1-31-2013

Q&A Platforms Evaluated Using Butler University Q&A Intelligence Index

Trent Ritzenthaler
Butler University

Richard Fetter
Butler University

Ginger Lippert
Butler University

White Paper

Q&A Platforms Evaluated

Using Butler University Q&A Intelligence Index

A Study by the Butler Business Accelerator - January 31, 2013

Trent Ritzenthaler,
Butler Business Accelerator

EXECUTIVE SUMMARY

A new study using the *Butler University Q&A Intelligence Index* measures how various mobile Q&A platforms deliver quality, accurate answers in a timely manner to a broad variety of questions. Based on the results of our analysis, ChaCha led all Q&A platforms on mobile devices.

Results of the study are based upon review of a large set of responses from each of the major Q&A platforms, coupled with a comparison of disparate Q&A platforms that serve answers in different ways. Our methodology included the creation of a new metric, termed the ***Butler University Q&A Intelligence Index***, which measures the likelihood that a user can expect to receive a correct answer in a timely manner to any random question asked using natural language. We asked questions via mobile services and randomized the questions to cover both popular and long-tail knowledge requests.

Butler University Q&A Intelligence Index

Rank Provider	Type	Coverage	Accuracy *	Intelligence Index
ChaCha	Human-assisted	99.4%	73.2	72.8
Ask.com	Crowdsourcing	100%	63.9	63.9
bing	Algorithm-only	100%	55.1	55.1
YAHOO!	Crowdsourcing	100%	54.5	54.5
iris.2	Virtual Assistant	98.2%	52.9	51.9
Google	Algorithm-only	100%	50.1	50.1
Answers.com	Crowdsourcing	70.8%	65.2	46.2
speaktoit	Virtual Assistant	78.6%	43.8	34.4
WolframAlpha	Structured Data	86.9%	35.9	31.2
Siri.	Virtual Assistant	43.5%	37.5	16.3
Quora	Crowdsourcing	23.8%	55.4	13.2

*Mean accuracy of responses originally graded on a 5 point scale.

Here are a few details about some of the platforms tested:

ChaCha delivered the highest quality responses consistently across the largest group of categories and question types, but did have occasional issues with Objective|Temporal and Sports questions. Ask.com performed best in the single category of questions tagged as Objective|Temporal.

Quora was proficient at answering difficult questions that require expert and extensive explanations, but it was generally unable to deliver answers within 3 minutes for most information searches on mobile devices. Quora answered only 24% of the questions at all, and often the match found did not include a viable answer.

Siri did not perform nearly as well on this random sampling of popular and long-tail questions as it did on a recent Piper Jaffray study, where results indicated that Siri correctly answered 77% of questions (Elmer-DeWitt, 2012). Our study found Siri only accurately answered 16% of the questions posed. The variance may be due to the types of questions asked and the testing conditions. Piper Jaffray notes that Siri's biggest strengths are in "local discovery and OS (operating system) commands" which were not highly represented in our study of more mainstream questions.

Google's response rate was 100%, but the first non-sponsored result on the search results page (which often times was not fully visible as an organic search result on the presented page on a mobile device) only presented an accurate answer about 50% of the time, according to the *Butler University Q&A Intelligence Index*. On a mobile phone, when accounting for the clutter of ads and the likelihood of extra clicks to achieve the answer, allowing for the answer to be within the first non-sponsored search result might be considered generous. Again, this study differs from the results found in the Piper Jaffray study, but differences are likely due to variations in methodology. For example Piper Jaffray found that Google scores highest in terms of navigation and information (Elmer-DeWitt, 2012).

This study's results support the hypothesis that Q&A platforms cannot rely on algorithmic search results alone to deliver quality answers. Search Engine Results Pages (SERP) lack deep semantic understanding of "natural language" human questions, and, therefore, cannot account effectively for long-tail questions like those posed in this study (De Virgilio, Guerra, & Velegrakis, 2012).

To achieve a score above 50% or 60% on the *Butler University Q&A Intelligence Index*, it would appear that Q&A platforms must supplement algorithmic document indexing with either:

- Utilization of structured data
- Semantic understanding via artificial intelligence (AI) or real humans

In terms of handling structured data more effectively, Google is promoting direct answers using its new Knowledge Graph and Google Now technology, "which tap into the collective intelligence of the web and understand the world a bit more like people do," (Google, 2012). The limits of Google's algorithmic technologies are evident in the empirical results of this study

and users' actual experiences. Other Q&A platforms in this study are also incorporating similar algorithmic solutions.

Improved machine learning may eventually push past the algorithmic limitations of document analysis. The efforts of DeepQA (IBM's Watson on Jeopardy) proved that intensive semantic processing can help, albeit without cost-efficient ability to scale using today's systems. While Ferrucci notes that "Computers cannot ground words to human experiences to derive meaning," Watson showed potential. The DeepQA project claimed 90% accuracy in answering 60% of questions and 70% accuracy in answering 100% of questions (Ferruci, 2010). These results translate to 54 and 70 *Butler University Q&A Intelligence Index* scores respectively.

We conclude that, without large advances in semantic processing and better utilization of knowledge graphs, Q&A platforms can benefit from the timely injection of human semantic understanding into the Q&A experience. ChaCha's top score on the Butler University Q&A Intelligence Index is an indication that just-in-time human-assisted Q&A can outperform algorithm-only solutions.

INTRODUCTION

More and more people in today's technology-driven culture are turning to the internet and smart phones to find the answers to their questions. Given this fact, the availability of accurate answers to these questions by Q&A platforms[1] is of growing importance. ChaCha Search, Inc (ChaCha) commissioned the Butler Business *Accelerator* (BBA) to conduct an independent assessment with qualitative and quantitative analyses to better understand ChaCha's competitive position with respect to Q&A platforms in the industry, as it relates to quality. We conducted a focused independent study examining comparative data across multiple Q&A platforms in the marketplace. This white paper outlines the methodology, results and conclusions of the study.

PROBLEM AND METHODOLOGY

We gathered our data in August of 2012 to assess the mobile experience when using Q&A platforms. Our study evaluated mobile applications (an "app") for each Q&A platform unless an app was not available, and in that case our analysts used a mobile website.

Our team evaluated ChaCha's mobile application services for Apple devices and ten other Q&A platforms. We selected these platforms because of their popularity and ability to provide answers to a variety of posed questions. Appendix A describes all of the Q&A platforms included in this study.

To assess the quality of the responses of these mobile Q&A platforms, we conducted mobile research, which included the following activities:

a) Posing a sample set of questions to each of the Q&A platforms using a mobile application when one was available and a mobile website when no application existed

b) Recording all responses

c) Rating the responses from each Q&A platform for coverage and accuracy

d) Analyzing data and tabulating it in a summarized format

Attributes of Quality:

For the purposes of this study, in order to evaluate the quality of answers in terms of coverage and accuracy for each Q&A platform, we defined the terms as follows:

Coverage: A binary decision based on whether a Q&A platform returned an answer within 3 minutes. For search engines, we only reviewed the first non-sponsored[2] result

[1,2] See definitions in Appendix C

provided because, on a mobile phone, the visual bandwidth is highly constrained, especially when there are advertisements and navigational elements competing for space on the page. In this study, we took the user's perspective that an "answer" must be easy to access without extra navigation.

Accuracy: The objectively determined correctness of answers, in relation to an answer key, which an independent third party developed. The answers were then scored blindly by Butler Business *Accelerator* analysts.

Measurement:

We developed a mechanism to score each Q&A platform's answer based on "coverage" and "accuracy." We then gave the answer to each question a cumulative score based on each variable.

Coverage: We determined coverage based on whether or not the Q&A platform gave an answer for each question asked.

1 No, answer not provided. If the question was not answered within 3 minutes, the analyst logged the result as "no answer."
2 Yes, answer provided within 3 minutes. The analyst logged the results as "answer provided."

Accuracy: We measured accuracy using an answer key, which a contracted independent third party created. Then, two analysts used a blind answer key, meaning that the scorers did not know which Q&A platform gave a particular answer. The analysts used a reasonableness test to determine accuracy by rating the answer using the interval scale below:

1 Incorrect answer
2 More incorrect than correct
3 Neither correct or incorrect
4 More correct than incorrect
5 Correct answer

When the answer to a question was subjective, we measured the answers using the following scale:

1 Very illogical, inappropriate or one-sided
2 More illogical, inappropriate or one-sided (rather than logical, appropriate, and impartial)
3 Neither illogical, inappropriate or one-sided or logical, appropriate, or impartial

4 More logical, appropriate, and impartial (rather than illogical, inappropriate or one-sided)
5 Very logical, appropriate and impartial

For each answer, we had a total of four accuracy scores (researcher 1/scorer 1, researcher 1/scorer 2, researcher 2/scorer 1/, researcher 2/scorer 2). We calculated the average of each of the four scores, and then used this data set for the analysis.

Question Selection:

Given the scope of this project, we tested 180 total questions during the Study. This number was large enough to be significant and compelling, yet small enough to keep the effort reasonable in scale.

Our study examined a random sample of both popular and long-tail[3] questions from mobile users. The long-tail nature of questions is described by David Ferrucci in his IBM DeepQA study: "In a random sample of 20,000 questions we found 2,500 distinct types*. The most frequent occurring <3% of the time. The distribution has a very long tail." (Ferruci, 2010).

The question corpus in this Study began with 20,000 randomly chosen questions asked by third-party users and received by either ChaCha's mobile app or IRIS during a 30-day period. From 3,000 randomly selected questions, filters removed those with sexual references or profanity and those that were confusing or incomplete. From the remaining questions, BBA again randomly sampled to arrive at 180 non-duplicative questions. For more details about the methodology BBA utilized to provide questions for this Study, see Appendix B.

Question Type and Category Assignments:

Once our project team selected 180 questions from the original corpus for testing on the various services, ChaCha then aided in classifying those 180 questions in two ways: 1) a category-based ontology that ChaCha uses as a base taxonomy and 2) a question-type matrix that looked at the type of answer required of the question and some attributes of the question.

Evaluated Questions:

Before our team completed its analyses, an analyst labeled each question in the study by its type and category in order to provide more detailed analysis. Because the Objective|Static type consisted of 63 more questions than the next highest type or category, we randomly selected 30 questions from this type to be evaluated in order to make the sample size more comparable. Using a smaller sample also helped to assure that we did not reach statistical significance simply because of a large sample size. We chose the questions to be evaluated for this type using the Excel random number generator in-between function. This assured that we selected

[3] See definition in Appendix C

the questions for evaluation completely at random. Because the function repeats numbers, we used the first 30 different randomly generated numbers for evaluation in the study. Based on this methodology, the total number of questions scored and analyzed was 168. For a full list of the 168 questions scored and analyzed, see Appendix D.

Answer Selection:

Many of the Q&A platforms gave one direct and definitive answer when posed with a question. When one direct and definitive answer was given, we evaluated the accuracy of this answer.

However, if one direct and definitive answer was not given, we scanned any information that was viewable in the first non-sponsored search result. If the answer was within a link, we only used the description of the link on the search page. Our researchers did not follow links in search of an answer.

If SIRI, IRIS or SpeakToIt automatically searched the web in response to being asked a question, we evaluated the first answer of the web search. However, if the assistant offered to perform a web search but did not search automatically, we did not evaluate the result of the web search, but we noted this response.

Testing Conditions:

- Our project team used the Apple iPhone 4 as the primary device for this study. We used this device because of its capability to search using both mobile web browsers and applications and its availability to our analysts conducting the research. The only time the iPhone 4 was not used was when needed software was unavailable on this device. When testing SIRI, we used the iPhone 4s; when testing IRIS, Google and SpeakToIt, we used an Android phone.
- We posed all questions from the same two iPhone 4 devices (with the exception of SIRI, IRIS and SpeakToIt testing) from the same location at Butler University in Indianapolis, Indiana, to ensure that internet signal strength was constant at all times.
- The same two researchers conducted all portions of the study to ensure that notation of answers was consistent.
- One question was posed to all Q&A platforms before our researcher moved on to asking the next question, so that the same question was answered by each platform within the same time frame.
- Our researchers posed all questions using the same wording. We typed all questions with the exceptions of the questions asked to SIRI, IRIS and SpeakToIt. Because SIRI, IRIS and SpeakToIt respond only to voice commands, we spoke the questions clearly using the same words that we typed to the other Q&A platforms.

FINDINGS

In order to evaluate the results of this study, our project team created the *Butler University Q&A Intelligence Index* using the scores for both Coverage and Accuracy. The *Butler University Q&A Intelligence Index* is a metric that represents the probability that a completely random search, using natural language, will return an accurate answer within 3 minutes and in a useful format.

Summary results in terms of the two metrics evaluated and the *Butler University Q&A Intelligence Index* are shown in Table 1.

Table 1: Summary Results

Butler University Q&A Intelligence Index

Rank Provider	Answers Provided	Coverage	Mean Accuracy	Accuracy *	Intelligence Index
ChaCha	167	99.4%	3.66	73.2	72.8
Ask.com	168	100%	3.19	63.9	63.9
bing	168	100%	2.75	55.1	55.1
YAHOO!	168	100%	2.72	54.5	54.5
iris.2	165	98.2%	2.64	52.9	51.9
Google	168	100%	2.51	50.1	50.1
Answers.com	119	70.8%	3.26	65.2	46.2
speaktoit	132	78.6%	2.19	43.8	34.4
WolframAlpha	146	86.9%	1.79	35.9	31.2
Siri.	73	43.5%	1.87	37.5	16.3
Quora	40	23.8%	2.77	55.4	13.2

*Mean accuracy of responses originally graded on a 5 point scale.

Coverage:

To measure coverage, our project team calculated the percentages listed in Table 1 based on the number of answers provided divided by the total number of questions.

From visual inspection of the data, the team concluded that there are not significant differences in coverage between ChaCha, Ask, Bing, Yahoo, Google and IRIS. The results also show that these Q&A platforms are superior to the other competitors listed with regards to coverage based on the percentage of questions answered.

In this case, we believe visual inspection is adequate, but difference of proportions testing to deliver inferential statistics could also be warranted.

Accuracy:

ChaCha ranks highest in terms of Accuracy as compared to its competition (see statistics in Table 2). On a 5 point scale, ChaCha was .40 points better than Answers & .466 points better than Ask, which is very strong given the sample size.

Table 2: Mean Accuracy Summary Analysis for All Questions

All Questions

One factor ANOVA

Mean	n	Std. Dev	
3.194	168	1.2349	Ask
2.754	168	1.2642	Bing
3.660	167	1.0974	ChaCha
2.506	168	1.1015	Google
2.771	40	1.2587	Quora
2.723	168	1.2709	Yahoo
3.260	119	1.3934	Answers
2.644	165	1.5336	IRIS
1.795	146	1.2369	Wolfram Alpha
2.189	132	1.2782	SpeakTolt
1.874	73	1.3399	SIRI
2.716	1514	1.3794	Total

Analysis shows that in terms of accuracy ChaCha is statistically superior to its major competitors. It is managerially superior to all competitors except Ask and Answers.

ANOVA table

Source	SS	df	MS	F	p-value
Treatment	443.3637	10	44.33637	27.36	2.05E-48
Error	2,435.6092	1503	1.62050		
Total	2,878.9729	1513			

Post hoc analysis

p-values for pairwise t-tests

		Wolfram Alpha	SIRI	SpeakTolt	Google	IRIS	Yahoo	Bing	Quora	Ask	Answers	ChaCha
		1.795	1.874	2.189	2.506	2.644	2.723	2.754	2.771	3.194	3.260	3.660
Wolfram Alpha	1.795											
SIRI	1.874	.6615										
SpeakTolt	2.189	.0099	.0900									
Google	2.506	8.71E-07	.0004	.0327								
IRIS	2.644	5.27E-09	1.82E-05	.0023	.3228							
Yahoo	2.723	1.53E-10	2.16E-06	.0003	.1180	.5700						
Bing	2.754	3.79E-11	9.19E-07	.0001	.0744	.4305	.8248					
Quora	2.771	1.84E-05	.0004	.0115	.2371	.5717	.8317	.9400				
Ask	3.194	1.10E-21	2.35E-13	1.67E-11	8.12E-07	.0001	.0007	.0016	.0591			
Answers	3.260	4.00E-20	4.02E-13	4.04E-11	8.58E-07	.0001	.0004	.0009	.0357	.6660		
ChaCha	3.660	2.30E-36	7.89E-23	1.64E-22	2.33E-16	5.64E-13	2.31E-11	9.91E-11	.0001	.0008	.0088	

Our team conducted additional in-depth analyses on the accuracy metric to better understand how ChaCha performs against the competition across different question classifications and categories. Table 3 gives a visual representation of the types and categories where ChaCha's performance was either statistically or managerially significant compared to its competition.

For Table 3, we only recognized managerial significance when a Q&A platform had first obtained statistical significance. Using this approach, the only time that there was managerial significance was when the results first cleared the statistical significance screen.

This conservative approach does not take into account the idea that managerially meaningful differences might exist, but are sometimes masked by either small sample size or high variance.

Table 3: Summary of ChaCha's Statistical and Managerial Significance Over Other Q&A Platforms

	All		Type									
			Advice| Personalized		Advice|Static		Objective| Static		Objective| Temporal		Subjective| Static	
Source	Stat.	Man.	Stat.	Man.	Stat.	Man.	Stat.	Man.	Stat.	Man.	Stat.	Man.
Answers	x				x	x						
Ask	x				x	x					x	x
Quora*	x	x									x	x
Bing	x	x					x	x			x	x
Yahoo	x	x	x	x	x	x	x	x			x	x
IRIS	x	x			x	x					x	x
Google	x	x	x	x	x	x	x	x			x	x
SpeakTolt	x	x	x	x	x	x	x	x			x	x
SIRI	x	x	x	x	x	x	x	x			x	x
Wolfram Alpha	x	x	x	x	x	x	x	x			x	x

	Category											
	Entertainment & Arts		Health		Language & Lookup		Sci Tech		Society & Culture		Sports	
Source	Stat.	Man.	Stat.	Man.	Stat.	Man.	Stat.	Man.	Stat.	Man.	Stat.	Man.
Answers												
Ask												
Quora*									x	x		
Bing	x	x	x	x	x	x						
Yahoo	x	x	x	x	x	x						
IRIS	x	x	x	x	x	x	x	x	x	x		
Google	x	x	x	x	x	x			x	x		
SpeakTolt	x	x	x	x	x	x	x	x	x	x		
SIRI	x	x	x	x	x	x	x	x	x	x		
Wolfram Alpha	x	x	x	x	x	x	x	x	x	x		

Sample size varied based on each category and type as well as the number of questions answered by each provider. *Quora data derived from a very small sample size per lack of coverage by the provider.

CONCLUSIONS

Based on the results of our analysis of Q&A platform performance related to quality results, ChaCha led all competitors in delivering accurate answers to varying types of questions asked on mobile platforms.

This conclusion is derived from an in-depth study that included the review and evaluation of a large sample set of actual responses, collected by the Butler Business *Accelerator* through the use of a defined method for comparing disparate Q&A platforms that offer answers in different ways. Our methodology included the creation and use of a unique metric we have termed the

Butler University Q&A Intelligence Index, which multiplies raw accuracy scores by coverage rates to get at the percentage of times end users received a correct answer to random questions asked using natural language. The questions we tested had been previously asked via mobile services and randomized to cover both popular and long-tail knowledge requests.

Below are a few details about some of the Q&A platforms tested:

ChaCha delivered the highest quality responses consistently across the majority of categories and question types, but did not have the top position with respect to Objective|Temporal and Sports related questions.

Quora was strong in answering difficult questions that required experts and extensive explanations, but it was unable to deliver answers within 3 minutes for information search on a mobile device. Only 24% of the questions were answered, and many times the match found did not include a viable answer.

Ask.com came closest to ChaCha's high mark on the Intelligence Index, but scored more than eight percentage points lower than ChaCha. Ask.com did score higher than ChaCha on questions from the Objective|Temporal category.

Our study found that Siri did not perform well on this random sampling of popular and long-tail questions. In clear contrast to a recent Piper Jaffray study, where results indicated that Siri correctly answered 77% of questions (Elmer-DeWitt, 2012), we found Siri only able to accurately answer 16% of questions. The variance in results of the studies is mainly due to the types of questions asked and the testing conditions. Piper Jaffray notes that Siri's biggest strengths are in "local discovery and OS commands," which were not highly represented in the randomized, empirical question sample set for this BBA study.

Google provided a response 100% of the time, but the first non-sponsored result on the search results page displayed an accurate answer only about 50% of the time, according to the *Butler University Q&A Intelligence Index*. Again, this study differs from the results found in the Piper Jaffray study, but differences are likely due to variations in methodology. For example Piper Jaffray found that Google scores highest in terms of navigation and information (Elmer-DeWitt, 2012), which may not have been tested at comparable levels in the two studies.

The results of our research support a hypothesis proposed by ChaCha that Q&A platforms cannot rely on algorithmic search results alone to deliver quality answers to mobile users. Search Engine Results Pages (SERP) lack deep semantic understanding of "natural language" human questions and, therefore, cannot account effectively for long-tail questions like those posed in this study (De Virgilio, Guerra, & Velegrakis, 2012).

To obtain a score above the 50-60% mark on the *Butler University Q&A Intelligence Index*, it would appear that Q&A platforms must supplement algorithmic document indexing either with utilization of structured data or with semantic understanding through artificial intelligence or real humans.

In terms of handling structured data more effectively, Google is providing direct answers using its new Knowledge Graph and Google Now technology, which "taps into the collective

intelligence of the web and understands the world a bit more like people do," (Google, 2012). The limits of Google's algorithmic technologies are evident in the empirical results of this study and users' actual experiences. Other Q&A platforms in this study are also incorporating similar algorithmic solutions.

Improved machine learning will eventually push past the algorithmic limitations of document analysis. The efforts of DeepQA (IBM's Watson on Jeopardy) proved that intensive semantic processing can help. While Ferrucci notes that, "Computers cannot ground words to human experiences to derive meaning," Watson showed potential. The Deep QA project claimed 90% accuracy in answering 60% of questions and 70% accuracy in answering 100% questions (Ferruci, 2010). These results translate to 54 and 70 scores for the *Butler University Q&A Intelligence Index* respectively.

We conclude that, without large advances in semantic processing and better utilization of knowledge graphs, Q&A platforms can benefit from the timely injection of human semantic understanding into the Q&A experience. ChaCha's top score on the Butler University Q&A Intelligence Index is an indication that just-in-time human-assisted Q&A can outperform algorithm-only solutions.

LIMITATIONS

This study was a limited study in that our analysts conducted research on primarily the iPhone4 (device) and not the newest version of the iPhone. However, our two analysts asked a total of 3,960 questions, which produced numerous data points for use in evaluating the performance of all key competitors in the Q&A platform space.

It is important to acknowledge that other mobile Q&A platforms are now available on devices, and newer versions of specific applications are now available. For example, a new version of SIRI is now available on the iPhone5. While these new solutions were beyond the scope of this study, the evaluation of new versions of all platforms in a subsequent study may be warranted.

The following activities were out of scope for this study but may be considered for additional potential research studies:

1. Hands-on development of, deployment of, and direct response compilation of the SMS text-based and app-based survey technology.

2. Surveys to the mobile or web-based community at large to understand and assess end user perceptions and experiences with Q&A platforms.

3. Repeat this study with newer or latest versions of technology and applications for key Q&A platforms in the industry.

REFERENCES

De Virgilio, R., Guerra, F., & Velegrakis, Y. (2012). *Semantic Search Over the Web.* Springer.

Elmer-DeWitt, P. (2012, December 20). *In Piper Jaffray re-test, Siri raises her grade from a D to a C.* Retrieved January 2, 2013, from CNN Money: http://tech.fortune.cnn.com/2012/12/20/google-now-vs-apple-siri/

Ferruci, D. (2010). *Building Watson: DeepQA@ IBM Research.* Retrieved December 20, 2012, from Whitehouse: http://www.whitehouse.gov/sites/default/files/ibm_watson.pdf.

Google. (2012, May 16). *Introducing the Knowledge Graph: things, not stringsl.* Retrieved January 2, 2013, from Google Blog: http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html#!/2012/05/introducing-knowledge-graph-things-not.html

About the Butler Business *Accelerator*

The Butler Business Accelerator, established in 2006, is an extension of the Butler University College of Business. The BBA employs full time leaders and consultants steeped in consulting experience across a vast range of industries. The BBA also utilizes both graduate and undergraduate student interns to assist in research, investigation and deliverable development. Finally, the BBA capitalizes on the breadth and depth of knowledge and experience within the faculty of Butler to provide subject matter expertise where needed and warranted.

The BBA and its staff of professionals have worked with other service-oriented companies to help bring clarity to their service offerings, bring new offerings to market, develop strategies and tactics to better market their services to customers and educate internal stakeholders on the value proposition of their service offerings. This result has been accomplished in companies in various industries, including both traditional and web-based services organizations.

The project team for this study consisted of Ginger Lippert, Butler Business Accelerator project manager, and Dr. Richard Fetter, Butler University associate marketing professor, as well as Butler Business Accelerator marketing student analysts.

The student analysts performed the research testing under the oversight of Ginger Lippert. Two student analysts engaged in the activities outlined in this study to establish "inter-rater reliability." Two different student analysts conducted the scoring process for the accuracy portion of the study, so that the scoring for all questions could be compared.

Ginger Lippert and the student analysts conducted the analysis portion of the study with the guidance and review of Dr. Fetter.

Appendix A: Q&A Platforms

Q&A Platform	Description	Type	Tested using
Answers.com	A question and answer website that gives direct answers to posed questions. Researchers used the mobile web version of Answers.com because no mobile application is available.	Crowdsourcing[4]	Mobile web
Ask.com	A question and answer website and app that provide direct answers to posed questions.	Crowdsourcing	Application
Bing.com	A search engine that provides direct answers to some of the most commonly asked questions and also provides a multitude of links to other websites.	Algorithm-only[5]	Application
ChaCha	A question and answer website and app that provide direct answers to posed questions.	Human-assisted[6]	Application
Google.com	A search engine that provides direct answers to some of the most commonly asked questions and also provides a multitude of links to other websites. Researchers used the voice recognition software of the Google Search application in order to test the latest and most advanced Google technology.	Algorithm-only	Application
IRIS	A personal assistant app for Android that responds to voice commands and answers questions.	Virtual Assistant[7]	Application
SIRI	An "intelligent assistant" that responds to voice commands and answers posed questions on select Apple devices.	Virtual Assistant	Built-in
SpeakToIt	Virtual buddy app for smart phones that answers questions in natural language, performs tasks, and notifies the user of important upcoming events.	Virtual Assistant	Application
Quora.com	Website and app that deliver answers and content from people who share similar interests.	Crowdsourcing	Application
Yahoo Answers	A search engine that gives direct answers to some of the most commonly asked questions and also provides a multitude of links to other websites.	Algorithm-only	Application
Wolfram|Alpha	Website and app that deliver answers to questions based on a collection of built-in data, algorithms and methods.	Structured Data[8]	Application

[4, 5, 6, 7,8] See definitions in Appendix C

Appendix B: Methodology for Question Selection

Our BBA project team used the following methodology to select questions for this study:

The corpus started with two sets of data, randomly sampled from incoming questions over a 30-day period using Hadoop queries[9]. The first set was from the Android app called "IRIS" from Dexetra, Inc. The second set was from incoming questions via the iPhone mobile app called "ChaCha" from ChaCha Search, Inc. The only requirement for a question to be randomly selected was that the query needed to start with an interrogative word such as "Who," "What," "When," "Where," "Why" or "How."

Initially, at the direction of our BBA project team, ChaCha sampled 10,000 questions from each source for a total of 20,000. From this set, ChaCha, with our direction, then discarded questions containing sexual references and profanity using a Regular Expressions-based filter. Then ChaCha, at our direction, re-randomized the data to select only 3,000 questions from the 20,000 corpus using an Excel randomization function. These 3,000 questions were then manually reviewed by two analysts who culled from the corpus only questions that 1) were not complete, understandable questions, 2) contained significant grammatical errors, 3) were duplicative or extremely similar to questions already in the resultant corpus.

From the resultant corpus, our BBA project team then randomized the questions again and selected the first 1000 items.

We then selected 180 questions from the corpus of 1000 items using the Excel random number generator in-between function. This technique assured random selection. We selected the first 180 non-duplicated questions using this method.

[9] See definition in Appendix C

Appendix C: Definitions

1. Q&A platform: A provider of answers to natural language queries.

2. Non-sponsored: A result not being displayed as a form of advertising.

3. Long-tail: Not focusing only on popular questions but rather the full universe of possible questions. For example, on one graph showing questions asked, roughly 3% were popular questions showing high volume while roughly 97% were low-volume questions that together resembled a long tail.

4. Crowdsourcing: A type of Q&A platform that primarily solicits answers from a large group of users. In most cases, the question and answer are asynchronous and the transactions are completed on the web.

5. Algorithm-only: A type of Q&A platform that primarily uses a set of computer instructions to search for specific data.

6. Human-assisted: A type of Q&A platform that primarily relies on human intervention to supplement algorithmic results.

7. Virtual Assistant: A type of Q&A platform primarily designed for voice use on a smartphone device. The user interface is generally a virtual personal helper, that attempts to understand user intent across a variety of functions including Q&A.

8. Structured Data: A type of Q&A platform that primarily delivers computational results from large sets of structured sets, generally based upon an entity–attribute–value model.

9. Hadoop query - A command sent to a data system that allows the storage and manipulation of very large sets of data to, in this case, randomly sample a group of questions from a larger corpus.

Appendix D: Study Questions Scored and Analyzed

The following questions were scored and analyzed as part of the Study.

Type *	Category *	Questions
Advice|Personalized		Why do i have to go to work?
Advice|Personalized	Society & Culture	What do you know about love?
Advice|Personalized		What things can i do in portland?
Advice|Personalized	SciTech	Whats better apple or android?
Advice|Personalized	Society & Culture	What do you do for fun with your boyfriend?
Advice|Personalized		If i was getting a room what should i get?
Advice|Personalized	Society & Culture	What are ways I could make my boyfriend happy?
Advice|Personalized	Society & Culture	Should i talk to my former fiance?
Advice|Personalized		Can you slap contractors in the face?
Advice|Personalized		What should i order from wendy's?
Advice|Personalized	Society & Culture	What should i get my wife for our anniversary?
Advice|Personalized	Health	What should i do about my chronic gas?
Advice|Personalized	Entertainment & Arts	how do I get married on skyrim?
Advice|Static	Society & Culture	What if a girl doesn't want to talk to you?
Advice|Static		What is a safe methed to bleach your hair at home?
Advice|Static	SciTech	What is the best cellphone?
Advice|Static	Health	What are some tricks to falling asleep faster?
Advice|Static	Health	What is the best age to get pregant?
Advice|Static	Health	Whats a great way to sleep through the night?
Advice|Static	Entertainment & Arts	Whats a fun game app to download?
Advice|Static	Health	What's the best temperature to go swimming in?
Advice|Static	SciTech	What dog is the best fighter?
Advice|Static		Whats the best foundation for broke out skin?
Advice|Static	Society & Culture	What do you do if a zombie attacks you?
Advice|Static	Health	Whats is a good diet to help lose 50lbs fast.
Advice|Static	SciTech	What type of cd does it take to burn a video game?
Advice|Static	Health	Should toddlers take naps?
Advice|Static		What removes white board marker from a fridge surface?
Advice|Static	Health	how can i get rid of my hangover!
Objective|Static	Entertainment & Arts	Who starred in the movie the dark knight?
Objective|Static	Entertainment & Arts	What is Channing Tatums phone number??
Objective|Static	Language & Lookup	What's the plural for flute?
Objective|Static		Who makes the best best cake?
Objective|Static	Entertainment & Arts	Whats nickolodeon slime made from?
Objective|Static	SciTech	Why do kittens sleep so much?
Objective|Static	Entertainment & Arts	Who are the sirens in the odyssey?
Objective|Static	Language & Lookup	Why do they call it dead time?

Type *	Category *	Questions (continued)
Objective|Static	Society & Culture	Who killed martin luther king?
Objective|Static	Society & Culture	Who found the dominican republic?
Objective|Static	Health	Can you overdose on methadone?
Objective|Static	Society & Culture	Who is benedict arnold?
Objective|Static	Language & Lookup	What is the jersey shore name for anthony?
Objective|Static	Health	Whats a fast way to lose weight?
Objective|Static	Entertainment & Arts	Can Americans participate in the X factor?
Objective|Static	Entertainment & Arts	What is spiderman 3?
Objective|Static	Entertainment & Arts	What zombie movie was filmed in rome, ga?
Objective|Static	SciTech	Whats the horsepower of a 2005 chevy 5.3?
Objective|Static	SciTech	What kind of pets have suicidal tendencies?
Objective|Static	Health	When does a baby start to sit up?
Objective|Static	Health	Why do we bleed?
Objective|Static	Entertainment & Arts	Who formed the blazing sword?
Objective|Static	Language & Lookup	What does it mean when people say they're on cloud 9?
Objective|Static	Health	What is the average height of a man?
Objective|Static	Entertainment & Arts	What are the lyrics to mercy?
Objective|Static	Entertainment & Arts	What is the name of the song at the beginning of the credits in the 2011 movie colombiana?
Objective|Static	Entertainment & Arts	When does the new family guy season start?
Objective|Static	Language & Lookup	What is a jewish guilt?
Objective|Static	Entertainment & Arts	Who is patrick star?
Objective|Static	SciTech	What is the biggest catfish caught?
Objective|Static		What year did president nixon resign?
Objective|Static	Entertainment & Arts	Who is david beckham?
Objective|Static	Health	Can you get drunk on oil?
Objective|Static	Society & Culture	Who was the leader of russian during ww1?
Objective|Static	Entertainment & Arts	What happens to Naruto?
Objective|Static	Language & Lookup	What are the clothing articles?
Objective|Static	Sports	Who is the best golfer?
Objective|Static	Society & Culture	What happened to flight 93 on 911?
Objective|Static	Entertainment & Arts	What football team did michael jordan play 4?
Objective|Static	Society & Culture	Who was the first dog to go to space?
Objective|Static	Language & Lookup	What does the name brandon mean?
Objective|Static	Society & Culture	What does it mean when a girl tickles a guy?
Objective|Static	Language & Lookup	What are whirlpool bathtubs?
Objective|Static	Society & Culture	Who was adolf hitler?
Objective|Static	Language & Lookup	What is a real doll?
Objective|Static	Language & Lookup	What does the name Michael mean?
Objective|Static	Entertainment & Arts	Who is hector's wife in the iliad?

Type *	Category *	Questions (continued)
Objective|Static	Sports	What is quinton rampage jackson full name?
Objective|Static	Entertainment & Arts	Who is ronnie radke?
Objective|Static	Entertainment & Arts	Who is the voice of peter, stewie and brian griffin?
Objective|Static	Language & Lookup	What are the chances of winning the lottery?
Objective|Static	Entertainment & Arts	What is the plot of unbroken?
Objective|Static	Society & Culture	Who is honest abe?
Objective|Static	Health	What is being hypoglycemic?
Objective|Static	Health	Can dogs survive parvo?
Objective|Static	Sports	Who is the world's strongest man?
Objective|Static	SciTech	What is facebook?
Objective|Static	Language & Lookup	What does the name alyssa mean?
Objective|Static	Health	Can a persons eyes slightly buldge?
Objective|Static	Health	Can you get a small pimple from a clam that itches?
Objective|Static	Language & Lookup	What is capitalistic democracy?
Objective|Static	Language & Lookup	What is dog water?
Objective|Static	SciTech	What is the difference between direct current and alternating current?
Objective|Static		What were the origins of world war I?
Objective|Static	Language & Lookup	If your 40 time is a 4.9 how fast are you running?
Objective|Static	Health	What causes autism in children?
Objective|Static	Society & Culture	Who did the us win its independence from july 4th?
Objective|Static	Language & Lookup	Who is veronica moser?
Objective|Static	Language & Lookup	What is the average score on the ap lit exam?
Objective|Static	Entertainment & Arts	What is the name of all the disney princesses?
Objective|Static	Entertainment & Arts	Did the tooth fairy 2 make it to the movie theaters
Objective|Static	Society & Culture	What is a Sikh temple?
Objective|Static	SciTech	Did we land on mars?
Objective|Static	Sports	Who won the 2012 Olympic gold medal in women's soccer
Objective|Static	Sports	What is a heptathalon?
Objective|Static	Sports	Who medaled in cycling in the 2012 London Olympics
Objective|Static	Entertainment & Arts	What's wiz Khalifas real name
Objective|Static	Entertainment & Arts	What is fifty shades of grey about?
Objective|Static	Sports	How many medals does Michael phelps have?
Objective|Static	SciTech	Has the higgs-boson particle been found yet?
Objective|Static	Sports	How many medals does the usa have in the 2012 Olympics?
Objective|Static	Sports	How Many Medals Did The US Get On The First Day Of Olympics 2012?
Objective|Static	Health	What are the symptoms of pregnancy?
Objective|Static	Entertainment & Arts	When does the new One Direction album come out?
Objective|Static	Sports	Where will the 2020 Olympics be
Objective|Static	Entertainment & Arts	How old is Alexandria Raisman?

Type *	Category *	Questions (continued)
Objective|Static	Entertainment & Arts	When is grand theft auto 5 coming out
Objective|Static	Sports	How old is michael phelps
Objective|Static	Entertainment & Arts	Why is the movie "Ted" rated R?
Objective|Static	Entertainment & Arts	What is the book 50 shades of grey about?
Objective|Static	Entertainment & Arts	when are the 2012 teen choice awards
Objective|Static	Language & Lookup	What is skeet shooting?
Objective|Static	Society & Culture	When did the shooting in colorado happen?
Objective|Temporal	SciTech	What's the cheapest Mazda?
Objective|Temporal	Sports	What high school has the biggest football stadium?
Objective|Temporal	Language & Lookup	What is the most common Italian girl name?
Objective|Temporal	Sports	Who is the number 1 ranked golfer in the world?
Objective|Temporal	Language & Lookup	What is the most popular name for a girl?
Objective|Temporal	Language & Lookup	What's the price of copper?
Objective|Temporal	Language & Lookup	What are the top 10 most common names?
Objective|Temporal		Who is the prime minister of england?
Objective|Temporal	Sports	Whos the qb for the 49ers?
Objective|Temporal		Who is the current president of united states?
Objective|Temporal	Entertainment & Arts	When is snooki's baby due
Objective|Temporal	Society & Culture	When is halloween
Objective|Temporal	Entertainment & Arts	When is Breaking Dawn Part 2 coming out?
Objective|Temporal		When does summer end
Objective|Temporal		How much is minimum wage?
Objective|Temporal		How much is a ticket for cedar point ?
Subjective|Static		What are some good margarita recipe?
Subjective|Static	Entertainment & Arts	What's the best tv show ever?
Subjective|Static	Health	Why are outlets cheap?
Subjective|Static	Language & Lookup	Whats a cute two word saying?
Subjective|Static	Society & Culture	Who makes a woman happy?
Subjective|Static	Entertainment & Arts	Who would win in a fight the hulk or superman?
Subjective|Static	SciTech	Why is holden better than ford?
Subjective|Static	Society & Culture	Can people with hearing loss have a relationship?
Subjective|Static	Entertainment & Arts	What's the best black keys song ever?
Subjective|Static	Society & Culture	What are goodluck signs?
Subjective|Static	SciTech	Why don't cdma based cellphone cariers use sim cards?
Subjective|Static		What hotels in michigan are haunted?
Subjective|Static		Why is chick fil a's mascot a cow?
Subjective|Static		Why are cigarettes legal?
Subjective|Static	Language & Lookup	What does it mean to be tied down?
Subjective|Static	Society & Culture	When is the zombie apocalypse going to happen?
Subjective|Static	Language & Lookup	What does the air force do?

Type *	Category *	Questions (continued)
Subjective|Static	Society & Culture	What year will the world end?
Subjective|Static	SciTech	Why do stars only come out at night?
Subjective|Static		What is the most illegal thing to do in the world?
Subjective|Static	Language & Lookup	what's the meaning of life
Subjective|Static		Mitt Romney or Barack Obama?
	Entertainment & Arts	Who is the best rapper alive?
	Health	What can make your appendix hurt?
	Language & Lookup	If two planes are parallel to the same plane, are they parallel to each other?
	Health	Why do i have muscles and i don't work out?
	SciTech	Can i repost someone elses picture on instagram?
	SciTech	Whats a website you can read books on?
	Language & Lookup	What's the weather like for this weekend in priceville new york?
	Language & Lookup	What's the current weather like in surrey bc?

* Not all questions could be categorized into a specific Type or Category (intentionally left blank in the table)