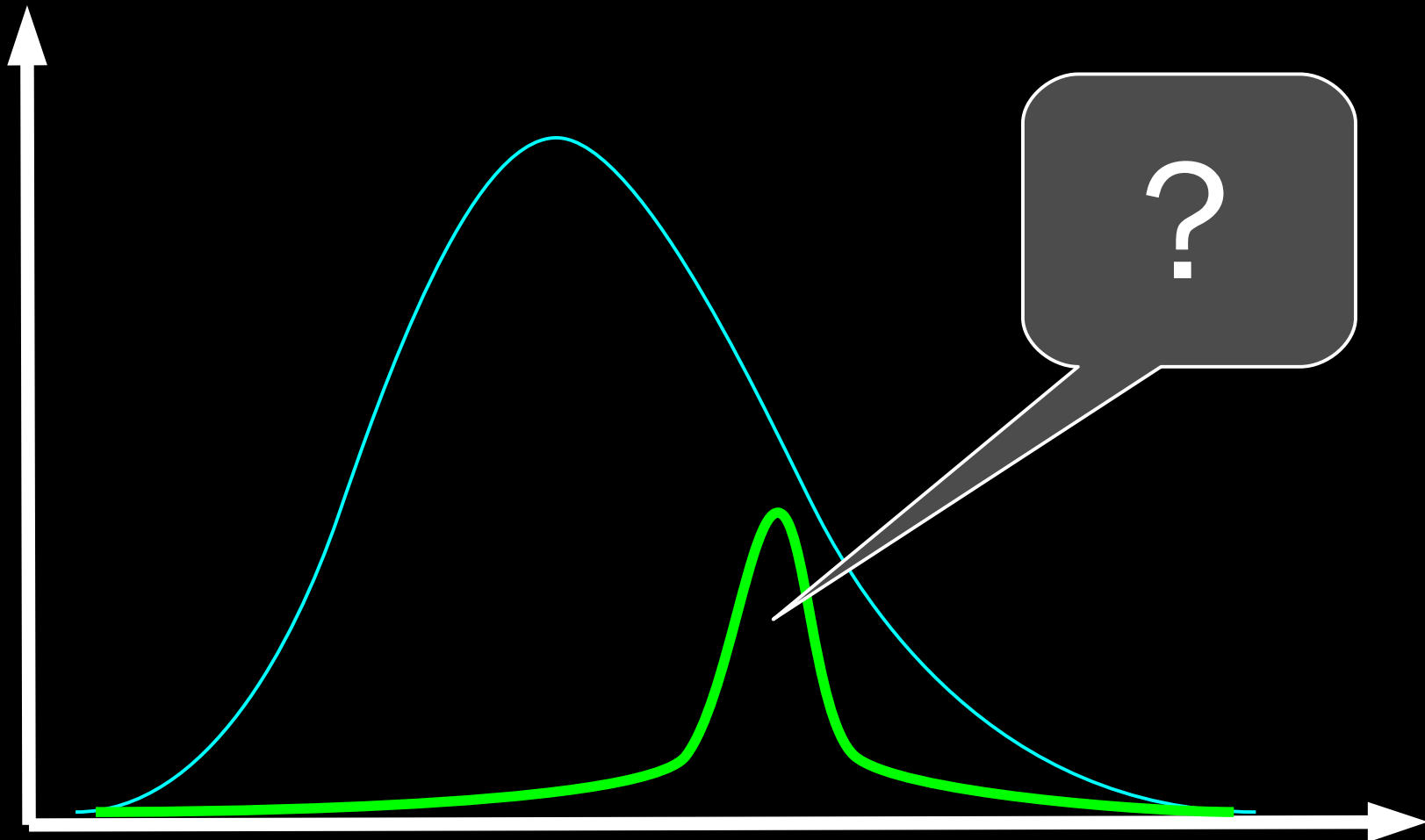
User-generated content \neq Traditional publishing







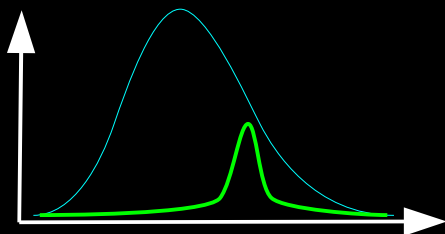
Quantity



Quality

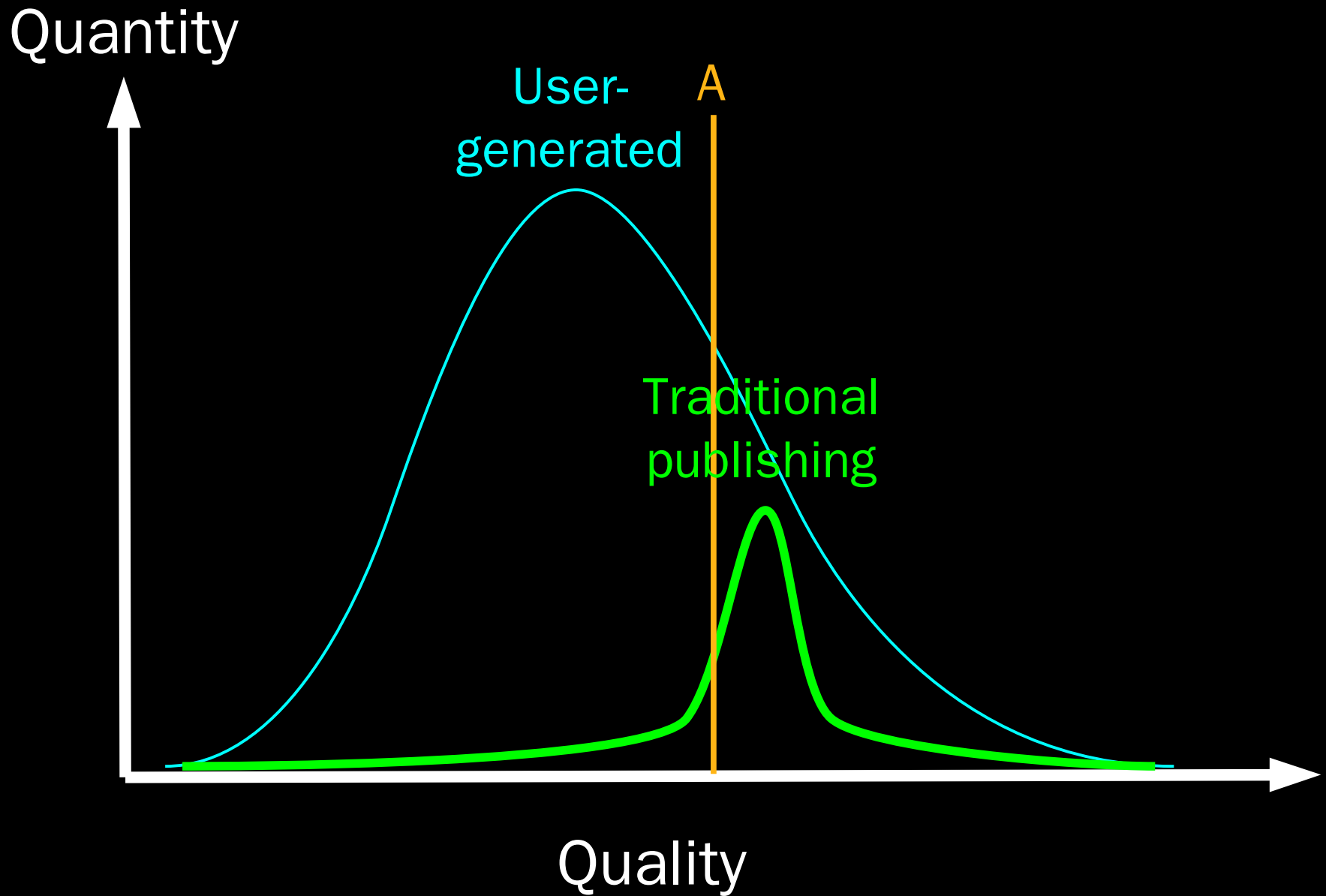
“We think it's all about quality over quantity now, because there's so much noise everywhere, there's no point in putting anything out unless it's fucking amazing.”

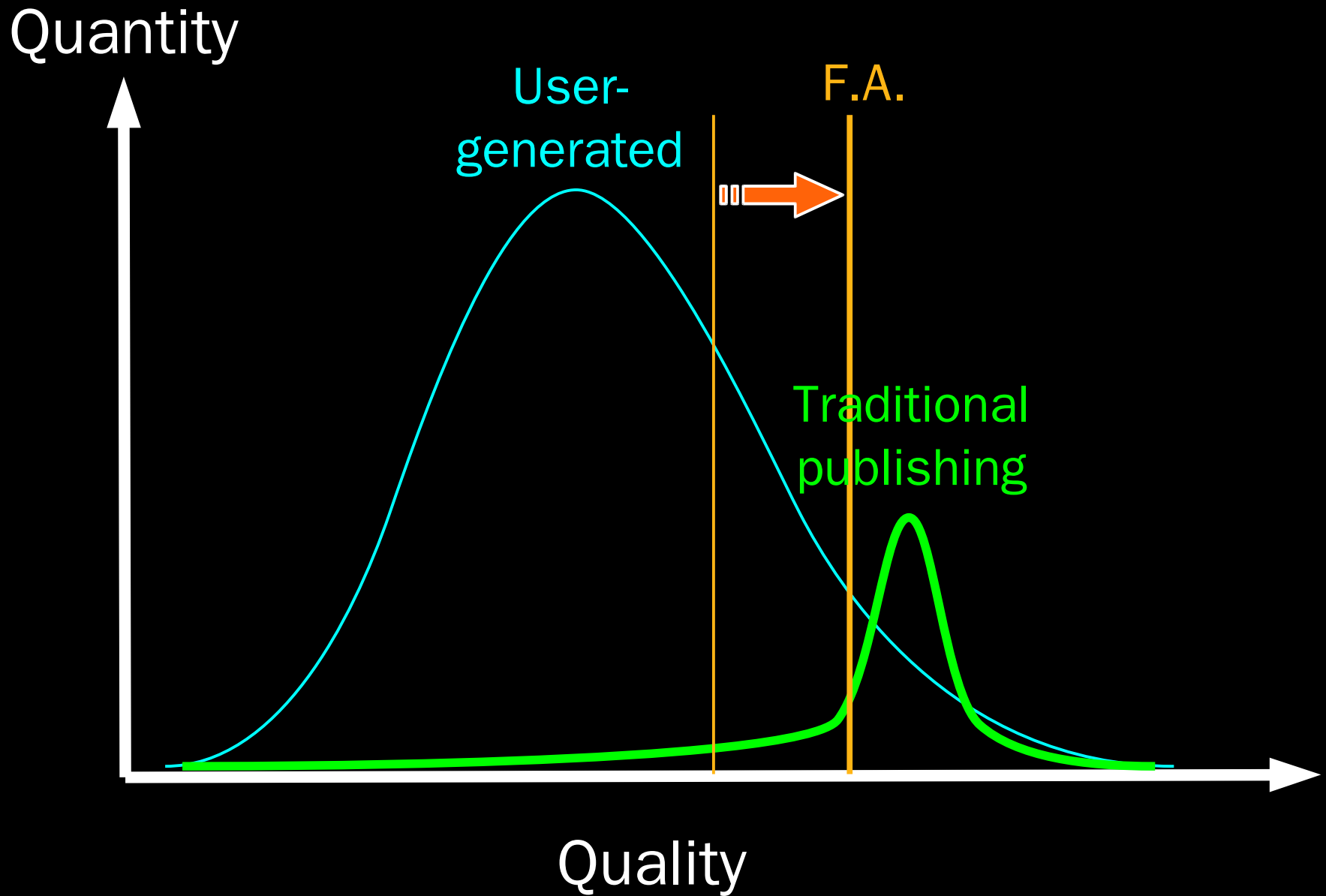
Quantity



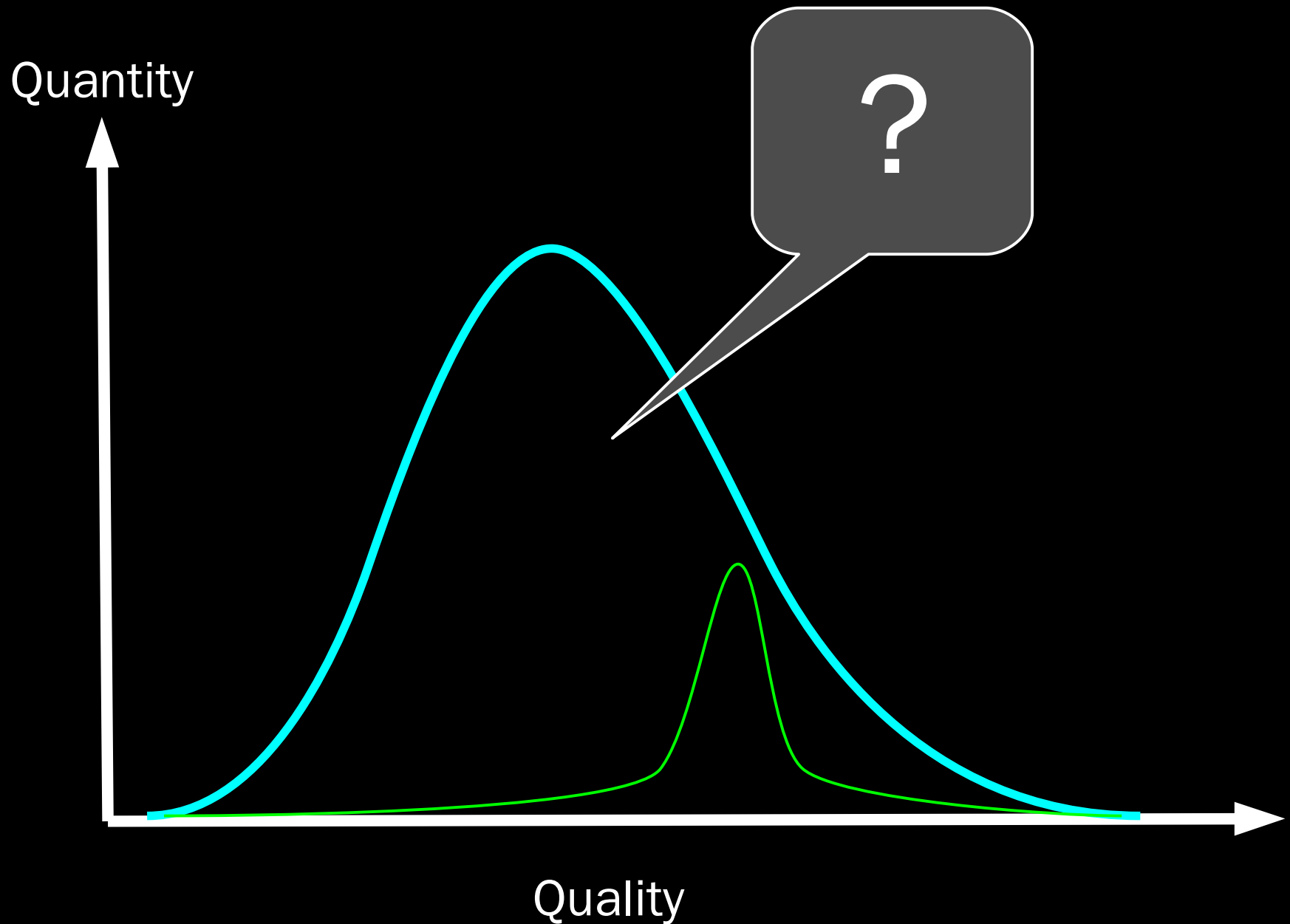
Quality











(Hard) problem



YAHOO! ANSWERS

Welcome, **chato**
[Sign Out, My Account]

ask.



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.



Share knowledge
Help others
Earn points

What people think of Answers
How does it work?

dis

Search for questions:

Search



ask.



answer.



discover.

Search for questions:

Search

Advanced

My Profile

[Home](#) > [Consumer Electronics](#) > [Land Phones](#) > Resolved Question



ndyou

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
-or-
Picked by asker

The screenshot shows a Yahoo! Answers page. At the top, there are navigation links for 'ask', 'answer', and 'discover'. The main question is 'What's the best way to get telemarketers off my back?'. Below the question, there are several answers from different users, each with a rating (e.g., 1/5, 2/5, 3/5, 4/5, 5/5) and a 'Best Answer' badge. The answers provide various tips and advice on how to handle telemarketers, such as registering with the Do Not Call list, reporting to the FTC, and using caller ID.

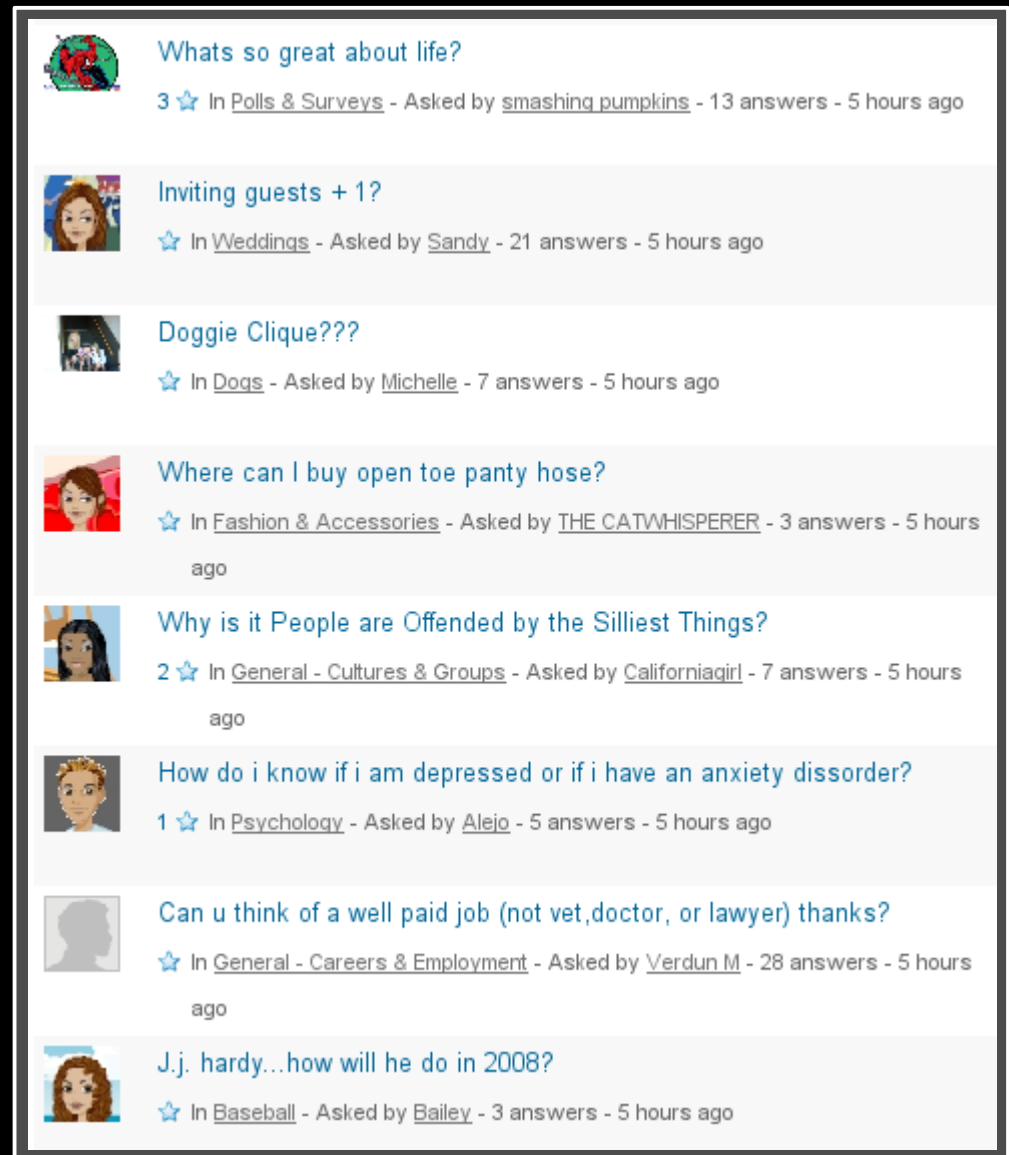
Question
+ "Stars"

All answers
+ "Thumbs up"
+ "Thumbs down"



$\frac{1}{4}$ questions want an **opinion: informal polls**

$\frac{3}{4}$ questions seek for **information or advice**



The screenshot shows a vertical list of eight questions from a social media platform. Each question entry includes a small profile picture on the left, the question text in blue, a star icon, the category name, the asker's name, the number of answers, and the time since asked. The questions are: 1. 'Whats so great about life?' (3 stars, Polls & Surveys, asked by smashing pumpkins, 13 answers, 5 hours ago). 2. 'Inviting guests + 1?' (1 star, Weddings, asked by Sandy, 21 answers, 5 hours ago). 3. 'Doggie Clique???' (1 star, Dogs, asked by Michelle, 7 answers, 5 hours ago). 4. 'Where can I buy open toe panty hose?' (1 star, Fashion & Accessories, asked by THE CATWHISPERER, 3 answers, 5 hours ago). 5. 'Why is it People are Offended by the Silliest Things?' (2 stars, General - Cultures & Groups, asked by Californiagirl, 7 answers, 5 hours ago). 6. 'How do i know if i am depressed or if i have an anxiety disorder?' (1 star, Psychology, asked by Alejo, 5 answers, 5 hours ago). 7. 'Can u think of a well paid job (not vet,doctor, or lawyer) thanks?' (1 star, General - Careers & Employment, asked by Verdun M, 28 answers, 5 hours ago). 8. 'J.j. hardy...how will he do in 2008?' (1 star, Baseball, asked by Bailey, 3 answers, 5 hours ago).



kieran.b...

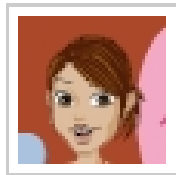
Resolved Question

[Show me another »](#)

Do girls like computer geeks / nerds?

2 weeks ago

Report It



tabitha c

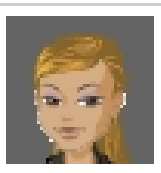
not really

2 weeks ago

0

1

Report It



Ella G


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

2 weeks ago

1

0

Report It



Resolved Question [Show me another »](#)


Melting point?

aiooii

which compound has a higher melting point? SiH4 or CH4?

1 month ago

[Report It](#)



Best Answer - Chosen by Asker

Gregg H
TOP CONTRIBUTOR

Silane has a melting point of -185C. Methane has a slightly higher melting point of -182.5C

1 month ago

[Report It](#)

Asker's Rating: *****
Thank You!

17%-45% of answers were correct

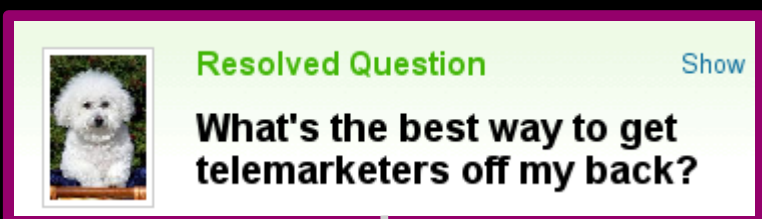
65%-90% of questions had at least one correct answer

There are top contributors ...

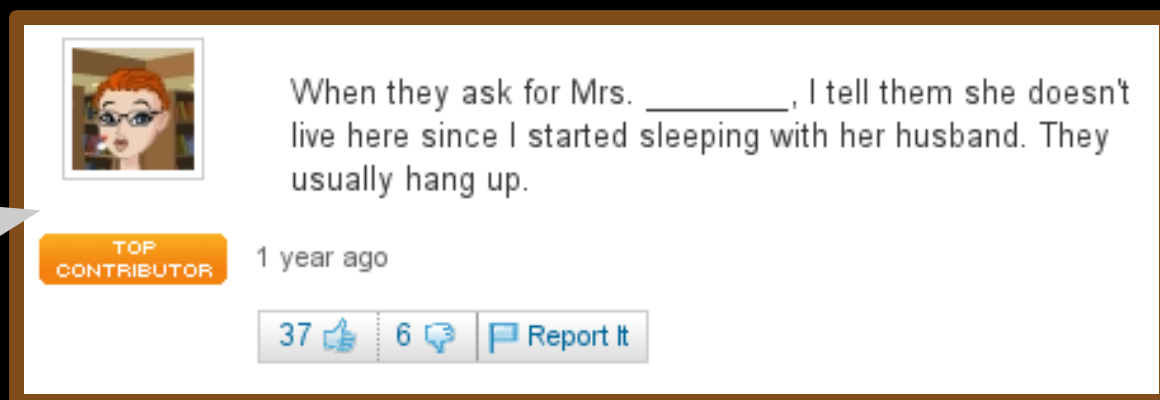


A user profile card for a top contributor. It features a profile picture of a woman with red hair and glasses. To the right of the picture is a green box containing the text "20% Best answer" and "31398 answers". Below the picture is an orange badge that says "TOP CONTRIBUTOR". To the right of the green box is another orange badge that says "TOP CONTRIBUTOR". Below the badges, the text reads "History Books & Authors", "Member Since: January 25, 2006", "Total Points: 175639 (Level 7)", and "Points earned this week: 549".

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

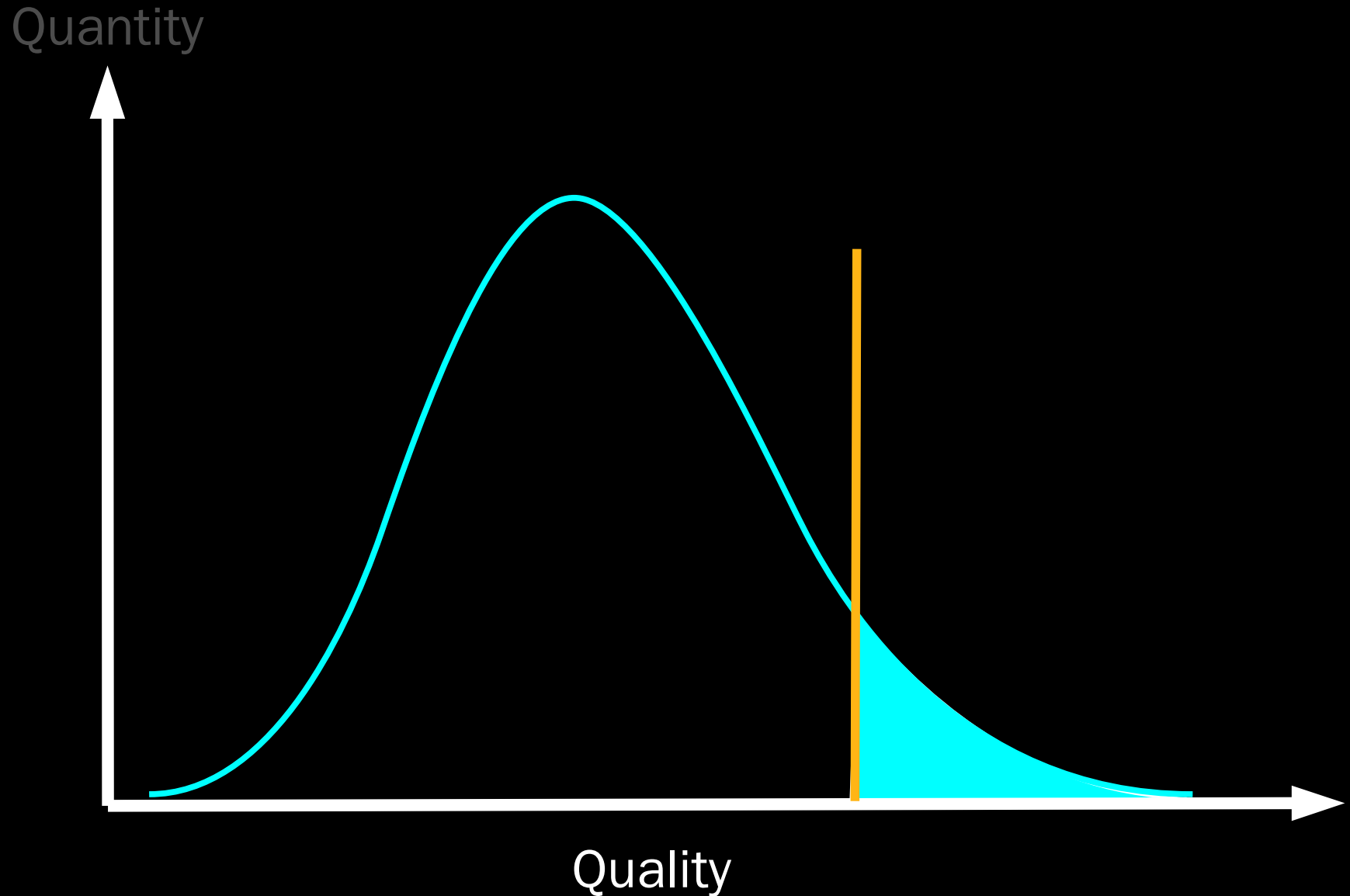


A "Resolved Question" card. It features a profile picture of a white dog. To the right of the picture is the text "Resolved Question" in green and "Show" in blue. Below the picture is the question text: "What's the best way to get telemarketers off my back?".



An answer card for the resolved question. It features a profile picture of the same woman with red hair and glasses. To the right of the picture is the answer text: "When they ask for Mrs. _____, I tell them she doesn't live here since I started sleeping with her husband. They usually hang up." Below the picture is an orange badge that says "TOP CONTRIBUTOR". Below the badge is the text "1 year ago". At the bottom of the card are three buttons: "37" with a thumbs-up icon, "6" with a speech bubble icon, and "Report It" with a flag icon.

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

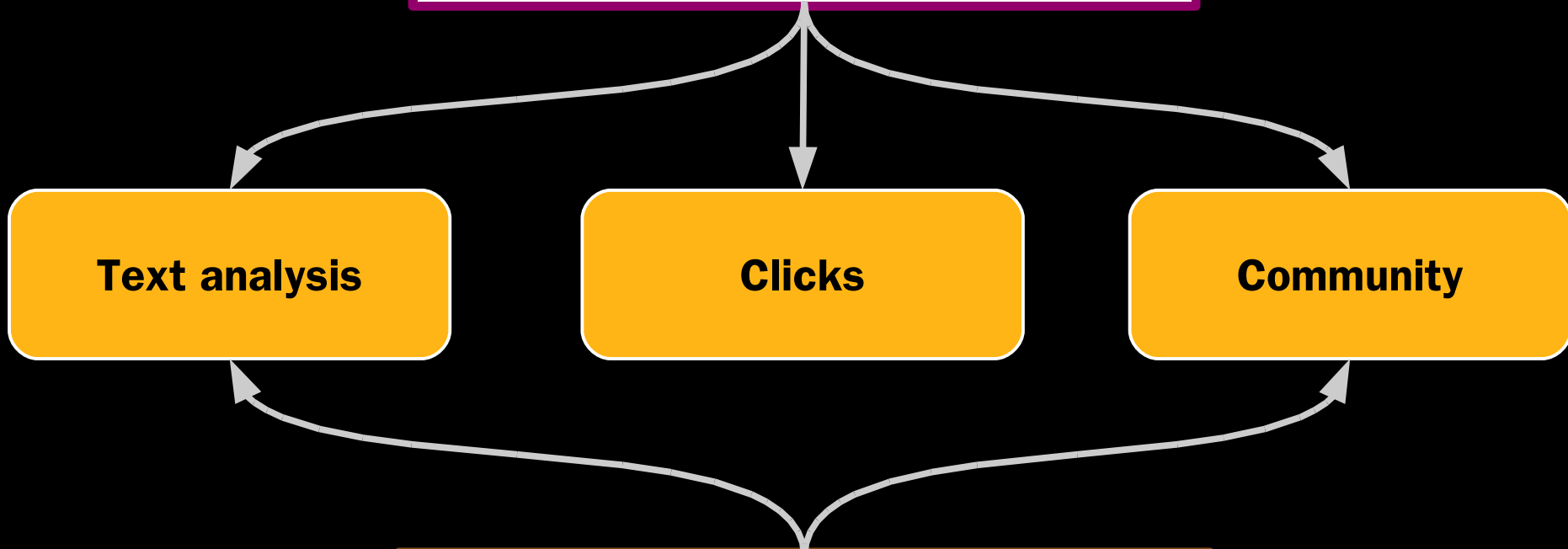
(sparse, with spam)




 **Open Question** [Show me another »](#)

I wonder.....how many megapixels have our eyes ?

4 hours ago - 3 days left to answer.



 Eyes are analog, they don't use pixels.

It's a hell of a lot higher than any current photographic standard being used though.

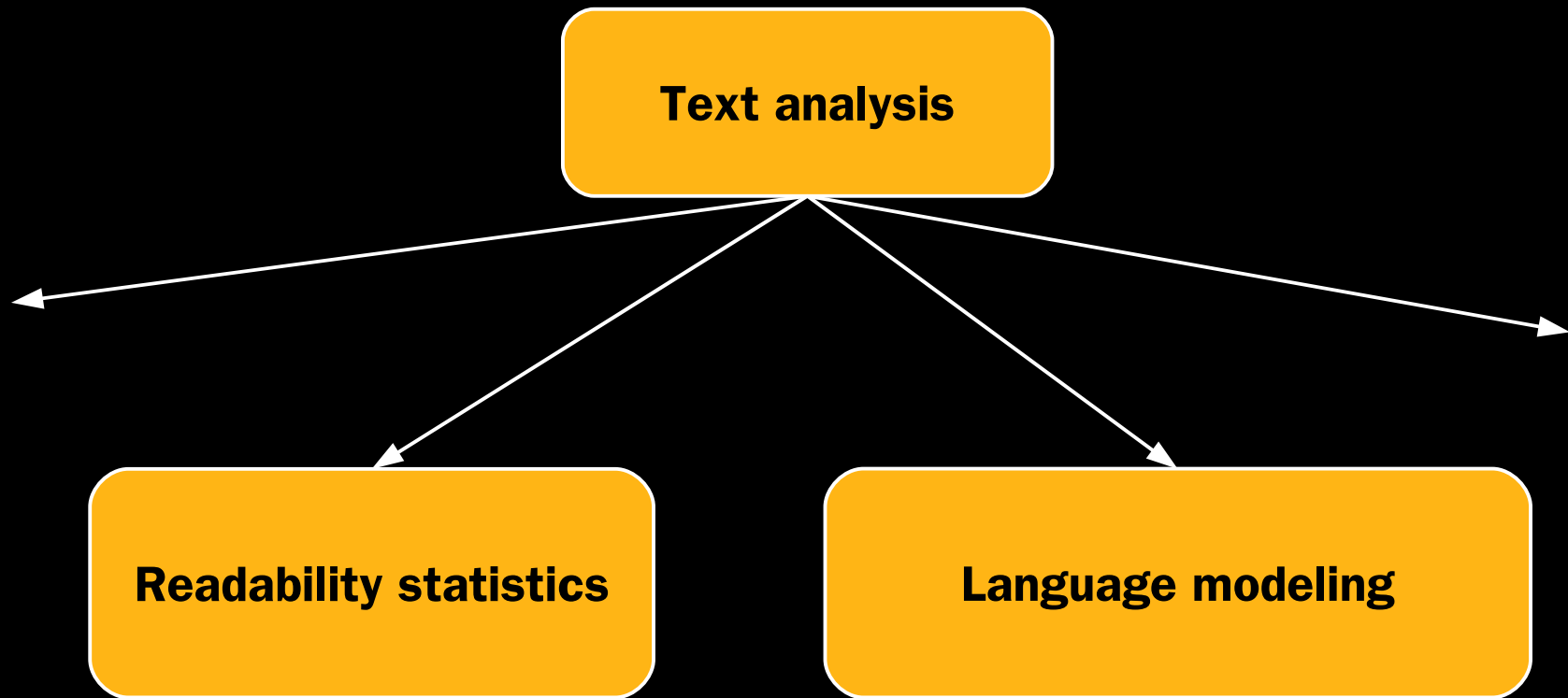
TOP CONTRIBUTOR

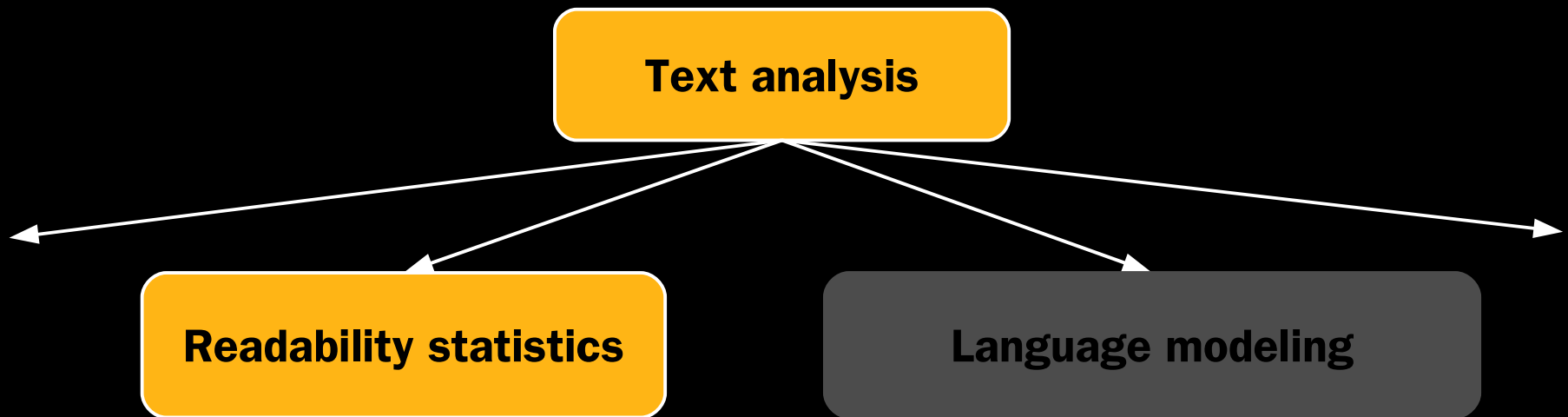


Text analysis

Clicks

Community





Punctuation density



Help! math! histogram! asap?

☆ In [Mathematics](#) - Asked by [Markyme123](#) - 0 answers - 3 minutes ago

Capitalization errors



WHAT is heidi montag thinking WITH THIS MUSIC VIDEO?

☆ In [Celebrities](#) - Asked by [chrls_bann88](#) - 0 answers - 3 minutes ago

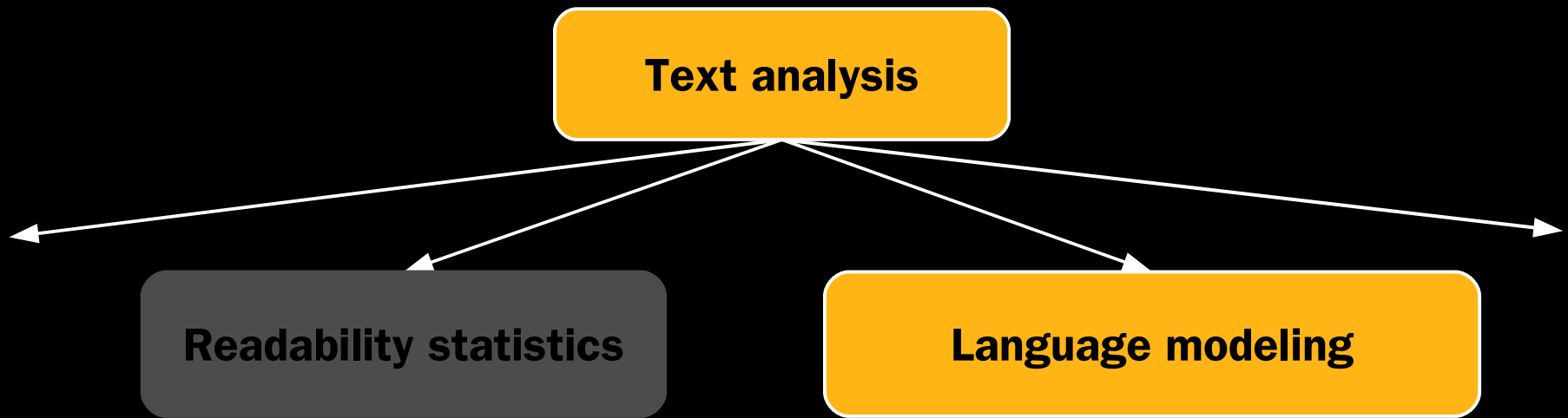
Number of words



Help!!!!!!!!!!!!!!?

☆ In [General](#) - Asked by [^So Confused^](#) - 1 answer - 6 minutes ago

+ spacing density, syllables per word,...



Language model disagreement

Distributions of word n-grams and part-of-speech sequences

when|how|why – “to” – verb

“how to identify ...”

when|how|why – verb – verb – pronoun – verb

“how do I remove ...”

Text analysis

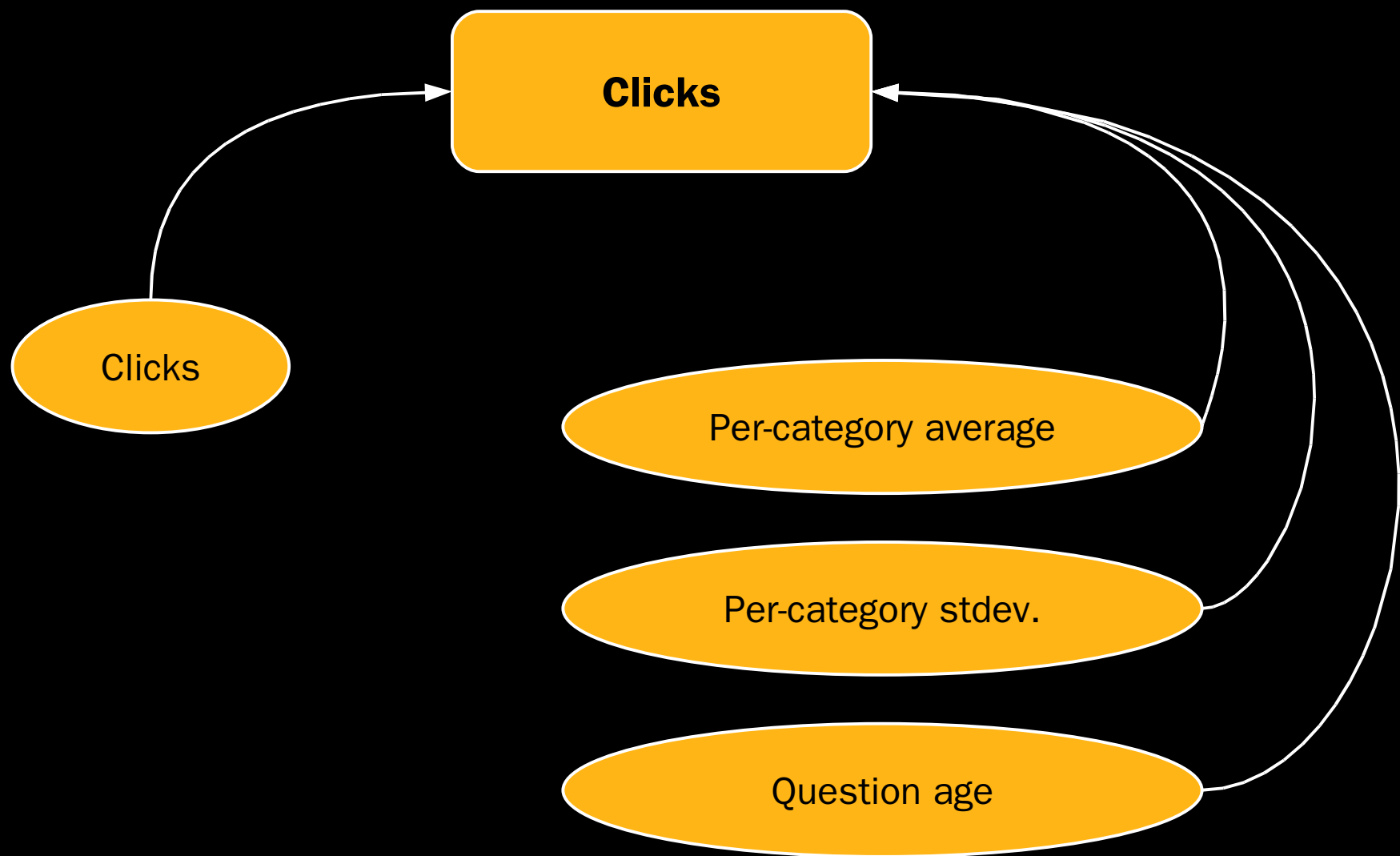
Clicks

Community

Clicks

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

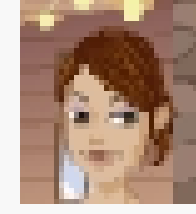
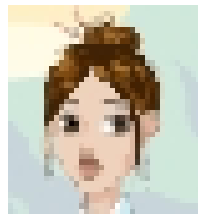
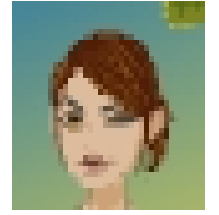
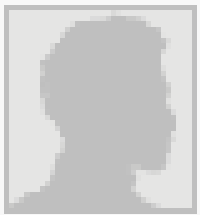
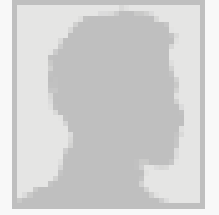
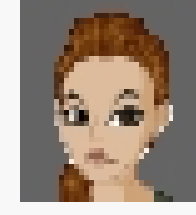
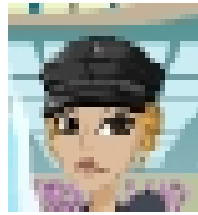
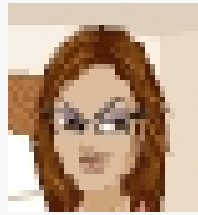
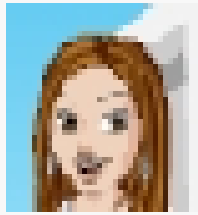
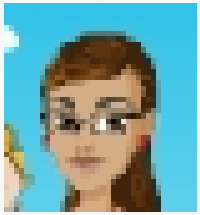
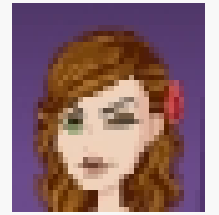
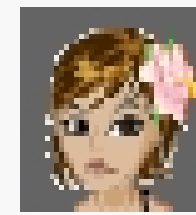
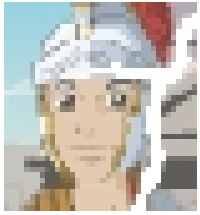
... we still know nothing

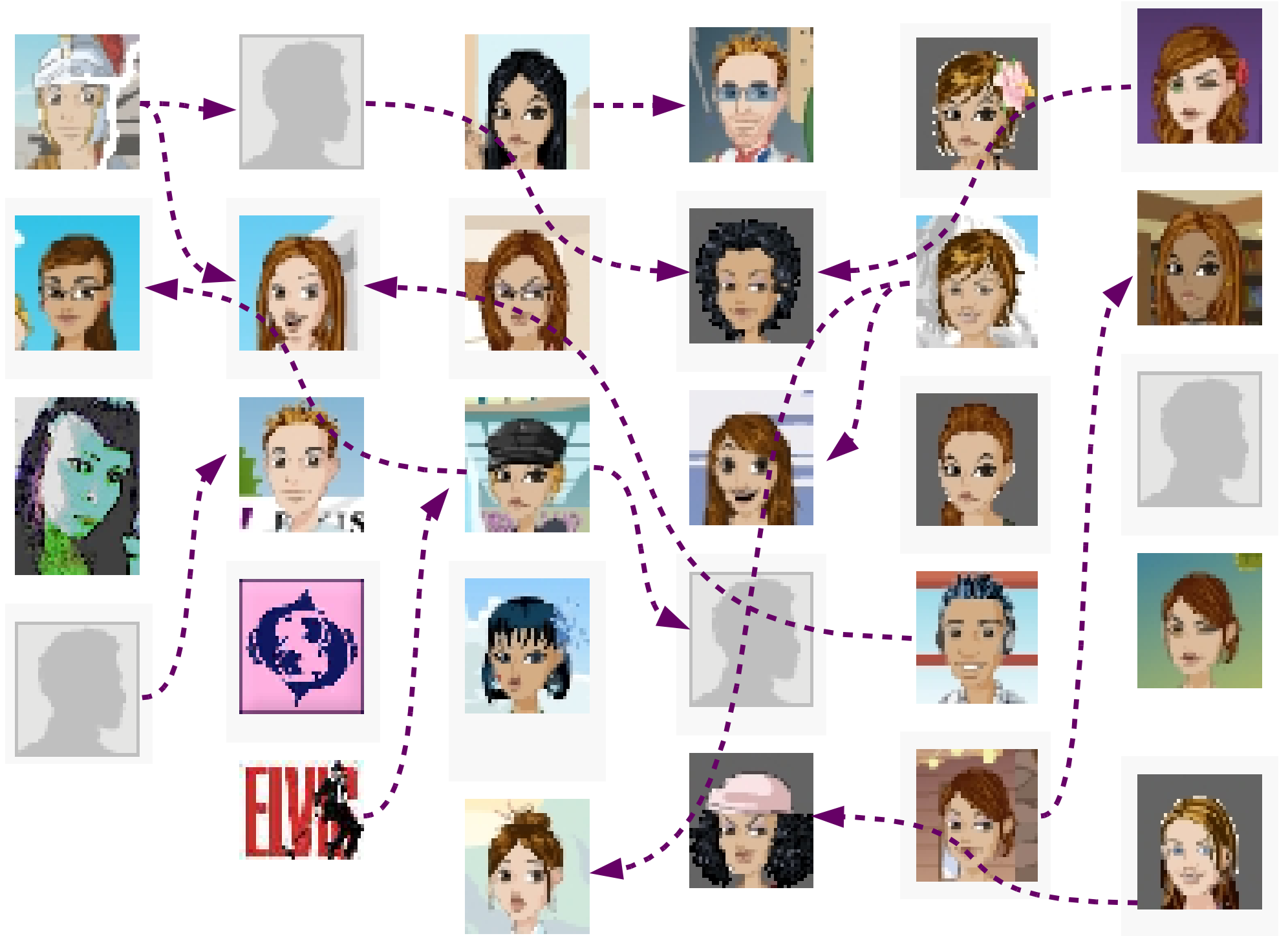


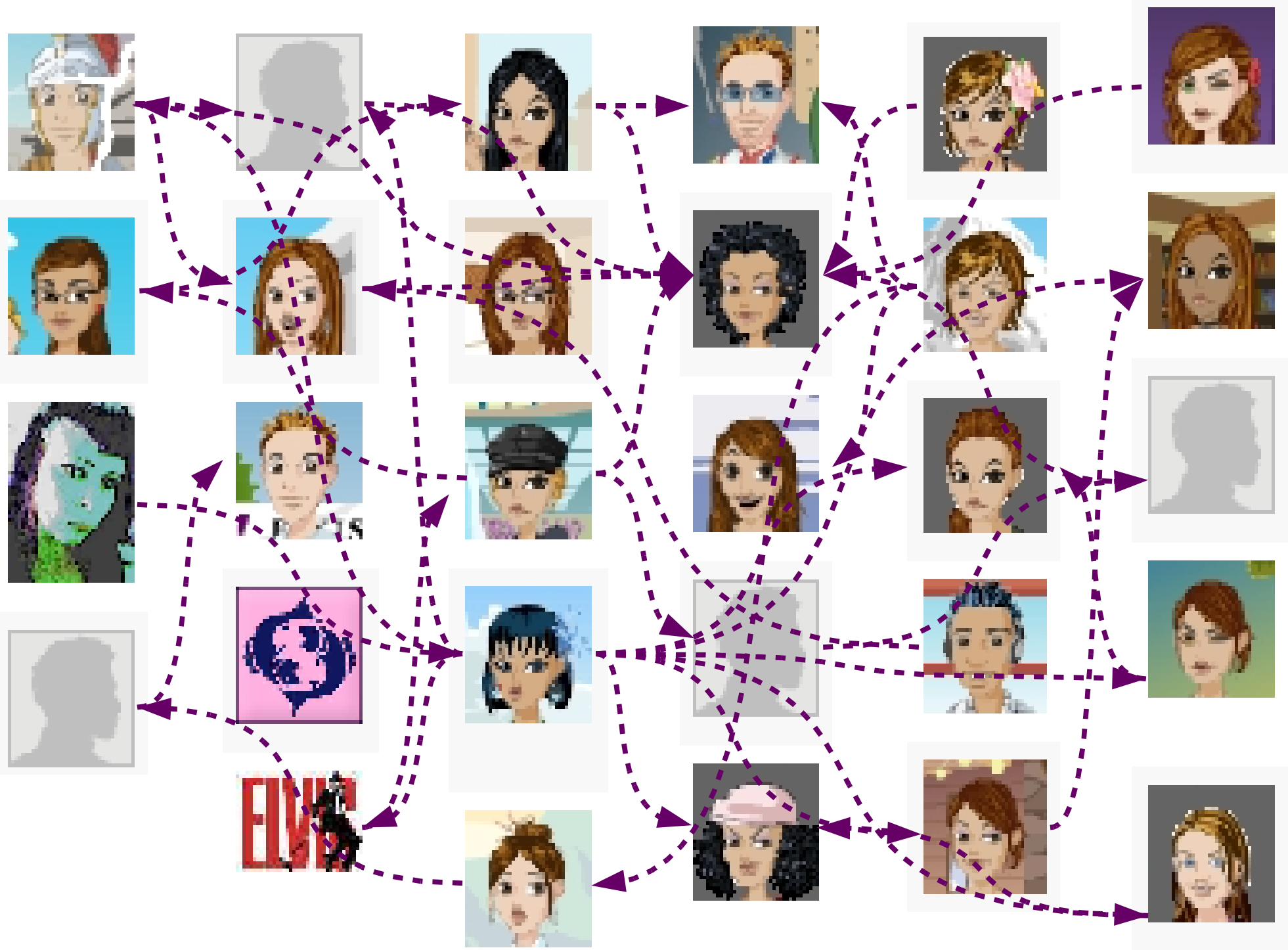
Text analysis

Clicks

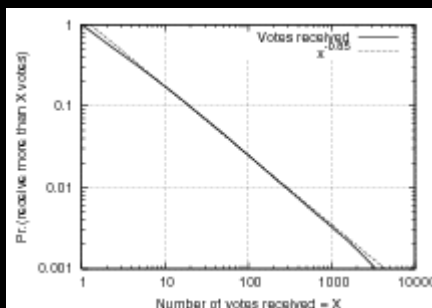
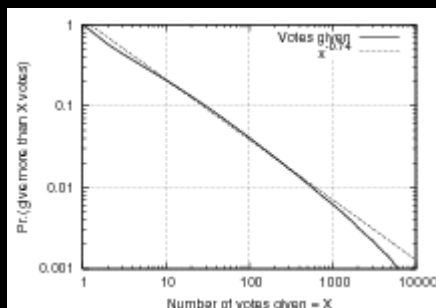
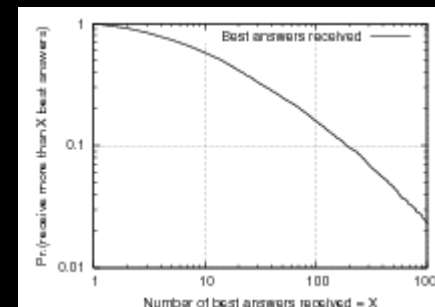
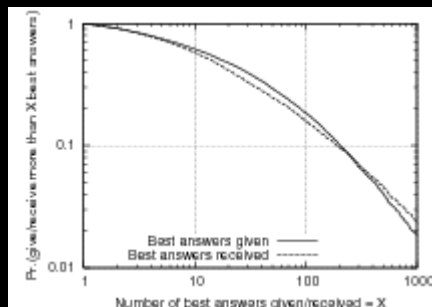
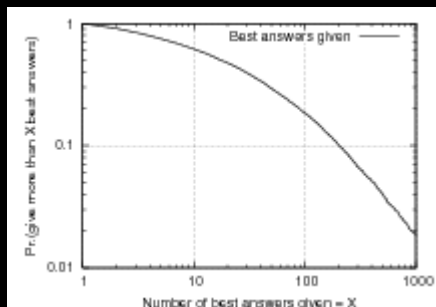
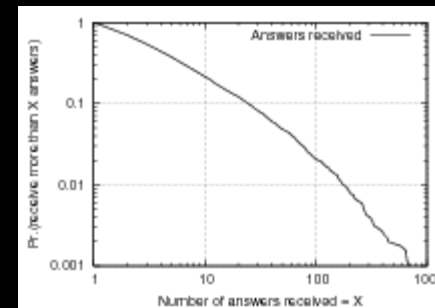
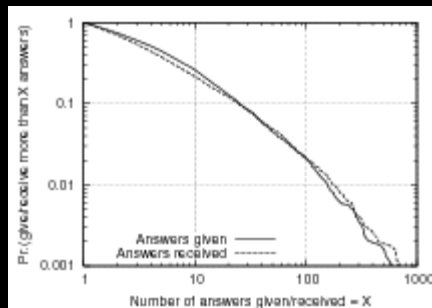
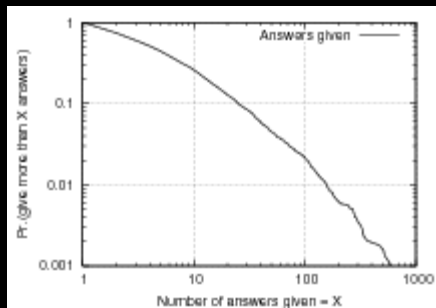
Community

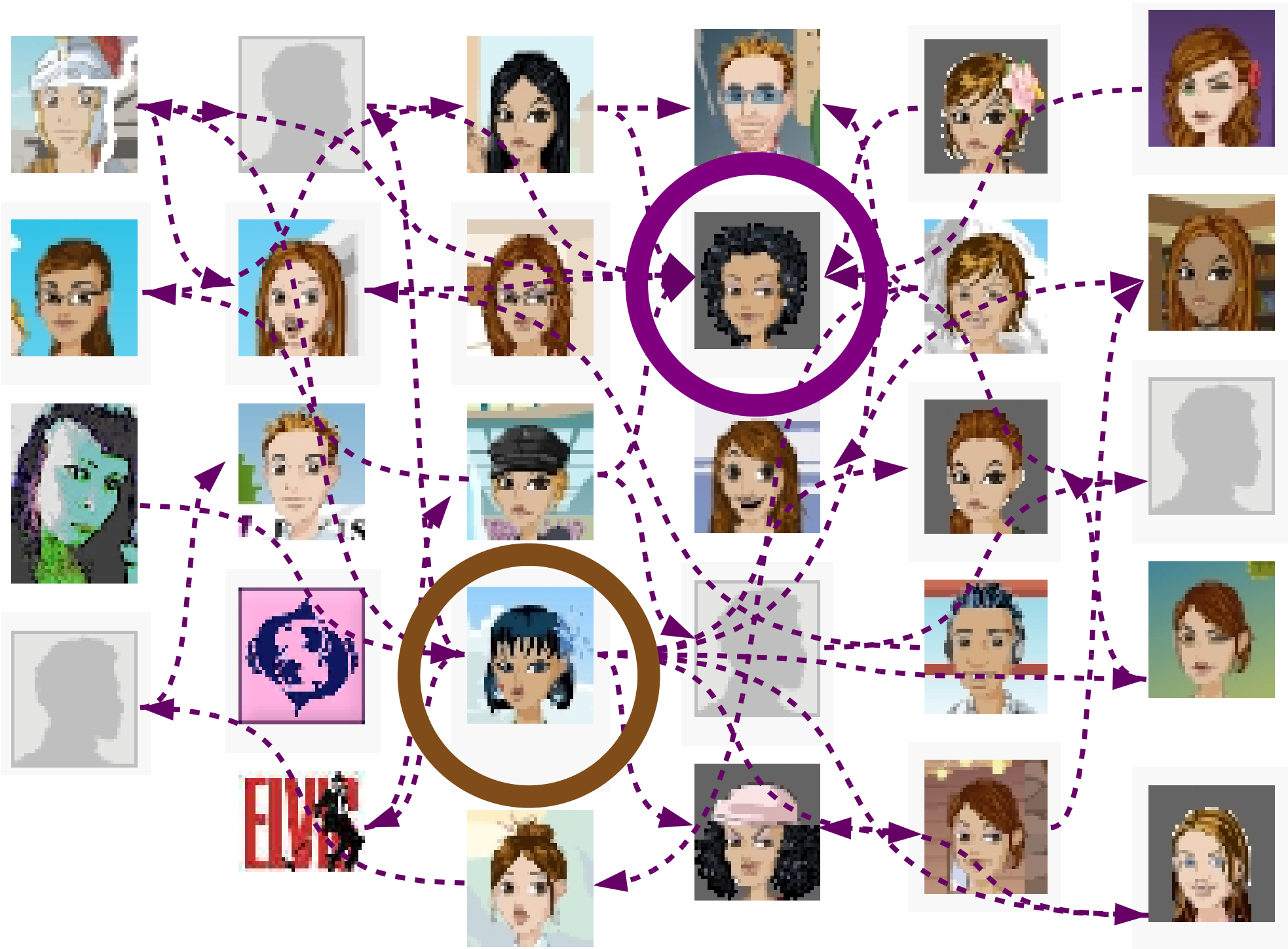


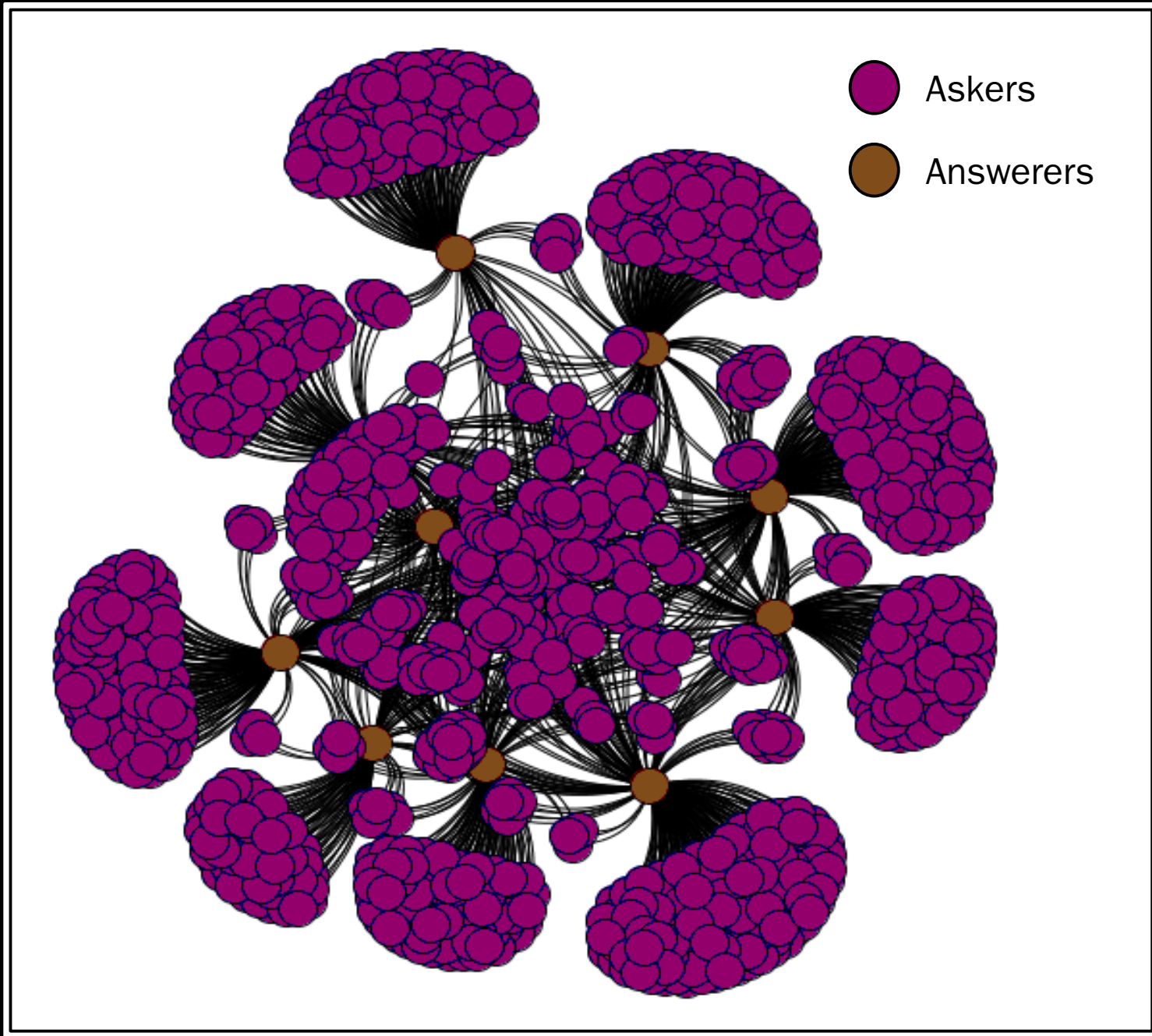




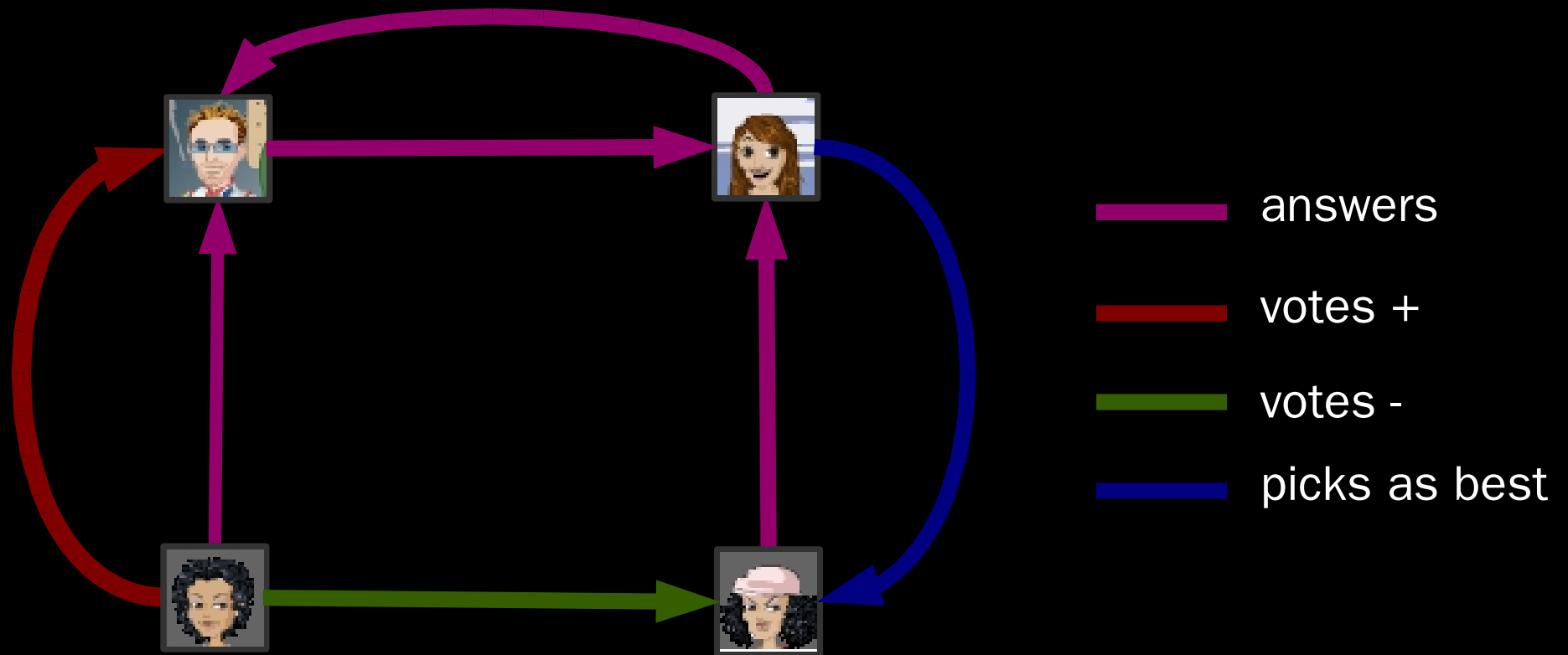
Power laws



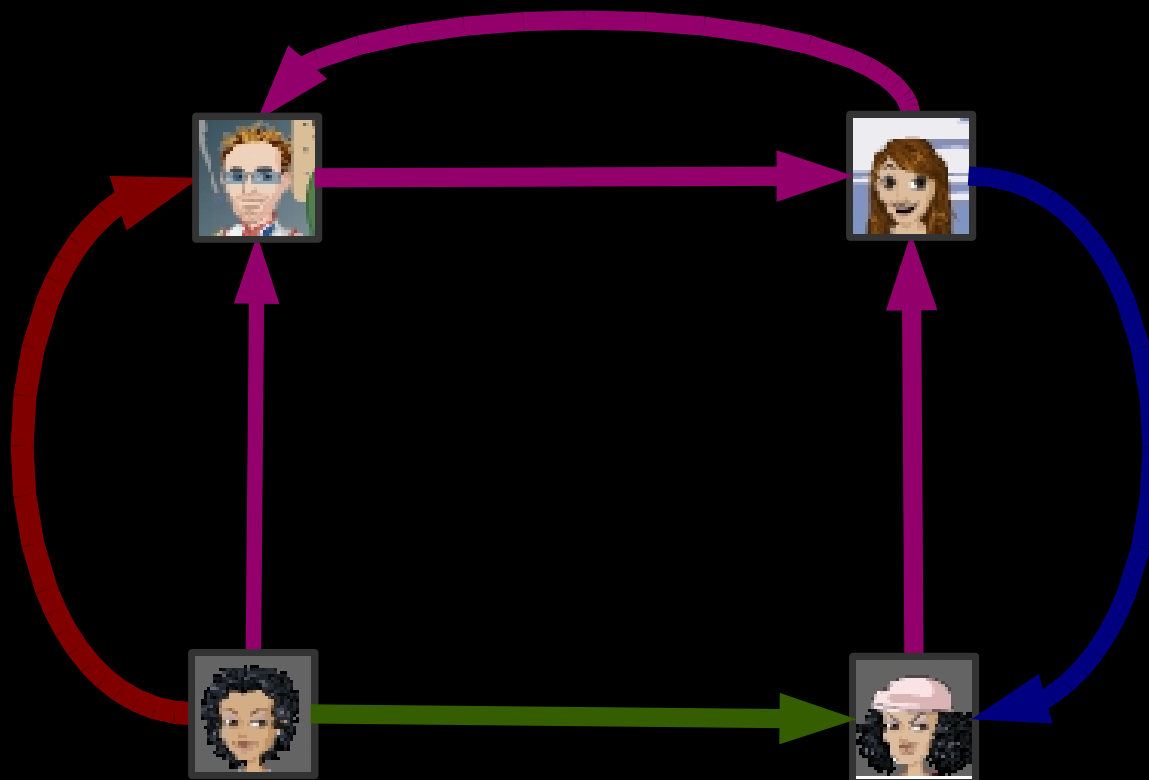




Community



Community



Degree-based metrics

answers given

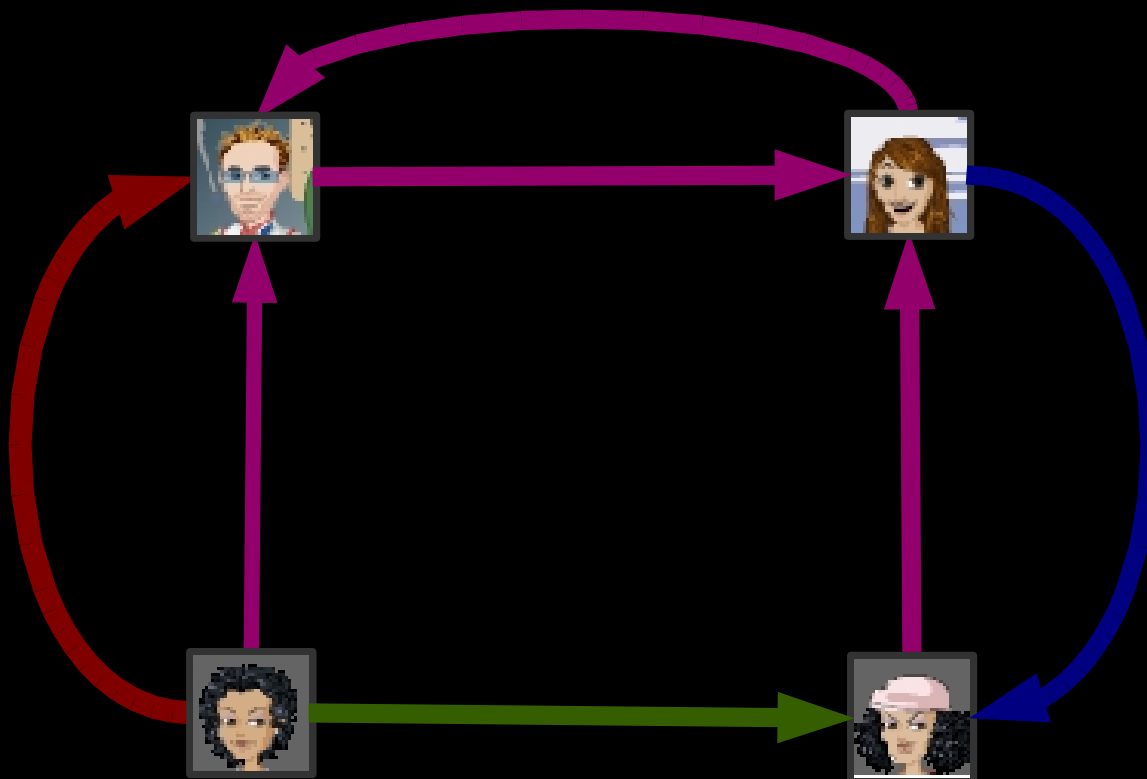
answers received

votes + given

votes + received

etc...


Community



Propagation-based metrics


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

Computed on each graph

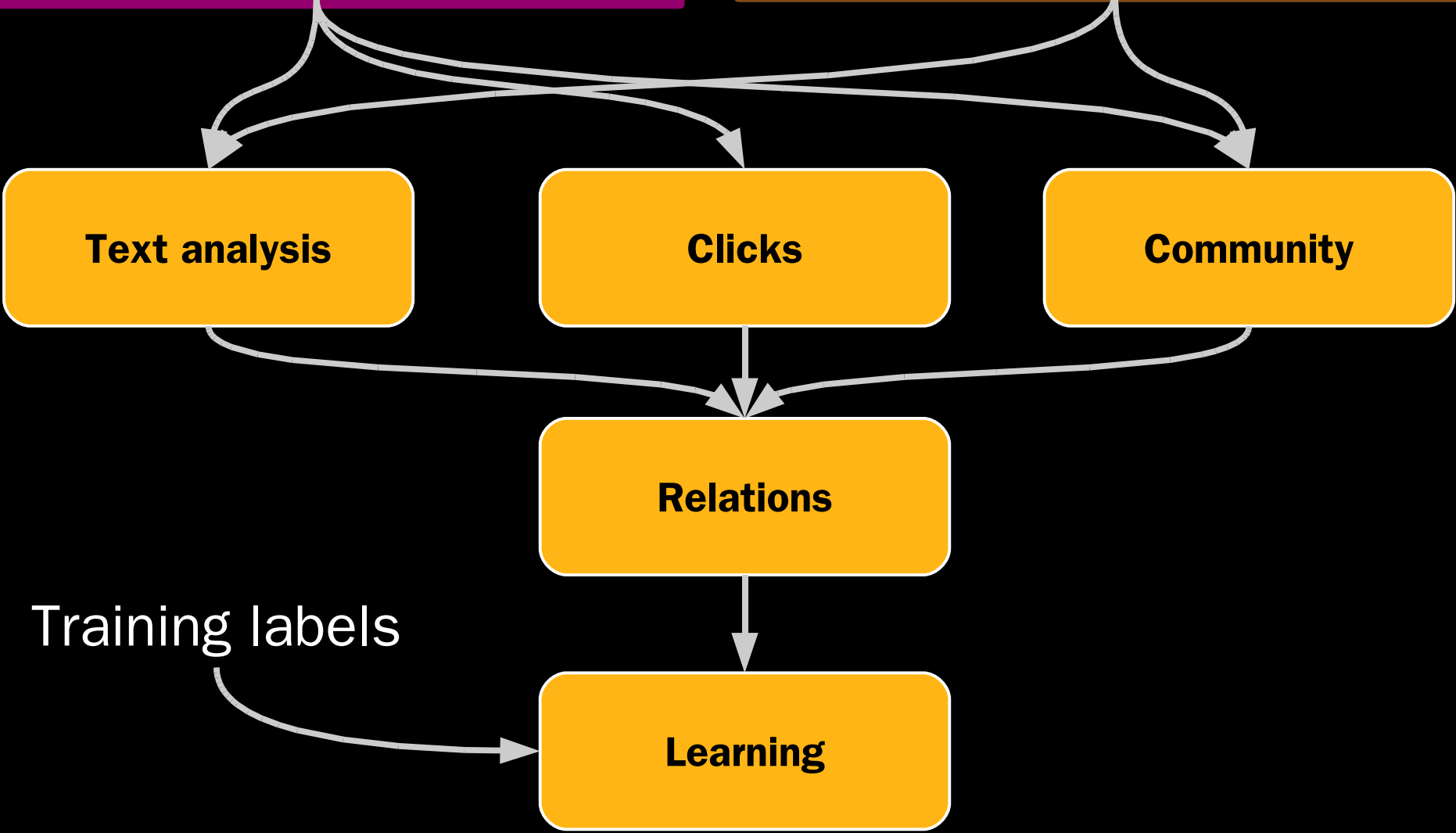
 **Open Question** [Show me another »](#)


I wonder.....how many megapixels have our eyes ?

4 hours ago - 3 days left to answer.

 Eyes are analog, they don't use pixels.


It's a hell of a lot higher than any current photographic standard being used though.



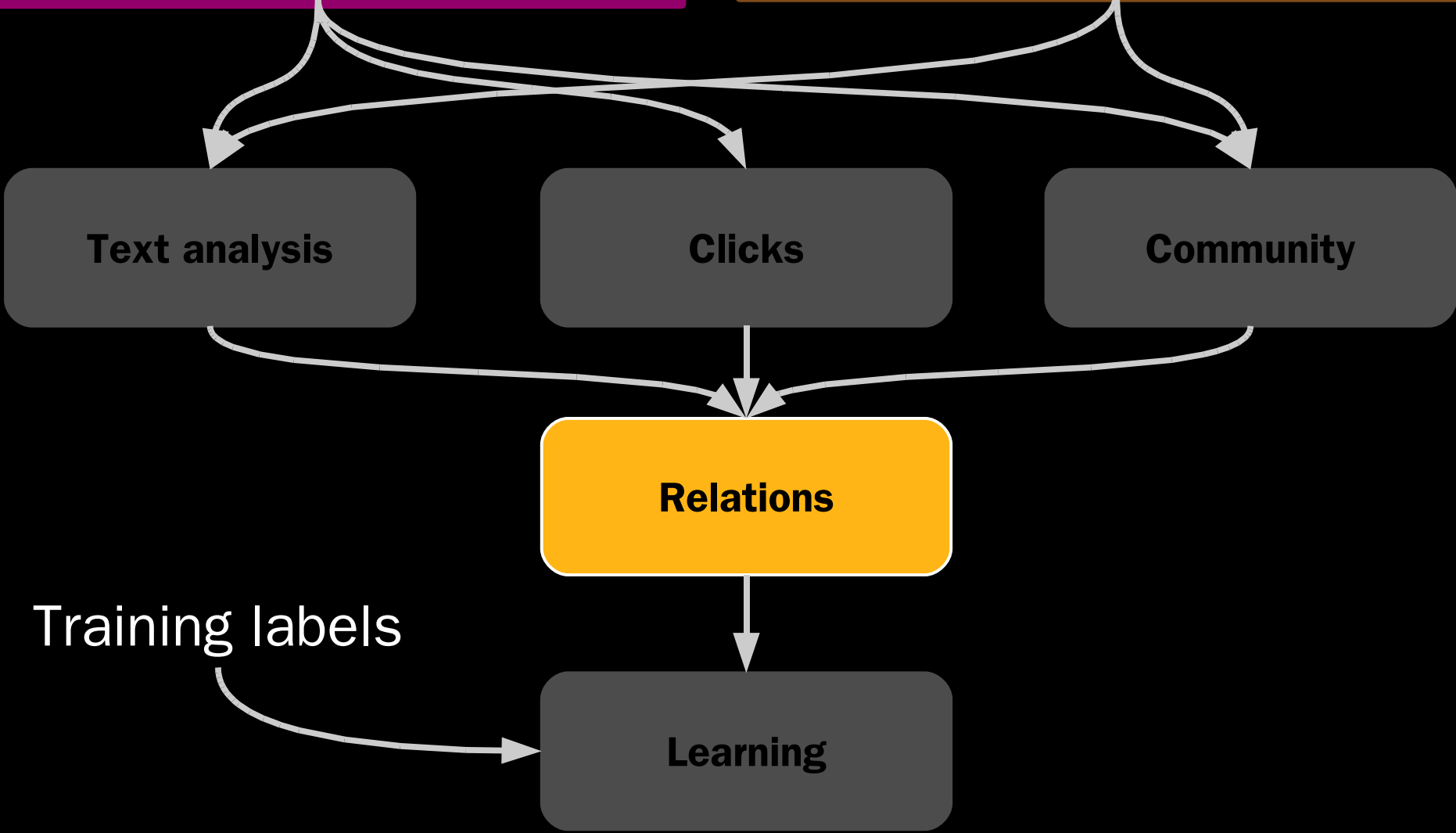
 **Open Question** [Show me another »](#)

I wonder.....how many megapixels have our eyes ?

4 hours ago - 3 days left to answer.

 Eyes are analog, they don't use pixels.

It's a hell of a lot higher than any current photographic standard being used though.



Question quality

Answer quality

	High	Medium	Low
High		15%	
Medium		76%	
Low		9%	
		100%	

Question quality

Answer quality

	High	Medium	Low
High		15%	8%
Medium		76%	74%
Low		9%	18%
		100%	100%

Question quality

Answer quality

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

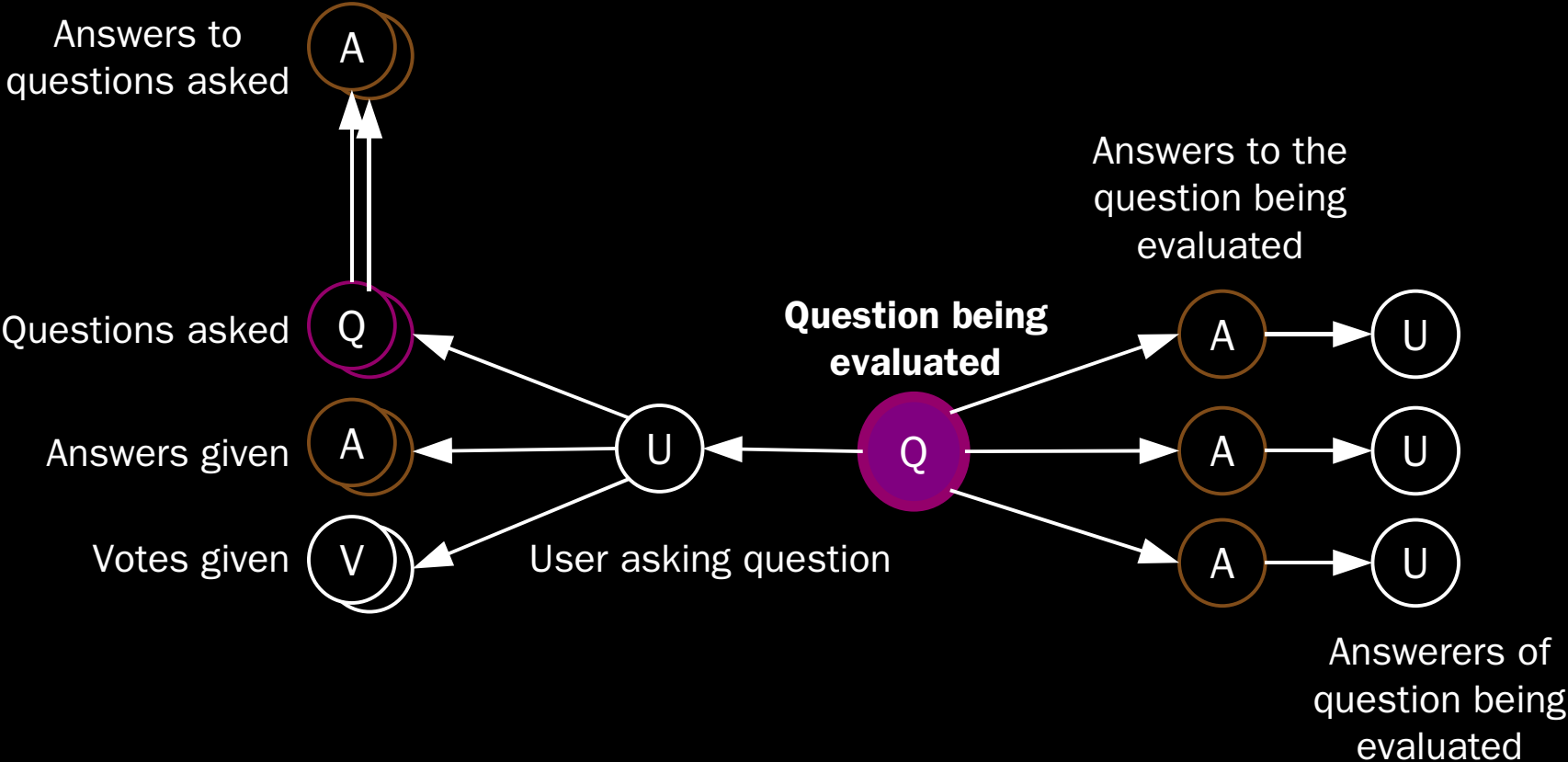
Question quality

Answer quality

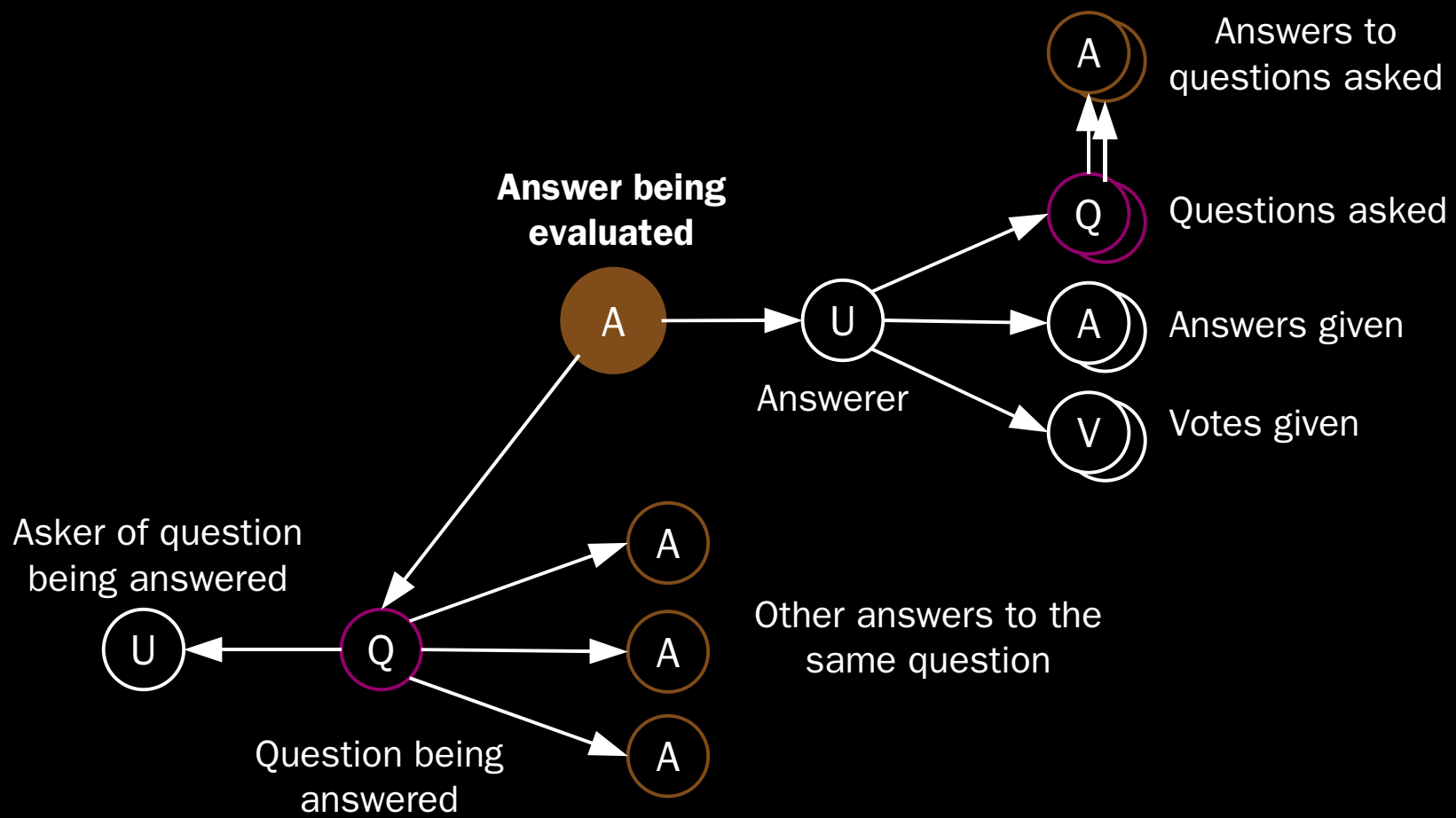
	High	Medium	Low
High	41%	15%	8%
Medium	53%	76%	74%
Low	6%	9%	18%
	100%	100%	100%


Question quality and answer quality are not independent


Relations: questions

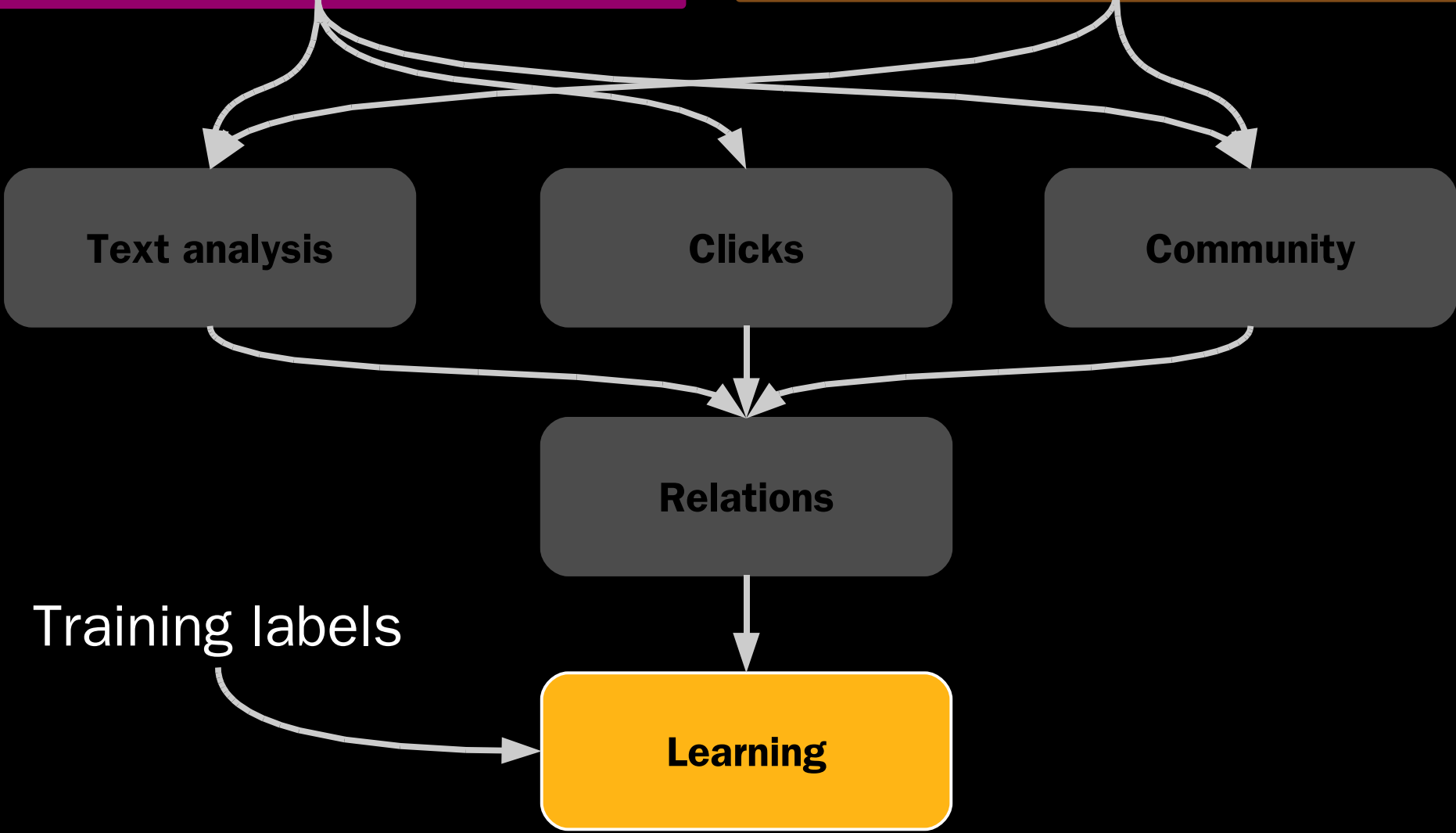


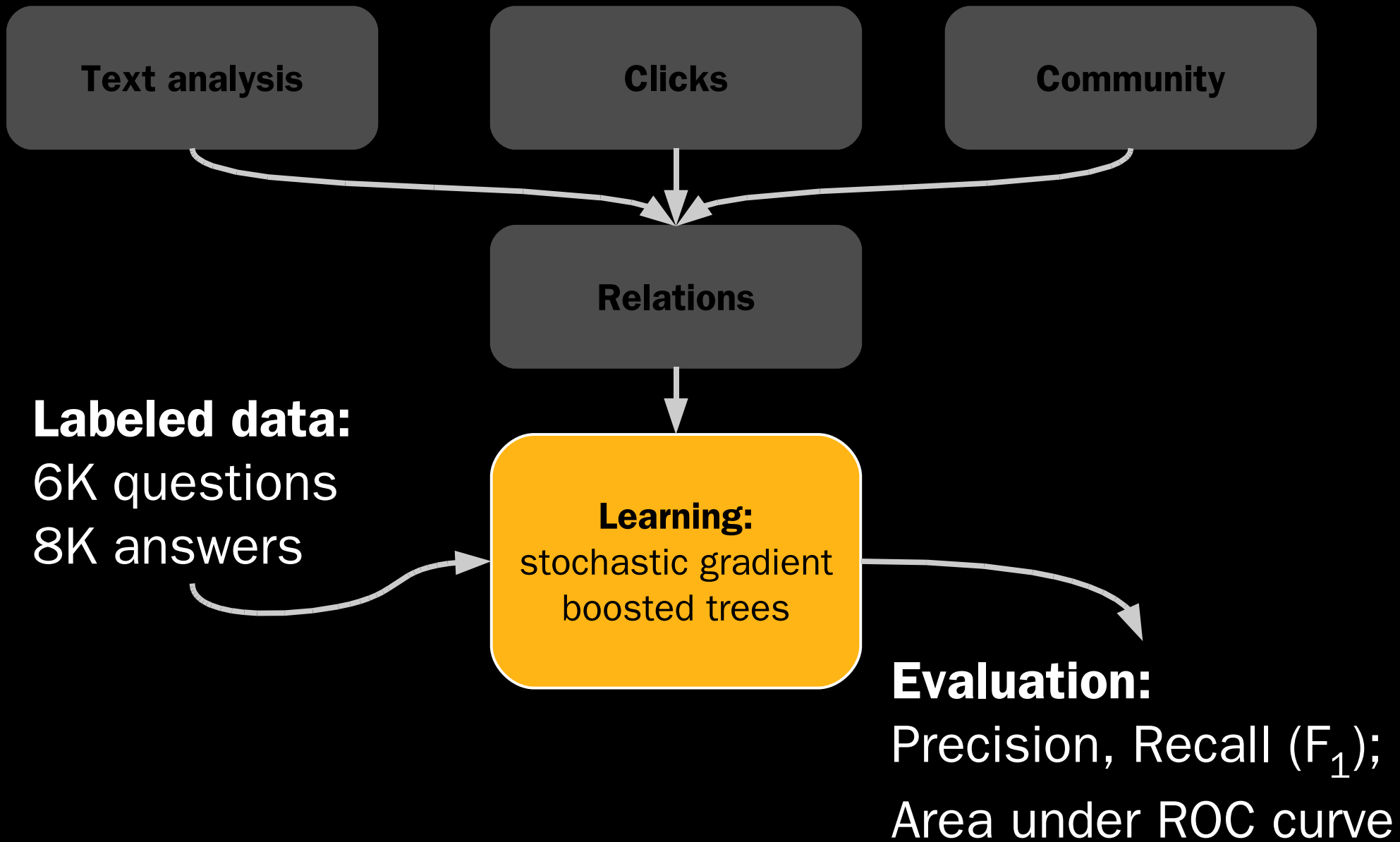
Relations: answers



 **Open Question** [Show me another »](#)
I wonder.....how many megapixels have our eyes ?
4 hours ago - 3 days left to answer.

 Eyes are analog, they don't use pixels.
It's a hell of a lot higher than any current photographic standard being used though.





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%	77%	0.76

Task: high-quality answers

	Precision	Recall	AUC
N-grams (N)	67%	86%	0.81
N + text analysis	71%	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

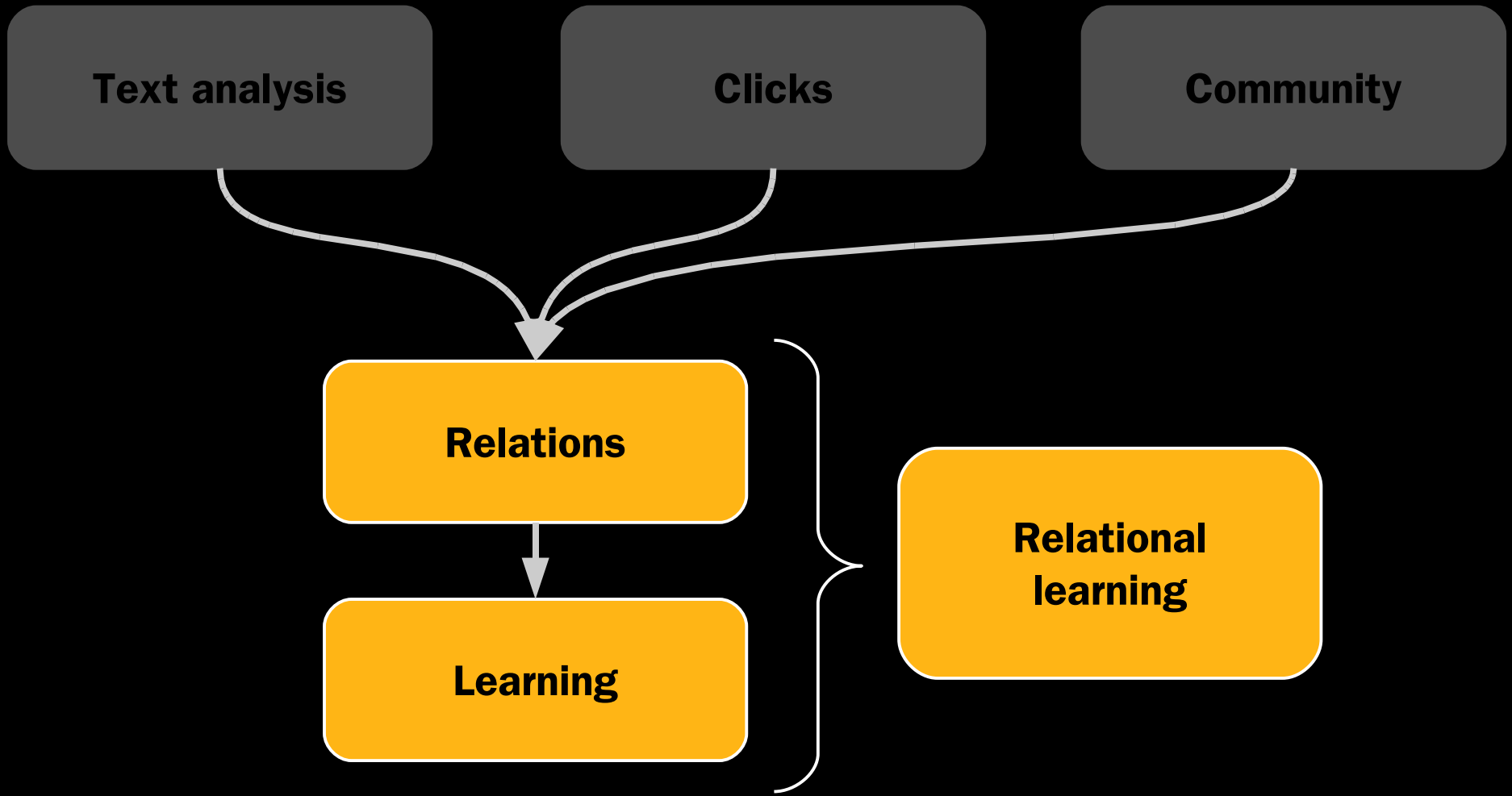
... have relatively low agreement

Classifying questions/answers

... is substantially different from
document classification

Look at orthogonal feature spaces

Future work





ask.



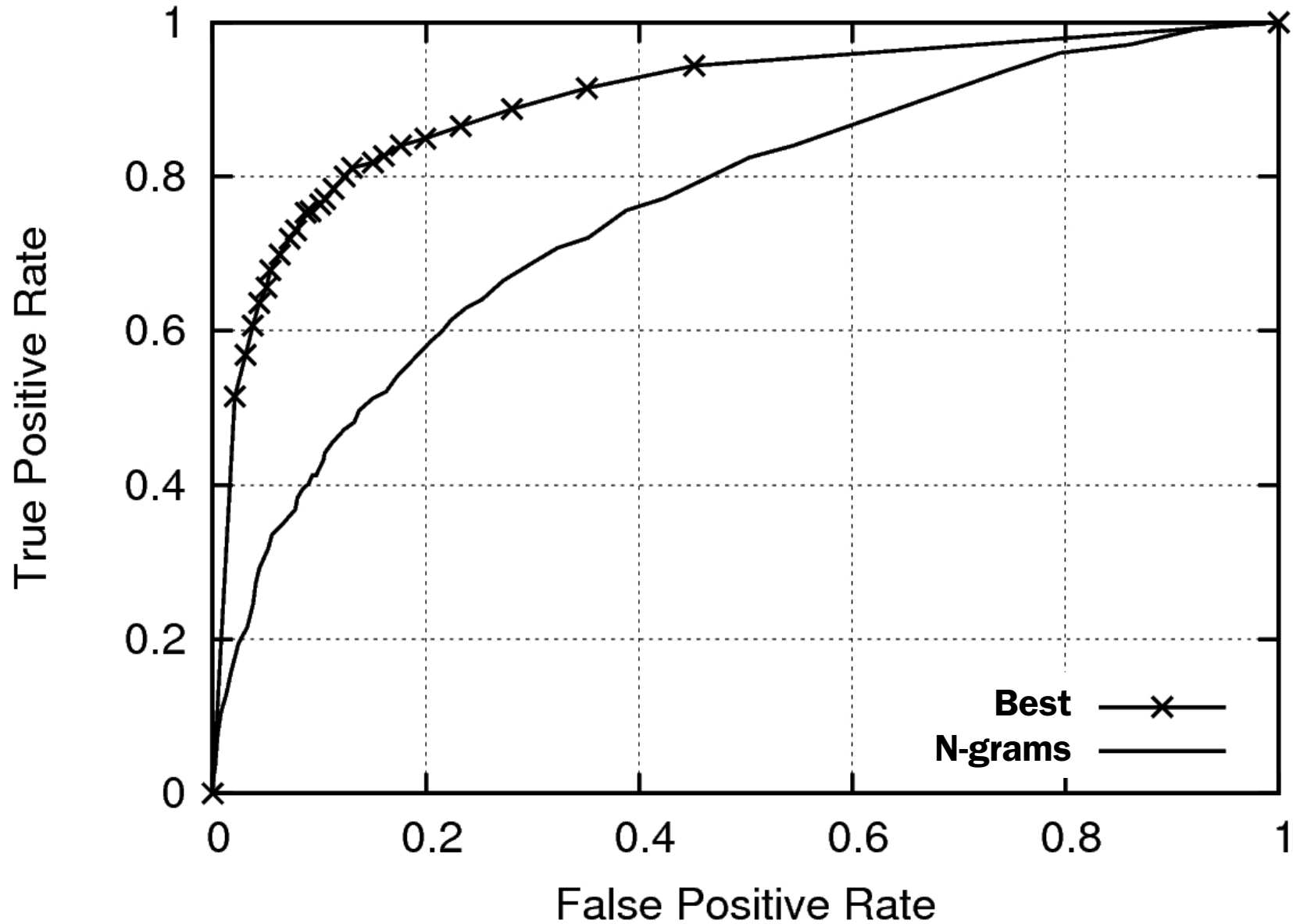
answer.



discover.

YAHOO! ANSWERS
answers.yahoo.com

ROC curve: high-quality questions



ROC curve: high-quality answers

