Feasibility study on the potential to gain insights on Kenyans’ access to finance through analysis of big data

USAID DEVELOPMENT CREDIT AUTHORITY & UN GLOBAL PULSE

SEPTEMBER 2013
# TABLE OF CONTENTS

3 Introduction  
5 Background on Digital Landscape in Kenya  
8 Trends in Social Media  
25 Google Trends  
29 Finding the Digital Footprint  
32 Recommendations and Next Steps

**ANNEX 1: Survey Questionnaire**

---

**Acknowledgements:**
The lead researcher and author of this report was Eva Kaplan (eva.kaplan@gmail.com). It was prepared with the support of the team at the United Nations Global Pulse (www.unglobalpulse.org), the USAID Development Credit Authority (DCA), and USAID-Kenya.

Additional acknowledgements are due to Anoush Tatevossian, René Clausen Nielsen, and Robert Kirkpatrick at the UN Global Pulse; Stephanie Grosser at DCA; and Linda Kagota at USAID-Kenya. Thank you also to Esther Ajambo for implementing the survey and Clarise Aduma from Kenya Commercial Bank for generously working with us to find clients to survey. The study was made possible through a Learning Improvement grant through USAID Training Support Fund.
INTRODUCTION

The goal of this study was to better understand how big data can support the work of the Development Credit Authority at USAID. This is a feasibility study and as such is intended to point to areas that might be appropriate for further investigation. As an initial foray, the analysis was confined to Kenya, and is targeted specifically to the question “What barriers to accessing loans do small businesses in Kenya face?” It is important to note that the purpose of this exercise is not primarily to answer that question, but rather to determine the feasibility of answering such a question using new sources of digital data. The purpose of this report is to paint a picture of the big data landscape in Kenya, show some preliminary findings, and lay the groundwork for further investigation by highlighting possibilities and challenges.

The report has two primary goals:

Describe the digital landscape of Kenya: Section I of the report lays out baseline data of ICT and social media use in Kenya with a particular emphasis on understanding gaps in baseline information.

Describe big data in relation to loans in Kenya: Section II and III dig in to available data—namely Twitter data and Google search trends—to better understand online conversations and information seeking behavior. Section IV takes the opposite approach, looking not at what data is currently available but rather at the digital footprint of DCA clients and priority sectors. Section IV first summarizes information collected from DCA clients themselves, and then uses two case studies—agriculture services and bank websites—to look at where else relevant digital signals might exist in Kenya.

What is Big Data?

“Big data” is typically defined as terabytes or petabytes of data that is generated as people go about their daily lives in the digital age. Big Data is characterized by the “3 Vs:” greater volume, more variety, and a higher rate of velocity. This includes information seeking behavior (i.e. Google searches), social data (i.e. Twitter or Facebook data), mobile phone data, and digital services (i.e. Amazon purchasing or mobile banking).

Big Data is both the information that is passively generated as by-products of people’s everyday use of technologies and the information people willingly communicate about themselves on the web. Even when individuals do not have direct access to mobile phones or other technologies, they may still be passively emitting information as they go about their daily lives (e.g., when they make purchases, access services, or when they interact with better-connected members of the community). These digital trails, or “smoke signals” can reveal changes in trends in our collective preferences, struggles and overall well-being. Big Data for Development refers to Global Pulse’s effort to determine how big data can be used to better inform the policy and planning of development programs.¹

In general, sources of Big Data that can be useful to development are those that can be analyzed to gain insight into to human well-being and change. A preliminary categorization of sources may reflect: WHAT PEOPLE SAY (i.e., international and local online news sources, publicly accessible blogs, forum posts, comments and public social media content, online advertising, e-commerce sites and websites created by local retailers that list prices and inventory) or WHAT PEOPLE DO

¹ For reference, UN Global Pulse’s introductory guide, “Big Data for Development: A Primer,” is available online and for download at: http://unglobalpulse.org/bigdataprimer
(i.e., passively collected transactional data from the use of digital services such as financial services (including purchases, money transfers, savings and loan repayments), communications services (such as anonymized records of mobile phone usage patterns) or information services (such as anonymized records of search queries).)

In some cases, the volume of data which could be analyzed to augment development planning may not be technically be “big” (for example, many of the digital services listed in the agriculture case study in Section IV do not generate massive amounts of data, relatively speaking). In these cases it might be more appropriate to use the label “new data” or “digital data.” However, the term Big Data has evolved to be a catch-all phrase describing the new age of digital data, and the innovations in technologies and methodologies now available to collect, store, and make sense of it all.

**Methodology & Purpose of Feasibility Study**

In order to fill in gaps in our understanding of the big data landscape in Kenya, the study was undertaken in different phases, which are reflected in the format of the report.

First, the report summarizes who in Kenya is generating digital data, looking specifically at what is known about the penetration rates and demographics of mobile, Internet and social media use in the country. In addition, to find out, how DCA’s target population (underserved entrepreneurs) generates digital data we did a non-representative survey of 10 DCA clients. All the clients were from Kenya Commercial Bank, and all of them were farmers. They were all from a similar region of Kenya, between Nakuru and Nairobi. The clients were asked about (1) words they use to talk about loans in informal and formal chatter; (2) how they use mobile phones, computers and social media; and (3) how they find information about their business, their existing loans, and new loans.

Second, the report explores finance-related data generated through use of the Internet. This analysis primarily relies on Crimson Hexagon’s proprietary social media analysis platform, called **Forsight™.** We also make use of the publically available tools on **Google Trends.**

Finally, in the last section we provide examples of other big data methodologies that might be useful to DCA, exploring what other types of digital data exist.

This report is intended to inspire new thinking in how DCA can use new sources of digital data to inform its work. Big data may help researchers ask the same types of questions they have traditionally researched in other ways—for example, by better understanding people’s behaviors based on observation. However, it may also facilitate entirely new ways of thinking about issues. For example, big data is often particularly well suited to understanding networks or information seeking behavior. In addition, it is worth noting that while some big data may be well-suited to track individual behavior, Global Pulse recommends that at this stage responsible uses of big data analytics for global development should target aggregate trends rather than individuals to protect the privacy of individuals.

Being imaginative about big data is thus important, in two key ways—first, being imaginative about where to “look” for relevant data sources; and second, being imaginative about how to use the information. While big data might not provide a silver bullet answer to key questions faced by DCA, it might be useful in ways that were not anticipated.

---

2 Excerpted from “Big Data for Development: A Primer” (June 2013) [http://unglobalpulse.org/bigdataprimer](http://unglobalpulse.org/bigdataprimer) (page 3)

3 For more information on UN Global Pulse’s privacy and data protection principles and recommendations, please visit: [http://unglobalpulse.org/privacy-and-data-protection](http://unglobalpulse.org/privacy-and-data-protection)
BACKGROUND ON KENYA’S DIGITAL LANDSCAPE

Over the past 5 years, Kenya has emerged as a tech leader in Africa. The rise of mobile phone use in Kenya is a well-told story, and the mobile money service M-PESA has changed the way many entrepreneurs globally think about providing services, from finance to health to energy provision.

In order to understand the implications of the Kenyan digital revolution for big data analytics, it’s important to understand in greater detail the demographics behind these trends. However, despite the great deal of interest in Kenya’s digital landscape, the exact make-up of who is accessing services is not very well understood. The purpose of this section is to explore what is known and not known about the demographics behind Kenya’s digital revolution.

Mobile and internet penetration
According the Communications Commisson of Kenya, as of December 2012, mobile phone coverage in Kenya had reached a total of 30.7 million subscriptions, or 78.0% of the adult population. This represents a growth of 19.6% percent since 2009, when 49.1% of the adult population had a mobile phone. As of September 2012, it was estimated that only 7% of these phones were smart phones, but this sector is projected to grow rapidly, with Safaricom’s launch of a new smartphone in January 2013 selling out in less than two weeks after launch. Data for internet accessed through the mobile networks can be purchased using the same pre-paid card that is used for mobile phones.

Internet, by contrast, stood at 9.4 million subscriptions as of December 2012. While this number clearly lags significantly in comparison with mobile penetration, it represents a growth of 75.1% of internet subscriptions over the same period the previous year. The vast majority of subscribers access the internet via mobile networks, representing 99% of total internet subscriptions. This includes both those who access the internet directly on their phones, and those who use mobile plug-ins to computers. Including non-subscribers, it is estimated 16.2 million Kenyans, or 41.1% of the population, were accessing internet by December 2012.

Social Media
Data on social media is less readily available. In a survey done in 2009 by AudienceScapes, 20% of those surveyed reported having used internet in the past year. Of those who had used internet, 71% reported using some sort of social media, and 50% reported using social media at least once a week.

A study by Portland Communications in 2011 which analyzed 3 months of tweets geolocated to sub Saharan Africa (representing 11.5 million tweets) showed Kenya as the second most prolific African country on Twitter, after South Africa. Africa-wide trends show that 81% of these “tweeps” are using Twitter to engage in social conversations. Sourcing information about news was also highly ranked. Indeed, across the continent, many Twitter users reported that Twitter has become a primary source of information for them, particularly for international news.

---

According to Facebook, as of May 2013, there were 1,934,980 users in Kenya. These represent those with Facebook accounts who have identified themselves as living in Kenya.

**Mobile Services**
Globally, Kenya stands out as a hub of innovative mobile services. These services are highly relevant to big data analytics because they leave digital trails. As people interact with mobile services, they are generating valuable data on anything from saving rates to food prices to health problems. This data is held privately by service providers.

M-PESA—or mobile money—is the most salient example of Keyna’s mobile services. M-PESA allows people to transfer money via their mobile phones. A plethora of financial services have also been built on top of the M-PESA platform. These have largely failed to have broad appeal. However, Safaricom’s new paperless banking service M-Shwari is experiencing rapid growth. Launched in November of 2012, Safaricom reported that M-Shwari had 1.6 million customers after just 2.5 months. M-Shwari allows customers to save money and collect interest. It also allows them to access emergency loans.

In addition to M-PESA related services, other types of information is transmitted via mobile phone. For example, information systems such as mFarm or iCow allow people to share information on their farming practices over the phone.

As mentioned, this data is not publically available. However, Section IV will explore the potential of analyzing the data generated by these services, as well as what would be required to access them.

**Demographics**
If big data is to be used for designing policy or programs—for example, for understanding unmet financing needs—knowledge of who is represented in the data would be paramount. While information on ICT penetration is published quarterly, the demographics of who is using ICT is published in an ad-hoc way, typically by researchers doing a one-off survey. Uptake of new technology is Kenya is happening faster than this type of research is apparently keeping up with.

Some impressive attempts to better understand the demographics of mobile use have been made. In 2009, a Financial Sector Deepening Kenya (FSDK) survey in 646 communities in Kenya sought to give shape to mobile trends, gathering demographic data on phone ownership and phone usage and sharing. In general, counties with higher urban populations reported higher phone ownership.

The survey found that there was some level of mobile phone ownership in every income bracket, although as might be expected there was a higher concentration of ownership at higher income levels. It was also found that women are more likely to be phone sharers than men. Among those who had never used a phone, the vast majority (81%) were women.

Except for those under 17 and those over 60, the survey found comparable rates of phone ownership across age brackets.7

While it appears that in the past year or so growth in mobile phone access may be leveling off, growth in other ICT sectors—for example smart phones— continues to expand rapidly. Thus, it is unclear how relevant evidence from 2009 is for today’s digital landscape.

---

Demographic information on Twitter usage is not clear and much less readily available. Again, ad-hoc studies for particular purposes may shed light on Twitter demographics. For example, iHub Research is seeking to uncover trends on who was tweeting during the 2013 elections. The study, once completed, is likely to be illustrative; however it does not represent a comprehensive assessment of demographic trends of Twitter users.

Relevance for big data analytics

All data has bias, and big data is no exception. However, without up-to-date demographic information, it is not possible to understand biases in the data. There might be some areas of investigation where having demographic trends is not critical—for example, bank-specific information like interest rates or different types of loan instruments. However, for some analytical questions, gathering baseline information on DCA clientele may be important.

The survey done for the purposes of this report is instructive. While it only covered 10 clients, all farmers from the same bank, it does offer some clues as to the digital footprint of DCA clients. The table below describes how many respondents use digital tools.

<table>
<thead>
<tr>
<th>Owns a mobile phone</th>
<th>Owns a computer</th>
<th>Has accessed internet on mobile phone</th>
<th>Has accessed internet</th>
<th>Has Facebook Account</th>
<th>Has discussed loan or business on Facebook</th>
<th>Has a Twitter Account</th>
<th>Uses Mobile Banking</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 1: Number of DCA clients who reported using digital tools.

While it is impossible to extrapolate broader trends from this data, this indicates that social media may be limited in its ability to represent these clients.
TRENDS IN SOCIAL MEDIA

Through an existing partnership with the company Crimson Hexagon, Global Pulse is able to use Forsight™, a social media monitoring and analysis platform, which allows access the full “Twitter firehose” (that is, all Tweets being published today, in real-time), and archival Twitter data going back to 2009. With the ForSight™ platform, it is possible to build “monitors,” which filter the billions of Twitter posts according to particular specifications. For this analysis, we built monitors to do the following:

- Estimate the country of origin of each Twitter post, and narrow the analysis to Tweets originating from Kenya
- Filter out Twitter posts with content/keywords related to the specified issue area (in this case, loans)
- Sort the posts according to specified sub-categories (for example “business loan,” “personal loan,” etc.)
- In some cases, we also did “sentiment analysis,” sorting posts according to positive and negative sentiments.

In this section, we describe the step-by-step process of building the monitors, along with analyzing the results. The process of building a monitor involves first developing a list of keywords to filter out relevant Twitter posts. Then, subcategories are defined and refined based on available data. The last step is to “train” the monitor to automatically categorize based upon the specifications. This involves hand-sorting roughly 20 posts per sub-category as model examples. The machine will then automatically categorize the remainder of the posts (and capture new Tweets as they are published in real-time) once it is trained. To provide a clear demonstration of this process, the process of building and analyzing the first monitor is described in depth.

To analyze the data, this report largely relies on the tools available in ForSight™. After the monitors are built, there are several types of investigations:

- Changes in how Twitter is used over time.
- Categorization of general Twitter chatter, including narrowing down to investigate particular tweets.
- Exploration of anomalous events, represented by sudden spikes in the volume of relevant Twitter data.

Throughout this section, it is important to keep in mind the initial question: What barriers to accessing loans do small businesses in Kenya face? However, the point of this exercise is not exactly to answer the research question, but rather to describe the type of conversation related to finance that are taking place on Twitter—to investigate whether, and where, information relevant to this question exists in social media.

To this end, the starting point for analysis is broad, and seeks to scan social media in different ways, looking for data clues. This question is approached from 3 vantage points, using three distinct monitors to filter Tweets related to each of: (1) General Loans; (2) Loans by Sector; and (3) Banks. In each area, the general goal is to understand what types of loan conversations Kenyans are having on Twitter. In order to do that, we used the survey to better understand (1) the words people use to discuss loans; and (2) the types of information people share in the public sphere.
When survey participants were asked who they discuss business loans with, most people said they discuss loans with loan officers and families. While some of those surveyed cited friends as a source of information about their business, several people explicitly stated that they do not discuss loans with friends due to issues of trust/privacy. If it is indeed the case that people feel that discussing loans is a private endeavor, this will impact what we can expect to find in Twitter.

Having said that, Twitter is a dynamic space, and usership in Kenya is growing. What we have discovered in this initial foray is likely to change over time as digital cultural norms evolve.

**Vantage One: General Trends in Loan Chatter**
The first monitor looks at general loan chatter. The goal was to get information on all loans except for university loans.

**Building the taxonomy**
A taxonomy is the list of keywords that is used to extract relevant posts from the full Twitter dataset. The taxonomy has two goals: (1) to include relevant posts and (2) to exclude irrelevant posts. When building the taxonomy, it is important not only to have a list of keywords that the post should include, but also to exclude particular words that result in irrelevant posts.

**Step one: survey bank clients to identify relevant keywords**
The first step for building this monitor was to survey 10 bank clients to get clues about the language that they use to discuss banks. This was done by asking the clients to recall specific conversations with bank officers, friends and colleagues. The field officer prompted the clients to recreate these conversations as she took notes on the keywords that were used. These words formed the backbone of the initial taxonomy.

**Step two: test and refine taxonomy**
The initial requirements were that the post had to include one of the following: loans, loan, mkopo, wakopo, financing, finance, or credit AND one of any of the keywords that DCA clients mentioned in the survey.

This did not have the desired results—many of the mined posts were not related to loans. Upon further investigation, the taxonomy was revised in two ways.

First, words related to finance and credit create a lot of “noise” (or, irrelevant results) in the data, without returning relevant posts. Most financing chatter refers to government or corporate finance:

- **Barclays Bank - Kenya, signs $13m financing deal with Marriot Drilling...**
  - [http://t.co/rk1sLbyJqu](http://t.co/rk1sLbyJqu)

Credit in particular creates a lot of noise. Most credit-related tweets refer to phone credit:

- **Yes @Nettoh_ken Yes. You can take home the credit, give to a friend, donate, purchase bundles. Whatever dya want 2do @NakuruRFC @wanyore**

An alternative definition of credit was also frequently used:

- **Politics is about seeing an opportunity, use the masses to achieve and taking the credit for it.**

---

8 Please note this only includes loans from Kenya’s university student loan authority, HELB. Loans from private banks for education are not explicitly filtered out.
Where the word credit does appear in a relevant manner, it typically also includes loan. The words related to financing and credit were thus removed from the taxonomy.

The second challenge was that the keywords generated from the surveys did not return many results. To refine the taxonomy, we took out the second set of keywords to see all of the posts related to loans without any additional specifications. We extracted additional keywords from the resulting posts that were relevant. This resulted in several additional words, but the words that made a significant difference were fairly simple—first was “bank” and second was “mshwari,” the mobile loan service. We also excluded common keywords that resulted in irrelevant posts, in particular those related to sports and student loans.

Step Three: Pre-viewing results and further refinement
At this point, the data was fairly clean, and there was enough data to train the monitor (described below). Once the categories were set and the monitor was run, another challenge arose. Because the daily volume of relevant posts is low, one joke or retweet can create a lot of noise in the data. For example, one Twitter user posted the same post 40 times over the course of a day, creating a dramatic spike in the Twitter data for that date. This user was thus excluded from the data. The following tweets in particular were shared widely, creating an outsized imprint in the data.

I want the bank to do two things for me.....give me a loan and leave me alone

#STOLEN Hooker walks into an Equity branch... "nataka loan ya kupanua biashara."

It is important to note that if the overall number of tweets were higher, even widespread retweeting would not throw the results. On the flip side, because of the low overall volume, it’s possible to easily find these disruptive tweets and exclude them.

Final taxonomy
For the purposes of this report, the final taxonomy is below. If USAID and Global Pulse continue to use this monitor, it is likely that further rounds of refinement will be required.

(lan OR loans OR mkopo OR wakopo) AND ("Top up" OR "Payback period" OR installments OR expansion OR mpesa OR mbesa OR financing OR "business financing" OR biashara OR dairy OR msoto OR red OR doh OR qualify OR stocking OR application OR maximum OR duration OR interests OR delay OR security OR "land title" OR deed OR "deposit dates" OR tembelea OR "fixed deposit receipts" OR secured OR "calculated interest" OR interest OR guarantees OR guarantor OR lawyer OR Agricultural OR agriculture OR development OR application OR procedures OR payback OR improvement OR n’gome OR wakora OR repay OR balance OR "agreement letter" OR period OR clear OR siri OR security OR sambaza OR defaulted OR "cooperative society" OR Faulu OR credit OR Agrovet OR mfugo OR zidisha OR "penalty charges" OR penalty OR Emergency OR "ketes temiship" OR inflation OR expectations OR capital OR terms OR payment OR "nilitemelela banki" OR farm OR status OR assets OR asset OR mshwari OR land OR animal OR animals OR "long term" OR "short term" OR "mini statement" OR "mini statements" OR ministatements OR "shamba shape ups" OR "fixed accounts" OR mshwari OR zidisha OR bank OR banki) AND -helb AND -@MweuDeh AND -hooker AND - @helbpage AND -Hooker AND -@HELBPAGE AND -“car-jacker” AND -Chelsea AND -Manchester
Defining the categories

Step one: detailed sorting
In the first round of establishing the categories, we started with 6 categories and two sub-categories:

I want a loan
  - Business
  - Personal
I have a loan, negative
  - Business
  - Personal
I have a loan, positive
  - Business
  - Personal
I have a loan, neutral
  - Business
  - Personal
Information seeking
Information provision

Step two: train the monitor and refine
After setting the categories, the next step is to train the monitors. This means sorting posts by hand into the relevant categories. This “trains” the machine to automatically categorize the rest of the posts.

When we tried to sort the posts into these categories, it became clear that the categories were too specific, and that most of them did not have enough data to use for training the monitor. We first removed the distinction between business and personal, because, using this taxonomy, most of the loan related tweets refer generally to loans, and do not distinguish between business and personal loans.

Next, there was not enough data to clearly distinguish between “I want a loan” and “I have a loan.” For example, the loan below does not relate to either, and rather is a general comment on M- shwari:

Ati Mshwari can only allow you loan equivalent to what you have in your account, why do you need the loan in the first place, is u withdraw

Final categorization
The final categorization is therefore rather general:
  General Loan, positive
  General Loan, negative
  Supplying information about loan
  Seeking information about loans
  Jokes and news items were excluded.

While this monitor does not allow for specific categorizations, for the general loan monitor it was important to cast as wide a net as possible on chatter related to loans. This monitor is useful for tracking the nature of chatter related to loans on Twitter, and to observe how it changes over time.
**Results**

Figure 2 below represents the volume of tweets from January 1, 2012 and August 25, 2013. As is clear, there is a steady rise in overall chatter over this time. Those posts are also becoming more relevant, meaning they are less related to things like international news (ie. World Bank loans to countries). On January 10, 2012 there were 2 relevant posts of 10 total posts. On January 10, 2013, there were 19 relevant posts of 28 overall posts.

Despite this growth, the volume is still very low. In July of 2013, there was an average of 10.7 posts per day.

![Volume graph](image)

**Figure 2: Volume of tweets, generally about loans, from January 1, 2012 and August 25, 2013**

**Categorizing general chatter**

It is clear that much of the growth in chatter about loans is related to M-shwari. Figure 3 below represents the tweets driving conversation in the month of July 2012 (before M-shwari was launched) and July 2013 (after M-shwari was launched). The inner circle represents the top words that co-occur with “loan” over the specified period. The outer circle displays the top words co-occurring with those words, in effect automatically creating a set of sub-categories. For example, “bank” was the top co-occurring word in July of 2012, with popular discussions breaking down further into 5 sub-categories, including “school fees” and “loan to start.” By contrast, the top topic in July 2013 is “Mshwari loan,” with five sub-categories. The overall number of relevant tweets in July 2012 was 82, while in 2013 it was 336.

![Keyword clouds](image)

**Figure 3: Key trends driving Twitter chatter from July 1- July 31, 2012 (left) and July 1-July 31, 2013 (right).**
More specifically, growth is even more pronounced after Safaricom launched its customer care service “Okao Fresha,” which includes direct engagement over Twitter. The service, launched in early February 2013, created several days of increased twitter activity in relation to loans.

![Graph showing increased Twitter activity](image)

**Figure 4:** A rise in chatter about loans upon the February 2013 launch of Safaricom’s customer care service, which includes direct engagement with customers over Twitter.

What is more, the general chatter about loans changed based on the new services. As can be seen in Figure 4 above, prior to the service being launched, most conversation on twitter was general commentating on loans, while after the service was launch there is a rise in information seeking and information provision. Much of this growth is not due to Safaricom’s informational posts, but rather to *retweets* of Safaricom’s informational posts.

![Graph showing categorization of general loan tweets](image)

**Figure 5:** Categorization of general loan tweets between January 1, 2013 and March 14, 2013.

The type of chatter, and by extension the type of analysis thus changes with the launch of M-shwari, and even more so with the launch of Okoa Fresha.
That said, of course the data is not exclusively related to M-shwari. The clustering below shows the breakdown of general chatter between March 28, 2013 and May 13, 2014. This is after the launch of Okoa Fresha, and represents a period with no dramatic spikes.

Figure 6: Main words driving Twitter loan chatter in Kenya between March 28 and May 13, 2013.

This shows that the nature of general chatter changes over time—in this case due to the relatively large Twitter presence of Safaricom.

*Investigating anomalous events*

In addition to the prolonged increase in Twitter volume explored above, there are two dramatic spikes in this period.

Figure 7: Anomalous behavior in Twitter. The date circled in blue is November 27, 2012. The date circled in red is December 22, 2012.

The first spike, on November 27, is the launch of M-shwari. The figure below represents the nature of the conversation on November 27-28. As can be seen, almost all of the conversation is about Safaricom and M-shwari, with a subset discussing M-shwari interest rates.
The second spike, on December 22, 2012, came with the announcement by vice-Presidential Candidate William Ruto that if elected, the government would offer an interest free loan to women and youth. As is clear from the break down below, many of the tweets were neutral in nature, related to sharing the news. However, it is also clear the announcement was met with some skepticism—in addition to “gullible” and “silly season,” one popular retweet was “Ruto anafikiria sisi ni wajinga” or, in English, “Ruto thinks that we are fools.”

Investigating particular tweets
With the low overall volume of tweets, it is possible to go through the sorted tweets of relevance. Some tweets do shed light on barriers to accessing loans:

*I need a bizness loan..interest is double*
Understanding the positives and negatives surrounding bank loans help people make the wisest decisions when they need loans.

The youth and women have no skill nor security to secure the loans that they need to start enterprises #OrangeNDC

However, there are not enough of these tweets to discern clear trends, and without demographic information if they represent the voices of underserved entrepreneurs.

Vantage Two: Loans by Sector
This portion of the analysis was focused on setting up monitors related to various sectors. In the first attempt, we concentrated on the sectors listed in DCA’s Market Assessment for Kenya\(^9\). While this led to a large number of overall tweets, very few of them were relevant. We therefore focused on *tweets specifically about finance and loans*. For this monitor, we left “financing” in the taxonomy to capture the chatter about financing in particular sectors, although they might not be related to individual loans. The rest of the taxonomy was combination of (1) keywords extracted from the surveys; (2) sectors identified as priority areas of DCA; and (3) sector-specific words surfaced from the posts themselves. Therefore, this taxonomy *only* surfaces posts which are related to a specific type of loan.

(loan OR financing OR mshwari OR mkopo) AND (water OR maji OR energy OR agriculture OR n’gombe OR business OR biashara OR bizness OR personal OR emergency OR cow OR dairy OR health OR crops OR resistant OR maize OR infrastructure OR equity OR home OR "short term" OR afya OR land OR animal OR animals OR farm OR shamba OR machine OR property OR machine OR auto OR vehicle OR car OR gari OR bank OR SME OR Enterprise OR stima) AND -*KIKUYU MIND* AND -*take that to the bank and get a loan*

Our first attempt at categorization was to parse between and within sectors. For the business category, we first tried to sort according to the DCA’s Market Assessment for Kenya priority areas:

Agriculture (Dairy, Drought-resistant crops, Horticulture, Maize value chain)
Clean Energy
Health
Water
Infrastructure

However, it was clear that this would not work because of the overlap between personal and business loan keywords. For example, chatter on Twitter related to energy (or “stima” in Kiswahili) are both related to the energy sector and personal loans, but none that came up when training the monitors were related to specific energy enterprises:

#WorldBank backs #Chinese firm’s loan for energy plants in #Kenya http://t.co/L8MkJd458h

@KenyaPower what’s the procedure of acquiring stima loan. What are the requirements?

The second attempt took a broader scope:

Personal
Emergency
Student

Home/ electricity
General
Business
Agriculture
Industry
General
Government
International Loans to government
Local loans to government
Government financing to sectors

However, the low volume of data made it impossible to define the categories with this level of specificity. The next attempt was to create sub-categories in each broad grouping of “Negative” and “Positive” and “Neutral.” However, most of the tweets were neither negative nor positive. In the end, the broader categories were used on their own:

Personal
Business
Government

Results
From January 1, 2012 to August 25, 2013, there have been 5,317 relevant posts, representing an average of 9.2 posts per day. There has been a growth in Twitter similar to the one seen in the first monitor, with a sustained growth after the launch of M-shwari. This growth is driven by chatter in business and personal loans, as opposed to government loans. The anomalous events in this graph are very similar to those described in the general loan category. The chatter, however, differs.

![Chart showing growth in relevant posts](image)

**Figure 10: Volume of Business, Personal and Government related loan data from January 1, 2012 to August 26, 2013.**

Categorizing General Chatter
It is possible to see trends between sectors. Figure 11 shows the nature of the conversation between March 26, 2013 and August 26, 2013.
Although the low level of data made it impossible to create specific sub-categories with each sector, it is possible to look at the actual Twitter data to see how relevant it is to assessing barriers to accessing loans. Here is a sampling of business related tweets.
Breaking the loan data down this way does allow for a deeper understanding of Twitter data that might relate to barriers faced by small businesses seeking to access loans, but the low volume of data prohibits monitoring those trends over time.

**Vantage Three: Banks**
The final vantage point was to look at how people discuss the banks themselves.

The first step was to look at how much data there is on each bank in Twitter. We first did an overall search for bank names listed as partner banks in the DCA Market Assessment for Kenya. The two with the highest amount of search data were Equity and Kenya Commercial Bank.

For these monitors, we relied heavily on sentiment analysis. Sentiment analysis relies on an algorithm that ascertains language clues to measure the overall “feeling” of the tweet. In addition to the predefined algorithm, the algorithm also employs a machine learning technique. The user sorts a subset of tweets by hand, in the process defined above.

**Equity Bank**
Equity Bank consistently has a small amount of Twitter data. The graphic below represents the volume of Equity Bank tweets from Jan 1, 2012 to August 25, 2013. Over this period, there is a clear growth in chatter about Equity Bank, both in terms of general daily chatter and in the frequency of spikes in conversation.

![Volume of posts related to Equity Bank from January 1, 2012 to August 25, 2013.](image)

Figure 12: Volume of posts related to Equity Bank from January 1, 2012 to August 25, 2013.

The volume of tweets ranges between zero and a peak of 385.

Sentiment analysis was run on this monitor, but because the level of data for general chatter is so low—most of the data is from news and events—we were not confident in the results. A further refinement of this monitor might include decisions regarding exclusion of certain news items to better focus on peoples’ discussions of the operations of the bank. For the purposes of this report, all news items were included to provide a broad overview of the type of Equity Bank related chatter.

---

Investigating anomalous events
There are two drivers of peaks during this period—one is news related to Equity, accounting for most of the peaks. This news might be related to a new service Equity is launching, or business news about Equity. The latter includes the largest peak on October 4, 2012, circled in red above. This is a news peak, Equity appointed Julius Kinng'etich as its new COO.

Major events can also be seen in this data. The second major peak, which occurs in a cluster between July 8, 2012 and July 13, 2012 (circled in blue above) is related to a widespread systems failure in Equity’s ATM services. This event can be seen very clearly in Twitter. Figure 13 illustrates this conversation.

![Figure 13: Break-down of tweets related to Equity Bank during the Bank’s three-day ATM failure, date range July 8, 2013 – July 13, 2013.]

General chatter
For Equity Bank, the level of general chatter—ie. posts that are not driven by specific events or the news—is too low to analyze with any confidence in trending behavior.

Kenya Commercial Bank (KCB)
The overall volume and trends in KCB twitter data is represented in Figure 14 below. KCB data in Twitter somewhat more interesting than Equity data for two primary reasons.

First, KCB has more tweets overall than Equity Bank. Over the same time period, KCB had 13,297 relevant posts, compared to Equity's 7,944. Second, KCB appears to have a more robust Twitter strategy than Equity. They post information about bank products, and they also engage somewhat with customers via Twitter, although as can be seen below the instance of customers reaching out with questions is low. This volume made it possible to do more detailed analysis of the content of the posts. Figure 14 shows the proportion of posts that are negative, positive, neutral/news related, or information related.

---

11 Equity Bank often advises Twitter users to contact a local bank when they ask questions; KCB does appear to answer them. Other companies are even more directly engaged with customers via Twitter—while this report does not include a monitor of Safaricom, they are very engaged with their customer base on Twitter.
One challenge working with the KCB data is that KCB sponsors sports team and sporting events. This creates a lot of noise in the data, with many KCB references being recaps of games. While most of these posts can be removed from the data by excluding posts with words like “football,” this still accounts for the large number of irrelevant posts—less than half of the posts linked to KCB are relevant. It also accounts for most of the dramatic spikes seen in Figure 14. On game days, up to 95% of the tweets are irrelevant to function of the bank.

Figure 14 represents a period of time when the majority of posts (those shaded in blue) were related to the bank. This period shows both a spike representing 215 relevant posts on the peak day of February 28 (circled in blue). The rest of the period is more representative of “normal” posting behavior, and appears to be general chatter.

Figure 14: Volume and sentiment of all tweets related to Kenya Commercial Bank from January 1- August 25, 2013.

Figure 14: Detail of February 26, 2013 to March 23, 2013.
**Investigating anomalous events**

The spike on February 28, 2013, circled in blue in Figure 14 above, is the day KCB released its investor report. Not only is this event clearly visible in Twitter, it’s also possible to break down the parts of the report the Twitterverse found the most relevant. This particular event may not be relevant to the initial research question; however, it is worth exploring what we can see in Twitter as an example of how one can analyze anomalous events.

Figure 15 groups the primary categories of posts on that day. Here, it is possible to see the elements of the report that people found interesting, including chatter about the CEO and about the 14% growth of the bank.

![Figure 15: Twitter reacts to KCB press release on February 28, 2013](image)

**Categorizing General Chatter**

KCB was mentioned much less in Twitter on the dates circled in green in Figure 14—from March 6, 2013 to March 20, 2013. However, because the Twitter trends on those days were more “normal”, they could potentially have baseline information on more typical trends in Twitter.

These dates include a total of 472 relevant posts, with a peak day of 83 relevant posts on March 9 and a low day of 12 relevant posts on March 10. One cluster—with the word “Masomoloan” is about education loans.
Another, much smaller, trend that can be seen in this word cloud is “smeloans,” short hand for small to medium size enterprises. Because small to medium sized enterprises are relevant to the research question, we investigated that trend further by looking at the actual data. Upon further investigation, it’s clear that this trend is related to a KCB/Safaricom announcement on March 6 and is largely driven by re-tweets.

Figure 16: Word cloud showing popular words in loan related tweets from March 6, 2013 to March 20, 2013.
This time period was chosen to show “general chatter”—however, the data is still driven by news and events. In essence, it is less a demonstration of general chatter, and more a series of smaller, news-driven peaks.

Sentiment analysis and the banks: Can we compare?
In looking at these bank-related monitors, a natural question arises of how we can compare the data between banks. This would require developing the methodology further, including understanding how the banks’ social media strategies impact related posts.

KEY OPPORTUNITIES
• Twitter use is growing rapidly, and the general monitor shows the emergence of a Kenya specific Twitter culture.
• In particular, Twitter is being used to seek, access and share information about loans, especially mobile loans.
• Twitter is also being used to share and comment on the news related to personal and business loans. Monitoring the news, as well as how the news is being perceived, might be of interest.
• Much of the data is related to M-shwari and other Safaricom related topics. A future iteration of the monitors could focus solely on non-traditional banking or exclude M-shwari data for deeper insight on traditional loans.
• Monitoring Tweets with a taxonomy of keywords related to sectors does a better job of discerning trends that might relate specifically to entrepreneurs.
• Redeveloping the sector-specific taxonomy such that it excludes personal and government loans might be worth exploring.
• While there is not enough data to train the monitors into further subcategories, the low number of posts makes it possible to filter the tweets and actually glance through them, making more qualitative methods viable. This could serve to bring in new ideas, similar to the role of some focus groups.
• For banks that have a robust social media strategy, monitoring the specific Twitter handle of the bank could provide insight into (1) products and services available at the bank, and, to a lesser extent, (2) information seeking behavior.

KEY CHALLENGES
• Across monitors, the overall volume of chatter is low, so small changes in Twitter behavior, for example due to a popular retweet or the behavior of one Twitter user, can create spikes in overall volume of conversations.
• There is very little general chatter about loans (i.e., chatter is largely driven by news/events or information seeking).
• There is little demographic data available, so even where there may be tweets that are directly related to barriers to accessing loans, it is difficult to know whether they represent underserved entrepreneurs.
• As seen by the launch of M-shwari, Twitter is a dynamic space and the analysis that is possible using Twitter will change as the way that people use it changes. It is important to keep an eye on these changes.
• There is a lot of noise in the data. For KCB, this noise includes sports chatter. For both banks, this includes news item related to the overall business of the bank, not necessarily directly related to bank services.
GOOGLE TRENDS

Information seeking behavior can shed light on people’s activities in a profound way. From the surveys we did, clients cited a wide variety of information needs, from new farming inputs to different loan products to definitions of keywords like “guarantee.” Google represents an enormous repository of data about information seeking behavior. While Google does not provide direct public access to the search data, they do provide some free online tools to analyze the data. This data is interesting in part because, although we have no individual demographic information about the user, it represents data that is not shared publically, ie. it might illustrate personal needs and interests that people do not wish to share publically.

In Kenya, the only available tool is Google Trends “Hot Searches.”12 With Google Trends, one can track search terms over time. As an analytical tool, Hot Searches is somewhat limited13. First, Google Trends only shows relative volumes of searches, it does not show real numbers. Second, while it is possible to create a graph that aggregates synonyms in one trend line, it is not possible to create subcategories within one search term—this is explored in greater detail in the examples below.

Another limitation is that it is difficult to use Google Trends to “learn”—typically in big data analytics taxonomies will be developed both by working with domain experts (in this case DCA clients) and by identifying keywords in the data itself. Scope for this type of iteration is limited in Google Trends—only one word that commonly occurs with the initial keyword is given. This both limits discovery and the ability to contextualize the trends.

This section explores what is possible to look at using Google Trends, starting with the keywords generated from the survey. It then moves on to look at loans generally, banks, and sectors.

In all of the experiments below, the time period is August 21, 2012 to August 21, 2013.

Taxonomy from surveys
In the surveys, people noted that they looked up definitions of banking terms. However, trends in searching for this data is very erratic, suggesting that the overall volume is low, allowing small changes in volume to dramatic spikes.

One of the survey respondents reported searching for the word “guarantee.” Because there are several variations of that word that could be used, the red trend line is an aggregation of “guarantee” and “guarantor.” Even with both terms, the search volume is too low to even register on the graph.

12 Hot Searches can be found at the website: http://www.google.com/trends/explore#cmpt=q
13 Google Correlate, while not currently available in Kenya, is somewhat more useful than Hot Searches. Google Correlate uses time series data sets—for example the unemployment rate—to detect search volumes that follow the same trends.
In addition, when compared with searches for loans, it appears as if the volume of searching for loans is much higher than for definitions.

Loans

There is no straightforward way to create sub-categories within overall loan searches. In Google Trends there are two ways to approximate this. First is to specify a full search phrase in quotes, for example “business loan” or “personal loan.” When we did this, no data was returned. Second is to exclude words from the search, for example “Loan -student.”

Google Trends shows the top co-occurring word with “loan,” which is HELB, the student loan authority. The following graph shows loans with and without the words “HELB” and “student.” This is essentially the extent to which it is possible to specify what types of loans people search for with the information available. It is important also to underline a subtle distinction—the red trend line below merely excludes those searches which include the words “HELB” and “student”—it is quite possible that people searching for simply the word “loan” are looking for student loans.
Figure 19: Google search trends, comparing “loan” and “loan,” excluding “HELB” and “student” in Kenya, August 21, 2012 to August 21, 2013.

From the above graph, it’s clear that general loan searches are fairly stable, with one spike in early August, 2013. Given that this spike occurs roughly one month before the start of school, it is possible that this is indeed related to student loans.

**Banks**
I also compared searches related to DCA partner banks. Here, despite the limited amount of information we have about this data, there are potential insights we can extract. One is anomalous behavior. In both of the graphs below, there is not much to suggest a spectacular event, with the exception of the drop on March 3 across all issue areas. This was Kenya’s presidential election.

The other is overall interest between search terms. It’s clear, for example, that more people search for Equity Bank and KCB than Faulu and Bank of Africa, perhaps related to the size of the banks.

Figure 20: Google search trends, comparing four banks in Kenya, August 21, 2012 to August 21, 2013.

**Sectors**
Finally, we compared the sectors listed in DCA’s Market Assessment for Kenya. As mentioned above, it is not possible to look at how searches for loans intersect with sectors other than student loans. The drop across sectors is again on March 4, the day of Kenya’s national election.
It’s clear that between sectors, people search for health more than any other sector. However, this could be because there is more information about health online that is directly relevant to people’s lives—compared, for example, with “infrastructure” which might be considered important but less relevant in terms of immediate information needs. In this case, defining infrastructure more precisely—for example “roads” or “municipal water” could potentially change this picture.

Figure 21: Google search trends, comparing sectors in Kenya, including “Health + Afya,” “agriculture + drought resistant + crop + horticulture,” “water,” “energy + stima + electricity,” and “infrastructure” in Kenya, August 21, 2012 to August 21, 2013.

**KEY OPPORTUNITIES**

- The initial tests with Google can provide baseline information to monitor for anomalies in search trends in the future.
- Events that are broadly shared—ie. the beginning of a school semester or the Kenyan elections—are clearly visible in the data. This could mean that other systemic events will be visible as well.

**KEY CHALLENGES**

- The analytical tools made available by Google for Kenya data are limited.
- Currently it is only possible to know the relative volume of tweets, it is not possible to analyze real numbers.
- Student loans seem to be driving the loan searches, even when words like “HELB” and “student” are excluded.
FINDING THE DIGITAL FOOTPRINT

Sections II and III starts from the available data to look for information that might be useful. Section IV takes the opposite approach: looking at where business owners (1) discuss and (2) seek business and loan related information. The section begins by discussing the results of the survey, which asked people where they get information about loans and their business. It then provides two case studies on where other digital data might lie.

Survey results
In the survey of 10 DCA clients, we asked where people got information about their business, existing loans, and new loans. Respondents all reported getting information from multiple information streams, listed in Table 2 below. People reported that, while interacting in public spaces is a primary way to access information about business, they are much less likely to discuss existing or new loans in public spaces. Across the board, those who mentioned that they use the Internet to access information mentioned that their children as helped them. While mobile phones were not cited as primary source of information about business or about new loans, most respondents used mobile phones to manage their existing loans. This includes paying, receiving alerts and other communications from loan officers via SMS, receiving mini-statements and deposit information.

<table>
<thead>
<tr>
<th>In person</th>
<th>Business Information</th>
<th>Loan information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Colleagues in markets, buyers who come to the farm, friends and family, trade shows,</td>
<td>Family, visit banks</td>
</tr>
<tr>
<td></td>
<td>Agrovet stores, Cooperative society</td>
<td></td>
</tr>
<tr>
<td>Media</td>
<td>TV programs (&quot;Shamba Shake-ups&quot;), Radio, Newspapers, Magazines</td>
<td>Newspapers, pamphlets,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>books (especially the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dictionary to look up</td>
</tr>
<tr>
<td></td>
<td></td>
<td>technical terms), Notice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>boards</td>
</tr>
<tr>
<td>Digital</td>
<td>E-banking, Wikipedia, Mobile phones (calling colleagues, communication with staff)</td>
<td>Bank website, Google,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M-shwari, SMS (ministatements), Call</td>
</tr>
<tr>
<td></td>
<td></td>
<td>banks</td>
</tr>
</tbody>
</table>

Table 2: Summary of survey results regarding where people self-reported getting information about their business, existing loans, and new loans

Case studies in digital data
Case Study 1: Agriculture
Kenya’s dynamic tech sector is developing new services across a number of sectors. These services are creating a great deal of “data exhaust,” but the data is often either proprietary or it is not collected in a way that makes it directly useable. This section looks at programs that may create data exhaust, using the case study of agriculture. It’s likely that not all of these programs will succeed; however, for those that grow, the data will likely be a valuable asset.

The key will be to have access to that data in order to conduct any analysis. Depending on the nature of the service, accessing the data will require different efforts. One approach might be to establish partnership agreements to access to the data. In some cases, the service itself is an aggregation of data, and a simple subscription will allow access to the data. The final suggested
way forward is a big data methodology called “scraping.” For data that is made available on-line, a tool can be built which will automatically “scrape” the data from a website. Scrapers are typically designed such that the data is pulled down from the web and is made available in a pre-designated format. Data from different websites can thus be integrated in such a way that it can be analyzed together. The table below lists programs related to agriculture that generate data exhaust, as well as what would be required to access that data.

<table>
<thead>
<tr>
<th>Program</th>
<th>Data Description</th>
<th>What's required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agrimanagr</td>
<td>Real time data all along the value chain, from crop health to distribution to staff management. Designed for medium to large agribusiness.</td>
<td>Partnership</td>
</tr>
<tr>
<td>Dairy Sacco App</td>
<td>Real time data on dairy credit within cooperative credit structures.</td>
<td>Partnership</td>
</tr>
<tr>
<td>DrumNet</td>
<td>General agricultural market data on producers, banks, agro-dealers and buyers. This integrates other data, like weather data and subscribers can access the aggregated forms of data.</td>
<td>Partnership or subscription</td>
</tr>
<tr>
<td>iCow</td>
<td>Sector information on cows, around things like cow health, calf management, disease management, milking methods and feeding.</td>
<td>For all data: partnership with Safaricom; alternatively, the basic messaging could be aggregated by signing up for the service.</td>
</tr>
<tr>
<td>Kilimo Salama</td>
<td>Weather data, basic micro-insurance data.</td>
<td>Partnership</td>
</tr>
<tr>
<td>Mfarm</td>
<td>Market information on agriculture, price information and supply and demand.</td>
<td>Partnership or scraper</td>
</tr>
<tr>
<td>Mkilimo</td>
<td>Data on what information farmers are seeking (Farmer help line)</td>
<td>Partnership</td>
</tr>
<tr>
<td>National Farmers Information Service</td>
<td>Aggregates information streams from government sources, markets, and weather data, including national livestock marketing information system.</td>
<td>Scraper, partnership</td>
</tr>
<tr>
<td>National Livestock Marketing Information System</td>
<td>Data on livestock producers and traders.</td>
<td>Partnership</td>
</tr>
<tr>
<td>Next2</td>
<td>Social data-- Next2 connects people and organizations with similar concerns or interests over mobile phone, a text-based social network.</td>
<td>Partnership</td>
</tr>
<tr>
<td>Regional Agri Trade Network</td>
<td>Agriculture market information for East Africa, includes trade data.</td>
<td>Scraper, partnership</td>
</tr>
<tr>
<td>SokoniSMS</td>
<td>Price information</td>
<td>Subscription</td>
</tr>
<tr>
<td>Sokopepe</td>
<td>Market information on agriculture, facilitates mobile transactions</td>
<td>Partnership</td>
</tr>
<tr>
<td>Soko Shambani</td>
<td>Potato supply and demand data, using SMS short term, integrated with Twitter.</td>
<td>Partnership</td>
</tr>
<tr>
<td>Ukulima.net</td>
<td>Aggregation site to integrate data from the web, mobile and specialized apps. (In development)</td>
<td>Partnership</td>
</tr>
</tbody>
</table>

Table 3: List of digital services related to agriculture in Kenya.
**Case Study 2: Banks**

Of course, the treasure trove of relevant big data would be to access Safaricom, M-PESA and M-shwari data. However, aside from the mobile banking data, banks put a great deal of information online which could be useful. Depending on what data could be relevant, and how often the data is changed, building scrapers to collect this data from websites could be another approach.

Some banks, for example Barclays Bank below, post interest rates, but this is not consistent across banks. However, most banks provide information about the types of loan products they provide. A scraper could help track changes in loan products and other bank information across banks.

Two examples of online data that could be scraped are below, from the websites of Barclays Bank, Kenya and Equity Bank.

**Interest Rates**

Barclays offers competitive interest rates pegged to each customer’s risk profile. Interest calculations are based on reducing balance basis as opposed to the flat rate basis which is the current common market practice.

<table>
<thead>
<tr>
<th>Year</th>
<th>Reducing Balance Interest Charged</th>
<th>Flat Rate Interest Charged</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>259,391</td>
<td>300,000</td>
<td>40,609</td>
</tr>
<tr>
<td>Y2</td>
<td>164,916</td>
<td>300,000</td>
<td>135,084</td>
</tr>
<tr>
<td>Y3</td>
<td>60,856</td>
<td>300,000</td>
<td>239,414</td>
</tr>
<tr>
<td>Total</td>
<td>484,856</td>
<td>900,000</td>
<td>415,144</td>
</tr>
</tbody>
</table>

**Figure 23:** Barclays Bank interest rates from the Barclays Kenya website ([http://www.barclays.com/africa/kenya/commercial/assetfinance.php](http://www.barclays.com/africa/kenya/commercial/assetfinance.php)), accessed on August 20, 2013.

**Figure 24:** Screen shot of Equity Bank Agricultural Loans from the Equity Bank Kenya website ([http://www.equitybank.co.ke/index.php/loans/agriculture-loans](http://www.equitybank.co.ke/index.php/loans/agriculture-loans)), accessed on August 20, 2013.
RECOMMENDATIONS AND NEXT STEPS

How do big data projects typically work?
Big Data projects work best when they involve several iterations and collaboration between topical domain experts (who understand both the context, and the programmatic information gaps or needs), and data scientists/analysts. This feasibility study was intended to lay out the landscape of finance-related big data in Kenya. The next step is for DCA to map out their own data needs in relation to these possibilities. At that stage, creative brainstorming with the Global Pulse team might be productive in connecting the opportunities with the needs.

To assist with discussions, some potential next steps are below. These are intended to inform a larger conversation on whether further research could support the work of USAID.

Laying the Foundation: Better Baseline data
The survey about DCA clients’ use of digital services, limited though it was, yielded insights that proved valuable to this report. To better understand how to work with new sources of digital data in Kenya, expanding the survey to more DCA clients or other underserved entrepreneurs could (1) provide more contextual information to inform the approach, including how underserved entrepreneurs use digital services; and (2) provide further ethnographic data on how and where people discuss finance or related issues. It might be the case that DCA would choose to continue working with the banks to target the surveys. It might also be valuable to explore other targeting options, for example by working with trade associations or cooperatives. Looking at where clients reported accessing information about their businesses (see Table 3 above) might be a good place to start thinking through targeting mechanisms. Another option would be to experiment with new ways of targeting, for example using Twitter itself, or perhaps Facebook ads or sending out opt-in mobile phone surveys.

Looking for low-hanging fruit in social data
While it’s clear that social media is on the rise in Kenya, it is also clear that for the purposes of informing research on financial inclusion, it is probably not saturated enough yet. If DCA is interested in using Twitter data to inform its work now, it might be a good idea to use this methodology in a country that DCA works which has a stronger social media culture. In some emerging markets, Twitter has become fairly mainstream. For example, it has been widely reported Jakarta is the top-tweeting city in the world—and the 6th most prolific city is another Indonesian city, Bandung. Indonesia, India and Mexico are the 5th, 6th and 7th most prolific countries, respectively14.

Looking ahead: staying in front of digital data in Kenya
That said, there are areas that might be worth exploring more to anticipate the rising relevance of this data in Kenya. This is the case both for social media and other digital data.

Social Media
For social media data, looking ahead could involve keeping an eye on the evolution of Twitter culture to understand how it could be informative for future research. By looking at the “general loan” monitor, it’s possible to see the emergence of a Kenya-specific Twitter culture. At this stage, the Twitter culture includes:

Information seeking behavior: Twitter is being used for customer care. For banks with robust social media strategies, in particular Safaricom’s mobile banking, customers are asking questions and receiving answers. This might be a good opportunity for further investigation to understanding people’s concerns.

M-shwari and non-traditional banking: Close to half of the chatter related to loans is related to M-shwari. In addition, 8 out of the 10 peri-urban or rural farmers that we surveyed mentioned that they use mobile banking. Monitoring trends in Twitter might be a good way to understand how this type of lending intersects with business finance.

News, events and institutions: Across all categories, news and events drive the conversation. Keeping an eye on these trends could provide information about new financial products. Sentiment analysis could be particularly useful to show how Kenyans perceive news, events, and institutions like banks or even USAID.

Institutional Innovation
Laying the groundwork for future work might also include thinking about DCA or USAID’s current capacities and practices, including:

Partnerships As shown in the agriculture case study in Section IV, new mobile services are being developed in Kenya all the time. All of these services create data—which is not publicly available. However, USAID funds some of these initiatives and should be thinking about data from the beginning of project development. In addition, it might be possible to access such data through partnership agreements, although the data would have to be properly anonymized.

New Capacities  A great deal of information is readily available on-line, but collecting it by hand doesn’t make sense. Building the capacity to automate collecting that data through scrapers might be able to contribute to market analysis, in particular around things like prices, products, and/or interest rates. Deciding whether to go in this direction will require identifying websites that hold information of interest of DCA and assessing how often the information is updated. These scrapers need some supervision—for example the scraper would need to be updated if a website is redesigned—so building a scraper is not a one-time expense. Therefore the next steps here are to (1) assess the information on-line that could be of use to DCA; (2) assess the requirements of the scraper to automate the data collection; and (3) assess the maintenance requirements of the scraper.

Combining new and old: “ground truth” data
New, digital data sources need not be considered on their own. There are several important reasons to bring together digital data with more traditional data sources.

Establishing relationships between on-line and off-line behavior  Evaluating digital trends in relation to off-line trends could provide more insight into the meaning of observed digital behavior. As a hypothetical example, it might be possible to find a correlation between loan application rates and Twitter chatter around loans. Once a pattern is established, detecting similar behavior in Twitter going forward could provide clues about loan seeking behavior. (However, because the level of general chatter about loans is so low in Kenya, for now this approach might not be very useful in that particular country.)

The same could be true of trends in mobile services, if such data can be accessed. Correlating anonymized mobile data with statistical data-sets or dedicated surveys could lay the groundwork for understanding, for example, what changes in M-shwari loan patterns mean in term of people’s lived experiences.
Integrating digital data with existing data sources

Just like any data source or information stream, digital data only paints part of the picture. When thinking through a big data research agenda, it might be worth considering how existing traditional data can support the work. Conversely, when thinking through a more traditional research agenda, it’s worth considering how big data might provide additional insights.

Conclusion

From this feasibility study, it is clear that certain big data methodologies could be leveraged to enhance USAID’s programmatic capacity to get early insight or warning of emerging trends, keep an eye on the “pulse” of a given issue area, and get more real-time feedback on programs as they are being rolled out. However, it is equally clear that further work is required to turn the possibilities for big data analysis into meaningful means of operation in practice.

To do so, creativity and imagination is key—and it is worth thinking both “big” and “small.” At some stage, big data could contribute to projects that we previously have struggled to undertake, for example understanding complexity in markets; or gaining new insights into social networks and financing. Big data might also be able to support relatively simple information needs, for example capturing the pulse on sentiment around news and events.

This report seeks to provide the information required to understand the opportunities for big data analytics around financial inclusion in Kenya. Determining which line of inquiry makes the most sense moving forward is a matter of matching the needs and opportunities, as well as exploring the internal capacity required to employ new methodologies.
ANNEX 1:

ID Number ________

Bank Questionnaire

Introduction
Hello, my name is Esther Ajambo. I’m doing a survey in conjunction with the United Nations Global Pulse initiative and USAID. We are trying to understand how people get information about loans and we have some questions for you. As someone who receives a loan, you were recommended by KCB bank (?). The survey will take between 30-45 minutes. Are you available to do the survey now?

If yes, then begin survey.
If no, then say
Is there another time that you are available to do the survey? I will call you at a later time.

I’m going to start by asking a few background questions.

1. How old are you? ________________

2. Esther please fill out without asking. Are you male/ female? M/ F

3. What town and county do you live in? ________________

4. What is your highest level of formal education? (circle one)
   (1) no formal education
   (2) primary school
   (3) secondary school
   (4) university
   (5) graduate school

5. What type of business do you do? ____________________________

6. Do you employ anyone? Y/N
   a. If yes, how many people? ___________

Loan Profile

Now I’m going to ask you some questions about loans that you have applied for or received in the past two years.

7. In the past 2 years, how many loans for your business have you applied for? ________
8. In the past 2 years, how many loans for your business have you received? 

__________

9. In the past 2 years, which banks did you apply for loans from? (list)

10. In the past 2 years, which banks did you receive loans from? (list)

11. Please think about all the loans you applied for in the last two years. How did you find out about the loan?
   a. (List, make sure their number of answers matches the number they said they applied for)

We are looking for words that people use to talk about loans—we’re interested in the actual words, the languages that you use, and the conversations that you have. Please try and think of the key words that you use when you talk to people about loans. For example, you might say to a loan officer “I have a question about whether I can top-up my loan?” And the loan officer might answer giving you a new interest rate, a payment schedule, or other things. We would be interested in the keywords in this discussion, like “top-up,” “payment schedule” etc.

12. Can you remember the last time you spoke to a bank official or a loan officer about a loan—like a friend, or a family member, or a colleague? Y/N

13. What is that person’s job title/relationship to you? ________________________________

14. What type of loan did you talk about? ________________________________

15. What language or languages did you use to talk about the loan to this person?
   1. English
   2. Kiswahili
   3. Local language, please list _______________________

16. What words did you use to talk about loans to this person? Please list all the words they use. You might have to explain this, asking them a little bit about the conversation that you had with them.

_____________________________________________________________________
_____________________________________________________________________
Esther, let them answer, and then prompt them.
   a. Did you use any technical words, like the type of loan?
      *List technical words:*

   b. Did you use any slang words—for example in a dialect?
      *List slang:*

17. Can you remember the last person who **was not** a bank official or loan officer that you spoke about loans with?  Y/N

18. What is that person’s relationship to you?  ________________________________

19. What type of loan did you talk about?  ________________________________

20. What language or languages did you use to talk about the loan to this person?
   1 English
   2 Kiswahili
   3 Local language, please list ________________________________

21. Including all languages that you spoke to this person in, what words did you use to talk about loans to this person?

Esther, let them answer, and then prompt them.
   a. Did you use any technical words? For example, type of loan?
      *List technical words:*

   b. Are there any slang words that you use to talk about loans—for example in sheng or another dialect?
      *List slang:
22. Have you ever discussed loans on (1) Facebook? Y/N
   (2) Twitter? Y/N

23. Have you every searched for loan information on Google? Y/N
   i. What search terms did you use?

Digital footprint
I'm going to ask you some questions about your access to technology.

24. Do you have a mobile phone? Y/N
25. Do you have a computer? Y/N
26. Have you ever accessed internet? Y/N
   If yes, then ask . . .
   - Mobile phone Y/N
   - Cyber Café Y/N
   - At home on a computer Y/N
   - Friend’s computer Y/N
   - School computer Y/N
   - Work computer Y/N
   - I don’t use the internet Y/N
   - Other, please describe ________________________________________

27. How often do you use facebook?
   (1) More than once per day
   (2) 2-3 times per week
   (3) Once per week
   (4) 2-3 times per month
   (5) Once or less per month
   (6) I don’t use facebook

28. Do you ever talk about your business on facebook?
29. Do you ever talk about loans on facebook?
30. Do you know what Twitter is? Y/N
31. Do you have a Twitter account? Y/N
32. Do you use banking services on your mobile phone? Y/N
   a. If yes, please list all banking services you use on your mobile phone.

Access to information
33. How do you access information about your business? This includes price information, market opportunities, and general queries. Please list all the places you get information for your business.

Esther, use these as prompts—

i. Do you talk to friends or family to get information that you use in your business? Y/N
   What type of information?
   ______________________________________________________________
   ______________________________________________________________
   ______________________________________________________________

ii. Do you access information on about your business on your mobile phone?
    What type of information?
    ______________________________________________________________
    ______________________________________________________________
    ______________________________________________________________
    ______________________________________________________________
    Please list all programs that you use on a mobile phone for business related information
    ______________________________________________________________
    ______________________________________________________________

iii. Do you access information about your business in the newspapers, pamphlets, or books? Please describe the type of information and the media.
    ______________________________________________________________
    ______________________________________________________________

iv. Do you ever seek business related information on the internet? Y/N
    Please list the sites you would use on the internet.
    ______________________________________________________________
    ______________________________________________________________

v. Other, please describe.

34. How do you access information about your existing loans?
Esther, use these as prompts—

i. Do you discuss your loans with friends or family?  Y/N

ii. Do you access information about your loans on a mobile phone?  

   Please describe what information about your loans you get on a mobile phone.

iii. Do you physically go to the bank?  Y/N

   Please describe what information you get by physically going to the bank.

iv. Do you use the internet?  Please describe what information you get via the internet.

   ______________________________________________________________

v. Other, please describe

35. If you were interested in getting a new loan, and you wanted information on what types of loans were available where would you look for that information?

Esther, use these as prompts—

i. Would you talk to friends or family to get information about new loans?  Y/N

   1. What type of information?

ii. Would you access information on about new loans on your mobile phone? Y/N
What type of information?

__________________________________

__________________________________

__________________________________

Please list all programs that you would use on a mobile phone to get information about new loans

__________________________________

__________________________________

__________________________________

iii. Would you access information about new loans in the newspaper, pamphlets, or books? Please describe what information and the type of media.

__________________________________

__________________________________

iv. Would you ever seek information about new loans on the internet? Y/N

Please list the sites you would use on the internet.

__________________________________

__________________________________

__________________________________

v. Would you physically go to the bank? Y/N

What type of information would you seek at the bank?

__________________________________

__________________________________

vi. Other, please describe