Two contributions about big data

A draft – Frédéric Lefebvre-Naré – November 6th, 2013

1 — Big data and us: revolution was just around the last corner

1a — What big data are, and are not

i. Big data are our daily experience: basically, they are the inputs and outputs of our phone, stored forever.

So we experience them as citizens, employees, consumers, travelers, readers and so much more:

- Billions of individual opinions are recorded and readable, forever or almost so, on the internet.
- Within factories or in every industry, myriads of data describe what's happening — machines and processes are not any more driven by physical actions, but by changing of some data on some electronic memory.
- When we visit a web shop, it may compute within microseconds a product we are expected to be interested in, based on our previous purchases or the other pages we visited — compared with the whole database of other consumers behaviors.
- Our daily or international travels are tracked at many steps, and anticipated by processing such data.
- Local computers and servers remember the web pages we read, the videos we launched and how many seconds we stayed there, the queries we made.

Maybe we did not really notice, as long as these experiences remained distinct from each other. Now that we initiate all of these experiences on our same phone[1] — typing in or reading opinions, buying products or services, being geo-located second by second even in our "private" home, and making "private" calls, and sending "private" messages, and so on — big data are our daily experience.

ii. Thus, **big data are a simple fact**. They are just another logical consequence, as the (formerly known as smart-)phone is, of one same fact: technological change in electronics divided the unit cost for collecting, storing and transmitting raw data by 2 every year, since more than 20 years[2]. Data became inexpensive, so they became big.

iii. Thus, **big data are pervasive, too.** When every behavior turned digital, it turned tracked, at least potentially. Education and communication, agriculture and industries,

[1] Daniel Kaplan specified the transformation of the local computer into an *initial*, not any more a *terminal*: April 24th, 2002, — http://www.internetactu.net/2002/06/24/du-terminal-linitial/ (in French)

[2] Matthew Komorowski and Ray Kurzweil, quoted in "L'écosystème du 'déluge de données'" (The data deluge ecosystem), Emmanuel Bellity and others, "Variances" n°46, February 2013, http://www.ensae.org/docs/2013164614_v46-focus1.pdf

war and police, tax and welfare, culture and policymaking itself — none of them stayed out of this change.

iv. Big data are all over the world. They bridged an information gap. We may have experienced this kind of gap in the 90's and still the 00's: companies in rich countries deployed global software to track and command their operations (ERPs), and the citizens of these countries had widespread access to the internet; while citizens and companies of poorer countries still lived most of their lives in the analogic age. Now, the phone is nearly the same everywhere, and it registers as many data. If you can track the European lorry driver to check his or here timetable, you can also track the employee of a West African NGO to check his or her own activities.

v. Yet **big data did not change, by themselves, physical reality**. You can ignore them as well — just switch off your phone. Oil companies still drill with real drills, teachers still teach with their own voice, the police still has to slap physical cuffs on robbers.

vi. Big data are just the persistent digital layer, or digital skin, that covers almost every physical reality, making it visible from everywhere at every time, sometimes actionable from everywhere, globally interconnected. Realities, and especially the realities we have to evaluate, the efficiency and impacts of policies, did not change — but they are not any more bare, lonely or hidden.

vii. So, **big data crunched the world**. They brought it back to the global village McLuhan announced. Privacy is over, at least from the point of view of the Matrix — this place that would centralize all data regarding us; our phone, after all.

viii. "Big data" are not a new idea or paradigm. It has been widely written that science fiction did not anticipate the worldwide web; but science fiction, or culture at large, described big data under all perspectives. From Big Brother's eyes and ears in George Orwell's "Nineteen Eighty-Four" (1949) to the pervasiveness of images in Wim Wenders' Lisbon Story (1994), from Philip K. Dick's "Minority Report" (1956) to Andrew Niccol's "Gattaca" (1997), the public hasn't been short of initiations to the Brave New World of big data.

ix. And even **the collection, storage and use of big data are not that new**. Many of us have this definite feeling of *déjà vu, déjà entendu* when we read and hear about big data. These stories about WalMart predicting this and this customer behavior, didn't we read quite the same in the mid 90's, the era of *data mining*[3]? Didn't the debate on *open data* take place 30 years ago[4]? This new job of *data crunching* that is required to transform mountains of raw data into sense and action, wasn't it already called so in the 70's[5]?

---

[3] "Beer and Nappies -- A Data Mining Urban Legend", on Donald Fisk's website http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html (retrieved on November 6th, 2013), this page is as of December 15th, 1997, according to Google index, and sources quoted are dated 1997.

[4] In France: under the Raymond Barre cabinet, 1976-1981.

[5] For example, in this presentation by Margaret D. LeCompte and Judith Preissle, « Data Crunching, or, What Do I Do with the Five Drawers of Field Notes? », at a meeting of the Council on Anthropology and Education of the American Anthropological Association, 1977.

x. And **big data don't track your mind.** They track behaviors. So far, they don't track thoughts; that would be another revolution. Big data respect quite well that old distinction from the Roman Catholic canon law, between the *internal forum* and the *external forum.* Big data respect it much better than they would respect a distinction between private spaces or times, and shared ones: isn't your phone switched on in both cases?

Okay — but when your behavior is tracked, are your thoughts really unknown?

When AOL published, as a research material, 20 million anonymous web search queries, that material told much, if not everything, of Ms Arnold's life, « a (by then) 62-year-old widow who lives in Lilburn, Ga. »[6]. But if you keep your thoughts for yourself and hide them from your computer and your phone, if your deeds are disconnected from your *internal forum*, then you are untracked. As Tracy Chapman put it, "All that you have is your soul"[7].

xi. Whether it is about behaviors or thoughts, **big data can't know anything;** just as water dams can't fly. Their electricity can power airplanes; data can feed algorithms that will generate useful information.

So, big data make, by themselves, little knowledge.

Crude statistics from big data are often highly disappointing:

- heterogeneous data collection makes processing, not to say interpretation, a nightmare;
- outliers make mean values irrelevant;
- geographical maps of anything make beautiful pictures and — so what[8]?

xii. Therefore, **big data aren't used so often to support decision making, including in evaluation**. When a decision maker, or an evaluator, looks for knowledge, why wouldn't he or she ask directly some person who knows, instead of consuming working time on fishing in oceans of raw data?

And that's where the revolution starts. That's where much is on the way.

If data oceans are not new (their volume just doubles every year[9]), building dedicated tools, methods, algorithms, science, in order to fish in such oceans, is rather new.

1b — What the big data revolution is about

The big data revolution is about becoming effective in handling data oceans, in "turning the data deluge into decisions"[10].

---

[6] http://www.nytimes.com/2006/08/09/technology/09aol.html

[7] In her « Crossroads » LP, 1989.

[8] The post « sentiment + agrégat + jolie carte = arnaque » (« Sentiment + Aggregate + Nice Map = Scam ») by Dominique Boullier on Feb 18th, 2013, http://opiniONsentiment.hypotheses.org/44 (in French)

[9] Or so. « 90% of the data in the world today has been created in the last two years alone », according to IBM, http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html (retrieved on November 4th, 2013).

[10] The motto of the Paris Big Data event 2013, http://www.bigdataparis.com/2013-fr-exposition.php

It involves a technological move, and a scientific one. Both moves started in the early 2000s.

Data Technology

During the 2000s, Yahoo!, Google, Amazon and other huge IT companies created algorithms and software able to handle these amounts of data[11]. Handling these data means, in practice, answering the "3V" requirements the META Group (now Gartner) specified in 2001 for the e-commerce data[12]:

- [processing] not only data [of large] **volume**,
- but also [providing] **velocity** — relevant data should be retrieved, processed, and used, at the speed of the user's click,
- and reconciling data **variety —** because big data are heterogeneous data, issued by independent sources.

The velocity issue highlights another important side of big data. Information made from raw data is not always expected to be read or understood by human people. Much of it is sent to algorithms only. So, the issue of effectiveness in the use of data is wider than the issue of "what can I do from that?" as a decision maker or evaluator; it's about controlling algorithms, their business models as well as their social impact at large[13].

For example, in the Stockholm county (Sweden), 9,500 volunteers receive a text message and a computer generated phone call if they are located within 500 m from a suspected out-of-hospital cardiac arrest (OHCA). In 44% of reviewed OHCAs, "one or more volunteers reached the location prior to ambulance arrival".[14] This is big data: heterogeneous data oceans (geo-location data, phone calls data including emergency calls, plus a small file of volunteers contact data), interfaced with real life events (OHCA, call reception, sending an ambulance) but processed much faster than human decision could do.

Another, and widely known, example, is the Amazon recommendation feature[15]. It was easier to do, in a sense, as online recommendations require no connection with the physical world.

And for this reason, e-commerce may have been a step ahead; in many industries, such as videogames, consumer retail, or health, being able to access data in a practical manner is quite new, starting in the 2010s, say 2012 or 2013.

---

[11] The MapReduce algorithm (2004), implemented in many NoSQL database management tools, such as Apache open-source sofware framework Hadoop (2005).

[12] Doug Laney, February 6th, 2001, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[13] « Surveiller les algorithmes » (Watch the Algorithms), a post by Hubert Guillaud on Internet Actu, November 2nd, 2013, http://internetactu.blog.lemonde.fr/2013/11/02/surveiller-les-algorithmes/ (in French)

[14] « Use of Mobile Phones to Dispatch Volunteers to Perform Early CPR, the First 303 Cases », David Fredman and others, communication to the Boston Medicine 2.0'12 congress, http://www.medicine20congress.com/ocs/index.php/med/med2012/paper/view/1223

[15] Patent on http://www.google.com/patents?id=dtp6AAAAEBAJ , May 7th, 2001.

And then arises another research perspective: what for? And then: how?

Data Science

Collecting, analyzing and using raw data raises lots of issues.

- About the data deposit itself: What are relevant information requirements? How can data collection be optimized for further — and yet unknown — uses? What is the best balance between closed and open-ended variables? How to assess the quality of a data source? How to define an optimum in "data cleansing"?
- About mining in this deposit: what is diamond and what is gangue? For example, how to detect the most promising variables in a database? What are the most effective algorithms to find connections between variables? How to detect false figures? How to manage the many sources of bias?
- About polishing and exposing, or industrially using, the outcomes: How to deliver results in order to optimize their use for decision, and reduce the risks of mishandling, wrong (automated or human) decisions? How to emerge from the global information overflow?

Yes, dealing with these issues is the very know-how of the statistician. More precisely, that's the kind of questions "statistical methodology" was intended to deal with, since the 18th century at least.

But the environment changed — from steppe to ocean. The statistician methodologist strived hard to make information grow where there was ignorance; the data scientist will know how to transform data oceans into informational fried fish[16].

Data science is expected to deal with these issues in that new environment.

Data science exists, in the sense that it has its publications, its conferences, some higher education programs. Yet it is still described as "an interdisciplinary subject"[17], meaning that this "science" is in its prolegomena, still waiting for its Hippocrates, or its Alan Turing. As was the case for medicine or computing, theory will change technology, but people used technology before science got solid foundations.

As long as data are the material of which information is made, the big data technological revolution, and their emerging scientific revolution, will change all jobs and processes dealing with information.

So, big data are not only a growing wave, coming from the deep 70's and described many times since. They constitute a new layer in the world, the persistent and interconnected digital descriptions of physical realities and human behaviors. They generate not only a new access to pre-existing physical and human realities, but also new business models, new physical and human realities. And they will change the evaluation discipline, too.

[16] This new approach to data science, as another perspective than the traditional one ins statistics, was introduced by William S. Cleveland, "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics', published in Volume 69, No. 1, of the April 2001 edition of the International Statistical Review / Revue Internationale de Statistique.

[17] http://en.wikipedia.org/wiki/Data_science, retrieved on November 5th, 2013.

2 — Big data, public policies and evaluation: 10 opportunities or challenges

This second part delivers very preliminary first thoughts by the author about what big data may change in public programs and in their evaluation.

The paper uses a 10-point grid, with which the author reviewed some business models and business practice of companies. It's applied here to public policies, on the basis of the author's experience in the 90's-2000's as co-CEO of evalua, then the #2 evaluation consultancy in France.

The structure for each of the 10 points is:

- The big data fact,
- An example, generally from the world of for-profit industry (sometimes from the author's own experience),
- *Perspectives for public policies and programs,*
- Examples of questions for reviewing public programs.

1. The economics of public service change: big data bring a cheap deposit of information.

The big data opportunity is about using a deposit that is, in part, already there, and sometimes available for free.

- Market and opinion research tools like net-conversations®[18] use the wealth of opinions and experiences people published on the internet, to discuss questions that would, beforehand, have required tailored qualitative surveys such as focus groups or in-depth interviews. That saves time, money, and brings new accuracy, as spontaneous expression is safe from researchers' preconceptions!

Does the program make the best use of existing data, be they stored within the government agency in charge, or accessible elsewhere? Has the economy of the project been revamped to save, on other resources (including fresh information collection), what can be found on existing sources?

- Review the true activity of (subsidized…) firms or NGOs via individual staff profiles on LinkedIn,
- Validate addresses via GoogleMaps / GoogleStreetView,
- Check public events calendars before setting the date of local similar events…

2. A newly found deposit makes the fortune of shovels and pickaxes suppliers

The business of data ate billions of investments during the 10 last years, and produces around 36 billions US$ turnover in 2013, according to estimates by the Gartner Group. That may change the business environment of many public policies.

[18] www.net-conversations.fr . The author is the CSO of net-conversations®.

- "The Green Button initiative is an industry-led effort that responds to a White House call-to-action to provide utility customers with easy and secure access to their energy usage information in a consumer-friendly and computer-friendly format[19]." Sure, it's about being green, but in particular, about aligning the efforts of industry stakeholders around a "common, machine-readable (data) format".

Did the program managers get in touch with relevant stakeholders in the data business? Are they aware of applications, either free, on sale or under development, that could change the way the public service is delivered to citizens?

- If the service delivery can be located on a geographical map: is it tracked / analyzed via some OpenStreetMap application?
- If the outputs of some local service should be visible to citizens: does the service use "Fix my city"?
- Can the Googles, Facebooks or others find an interest in pushing, subsidizing, feeding the program in any way?

3. Big data allow mass customization of service…

That's the less new side of big data, and more precisely, the core promise of "data mining" in the 1996's. But it's still relevant!

- When a client of perfumes retailer Sephora wonders what she might buy, the floor staff may help her — well, the staff's iPod touch may help her. By scanning the client's loyalty card or typing her name, the staff member will see customized recommendations appear on the screen, based on the previous purchases of the same client, and correlations computed from a 8-million-clients database[20].

Is the way the service is delivered, optimized under the (beneficiary, customer, citizen) knowledge that the databases include?

- Is the information sent to people, filtered on an "only if relevant" basis?
- Do social welfare services target their individual controls on the most effective "predicted detected frauds amounts per working hour" basis?
- Are new residents of an area quickly informed of the local services that would be most relevant to their individual situation?

4. …and this customization gets in touch with people within their place and time

Big data added a new perspective to traditional data mining: the search for bridging variables, meaning, concepts or units that could be applied to distinct data. And the most valuable discovery big data practitioners made is a return to Emmanuel Kant: all the people, and much of what happens to them, are somewhere at each moment. Written otherwise: space-time coordinates are a quite effective way to unify databases as well as to display information.

---

[19] http://energy.gov/data/green-button , retrieved on November 6th, 2013.

[20] http://www.lefigaro.fr/societes/2011/11/16/04015-20111116ARTFIG00665-sephora-lance-le-conseil-d-achat-sur-mesure.php , November 16th, 2011. The author provided consultancy to design the recommendation algorithm.

- "Predictive policing" automates the route of police field patrols, according to the probabilities of wrongdoing at each place and time[21]. These probabilities take into account the most recent registered events, esp. robberies, and a time-space correlation between these robberies, measured on data tracked in the long term.

Does the service reach the right persons at the right time in the right place?

- Does the service provider know who / which kind of people / with which motivations, use the service at each moment of the day / week / year? Can service delivery be optimized or smoothed under this respect?
- Are "pushed" actions located at the right place and time, according to the predictable outputs?
- Can and does the mobile app of the service, take advantage of the geo-localization of the user?

5. Big data add power to the monitoring of internal processes

Big data are not only about data storage and processing, but also and primarily about data collection, about sensors. Each time a physical process or an online process takes place (that means: quite often), tracking this process can bring value.

- "On-board sensors" in trains "and predictive analytics have the potential to make train maintenance more timely and more efficient. Problems could be detected before they cause disruption, easing pressure on the rail network", according to a research by consultancy Amadeus[22].

Are the physical components of a program — machines, vehicles…— tracked the most effective way? Are online actions tracked in order to improve service targeting and delivery?

- Are the vehicles used for the program continuously geo-located, and are these data used to evaluate the delivery and efficiency?
- Are the machines a program subsidized, continuously monitored to allow predictive maintenance and avoid costly emergency repairs?
- As far as the program delivers physical products (such as drugs, tools, agrochemicals…), are local inventories continuously available? Does the program use an effective inventory forecast tool, in order to improve availability at lowest cost?

6. Big data can bring added intelligence to decision-makers, stakeholders and evaluators

Intelligence for decision-makers is not a core big data issue, even if it's the perspective under which "big data delusion" is most often raised. But if intelligence could be added, the opportunity should not be neglected. And the evaluator can be one of the users of information processed from big data.

---

[21] "Predictive Policing: Using Technology to Reduce Crime", Zach Friend, crime analyst with the Santa Cruz, Ca., Police Department, April 2013 http://www.fbi.gov/stats-services/publications/law-enforcement-bulletin/2013/April/predictive-policing-using-technology-to-reduce-crime

[22] http://info.amadeusrail.net/big-data-crossroads-download , retrieved on November 6th, 2013.

- Standard "smart meters" of electricity measure consumption every 10 or 15 minutes. That's enough to determine if you use more or less power than your neighbors, and when; but not enough to detect which appliances you used, when and for how long. Five-second metering on a representative sample of households, as enforced by WattGo-Powermetrix[23] unleashes a broad potential for detecting and statistically analyzing uses, and therefore, for improvement in long-term forecasts of demand.

After decision-makers, stakeholders, or evaluation commissioners, asked questions (for example in the terms of reference of an evaluation assignment), three checks should be made for each question:

a. Is there some behavior tracking that would bring answers to that question? For example:
   - energy consumption tracking, if the program promoted energy saving solutions,
   - taxpayers data, as far as economic development is concerned,
   - professional profiles on LinkedIn, Facebook, hi5…, to evaluate impacts of higher education programs…
b. Do some stakeholders already express views on the evaluative question? Is there a past or ongoing debate on Twitter, on bulletin boards, or via broadcasted emails?
c. What should the answers look like? Is there any kind of information visualization (*dataviz*) that would make conclusions more transferable, and foster ownership by the recipients of the information?

7. Big data can bring added intelligence to users and citizens too

That's a big difference to Orwell's "1984": information flows back and forth. An effective business model in big data is often the one that provides shared benefits to the central service provider and to the end user.

- Google Maps' traffic estimates are, of course, a geo-localized service — it shows you first the situation around your position. They are also a crowd-sourced service, based, in part, upon the geo-locations of Google Maps for Mobile phone users[24]. "If you find the live traffic info in your area isn't particularly accurate, you can help improve that by turning on My Location in the Maps for Android app", as the columnist at Engadget comments[25].

Within a program, are opportunities taken to distribute relevant information to each potential user / customer / citizen where and when he or she can use it?

---

[23] http://www.powermetrix.fr/en ; retrieved on November 6th, 2013. The author is the CSO of WattGo Powermetrix.

[24] "The bright side of sitting in traffic: Crowdsourcing road congestion data" Google official blog, August 25th, 2009, http://googleblog.blogspot.fr/2009/08/bright-side-of-sitting-in-traffic.html

[25] Terrence O'Brien, http://www.engadget.com/2012/03/29/travel-in-traffic-estimates-return-to-google-maps-promises-not/ March 29th, 2012.

- Professional education: do the actual or potential students receive all relevant information about the careers, challenges, required skills, of their predecessors?
- Events programming, or service delivery at specific hours (library, sport training, medical, and so on): are all potential users connected and does the service send SMS at each schedule change, or to recall each appointment?
- If the service is delivered for a specific area (transportation, culture, local information…), is some advertising or information displayed on the phones of people who happen to enter the area for the first time?

8. Crowd-sourced data transform collective behaviors into information

Many program managers would be happy to anticipate needs of beneficiaries. If big data don't always provide anticipation, they can often provide early warnings, generally through crowdsourced information.

- Statistics of queries on Google are a faster way to detect flu epidemics, than (previously used) reports from doctors' offices[26]. But this source, Google Flu, includes bias, too[27]: "at the flu season's peak in mid-January, Google's algorithms results were double the actual estimates by the Centers for Disease Control and Prevention: 'Several researchers suggest that the problems may be due to widespread media coverage of this year's severe U.S. flu season', Declan Butler wrote in Nature."

How far does the information system design within the program, take advantage of spontaneous data emission by people?

- What use does the program make of search histories (or of relevant Adwords analytics) that could tell what topics are hot for potential users?
- If the program management detected some risks bound to one-time events: did they enforce a Twitter posts tracking (if relevant) to be more quickly informed if the event occurred?
- Are (relevant) communications, comments, reports… by beneficiaries, local agents or partners, gathered in a repository allowing full text search, noteworthy for evaluation purposes?

9. Open data can make added value

"Open data" are one of the keywords with some connection to "big data", even if "open data" are often small data. Open data are those data an organization makes public under some form: individual data, anonymous data, aggregated data. Many public authorities, and many stakeholders around them, found that opening data brings, in one hand more transparency and accountability, and on the other hand, opportunities for new data-based services the government wouldn't have provided alone.

[26] See for example (in French) "Big Data : le grand déséquilibre ?" Hubert Guillaud on Internet Actu, October 4th, 2012, http://www.internetactu.net/2012/10/04/big-data-le-grand-desequilibre/

[27] Nick Bilton, February 24th, 2013: http://bits.blogs.nytimes.com/2013/02/24/disruptions-google-flu-trends-shows-problems-of-big-data-without-context/

- In poor neighborhoods of the Third World, some inhabitants may wonder if the local authorities will take care of their streets, of their safety, or will plan services supply in their area. Being registered, with names and individual photographs, may be reassuring[28]. "For many years, Kibera, the informal settlement on the outskirts of Nairobi, was believed to be the largest slum in Africa. Despite being one of Africa's best-known settlements, Kibera has always been a gray area: nonexistent on maps, including Google maps. This is now changing, thanks to the Map Kibera Project, which has produced the first public digital map of the slum. Using open-source technology [OpenStreetMap] to document the physical landscape of the settlement and its resources"[29].

Does the program take the best advantage of already opened data, either in its jurisdiction, or in other jurisdictions on topics similar to the ones the program deals with? Do the program stakeholders push public authorities, or other data owners, to make them public as far as that could improve the services to the beneficiaries? Does the program open its own data and contribute to the information ecosystem?

- Much information can be found, and also provided, on information-sharing platforms, including Wikipedia for knowledge, or OpenStreetMap for geo-located data.
- Added value can be brought to users and citizens through data accessibility, in the fields of elementary or further education, of public transportation, of environment monitoring, and so on.
- Do the evaluators take into account the wealth of public data that could provide reference marks to assess the impacts of programs? For example, compare the development of their target area, with the development of other areas with similar profiles. Sometimes it requires a phone call to the National Statistics Institute to open the door to these data, or to find the door that was already open… But often, it's worth the effort, because getting these reference points changes conclusions.

10. Closed data can make added value, too

Most of the big data are not open, after all, and most of them are not really used (which is unsurprising, as their collection and storage is less and less expensive). Data owners should nevertheless not sleep on deposits of potential value!

- That's the view of commenter Marcos Menendez on Doc Searls weblog[30] — in a discussion about privacy and data ownership: "there will be more and more services around data, based on business models where data is a source of revenue and therefore trying to convince people to use them in exchange of people's new oil. Cloud services, wearable computers, glasses, whatever. (…) In a "privacy rules" world, Facebook and services alike that have been built around the value of people's data would not exist anymore. (…) In a world where people

[28] The author noticed such reactions during an evaluation of urban social projects in Nouakchott, Mauritania, in 2002.

[29] Dinfin Mulupi, January 27th, 2011, http://www.audiencescapes.org/putting-kibera-map-kenya-OpenStreetMap-AMREF-Plan-International

[30] Post: "Thoughts on privacy", August 31, 2014. Direct link to the comment : http://blogs.law.harvard.edu/doc/2013/08/31/thoughts-on-privacy/comment-page-1/#comment-320055

(citizens) are true owners of the data (…), people that produce more valuable data may even ask Facebook a share of the revenues generated with it or even ask a fourth party to exploit that data and share it with Facebook and the user."

Sure, "the value of closed data" may be a more intuitive approach for private businesses, than for public programs and services. But the evaluative question is the same in both cases: is there value, and for whom, in the data the program is collecting? Can the ownership of these data bring competitive edge, or added value to the beneficiaries? From the opposite point of view: can the collection of (closed) data by other service providers or authorities, threaten the sustainability of the evaluated program?

- A first issue for evaluation is the data security issue: who can access to the data the program collected? Are the risks of leakage, if relevant, countered?
- What competitive advantage does the holding of data give to the program (or to competitors)? That may be especially relevant for microfinance institutions, agricultural projects, education programs within the State-run sector…
- Does this deposit of data open the way for other kinds of services? Including academic research, better design of similar programs in other regions, individual facts-based recommendations applications…

Not all these points will be relevant regarding a given program and environment. More discussion and examples would help to make this list more specific, and more clearly oriented towards evaluation. But anyway, on specific programs, enforcing such orientations will require time and competencies.

As columnist Steve Lohr put it in the New York Times in 2012, *« Big Data does seem to be facing a work-force bottleneck* ». Yes, big data skills may be missing sometimes; then, training and hiring are adequate solutions, while ignoring the issue is probably not. Big data are here to stay.

As the same Steve Lohr put it in the same column[31], *« Big Data is great. But so is intuition.* » Sometimes, senior program managers or evaluators will feel that intuition skills are missing among junior staff, even if they were raised in the era of big data. Then, training and hiring are adequate solutions, while ignoring the issue is probably not. The need for intuition is here to stay, too.

[31] http://www.nytimes.com/2012/12/30/technology/big-data-is-great-but-dont-forget-intuition.html