The value of data

Thierry Berthier Mathematics Lecturer, Université de Limoges

May 2014 –Article n° IV.3

Translated from French

A world of data

The amount of data created in the world has increased from 1.2 zettabytes (1 ZB = 10^{21} bytes) in 2010 to 1.8 ZB in 2011, 2.8 ZB in 2012 and it should reach 40 ZB in 2020. It is estimated that the global volume of data doubles every 18 months. For example, the social network Twitter produces 7 terabytes daily (1 TB = 10^{12} bytes) and Facebook generates over 10 TB every day. The large radio telescope Square Kilometre Array (SKA) that will be operational in 2024, will produce more than a billion gigabytes of data per day or between 300 and 1,500 petabytes each year (1 PB = 10^{15} bytes). The Large Hadron Collider (LHC) of CERN produces about 15 PB of data annually. The volume of data produced by such systems will soon exceed that produced by humans.

To deal with this deluge of data, Big Data technologies evolve very quickly across three classic dimensions called the "three Vs": volume, variety, and velocity. These are sometimes complemented by two other Vs: visualization and veracity. The exponential increase in the volume of data to process necessitates the creation of increasingly high-performance "Data Centers." Variety describes the heterogeneity of the often unstructured raw data; it is usable by a complex algorithmic infrastructure capable of interpreting the information regardless of its format. Velocity, in turn, refers to the need for increasingly high processing speeds related to real-time data analysis ("in-memory" technologies) and other "high frequency" digital systems. The sixth V could affect the value of data, whether linked to a big data or not.

Is there a definitive definition for the value data that is compatible with the environment in which it is interpreted or, on the contrary, should we narrow our definition to take into account local and instantaneous data types?

Using two recent examples, we show that the value of data is above all volatile and strongly dependent on time and on the context in which it is evaluated. In the second half of the paper

we propose a systemic approach to the problem based on a reduced formalism that helps to define the instant value of data in relation to the algorithm that interprets it.

I -The price of data in two examples

1.1. The tweet that cost $136 billion dollars

The Syrian Electronic Army (SEA) [1] is a cell of hackers that appeared at the beginning of the Syrian conflict in 2011 and supports the Bashar al-Assad regime. Over three years it increased the number of digital aggressions it carried out against targets identified as enemies of the Syrian nation. Its primary mission is establishing their message about the Syrian conflict through a structured infrastructure of counter-information deployed on social networks (Facebook and Twitter) and on the internet (sea.sy website). SEA has carried out more than 200 cyberattacks against a variety of western digital targets such as media (TV, major US and European newspapers), government sites (European, American, Arab, Israeli), and large organizations such as Microsoft, PayPal, Facebook, Twitter, and the US Army, etc.

The attacks most often use social engineering tools like hacking (through phishing) and account takeovers, where the attacks are carried out. Victim sites are regularly "defaced" by redirecting them to a similar page containing a counter-claim and or justification of actions. Depending on the level of attack, sometimes SEA captures very large databases. For example, during the attack against the Forbes website in 2014, over one million account credentials were hacked. The attack against PayPal-UK allowed them to get their hands on an online payment service database. SEA sometimes uses denial-of-service (DDoS) attacks or injections of malicious code to collect more sophisticated information especially on Syrian rebels that are intelligence targets. On April 24th, 2013 SEA attacked the Twitter account of the Associated Press (AP). It temporarily took control and at 13:07 published the message, "Breaking: Two Explosions in the White House and Barack Obama is Injured." Some 1.9 million Associated Press Twitter account subscribers received the fake message posted by SEA and it was considered genuine. Financial markets reacted almost immediately. Between 13:08 and 13:10 the main index of Wall Street, the Dow Jones (DJIA), lost 145 points and wiped out the equivalent of $136 billion dollars (€105 billion euros) in part because of high-frequency trading (HFT), which interpreted and "reacted" to the false tweet. Shares of Microsoft, Apple, and Mobil lost more than 1% almost instantly. A few minutes later, the Associated Press regained control of its account and immediately published a tweet announcing that the message was a fake and that the account had been hacked. Immediately, the Dow Jones recovered from the drop in points and quickly returned to normal. The brief wide-spread belief in the false message published by SEA was sufficient to alter a market index with strategic importance. Automated high-frequency trading, which is able to place orders in microseconds, changed the methods decision-making, pushing human control to the end of the operation.

The automatic validation and inclusion of false information can therefore have a significant impact in an interconnected environment. One can therefore question the real value of the SEA tweet as piece of data that was believed to be true for an instant then denied a few minutes later. It is clear that the value depends on time as much as it does the validity that it has been granted, and finally on the context in which it is interpreted. We must also agree on the meaning of the word value: is it the best selling price of the data from a seller to a buyer or should we take into account the "impact value" of this data in the context or in the broader environment? In the case of the false SEA tweet, the impact value is high since the cost of turbulence experienced in the markets while the data was considered valid should be taken into account.

1.2. The sale of Microsoft customer data to the FBI

In January 2014, the highly active Syrian Electronic Army attacked the official Microsoft website several times and managed to get hold of several databases, emails and Microsoft invoices regarding the selling of "client" data to the Federal Bureau of Investigation (FBI). On January 21st, 2014, SEA published on its website copies of many Microsoft invoices sent to the FBI as well as listings of the personal data sold. These were concerning Outlook and Skype users and contained identities, usernames, IP addresses, account names in hotmail.com and passwords. According to the invoices published by SEA, the unit cost of the user data is between $50 and $200 dollars, depending on the transmitted content.

The bill for only November 2013 from Microsoft reached $281,000 dollars. That of August 2013 reached $352,000 dollars. A dataset that contained a Microsoft product user's password was billed at $200 (the maximum cost).

Note that this type of transaction is perfectly legal in the context of a rogatory body in a criminal investigation. Other invoices were billed by Microsoft alongside private foreign companies based in South America in the customer sales data framework. These transactions concretely help to answer our questions about the value of data.

As part of the management a massive volume of data, Microsoft manages to define the data unit price based on content and format. In this case, the impact value of the data is not taken into account by Microsoft in the determination of the price and only the cost of processing and structuring of data is taken into account.

These two examples of cyberattacks carried out by SEA highlight the variety of contexts that add value to data and, finally, the real difficulty finding a canonical definition for the price of data. A systemic approach can help to define the parameters and components that underpin the value of a given piece of data.

II A Systemic approach

We propose using a minimal formalism for finding a functional definition of the instant value of a piece of data in a context and relative to the algorithm that uses this data.

Definition 2.1. Data and a binary word

- Data is represented by a finite set of binary words.
- A binary word is a finite binary sequence, that is to say, a finite sequence consisting of 0s and 1s that is readable to a computing system.
- This allows one to be free of the original data type (text, image, sound, video, signals or measurements from sensors, etc.). All the information contained in the initial piece of data is translated into binary words in a format compatible with future algorithmic processing.

Notation 2.1

- Note that D is the data defined by $D = \{M_1, M_2, ..., M_n\}$ where M_j are binary words, with $M_j = b_1b_2\ldots b_k$ and $b_i = 0$ or 1.
- vol(D) is the volume (in bytes) of the uncompressed data D. Also known as the size of data D. When we compress data D using the compression algorithm K, then $vol_K(D)$ is the volume of data D after compression by K: $vol_K(D) = vol(K(D))$.

Definition 2.2. Context, sub-context and system

- We consider that the context refers to a set of human infrastructures, physical and algorithmic linked by relationships and information transfers, assuring a global systemic

coherence. A context consists of human groups and physical and algorithmic systems, assuring their interconnection.
- Any subset of a context is called a sub-context and can be considered a more restricted context.

Examples

The international commodity market is a context and the cocoa market is a sub-context. The art market or the energy sector are global contexts. The defense and security infrastructure of a nation is a national context. Silicon Valley, even as a technology hub, is a local context although it is highly globalized. A datacenter located in Silicon Valley is then a sub-context of this context.

Notation 2.2

- Note that C is a context, S is a physical and algorithmic system, a component, and A is an executable algorithm on the system S in context C.
- Data is assessed in the context C, at time t, according to the algorithm A used to interpret it. The assembly of all contexts is denoted Ω.
- For all data D, at time t, we have $\Omega = O_{D,t} \cup F_{D,t}$ ($\cup$ being the symbol of assembly) where $O_{D,t}$ denotes the assembly, at time t, of contexts with access to the informational content of D and $F_{D,t}$, denotes the assembly, at time t, of contexts without access to the content of D.
- We define the instant indicator function as the following:
  $I_{C,t}(D) = 1$ if the context C has access to the data D, at time t
  $I_{C,t}(D) = 0$ if not
  Thus, $O_{D,t} = \cup C$ such as $I_{C,t}(D) = 1$ and $F_{D,t} = \cup C$ such as $I_{C,t}(D) = 0$

- Data is called public when it is known and accessible to all contexts ($O_{D,t} = \Omega$ and $F_{D,t}$ is empty). Data is called private, at time t, if $F_{D,t}$ is not empty at time t.
- Over time, private data can become public, but not the reverse. The set $O_{D,t}$ is generally increasing over time (in the sense of inclusion) while $F_{D,t}$ never increases with time because it is assumed that the information acquired is not lost over time.

Definition 2.3. Instant functional value of data in a context according to an algorithm

- If D is accessible data in context C, and A is the algorithm interpreting D, executable on computing system S of the context, then it should be noted that $Val_t(D / C, A)$ is the value at time t of D with respect to context C and algorithm A uses D in C.
- $Val_t(D / C, A)$ is an instant numeric value, positive or null depending on the context and the operating algorithm.

Definition 2.4. Initial value of data in a context

At the initial time t = 0, context C first has access to the informational content of data D and operates according to algorithm A (which may just be a simple reading algorithm). This acquisition of knowledge of the content of D results in:
- Production of D by a component of C which is made public on C.
- Or a simple purchase of data sold to C by an element in another context.

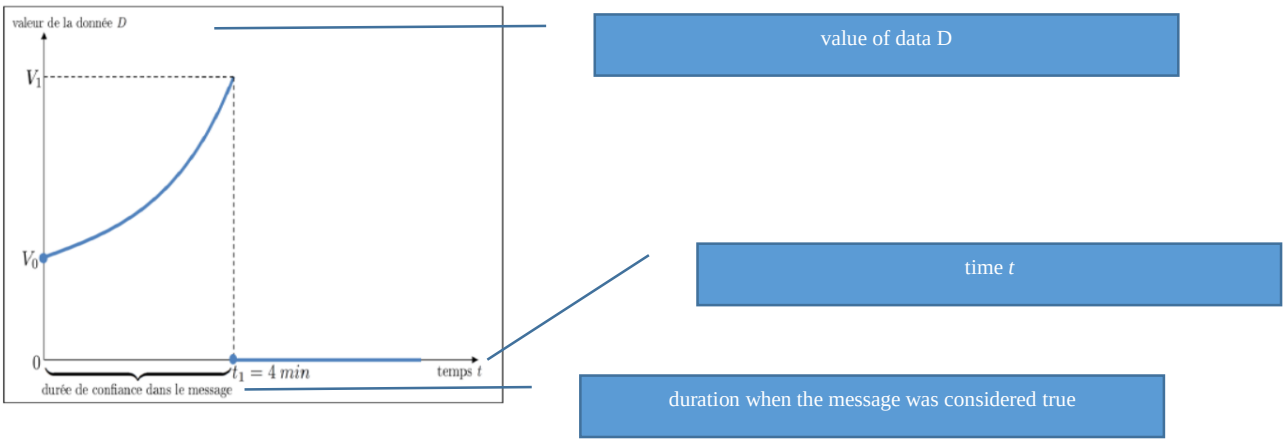
- $Val_0(D / C, A)$ is the initial value of the data D in C according to A. It can be equal to the purchase price in the context of data D or at production cost if the context itself produced this data.
- When data D is public, its instant value may not be zero, it is referred to as the residual value of the data: a context may indeed have interest in buying a set of public data previously structured and refined in another context [2]. The cost of refining and structuring of public data creates value for other contexts.

The nature of the data interpretation algorithm A of data D in the context C directly influences its actual value [3]. For example, suppose A is an algorithm that starts by reading the data D and then calculates the probability p(D, t) that this data is true at time t. This program performs a test on the veracity of data before it is used in a big data environment or semantic analysis. If p(D, t) turns out to be close to zero after the calculation, the value of the data will also be close to zero in a rational context. If on the contrary, this probability is close to 1, the context will consider the data as true or almost true and can then assign an instant value that is a function of the economy and the interactions between contexts.

When D, C and A are fixed, the function that, at time t, corresponds to Val_t (D / C, A) shows the instant variations in the value of data D in the context C according to the interpretation algorithm A. This value changes over time, from an initial value val_0 (D / C, A), corresponding to the production cost of D in C or to its purchase price at time t = 0, to its residual value denoted as Val_∞ (D / C, A). Such a function can have strong discontinuities as shown in example 1 (fig 1) the false tweet created by SEA. It may instead be constant in the context that produced and structured it, as in example 2 (fig 2) the Microsoft customer data sold to the FBI for $200 dollars each. The instant value depends directly on the demand from contexts that do not yet have access to D and want to buy it. Asymptotically, the more the contexts that know the content of D the more the value of D approaches its residual value. When data becomes public (that is to say it is known to all contexts), the residual value is reached.

Fig.1 -Example 1 -The data is a false tweet posted on the AP account

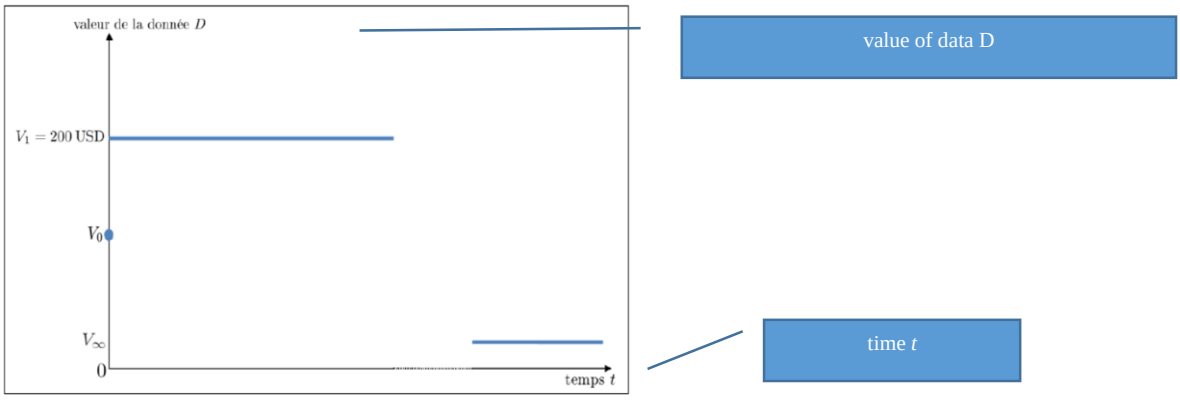
D = {Explosion at the White House, President Obama is hurt}



At time t = 0, the tweet from SEA is published on the AP account and remains accessible and is regarded as credible for four minutes. At time t_1, AP and the White House publish a retraction that immediately voids the instant value of data D.

V_0 is the value of the production and insertion of the data on the AP account. This value takes into account the overall cost of the hacking of the account by SEA.
V_1 is the maximum value of the data before retaking control of the AP account. It can take into account the value of the impact of the fake tweet on markets.

Fig.2 -Example 2- Customer data sold by Microsoft to the FBI



The instant value of customer data D sold by Microsoft to the FBI verifies:

- Val_t (D / C, A) = 200 USD for t > 0 in the Microsoft production context.
- A is the structuring algorithm (or formatting) and data reading.
- V_0 is the cost of structuring, formatting and storing of the data for Microsoft.
- V_1 is the unit selling price set by Microsoft to the FBI.
- V_∞ is the residual value of the data.

Refining data

When D, C, and t are fixed, we say that the algorithm A' refines the data D in C at time t better than the algorithm A if: Val_t (D / C, A') ≥ Val_t (D / C, A)
This is the case for the example of when A just reads data D in the context, while A' reads this data and it evaluates the probability of the veracity at time t and shows that it is close to 1. The second algorithm brings confidence to the data, thus increasing its instant value and it refines this data in C, better than in A at time t.
On the contrary, if the probability calculated by A' is close to 0, then Val_t (D / C, A') will be close to 0 and may in this case be increased by Val_t (D / C, A) which does not take into account of the veracity of D.

Instant value and sub-contexts

If C_1 is a sub-context of C_2, then for fixed data D, for algorithm A and a fixed time t, it is generally not possible to compare Val_t (D / C_1, A) and Val_t (D / C_2, A).
As a matter of fact, algorithm A may be more effective at valuing the data in the sub-context or on the contrary in the broader context.

The origin and nature of data

When the data D is generated in the context C, we say that C is its original context. This is the case when D is produced by a computing system installed on C (like connected objects, automated monitoring systems, and measuring instruments). This origin gives it a systemic nature. The data can also be produced by a human as a result of an interaction with a computing system. This is referred to as the algorithmic projection of an operator, under an algorithm executed on a system in the context. [4] In this case, the data is referred to as projective. Today, the global volume of systemic data is increasing very quickly and will soon exceed that of projective data. What impact will there be on the value of these types of data? Will systemic data be less valued than projective data?

Dissemination of data

According to Philippe Davadie,[1] the formalism used for defining the instant value of data can use three effective data diffusion measures in a context. He introduces the audience, echo, and impact of data D.

The instant audience of data D in context C, denoted $Aud_t(D/C)$, measures the fraction of the population of operators in context C who have access to the data D. It is a real number between 0 and 1 that is worth 1 if any of the context operators have access to the data without particular restrictions and 0 if, instead, the context operators have no access. Open public data is by definition accessible to every operator with an interconnected computing system; its audience is then worth 1 in any context.

The instant echo of data D in context C, denoted as $Echo_t(D/C)$, measures the fraction of the population of operators in context C with access to this data and that actually use it.

Finally, the instant impact of data D in context C, denoted as $Imp_t(D/C)$, measures the effect of the data on the context, that is to say, its ability to change the status of the context such as its parameters, responses, and ultimately, its entropy. The impact is certainly the quantity most difficult to determine and formally define as a function of the characteristics of the context and its interconnection to other contexts. The overall impact of data when all the contexts come together determines its impact value. Likewise, the instant audience and echo of the data influences its instant value. Are there simple functional relationships between these three measures and the instant value of data? This question remains open.

Conclusion

The formalism introduced in this paper allows us to establish some first definitions for the instant value of data in a context in relation to the algorithm that interprets it. This relative and functional approach deliberately avoids an absolute description while providing a flexible representation adapted to the volatility of the concept of value. It is still necessary to continue its development towards a more dynamic model, with equations or not, that is able to describe the changes in the value of data in order to better anticipate future events. Data is a resource, let us measure its value!

[1] Philippe Davadie, Colonel, Centre d'Enseignement Supérieur de la Gendarmerie, is also a member of the working groups of the Chair of Cyberdefense and Cybersecurity. He published "*L'entreprise: nouveaux défis cyber*," éditions Economica, collection Cyberstratégie, 192 pages, May 2014.

Bibliography

[1] Kempf Olivier and Berthier Thierry –"L'armée syrienne électronique: entre cyberagression et guerre de l'information" RDN –revue de la défense nationale –"Guerre de l'information" Vol. May 2014.
[2] Janert Philipp K. –"Data Analysis with Open Source Tools" – O'Reilly.
[3] Bulusu Lakshman –"Open Source Data Warehousing and Business Intelligence"– CRC Press
[4] Berthier Thierry –"Projections algorithmiques et cyberspace" R2IE –revue internationale d'intelligence économique –Vol 5-2 2013 pp. 179-195.

Online sources

Website and Twitter account of the Syrian Electronic Army:
https://twitter.com/Official_SEA16
http://sea.sy/index/en

Website of Simon Chignard on l'Open Data:
http://donneesouvertes.info/

Website for public data:
http://www.data.gouv.fr/

Website of Thierry Berthier –"cyberdéfense, cyberstratégies":
http://cyberland.centerblog.net/

Simulator "Financial Times" calculates the value of personal data:
http://www.challenges.fr/entreprise/20130711.CHA2303/combien-valent-vos-donnees-personnelles-sur-internet.html
http://www.ft.com/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html#axzz2WfFmKwic

"Théories de la valeur"- Theories of value:
http://www.pise.info/eco/valeur.htm

The price of public data in France:
http://www.data-publica.com/content/2012/09/les-donnees-publiques-payantes-en-france-ce-quil-faut-retenir/#

Open Data France:
http://opendatafrance.net/category/donnees-publiques/

Public data - intellectual property–Sciences-Po
http://www.sciences-pi.fr/content/%C3%A0-qui-appartiennent-les-donn%C3%A9es-publiques

Website of Philippe Davadie: "Informatiques orphelines" and upcoming book -May 2014:
http://informatiques-orphelines.fr/

Chaire Cyber-Défense et Cyber-sécurité (Chair of Cyberdefense and Cybersecurity)
Fondation Saint-Cyr, Ecole militaire, 1 place Joffre, 75007 Paris
Phone number: 01-45-55-43-56 - email: contact@chaire-cyber.fr;
SIRET N° 497 802 645 000 18

The Chair thanks its partners