Lecture 4b
Big Data in Development Economics

Pierre Biscaye
Université Clermont Auvergne

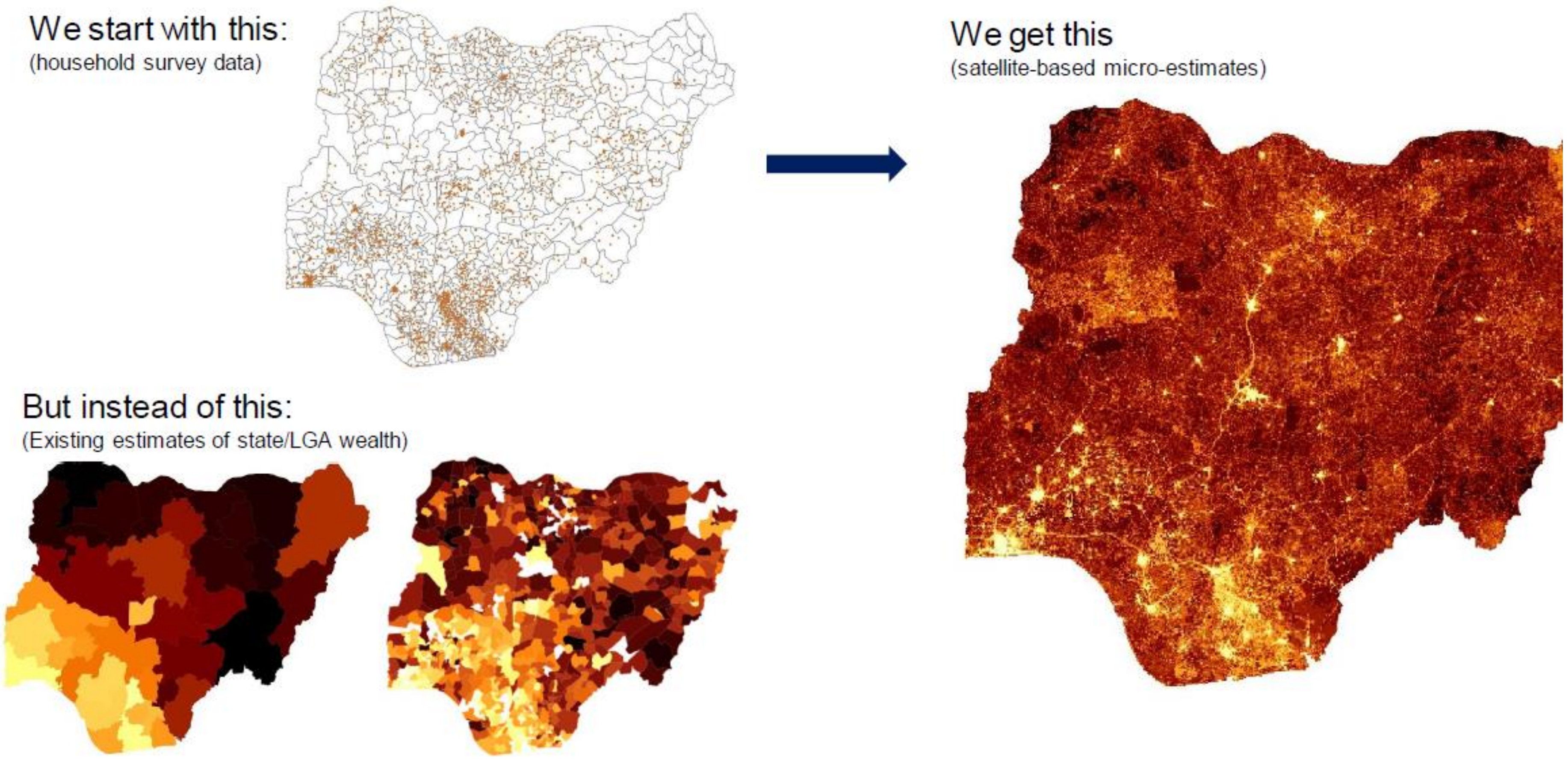
Data Science for Economics

Note: Materials for this lecture are drawn from Josh Blumenstock's Big Data & Development course at UC Berkeley.

Agenda

1. Big picture: Big data and development
2. Motivating example: Measuring poverty
3. Major “big data” sources in development economics
    1. Remote sensing
    2. Mobile phones
    3. Others

1. Big picture: Big data and development



Smythe & Blumenstock 2021

Big Data and the SDGs

1 NO POVERTY
Spending patterns on mobile phone services can provide proxy indicators of income levels

2 ZERO HUNGER
Crowdsourcing or tracking of food prices listed online can help monitor food security in near real-time

3 GOOD HEALTH AND WELL-BEING
Mapping the movement of mobile phone users can help predict the spread of infectious diseases

4 QUALITY EDUCATION
Citizen reporting can reveal reasons for student drop-out rates

5 GENDER EQUALITY
Analysis of financial transactions can reveal the spending patterns and different impacts of economic shocks on men and women

6 CLEAN WATER AND SANITATION
Sensors connected to water pumps can track access to clean water

7 AFFORDABLE AND CLEAN ENERGY
Smart metering allows utility companies to increase or restrict the flow of electricity, gas or water to reduce waste and ensure adequate supply at peak periods

8 DECENT WORK AND ECONOMIC GROWTH
Patterns in global postal traffic can provide indicators such as economic growth, remittances, trade and GDP

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
Data from GPS devices can be used for traffic control and to improve public transport

10 REDUCED INEQUALITY
Speech-to-text analytics on local radio content can reveal discrimination concerns and support policy response

11 SUSTAINABLE CITIES AND COMMUNITIES
Satellite remote sensing can track encroachment on public land or spaces such as parks and forests

12 RESPONSIBLE CONSUMPTION AND PRODUCTION
Online search patterns or e-commerce transactions can reveal the pace of transition to energy efficient products

13 CLIMATE ACTION
Combining satellite imagery, crowd-sourced witness accounts and open data can help track deforestation

14 LIFE BELOW WATER
Maritime vessel tracking data can reveal illegal, unregulated and unreported fishing activities

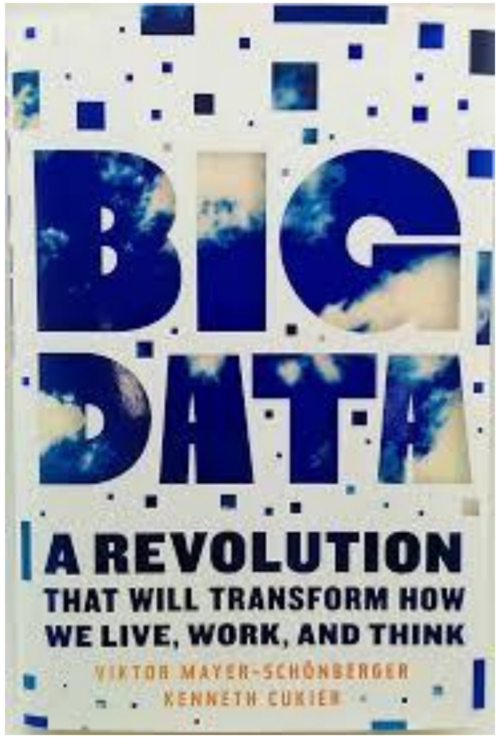
15 LIFE ON LAND
Social media monitoring can support disaster management with real-time information on victim location, effects and strength of forest fires or haze

16 PEACE, JUSTICE AND STRONG INSTITUTIONS
Sentiment analysis of social media can reveal public opinion on effective governance, public service delivery or human rights

17 PARTNERSHIPS FOR THE GOALS
Partnerships to enable the combining of statistics, mobile and internet data can provide a better and real-time understanding of today's hyper-connected world

Source

Context: “Big Data Revolution”



- Mobile phones: 96% penetration globally
- Facebook: 3.07 billion monthly active users
- X/Twitter: 611 million monthly active users
- Whatsapp: over 2 billion monthly active users
  - Stats as of January 2025
- Sensors: millions of satellites, traffic cameras, infrastructure monitors, etc.

Big data in developing countries

- Access to fewer sources of big data
- Prominent exceptions:
  - Mobile phones: Over 6.5 billion subscriptions in LMICs as of 2022
  - Satellites: 1000s in Earth orbit

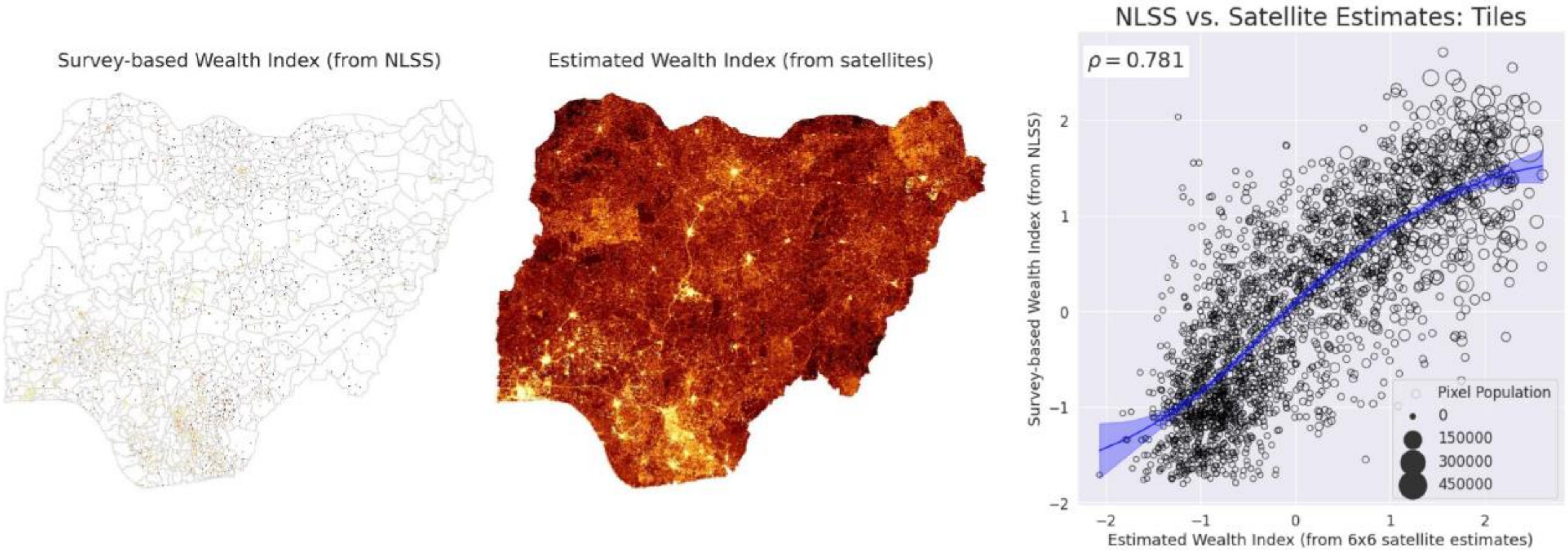
Combining satellite imagery and machine learning to predict poverty

Neal Jean,[1,2*] Marshall Burke,[3,4,5*†] Michael Xie,[1] W. Matthew Davis,[4] David B. Lobell,[3,4] Stefano Ermon[1]

Predicting poverty and wealth from mobile phone metadata

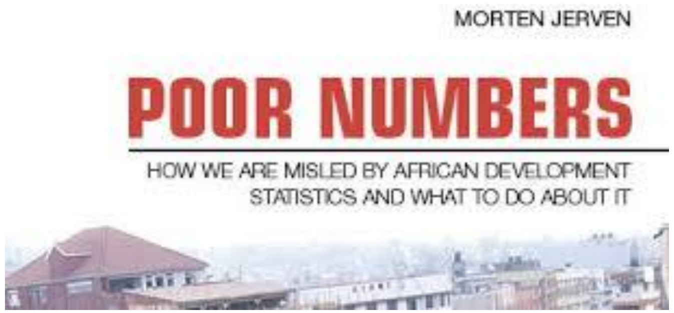
Joshua Blumenstock,[1*] Gabriel Cadamuro,[2] Robert On[3]

2. Motivating example: Measuring poverty at high resolution



Smythe & Blumenstock 2021

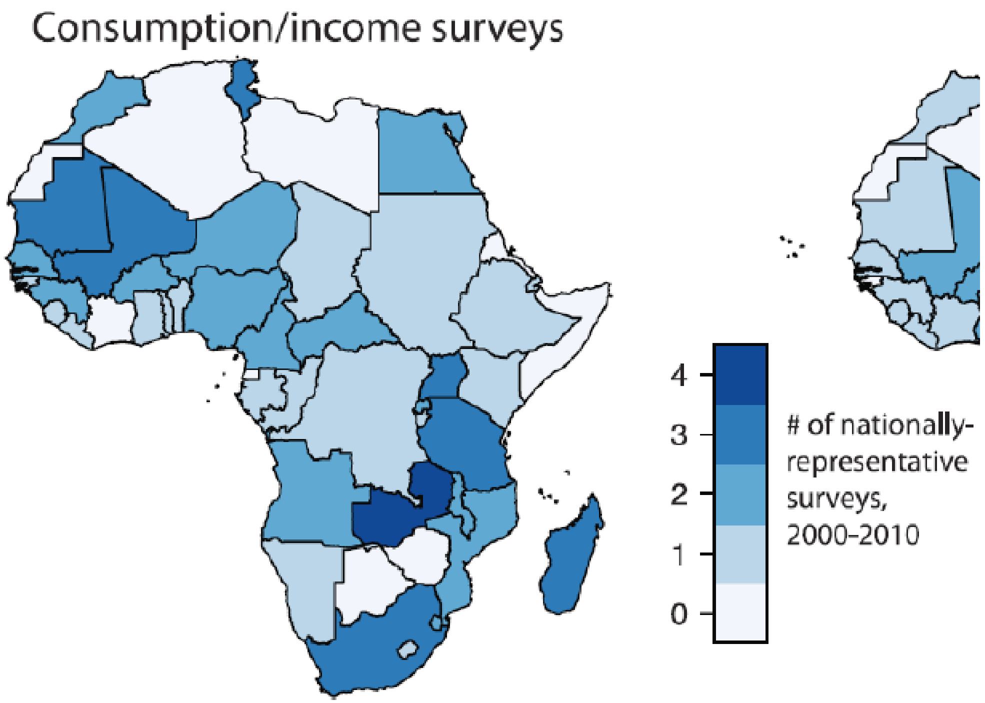
Measuring poverty



- Limitations to official estimates in many developing countries
  - Based on outdated data and unclear assumptions
  - Limited disaggregation
  - Limited temporal frequency
- Examples
  - Ghanaian GDP went from US$6.9B on 4/11/2010 to US$11.8B on 5/11/2010
  - Nigerian GDP went from US$270B to USB$510B in 2014 after rebasing
- How do we find the poor? Data gaps are a problem
  - Need information to target resources, develop policies, track accountability

Traditional approaches to measuring poverty

- Income, consumption, expenditure data
  - Other measures: subjective well-being, capabilities, cortisol, …
- Expensive and time-consuming
  - Single LSMS survey takes 1-3 days to complete
  - $10-50M for a standard Demographic and Health Survey (N=15,000)
- Infrequent data collection



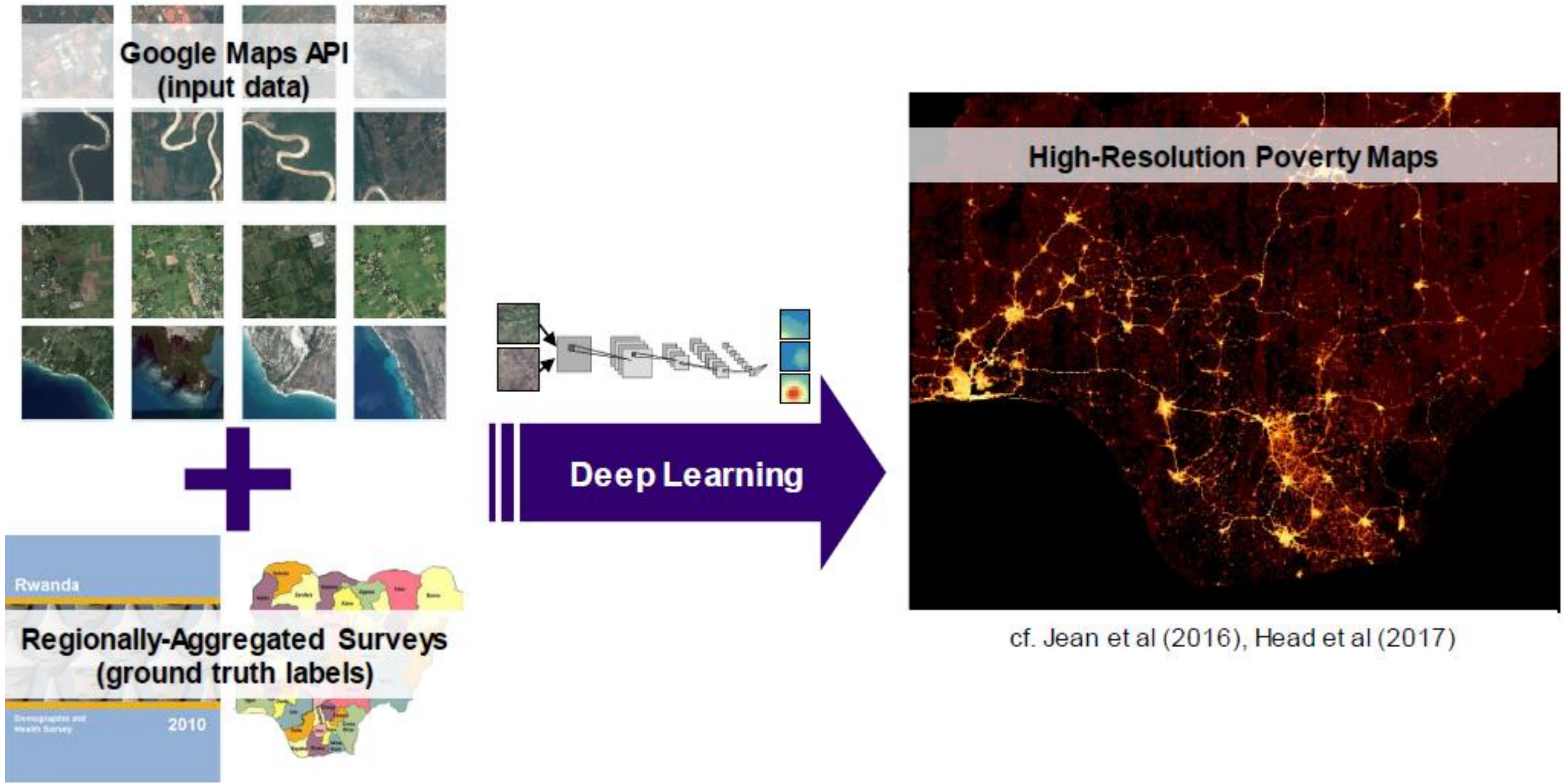
Source: Jean et al 2016

Recent advances and state of the art

- Satellite night lights (Henderson et al 2012, Chen & Nordhaus 2011)
- Google, Social Media (Choi & Varian 2012, Llorente et al 2015)
- Remote sensing/satellite imagery (Jean et al 2016, Pokhriyal et al 2017, Head et al 2017, Hersch et al 2018)
- Mobile phone trace data (Blumenstock et al 2015, Jahani et al 2017, Douglass et al 2015)

Remote sensing and poverty

1. Train neural network to estimate “village” wealth from daytime satellite imagery of “village”



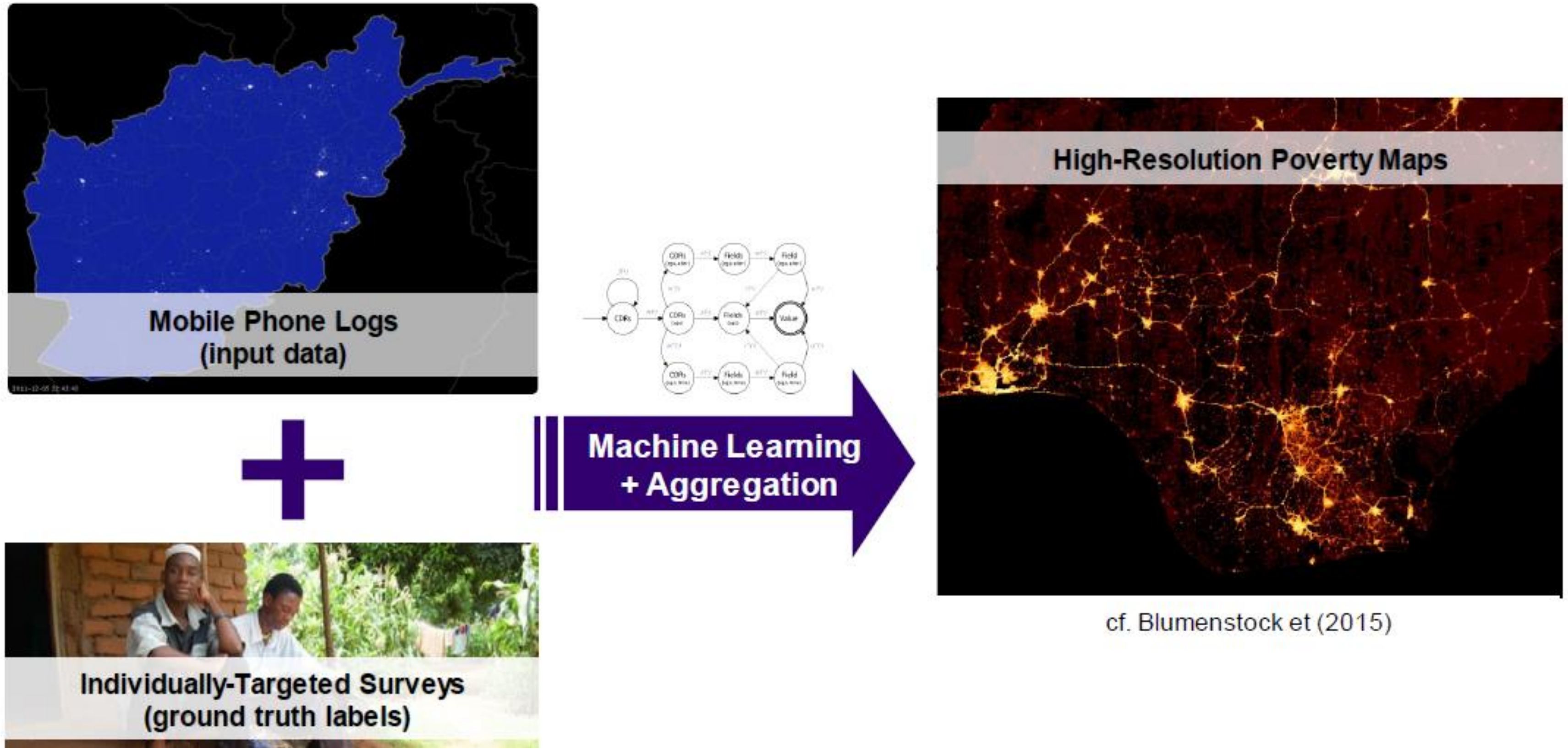
cf. Jean et al (2016), Head et al (2017)

Tradeoffs with satellites

- Advantages:
  - Global coverage
  - Publicly available sources
  - Rapidly improving technology
  - Fewer privacy concerns
- Limitations:
  - Not everything is visible in overhead imagery
  - Uncertain ability to detect dynamic changes
  - Cannot identify individuals

Mobile phones and poverty

2. Match individual mobile phone data to individual surveys, predict poverty



cf. Blumenstock et (2015)

Tradeoffs with mobile phones

- Advantages
  - Individual-level data
  - Potentially track dynamic changes
- Limitations
  - Privacy concerns
  - Limited availability; need access agreements with providers

Big data and poverty measurement

- Huge advances in measuring cross-sectional static wealth
- Less progress in measuring other development indicators
    - But rapidly changing!
- Constraints to dynamic prediction, estimating changes in welfare over time
    - Limited variation in outcomes within subjects over time
    - Data sparsity
    - Stationary ML models inappropriate
    - Etc.

3. Major “big data” sources in development economics research

1. Remote sensing/Satellite imagery
2. Mobile phone data
3. Internet and social media sites
4. Others
    1. Connected sensors
    2. Financial transactions
    3. Utility data
    4. Etc.

Remote sensing (see Donaldson & Storeygard 2016)

- Many types of satellite sensors: imagery (multispectral), radar, LiDAR
- Primary advantages:
  - Access to information difficult to obtain by other means
  - Unusually high spatial resolution
  - Wide geographic coverage
  - Increasingly greater temporal frequency
- Primary disadvantages
  - Dataset size
  - Spatial dependence
  - Measurement error
  - Privacy concerns

Source

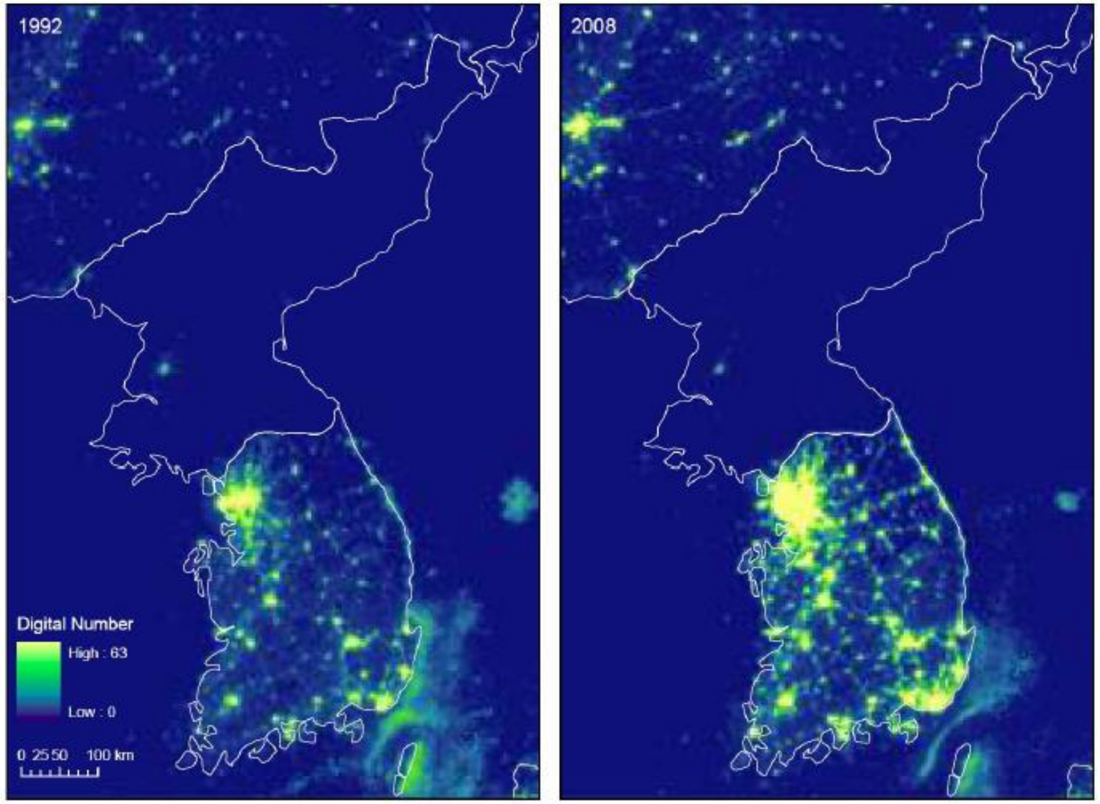
Main satellite sources used in economics

- **Landsat**
  - Resolution: Up to 30 meters; Free; Years: 1972–present
- **Sentinel-1 & Sentinel-2 (ESA)**
  - Resolution: 10–60 meters (Sentinel-2), 5–20 meters (Sentinel-1); Free; Years: Sentinel-1 (2014–present), Sentinel-2 (2015–present)
- **MODIS (NASA)**
  - Resolution: 250 meters–1 kilometer; Free; Years: 1999-present (Terra), 2002-present (Aqua)
- **WorldView (Maxar)**
  - Resolution: 31 centimeters; Not free; Years: 2007–present
- **VIIRS (NASA/NOAA)**
  - Resolution: 375 meters; Free; Years: 2011–present
- **Digital Globe**
  - Resolution: <1m; Not free; Years: 1999-present

Nightlights (Henderson et al 2012, Chen & Nordhaus 2011)

Correlated with GDP across and within countries

Issues: More limited spatial resolution, issues of oversaturation and leakage, more correlated with population density than welfare within country



Daytime imagery

High frequency, high resolution, more details than nightlights

How to use?

1. Use raw information
2. Hand-code features
3. Something else, often some form of machine learning

Image of Mumbai

Working with satellite imagery: poverty mapping in Jean et al 2016

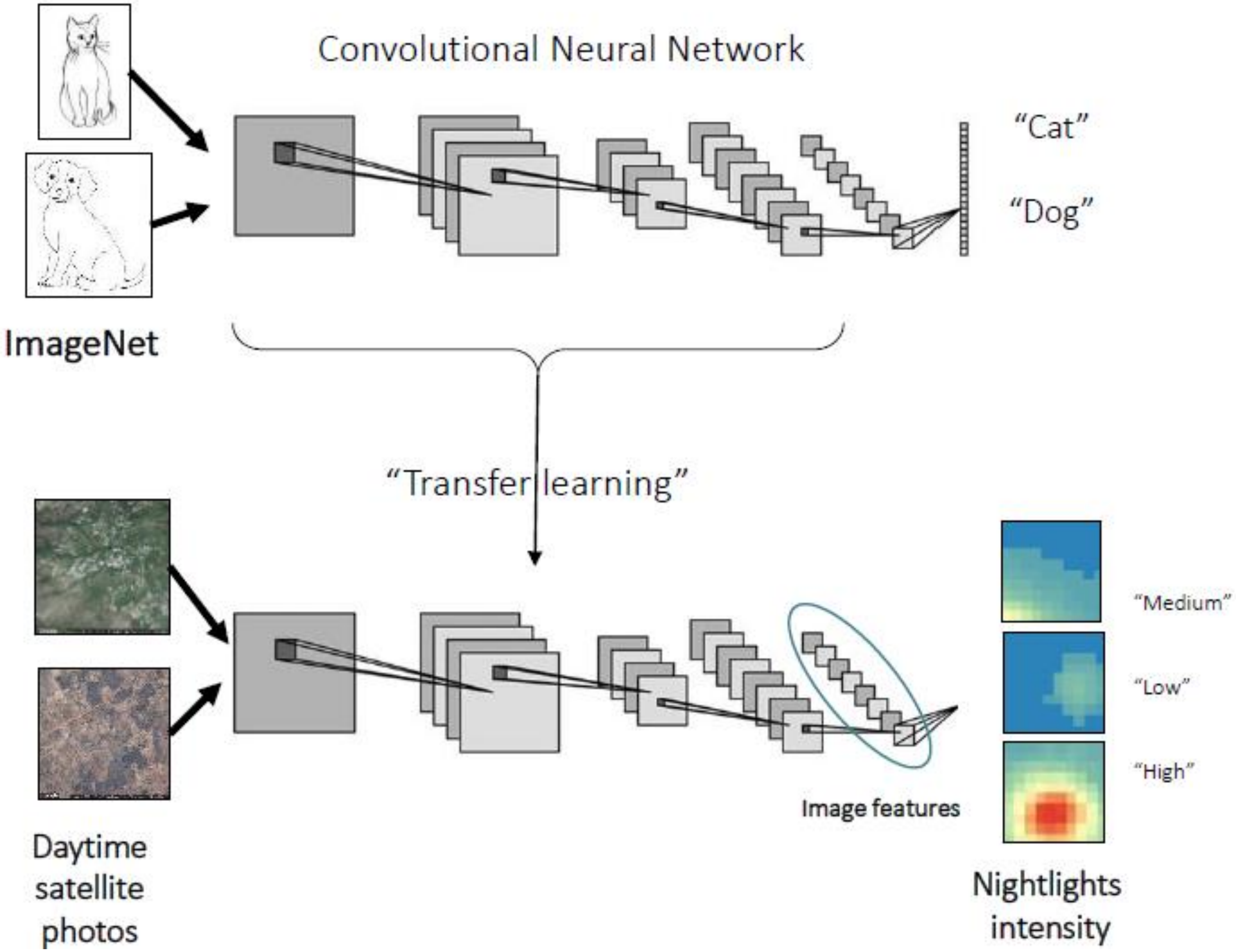
Data:

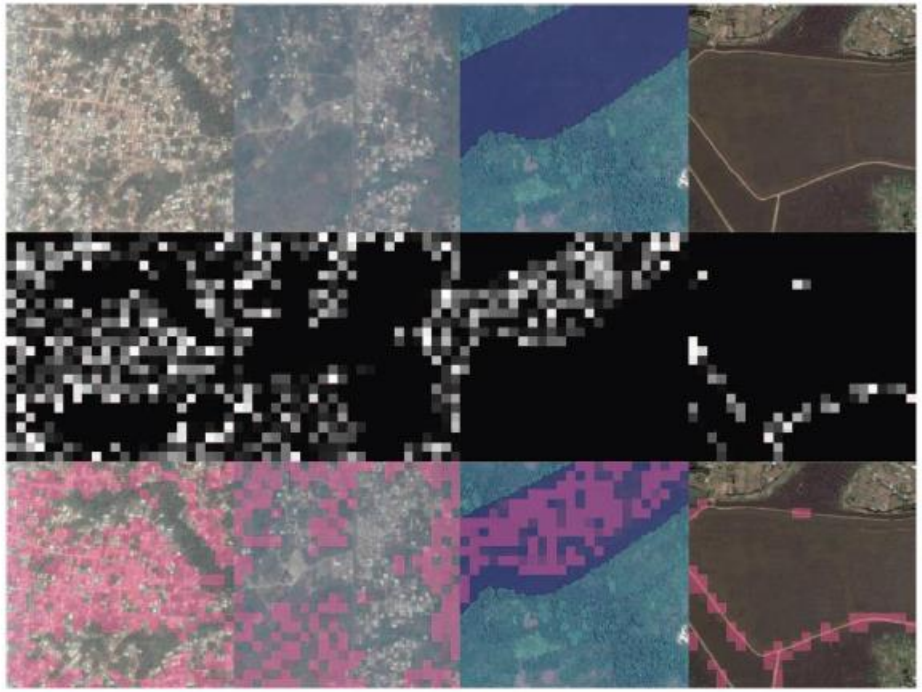
1. Satellite imagery from Google Maps API
2. Nightlights from DMSP-OLS v4
3. Poverty: consumption from LSMS and assets from DHS

Approach:

1. Feature extraction: transfer learning
    1. Better performance than raw features or PCA
2. Spatial join at “cluster” level of satellite and survey data
3. Modeling: ridge regression
4. Prediction: repeated cross-validation

Feature extraction and deep/transfer learning (Jean et al 2016)

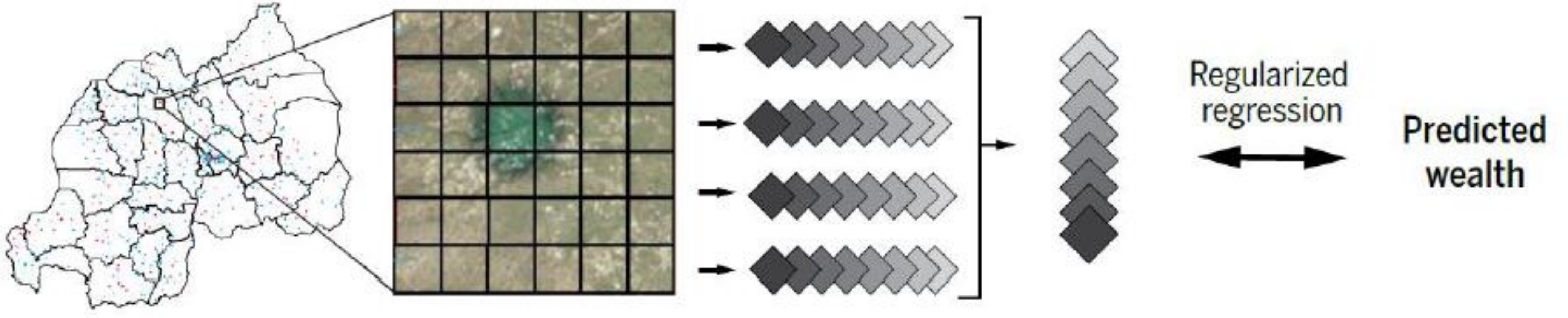




Extracted features

Modeling relationship (Jean et al 2016)

1. Match satellite features to survey data
   1. Use images from area around survey location
   2. Take average of images across a desired time period
2. Model relationship with ridge regression



More recent satellite imagery applications

1. Changes over time (Yeh et al 2020, Huang 2021)
2. Education, health, drinking water (Head et al 2017)
3. Slums/informal settlements (Gadiraju et al 2018, Helber et al 2018)
4. Road quality (Cadamuro et al 2018)
5. Buildings (Nieves et al 2018, Liu et al 2019)
6. Population (Tiecke et al 2017)
7. Crop yield (Lobell 2013, Lobell et al 2015)

Mobile phone data

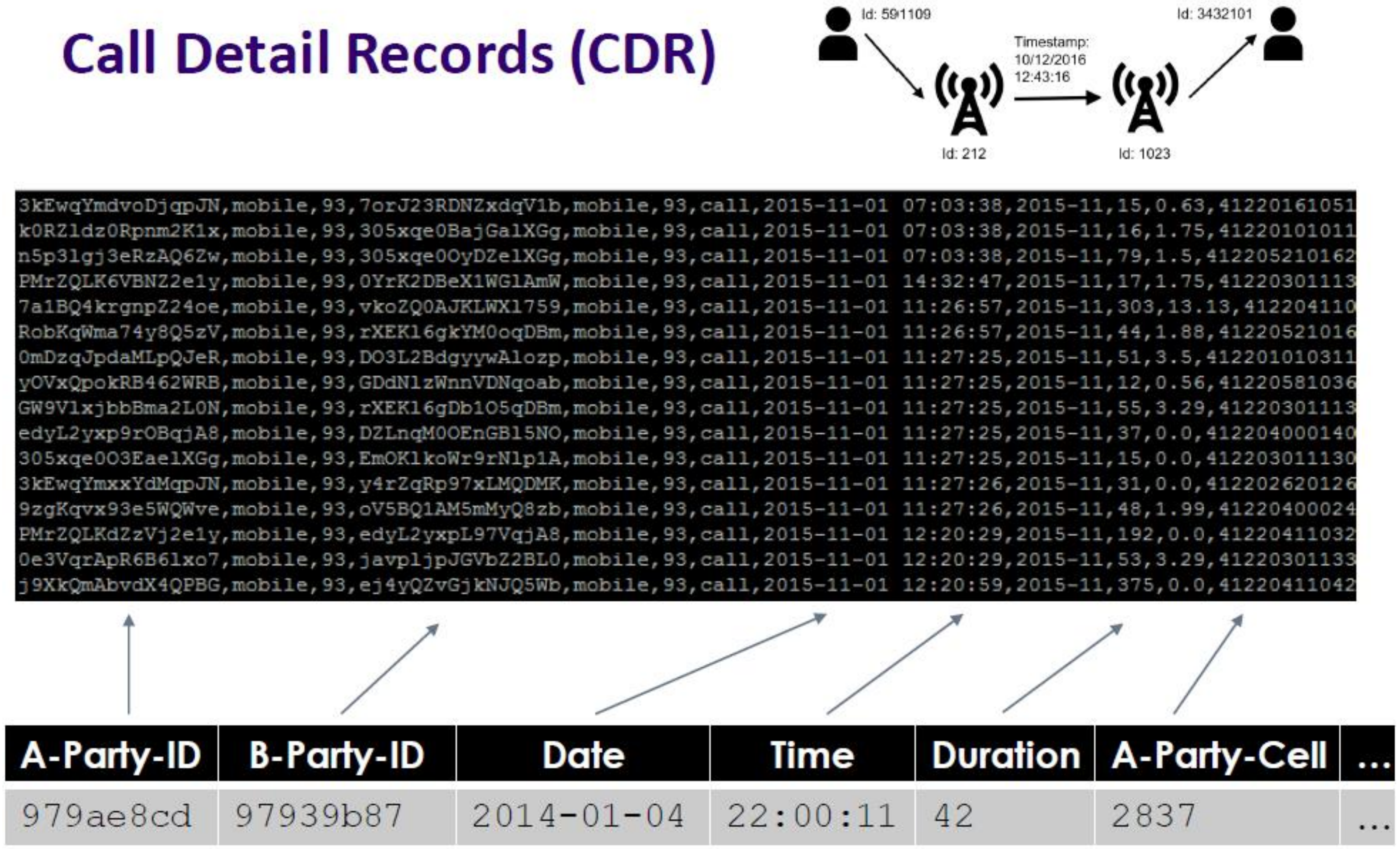
Source

1. Most of world population lives within research of mobile network, rapid increases in adoption globally over time
    1. Phone data: timing, location, other metadata on voice call, text messages, internet/data use, proximity (Bluetooth)
2. Mobile money spreading rapidly in many developing countries
    1. Metadata: timing, location, amount of account use, recharges, etc.
3. Typically limited sociodemographic information collected on users

Call detail records

Can use these to infer:

1. Communication frequency and timing
2. Location
3. Mobility and migration
4. Social network
5. And more



Phone use and religiosity (Dube et al 2022)

Figure 2: Intensity of mobile phone calls over time

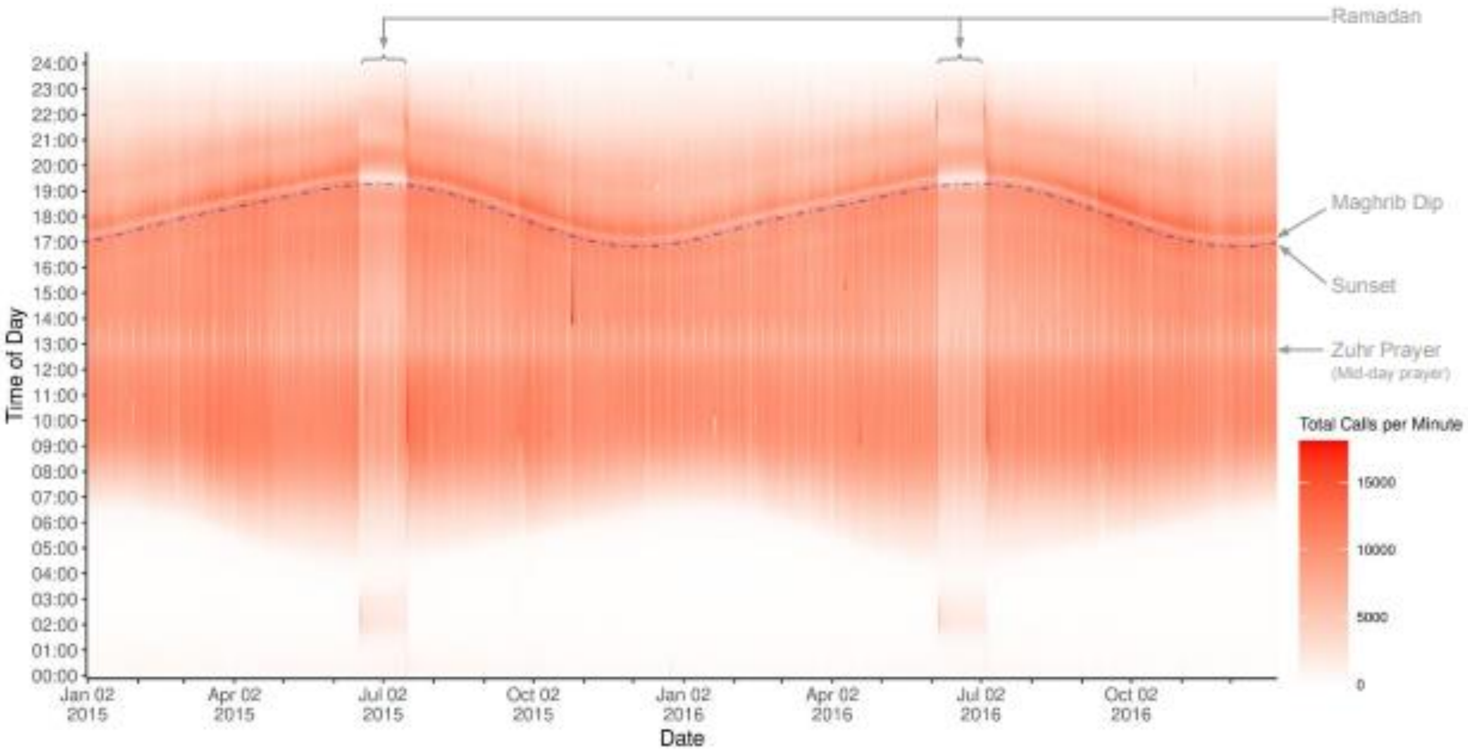


Figure 3: Religious Adherence by District

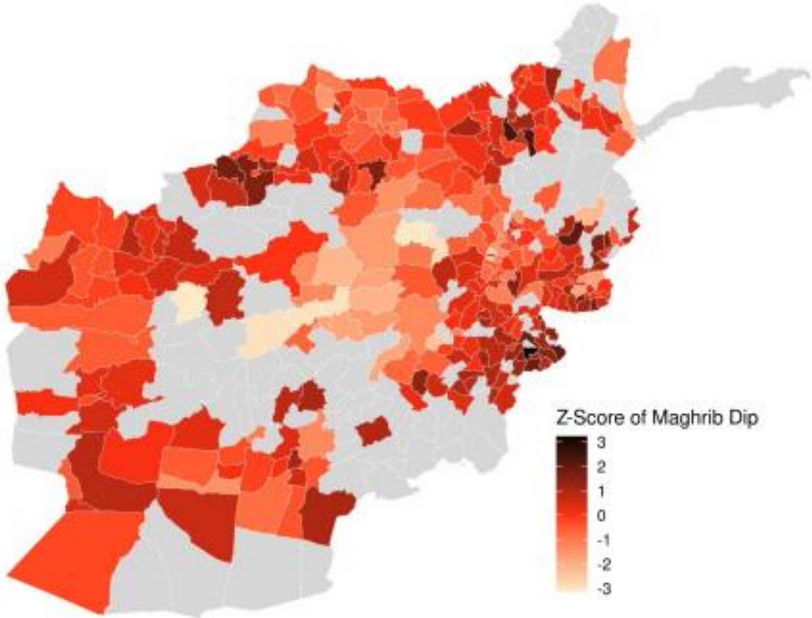


Table 3: Climate and Religious Adherence by Land Type

	Maghrib Dip		
	(1)	(2)	(3)
SPEI (6 months)	0.741		
	(0.456)		
SPEI (6 months) x Cropland	-1.251**		
	(0.580)		
SPEI (6 months) x Rangeland	-0.869*		
	(0.500)		

Data access: not just MNOs

Considerations with phone data

1. Representativeness, population heterogeneity
2. Data access and privacy
3. Connection with other data
   1. Individual: conduct surveys
   2. Spatial: match to ‘home’ locations
4. Extracting features from phone metadata: feature engineering
   1. Communications: Call volume, duration, entropy, incoming vs outgoing, timing, top-ups, contacts
   2. Mobility: distance traveled, areas visited, radius of gyration
   3. Network: measures of centrality, clustering, diversity
5. Predicting outcomes: machine learning, cross-validation

Internet and social media

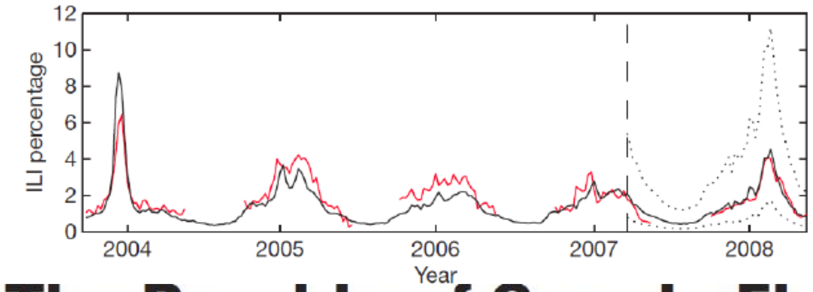
1. Internet: still low use in many developing countries
   1. Increasing but with gender, urban/rural, wealth, language gaps
2. Social media: data on networks, locations, ads, etc.
   1. Kondmann et al 2020 combine tweet counts, remote sensing, and DHS for local poverty mapping in SSA
   2. Patel et al 2017 combine tweet locations and densities and admin data for population density mapping in Indonesia
3. Other internet sources: search, maps, news, IP addresses, Yelp, etc.
4. Concerns: access, privacy, representativeness, measurement and construct validity and reliability

Illustrating measurement issues: Google Flu Trends

- GFT built to predict CDC reports
- Problems of overfitting and ad hoc modeling + endogeneity of Google search algorithm
- Overpredicts most periods but misses others
- Lessons (Lazer et al 2014)
  - "Quantity of data does not mean one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data"
  - Core challenge: most big data "are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis"

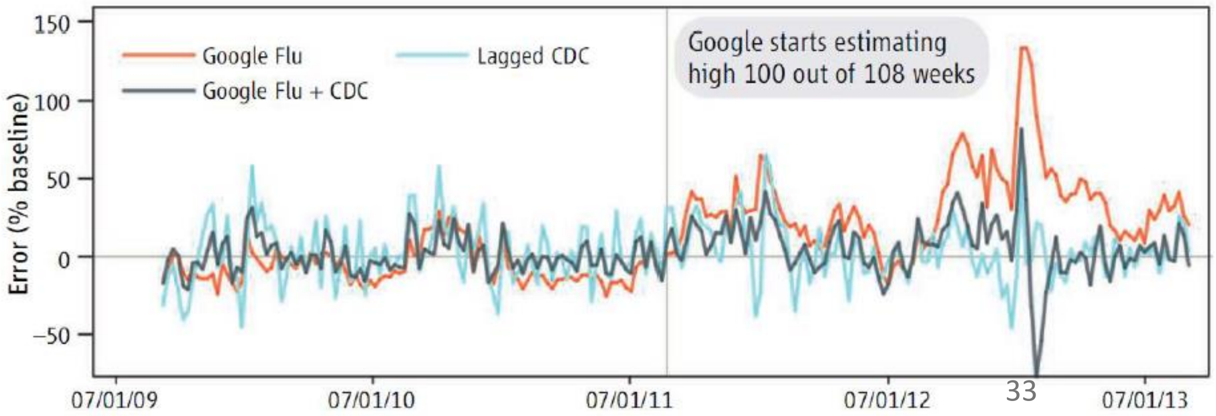
Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]



The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]



Other data: financial transactions, sensors, utility data, etc.

Example: Cisse (2024 JMP)

- Combine electricity grid spatial data, utility data on local-level outages, utility data on customer consumption and billing, and survey data to estimate value of electricity reliability

Figure A3: Map of feeder lines by timing of reliability project in Senegal

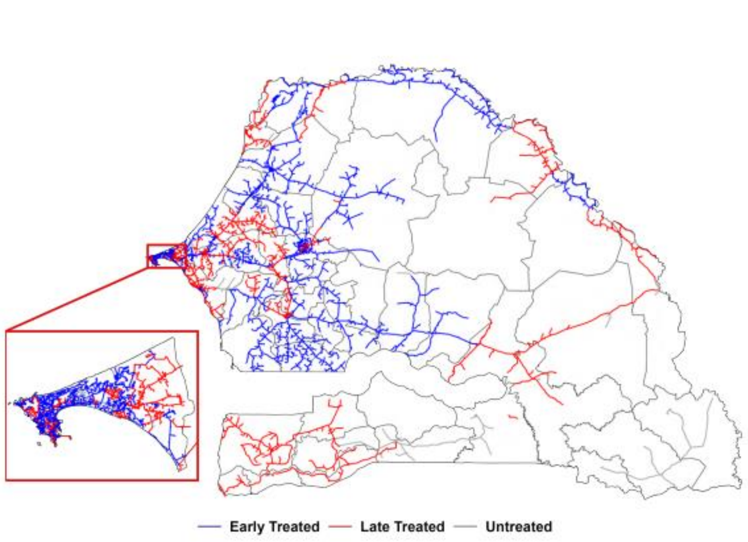


Figure A4: Enumeration Areas in Household Surveys vs. Stations in Electricity Network vs. Customer Locations in Billing Data

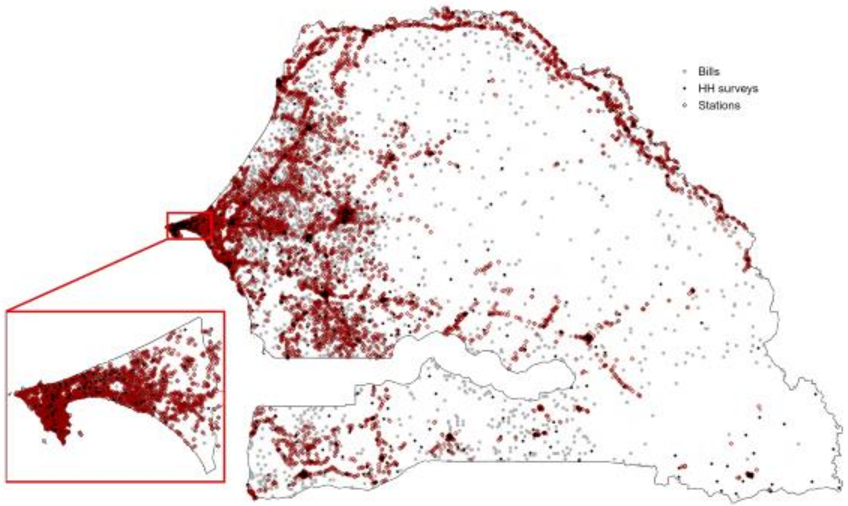
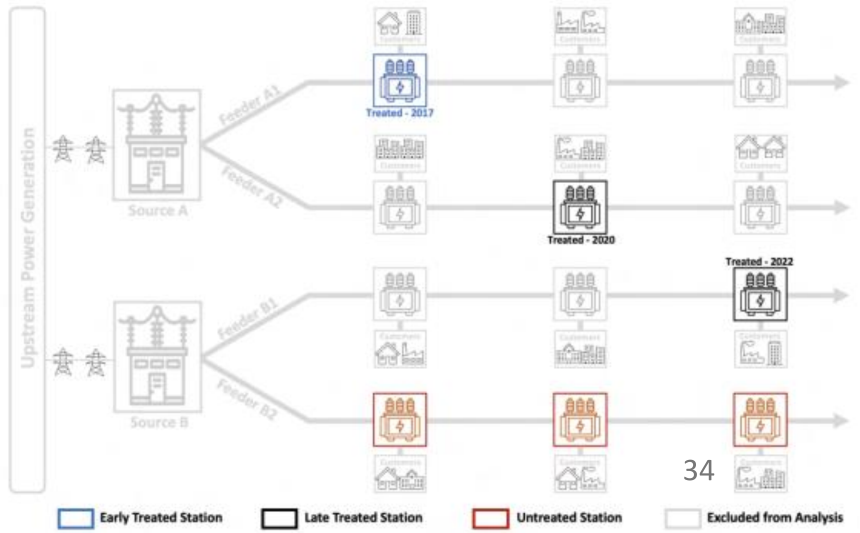


Figure A5: Illustration of Variation Used to Estimate Direct Effects of Reliability Projects



Cisse (2024) electricity outages and consumption in Senegal

Figure C3: OUTAGES BY HOUR OF THE DAY AND OVER TIME FROM 2016 TO 2022

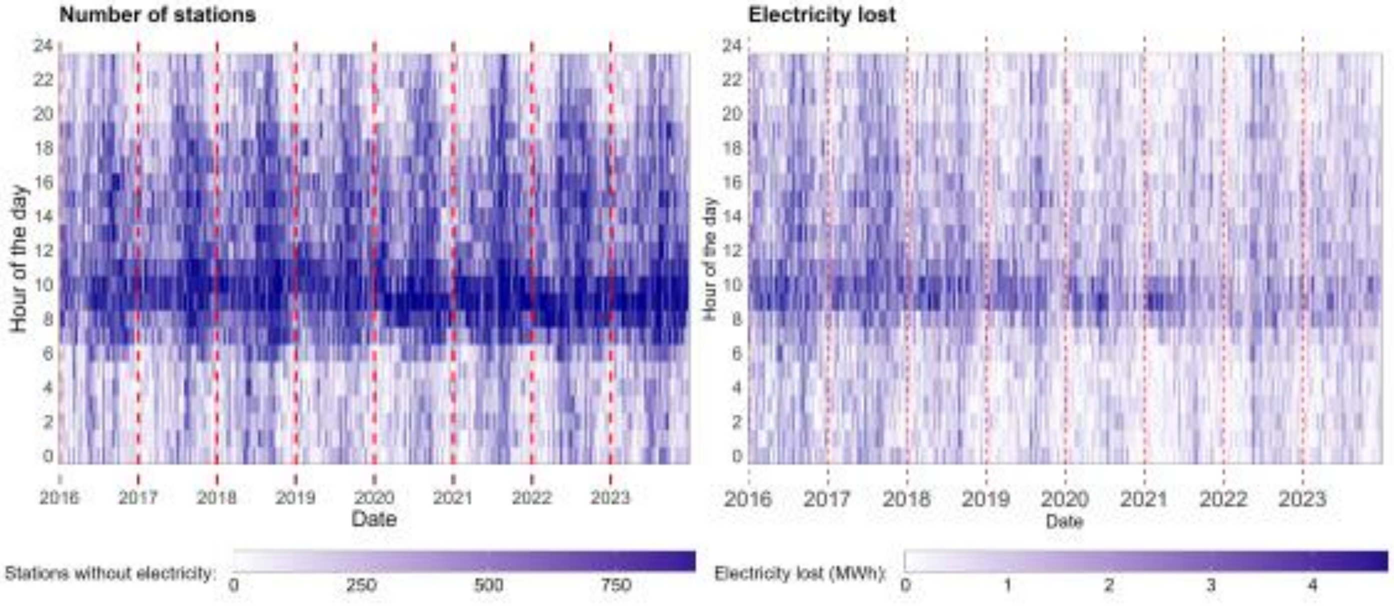
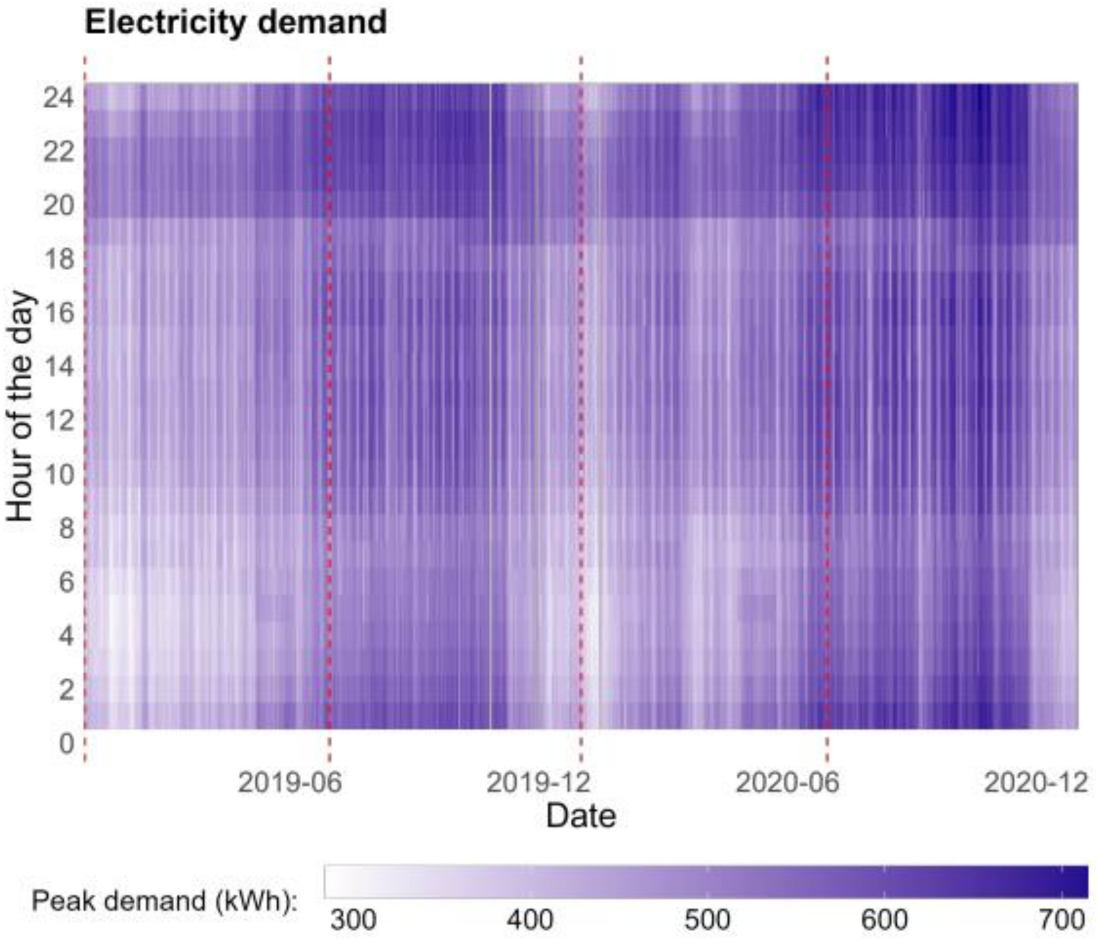


Figure C5: PEAK ELECTRICITY DEMAND THROUGHOUT THE DAY AND OVER TIME



Big data ethics in developing contexts

1. Privacy, consent, and data security
    1. Big data can inadvertently reveal sensitive identifiable information
    2. Anonymize rigorously to ensure re-identification not possible, adhere to data protection laws, use secure storage and encryption, justify ethical data use under IRB guidelines
2. Bias and representation
    1. Analyze and document coverage limitations in the dataset and supplement with additional data sources or methods to address gap
3. Misuse of findings
    1. Be transparent about the potential implications of the research, and ensure findings are contextualized to avoid misuse. Partner with trusted organizations and stakeholders to guide ethical applications.
4. Context and power dynamics
    1. Center research in local context and norms, engage local stakeholders, prioritize studies with tangible local benefits