Opportunities and Requirements for Leveraging Big Data for Official Statistics and the SDGs in Latin America

Emmanuel Letouzé
Director and co-Founder, Data-Pop Alliance
Visiting Scholar, MIT Media Lab

Eight meeting of the Statistical Conference of the Americas of ECLAC
Substantive seminar: "The data revolution and the 2030 Agenda for Sustainable Development: Challenges and Opportunities for National Statistical Institutes"
Quito, November 17, 2015

Thanks

- **ECLAC:** Alicia Barcena, Romain Zivy, Pilar Arturo, Paula Fuenzalida, Maria Eugenia Johnson…
- **UNFPA:** Pablo Salazar, Sabrina Juran…
- **INEC:** Carlos Riva,…
- All other organizers..

DATA-POP ALLIANCE

WHITE PAPER SERIES

Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America

FULL DRAFT V1 FOR DISCUSSION

November 2015

Please send comments & suggestions to eletouze@datapopalliance.org

DATA-POP ALLIANCE
WHITE PAPERS SERIES

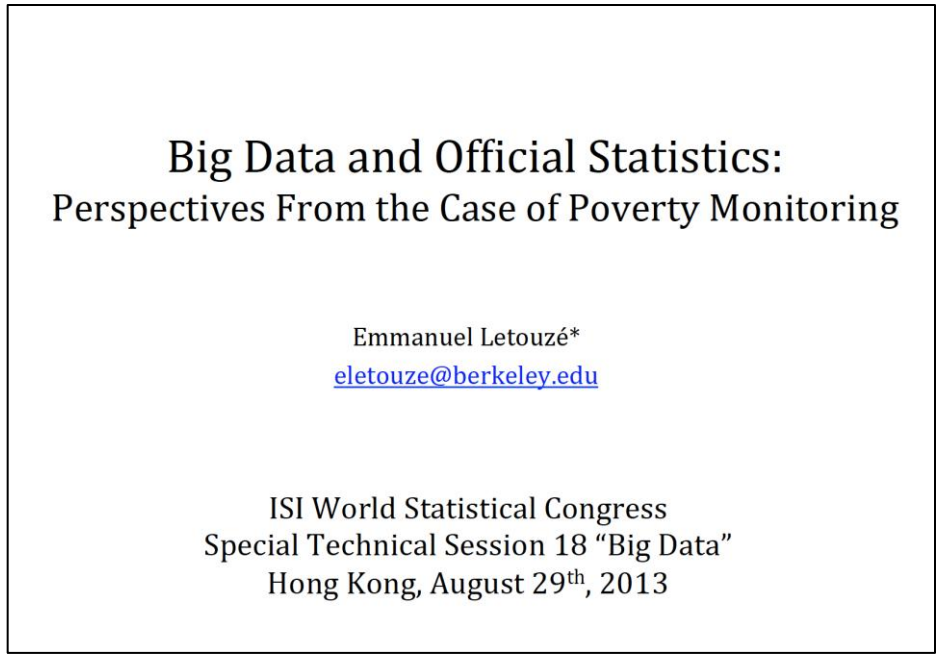
OFFICIAL STATISTICS, BIG DATA AND HUMAN DEVELOPMENT

March 2015

Emmanuel Letouzé
Johannes Jütting

Both a new and old topic…

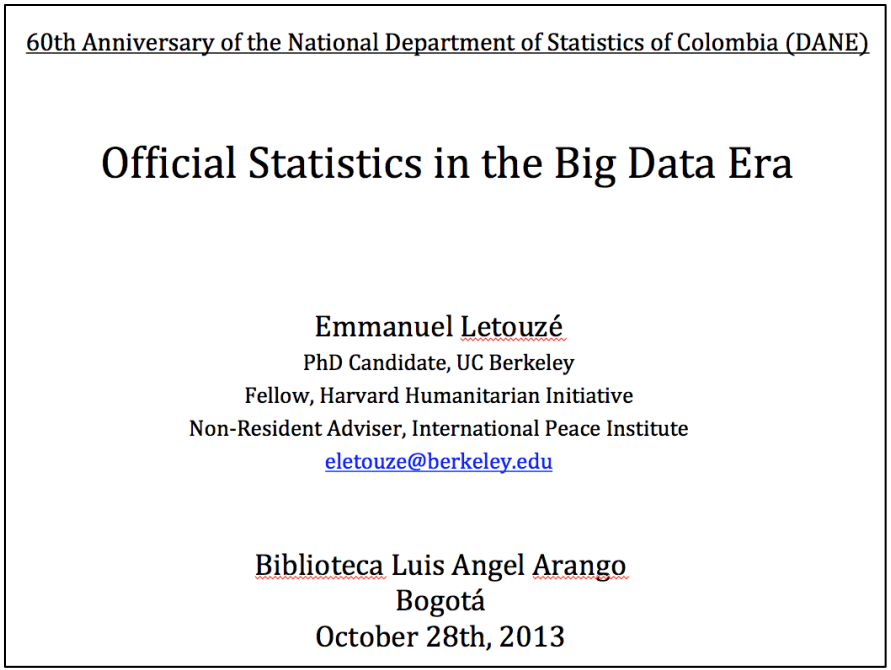
- 2 years ago…feels like 10 years ago..



Big Data and Official Statistics:
Perspectives From the Case of Poverty Monitoring

Emmanuel Letouzé*
eletouze@berkeley.edu

ISI World Statistical Congress
Special Technical Session 18 "Big Data"
Hong Kong, August 29th, 2013



60th Anniversary of the National Department of Statistics of Colombia (DANE)

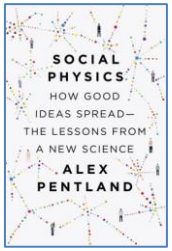
Official Statistics in the Big Data Era

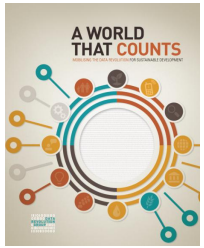
Emmanuel Letouzé
PhD Candidate, UC Berkeley
Fellow, Harvard Humanitarian Initiative
Non-Resident Adviser, International Peace Institute
eletouze@berkeley.edu

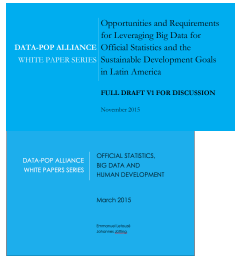
Biblioteca Luis Angel Arango
Bogotá
October 28th, 2013

‘Data is the new oil’—the rush

“We are at the beginning of what I call The ***Industrial Revolution of Data.***” Joe Hellerstein , November 19, 2008

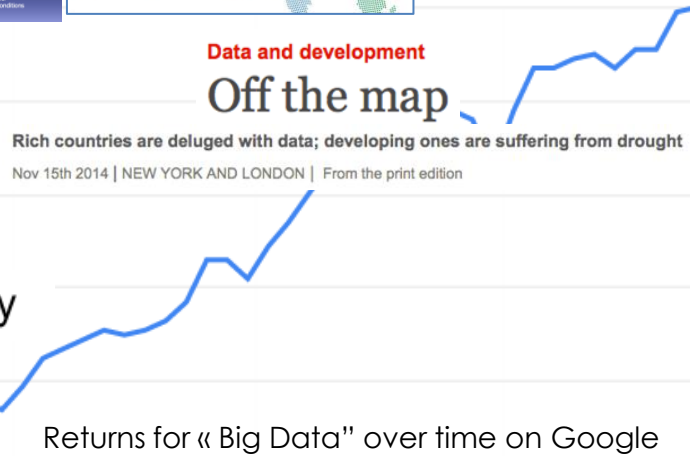




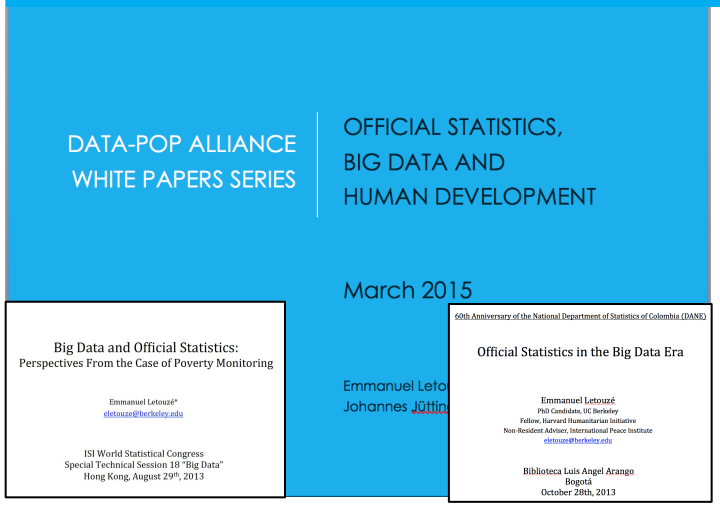


6 provocations

Africa’s statistical tragedy



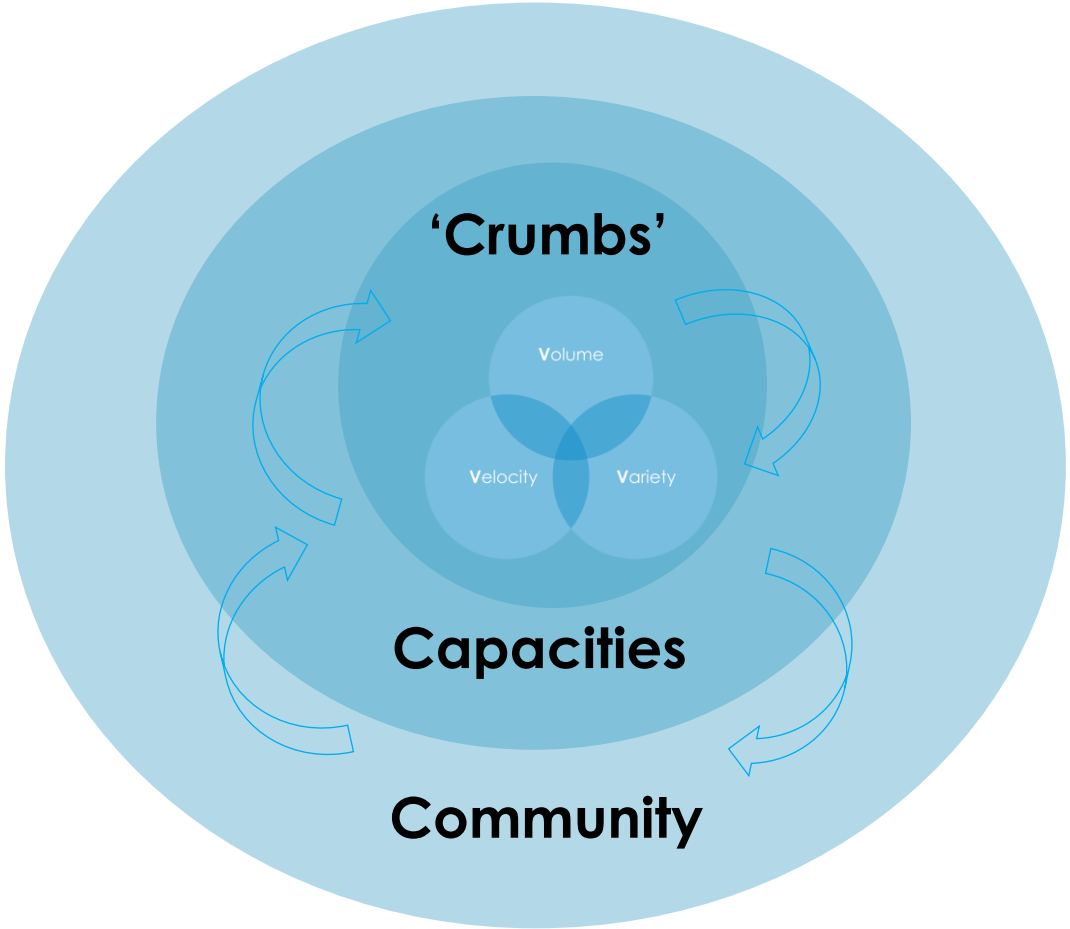
Returns for « Big Data” over time on Google



Main message:

"Big Data as a new ***ecosystem*** *presents the* ***official statistical community*** *with* ***a historical opportunity and a political obligation*** *to* ***engage with it*** *to retain or regain* ***its role as the legitimate custodian of statistical knowledge*** *and creator of a* ***deliberative public space*** *to discuss and drive human development in and about societies on the basis* ***of sound democratic and statistical principles****."*

What is Big Data? *An ecosystem*



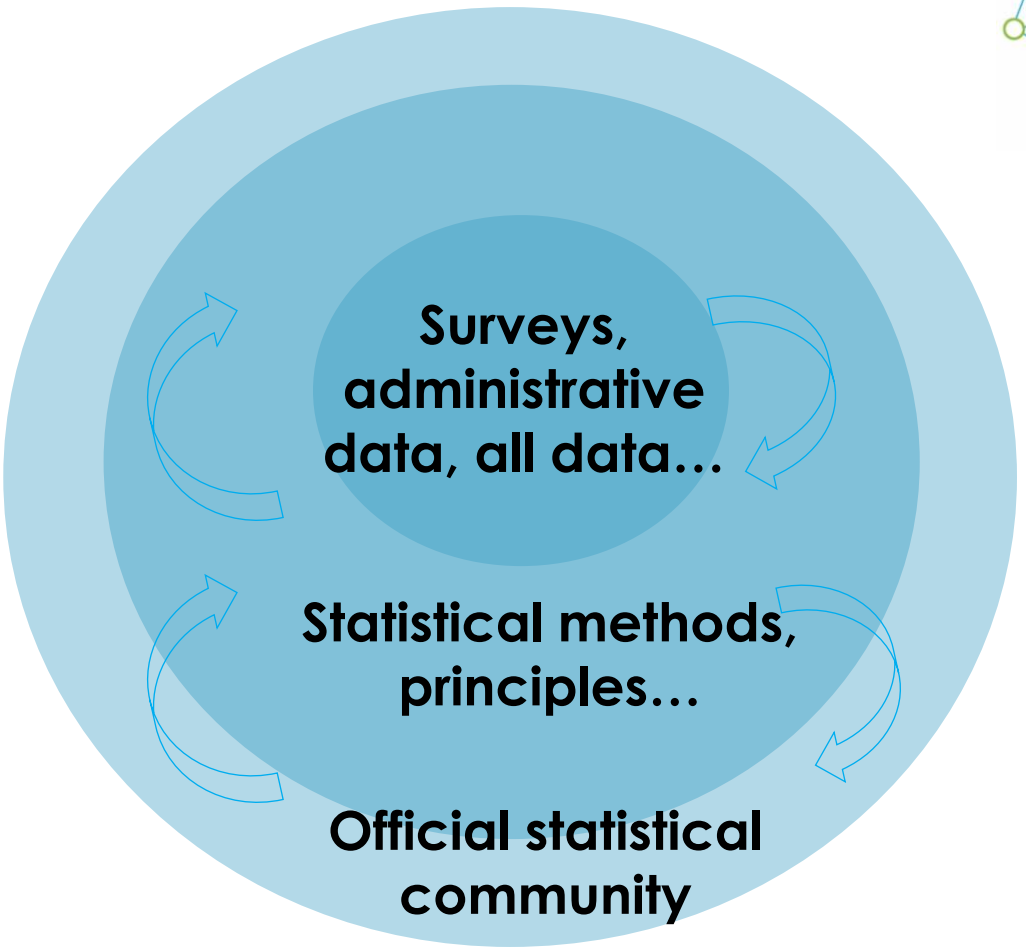
Made up of the **3Cs of Big Data**

Serving **4 main applications**:

1. ***Descriptive*** *(e.g.* maps);
2. ***Predictive***
    1. now-casting' or inference
    2. forecasting
3. ***Prescriptive***, by establishing *causal* relations
4. ***Discursive****;* by spurring and shaping dialogue within and between communities

What is Official Statistics? also an ecosystem

An ecosystem that turns **data into knowledge** and provides a **space to drive and discuss** development outcomes **according to sound democratic and statistical principles**



What is Big Data for Official Statistics?…

"For the Official Statistical Community, ***Big Data is a wake up call****"*

John Pullinger
UK National Statistician, Chair of the 46th UN Statistical Commission Session,
Data-Pop Alliance—ODI-RSS event,
January 19th, 2015

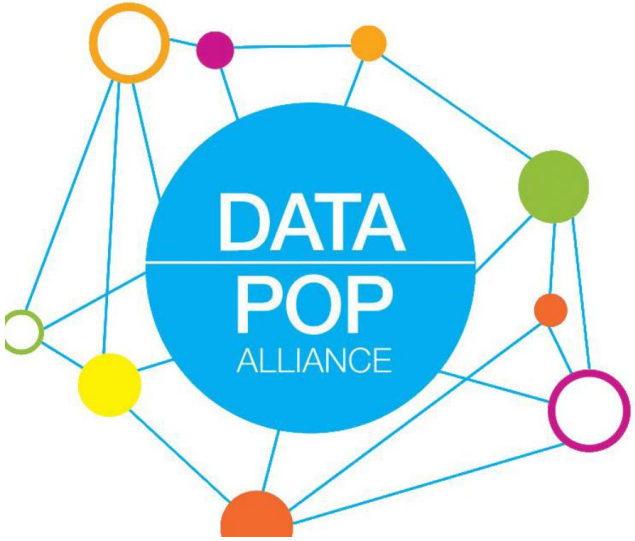
Another way to say it: ***a historical opportunity and political obligation***

DATA-POP ALLIANCE
WHITE PAPER SERIES

Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America

FULL DRAFT V1 FOR DISCUSSION

November 2015



HARVARD HUMANITARIAN INITIATIVE

mit media lab

Table of Contents

The national statistical offices of Latin American and Caribbean countries

A selection of non-NSO actors and projects involving Big Data in Latin American and Caribbean countries

1. Government
Agência Nacional de Águas + Ministério do Meio Ambiente (Brazil): Installing smart meters to help the NSO monitor customer water bills in Brazil.

2. Government
Secretaría de Turismo (Mexico): Using geo-located Twitter date to track domestic tourism in the states of Guanajuato and Puebla.

3. Government
Instituto Nacional de Vías (Colombia): Using GPS data to improve traffic circulation and to serve as input for transport statistics in Colombia.

4. Government
Ministerio de Hacienda y Crédito Público (Colombia): Using Google Trends data to nowcast economic activity in Colombia and create better indicators for the NSO.

5. Government
Ministerio del Trabajo de la República de Colombia: Using webscraping techniques to monitor job vacancies in Colombia.

6. Academic, Government
University of Chicago + Office of the President (Mexico): Scraping public records to study maternal mortality in Mexico.

7. Academic
Michele Coscia and Viridiana Rios: Using data on Google search to track Mexican drug trafficking organizations.

8. NGO
World Bank: Using Twitter data to analyze public perception of a gas subsidy reform in El Salvador.

9. Government
Rio Operations Center: Using satellite and weather data to improve emergency management in Rio.

10. Government, Private
Municipality of Halisco (Mexico) + Telefónica: Using cellphone data to monitor displacement after natural disasters in Mexico.

11. Private
PriceStats: Tracking website visits to build a consumer price index and monitor inflation in Argentina.

12. Private
Ecovec: Monitoring Dengue fever vectors in individual homes through SMS reporting in Brazil.

13. Private
Microsoft Research: Assessing social media trends and in particular Twitter hashtags to examine the drug war in Mexico.

14. Academic
Fundação Getúlio Vargas: Monitoring population movements through cell-phone use to help plan transportation and control epidemics in Brazil.

15. Academic, NGO
Universidad TecMilenio + Data2x: Analyzing subjective wellbeing in Tweets to create a measures that can help disaggregate poverty data with regard to gender in Mexico, in partnership with the NSO.

16. NGO
International Center for Tropical Agriculture + Federación de productores de arroz (Colombia): Using information on farmers' crop and weather patters to identify what kind of agricultural practices have historically worked well in specific locations during particular spells of weather.

17. Government, NGO
Ministerio del Interior y Justicia (Colombia) + UN Office on Drugs and Crime: Using satellite data to monitor coca crops in Colombia.

18. Private, NGO
Telefónica + Worldbank: Using cellphone data to monitor poverty levels in Guatemala, in partnership with the NSO.

19. Private, NGO
Telefónica + World Bank + Data-Pop Alliance: Using cellphone and survey data to predict crime and socioeconomic indicators in Bogotá.

20. NGO
UN Women: Predicting malaria risk levels through satellite remote sensing of geophysical and demographic data in Brazil in Latin American and Caribbean countries.

21. NGO
Paris21: Using luminosity data from satellite imagery as a proxy for economic statistics in Latin American and Caribbean countries.

22. NGO
Humanitarian Data Exchange: Facilitating information exchange among UN agencies, NGOs, and government by establishing standards and hosting a data sharing platform worldwide.

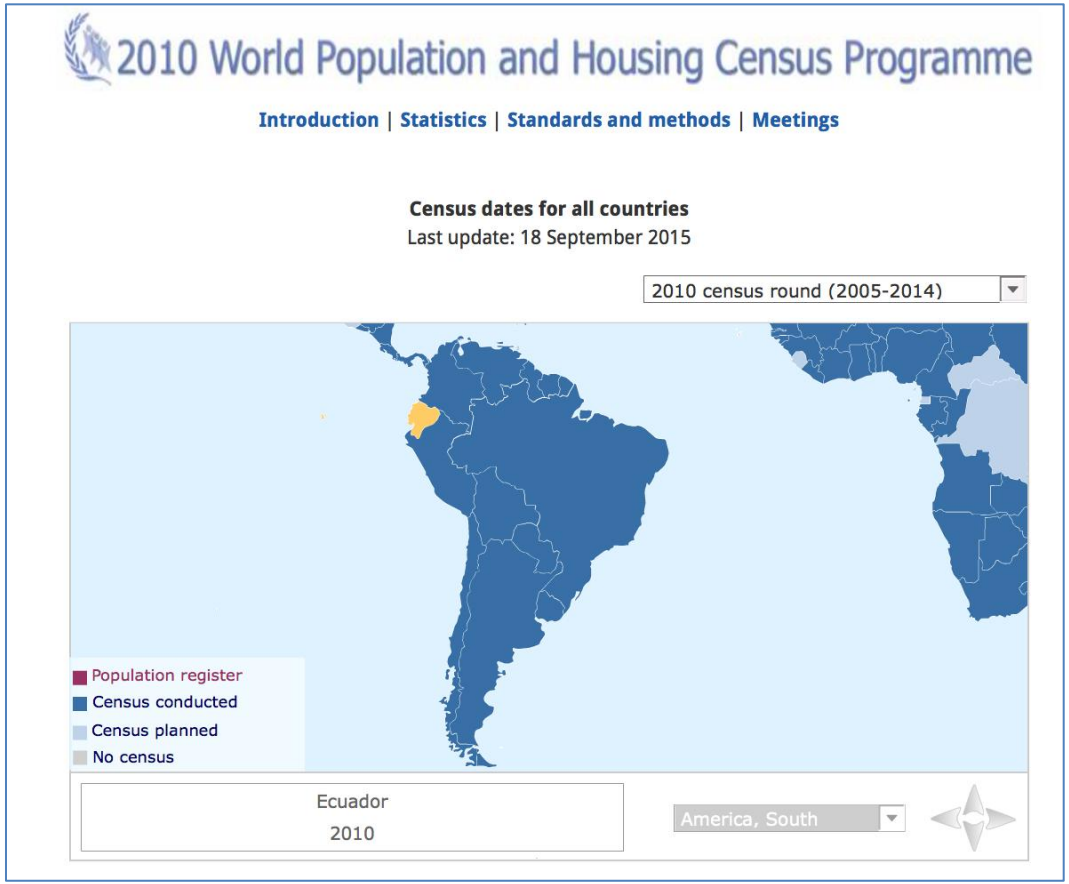
23. NGO
World Resources Institute: Using satellite imagery for real-time monitoring of forest resources worldwide.

24. NGO
UN Global Pulse: Analysis official data from several countries to study crime patters in relation to economic shocks.

DATA POP ALLIANCE

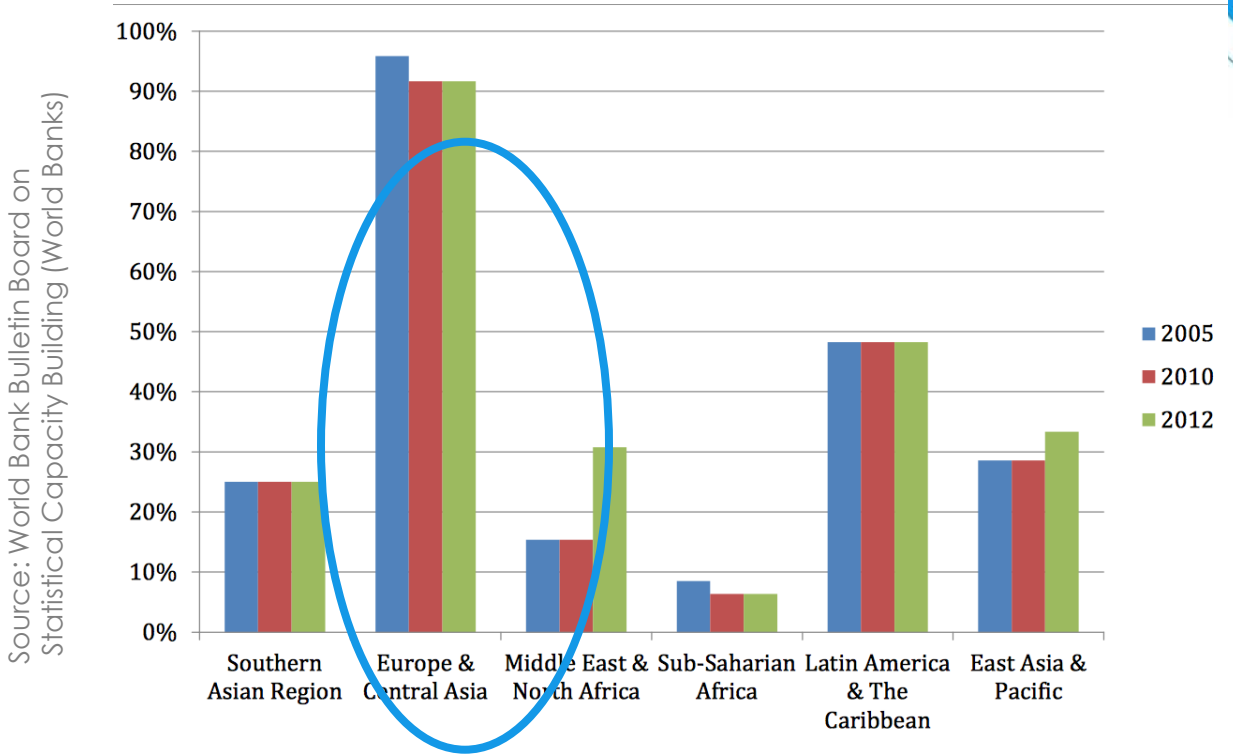
The State of Official statistics in LAC: B+?

Censuses: A-



The State of Official statistics in LAC: B+?

Other aspects: B



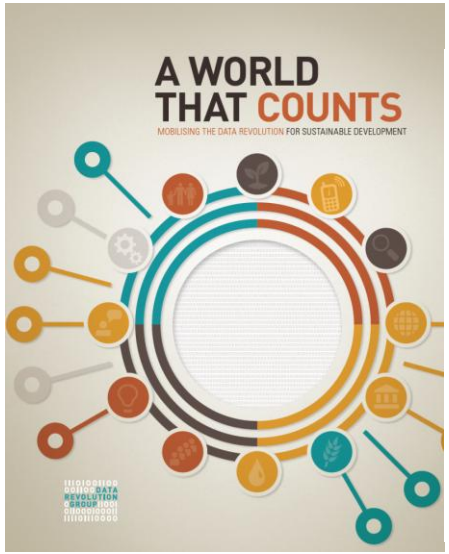
➔ CEPAL project: hard to estimate maternal mortality because of lack of certification or registry in areas inhabited by indigenous populations or remote

Main obstacles for LAC NSOs to leverage Big Data

1. **Institutional and cultural barriers to innovation and change**, e.g. lack of internal digital culture, scepticism vs. on Big Data, political will & coordination
2. **Constraints to data access and completeness**, particularly in access and continued use of data held by the private sector, limited ownership rights involving people and their relationships with data;
3. **Technical challenges**, including infrastructure for capturing, cleaning, processing, analyzing and visualizing structured and unstructured data
4. **Human capacity gaps**, including talent discovery, data literacy, limited data science training programs and involvement of academia
5. **Methodological challenges**, including challenges in data representativeness, biases, and the lack of standards and guidelines;
6. **Ethical and political risks**, including risks to privacy and weak legal frameworks

But huge window of opportunities and assets too

Post-2015 development / SDG / Big and Open Data agendas place NSOs at their core




CYBERSECURITY

Saving Big Data from Itself

A three-step plan for using data right in an age of government overreach

By Alex "Sandy" Pentland

For the first few decades of its existence, the National Security Agency was a quiet department with one primary job: keeping an eye on the Soviet Union. Its enemy was well defined and monolithic. Its principal tools were phone taps, spy planes and hidden microphones.

After the attacks of September 11, all of that changed. The NSA's chief enemy became a diffuse network of individual terrorists. Anyone in the world could be a legitimate target for spying. The nature of spying itself changed as new digital communication channels proliferated. The exponential growth of Internet-connected mobile devices was just beginning. The NSA's old tools apparently no longer seemed sufficient.

In response, the agency adopted a new strategy: collect everything. As former NSA director Keith Alexander once put it, when you are looking for a needle in a haystack, you need the whole haystack. The NSA began collecting bulk phone call rec-

GLOBAL PARTNERSHIP
FOR SUSTAINABLE DEVELOPMENT DATA

Table 2: Overview of Big Data projects in selected LAC NSOs

Type of Big Data	Data Used in Current NSO Activities	Projects	Other Organizations Involved	Project Status
		Argentina (INDEC)		
Exhaust data	Web scraping	IPC Online		Planned
		Brazil (IBGE)		
Digital content	Google Maps	Developing Water Accounts	National Water Agency	Implemented /On-Going
	CDRs	Tourism Monitoring	IBGE	Planned
		Colombia (DANE)		
Exhaust data	Web scraping	IPC Online SIPSA		Planned Implemented/On-going
Digital content	CDRs	Monitoring crime activities Socio-economic levels and networks	World Bank Data-Pop Alliance TransMilenio	Pilot stage Pilot stage Pilot stage
Sensing data	Satellite	Complementing the National Agriculture Census		Completed
		Ecuador (INEC)		
Exhaust data	Web scraping	IPC Online		Pilot stage
Digital content	Twitter CDRs	Measuring Subjective Wellbeing Daytime Migration		Pilot stage Planned
		Guatemala (INE)		
Digital content	CDRs	Monitoring Poverty Levels	World Bank Telefónica	Pilot stage
		Mexico (INEGI)		
Digital content	Twitter	Subjective Wellbeing	InfoTec and Tec Monterrey	Completed
		Subjective Wellbeing of Women	Data2x, the University of Pennsylvania	Pilot
		Tourism Monitoring Movements Across Borders	Ministry of Tourism	Completed Planned

Latin American NSOs are active!

Overview of selected Big Data projects involving regional NSOs

Overview of selected Big Data projects involving regional NSOs

Box 4: Twitter for Tourism Monitoring in Mexico

In 2014, a working group on Big Data at INEGI conducted a pilot study to track domestic tourism from Twitter data, in order to contribute to the empirical modelling of individual tourist behavior.The objective of this pilot program was to identify the characteristics of an average Tweeting tourist in order to identify how many people travelled to Puebla and Guanajuato during the holiday weekend of February 1-3, 2014. The team of researchers from INEGI, in collaboration with the Mexican Ministry of Tourism, analysed 60 million Tweets from January to July 2014, from the continuous 1% georeferenced sample that Twitter makes available for free.[65] From this data, INEGI collected Tweets from the 7,955 Twitter users who Tweeted in Guanajuato (48%) and Puebla (52%) during the holiday. They then gathered all the Tweets sent by those users in the remainder of the target period (amounting to 827,424 total Tweets), and identified which users Tweeted from another state (presumably their homestate) after being in Guanajuato or Puebla, in order to map the origin of domestic tourism to those two areas during the holiday.[66] The resulting estimates of domestic tourism to Guanajuato and Puebla were compared to estimates made by the respective offices of tourism of those two states.[67]

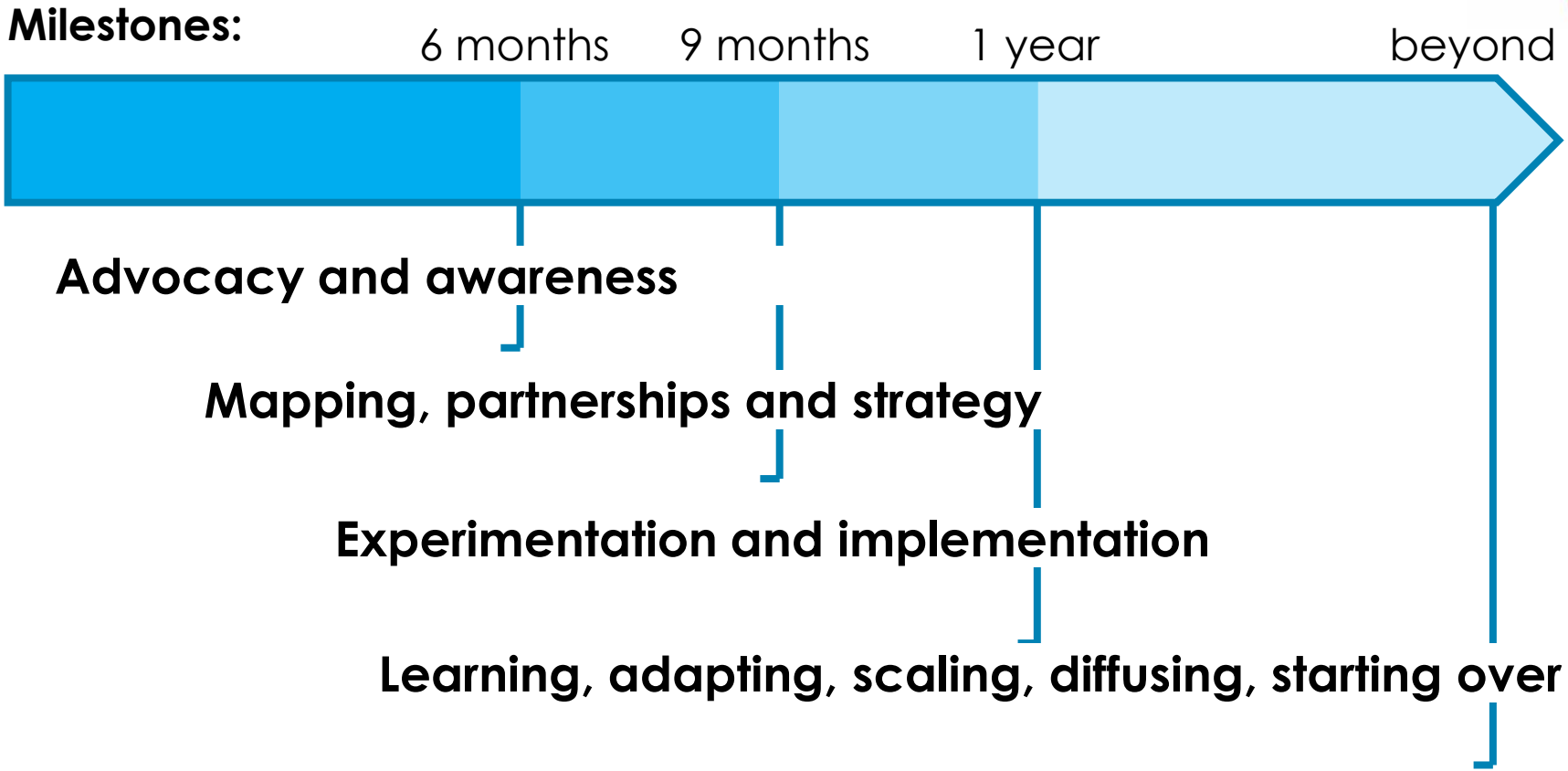
Box 5: Maternal Morbidity and Remote Sensing of Malaria in Brazil

Remote sensing satellite data on vegetation density, soil moisture, population density, and spatial pattern of human infrastructure have long been used to predict levels of malaria risk. Advances in computing now allow more powerful use of these big datasets, including analysis of extreme spatial and temporal heterogeneity and inclusion of greater numbers of explanatory variables. This project seeks to create malaria risk maps for the Amazon Basin, focusing first on urban and peri-urban zones along the Brazil-Guyana border, which are areas with highly variable vector habitats and elevated incidences of illness. At least two vector distribution-mapping studies in this region exist, but to our knowledge there is no high-resolution dynamic mapping of malaria risk. The first phase of the project will use remote sensing data and existing health records, in combination with information about the economic, cultural, and health system, to estimate a spatial regression model that predicts morbidity burden in pregnant women, using DALYs (Disability-Adjusted Life Year) as the principal metric. The second phase will then test the accuracy of this model using data collected in real-time. UN Women and IBGE and the leading institutions piloting this study, drawing support from partner institutions The Vargas Foundaiton and the Amazon Malaria Inititatve.

LAC countries and their NSOs have many assets too

1. An **urban, relatively young, innovative and technology hungry population** that share 4 major languages (Sp., Port., Fr., Eng.)
2. A long experience of the **Open Data movement**
3. The presence of strong **region-wide civil-society groups, institutions, academic networks, private sector, and working groups**
4. The emergence of **several pilots and public-private partnerships on Big Data**
5. The active involvement of **governments and public actors** that devise new **strategies and policies** and participate in **global forums**

Strategies must be *multi-partner*, multi-sector, *multi-year* and *regional*



The paper was part of multi-partner research, training and strategic assistance program funded by the World Bank

1. **Technical assistance:** strategy paper + strategic support

2. **Research:** 2 pilots on crime and poverty

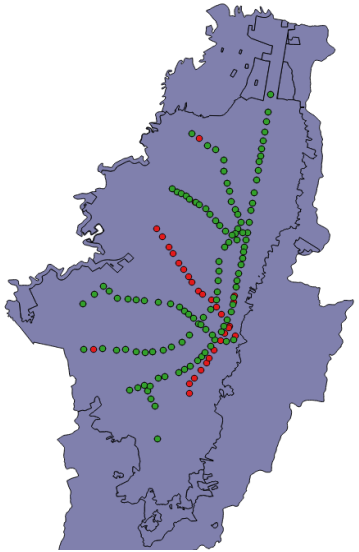
3. **Regional training program** seed-funded by the Hewlett Foundation wigth other partners (Paris21, CEPAL..UNFPA…)

WORLD BANK
FONDAZIONE BRUNO KESSLER
Telefónica
Universidad de los Andes
DANE
TODOS POR UN NUEVO PAÍS
PAZ EQUIDAD EDUCACIÓN

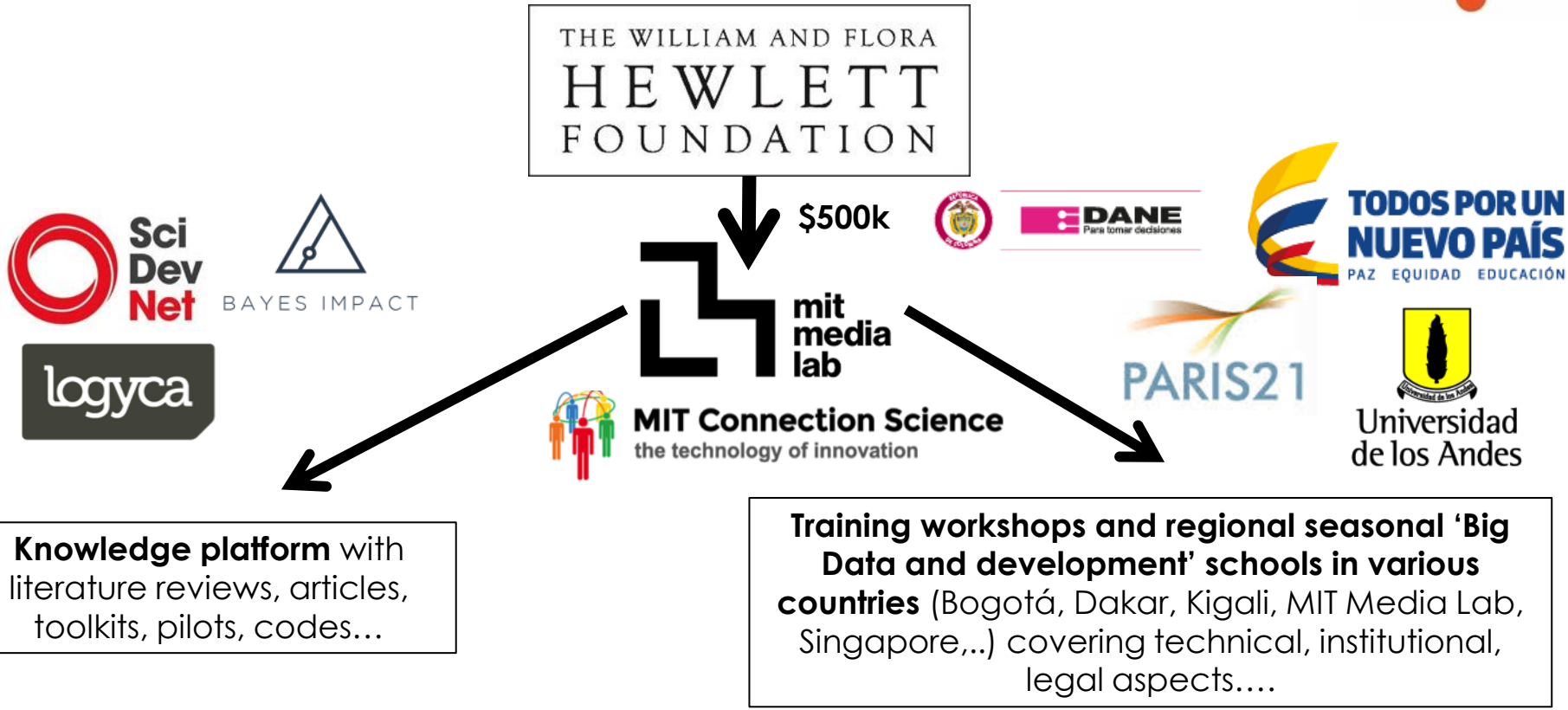
Piloto 2: predicting crime trends and patterns using aggregate data on people's dynamics

Research questions:

- Which areas are more likey to experience crime based on their digital traces?
- Can a public transportation system act as a public safety system?
- How do social networks in CDRs and crime interact?



Regional professional training program in Big Data en development 2015-2017



Regional Data Spaces

Brokering and fostering regional synergies to foster a people-centered Big Data revolution


SNAPSHOT January 15, 2015 Africa Americas

Fixing Fragile Cities

Solutions for Urban Violence and Poverty

DATA POP ALLIANCE

IGARAPÉ INSTITUTE
think connect transform

MIT Connection Science
the technology of innovation

UNFPA

SUSTAINABLE DEVELOPMENT SOLUTIONS NETWORK
A GLOBAL INITIATIVE FOR THE UNITED NATIONS

Cepei

SOCIALTIC

PARIS21

HDX Humanitarian Data Exchange

Introducing the Humanitarian Data Exchange

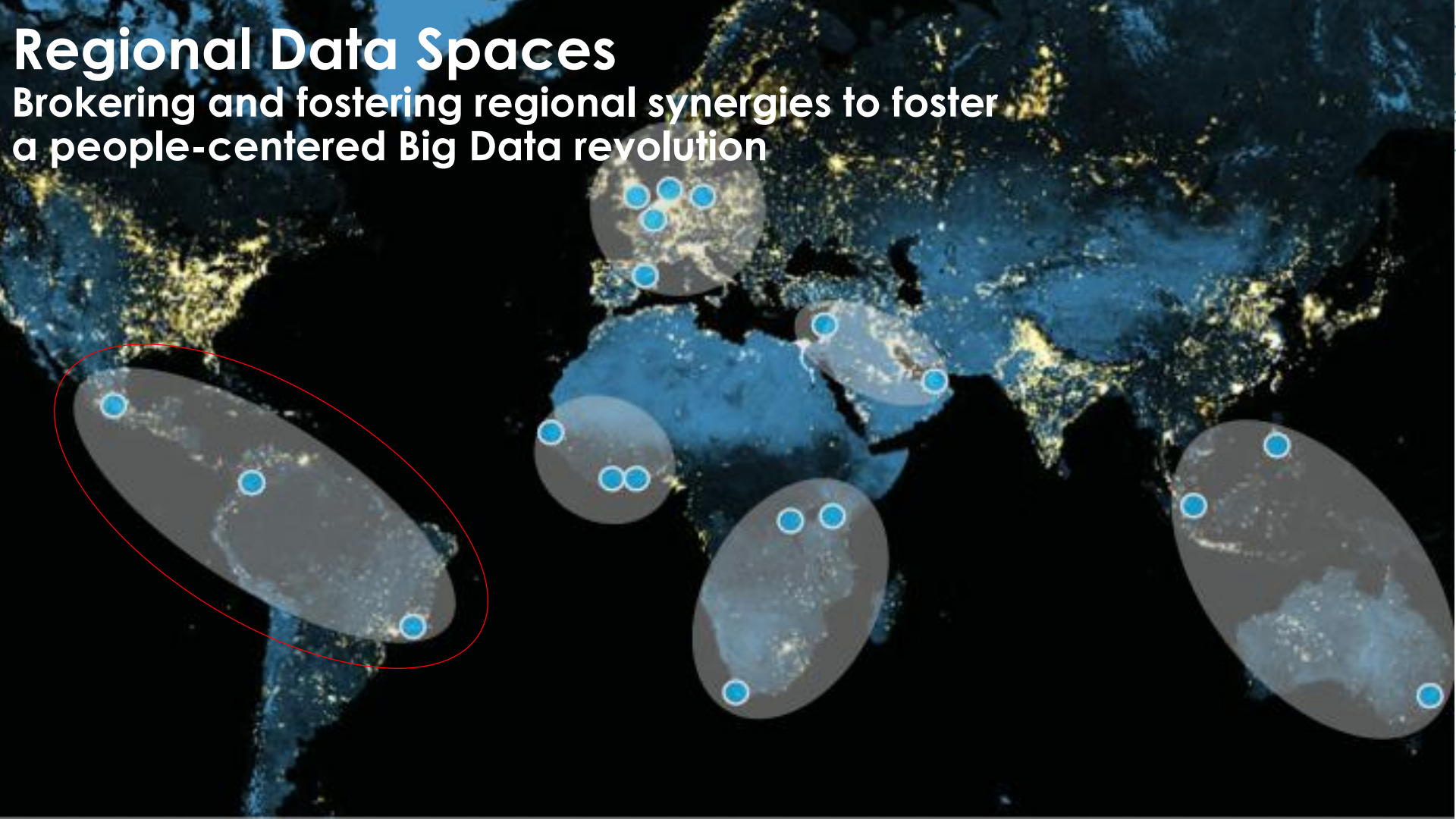
UNITED NATIONS

ECLAC

OXFAM México

Regional Data Spaces

Brokering and fostering regional synergies to foster a people-centered Big Data revolution




Gracias!
www.datapopallliance.org
eletouze@datapopalliance.org