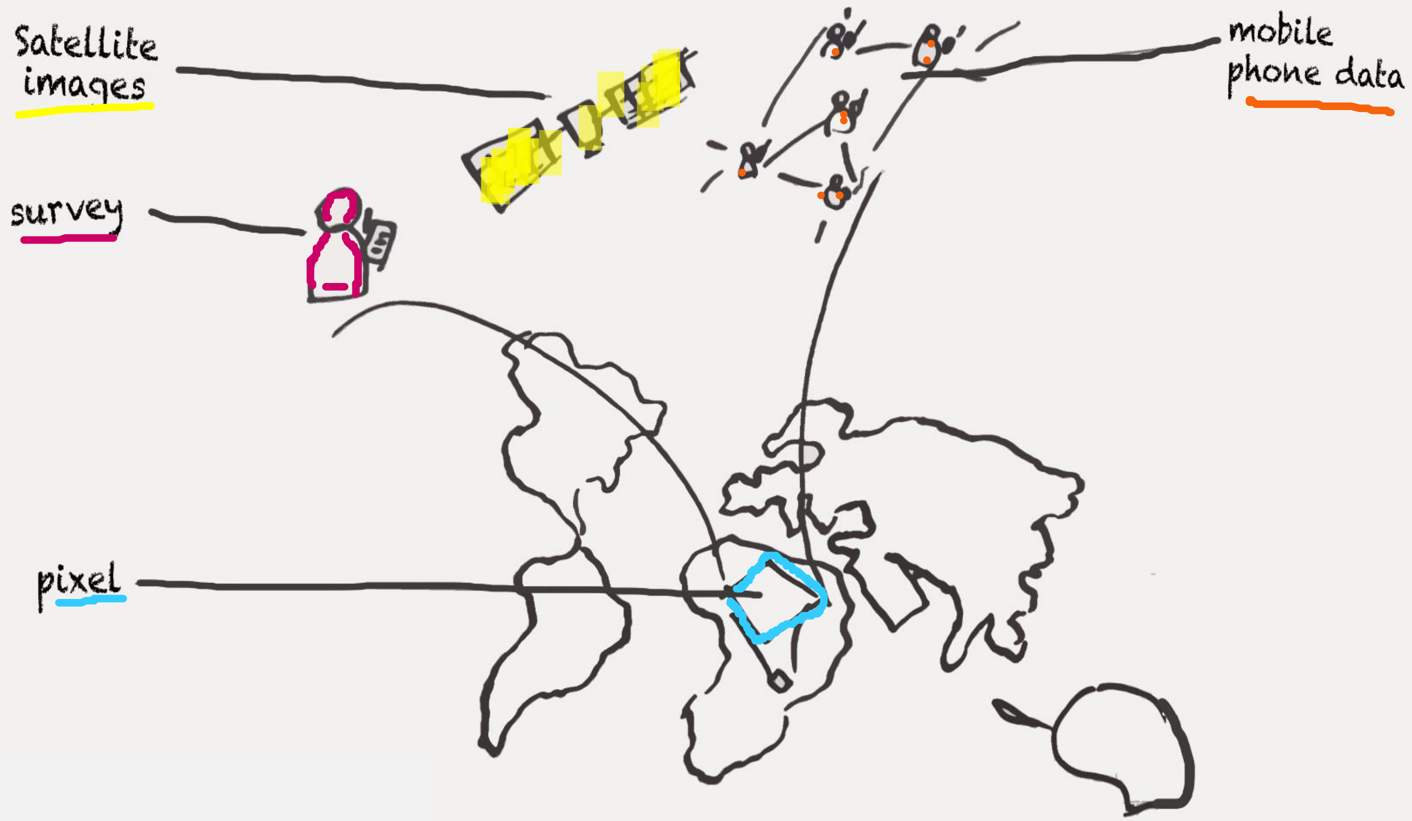
TREVOR MONROE &
AIVIN SOLATORIO

DEVELOPMENT DATA
GROUP, WORLD BANK

MEASDEV2020

TOOLS & TRADECRAFT DATA FUSION



THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Data

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) →

Enriched

BIG DATA ⟷ **Record Linkage** ⟷ Survey

Amplified

Big data source → Estimate Model ← Survey Data

Estimate Model ↓ Predict

Big data source → Predict → Imputed Survey Data

=

More robust & dynamic data product than either data source alone

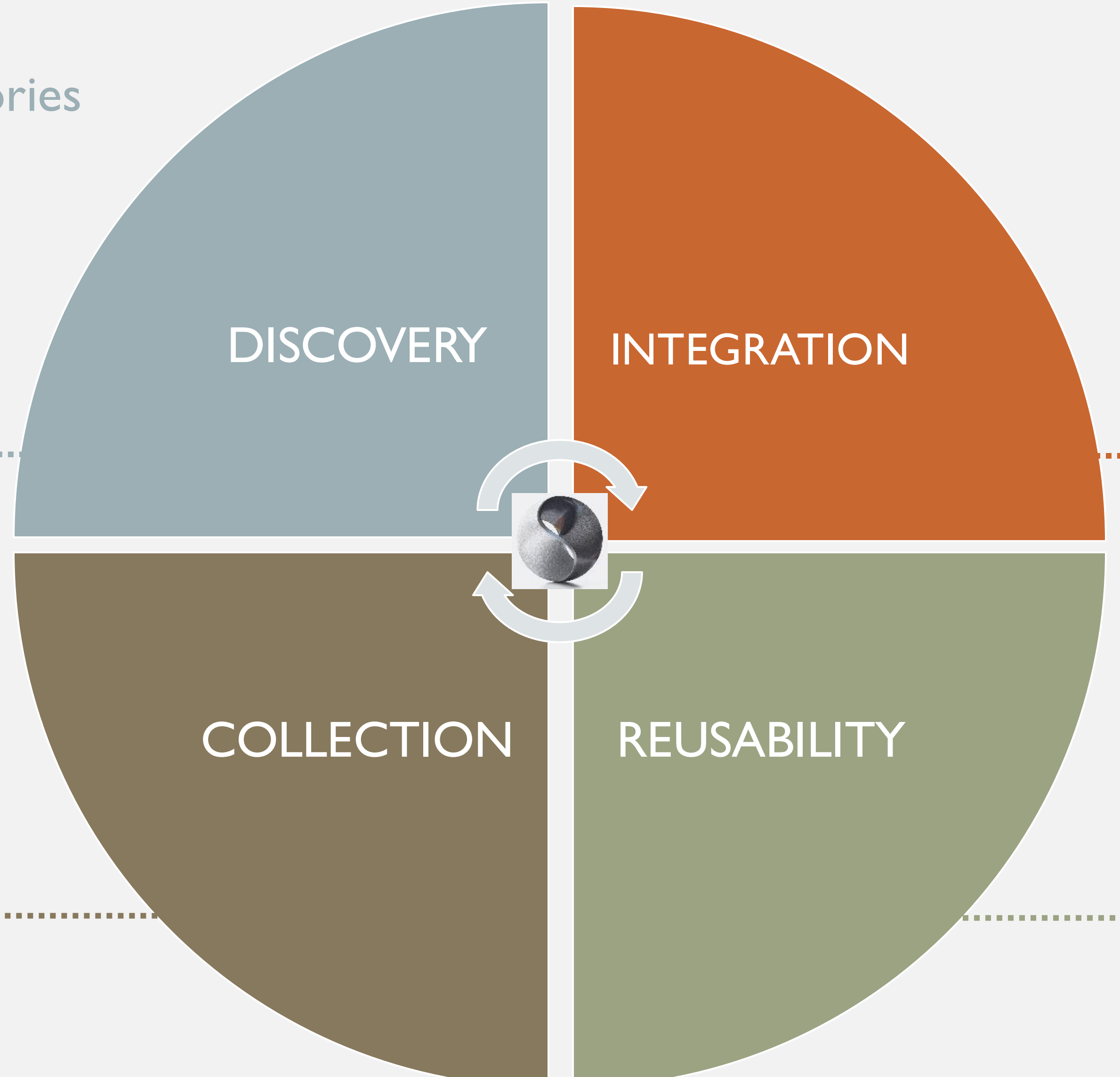
“THE DATA REVOLUTION IS
MAKING FREE DATA CHEAPER”

\- **JED SUNDWALL**, AWS OPEN DATA

MAKING DATA FUSION EASIER

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

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



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

- 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

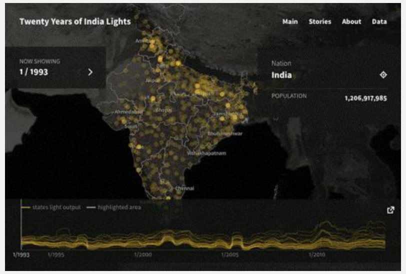
Data

- 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

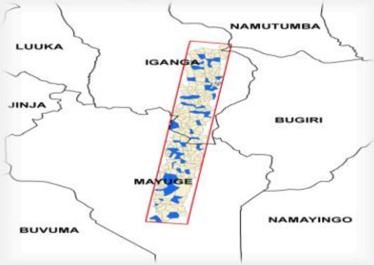
Challenges

- 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)

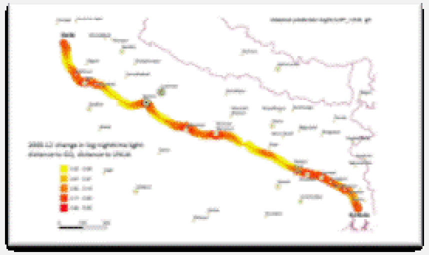
World Bank Examples



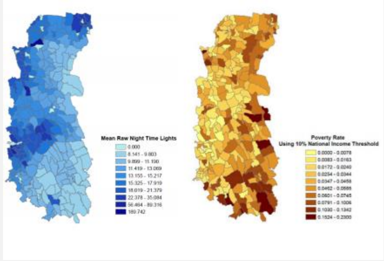
High Resolution Electrification Access measurements; Nightlight + ground surveys– *Kwawu Gaba (WB); Brian Min, UM*



High Resolution Crop Predictions measurements, satellite + survey, – *Talip Killic (WB); Lobell, Burke (Stanford)*

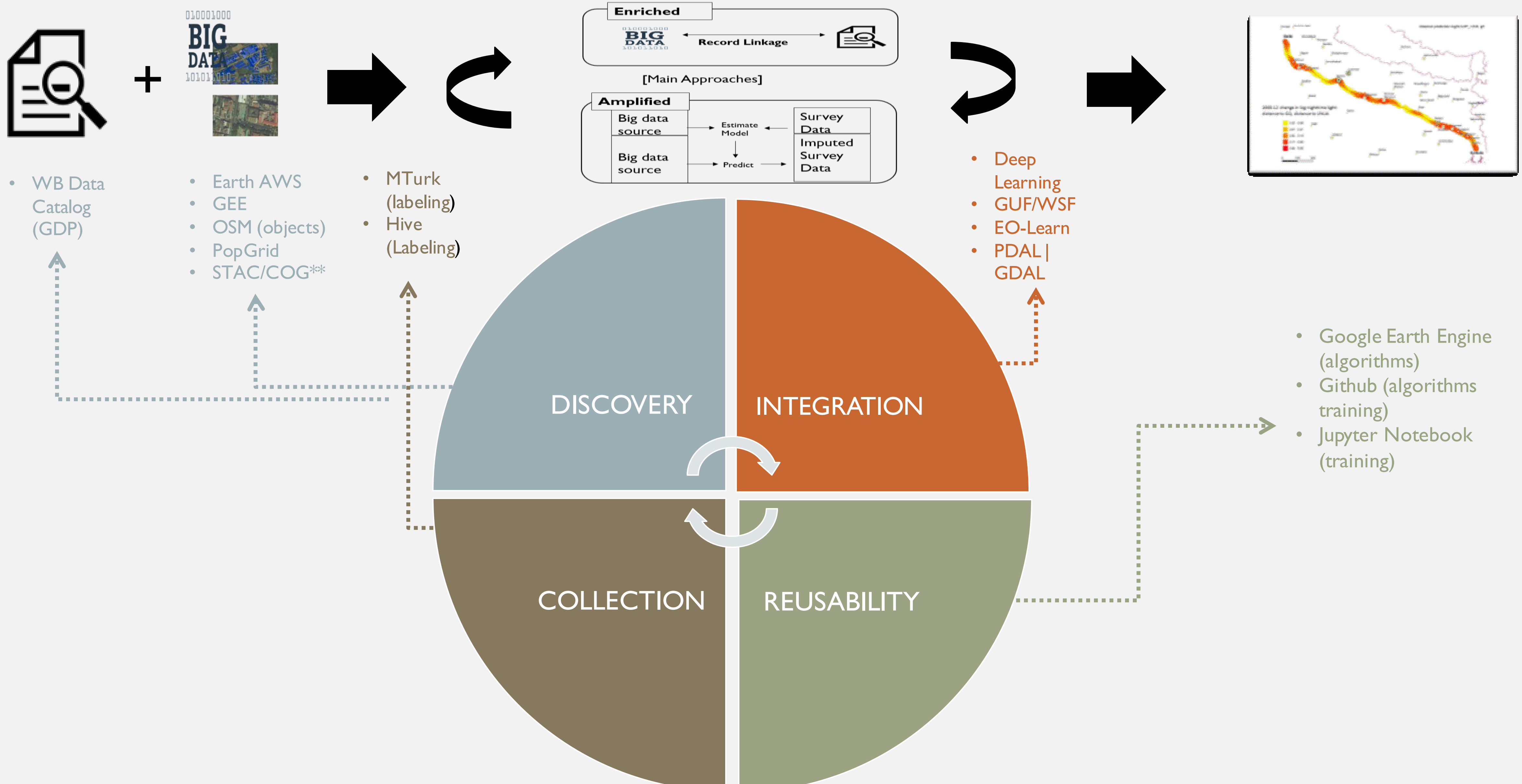


Pixel measures of economic activity using satellite + economic, aux data – *Somik Lal (WB); Gordon Hanson; Ran Goldblatt*



Small Area Estimates of Poverty using census + survey – *Newhouse (WB); Hersh; Engstrom (GW)*

SATELLITE APPLICATION – ECONOMIC ACTIVITY



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.

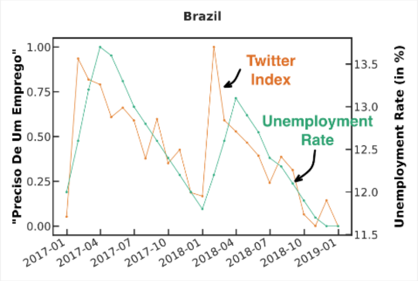
Data

- 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.

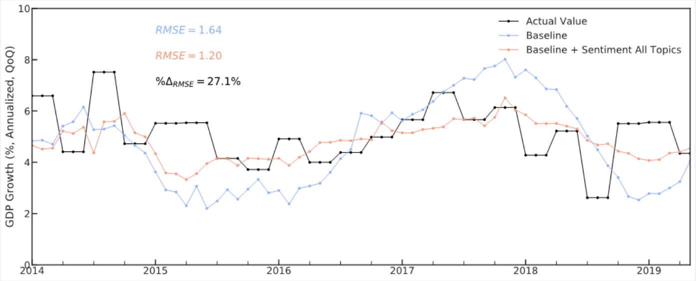
Challenges

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

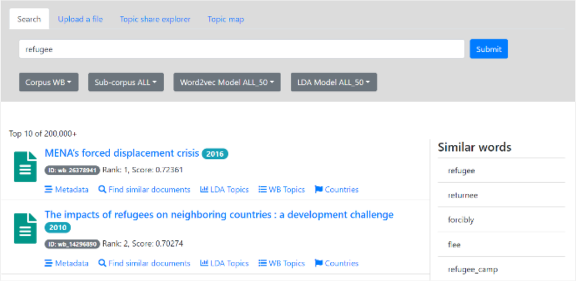
World Bank Applications



Predicting Unemployment – *Sam Fraiberger (WB)*



Predicting Leading Indicators – *Sam Fraiberger (WB)*



Rich Context Micro Data Search– *Olivier Dupriez (WB)*

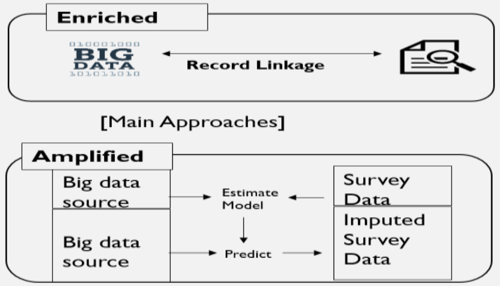
TEXT DATA APPLICATION

+

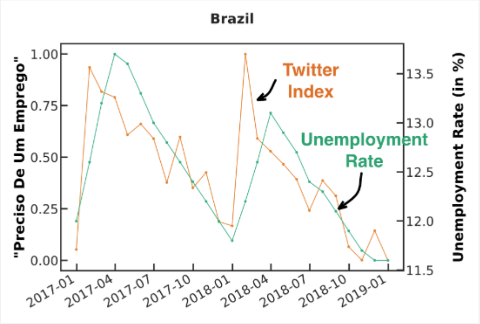
- Twitter
- News

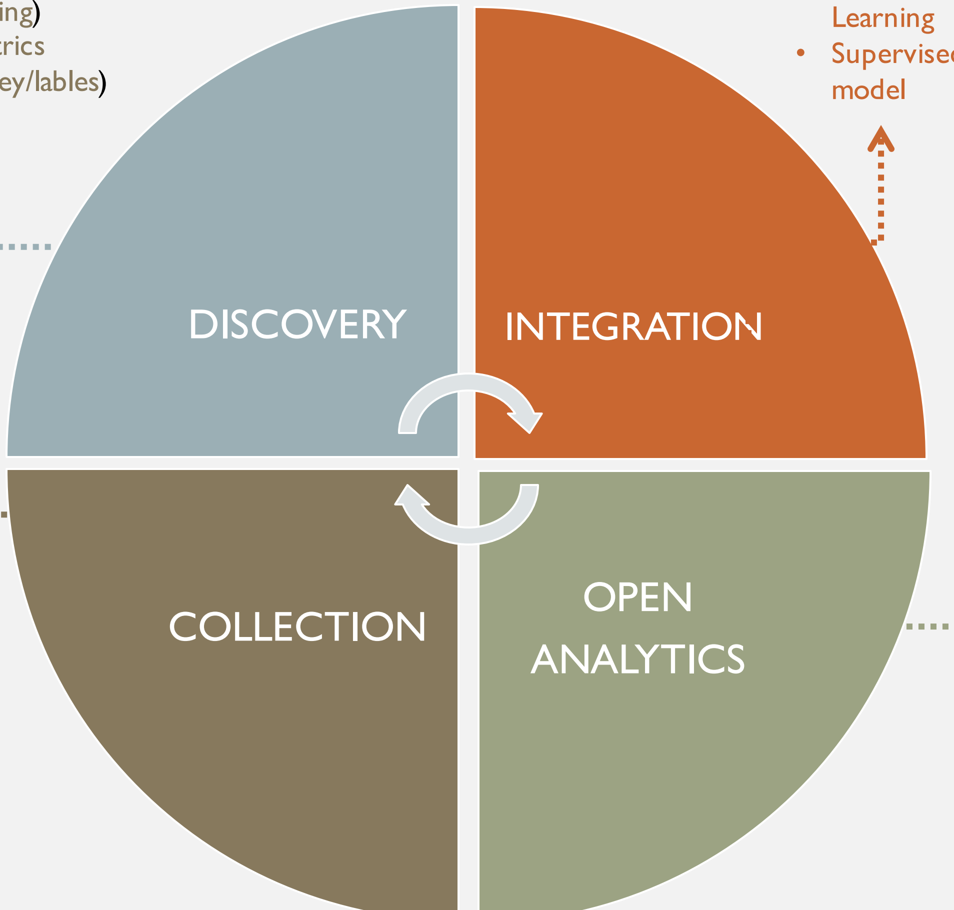
Emplyment Statistics

- MTurk (labeling)
- Qualtrics (survey/lables)



- BERT
- Deep Learning
- Supervised model





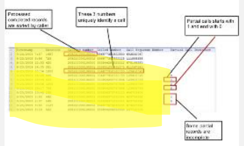
- Github (algorithms training)
- Jupyter Notebook (training)

MOBILE DATA OVERVIEW

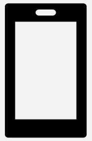
Overview:

- 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



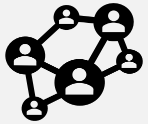
Call Detail Records



Smart Phone Traces (App)

Cell Signals

Telemetry



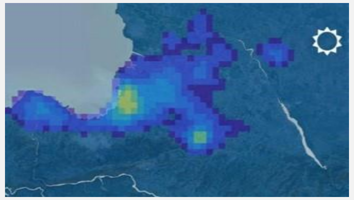
Geo-referenced Social-Media

In-situ sensors

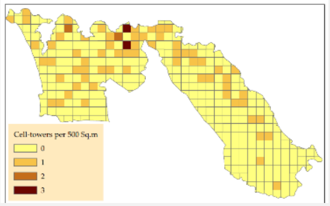
Challenges

- **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

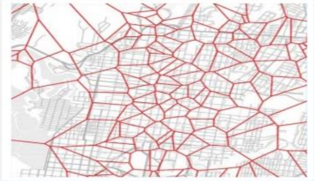
World Bank Examples



Access to opportunity measures - Mobile + satellite– *Nancy Lozano (WB); Flowminder*

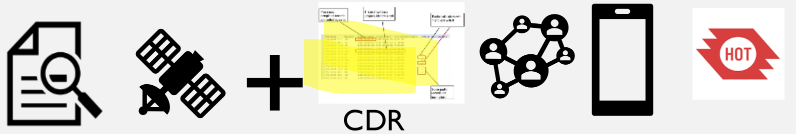


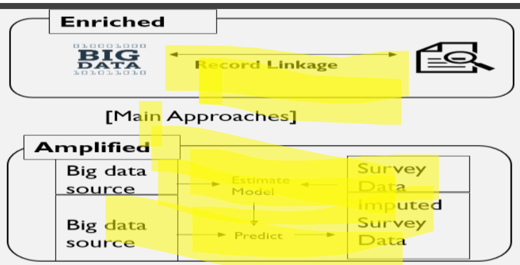
Transport Planning- Mobile + survey– *Fatima Arroyo; Dunstan Matekenys (WB); Marta Gonzales, UCB*

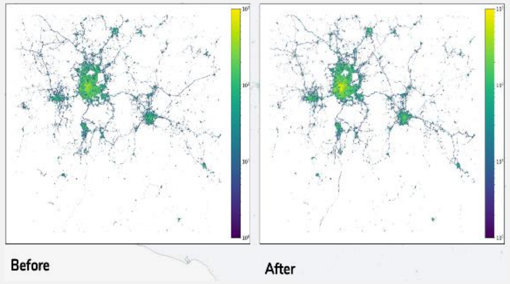


Estimating Poverty - Mobile + admin data– *Marco Hernandez; (WB); Friaz-Martinez (UMD)*

MOBILE DATA APPLICATION – CRISE RESPONSE







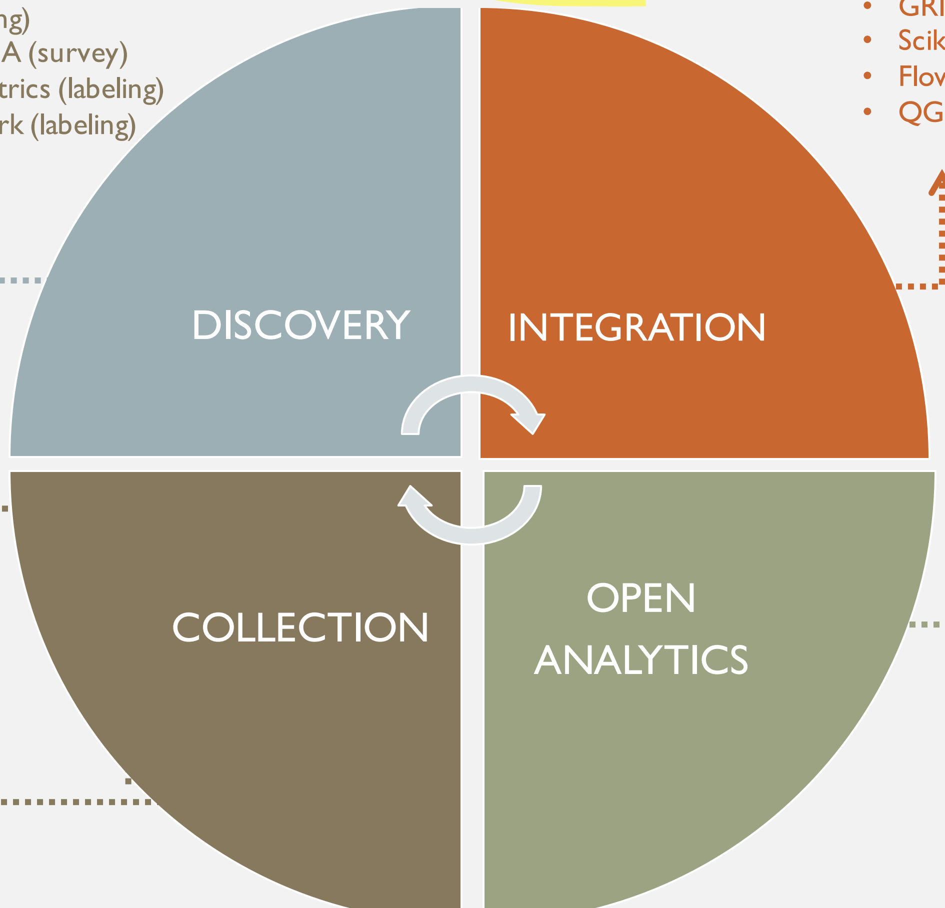
Micro-Data Catalog (WB)

- Earth AWS
- OSM
- PopGrid

- 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



- 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

DATA FUSION PRODUCT *STARTER KIT*

* FOR MORE TOOL SEE ODSC | FOSS4G

GENERAL PURPOSE	RESOURCES | TOOLS | PRACTICES
Open Data Science	Python | R |Gitub | GitLab | | Jupyter | Anaconda | Kaggle | Cookie Cutter Data Science | Docker | BDAS | BITTS | PySal |
Interoperability	Tidy Data | Slippy | GeoStat | World Pop – GRID3 | NIST | Analysis Ready Data | IPUMS | SDMX| DDI | DublinCore | Schemas.org | ONS-ML
Ethics & Privacy Preserving Methods	UN Handbook of Privacy Preserving Methods | Differential Privacy | UNGP Risk Assessment Tool | Deon DrivenData Checklist
Rapid Data Collection, Tasking, Labeling	MTURK | AWS Sagemaker Ground Truth | Figure8 | Hive Data | Samasource |Qualtrics Prolific | Premise | Native | Snorkle | Kobo | Qfield |RapidPro SurveyCTO| Survey Solutions | Geo-referencing data for ML
Data Catalogs	WB Data Catalog | Google Data Search | Open Street Map | Enigma | AWS Open Data | WorldPop | Kaggle| Awesome Satellite Data

STEP	SATELLITE	MOBILE	TEXT
COLLECTION - High Frequency Data Collection; Data Labelling;	Label Maker | Cumulus | MTURK |Hive Data	Flowkit | UTSDC | OPAL | Positium | FB MTK API	Google API | Twitter decahose stream | FB Marketing API, LinkedInDevAPI | GDELT | Factiva Content API |
DISCOVERY – Training Data Sets; Rich Context Data Search; Knowledge Products | Primitives	Earth on AWS | GeoNet | Nasa Earth Science | Maxar Open Data | Planet Explorer | GEOSTAT | Global Change Master Directory | Radiant ML Hub| Carto Observatory |WSF | ELA| Sentinel-Hub | PopGrid | STAC | COG | ASL SpaceNet | UCI	Cubiq | Mapbox Telemetry | SmartGraph (M-SDK) | Orbital Insights Go | OpenMobile | Mirage | Uber-D | OMF | OpenCellID |	Coleridge Rich Context Scholarly articles API | Kaggle-Text | Awesome-NLP
INTEGRATION & ANALYTICS Analytic tools; privacy preserving methods; integration frameworks	Grid3 | GEE | eo-learn | Solaris | Raster Vision | Pangeo | RoboSat | Sat StatePlay | G-DIF | GeoPandas | Orbital Insights Go | PDAL | Sentinal Toolbox | Orfeo | mosaicJSON | GeoNode | SNAP | RasterFoundry | PPovRepo | Spfeas	Bandicoot | Flowkit | UT-CDR Analysis Kit | Mobile Privacy Model | Movesense | Grid3 | Scikit-Mobility	BERT | GLUE | GPT-2 | Microsoft Turing | Pytorch | StanfordNLP Stanza | NLTK | | OpenNLP

@TREVMON28

@AVSOLATORIO

"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**

Credits: @ingmarweber | #TillKoebe | #RanGoldblatt | @HamedAlemo | @manuletouze | @jed
List of tools included in Toolkit maintained at https://github.com/trevmon28/Data-Alchemy-Guide