TREVOR MONROE & AIVIN SOLATORIO

DEVELOPMENT DATA GROUP, WORLD BANK

MEASDEV2020

TOOLS & TRADECRAFT DATA FUSION

Credit - Damien Jaques; Trevor Monroe
SURVEY DATA VS BIG DATA

**SURVEY DATA**

**The Good:**
- Representative
- Standard Errors known
- Fit for Purpose

**The Bad:**
- Costly
- Gaps & Lags in Coverage

**BIG DATA**

**The Good:**
- Big
- Always on
- Non-reactive

**The Bad:**
- Non-representative
- Confounded
- Drifting
- Incomplete
“BIG DATA INCREASES THE VALUE OF SURVEY DATA”

~Mathew Salganik, Bit by Bit

Common approaches for data integration and fusion

BIG DATA (wide, thin) + Survey (narrow, thick) = More robust & dynamic data product than either data source alone

Enriched

Record Linkage

Amplified

Big data source

Estimate Model

Survey Data

Imputed Survey Data

Predict

More robust & dynamic data product than either data source alone
“THE DATA REVOLUTION IS MAKING FREE DATA CHEAPER”

- JED SUNDWALL, AWS OPEN DATA
Making Data Fusion Easier

- Data Catalogs
- Training Data Repositories
- Rich Context Search
- Standards (STAC, DDI, BDGMM, IPUMS)

- Micro-tasking (labels)
- Distributed Collection
- Embedded Surveys
- Social APIs

- Interoperability standards, practices
- Packages for feature engineering, modelling
- Privacy Preserving Methods

- Collaborative Coding Practices, Repos, Tools
- Open Data Science
# Satellite Data Fusion Overview

## Overview
- Seeing dramatic improvements in Analysis Ready Data (COG, STAC)
- Integration of satellite survey-based and ground data mainly happens via geographic matching
- Pixel values are matched or used to train classifiers to proxy socio-economic activity

## Data
- Proliferation of public and commercial satellite data (Earth on AWS, GEE)
- Growing repositories of labeled data catalogs (Spacenet, UCI, OSM)
- Growing collections of geo-referenced ground data (LSMS, DHS HFS), and good practice for geo-referencing ground data

## Challenges
- Use of geo-referenced household survey data as training data is constrained by privacy issues
- Ability for the masses to readily use satellite data (ARD)

## World Bank Examples

<table>
<thead>
<tr>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Resolution Electrification Access</td>
<td>Measurements; Nightlight + ground surveys — Kwawu Gaba (WB); Brian Min, UM</td>
</tr>
<tr>
<td>High Resolution Crop Predictions measurements</td>
<td>Satellite + survey, Talip Killic (WB); Lobell, Burke (Stanford)</td>
</tr>
<tr>
<td>Pixel measures of economic activity</td>
<td>Using satellite + economic, aux data — Somik Lal (WB); Gordon Hanson;</td>
</tr>
<tr>
<td></td>
<td>Ran Goldblatt</td>
</tr>
<tr>
<td>Small Area Estimates of Poverty</td>
<td>Using census + survey — Newhouse (WB); Hersh; Engstrom (GW)</td>
</tr>
</tbody>
</table>
SATCHEL APPLICATION – ECONOMIC ACTIVITY

- WB Data Catalog (GDP)
- Earth AWS
- GEE
- OSM (objects)
- PopGrid
- STAC/COG

- MTurk (labeling)
- Hive (Labeling)

- Deep Learning
- GUF/WSF
- EO-Learn
- PDAL | GDAL

- Google Earth Engine (algorithms)
- Github (algorithms training)
- Jupyter Notebook (training)
## TEXT DATA ANALYTICS

### Overview:
- Utility of unstructured yet massively available content of useful information
- Traditional Natural Language Processing and application of Deep Learning techniques using transformer-based models, e.g., BERT
- Unsupervised and supervised techniques to generate value, e.g., recommendations engine, predict famine, and estimate employment statistics.

### Data:
- Web-scale data of text content available through web scraping.
- Availability of publicly contributed text content on social media platforms such as Twitter.
- Internal archive of documents of companies and institutions.
- Tagged and curated content from archives and posts.
- Manually labelled data sourced via platforms like MTurk.

### Challenges:
- Text data is unstructured
- Content extraction is largely dependent on the corpus and kind of document
- Labelled data, need more training data

### World Bank Applications
- **Predicting Unemployment** – Sam Fraiberger (WB)
- **Predicting Leading Indicators** – Sam Fraiberger (WB)
- **Rich Context Micro Data Search** – Olivier Dupriez (WB)
TEXT DATA APPLICATION

- Discovery
- Integration
- Collection
- Open Analytics

Main Approaches:
- Enriched
  - Record Linkage

Amplified
- Big data source
  - Predict
  - Imputed Survey Data
- Survey Data
- Demographic Data

Main Tools:
- Twitter
- News
- MTurk (labeling)
- Qualtrics (survey/tables)
- Github (algorithms training)
- Jupyter Notebook (training)
- BERT
- Deep Learning
- Supervised model

Unemployment Statistics

- Github (algorithms training)
- Jupyter Notebook (training)
### Overview:
- Development applications typically use mobile data for **real-time awareness, real-time feedback, and prediction** towards crisis response, urban/transport planning, dynamic population & poverty estimates.
- Data integration often done via geographic matching. Individual-level mobile metadata is aggregated (e.g. via tower locations) to match socio-economic phenomena on local levels.
- Machine Learning and high frequency surveys used to improve features, train classifiers, predictions.

### Top Sources
- Call Detail Records
- Smart Phone Traces (App)
- Cell Signals
- Telemetry
- Geo-referenced Social-Media
- In-situ sensors

### Challenges
- **Access:** call record data in developing countries is ubiquitous but not consistently accessible. Need pathways to scale.
- **Privacy:** need capacity for wide use of privacy preserving methods, homographic encryption.
- **Interpretability:** transparency and interpretability of algorithms.
- **Representation:** mobile-based survey data underrepresents population, gender.

### World Bank Examples
- **Access to opportunity measures**
  - Mobile + satellite—Nancy Lozano (WB); Flowminder
- **Transport Planning—Mobile survey**
  - Fatima Arroya; Dunstan Matekenys (WB); Marta Gonzales, UCB
- **Estimating Poverty**
  - Mobile + admin data—Marco Hernandez; Friaz-Martinez (UMD)
MOBILE DATA APPLICATION – CRISIS RESPONSE

- Flowkit (mobile)**
- UT CDR toolkit (mobile)**
- OPAL**
- Social Marketing API
- Cubiq/Waze/Mapbox
- OpenCellID
- HotOSM (map tasking)
- PDNA (survey)
- Qualtrics (labeling)
- MTurk (labeling)
- G-DIF
- GRID 3
- Scikit - mobility
- Flowkit
- QGIS/GEE
- UT CDR toolkit (mobile)**
- OPAL**
- Social Marketing API
- Cubiq/Waze/Mapbox
- OpenCellID
- HotOSM (map tasking)
- PDNA (survey)
- Qualtrics (labeling)
- MTurk (labeling)
- G-DIF
- GRID 3
- Scikit - mobility
- Flowkit
- QGIS/GEE

- QGIS (training)
- Google Earth Engine
- Github (algorithms training)
- GEE (algorithms/training)
- Jupyter Notebook (training)
“SOCIAL RESEARCH IS BORROWING PRACTICES FROM SOFTWARE DEVELOPMENT FOR REUSABILITY”

– SUSAN ATHEY

MEASURE DEVELOPMENT CONFERENCE, 2018
<table>
<thead>
<tr>
<th>GENERAL PURPOSE</th>
<th>RESOURCES</th>
<th>TOOLS</th>
<th>PRACTICES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Data Science</td>
<td>Python</td>
<td>R</td>
<td>GitLab</td>
</tr>
<tr>
<td>Interoperability</td>
<td>Tidy Data</td>
<td>Slippy</td>
<td>GeoStat</td>
</tr>
<tr>
<td>Ethics &amp; Privacy Preserving Methods</td>
<td>UN Handbook of Privacy Preserving Methods</td>
<td>Differential Privacy</td>
<td>UNGP Risk Assessment Tool</td>
</tr>
<tr>
<td>Rapid Data Collection, Tasking, Labeling</td>
<td>MTURK</td>
<td>AWS Sagemaker Ground Truth</td>
<td>Figure8</td>
</tr>
<tr>
<td>Data Catalogs</td>
<td>WB Data Catalog</td>
<td>Google Data Search</td>
<td>Open Street Map</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STEP</th>
<th>SATELLITE</th>
<th>MOBILE</th>
<th>TEXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLLECTION - High Frequency Data Collection; Data Labelling;</td>
<td>Label Maker</td>
<td>Cumulus</td>
<td>MTURK</td>
</tr>
<tr>
<td>DISCOVERY – Training Data Sets; Rich Context Data Search; Knowledge Products</td>
<td>Earth on AWS</td>
<td>GeoNet</td>
<td>Nasa Earth Science</td>
</tr>
<tr>
<td>INTEGRATION &amp; ANALYTICS Analytic tools; privacy preserving methods; integration frameworks</td>
<td>Grid3</td>
<td>GEE</td>
<td>eo-learn</td>
</tr>
</tbody>
</table>

* FOR MORE TOOL SEE ODS | FOS4G

**Data Fusion Product Starter Kit**

[**FOR MORE TOOL SEE ODS | FOS4G**](#)
“IN A WORLD OF BIG DATA, WE WILL NEED MORE SURVEYS”
– MATHEW SALGANICK

“We need to make free data cheaper”
– Jed Sundwall, AWS

“Social science research is becoming more like software development” – Susan Athey

“Institutional innovations are needed to ensure algorithms are responsibly used” – Duncan Watts