The Big Data Revolution for Sustainable Development

XI MEETING OF THE LATIN AMERICA AND THE CARIBBEAN MONITORING AND EVALUATION NETWORK

Santiago de Chile, June 28-30, 2016

Anupam Anand
Evaluation Officer
WHAT WE WILL TALK ABOUT

- What is big data?
- Why do we want big data for sustainable development?
- What questions can we answer with big data?
- Challenges, limitations and lessons from using big data
What is BIG DATA?

- No fixed definition
- Data sets that are so large or complex that traditional data processing applications are inadequate
- Characterized by
  - Volume from various sources needing large storage
  - Velocity at which they are generated
  - Variety of unstructured formats needing additional processing
  - Value or meaning not immediately apparent

**Volume**

- 40 Zettabytes (40,000,000,000,000,000,000 bytes)
- 2.5 Quintillion bytes

**Velocity**

- 1 TB of trade information
- 100 sensors

**Variety**

- 30 billion pieces of content shared on Facebook every month
- 4 billion+ hours of video are watched on YouTube each month
- 400 million tweets sent per day by about 200 million monthly active users

**Veracity**

- 27% of respondents

---

**The FOUR V’s of Big Data**

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that we rely on every day. But what exactly is Big Data and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015, 4.4 million IT jobs will be created globally to support big data with 1.9 million in the United States.

1 in 3 business leaders don’t trust the information they use to make decisions.

---

**Sources:**
- McKinsey Global Institute
- Twitter, Cisco, Gartner, ENR, SAS, IBM, NIST, EE, QRS
What can we use BIG DATA for?

- Foster Decision Making and Accountability
- Where are the funds going?
- Is funding going to the right places?
- Monitoring & Evaluation
- What changes occurred over time?
- Did the intervention cause the change?
- What other factors might have led to the outcome?
Q1: How many SDG Goals, Targets and indicators are there?
A: SDGs- 17 goals, 169 targets and 230 indicators
Why use BIG DATA for SDG?

- Scarcer financial resources
  - Need to target interventions where most needed
- Greater demand for transparency and country ownership
- Monitoring of the progress
- Need objective evidence base for decision-making
Big data such as from satellite imagery and sensor networks make environment and development indicators increasingly measurable.
The GEF and the SDGs

- GEF support closely aligns with the SDGs on climate, oceans and marine resources, terrestrial ecosystems, forests, biodiversity and land degradation.
- The creation of more than 3,300 protected areas covering 860 million hectares.
- Conservation-friendly management of more than 352 million hectares of productive landscapes and seascapes.
- 790 climate change mitigation projects contributing to 2.7 billion tonnes of GHG emission reductions.
- Sustainable management of 34 transboundary river basins in 73 countries.
How are we leveraging Big Data at GEF-IEO
Big data for Biodiversity

- **Goal 15**: Sustainably manage forests, combat desertification, halt and reverse land degradation, **halt biodiversity loss**

- **Indicators**
  - Annual change in forest area and land under cultivation* - Geospatial data
  - Area of forest under sustainable forest management as a percent of forest area - Geospatial data/Administrative data
  - Red List Index - Telemetry Tracking Data/International monitoring
  - Protected areas overlay with key biodiversity areas (KBAs)
Visualization of geographical context

1292 GEF-supported protected areas

~2.8 million km$^2$ in 137 countries

Where are the funds going?
Is funding going to the right places?
Overlay of project sites with scientific criteria

Use of global datasets + GIS analysis to determine overlaps of GEF support with critical sites
What changes occurred over time?

Analysis of forest cover change

Extraction of satellite data for 30,000 GEF and non-GEF sites

30-m resolution (LANDSAT) for 12-year period
Did the intervention cause the change?
Quasi-experimental analysis

Propensity score matching found appropriate counterfactuals using 9 socioeconomic and biophysical variables
Identify the drivers

Images at 2.5 to 0.5 m resolution used to identify drivers of change that hinder success of GEF support.
What other factors might have led to the outcome?

Use of contextual variables in different formats to assess correlations with changes
Big data for Land degradation

- **Goal 15**: Sustainably manage forests, combat desertification, halt and reverse land degradation, halt biodiversity loss
  - Indicator for Goal 15
    - Annual change in degraded or desertified arable land (% or ha) - Remote sensing/satellite and administrative data.

- UNCCD Indicators for Land Degradation Neutrality (LDN)
  - Vegetation productivity (NDVI)
  - Landuse and landcover change
  - Carbon sequestration
Geospatial Impact Evaluation

“Treatment”

GEF Project Locations (i.e., area under restoration project)

Candidate Control Locations (can search across any set of relevant geographies)

“Best Match Control”

Image: Aiddata
Big data for Climate Action

Areas vulnerable to sea level rise

Aboveground biomass

Image: NASA

Saatchi et al, PNAS, 2011
Big data needs big tools
Planetary level cloud computing with Google Earth Engine

10 years desktop computing = 7 days cloud computing
Machine learning and modelling

Data-hungry algorithms required multiple global datasets of

<table>
<thead>
<tr>
<th>Data</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>agricultural production</td>
<td>FAO, 2012</td>
</tr>
<tr>
<td>export of agricultural product</td>
<td>FAO, 2012</td>
</tr>
<tr>
<td>trade of agricultural product</td>
<td>FAO, 2012</td>
</tr>
<tr>
<td>urban population</td>
<td>FAO, 2012</td>
</tr>
<tr>
<td>rural population</td>
<td>FAO, 2012</td>
</tr>
<tr>
<td>Gross domestic product</td>
<td>world bank, 2015</td>
</tr>
<tr>
<td>rule of law</td>
<td>world bank, 2013</td>
</tr>
<tr>
<td>control of corruption</td>
<td>world bank, 2013</td>
</tr>
<tr>
<td>monitoring capacity</td>
<td>Romijin et al (2012)</td>
</tr>
<tr>
<td>International aid</td>
<td>Aid data (2010)</td>
</tr>
</tbody>
</table>

Variables and its sources used in the regression
and decision tree analysis
Challenges and Limitations

- High computing power and technical skills needed
- Uneven availability and accuracy of contextual variables
  - often vary widely across countries and sites
- Cannot answer “how” and “why” questions
- Data only as good as available resolution
  - still need to do field verification/ ground truthing
- Still need to account for possible biases in data collection methods
- Legal issue
Solutions and Lessons

- Partner with global institutions with access to and infrastructure for using big data
- Used mixed approaches and methods
  - complemented global analyses with case study and portfolio analyses to triangulate findings
- Continue exploring use of new technology
  - drones, deep learning, internet of things, sentiment analysis, social media analysis, etc.
- Approach evaluation as a dynamic learning process
  - new data sets, approaches, issues will always emerge!
PARTNERS

• University of Maryland

• WCPA-SSC Joint Task Force on Biodiversity and Protected Areas at IUCN

• National Aeronautics and Space Administration (NASA)

• AidData
Thank you!

For more information, visit www.gefieo.org