Research and Training on Big Data

Seminar on Statistical Capacity Building for New Data Sources

Keio Plaza Hotel, Tokyo, Japan
8 December 2017

Kaushal Joshi
Asian Development Bank

Outline

- Conventional vs Big Data sources
- ADB's Innovative Data Collection for Agriculture and Rural Statistics
- ADB's forthcoming Data for Development Initiative
- Concluding Observations

Conventional vs Big Data sources

Conventional Data Sources of Official Statistics

- SURVEYS,
- CENSUSES,
- ADMINISTRATIVE REGISTERS.

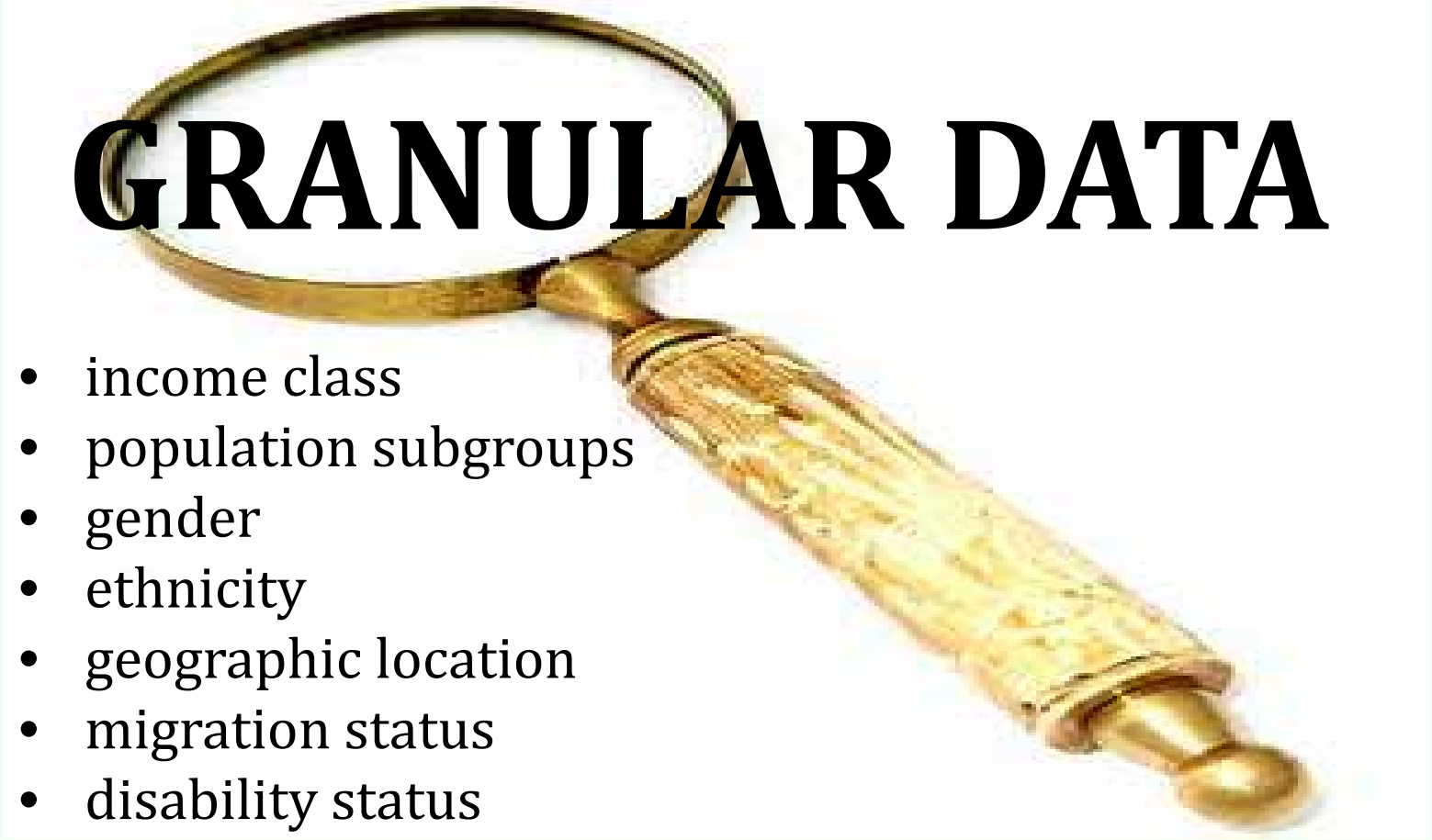
Innovative Data Sources

- SATELLITE IMAGES,
- MOBILE PHONE RECORDS,
- SENSORS AND SCANNER DATA,
- SOCIAL MEDIA DATA, etc.


SDGs call for no one is left behind.

Conventional vs Big Data

Leave no one behind principle requires



- income class
- population subgroups
- gender
- ethnicity
- geographic location
- migration status
- disability status
- etc.

Conventional vs Big Data

Challenges for Data disaggregation from traditional sources

- Limitations of sample surveys
- Increasing costs to collect and analyze
- Potential loss of quality
- Pressure to collect more information
- Response burden
- Politics of/over data
- Transparency, etc

Innovative Data Collection for Agriculture and Rural Development - R-CDTA 8369

- **Source of Funds:** Japan Fund for Poverty Reduction
- **Pilot Countries:** Lao PDR, Philippines, Thailand and Viet Nam
- **Implementation Period:** June 2013 to October 2017
- **Objectives:**
  - Development of customized software applications and methodology to estimate paddy rice cultivation area and crop production using satellite data,
  - Training of counterpart staff in the four pilot countries, and
  - Development of an online training program on the use of satellite data for agricultural and rural statistics.

Innovative Data Collection for Agriculture and Rural Development - R-CDTA 8369

- Developed customized versions of **IN**ternational **A**sian **H**arvest m**O**nitoring system for **R**ice (INAHOR-AD)
- Trained staff from Lao PDR, the Philippines, Thailand, and Viet Nam
  - basic remote sensing, use of INAHOR-AD software, use of QGIS, crop cutting, farmer recall survey, and
  - geospatial technologies (e.g. SNAP) and computer-assisted personal interviewing (e.g. Survey Solutions)

Innovative Data Collection for Agriculture and Rural Development - R-CDTA 8369

- Developed an online training con Estimating Rice Paddy Extent and Production with ALOS-2/PALSAR-2 and INAHOR-AD
- Promotional video for the course https://youtu.be/SSwg000ooHc
- Link for the course: http://adbx.online/

Innovative Data Collection for Agriculture and Rural Development

Methodological Research - 1

Using Area Frame for Paddy Rice Statistics: Methodology and Weighting Procedures, Results of Survey Estimates and Sampling Errors.

- Area frame approach in conjunction with crop cutting technique is used to estimate paddy rice area, yield, and production for the 2015 cropping season (July 2015 – November 2015) in the provinces of Savannakhet, Lao PDR; Ang Thong, Thailand; and Thai Binh, Viet Nam.
- Results obtained are compared with existing administrative data sources. Significant deviation for rice area between the two estimates. Yield estimates are similar for both methods in all countries except in Lao PDR.

Innovative Data Collection for Agriculture and Rural Development

Methodological Research - 2

Land Measurement Bias: Comparisons from Global Positioning System, Self-Reports, and Satellite Data

- This research looks at differences in farmer reported area versus GPS (gold standard) and Google Earth for agricultural plot area.
- Farmer reported plot area estimates are found to be statistically different when compared with the two methods in three out of four countries.
- Google Earth performs just as well as GPS (no statistically significant differences).

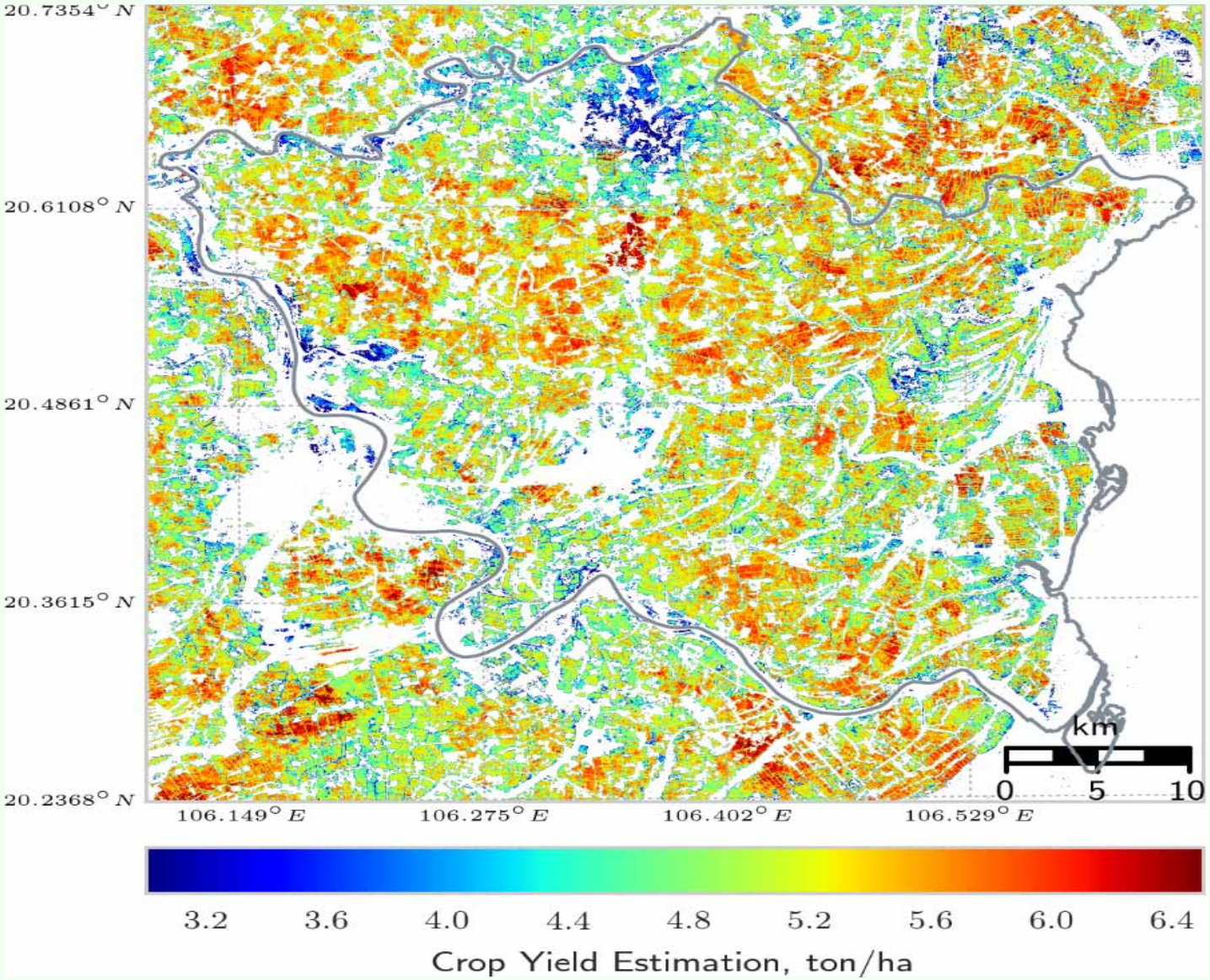
Innovative Data Collection for Agriculture and Rural Development

Methodological Research - 3

Measuring Rice Yield from Space

- Compared area and yield estimates between ALOS-2 satellite and Landsat-MODIS fusion data.
- ALOS-2 satellite and Landsat-MODIS fusion data are equally efficient for paddy rice area estimation, but Landsat-MODIS provides better results for crop yield estimation.
- Ground data (crop cutting) and fusion Landsat-MODIS data used to create a spatially delineated rice yield map for Thai Binh province in Viet Nam to permit spatial analysis.

Innovative Data Collection for Agriculture and Rural Development - R-CDTA 8369




TA 9356-ADB's Data for Development Project aims to **strengthen** the capacity of **NSOs** meet the **disaggregated data** requirements of the **SDGs.**

TA 9356-REG: Data for Development

Target Outputs

- Training workshops on SAE and Big Data analytics for targeted to NSO staff
- Training Manual on Disaggregation of Official Statistics and SDGs
- Online Course Modules on SAE and Big Data Analytics
- Country-Specific Case Studies on SAE and Big Data Analytics

TA 9356-REG: Data for Development

Country-Specific Case Studies

- Explore the potential of using big data as an alternative source.
- Facilitate comparison of estimated indicators based on methods using traditional data sources and big data complemented techniques.
- Help NSOs identify their operational resource requirements in integrating big data analytics in their work programs.

TA 9356-REG: Data for Development

TA focus areas can capitalize on the following innovative data sources

- Satellite images
  - Publicly accessible
  - Has various applications
  - Developed methods for estimates already existing

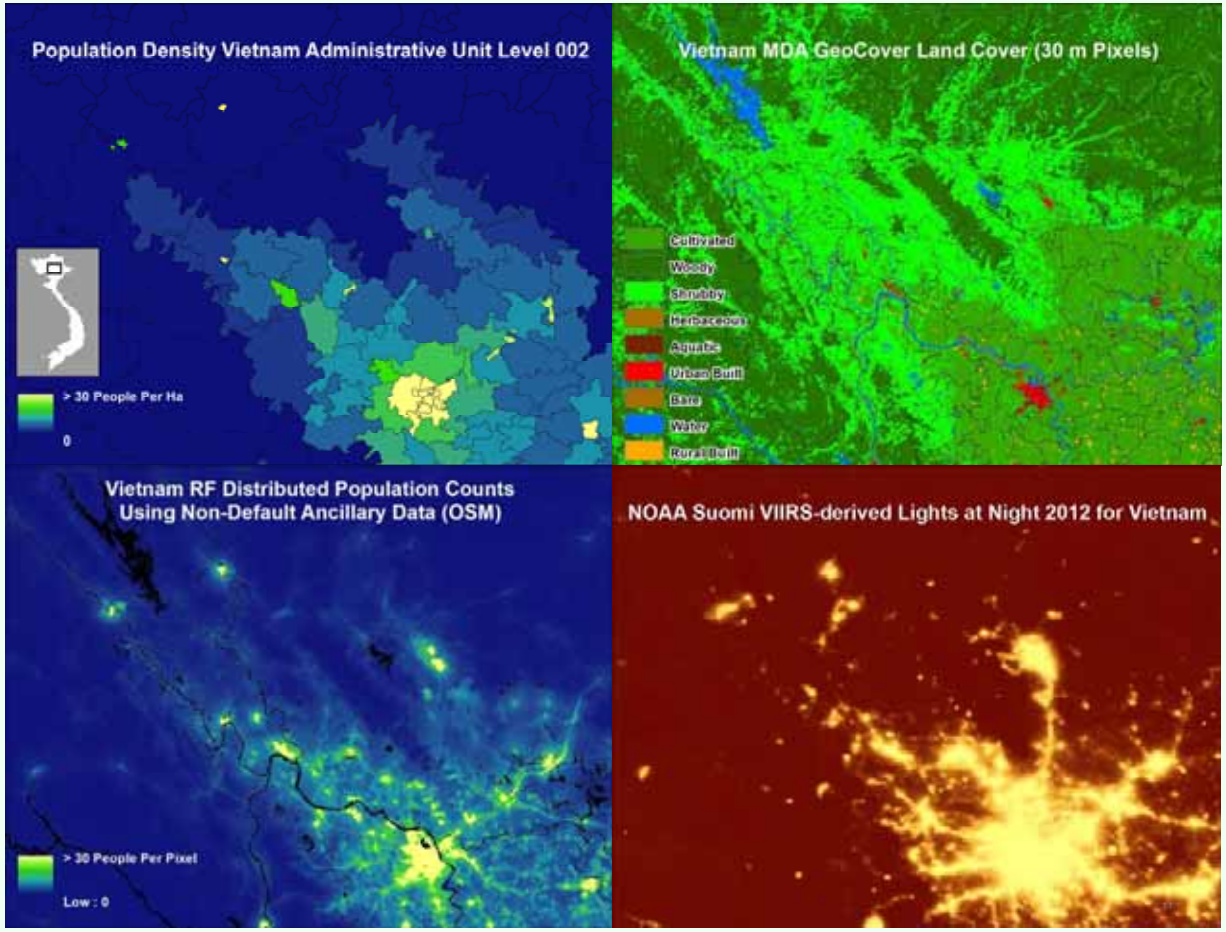
- Mobile phone call detail records (CDRs)

TA 9356-REG: Data for Development

Focus of Country-specific Case Studies

- Population Mapping
- Poverty Mapping

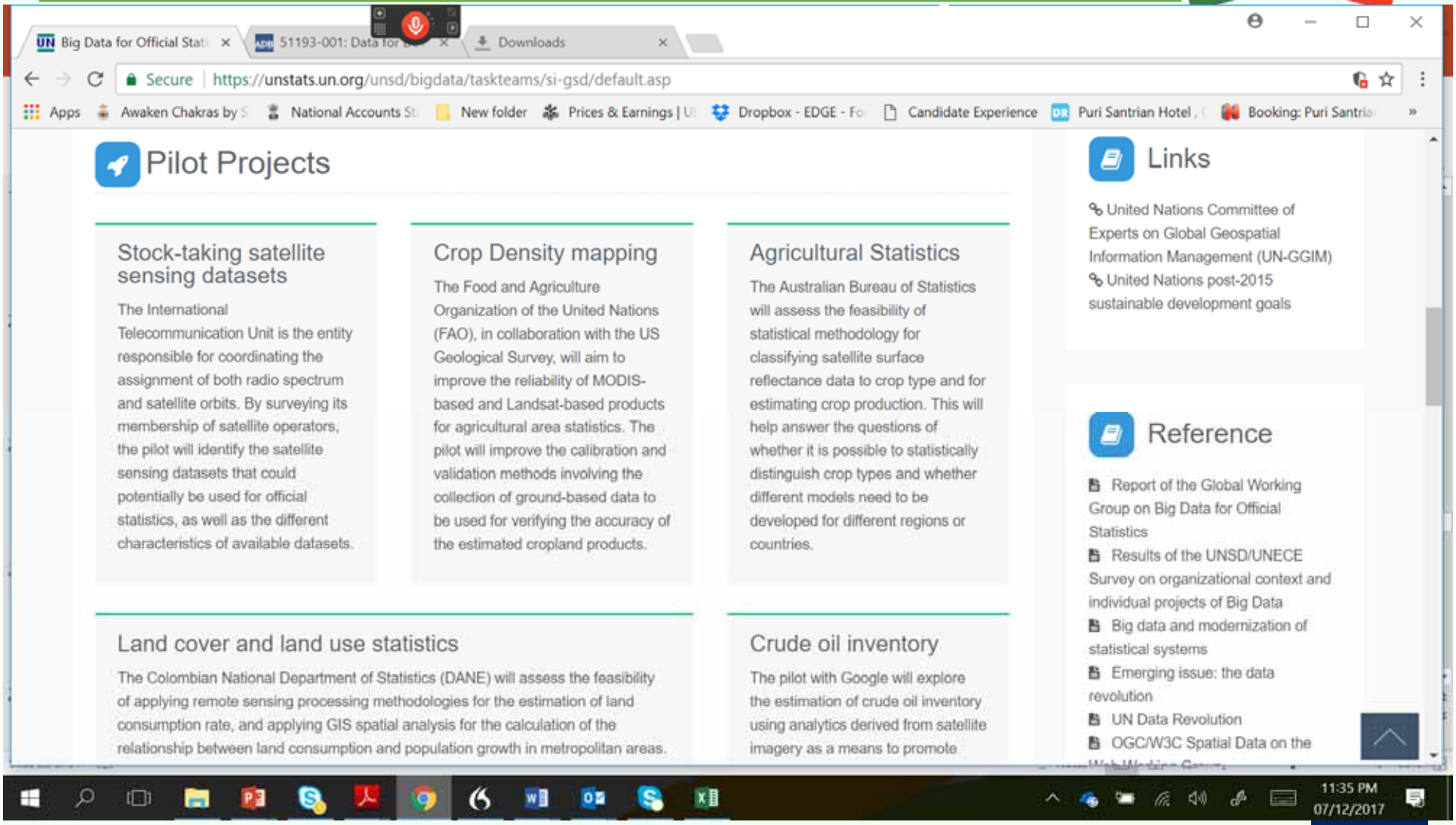
TA 9356-REG: Data for Development



Population mapping example: (Top-left) Population density from census data for each administrative level 2 unit in an area of northern Vietnam, (Top-right) Land cover dataset for the same area, (Bottom-left) Satellite image of the area at night, (Bottom-right) WorldPop population modelling methods take the census data as input, then use machine learning methods to exploit the relationship between population density and high resolution landscape features, such as those from land cover and satellite data, to predict population densities for each 100x100m grid cell on the landscape.

Source: http://www.worldpop.org.uk/about_our_work/case_studies/

Bigdata – UN Global Working Group



Source: https://unstats.un.org/unsd/bigdata/taskteams/si-gsd/default.asp

ADB

Concluding Observations

- Big data with advantages of timeliness and geo-spatial details offers immense potential in generating quick and more granular estimates.
- Methods and results from big data applications however, need to be tested and cross validated with traditional surveys results for robustness.
- Proxy indicators correlated with traditional indicators (like night time lights) provide opportunities to generate more frequent estimates and can complement traditional databased estimates for early estimates and forecasting trends.

Concluding Observations

- Big data pose challenges - privacy issues, costs, sharing of data by holders of big data, capacity to use.
- While big data should be embraced, but traditional sources will remain important.
- More methodological research needed to adopt big data in official statistics.

Thank you!

Email: kjoshi@adb.org

ADB

Innovative Data Collection for Agriculture and Rural Development

SATELLITE	SOURCE	SPATIAL RESOLUTION	TEMPORAL RESOLUTION	COST	SENSOR TYPE
MODIS	NASA	1km/ 500m/ 250m	1-2 days	FREE	Optical
Landsat	USGS/ NASA	30m	16 days	FREE	Optical
ALOS-2	JAXA	100m	14 days	Paid	SAR
Sentinel -2	ESA	10m	5 days	FREE	Optical

ADB

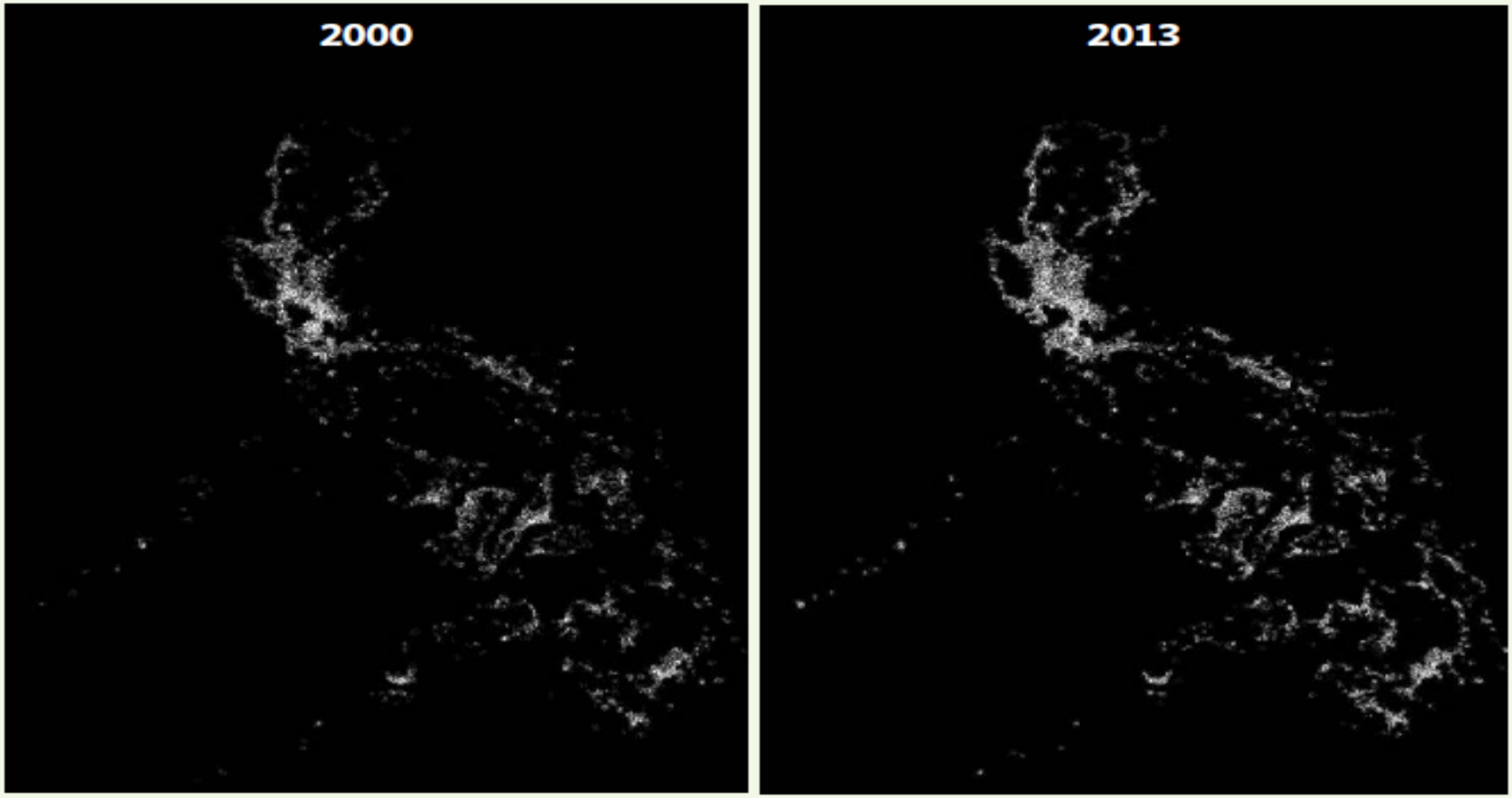
DATA FOR DEVELOPMENT

Appendix: **Big Data Analytics**

Source: Google Images

DATA FOR DEVELOPMENT

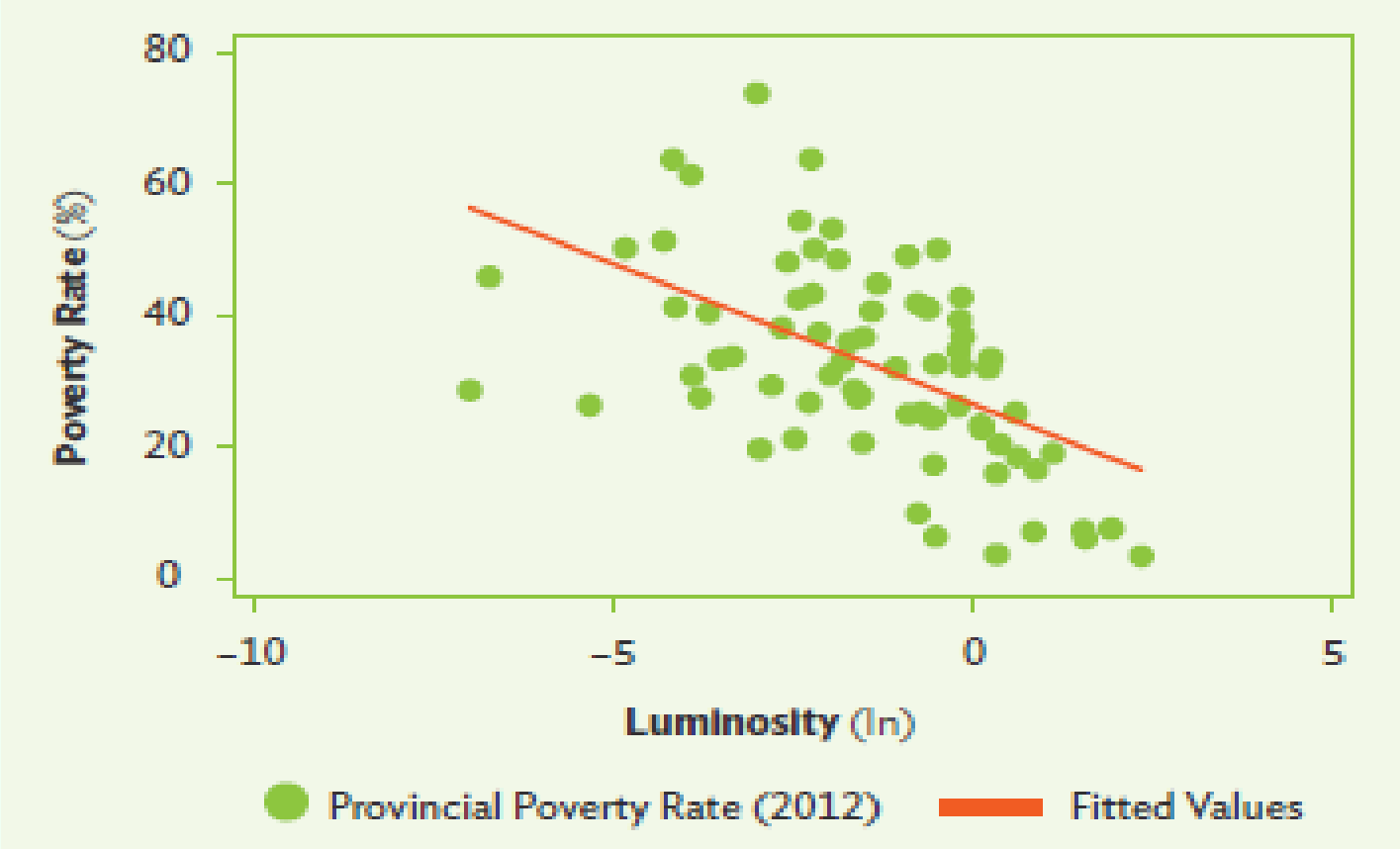
Appendix: **Big Data Analytics**



Source: ADB's Key Indicators for Asia and the Pacific 2016

DATA FOR DEVELOPMENT

Appendix: **Big Data Analytics**

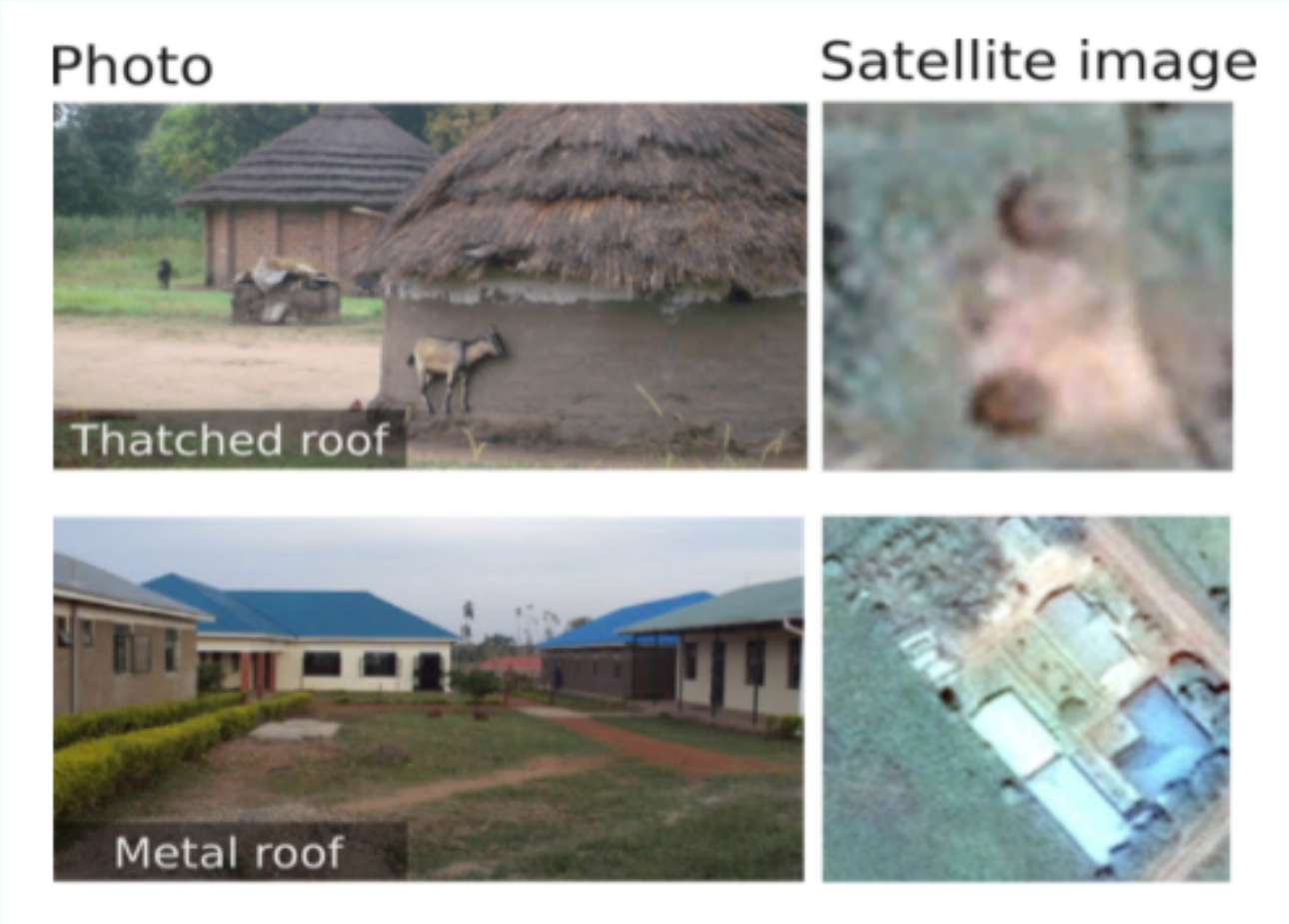


https://blogs.adb.org/blog/how-nighttime-lights-help-us-study-development-indicators

Source: ADB's Key Indicators for Asia and the Pacific 2016

DATA FOR DEVELOPMENT

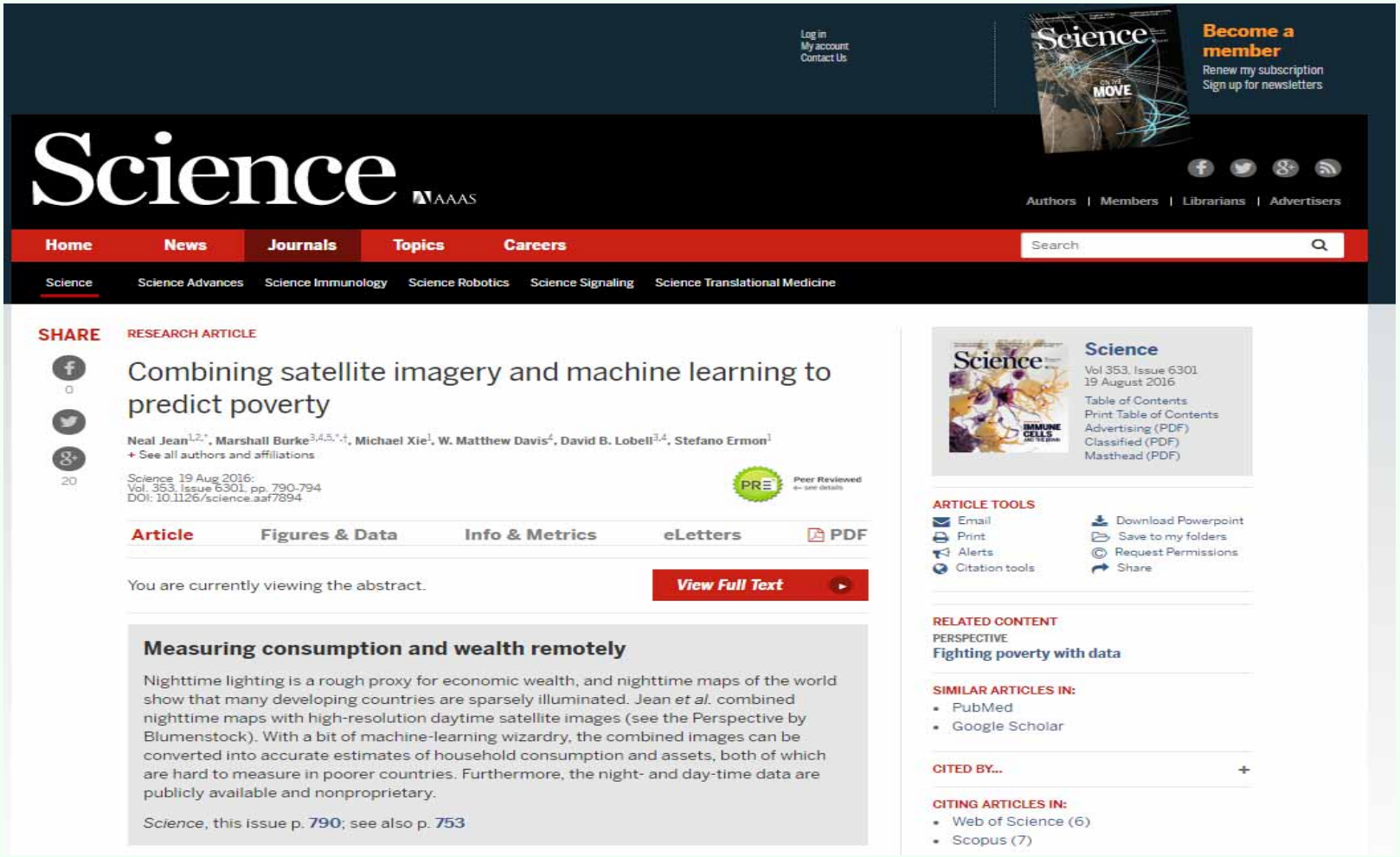
Appendix: **Big Data Analytics**



Source: www.unglobalpulse.org

DATA FOR DEVELOPMENT

Appendix: **Big Data Analytics**



DATA FOR DEVELOPMENT

Appendix: **Big Data Analytics**

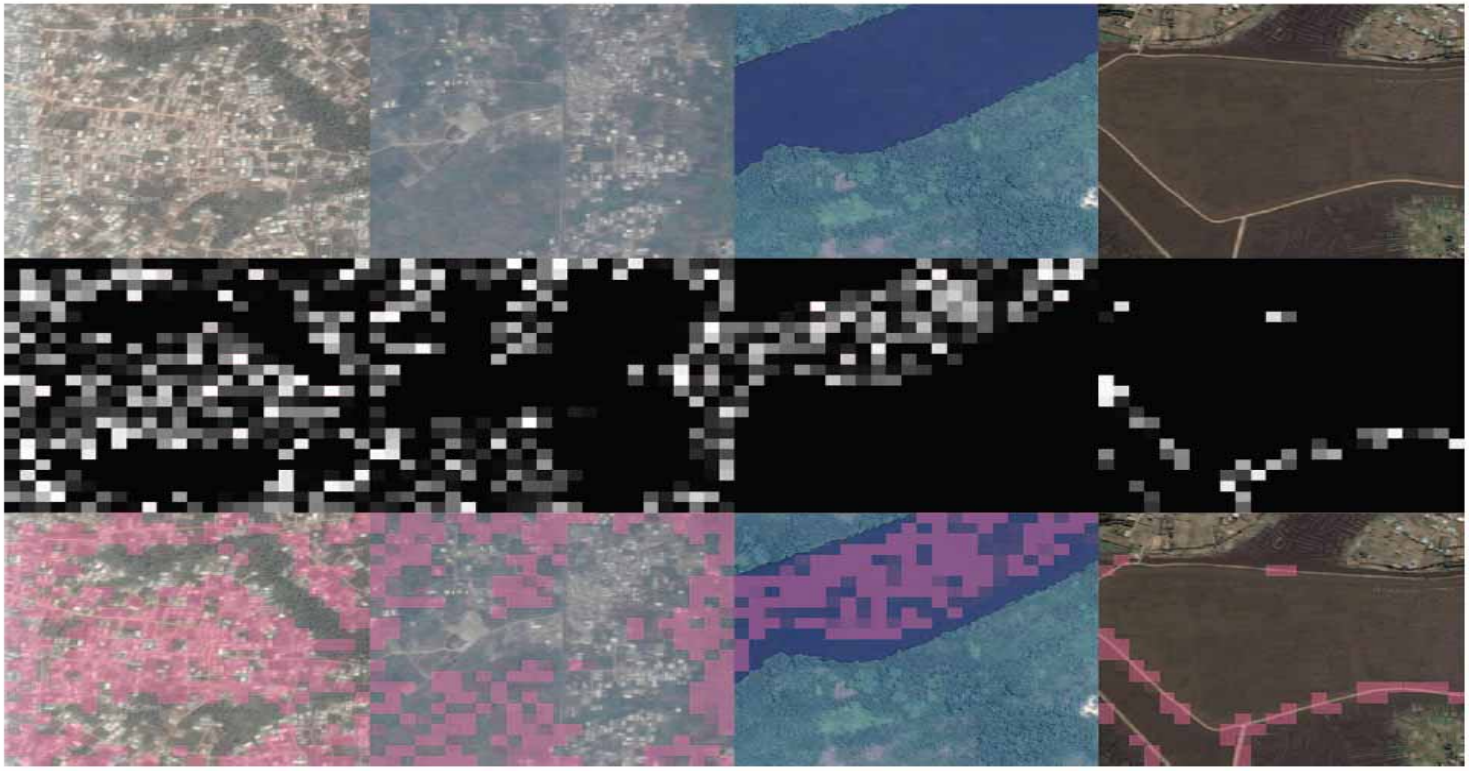
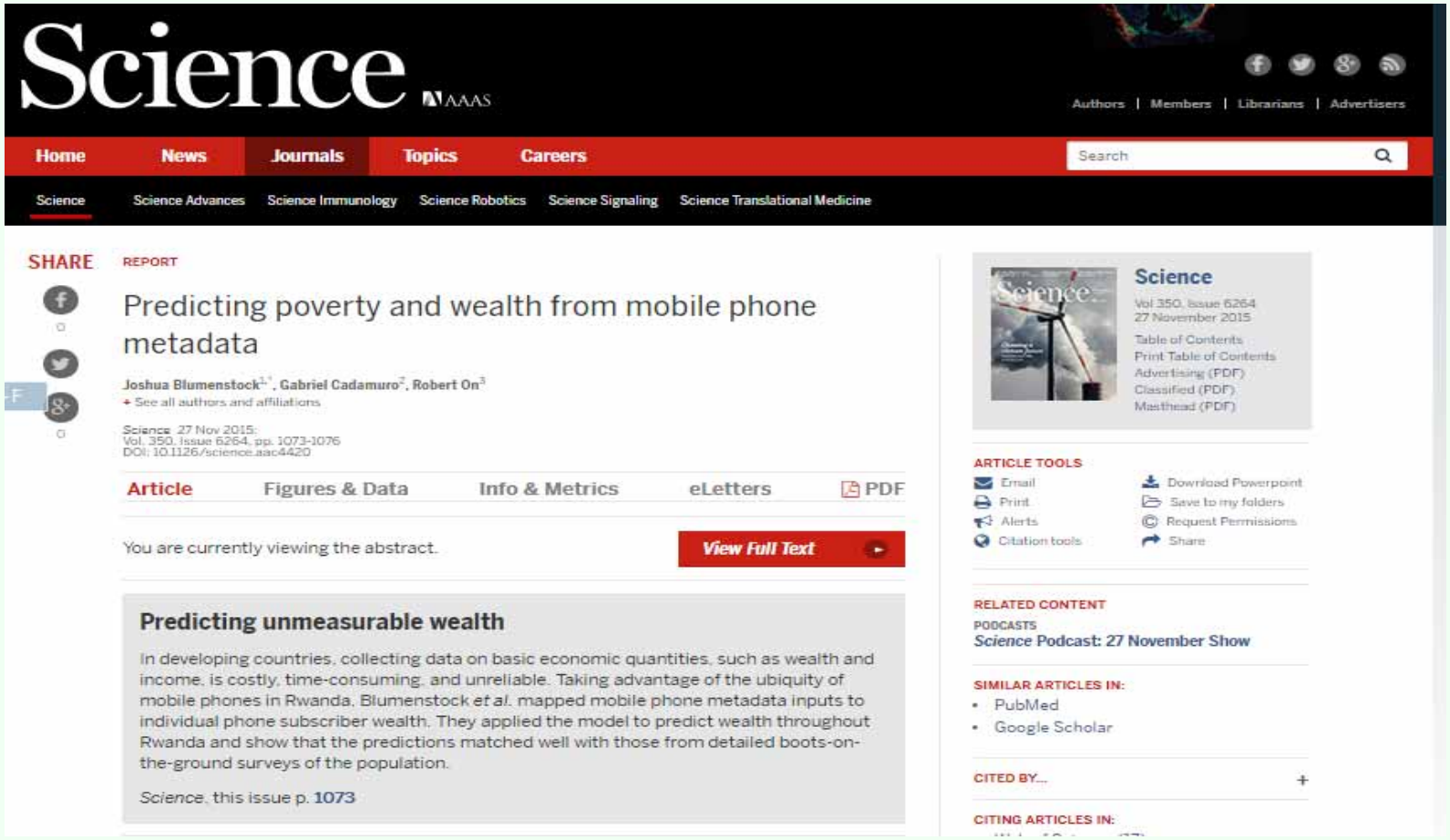


Fig. 2. Visualization of features. By column: Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, nonurban areas, water, and roads) in the convolutional neural network model used for extracting features. Each filter "highlights" the parts of the image that activate it, shown in pink. By row: Original daytime satellite images from Google Static Maps, filter activation maps, and overlay of activation maps onto original images

Source: (Science) – Combining satellite imagery and machine learning to predict poverty

DATA FOR DEVELOPMENT

Appendix: **Big Data Analytics**



Science AAAS

Authors | Members | Librarians | Advertisers

Home | News | Journals | Topics | Careers | Search

Science | Science Advances | Science Immunology | Science Robotics | Science Signaling | Science Translational Medicine

SHARE

REPORT

Predicting poverty and wealth from mobile phone metadata

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

+ See all authors and affiliations

Science 27 Nov 2015:
Vol. 350, Issue 6264, pp. 1073-1076
DOI: 10.1126/science.aac4420

Article | Figures & Data | Info & Metrics | eLetters | PDF

You are currently viewing the abstract. View Full Text

Predicting unmeasurable wealth

In developing countries, collecting data on basic economic quantities, such as wealth and income, is costly, time-consuming, and unreliable. Taking advantage of the ubiquity of mobile phones in Rwanda, Blumenstock *et al.* mapped mobile phone metadata inputs to individual phone subscriber wealth. They applied the model to predict wealth throughout Rwanda and show that the predictions matched well with those from detailed boots-on-the-ground surveys of the population.

Science, this issue p. 1073

Science
Vol 350, Issue 6264
27 November 2015
Table of Contents
Print Table of Contents
Advertising (PDF)
Classified (PDF)
Masthead (PDF)

ARTICLE TOOLS
Email | Download Powerpoint
Print | Save to my folders
Alerts | Request Permissions
Citation tools | Share

RELATED CONTENT
PODCASTS
Science Podcast: 27 November Show

SIMILAR ARTICLES IN:
- PubMed
- Google Scholar

CITED BY... +

CITING ARTICLES IN:

ADB

DATA FOR DEVELOPMENT

Appendix: **Big Data Analytics**

How data science and analytics can contribute to sustainable development

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

GLOBAL PULSE

www.unglobalpulse.org
@UNGlobalPulse 2016

ADB