Social Big Data

3.1 The Current State of Big Data

Like "cloud computing," the term "big data" is extremely conceptual. Its definition remains vague, and what it means can change over the course of time, so it is often used in a very abstract way. For this reason I believe the term will survive for a long time. However, when it comes to "big data" in the constantly changing world of buzzwords, I get the feeling that as of 2015 it is coming to the end of its shelf life. This impression was also given in the Hype Cycle for Emerging Technologies published each year by Gartner, who stated in 2014 that big data was moving toward the Trough of Disillusionment, so there is likely some substance to this opinion (Figure 1).

However, because it is still towards the end of the Peak of Inflated Expectations in the Hype Cycle limited to Japan, it seems "big data" as a buzzword occupies an extremely tenuous position (Figure 2).

■ Big Data Trends in Japan

In any case, while "big data" has remained a buzzword that symbolized technology trends over the past few years, it has begun to show its age. I feel that other slightly more specific buzzwords, such as the "Internet of Things (IoT)", are now touted more often.

The Internet of Things (IoT) is a term put forward in 1999 by Kevin Ashton, who contributed to the creation of a global standard system for RFID. The concept involves networking a range of objects by assigning IDs to them, literally building an Internet of things. Another extremely similar concept is the "Cyber-Physical System (CPS)," which is defined as a system in which computational elements that control physical entities collaborate by sharing information. Both concepts involve bringing the physical world and cyber world together*1.

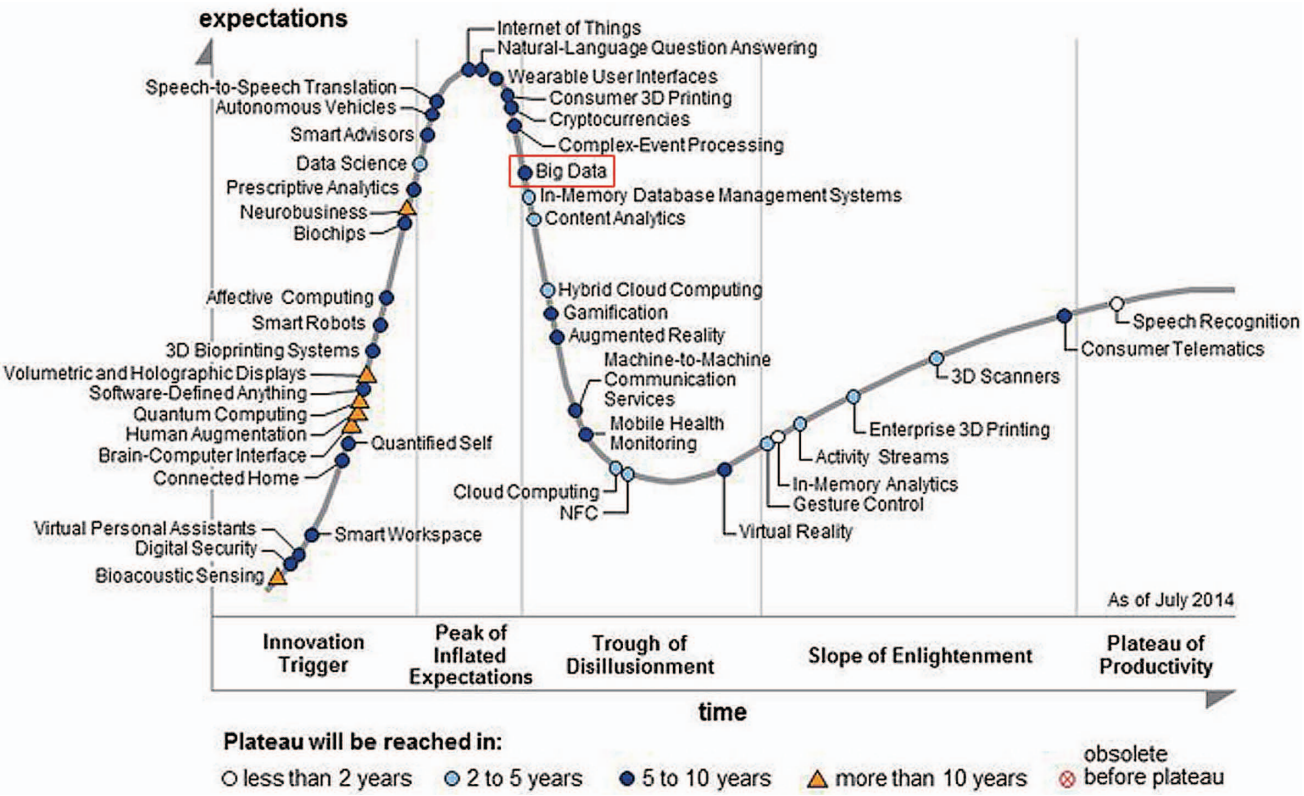


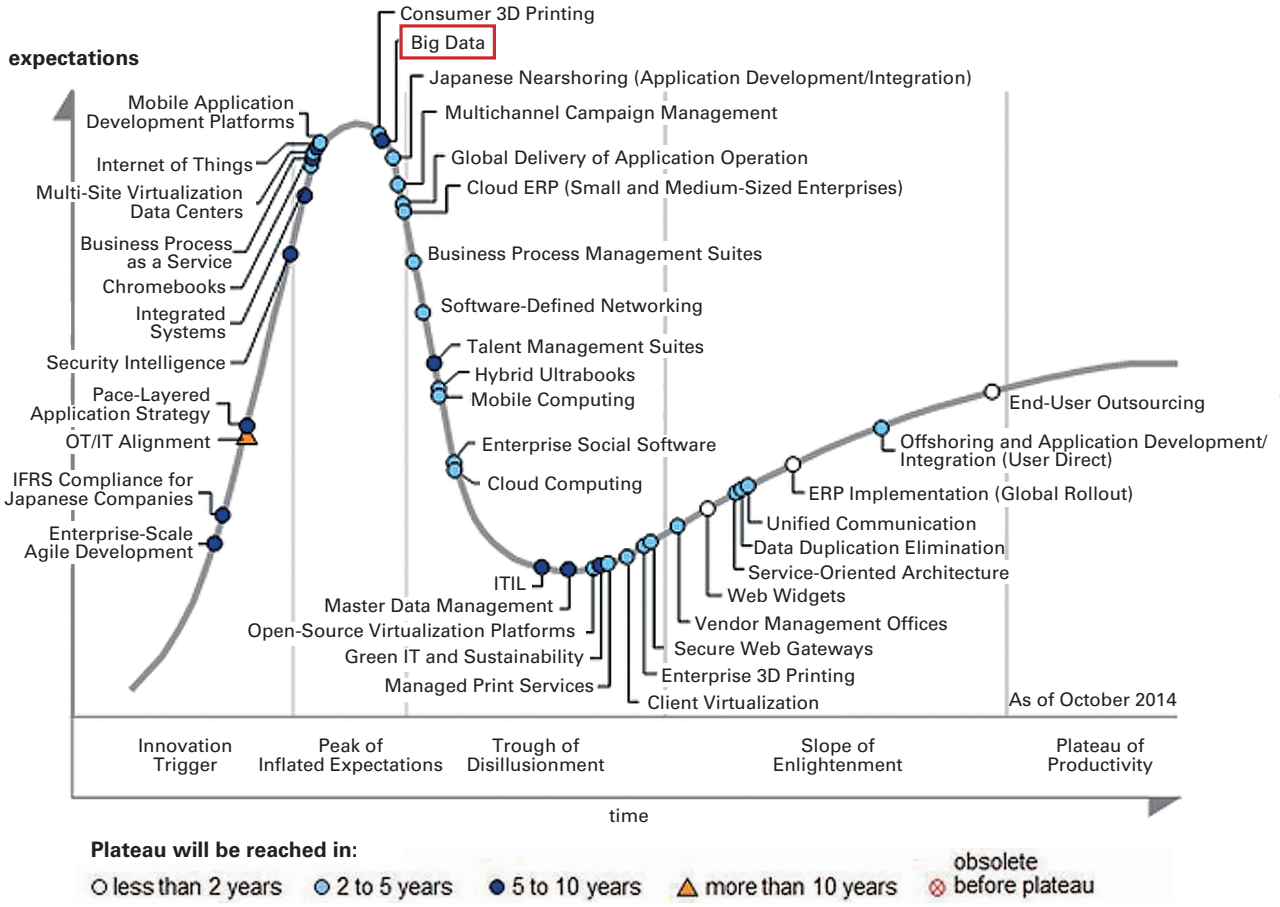
Figure 1: 2014 Hype Cycle for Emerging Technologies

*1 "Cyber Physical Systems and Internet of Things: Connecting the real world with the cyber world" Kazuo Iwano, Yosuke Takashima, Journal of Information Processing and Management 2-15 vol.57 no.11 (https://www.jstage.jst.go.jp/browse/johokanri/57/11/_contents/-char/en/).

Today, IDs are already assigned to objects such as parcel delivery tags, contactless IC cards used as commuter passes, and smartphones, making it possible to track their movement, etc. Accordingly, the location information continually generated for each ID is big data in a literal sense, and by analyzing this it is possible to perform a range of optimizations. This is the "brighter future" that IoT and CPS are aimed at bringing about. For example, I would hazard a guess that many people have seen proposals for projects such as smart grids and smart cities.

In other words, IoT and CPS are massive paradigms aimed at revolutionizing and reforming social systems in a fundamental way. Research and development related to these technologies has the potential to directly impact many people living in society. One example is the issue of privacy for ordinary citizens*2. For researchers interested in methods for analyzing big data such as myself, I believe that research into IoT and CPS is an area fraught with difficult challenges, considering the social aspects that must also be taken into account.

I think the manifestation of social issues such as these is evidence that progress has been made in big data research and development. In 2015, research into big data is currently in a temporary lull where we must carefully consider the direction in which we take it.



Source: Gartner (October 2014)

Figure 2: 2014 Technology Hype Cycle in Japan

*2 NICT (November 25, 2013, press release), "Proof-of-Concept Tests for Using ICT Technology at Large-Scale Complex Facilities Conducted at Osaka Station City" (http://www.nict.go.jp/press/2013/11/25-1.html) (in Japanese).
Nikkei Computer (March 11, 2014, Naoki Kiyoshima), "Facial Recognition Proof-of-Concept Tests at JR Osaka Station Building Postponed Due to Concerns About Invasion of Privacy" (http://itpro.nikkeibp.co.jp/article/NEWS/20140311/542723/) (in Japanese).
Nikkei Computer (November 7, 2014, Naoki Kiyoshima), "Facial Recognition Tests Using Security Cameras at JR Osaka Station Building to be Resumed with Limited Scope" (http://itpro.nikkeibp.co.jp/atcl/news/14/110701801/) (in Japanese).

I think the approach of linking the real world and cyberspace that IoT and CPS are oriented towards actually opens up new possibilities for big data analysis, in the sense that it creates data from facts. For example, if we could utilize data indicating facts such as "a certain person was in a set place at a given time," or "a certain location at a certain plant reached a set temperature at a given time" in analysis, we could treat at least these facts as conclusive information (logically unvarying). I surmise that this would contribute to raising the accuracy of analysis results. Furthermore, there are social demands requiring analysis and prediction to be carried out based on this kind of reliable data. For example, tasks such as the monitoring of nuclear power plant operation and the tracking of criminals have high urgency, and data that indicates facts is immensely valuable for analysis when certainty is regarded as important.

3.2 Social Big Data

Here I will refer to the data indicating facts that is created by IoT as "factual big data." Data with contrasting properties, for example the data used in analysis that takes into account individual impressions or preferences like "things that enrich your lifestyle" or "things to your taste," can be called "social big data." Messages circulated via social media such as Twitter and Facebook would be typical examples of this.

Social big data is generally interpreted as a hodgepodge of miscellaneous information. However, according to former Amazon Chief Scientist Andreas Weigend, the amount of social big data is growing rapidly, with the total amount of data generated in the year of 2009 exceeding the total amount generated up to 2008 since the dawn of history. He reports that this has brought about revolutionary change for the search and collection of customer product information*3.

Weigend points out that more and more social media users are sharing posts with explicit details on impressions of manufacturers and sites, or products they or their friends have purchased. This tends to be more effective than information provided by manufacturers as part of their advertising or public relations. Some manufacturers have also put in place systematic methods that encourage users to provide candid feedback, and reward them for valuable data.

Against this backdrop, the world of online marketing is currently shifting to a model in which specific data beneficial to third parties is created through collaborations between users (from e-business to me-business). Weigend calls this the "Social Data Revolution." The analysis of social big data may present answers to questions such as, "What can customer expectations bring about?" and "What can companies do to meet such expectations?" It is clear that the amount of big data on the Internet will continue to grow, and I believe that the spread of IoT and rising prominence of social media is accelerating the factors that are prompting the bipolarization of specific characteristics for the big data in circulation.

3.2.1 Properties of Social Big Data

Comparing social big data with factual big data from the perspective of big data analysis, its unique properties include the difficulty of machine generation, its multifactorial implications, the fact that ambiguity is acceptable, and the fact that conjecture is allowed (errors are tolerated). To use a slightly abstract expression, you could say that while factual big data is hard data generated by machines, social big data is soft data generated by people.

Normally, when people are involved in data generation, it becomes harder to analyze that data. For example, imagine you are surveying the most pleasant area in a large park. If you assume that "pleasant" means that it has a suitable temperature, you could place 10,000 devices with temperature sensors in the park, and build a system that collects this data around the clock. Analyzing the factual big data accumulated in this way would provide extremely clear results. Devices could be placed in every nook and cranny of the park, and once installed you could obtain temperature data for that location with absolute certainty. With 10,000 devices, a few are likely to fail, but as long as this can be confirmed you could simply replace them straight away. This would allow you to thoroughly investigate which areas have a temperature that people find pleasant.

*3 The Social Data Revolution(s) Andreas Weigend May 20, 2009 (https://hbr.org/2009/05/the-social-data-revolution.html).

In contrast, hiring 10,000 surveyors and asking them to measure the temperature at each point would be the social big data method of surveying and analyzing this data. Honest surveyors may use a thermometer to produce reports stating that "point XX was YY degrees," resulting in accuracy approaching that of devices fitted with temperature sensors. That said, there would likely also be surveyors that eschewed the use of a thermometer, and produced extremely vague reports stating things like "the area around the lake is cold." Others may check a Fahrenheit thermometer and answer "72 degrees," or just lie in their report and not make the trip to each point. There could even be surveyors who make irrelevant reports, such as "the wooded area is pleasant." This would make it very difficult to find points with a suitable temperature within the park.

However, if you asked the surveyors to find points that seem pleasant, the results of the survey would probably be very different. The surveyors would start by heading to places that look pleasant. They may run into other surveyors on the way, compare opinions, and seek likely places together, eventually settling on the place that seems the best when they find somewhere they both agree on. As a result, over time we could expect surveyors to gather at certain points in the park. Judging the right time, we could interview individual surveyors on their findings. By asking questions such as where they are, whether that place is pleasant, and why they think so, you could obtain information on pleasant points in the park that third parties can sympathize with.

The question of which survey result is more applicable depends on the purpose of the survey. For staff in charge of park management or environmental conservation, factual big data is probably all that is required. However, for the owner of a kiosk in the park, survey results based on social big data may actually be more desirable. The decisive difference between these two examples is that factual big data produces results comprised of objective data based on homogenous facts, while social big data is an aggregation of data based on the subjective view of surveyors (people). Both are big data, so consideration must be given to reliability to prevent erroneous data. Social big data also requires analysis techniques that take into account issues with trustworthiness, such as whether the data is correct or contains errors, and how much it can be trusted.

3.3 Wikipedia as Big Data

Many people may think of the messages circulated over Twitter when social big data is mentioned, but my focus is on Wikipedia.

As Wikipedia is the most popular electronic encyclopedia service, I doubt there is any need for me to give an overview of it here, but Wikipedia is a notable site when it comes to big data as well. On the Japanese[*4] or English[*5] "Wikipedia:Statistics" pages, you can view an up-to-date pageview count for articles on Wikipedia, etc.

As of July 20, 2015, Wikipedia has a total of 4,920,887 articles and 36,748,410 pages (including peripheral information). Even limiting results to Japanese, the total number comes to 974,894 articles and 2,785,007 pages. This data can also be modified or reproduced for secondary use based on the terms of the CC-BY-SA 3.0[*6] license. In terms of licensing, you could say a site like this that can be used in the same way as open source software is the most secure big data.

Wikipedia data can be traced from a page called "Wikimedia Download"[*7]. Wikipedia constantly has data backup tasks operating, and articles in each language are regularly backed up. To my knowledge, new backups are created around once a month.

*4 https://ja.wikipedia.org/wiki/Wikipedia:%E7%B5%B1%E8%A8%88
*5 https://en.wikipedia.org/wiki/Wikipedia:Statistics
*6 Creative Commons (https://en.wikipedia.org/wiki/Creative_Commons)
*7 http://dumps.wikimedia.org/

This Wikipedia data is already utilized in a variety of areas. One prominent example is DBpedia*8*9, a project that generates Linked Open Data (LOD) from Wikipedia data to create a database. This database is apparently being utilized in areas such as natural language processing and text mining.

3.3.1 Wikipedia as Social Big Data

Wikipedia maintains its electronic encyclopedia content via human wave tactics using a wiki system. So, is it social big data? In my opinion, Wikipedia article data should be classified as factual big data, as a lot of effort has been put into the system and its operation to improve its utility as an encyclopedia, and encourage writers to base their work on objective facts. Of course, because the data is written and edited manually using a wiki system, compared to data generated by machines there are more issues with data integrity, such as typos, factual errors, and inconsistencies between articles. On the other hand, it can also be thought of as extensive factual big data, which can cover matters for which the facts cannot be confirmed or circumstances without an established theory.

It must, however, be said that Wikipedia has a social big data aspect to it, too. Namely, the "Page view statistics for Wikimedia projects"*10, which have been published since 2013. This data is a tabulation of the page view count in hourly increments for all pages of the Wikimedia project. Going back you can obtain page view information from 2008 onward (or December 2007 to be exact).

In my research group we have kept a close eye on this data ever since we launched a service*11 based on it in June 2013, and it is fascinating to see how the rankings change, or in other words which hot topics show up in the top rankings (Figure 3).

3.3.2 Research by Tobias Preis

Tobias Preis, an Associate Professor of Behavioral Science & Finance at the University of Warwick, explains that individual Internet users seeking information is driving this kind of behavior. For example, after hearing about major news topics reported by the mass media, etc., individuals who use the Internet on a daily basis tend to gather information related to that topic by performing

Figure 3: Wikipedia Rankings

*8 http://wiki.dbpedia.org/
*9 http://ja.dbpedia.org/
*10 http://dumps.wikimedia.org/other/pagecounts-raw/
*11 http://www.gryfon.iij-ii.co.jp/ranking/

searches on search engines or viewing Wikipedia. It is thought that traces of this behavior are recorded as search engine query data or Wikipedia page views. Focusing on this phenomenon, Preis ascertained that search engine query data and stock market fluctuations were correlated in his 2010 paper, "Complex dynamics of our economic life on different scales: insights from search engine query data"*12.

Next, in his paper, "Quantifying Trading Behavior in Financial Markets Using Google Trends"*13, he suggested that an increase in searches for 98 terms related to finance included in query data obtained from Google Trends tended to precede major crashes in the financial market.

Furthermore, in his paper, "Quantifying Wikipedia Usage Patterns Before Stock Market Moves,"*14 he used knowledge gained through Google Trends analysis to identify that Wikipedia page view counts correlated with large-scale fluctuations in the stock market.

In these papers, Preis concluded that it was possible to gain new insight regarding the information gathering behavior of people pressed to make decisions from online data such as search engine query data and Wikipedia page views. For example, phenomena such as stock market crashes result from the decision-making of individual investors, but paying attention to online data makes it possible to discover signs of this at an early stage. This is a typical example of a solution that utilizes social big data.

3.4 Conclusion

In this report, I have focused on social big data and its characteristics. The Social Data Revolution advocated by Andreas Weigend suggests that in certain fields such as e-commerce, vast quantities of data related to human behavior for product purchases is beginning to accumulate, and there is a need for a new approach based on analysis of this data.

Because social big data is data that relates to people's behavior, development of analysis methods will fall under the domain of social science or behavioral science. However, given that we have never before been in a position to obtain such wide-ranging and detailed data regarding human behavior, it is likely to require a considerable amount of time to find effective techniques for extracting valuable knowledge. As far as I am aware, the approach of tracking and analyzing individual behavior based on data obtained from microblogs or SNS has not been very successful, despite the ability to produce analysis results of some kind, due to the difficulty of identifying meaningful knowledge in these results.

On the other hand, I believe Tobias Preis's approach of treating Internet user behavior as a macro phenomenon, and applying complex analysis techniques, may make it easier to obtain significant findings. Factors such as query data and Wikipedia PVC are used for this analysis, and Preis's papers seem to show that short-term forecasting will be possible through predicting group behavior by independent individuals. Additionally, by analyzing notable fluctuation phenomena identified through this analysis in further detail, I believe it should be possible to gain insight into root causes, etc. by tracking the behavior of individuals using data obtained from social media.

Author:
Akito Fujita
Chief Architect, Business Strategy and Development Center, IIJ Innovation Institute Inc. (IIJ-II). Mr. Fujita joined IIJ in 2008.
He is engaged in the research and development of cloud computing technology utilizing knowledge gained through structured overlay research.

*12 http://www.tobiaspreis.de/publications/prs_ptrsa_2010.pdf
*13 http://www.nature.com/srep/2013/130425/srep01684/full/srep01684.html
*14 http://www.nature.com/srep/2013/130508/srep01801/full/srep01801.html