Data in the First Mile

Kuang Chen
UC Berkeley
kuangc@cs.berkeley.edu

Joseph M. Hellerstein
UC Berkeley
hellerstein@cs.berkeley.edu

Tapan S. Parikh
UC Berkeley
parikh@ischool.berkeley.edu

ABSTRACT

In many disadvantaged communities worldwide, local low-resource organizations strive to improve health, education, infrastructure, and economic opportunity. These organizations struggle with becoming *data-driven*, because their communities still live outside of the reach of modern data infrastructure, which is crucial for delivering effective modern services. In this paper, we summarize some of the human, institutional and technical challenges that hinder effective data management in "first mile" communities. These include the difficulty of deploying, cultivating and retaining expertise; oral traditions of knowledge acquisition and exchange; and mismatched incentives between top-down reporting requirements and local information needs. We propose a set of directions, drawing from projects that we have implemented. They include 1) separating the capture of data from its structuring, 2) applying intelligent automation to mitigate human, institutional and infrastructural constraints, and 3) deploying services in cloud infrastructure, opening up further opportunities for human and computational value addition. We illustrate these ideas in action with several projects, including Usher, a system for automatically improving data entry quality based on prior data, and Shreddr, a hosted paper form digitization service. We conclude by suggesting next steps for engaging in data management problems in the first mile.

1. INTRODUCTION

International development organizations aim to improve health, education, governance and economic opportunities for billions of people living in sub-standard and isolated conditions. In many places, this process is becoming increasingly *data-driven*, basing policies and actions on context-specific knowledge about local needs and conditions. Leading development organizations, with the help of research specialists like the Poverty Action Lab, undertake rigorous impact evaluations of development interventions, driven by "belief in the power of scientific evidence to understand what really helps the poor."[1]

Unfortunately, the most under-developed communities are still beyond the reach of modern data infrastructure—in areas with limited power, bandwidth, computing devices, education and purchasing power, among other constraints. Networking researchers often refer to this problem as bridging the "last mile". Even so, as the adoption of mobile phones drive rapidly-expanding network coverage, all but the most remote places seem poised to be connected. While connectivity improves the potential for effective data infrastructure, it alone does not ensure data availability. For database researchers, this last mile is our "first mile" – where essential local data is created, and the hard work of building modern data-pipelines is just beginning.

Our experience working with development organizations around the world has shown that the "first mile" still lacks critical human and institutional capacity for creating modern data-pipelines. [12]. In public health, even basic vital statistics are still not reflected in data-driven processes that affect billions of lives: for example, only 24% of children born in East and Southern Africa are registered [17].

First mile data infrastructure is crucial for delivering effective modern services. Without it, development practitioners, policy makers and communities rely on incomplete, inaccurate and delayed information for making critical decisions. The international public health community warns of an "innovation pile-up": scientific advances, such as new vaccines, will sit idle, awaiting efficient local delivery and adoption [5]. Advances in database technology suffer from a similar innovation pile-up. For want of data, some our best technologies, particularly those in data analytics, are sidelined.

In doing development-minded research, we have observed firsthand that there are many data management challenges that must be addressed to provide for effective data acquisition and interpretation within the first mile. As database researchers, we can provide tools and methods to meet these challenges. However, this requires a shift from our traditional focus on backend infrastructure and algorithms, to the needs of *local data processes* (LDPs) in data capture, quality, throughput and availability in the context of limited human, organizational and technical resources.

In this paper, we first lay out specific data management challenges that we have observed in the field. Next, we discuss promising approaches for addressing these challenges, including concrete examples from our current work. Finally, we suggest some practical next steps for the database community to engage in the first mile.

2. CHALLENGES

In organizations across the developing world, we have witnessed many first mile data challenges. Here we summarize several, with perspective from the first author's work in Tanzania and Uganda with public health and international development organizations.

2.1 Expertise, Training and Turnover

In low-resource organizations, even office-based administrative staff lack expertise in critical areas like database and computer sys-

[1] http://www.povertyactionlab.org

5^{th} *Biennial Conference on Innovative Data Systems Research (CIDR '11)* January 9-12, 2011, Asilomar, California, USA.

tems administration, form design, data entry, usability and process engineering. This is especially true for small grassroots organizations and the local field offices of international organizations, which are assigned the most critical and challenging task of actual service delivery.

It is expensive to provide training and expertise in remote and unappealing locations. For the same reason, it is difficult to recruit and retain high-quality talent. The best staff almost always leave to climb the career ladder; eventually ending up with a job in a major city, or even abroad. Turnover is very high, especially among the young, English-speaking and computer literate. This means that even those organizations that invest heavily in training see limited returns.

2.2 Storytellers versus Structured Data

The field staff of low-resource organizations often have limited formal education. Previous empirical work has shown that uneducated users have difficulty organizing and accessing information in an abstract manner [15]. These characteristics have in turn been associated with a culture of "orality" [11]. According to this theory, oral cultures are characterized by situational rather than abstract logic; preferring nuanced, qualitative narratives to quantitative data. Oral knowledge processes are also aggregative rather than analytic, preferring to assemble complex and potentially conflicting "stories", as opposed to noting down experiences as individual "correct" measurements. Finally, oral communication is usually two-way, with a concrete audience, as opposed to writing, for which the audience can be abstract, temporally and spatially removed, or not exist at all. These characteristics do not translate naturally to field workers capturing structured data using constrained forms destined for a distant, abstract recipient.

2.3 Mismatched Incentives

Like enterprises in the developed world, monitoring organizations are becoming increasingly data-driven. Indeed, the World Bank reports, "Prioritizing for monitoring and evaluation (M&E) has become a mantra that is widely accepted by governments and donors alike." [10]. On-the-ground organizations face, on one hand, growing data collection requirements and, on the other, the mandate to minimize "administration" overhead, the budget that sustains data management. Some international funders assume that if an on-the-ground organization is effective and worthy of their help, then their reporting requirements can be met with data already being collected [6]. This is wishful thinking and far from reality. Organizations in the local communities are often several rungs down on the sub-contracting or delegation ladder, and are disconnected from the rationale behind reporting requirements. Ironically, local organizations create one-off, haphazard, heavily tailored "point solutions" that minimally meet essential reporting requirements, often at the expense of local information needs. The notion of data independence is painfully missing, leading to processes that are inefficient, inflexible to change, and hard to staff.

In one large urban Tanzanian health clinic, we observed that patient visit data was recorded by hand twice and digitally entered twice. Staff wrote by hand first, in a paper filing register, which the clinic used for day to day operations, and next, on a slightly different carbon-copy form. The first copy, for the local ministry of health, was digitally entered onsite; the second copy, for an American funder, was shipped to headquarters and entered there.

Another misaligned incentive is that generating clean, aggregated, long-term data (months to years) that is useful for top-down evaluation and policy is very different from generating the more nuanced, individual, short-term data useful for decision making at the local level. For example, a funder may be interested in a quarterly count of patients with malaria, while a health clinic wants to know which malaria patients from yesterday require follow-up. In emphasizing the former, the latter is often ignored. In the example from Tanzania described above, the local health clinic had no access to digitized records, despite onsite data entry. They could only rely on searching through paper forms. In a busy, resource-constrained environment, this means that patient records were often not referenced during treatment. In turn, this lack of direct benefit creates no incentives for local practitioners to generate quality data consistently. Finally, reporting to funders means emphasizing one's successes, while improving operations often requires learning from your own mistakes. This subtle bias suggests that the most important insights from the data probably do not surface.

3. EMERGING DIRECTIONS

In this section, we propose some technical directions for addressing the challenges listed above. The general approach is to better segment the data workflow, and to either automate certain high-skill tasks, or to delegate work in ways that better suit the incentives and capabilities that are available.

3.1 Separate Capture from Structure

First of all, we believe it is important to distinguish between data *capturing* and data *structuring* tasks. The first refers to extracting some bit of information or knowledge from the real world and recording it in a persistent form. The second refers to organizing, categorizing and quantifying this information, often according to some pre-ordained structure. Our experience suggests that frontline field workers are the best suited to capturing important local information, due to their local contextual knowledge, familiarity with the community and in some cases, oral culture. On the other hand, structuring tasks require more literacy, training and knowledge about increasingly specific data vocabularies and schemas. The goal should be to move structuring tasks to where the incentives and capabilities are most appropriate.

This suggests a number of directions for future research. One project, Shreddr, described in further detail in the next section, allows field workers to capture information using existing and familiar paper forms. These forms are iteratively digitized, using a combination of automated and human-assisted techniques. Another project, Avaaj Otalo, is extracting important statistics about farm cultivation, pest infestation and mitigation directly from farmers' own recorded questions and answers [14]. Increasingly affordable technologies like GPS-enabled camera phones or digital paper[2] suggest even more powerful possibilities [9].

In general, these techniques trade off more contextually appropriate input techniques, for more uncertainty in the initial results. Capture by field agents is only the first step in a multi-stage "entropy-reduction" or "denoising" process. Down the data-pipeline, we can interleave a sequence of automated and human-assisted steps to progressively reduce noise, generating increasingly accurate statistics for decision makers, leaving intermediate results explicitly available for local analysis.

3.2 Intelligent Automation

By applying automated techniques such as optical-character recognition, voice recognition and statistical prediction, we can reduce local expertise and training requirements. For example, intelligent prediction can be used to simplify data entry. Instead of requiring a user to type in a field value, we could ask whether the most probable value is accurate. The approach of converting entry into validation can improve efficiency and quality, and can potentially even remove the requirement for keyboards and computers in some settings.

We can also use statistical techniques to more effectively organize tasks, including automatically deriving form designs based

[2]http://www.anoto.com

on prior data, including appropriate field constraints. The Usher project, described in the following section, applies these and other techniques to improve the quality of entered data. Essentially, these adaptive techniques learn from existing data, applying the results to mitigate the lack of management, formal processes, staff expertise and high turnover that can stifle other forms of organizational learning.

3.3 Leverage the Cloud and Crowd

Separating capture from structure also allows us to host more of the structuring activities directly in "the cloud", further reducing local data management requirements, and creating opportunities for more intermediaries to provide value. Both the Shreddr and Avaaj Otalo systems are positioned to be hosted services, allowing local organizations to focus on capturing paper scans and audio recordings, respectively, while the structuring tasks are distributed across the Internet. Workers can include staff at headquarters who often have the most direct incentives and motivation to obtaining timely, high-quality data. Moreover, given that many of these projects are directly in support of social goals, we may even be able to rely on "cognitive surplus" in the developed world, in the form of crowd-sourced workers working for social or other incentives [16].

In general, aggregating many "little-data" processes into a smaller number of more traditional big-data activities achieves economies of scale, and better facilitates a variety of value-adding services, including: (1) automatic value estimation, such as OCR/OMR; (2) incremental and elastic scaling of workers with crowd-sourcing; (3) dynamic task assignment according to workers' skills and incentives; (4) reporting and analytics for multiple recipients, including returning data back to the first mile for local usage.

3.4 Bottom-up Optimization

The above discussed directions on optimizing data-pipelines will make timely data available to on-the-ground staff. If we develop contextually appropriate tools that allow these users to perform analysis, we can help organizations self-optimize quantitatively.

In a rural Ugandan village, we developed a simple Excel tool allowing clinicians to view data visualizations of health trends in their community. The key idea was leveraging "found data" from the intermediate results of fulfilling external data collection requirements. The tool was simple: an Excel workbook with macros that tapped into existing data collected from community health workers (CHWs) reports. We created a workbook tab of visualizations featuring PivotCharts like "Patients under 5 years old with malaria by village.", and taught the village doctor in charge of CHWs to create his own PivotCharts. The village doctor delighted in his new-found ability to monitor CHWs through visualizations. We saw that the ability to see and benefit from CHW collected data immediately improved incentives and feedback loops for CHW data collection. Motivated by this simple tool's adoption, we have proposed, as future work, a framework for automatically generating and identifying visualizations that contain "actionable anomalies" [2].

This type of *bottom-up* analytics can improve local decision-making. It allows practitioners to surface locally-important outliers and trends, which may otherwise get lost in higher levels of aggregation. As well, it has the additional benefit of aligning mismatched incentives between local and oversight organizations. Encouraging local data-consumption feeds a virtuous cycle: growing data usage increases the desire to collect higher quality and lower latency data, which then makes the data more useful, and so on.

4. CURRENT WORK

In this section, we describe two of our projects aimed at improving the quality, efficiency and utility of data entry from paper in public health organizations in sub-Saharan Africa. We have found that batch data entry is a key choke-point for such organizations in the first mile, and an early opportunity to improve efficiency, to catch (and correct) errors in the data-pipeline, and to directly and immediately apply lessons learned.

4.1 Usher

Usher is a tool for automatically improving the accuracy of data entry interfaces. The survey design literature provides a number of existing best practices for form design [8]. However, most of these are still heuristics, and implementing them in any given context is still more of an art than a science. Drawing from these best practices, and an information-theoretic entropy reduction model of data entry, Usher seeks to automatically generate a form layout and digital data entry interface that can maximize information gain, input efficiency, and accuracy, for any arbitrary form and dataset.

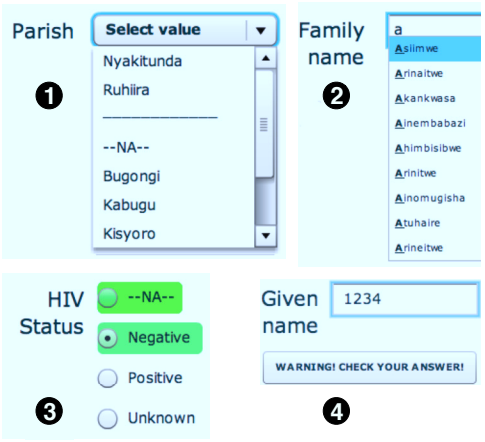

Figure 1: (1) drop down split-menu promotes the most likely items (2) text field ranks autocomplete suggestions by likelihood (3) radio buttons highlights promote most likely labels (4) warning message appears when an answer is an outlier.

Usher is driven by a probabilistic model of relationships between a form's questions and values derived from prior data. Leveraging this predictive ability, Usher provides algorithms for: (1) *reordering* the sequence of form questions to maximize information gain at every point in data entry, allowing for better prediction of remaining fields – similar to what a good form designer might do, (2) *re-formulating the presentation of questions* to make it easier to select more likely choices, and more difficult to select less likely ones (Figure 1 shows a subset of Usher-powered feedback mechanisms that we tested with users), and (3) *re-asking* questions that are likely to be wrong – approximating double-entry (the practice of having two data clerks enter the same form and comparing), but only for values likely to be incorrect, and thus at a fraction of the cost.

Our user experiments working with real datasets and real data entry clerks in rural Uganda demonstrated that Usher can significantly improve input efficiency and accuracy [3, 4].

4.2 Shreddr

Our more recent work on Shreddr takes a "column-oriented" view of data entry, with the hypothesis that automatic decomposition and information theoretic redistribution of data entry tasks, along with novel entry interfaces, can provide significant gains in data entry efficiency. Shreddr works as follows: (1) *extract schema and physical locations* of schema elements semi-automatically from a scanned form, via a simple web interface. (2) *align and shred* images of completed forms into image fragments according to physical locations, and estimate field values via optical character and

mark recognition (OCR/OMR) and Usher. (3) *dynamically re-batch* image fragments by data type and value estimate into worker tasks, and present with Usher interfaces. (4) *crowd-source or in-source* tasks to workers in an elastic labor pool (such as Amazon's Mechanical Turk), or an organization's own workers.

The Shreddr approach to data entry has several advantages. It enables low-fidelity automation to greatly simplify a large percentage of tasks. Since confirmation is often much less difficult for humans than *de novo* entry, it focuses the limited attention of human workers directly on entering the most difficult to guess values. As well, the freedom to order tasks in a "column-oriented" fashion allows control of latency and quality at field-by-field granularity. This means time-sensitive fields can be given priority, and important fields can be confirmed and re-confirmed.

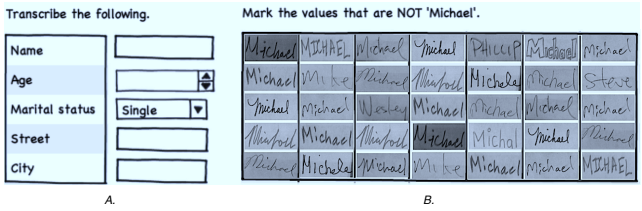

Figure 2: Two example interfaces for data entry.

Columnar-orderings enable several mechanisms for better efficiency. First, workers can better retain mental focus by transcribing similar values, without switching question context—for example, a sequence of only "firstname" values. Second, a column can be sorted by its predicted value, allowing workers to verify sequences with roughly a single value, like "Michael". We can provide user interfaces that essentially allows batch confirmation of several values at one time. We put these techniques together in Figure 2: interface A is traditional row-order entry; interface B is column-ordered validation of sorted "firstname" values predicted to be "Michael". We suspect that batch entry of pseudo-sorted sequences will yield much higher digitization throughput, much like run length encoding in a compressed database column.

5. GETTING STARTED

As we have illustrated, there are a number of first mile data challenges that can be directly addressed by re-organizing and optimizing the local data infrastructure. We believe the database community is well-positioned to make significant contributions in this area. However, to do so, we must recognize some of the implicit assumptions in current database research. We list some of these below in the hope of stimulating discussion that can advance notions about our field, of Computer Science in general, and its applicability to a number of important real-world contexts.

The notion of "too much data": William Gibson observed that "The future is already here, it is just unevenly distributed" [7]. This insight applies to data as well. While we in the database community often talk about the data deluge occurring in the developed world, there is, ironically, far too little data available about conditions in the developing world – data that is relevant to some of the most important challenges and opportunities of the 21st century. While we are very comfortable with issues like scale and privacy in data-rich environments, we are less familiar with circumstances where even the most basic improvements in data availability can enable significant progress in meeting local needs.

The infatuation with "big data": Database researchers take more interest in problems that center on large data volumes. But, because low-resource organizations use tools like Microsoft Access, they tend to fly under our radar. However, their multitude of "little data", each different by culture and environment, also presents an interesting scale problem: the challenge of wide-scale in *contextual diversity*, rather than large-scale in volume.

The myth of expertise: We often assume that competent staff is on hand to implement, administer and use our systems. This thinking is reasonable for many office-based, developed world environments, but if we want to extend the reach of our systems to more people and organizations, we must go further in terms of making our solutions more appropriate for a broader range of skill levels and familiarity with technology.

The most direct route to engaging with global problems is pragmatic. As several early researchers in this emerging field have highlighted, there is a simple formula for achieving success [1, 13]: go to the field, find a good partner organization, and solve their real problems in a empirically demonstrable and hopefully broadly generalizable way. This path leads to interesting and unexpected solutions, including some we may never have thought of otherwise.

6. REFERENCES

[1] E. A. Brewer. VLDB Keynote Address: Technology for Developing Regions, 2007.

[2] K. Chen, E. Brunskill, J. Dick, and P. Dhadialla. Learning to Identify Locally Actionable Health Anomalies. In *Proc. AAAI Spring Symposium on Artificial Intelligence for Development*, 2010.

[3] K. Chen, H. Chen, N. Conway, T. S. Parikh, and J. M. Hellerstein. Usher: Improving data quality with dynamic forms. In *Proc. ICDE*, 2010.

[4] K. Chen, T. Parikh, and J. M. Hellerstein. Designing adaptive feedback for improving data entry accuracy. In *Proc. UIST*, 2010.

[5] C. J. Elias. Can we ensure health is within reach for everyone? *The Lancet*, 368:S40–S41, 2006.

[6] B. Gates. ICTD Keynote Address, 2009.

[7] W. Gibson. The science in science fiction. *Talk of the Nation*, 1999.

[8] R. M. Groves, F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey Methodology*. Wiley-Interscience, 2004.

[9] C. Hartung, Y. Anokwa, W. Brunette, A. Lerer, C. Tseng, and G. Borriello. Open data kit: Building information services for developing regions. In *Proc. ICTD*, 2010.

[10] I. E. G. (IEG). *Monitoring and Evaluation: Some Tools, Methods and Approaches*. World Bank, Washington, DC, 2004.

[11] W. J. Ong. *Orality and Literacy: The Technologizing of the Word*. Routledge, 2002.

[12] T. S. Parikh. Engineering rural development. *Commun. ACM*, 52(1):54–63, 2009.

[13] T. S. Parikh, K. Ghosh, and A. Chavan. Design studies for a financial management system for micro-credit groups in rural India. In *Proc. Conference on Universal Usability*, 2003.

[14] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india. In *Proc. SIGCHI*, 2010.

[15] S. Scribner and M. Cole. *The Psychology of Literacy*. Harvard University Press, 1981.

[16] C. Shirky. *Cognitive Surplus: Creativity and Generosity in a Connected Age*. Penguin, 2010.

[17] UNICEF. The state of the world's children 2008: child survival, 2008.