# 7 “Big Data – Introduction and Definition”

Prof. Dr. Johann Günther
Professor at: Danube University Krems, Austria, Bonch-Bruevich Saint Petersburg State University of Telecommunications, Russia, Jianghan University, Wuhan, China
Johann@johannguenther.at

In 2012 a 16 years old girl in the Netherlands had a birthday party invitation posted on Facebook. Mistakenly she did it „public“. 3000 people turned up! She wanted just a dozen guests. At the end the father called police to rectify the situation. These are consequences of Big Data and wrong use of this powerful tool.

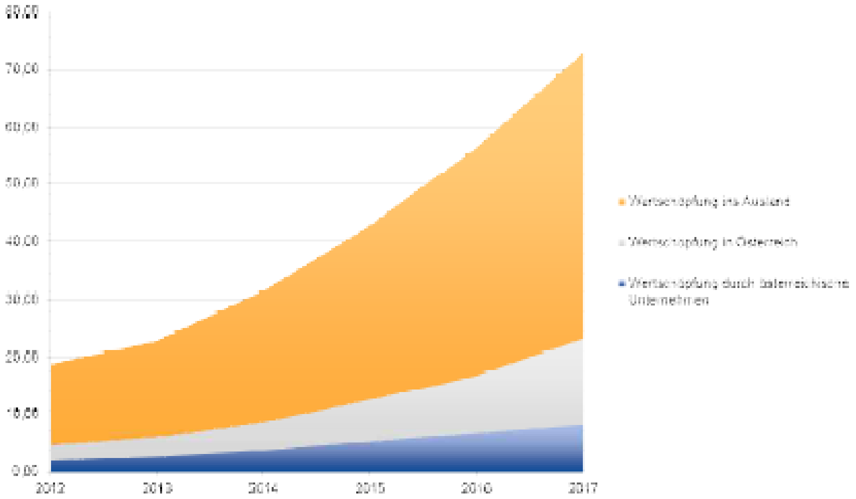
The digital data in the network are exploding. Also experts are not aware, how much it is already. I just give some figures, that the reader can get some feeling for this situation:

## 7.1 Data Worldwide[1]

- Beginning recorded history till 2003: 5 billion Gigabytes
- 2011: 5 billion Gigabytes every 2 days
- 2013: 5 billion Gigabytes every 10 minutes
- 2015: 5 billion Gigabytes every 10 seconds

These are figures no human being has an imagination.

There is also business behind Big Data. In 2012 the international turnover was 9,8 billion $. In 2017 we expect four times more: 32,4 billion $.[2]
2012: 9,8 billion $
2017: 32,4 billion $

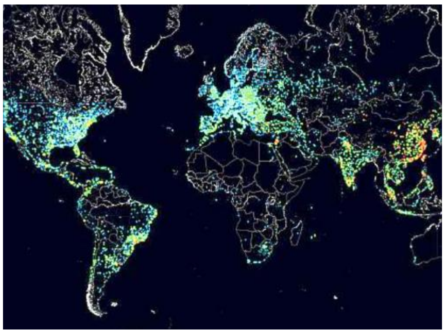


[1] Source: Smolan & Erwitt 2012
[2] Source: IDC

## 7.2 Definition

Big Data is a large collection of data. Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.“[3]
Big Data includes also data sets „with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.“[4]
Data Set is the old term from Main Frame Computing. Today Data Sets can not be handled with traditional applications.
In the first impression the development in data processing went back from the centralized main frame computing to the centralized Cloud Computing. BUT: there is a development like in spiral stairs. When you go up, all stories look the same, but you get higher and higher. The same is and was in the devlopment of data processing.

- It started with centralized main frame computers.
- With Personal Computers it became a decentralized architecture. Every user stored his data by himself localy.
- With cloud computing the size of local CPU is not so important any more. Big Data is stored somewhere outside in the „Cloud“. This is a server which stays in a place most users do not know. It looks like we are back at the stage of the beginning of Data Processing, but it is far more developed.



A very important impact to Big Data was globalization. The link between all countries of the world brought us the volume, which one country would not be able to produce. This also increases the quality of information. It can be said: „The world became a globale village“.

## 7.3 Characteristics of Big data

- First criterion – which also gaves the name „Big Data“ – is volume, the quantity of data.

[3] http://en.wikipedia.org/wiki/Big_data
[4] Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science 7: 1–5.

- Second criterion is variety. When we ask in Google or another search engine a question, we get so many answers, that often we do not know what to do. In former times a teacher had to train his students with facts. Today he must train them to find the right information in the right place.

- Third criterion is velocity. Due to the big volume of data a high speed of processing is needed. To guarantee high speed in the network is a new challenge of governments. With the quality of the network the quality of the economy is defined. Countries that do not invest in a highspeed network will be the loosers in the long run.

- Fourth criterion is veracity. For one question we get so many answers, that it is not easy to find out which one is the correct one. Everybody can offer some information in the net. The quality of all these offers is very different. To define the quality of information is a new and big problem.

- Last but not least is complexity an important criterion. To manage the data with large volume and from different sources is a challenge and an important instrument.

## 7.4 Politics and Big Data

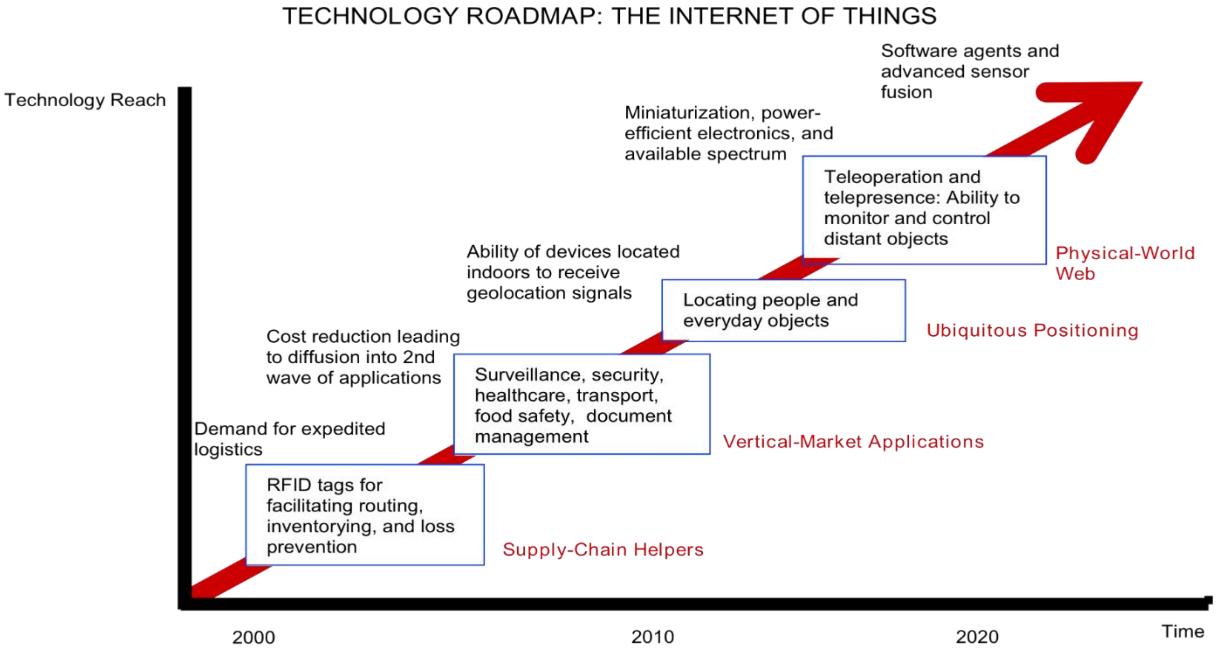
Law must be adapted to this new situation. Data from public cameras, different networks deliver information. Who is allowed to get which information. A regulation is needed. In the country and international. In a global network national legislation not enough. International regulations are needed. Information and Data became a power-tool. A power tool which is manly used by devoloped countries.
Politicians should make the regulation, but they see in Big Data also chance for their own career. United States of America are a leading country in this respect. Most data is stored in US. Most of spying is done from there. Obama´s administration announced in 2012 a Big Data research and development Initiative. Big data analysis was also an important fact for Obama's re-election in 2012.[5]
The US Federal Government owns six of the ten most powerful supercomputers in the world. Another one – the so called „Utah Data Center“ – is under construction by „United States National Security Agency“. It will handle a large amount of information, which was collected by the NSA.

Nowadays the war and terror from small organisations and single persons needs new methods how to handle data and information. Innumerable cameras are watching the public world and are used against terrorists.
Information from airlines are used for suppression of terrosism. Individual rights and privacy are less important than public interests.

[5] Picture: http://blogs.reuters.com/great-debate/2013/07/25/obamas-plan-one-nation-under-government/

## 7.5 Internet and Things

In a next step „things“ will bring data to the networks. This will make a new jump in the development. Already today end users collect digital photos, they very often can not be stored localy. The need a cloud. They store the pictures in Drop Box or similar. At the beginning of this century a new technolgy came for logistic and administration: RFID – Radio frequence Identification. The former BAR-code will be replaced by small computers, they are able to intentify in an automated way the characteristics of a good. These are the new supply chain helpers.
Vertical market applications came in the first decade of the 21st century in operation. Surveillance, security, health care, transportation, document management and many other fields produced data in big volume and started to select and interprete.
With the help of geometricals signs it was possible to locate things and Ubiquitous positioning


TECHNOLOGY ROADMAP: THE INTERNET OF THINGS
Technology Reach
Software agents and advanced sensor fusion
Miniaturization, power-efficient electronics, and available spectrum
Teleoperation and telepresence: Ability to monitor and control distant objects
Physical-World Web
Ability of devices located indoors to receive geolocation signals
Locating people and everyday objects
Ubiquitous Positioning
Cost reduction leading to diffusion into 2nd wave of applications
Surveillance, security, healthcare, transport, food safety, document management
Vertical-Market Applications
Demand for expedited logistics
RFID tags for facilitating routing, inventorying, and loss prevention
Supply-Chain Helpers
2000
2010
2020
Time

Source: SRI Consulting Business Intelligence

Gartner expects, until 2020, 25 billion devices in the net. But also tradional things like computer and notbooks we register an enormous growthrate. Alone between 2010 and 2016 we will have 8% more notbooks and desktops – all in all 3 billion units – and 33% more tablets and smartphones (6 billion).
Since 2008 people have downloaded 200 billion Apps. Out of this 50% were done in 2013, which is 100 billion in one year.
This is possible, because we have more and more smartphones. In 2014 it were 2 billion smartphones and in 2020 experts expect, that 8 out of 10 mobilephones will be smartphones[6].
The market for mobile phones is a heavy growing, but also very competitive market. In the last years we had several changes in the leading companies: from the american Motorola Nokia took over first place and after a technical lack in Finland

[6] Source: Gartner, IDC

Samsung became number one. But market leader never means technological leader. Apple is still the inovator, but others make the big business. In 2015 the ranking in the worldwide mobil phone market looks like this:

* Samsung
* Apple
* Lenovo with Motorola
* Huawei
* LG Electronics
* Xiaomi (Chinese “Apple”

The Far East became, after USA and Finland, the leading area for this technology. China started as producer and now they took over the technological leadership.

A 2 person household in Europe had in 2015 1 Desktop, 1 Notebook (+2 company Notebooks), 2 Tablets, 2 Smart Phones, 1 Music Player, 1 Media Box, 1 Blue Ray Player, 1 Camera, 1 SAT Receiver, 2 Printer, 1 NAS, 1 Switch, 1 Router and 1 Access Point. Every person had several IP adresses.

All this infrastructure makes it possible, that „BIG DATA“ can be.
Big Data in a dual, interconnected manner:

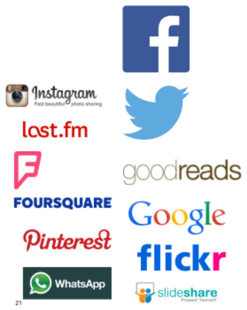
- Targeting of consumers
- Data-capture

Big Data is a modern form of communication. It is driven by Social Media and leads to a hyper-networked society in which individual acts are spread via the network. The society is dependent from social media. To close an account at a social network is like a „Virtual Identity Suicide“. Social networks are very powerful. Their size is enormous. By end of February 2014 Facebook had 1,228 billion user. Out of this 282 million were in Europe, 27,38 million in Germany and 3,24 million in Austria.

The competion of Facebook is very much smaller, but still big volume of users:

* Instagram: 200 million user
* Twitter; 232 million active user
* Google+ 300 million
* Xing: 12,65 million

The list of Social Network alternatives is long:
* Facebook
* Instagram
* Twitter
* Google
* Xing
* Flickr
* Foursquare
* Last.fm
* Delicious
* Scribd
* Slideshare



* Covestor
* Pinterest
* tumblr
* Goodreads
* WhatsApp

## 7.6 Open Data

The volume of data, which is offered now is one aspect. The other one is the public availability. It is called „Open Data“.
Open Data are Data, which is

- freely available,
- can be used by everyone,
- can be republished and
- has no copyright.

Everyone can use these data for free and can use them also for republishing. There is no copyright on „Open Data“.

What is the difference between „Open Data“ versus „Closed Data“?

- The access is with open data for everyone and for closed data it is restricted to persons or organisations.
- To take data from „Open Data“ is free (no costs). Closed Data are allowed to take fees for the access.
- Open Data have legaly no licence fees. Closed Data can have usage rights.

For many applications we use today we need open data. Our world is observable and has a configurable infrastructure with

- Social Web
- Smart Phone
- Smart Home
- Smart City

The objective is efficiency. For example traffic control and traffic lights make the public traffic more flowing.

## 7.7 Users of Big Data

### 7.7.1 eBay.com

Has two data warehouses with 7.5 petabytes and 40PB. In addition 40PB Hadoop cluster for search, consumer recommendations and merchandising.
Inside eBay a 90PB data warehouse is in use.

### 7.7.2 Amazon.com

Amazon handles millions of back-end operations every day. They have more than half a million third-party sellers.
Amazons IT is Linux-based. In 2005 it was the world’s three largest Linux databases with 7.8 TB, 18.5 TB, and 24.7 TB.

### 7.7.3 Facebook

Facebook was descripte already before. In addition to this information they handle 50 billion photos.

## 7.8 The Future

Where all this will lead us?

Many changes. Teachers are not any more instructors, they are leaders. Normal consumer can not get the answer directly. A sentence like „Google will know the answer to your question before you ask it“ sounds funny, but is realistically a problem.

The future with Big Data is open, but the past is past.

Technology is neutral - they can both good and bad. It is, what human beings make with all this information. There can be still more progress and more data, but a way back to a "Digital Biedermeier“ seems is not possible any more.