DATA-POP ALLIANCE

MIT Connection Science
the technology of innovation

HARVARD HUMANITARIAN INITIATIVE

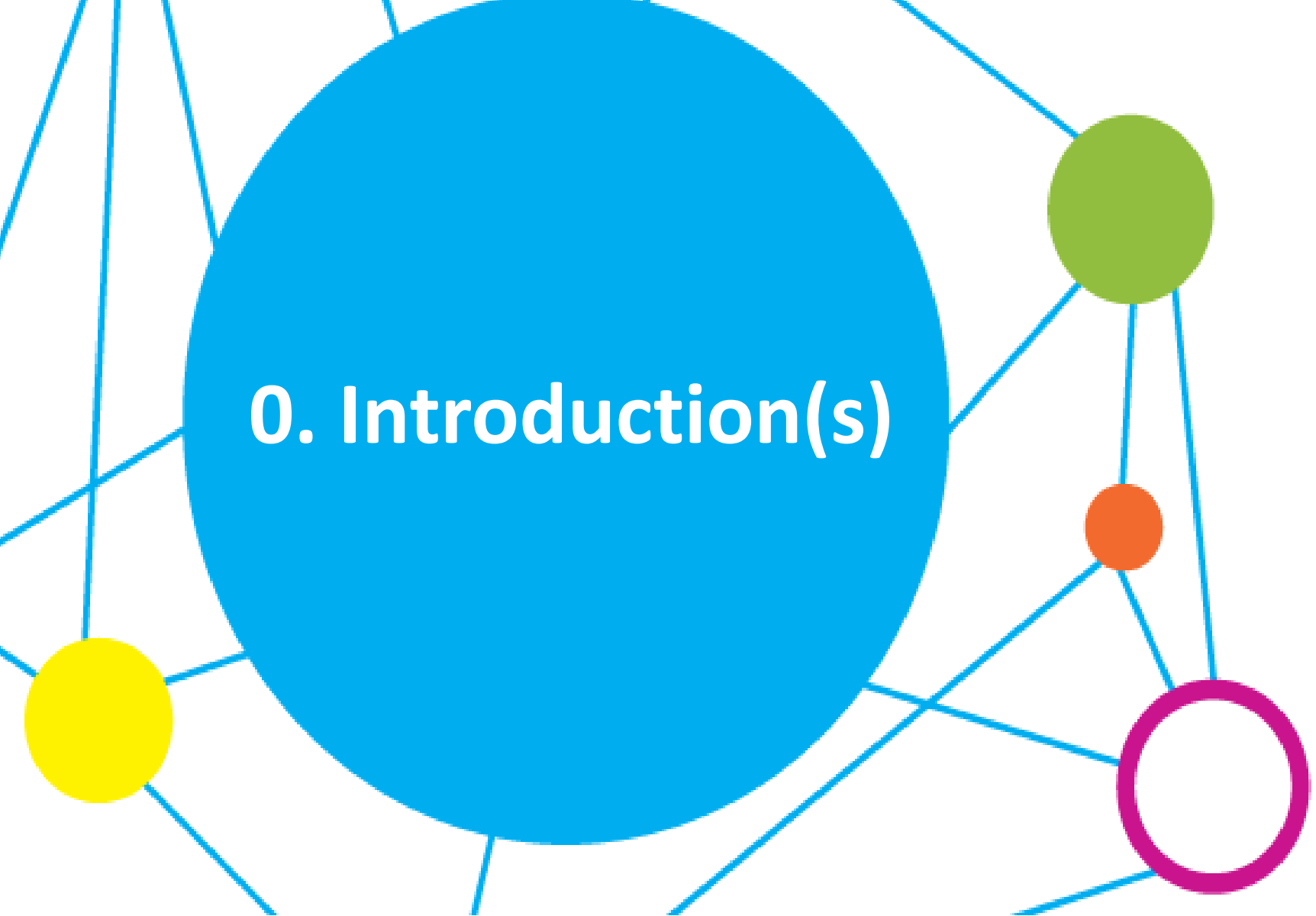
mit media lab

FLOWMINDER.ORG

Harnessing Big Data in Latin America

ECLAC-MIT Seminar
Santiago, Chile
March 6, 2017

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


0. Introduction(s)

Thanks

- **CEPAL**: Alicia Barcena, Alejandro Patiño, Claudia Guerrero, Sebastian Rovira, Romain Zivy, Pascual Gerstenfeld …
- **MIT**: Lee Ullman, Marissa Cannon, Sandy Pentland, David Shrier, Julie Hall…
- **Data-Pop Alliance**: Andrés Clavijo, Ana Lucia Martinez, Rodrigo Lara Molina, Natalie Shoup, Gabriel Pestre, Nuria Oliver, Julia Manske…

Data-Pop Alliance

is a global coalition on **Big Data & development** created by the **Harvard Humanitarian Initiative, MIT Media Lab**, and **Overseas Development Institute** joined by **Flowminder,** bringing together researchers, experts, practitioners and activists to "**promote a people-centered Big Data revolution"** by **locally co-designing and deploying collaborative research, training,** and **engagement activities**

CYBERSECURITY

Saving Big Data from Itself

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

By Alex "Sandy" Pentland

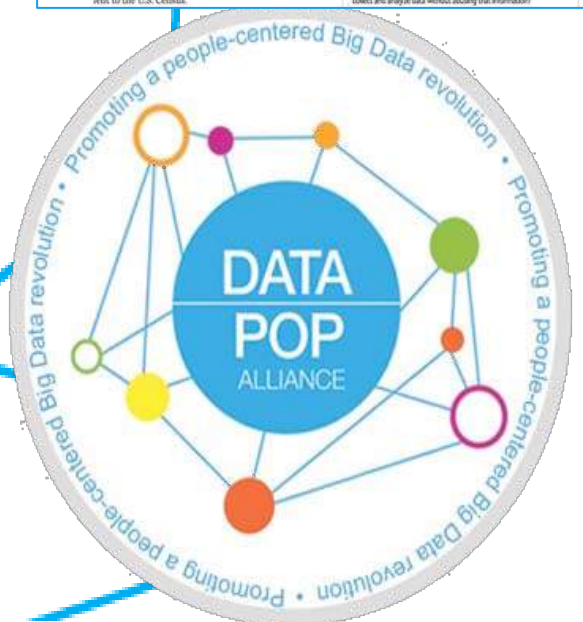
For the first few decades of its existence, the National Security Agency was a quiet department with one primary job: keeping an eye on the Soviet Union. Its enemy was well defined and monolithic. Its principal tools were phone taps, spy planes and hidden microphones.

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 records from virtually every person in the U.S.; soon it was gathering data on bulk Internet traffic from virtually everyone outside of the U.S. Before long, the NSA was collecting an amount of data every two hours equivalent to the U.S. Census.

IN BRIEF

Data about human behavior have always been essential for both government and industry to function. But how do we enable institutions to collect and analyze data without abusing that information?



Leadership

DATA POP ALLIANCE

Prof. Alex 'Sandy' Pentland
Academic Director

Dr Nuria Oliver Ramirez
Chief Data Scientist

mit media lab

Dr Linus Bengtsson
FLOWMINDER.ORG

Prof. Patrick Vinck
Co-Director &
Co-Founder

Elizabeth Stuart
Co-Director for ODI

HARVARD HUMANITARIAN INITIATIVE

ODI

Dr Emmanuel Letouzé
Director & Co-Founder

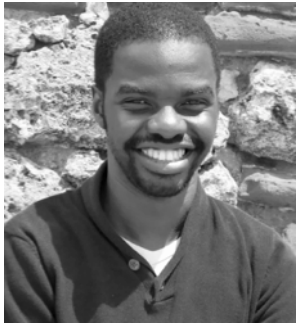
Prof. Phuong Pham
Co-Director for HHI

Dr Emma Samman

Core Team in NYC, Cambridge and LAC

Emmanuel Letouzé
Director & Co-Founder

Natalie Shoup
Training and
Engagement Manager



David Sangokoya
Research Manager
.

Gabriel Pestre
Research Scientist

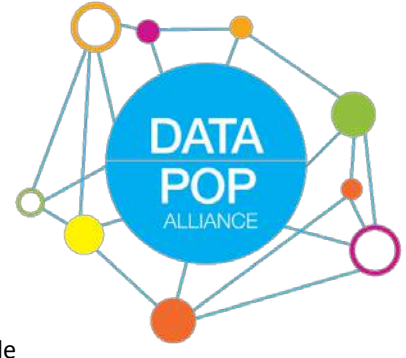
Julie Ricard
Research and
Communications Officer

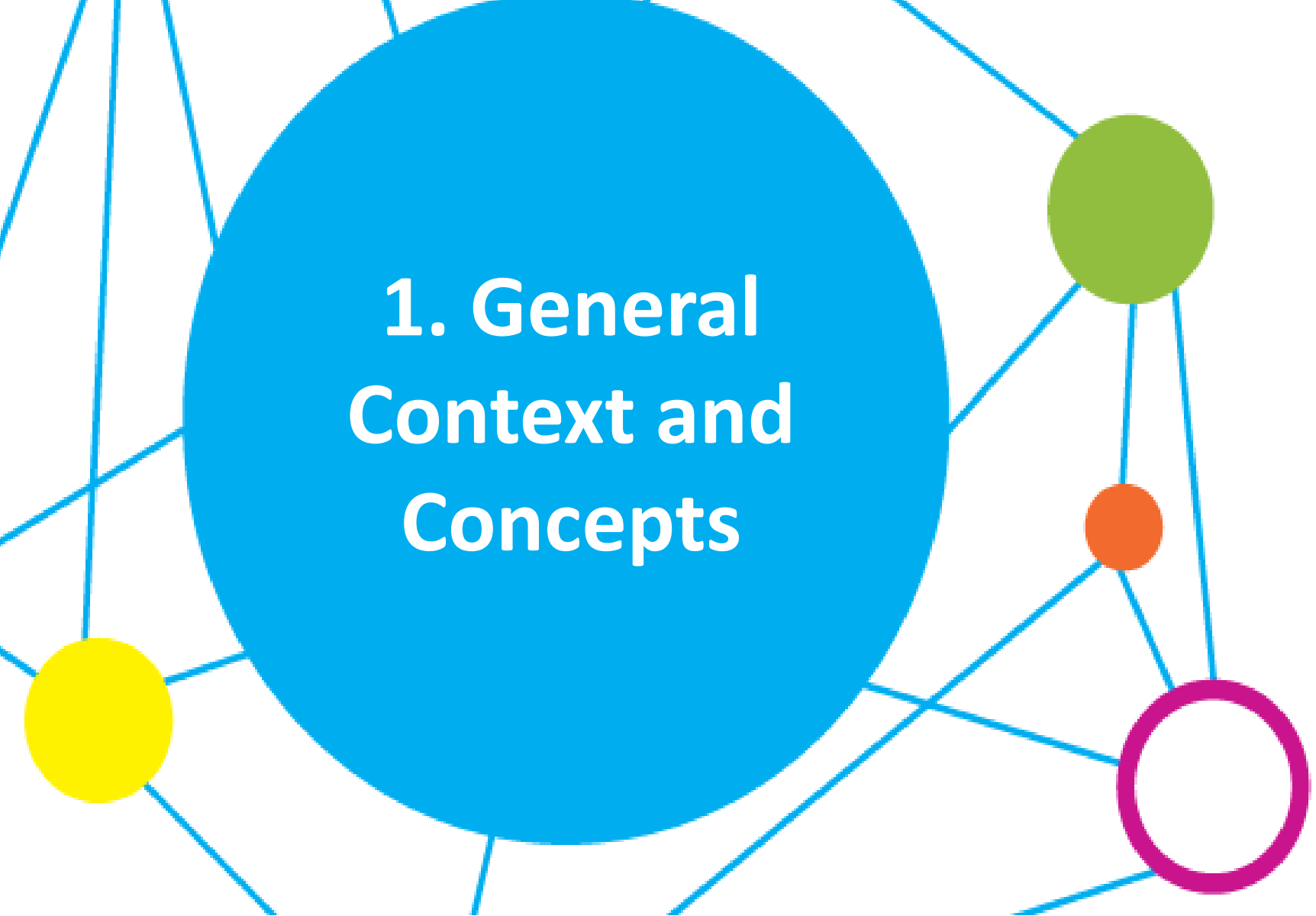
David Shrier
Managing Director
MIT Connection Science

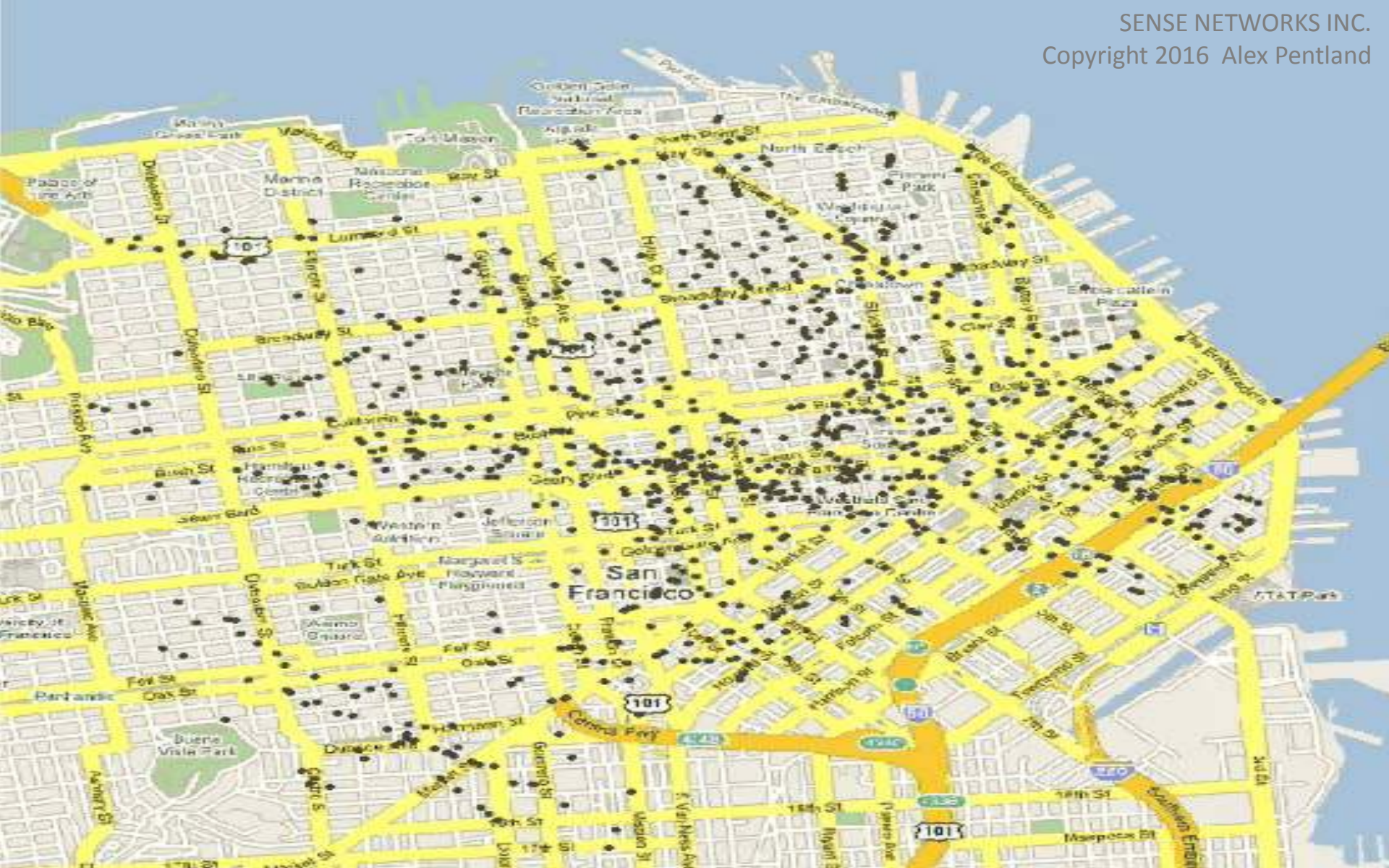
Andrés Clavijo
Lead Researcher
and Coordinator for Colombia

Ana Lucia Martinez
Program Officer in Colombia

Rodrigo Lara-Molina
Research Assistant in Chile




1. General
Context and
Concepts



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

SOCIAL SCIENCE

Computational Social Science

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors. *(Science, 2009)*

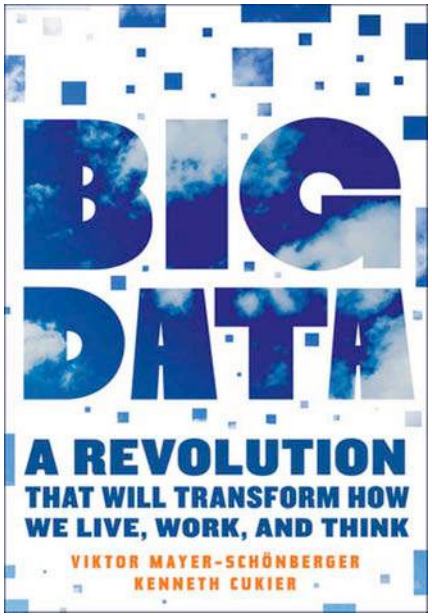
(2012)

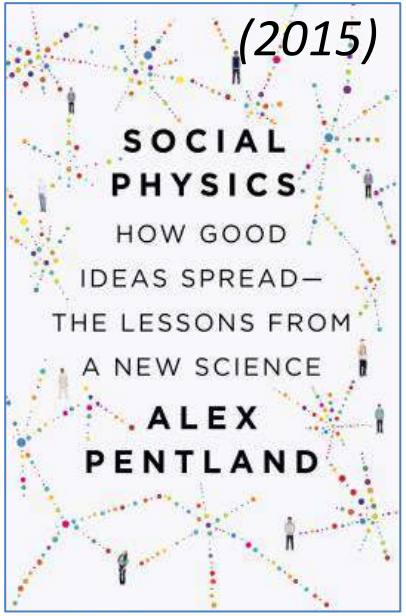
Data and development

Off the map

The Economist

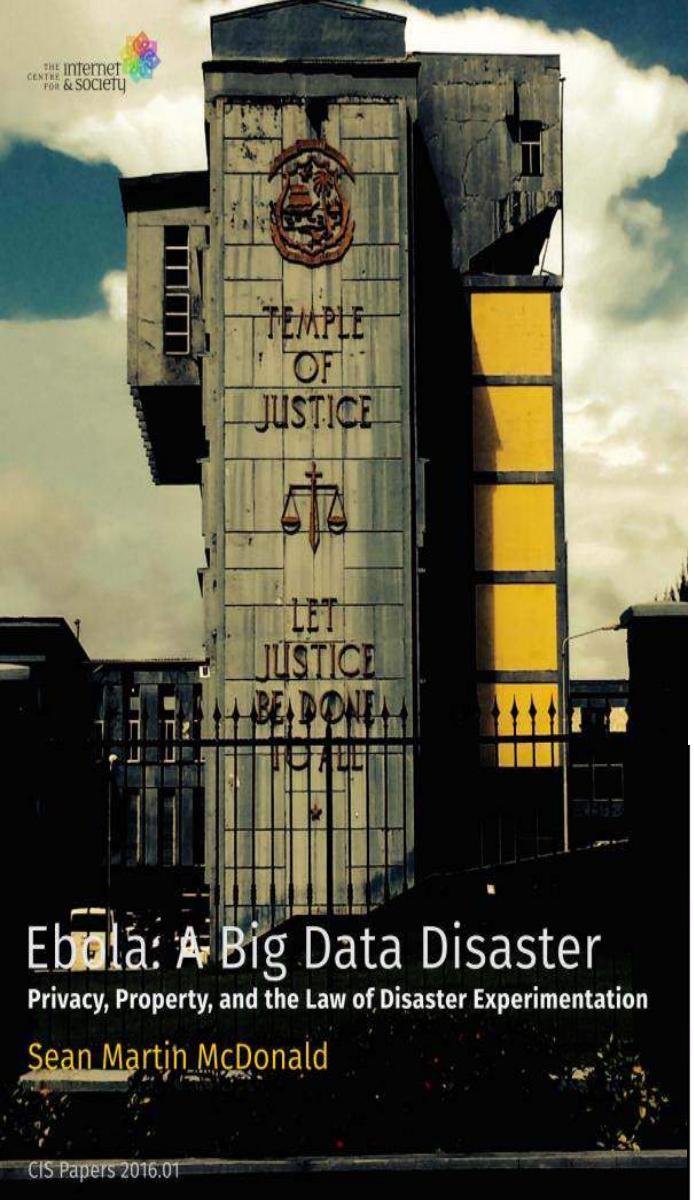
(2014)

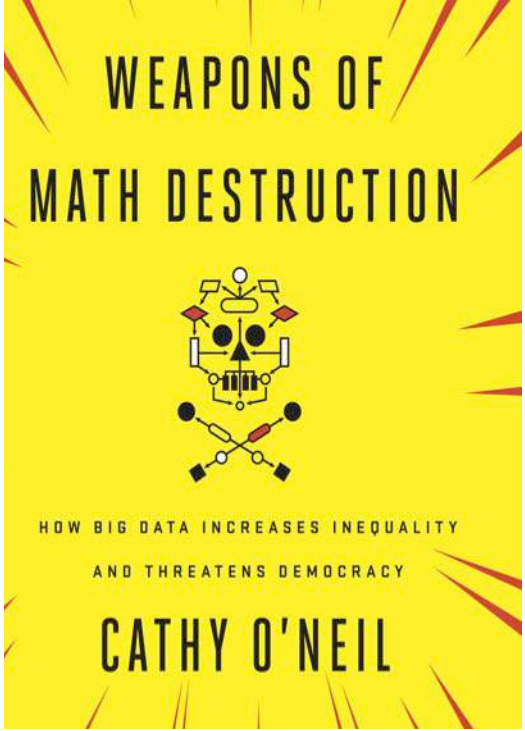




(2015)

Rich countries are deluged with data; developing ones are suffering from drought

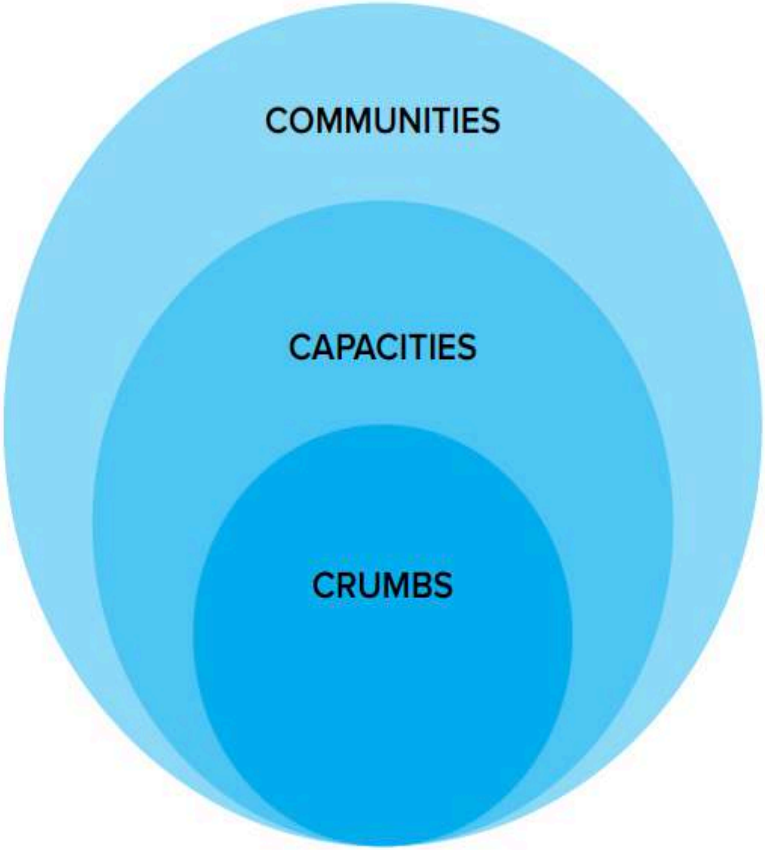




Science without conscience is the soul's perdition.

(François Rabelais)

THE 3 C's OF BIG DATA



Originally framed as the "3 V's" (volume, velocity and variety) in the early 2000s, Big Data has emerged as an ecosystem of "3 C's": digital "crumbs" (digital translations of human actions and interactions captured by digital devices); powerful capacities to collect, aggregate and analyze data; and communities involved in generating, governing and using data, including data generators, end users, policy-makers, experts, privacy advocates and civic hacker communities.

How can Big Data spur development and democracy? A taxonomy of 4 functions

1. Descriptive function

Showing features and patterns of human societies

2. Predictive function

a) 'Nowcasting' through proxies

b) Forecasting what may happen next

3. Prescriptive function

Based on causal relationships

4. Discursive function

Social dialogue and engagement through and about Big Data

2. Big Data and/in Latin America*

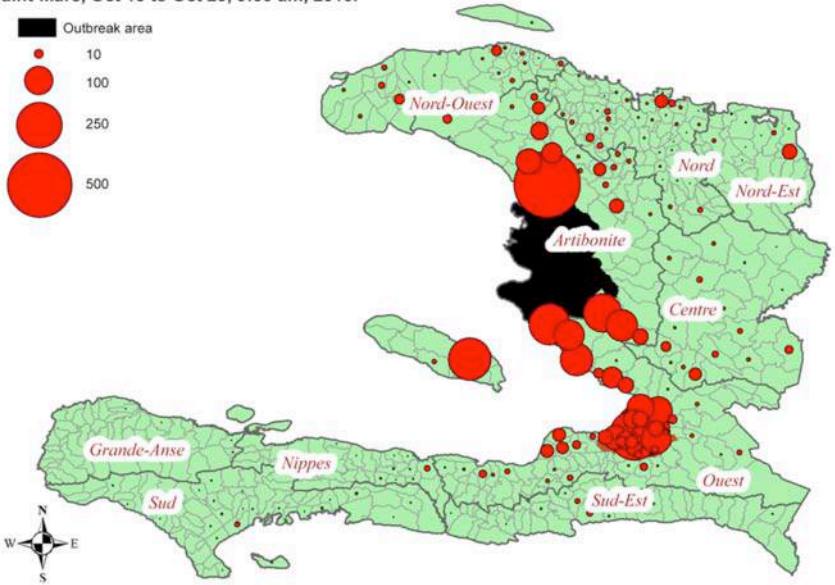
*including the Caribbean

Many early work focused on LAC

Using Mobile Phone Data to Predict the Spatial Spread of Cholera

Linus Bengtsson[1,2], Jean Gaudart[3], Xin Lu[4,1,2], Sandra Moore[5], Erik Wetter[2,6], Kankoe Sallah[3], Stanislas Rebaudet[5] & Renaud Piarroux[5]

Average daily numbers of sims that moved out from the communal sections surrounding Saint-Marc, Oct 15 to Oct 23, 9:00 am, 2010.



Movements of mobile phones out of the early outbreak area.

FLOWMINDER.ORG

FLOWMINDER.ORG Digicel

Hurricane Matthew: Estimated population movement, 24 October 2016

- The map shows the estimated distribution of people for whom their home Section Communale in the pre-hurricane period was in either Grande Anse, Sud or Nippes département, and as of 24 October had moved to another Section Communale.
- Estimates are based on movements of de-identified SIM cards which made or received at least one call pre-hurricane and on 24 October 2016.
- The SIM card movements combined with available population data derived from estimates for the year 2015[1].
- The table lists the locations with the largest number of arrivals

Location	Population pre-hurricane	Persons arrived	Ratio (%)
Port-au-Prince (Metropolitan area)	2870000	85700	3
Bourdet (Les Cayes)	71600	30100	42
Fond Rouge Daiyer (Jeremie)	27100	6950	26
Fond Rouge De Torbec (Jeremie)	26800	6790	25

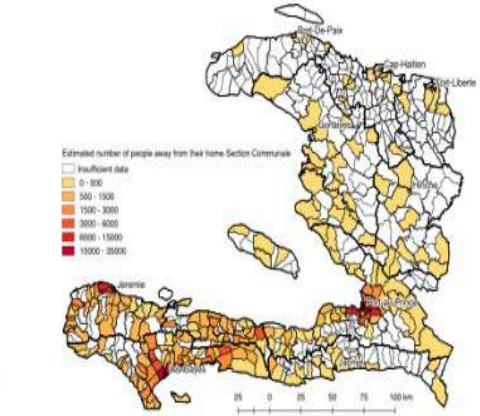
[1] https://data.humdata.org/dataset/estimated-population-of-haiti-2015

Contacts:
linus.bengtsson@flowminder.org +41 78 964 88 28
chris.brooks@flowminder.org +44 7815 944 012

Estimated population away from their home Section Communale[2]:

HOME DEPARTMENT:	GRANDE ANSE	SUD	NIPPES
POPULATION AWAY FROM HOME:	77500	132000	51000
% AWAY FROM HOME:	18%	17%	15%

24 October 2016, location of people away from their home Section Communale (out of those living pre-hurricane in Grande Anse, Sud and Nippes only)[3]



FLOWMINDER.ORG Digicel

[2] Of the people normally resident within the given Département, we estimate the total number away from their home Section Communale on the given day.
[3] Section Communales are left blank where insufficient data is available.

Flowminder.org is a non-profit organization registered in Stockholm, Sweden. Digicel is a mobile operator in Haiti

Produced 26 Oct 2016 v1.0

Many early works focused on LAC

Prediction of Socioeconomic Levels using Cell Phone Records

V. Soto, V. Frias-Martinez, J. Virseda and E. Frias-Martinez

Telefonica Research, Madrid, Spain
{vsoto,vanessa,jvirseda,efm}@tid.es

Studying Human Behavior through the Lens of Mobile Phones during Floods

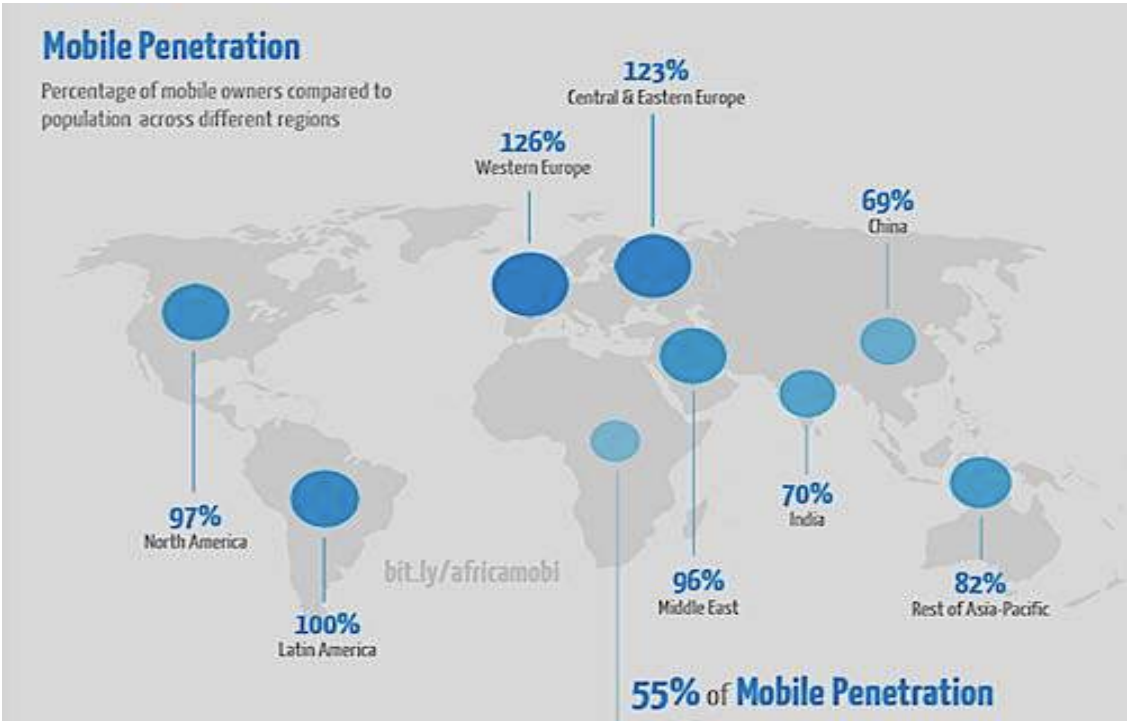
A. J. Morales[1], D. Pastor-Escuredo[1], Y. Torres[1], V. Frias-Martinez[2], E. Frias-Martinez[3], N. Oliver[3], J. M. Bauer[4], A. Wadhwa[4], A. Rutherford[5], T. Logar[5], R. Clausen-Nielsen[5], O. De Backer[5], M. A. Luengo-Oroz[5]
[1] Universidad Politecnica de Madrid, [2] University of Maryland, [3] Telefonica Research,
[4] United Nations World Food Program, [5] United Nations Global Pulse

USING FINANCIAL TRANSACTION DATA TO MEASURE ECONOMIC RESILIENCE TO NATURAL DISASTERS

PARTNERS: BBVA DATA & ANALYTICS

PROGRAMME AREA: HUMANITARIAN ACTION

Big Data's relevance in/for Latin America



FEATURE STORY

Latin America leads Global Mobile Growth

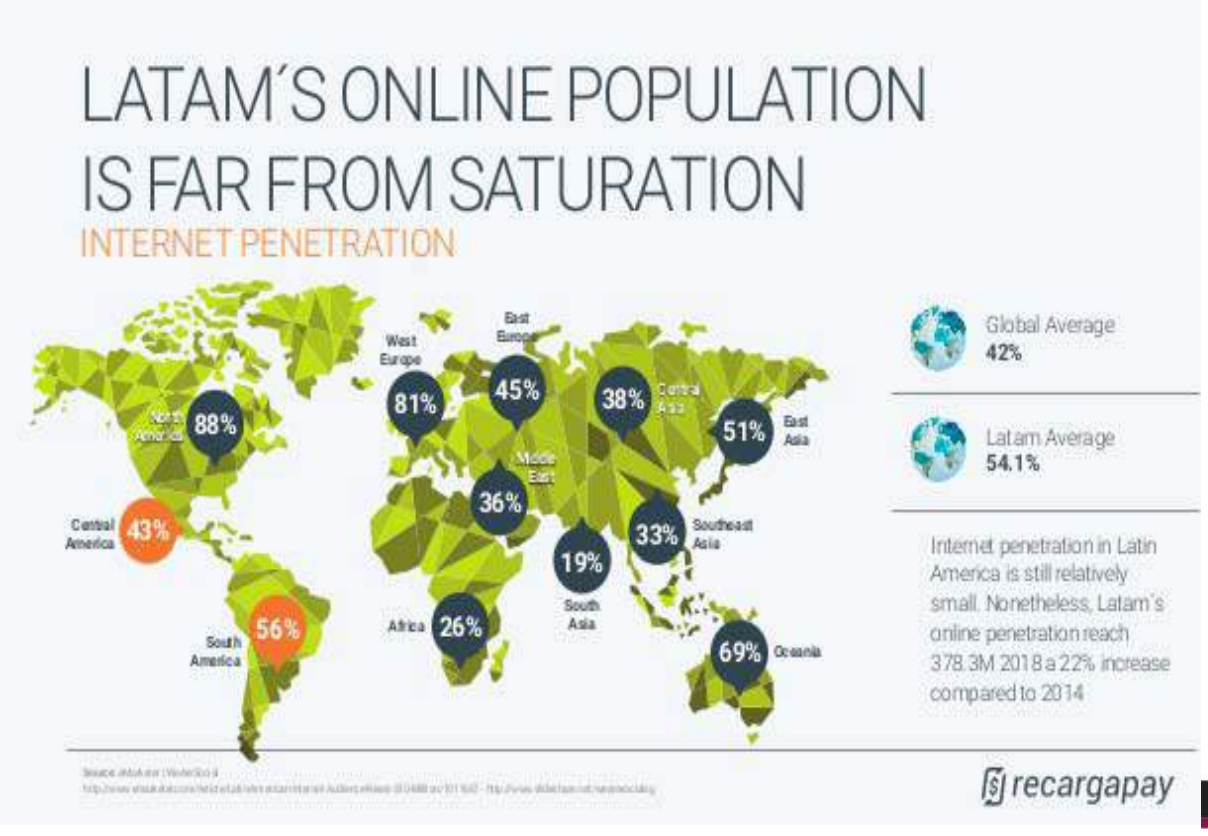
July 23, 2012

- *98% of Latin American population has mobile cell signal.*
- *In Brazil, 92% of the households have a mobile telephone.*

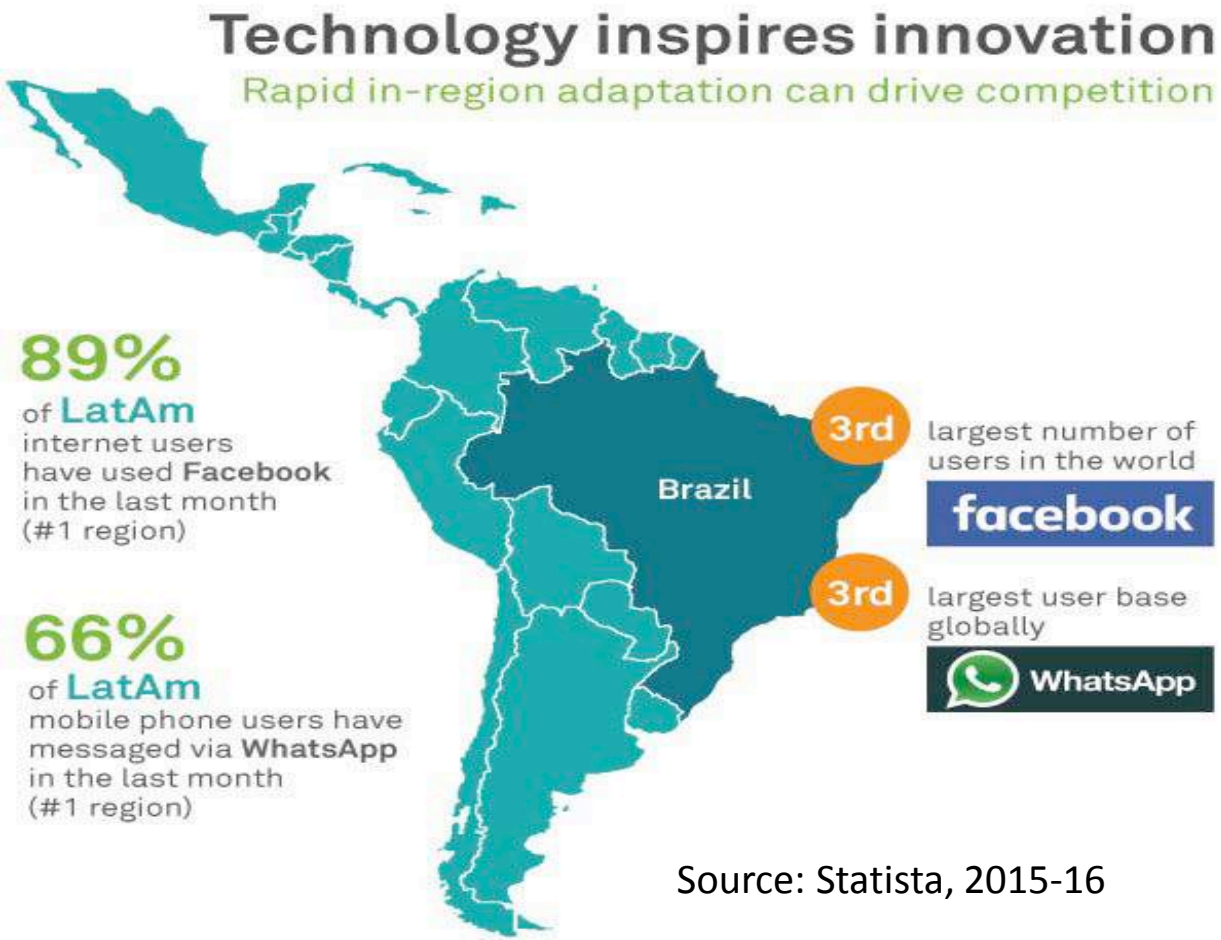
The 10 Countries With the Highest Smartphone Penetration

Country	Population with a smartphone
South Korea	88%
Australia	77%
Israel	74%
U.S.	72%
Spain	71%
U.K.	68%
Canada	67%
Chile	65%
Malaysia	65%
Germany	60%

Big Data's relevance in/for Latin America



Big Data's relevance in/for Latin America

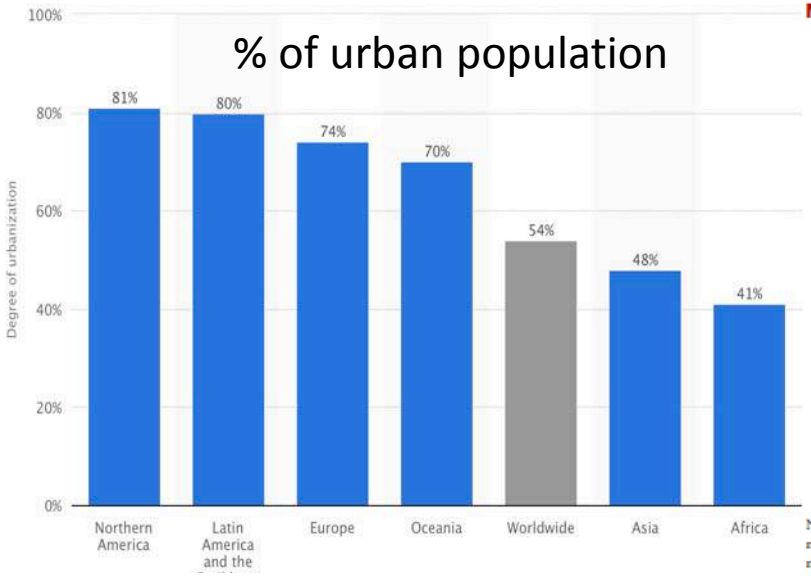


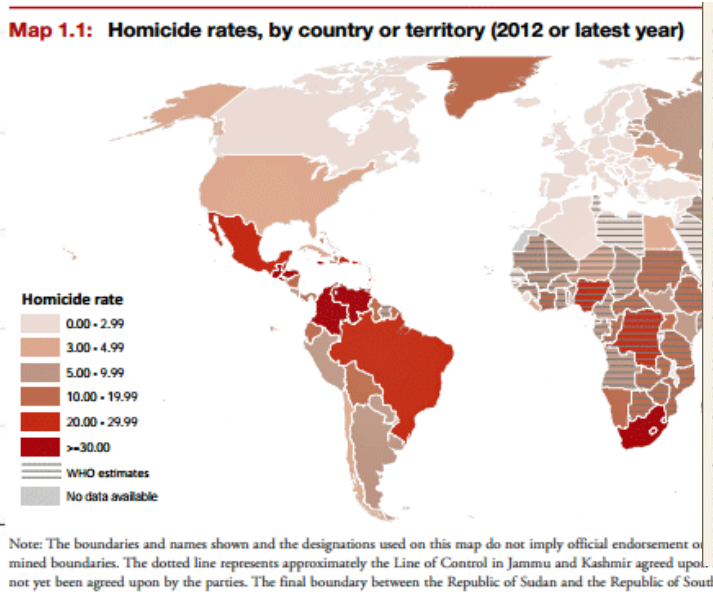
América Latina Q1-Q3 2015	A/A Crecimiento
Facebook	+102%
Twitter	+111%
Instagram	+123%
Overall	+104%

Source: Statista, 2015-16

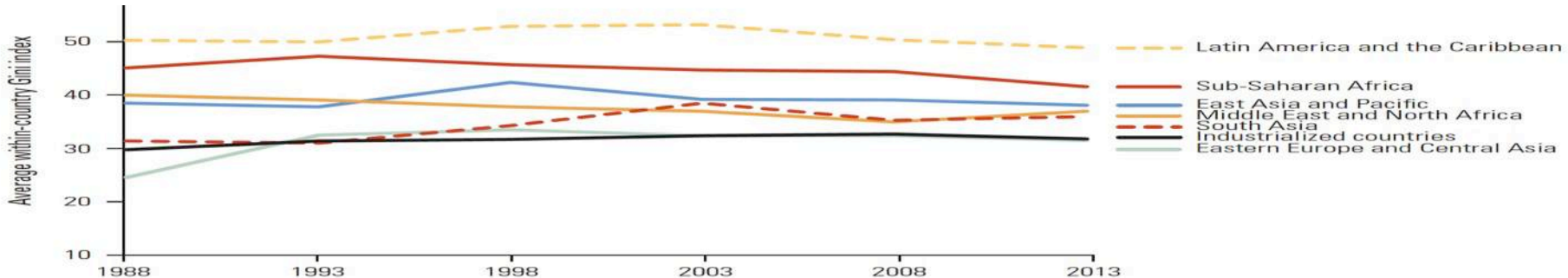
comScore Shareablee

Big Data's relevance in/for Latin America





Trends in economic inequality by world region 1988-2013



Source: The World Bank (2016) – Poverty and Shared Prosperity 2016: Taking on Inequality. Reformatted for OurWorldInData.org
Original data source: World Bank calculations based on data in Milanovic 2014; PovcalNet (online analysis tool).

Big Data's relevance in/for Latin America*

*NON EXCLUSIVE/FOR ILLUSTRATION PURPOSES…

Big Data's relevance in/for Latin America

➔ Despite wide variations, LAC countries share a few features:

1. A **urban, relatively young, innovative and technology hungry population** that share 4 major languages (Sp., Port., Fr., Eng.)
2. **Region specific socio-political economic features**: urbanization, inequality, violence, democracy, overall HD / GDP growth…
3. A long experience of the **Open Data movement + activism**
4. The presence of strong **region-wide civil-society groups, international institutions, academic networks, private corporations**
5. The active involvement of **governments and public actors** that devise new **strategies, policies, pilots** and participate in **global forums**

Ex: Big Data, SDGs & Official Stats in LAC

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.

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]

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 socioeconmic 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.


3. Our Vision, Plans and Featured Projects to Harness Big Data in Latin America

Our Global Vision & Strategy Started in LAC

DATA POP ALLIANCE

HARVARD HUMANITARIAN INITIATIVE

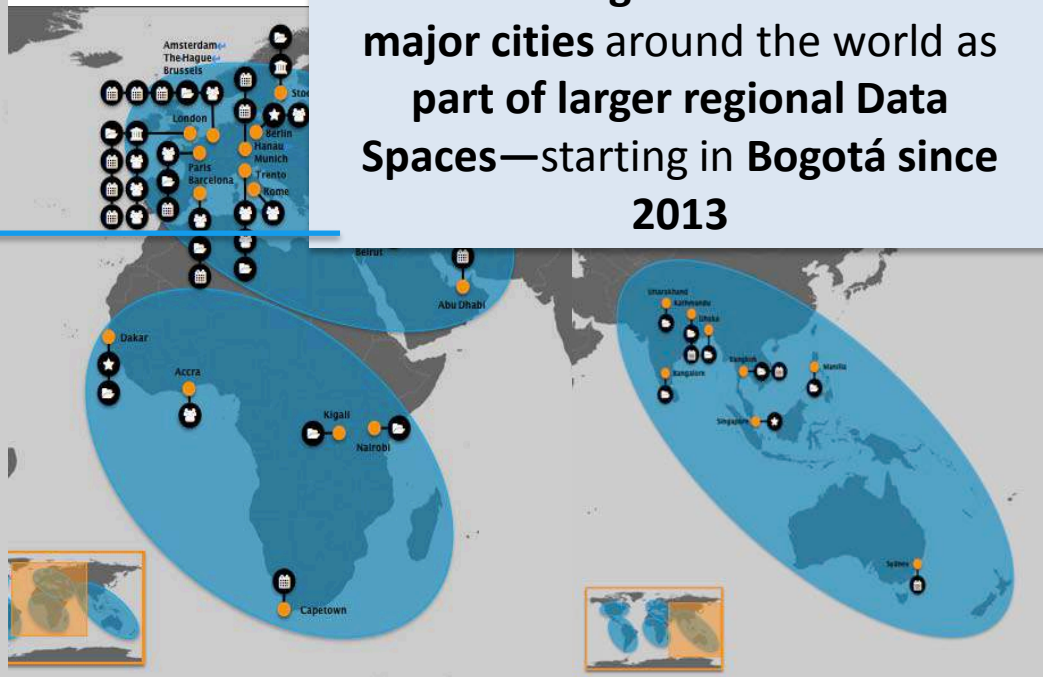
mit media lab

FLOWMINDER.ORG

MIT Connection Science
the technology of innovation

FONDAZIONE BRUNO KESSLER

Data-Pop Alliance and MIT Connection Science with key partners are working to develop and deploy connected research, training and policy activities **in Urban Living Labs anchored in major cities** around the world as **part of larger regional Data Spaces**—starting in **Bogotá since 2013**



Data Space LAC:

Locally Co-Design and Connect Coherent Research, Training and Engagement Projects and Partnerships to Leverage Big Data for Development and Democracy

From Bogotá-Colombia to Latin America

60th Anniversary of the National Department of Statistics of Colombia

Official Statistics in the Big Data Era

Emmanuel Letouzé
PhD Candidate, UC Berkeley
Fellow, Harvard Humanitarian Initiative
eletouze@Berkeley.edu

Biblioteca Luis Angel Arango
Bogotá October 28th, 2013

Data-Pop Allliance is hired to help Colombia's National Statistical Office (DANE) to develop its **Big Data Strategy**

Delivery of **Big Data strategy** analysis to DANE

Data-Pop Alliance gets **$230k from the World Bank** to develop a research program in Colombia

Cartagena Data Festival, April 20-22

Data-Pop Alliance gets **$500k from Hewlett Foundation** to develop professional training workshops in Senegal, Rwanda, Kenya and Colombia

DATA-POP ALLIANCE Report | **Opportunities and Requirements for Leveraging Big Data for Official Statistics in Latin America**

May 2016

March 15th 2016, Kick-off of **Ciudad Laboratorio** in Bogotá

Development of **Colombia's National Big Data strategy**

Regional **Research on Citizen Security with AFD, IDB…**

Agreements with

2013 | 2014 | 2015 | 2016 | 2017

Featured Project #1: *Laboratorio Urbano*

Networking

Citizen engagement

Use of Big and Open Data

Cámara de Comercio de Bogotá

DATA POP ALLIANCE

HARVARD HUMANITARIAN INITIATIVE

mit media lab

FLOWMINDER.ORG

ALCALDÍA MAYOR DE BOGOTÁ D.C.

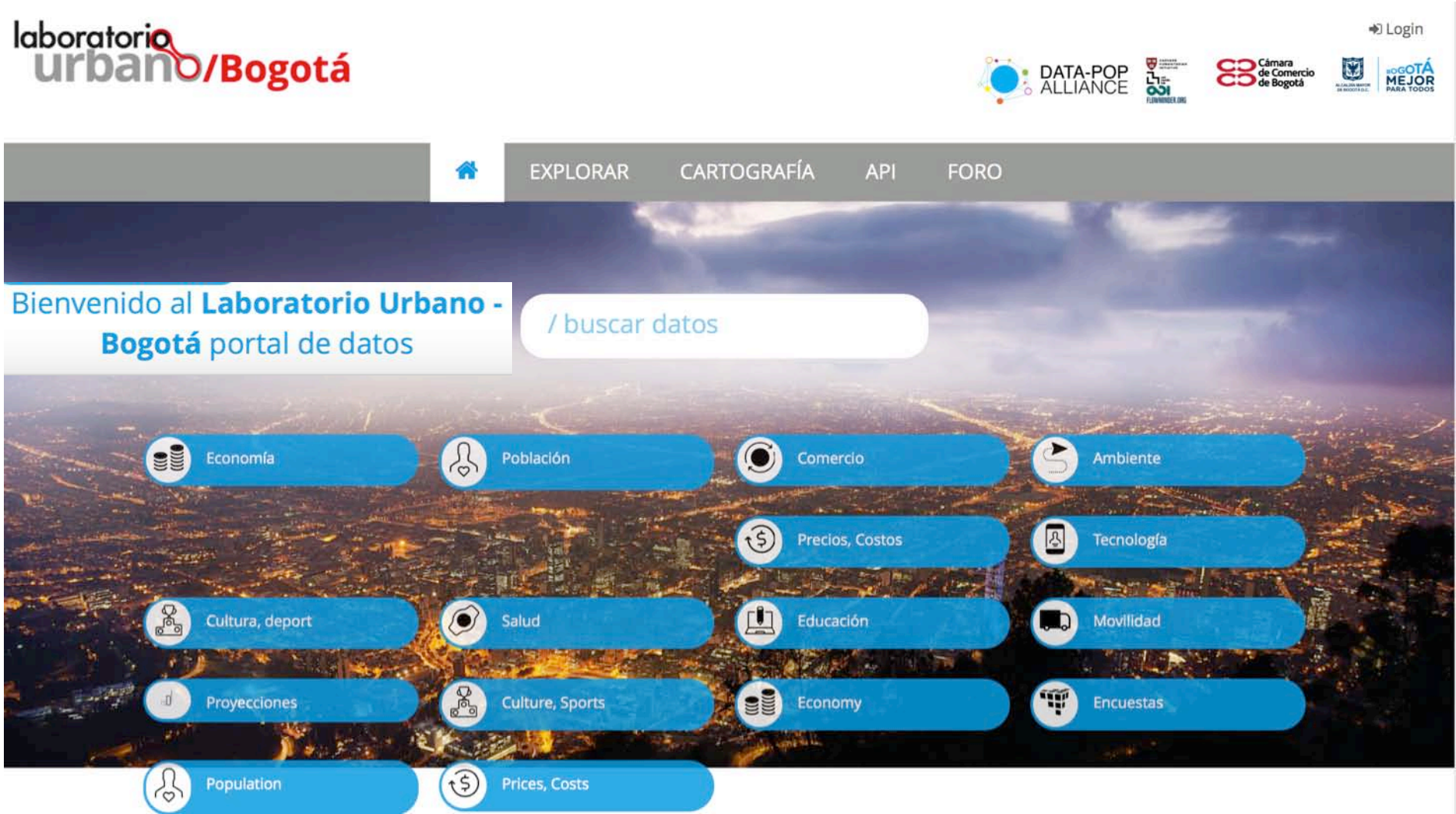
BOGOTÁ MEJOR PARA TODOS

SECRETARÍA DE DESARROLLO ECONÓMICO

MIT Connection Science

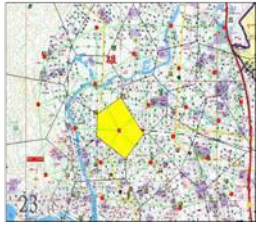
the technology of innovation

Featured Project #1: *Laboratorio Urbano*



Featured Project #2: Research and Pilots on Citizen Security & Social Cohesion

CALLER ID	CALLER CELL TOWER LOCATION	RECIPIENT PHONE NUMBER	RECIPIENT CELL TOWER LOCATION	CALL TIME	CALL DURATION
X76VG588RLPQ	2°24' 22.14", 35°49' 56.54"	A81UTC93KK52	3°26' 30.47", 31°12' 18.01"	2013-11-07T15:15:00	01:12:02



What makes a city vital and safe: Bogotá case study

Marco De Nadai
FBK and University of Trento

Andrés Clavijo
Data–Pop Alliance

Rodrigo Lara Molina
Data–Pop Alliance

Bruno Lepri *
FBK

Emmanuel Letouzé †
Data–Pop Alliance

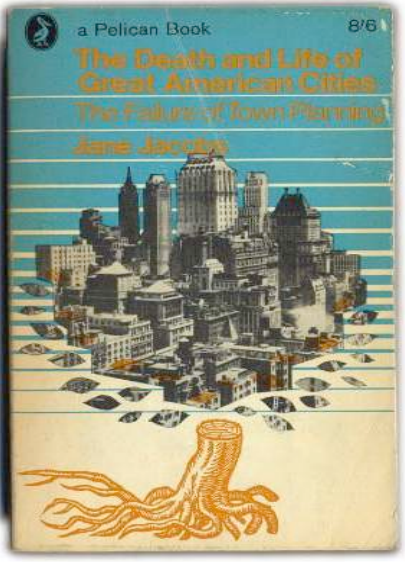
Nuria Oliver ‡
Telefónica Research

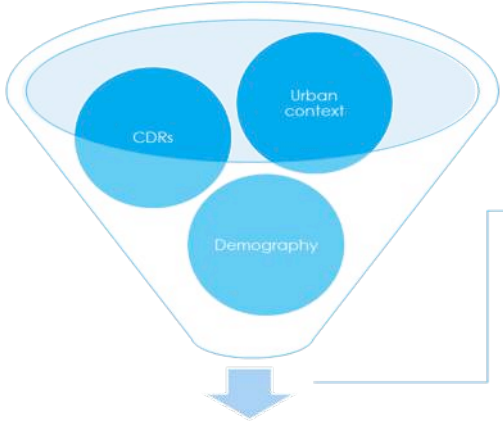
Gabriel Pestre
Data–Pop Alliance

Joan Serrà
Telefónica Research

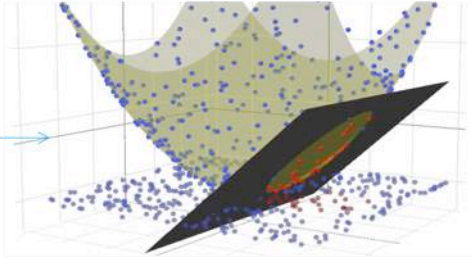
Natalie Shoup
Data–Pop Alliance

Alvaro Ramirez Suarez
Telefónica Colombia





Criminality?



WORLD BANK GROUP

AFD

Telefónica

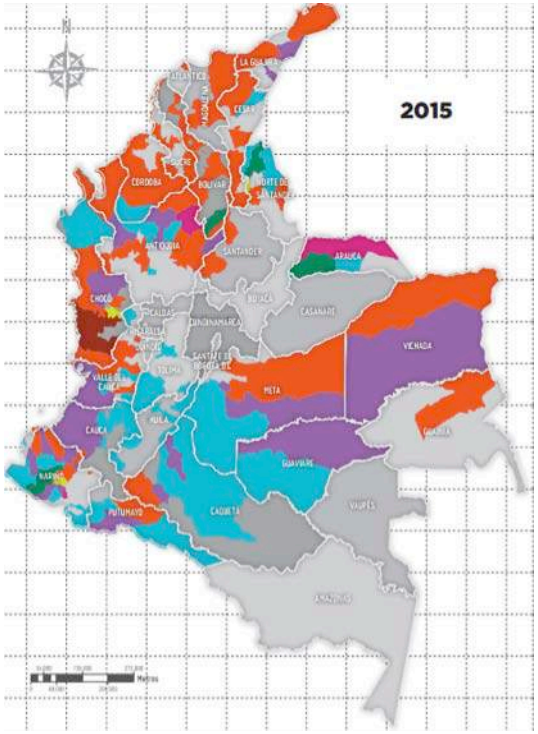
FONDAZIONE BRUNO KESSLER

Featured Project #2: Research and Pilots on Citizen Security & Social Cohesion

Next: **Regional program starting in Colombia**

Could CDRs and social network data help understand and improve:

- The level of socioeconomic vitality and resilience of selected municipalities (Bogotá, Cartagena, Medellin, Santa Marta, Pasto…) in Colombia?
- The trajectories of the people displaced by the conflict and improve their reinsertions into the labor market?
- People's perception and positions on the current peace process?



FIP FUNDACIÓN IDEAS PARA LA PAZ

mit media lab

FONDAZIONE BRUNO KESSLER

BID Banco Interamericano de Desarrollo

AFD

Featured project #3: Open Algorithms (OPAL) project

OPAL

bring the code to the data

DATA POP ALLIANCE

Imperial College London

mit media lab

orange™

WORLD ECONOMIC FORUM

Featured project #3: Open Algorithms (OPAL) project

Certificación de algoritmos

CODE: Comité de Orientación para el Desarrollo et la Ética

Cohesión social

Densidad poblacional

Proxy de pobreza

Precios

Solo se exportan resultados seguros y acordados

REDES SOCIALES

TELCO LOCAL

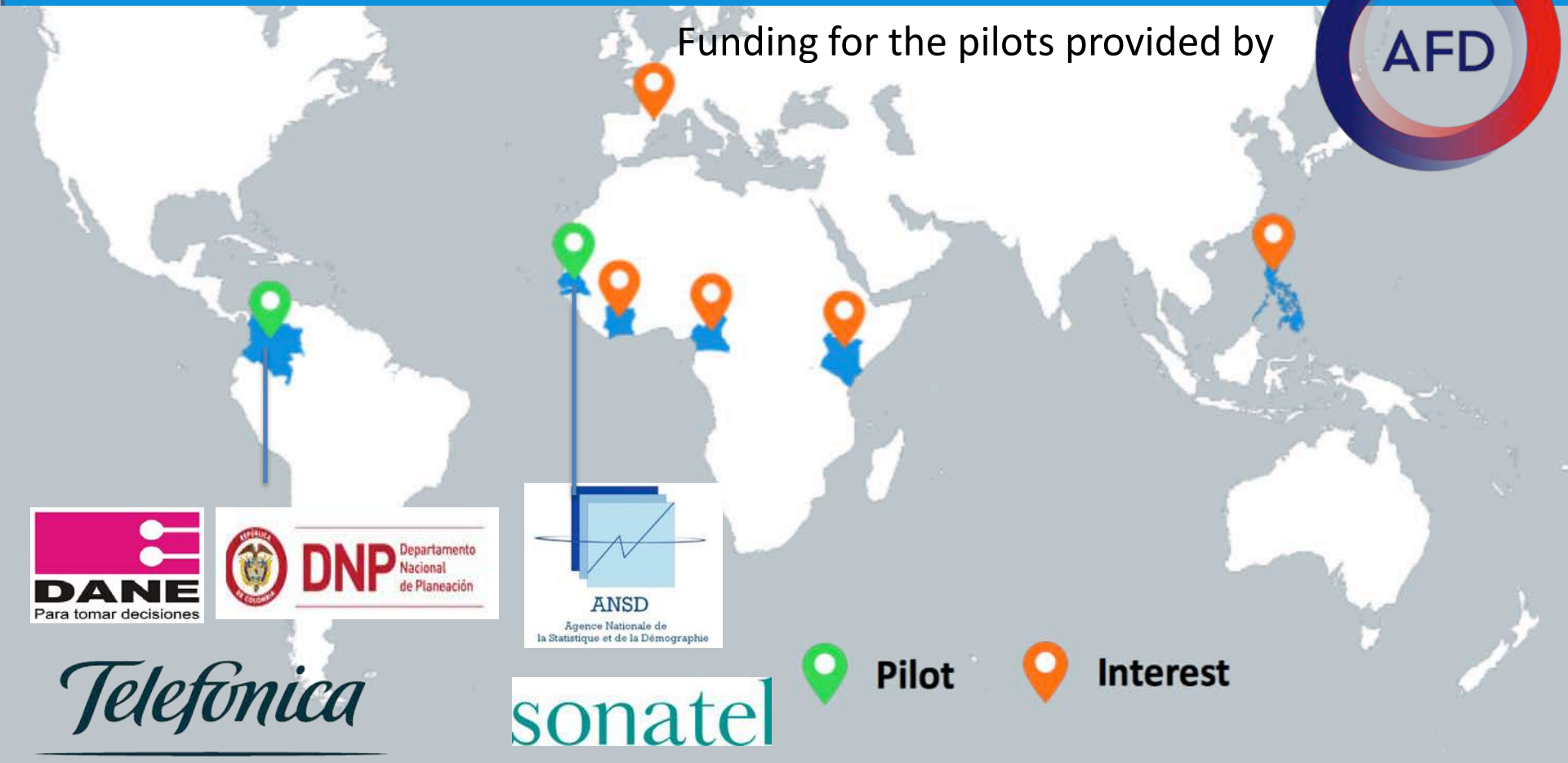
COMERCIO LOCAL

La plataforma accede a los *DATOS* que permanecen detrás del *FIREWALL* de los provedores

Iconos por Sergey Novosyolov,, Takao Umehara, aguycalledgary, Gregor Cresnar

Featured project #3: Open Algorithms (OPAL) project

Funding for the pilots provided by AFD



Featured project #4: Developing Big Data Capacities & Literacy for the SDGS in LAC

Professional training workshops on Big Data and development held in Bogotá, MIT and NYC with and for Colombian partners over 2016-17

+

A curated knowledge platform in English and Spanish (and French)

Main funders/partners

THE WILLIAM AND FLORA HEWLETT FOUNDATION

Selected technical partners

PARIS21

BID Banco Interamericano de Desarrollo

SCHOOL OF DATA

cetic.br

Universidad de los Andes

UNITED NATIONS SYSTEM STAFF COLLEGE

UNITED NATIONS ECLAC

Selected target audience

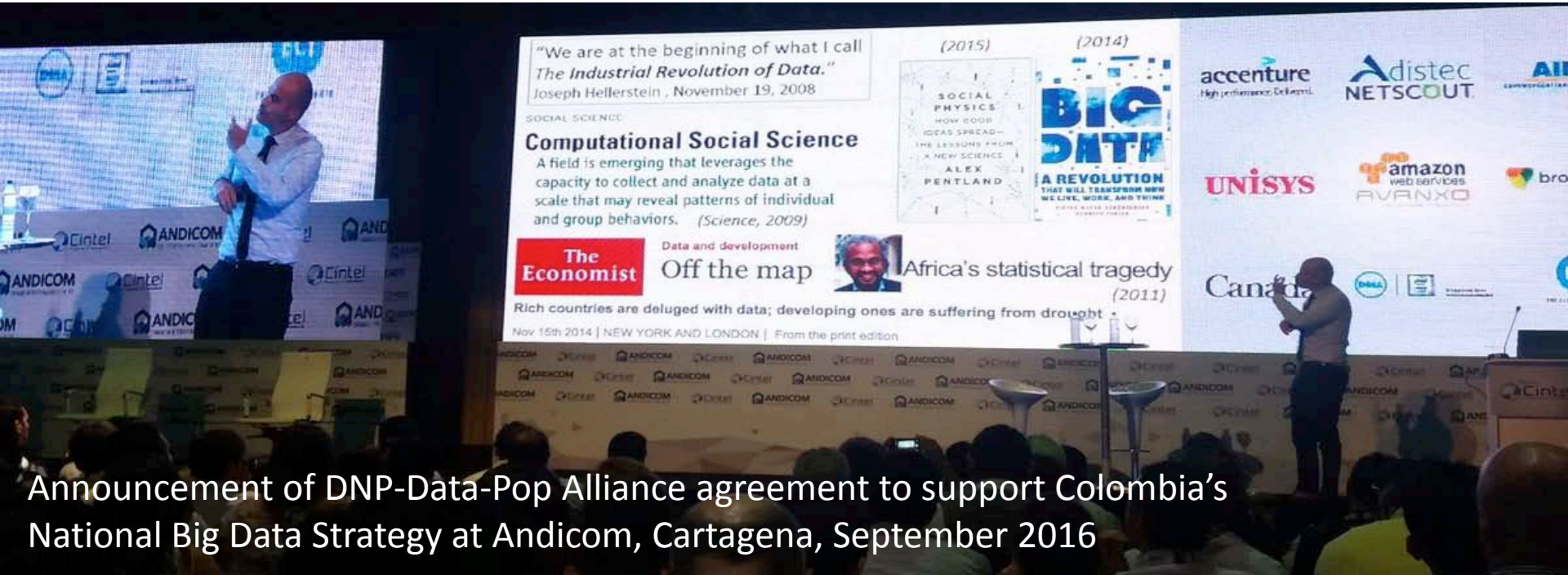
- UN Officials
- Gov officials
- Official Statisticians
- Journalists
- NGO/CSO Staff
- Academics/

Featured project #5: Support to Colombia's National Big Data strategy

DNP Departamento Nacional de Planeación

TODOS POR UN NUEVO PAÍS
PAZ EQUIDAD EDUCACIÓN

mit media lab



Announcement of DNP-Data-Pop Alliance agreement to support Colombia's National Big Data Strategy at Andicom, Cartagena, September 2016

My 3 main messages to "Think Big" when "Harnessing Big Data in Latin America"

1. **Think crafting Big Data** (ecosystem) not harnessing big data (resource): craft incentives, investigations, regulations, capacities, partnerships, trust…
2. **Think transformation, purpose**: what isn't right? What do you / I want to change fundamentally? (hint: governance)
3. **Think strategically in a temporal and geographic matrix**
    1. long term → medium → short-term → medium→ long term…
    2. regional → national → local→ national→ regional…

Thank you

For more information
www.datapopalliance.org

Contact for feedback
and expressions of interest to work with us:

eletouze@datapopalliance
aclavijo@datapopalliance.org

Follow us on Twitter & Facebook
@datapopalliance
www.facebook.com/datapopalliance

Hosted at ThoughtWorks NYC
99 Madison Avenue, 15th Floor
New York, NY 10016