GLOBAL PULSE

BIG DATA FOR DEVELOPMENT: A PRIMER

Harnessing Big Data For Real-Time Awareness

WHAT IS BIG DATA?

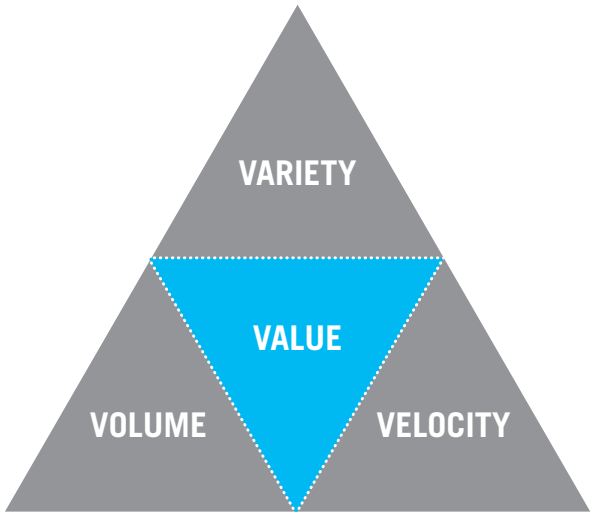
Big Data is an umbrella term referring to the large amounts of digital data continually generated by the global population. The speed and frequency by which data is produced and collected—by an increasing number of sources—is responsible for today's data deluge: the amount of available digital data is projected to increase by an annual 40%.

A large share of this output is "data exhaust," or records generated as a by-product of everyday interactions with digital products or services.

The private sector—including mobile phone carriers, credit card companies and social media networking sites—manages enormous data sets that hold rich insights. Companies analyze this data to support decision-making or provide market intelligence. More recently, public sector institutions have begun leveraging similar techniques to generate actionable insights for policymakers.

DATA EXHAUST
Passively collected data deriving from daily usage of digital devices.

HOW TO CITE THIS DOCUMENT: United Nations Global Pulse (2013) Big Data for Development: A primer.



Big Data is characterized by the "3 Vs:" greater volume, more variety, and a higher rate of velocity. A fourth V, for value, can account for the potential of Big Data to be utilized for development.

Advances in computing and data science now make it possible to process and analyze Big Data in real time. However, due to its size and often complex and unstructured nature, Big Data presents several analytical challenges that demand continually updated tools and expertise. Legitimate concerns about privacy and the digital divide also present new obstacles to harnessing Big Data sets for public benefit.

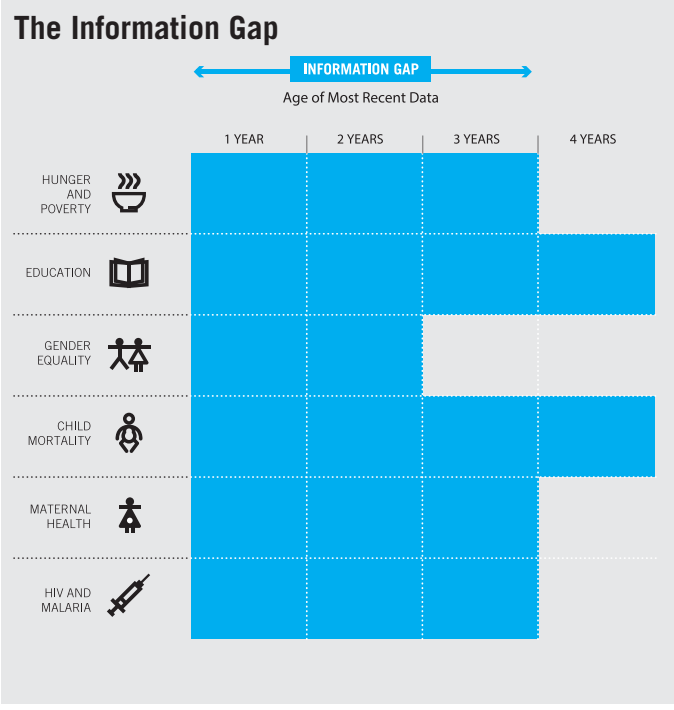
This increasing volume of real-time data is driving a global data revolution—a phenomenon that is recent (less than a decade old), extremely rapid (growth is exponential) and immensely consequential for society.

Big Data is different from Open Data. Open Data refers to data that is free from copyright and can be shared in the public domain. That is not a defining characteristic of Big Data, which can be privately owned or have varying levels of access control.

BIG DATA IN THE DEVELOPING WORLD

The data revolution is not restricted to the industrialized world. The spread of mobile phone technology to the hands of billions of individuals may be the single most significant innovation that has affected developing countries in the past decade. Across the developing world, mobile phones are used daily to transfer money, buy and sell goods, and communicate information including test results, stock levels and prices of commodities. Mobile technology is used as a substitute for weak telecommunications and transport infrastructures as well as underdeveloped financial and banking systems.

The numbers of real-time information streams and people using social media are growing rapidly in developing countries as well. Tracking trends in online news or social media can provide insights on emerging concerns that can be highly relevant to global development.



Household-level data is challenging to collect on a real-time basis, making development progress difficult to track. (Source: Millennium Development Goals Report, 2011)

The recent waves of global shocks—food, fuel and financial—have led to greater volatility, and policymakers are increasingly aware of the costs. Despite greater interconnectivity, local impacts of shocks like food crises or natural disasters may not be immediately visible and trackable.

These are important issues that often unfold beneath the radar of traditional monitoring systems, and by the time hard evidence finds its way to the front pages of newspapers and desks of decision makers, it's often too late and more expensive to respond.

While Early Warning Systems and data collected through "traditional" methods (surveys and statistics) continue to generate relevant information, the Digital Revolution presents a tremendous opportunity to gain richer insight into the human experience, and Big Data can complement the existing indicators.

WHAT IS "BIG DATA FOR DEVELOPMENT"?

"Big Data for Development" is a concept that refers to the identification of sources of Big Data relevant to policy and planning of development programmes. It differs from both "traditional" development data and what the private sector and mainstream media call Big Data.

In general, sources of Big Data for Development are those which can be analyzed to gain insight into to human well-being and development, and generally share some or all of the following features:

1 DIGITALLY GENERATED
Data is created digitally, not digitized manually, and can be manipulated by computers.

2 PASSIVELY PRODUCED
Data is a by-product of interactions with digital services.

3 AUTOMATICALLY COLLECTED
A system is in place that automatically extracts and stores the relevant data that is generated.

4 GEOGRAPHICALLY OR TEMPORALLY TRACKABLE
For instance, this is the case in mobile phone location data or call duration time.

5 CONTINUOUSLY ANALYZED
Information is relevant to human well-being and development, and can be analyzed in real time.

Big Data for Development is constantly evolving. However, a preliminary categorization of sources may reflect:

WHAT PEOPLE SAY

Online Content: International and local online news sources, publicly accessible blogs, forum posts, comments and public social media content, online advertising, e-commerce sites and websites created by local retailers that list prices and inventory.

WHAT PEOPLE DO

Data Exhaust: Passively collected transactional data from the use of digital services such as financial services (including purchases, money transfers, savings and loan repayments), communications services (such as anonymized records of mobile phone usage patterns) or information services (such as anonymized records of search queries).

Before it can be used effectively, Big Data needs to be managed and filtered through data analytics—tools and methodologies that can transform massive quantities of raw data into "data about the data" for analytical purposes. Only then it is possible to detect changes in how communities access services that may be useful proxy indicators of human well-being.

It is important to note that, for the purpose of global development, "real time" does not always mean occurring immediately, but rather refers to information that is produced and made available in a relatively short and relevant period of time and within a timeframe that allows action to be taken in response, creating a feedback loop.

If properly mined and analyzed, Big Data can improve the understanding of human behavior and offer policymaking support for global development in three main ways:

1 **EARLY WARNING**
Early detection of anomalies can enable faster responses to populations in times of crisis.

2 **REAL-TIME AWARENESS**
Fine-grained representation of reality through Big Data can inform the design and targeting of programs and policies.

3 **REAL-TIME FEEDBACK**
Adjustments can be made possible by real-time monitoring the impact of policies and programs.

In general, the utility of Big Data for Development is maximized when traditional data sets and statistics are compared or correlated for context.

BIG DATA ANALYTICS
A type of quantitative research that examines large amounts of data to uncover hidden patterns, unknown correlations and other useful information.

Big Data analytics is not a panacea for age-old development challenges, and real-time information does not replace the quantitative statistical evidence governments traditionally use for decision-making. However, it does have the potential to inform whether further targeted investigation is necessary, or prompt immediate response.

FURTHER READING

Big Data for Development: Challenges and Opportunities (UN Global Pulse, 2012)
http://www.unglobalpulse.org/projects/BigDataforDevelopment

Big Data: The Next Frontier for Innovation, Competition, and Productivity (McKinsey Global Institute)
http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Big Data, Big Impact: New Possibilities for International Development (World Economic Forum)
http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development

Data for the Public Good (O'Reilly Media)
http://shop.oreilly.com/product/0636920025580.do

New Data for Understanding the Human Condition: International Perspectives (OECD)
http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.htm

RESEARCH EXAMPLES

To evaluate the effectiveness of harnessing Big Data for Development, UN Global Pulse has worked on several research projects in collaboration with public and private partners. Such proof-of-concept projects and prototypes demonstrate how Big Data analysis can be beneficial to the work of policymakers in different contexts—from monitoring early indicators of unemployment hikes to tracking fluctuations of commodity prices before they are recorded in official statistics.

Social media as early indicator of an unemployment hike

PARTNER: SAS

Can social media add depth to unemployment statistics?

Method:
1. Collect digital data (social media, blogs, forums and news articles) related to unemployment.
2. Perform sentiment analysis to categorize the mood of these online conversations.
3. Correlate volume of mood-related conversation to official unemployment statistics.

Findings: Global Pulse and SAS International found that increased social media conversations about work-related anxiety and confusion provided a three-month early warning indicator of an unemployment spike in Ireland.

Monitoring the evolution of food security issues through news media

PARTNER: COMPLEX SYSTEMS INSITUTE OF PARIS (ISC-PIF)

Is it possible to track thematic shifts in media attention through the automatic analysis of news articles?

Method:
1. Collect a corpus of news media on topics of interest based on keywords (i.e. food security).
2. Cluster articles in theme-based categories through semantic analysis.
3. Visualize the information both geographically and over time.

Findings: Thanks to the machine classification of millions of documents without human intervention, it's possible to visualize the shift in media attention related to a specific topic (in this case, food security) over time and thematically.

Real-time tracking of commodity prices: the e-bread index

PARTNER: PRICESTATS

Can mining online food prices provide real-time information on commodity price dynamics?

Method:
1. Use web scraping technologies to create a real-time price index ("e-bread index") by extracting bread prices from online supermarkets and retail websites.
2. Compare this e-bread index with the official food price index (CPI).

Findings: The relationship between web-extracted prices and official statistics on food prices (i.e. bread) proved to be closely correlated, allowing for price forecasting and additional real-time indicators of inflation activity.

Twitter and perceptions of crisis-related stress

PARTNER: CRIMSON HEXAGON

Can Twitter data be used to provide insight about how people perceive issues related to food, fuel, housing and the economy in Indonesia?

Method:
1. Develop a taxonomy of keywords related to food, fuel, housing and the economy, as well as keywords reflective of concern (i.e. "afford").
2. Classify Twitter messages into categories and quantify sentiment of relevant messages.
3. Correlate or compare the volume of keywords from Twitter against official statistics (e.g. unemployment or inflation statistics), and significant events.

Findings: The number of tweets discussing the price of rice in Indonesia closely matched the official inflation statistics, showing how the volume and topics of Twitter conversations can reflect a population's concerns in close to real time.

Detailed reports for these projects are available at **www.unglobalpulse.org/research/projects**.

CHALLENGES

There are important issues specifically relevant to Big Data for Development that must be acknowledged.

PRIVACY

Privacy, defined as the right of individuals to control what information related to them may be disclosed, is a pillar of democracy, and protections must be put in place to avoid compromising this basic human right in the digital age. Privacy is an overarching concern for anyone wishing to explore Big Data for development, since it has implications for all areas of work, from data acquisition and storage to retention, use and presentation.

In many cases, the production of data itself raises concerns, as people may be unaware of the sheer quantity or types of data they are generating on a daily basis, as well as that data they unknowingly consent to the collection and usage of without understanding how it may be used.

In this context, it is important to note that suitable legal frameworks, ethical guidelines and technological solutions for protected data sharing are at the center of efforts to leverage Big Data for development.

The Big Data analyzed for development purposes does not include personal or personally identifiable information. Instead, data sets are anonymized and aggregated to ensure full protection of individual privacy. And, after all, policy decisions by definition must be made to address issues at the community level.

The public sector cannot fully exploit Big Data without leadership from the private sector. With this in mind, the concept of "Data Philanthropy" has emerged as a partnership by which private sector companies share data for public benefit, taking the initiative to anonymize their data sets and provide them to social innovators to mine for real-time insights, patterns and trends.

For more information on Data Philanthropy, visit http://bit.ly/dataphil2013.

DIGITAL DIVIDE

Although the data revolution is unfolding around the world in different ways and at different speeds, the digital divide is closing faster than many had anticipated. The availability and types of digital data, however, differ from country to country. For instance, countries with high mobile phone and Internet penetration rates will produce more data directly generated by citizens, while nations with large aid communities will produce more programme-related data. Data also varies between age groups, economic income brackets, gender and geographic location.

These types of biases must be addressed in the way Big Data for Development research projects are designed and conducted, and particular attention must be given to the countries that are producing less data and/or have less capacity in data analytics to avoid adding new facets to digital divide.

ACCESS

Although much of the publicly available online data has potential utility for development purposes, private sector corporations hold a great deal more data that is valuable for development. Companies may be reluctant to share data due to concerns about competitiveness and their customers' privacy. It is therefore critical to ensure a legal framework that defines rules for privacy-preserving analysis and protect the competitiveness of the private sector companies willing to share data. For example, aggregating data belonging to companies operating in a similar sector in a data commons may prevent the attribution of a certain data set to a specific company.

ANALYTICAL CHALLENGES

The process of mining Big Data for Development (using Big Data analytics techniques to extract relevant information) contains certain analytical risks that may reduce the accuracy of the results. Analyzing Big Data for development poses different challenges that are in part methodological, or related to interpretation accuracy, methods of analysis, and detection of anomalies.

DATA SCIENTISTS VS DATA ENGINEERS

Data scientists clean, organize and analyse data through different techniques looking for patterns that can generate actionable information.

Data engineers design and develop information systems to scrap, collect, transfer and sort data.

Some methodological challenges may include:

- **Sentiment analysis (or opinion mining):** Refers to the study of emotions and opinions expressed in digital messages and translating those sentiments to hard data. Quantifying moods and intents is difficult, and obstacles such as slang, sarcasm, hyperbole, and irony may impede data analysis.
- **Text mining:** Such as going beyond sentiment analysis to extract keyword and events, text mining faces difficulties the true significance of the statements in which the facts are reported.
- **Falsification:** Data can be false, fabricated with the intention of providing misleading information.
- **Perceptions versus facts:** Perceptions aren't necessarily accurate and may differ even significantly from actual facts. For instance, this happened with Google Flu Trends, an analytics platform that was meant to predict actual flu, but instead proved useful only for general public health surveillance. Without recognizing this key insight, doctors using Google Flu Trends may be inclined to overstock vaccines or misdiagnose their patients.
- **Sampling selection bias:** The people who use mobile or digital services may not be a representative sample of the larger population considered.
- **Apophenia:** Seeing patterns and correlations where none actually exists is a risk, and the massive amount of data available for analysis intensifies the search for interesting correlations that may not actually exist.
- **Correlation does not mean causation:** Even where an actual correlation is found, it doesn't necessarily signify that there is a causation link between the data and theory and context are relevant to reduce the risk of self-fulfilling prophecies.

CONCLUSIONS AND OPPORTUNITIES

Beyond the availability of raw data alone, and beyond the intention to utilize it, there needs to be capacity to understand and use data effectively. Big Data for Development is about turning imperfect, complex and often unstructured data into actionable information. In order to do so, there are specific requirements in terms of skill sets and technologies, aside from getting access to the right data sets.

Despite the many challenges that Big Data for Development presents, understanding the growing amount of digital information human communities produce can be invaluable in providing them with support and protection.

The technologies for data analysis are constantly improving and developing, and it is difficult to share a comprehensive taxonomy of tools with related costs and benefits. Global Pulse is, however, available to provide support and direction to UN offices willing to embark in Big Data analysis and to advise them on the most effective technologies and skills they should seek according to their specific analytical needs.

The question is no longer if Big Data can provide insights useful to global development and resilience, but how.

At this stage, it's fundamental to focus on implementing new norms and ontologies for the use and sharing of data as well as forming innovative public-private sector partnerships to facilitate analysis and research. This will determine whether development organizations and policymakers will be able to use Big Data to its full potential to enhance the public good.

ABOUT GLOBAL PULSE

Global Pulse is a United Nations innovation initiative of the Secretary-General exploring how new, digital data sources and real-time analytics technologies can help policymakers gain a better understanding of changes in human well-being and emerging vulnerabilities. Through strategic public-private partnerships, innovative analysis and open source technology development across its network of Pulse Labs, Global Pulse is developing approaches for applying Big Data to 21st century humanitarian and development challenges.

The three-fold project strategy includes:

1. **Research & Development:** Conducting research to discover new proxy indicators for tracking development progress and emerging vulnerabilities in real time, and assembling a toolkit of technologies for analyzing real-time data.

2. **Big Data partnerships:** Forging partnerships with companies, organizations, researchers and academic institutions that have the data, technology and analytical expertise needed for Big Data for Development projects and advocacy.

3. **Pulse Lab network:** Establishing an integrated network of country-level innovation centers that bring together government experts, UN agencies, academia and the private sector to prototype and pilot approaches at country level and support successful adoption.

For more information please visit www.unglobalpulse.org.