Data in the First Mile

Kuang Chen
kuangc@cs.berkeley.edu
With Joe Hellerstein and Tapan Parikh

CIDR 2011
Mountains Beyond Mountains

The Quest of Dr. Paul Farmer, A Man Who Would Cure the World

“INSPIRING, DISTURBING, DARING AND COMPLETELY ABSORBING.”
—ABRAHAM VERGHESE, THE NEW YORK TIMES BOOK REVIEW
“database research and the data management industry are at a turning point, with unusually rich opportunities for technical advances, intellectual achievement, entrepreneurship and impact on science and society”

- Claremont Report, 2008
Talk outline

Data management in developing regions

Challenges

Research directions (and examples)
76% of children born in sub-Saharan Africa are unregistered

The last mile
The first mile
Talk outline

Data management in developing regions

Challenges

Research directions (and examples)
Lack Data Expertise
Storytellers vs. Structured Data
Access DBs everywhere...

Periodic report vs. Actionable datum

Report success vs. Analyze failure
## Database Research Assumptions

<table>
<thead>
<tr>
<th>Database research</th>
<th>Developing world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Too much data – privacy</td>
<td>Not enough data – data collection</td>
</tr>
<tr>
<td>Competent DBAs</td>
<td>Lacking technical skills</td>
</tr>
<tr>
<td>Many research opportunities</td>
<td>No novel research in applications</td>
</tr>
</tbody>
</table>
Talk outline

Data management in developing regions

Challenges

Research directions (and examples)
Goal:
Automatically **increase quality** and
**reduce expertise required** in local
data management
USHER
- CHEN ET AL. ICDE 2010, UIST 2010
Direction: Intelligent Automation

Use data-driven techniques to:

- Reduce expertise and training required
- Improve data entry efficiency & accuracy
- Identify actionable anomalies
Goal:

Improve **digitization efficiency** and local **data access**
Column-oriented Data Entry

- worker 1
  - form 1: ID
  - form 2: ID
  - form 3: ID

- worker 2
  - form 1: Village
  - form 2: Village
  - form 3: Village

- worker 3
  - form 1: Age
  - form 2: Age
  - form 3: Age
Column-oriented Data Entry

Select the values are not: Michael
Direction:
Break apart the workflow

More appropriate capture interfaces for field workers:
paper, audio, sensors

Move structuring to appropriate workers

Pool little data to make big data

Create new opportunities to add value
Summary

First Mile Challenges

Lack of data expertise

Oral cultures

Local data usability

Data research assumptions

Directions

Intelligent automation

Breaking up workflow to leveraging the cloud and crowd
How to Get Started

Find a local partner

Collaborate

Solve a real problem
Backup
Innovation pile-up