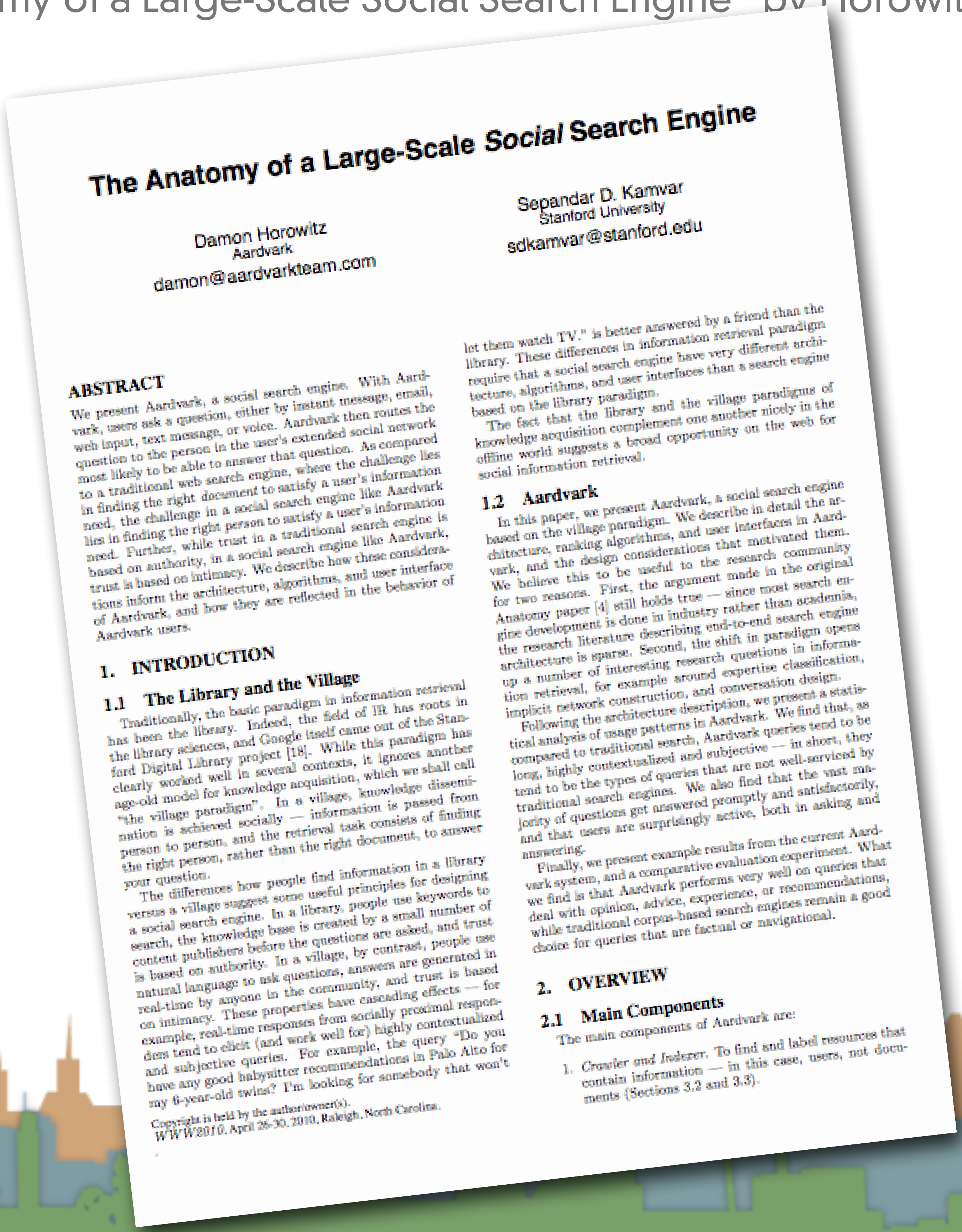
SOCIAL SEARCH

Introduction to Information Retrieval
CS 150
Donald J. Patterson

Aardvark

"The Anatomy of a Large-Scale Social Search Engine" by Horowitz, Kamvar WWW2010



The Anatomy of a Large-Scale *Social* Search Engine

Damon Horowitz
Aardvark
damon@aardvarkteam.com

Sepandar D. Kamvar
Stanford University
sdkamvar@stanford.edu

ABSTRACT

We present Aardvark, a social search engine. With Aardvark, users ask a question, either by instant message, email, web input, text message, or voice. Aardvark then routes the question to the person in the user's extended social network most likely to be able to answer that question. As compared to a traditional web search engine, where the challenge lies in finding the right document to satisfy a user's information need, the challenge in a social search engine like Aardvark lies in finding the right person to satisfy a user's information need. Further, while trust in a traditional search engine is based on authority, in a social search engine like Aardvark, trust is based on intimacy. We describe how these considerations inform the architecture, algorithms, and user interface of Aardvark, and how they are reflected in the behavior of Aardvark users.

1. INTRODUCTION

1.1 The Library and the Village

Traditionally, the basic paradigm in information retrieval has been the library. Indeed, the field of IR has roots in the library sciences, and Google itself came out of the Stanford Digital Library project [18]. While this paradigm has clearly worked well in several contexts, it ignores another age-old model for knowledge acquisition, which we shall call "the village paradigm". In a village, knowledge dissemination is achieved socially — information is passed from person to person, and the retrieval task consists of finding the right person, rather than the right document, to answer your question.

The differences how people find information in a library versus a village suggest some useful principles for designing a social search engine. In a library, people use keywords to search, the knowledge base is created by a small number of content publishers before the questions are asked, and trust is based on authority. In a village, by contrast, people use natural language to ask questions, answers are generated in real-time by anyone in the community, and trust is based on intimacy. These properties have cascading effects — for example, real-time responses from socially proximal responders tend to elicit (and work well for) highly contextualized and subjective queries. For example, the query "Do you have any good babysitter recommendations in Palo Alto for my 6-year-old twins? I'm looking for somebody that won't let them watch TV." is better answered by a friend than the library. These differences in information retrieval paradigm require that a social search engine have very different architecture, algorithms, and user interfaces than a search engine based on the library paradigm.

The fact that the library and the village paradigms of knowledge acquisition complement one another nicely in the offline world suggests a broad opportunity on the web for social information retrieval.

1.2 Aardvark

In this paper, we present Aardvark, a social search engine based on the village paradigm. We describe in detail the architecture, ranking algorithms, and user interfaces in Aardvark, and the design considerations that motivated them. We believe this to be useful to the research community for two reasons. First, the argument made in the original Anatomy paper [4] still holds true — since most search engine development is done in industry rather than academia, the research literature describing end-to-end search engine architecture is sparse. Second, the shift in paradigm opens up a number of interesting research questions in information retrieval, for example around expertise classification, implicit network construction, and conversation design.

Following the architecture description, we present a statistical analysis of usage patterns in Aardvark. We find that, as compared to traditional search, Aardvark queries tend to be long, highly contextualized and subjective — in short, they tend to be the types of queries that are not well-serviced by traditional search engines. We also find that the vast majority of questions get answered promptly and satisfactorily, and that users are surprisingly active, both in asking and answering.

Finally, we present example results from the current Aardvark system, and a comparative evaluation experiment. What we find is that Aardvark performs very well on queries that deal with opinion, advice, experience, or recommendations, while traditional corpus-based search engines remain a good choice for queries that are factual or navigational.

2. OVERVIEW

2.1 Main Components

The main components of Aardvark are:

1. *Crawler and Indexer.* To find and label resources that contain information — in this case, users, not documents (Sections 3.2 and 3.3).

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

Aardvark

- Web IR
  - Input is a query of keywords
  - Search is over documents
  - Trust is based on authority
  - Mental model is a library

Aardvark

- Web IR
  - Input is a query of keywords
  - Search is over documents
  - Trust is based on authority
  - Mental model is a library

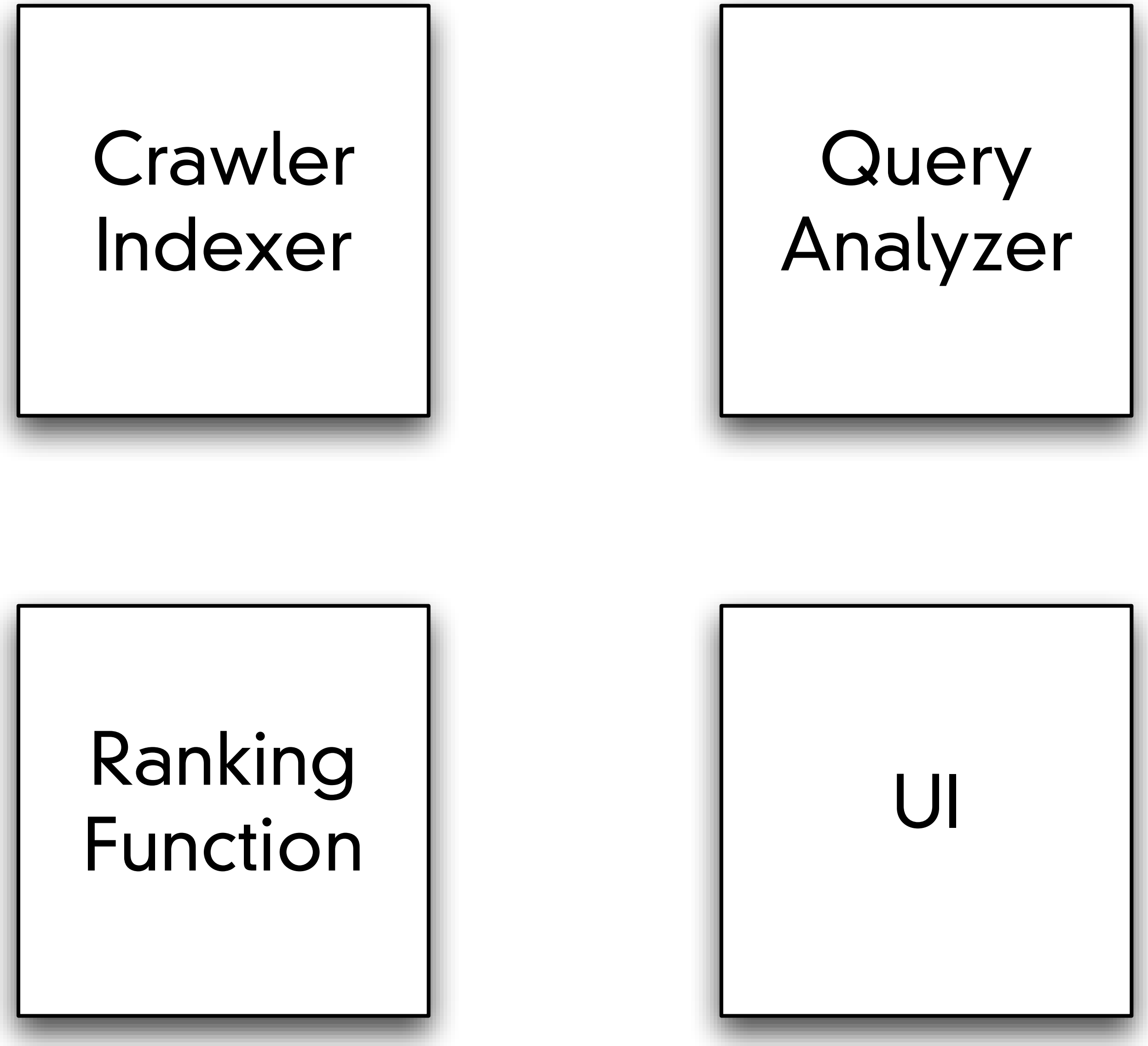
- Social Search
  - Input is a question
  - Search is over people
  - Trust is based on intimacy
  - Mental model is a village

Aardvark

- Web IR
  - facts
  - navigation
  - transactions

- Social Search
  - opinion
  - advice
  - experience
  - recommendations

Components



Crawler Indexer

Query Analyzer

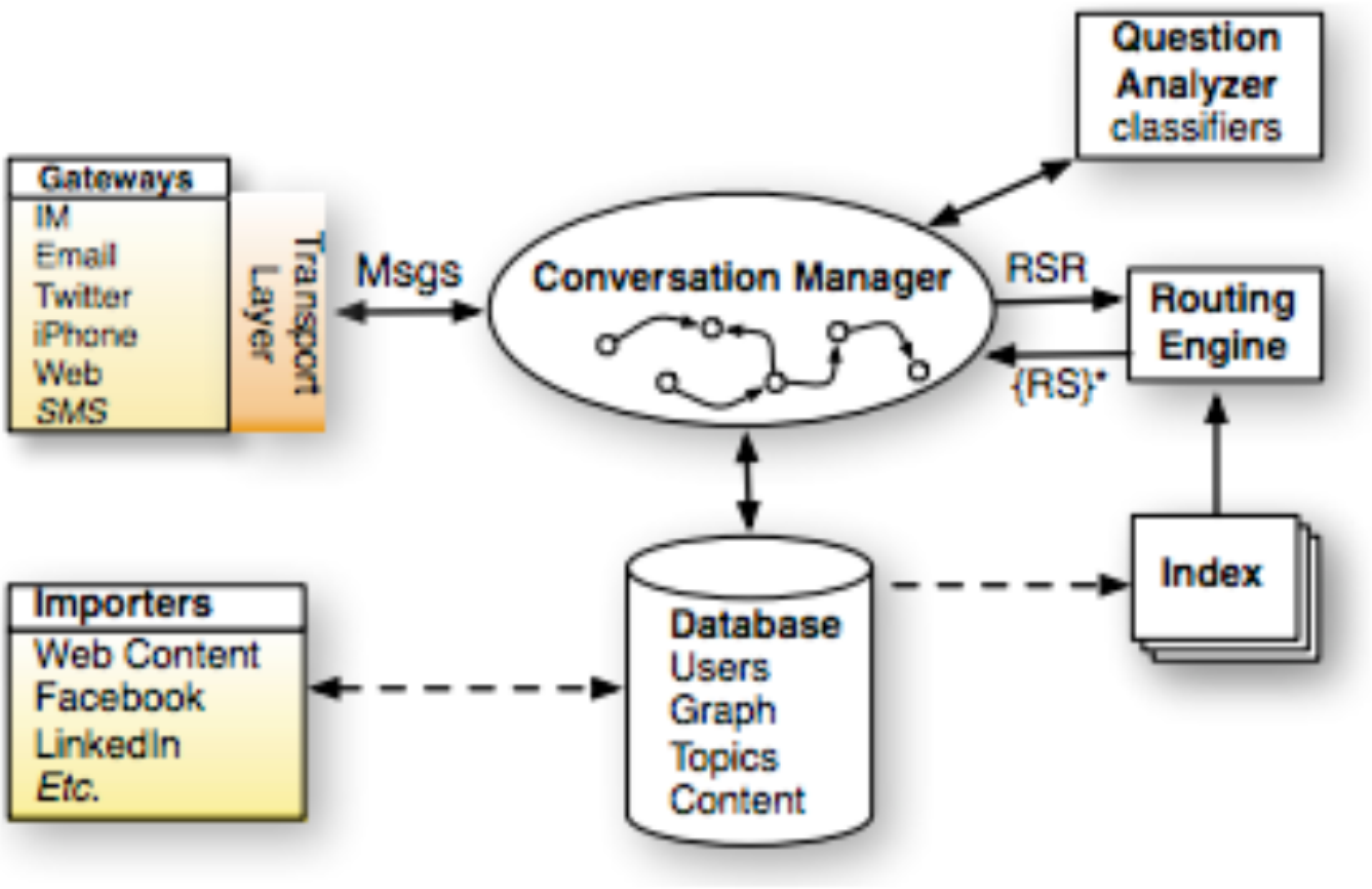
Ranking Function

UI

Aardvark

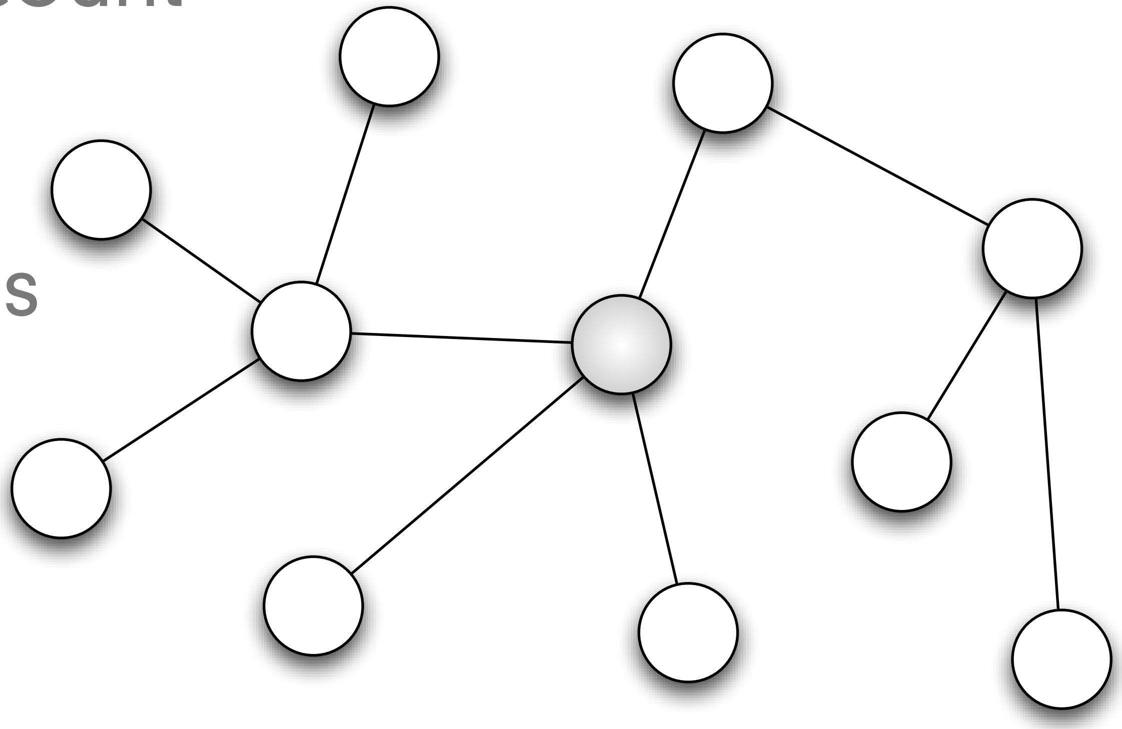
- Crawler/Indexer
  - Users not documents
- Query Analyzer
  - Understand the information need
- Ranking Function
  - Pick the best resources
- UI
  - To manage the conversation

SCHEMATIC OF THE ARCH. OF AARDVARK



Welcome to Aardvark - Sign on process

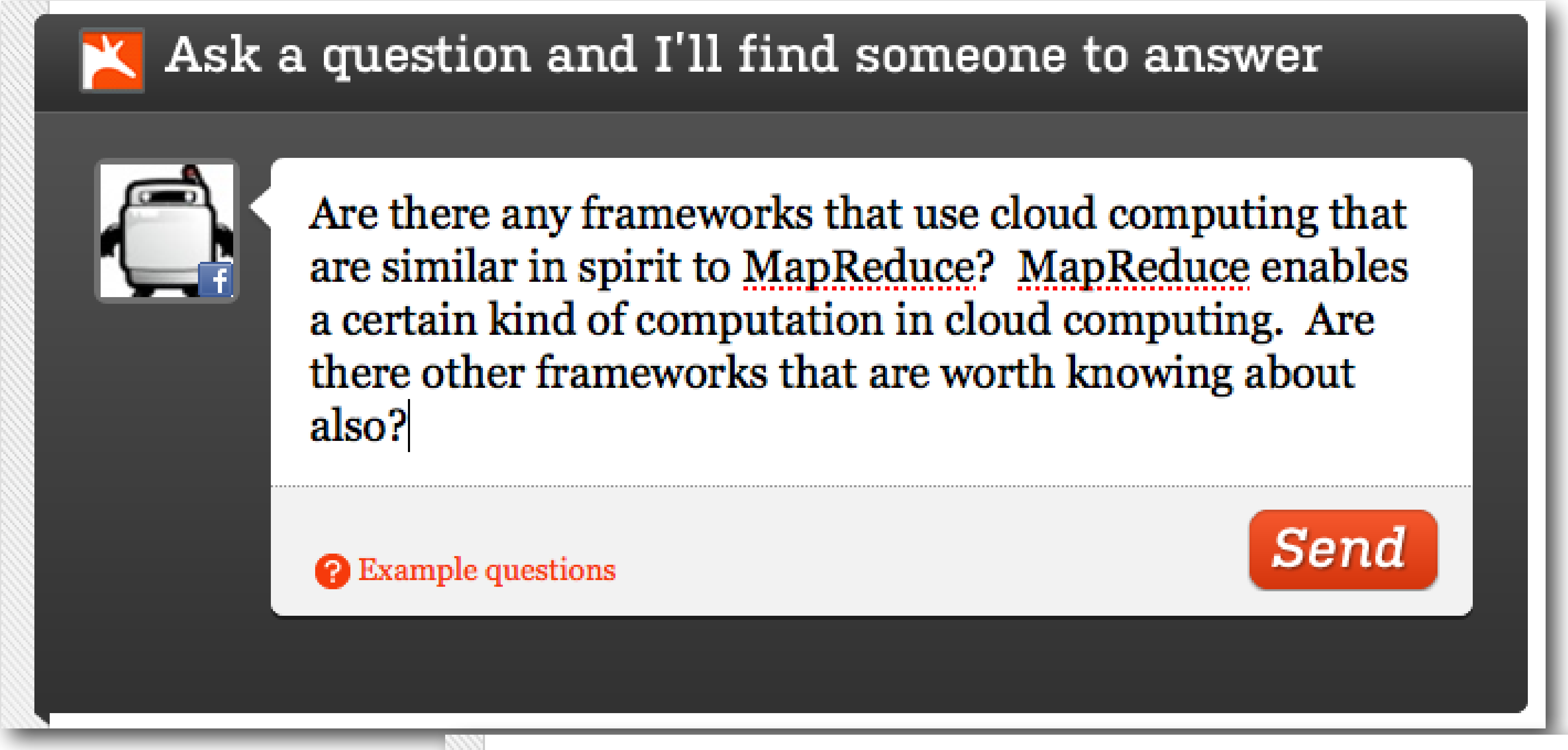
- After confirming a new user’s account
- A Social Graph is built
  - Facebook/LinkedIn connections
  - webmail connections
  - manual email invites
  - “group” aware
    - This is a work colleague, college friend, etc.



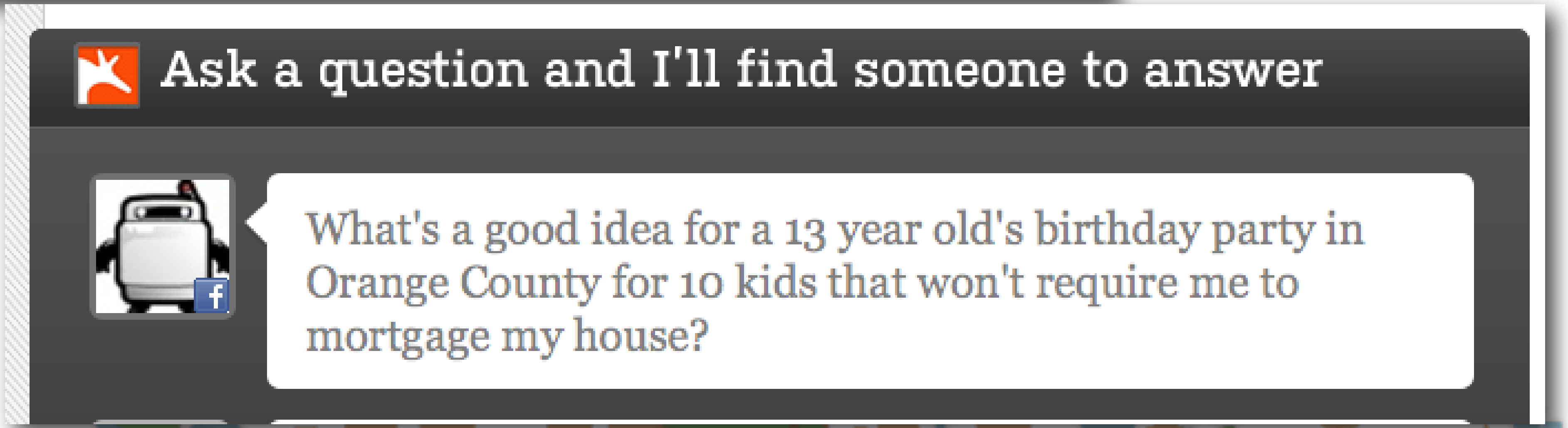
Welcome to Aardvark - Sign on process

- A knowledge bank is built on
    - self-identified expertise
    - friend identified expertise
    - home page identified expertise
    - facebook status update analysis
    - twitter status update analysis
    - observed Aardvark usage
- knowledge bank’s inverted index maps topic -> user

The Question



Are there any frameworks that use cloud computing that are similar in spirit to MapReduce? MapReduce enables a certain kind of computation in cloud computing. Are there other frameworks that are worth knowing about also?



What's a good idea for a 13 year old's birthday party in Orange County for 10 kids that won't require me to mortgage my house?

The Question

- The question is acquired by various input channels:
  - text, web, IM, mobile, etc.
- The question is screened for obscenity
- The question is topic analyzed
  - topic is presented to asker for confirmation
- The question is passed to a routing engine
  - which ranks potential answers based on
  - social graph and expertise

Routing Engine

- Pick the best answerer

  $$p(u_i|q) = \sum_{t \in T} p(u_i|t)p(t|q)$$

  - What is the probability that user i will answer question q?
  - Marginalize over the topics in the question and the topic expertise of the user
- Pick the best pair of users

  $$p(u_i|u_j)$$

  - Which user i is the most likely to give a good answer to user j?

Routing Engine

$$s(u_i, u_j, q) = p(u_i|u_j) \cdot p(u_i|q) = p(u_i|u_j) \sum_{t \in T} p(u_i|t)p(t|q)$$

- Ranking function combines the two
    - For a given query, q, by user j
        - what is the best user i to ask?
- Biases intimacy over authority
- Notice there is nothing like PageRank here
- The only real-time component is $p(t|q)$

Indexing People

- Figuring out $p(t|u_i)$
- Figuring out $p(u_i|u_j)$

Indexing People

- Figuring out $p(t|u_i)$
  - Users self-identify topics they are “experts” in
  - Others identify topics you are “experts” in
  - Topics are mined from
    - Facebook
    - Twitter
    - Homepages
    - Blogs

Indexing People

- Figuring out $p(t|u_i)$
  - In unstructured text
    - an SVM classifies the general topic
    - an ad-hoc entity parser figures out specific topics
      - scaled by TF-IDF score
  - topics are also mined from aardvark conversations

Indexing People

- The mined information is not for answering questions
- It's for identifying people who can answer questions
- Topics are enhanced with social network information

$$s(t|u_i) = p(t|u_i) + \gamma \sum_{u \in U} p(t|u)$$

Indexing People

- All of the scores for topics given a user are normalized

$$\sum_{t\in T} p(t|u_i) = 1$$

- Bayes Law is used to invert the probability

$$p(u_i|t) = \frac{p(t|u_i)p(u_i)}{p(t)}$$

Indexing People

- Probability that u_i will respond to u_j $p(u_i|u_j)$
  - Social Connections
  - Demographic similarity
  - Profile similarity
  - Vocabulary similarity
  - Chattiness similarity
  - Verbosity similarity
  - Politeness similarity
  - Speed match

Pulling topics out of a question

$p(t|q)$

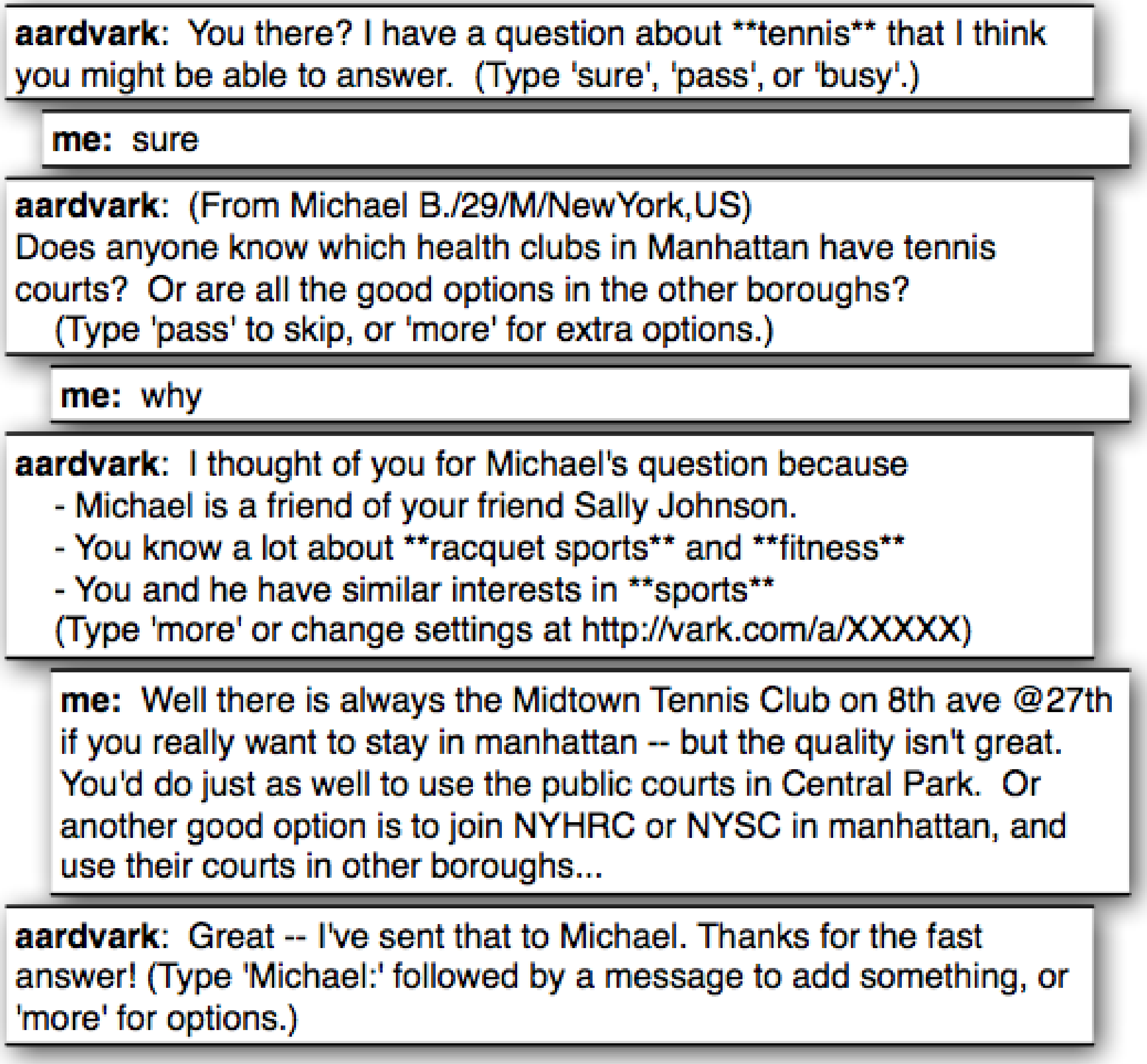
- Real-time response needed
- Doesn’t have to be perfect, people pick up the slack
- Screening
  - Is it a question? (No -> reject)
  - Is it inappropriate?Spam? Sex? Commercial? (Yes -> reject)
  - Is it trivial? (Yes -> answer it)
  - Is it location sensitive?
    - Location is treated differently than topic

Pulling topics out of a question

$p(t|q)$

- topics
  - Keyword to Topic Matcher
  - Taxonomy Topic Mapper
    - SVM classifier
  - Phrase to Topic Matcher
  - User Tag to Topic Mapper
    - If the user tags the question

User Interface



aardvark: You there? I have a question about **tennis** that I think you might be able to answer. (Type 'sure', 'pass', or 'busy'.)

me: sure

aardvark: (From Michael B./29/M/NewYork,US)
Does anyone know which health clubs in Manhattan have tennis courts? Or are all the good options in the other boroughs?
(Type 'pass' to skip, or 'more' for extra options.)

me: why

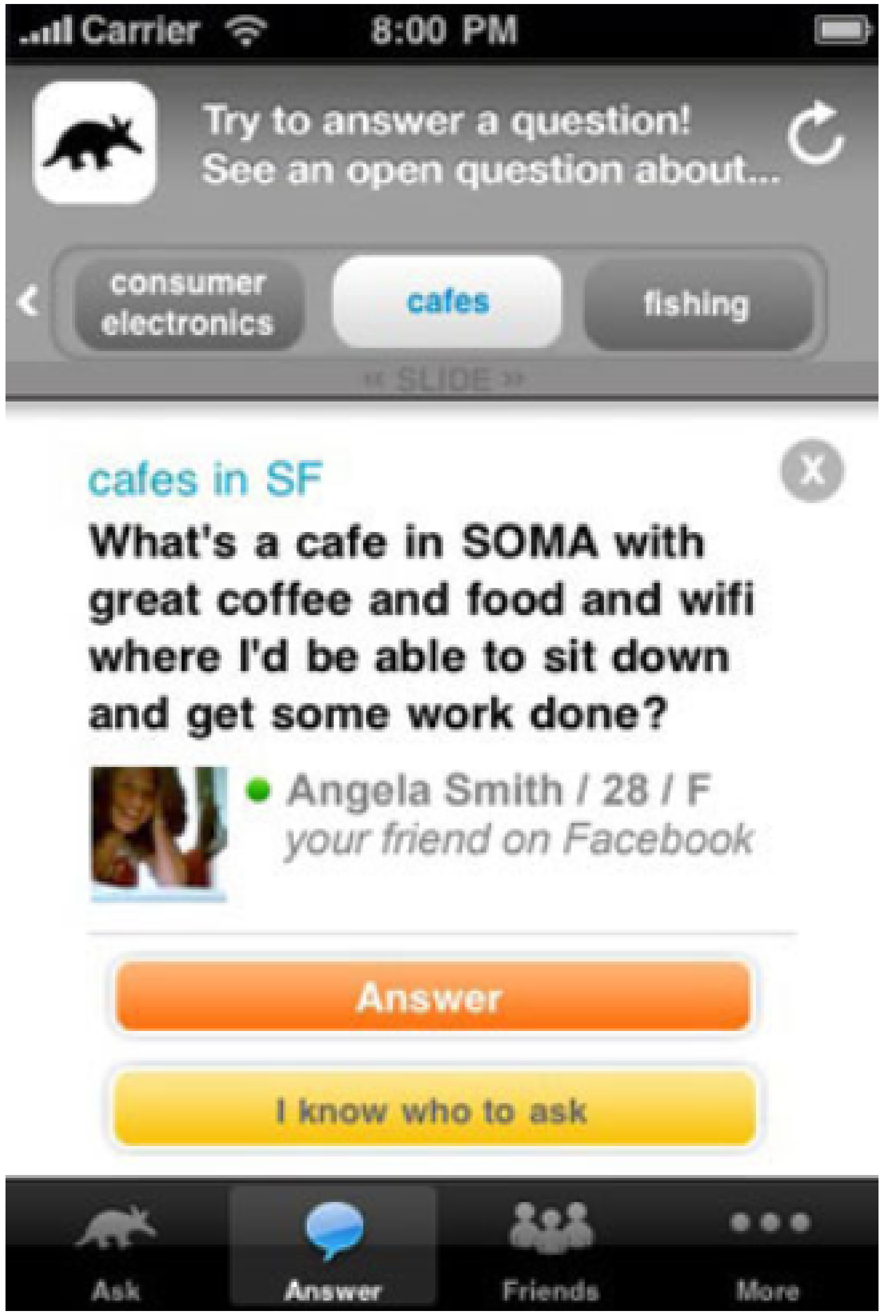
aardvark: I thought of you for Michael's question because
- Michael is a friend of your friend Sally Johnson.
- You know a lot about **racquet sports** and **fitness**
- You and he have similar interests in **sports**

(Type 'more' or change settings at http://vark.com/a/XXXXX)

me: Well there is always the Midtown Tennis Club on 8th ave @27th if you really want to stay in manhattan -- but the quality isn't great. You'd do just as well to use the public courts in Central Park. Or another good option is to join NYHRC or NYSC in manhattan, and use their courts in other boroughs...

aardvark: Great -- I've sent that to Michael. Thanks for the fast answer! (Type 'Michael:' followed by a message to add something, or 'more' for options.)

User Interface

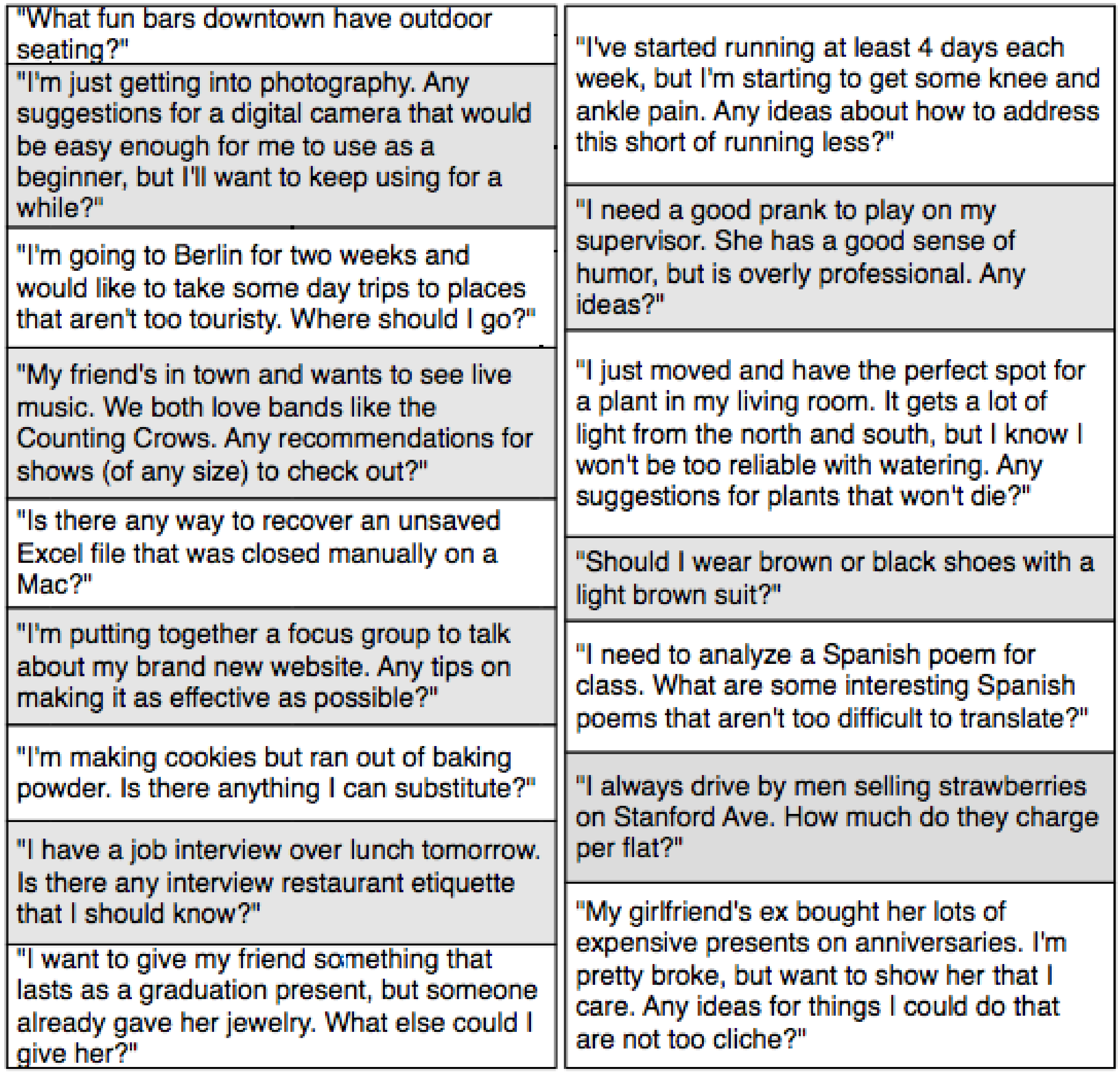


cafes in SF

What's a cafe in SOMA with great coffee and food and wifi where I'd be able to sit down and get some work done?

Angela Smith / 28 / F
your friend on Facebook

Samples



"What fun bars downtown have outdoor seating?"

"I'm just getting into photography. Any suggestions for a digital camera that would be easy enough for me to use as a beginner, but I'll want to keep using for a while?"

"I'm going to Berlin for two weeks and would like to take some day trips to places that aren't too touristy. Where should I go?"

"My friend's in town and wants to see live music. We both love bands like the Counting Crows. Any recommendations for shows (of any size) to check out?"

"Is there any way to recover an unsaved Excel file that was closed manually on a Mac?"

"I'm putting together a focus group to talk about my brand new website. Any tips on making it as effective as possible?"

"I'm making cookies but ran out of baking powder. Is there anything I can substitute?"

"I have a job interview over lunch tomorrow. Is there any interview restaurant etiquette that I should know?"

"I want to give my friend something that lasts as a graduation present, but someone already gave her jewelry. What else could I give her?"

"I've started running at least 4 days each week, but I'm starting to get some knee and ankle pain. Any ideas about how to address this short of running less?"

"I need a good prank to play on my supervisor. She has a good sense of humor, but is overly professional. Any ideas?"

"I just moved and have the perfect spot for a plant in my living room. It gets a lot of light from the north and south, but I know I won't be too reliable with watering. Any suggestions for plants that won't die?"

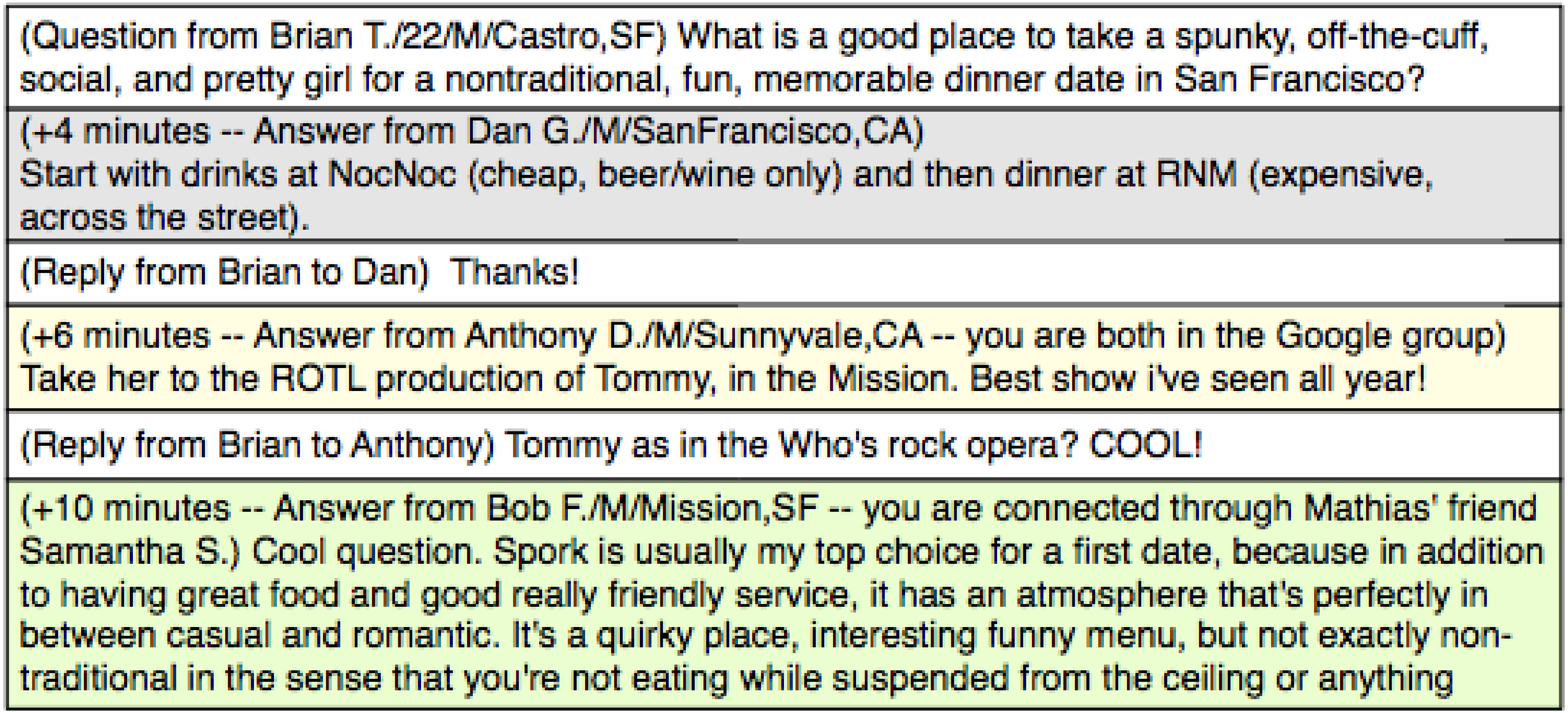
"Should I wear brown or black shoes with a light brown suit?"

"I need to analyze a Spanish poem for class. What are some interesting Spanish poems that aren't too difficult to translate?"

"I always drive by men selling strawberries on Stanford Ave. How much do they charge per flat?"

"My girlfriend's ex bought her lots of expensive presents on anniversaries. I'm pretty broke, but want to show her that I care. Any ideas for things I could do that are not too cliche?"

Samples



(Question from Brian T./22/M/Castro,SF) What is a good place to take a spunky, off-the-cuff, social, and pretty girl for a nontraditional, fun, memorable dinner date in San Francisco?

(+4 minutes -- Answer from Dan G./M/SanFrancisco,CA)
Start with drinks at NocNoc (cheap, beer/wine only) and then dinner at RNM (expensive, across the street).

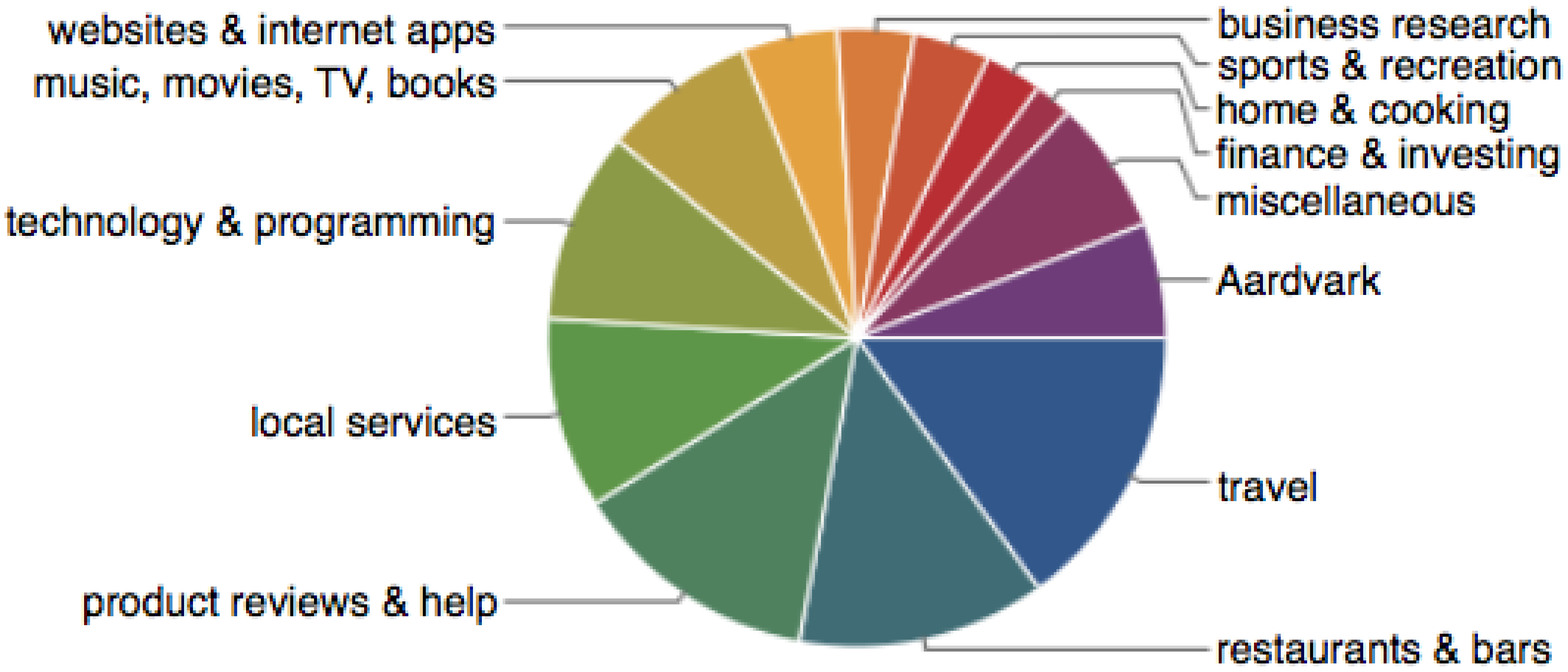
(Reply from Brian to Dan) Thanks!

(+6 minutes -- Answer from Anthony D./M/Sunnyvale,CA -- you are both in the Google group)
Take her to the ROTL production of Tommy, in the Mission. Best show i've seen all year!

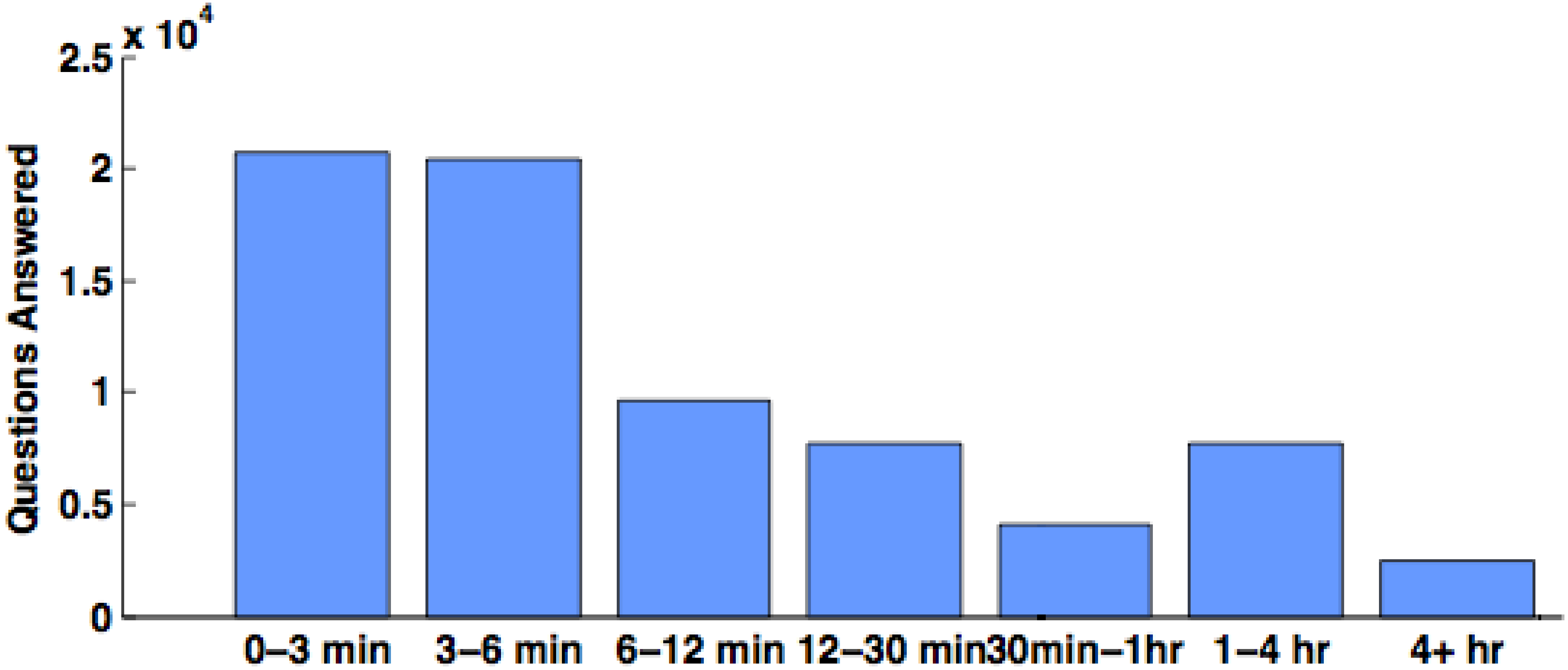
(Reply from Brian to Anthony) Tommy as in the Who's rock opera? COOL!

(+10 minutes -- Answer from Bob F./M/Mission,SF -- you are connected through Mathias' friend Samantha S.) Cool question. Spork is usually my top choice for a first date, because in addition to having great food and good really friendly service, it has an atmosphere that's perfectly in between casual and romantic. It's a quirky place, interesting funny menu, but not exactly non-traditional in the sense that you're not eating while suspended from the ceiling or anything

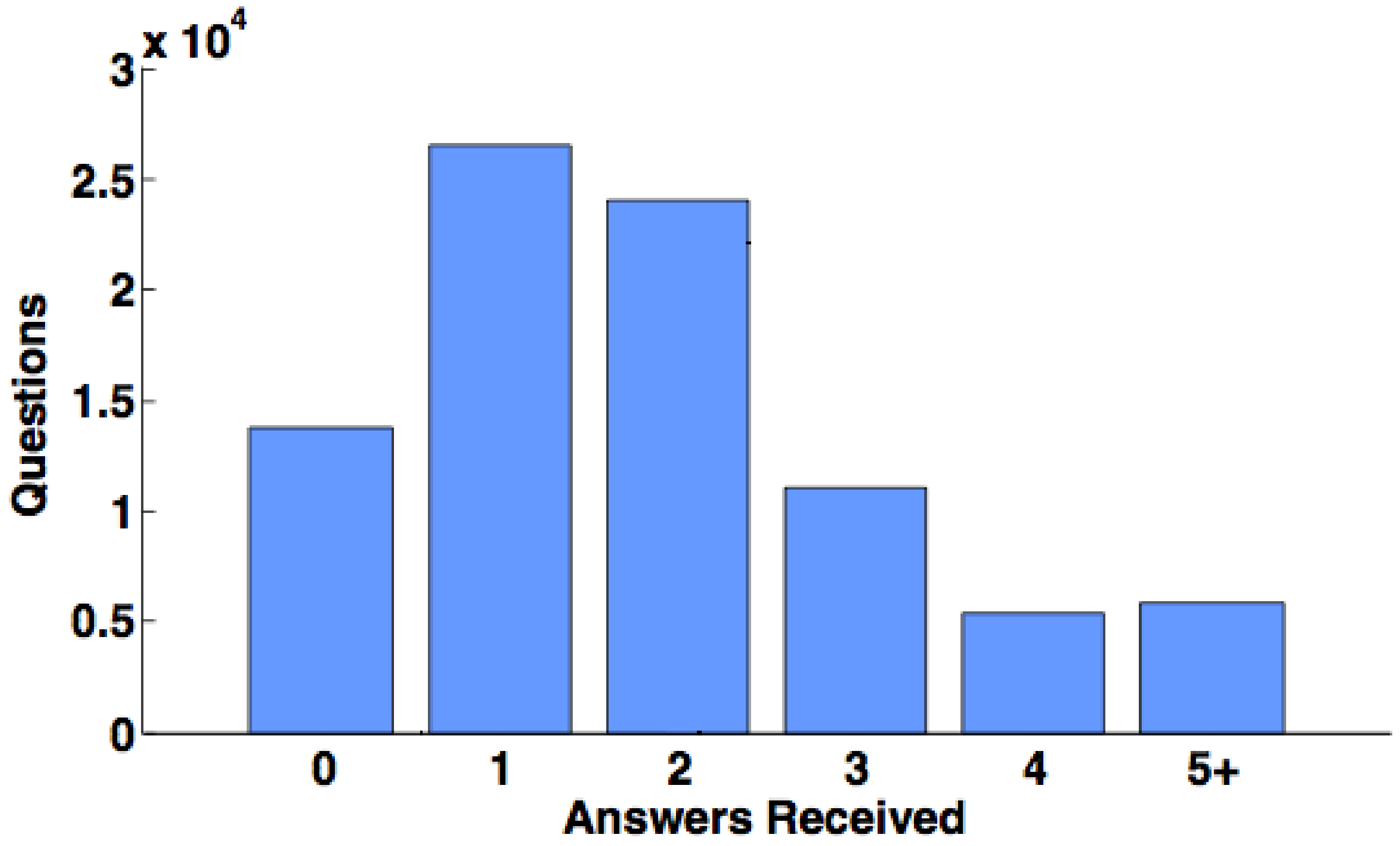
Empirical Results



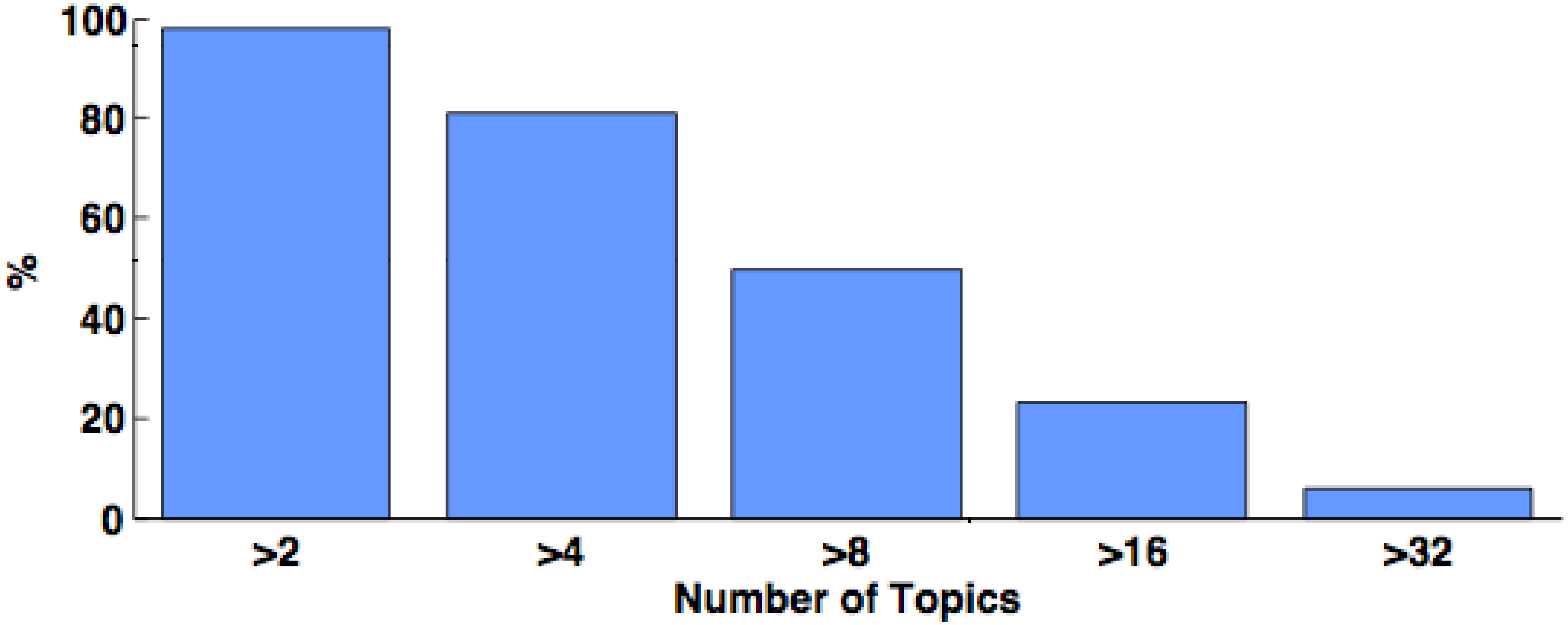
Empirical Results



Empirical Results



Empirical Results



ANALYSIS

- Aardvark is actively used
- Mobile users are particularly active
- Questions are highly contextualized
- Questions are often subjective
- Questions are answered quickly
- Answers are high quality
- Many people answer
- Social Proximity matters
- People are indexable

WESTMONT INSPIRED

COMPUTING LAB