Big Data, Small World

Rachit Tomar
B.Tech 4th year (computer science)
Maharaja Surajamal Institute of Technology (GGSIPU)

Shantanu Choudhary
B.Tech 4th year (computer science)
Maharaja Surajamal Institute of Technology (GGSIPU)

Aditya Vikram Bisht
B.Tech 4th year (computer science)
G B Pant Engineering College (GGSIPU)

ABSTRACT

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. In this paper, we have discussed about the various characterstic’s of big data and how data is increasing day by day. There are various aspects of knowledge discovery that are discussed. Moreover, the small world phenomenon with various examples and six degrees of separations principle has been discussed in this. Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, combat crime and so on. Thus we have concluded that our world is shrinking to the advent application of big data.

Keywords

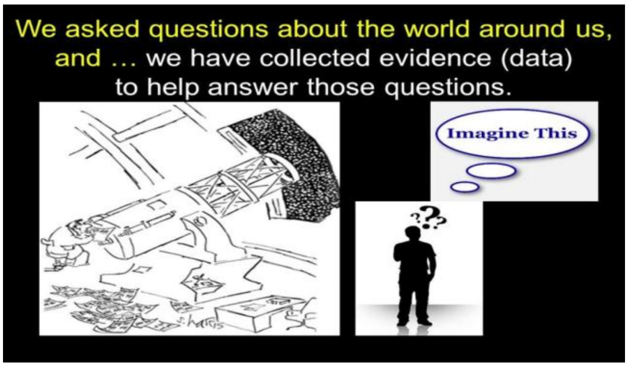
Big Data, Small World Phenomenon, Six degrees of Separation.

1. INTRODUCTION

Let us discuss „Big Data‟ in the context of the Small World Phenomenon, that we are not familiar with. In our daily life, we meet people from different walks of life, who turn out to be strangely related to us in some way or the other. Some are related to us through a third party who is a mutual acquaintance, some by families ties, some may even share similar places of origin, hometown or even the same neighbourhood as us. This is truly a small world phenomena.

The Small World phenomenon was introduced by Stanley Milgram in the 1960‟s. Milgram's basic small-world experiment remains one of the most compelling works of all times, for us to think about the problem. The goal of the experiment was to find short chains of acquaintances linking pairs of people in the United States who did not know one another. In a typical instance of the experiment, a source-person in Nebraska would be given a letter to deliver to a target person in Massachusetts. The source would initially be told basic information about the target, including his address and occupation; the source would then be instructed to send the letter to someone she knew on a first-name basis, in an effort to transmit the letter to the target as efficaciously as possible. Anyone subsequently receiving the letter would be given the same instructions, and the chain of communication would continue until the target was reached. Over many trials, the average number of intermediate steps in a successful chain was found to lie between five and six, a quantity that has since entered popular culture as the “six degrees of separation" principle. No wonder, big data is changing our world and decreasing the number of steps that connect one person to another. Hence, strengthening ties amongst our kind.

About two to three decades ago, we don’t call it big data; but we often talked about various things we could discover with the help of it; for example; we could discover things that we already knew about, hence these were called the “***known knowns***” or things that we knew about but had never seen or witnessed before, hence these were in tern called “***known unknowns***”. There are other things that we are completely unaware of and which might only surprise us on being discovered, those were named as “***unknown unknowns***”.


We asked questions about the world around us, and ... we have collected evidence (data) to help answer those questions.
Imagine This
??

[Fig 1]. Collecting evidence to answer our questions

It is in the human nature to be inquisitive. Humans have been scientists since time immemorial. We have explored, observed and experimented on almost all the thing that surround us. This inquisitive factor is present by virtue of our birth. It can be seen even in a small child, when he tries to put random objects in his mouth without having even a bare minimum knowledge about that object, or the constant and repetitive questioning by children in various subjects. We collect information about our world, we ask questions and those questions lead us to asking new questions. By this ongoing process, we tend to collect a whole lot of data, which we call as evidence [fig.1]. This is a never-ending and a dynamic process, which in turn leads to the formation of „big data‟. The more evidence we collect, more questions we have and we go out collect more data and this concept as we can see, is a universal and an omnipresent concept. It can be seen as primary in running businesses, Organisations, governments and even social networks.

Now... Big Data is everywhere and growing
DATA
priority for most businesses, social networks, and others...

[Fig 2]. Prioritizing big data

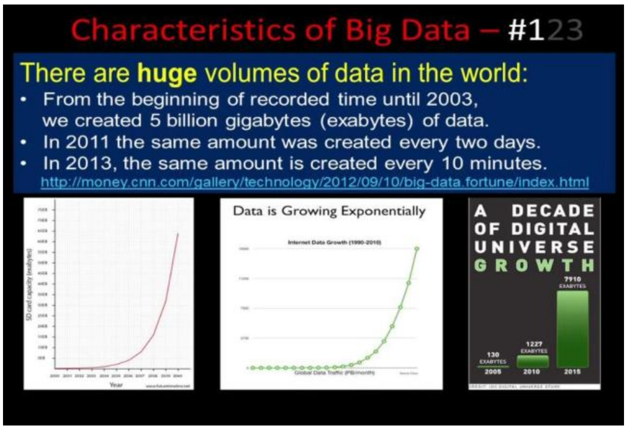
So we can say that the social media is contributing to the flood of data in the world but it's everywhere. It is on the top of the priority lists of business organisations, federal agencies. It is the topic of conversation at the highest levels in the US government [Fig 2]. As it is evident from the recent initiative announced by President Obama, about three years ago, also known as the "*the National Big Data initiative*" , It is one of the major drivers of discovery and that of revenue in today's time. So big data is present everywhere, it is available for everyone and it is bringing us closer to one another.

2. CHARACTERSTICS

2.1 HUGE VOLUMES OF DATA

In 2003, scientist at University of California Berkeley, conducted a study of the total amount of data in the world and came up with a number. From the beginning of human history all recorded history up to the year of 2002 amounted to, approximately, "5 Exabyte",i.e., five billion gigabytes of data, has been created and accumulated by humans.

In order to understand it better, we may relate it to the amount of data that we download on a daily basis, be it a movie or a game, which hardly amount to a single gigabyte. Now, if we put together a billion of those, it would be make an Exabyte. Imagine burning it all down into CDSs or DVDs and laying them out on a football field or stacking them one on top of the other. It might even cover an entire stadium!! That is a representative of the amount of data that equals 5 Exabytes. That's all the data created up through beginning of time to 2002, and humanity created that same amount of data in one consecutive year of 2003



Characteristics of Big Data – #123

There are **huge** volumes of data in the world:

- From the beginning of recorded time until 2003, we created 5 billion gigabytes (exabytes) of data.
- In 2011 the same amount was created every two days.
- In 2013, the same amount is created every 10 minutes.

http://money.cnn.com/gallery/technology/2012/09/10/big-data.fortune/index.html

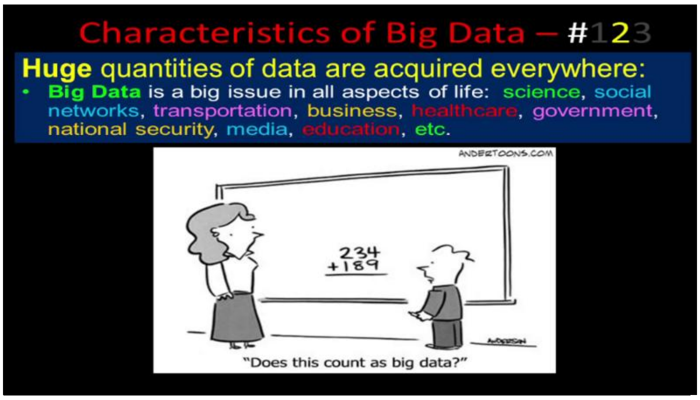
[Fig 3]. Universal data growth

In 2011, it accelerated to creating the same amount of information every two days and in 2013, the number has changed drastically, 5 Exabytes of data was created every 10 minutes. Surprisingly, these numbers or rates are only expected to rise up to a few seconds, in the next few years [Fig 3].

2.2 Big Data Is Ubiquitous Thing

This a ubiquitous problem of big data, it is used in businesses, education, science, social media, homeland security, network security, cyber terrorism, it's everywhere.

Everyone is busy collecting information, be it about the objects around us, our universe, about the human body, ranging from genomics to astronomy to searching for the Higgs boson, big data is collected everywhere [Fig 4].



Characteristics of Big Data – #123

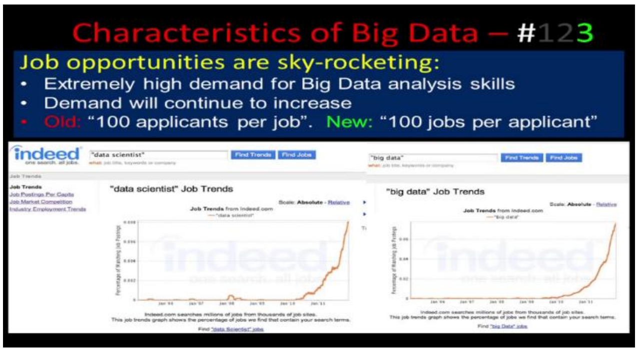
Huge quantities of data are acquired everywhere:

- **Big Data** is a big issue in all aspects of life: science, social networks, transportation, business, healthcare, government, national security, media, education, etc.

[Fig 4]. Big data is ubiquitous thing

2.3 JOB OPPORTUNITIES

It's the most secured job on the planet in the current time. About 2-3 decades ago, there were very few job opportunities in this regard and this field consisted of a great amount of competition, but the same equation has been completely reversed today. There are probably about a hundred jobs for every applicant looking for work in this field. Despite there being thousands of job and employment opportunities, there is good number of them that are still unoccupied, and the real problem is that this number of vacancy might reach up to a few millions in the next few years to come [Fig 5].



Characteristics of Big Data – #123

Job opportunities are sky-rocketing:

- Extremely high demand for Big Data analysis skills
- Demand will continue to increase
- Old: "100 applicants per job". New: "100 jobs per applicant"

indeed
one search. all jobs.
"data scientist"
Find Trends
Find Jobs
"big data"
Job Trends
Job Postings Per Capita
Job Market Competition
Industry Employment Trends
"data scientist" Job Trends
"big data" Job Trends
Scale: Absolute - Relative
Job Trends from Indeed.com
Percentage of Matching Job Postings
Indeed.com searches millions of jobs from thousands of job sites.
This job trends graph shows the percentage of jobs we find that contain your search terms.
Find "data Scientist" jobs
Find "big Data" jobs

[Fig 5]. Job opportunities in big data

3. MEDIA POINT OF VIEW

3.1 Promising News

The promising news is that the potential for big discoveries and insights is enormous now. We have the ability to discover treatments for diseases, new drugs just by exploring data sets [Fig 6]. When we have the full genomic sequence of a human, we can discover the likelihood of different diseases that we are susceptible to and to what extent can they be treated, after using a certain drug. These things require the collection of the Big Data along with the right amount of expertise in the concerned field.

PROMISING ASPECT: Big Data leads to Big Insights and New Discoveries

http://news.nationalgeographic.com/news/2010/11/photogalleries/101103-nasa-space-shuttle-discovery-firsts-pictures/

[Fig 6]. Promising aspect of big data

3.2 Scary Aspect

The scary part about all of this is that, we're also reaching a tipping point[Fig 7].It is a very difficult task to handle , manage , learn , discover, build revenue , fight cyber terrorism, seclude the relevant from the irrelevant, from such a vast ocean of data. This is nothing but a bitter truth that surround big data.

SCARY

Big Data is taking us to a Tipping Point

http://bit.ly/HUqmu5

http://goo.gl/Aj30t

[Fig 7]. Scary part of big data.

4. BIG DATA BENEFIT'S: KNOWLEDGE DISCOVERY

4.1 Novelty Discovery

Novelty discovery is essentially about finding data related to rare and interesting things, which points to the unknown unknowns [Fig 8]. For example, who could have imagined before the unpopular 9/11 attacks in the United States, that such people would do such things so as to fly planes into the buildings?

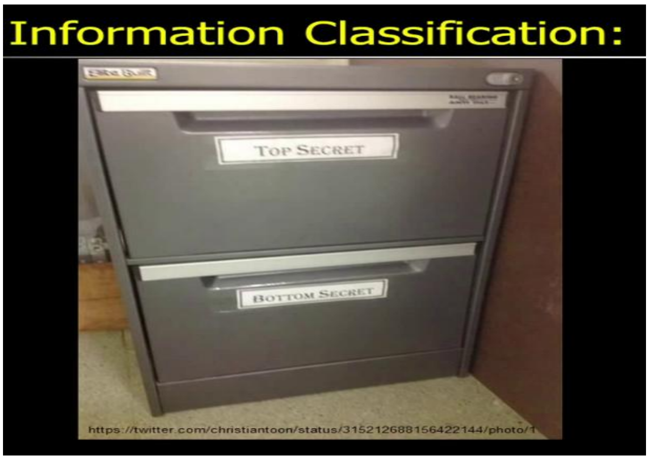
Novelty Discovery:

Discovering the one that is different

[Fig 8]. Novelty discovery

4.2 Class Discovery

Class discovery is about finding new classes of data needed. For example, if you are in business then finding new classes of customers, in case of an astronomer, finding new classes of galaxies or in case of a genomist then finding new classes of drugs.



Information Classification:

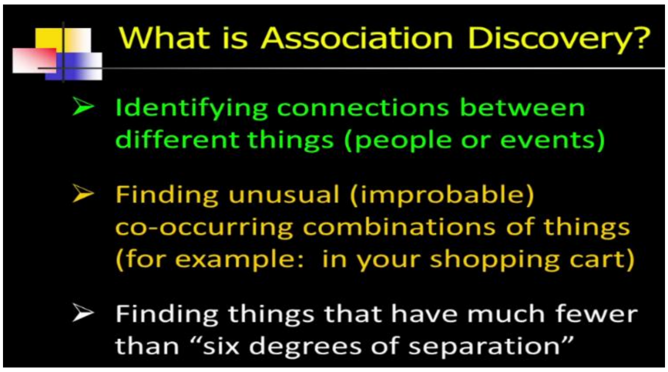
https://twitter.com/christiantoon/status/315212688156422144/photo/1

[Fig 9]. Class discovery

Here's a simple example, there are primarily two types of information that this person is classifying, by storing in a file cabinet; there's a top-secret and a bottom secret [Fig 9]. This implies a class discovery or information classification.

4.3 Association Discovery

Association discovery is related to finding the unusual interesting things that co-occur simultaneously without any expectation. It is finding the number of degrees of separation between these things. This is the small world effect. For example, we come across a person who grew up in the same hometown as ours [Fig 9]. The small world connection through big data is now drawing us closer together that is discovered to this process of finding associations.



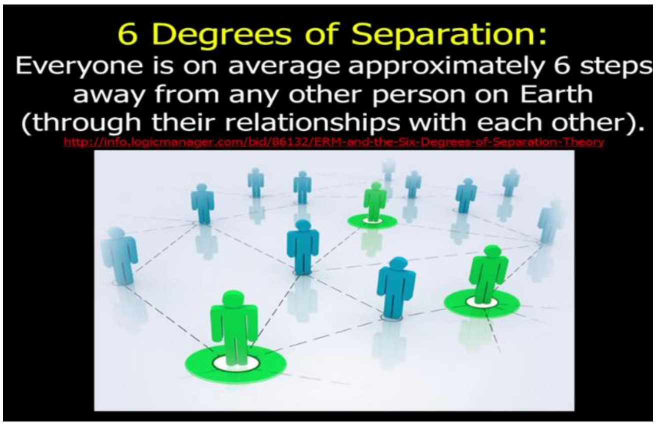
What is Association Discovery?

- Identifying connections between different things (people or events)
- Finding unusual (improbable) co-occurring combinations of things (for example: in your shopping cart)
- Finding things that have much fewer than "six degrees of separation"

[Fig 10]. Knowledge discovery ways

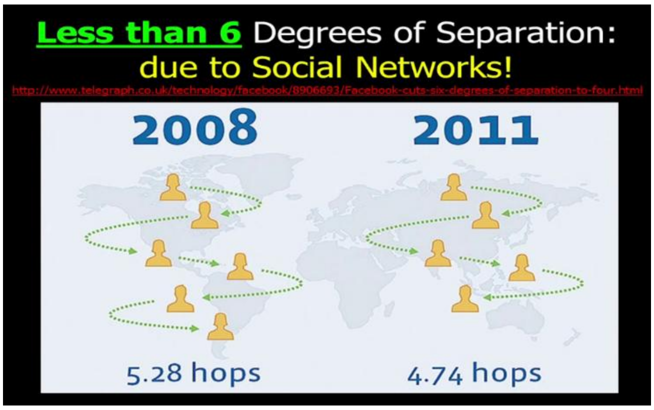
5. SIX DEGREES OF SEPARATION

This is the concept of you being connected to just about everyone in the world by no more or averaged six degrees of separation. This means that everyone is, on average approximately 6 steps away from any other person on the earth (through their relationship with each other either digitally or physically). Six degrees of separation works with people, and any type of network, including the network of the things that we like or purchase and is linked to the things that other people buy or like.



6 Degrees of Separation:
Everyone is on average approximately 6 steps away from any other person on Earth (through their relationships with each other).
http://info.logicmanager.com/bid/86132/ERM-and-the-Six-Degrees-of-Separation-Theory

Six degrees of separation works not just with people but with products as well, say if we like or follow something on Facebook, then we tend to share the same with other people as well. Therefore, we end up having something in common with somebody that we do not even know and if we start discovering several number of things in common, we might become friends with that person. Now, we have made a new friend primarily through big data, even though we might have not met before, both of are connected to internet and are sharing things. With the advent of social networks this number has actually been shrinking[Fig 10]. So when this concept was first introduced it was six degrees of separation, but now it’s even smaller and fewer than six, due to the collateral effect of social networking.



Less than 6 Degrees of Separation:
due to Social Networks!
http://www.telegraph.co.uk/technology/facebook/8906693/Facebook-cuts-six-degrees-of-separation-to-four.html

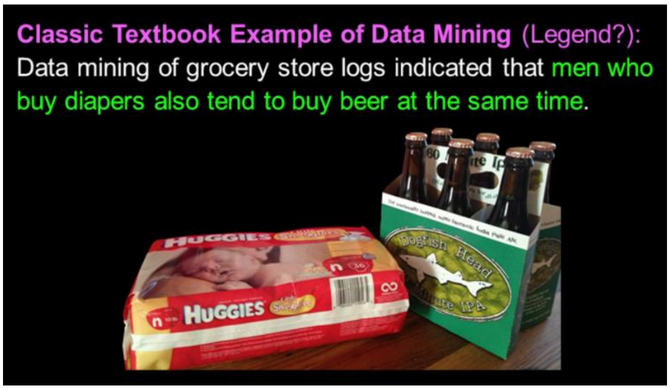
[Fig 11]. Six degrees of separation.

6. EXAMPLES

Four examples of a small world phenomenon in the business world. Showing how big data is shrinking our world through associations.

6.1 Beer And Diapers Example

The infamous example from data mining text books called beer and diapers example. The story goes like this. Some retail store analysed the transactions to discover that do men who visit the store to buy diapers, also tend to buy beer at the same time or not. Now we can analyse but one of the great things about statistics is that it warns us that correlation does not imply causation, so buying the diapers does not cause the man to buy the beer and vice versa[Fig 11]. But we have also learned that there are some things called the hidden variables. In this case, hidden variable is the crying baby at home which may have caused both of these things to happen.

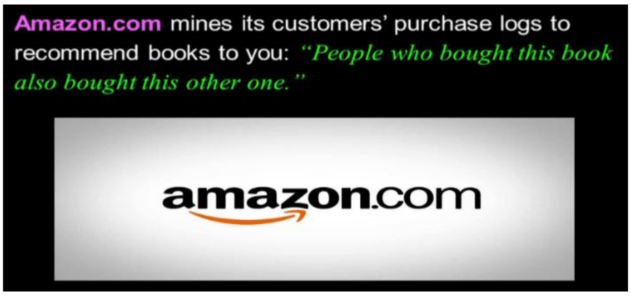


Classic Textbook Example of Data Mining (Legend?):
Data mining of grocery store logs indicated that men who buy diapers also tend to buy beer at the same time.

[Fig 12]. Beer and diapers example.

6.2 Online Store Example

We often come across instances on various online stores like Amazon, Flipkart etc, where we pick up a certain product and they mention to us as to what other products did people buy along with the product chosen. This may also be in the form of recommendations or popularly bought stuff. They're using the data to make their recommendation. Purchase histories, purchase transactions but also which pages people have viewed on their website. So basically this is a win-win situation, we say Amazon wins if we buy their product, they make some more money [Fig 12]. On the other hand we win because we discover something interesting that we didn't previously know about and how did we find out about it? Solely because someone else has similar interests as us. Therefore the world is shrinking, “I like this book and wow! look at that I didn't know this one existed”.



Amazon.com mines its customers’ purchase logs to recommend books to you: *“People who bought this book also bought this other one.”*

[Fig 13]. Online store example.

6.3 Online Media Example

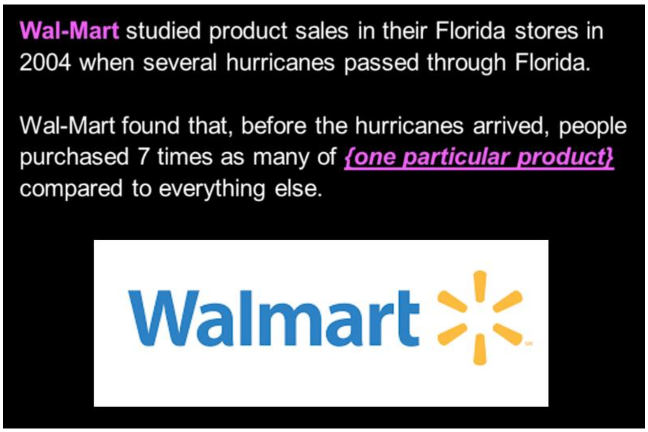
Same thing is happening at Netflix, recommending movies to people, based upon your rental history and people who have similar rental histories [Fig 13]. This is again a similar situation as mentioned above. Us discovering a new and interesting movie that we didn't know about, and Netflix making a little money out of it.

Netflix mines its video rental history database to recommend rentals to you based upon other customers who rented similar movies as you.

[Fig 14]. Online media example

6.4 Retail Store Example

Let’s take an example of famous retail store such as Walmart. So a few years ago, probably before the birth the Facebook, Walmart had the biggest data warehouse in the world. They cut down all information on all the transactions and other customers in all the stores, across the world and then in 2004, there was a series of hurricanes that hit the state of Florida. Walmart decided to examine as to what did the customers really want? They see this other hurricane coming by the way, and then they see another one coming. So they analysed their database, their purchase logs, their customer histories and discover that there's one single product that customers bought, at a seven times higher rate than everything else, so lots of things increase in sales prior to the advent of a natural disaster which is predictable in the case of a hurricane [Fig 14].



Wal-Mart studied product sales in their Florida stores in 2004 when several hurricanes passed through Florida.

Wal-Mart found that, before the hurricanes arrived, people purchased 7 times as many of ***{one particular product}*** compared to everything else.

[Fig 15]. Retail store example.

Water sales, toilet paper sales, plywood, generators, flashlights, batteries, the one thing that increased more than all those by factor of 7 (that is not 7%, it is 700%) was strawberry pop-tarts. And this huge amount of information they get, just by analysing the data sets they had.

7. CONCLUSION

Hence, the websites we visited, the places we go to, things we purchase, the friends that we like; all imply one thing, that the world is becoming a smaller place now. All these things are narrowing down our connections with other people in the world and businesses are using the same to make retail sales, governments are using this to discover terrorists, scientists are doing this to discover new drug & the new properties of things about our universe. Our world is shrinking to the advent application of the Big Data.

8. REFERENCES

[1] IDC, 2011. Big Data: What It Is and Why You Should Care. [online] IDC. Available at: http://sites.amd.com/us/Documents/IDC_AMD_Big_Data_Whitepaper.pdf.

[2] The Age of Big Data. Steve Lohr. *New York Times*, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[3] IDC Press Release, 2012. Worldwide Big Data Technology and Services 2012-2016 Forecast. [online] IDC. Available at: http://www.idc.com/getdoc.jsp?containerId=prUS23355112

[4] “Big Data”, Wikipedia. available at: https://en.wikipedia.org/wiki/Big_data

[5] "Data, data everywhere". *The Economist*. 25 February 2010. Retrieved 9 December 2012.

[6] Laney, Douglas. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.