THE DATA ISN’T BIG ENOUGH

The story of the ever growing hollow haystack
A FEW ESSENTIAL CAVEATS

I work at the Trade & Competitiveness practice at the World Bank and focus on data driven products and services

- I founded the World Bank’s Open Finances program a few years ago; one of the Bank’s earliest forays into big data
- I am keenly interested in data driven entrepreneurship and innovation in emerging countries

What follows is an individual perspective on the state of open/big data at the World Bank and beyond

This is an incomplete view but I hope you will find it helpful

I use the term ‘data’ loosely in the presentation but favor examples that feature either non-traditional sources of data, or proxy data sources, or high frequency data to create insights that are either expensive, impractical, or impossible to obtain using traditional data sources or practices

My thanks to the Bank’s innovation and open data teams for their contributions to this presentation
TALKING HAYSTACKS
IN 1 MINUTE ON THE INTERNET*

- 204 million email messages
- 2 million Google searches
- 684,000 pieces of content on Facebook
- 100,000 tweets
- 47,000 app downloads in the Apple app store
- 48 hours of new video on YouTube
- 36,000 new photos on Instagram
- 34 million messages on WhatsApp

*Future Crimes – Marc Goodman
YES THE WORLD IS AWASH IN DATA

Doubling time: ~18 months

Exponential growth in computation, storage, communication capacity.

Continually **easier, cheaper, faster** to collect, store and process data.

1,000 Libraries of Congress. _Per day._

90% of the data in the world today has been created in the last two years alone. (IBM)
AND SOME OF IT IS EVEN OPEN
THE BENEFITS ARE EVERYWHERE

In the field

Nigeria – BudgIT simplifies government budgets for ordinary citizens and provides tools to visualize and share data.

India – IndiaSpend is a data journalism initiative that uses open data and more to foster better governance, transparency, and accountability in the Indian government.

Jamaica – CrimeBot alerts citizens to crimes in their neighborhood and provides different kinds of analysis of crime data.

UK – FixMyStreet lets citizens report local problems directly to relevant government agencies.

Macedonia – a UNDP initiative combined SMS, open data, and other data sources to create a service that helped farmers reduce pesticide use by 30%.

Malawi – Malwivote helps voters find their registration centers.

Kenya – an NTV journalist produced a series of stories linking grades with sanitation facilities.

Philippines – CheckMySchool provides information on services and facilities in schools.

USA – BuildingEye provides ‘mashed-up’ information to government agencies and citizens about building permits, planning applications, business licenses, public events, etc.

In the boardroom

USA – Climate Corporation sells highly localized crop insurance products (the company was recently sold for almost a billion dollars).

Mexico – Medii.co helps consumers compare drug prices in different pharmacies in Mexico.

India – MandiTrades provides commodity price information to farmers/agribusiness.

Indonesia – UrbanIndo connects property buyers and sellers.

Chile – Junar is a cloud based open data publishing platform.
A DEFINITE ROLE IN ENDING EXTREME POVERTY AND PROMOTING SHARED PROSPERITY

Measuring the world
Daily satellite nightlights imagery to monitor power outages/load shedding in Indian villages

Understanding the world
Combining vehicle GPS data and open source software to lower the financial and technical barriers to evidence-based traffic management and planning

Changing the world
Helping mobile money operators use cellphone behaviors to better target potential customers, increasing financial inclusion, bring banking to the ‘unbanked’.
A GROWING PORTFOLIO AT THE BANK

Mapping basic amenities in informal settlements

Sensor Generated Big Data for Safer Mass Tourism in Uttarakhand: Preparing for the Next Himalayan Disaster

Using Geo-spatial Data to Secure Property Rights

Using big data to conduct rapid disaster damage assessments

Big Data Harmonization for Improved Governance in Ukraine

CommunityCollect: Increasing Transparency and Integrating Project Data for CDD Projects through a Cloud-Based Application

The Livestock Early Warning System (LEWS) in Mongolia: Where Big Data meets Social and Cognitive Realities

Making Sense of Global Urban Data
OPENING THE WORLD
A REVOLUTION IN SLOW MOTION
With all the data in the world, we still can’t provide reliable estimates of –
- Women who are impoverished or who die in childbirth
- The size of the non-formal economy (or sometimes even sectors such as agriculture)
- The number of births and the causes of death

When we think we have the numbers, we frequently get them wrong –
- In 2007—the start of the global financial crisis—estimates of the personal savings rate in the United States were negative, at −2.5% (Federal Reserve Bank, 2014). The estimate was later revised to +2.5%
- GDP revisions are understandable but by 60% (Ghana and Nigeria)?

When we really need the data, we can’t produce it –
- Demographics of exporters in a country
- Measures of global value chains
- Mapping of trade logistics

A WORLD OF DATA MOATS

A shift in the data balance – the most valuable data is now private

When data does exist, it can be hard to access or share –

- Privacy concerns
- Regulatory impediments
- The lack of standards for sharing data
- Nascent understanding by most data owners of the benefits of sharing
- The cost of the resources needed to carry out the process of data sharing

Image source – http://xkcd.com/802/
A SPECIAL CHALLENGE IN EMERGING ECONOMIES

Data
- Most big/open data programs are superficial; too much data – especially local, granular, current – is still unavailable

Policy
- Significant policy gaps discourage both the publication and use (especially commercial) of data

Capacity/Literacy
- Few government policy makers are trained to understand the potential of data in governance
- Limited understanding of open data within critical communities like entrepreneurs, investors, journalists, CSOs, and development practitioners

Innovation
- Limited support/infrastructure for innovation and incubation of new data driven ideas and companies

Financing
- Many government data programs face significant financing challenges
- Access to capital, especially post-seed stage capital, is a significant challenge for open data driven companies

Result measurement
- The current focus is on supply, rather than demand
- Measurements are focused on inputs, rather than outputs

Data puritanism
- Open data purists are unable to see the larger data landscape
- Most data initiatives focus only on data, ignoring ‘ecosystem’ challenges associated with policy, engagement, capacity, use, financing, and the like
ONE NEEDLE AT A TIME
THERE’S A ROLE FOR ALL OF US

The World Bank is a leading proponent of and contributor to the ‘data revolution’ within UN agencies, international organizations, NSOs, and more.

There is increasing demand from clients who are overwhelmed by the abundance of data in governance.

- ‘Conflict countries were once considered data deserts and today they are data blizzards’*

The World Bank recognizes the value of data in achieving better program results.

The competitive landscape of development is changing and data is becoming a major differentiator.

AREAS OF INTEREST

What are the policy implications of open and big data

What is the specific impact and potential of data driven interventions in developing countries

How do you develop data smart skills at all levels within our partner countries
  • Policy makers
  • Civil society
  • Media
  • Private sector
  • Academia

Does the Bank leadership itself have well rounded data skills

How do we leverage the skills of the data community at large around the world
THANK YOU
## FUTURE DATA OPPORTUNITIES

<table>
<thead>
<tr>
<th></th>
<th>Satellite/Aerial</th>
<th>Smartphone</th>
<th>Fixed sensor</th>
<th>CDR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cost of collection</strong></td>
<td>Moderate</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Ease of collection</strong></td>
<td>High</td>
<td>Moderate</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Current use</strong></td>
<td>Moderate</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Future changes</strong></td>
<td>Higher resolution</td>
<td>Falling costs =&gt; much higher penetration</td>
<td>Falling hardware costs</td>
<td>Shift from calls/SMS to ‘hidden’ instant messaging</td>
</tr>
</tbody>
</table>
| **Example uses**       | • Electrification (nightlights)  
                         • Urban extent  
                         • Formal/informal settlements  
                         • Damage assessment  
                     | • Mobility  
                         • Citizen feedback  
                     | • Traffic (foot/vehicle)  
                         • Environmental quality  
                         • Service quality (utilities)  
                     | • Granular population distribution  
                         • Patterns of mobility  
                         • Social network |