Big data & development

The whats, whos & whys

SIMONE SALA - @HEREISSIMONE - WWW.SIMONESALA.IT

* ASSOCIATE DIRECTOR OF THE DR. STEVE CHAN CENTER FOR SENSEMAKING, AIRS (HAWAI'I PACIFIC UNIVERSITY AND SWANSEA UNIVERSITY)
* RESEARCH AFFILIATE, DATAPOP ALLIANCE
* FELLOW, UNIVERSITY OF MILAN GEOLAB

ICTP – Trieste – March 26, 2015

Big data & development

- What?
- Who & Where?
- Why?

What is big data (for development)?

- Many definitions: is there a right one?
- What is big data (as of today)?
- What is big data in/for development?

What is big data?

Definitions

- Is there a single definition of big data?

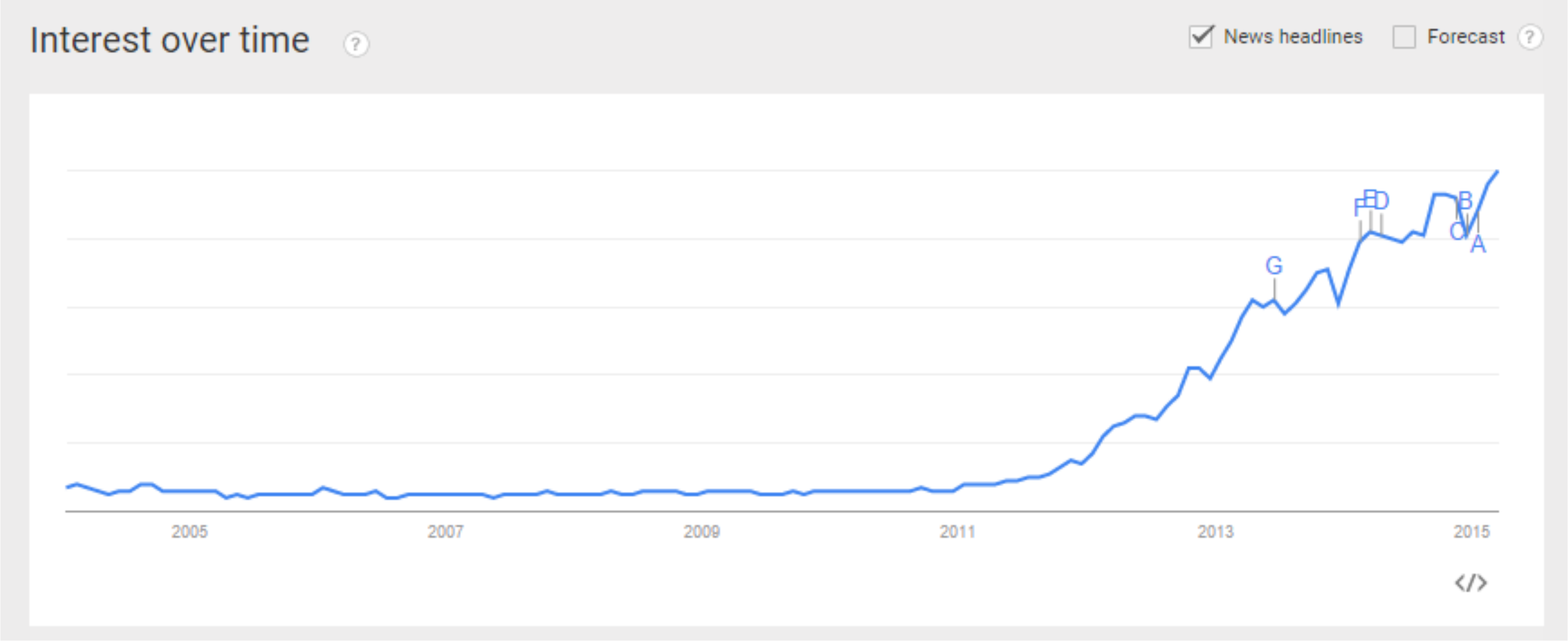
What is big data?

Definitions

- Is there a single definition of big data?
- No single agreed definition of big data.
- Though a lot of people is talking about big data… isn’t it?

What is big data?
Definitions

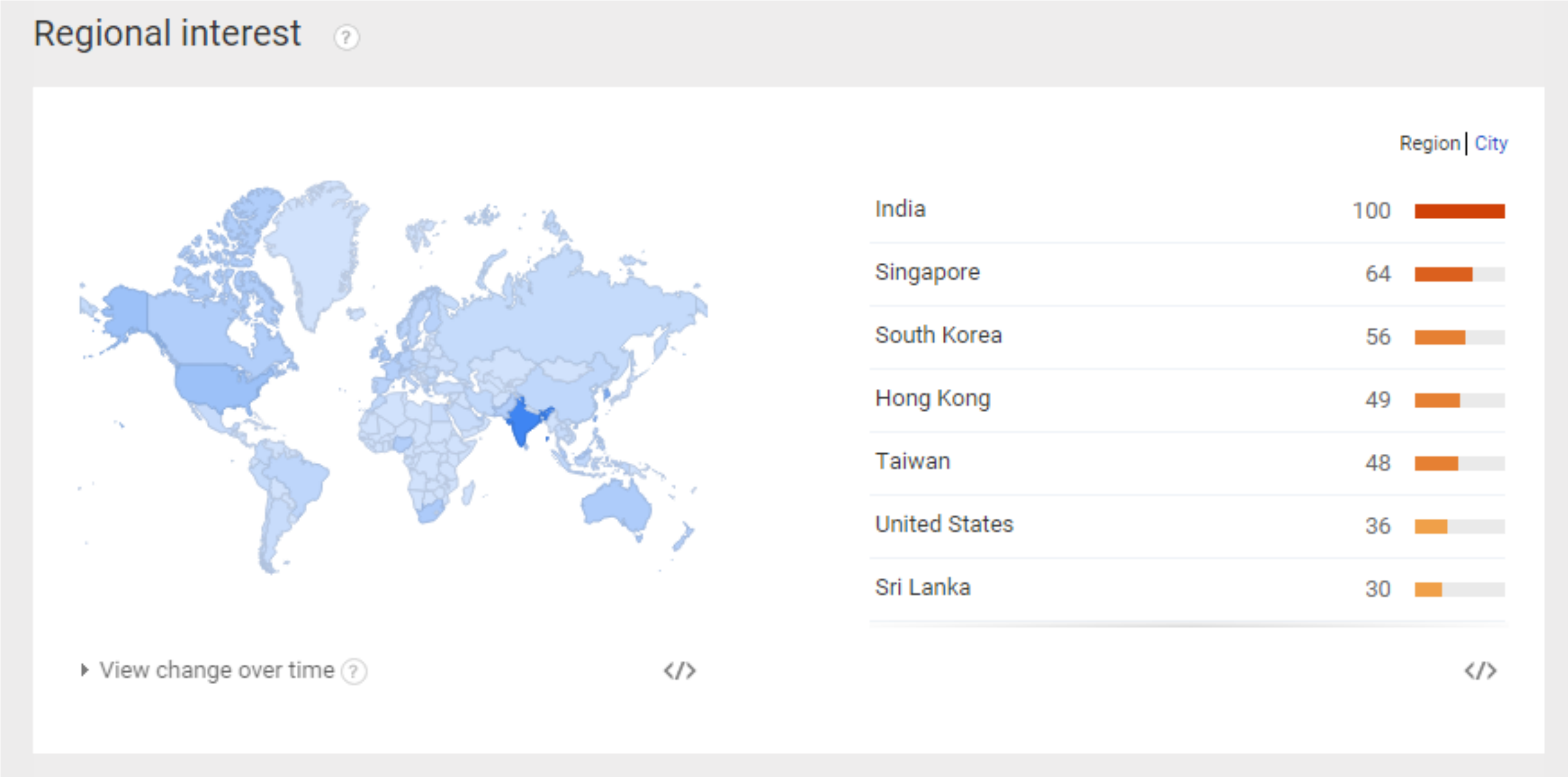
Let’s see what data is saying about big data!



Source: http://www.google.com/trends/explore#q=big%20data

What is big data?
Definitions

Let’s see what data is saying about big data!



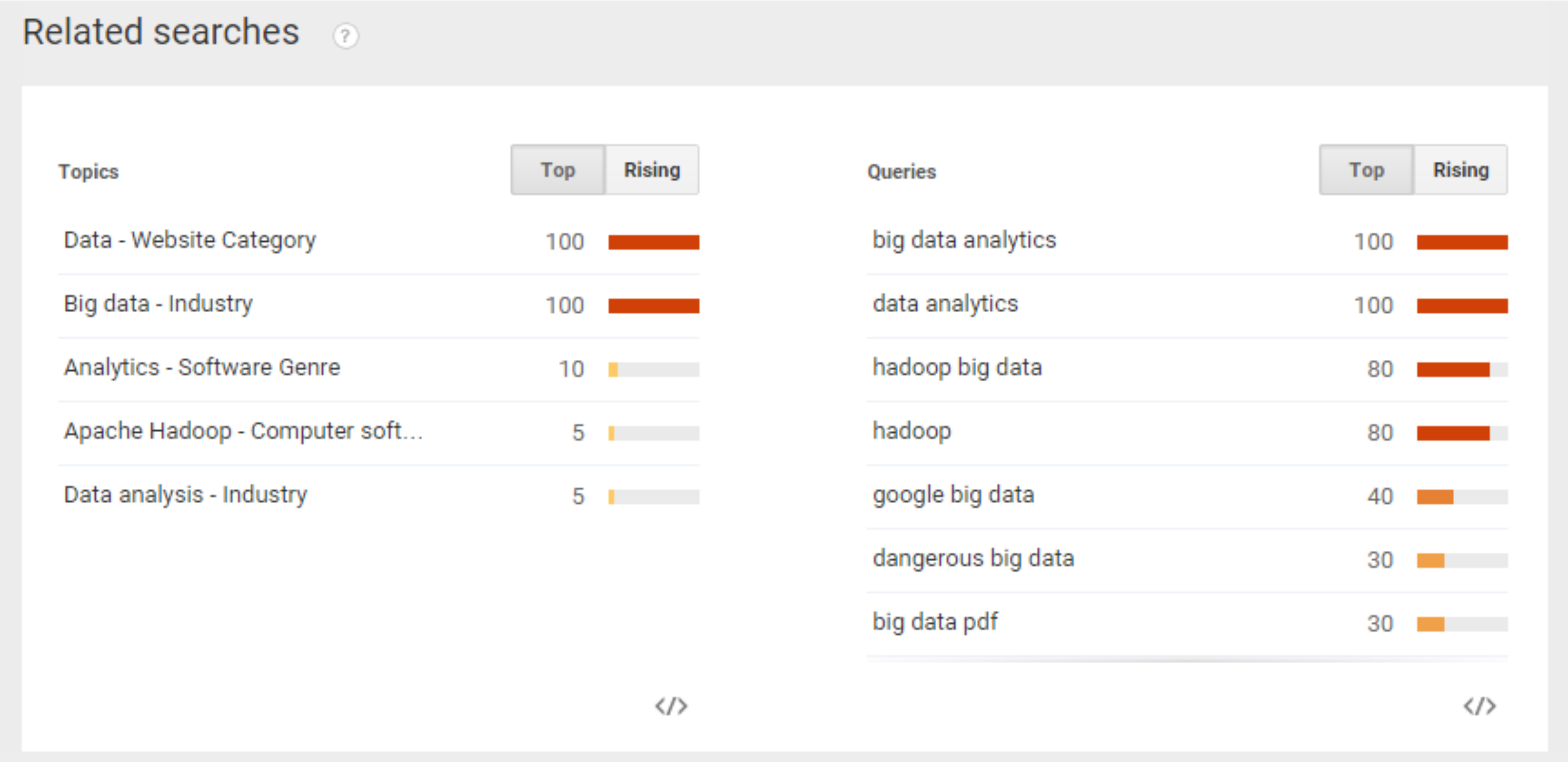
India	100	🟥
Singapore	64	🟥
South Korea	56	🟥
Hong Kong	49	🟥
Taiwan	48	🟥
United States	36	🟥
Sri Lanka	30	🟥

Source: http://www.google.com/trends/explore#q=big%20data

What is big data?

Definitions

Let’s see what data is saying about big data!



Related searches

Topics	Top	Rising
Data - Website Category	100	
Big data - Industry	100	
Analytics - Software Genre	10	
Apache Hadoop - Computer soft...	5	
Data analysis - Industry	5	

Queries	Top	Rising
big data analytics	100	
data analytics	100	
hadoop big data	80	
hadoop	80	
google big data	40	
dangerous big data	30	
big data pdf	30	

Source: http://www.google.com/trends/explore#q=big%20data

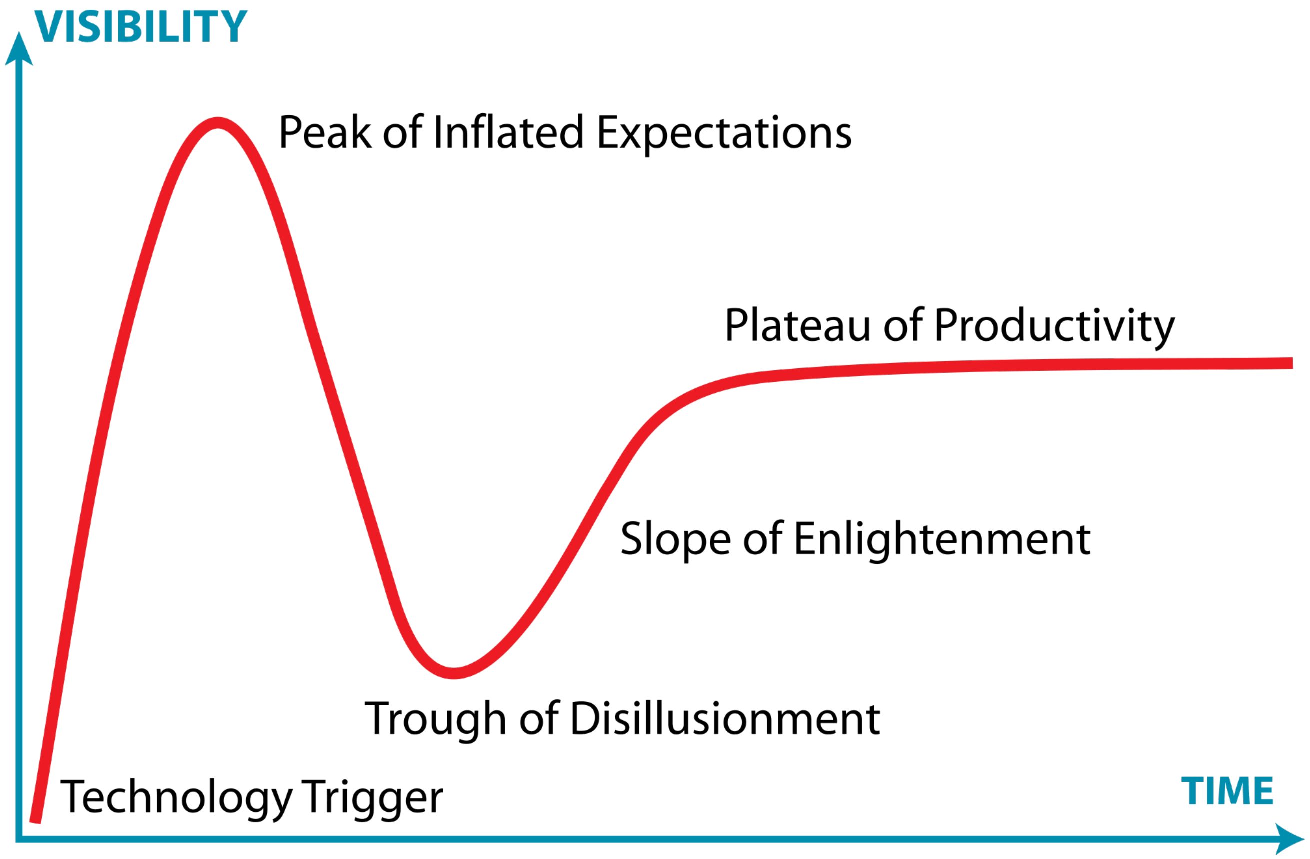
What is big data?

Definitions

- Is there a single definition of big data?
- No single agreed definition of big data.
- Though a lot of people is talking about big data!
- So is that just hype?

What is big data?

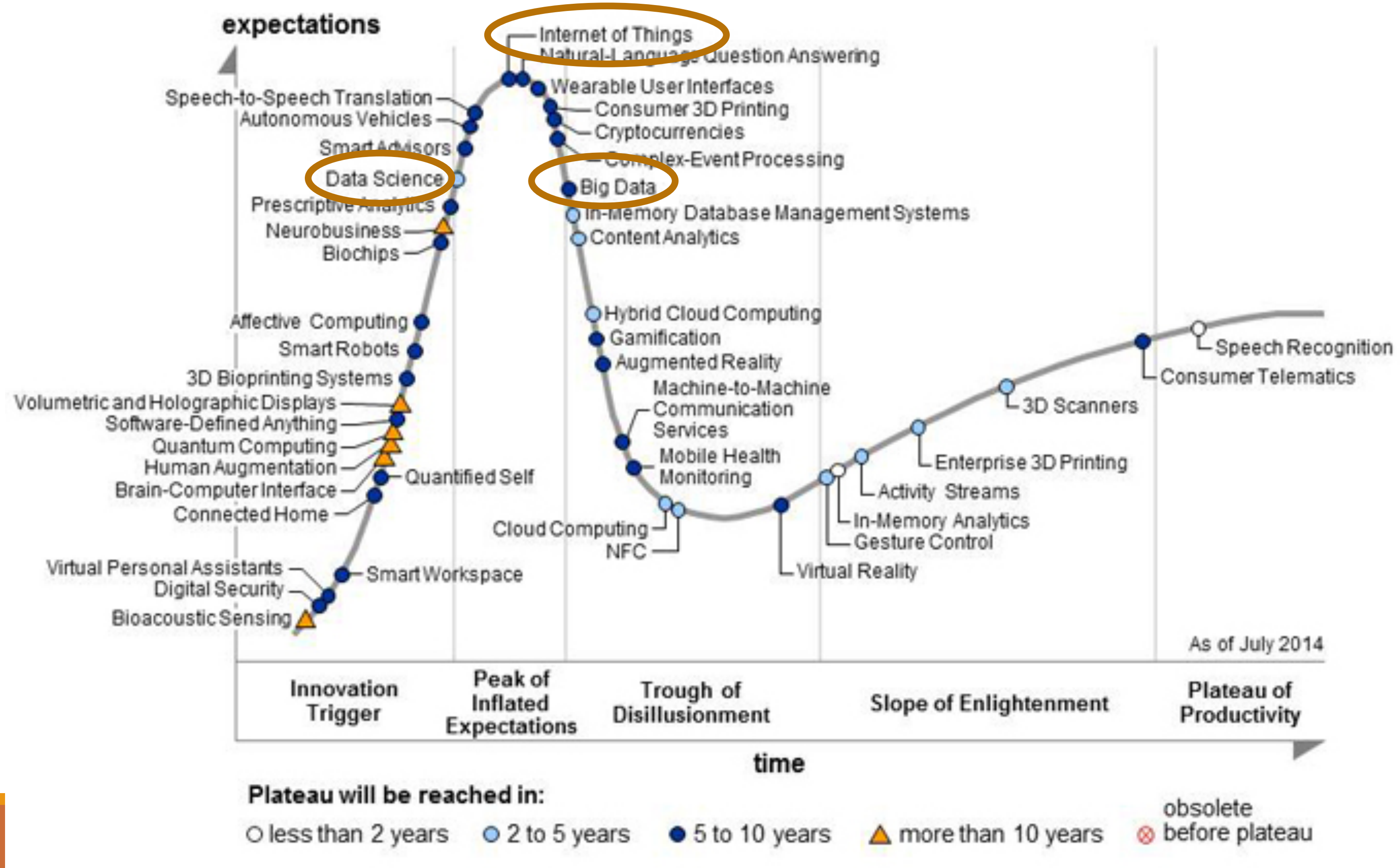
Definitions & Hype



http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

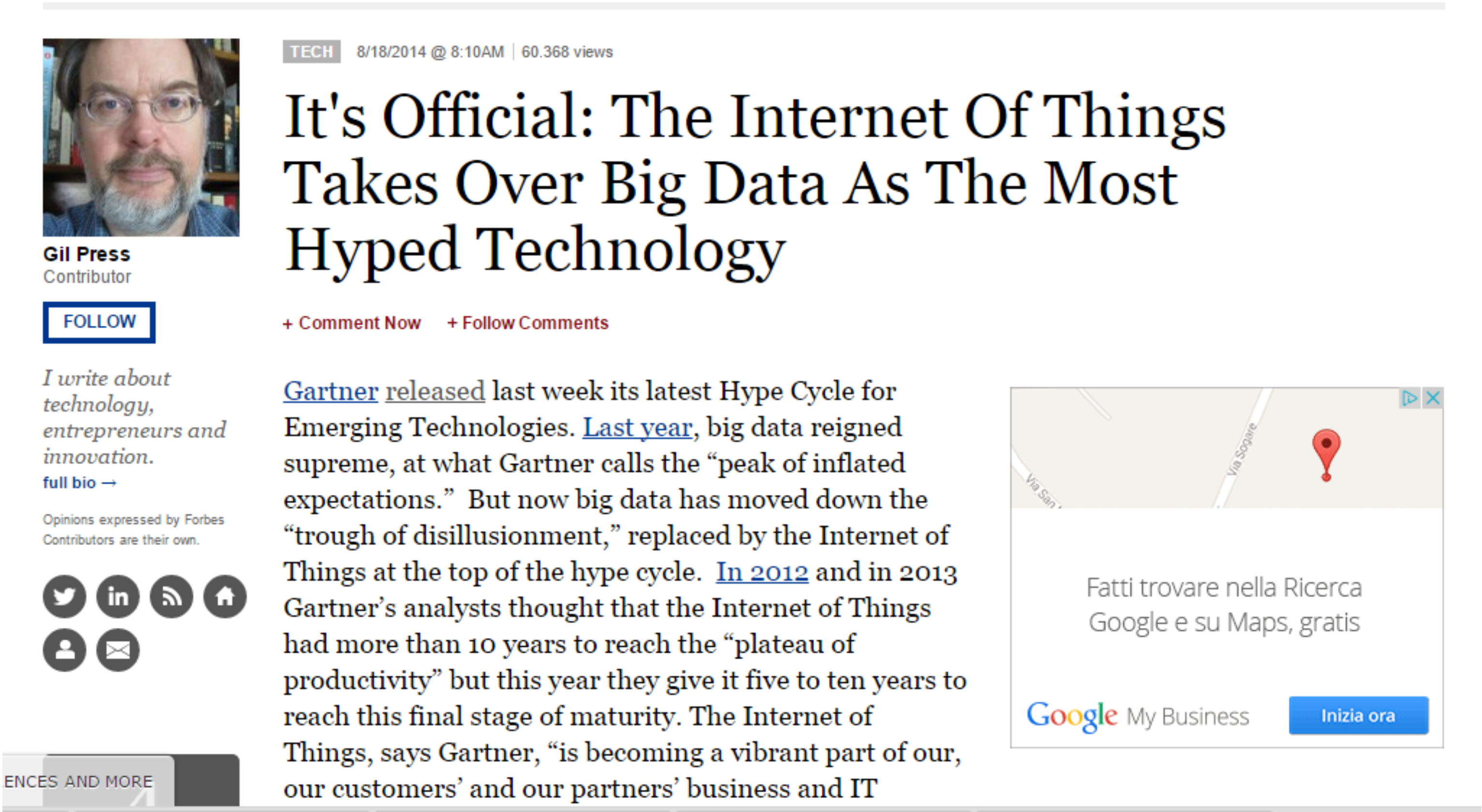
What is big data?

Definitions & Hype



What is big data?

Definitions & Hype



TECH 8/18/2014 @ 8:10AM | 60.368 views

It's Official: The Internet Of Things Takes Over Big Data As The Most Hyped Technology

+ Comment Now + Follow Comments

Gil Press
Contributor

FOLLOW

I write about technology, entrepreneurs and innovation.

full bio →

Opinions expressed by Forbes Contributors are their own.

Gartner released last week its latest Hype Cycle for Emerging Technologies. Last year, big data reigned supreme, at what Gartner calls the "peak of inflated expectations." But now big data has moved down the "trough of disillusionment," replaced by the Internet of Things at the top of the hype cycle. In 2012 and in 2013 Gartner's analysts thought that the Internet of Things had more than 10 years to reach the "plateau of productivity" but this year they give it five to ten years to reach this final stage of maturity. The Internet of Things, says Gartner, "is becoming a vibrant part of our, our customers' and our partners' business and IT

ENCES AND MORE

What is big data?

Back to definitions

- Is there a single definition of big data? There is no single agreed definition of big data.
- Let’s take a look at a few of them…
- “Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.” – Wikipedia

What is big data?

Definitions

- Is there a single definition of big data? There is no single agreed definition of big data.
- Let’s take a look at a few of them…

- “Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.”
  - Snijders, C.; Matzat, U.; Reips, U.-D. (2012)

- “Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.”
  - Ibrahim Abaker Targio Hashem et al. (2015)

What is big data?

Definitions

- Size matters?

- When is data big enough to talk about big data?

What is big data?

Definitions

- Size matters? When is data big enough to talk about big data?
- A.k.a. is this big data?

What is big data?

Definitions

- Size matters? When is data big enough to talk about big data?
- The big thing about big data is that we have new data coming from a variety of machines/ people through electronic sources into databases for no statistical inference purposes
- It’s not (only) about size

What is big data in/for development?

- Is this a brand new story?

What is big data in/for development?

- Is this a brand new story?

In short – nope.

CyberSyn, Chile - 1971

What is big data and development?

- Is this a brand new story?
In short – nope.

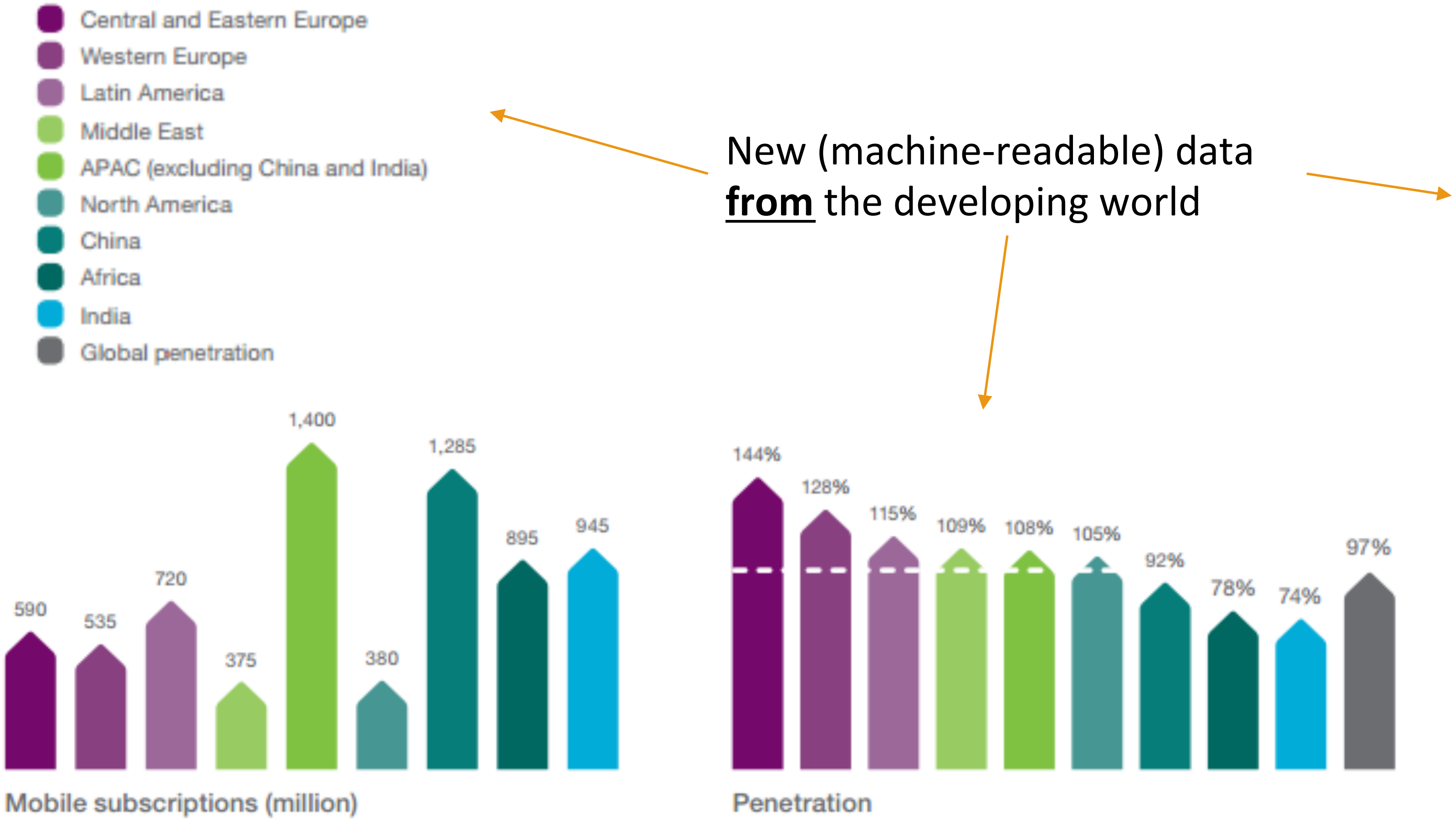
- So what is different?

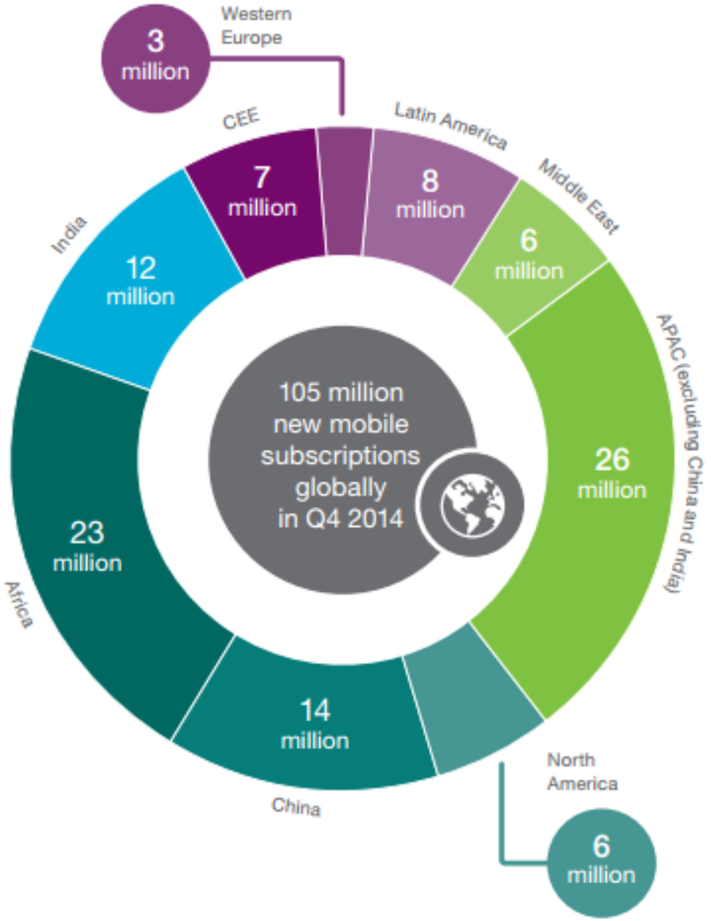
What is big data in/for development?

- Many reasons why *big data* is different and matters for development and in developing regions
- Thanks to social media & mobiles we have **new (machine-readable) data**, generated about and by people that was compeletely unavailable 10 years ago

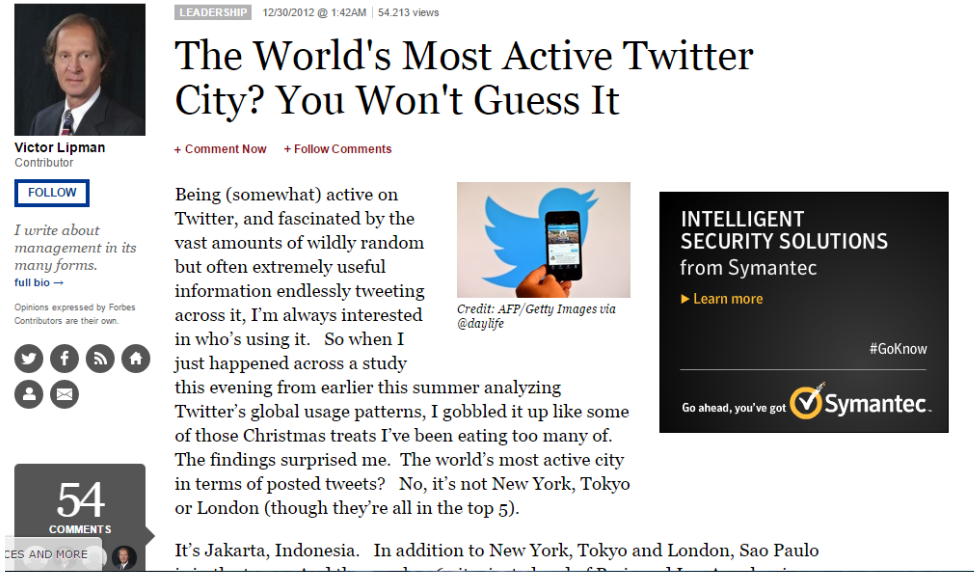
What is big data in/for development?

New (machine-readable) data **<u>from</u>** the developing world





What is big data in/for development?



LEADERSHIP 12/30/2012 @ 1:42AM | 54.213 views

The World's Most Active Twitter City? You Won't Guess It

Victor Lipman
Contributor

FOLLOW

I write about management in its many forms.
full bio →

Opinions expressed by Forbes Contributors are their own.

+ Comment Now + Follow Comments

Being (somewhat) active on Twitter, and fascinated by the vast amounts of wildly random but often extremely useful information endlessly tweeting across it, I'm always interested in who's using it. So when I just happened across a study this evening from earlier this summer analyzing Twitter's global usage patterns, I gobbled it up like some of those Christmas treats I've been eating too many of. The findings surprised me. The world's most active city in terms of posted tweets? No, it's not New York, Tokyo or London (though they're all in the top 5).

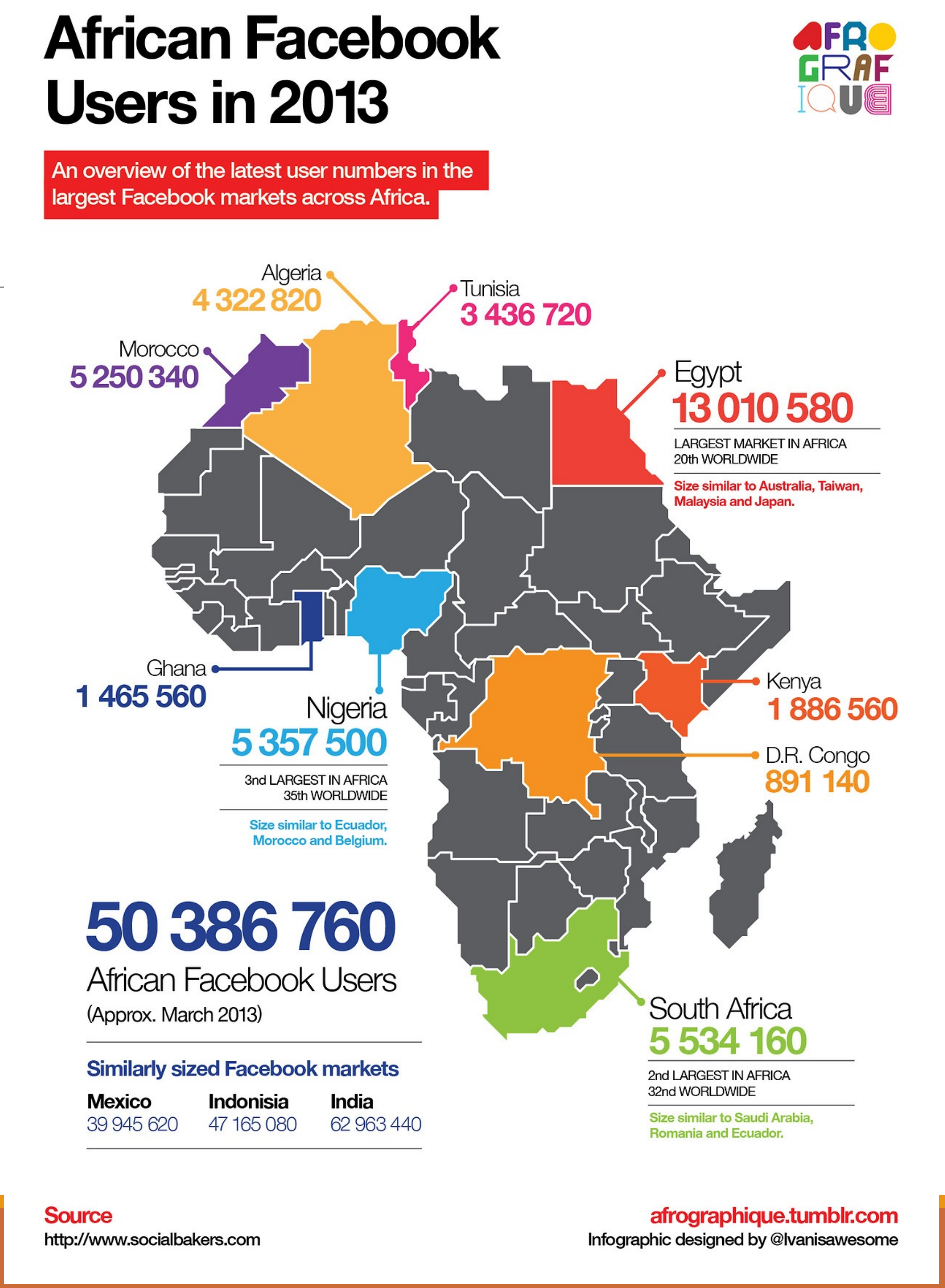
Credit: AFP/Getty Images via @daylife

It's Jakarta, Indonesia. In addition to New York, Tokyo and London, Sao Paulo

54 COMMENTS

New (machine-readable) data **from** the developing world

"Egypt, Russia, the Philippines and 14 other DCs outpace the U.S. in the proportion of Internet users who log on to social sites." [PEW Research Center]



What is big data in/for development?

- Many reasons why *big data* is different and matters for development and in developing regions
- Thanks to social media & mobiles we have **new (machine-readable) data**, generated about and by people that was compeletely unavailable 10 years ago
- It's the analytics beyond data: we maximized our power to collect + manage/analyze + transmit data in the past decade, and we will further do it
- The ~~medium~~ community is the message: open data & big data communities behind deployments
- E.g. Call Data Record vs World Bank database of indicators (again – size is not the key attribute!)

What is big data in/for development?

- It is not necessarily a technology-push that enables data-driven innovations in/for development…

What is big data in/for development?
The Toilet Gap

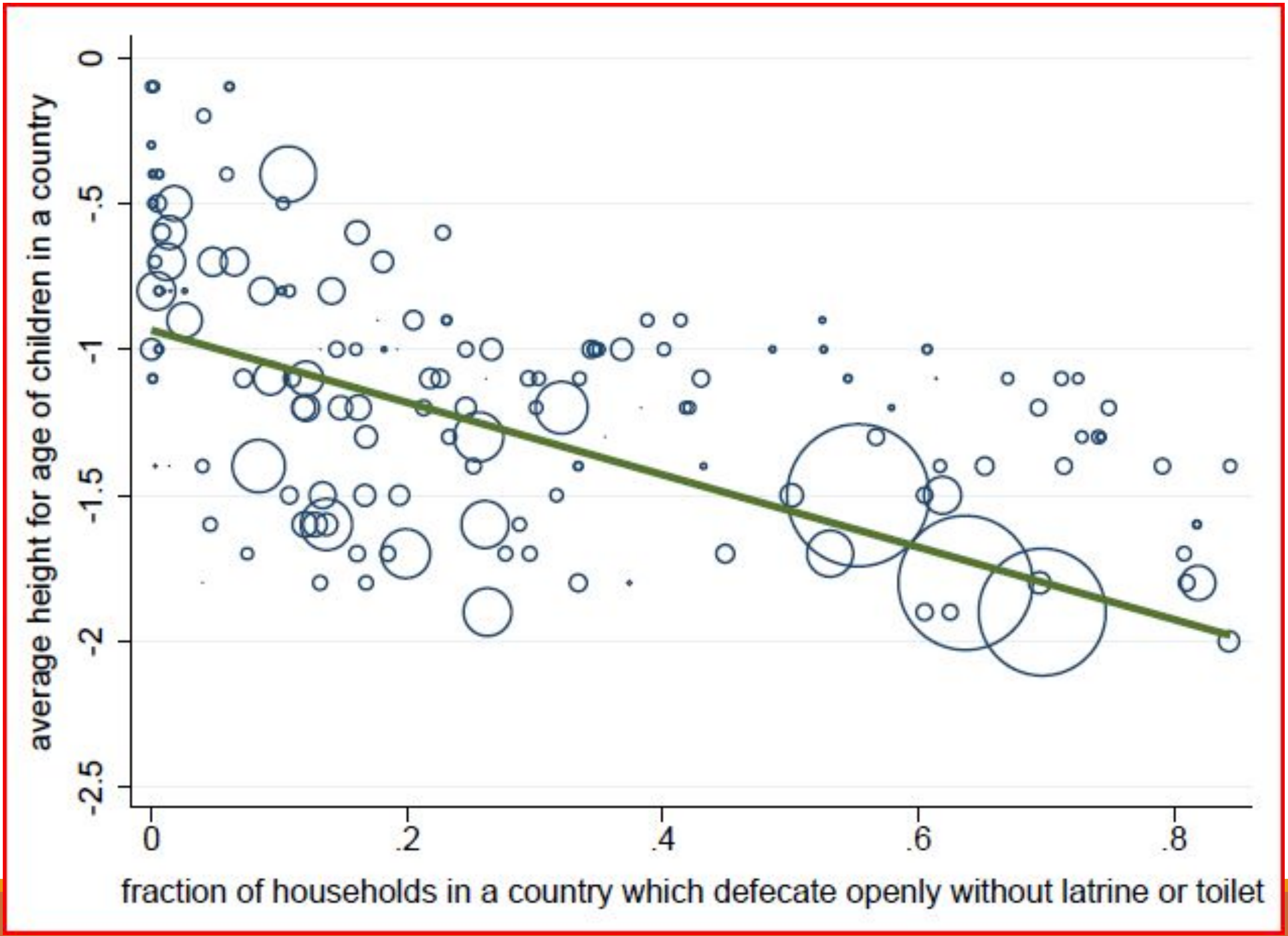
- Here is a story to think about…

- **Toilet Gap**
- Height of a child below average is among main proxy to describe her/his bad health status
- International policies against malnutrition privileged food security interventions
- Recently, something was found out…

What is big data in/for development?

- The **Toilet Gap**!
- Sanitation is heavily correlated with children malnutrition
- [Through big data?] new policies could have been formulated

Source: World Bank
http://goo.gl/qkzj1X



What is big data in/for development?

Main data types

3 main types of data being already used

1. Structured datasets – e.g. Call Detail Records (CDRs) collected by mobile phone operators, metadata capturing mobile phone subscribers' use of their cell-phones (identification code, location of the phone tower that routed the call for both caller and receiver, information about the call)

2. Media content: both in traditional media (e.g. videos, documents, blog posts) and social media.

3. Data from digital sensors – e.g. vegetation indexes from satellite sensors, smart meters to record water consumption

Who is actually doing «big data for development»?

- Main actors

- Main experiences

Who is doing that?

Various visions

- Private sector (not just tech corporations…) as a market strategy – e.g. IBM, Monsanto, etc.
- Private sector as a CSR strategy – e.g. Orange Telecom, etc.
- Governments – e.g. Governments of low- and middle-income countries, USAID, etc.
- Inter-governmental organizations – e.g. United Nations, World Bank, etc.
- Academia & research institutes – e.g. DataPop Alliance, Columbia University IDSE, MIT Media Lab, Université Catholique de Louvain, etc.
- Non profits – e.g. Rockefeller Foundation, Knight Foundation, Bill & Melinda Gates Foundation, DataKind
- …

Who is doing that?

Clashing (sometimes) narratives

- Big data as panacea VS Big data as evil

- Kenneth Cukier, data editor at The Economist
  http://www.economist.com/multimedia?bclid=0&bctid=2380227548001
- Andreas Weigend, former chief scientist at Amazon, introduced the ’data is the new oil’ concept
- E. Morozov “The rise of data and the death of politics”

  http://www.theguardian.com/technology/2014/jul/20/rise-of-data-death-of-politics-evgeny-morozov-algorithmic-regulation
- Danah Boyd and Kate Crawford, “Six provocations for big data”
  http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

Who is doing that?

- United Nations: **Global Pulse** (UN’s Big Data innovation lab)
- Launched in 2009
- Aims at developing the capacity to responsibly use new data sources and analytical tools
- Bridge among data scientists, data providers, governments, and development sector practitioners
- Both evangelization and practice
- Three labs: NYC, Kampala & Jakarta

Source: http://www.unglobalpulse.org

Who is doing that?

- Governments:
- Kenya launched its new Open Data Portal in 2011 - full digital edition of 2009 census, 12 years of government expenditure data, government household income surveys, location of schools and health facilities
  https://opendata.go.ke/
- The Philippines decided to establish a Big Data Center in 2014 to improve disaster response
  https://www.senate.gov.ph/press_release/2014/0625_aquino1.asp
- Mexico leveraged CDRs to analyze the country's response to the 2009 swine flu outbreak as well as to strengthen future response to flooding
  http://goo.gl/j3w9mA
- Ghana started mining social web to keep track of citizens' satisfaction on their government

Who is doing that?

United Nations: **Global Pulse**'s main taxonomy

- **Data Exhaust** - passively collected transactional data from people's use of digital services like mobile phones, purchases, web searches, etc., these digital services create networked sensors of human behavior;
- **Open Web Data -** Web content such as news media and social media interactions (e.g. blogs, Twitter), news articles obituaries, ecommerce, job postings; sensor of human intent, sentiments, perceptions, and want.
- **Citizen reporting -** information actively produced or submitted by citizens through mobile phone-based surveys, hotlines, user-generated maps, etc
- **Physical Sensors** - satellite or infrared imagery of changing landscapes, traffic patterns, light emissions, urban development and topographic changes, etc; remote sensing of changes in human activity

Who is doing that?

United Nations: **Global Pulse**’s 3 main domains of application

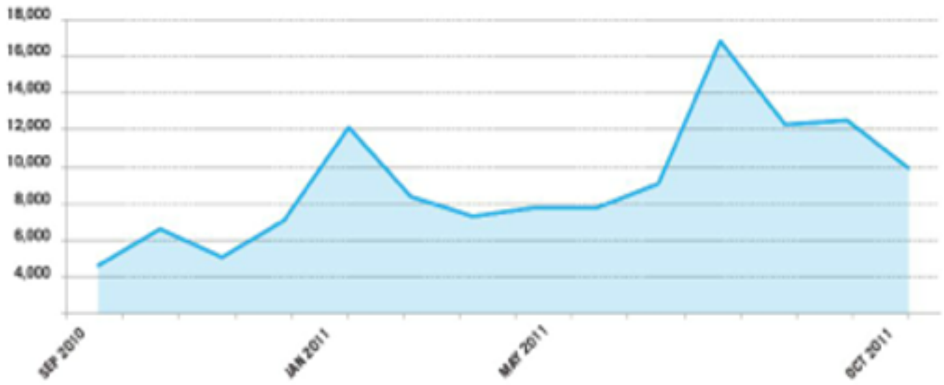
- **Early warning** - “early detection of anomalies in how populations use digital devices and services can enable faster response in times of crisis”;
- **Digital awareness** - “Big Data can paint a fine-grained and current representation of reality which can inform the design and targeting of programs and policies;”
- **Real-time feedback** - “monitor a population in real time makes it possible to understand where policies and programs are failing and make adjustments.”

Who is doing that?

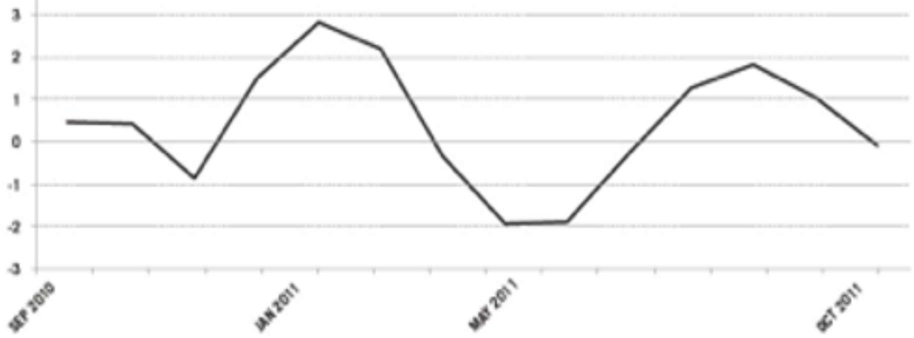
United Nations: **Global Pulse**’s mining Indonesian tweets to understand food price crises (2013)

- Tweets about the price of rice
vs. actual price of rice in Indonesia

Tweets about the price of rice
(per month)



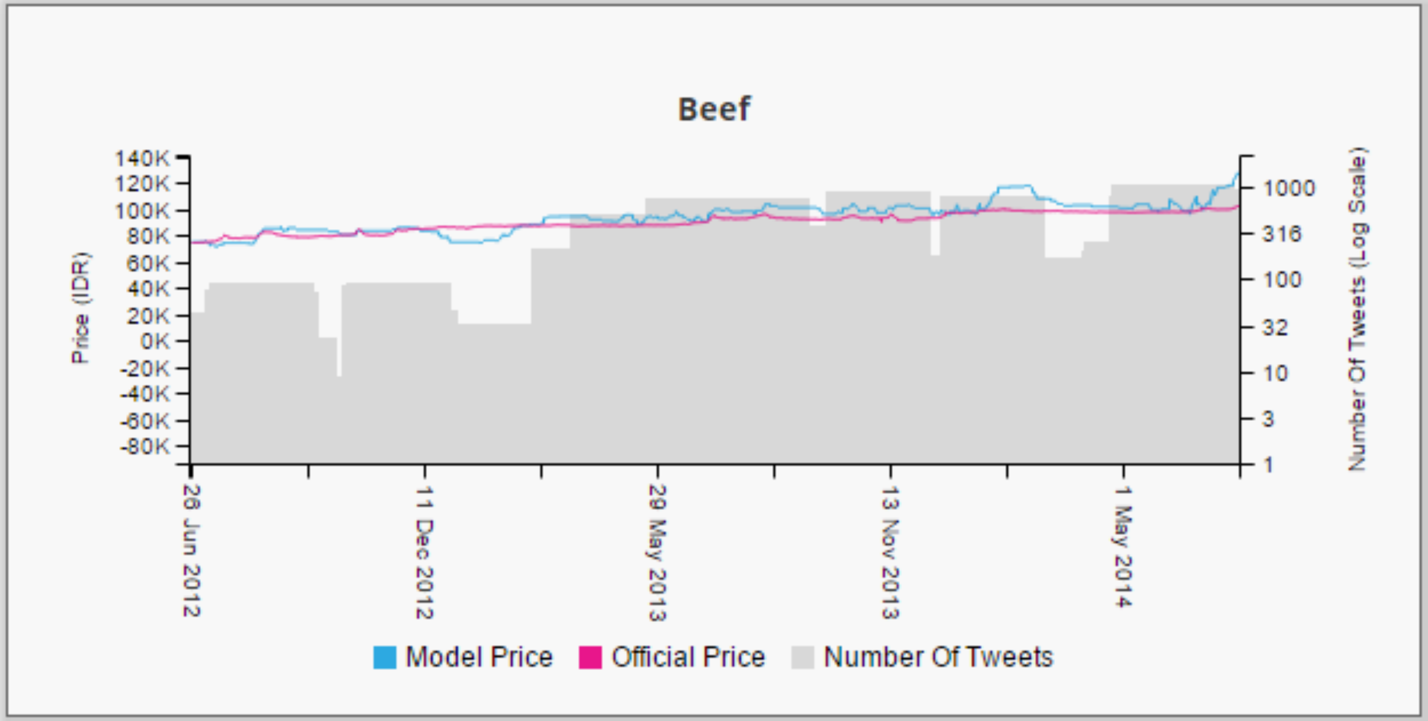
Food Price Inflation



Who is doing that?

United Nations: **Global Pulse**'s mining Indonesian tweets to understand food price crises (2013)

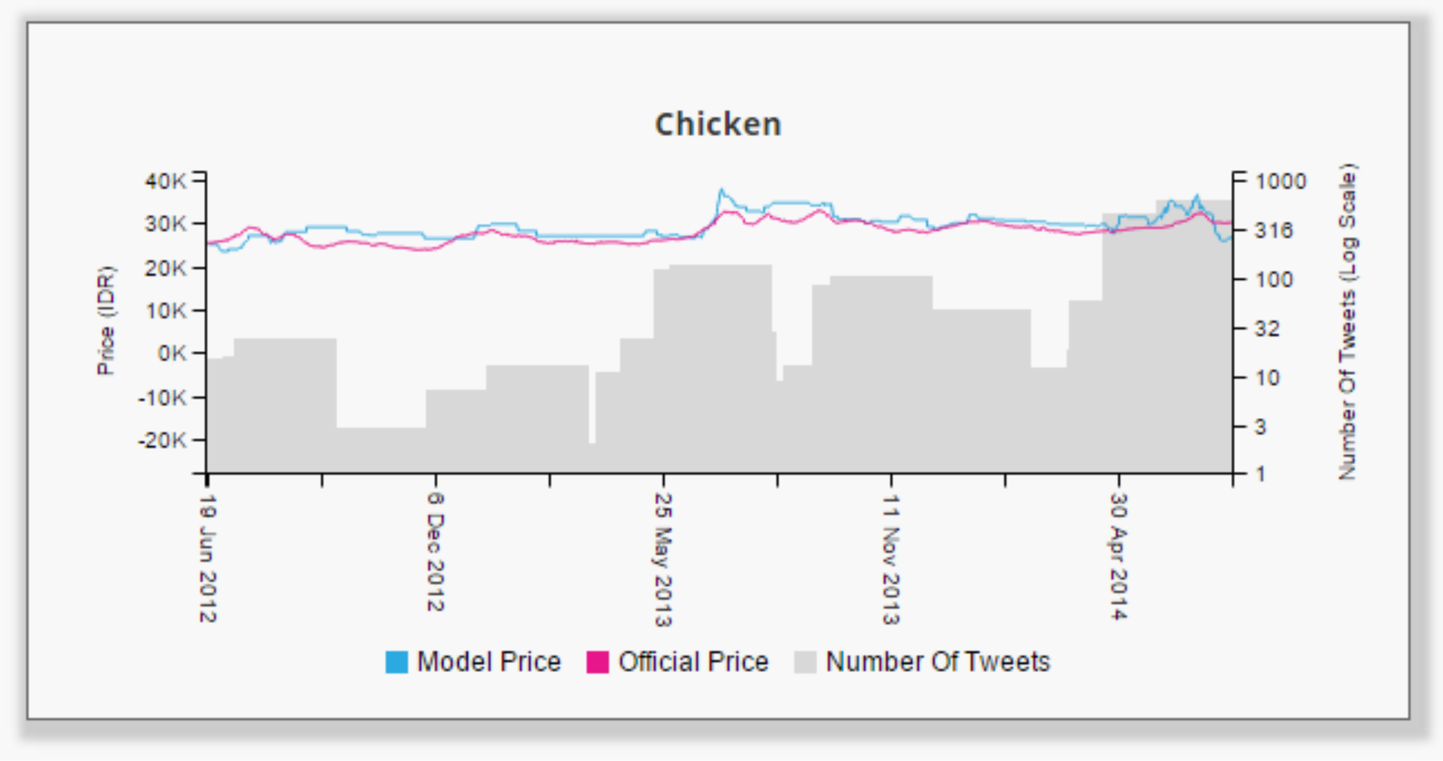
- Tweets about the price of beef vs. actual price of beef in Indonesia



Who is doing that?

United Nations: **Global Pulse**’s mining Indonesian tweets to understand food price crises (2013)

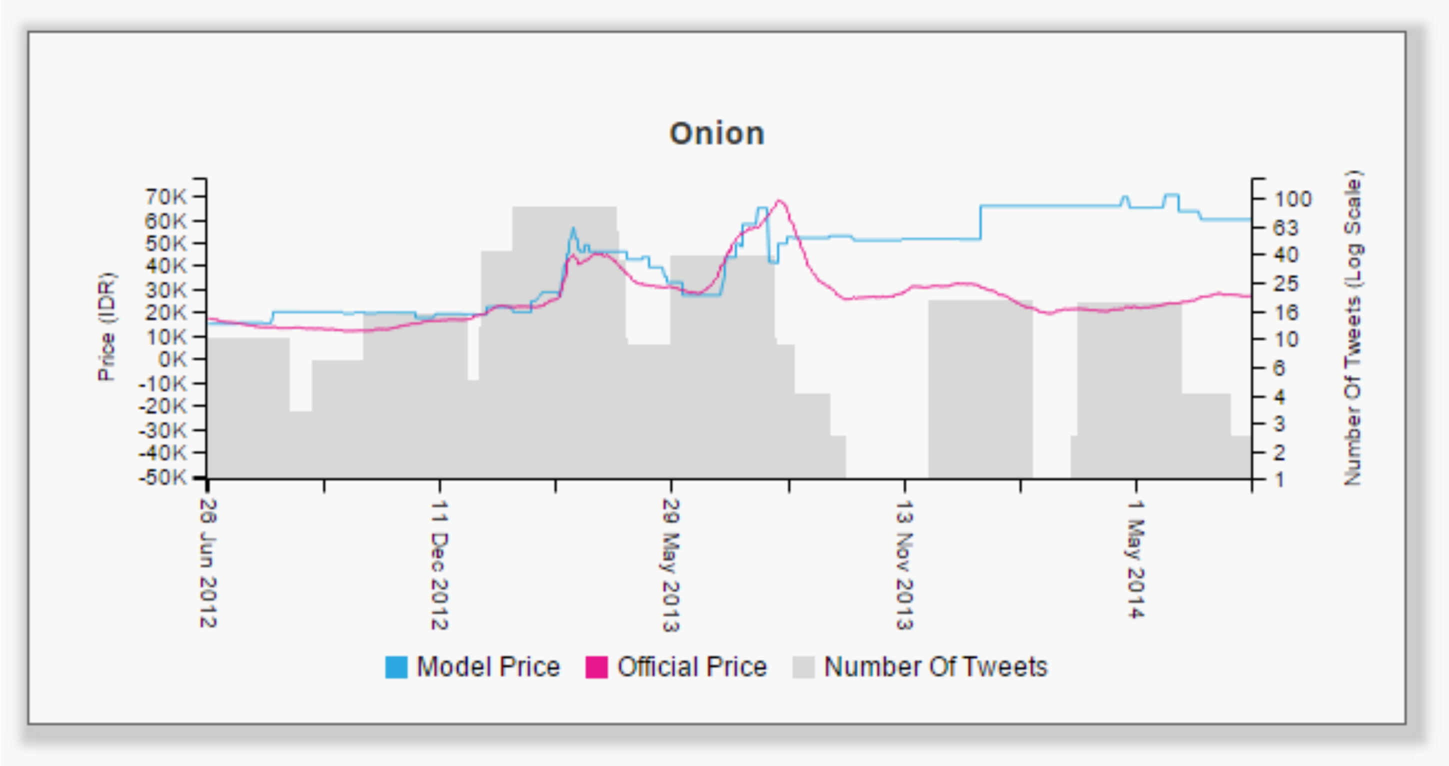
- Tweets about the price of chicken vs. actual price of chicken in Indonesia



Who is doing that?

United Nations: **Global Pulse**’s mining Indonesian tweets to understand food price crises (2013)

- Tweets about the price of onions vs. actual price of onions in Indonesia



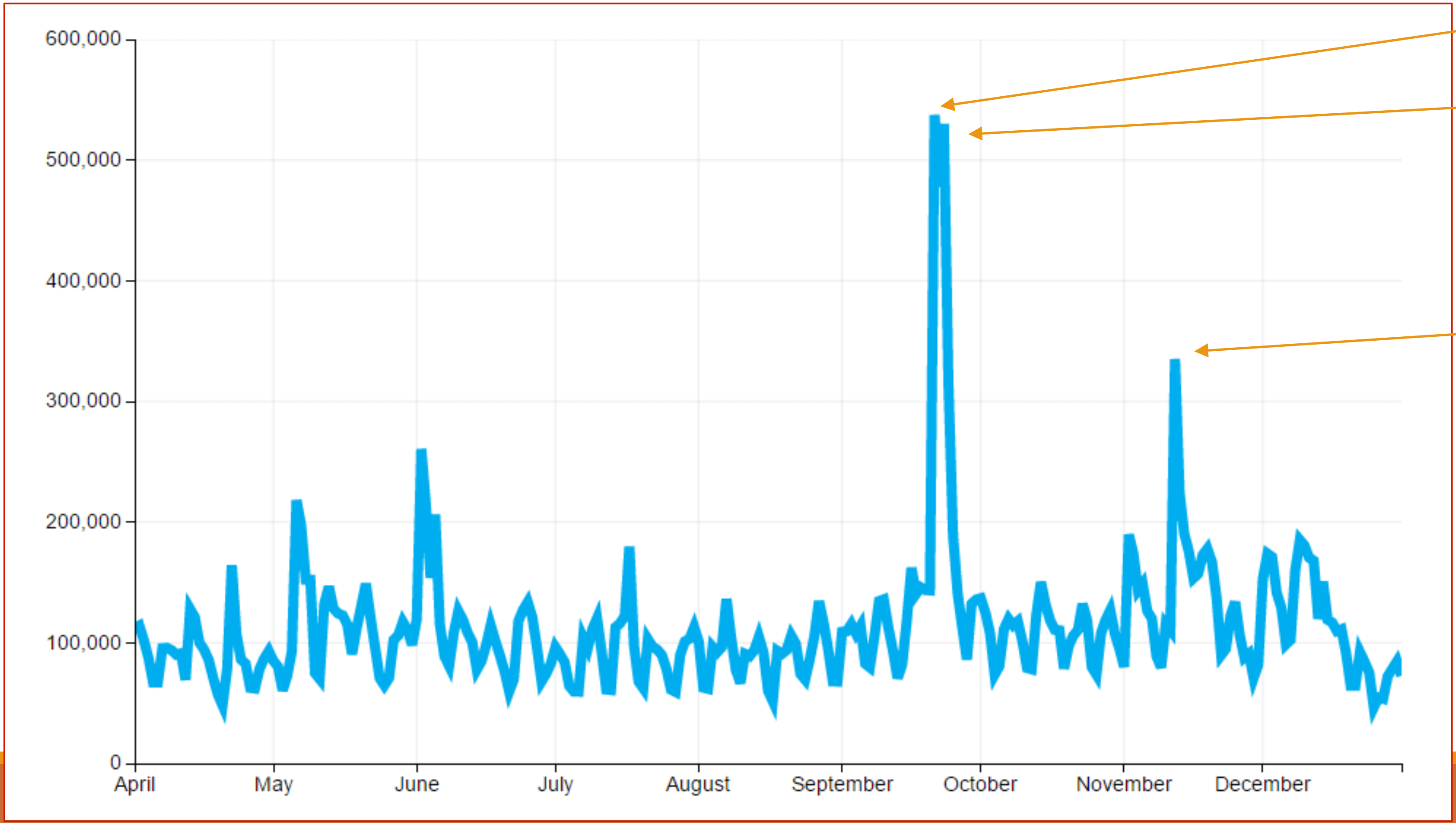
Who is doing that?

United Nations: **Global Pulse**’s analysis of global engagement on climate change via twitter (2014)

- Towards 2014 **UN Climate Summit**: real-time social media monitor to measure and explore online discourse about climate change
- Goal: Assess volume & content of tweets about climate change (in 3 languages) across a range of topics (e.g. economy, energy)

Who is doing that?

United Nations: **Global Pulse**'s analysis of global engagement on climate change via twitter (2014)



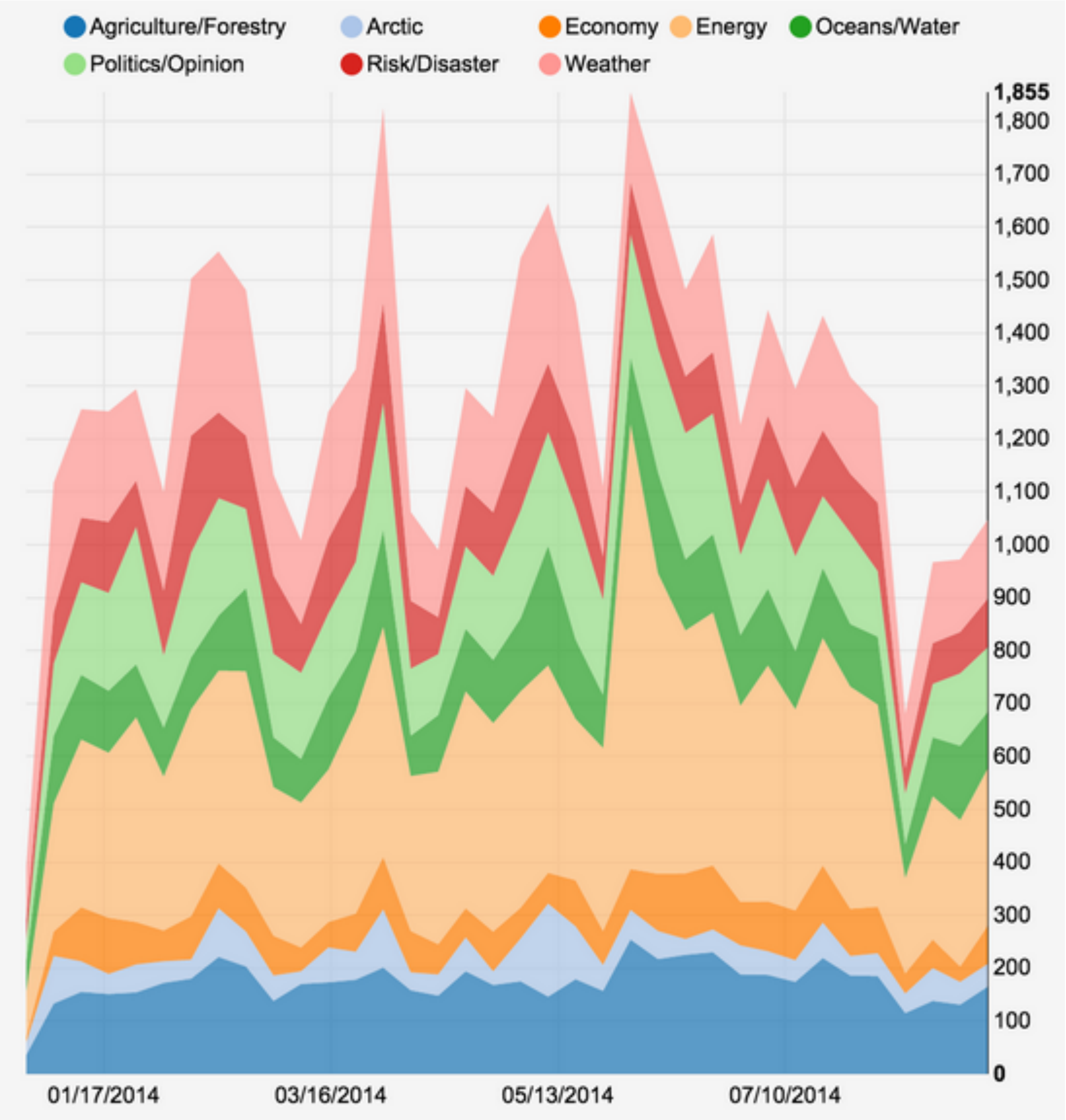
Climate March

Climate Summit

US – China
deal on carbon cuts

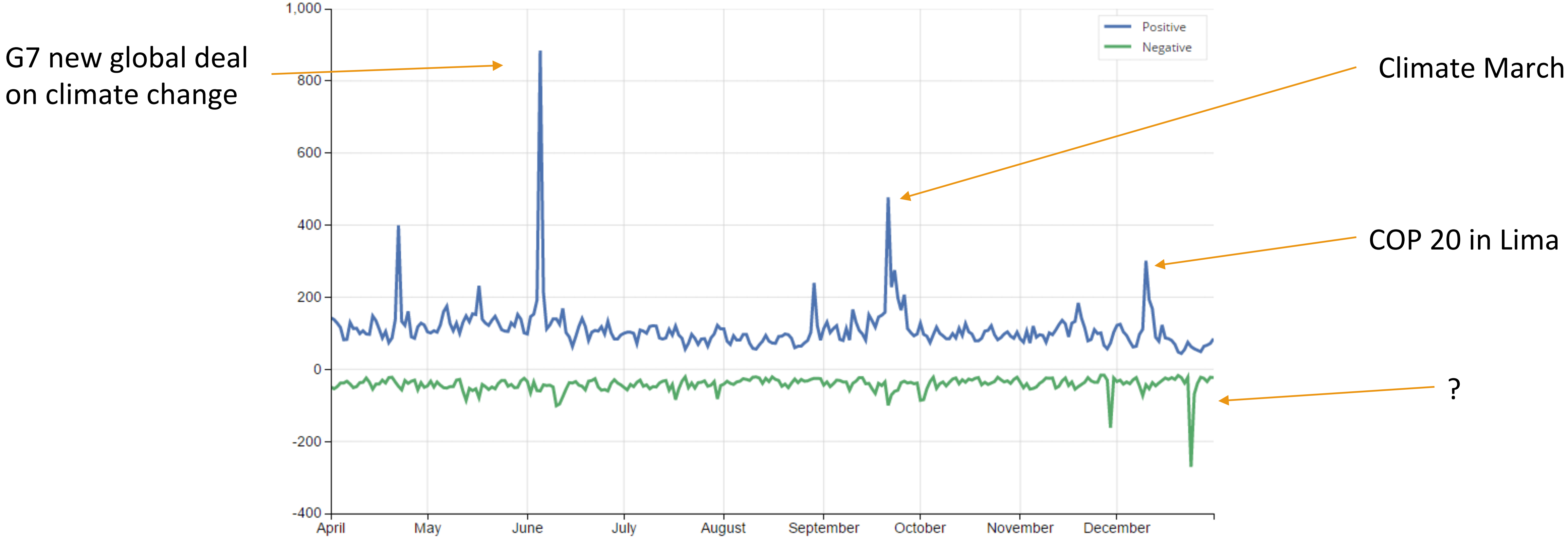
Who is doing that?

United Nations: **Global Pulse**’s analysis of global engagement on climate change via twitter (2014)



Who is doing that?

United Nations: **Global Pulse**'s analysis of global engagement on climate change via twitter (2014)



Who is doing that?

United Nations: **Global Pulse**'s analysis of global engagement on climate change via twitter (2014)

- Towards 2014 **UN Climate Summit**: real-time social media monitor to measure and explore online discourse about climate change
- Goal: Assess volume & content of tweets about climate change (in 3 languages) across a range of topics (e.g. economy, energy)
- Outputs: development of a baseline of engagement by measuring public tweets over time (highlighiting spikes during the Summit); identification of main sources of information in the social web; …
- Outcomes: improved ability to compare interest level between topics/regions; improved capacity to monitor impact of social media impact of climate-related communications (thus higher capacity to measure awareness, support climate policy decision-making, foster public engagement, …).
- Source: http://www.unglobalpulse.net/climate/

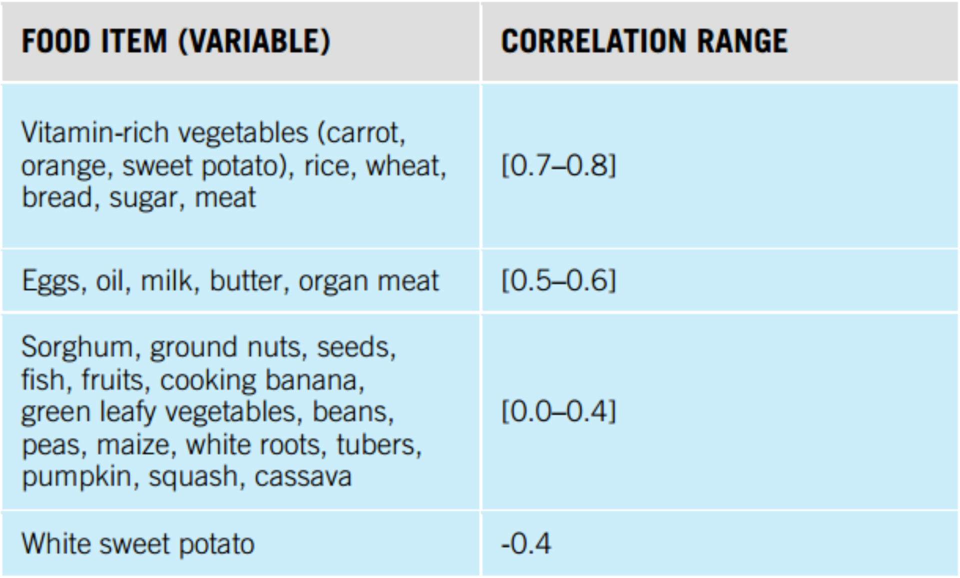
Who is doing that?

United Nations: **Global Pulse**’s development of food security estimates via mobile phone data & airtime credit purchases in East Africa (2015)

- Research conducted with UN WFP, Université Catholique de Louvain and Real Impact Analytics
- Comparison of data extracted from airtime credit purchases (or “top-ups”) and mobile phone activity with nationwide household survey conducted by WFP at the same time.
- Outcome: high correlations between airtime credit purchases and survey results referring to consumption of several food items – e.g. vitamin-rich vegetables, meat or cereals.
- Source: http://www.unglobalpulse.org/mobile-CDRs-food-security

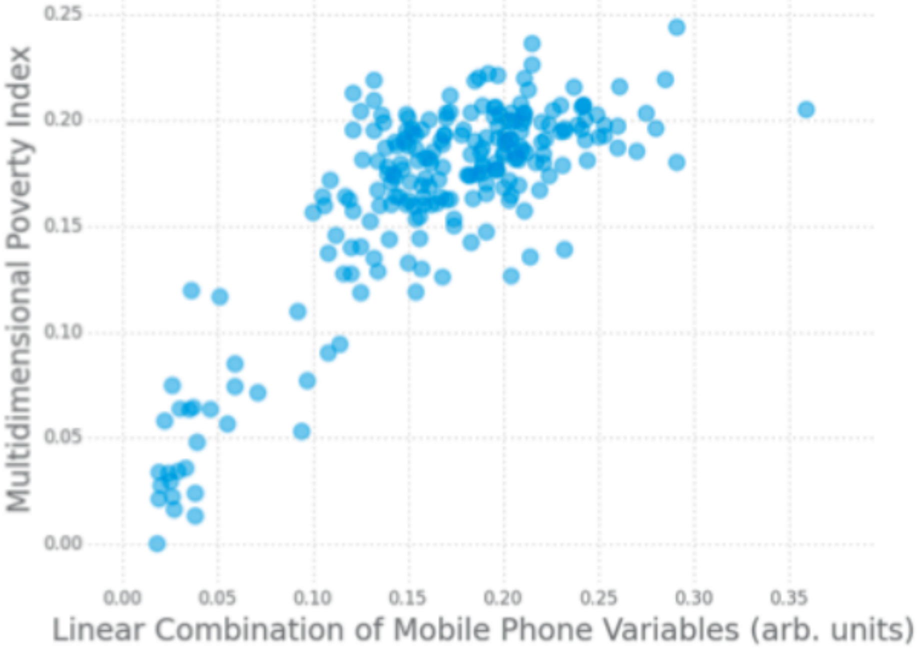
Who is doing that?

United Nations: **Global Pulse**’s development of food security estimates via mobile phone data & airtime credit purchases in East Africa (2015)



FOOD ITEM (VARIABLE)	CORRELATION RANGE
Vitamin-rich vegetables (carrot, orange, sweet potato), rice, wheat, bread, sugar, meat	[0.7–0.8]
Eggs, oil, milk, butter, organ meat	[0.5–0.6]
Sorghum, ground nuts, seeds, fish, fruits, cooking banana, green leafy vegetables, beans, peas, maize, white roots, tubers, pumpkin, squash, cassava	[0.0–0.4]
White sweet potato	-0.4

Correlation between consumption of foods and the sum of airtime credit purchases



Correlation between mobile phone variables and the developed Multidimensional Poverty Index

Who is doing that?

United Nations: **Global Pulse**’s additional activities

- Assessment of indicators of various socio-economic variables – e.g. unemployment through big data
  Source: http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics
- Feasibility study with USAID to identify barriers to financial inclusion in Kenya via new digital data sources
  Source: http://unglobalpulse.org/Kenyan-access-finance
- Mining the social web (blog posts, online comments and tweets) with UNICEF to gain insights into the attitude of parents towards immunization of their children. Leverage the findings to improve communication with parents to allay their fears.
  Source: http://www.unicef.org/ceecis/media_24017.html

Who is doing that?

- United Nations (post2015 agenda)

“Better data and statistics will help governments **track progress** and make sure their **decisions are evidence-based**; they can also strengthen **accountability**.

This is not just about governments. International agencies, CSOs and the private sector should be involved.

A true *data revolution* would draw on existing and new sources of data to fully **integrate statistics into decision making**, promote **open access** to, and use of, data and ensure increased support for statistical systems.”

U.N. High -Level Panel report on the post-2015 agenda

Source: http://www.post2015hlp.org/featured/high-level-panel-releases-recommendations-for-worlds-nextdevelopment-agenda/

+ World Bank’s work on the same topic: Shanta Devarajan, Africa’s statistical tragedy
http://blogs.worldbank.org/africacan/africa-s-statistical-tragedy

Who is doing that?

Academia & research institutes - Engineering Social Systems lab, Harvard

- Smart (?) slums: Kibera (Nairobi, Kenya)
- Coupled analysis of (a few terabyte of) data gathered from mobiles and information extracted from national census
- Mobile phone call logs from June 2008 to June 2009
- Development of a growth model of the slum, inferring places of work and migration out of Kibera
- Ultimate goal: make municipalities aware about the places where water pumps & health services will mostly be needed

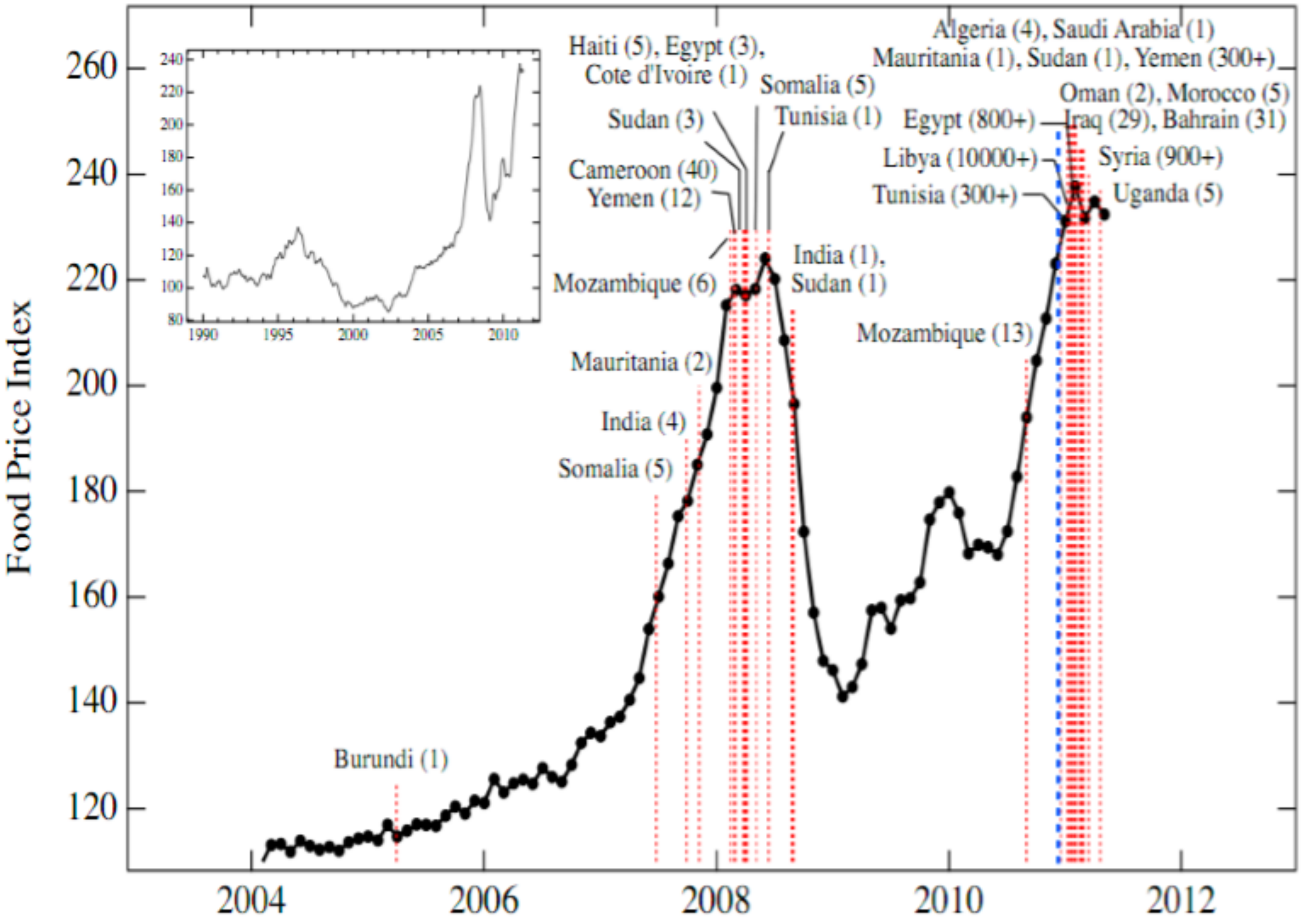
Source: http://www.hsph.harvard.edu//ess/

Who is doing that?

Academia & research institutes –
New England Complex Systems Institute

- Model and predict food riots based on historical data about food prices

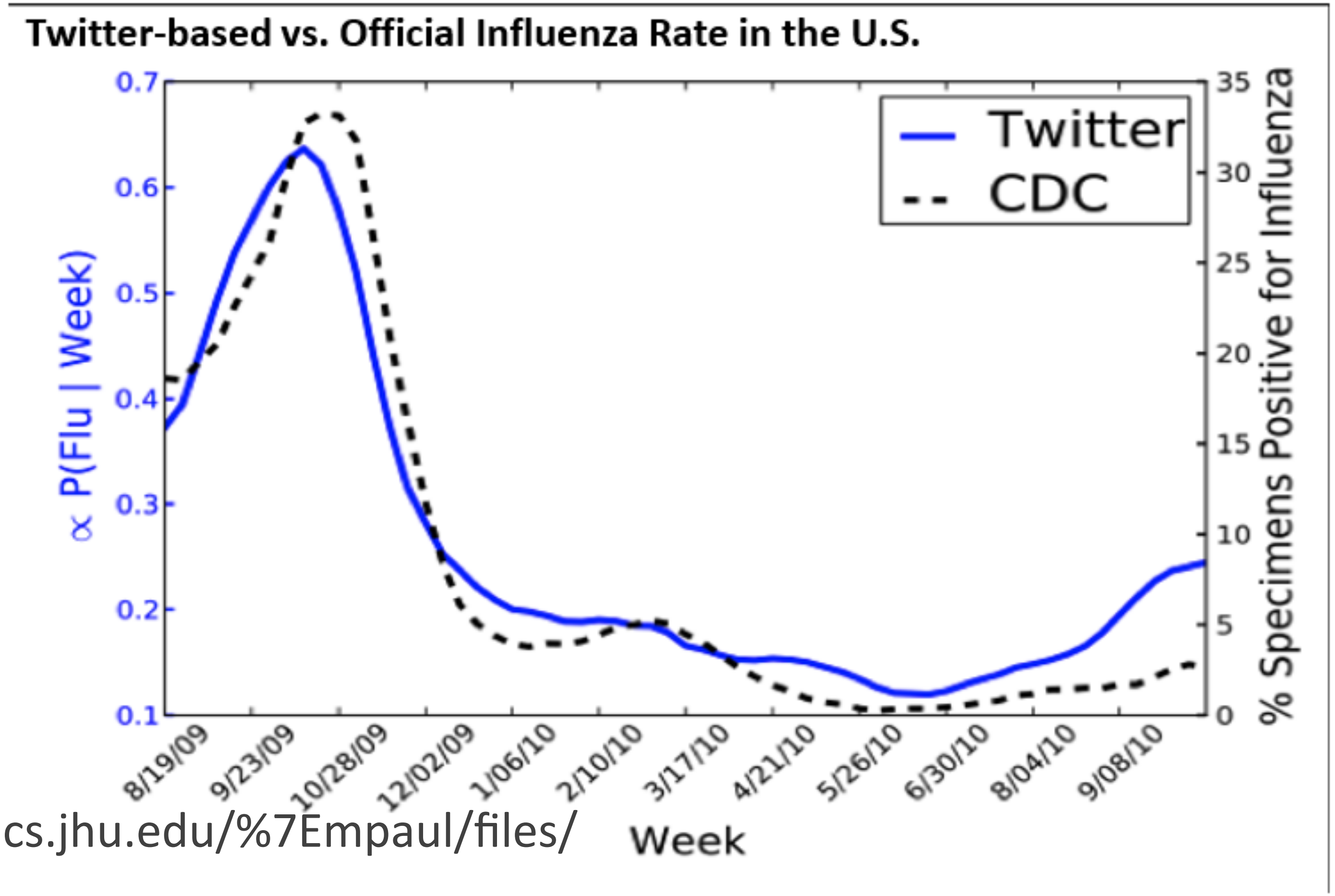
Source:
http://necsi.edu/research/social/food_crises.pdf



Who is doing that?

Academia –
Computer Science @ JHU

- Model and predict flu outbreaks in the US based on tweets



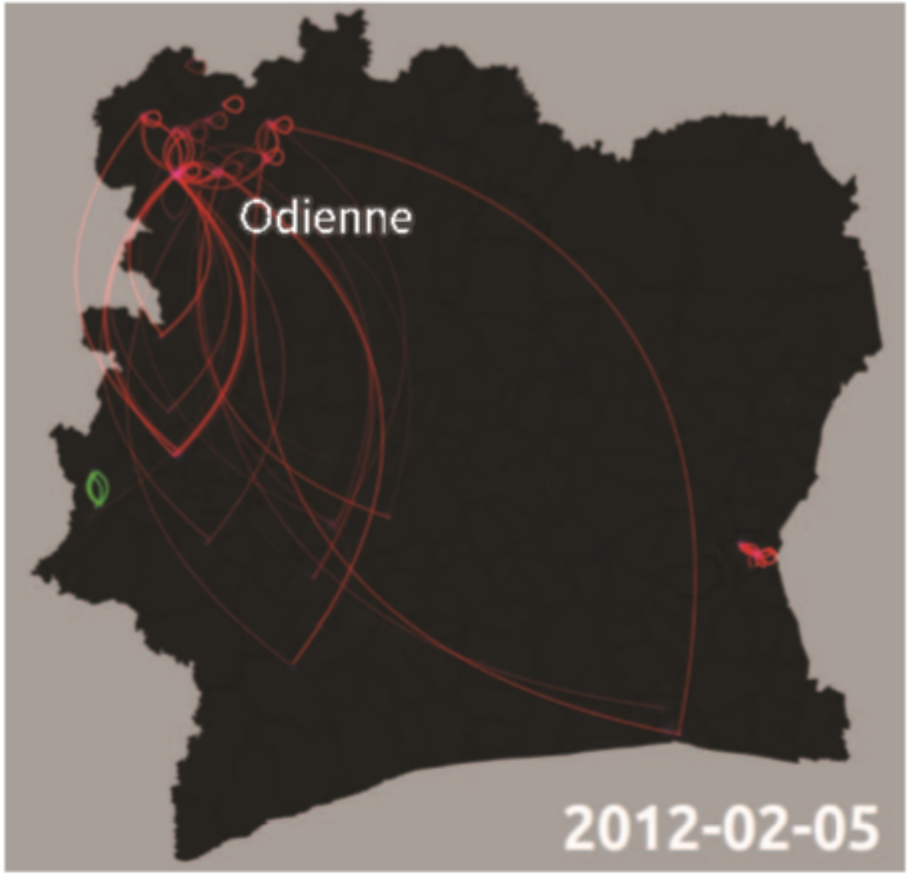
Source: Paul and Dredze (2011) http://www.cs.jhu.edu/%7Empaul/files/2011.icwsm.twitter_health.pdf

Who is doing that? And where?

Academia – joint research groups for D4D Challenge (2013)

- Exploration and analysis of 2.5 billion calls and SMS exchange between around 5 million users located in Ivory Coast over a period of 5 months.
- Analysis of linkages between localized increase and decrease of calls with major local events – e.g. increased phone calls correlated with conflict events

Source: Van den Eltzen et al. (2013) http://goo.gl/DNPmg4

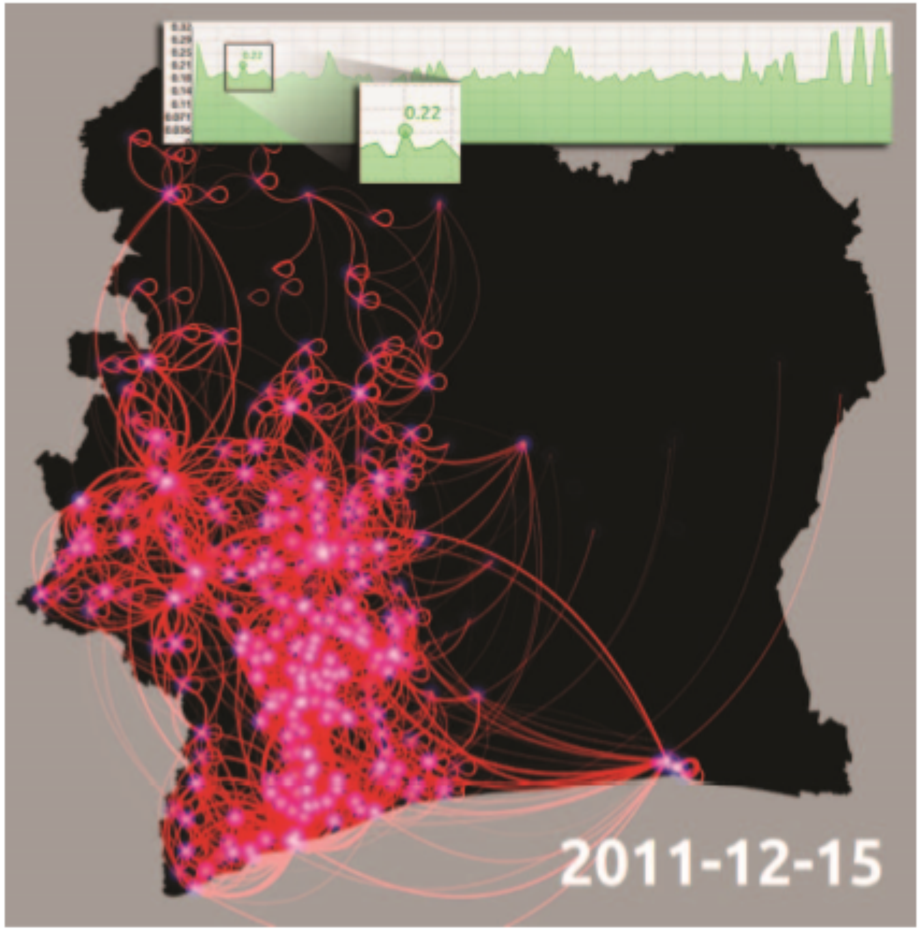


Who is doing that? And where?

Academia – joint research groups for D4D Challenge (2013)

- Exploration and analysis of 2.5 billion calls and SMS exchange between around 5 million users located in Ivory Coast over a period of 5 months.
- Analysis of linkages between localized increase and decrease of calls with major local events – e.g. shut-down of network after the elections in selected areas of the country

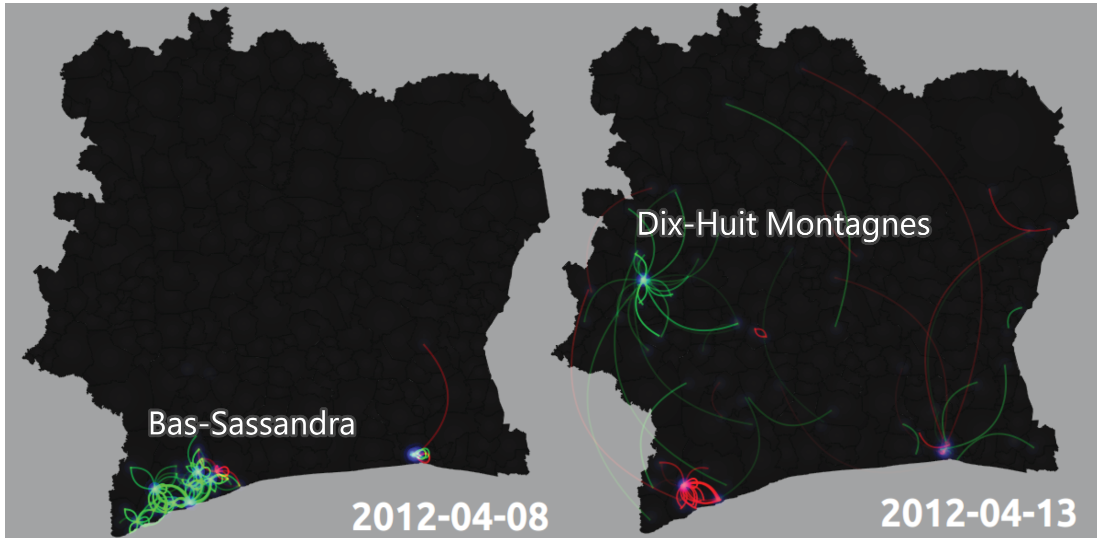
Source: Van den Eltzen et al. (2013) http://goo.gl/DNPmg4



2011-12-15

Who is doing that? And where?

Academia – joint research groups for D4D Challenge (2013)



- Exploration and analysis of 2.5 billion calls and SMS exchange between around 5 million users located in Ivory Coast over a period of 5 months.
- Analysis of linkages between localized increase and decrease of calls with major local events – e.g. increased phone calls correlated with weather anomalies
- Possible way forward:
  Optimization of irrigation, logistics, harvesting
  Early warning (e.g. calls correlation with weather events, pests, diseases)
  Yield performance (e.g. calls at marketing time in cocoa-producing region)

Source: Van den Eltzen et al. (2013) http://goo.gl/DNPmg4

Why are we doing this?

Access issues

- New digital divides
- ~~Can we trust data?~~ At what extent can we trust data?

Why are we doing this?

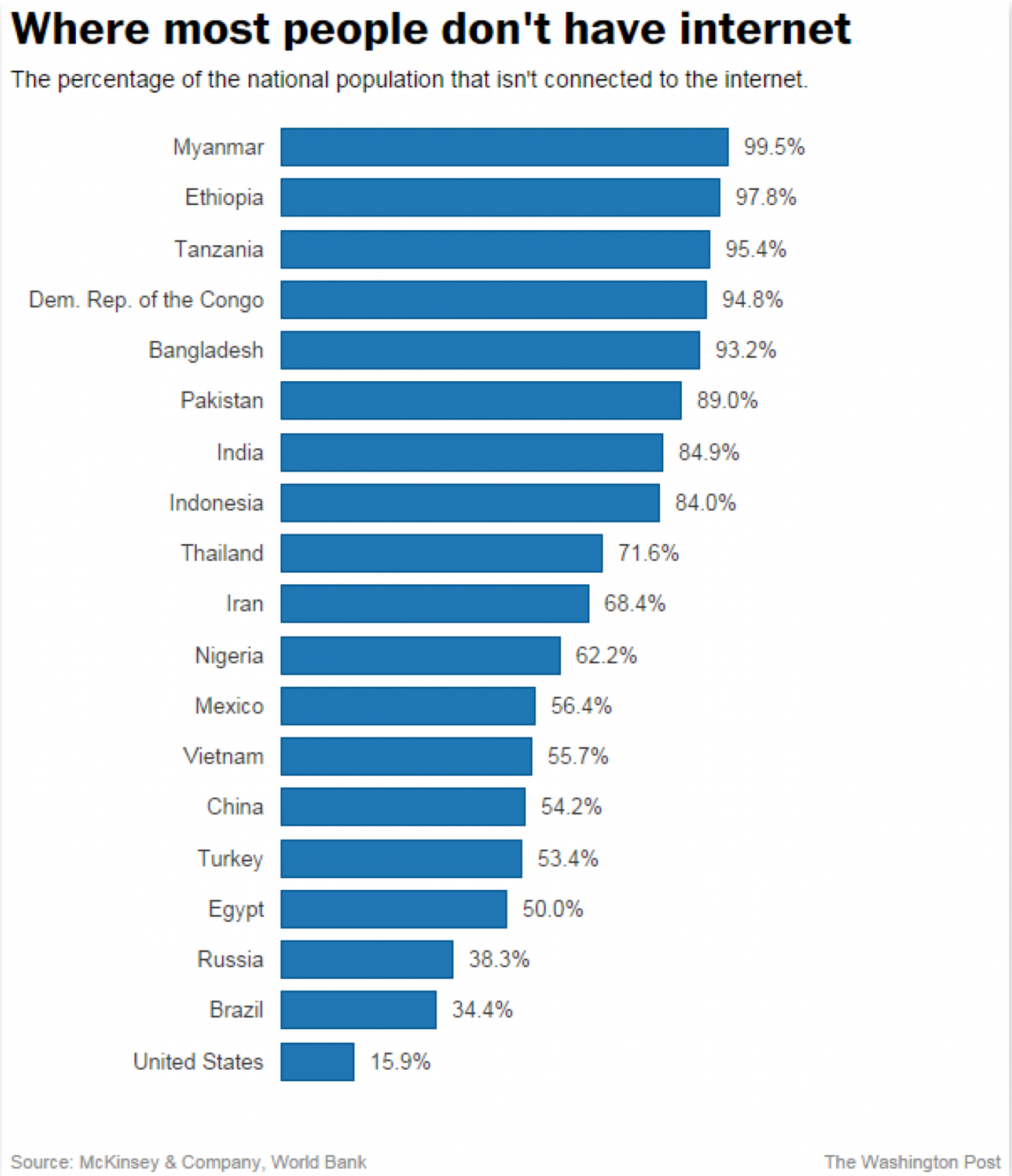
Access issues

- Who are we leaving behind?

Why are we doing this?

Access issues

- Who are we leaving behind?
- Quite a lot of persons, actually.



Where most people don't have internet

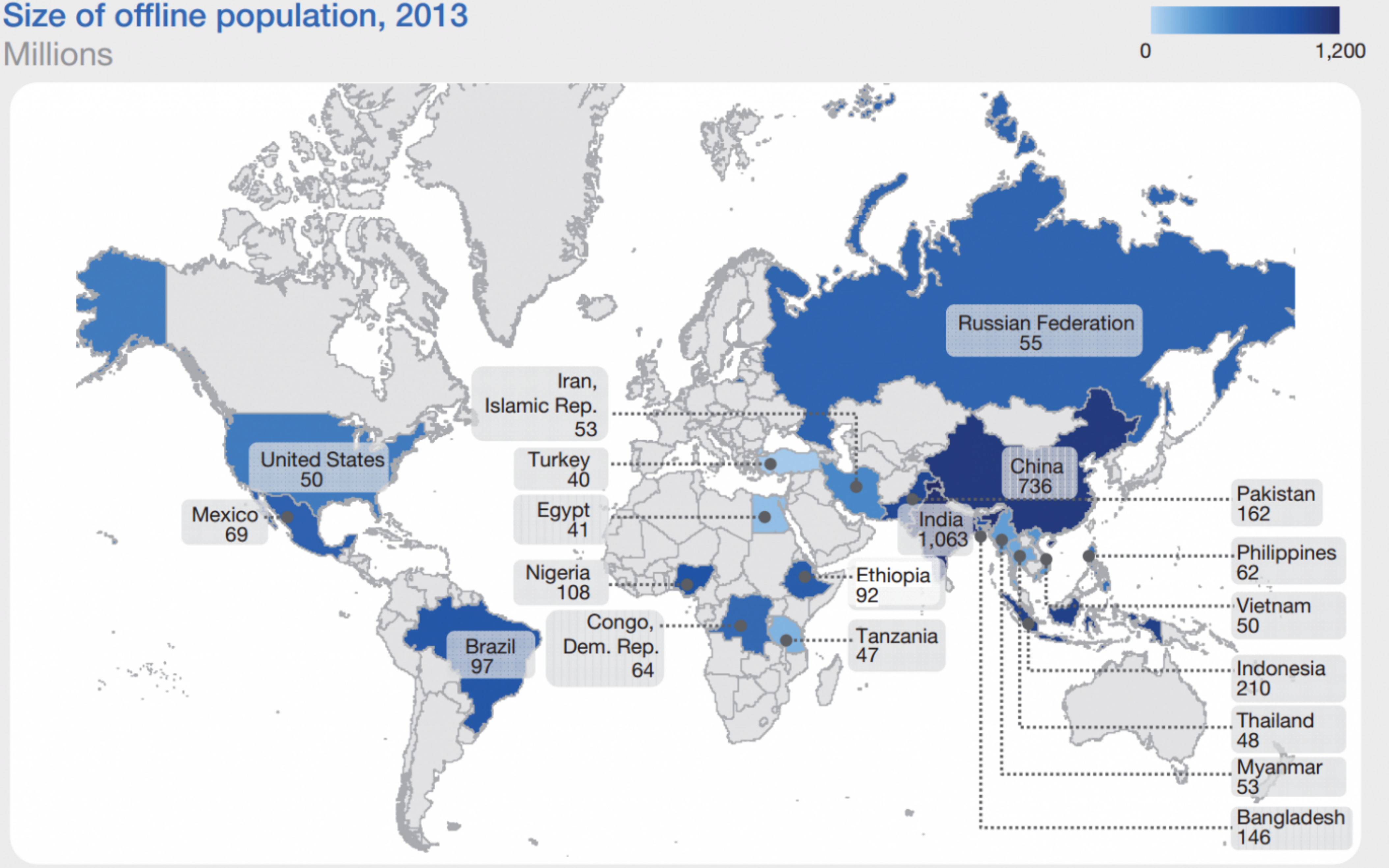
The percentage of the national population that isn't connected to the internet.

Source: McKinsey & Company, World Bank

The Washington Post

Size of offline population, 2013

Millions



Why are we doing this?

Access issues

- Who are we leaving behind?

- Even when it is just for a while – it matters

- Sub-Saharan African countries suffer 5-hour power cuts on average every 4 days. This is 25% more frequent and almost 2x as the overall average of the 135 predominately developing nations investigated [World Bank's Enterprise Survey]

Why are we doing this?

Access issues

- Who are we leaving behind?

- Access to data is not always granted: researchers’ and policy makers’ ability to access new sources of data is critical (as much as their ability to properly maintain their old sources)
- Most of new data sources are proprietary

Why are we doing this?

Capacity issues

- Can we actually use all our data?

“The explosion of big data has far-outpaced our ability to make sense of it in poorer nations that already lack human and technical capacity.” - Claire Melamed, Overseas Development Institute

“The real problem is that governments are swimming in more data than they have ever had but they lack the capacity in their staff to do anything with it” - Jon Gosier, D8A Group CEO

- Strengthen capacities is critical

Why are we doing this?

Interpretation issues

- Can we trust data ?

"Nobody believes in simulation models except their developers…

Everybody believes in experimental data except who collected them"

- Prof. Gaylon S. Campbell

Why are we doing this?

Interpretation issues

- At what extent can we trust data ?
- **Interpretation issues**
- Correlation is not causation

"When you have a billion observations, everything's significant" – Prof. Hal Varian, Chief Economist at Google

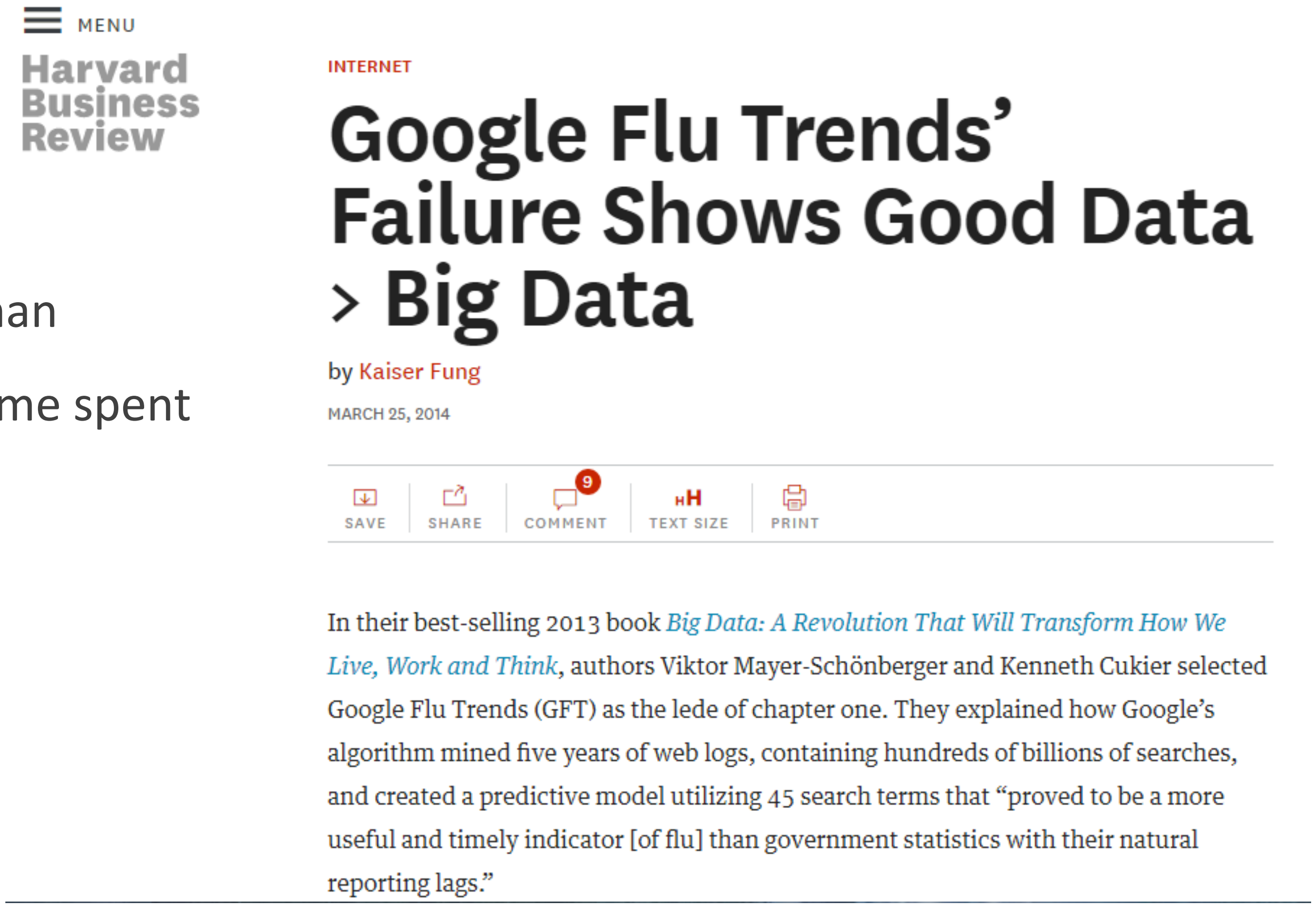
- Data cannot always speak for itself: expert knowledge is often needed (meaning of data is not stable, and it is hard to detect bias in data)

"[…] still you will need theory to understand the mechanisms or even to suggest what you might hope to find in the first place." Prof. Sascha Becker, University of Warwick

Why are we doing this?

Interpretation issues

- At what extent can we trust data ?
- Still a long, long way to go…
- Behind every data/algorithm there is a human
- Quality matters: 80% of a data scientist's time spent cleaning up data
- Good data >>> Big data:
  E.g. GFT has over-estimated the prevalence of flu for 100 out of the last 108 weeks;
  it's been wrong since August 2011.



INTERNET

Google Flu Trends' Failure Shows Good Data > Big Data

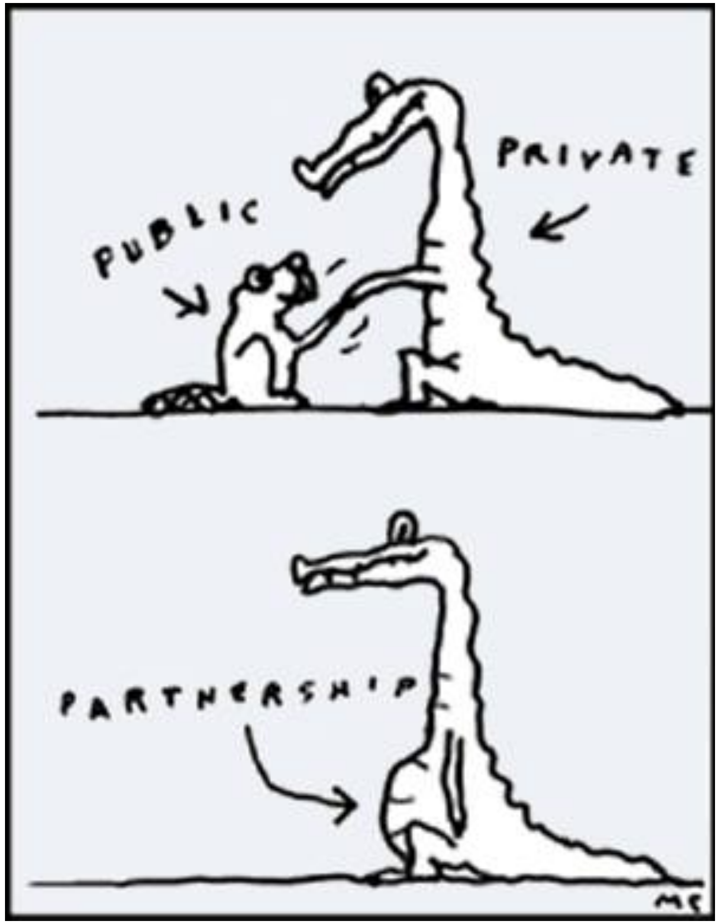
by Kaiser Fung

MARCH 25, 2014

In their best-selling 2013 book *Big Data: A Revolution That Will Transform How We Live, Work and Think*, authors Viktor Mayer-Schönberger and Kenneth Cukier selected Google Flu Trends (GFT) as the lede of chapter one. They explained how Google's algorithm mined five years of web logs, containing hundreds of billions of searches, and created a predictive model utilizing 45 search terms that "proved to be a more useful and timely indicator [of flu] than government statistics with their natural reporting lags."

Why are we doing this?

Ethical issues



Why are we doing this?

Ethical issues

- When Big Data ends and Big Brother begins?
- Are we sure hyper-transparency is always a great thing (e.g. for individuals who belong to a minority)?
- Risks behind unilateral models of intervention in development settings
- What about the open data dividend concept? http://www.ictworks.org/2015/03/18/how-can-we-all-profit-from-development-data/
- …

Further information on many of these issues: https://linnettaylor.wordpress.com/

Thank you very much!

Feedback welcome: *salas@mit.edu*