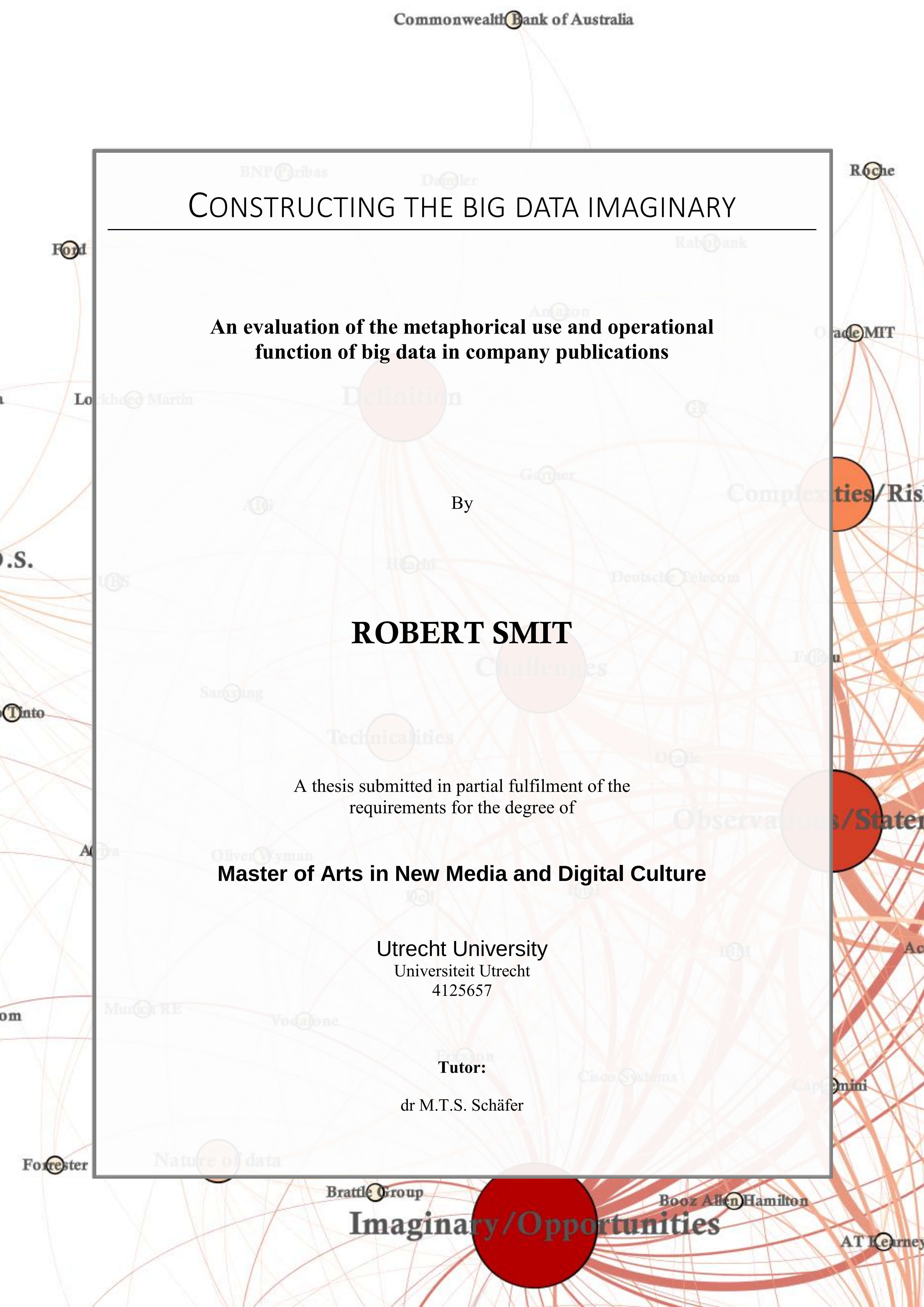


CONSTRUCTING THE BIG DATA IMAGINARY

An evaluation of the metaphorical use and operational function of big data in company publications

By

ROBERT SMIT

A thesis submitted in partial fulfilment of the requirements for the degree of

Master of Arts in New Media and Digital Culture

Utrecht University
Universiteit Utrecht
4125657

Tutor:

dr M.T.S. Schäfer

CONTENTS

ABSTRACT

This research investigates the metaphorical use of big data in 275 company whitepapers published from 2011 until 2015. Lacking a clearly demarcated definition the term functions as a material metaphor connecting data practices and its proposed benefits. Qualitative analysis of the corpus revealed that the big data is promoted as a tool that resonates corporate values such as cost saving, effectiveness and decisiveness amongst others. Colorful imagery and lively models are used to exemplify these benefits. Quantitative analysis uncovered that big data is most often used in quotes on the opportunities of big data, that a business lexicon is used to explain these opportunities and that the sentiment of these quotes is predominantly positive. All documents in the corpus are informative, describe problems, explaining complex issues and propose solutions to these problems. More importantly, they serve to persuade customers and partners by promoting ideas and concepts. Big data is very carefully constructed metaphor expressed through authoritative and positive language. It creates a self-referential system because it does not require meaningful statements to be constructed using outside references. Without big data a company would have to undertake a multi-threaded conversation in which every element of every data practice is connected to every trait. Using big data metaphorically it is possible to group these traits together and communicate them under one banner.

MANY THINGS TO MANY PEOPLE

CORRELATION OVER CAUSALITY?

The term big data is pervasive yet its meaning also brings about confusion. It adheres to much more than technological concepts as, with the advent of new digital devices and platforms, big data seems to have seeped into the daily lives of individuals, researchers and companies. In the wake of these digital innovations data-driven business models are becoming more prevalent, and people are starting to notice. In their popular scientific book *Big Data: A Revolution that Will Transform how We Live, Work and Think* data-editor Kenneth Cukier and Viktor Mayer-Schönberger (2013) go even further; they state that “Big Data’s ascendancy represents three shifts in the way we analyze information that transforms how we understand and organize society”. Big data ought to present a future in which all the data related to a specific phenomenon is available. Eventually we would start going to data-based research that underline explanations for large phenomena, instead of focusing on one particular event. Finally this might lead to a situation in which causality is no longer after in the data. Rather patterns will be found which are pointing out that things are happening, not why they happen. The story goes that an American retail chain used big data to predict when women are in their third trimester, and then sends them special coupons. Upon getting these coupons in the mail a teenage girl’s father was enraged with the company, accusing it of encouraging teen-pregnancy. But after finding out that his daughter had been withholding this information about her pregnancy from him, he made his apologies to the company (Duhigg, 2012).

For companies big data is presented as the next big thing in marketing, forensic analytics, capital management, decision making, and so forth. On the other hand people are starting to get worried that these developments might infringe on their freedom of choice and right to be left alone (Krin, 2015). But amidst the vast media coverage a clear definition of big data and its implication seems lost. Yet the term is pervasive in company publications as well as in popular media. And although neither a definition nor the implications of big data are evident, the term itself seems to have a prominent place in business, scientific and popular discourses. So superficially big data seems to be many things to many people and seeking to define big data is virtually impossible. Yet the term is used frequently. So what is its function?

BIG DATA AS A METAPHOR

Big data introduces the possibility for people to talk about data practices with a degree of mutual understanding that was not possible until now. It is a bridge between subjects in specific (technical) fields and the public discourse. While allowing for the transference of meaning from one domain to the other it allows for people with different backgrounds to have conversations on the topic. By using big data as a concept in conversations our understanding of the world around us, and as such its function, may be described as metaphorical. And transference of meaning does not only happen between words (or signs) but also between these signs and the tools that it describes and utilizes. Big data not only shapes understanding of the world it also has a pragmatic dimension that operationalizes the concept through tools and procedures. In *Transcoding the Digital: How Metaphors Matter in New Media* van den Boomen (citing Hayles) describes this function of metaphors as 'physical objects that, through its construction and functioning, act as a crossroads or a juncture point for the traffic between the physical and the verbal (Van den Boomen, 2014). Traditionally metaphors connect concepts that share conceptual similarities, and to investigate them would entail how one can be understood in terms of the other. With big data this is not the case as it there no other concept which it references, no other domain that it utilizes to be understood as. But its metaphorical function is also evident; without it a large discourse would not exist. The aim of this thesis is to investigate the function of big data as it seems to fall outside functional domain of traditional metaphors by investigating its 'materiality'. This moves the term away from an abstract concept by exploring the media in which it is embedded, which words are used to describe it and the topic of texts in term is used. Rather than investigating the definition of big data the focus is put on how big data as a term is constructed. Overall this leads to a better understanding of how data practices are portrayed and operationalized. Furthermore insights will be generated that cast a light on how metaphors, such as big data, are used in the discourse to create a conceptual and pragmatic understanding of the world.

CORPORATE METAPHORS

Companies use big data actively to promote new business opportunities, introduce features that personalize technology or reactively to reassure that the data their consumers generate are treated ethically correct. New business models become more data-driven and the product and its data aggregation more symbiotic. But media studies in general do not often consider corporate language from an inside perspective. Marketing, PR and advertisements have been studied from a consumerism perspective in terms of framing, representation and possibly manipulation (Long & Wall, 2012). Yet there a few accounts of research that report on the workings of the corporate world, or how it makes sense and constructs meaning towards itself. However, companies are often opinion leaders whose expressions can be read as explaining the corporate world, whose research affects policy making and reporting in the media. They have a big impact on our daily lives and their announcements are reported on by popular media. And the media they convey can be studied like any other corpus in media studies. Studies by Franco Moretti and Dominique Pestre (2015) who conducted an analysis of how the World Bank frames its missions' statements by analyzing their annual reports, or Ramón Reichert (2009) who investigated how the use of the media influence the stock exchange through metaphors and how this use is shaping public opinion, these studies and statements serve as an example of the thesis under investigation. To study the way big data is used by companies a corpus consisting of corporate publications will be used. The problem this research will try to answer is not '*what is'* big but the research focusses on the way big data is used in communication. This yields insights into the workings of the concept and which ideas and functionality is attached to it. To answer this question is to investigate the question:

What is the metaphorical function of big data in company publications?

Sub questions:

1. Which words are associated with 'big data' to frame data practices in company publications?
2. Are there words and images that re-occur often, or is there a difference between publications?
3. How are the characteristics of big data (technical, methodology) connected to its portrayed possibilities?

In addition to revealing the functional value of big data answering these questions will shed a light on the use of corporate publications in general. As they can be opinionating in the business ecosystem and they yield the power to influence how businesses are run. Rather than looking from the outside in, at how companies portray themselves to society, this research will look at how companies talk to each other.

INVESTIGATING A METAPHOR[1]

Analyzing the company publications on big data will happen in the same way that Moretti and Pestre conducted their research, although the analytical scope will be different. Usage of the term big data has only been widespread and 'trending' since early 2010 and therefore there exists not enough material to do a comparative analysis. Moreover, as will be delineated in the next paragraph, the definition of big data remains too vague to distill a clear research object from. Big data will be studied from a metaphorical angle to investigate what its function is in constructing a (technological) imaginary.

SELECTING THE CORPUS

Analysis of how corporates use big data will be done by analyzing their publications. These are selected by two rankings:

1. 'Global Fortune 500' list 20122
4. 'Vault Consulting 50' list 20153

The Global Fortune 500 (GF5) is a global company ranking list based on annual profits across industries sectors. The Vault Consulting 50 (VC5) is a list of consulting firms ranked by their employee satisfaction status. Their survey methodology included metrics such as 'prestige', 'satisfactions' and 'compensation' amongst others. Using both lists ensures that the corpus-data represents at least a majority of the publications that are currently circulating on the topic.

DATA COLLECTION

Several steps are taken to ensure that the most relevant publications are selected for the corpus. First the central company website was determined by a Google search using a specific google search query[4]. The international company website was always selected over the national website to increase the likelihood of a high amount of publications in the English language. By following the specific google query only the files that where published in the .pdf format are shown as results. If this query returns with zero results, the .pdf parameter is dropped and all file types are queried on the domain. Website pages, .doc & .ppt files containing articles or slideshows on big data will be converted to .pdf files to maintain uniformity. If a publication is found it is given a filename using the following format:

[year of publication] – [Company name] – [Publication Title].

CREATING THE CATEGORIES

Analysis of the corpus will be done by selecting all the sentences in the publications that contain the phrase 'big data'. This is done in a supervised manner, meaning these are categories in which the topics are sorted and so they will be pre-determined to control the analytical scope. The scope is determined

[1] For an in-depth overview of the methodology, see appendices

[2] http://topforeignstocks.com

[3] http://www.vault.com/company-rankings/consulting/vault-consulting-50/

[4] "big data" site:[website URL] filetype:pdf

by the assumption that big data is a container concept that is used to frame a technological imaginary. Categorization is done based on description of how an earlier imaginary, that of the internet, was similarly constructed. This refers to the notion of technological imaginary, which is constituted in the collectively shared visions on the social impact of technology. To exemplify: in framing the ‘imaginary of the internet’ Partice Flichy (2002) investigated the collective vision that shaped the emergence of the internet. Revisiting the promises of new media as enabling technologies, Schäfer developed the notion of a “rhetoric of progress” that has been associated with new technologies (Schäfer, 2008). Themes that can be condensed from this text are (amongst others); ‘collaboration’, ‘access’, ‘communities’ and ‘problem solving’. The categories used here are distilled from a sample of the corpus. Consequently, as the categories are a distillation of the corpus, they will inherently resonate the terms and codes that are relevant within this discourse.

Sampling the texts on which the categorization is based has been done using the following parameters:

1. 1 text from every year of publication (2011 – 2015)
2. Text length = >1500 words
3. Content quality should be sufficient (density = >0.10)

Quality in the sense means that the texts offer in-depth information on big data to base the categorization on (determined by big data references). The following tables detail the chosen texts.

Year	Company	Title	Length	Density[5]
2011	McKinsey & Company	*Are You Ready for the Era of Big Data?*	3456w	0.128
2012	IBM	*The Real-World uses of Big Data*	9747w	0.346
2013	PWC	*Capitalizing on the Promises of Big Data*	3333w	0.342
2014	Dell	*Big Data: Unlocking Strategy Dimensions*	1923w	0.171
2015	Gartner	*Big Data Business Benefits Are Hampered by ‘Culture Clash’*	2607w	0.253

Table 1: Publications selected for first categorization

CODING THE DATA

Extracting relevant data from the text is done by coding certain parts of the text like phrases or titles and turning them into ‘quotations’. This is the term that is used by the qualitative text analysis program atlas.ti (v.7) which is used in this step. The codes act as containers in which quotes are stored if they meet the criteria that demarcate the code. Finding quotes is done by searching for all the sentences in which big data appears by using a specific search query[6]. The code matrix below shows the possible categories that can be ascribed to a sentences.

[5] $density = \frac{(big\ data\ references)20}{total\ words}$

[6] big data|‘big data’|”big data”|big data = |big data: (the search is not case sensitive)

Code	Keywords	Quotes referencing:
Challenges	challenges	Obstacles that are identified before data-based projects or business models can come into fruition. This category is different from complexities/risks in the way that its sentiment is neutral where the latter is negative. For example: “in order to capitalize on the possibilities of big data companies will have to invest in hiring people with new skillsets.”
Complexities	privacy, risks, dangers, problems	Inherent properties of data practices that make them a complex and multi-disciplinary endeavour. Statements that shed light on the risks/dangers of using big data are also filed here.
Organizational factors	management, leaders, owners	Factors that companies need to take into consideration apart from data or technically related issues. For example: ”in order for big data projects to come into fruition business leaders that drive the idea forward in an organization have to be identified.”
Definition	defining, meaning	Attempts at explaining what big data is. This can either be done in a general matter (e.g. ‘big data means […]’) or specific to the subject of the publication (e.g. ‘within these parameters we define big data as being […]’).
Nature of data	geographic, social media, real-time, (un)structured	Everything that is referencing the nature of the data that is big data sets. Statements on the aggregation, access and dissemination of data, although sharing similarities with the technicalities, can also be stored here.
Observations/ Statements		Here companies make a statement or observation that are related to big data but not otherwise fall into one of the categories. For example: ‘In particular, companies routinely lose opportunities because they use poor-quality big data to make major executive decisions.’ (Boston Consultancy Group, 2015)
Opportunities/ Imaginary	possibilities, opportunities, vision	Companies express their ideas and vision with regard to data practices and the place that it will have in their company/business-model/society at large. These quotes represent the gospel of big data, its imagined impact, possibilities and opportunities that are the result of implementing new ways of working with data.
Technicalities		Everything involved with the technical requirements to analyse and operationalize data sets (from storage to infrastructure and analytical tools).
Titles		Titles or headers in which big data is mentioned
Questions		Sentences in which there is a question regarding big data (usually in interviews)

Table 2: Coding categories

LIMITATIONS

Working ‘from the ground up’ means that the categorization is intricately intertwined with the data. Theories created through analysis of the data can therefore not be refuted by external data, as its scope would inherently introduce different parameters into the sample (Glaser & Strauss, 1995). Because of this the scope to which the results can be applied only extends as far as the sample size. Simultaneously, because the data and categories are linked the results following the analysis are only applicable within the parameters of the data. Meaning that changes or additions to the data would inherently undermine the research results rather than expand them. So the results are only relevant for company publications on big data that fall within the timeframe of 2011 to 2015.

The categorization is based on a small portion of the case (see below) and is definitive once coding starts in atlas.ti. It is not practical to change this once quotes get appended. Yet is may happen that due to new insight the categorization needs to be revised. For the sake of keeping the process manageable the categorization will be revised only if new evidence presents itself that makes it totally unjustifiable to stick to the old scheme. If this happens all the previously coded texts also have to be revisited.

EXTRACTING & CLEANING THE DATA

Data parsing with Python

After all of the data has been coded to categories in atlas.ti it is possible to generate report of all the quotes per categories[7]. From this report the following information needs to be extracted (or parsed): year of publication, company, category and the quote. Doing this by hand would be a tedious and prone to human error. As it is a straightforward procedure the task will be scripted using the programming language Python 3.5.0[8]. The data is parsed to a .csv file so that it can be used in the next phase of the analysis. Afterwards the category 'industry' is manually added to the sample. The following table show the header and first row of the da**ta**.

Year	Company	Industry	Category	Quote
2014	Amazon	Products	Big Data N.O.S.	This whitepaper provides an overview of the different data options available in the aws cloud for architects data

Table 3: Data entry example (one node)

Word frequency analysis in R

Using the R programming environment all the codes are fed into an algorithm that preform the following steps by year per category[9]:

1. Read .csv data into memory
2. Transform string data into vector
3. Produce corpus. This puts all the words into a 'bag of words' where they can be counted
4. Transform to lower case; remove punctuation/stop-words and whitespaces
5. Transfer all words to a data frame
6. Count words and sort decreasing in frequency of appearance
7. Produce top 5 words

[7] For an example of this report see appendix 2

[8] See appendix 3

[9] See appendix 4

The top 5 words is manually transferred into a new data table. After the first category has been analyzed the an excess of the words ‘big’, ‘data’, ‘can’, ‘also’ and ‘will’ produced worthless data so they were also added to a list of words to be removed.

Sentiment analysis in R using naïve Bayes classifier

Classifying a quote’s sentiment as either positive, negative or neutral requires a mathematical algorithm. To do this a naïve Bayes classifier will be used because of its simple application and efficiency in dealing with large datasets[10]. It is a supervised learning algorithm that uses probabilistic classifiers and assumes all the variables (or predictors) to be independent. The algorithm needs to be trained before it can operate properly by being shown what certain sentiment polarity means. So it regards words as ‘good’, ‘excellent’, ‘fun’ etc. as positive, and ‘bad’, ‘depressing’, ‘loss’ etc. as negative. The lexicon that is used for training is Janyce Wiebe’s subjectivity lexicon[11] that lists positive and negative words and is integrated into the ‘sentiment’ package in R.

The algorithm calculates the probability factor by word per class, adds up the totals to create a score per polarity class, normalizes these scores, and then compares them to get a best fit. A higher probability in the positive class results in a positive polarity, and similar for the negative class. If the calculation results in equal negative and positive polarity scores the algorithm assumes the quote to be of neutral polarity.

CORPUS DATA

Year	*n* publications	*n* quotes
2011	5	67
2012	34	674
2013	52	758
2014	91	1006
2015	92	744
Total	**274**	**3249**

Table 4: Metadata

Metadata on the sample reveals that the increased publication on big data follows the same trend as can be seen in academia, with the rise plateauing after 2014 (table 3). It is premature to conclude that big data as a trend is already on the decline, but corporate attention seems to be wavering.

***n* companies in sample** 79

Top 5 most quoted companies & industries	**Companies**		**Industries**	
	EY	305	Consultancy	1350
	IBM	300	Technology	802
	Oracle	269	IT	742
	Cisco Systems	195	Telecom	127
	Capgemini	179	Finance	75

Table 5: Companies in case sample

[10] See appendix 5

[11] Available via http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

The consultancy, technology and IT industries are by the most prevalent in de data sample. This is reflected by the top three companies (EY, IBM and Oracle) which are consultancy and IT companies. After the top 3 there is a relatively large gap in both categories. Because of this some of the analysis will be done using only the companies or industries from the top five list. Other amounts in the sample are so small that they can be neglected in terms of relevance to the amount of quotes being less than 1 of the total quotes[12].

THE LANGUAGE OF WHITEPAPERS

DESCRIPTIVE FEATURES

Like other media and publications company publications are distinct in their features which make them stand out and allow people to recognize them as such. Their makeup is largely irrespective of the topic but rather is determined by its type. An annual report looks different than a point-of-view or whitepaper report, but all whitepapers look alike. Together all publications represent a discourse which reflects corporate areas of interest and provides insights into its cultural values and motives. In essence all documents are informative, describing a problem or explaining complex issues. Additionally, they serve to persuade customers and partners by promoting ideas and concepts. As such they are recognized as 'grey literature' i.e. literature not published in indexed peer-reviewed journals and publications and where publishing is not the primary activity of the producing body (Van Cauwenberghe, et al., 2010; Schopfel, 2010). As such the quality of the publication can vary greatly as well as the price. Mostly they are freely available as promotional material and as such the quality is often very high. Only Gartner's industry reports had to be licensed. There is a difference between corporate content and 'sponsored' content. The latter are reports published by companies that specialize in research. Their language is more academic and less activating with more references to academic sources to back up argumentation. Like academic publications the tone of voice is intelligent with a strong emphasis on business intelligence. The recommendations sections is often employed to ty research observations to a business offering from the sponsoring company.

CONTENT

Most publications start off with a testimonial by a CEO or statement from a senior staff member commenting on the industry at large, or the state-of-affairs concerning the leading topic. These statements are underlined by the executive summaries; they are the key finding of the report and almost exclusively presented as short bullet point statements. Following is a section along the lines of 'what is big data?', where companies explain their idea of big data and how it came the subject of the report. Tone of voice is usually authoritative and optimistic.

[12] For a complete overview of the total relative amount of quotes per industry per category see appendix 6

This is illustrated by the following quote from a Deloitte report:

> "Don't pursue your big data dreams at the expense of small data value. There's plenty of room for both—and each offers different strengths for different challenges. The guide below shows how they can be put to their best use." (Deloitte, 2012)

Visually the reports are very appealing. Color usage can either be in line with the company's brand identity or fitting with a series of publications. The reports are stylized in such a way that its textual and visual content complement each other, making for an elaborate business card. This is emphasized by the free distribution of these reports. Online corporate portals often have a special 'insights' sections dedicated to distributing them. The visual appeal of the whitepapers is demonstrated below by a diagram taken from an IBM report and a model from an Ericson report.

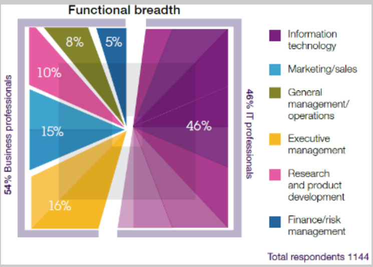
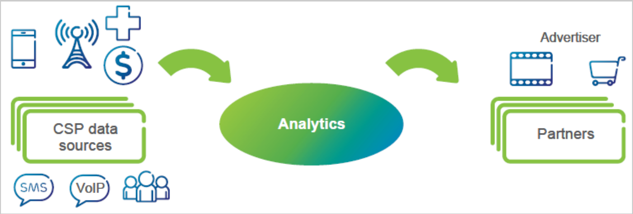
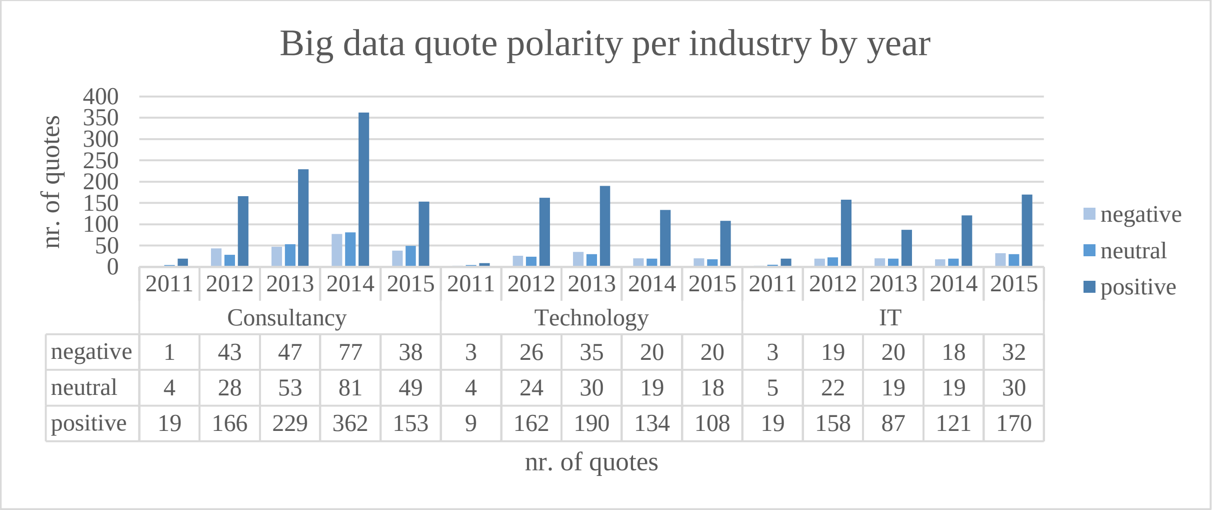
IBM Interview data diagram	Ericsons' 'brokering network data insights' model
	

Table 6: Examples of big data models

The main argument of the text is commonly illustrated by using either interview data or models. Interview results are presented quantitatively like the IBM example above. Rather than reference sources from other authors, insight gained from surveying 'industry leaders' is used to substantiate claims and arguments. Looking from the outside one could get the impression that, much like academia, corporate publications construct their own discourse where facts have been build up in a distinct (yet structured) fashion. If surveys are not the crux of the report models form cornerstone information elements. The textual explanation they are embedded in serves to underpin what is shown in the model. Again one hardly refers to sources outside the company outside the company to establish theoretical context. The method used to survey an audience is mentioned only when relevant to the context of the report. IBM for example states the timeframe, scope and sample audience in their *'Real World Uses of Big Data'* report but does not include the questions or statistical methods so that the findings are reproducible. In total the explanation is one paragraph long and serves more as a disclaimer than a methodological reflection.

STRONG POSITIVE SENTIMENT



	Consultancy					Technology					IT				
	2011	2012	2013	2014	2015	2011	2012	2013	2014	2015	2011	2012	2013	2014	2015
negative	1	43	47	77	38	3	26	35	20	20	3	19	20	18	32
neutral	4	28	53	81	49	4	24	30	19	18	5	22	19	19	30
positive	19	166	229	362	153	9	162	190	134	108	19	158	87	121	170

Figure 5: Big data quote polarity per industry by year

By using the word ‘new’ the general statement derived from the word frequency analysis already constructs positive meaning around big data. By analyzing the individual quotes on their sentiment it is possible to determine the portrayed predisposition towards big data per industry. The following graph provides an overview of the three most prevalent industry and the sentiment that can be derived from their quotes on big data13.

Consultancy represents the biggest part of the case followed by technology and IT companies and the graph is a faithful representation of distribution. It shows that, on average, companies are split on their neutral or negative quotes on big data, but that the majority has a positive polarity. The difference between consultancy companies and IT companies is noticeable. Where the first had a steady increase in positive quotes until 2014 the latter shows a decline in positive comments in 2013 where negative and neutral quotes remained similar. Small differences aside, the overall polarity of quotes is distinctly positive.

[13] For a complete overview of the quotes’ polarity scores per industry see appendices 7.

CONSTRUCTING THE BIG DATA IMAGINARY

As big data is a multidisciplinary endeavor the following chapter is dedicated to outlining the phenomenon from three different perspectives. Firstly, there is the data[14] itself, investigating which is mostly an ontological endeavor. Data is a subset of reality, it tells something about the world and can be used for the generation of new information and knowledge. Investigating what data is shapes understanding of what is told in the corpus and is the first step towards understanding how big data is portrayed. The corpus contains company publications which are assumed to be selling, promoting or to be otherwise convincing the reader of the benefits of big data. The second part of this chapter will therefore investigate what the texts imagine the corporate implications of big data to be, and how this imaginary is constructed. A model will be put together that demonstrates how big data is the metaphorical conduit that ties together the data and its imaginary. The final part of this chapter will demonstrate how the model reflects the theoretical function of big data as a metaphor in company publications.

THE DATA

At first sight there seems to be a gradual increase in the popularity of publications on big data from 2011 onwards. Google's N-gram viewer, which uses text or speech input to query a database of *n-grams*[15], shows that usage of the term first arose in book publications at the end of the 1920's. If one reads the publication abstracts of these books it quickly becomes apparent that 'big data' was only used in the literal sense, namely to delineate the size of a dataset as 'big'. And even though that definition has not gone unused, nowadays the term is used to outline a broader set of definitions. The first company to expand on the definition of big data was the IT & Technology research company Gartner Inc. In 2001 their analysis Dough Laney (as part of the META Group) published a short paper on the increasing demand for proper data management, outlining data- 'Volume', 'Velocity and 'Variety' as its key-characteristics (Gartner/META, 2001). This terminology remained largely unused until in 2011 companies started to use the term 'big data' in publications on a larger scale. As can be seen in the methodology chapter the publication amount spiked in 2012. One sees a similar trend in academic publications. According to the insight driven website (that has strong ties with publishing firm Elsevier) *researchtrends.com* the phrase 'Big Data' was first observed in a 1970's publication on atmospheric and oceanic surroundings. Until the 2000's big data was predominantly used in the fields of engineering and computer engineering research. From 2000 onwards this shifted more and more towards the field of computer science, and mathematics (Halveli & Moed, 2012).

Collecting data is often a costly and time consuming and therefore the size of the sample from the total sum of data has to be adapted to fit research goals or match computational power. But with the introduction of new software platforms and technologies the aggregation of data has increased

[14] 'Data' in this paper is used as a mass noun referring to collections of information. Although technically incorrect, sentences like "data was collected over a number of years" are widely accepted in Standard English (Rogers, 2012)

[15] A N-gram is an N-character slice of a longer string (Cavnar & Trenkle, 1994)

explosively to the point where sampling could become unnecessary (Cukier & Mayer-Schönberger, 2013). Or, as worded by consultancy firm EY:

> "Ernst & Young defines "Big Data" as: Very large data sets volume that are being produced at a tremendous speed by the growing digitization of the society velocity and consists of data from all possible sources from structured to unstructured variety" (EY, 2013)

For corporations it is promising to think that the increased influx of data from the digitization of society represents new business opportunities. In the case of closed systems such as Facebook or Amazon it is even possible to capture all the transactions and interactions. Big data is portrayed as playing into the shift of digitization by enabling all the data to be captured (where n = all) and sort and sift through it later. Financial company BNP Paribas describes the possibilities of this new type of data management as leading to new marketing model:

> "Having built the infrastructure that supports the digitalization of information the big data industry created marketing models based on the management of data and of user behavior." (BNP Paribas, 2015)

More equals better, and big data is presented in such a way that a company's ability to capture and store data and the value of the analysis derived from that data increase in parallel. But as soon as all data is captured new challenges will present themselves. Ethical and technical questions with regard to storage aside, if big data is presumed to be a byproduct of digitization where every action is captured, then sorting useful data from useless data is still a challenge. Simply dumping more variables into a predictive model does not increase its effectiveness. Paradoxically the chaotic nature of big data is presented as an opportunity equally often as a challenge[16] as exemplified by a quote of consultancy firm Boston Consultancy Group:

> "Using unstructured and imperfect big data can make perfect sense when companies are exploring opportunities such as creating data driven businesses and trying to better understand customers products and markets" (Boston Consultancy Group, 2015)

Knowledge and wisdom that can be derived from data analysis seem inherently connected to each other, where the likeliness that the results represent the studied object faithfully increment with the size of the data used. But the data scientist has to do more than merely reach out his hand and wait until the data accumulates in it. Either type of data still has to be collected, funneled through hypothesis, sorted, analyzed, categorized, and reproduced faithfully as 'finding' (Markham, 2013). But regarding data as the consequence of usage of other products and services rather than the primary output of explicit data gathering might proliferate the notion that it is simply out there for the taking. Unfortunately this aspect of big data is rarely portrayed in corporate publications.

Dana Boyd and Kate Crawford expand on the epistemological idea that by thinking along these lines, big data introduces a new way of thinking about analysis. Not only does the term rest on the interplay

[16] See appendix 8 for a complete list of quotes referencing the unstructured nature of big data in the categories Challenges and Imaginary/Opportunities .

between technical and analytical capabilities, it also constructs the belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity and accuracy (Boyd & Crawford, 2012). In the publications this belief has been solidified in three pillars that underline the shift from traditional data practices to big data. At first there is a shift in the volume of data that is generated, second there is the variety of data that is in datasets and lastly there is the velocity at which new data is generated. Doug Laney was the first to ascribe these characteristics to data and Gartner adapted the definition to fit big data by stating that as a phenomenon it was about 'High Volume', 'High Variety' and/or 'High Velocity' (Gartner, 2012) and I.T. In 2012 IBM added 'Veracity' to the definition where they accounted for the inherently unstructured nature of the data (IBM, 2012) as can be seen in the following quote:

> "[…] and while they cover the key attributes of big data itself we believe organizations need to consider an important fourth dimension veracity." (IBM, 2012)

The 3V's model is mentioned in 18 percent of the company's quotes on the definition of big data[17]. Other attempts have been made to group quotes on the definition of big data yet the descriptions were too dispersed in meaning to distill clear categories from it. Although the (x)V's model gets referenced most commonly, it deals with the ontological questions of big data only superficially. Initially the publications are informative to the degree that they provide the reader with a new perspective on data practices and how they could be transformative for either the industry or company processes. But definitions are kept universal, the tip of the iceberg that is non-technical and applicable to a broad audience. It should entice the reader to get to know more about big data; now the company may step in with a paid consultancy service or technical solution. Moreover, by keeping the definition broad and on the surface, big data can be presented as a 'solution' to a broad range of problems. Adding more specificity might deter customers who fall outside of the presented scope from becoming interested. It is important that the image that is created of big data is enticing. How this works in company publications is the focus point of the next section.

THE IMAGINARY

The corporate definitions reveal an instrumental proposition of how big data should be viewed. Output, or output delimiting metrics such as satisfaction-, and adoption ratings and cost-saving, are used as reasons why other companies should also start 'using' big data. The following two quotes are selected from the Opportunities/Imaginary category to exemplify this tendency:

> "A roughly equal percentage (37 percent) thinks organizations can achieve extremely large cost savings with big data" (Accenture , 2014)
>
> "Big data tools and technologies offer ways to efficiently analyze data to better understand customer preferences to gain a competitive advantage in the marketplace and to use as a lever to grow your business" (Amazon, 2014)

[17] 36 out of 193 in quotes on the definition of big data mention the 3V's model, see appendix 9 for the complete list of all quotes.

By speaking of big data in terms of its usefulness, or its ethical considerations we portray it as if it has been put into the world with a clear purpose. As a means to an end. And like many socio-technological phenomena before it, if big corporates start spreading the word it eventually becomes a trend. By repeatedly communicating that big data is going to be the new tool of business, companies have the power to make it a reality. The internet is a good example of how this happened before. When it was first conceived of as a communication tool the word started spreading of how it could lead to new forms of corporation and effectiveness in using valuable computer time (Flichy, 2002). It started small with just a handful of scientists united in groups like the Whole Earth Society that would lay the foundations of the internet or the Tech Model Railroad Club where the first computer prototypes where built (Turner, 2006). Both groups had strong shared values that were fundamental to the innovative processes that later spawned into industries of their own. Through their actions, innovations and visions were able to create a technological imaginary. Patrice Flichy proposes that this kind of imaginary plays a role in the creation of a new medium. But rather than being grounded in hard mathematical calculations and predictions it owes more to certain ideologies and desires that circulate within a specific group (Flichy, 1999). Like with the internet and personal computer before it the corpus on big data shows strong sign of describing a technological imaginary. The comparison with the internet is put literally in a publication by Accenture:

> “Many organizations believe big data will revolutionize business operations in the same way the internet did and feel big data will dramatically change the way they do business.” (Accenture, 2014)

Exemplifying how big data can be a valuable endeavor is the business model of Farecast. It employs a method for scraping and analyzing 200 billion price-flight records to forecast the best moment for buying a flight ticket (Cukier & Mayer-Schönberger, 2013). The system it employed provides no insight into the workings of ticket pricing, nor does it need to. If the savings are adequate the company has a valuable business proposition to its customers. In other situations more contextual information is needed to generate valuable insights. But whatever the situational requirements, the corporate discourse strongly resonates its values through its quotes in the category of Opportunity/Imaginary. For example quick and decisive decision making as put forward by high-tech company Cisco systems:

> “Innovative software has taken command to capture all this and provide new insights in organizations that are accessing big data, leading to faster better decisions and quicker responses” (Cisco Systems, 2012)

Meeting customer demand as outlined by financial company Credit Suisse:

> “Accordingly retailers offering multichannel services can best leverage big data to meet different customer needs either from their online stores social media platforms or personal instore interactions” (Credit Suisse, 2013)

Or to make medicine work more effective as worded by pharmaceutical company GSK:

> “But this analysis of so called ‘big data’ is the first of many small steps towards making medicines more targeted to a person’s genetic makeup and the way their bodies tackle disease” (GSK, 2015)

Big data is always cast in the light unique selling points that are relevant to the industry. The imaginary is the linguistic vehicle that carries ideas forward and makes them operational in conversations while allowing for the construction of a discourse surrounding it. Meanwhile it also allows us to imagine it capable of things that were not possible without it. It is this imaginary that can transform the formation of reality, bringing closer that what is poetic or conceptual (Malbreil, 2007). Given specifics of a phenomenon we might imagine which benefits it yields, or how it can transform our daily lives, for the good or the bad. The internet for example, during its inception and explosive growth in the 90s, was described as the 'Information Superhighway'; a metaphor that framed it as a transportation infrastructure for international commerce. Nowadays we do not use the description anymore nor do we mention using 'network computing and communication retrieval tools' (Flichy, 2002) even though that description is just as accurate. We use the internet as a metaphor, and by doing so we consciously or non-consciously shape how we make sense of an experience or phenomenon (Law & Pepperell, 2015).

FRAMING BIG DATA

The use of style-elements such as metaphors or analogies can help the reader to understand what a concept entails, but before that the writer has to decide *what* to tell. Or to put it differently: what is highlighted and what is left out by the writer? It is this process of selection and salience that is referred to as 'framing'. More precisely: to frame is to "select some aspects of a perceived reality and make them more salient in a communication text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described" (Entman, 1993). Framing determines the way something is portrayed to the public, and consequently how opinions about it may arise. Consequently this process is susceptible to change as new discoveries are made or (public) opinion shifts.

THE CORPORATE BIG DATA IMAGINARY

In many reports big data is framed in relation to industry specific topics. For example, a telecom company may use it to illustrate how they bundle their subscription packages more effectively, and for an insurance company using big data is beneficial in predicting market fluctuations. Observations within a specific field are described and the imagined benefits of big data are superimposed over these observations, usually as a solution to a problem. This is reflected in a trend-line plot of the categories per year as seen in the graph on the next page.

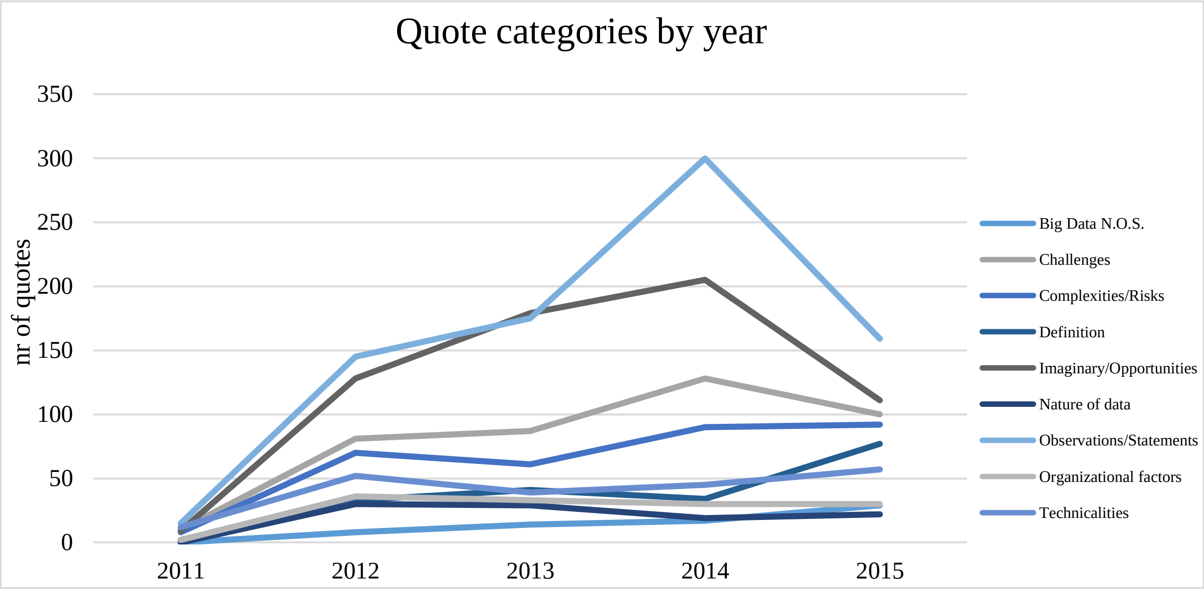


Figure 1: Overview of big data quotes per categories by year

The reports' focus is mainly on the observations made on big data practices and the possibilities that are created. 2014 marks a peak in publications as well as observations made on big data. Close observation reveals that the slope of imaginary/opportunities has flattened in 2014 in comparison to 2013 where it rose steeper. On the contrary, the slope of challenges and complexities/risks shows a steeper rise in slope and reversal from decline to increase respectively. In tandem with more observations on big data, companies also seem more aware of the challenges big data introduces and the complexities and risks that come with it. In 2015 this might explain the sharp decline in quotes[18] on the topic as well as a relative increase in quotes on the definition and of big data. The technicalities, complexities and risks categories also show an (although with a decreasing slope) increase in quotes contrary to observations, imaginary/observations and observations. In 2015 companies have become more careful with constructing a big data imaginary that only reflects its possibilities. Challenges and the complexities/risks associated with the phenomenon have become more prevalent warranting a more cautious attitude.

DIGITAL MATERIALITY

The big data metaphor has the power to make things happen in the real world, for it is embedded in a discourse that manipulates corporate methodologies as well as socio-material constructions and economic transactions. It surpassed the metaphor as a verbal figure as it is inherently integrated with operational processing units that can actually make something happen in the world. The transference of meaning that is established in the publications between the digital objects it describes and its proposed ramifications can be regarded as a material metaphor, a term that foregrounds the traffic between words and physical artefacts (Hayles, 2002). To evaluate big data in terms of materiality is to strip it down to

[18] Although this can also be because the data was gathered in November of 2015

its binary make-up, down to analog processing cycles of the processors that power computers. Technically speaking it makes sense to (visually) describe the multilayered web of interconnected systems, subsystems and languages expressing the functionality of those systems that constitute big data practices. Utilizing such a style of communication could be done in the shape of an online repository where code is shared as well as access to virtual processing machines. But in terms of the promotional nature of company whitepapers, this would be a less fruitful endeavor. Not only would the technical nature of this type of promotion discourage those who are less technically endowed, it is also contradicting their business model. So technicalities get abstracted into functions. Function into objects that get picked, shuffled, re-ordered, or left out as to become 'depresented' (Van den Boomen, 2009). When communicating about data the sender picks and chooses how to frame the message so that it fits his specific goal. This way the corporate big data discourse constructs what companies wish big data to be as detailed in the model below.

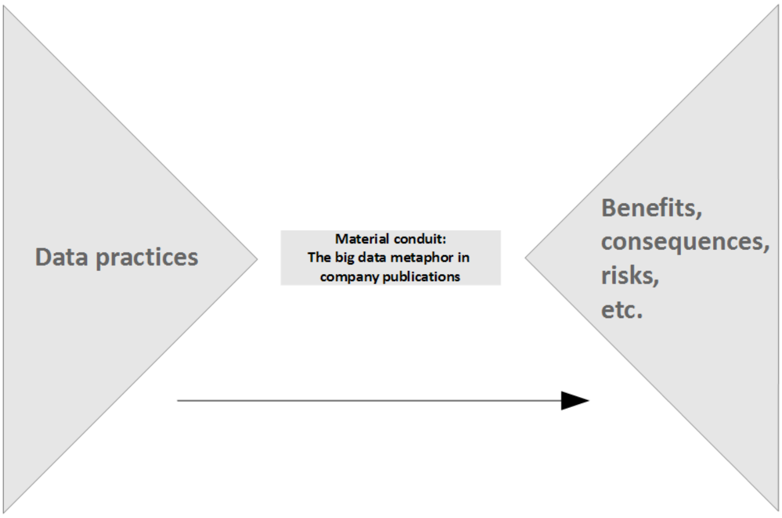


Figure 2: The big data metaphor as material conduit

Publishing information about big data in the shape of a company whitepaper is a deliberate choice. Its material implications being, as detailed in previous chapters, that colorful models, fruitful textual descriptions and an authoritative tone of voice is are explicitly used stylistically. Meanwhile, distribution through online channels ensures that the documents will reach targeted recipients without difficulty. The pyramids on both side of the model represent the source domain which is referred, and the target domain, upon which the references are applied.

Using metaphors in the sign domain implies the source domain to be equal to the target domain. George Lakoff and Mark Johnson's (1980) well known ARGUMENT IS WAR comparison is often used to exemplify this point. To make sense of what an argument is we view it in terms of war. We can *win* an argument like we win a war, we *attack* each other's statements like two sides attack each other, criticism can be *right on target* like a bomb or tactical maneuver, and so forth. Although having the similar function the big data metaphor works differently. Its meaning is inferred from the discourse on data practices, but it presents a more selective description and expands on this with its own features. The term becomes a linguistic container through which we communicate and construct concepts to others (Lakoff & Johnson, 1980). For this reason big data is placed in the middle of the model, functioning as the conduit through which ideas are communicated. It details the process of speaking about data practices in terms of big data in order to propose certain benefits. To demonstrate this process, and point out where the attention of the analysis will be, I have applied the Data/Information/Knowledge/Wisdom to the model as part of the source domain the in the adapted model below.

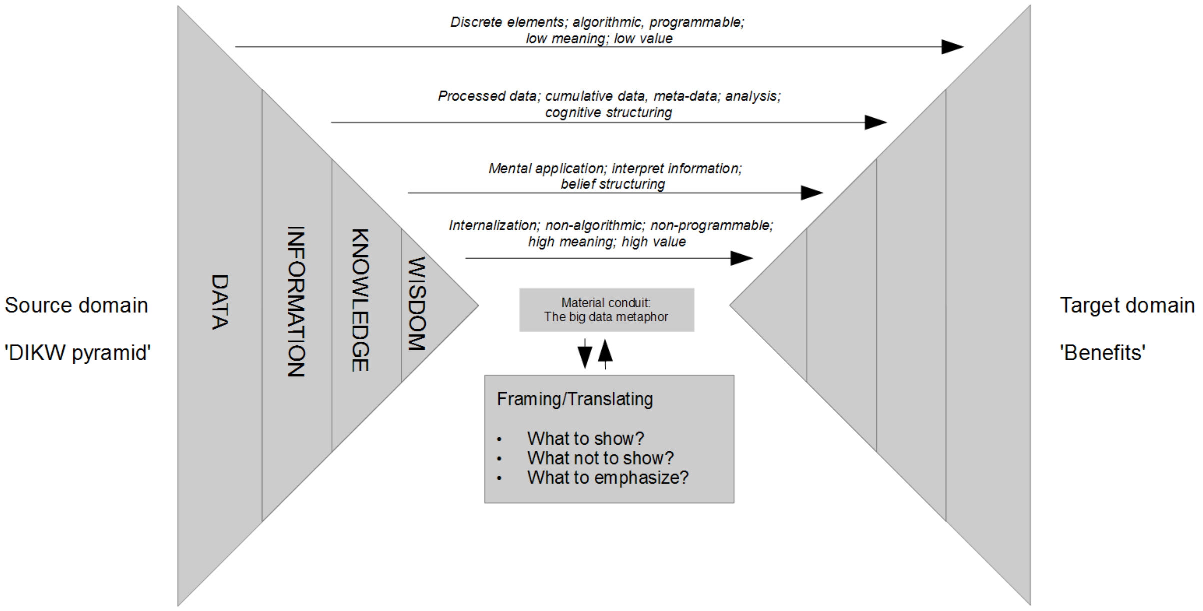

Discrete elements; algorithmic, programmable; low meaning; low value
Processed data; cumulative data, meta-data; analysis; cognitive structuring
Mental application; interpret information; belief structuring
Internalization; non-algorithmic; non-programmable; high meaning; high value
Source domain
'DIKW pyramid'
DATA
INFORMATION
KNOWLEDGE
WISDOM
Material conduit: The big data metaphor
Framing/Translating
What to show?
What not to show?
What to emphasize?
Target domain
'Benefits'

Figure 3: The big data ***imaginary*** *with DIKW pyramid as example*

The pyramid originated from both the philosophy of knowledge and information management literature. The hierarchy it implies originated from a T.S. Elliot poem but in recent literature Russel Ackoff's *From Data to Wisdom* (1989) is most often cited as the source of the pyramid (Rowley, 2007). In case of big data its relevance is in the description of how discrete data points can be turned into wisdom through a series of refinements in thinking and believing. This model should not be read as the pictogram to which all big data is referring. Any theory or model describing data practices that can be referenced to depending the proposed implications. If the sender wants to communicate the benefits of the data practices he could opt to use the DIKW pyramid as a source, pick some of its characteristics and through the big data conduit he communicates this to the receiver. For example in the quote:

> "Getting big data right means aligning information capital, human capital, and organizational capital to build a culture of disciplined decision making." (Deloitte, 2012)

The big data metaphor is used to indicate that information flows may benefit decision making. Here big data is the conduit that bridges the gap between what is in the language domain (data, information, capital) and the practical domain (decision making). What is constructed is a reality in which information and capital equal disciplined decision making *through* big data.

The model differs from what Lakoff & Johnson and Van den Boomen describe. According to them the function of a metaphor is: a conceptual understanding of two things sitting in-between the both domains, instead of the source and target domain will being one and the same. Or as Van den Boomen describes it:

> "Their [metaphors] mechanism can be defined quite simply: metaphors compress two (or more) references or associations by transferring and incorporating qualities from the one into the other. As a result, they do not just refer or represent something which cannot be expressed otherwise, they also differentiate, predicate and qualify by transferring specific qualities from one domain to another." (Van den Boomen, 2014)

The metaphor does not only construct the way we talk, it influences the way we view objects and phenomena too. Consequently the most fundamental structures of our culture are coherent with the metaphorical structure of the most fundamental concepts in the culture (Lakoff & Johnson, 2003). Companies use the big data metaphor to package ideas in a friendly fashion: receiving it the customer may unpack it with relative ease as the materiality of the publications (be it digital, or analog) dictates to the receiver what he will be presented with. The conduit operates seamlessly, ruptures only become visible if there is misunderstanding between parties. It ignores common situations of partially successful communication, of noise, incompatibility, and failure, based as it is based on unproblematic transference of meaning through a non-intervening transport medium (Van den Boomen, 2009).

BUSINESS LEXICON

Whitepapers follow a general structure while adhering to textual and visual conventions. Inadvertently what is communicated to the reader is: 'we want to convince you off the relevance of our expertise and persuade you to do business with us'. Attractive visual imagery and models aim to show explicitly the creative competency of the company and the imagery and models highlight graphically key points of what was expressed in words. Earlier chapters showed that the tone of voice of the big data quotes where mostly positive as the negative and neutral quotes often making up less than 50 percent combined. And analysis of the quote categories showed that most of the quotes are observations made about big data or quotes constructing a technological imaginary by stipulating the opportunities big data creates. Finally this last chapter aims to disentangle the quotes by investigating which words are used most frequently to construct them.

If you review the popular categories you get a general impression of the tone of voice. And if you look at the contents of these categories or you read the quotes, they will give you an in depth view of what is passed through the big data conduit. It also provides a first insight into which words often co-occur with big data in the same sentence. In extracting the quotes from the text they are stripped from the context leaving only the semantic value of the sentence. The following step involves disentangling these sentences: now only their building blocks are left. Every word has the same 'weight' and is counted only once. Therefore also syntactical make-up such as subject/object relations become meaningless. From the stripped down 'bag-of-words' the most prevalent words per category per year are selected.

Essentially all the words are sorted in a list by the sum of their occurrence and from that list the top 6 are selected.

New meaning arises as soon as they are paired with their category and their publication year. In the following table a top five is displayed of the most used words per year, where the bar is a stack of how much each category contributed to the total amount of a words' occurrence. In the year 2011 not much was published. As a consequence it had a very low word-count, therefore it is dropped out from this analysis.

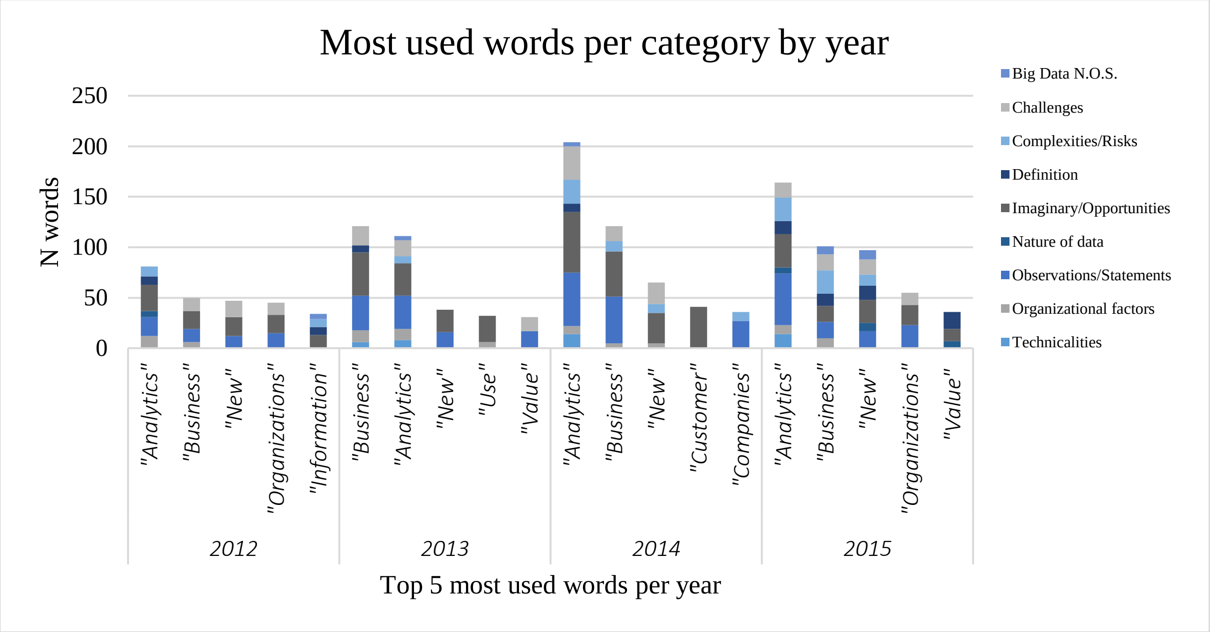


Figure 4: Most used words per category by year

In accordance with the meta-data 2014 displays a peak in word count after which 2015 shows a slight drop. Far outnumbering other are the words *'Analytics'* and '*Business'* going for first and second place in all the years**.** The message is: if you are doing business you should use big data analytics. But 'analytics' in and of itself does not demarcate a specific methodology or predisposition towards data. Neither is it something that exclusively belongs to the realm of big data, nor to any type of business. What is established is a strong relationship between 'analytics' and big data that might or might not have existed previously or outside the sample scope. 'Business' is possibly stronger connected to the overlapping category of corporate publications than it is to big data with respect to 'organizations' and 'companies'. In 2012 'information' makes its first and only appearance. The ties to the discourse on data practices is stronger because it is part of the DIKW pyramid as fundamental to operationalizing data. The words 'New' and 'Value' accentuate the big data imaginary. 'New' is the third most used word irrespective of year where 'value' is the 5th most used word in both 2012 and 2015. They underline the imaginary by stating that big data is a new phenomenon. There is strong sense in which this 'new' carries the strong ideological connotation of 'new = better' and it comes with a whole range of exciting meanings attached to it (Lister, Dovey, Giddings, Grant, & Kelly, 2009). By stating that big data is something which is new the corporate discourse, through positive connotation, entices an optimistic reading. Finally the word ‘customer’ implies that the benefits of big data are portrayed as outwards facing. This word only appears in the top list of 2014 and all within the same category of Imaginary and Opportunities. Similar as with ‘business’, ‘organizations’ and ‘companies’ it might be more inherent to the corporate nature of the sample then it is to big data in itself. Yet the result of big data is that a company may create value for its customers.

The words are the embodiment of meaning, the material signal flows from the sender to the receiver. By disentangling the signal into little bits allows for insight into how this meaning is constructed. If the words were to be put back into a coherent statement it would read something like ‘Big data *analytics* allow *businesses/companies/organizations new* ways of creating *value* for *customers’.* And even though there are plenty of variations possible with the same set of words their coherent meaning will probably not differ very much if one practices them to create a different sentence.

A SELF REFERENCING CONCEPT

When speaking of big data one thing is certain: the data-set is supposedly large, or at least larger than what was previously assumed to be a normal sized dataset. Any other characteristics implied by the term have to be explained as they cannot be inferred from the term itself. The data itself is most likely unstructured and generated at a quick pace by sources outside of a company’s direct influence. Integration of these data-streams into the analytics pipeline will most likely mean overhauling both digital infrastructure and decision making trees while also getting management to carry the concept from idea to execution. Yet big data as a metaphor fails to explain all of these traits. So what is its redeeming factor? How did the term become so dominant in the conversation on data practices from 2011 on? Unlike the email icon it does not transfer material meaning from one concept to the other. Unlike the ARGUMENT = WAR big data does not explain itself in terms of either its ‘bigness’ or ‘dataness’. Rather than SOURCE DOMAIN = TARGET DOMAIN, big data sits in between the two domains. But it does not only act as a transference conduit like how van den Boomen & Hayes (2014) picture a material metaphor to operate. Rather it simultaneously allows for communication to happen and is also what is being communicated. It is the commensuration of all the data practices traits that were hitherto confined to specific (most technical or scientific) discourses, into one linguistic container. If these traits were to change, for example when privacy concerns make some practices undesirable, big data will change with it. Big data does not explain the reality in terms of referencing to other concepts, it is a carefully constructed label that constructs reality in through the discourse.

BIG DATA CONSTRUCTS REALITY

Companies use the big data metaphor and materialize it by visualizing with models that speak to the imagination. The image is very carefully constructed by using authoritative and positive language and colorful models wrapped up in beautifully designed publications. These practices have created a self-referential system which do not require meaningful statements to be constructed using outside references. Big data is not used to explain the world but it allows companies to communicate traits of new data practices. The traits can be both negative and positive and the topics that are discussed may range from technical to managerial as this does not influence the function of the term. However it does not function as a traditional metaphor in the sense of what Lakoff and Johnson (1980) portray it to be. There is no equal source domain that big data can be the target domain of. Rather the term sits in between both the source and target domain, mediating traffic as a metaphorical conduit. Consequently it may fails an explanatory device as transference of concepts is not guaranteed. Within the big data discourse terms like 'Cloud Computing' would better fit this the description as metaphor as they do allow for comprehensive transference of meaning from one concept to the other. What big data allows companies to do is group related practices under a similar banner that intrigues a larger audience. It seduces people who are not directly familiar with the ins-and-outs of data related products and services but at the same time works as a veil that covers up more than it reveals. Companies can use it to entice new customers to whom the magnitude of employing and implementing data practices is only revealed once their interest is aroused.

On the other hand, big data allows for two people of different backgrounds to have a meaningful conversations on a topic that would otherwise not have been impossible or only in a limited way. A marketer might still be flummoxed by the answer if he asks a data-scientist on how big data may help him but it does allow them to have the conversation in the first place. Similarly, related issues such as privacy, data-policy, data-driven decision making and so forth can now be addressed in popular media by using big data as a container term. To name an example, a news broadcast would not speak of the 'blurring of personal and company boundaries as data becomes a common resource in the marketing pipeline' yet would address the 'privacy concerns of big data driven marketing'. But as big data only enlightens the audience on a high abstraction level, both parties would still have to engage more in-depth conversations to get to essence of the implications big data for them. It therefore remains to be seen if big data will remain on the radar of companies or whether the term will eventually dissipate into separate concepts. In five years' time, if big data has been dismantled, the discussion might have shifted to talking about a 'data-delta' to describe a distributed data management system or 'raid-data' when privacy is concerned.

RECOMMENDATIONS FOR FURTHER RESEARCH

To further understand the function of big data the sample of publications should be made larger. In that case it would also include academic and popular publications, cover a wider timespan, or include research on metaphors that are used in close relation to big data such as the 'internet of things', 'smart cities' and 'machine learning'. The results as they are now only speak for the corporate world, and it could be valuable to compare these to other fields. Metrics such as a co-occurrence matrix would tell more about the relative position of words compared to big data or of big data quotes relative to the entire publication. This would tell more about the content of the publication as a whole, and how much of that content is actually dedicated to big data.

Further research could lead to more insights into understanding how these concepts are connected to the corporate discourse and how they help frame company practices.

What will be interesting to see: in what way the term big data will develop over the years. According to the Gartner's Technology Hype Cycle big data was already over the peak of inflated expectations in 2014 and seems to have fallen off completely in 2015 (Gartner, 2015). It is possible that those practices first labeled as big data will move to the background and crystalize into clearly demarcated professions (Data Scientist, Forensic Data Analyst). But companies will not linger on this demise. Like with many technological innovations before it, the rhetoric of progress will continue to be a dominant factor in company publications. Big data's popularity may be short lived but after the dust has settled down, companies will still be developing new techniques to deal with the increasing demands and opportunities digital technologies bring along. After all, if anything, big data is presented as technological progress, and you can't stop progress.

BIBLIOGRAPHY

Accenture. (2014). *Big Success With Big Data.* Retrieved from Accenture Plc: https://www.accenture.com/us-en/insight-big-data-research.aspx.

Accenture. (2015). *Retail Hyperpersonalization: Creepy v.s. Cool.* Retrieved from Accenture.com: https://www.accenture.com/t20150914T131203__w__/us-en/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Dualpub_8/Accenture-Technology-Labs-Hyperpersonalization.pdf#zoom=50

Amazon. (2014). *Big Data Analytics Options on AWS.* Retrieved from Amazon Web Services: https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf

Baym, N. K. (2013, October). *Data not seen: The uses and shortcomings of social media metrics.* Retrieved from First Monday: http://firstmonday.org/ojs/index.php/fm/article/view/4873/3752

BNP Paribas. (2015). *Big Data and the New Economy.* Retrieved from BNP Paribas: http://www.bnpparibas.com/en/node/19553?page=34.

Boston Consultancy Group. (2015, 10 15). *How To Avoid the Big Bad Data Trap.* Retrieved from BCG Perspectives: http://www.bcgperspective.com

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data . *Information, Communication & Society*, 662-679.

Bryant, A., & Raja, U. (2014, 02 03). *In the realm of Big Data...* . Retrieved from First Monday: http://firstmonday.org/article/view/4991/3822

Cauwenberghe, van , E., Maes, L., Spitteals, H., Lenthe, van, F. J., Brug, J., Oppert, J.-M., & De Bourdeaudhuij, I. (2010). Effectiveness of school-based interventions in Europe to promote healthy. *British Journal of Nutrition* , 781-797.

Cavnar, W. B. & Trenkle, J. M. (1994). N-Gram based text categorization. *Ann Arbor MI, 48113(2)*, 161-175.

Schäfer, M.T. (2008) Chapter 1: Promoting Utopia/Selling Technology. In M. T. Schäfer, *Bastard Culture!: How User Participation Transforms Cultural Production* (pp. 25-40). Amsterdam: Amsterdam University Press .

Cisco Systems. (2012). *For Big Data Analytics There's No Such Thing as Too Big* . Retrieved from Cisco Systems: http://www.cisco.com/c/dam/en/us/solutions/data-center-virtualization/big_data_wp.pdf.

Credit Suisse. (2013). *Big Data: Reinventing Shopping* . Retrieved from Credit Suisse: https://www.credit-suisse.com/nl/en/news-and-expertise/economy/articles/news-and-expertise/2013/06/en/big-data-reinventing-shopping.html.

Cukier, K., & Mayer-Schönberger, V. (2013). *Big data: A Revolution that Will Transform how We Live, Work and Think.* Boston/New York: Houghton Mifflin Harcourt

Cukier, K., & Mayer-Schönberger, V. (2013). The Rise of Big Data. *Foreign Affairs*, pp. 27-40.

Deloitte. (2012). *The insight economy: Big data matters, except when it doesn't.* Retrieved from Deloitte.com: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Deloitte-Analytics/dttl-analytics-us-ba-insight-economy-10012012.pdf

Duhigg, C. (2012). *How Companies Learn Your Secrets.* Retrieved from nytimes.com: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Reichert, R. (2009) Einleitung. In R. Ramón, *Das Wissen der Börse* (pp. 11-19). Bielefeld: Transcript.

Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *McQuail's reader in mass communication theory*, 51-58.

EY. (2013). *Big Data and Enterprise Mobility* . Retrieved from EY.com: http://www.ey.com/Publication/vwLUAssets/Big_Data_and_Enterprise_Mobility/$FILE/Big_Data_Enterprise_Mobility_LR.pdf.

Flichy, P. (1999). The construction of new digital media. *New Media & Society.* 33-39

Flichy, P. (2002). The Imaginary of the Internet. *Invited lecture at the 8th International Summer School of the European PhD on social representation and communication: The Net and the Internet.*

Flichy, P. (2007). *Understanding Technological Innovation.* Cheltenham: Edward Elgar Publishing Limited.

Gartner. (2012, 06 21). *The Importance of 'Big Data': A Definition.* Retrieved from Gartner.com: https://www.gartner.com/doc/2057415?ref=clientFriendlyURL

Gartner. (2015). *Hype Cycle for Emerging Technologies, 2015.* Gartner Inc.

Gartner/META. (2001, 02 06). *3D Data Management: Controlling Volume, Velocity and Variety.* Retrieved from Gartner Blogs: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Glaser, B. G. & Strauss, A. L. (1995). *The Discovery of Grounded Theory: Strategies for Quantitative Research.* Aldine Transactions: New Brunswick.

GSK. (2015). *Precision medicine: science fiction becomes science fact.* Retrieved from GSK: http://www.gsk.com/en-gb/our-stories/everyday-health/precision-medicine-science-fiction-becomes-science-fact.

Halveli, G. & Moed, H. (2012, 09 30). *Section 1: The Evolution of Big Data as Research and Scientific Topic.* Retrieved from researchtrends.com : http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf

Hayles, K. N. (2002). *Writing Machines.* Cambridge: MIT University press.

IBM. (2012). *Analytics: The real-world use of big data.* Retrieved from IBM Business Consulting: http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html

Kitchin, R. (2014). *The Data Revolution,* London: SAGE Publications.

Krin, W. (2015), *If Your're Not Paranoid, You're Crazy.* Retrieved from The Atlantic: http://www.theatlantic.com/magazine/archive/2015/11/if-youre-not-paranoid-youre-crazy/407833/

Kroes, N. (2013, 11 07). *Big data for Europe.* Retrieved from European commision press release database : http://europa.eu/rapid/press-release_SPEECH-13-893_en.htm

Cukier, K., & Mayer-Schönberger, V. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think.* New York: Houghton Mifflin Harcourt Publishing Company.

Lakoff , G., & Johnson, M. (2003). Metaphores We Live By . *Language, Thought, and Culture* , 1-11.

Lakoff, G., & Johnson, M. (1980). *Metaphores we live by.* Chicago: University of Chicago press.

Laney, D. (2001, 02 06). *Application Delivery Strategies.* Retrieved from blogs.gartner.com: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Law, D., & Pepperell, N. (2015, April). *The Internet Imaginary: Between Technology and Technique.* Retrieved from Journal of Media & Culture : http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/957

Lister, M., Dovey, J., Giddings, S., Grant, I., & Kelly, K. (2009). *New Media: a critical introduction .* London and New York: Routledge.

Long, P., & Wall, T. (2012). *Media Studies* . London: Pearson .

Malbreil, X. (2007). *About the internet imaginary and its evolution.* Retrieved from <REVISTA TEXTO DIGITAL>: https://periodicos.ufsc.br/index.php/textodigital/article/viewFile/1397/1093

Markham, A. N. (2013). Undermining 'data': A critical examination of a core term in scientific inquiry. *First Monday*, pp. 0-13.

Mattelart, A. (2002). An archeology of the global area: constructing a belief. *Media, Culture & Society*, 591-612.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform thy Way We Live, Work and Think .* New York : Eamon Dolan .

PWC. (2014). *Capitalizing on the promise of Big Data.* PWC.com.

Rogers, S. (2012). *Data are or data is?* Retrieved from The Gaurdian: http://www.theguardian.com/news/datablog/2010/jul/16/data-plural-singular

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW pyramid. *Journal of Information Science OnlineFirst*, pp. 1-18.

Schopfel, J. (2010). *Towards a Prague Definition of Grey Literature.* Retrieved from Opengrey.eu: http://www.opengrey.eu/item/display/10068/700015

Turner, F. (2006). *From Counterculture to Cyberculture.* Chicago: University of Chicago Press.

Van den Boomen, M. (2014). *Transcoding the Digital: How Metaphors Matter in New Media.* Amsterdam: Amsterdam University Press.

Van den Boomen, M. (2009). Interfacing by material metaphores. In M. van den Boomen, S. Lammes, A.-S. Lehmann, J. Raessens, & M. T. Schäfer, *Tracing New Media in Everyday Life and Technology* (pp. 253-265). Utrecht: Amsterdam University Press .

APPENDICES

1. METADATA NR OF QUOTES PER INDUSTRY SUBDEVIDED BY COMPANY

Industry - Company	Nr. of quotes per company
Automotive	**15**
Daimler	2
Ford	7
Mitsubishi	3
Toyota	3
Consultancy	**1350**
Accenture	136
Analysis Group	11
AT Kearney	140
BAIN	53
BCG	67
Booz Allen Hamilton	47
Brattle Group	17
Capgemini	179
Cornerstone	3
Delloite	88
EY	305
Forrester	5
FoxConn	1
Gartner	72
Kurt Salmon	31
MCKesson	14
McKinsey	47
NEC	32
Oliver Wyman	25
Point B	32
PWC	38
PWC pathways	7
Finance	**75**
Aviva	10
BNP Paribas	2
Commonwealth Bank of Australia	2
Credit Suisse	18
Munich RE	21
Rabobank	5
Societe Generale	3
UBS	14
Food	**3**
Unilever	3

Insurance	**13**
AIG	13
IT	**742**
Dell	80
HP	29
IBM	300
ING	4
Nucleus	22
Oracle	269
Oracle MIT	28
Tableau	10
Logistics	**26**
Airbus	9
Lockheed Martin	15
Maersk	2
Manufacturing	**4**
FocConn	4
Pharmaceutical	**11**
GSK	2
MetLife	1
Novartis	2
Roche	6
Products	**22**
Amazon	22
Resources	**59**
Baosteel	6
BP	13
ENI	1
GE	28
Iberdrola	1
Repsol	2
Rio Tinto	7
Sinopec	1
Technology	**802**
BASF	4
Cisco Systems	195
Ericsson	107
Fujitsu	168
Hitachi	62
Ingram Micro	4
Intel	201
Samsung	33
Siemens	25
United Technologies	3
Telecom	**127**

AT&T	1
China Telecom Corporation	4
China Telephone Corporation	2
Comcast	6
Deutsche Telecom	63
Orange	15
SK Telecom	7
Telefonica	3
Vodafone	26
Total	**3249**

2. EXAMPLE OF ATLAS.TI CODE REPORT

Atlas.ti allows for a collection of elements from a corpus to be coded and stored. All the quotes that are extracted from are stored this way. Finally a list of collected quotes per category can be collected from the program and copied to a text file for further processing. The list below is an example of what the text output looks like. It shows metadata (user, filepath, date of creation, etc.) as well as the amount of quotes listed in de category, the position in the text from where the quote was extracted, which code was used, whether the user attached memo's to the quote and the quote itself.

```
Report: 159 quotation(s) for 1 code
______________________________________________________________________

HU:   Big Data 2015
File:   [E:\Thesis\2015\Big Data 2015.hpr7]
Edited by:          Super
Date/Time:          2015-11-23 14:34:29
______________________________________________________________________

Mode: quotation list names and references

Quotation-Filter: All

Observations/Statements

P 7: 2015 - Airbus - NEW_FORUM85_EN.PDF - 7:1 [Fast forward to today: informa..]
(2:1166-2:1372)   (Super)
Codes:              [Observations/Statements]
No memos

Fast forward to today: information technology and big data are
driving demand and disruptors are front and centre. Rather
than governments, it's now entrepreneurs and innovators
exploring new horizons.


P 7: 2015 - Airbus - NEW_FORUM85_EN.PDF - 7:2 [Whether we consider this rapid..]
(2:1572-2:1856)   (Super)
Codes:              [Observations/Statements]
No memos

Whether we consider this rapid pace of innovation promising
or overwhelming (or both), it is clear that the rate of change
is set to increase exponentially as new technologies and the
exploitation of big data disrupts how we sell, deliver, buy and
access products and services.
```

3. PYTHON CODE FOR DATA PARSING

The (unstructured) text-data from atlas.ti needs to be parsed to a structured format before it can be analyzed. Using the python script shown below this procedure was automatized. It works as follows: first all relevant files are identified through by recursively crawling the folders in which the text files are stored. Files are identified by their .txt extension. The search function will stop as soon as there are no longer .txt files in the folder above the current folder and error handling is done by printing invalid file path to the console. All files are stored in the function pathList using the 'glob' operator. The first file in pathList is opened and its contents stored into memory as a string. First the year and code category are extracted as they remain the same for every quote. The script then cycles through the string using identifiers to extract the company and quote. Results are stored in a block variables and appended to a bigdata.csv file. Once this procedure is finished the file is closed and removed from memory. Finally a the next file in pathList is opened, the script ran again, until all files have been parsed into bigdata.csv.

Script:

```
import os
import sys
import glob

FILETYPE = ".txt"
SEARCHPATHNAME = "C:/Users/robert.smit/Google Drive/Thesis/"
OUTPUTPATHNAME = "C:/Users/robert.smit/Google Drive/Thesis/"
OUTPUTFILENAME = "bigcsv.csv"

# Strip off trailing slashes and return last part of path
directoryName = os.path.basename(os.path.normpath(SEARCHPATHNAME))

# Try changing to the directory above the search path
try:
   os.chdir(SEARCHPATHNAME)
   os.chdir("..")
except (WindowsError, OSError) as e:
   print("[Invalid pathname for search]")
   print(SEARCHPATHNAME)
   print(e)
   sys.exit(1)

# Get a list of files of the specified file type
pathList = glob.glob(directoryName + "/**/*" + FILETYPE, recursive=True)

# Try to open output file and specify the delimiter ";" (overwriting previous versions)
try:
   outputFile = open(OUTPUTPATHNAME+OUTPUTFILENAME, "w")
   outputFile.write("sep=;\n")
   outputFile.close()
except (WindowsError, OSError, IOError) as e:
   print("[Invalid pathname for output file]")
   print(OUTPUTPATHNAME+OUTPUTFILENAME)
   print(e)
   exit(1)

# For every file found in the search path
for path in pathList:
   year = 0
   code = ""
   rawString = ""
   data = []

   # Read file to string, ignoring encoding errors
```

```
    dataFile = open(path, errors='ignore')
    while True:
        nextLine = dataFile.readline()
        if nextLine == "":
            break
        else:
            rawString += nextLine
    dataFile.close()

    # Extract the year
    yearEnd = rawString.find(".hpr7")
    yearStart = yearEnd - 4
    year = rawString[yearStart:yearEnd]

    # Extract the code type
    codeStart = rawString.find("Codes:")
    codeStart += rawString[codeStart:].find("[") + 1
    codeEnd = codeStart + (rawString[codeStart:].find("]"))
    code = rawString[codeStart:codeEnd]

    # Extract the company/quote combinations
    currentString = rawString
    startIndex = rawString.find("\nP")

    # While there are still quotes left in the file
    while startIndex != -1:
        currentString = currentString[startIndex+1:]

        # Extract company name
        companyStart = currentString.find("-") + 1
        companyEnd = currentString.replace("-", "+", 1).find("-")
        company = currentString[companyStart:companyEnd]
        company = company.strip()

        # Extract quote
        quoteStart = currentString.find("No memos") + 8
        while not currentString[quoteStart].isalnum():
            quoteStart += 1
        quoteEnd = quoteStart + 1
        while (currentString[quoteEnd] + currentString[quoteEnd+1]) != "\n\n":
            quoteEnd += 1
            # If at end of file stop to prevent IndexError
            try:
                tmp = currentString[quoteEnd+1]
            except IndexError as e:
                break
        quote = currentString[quoteStart:quoteEnd]
        quote = quote.replace("\n", " ")
        quote = quote.strip()

        # Add to list
        block = [year,company,code,quote]
        data.append(block)

        currentString = currentString[quoteEnd:]
        startIndex = currentString.find("\nP")

    # Write to output csv file, ignoring encoding errors
    outputFile = open(OUTPUTPATHNAME+OUTPUTFILENAME, "a", errors='ignore')
    for block in data:
        for item in block:
            outputFile.write(item)
            if item != block[3]:
                outputFile.write(";")
        outputFile.write("\n")
    outputFile.close()
```

4. WORD FREQUENCY ANALYSIS SCRIPT IN R

Word frequency analysis was done using different individual .csv for every year in which all the all the quotes per category were stored in separate columns. The function wordFreq that uses the qdap package in R to parse the quotes into a large ‘bag of word’ stripping them of their syntactical make-up and. Redundant words, as well as whitespaces, punctuation and stop words were than stripped from the collection. Finally the words were parsed into a data matrix that stores every word as a string value, counted and sorted by frequency of appearance. The function uses a filename$collumnName argument to identify which column to scrape the data from. These arguments were adjusted for every column until all the categories where analyzed. The console output was manually transferred into an excel file for further analysis.

Script:

```
install.packages('tm')
install.packages('qdap')

#load library
library(tm)

#set working directory

setwd("C://Users//Robert//Google Drive//Thesis//Data//2015//Output")
data2015 <- read.csv("Data 2015.csv",row.names = NULL, header = TRUE, stringsAsFactors = FALSE)

wordFreq <- function (data2015) {
  #make 'bag of words'
  data1 <- paste(data2015, collapse = " ")

  #make corpus data
  data_Source <- VectorSource(data1)
  corpus <- Corpus(data_Source)

  #remove big & data
  words1 <- c("big", "data", "can", "will", "also")

  #data cleaning
  corpus <- tm_map(corpus, content_transformer(tolower))
  corpus <- tm_map(corpus, removePunctuation)
  corpus <- tm_map(corpus, stripWhitespace)
  corpus <- tm_map(corpus, removeWords, stopwords("english"))
  corpus <- tm_map(corpus, removeWords, words1)

  #turn corpus into data frame
  dtm <- DocumentTermMatrix(corpus)
  dtm2 <- as.matrix(dtm)

  #word counting & reporting
  frequency <- colSums(dtm2)
  frequency <- sort(frequency, decreasing = TRUE)
  frequency
}

frequency <- wordFreq(data2015$Observations.Statements)

head(frequency)
```

5. EMOTION & POLARITY ANALYSIS SCRIPT IN R

This script uses the bigdata.csv that resulted from running the Python script described in appendix 2 (renamed to Big data quotes – copy.csv). All text in the ‘Quote’ column is parsed into a dataframe to prepare it for analysis. The text is cleaned by removing punctuation, and numbers. After which the sentiment and polarity are analyzed using a naïve Bayes algorithms trained on specific lexicons (as mentioned in the thesis). Finally the results are parsed to a dataframe that gets added to the bigdata.csv file as a new excel sheet. This results in an overview of the emotion and sentiment per quote.

Script:

```
library(sentiment)
library(xlsx)
library(plyr)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)

setwd("C://Users//Robert//Google Drive//Thesis//Data")
data <- read.csv2("Big data quotes - Copy.csv", header = TRUE, stringsAsFactors = FALSE, blank.lines.skip = TRUE)

quotes <- data.frame(data['Quote'])


cleanCharacters <- function (quotes) {

  # remove punctuation
  quotes <- gsub("[[:punct:]]", "", quotes)

  # remove numbers
  quotes <- gsub("[[:digit:]]", "", quotes)
}

#clean text
quotes <- sapply(quotes, cleanCharacters)

#remove capitalization
quotes <- tolower(quotes)

# classify emotion
class_emo = classify_emotion(quotes, algorithm="bayes", prior=1.0)
# get emotion best fit
emotion = class_emo[,7]
# substitute NA's by "unknown"
emotion[is.na(emotion)] = "unknown"

# classify polarity
class_pol = classify_polarity(quotes, algorithm="bayes")
# get polarity best fit
polarity = class_pol[,4]

#make dataframe from sentiment
sent_df <- data.frame(text=quotes, emotion=emotion, polarity=polarity, stringsAsFactors=FALSE)

#combine df with data
sentiment_data <- cbind(data[c('Year', 'Company', 'Category')], sent_df)

Sentiment_data_statistics <- data.frame(quotes, class_emo, emotion, class_pol, polarity)

#write new file

Sentiment_Data <- createWorkbook(type = "xlsx")
quotes_Sheet <- createSheet(wb = Sentiment_Data, sheetName = "Quotes & Sentiment")
statistics_Sheet <- createSheet(wb = Sentiment_Data, sheetName = "Statistics")
addDataFrame(sentiment_data, sheet = quotes_Sheet)
addDataFrame(Sentiment_data_statistics, sheet = statistics_Sheet)
saveWorkbook(Sentiment_Data, "Quotes & Sentiment.xlsx")
```

6. OVERVIEW OF RELATIVE QUOTES PER CATEGORY PER INDUSTRY

	Industries >				Percentage of quotes per category per industry								
Categories	Consul.	Tech.	IT	Telco.	Fin.	Resour.	Logis.	Prod.	Auto.	Insur.	Pharma.	Manu.	Food	Total
Observations/ Statements	49.8	20.8	20.4	3.9	1.5	1.3	1.0	0.6	0.3	0.0	0.3	0.0	0.0	100.0
Imaginary/ Opportunities	39.4	27.0	20.0	4.0	3.9	2.5	0.6	0.4	0.3	0.6	0.7	0.3	0.3	100.0
Challenges	37.7	27.0	27.9	2.6	1.3	1.8	0.4	0.7	0.0	0.7	0.0	0.0	0.0	100.0
Complexities/ Risks	47.8	22.8	21.8	3.2	1.3	1.6	0.3	0.6	0.3	0.0	0.3	0.0	0.0	100.0
Titles	45.0	25.4	20.0	3.8	2.1	0.4	1.3	0.8	1.3	0.0	0.0	0.0	0.0	100.0
Definition	28.9	26.8	23.7	4.6	6.7	2.1	1.0	0.5	1.5	3.1	0.5	0.0	0.5	100.0
Organizational factors	41.7	27.1	24.3	4.9	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0	0.0	100.0
Questions	59.2	13.8	23.8	3.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
Technicalities	10.8	41.4	42.3	0.0	0.0	4.5	0.0	0.0	0.0	0.0	0.9	0.0	0.0	100.0
Big Data N.O.S.	11.5	19.7	4.9	23.0	11.5	8.2	8.2	4.9	4.9	0.0	0.0	3.3	0.0	100.0
Nature of data	15.1	30.2	45.3	1.9	1.9	5.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
Total	**41.6**	**24.7**	**22.8**	**3.9**	**2.3**	**1.8**	**0.8**	**0.7**	**0.5**	**0.4**	**0.3**	**0.1**	**0.1**	**100.0**

7. OVERVIEW OF ALL POLARITY SCORES PER INDUSTRY

	Industries >				Polarity amount of quotes per industry								
Polarity:	Consul.	Tech.	IT	Telco.	Fin.	Resour.	Logis	Prod.	Auto.	Insur.	Pharma	Manu.	Food	Total
negative	206	104	92	19	13	6	2	7	1	4	4	0	0	458
neutral	215	95	95	22	12	15	4	2	0	2	1	0	1	464
positive	929	603	555	86	50	38	20	13	14	7	6	4	2	2327
Total	**1350**	**802**	**742**	**127**	**75**	**59**	**26**	**22**	**15**	**13**	**11**	**4**	**3**	**3249**

8. LIST OF QUOTES REFERENCING THE UNSTRUCTURED NATURE OF BIG DATA IN CHALLENGES AND OPPORTUNITY/IMAGINARY CATEGORIES

8A CHALLENGES CATEGORY

Year	Company	Quote
2012	Intel	“They are being asked to manage a wide range of data sources and most have already gone beyond structured data to begin processing unstructured and semi structured data”
2012	NEC	“However in the case of exploiting big data it is necessary to mine data for knowledge and added value a purpose for which the data were not originally envisioned to be used and to set up a flexible processing flow that can handle unstructured and irregular data and perform real time or batch processing according to the form of utilization”
2012	Intel	“Emerging technologies such as the Hadoop framework and mapreduce offer new and exciting ways to process and transform big data—defined as complex unstructured or large amounts of data—into meaningful insights but also require it to deploy infrastructure differently to support the distributed processing requirements and real time demands of big data analytics”
2013	Fujitsu	“Many proposed approaches to the management of big data could potentially result in organizations creating new ‘silos’ separate self contained islands of information for transactional systems business intelligence data warehouse systems and unstructured big data solutions”
2014	EY	“Those leading the change are now including big data from both within and outside the enterprise including structured and unstructured data machine data and online and mobile data to supplement their organizational data and provide the basis for historical and forward looking statistical and predictive views”
2014	EY	“Real time fraud monitoring is a classic big data challenge demanding the integration of large amounts of diverse structured and unstructured high velocity data that needs to be analyzed in near real time to realize the benefits”
2014	Deutsche Telecom	“key to unlocking the power of big data is leveraging both structured and unstructured data and drawing the right conclusions”
2014	EY	“For descriptive analytics to assess past activities the value chain needs to be able to consistently collect blend and distribute structured and unstructured data from both internal and external sources manage big data”

2014	Capgemini	“Deutsche bank has been working on a big data implementation since the beginning of in an attempt to analyze all of its unstructured data”
2014	Dell	“Current market hype would have decision makers believe that this mountain of big data primarily encompasses unstructured and semistructured data — such as social media sensor data and machine to machine data — which cannot be handled by traditional tools and skill sets”
2014	Capgemini	“Big data requires new technologies and processes to store organize and retrieve large volumes of structured and unstructured data”
2014	EY	“Big data includes information garnered from social media data from internet enabled devices including smartphones and tablets machine data video and voice recordings and the continued preservation and logging of structured and unstructured data”
2014	Kurt Salmon	“Any big data program must account for both existing and missing resources technological and human alike and be flexible enough to adjust and scale to new sources of unstructured data and new regulations”
2015	Brattle Group	“The mining of big data is a critical component of an effective anti tbml program which involves extracting and analyzing data that is both structured and unstructured and that resides both inhouse and externally”

8B OPPORTUNITIES/IMAGINARY CATEGORY

Year	Company	Quote
2012	Cisco Systems	“The collection of this data both structured and unstructured opens the door to and mandates development of new approaches and applications grounded in emerging analytics techniques and emerging processing technologies within the big data paradigm”
2012	HP	“The big data opportunity is about driving actionable insight and timely enterprise decisions from the combination of different types of data – unstructured semistructured and structured data”
2012	IBM	“Using a big data solution on a supercomputer that is one of the world’s largest to date and a modeling solution designed to harvest insights from an expanded set of factors including both structured and unstructured data the company can now help its customers optimize turbine placement and as a result turbine performance”
2012	Nucleus	“Big data also improves decision making by aggregating and analyzing large volumes of unstructured data”
2013	AT Kearney	“Existing bi vendors are likely to consider a play for big data solutions that combine the best technologies for structured and unstructured data”
2013	Cisco Systems	“Big data in its raw form is typically unstructured streaming and often imprecise”
2013	Dell	“While big data analytics holds the promise of unprecedented insights the amount of useful data from social media blogs and other unstructured object repositories is often quite small relative to the amount of data stored in them”
2013	EY	“With respondents expecting unstructured data to grow at faster rates – than unstructured information will grow out of control and will drive big data explosion”
2013	IBM	“To use big data in this way organizations need solutions that can quickly search and analyze a wide variety of internal and external data including the unstructured data of news feeds”
2013	IBM	“In other words big data can include structured unstructured and semistructured data”
2013	IBM	“In addition to sensor data this big data includes large volumes of semistructured and unstructured data—ranging from high frequency drilling and production measurements to daily written operations logs—that quickly add terabytes of new data”
2013	Intel	“Big data includes three types of data—structured semistructured and unstructured—and intel’s it manager survey of it professionals found that four of the top five data sources for it managers today are semistructured or unstructured many businesses are simply unable to analyze these emerging forms of data which include everything from emails photos and social media to videos voice and sensor data”
2014	EY	“As companies become more sophisticated around big data combining both structured and unstructured data sources into their fda programs we see countries such as brazil leading the charge with the adoption of text analytics”
2014	EY	“By some estimates more than of the data within organizations is unstructured and unfit for traditional processing using big data will enable the processing of this unstructured data and increased system intelligence which can be used to improve performance in sales increase understanding of customer needs reinforce the internal risk management function support marketing initiatives and enhance fraud monitoring”
2014	EY	“Big data has brought a new paradigm to data architecture in the past data systems were built with a predetermined set of data requirements in the big data world data storage platforms are not restricted to a predefined rigid data model and data systems are capable of handling all kinds of structured and unstructured data”
2015	Oracle	Big data—information gleaned from nontraditional sources such as blogs social media email sensors photographs video footage etc and therefore typically unstructured and voluminous—holds the promise of giving enterprises deeper insight into their customers partners and business”

9. QUOTES REFERENCING THE 3V’S MODEL IN THE DEFINITION CATEGORY

Year	Company	Quote
2015	AIG	“Big data is commonly defined in terms of the ‘three vs’ developed by the analyst doug laney in 2001”
2013	AT Kearney	“Big data defined big data typically refers to the three vs—volume variety and velocity—of structured and unstructured data pouring through networks into processors and storage devices and the conversion of that data into usable business information see figure”
2013	AT Kearney	“In essence big data is the assemblage of infrastructure data sources software and skills that support the three vs allowing companies to undertake more relevant and timely analyses than is possible with traditional business intelligence methods”
2015	AT Kearney	“Traditionally big data has centered on data sets with significant volume variety and velocity of data”

2014	BP	“When data becomes too difficult to manage by conventional means either because of the volume the speed or the variety it becomes what is known as big data not necessarily ‘big numbers’ – although many certainly are – but numbers that as well as being present in huge volumes also come at high velocity”
2013	Capgemini	“The three common characteristics of big data—volume variety and velocity—are both relevant and challenging for fs companies”
2012	Cisco Systems	“In todays it marketplace big data is often used as shorthand for a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high velocity capture discovery and or analysis”
2012	Cisco Systems	“Big data represents a revolutionary step forward from traditional data analysis characterized by its three main elements variety volume and velocity”
2013	Dell	“Big data is often defined by three fundamental characteristics volume variety and velocity”
2013	Deutsche Telecom	“In large parts of the it industry big data is characterized by the vs“
2015	Ericsson	“The volume velocity and variety – or the three vs – of big data were simply overwhelming”
2013	EY	“Very large data sets volume that are being produced at a tremendous speed by the growing digitization of the society velocity and consists of data from all possible sources from structured to unstructured variety”
2014	EY	“For the purpose of this survey we have adopted gartner research’s definition “big data is high volume high velocity and high variety information assets that demand cost effective innovative forms of information processing for enhanced insight and decision making”
2014	EY	“According to gartner research “big data” is high volume high velocity and high variety information assets that demand cost effective innovative forms of information processing for enhanced insight and decision making”
2014	EY	“Big data refers to the huge and increasing volume of the data now available as well as the variety of it and the velocity at which it can be processed”
2013	Fujitsu	“Many analysts use the v model to define big data the three vs stand for volume velocity and variety”
2013	Fujitsu	“The application of new analytical techniques to large diverse and unstructured data sources to improve business performance data sets that grow so large that they become awkward to work with using traditional database management tools data typically containing many small records travelling quickly data characterized by its high volume velocity variety or variability — and ultimately its value”
2015	Gartner	“The three vs of big data volume velocity variety are not all that needs to be managed”
2015	Gartner	“However the data crunching power required to manage the big data characteristics of volume velocity and variety does not inherently require any more sophisticated algorithmic processing”
2013	Hitachi	“Idc defines big data as a set of new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high velocity capture discovery and or analysis”
2013	Hitachi	“In other words sometimes the “big” in big data is not the sheer volume of data but the variety and or the velocity”
2012	IBM	“I started this book with a definition of big data using the four v’s velocity volume variety and veracity”
2012	IBM	“These results align with a useful way of characterizing three dimensions of big data – “the three vs ” volume variety and velocity”
2012	Intel	“The three vs characterize what big data is all about but also define the major issues it needs to address”
2012	Intel	“The confluence of the three vs drives a fourth value for any enterprise to succeed in driving value from big data volume variety and velocity have to be addressed in parallel”
2013	Oracle	“Because big data refers to data streams of higher velocity and higher variety the infrastructure required to support the acquisition of big data must deliver low predictable latency in both capturing data and in executing short simple queries”
2015	Oracle	“Big data has also been defined by the four “v”s volume velocity variety and value these become a reasonable test to determine whether you should add big data to your information architecture”
2015	Oracle	“We stated earlier that data volume velocity variety and value define big data but the unique characteristic of big data is the process in which value is discovered”
2015	Oracle	“The defining processing capabilities for big data architecture are to meet the volume velocity variety and value requirements”
2012	Point B	“Big data relies on emerging high powered data storage and processing technologies that capture store process and analyze data in which there is significant volume velocity and variety”
2013	PWC	“Big data has often been described by its attributes – notably volume velocity variety and veracity”
2015	Siemens	“The it research firm gartner defines big data as high volume high velocity andor high variety information assets that require new forms of processing to enable enhanced decision making insight discovery and process optimization”