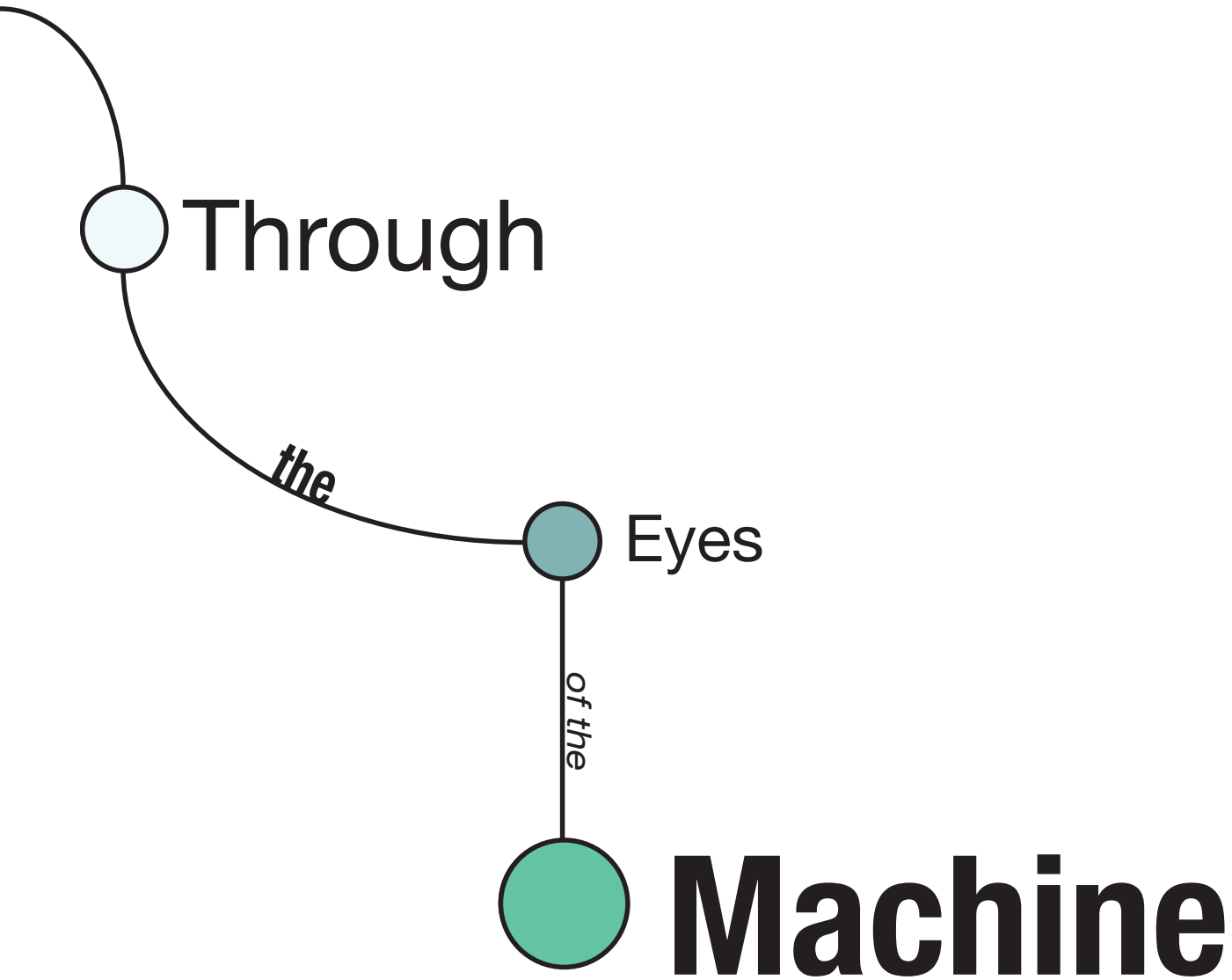


Exploring the emergence of Big Data through mapping Digital Expectations

By
Tao Legêne Thomsen
Tobias Bornakke

Through the Eyes of the Machine

Exploring the emergence of Big Data through mapping digital expectations

Authors
Tao Legêne Thomsen
Tobias Bornakke

Danish title:
Set med maskinens øjne: Digitale forventninger og fremtiden for Big Data.

Number of pages / characters: 130 pages / 270.894 characters.

Copenhagen Business School 2013,
Cand Soc. PKL, Department for Management, Politics and Philosophy

Submitted: 25/7-2013.

Layout: Tobias Bornakke Jørgensen

For more information please contact:
Phone.: 6065 5538
E-mail: info@ogtal.dk

Supervisor
Helene Ratner, Assistent professor, PKL CBS.

With special thanks to
Anders Kristian Munk, Assistent professor, Aalborg University
Anders Koed Madsen, Assistent professor, Aalborg University

ABSTRACT

This thesis explores how the trending phenomena Big Data came into being. It positions itself between the Sociology of Expectations and the emerging field of Digital Methods to conduct a large-scale, longitudinal study on how expectations and visions of the future of Big Data, are enacted in digital settings. Theoretically, the thesis draws on Actor-Network Theory, and employs a quali-quantitative approach to account both for patterns at a global scale and the particularities of locally constructed arrangements. Furthermore, it relies on Digital Methods to extract information by mining different databases, to map the development of different aspects of the networks in which Big Data resides.

The thesis studies the role of expectations in three ways: Firstly, it accounts for how expectations change over time. Secondly, it looks at how these expectations enroll a range of actors by associating them with Big Data. Lastly, it shows how expectations in the form of coherent visions propose a particular future shape for Big Data to take, thereby shaping the technological development trajectory.

We propose three contributions from this study. Firstly, the range and scale of data that Digital Methods allows this study to encompass provides a thorough account of the emergence of Big Data, providing empirical insight into one of the biggest buzzwords today and throwing light at how such large-scale trends and hypes arise. Secondly, the method we have created for the study is a proposal for how the Sociology of Expectations can be empirically operationalized for a digital reality. Lastly, by taking a critical look at the tools we employ we show the limitations and shortcomings of Digital Methods, augmenting the development of a field still in its formative phase and proposing venues for further development of large-scale digital studies of emerging phenomena.

PREFACE

This is the story of an idea. An idea that got very big, very fast.

This big idea is Big Data, an amorphous concept that went from a highly technical niche, a new method for parallelisation of computation, to an all encompassing umbrella term for the brave new world of ever increasing information flows.

In the process Big Data was heralded as a messianic answer to the challenges of mounting complexity, increased competition and scientific progress.

Thus it is also the story of Hype. Of exaggerated expectations, of hopes beyond measure. Of an idea in which people saw the future - before an understanding of it was even achieved. Of how these expectations thrust an idea into the forefront of future oriented thinking long before it had reached maturity in the present. Of how the desire for it to come true allowed it to appear and of how visions of it provided a form in which it could take place.

Faced with ever increasing complexity, with delugian data, we dreamed that the data itself contained the answer to the questions that arose from it, that a rosetta stone could be unravelled from its knots and tangles.

It is thus also the story of the idea that we live in a world, in which we can no longer make sense of our surroundings. It is a story of a world where most is unseen by man, endless arrays of values and cyphers- a landscape unsurveyable by human eyes, where most is only visible to algorithms and computer programs. Therefore, it is also a story we will tell through the eyes of machines.

But this is more than just a story. It is also a question - a host of questions actually. As we do not yet know how this story unfolds, nor how it came into being, the narrative is that of a mystery, and we set out to uncover the identities of those involved.

How did this conglomerate of ideas come into being? How was it shaped by the propulsion of the expectations vested in it? What can machines see? And how can they help us tell this story?

TABLE OF CONTENTS

1 ENTER ACADEMIA

IN THE FOLLOWING SECTION WE SET THE SCENE FOR OUR STUDY IN THE INTERSECTION OF BIG DATA, SOCIOLOGY OF EXPECTATIONS AND DIGITAL METHODS

1.1 PROBLEMATISATION

That Big Data as a term of interest has ascended in business, scientific and popular imaginations for some time can scarcely have escaped anyone who pays attention to such things. Big Data has been talked about in a score of different contexts; it has been heralded as the next frontier of business optimisation (Mckinsey 2011), as a harbinger of paradigmatic change in science (Anderson 2008) and as an oracle-like predicting power (IBM 2012). With enough data, anything can be known! Big Data has, however, also been criticised (Boyd and Crawford 2012; Crawford 2013), decried (Brockmeier 2012) and contested (Peters 2012).

Not much has been said about the technological underpinnings of Big Data, however, and its workings are mostly left black-boxed. The majority of discourse concerns what Big Data can contribute, what to expect from it and what its future role in society will be - instead of what it is on a material level, or what it is used for at the moment. An example can be seen in McKinsey's widely recognized report on Big Data, where the chosen metric to account for its relevance is a survey of how much CTO's expect to spend on Big Data in the future, not what they do now (Mckinsey 2011). So relevance is framed in terms of potentiality rather than actuality.

The common denominator for all these discussions is that they all point to the future - they are about expectations concerning what Big Data can be and what it will be able to do, not what it is now. While some deployments of this technology can be found already, it lives mostly as fragments of imagination, as visions of a future state, as expectations concerning its potential uses and contributions.

We therefore find it relevant to ask about how this future arose. Who are the actors driving these visions of the future forward? How are the expectations translated over time? How are expectations enacted differently in different contexts? And what happens when these often divergent expectations encounter one another?

1.2 SOCIOLOGY OF EXPECTATIONS

The impact of expectations on technological development has previously been addressed under the mantle of The Sociology of Expectations (Van Lente 1993; Brown & Michael 2003, Pollock & Williams 2010). A common feature of this discipline is looking *at* the future (and how it is constructed) rather than attempting to look *into* the future (Borup et al. 2006). So rather than something to be predicted, the future is seen as something enacted in the present, as an imagination or expectation. This does not make the future irrelevant, but rather opens up for considering how such expectations affects current actions and actors. It implies looking at the performativity of futures and their constitutive effect on the development of technological trajectories, e.g. how positive expectations for the future potential of a technology can mobilize resources for its development.

This approach is however largely theoretical and when venturing into empirical territory it has most commonly relied on a narrow spectrum of qualitative methods such as field studies, interviews and textual analyses (e.g. Brown 2003, Brown & Michael 2003, Lösch, 2006, Konrad 2006, Hedgecoe & Martin 2003). While these are undoubtedly valid methods, they are often rather time consuming. Given the limits of this study, they would not allow us to cover the temporal and geographical span we wish to embrace to account for the emergence of a widespread phenomenon like Big Data.

More quantitative approaches can be seen in the related studies of hypes. Hypes are temporal patterns of increased exposure and highly positive expectations to new technologies followed by disappointment and disinterest. In these studies a tool from the consulting firm Gartner, the so-called hype cycle, stands as a central pillar - both contested (Steinert & Leifer 2010) and corroborated (Jarvenpaa & Makinen 2008). This approach to studying expectations is often purely quantitative and built on relatively simple metrics such as media salience and search volume. In this way, it is able to conduct longitudinal studies across a large range of contributors - but tells nothing of the underlying changes in meaning or specific constructions of actors.

Both of these approaches can contribute to illuminating the story of how Big Data came into being, but from quite different angles: the empirical myopia of expectations studies can tell us about the intricacies of construction, while the birds-eye view of hype studies allows us to follow our subject over temporal and geographic expanses. In order to combine these two approaches, the rich qualitative constructionism of expectations with the sparse but wide spanning data of hype studies, we set out to find a way of combining and integrating the qualitative and quantitative approaches.

1.3 DIGITAL METHODS

A possible answer to combining the qualitative and quantitative approaches mentioned above can be found in emerging digital methods. The expectations to Big Data travel across common sociological distinctions such as *institutions* (DiMaggio & Powell 1983) or *fields* (Bourdieu 1993). This is illustrated by Chris Anderson's polemic editorial The end of theory (2008), that was published in a popular magazine but has been widely referenced in the academic debate (Madsen 2013:48).

This unruliness places two distinct demands on the research design. Firstly it stresses the need to put associations into the centre of the investigation. When the spread and development of expectations transverse the boundaries of classifications, we must follow. This thesis does this through the imperative in ANT to follow associations, letting relations be the decisive parameter guiding our analysis.

Secondly, the crossing of classical boundaries demands vast amounts of data and data sources. We need information on large amount of actors participating from widely different institutions, on different periods of the development, and located in many parts of the world - a quantitatively large span. It also demands qualitative knowledge on the constant translations and redefinition of the expectations that occur when they travel between the many different actors.

Due to resource constraints, data of this scope

has earlier been practically unattainable for projects of limited scope such as ours. But recent development in computational power and methods has opened a range of new ways to generate knowledge of the social, all sharing the feature of making data granular and scalable (Latour et al. 2012). This potentially opens a way of conducting science where one is neither locked to the micro nor macro scale, but can continuously zoom back and forth between the levels. This thesis will apply such a so-called quali-quantitative approach (Venturini 2010; Venturini & Latour 2010). Combined with new computational visualization tools to navigate the data (Madsen 2013) we wish to map how expectations to Big Data evolve, spread and change.

While we thus curiously employ new digital methods to chronicle the emergence of another digital method (Big Data), we stress that both in scale and techniques our approach does not qualify to be described as Big Data. So we do not study Big Data with Big Data, but with an approach that still relies on data and digital methods. The two approaches do, however, raise some of the same questions regarding ontological and epistemological positions - questions that we find to be inadequately addressed in the current literature and which will therefore also be scrutinized in this thesis.

1.4 RESEARCH QUESTION

Earlier studies have looked at how expectations change over time (Brown & Michael 2003), how they steer the development of a technology's trajectory and how interest for a product waxes and wanes (Van Lente et al. 2013).

We attempt to trace how Big Data as an idea and a term spreads over time and what role expectations play in this. To do this we will look at how expectations differ over time and space and how they provide propulsion for the technological trajectory of Big Data.

We will explore how digital methods can help us do that and what the consequences of using these tools are - to throw light on both this emerging field of social science studies and more generally the mode of knowledge production inherent in Big Data.

We thus aim at 3 theoretical contributions: firstly, we wish to account empirically for the emergence of Big Data, a phenomenon that attracts much attention but is still surrounded by uncertainty. Secondly, we wish to conduct an empirical application of the sociology of expectations that bridges the field's current quali-quantitative divide. And thirdly, we will develop a critical account of digital methods by disassembling the tools to look closer at what they can provide the social sciences.

This leads to the following research question:

How can digital methods generate knowledge of the hypes, expectations and visions surrounding Big Data?

With the following sub-questions:

- *How does the idea of Big Data emerge, evolve and spread?*
- *How can digital methods contribute to the empirical application of the sociology of expectations and the study of hypes?*
- *How do digital methods produce knowledge?*

2 METHOD ASSEMBLAGE

THROUGH THIS CHAPTER WE OUTLINE OUR METHODOLOGY AS A MESSY ASSEMBLAGE OF ACTOR NETWORK THEORY, DIGITAL METHODS AND THE SOCIOLOGY OF EXPECTATIONS ENACTED IN A MULTIPLE AND CO-FABRICATED REALITY.

2.1 ACTOR-NETWORK THEORY

Our research interest touches upon two separate research traditions in Science and Technology studies (STS): the sociology of expectations and digital methods. In order to establish commensurability between the two, we will present another branch of STS, Actor-Network Theory (ANT) that will provide us with a theoretical vocabulary[1] to discuss their relationship. We will then present an overview of more traditional ontological and epistemological considerations (chapter 2.2, 2.3 and 2.4), before we introduce our Digital tools (chapter 2.5) and a literature review of the Sociology of Expectations (chapter 2.6)

The term ANT has been both praised (Latour 2005) and dismissed (Latour in Law and Hassard 1999), with special regard as to whether it should be understood as a theory or a method, as ANT does not seek to provide an explanatory framework per se (Latour 1988) and thus is not a theory "of" anything. On the contrary, later work in ANT, sometimes referred to as Post-ANT, describes it as a method of studying relations between entities regardless of their nature. It is called a travel guide (Latour 2005:17), a method assemblage (Law 2004) or a technique, disposition or attitude (Gad & Jensen 2009:62). As Law summarizes it:

> *"Actor-network theory is a disparate family of material-semiotic tools, sensibilities and methods of analysis that treat everything in the social and natural worlds as a continuously generated effect of the webs of relations within which they are located"* (Law 2007).

Post-ANT is the result of attempts to inquire into ANT with its own methods, and is related to what has been called the ontological turn, a movement away from the pluralist perspectivism common to post-structural social theory that studies the effect of different culturally constituted perspectives while viewing the underlying reality as a passive element (Gad & Jensen 2009; Ratner 2012:72; Ratner 2009). In contrast to this, ANT tries to put reality, and the myriad of ways in which it is constructed and enacted, in the foreground (Latour 2005:91) - to add to reality rather than to subtract from it (Latour 2004a:246). In this way, things are made more real by showing the multitude of connections they draw together: the construction of reality is exactly what makes it real (Latour 2005:89), and the world is not seen as a passive object, but as an engaged co-participant in this construction of reality (Stengers 1997).[2]

We will approach ANT and its implications for our research through John Law's concept of method assemblage (Law 2004) presented in the next chapter. But first we will start by introducing a few ANT core concepts and enunciate our definitions of them.

2.1.1 Actors

A central tenet in actor-network theory is the insistence on extending agency to non-human actors. This controversial conferring of privileges to objects, normally reserved for humans, is seen in the concept of the actor, of which Latour says:

> *"an 'actor' in ANT is a semiotic definition -an actant-, that is, something that acts or to which activity is granted by others"* (Latour 1998:5).

Borrowing from Greimas' semiotics, an actor is anything - humans, objects, institutions, ideas, measurements and calculations - that makes a difference to a state of affairs.[3]

As we shall see later, our study assembles digital traces of a host of obvious actors including authors, researchers, firms and universities. Another important group arises from the technological underpinnings of Big Data in the form of hardware, such as server farms, fibre cables, RFID tags and ubiquitous sensors, as well as software including computation architectures, machine learning protocols and predictive algorithms. Other actors could be called ideas: expectations and visions of the future, enactments of the past, diagnoses of contemporary society - often encapsulated in the form of buzzwords[4].

These are all fairly traditional candidates for the constitutive elements of ANT studies. But our method introduces still other digital actors: Firstly, the use of digital methods means that we will only ever see traces of the aforementioned actors, as they are only visible to our machinery when they take the form of digital traces: keywords, search terms, articles, web domains and hyperlinks. By digital traces we refer to the recording and archiving of digital activities that makes it possible to follow their tracks (Venturini 2010). These digital traces are themselves actors.

Though it should be noted that these actors, the digital traces, differ from the aforementioned in that the first affect the emergence of Big Data, while the latter only our depiction of it, we award them the same status in our analysis. Just like a microscope is as much part in constructing germ theory as bacteria are, our tools are placed on the same level as what they are studying.

Secondly, our own devices for calculating, gathering and inscribing emerge as potent actors on their own, particularly to the extent that their inner workings are hidden from us. Finally, our study is shaped monumentally by the architectures of archiving, access and retrieval in the places from where we get our data: search rankings at Google, scientific databases such as Scopus and newspaper archives. Thus our study is also partially bound by these actors.Therefore we will analyse our analytical apparatus, including our tools among the actors that construct the phenomena we are studying.

2.1.2 Networks

Network is the other central term and is easy to misunderstand, especially in a digitally situated study due to the number of already well-established meanings ascribed to the term within computer science. Both in an engineering sense, as in the network of cables and servers that make up the Internet, and in the sense of user generated social networks.

What ANT refers to as a network is instead the association of several actors into a temporary arrangement (Ratner 2012). The relational ontology of ANT (Latour 2005) thereby posits that any given actor can be described by the associations between the components collected in it (as any actor is itself a network) and every network described as the assemblage[5] of actors that make it up.

The network is thus not a stable arrangement, but must continuously be enacted, and the term is not used to describe these stabilised assemblages, but rather the portrayal of the work done to stabilise it:

> *"No net exists independently of the very act of tracing it, and no tracing is done by an actor exterior to the net. A network is not a thing but a recorded movement of a thing"* (Latour 1996:14).

This definition is also central to our understanding of the representational power of maps and other data visualisations, which we shall return to later.

We will thus neither use the term network to denote social networks nor the Internet itself, but as the emergence of Big Data as the continuous work of associating and reassembling relations between actors. Analysing the network of Big Data is not about finding a shadow cabinet of powerful individuals who brought it into being, but rather to try and span the works done across actors - high and low, human and material - that goes into creating, shaping and spreading Big Data. This process is described through the concept of *translation*.

2.1.3 Translation

If actors are described by their associations and networks are formed by the continuous drawing of new connections, it follows that actors will change when the network evolves (Gad & Jensen 2009). When something moves or grows, it changes. This transportation of transformation is called translation (Latour 2005 & 1994; Callon 1986).

So when we say we are tracking how the idea of Big Data grows, spreads and changes, we are in effect looking at the translation of Big Data. Translation is the process by which networks are established and stabilised by some actors on account of being able to represent, or talk on behalf of, other actors (Callon 1986). From Callon we will also borrow the terms *interessement* and *enrollment*. The former is the enactment of a network in such a constellation that associating with it becomes alluring to other actors (Callon 1986). The latter we take simply as the process of drawing connections to actors, and as such use it interchangeably with association and mobilisation for linguistic variation.

Though we draw on parts of Callon's vocabulary of translation, it should be clear that our approach to the translation concept differs somewhat from that of Callon. Our use will be more generalised and open ended than Callon's usage, by not sticking to a four-phased model of translation[6] and by viewing the modus operandi of translation more as a fluid nonlinear processes than a number of fixed consecutive phases.

In the following we will touch upon one process by which actors can represent others; inscriptions.

2.1.4 Inscription

An important part of translation processes is when it is made possible to act upon something from afar (Latour 1988). This can be done by letting them be represented in text - translated into textual forms that can circulate and create present things that are otherwise absent (Justesen & Mouritzen 2008). Latour calls such machines inscription devices, and their work of rendering materials into inscriptions are found to be integral to the fabrication of scientific facts (Latour 1987; see also Law 2004:20).

By inscribing actors into text, other actors can represent them - talk on their behalf, stabilise their meaning, assemble a collective and act on them, all from a distance.

In our work, the digital tools and especially visualisation tools are ways of inscribing digital traces and making them present in our text. And it is precisely these devices ability to inscribe, collect and trace a large amount of actors with relatively low cost that makes it possible for us to cover the temporal and geographical span we desire to account for the translation of Big Data.

2.1.5 Summary

In this chapter we have introduced a conceptual vocabulary for describing both digital methods and the sociology of expectations. We described Actor-network theory as an assemblage of material-semiotic and methodological notions with a relational ontology of association at its center. Afterwards, we ventured further into this assemblage by introducing a number of its core concepts. Here, we presented the heterogeneous network of actors we expect to meet in our exploration of Big Data, ranging from human actors and organisation to computer hardware and software. Special emphasis was placed on the digital traces, along with ideas of future visions and shared expectations.

We then introduced how these actors are translated into never completely stabilised actor-networks. Finally we dived into a specific type of translations called inscriptions where actors can represent or talk on behalf of other actors through e.g. visual representations.

2.2 MULTIPLICITY & METHOD ASSEMBLAGE

The constructivist turns have led to a movement away from more traditional theory, method and methodology (e.g. Andersen 2003; Watson 2003; Esmark et al. 2005). Portrayed slightly generalized, this movement have attempted to reconstruct the relationship between observer and the observed when faced with the knowledge that the observer and the observation itself, always participates in the construction of the observed. One often cited example is Andersen 2003, which proposes an 'analytical strategy' as an alternative approach to theory and method- a program for (second order) observations that brackets ontology to elucidate on how epistemological perspectives are constructed and in turn construct their observations (Andersen 2003). Binary oppositions are central to this program, and the reservoir of these binary pairs constitutes the analytical toolbox. The empirical is seen as an effect of the particular binary difference employed in observation. By enforcing strict discipline in the deployment and conditioning of binary pairs, the analytical strategy strives to gain a more reflexive, critically distant and accountable depiction of how a given observation is constituted (Ratner 2012). Thou this epistemological constructivism aims at making sense of an admittedly contingent world by purity of method[7].

In contrast to the analytical strategy recent developments in Post-ANT have been centred on developing messy methods to handle a multiple and fluid reality with multiple and fluid methods (Law & Urry 2004; Law 2004; Sommerlund and Jespersen 2008). The underlying rationale is that for a given phenomena, different practices generates different material realities (Law 2007). Rather than a plurality of perspectives on a singular and inert world, ANT views a multiplicity of enactments of reality, where the world itself contributes to the enactment (Latour 2005:88).

The empirical is thus seen as co-fabricator (Stengers 1997) and not an inert object of observation, in the sense that knowledge is neither purely objective facts residing in the world, nor perspectives of the subjective mind. Instead, they are seen as the result of an engagement between both the researcher and the researched, both of whom contribute to the generation of knowledge. As Callon shows (1986) both scallops and the tools used for breeding them play a crucial role in the scientist's work of producing knowledge of them, so knowledge is the result of continuous work on both parts to draw connections between the different actors.

The world is thus made up of continuous engagements that each mobilises different elements (Mol 2002). The different engagements enact reality in different ways. What seems singular at first - a disease such as arteriosclerosis, or a technological development such as Big Data - emerges as a set of multiple material realities enacted by different actor-networks: one is composed of servers and cables, another of bits and computations, others still of business plans and roadmaps. So rather than seeing reality as singular, different realities are enacted by different engagements.

When these realities merge into a singular account, it is only by the extensive work of the actors - and then only momentarily. This is not the same as the perspectivism suggested by post-structuralism (Gad & Jensen 2009). Knowledge, rather than being multiple perspectives on one world, is itself the result of engagements with different collectives that enacts multiple realities (Mol 2002; Ratner 2009; Ratner 2012), and the current discussion has therefore largely been focused on developing new approaches capable of co-constructing and navigating these multiple realities.

One example of such an approach is Law's own concept of *Method assemblages* (2004). This concept is developed from the argument that reality - and knowledge of it - is neither pure perspective arising in the subjective mind nor simply objective facts waiting "out there" to be discovered, but something that emerges through co-fabrication between researchers, their tools, methods *and* the things they are studying. With the concept of Method assemblages Law argues both that different methods assemble different realities (ibid.:21) and that methods themselves are assemblages. Methods are always embedded and co-inscribed in larger networks of scientific practices, theories and earlier results, what he denotes as their hinterland (ibid.: 28). So rather than viewing method as a singular dogmatic guideline, we should view scientific

practices as assemblage of tools and protocols that enact a particular reality, and be aware of the ramifications of their construction - both the networks they are dependent on, but also the particularities of how they produce knowledge.

In other words, the method assemblage is the collection of relations that makes a specific reality come into being. Working with method assemblages is to present and make visible the objects and contexts that makes something visible (presence) and how it deliberately or un-deliberately leaves other parts hidden (absence) (Law 2004:55).

We will lean on this notion of Method Assemblage in the construction of our own research design, both at a strictly descriptive level - we will construct our study as a bundle of methods, tools and protocols that each enact particular effects - and as mode of inquiry into how our methods generate particular realities. This will answer our third research question (how digital tools enact reality). It will be the subject both of chapter 2.4, and also subsequent discussions of our findings and their implications in Ch. 4, 5 and 6.

2.2.1 Summary

In this chapter we have positioned our method in the recent Post-ANT discussions of messy methods as a break with constructivist and post-structuralist conceptions of purity of method. We hereby situate our study in a co-fabricated and multiple reality where every engagement enacts yet another material reality.

To navigate in these realities we introduce the concept of method assemblages as a way of directing attention towards the relations enacting presence and absence in our analysis. In the following chapters we will bundle our own method assemblage with this focus. We will do this through relating the brief summary of the core concepts of ANT to the specific ontological, epistemological, and methodological conditions for our digital venture, the tools we use and the theory from the sociology of expectations.

2.3 TRACES OF A DIGITAL ONTOLOGY

In this chapter, we will outline some considerations on the ontological status of the digital world responding to ontological considerations and its relation to the physical, analogue world. How is the digital composed and ordered? What consequences and possibilities does it offer to social science? And how it should be navigated?

We thus venture into more classical methodological territory.

2.3.1 There is no Space in Cyberspace

Quotidian observations and common sense commentary often describe the digital as an alternative dimension, a virtual world, by the telling synonymous neologism *cyberspace*, implying that the digital world is in some way spatially separate from the "real world" (Graham 2012). Graham contests this spatial interpretation as a constraint on the internet's potentiality for mediating information, since it is exactly the non-spatiality of the internet that allows it to transverse geographical boundaries (ibid.).

A similar argument is authored by Nathan Jurgenson, who tries to collapse the divide between a digital and a physical reality and argues for an intermeshing between the two he calls *Augmented Reality*: digital and physical exist on the same level of reality. They are not mirrors of each other, but intermingle in a heterogeneous extension of each other (Jurgenson 2011). This conceptualisation mirrors the ambitions in ANT to avoid bifurcations, to not separate reality into divisions such as nature/society (Latour 2004a), subject/object or micro/macro (Venturini and Latour 2010).

As such we accept that no ontic separation is drawn to the digital - although the digital is certainly still engaged in mobilising a wide range of ontologies. That does not however answer the question of what implications the digital world has for research. To answer this, we will first look at the sources of digital data.

2.3.2 Digital Traces

In a slightly simplified characterisation, the first meeting between analogue and digital has been characterised as a migration of existing practises into new territories (Rogers 2009). In what Richard Rogers calls the era of the virtual, researchers moved to the world of the digital. What social science had earlier done in the 'analogue' world was now done in the digital, as virtual methods. Surveys were turned into digital surveys, interviews were conducted in chat rooms and researchers on group dynamic began studying virtual group dynamics on Internet forums (Schroeder and Meyer 2012). Though the digital was treated as a realm of its own, it was always approached with the tools from the analogue world.

This subordination of the digital to the physical may be changing. On average since the 1970's, the price of storage and processing power has halved every second year[8]. Concurrently we have witnessed an increase in time spent on digital interaction bound to leave behind traces. When we turn on our GPS, visit a blog or interact with the government, traces are left behind. This has paved the way for a situation where, as Tommaso Venturini describes it, data is *"easily recorded, massively stored and inexpensively retrieved"* (2010:6).

As part of this development we have witnessed a gradual turn toward *native digital* data; data born in the digital world as a result of our digital behaviours rather than through the digitising of analogue phenomena (Rogers 2009). A key feature of these digital interactions is that they are imminently traceable; activity always leaves a trail. This makes it very cheap, almost free, to acquire large amounts of data (Venturini 2010:804).

The availability of these digital traces is the foundation for both computational social science (Lazer 2009), the digital method initiative led by Richard Rogers and Macospol Medialab, a project in digital methods led by Bruno Latour (Latour et. al 2012; Venturini 2009; Madsen 2013).

So native digital data is imminently traceable, and is therefore central to different approaches in digital methods. But if the web, and digital traces, are not ontologically different from the physical world, yet allows for the mobilisation of new ontologies, then from where does this difference arise? We venture a proposal: from the manner of their ordering, from their assemblage - because the web is unordered, yet eminently traceable. In the following chapter we will argue that this has important implications regarding how it ought to be portrayed.

2.3.3 Ordering the Web

A fundamental principle in the organisation of digital information is that the internet has no centrally constituted control, no governing body and no hierarchical mode of ordering (Flyverbom 2011; Castells 2003). It has been called a self-organising entity due to this lack of central archiving or indexing (Fuchs 2003), where any order that exists is only due to the work of local and distributed actors: homepages linking to each other, curatorial collections, aggregated news streams etc.

The Internet, in this regard, is in effect like what Venturini calls magma - neither completely solid nor totally fluid, but flowing in streams and eddies, temporarily and locally achieving solidity (2009). For Venturini this is, of course, a feature of all social affairs that are not yet completely stabilised. But the degree to which the technological underpinnings of the internet exemplifies this understanding of order as ephemeral emergence, forces us to stay true to the admonition of not resorting to predefined categories, but let the work of the actors themselves formulate the order; to take their own world-building activities into account (Latour 2005).

So how is the Internet organised, how does it assemble into order? This is even closer to ANT conceptions of ontology. To answer it we will take a closer look at the activities of one actor whose ordering work we will draw on as input in several of our analyses later: Google Search and its pagerank algorithm. Its work of ordering digital information has been so successful that it appears as a de-facto index of the Internet for many users. As we shall see, that is however not the case: Google is an actor on par with others, and its reach relies only on the number and strength of associations it is able to manage, not on a shift to another level. As we shall see, in our tracing of its ordering mechanism, we are able to follow Latour's admonition to not jump between local sites and global forces, but rather follow the connections of local actors that allow them to reach out globally (Latour 2005:1976). Similarly, we shall show that Google's global reach is the result of an arrangement that lets it feed off of local connections, and it is only because of the fact that it lets these associations take the foreground that it is able to attain its strength. How it achieves this will be elaborated in the next chapter.

2.3.4 Feeding of Local Connections - The Global Span of Google

While simply entering a query into Google seems like a quotidian operation for most of us, a brief familiarisation with the inner workings of the software reveals that Google's search and ranking algorithms are indeed tremendously powerful actors with which the researcher can ally him or herself with relative ease. Though the algorithms' precise workings are black-boxed (Latour 1987), the underlying concepts are known in a general outline. Thus we will argue that repurposing the digital traces collected by Google in various ways is not only a legitimate research tool, it is also in close accordance with the concept of association (Latour 2005).

Google's search results and rankings are composed and constructed by the intricate work of millions of automated minions called web crawlers (affectionately known as ants or spiders). These are simple scripts that trawl the textual data of the web, single-mindedly reporting back on the occurrence of keywords, and more importantly, tracing the hyperlinked associations between entities on the web (Madsen 2013:22). They do not, in any way discern as to what the pages contain or the quality of their content, nor do they try to categorise them in accordance to any preconceived taxonomy[9]. In this way, crawlers are thoroughly agnostic and treat all entities symmetrically (Latour 2005; Callon 1986).

They inscribe order only by drawing on the co-occurrence of keywords with references by hyperlinks between pages. So for any given keyword, an assemblage is traced that enacts a corpus of web pages *only from the hyperlink references between the pages themselves*[10]. This is the page of search results that emerges from a query.

In this way the activity of a crawler follows surprisingly close to the dictum of Latour (2005:176): to refrain from trying to achieve a higher order meta perspective and instead meticulously depict local ordering arrangements by the tireless tracing of associations and nothing else.

This argument is important enough to warrant repetition: this ordering of the internet is not achieved by a transcendent or objective perspective categorising every page, but rather by tracing associations between different homepages in the form of hyperlinks.[11] A site that receives a lot of links related to a particular keyword is deemed central and relevant to that particular keyword.

The relative ranking of entities is based on the position of the particular page in relation to the entire network of pages, i.e. the one that receives the most links (and the most important links, from other sites that are also highly ranked). The top ranking results are thus those that most other pages pay homage to in the form of linking, the ones that can assemble the longest chains of associations. So the crawlers create an order that makes visible which of the billions and billions of web pages in existence are able to muster the largest networks.

Is this an acceptable definition of relevance for our needs - can we use this enactment? That power and influence arises from the ability to draw on a large network of connections is congruent with findings in many ANT studies (e.g. Callon 1986 &1987). So if we follow the actors' accounts of relevance, and accept that hyperlinking is indeed a meaningful association, these tools fit the bill nicely. We thereby lean on Rogers' (2009) idea that criteria of judgement should be extracted from the same digital reality as the research is conducted in.

Regarding the reliability of web page ranking, inquiries made into their democratic qualities and representability of search engines indicate that niche subjects are underrepresented compared to commercial and mainstream subjects (Van Couvering, 2007). Other critiques have addressed a perceived homogenisation of web results by personalised algorithms (Pariser 2011)[12]. While this is undoubtedly a legitimate critique in terms of notions of political fairness and democratic representation, in light of the previous discussion we can see that rankings are indicative of the degree to which actors can assemble associations, be they commercial or what not. That relinquishes our reservations as to their compatibility with our chosen approach.

In summary, while Google Search results seem academically trite at first sight, they offer a valuable opportunity to engage an actor of tremendous brute power in collecting associations[13].

They do however have their weakness: although their underlying principles are known they remain operationally black boxed. By using them we surrender epistemologically to unknown forces. When dealing with information on the scale that contemporary society presents us with, this is an insurmountable condition for leveraging the reach of our observations. This invisibility must however be taken into account as a limitation, and the results seen as a approximations, indicative of scale and trends more than precise metrics. We must accept that this number does not show the meaning or relations between terms. These subtler measures can only be explored by zooming in to the higher resolution of individual entities[14]. Instead, the number merely serves as an indicator of the extent to which a particular idea has succeeded in gaining traction as a subject of conversation.

2.3.5 Summary

We have tried to outline some of the current discussions on the ontological status of the digital. We started out by discussing attempts to collapse the divide between the digital and analogue, and introduced the two as heterogeneous extensions of each other. We then presented digital traces as a native form of digital data, marked by its non-hierarchical ordering, its sheer scale and its intrinsic traceability. Based on these observations, we discussed how the basic organising principles of the digital offers both possibilities and drawbacks when evaluating relevance and identifying the strongest relation. From this discussion we concluded that relying on Google's ordering mechanisms would be an acceptable and potent approach in later analyses. These possibilities will be further explored when looking at the representational and epistemological context of digital traces in the following chapter.

2.4 REPRESENTATION

A central discussion in nearly all digital approaches is the questions of representation.

This question of representation is of course not new to the social sciences, and has also been debated elsewhere in the field of ANT. Latour e.g. sees the question of representation as a general question for all of science, since scientific texts always try to portray something that is absent in the text, whether that is tissue samples, scientific practices, indigenous populations or atomic nuclei. Therefore, a connection between the absent object of investigation and the text must be stabilised in form of information to provide a convincing account. Establishing such relations between the absent entities described and the text in which they are represented is done through inscription devices translating them into information (Latour 1988:159). Knowledge production, or the convincing argument, is then the act of building networks of such connections between the represented and the representations (ibid:160)

So how does one make such inscriptions? How do we represent what do we represent, and what is the relation between the represented and representation?

First we will discuss how digital traces allow for different scopes of representation. Then we will position ourselves in relation to realist and constructivist approaches. Lastly, we will consider Latour's 'politics of explanation', and describe the format our own contributions to the reality we are studying.

2.4.1 Quali-Quantitative Methods

As mentioned in the previous chapter, one of the primary affordances of digital data is its inherent traceability. This traceability has been argued to offer the possibility of transcending the need of traditional distinctions such as micro-macro, analogue-digital or quali-quantitative (Latour 2007; Venturini & Latour 2010; Venturini 2010; Latour 2011), of which especially the latter is deemed important for our thesis. This discussion has addressed the ANT problematic of *"how to follow stronger, wider and longer lasting associations"* (Latour et al. 2012:2) by suggesting a solution in the form of a quali-quantitative approach. In this approach, the traditional divide between qualitative/quantitative and micro/macro is not seen as fundamental dichotomies, but rather as a mere question of methodological approach (Latour 2011), building on an age-old sociological discussion on aggregation (see e.g. Tarde 1903).

Latour and Venturini propose that the distinction between micro and macro (and the related distinction between qualitative and quantitative) arose out of methodological restrictions in early sociology. They argue that the complexity of human interaction is too high for researchers to maintain both breadth and depth in their studies, but are forced to focus either on the intricacies of local phenomena *or* the broad strokes of global structures through a mathematical leap of statistical aggregation, without the possibility of connecting the two levels of analysis (Venturini & Latour 2010:4).

Instead they argue that the aid of digital data and its inherent scale- and traceability offer ways of practically bridging these dichotomies (Venturini & Latour 2010). When every aggregated point can be pinpointed, its traceability allows it to be zoomed in on and unfolded. This ability allows researchers to traverse the data across the perspectives of both micro and macro, qualitative and quantitative, allowing one to zoom between levels (Venturini & Latour, 2010; Madsen, 2012; Venturini, 2010). E.g. we do not only get a purely quantitative growth curve when we perform our large scale tracings on the mentioning of Big Data, but are parallel able to follow every single contribution in detail, as the list of websites that make up these aggregated sums are simultaneously presented as links no more than a click away. Thus our method allows us to both disaggregate and reaggregate, viewing our empirical data as both qualitative and quantitative while retaining the ability to trace the connection between singular local actors and global networks. Because these two levels of analysis can be mapped onto a singular framework, we can combine the insights.

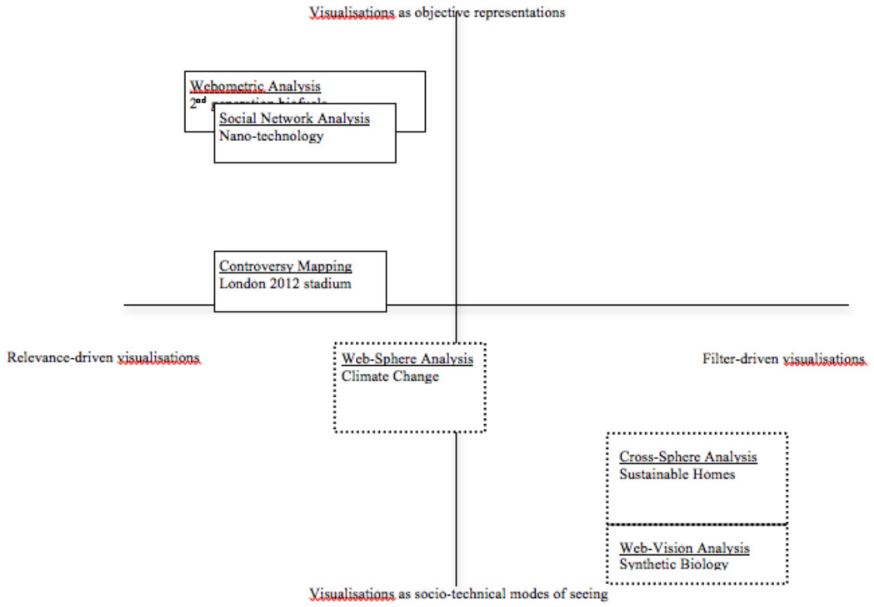


Figure 1: Meta-study of approaches in digital methods (Madsen 2012:8)

2.4.2 A POSITIVIST DRIFT

We would however like to rectify a tendency to commit a slight misconception. Both Venturini and Latour, the main proponents of the quali-quantitative approach, often describe the connections between traces as *a priori* relations (See e.g. Venturini 2010:11; Latour et al. 2012:3, Latour 2007). This poses a risk of succumbing to quasi-positivism. While we are fully aware that Latour and Venturini would theoretically never commit to such a positivist view, we still contend that the aforementioned articles have a tendency to drift towards it, inadvertently or not. In the excitement of the new digital possibilities, they seem to break with earlier notions in ANT that associations, even the digital, are always situational constructs, and that we ourselves are co-fabricators of them (Stengers 1997).

Our position is that while the digital traces do in fact offer a previously unknown degree of traceability, the relations are never generated by themselves, but are always the product of an actor. A more fitting depiction than the intrinsically interrelated digital traces would be to underline ohow digital traces being native digital data are easily manipulable and connectable (Rogers 2009), and how this offers ways of building relations between even strongly heterogeneous data sources. We will draw on this ability to construct a host of maps, each enacting a particular form, and ground our analysis in the juxtaposition of these maps to avoid falling into the same trap. So instead of stabilising our data as a singular object, we will enact it in a multiplicity of ways. These clarifications are important in relation to how the 'shock and awe' of vast data numbers and the increasing appearance of practitioners without sociological basic knowledge have seemed to reintroduce a naïve positivism to social science (Crawford 2013; Madsen 2013).

The positivistic tendency that we posited above has been fully embraced in other contexts, prevalent in e.g. the idea of an *end of theory* (Anderson 2008). In a widely cited Wired article, Chris Anderson points out how the digital era is an era where access to digital traces, instead of sociological theory or research interest, form the decisive parameter when gathering knowledge on social matters. He claims data is slowly making sociological theory obsolete since: *"With enough data, the numbers speak for themselves"* (ibid.). This idea has not only gained public attention, but has also been put to practice by e.g. the so-called social physicists; natural scientist who have specialised in social studies deprived of traditional sociological theory but employing massive behavioural data (See e.g. Newman, Barabási and Watts 2006; Lazer 2009; Freeman 2011).

So in summary we have argued that the traceability of digital traces and subsequent ability to zoom continuously between micro and macro levels allow us to employ a quali-quantitative method. We do however distance ourselves from some properties that others have argued emerges from this - such as the redundancy of theory – which will be a topic of the following chapter.

2.4.3 Entanglement

The following discussion tries to address in what respect digital traces should be seen as representations of the phenomena they are aligned to depict, and how this affects the fabrication of knowledge.

Digital Methods have generally been split between two epistemological positions, seeing maps and digital traces as either *'objective representations'* or *'socio-technical modes of seeing'* (Madsen 2013 - see figure 1). In one end of the scale digital researchers, especially social network analysts (Madsen 2012), occupy traditional realist positions with a clean division between reality and the observer, and a fundamental understanding of digital traces as objective representatives of an underlying reality. In this line of thought the idea of digital traces as 'honest signals' less prone to research bias has been dominant (Pentland 2008). The investigators should therefore in this perspective avoid contaminating the digital traces through human biases.

On the other end of the scale, digital traces are depicted as *socio-technical modes of seeing.* Following constructivist tradition, the digital traces are not to be taken as a consequence of some underlying phenomena, but merely as perspectives arising because of a mix of *"technological, human, and social influences"* (Madsen 2012:2).These approaches are positioned toward the bottom of figure 1.

2.4.4 Co-fabrication

As described in chapter 2.2, we align ourselves with approaches that try to avoid this schism between realism and (idealist) constructivism (Whatmore 2003; Stengers 1997; Latour 2004). Rather, we avoid distinction between the observer and the observed as they are always entangled, adopting an approach Latour has described as a *realistic constructivism* (Latour in Ratner 2012:80)[15]. The product of our observation is always a *shared* accomplishment (Marres 2012) between a myriad of actors entangled in the situation, making knowledge production a situated *co-fabrication* (Stengers 1997; see also Haraway 1988).

In our study, the digital traces represent neither reality nor complete construction, but instead a concrete mobilisation of reality as viewed through the eyes of the co-fabricators. Epistemologically the perspective hereby implicates a possibility of coexisting contradicting results, since disagreements should always be seen as mere differences in which actors are assembled in the co-fabrication (Law & Urry 2004:397). The results of our experiments are co-fabricated ontologies - neither our interpretation nor an objective account, but a potential reality brought into being by the tools, infrastructure and theories available to us, by the work of the actors we studied and our own work.

2.4.5 Redistribution

This entanglement also leads to what Marres has described as a redistribution of research: the access to data and capability of using it is no longer the sole domain of social scientist, but is also in the hands of commercial actors (who gather much of the data) and other scientific disciplines. This has led both to optimism among those who argue that digitalisation will improve the world (E.g. Newman, Barabási & Watts 2006; Anderson 2008; Lazer 2009) and pessimism in those seeing it as a sociological crisis (E.g. Savage & Burrows 2007; Boyd & Crawford 2011; Crawford 2013).

It is not just the production of data that is redistributed. It is the whole chain of research skills—from the data collection to analysis and visualisation—that is distributed across online platforms, web users, meta-data providers, algorithms and professional analysts (Marres 2012; Madsen, 2012). Instead of judging the general consequences of a reconfigured social science she argues for an empirical inquiry into new centres of research capacity: *"...by concentrat-*

ing on this overarching issue of the displacement of research capacity – to society at large, or the IT industry – we risk losing from view another, more fine-grained dynamic: the redistribution of social research between actors involved in social research (Marres 2012: 143).

For our own study, this redistribution has a big impact on the distribution of our analytical work. Since much of our analysis will be centered on the manipulation and visualisation of data - the method by which we assemble the multiplicity of enactments we have discussed in this chapter - a large portion of our analytical work might appear as what would be seen as mere empirical gathering and presentation in traditional sociological research. In contrast, in this redistributed mode of research in which we are engaging, is the analytical mode not confined to mere interpretation of presented data, but is prevalent *throughout* the entire chain of translations from data harvesting through inscription to discussion.

2.4.6 Explanation

As mentioned earlier, Latour's understanding of representation is closely linked to what he has termed *politics of explanation* (Latour 1988). In this approach, representation is modelled as two lists of elements - one of which (B) are the absent elements, and the other (A) their representations in the form of inscriptions (Ibid:157). Explanatory power can be mapped on a scale from descriptions to deductions. Deductions, which have the highest power, are situations where list A contains only a few elements that can account for all possible elements appearing in B. Descriptions, on the other hand, provides neither deduction nor correlation between list A or B: list A contains a large number of elements and can claim no singular causality as to their relation to the elements of B (ibid.:158). Descriptions, or narratives, thus do not reduce the complexities of the matter they are trying to depict, while deductions work by substituting a large number of elements with singular, monocausal representations. Deductions have larger explanatory power, because they can stabilise an arrangement between a centre and the absent represented setting (ibid.:159).

But should we even aim for explanations? Rather than abolish the distance between the represented and representation by reductionist means, Latour admonishes that our accounts should abstain from *adding* to the text additional reflexivity (Latour 2004a & 1988) or causality, but rather leave room for the represented itself.

The maps we draw are simply a way of trying to stabilise a portrayal of the actors we are following - the engineers, entrepreneurs, scientists and commentators who propel the issue of Big Data.

In our study, should the notion of hypes and expectations e.g. not be seen as explanations of the activities of the actors; they do not reduce the sum of associations to a singular placeholder. We are not trying to explain their translations because of hype, since *hype is nothing more than those very translations.* We are also not trying to impart reflexivity on presumably unknowing subjects: they are the ones who have articulated the idea of hype around the technological development of Big Data, not us.

So if not explanation or reflexivity, what is then our contribution?

Tommaso Venturini proposes *Second degree objectivity* as a possibility of using digital tools in the mapping of controversies (Venturini 2010). This entails ensuring representation to all the involved actors and viewpoints in a given controversy, rather than trying to determine which are "right". The way he proposes to do so is by drawing successive maps of involved actors and layering them so the relative strengths, positions occupied and relations drawn between them becomes visible - forming what he calls monads (Latour 2004; Latour et al. 2012). While we do not fully ascribe to this understanding of the objectivity achieved thus, vis-a-vis our earlier discussion on the posivistic drift in digital methods (see chapter 2.4.1), we do agree that the ability to represent exceedingly large numbers of actors is key to understanding the explanatory and representational power of digital methods.

The visual representation of large datasets in themselves seem to have quite a convincing effect (Madsen 2013), and it is our personal experience that their explanatory power is often

overestimated. Latour stresses the importance of academic text rendering themselves believable, providing a convincing account for how and why we should accept that A can represent the absent elements B. He proposes that we should throw light on the work needed to stabilise the relations between A and B, to make our depictions more believable, rather than try to hide or interrupt the distance between them (Latour 1988: 173).

It is precisely because digital mappings have the ability to represent such large amounts of actors, that they achieve their convincing power. Because they can combine so many B's into a singular graphical depiction A by processes quite inaccessible for laymen glossing over the distance between A and B. But rather than succumb to temptation to claim objectivity for our accounts, we remember that both our tools and our choices are co-fabricators in the accounts we produce (Stengers 1997). Therefore we will continuously disassemble our tools to account for the redistribution of agency they entice (Marres 2012); not to diminish the credibility of our mappings, but to add to the reality of them by showing the work that goes into representing what is absent by translating it into inscriptions in our text.

In extension of the earlier discussion in this chapter, we are merely finding a way we can represent mess, account for large amounts of actors and the work they do in stabilising networks. Rather than bloated claims of reaching objectivity, we venture a humbler approach: we are simple tailoring our methods to provide the most room for the largest number of actors.

2.4.7 Summary

In the above chapter we have elaborated on the epistemological consequences of what can be described as a digital representation. We started out in classic ANT theory, defining scientific representation as the attempt to portray the object of investigation, always absent from the final text. In our study, digital traces represent therefore neither reality nor complete construction, but instead a concrete mobilisation of reality as viewed through the eyes of the co-fabricators.

From here we zoomed in the discussion on quali-quantitative methods. Though we contended some of its overreaching claims, we see the potential of a quali-quantitative approach as a potent way of easily "zooming" between the levels of aggregation. Empirically applying this idea will therefore be of important interest to the study, by adhering to the fundamental ANT dictum to always follow the strongest relation (Latour 2005).

Regarding the redistribution of agency and method in digital research, we emphasised that the analytical mode we employ does not follow a split between empirical and analytical phases, but rather intermeshes the analytical work throughout the process.

Finally we changed focus from what was being represented to the explanatory power of the representation. Following traditional ANT literature we positioned our maps as attempts to stabilise portrayals of the actors we are exploring while providing room for large amount of the actors thus enacted.

2.5 DIGITAL TOOLS

From these more abstract considerations we will now present the type of tools applied in the analysis. The goal is not to give a complete list of tools, but to offer readers unfamiliar with digital methods an overview of the type of tools[16]. The tools are highly technical, so while we do not assume that common readers will gain a full understanding of them from simply reading this text, we will simply try to present their overall workings. Finally we hope that clarifying the technological underpinnings can help to illustrate some of the aforementioned theoretical concepts.

The tools we work with fall into two groups: *collection tools*, which harvest and make available the raw datasets, and *visualisation tools*, which translate the raw data into a visual decodable by humans through spatial arrangements and colouring. To distinguish between these different phases is often a difficult task since collection, analysis and visualisation often melt together because of the redistribution of method (Marres 2012; Madsen 2013:71). Any categorisation of the tools will therefore produce some amounts of categorical "bastards". Also outside of these two categories are host of smaller support tools,such as Excel, CSV converters and list sorters, which structure datasets and act as mediators between different standards.

2.5.1 Software Agents and Data-collection

As argued earlier, Digital traces are often presented as a store of ready-made knowledge awaiting the researcher to stumble upon it and and *harvest* the information. In opposition to this picture we argue that most digital traces live isolated and unknown lives at the bottom of our digital infrastructure, and that their networks and relations rarely appear until they are constructed during the process of collecting, tracing, comparing or visualising. The collection of data and choice of data sources is therefore to an even higher degree than traditional method a decisive parameter for the outcome of research (Anderson 2008).

Two types of approaches are employed in gathering data: *Software agents* and *Application Programming Interfaces* (API).

Software agents[17] or bots as they are often nicknamed by programmers, are small programs or scripts that stroll through digital information extracting information based on some pre-given criteria. In the field of digital methods software agents are often categorised as Crawlers or Scrapers.

Crawlers have also traditionally been used to make patterns in digital information. The main usage of crawlers is thus to make the Internet visible to us through search engines, which archives are built by bots who tirelessly crawl through the pages of the web, extracting and storing possible search words. Whenever we make a search on Google, we are actually not searching the web, but the data extracted by Google's crawlers.

Crawlers used in digital method functions essentially the same way as the bots of Google, except that there is usually only one of two in contrast to the millions of bots controlled by Google. Crawlers always departs from a pre-given point, e.g. a blog, Facebook profile or a journal article. From here they crawl onto related entities based on a number of pre-given criteria for the relations to follow. If one wishes to uncover a blog universe, hyperlinks will be followed. From Facebook profiles the crawler could follow common friends or likes, and from journal articles the crawler would follow citations. Based on these criteria of relations the crawler weaves it web like a spider (*spider* or *ant* being other common nicknames), building layer upon layer until the relations are exhausted or some pre-given limit is reached. In our second analysis we will use the crawler Hypher to collect digital traces from Big Data web sites.

Scrapers, the other category of software agents, is automated operations that extract information from data, or in relation to digital method, automated operations that extract structured (manipulable) data from unstructured data (Marres &

Weltervrede 2012: 4). Digital traces are often characterised by their lack of structure, which has made scrapers central tools in the data collection. Scrapers are e.g. used to search through a number of pages and extract predefined keywords, specific type of data (e.g. location or economic data) or special types of media (e.g. all blue pictures). Where crawlers followers relations (e.g. hyperlinks or citations) already available in the data, scrapers can be seen as constructing new relations that can be combined with other data sources, e.g. by comparing location information in thousands of different reports. In our second analysis we will use the scraper and text analysing tool ANTA (Actor Network Text Analyser).

A final remark on scrapers and crawlers is that while the distinction has been fruitful for categorising tools, the terms increasingly seem to link to practices rather than tools. With the ongoing advancement in software agents' development, there is a tendency to combine the techniques in the same tools. Crawlers seem increasingly to relate and structure data in new ways when crawling the net while scrapers increasingly are crawling from data source to data source when scraping[18]. Rather than making the terms obsolete, the transformation suggests a new usage, where the categorisation is used to describe the practice of crawling or scraping, instead of concrete software tools.

The other group of tools for data harvesting, Application Programming Interfaces (API), are protocols or interfaces through which stored data is made available outside of closed structure while minimising security risk by limiting access and administrative rights. APIs function a little like automated coffee machines where information is made avaiable based on a number of predefined and delimited options. The main difference between the software agents and the APIs is that while the producer of the digital traces has no direct participation in how crawlers and scrapers draw their net, do APIs only deliver the information that the provider has choosen to supply in the first place. Furthermore are the process of data extraction blackboxed the same way as the coffee mixing inside the coffe machine cannot be traced from outside; whenever you have made a request for data you are left waiting until suddenly coffee is pouring out and you will never know for sure if this is in fact *café latte*. Though this limitation can in part be met through extensive documentation, a direct tracing of the collection process will normally be seen as posing a security risk and will therefore be undesirable for the data provider.

These limitations have led to serious critique of the usage of APIs in science, which critiques argues will lead to the accept of black boxed data-collection as scientific foundation (Boyd & Crawford 2011; Marres & Weltevrede 2012:13) and for conflicting with fundamental scientific ideals (Venturini and Latour 2010).

Though data collected through API's is sometimes seen as mediocre, they still play a central role in most digital methods because of the ease of accessing them compared to crawlers. One can also argue that while you cannot ask an API for available data beyond some predefined options, you are at least sure to receive what you asked for unlike a crawler that can be configured for almost anything, but often responds with highly unexpected results.

In our first analysis we use Google Trend and the tool Google Autocomplete from the Digital Methods Initiative whcih are both simple visual interfaces built on top of a Google API.

2.5.2 Visualisation tools and Algorithms

One very distinct ANT contribution to the STS field was the discussion on visualisations as powerful actors (e.g. Latour and Woolgar 1979; Latour 1986; Latour 1987). Under the headlines of inscription devices and immutable mobiles ANT scholars has attempted to remove the innocent understanding of visualisations as simple representations of reality.

In the final chapter of this thesis we will follow up on this discussion in a new digital reality. But already now it is important to stress that we view visualisation tools as anything but innocent and that we therefore find it important to dedicate some space to these tools and their role in digital methods.

This thesis applies the open source visualisation tool *Gephi* developed primarily by the Latour led institute Medialab. The choice of Gephi were in part based on the ethical and efficiency advantages of open source software, and Gephi's focus on rapid visualisation through *WYSIWYG* navigation (Bastian et al.: 361). Another important argument is however its direct relation to central ANT scholars, evident already in the description of Gephi as a *"network exploration and manipulation software"* (ibid.) rather than simply a visualisation tool. Gephi is in other words not built to perform statistical and mathematical evaluations of reality like most visualisation tools, but as a tool to 'explore' and 'manipulate' reality.

To illustrate what we mean by exploring data, and also to provide an example on the sort of situated co-fabrication that we touched on in 2.4, we will briefly show how data is made legible in Gephi.

In Gephi the starting point of a visualisation after importing a data set is little more than a bundle of equal size grey dots placed in a quadric square; utterly non descript and providing little information (as seen in figure 2). While Anderson argued that the numbers would speak for themselves with enough data (2008), what we see here is quite the opposite situation: get enough data, and the numbers say nothing at all - we simply get noise.

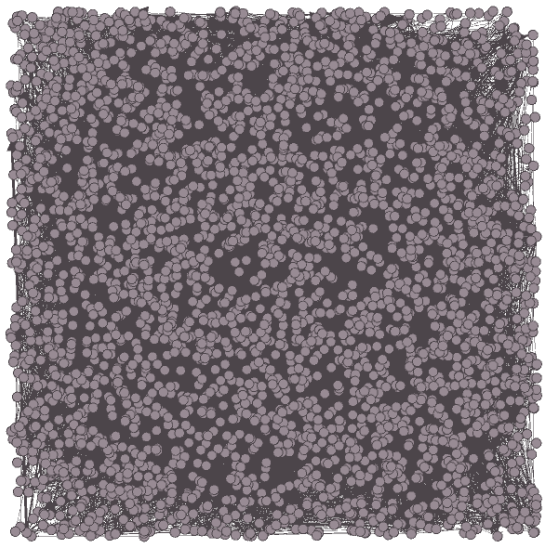


Figure 2: A unmanipulated Gephi visualization.

This serves as a stern reminder that reality does not present an obvious interpretation and that you are always a co-fabricating part of producing meaning in ANT (Stengers 1997).

The sense-making first starts with the manipulative techniques, which are ways of visualising *patterns* through spatialization, grouping and colouring. In this process we make use of four techniques, which we will outline below:

1) Layouts: Layout consists of sets of spatialization algorithms. Though Gephi comes pre bundled with a number and more can be added, the very backbone of the layout functionality (and to some degree of entire Gephi) is the *Force Atlas*[19] algorithm used to spread nodes based on their relation. Force Atlas is described as a force directed algorithm, which means that it simulates a physical system in order to spatialize the network (Jacomy et al. 2011:4). By adjusting Newtonian variables such as gravity and repulsion, the networks are spread out based on the nodes' interrelatedness, turning spatial proximity into a representation of coherence[20]. We use this algorithm to separate nodal points from outliers and construct clusters of related entities, whether they are websites, names of researchers or scholarly articles.

2) Ranking and **Partition:** The simplest but also of the two most often-used manipulation techniques. The ranking or partition function is used to respectively add size or colour to nodes (or edges) based on a variable e.g. type, number of occurrences, relations to other nodes or its grouping with other nodes. Running these allows us to discern which nodes are central with regard to particular parameters. We can e.g. show the amount of inbound associations by relative size, and the degree to which particular nodes belong to a community by colour.

3) Filters: The filter techniques are as the name indicates techniques for filtering out nodes and edges based on a range of different criteria such as occurrence count or ingoing relations. As our maps often have many thousands of data points, we commonly filter out any points that do not reach a minimum threshold, in order to separate central nodes from noise.

4) Statistics: The statistics techniques are integrated statically computations derived especially from traditional Social Network analysis (see e.g. Carrington & Scott 2011). In our work we primarily make use of *Modularity*, which algoritmically identify clusters based the relation of the nodes and *eigenvector* which caculating nodes connections to other well-connected ('influential') nodes. Once run, these statistical measures become a possible parameter for any of the above-mentioned functions.

The visualisation of data is an important step in transforming raw data into meaningful patterns, and our description of its functions should make it clear that rather than providing unmediated, objective accounts, these visualisation tools are actors on their own accounts. They work as mediators employing a chain of inscriptive and calculative devices to translate the network data into meaningful, decodable text accessible for sense making.

2.5.3 Summary

In our last methodological chapter we zoomed in on the actor group of digital tools that shape our methodological gaze. We here distinguish between data collection tools, used to harvest the raw data and the visualisation tools, translating the raw data into human decodable visuals. In the first group we presented how bots (primarily crawlers and scrappers) and APIs offers ways of automating and scaling the data extraction and how they each hold different strengths and weakness in the reality they depict.

In the second group, we introduced and discussed the vocabulary surrounding the use of visualisation such as layout algorithms, ranking and partition, filters and statistics. By doing so we addressed the earlier stated question of disassembling the digital tools in order to question how they produce meaning. We will return to this point during our analyses.

2.6 LITERATURE REVIEW: SOCIOLOGY OF EXPECTATIONS

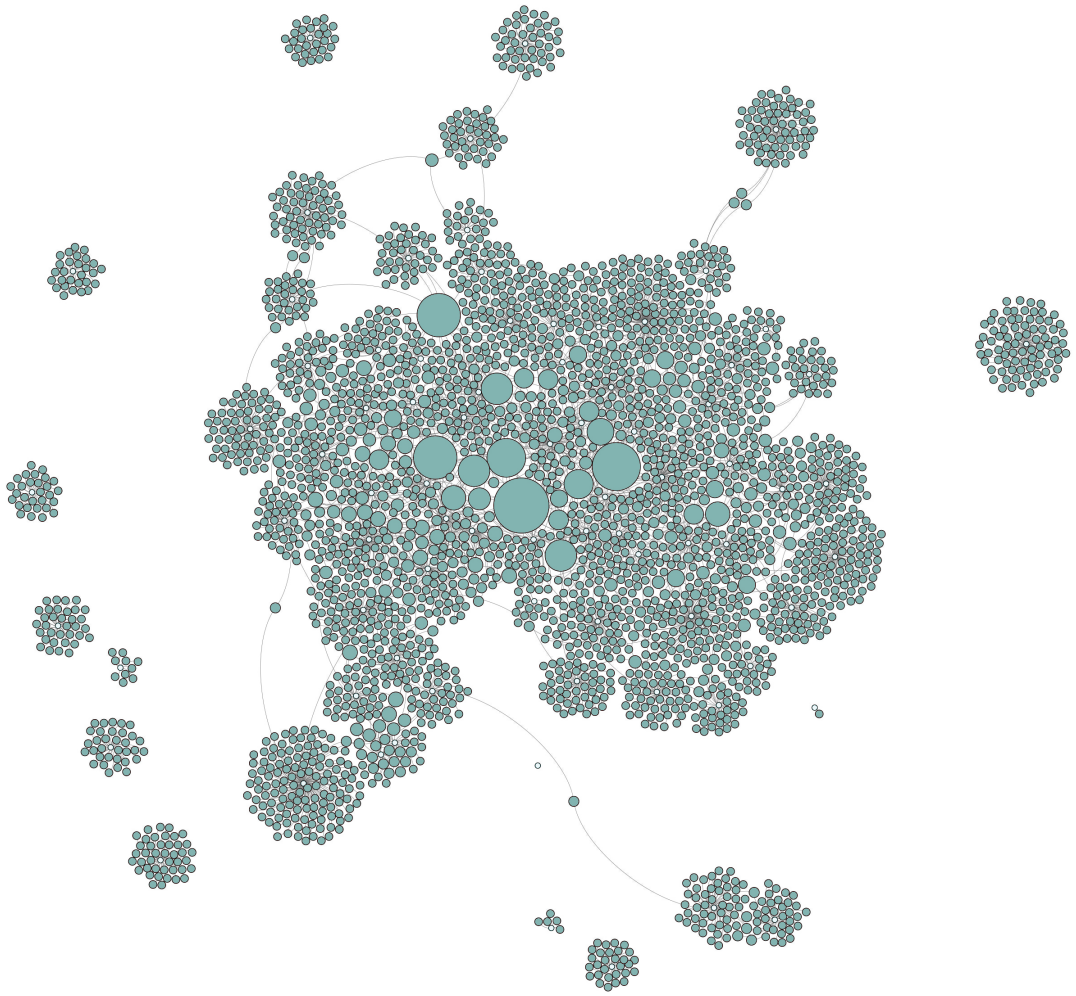
In the preceding chapter we have discussed methodological questions both generally and in specific relation to digital methods. In this chapter we will introduce some theoretical concepts from the sociology of expectations through a literature review.

The study of expectations is by no means unprecedented in the STS field. During the last two decades, this interest has especially clustered around what has been described as the *sociology of expectations* (Borup et al. 2006). Though hardly stabilised as a research field, due to largely intermittent publication, a network of referrals between authors, articles, institutions and journals has emerged during the last 10 years.

In the following we will explore this network through a literature review to uncover established knowledge on expectations, while simultaneously identifying limitations and a possible contribution on our part.

To guide our reading, we have completed a scientometric analysis of the citation network (see protocol C in appendix). Through this analysis 179 articles explicitly concerned with the sociology of expectations were selected from two online repositories of academic journals, Scopus and Web of Science. From these, a crawler identified 3,578 references, from which we constructed the map in figure 3. This preliminary analysis will serve as an exploratory tool to gain an overview of the field - an assistance that is particularly demanded when engaging with an emerging and not yet formalised field.

Figure 3: Scientometric analysis of the Sociology of Expectations.



The green circles represent citations and the size its relevance which is given based on the number of articles sharing this citation. While one could point to other criteria of relevance, we deemed that the most frequently cited literature would also be expected to play a central role in shaping the field and therefore a reasonable way of following the strongest relations (Elgaard 2005; Latour et. al. 2012).

Several important observations can be drawn from the mapping: firstly, that there is in fact a core group of articles that share a large number of references (a finding also confirmed by our subsequent reading). These will be taken to be the central texts in the sociology of expectations. Around this core floats a number of small clusters, indicating that the field has produced a fringe of related research beyond its immediate core.

From the map we then select the 12 most cited articles and use these as the starting point for our exploration of the field. These articles are then carefully read forming the base of our review. They are not seen as exhaustive, but rather as a starting point from which we followed reference to other promising works contesting or expounding on central concepts, assumptions or methods. We hereby try to assure that our scientometric preliminary exploration does not limit our reading by fencing in small groups of articles, but opens up the field with a number of relevant starting points.

2.6.1 Delineating the Field

In the shaping of the field a number of influential nodal points emerges: two special editions of the journal "Technology Analysis and Strategic Management", the journals "Social studies of science" and "Science, Technology & Human Values", authors such as Harro Van Lente, Mike Michaels, Andrew Webster and Nik Brown and several universities, in particularly Goldsmiths University.

The tradition is a clear offspring of Science and Technological studies (STS), but varies in the degree to which they adopt the epistemological and ontological leanings of actor-network theory. They are however unified in their approach to studying the emergence of technologies through the lens of expectation as a generative force, as well as the viewing actions such as prospecting and forecasting as constructionist "world-building". Using expectations as an analytical entry point and operationalisation of future-oriented communication *"shifts the analytical gaze from looking into the future to looking at the future as a sociological phenomenon in its own right"* (Brown & Michael 2003:2 our emphasis).

On a general level the term *expectations* is an attempt to grasp in a broad sense how conceptions of the future are enacted in the present. But expectations also serve as an umbrella term in many articles, an overarching category covering a number of different discussions with interests in future-oriented communication and their influence on technological development.

In these discussions we will focus on two subfields: hypes and visions. Below we will introduce some general characteristics of the two before we explore some of their central assumptions, arguments and findings.

2.6.2 Hype

Slightly simplified, hype studies can be said to be the study of expectation dynamics, answering how expectations to a technology gain and lose momentum (Van Lente et al. 2013). Most studies follow spectators' interests towards a given technological development as measurement for the degree of hype. The most common object of investigation in hype studies is therefore quantitative studies of the saliency of a given term for the technologically developing phenomenon, through e.g. studies of search traffic or media coverage (see e.g. Järvenpää & Mäkinen 2008; Jun 2011; Jun 2012).

An interesting discussion in the field relates to how to *measure* expectations. Often metrics such as the volume of either search traffic (Jun 2011), patent applications (Ruef & Markard 2010) or newspaper articles (Järvenpää & Mäkinen 2008) are seen as equal to the level of optimism - a claim we will contend in chapter 4.

To some degree these studies look at the constitutive effect of expectations, but often only in numerical terms, for example how the salience surrounding technology in different media arenas leads to increased funding. Thus, they often overlook the intricacies of how expectations enable broader techno-social arrangements beyond mere metrics and how a particular *form* of expectation shapes the development of technology in a particular direction.

Another common feature of many hype studies is to look at hype as something distinct from the underlying potential of the technology - positing that a realistic or in some way "truer" level of expectations exist, but that neophilia, hysteria and bandwagonism distorts this (Van Lente et al. 2013:2). This idea seems to be closely related to Gartner's Hype Cycle, a consultancy tool to diagnose the level of hype in order to predict the technological potential (Järvenpää & Mäkinen 2008, see also figure 4). This model is also the origin of the idea of a standard curve of expectations where early promise leads to very high expectations while later obstacles encountered in the development of new technology sends the expectations plunging (Järvenpää & Mäkinen 2008,Van Lente et al. 2013:2, Ruef and Markard 2010:1, Borup et al. 2006). While both widely disseminated and heavily criticised (Steinert & Leifer 2010), the Hype Cycle has stayed a focal point of the discussion for most hype studies.

Though many of the central actors criticise the Gartner model for its oversimplification and heavy reductionism (Steinert & Leifer 2010), they partly too seem to fall into the trap of reductionism (Pinch & Bijker in Hedgecoe & Martin 2003:330), exemplified by statements such as: "*Only rarely are the initially high-rising expectations met*" (Van Lente et al. 2013:2). While such claims are not necessarily false, their broad generalisations of relatively narrow and limited casework are problematic since they in their search for regularities fail to take into account the unique effects of how *specific* expectations fuel *particular* developments.

This tendency is not only limited to quantitative studies, but also qualitative reflections on our dynamic and changing relations to specific expectations. A clear cut example of such search

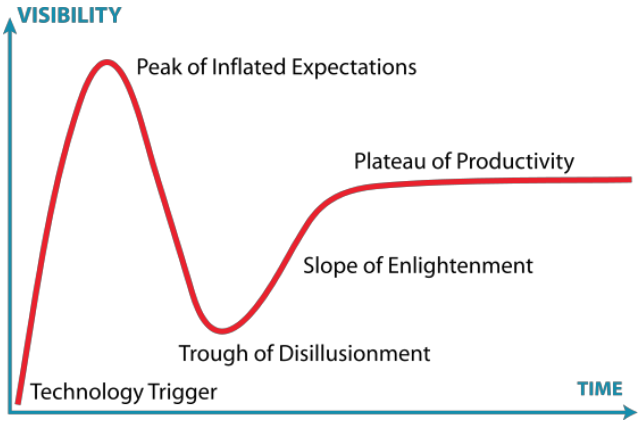


Figure 4: Depiction of Gartner's hype cycle (Kemp 2007).

for regularity in the dynamics of expectations are seen in Brown and Michaels, who concludes an otherwise great article with the highly generalised claim that actors always repeat their mistakes in spite of prior experience (Brown & Michael 2003).

Another example would be the claim of Hedecoe and Martin, that technologies in their early states have various expectations to their future contribution, but which are crystallise and condense into a single vision for their application later in their development (Hedgecoe & Martin 2003:330). While we will explore the same theme in chapter 6, we will not prematurely conclude that they follow a pattern from multiple to singular accounts.

2.6.3 Visions and Shared Expectations

Another discussion in the sociology of expectations is that on visions[21]. Briefly put, visions can be seen as *"internally coherent pictures of alternative future worlds"* (Eames et al. 2006) or "*expectations shared by multiple actors*" (Merkerk & Robinson 2006:416). In these studies, visions are seen as evolving from the negotiations of expectations between different actors, aligning future expectations and building generalised coherent depictions of the world to be. Visions therefore constitute a particular class of expectation which both project and anticipate how the future might emerge, and provide a strategic framework for actors as they attempt to construct particular socio-technical networks (Hedgecoe & Martin 2003: 331).

The study of Visions are thereby a subfield of expectations studies focusing on how expectations are bundled into visions and how these visions co-construct particular socio technical developments by mobilising perspectives on the future as strategic premises for the present. The vision can then function as an ordering device, rearranging the assemblage of actors and offering them a place in a possible future similarly to the concept of *Actor-Worlds* (Callon 1987). Another way visions can function strategically is by condensing and translating promises to requirements (Van Lente 1993, Brown 2000). In contrast to hype studies, the vision studies relies mainly on qualitative studies.

2.6.4 The Function of Visions

On the most general level the discussion of visions can be seen as growing from the idea of expectation as a constitutive and performative aspect of the development of new technologies (e.g. Van Lente et al. 2013:1; Borup et al. 2006:289; Brown 2003; Van Lente & Rip 1998:225). In contrast to the hype discussion, the expectations underlying the visions are in other words just as real and important actors as the technology itself and not something separated from some underlying innovation (Borup et al. 2006:289). Not only do visions actively interfere in the innovation, the dual processes of prospecting and producing a technology are inseparable from each other, and trying to detach one from the other is deemed impossible (Brown 2003:17).

Visions are seen as active participants in the shaping of the innovation in a number of ways. Several authors points to the function of expectations as a mobilising factor, building momentum for the innovation through attracting attention and resources to the field of innovation (Ruef & Markard 2010:1; Brown 2003; Geels & smit 2000). Analysis has also shown how expectations build legitimacy and allow for decision-making through reduction of uncertainty (Van Lente 1993). On a more general level visions are seen as fundamental in the creation of new socio-technical networks due to their ability to stabilise long term development through the association of actors such as institutions, investors and researchers (Ruef & Markard 2010:1; Hedgecoe & Martin 2003). Visions are even shown to be strong enough to enable innovation long after the hype has died (Ruef & Markard 2010). Hedgecoe and Martin have broadened the scope of actors by showing how expectations also construct ethical and legal order to enable the expected innovations (2003:329). In our study we also wish to follow this approach, searching for a broad scope of possible effects expectations can play.

Not only voluntary actors are linked by visions. Through tying together the different actors, the visions can be said to function as coordination device (Van Lente & Rip 1998; Borup et al. 2006; Ruef & Markard 2010). This becomes apparent in e.g. the negotiation of the problem the innovation are articulated as the solution to, where actors are tied to each other through specific expectations to the future (Hedgecoe and Martin 2003:329). Visions can here also function as an enabler of strategic alliances, contribution to the building of coalitions (ibid.). So as visions and expectations develop, new possible positions are enabled in the network, and conversely, visions must change to accommodate new members joining an organisation. The visions also determine the way different research options can be pursued: Visions provide a framework within which the future shape and application of a technology is constructed, as they act as both an aid for decision-making and a focus for the mobilisation of actors and resources (Hedgecoe & Martin, 2003). The visions hereby contribute to agenda building and to the transformation of requirement from the problem and into the innovation processes, aligning resources to the development of the technology (Van Lente 1993; Van Lente and Rip 1998 Hedgecoe and Martin 2003:330).

2.6.5 The dark side of Expectations

From this general outline of the field appears a tendency to focus on the positive effects of the expectations. A few articles do, however, take up some of the more problematic and dark sides of expectations, especially the risk of hype and overestimated expectations. In the perspective of Brown and Michael this risk is tightly connected to the nature of future visions as: *"potent mobilisers, but also fragile constructions"* (Brown & Michael 2003). This fragility is problematic since enduring disappointment and destruction of

visions constructs "a bad reputation", hurting the long term innovation while undermining the assemblage of new expectations into visions (Brown et al. 2003; Ruef & Markard 2010:2; Van Lente et al. 2013)[22]. Though dangerous, other studies have shown that not all disappointments carry a potential risk and how underlying hype can often die out without affecting the more general vision (Ruef & Markard 2010: 319; Van Lente et al. 2013). This also raises the question of whether interest (measured by media salience) can be taken as a measure for the level of expectations, or if it just gauges the novelty value.

2.6.6 Entering the Discussion

The literature summarised above paints a broad picture of how expectations, visions and hypes shape and take part in developing new technologies.

In the following we wish to take a more critical stance against the presented thoughts, as well as highlight the shortcomings and limitations that our exploration of Big Data can contribute with new perspectives to.

Overall, the empirical foundation, and especially its quantitative parts, often appears rather sparse. As Lazar et. al concludes in a recent heavily cited article, has the application of emerging digital data sources in social sciences has fallen behind competing fields of science (Lazer et al. 2009:721). A similar pattern is found in the study of hypes. Most of the times only qualitative data supports the claims of the analysis (e.g. Brown & Michael 2003; Hedgecoe & Martin 2003) and in the cases where quantitative data is used to support the case most often only one source of data is used in the argumentation (eg. Järvenpää & Mäkinen 2008; Van Lente et al. 2013). In a few cases the lack of ability to produce supportive data appears to push authors to stretch the claims of their data. An example of such would be (Jun 2012) who ends up interpreting search traffic as fundamentally equal to consumer interests; two variables which without a doubt are related, but which hardly can be said to be identical (Ruef & Markard 2010; Van Lente et al 2013:4).

In response to Lazer's critique we instead suggest to follow some of the recent work done in *social network analysis*. Here a number of scientists have suggested basing science more on multi-dimensional measurements and the application of different communication channels when interpreting social phenomena (Lehmann 2012:66; Blok et al. 2011); a proposal that is in line with our suggested quali-quantitative approach (Crawford 2013). Through this approach we wish to push the current empirical limitation of the sociology of expectation, exploring how analysis based on multiple data sources can contribute to new and deeper understandings of the expectation dynamics (hype) and the construction of visions. In this way, we hope to improve on the qualitative lack of detail in hype studies and the quantitative lack of overview in vision studies.

Another important critique of especially hype studies addresses the tendency to see expectations as a distortion of a "true" level of potential of a particular technological development[23]. This conception of hype appears to come from the heavily criticised but ubiquitous Gartner Hype Cycle - a theory in which our optic most fittingly has been described as a *"folk-theory par excellence"* (Rip in Pollock & Williams 2010: 530). We find it difficult to believe that expectations are mere irrational phantasms, extraneous to the "reality" of the technology. Instead we see the ability of technology to attract attention as a constitutive and intrinsic parameter of technological development, and the expectation dynamic as a subject in desperate need of more empirical exploration of the variety of effects expectations have (Pollock & Williams 2010:530).

For this reason we also wish to simultaneously explore the interrelatedness of hype, expectations and visions. Though main practitioners in the sociology of expectations have asked for more integrated approaches focusing more broadly on the variety of effects and dynamics resolving around expectations, a schism of quantitative hype and qualitative visions seems to be prevailing in the empirical work. Also there has been a tendency to singularise the phenomenon, discussing *the* vision or *the* hype.

We want to stretch the empirical scope to work concurrently through the perspective of hype and that of visions, and how these differ between different groups of actors. More concretely, we propose to study how the idea of Big Data exists as different visions and hypes, how they change over time and how they enrol different actors. What we propose is not only to look at how a vision can drive expectations, but also how visions change over time. This is a cyclical process: visions generate expectations that propel an idea forward, that interests, enrols and mobilises, but at the same time the movement across the social fabric also changes the vision itself. Concrete analyses and how we bundle them together to address these questions will be presented in chapter 3.

2.6.7 Summary

We have dived into the sociology of expectation through a preliminary scientometric analysis identifying 12 heavily cited articles as our starting points. Based on this we delineated the field based on central authors, journals, universities, its origin in STS and partly ANT and a shared interest for looking at how the future was enacted rather than trying to predict the future.

From this basis we identified two sub-fields, Hype and vision studies. Hype, the study of expectation dynamics, measured interest as the quantitative saliency in media, patents or inquiries on the developing technological phenomena. In contrast the studies of visions were primarily founded on qualitative studies addressing how expectations through negotiation are merged into shared visions of the future. In both areas we encountered an on-going discussion of the ontological status of expectations as something detached from an underlying technological development. In this discussion we position our study in opposition to the idea of expectations detangled from reality.

Finally we took a more critical stance in search for holes in current litterature for us to fill out. Here we identified a need for more multi sourced and integrated methodological approaches (such as quali-quantitative methods) as well as a need for exploring the interrelatedness of hype, expectations and visions. In particular we wanted to bridge the highly qualitative approach of vision studies with more quantitative hype studies in order to trace both the global translation of Big Data as the buzzword du jour and the local and particular effects it had in reassembling socio-technical arrangements

Leaving the methodological level, we now proceed to operationalisation and present our research design.

Notes

1. Allthough Latour (2004) would prefer the term infra-language.

2. We will return to this discussion in chapter 2.2

3. Latour sometimes draws a difference between actant and actor with regard to their figuration (2005:71). We will not use this distinction, but only use the term actor.

4. Which, on a side note, is highly advantageous for us since the semantic singularity of buzzwords allows us to track them much easier

5. A later synonym for network. Other examples are arrangement or collective. We will use them interchangeably to the same effect.

6. E.g. are the phases of *problemisation* completely omitted, while *mobilisation* is used interchangeably with enrolment throughout the study.

7. For a more thorough account see Ratner 2009 and Ratner 2012.

8. Moore's law

9. In contrast to earlier search engines that worked by manually categorising every webpage into a taxonomical directory. These engines were terrible, might we add.

10. It should be noted that ANT normally advocates following *all* types of associations. Google, though, follows only this one type, that is however powerful enough to warrant serious attention. We will try to remediate the limitations of this by employing multiple shifts to a qualitative level to find other associations. More on this in chapter 3.

11. This is - broadly outlined - the publicly known principle behind the baseline algorithm. The ranking is thought to be influenced by myriad other factors, including time, location and search history.

12. On a side note, we personally find that the problem with search algorithms creating homogeneity is not so much the slight personalisation we each get through our search results, but rather that nobody ever looks beyond the first few pages of results. This creates a strong centering on a few central sources, leaving little room for the outliers.

13. A point brazenly stated by the search page itself when it declares to have found millions of results and offhandedly mentioning the fraction of a second this operation took.

14. This is luckily surprisingly easy: one needs to look no further than the links provided beneath the aggregate count. We will later elaborate further on this granularity and the possibilities of zooming.

15. For similar discussions see e.g. (Latour 2005:88ff)

16. For a detailed walkthrough of our usage of the different tools, see appendix 9.1.

17. We purposely use the vague term 'Software agent' to capture a broad range of agent based tools (Nwana 1996).

18. An example of a crawler emulating scrapers could be the recently released tool Hypher which is discussed further in this project, see protocol D, appendix. An example of a scraper emulating the autonomous movement of the crawler is ScholarScape, see http://github.com/medialab/scholarScape.

19. Force atlas exists in two versions. We have decided not to distinguish between the two since the underlying techniques are the same and Force atlas II therefore should be thought of as an optimisation.

20. For an overview of the theoretical assumptions of the Force Atlas algorithm see Jacomy et al. 2011.

21. What we here define as visions also includes the related concepts of *frames* and to some degree *overreaching expectations, which* are also sporadically referred to in the field of sociology of expectations (e.g. Van Lente 1993; Ruef and Markard 2010; Van lente et al. 2013). We have also included a number of authors whom uses expectations as a synonym for what is described here as visions.

22. As articulated by Brown and Michael the main problems is that entrepreneurs only carry minor risk related to disappointment, while investors, patients and public policy makers are hit hard when e.g. life saving medicine shows not to work as expected (Brown & Michael 2003; Brown 2003).

23. It is important to stress that this critique is not new and has also been put forward in the sociology of expectation (e.g. Van Lente et al. 2013)

3 RESEARCH DESIGN

IN THE FOLLOWING WE WILL INTRODUCE OUR ANALYTICAL APPROACH THROUGH A OUTLINE OF THE SCOPE AND AMBITION OF OUR STUDY. WE WISH TO ELABORATE ON THE REQUIREMENTS OUR RESEARCH DESIGN NEEDS TO FULLFIL AND HOW OUR DIFFERENT EXPERIMENTS ATTEMPTS TO FULLFIL THESE REQUIREMENTS.

In the following we will outline the general scope and ambition of our study. We will then elaborate on the requirements our research design needs to fulfil in order to attain these ambitions in the form of three considerations. Lastly we will outline the parts of our analysis and how they will be conducted. We draw on presented methods and concepts from both the field of the sociology of expectations and digital methods to construct our own particular method assemblage (Law 2004). First, we will introduce three preliminary considerations that we find must be embraced if the limitations in the field of sociology of expectation presented in previous chapter are to be overcome.

3.1 PRELIMINARY CONSIDERATIONS

In consideration of the limitations we conceive in the current sociology of expectations, we propose to study the evolution of Big Data through a *large scale, longitudinal quali-quantitative approach*. We will look at how the concept Big Data changes and moves through society - how it is translated. To do that we will focus on a particular chain of translations divided into three analytical phases: Firstly, how the use of the term Big Data (and related synonyms) grows over time, secondly, how this term enrols a myriad of actors (organisations, technologies, field terms etc.) and thirdly, how these expectations are translated into generalised coherent depictions of worlds to be. These analytical phases will be further specified in the end of this chapter.

Our ambition is in other words to provide a thoroughly empirical account of the chain of translations from *hype* over *expectations* to *visions* with the goal of tracking how this chain propels Big Data from the obscurity of niche interest into the current massive mainstream attention. In the process we wish to map not only the growth in the actor-network (Latour 1987) around Big Data, but also zoom in on the semantic context of expectations in which Big Data resides. We will study how expectations serve to drive forward this expansion, and how this expansion in itself begets new translations of the generalised expectations to Big Data.

3.1.1 Consideration 1: Semantic Multiplicity

Our research design must be able to handle the semantic multiplicity associated with Big Data. The use of digital methods makes this all the more important, since unlike humans, software agents and APIs require exact semantic specifications to function because of their inability to see even obvious semantic correlations[1]. Therefore, a study done without consideration of e.g. abbreviations, shorthand, slang, synonyms and even misspellings[2] remains blind to these facets of the empirical. As a consequence we must preliminarily track the central terms of Big Data and include these in our research designs.

To some degree the knowledge of these terms comes through the familiarity with Big Data that we have established in the course of our research. Another important source to establish this semantic multiplicity will be didactic articles created by "experts" presenting key terms on Big Data. Lastly, we will engage a technology in finding these correlations for us through the tool Google Autocomplete[3]. Based on the Google Search API, the tool offers ways to extract a series of related keywords. While the exact algorithm underlying it is unknown, we can assume that it is derived from the same type of rationality that governs search ranking. It thus gives us insight into not only related keywords that may serve as synonyms – i.e. *data analytics, data science*, but also terms that indicate behaviour and relations such as searches for *Big Data jobs, Big Data conferences and Big Data services.*

3.1.2. Consideration 2: Spatial Multiplicity

The second consideration is that our study must span a multitude of worlds (Mol 2002; Law 2004; Ratner 2012). It must be able to handle a large range of actors and the network they may form. The expectations levelled at Big Data are hardly singular, and to be able to account for them we must set up our research design so it addresses not just a single set of actors, e.g. data scientist or technological entrepreneurs, but account for their relations and how Big Data is translated as it moves between different settings. Only by layering the different ontologies that actors construct around Big Data can we hope to achieve a comprehensive account. Through the successive use of different methods that each enact a particular assemblage, we will juxtapose these findings in order to give representation to a large range of actors, similar to the approach advocated for under the heading of second-degree objectivity (Latour et al. 2012, see also chapter 2.4.5).

3.1.3 Consideration 3: Temporal multiplicity

In order to encompass the evolution of the term Big Data and the translations it undergoes, we must of course set up our study to follow this progression over a certain chronological span. But this timeline is not the only aspect of temporality we should account for. As we are tracing expectations, we are also tracing a range of imagined futures or visions. In line with the sociology of expectations, we do this not to predict a singular future, but rather to account for the enactment of multiple futures and how they set up strategic conditions for the present (see chapter 2.6).

Tracing the history of expectations to Big Data, is thus tracing a range of futures as they were enacted in the past (Brown & Michaels 2003). So we study futures past and must be able to account for not only the form Big Data takes in chronological key points, but also the range of futures portrayed and how they change over time. In this way our research design should encompass not a plurality, but a multiplicity of temporalities.

To do so we need digital methods that allow us to trace an archive over time, and also zoom in on key periods to see what expectations were enacted at a given time. As not all of our available methods can account for temporal parameters, we will have to construct a coherent chronology as a patchwork of snapshots anchored in a timeline. By using the quali-quantitative approach that the traceability of digital data affords us, we are able to zoom in on nodal points and explore how expectations for Big Data are translated into visions - unfolding temporalities from discrete points on a timeline.

As mentioned earlier, these three considerations require that we construct our analytical assemblage as a patchwork, as a bundle of methods each of which enacts particular realities (Law 2004:42). How we hope to assemblage these three considerations in our bundle of methods will be delineated in the following section where we put forward three distinct phases, their methodological approach and epistemological interest and interrelations.

3.2. ANALYTICAL PHASES

Our research design proposes 3 phases that use largely technical means of tracing the translation of Big Data through first *time*, then *space* (understood as metaphor for its geographical expansion and the enrolment of actors) and finally *meaning*, where we attempt to tie the findings from our inscription devices into a set of coherent visions for the emergence and evolution of Big Data. Following the basic ANT premise of a flat ontology introduced in the previous section there are not differences in the levels of analysis - we do not study the spread of the *term* first, and then subsequently jump to another order to look at the expectations and visions: the spread of the term is in itself a way to look at expectations in the form of hype (Ruef & Markard 2010, Van Lente et al. 2013) - albeit an approach that at this time could benefit from a more elaborate operationalization.

These phases are constructed as an analytical relay, where the output of each phase generates input for further analysis in the next. Thus, in the first phase we look at how Big Data is translated through time to pinpoint nodal points in the chronology for further inspection. The major contribution from this chapter will be a rough timeline of the interest in the subject. Its peaks and valleys (as well as the geolocative parameters that emerge as a secondary result) will provide us with a inkling of when and where to continue the investigation.

The second phase will use the periods and moments identified and zoom in on how Big Data is translated through socio-spatial enrolment of actors, generating ever-larger multiplicities of expectations to the possible arrangements of Big Data. Taking the expectations in this regard as interessement devices, we will look at how the contestation and negotiation between actors serves as a strategic factor in the evolution of Big Data as a billion dollar industry (Van Lente & Rip 1998:245). The contribution from this chapter is an outline of when which actors are involved, and where the jumping points are from one area to another.

In our third and last phase, the meaning vested in Big Data will be the subject of our investigations through a narrative reconstruction of the visions enacted of Big Data. We will in other words identify how these expectations tie a host of actors together in a vision of a world to be - analogous to what Callon called an actor world (Callon 1987). Here we follow how the actors attempt to singularise the multiple expectations to the future as a vision. These visions' constitutive effect on technological development is the subject of the sociology of expectations. Although a process, which is never finalised or completed with a multiplicity of visions always coexisting, we trace how specific moments of translations, actors and relations are stabilised in a dominant future vision.

This chapter will tie the results from the different phases into a meaningful whole. We find this phase particularly important, since the dominant mode in the digital methods field has often been to *let the maps talk for themselves,* an approach we find to be unfitting. In this role, the cartographic technique's major explanatory power has been stabilisers of claims, as impenetrable inscription devices black boxing their hinterland (Law 2004). We propose instead to let them play the role of narrative devices, as building blocks around which meaning can be guided rather than as explanations unto themselves. Maps only tell you where to look. By frankly laying open the practices of their construction and denying them ultimate explanation power, we hope to both refine their use as argumentative political tools and throw light on the ways in which these digital tools can benefit from further development.

3.2.1 Translation through Time: Establishing the Chronology of Hype

First we will look at the emergence of the term and the hype surrounding it. This is a quite simple analysis, built by highly quantitative and linear tools from Google's services. It will produce a general timeline and also present us with secondary data on geographical distribution, related keywords and such. Additionally, we will be able to contrast these highly linear findings with the traditional hype cycle and contest it's predictive power. This analysis gives our research a digital grounding, lets us scope out the magnitude of the hype and pinpoint chronological points for further investigation. This is about expectations as hype.

The first step is to expand the number of search terms from simply Big Data to related synonyms and qualifiers. We do this through the tool Google Autocomplete previously introduced, querying Google search API for it's autocomplete function. The autocomplete function is nominally used to fill in the search bar for users of Google's search bar to save them time. The Autocomplete function tries to predict what users are searching for before they write it in full on the basis of historically co-related search terms. Here we repurpose (Rogers 2009) it to find related terms of interest[4]. Further related terms are also retrieved from Google trends, though these contain more specific information, which we will delve into later.

Armed with this list we conduct our experiments. Operationally, our tools for this phase will be quite directly interfacing with Google's search rankings and the service Google Trends. This is an attempt to gauge the chronological development in interest in Big Data from what we define as both a supply- and demand side.

The first tool gauges the "supply" of information on Big Data. Here we limit the time series for Google searches for Big Data and the earlier found related terms to year periods and conduct a search for each period, noting the number of returned results for each period. These numbers are transferred to an excel worksheet where we calculate yearly growth rates and plot the results into graphs. The results are then compared between search terms. This gives us a timeline in the number of mentions actors make of the term Big Data: websites dedicated to the subject, news articles, blog posts, discussions in online forums etc. We do not discern between the different sources at this point, but take it as an expression of the extent actors are supplying information about Big Data over time.

These results say nothing about how often actors are querying for information on the term: this is the amount of information supplied. To gauge the demand for such information we turn to Google Trends, a graphical interface drawing on Google's Search API that measures the amount of search traffic for a given term over time.

The service is closed - data cannot be exported and absolute numbers are not disclosed, with the relative volume of search traffic indexed to peak volume. This output is presented as a graph. So while we can only use this tool heuristically, it allows us to compare relative growth patterns and displays the demand for information on Big Data based on number of searches conducted.

This means we can compare both overall volume and the prevalence of different search terms between providers of information (number of sites) with the consumers of information (number of searches). From here we can analyse whether there is incongruence between the developments in the two sides. Additionally, this tool displays the geographic distribution for different search queries over time, providing us with geographical insight in the evolving hype.

From this we have a rough timeline, which will be the outset for the next phase - and a host of secondary observations, associations and translations that will feed into the narrative reconstruction.

3.2.2 Translation through Space: Mapping the Enrolment of Actors

In the second phase, we will look at how networks of references tie different actors to search terms. To the extent that the tools allow us to do so, we will seek to map the development of networks formed over time. This will show us in what context the term originates and how it is translated from there. From this analysis we can pinpoint which actors are assembled in Big Data at different times. This phase regards expectations as propellers of technical development and enroller of actors.

Operationally, our main tools for this phase are crawlers, scrapers and semantic text analysers.

First we will use the tool Hypher[5] by Science Po Media Lab, developed to make web corpuses available for social scientists (See protocol D). Hypher is a crawler, which crawls web pages for the aforementioned hyperlinks[6]. This tool can be configured to look for single types of sources: news outlets, blogs, Internet forums and regular websites. We will maintain an agnostic approach as to whether a particular source represents a particular sphere of actors. As mentioned in 2.3, we find the modus operandi of this form of crawler to be remarkably close to the methodical ambitions put forward by ANT (Latour 2011).

Hypher departs its crawls from a predefined list of addresses and records every single hyperlinked reference on these pages. It then scans the pages linked to in these references for their hyperlinks, follow each of these and repeat the entire exercise for a predefined number of rounds (the crawls *'depth'*). By tracing the sum of these references back and forth between sites, hubs emerge based on the number of links they get in the network of references. In this way, we can discern both the relative importance given to a site by the arrangement of links it resides in, and put it in a semantic context based on its neighbours in the hub.

Secondly we will zoom in on the semantic vocabulary associated to Big Data through the conduct of textual analysis (see protocol E). Operationally we will run articles, reports and websites identified through the previous tools through a text analysis tool - ANTA. ANTA uses word co-occurrence analysis to find the central actors, institutions, keywords and relations in texts collections. From this data we will ask as to what interessement it offers and how these objects presents possibilities of enrolling different actors and (Callon 1986).

As an extension of Hypher and ANTA we finally conduct a scientometric analysis of two academic databases of papers and journals, *Scopus* and *Web of Knowledge*, that offers an API-based interface for extracting citations as references between text, authors and keywords.

Common to the data harvested from these sources - Hypher, ANTA, scientometrics - is that they are scarcely decodable in their present form by just reading the tables. As the number of sources included quickly reach astronomical proportions (50 starting points crawled in 3 successions by Hypher produces 500.000 nodes, and the possible configuration of links between them offers so many permutations that we cannot even begin to account for them). Therefore, to render it readable we code it into Gephi, a visualisation tool that maps the connection of these nodes and sorts them according to their individual interconnectedness as calculated by the Force Atlas algorithm. We will in other words layer the networks of references on top of each other to construct a representation of a large sum of actors (Latour et al. 2012; Venturini 2010). We will then let the relative importance of actors emerge as the result of the groupings and weighting ascribed by the networks they reside in.

The end product is a series of maps that portrays the network of actors connected by Big Data at different points in time. We will use these to analyse at what points (both in time and web-space) that major translations occur.

This will leave us with two accounts of the translation of Big Data over time and space; a quantitative and qualitative. But as separate pieces they only tell so much - they do not provide an obvious singular interpretation. So how can these produce knowledge on the questions we are trying to answer: how did Big Data get so big, and what did it become in order to get so big? To answer that, we will use the inscriptions we have produced as input to reassemble coherent visions of Big Data.

3.2.3 Narrative Reconstruction: Telling the Story of Big Data

Finally, we will weave the threads thus spun into a narrative account of the emergence Big Data. We will analyse the expectations through the formulation of a set of five visions - coherent worldviews presented by actors to stabilise a network of expectations to a given technologies development trajectory (Lösch 2006). To do this we draw upon a text corpus of 50+ articles identified in the prior analyses. Operationally is this phase less leveraged by technological tools. Working on a purely qualitative level we will manually search these articles for statements articulating expectations to the future of Big Data in form of impending changes, potentialities and statements contrasting current or future states to the past (*we used to x, but now we y*). Based on these statements and the results of our previous analysis, we will retell the story of Big Data in the form of five coherent visions - each stabilised through the assemblage of a number of future expectations in what Callon has conceptualised as actor-worlds (Callon 1987). We will also discuss the different visions' possible effects as interessement devices in order to look at how different actors are enrolled.

Notes

1. I.e. by not grasping that IBM and International Business Machines are the same organisation.

2. As e.g. seen with the Danish journalist tracing 'Wilhelmsen' instead of 'Vilhelmsen' (Elkjær 2012).

3. https://wiki.digitalmethods.net/Dmi/ToolGoogleAutocomplete

4. We are not the only one repurposing the autocomplete function. The tool is also commonly repurposed by regular users as a impromptu spell and fact checker: writing the first few letters of a long word will write it out in full, and querying for e.g. lead role titanic will suggest Leonardo Dicaprio faster than any dictionary reference.

5. https://github.com/medialab/Hypertext-Corpus-Initiative

6. See Girard 2011 or Venturini 2011 for an overview of the tool, its potentials and history.

4 ANALYSIS I: TIME

THIS IS THE FIRST OF 3 CHAPTERS CONTAINING OUR ANALYTICAL PHASES. WE INITIATE OUR ANALYSIS BY TEMPORALLY TRACING THE GROWTH IN THE USE OF THE TERM 'BIG DATA', ESTABLISHING THE NODAL POINTS OF THE EMERGENCE OF BIG DATA.

The following chapters contain our three analytical phases. Before we start, we will shortly rekindle a point made in chapter 2.4.2 - that the emergence of digital methods redistributes the roles and functions of research (Marres 2012). As we pointed out, the boundaries between analysis, data collection and visualisation is blurred in digital methods (Madsen 2013:71). Therefore we will skip between the different levels, accommodating narrative and logical cohesion rather than clear methodological boundaries. Furthermore, we stress that while the operations with which we have produced the following maps and visualisations might be opaque to our readers, the manipulation and construction of these maps are as much of an analytical endeavour as is the subsequent interpretation and collation to theoretical models.

As mentioned earlier, we initiate our analysis by temporally tracing the growth in the use of the term 'Big Data'. We do this both to follow its trajectories peaks and valleys, but also to establish nodal points in the general timeline in which we can zoom in and continue our investigation. As such this analysis serves both to confirm our initial suspicion that Big Data is in fact a hype, a term subject to rapid increase in interest and optimism, but also as a component in an analytical relay, where the output of each analytical phase serves as input for the next hereby confining the scope of our investigation. We will also present some of the secondary findings from these experiments to add further richness to this depiction.

Finally we compare our analysis with the existing tradition of hype studies to provide criticism for both our own and other studies methodological shortcomings - shortcomings which we will propose ways to alleviate in the subsequent analytical chapters.

4.1 THE GROWTH OF AN IDEA

How can we trace the growth of an idea? Before we start talking about hype and exaggerated expectations, we will try to confirm that Big Data has indeed gained increased attention; that it is a trending topic in the parlance of our times.

In order to qualify this assumption we turn to batch-querying Google's search interface to return the total number of pages for the term Big

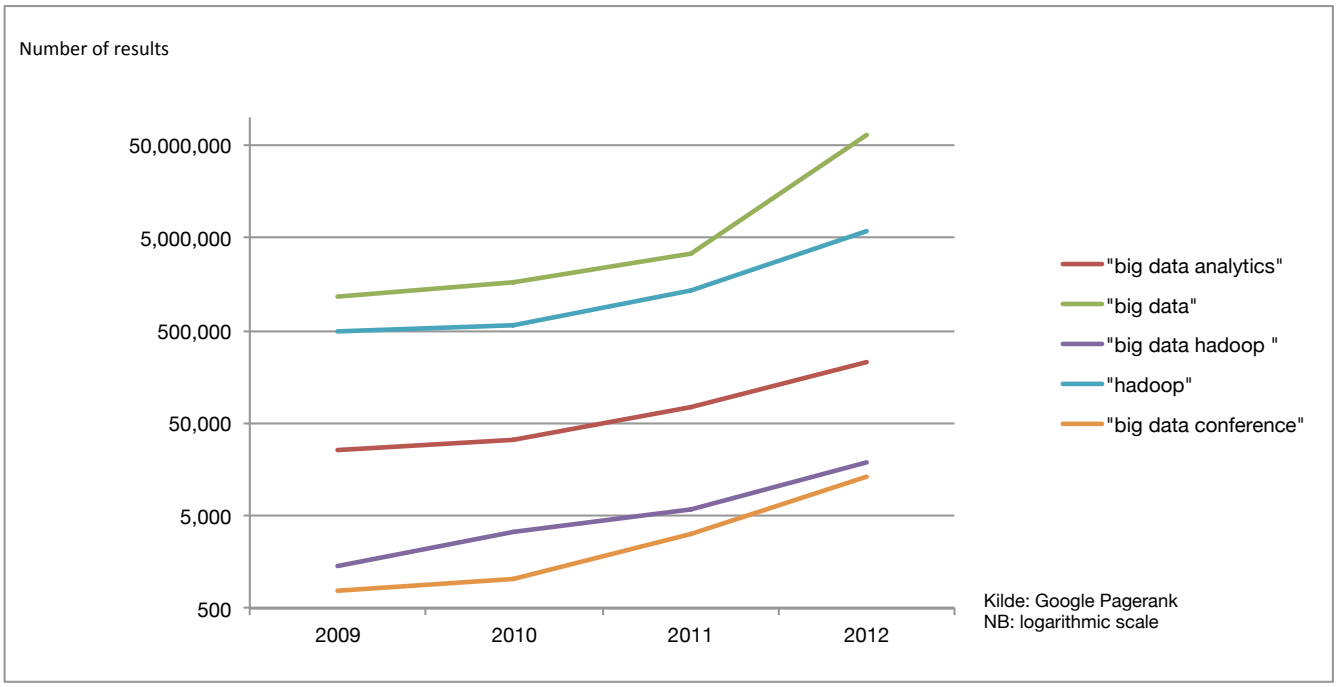


Figure 5: Number of Search Results over time.

Data (and related terms) for time periods of one year starting 2008. This is done as a quick reality check to ascertain whether Big Data actually was increasing in exposure as measured by the amount of websites mentioning the subject. As a simple proxy, we try to plot the number of search results returned from Google in a timeline[1]. We thus use Google search as an inscription device that assembles digital traces to enact a picture of the interest in Big Data.

The results as seen in figure 5 are staggering and quite indicative. Not only do we witness a overall growth, but the plotting is based on a logarithmic scale, which visually flattens the growth curves to accommodate for a explosive growth. The term "Big Data" itself generates the vast majority of results, starting from under 1 million results in 2008 but increasing during the next four years to a staggering 65 million homepages in 2012, with 2012 alone exhibiting a growth rate of 1847%. Similar patterns are found for the terms "Big Data conference" and "Big Data analytics", which both experience explosive growth rates during both 2011 and 2012 with Big Data conferences experiencing the overall highest and the next highest growth rates of the experiment. In contrast to these, we observe how the more technical queries on "Hadoop", the most widespread Big Data application, follows much more constant growth rates. So in regard to the patterns of expectation dynamics mentioned in 2.6, it seems that more technically minded translations are less subject to abject hype than generalised or business oriented terms. This is also in line with Brown and Michaels' application of Mackenzie's uncertainty through to expectation dynamics: the closer an actor is to the actual development of technologies, the higher their perceived uncertainty as to its future potential (Brown & Michael 2003).

2012 seems to be the tipping point for all queries, with all terms growing by over 200%. Especially the more open and less specific queries "Big Data" and "Big Data conference" have a marked increase in growth rates. Is this a significant trend? Statistically speaking, resoundingly so. While the exact degree of uncertainty around these results can only be guessed at, growth in the use of a term on such a massive scale allow us to safely say that Big Data and the related terms are exploding as a subject of conversation across the internet as measured by the number of web pages mentioning the term. So to answer our initial suspicion: yes, Big Data is indeed a trending topic.

A bit of informed guesswork makes the difference in growth curves between *Hadoop* and *Big Data* a notable clue as to the patterns of translation. Assuming that "Hadoop" and related terms are mostly referencing more technically specific content, we see that the technical and scientific interest is indeed on the rise, but an early, more constant and stable rise without the jagged peaks of the other search terms.

Oppositely, we find the increased mentions of *Big Data conference* points in direction of a rise of the term in regard to business and management. A quick qualitative glance on a number of central Big Data conferences, their programmes and many sales pitches, tells us that the conferences are in fact more focused on application and possible benefits than technical and scientific interests. The exponential growth of search results for the query on "Big Data conference" could thereby be an indicator that the term is translated into a subject for business and organisational literature in 2011.

Thus we arrive at a possibility for further exploration: *that the interest is led by a small cadre of technologically competent adopters and developers, while business and mainstream interest follows only later, but usher in much higher volume.*

As per the discussion of uncertainty introduced above, it will be interesting to find out if there is a contrast in the expectations these two groups enact that might account for this difference in growth patterns.

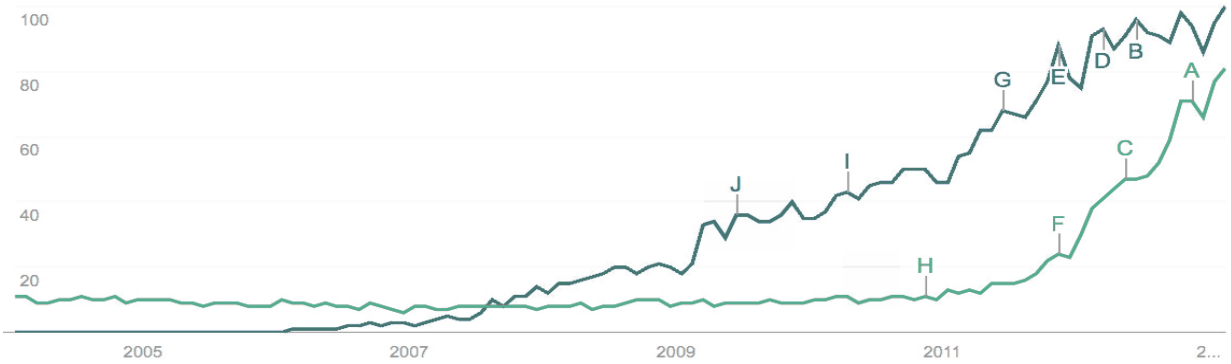
4.2 GOOGLE TRENDS - LOOKING AT SUPPLY

The numbers above only reveal the amount of sites that cite Big Data as a subject – they are the supply side of the equation. They give us an idea about the amount of information offered about Big Data. As stated in our research design, however, we also wish to explore the development of interest in Big Data, what we have called the demand side (see chapter 3.2.1). To do so we turn to Google Trends to survey the amount of queries for the terms made to Google's search engine. This gives us an estimate of how big the interest is in reading about the terms[2].

This inscription of course enrols different entities than the one in 4.1. By assembling search queries rather than number of web pages, we thus enact a picture of the demand for information. By comparing these two different inscriptions we hope to find insight into underlying dynamics.

In figure 6 we see the temporal distribution of results, and here we also find the same gentle curve for Hadoop and sharper incline for Big Data. The growth in queries for Big Data is especially marked for the period from 2011 and onwards, a possible indicator of massive public interest. Comparing this inscription with the one in 4.1, we find again on the demand side that the first attention is instigated by a technical interest, while the more generalised and broad interests appears together with the big peak from 2011.

Figure 6: Google Trends - Hadoop vs Big Data.



Comparing the trends of Big Data with Hadoop yields an interesting juxtaposition to the earlier results in 4.1: while the number of search results for "Big Data" was 10 times higher than that of "Hadoop", the number of searches for "Hadoop" is actually higher than the one for "Big Data". So the supply and demand for information is thoroughly mismatched: more people are talking about Big Data, but more are searching for Hadoop. This could be because Big Data is a word of commercial connotations, a buzzword for sales pitches, while Hadoop (open source in nature) is not something to be sold, but something to learn and apply. They two terms are thus able to assemble different entities: while commercial actors and purveyors flock to Big Data and present information on it, Hadoop gathers the queries of the technically minded.

On the demand side, we find that Big Data is indeed a trending topic, a topic of conversation growing at a tremendous rate. We also find a preliminary division of our timeline: a major break seems to occur in 2011, so we will divide the timeframe in two periods: from 2008-2010 and 2011 to 2013. Furthermore, we will add a period from 2001 to 2007 to try and capture any early indications in the development of the term. These periods will form the basis of the analysis in later chapters, and will be denoted as the *early, middle* and *late period* of Big Data.

Lastly, by comparing the growth rates of the supply and demand of different search terms, we find the first indications of a pattern. One, that the more technical terms (Hadoop) have a flatter growth curve. They are more evenly distributed through the period, and while they exhibit significant growth, they are nowhere as skewed as the distribution of the less technical terms (e.g. "Big Data conference") that have a markedly more pronounced peak and occurs later in the period. We take this as a possible indicator that technical interest started the development process, but more commercial interest contributed to the majority of activity in later stages. This hints at a pattern for the dynamics of translation, and the mobilising power of different terms. We will return to this pattern later to see whether the expectations vested in the different terms can account for this pattern.

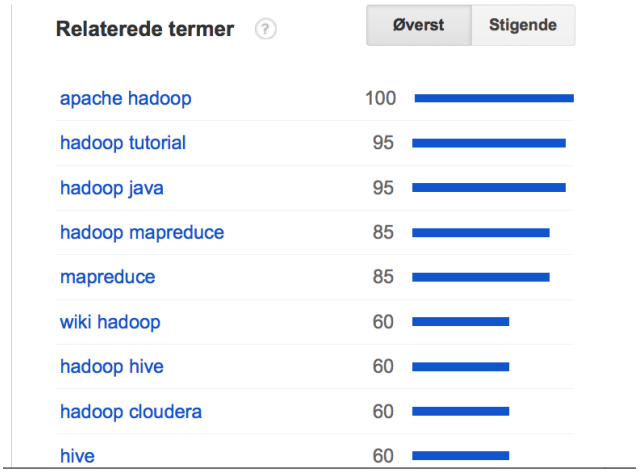
That generalised terms like Big Data are predominant in the supply of sites relative to search traffic, while the more specific technical terms like Hadoop figure primarily in demand as measured by search could support the notion that Big Data is a term with more commercial connotations.

All these inklings will be further explored in subsequent chapters.

4.3 GEOGRAPHIC AND RELATIONAL TRACES

Apart from these primary results, our experiment with Google Trends provides two additional traces to follow. The first is related searches; the second is the geographic distribution of searches. These two provide additional contributions by allowing us to trace associations to both geographical and semantic entities. Related searches (figure 7) add to the intuition that Hadoop is a query more technical in nature: the related searches are into tutorials, wiki databases and specifications of software variants. As such, we can see that this tool allows us to uncover the meaning vested in a term by positioning it in a network of associated terms.

Figure 7: Related searches - Hadoop.



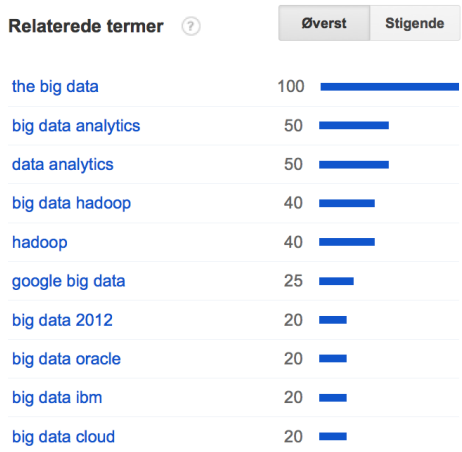


Figure 8: Related searches - Big Data.

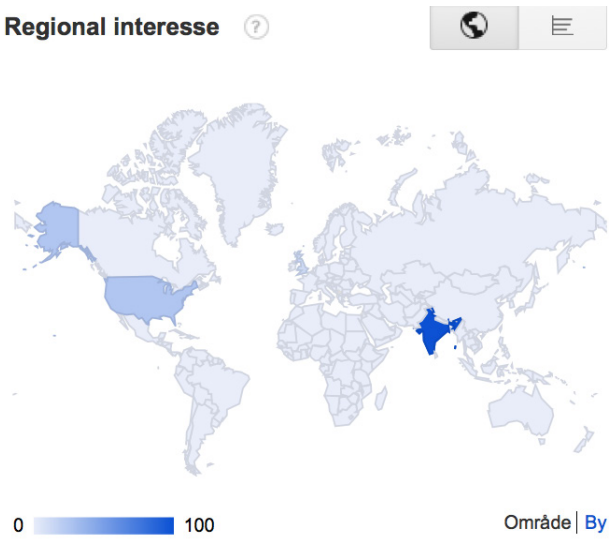
The related searches for Big Data (figure 8) tends to go more towards specific firms and organisations, with conferences also make an appearance - thus strengthening our claim on the commercial nature of the term.

The geographical distribution is however revealing: as seen in figure 9 are the more technical searches, e.g. *Hadoop*, are heavily skewed towards India - and Bangladesh in particular.

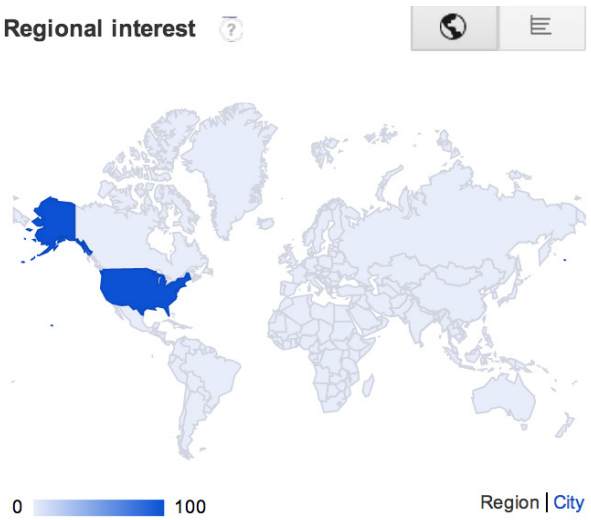
In contrast, figure 10 illustrates that queries for both "Big Data conference" and "Big Data companies" are an entirely American phenomena. Scanning the timeline shows they rise rapidly from obscurity in 2011. They thus follow the trajectory for the supply side more closely. The same pattern is observable for "Big Data jobs".

This points to markedly different translations of Big Data in the United States versus India. Is the arrangement thoroughly asynchronous, only allowing Indians to associate with Big Data as producers, and Americans as buyers and sellers?

Figure 9: Regional interest - "Hadoop".



Figur 10: regional interest - "Big Data conference".



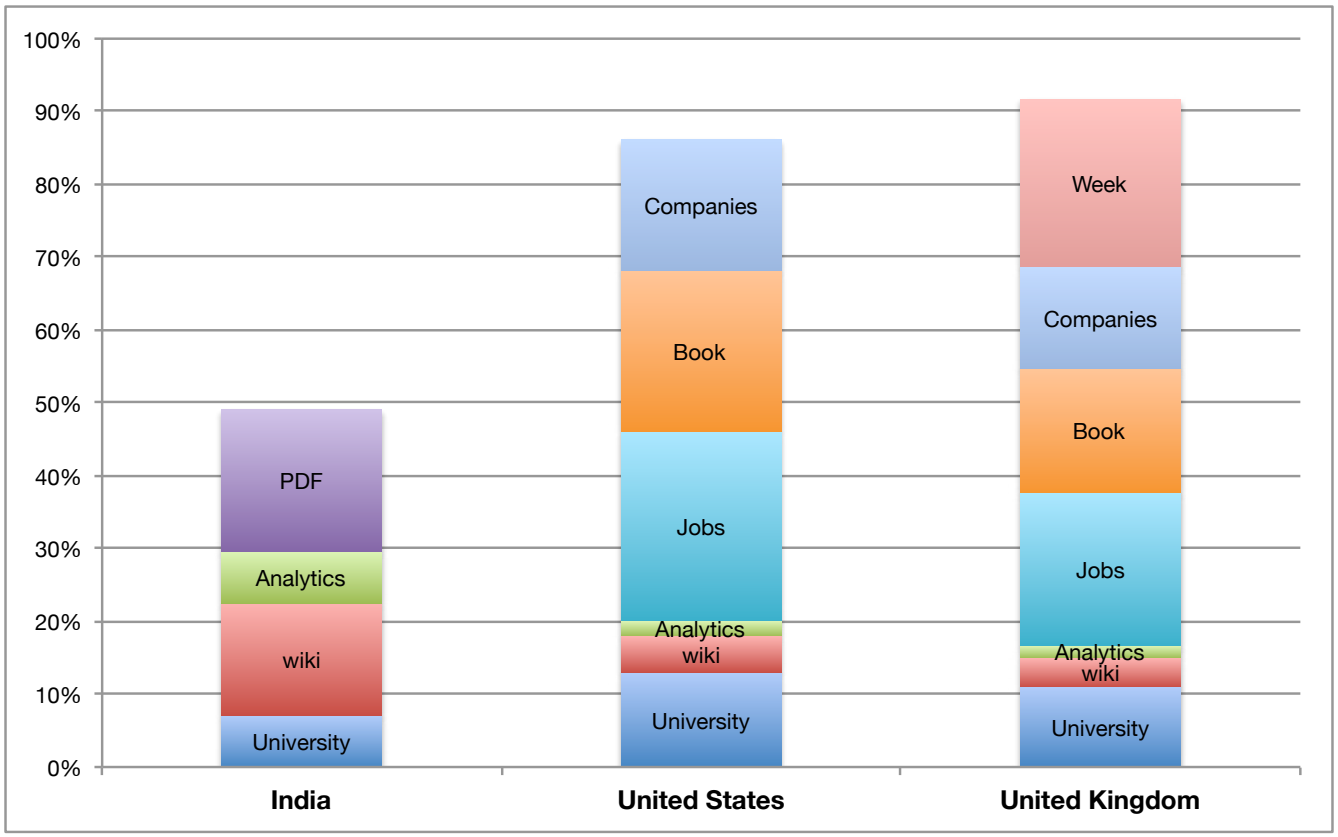


Figure 11: Co-Occuring Searches.

Meanwhile, Bangladesh has a reputation for being a site of IT offshoring activities. Such a deviation certainly warrants closer inspection. Is the actual computation behind Big Data carried out in Bangladesh, while the network of Big Data is asymmetrically distributed across the globe? To answer this, we have compiled a list of regional co-occurring search queries for 3 locations by using the tool autocomplete.

We compare searches across India and the United States, using the United Kingdom as a control measure.

What we find is that the related terms in India are *University*, *wikis*[3] and *PDF reports*, all indicative of a situation where Big Data is a subject of study first and foremost.

Contrary to this the two western countries are largely similar (with the exception of Big Data week - a *"global festival of data"*[4]) focusing on jobs and companies, with companies and jobs taking up a significant portion of searches and books occurring instead of PDFs - possible reflecting the difference in economic opportunity; PDFs are often free or pirated. Once again we see that the more technical term, Big Data analytics, occupies a significantly larger fraction in India. These networks of associated terms show how Big Data is enacted in two very different ways in India and the United States, hinting at the multiplicity we mentioned in chapter 2.2. What this means for us is that we cannot hope to stabilise Big Data as a singular network - already here we see that multiple enactments each assemble markedly different versions of Big Data: one as a subject of business and conferences, and one as a subject of learning and education. We will delve further into this in the two other analytical phases.

In summary, these secondary sources of related traces contribute with two things: the co-occurring search terms confirm our interpretation of Hadoop as being a specifically technical term, and Big Data a more generalised and commercial term - a buzzword. This apparent difference

in meaning between the two terms points to an important limitation of the purely quantitative measures of aggregated supply and demand for information: they can not tell us about the nature of the underlying interest. We can not conclude a high level of expectation purely from high attention. As we shall discuss shortly, this connection is often assumed in Hype studies.

Additionally, the geographical disparity in search terms point to an important consideration for our tracing of the rise of expectations towards Big Data: that interest (and by extension expectations) should not be seen as a singular mass, but as locally constructed, situated emergence (Haraway 1988) that differs widely across different actors and regions. With that in mind we will look at earlier attempts to operationalize the notion of hypes, and consider their limitations and shortcomings with specific attention to the relationship between interest and expectations.

4.4 HOW TO MEASURE HYPE?

As mentioned earlier, what we just studied in the preceding chapter has a certain overlap with so-called hype studies: a longitudinal tracing of the interest in a particular term. In this part we will discuss the existing approaches to hype studies and propose venues of developing their operationalization. We have examined these approaches in the literature review with regard to their explanatory ambitions, but will now delve further into their methodological operationalizations.

In our literature review we defined Hype as a specific pattern of expectation dynamics, characterised by a sudden and marked peak in positive expectations to the future potential of a technology, often followed by an equally marked decline in expectations once actors start questioning the hype and view it as exaggerated (Van Lente et al. 2013). But how should we measure hype? Apart from Gartner's seminal Hype Cycle, the composure of which is a trade secret and thus not attainable to us, we find a number of published studies trying in one way or another to provide an empirical operationalization.

We suggest distinguishing between two approaches. The first one is exemplified by the type of studies that aim to corroborate the hype cycle through empirical testing. Among these we find Jun (2011 & 2012) who puts forward a distinction between consumer, media and producer hype cycles. To this end he operationalizes the measurement of producer hype cycles as the intensity of patent applications, consumer hype cycle as intensity of search traffic and media hype cycle as intensity of media coverage (2012:1418). In another study he tries to relate this consumer hype cycle to purchasing behaviour (2011). His approach mirrors our own distinction between supply and demand of information. However, we contend his equation of search traffic with high expectations: *"Secondly, it is possible to measure consumers' expectations using search traffic"* (Jun 2011:97), a critique we will develop below.

The same approach is also seen in a devastating critique of Gartner's hype cycle centred on the inconsistency of its year-to-year rankings of different technologies (Steinert & Leifer 2010). Here, the rankings are compared with hype measured as media visibility and search traffic, thus again equating the level of expectations with the intensity of information.

A related take is seen in a study of DVD technology that also measures hype solely by media visibility (Järvenpää and Mäkinen 2008:4). In contrast to the work of Jun it does however acknowledge the heterogeneous distribution of hype by comparing different type of publications.

All of these studies thus equate high levels of expectations with high amount of visibility or search traffic. In response to this we will deny that sentiment, how positive expectations are, can be deduced from salience, how much attention a subject is given by either providers or consumers of information[5]. While we use similar measures for what we call the supply of information, we do not claim them to synonymous with expectations - it is possible for a subject to generate large amounts of public discussion without positive expectations - just look at wars or policy failures such as Cop15.

This critique is shared by the other approach, studies in the STS tradition. These include Ruef and Markard who propose to "*conceptually separate attention and expectations. We argue that hype, and especially disappointment, cannot be deduced from a peak (and decline) of media attention as attention and expectations are not necessarily related"* (2010:320). Their study combines these quantitative measures with qualitative analysis of statements of expectations, but on the other hand does not take search behaviour or other indicators of demand for information into account. Other studies follow this approach, including an attempt to construct a typology of hypes that also combines qualitative and quantitative analysis (Van Lente et al. 2013), but does not account for differences in the enactment of expectations between different actors. Lastly, Bakker also references the distinction between visibility and expectations, and cleverly uses the number of hydrogen car prototypes as a gauge of expectations by auto manufacturers, but also refrains from discerning between how expectations differ between actors. As we found in 4.3, the same term can cover markedly different enactments, and trying to establish a global or aggregated level of attention of expectations glosses over these variations.

In summary, we find that both of the two approaches presented above contained limitations: first approach acknowledges the difference in enactment between different actors by discerning between what we call the supply and demand of information, but does not separate what we call the salience and sentiment - the level of attention with the degree of optimism. The other approach separates salience and sentiment, but does not consider how different actors enact expectations differently, nor does it embrace multiplicity.

Both of these approaches seem inadequate to us - especially if the epistemological interest is not just the empirical phenomena of media hypes, but an attempt to look at expectations as a driver of development in a broader perspective. A core shortcoming, which we have yet to address ourselves, is the conceptualisation of hype as a singular, aggregated pattern, neglecting how expectations differ between actors and how a hype starts in one place and gradually enrols more actors in a process of translation. We tentatively pointed out this issue when we looked the geographical distribution of related search terms, and explored it here through a critique of earlier studies. We hope to rectify these shortcomings by looking at how Big Data assembles different arrangements of actors in chapter 5 to address the multiplicity of enactments, and how expectations are drawn together in visions to try to account for not only the level of expectations, but also *what* expectations are enacted.

4.5 INTERIM CONCLUSION

From our investigation we sketched a rough and primarily quantitative outline of the growth in Big Data. First and foremost, we confirmed our initial suspicion of Big Data as hype, with our exploration of both the supply side (Google Search result) and the demand side showing an exponential growth rate. Relying on the same data we also found indications that the development process of Big Data was founded on mostly technical interest, while more commercial interest contributed to the explosive activity in later stages; a hypothesis which we will trace further the second analytical phase.

Finally we constructed three periods - early (2001-2007), middle (2008-2010) and late (2011-2013) - to form a temporal basis for the analysis in the later chapters.

In conclusion we find that while hype is undoubtedly a valid description of an important empirical phenomena, hype studies in their current incarnation not only provide an incomplete operationalization of the phenomena they try to depict, they are also inadequate if we want to account for a more comprehensive range of ways in which expectations serve to propel the translation of trends in technological development.

Firstly, they do not consider the distinction between what we call the supply and demand for information on a subject. This problem we have already touched upon by comparing the difference in patterns between the two. Secondly, the majority does not consider the difference between salience and sentiment, between public attention and expectations. We will try to approach this problem in chapter 6, where we look closer at the semantic composition of expectations to Big Data. Thirdly, they often fail to consider the difference in local enactments of interest and expectations, instead aggregating everything into a singular representation of total interest. They do not address the different roles of the actors enrolled in the network around a term.

In the next chapter, we will explore how the network of associations to Big Data is translated over time to try and address these three problems.

Notes

1. As discussed in chapter 2.6, one should be cautious to judge these numbers as directly related to the level of expectations since increased interest is not necessarily an indicative of increased optimism.

2. The results generated by Google Trends come from a slightly more complex procedure, especially the geographic dispersion because of technicalities such as normalising for regional search volume. The gist of the function though, is displaying the growth in the number of times a given search query is performed over time. This is elaborated in the protocol for the experiment, see appendix 9.1.

3. Open access, co-created repositories of information, often highly specialised around niches for independent learners. A more generalised example is wikipedia.org

4. See bigdataweek.com

5. Here we borrow the nomenclature of agenda setting studies, not to establish a theoretical connection, but merely because we need an umbrella term for both the supply and demand for information (McCombs & Shaw 1972)

5 ANALYSIS II: ACTORS

THIS ANALYTICAL PHASE REGARDS EXPECTATIONS AS PROPELLERS OF TECHNICAL DEVELOPMENT AND ENROLLER OF ACTORS. WE WILL LOOK AT HOW NETWORKS OF REFERENCES TIE DIFFERENT ACTORS TO BIG DATA IN THE FORM OF ACADEMIC CITATIONS, HYPERLINKS OR WORD CO-OCCURENCES.

In the following we will look at how networks of references tie different actors to Big Data. To the extent that the tools allow us to do so, we will seek to map the development of networks formed over time and the origin of the translations: in which context the term emerges and at what speed and direction it is translated. From this analysis we hope to pinpoint the decisive points where new actors are enrolled in the network of Big Data and how these enrolments rearrange the networks.

The basic approach will be to map networks of associations, either in the form of academic citations, hyperlink references or word co-occurrence in documents.

We will start by mapping the academic citations in a scientometric analysis to account for the emergence of the term as a subject of research (See protocol C), based on the assumption that the underlying technical nature of the term points to an origin in science. This experiment will pinpoint the most cited articles and their topics (represented by article keywords) to draw a picture of the central contributions to Big Data.

We will then compare these findings to the next experiment, a mapping of hyperlink references between websites (Protocol D) to embed the scientific contributions role in a broader context. This experiment will show how the relative strength (understood as visibility in the discussion) of different actors varies over time. It will also lead to us to propose a typology of different types of actors, demarcated as different spheres based primarily on the different data sources. Lastly, we will explore how these different spheres enact Big Data in different ways. We do this by analysing the occurrence of different keywords, organisations and terminologies in documents from each sphere, and comparing the results across them (Protocol E).

In continuation of our discussion in 2.4 on the representational modus of digital methods, these mappings should be seen as socio-technical modes of seeing, each of which enact a different ontology - conjuring forth different actors by their different ways of inscribing order. The ordering, ranking and selection of actors might in this perspective differ from map to map due to the different modes of seeing.

The multiplicity in our modes of seeing thus not only sheds light on how the different tools of our method assemblages enact different realities. But by comparing these different inscriptions we can extrapolate an idea of how the different actors are mobilised and associated in the network around Big Data.

5.1 SCIENTOMETRICS

In this experiment we harvest and visualise data on citations from repositories of academic journals to map networks of academic articles whose associations are composed of citations (Protocol C).

In the centre of our map (figure 12) we find an article written by authors affiliated with Google[1], *Mapreduce* from 2008. In the same genre we also see that the third most referenced article is attributed to authors from *Apache's Hadoop.* These two articles detail the development of specific software solutions to handle large datasets by distributing the calculations among swarms of servers. Mapreduce is a programming model developed by Google to drive its search engine. Hadoop, in its various incarnations, is a later open source adoption of Google's

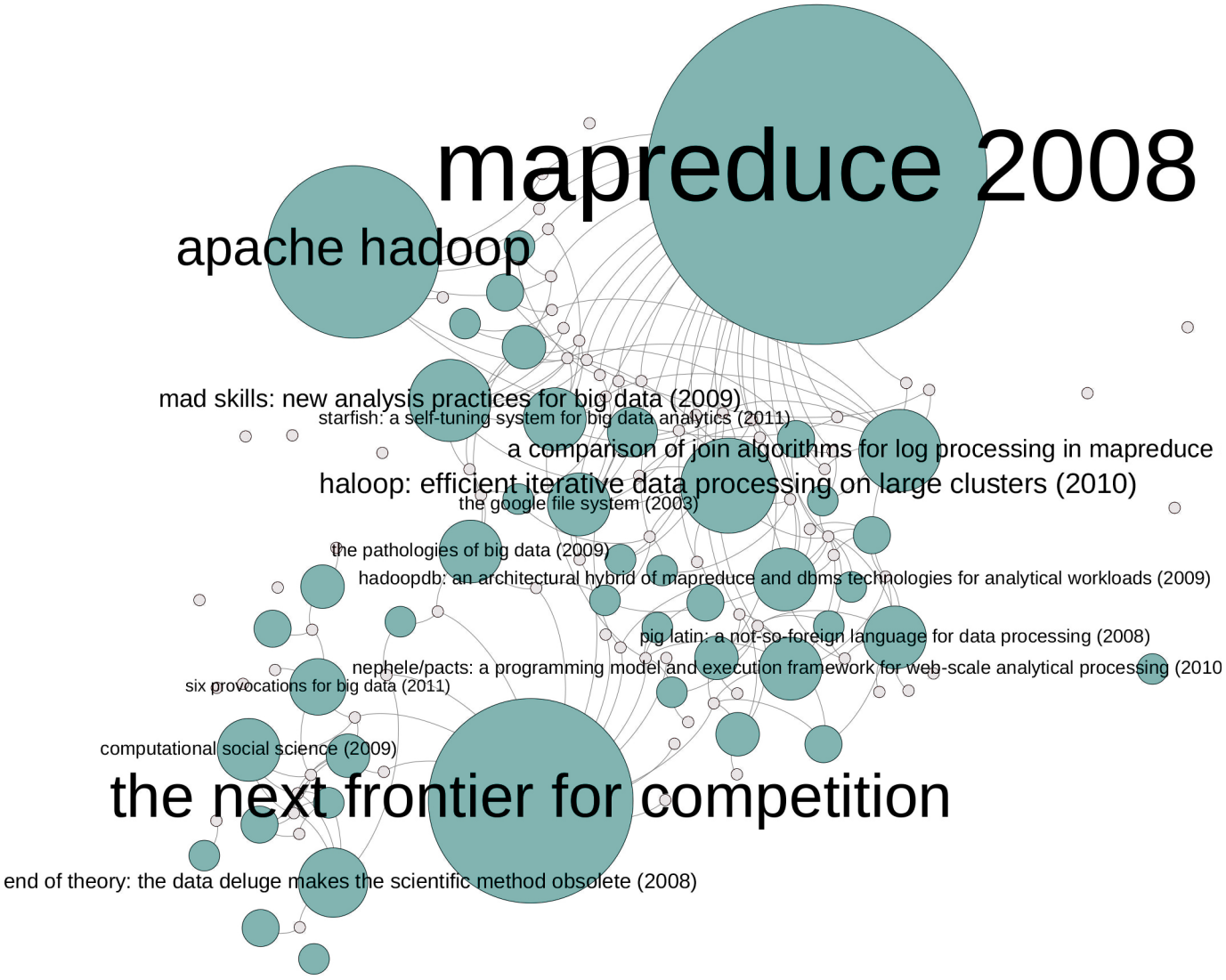


Figure 12: Big Data citations 2001-2013. Green: references, Grey: original articles. Visualization filtred by degree.

Mapreduce. These two technologies appears to form the technological basis of handling Big Data, and as such their central position is entirely warranted.

Surrounding these two nodal points in the upper part of the map are a host of other articles from computer science dealing with different aspects of calculating on these data sets[2]. In the opposite corner of the map we encounter another group of articles on the social sciences, e.g. Lazer's *Computational social science*, Savages and Burrows' *The Coming crisis of empirical sociology*, and Crawford and Boyd's *Six provocations for Big Data*, of which we have cited many in our own methodological reflections (see chapter 2.4). These articles all points to an emergent tension between sociology and computer science evolving from the behavioural data of computer science encroaching on the habitat of social science (a tension we have already discussed at length in chapter 2.4). While the articles take different stands in this discussion, their spatial closeness in the map points to a common corpus of citations.

Another type of text which attracts our attention is Mckinsey's 2012 report *The next frontier for competition* and Chris Anderson's 2008 article *The end of theory*. While the appearance of these decidedly un-academic articles are expected, their presence and centrality (the Mckinsey's report figuring as the second most cited article) in scientific publications is surprising. By tracing the associations drawn by citations, our scientometric method thus enacts Big Data as an assemblage of actors both in and outside of academia. This intermingling of science with both popular discourse (Wired) and commercial white papers (Mckinsey), as well as the central position occupied by researchers affiliated with commercial actors (Google and Hadoop), shows us a picture of Big Data as a hybrid that draws associations across traditional borders of science (Callon et al. 2002). Even when looking only at citations, commercial actors take central positions in the network. We can thus not trace its emergence solely from academic research, but have to follow relations as we encounter them. To further explore this hybridisation, we will look at an additional trace afforded us by the academic databases: the affiliation of articles.

5.1.1 Affiliations and Sponsorships

Our data further allows us to trace the sponsors of scientific articles on Big Data. To dig further into the actors behind the science, we therefore take a look at the organisational affiliation (see figure 13).

Once more we witness a strong appearance of big tech firms in the science. The top three supporters are *IBM*, *HP* and *Microsoft*, surpassing even universities with strong reputations for computer science such as *MIT* and *Berkeley*. So not only are commercial actors behind the most widely cited articles, as we saw above, they also produce the bulk of the research. While we expected to find IBM and Microsoft, the appearance of HP in the top is surprising since we had not encountered them in our work so far. To inquire into this, we shifted to a qualitative mode, scouring their website for mentions of Big Data by a customised search query. By zooming in on them thus, we find that HP acquired Vertica in 2011, a firm we had earlier seen referred to as a major purveyor of Big Data services. This introduces another way by which actors can engage in the network, gaining representation through the acquisition of smaller firms and start-ups.

Finally we notice to a surprise that Google appears below top 100; an absence we will dig further into later in this chapter.

5.1.3 Discussion

In summary, we find that the central text among the computer science articles are Google's Mapreduce and Apache's Hadoop - both articles detailing the basics of the architecture that allows for computing Big Data. As Google's Mapreduce forms the basis of Hadoop (and thus the majority of distributions) for handling large datasets, their contribution to the development of Big Data is immense and significant despite only being affiliated with two articles.

Social sciences are also taking notice of these recent developments, with discussions clustering around the quantification of social behaviour and meta-discussions on the social potentiality and problems of Big Data. Lastly we encountered a number of decidedly un-academic articles central in the academic debate, pointing to

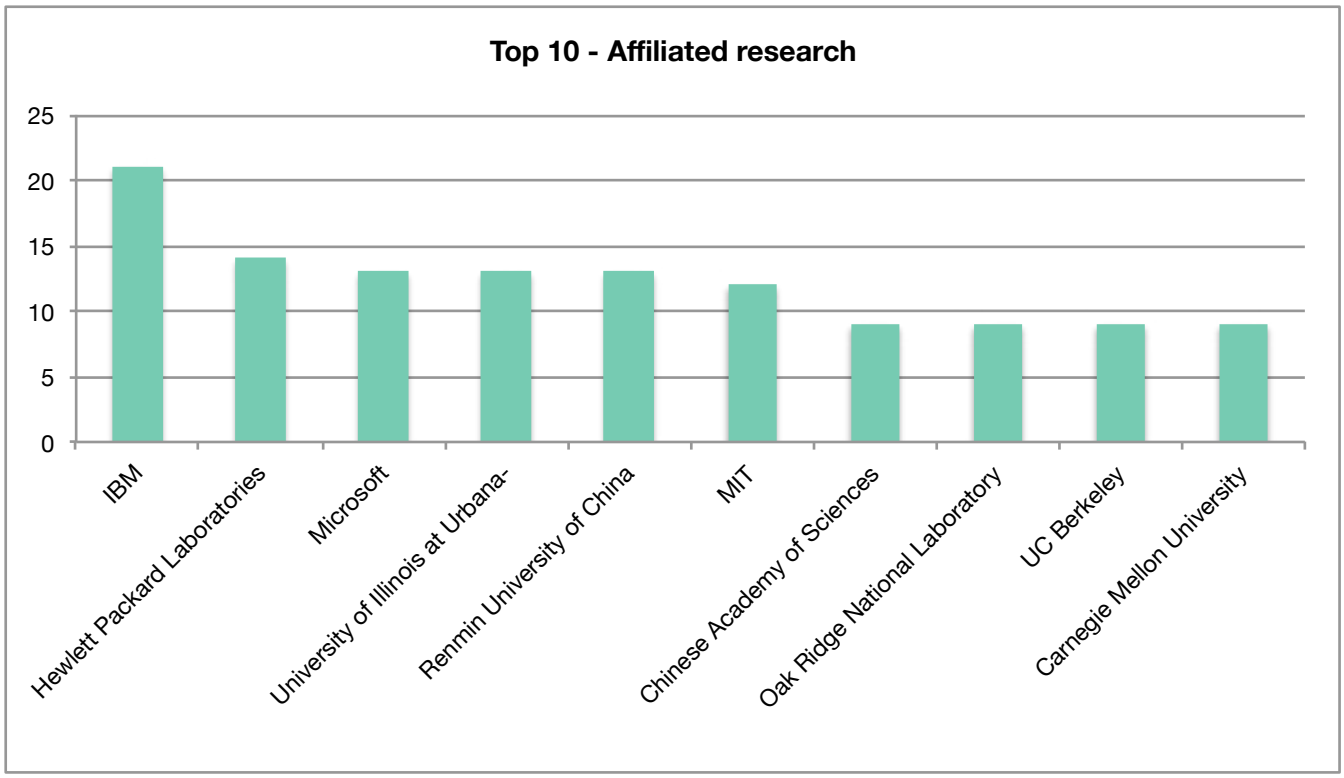


Figure 13: Number of affiliated research articles.

a hybrid intermingling between commercial and academic actors, a point we further explored by a short venture into the sponsorships underlying research.

Our methodological position on how different methods (as socio-technical modes of seeing) enacts different ontologies, was also exemplified through Google's strong appearance in the citation map and simultaneous disappearance as a sponsor of science. The two measures (citation map and affiliation) each make different aspects notable, and by juxtapositioning the results we can see how different actors are contributing in different ways.

This is however just two measures of which actors are related to Big Data. To extend the range of our vision we proceed with tracing which actors - both in and outside of academia - appear central to Big Data on the web.

5.2 HYPHER: CRAWLING WEB CONNECTIONS

Having found the central academic articles, we turn to the wider question: who are the central actors related to Big Data on the Internet at large? And how do they change over time? To do that we turn to a dataset we have constructed by assembling hyperlinks from the 50 websites ranked highest by Google's search engine from our three periods (See Protocol C for a detailed account). This corpus will be analysed in the following chapter by manipulating and visualising different parameters to construct a range of inscriptions that enact different measures of relevance. For each of these enactments we will compare the inscriptions between time periods, but also infer analytical points by comparing the different enactments.

5.2.1 Out-degree: The Propagators of the Web

We start by asking *who the main propagators of information are,* since hype is often seen, as we mentioned in chapter 4, as heavily influenced by media agendas. More specifically, we rank our network by out-degree, the amount of times they link to other sites in the map. This results in an inscription portraying who refers the most to other actors. We also remove all actors who do not get any inbound links, to assure that random "spammers" do not skew our results. As discussed in 2.3, we take this as a measure of degree of association to other actors in the network.

We however learned that cuation to this temporal setup had to be shown. During our experiments we find sites in our data which should not appear as they did not exist in the given time period, e.g. the strata conference which was not established until 2010. This temporal noise entered our data because of our technical setup, in which only our seed URLs could be temporally delimited by Google's filters.

This is illustrated in the adjacent figure (14). The initial starting points (seed URLs) were clean from temporal noise, as they were manually filtered. The links they pointed to, the 1. degree sites, were also relatively clean since they themselves must have existed at the time of writing to receive a link. However new links and content could have been added to these 1. degree sites, so that the 2nd degree sites might be from different time periods.

So the temporal contamination increases as the crawler moves away from its starting point. While sites identified to be from a wrong period were filtered out manually, the method used in the following can still only serve as an approximation in regard to temporal developments, as the noise risks contaminating the order this specific mode of seeing is trying to establish. That being said, the inscriptions it results in displays many overlaps with our other experiments, and the many plausible results suggest it is still a highly useful approach.

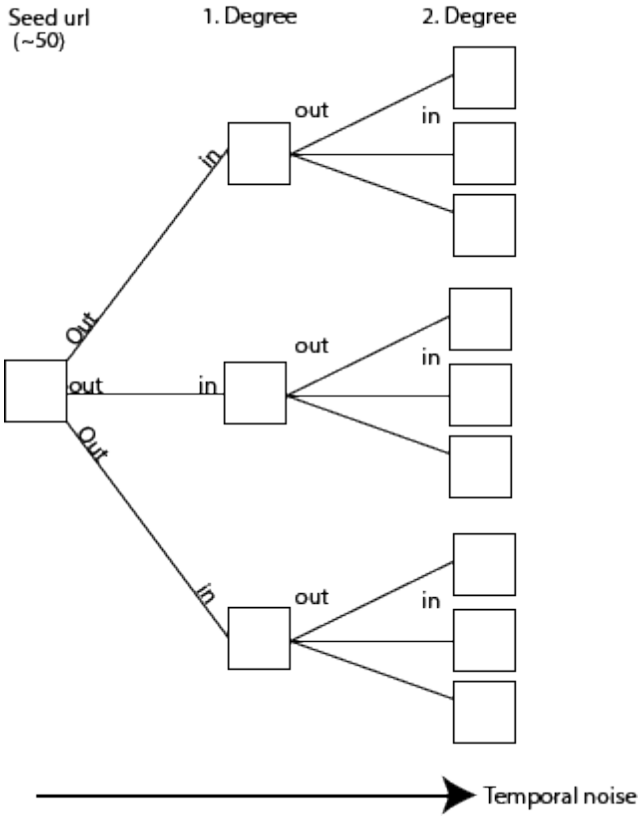


Figure 14: Temporal noise in Hypher studies. The squares represent linked websites.

5.2.2 Exploring the map

Our first map (figure 15) reveals 3 central nodal points: *Microsoft*, *OECD* and a cluster of somewhat similarly named sites around *IDG* and *IDC-connect*[3]. In between these lies a strata of smaller weakly connected clusters, like an asteroid belt between larger planets. Our quali-quantitative approach allows us to further explore the role of these entities by zooming in on the individual actors by simply visiting the URLs provided in a browser to take a closer look at web page. Through exploring this on a qualitative level we learn that the "asteroid belt" is composed mainly of middle size IT firms like *Greenplum* or *Paracel* founded in the early period of Big Data and dedicated to the then new discipline of data analytics. In the early period, these firms were mostly identified as data analytics (not *Big Data*), but have since slowly renamed their services as Big Data analytics. This change in terminology points to a development where certain words are subsumed by umbrella terms over time. This semantic development will be further explored in 5.4, but tentatively this points to one mechanism

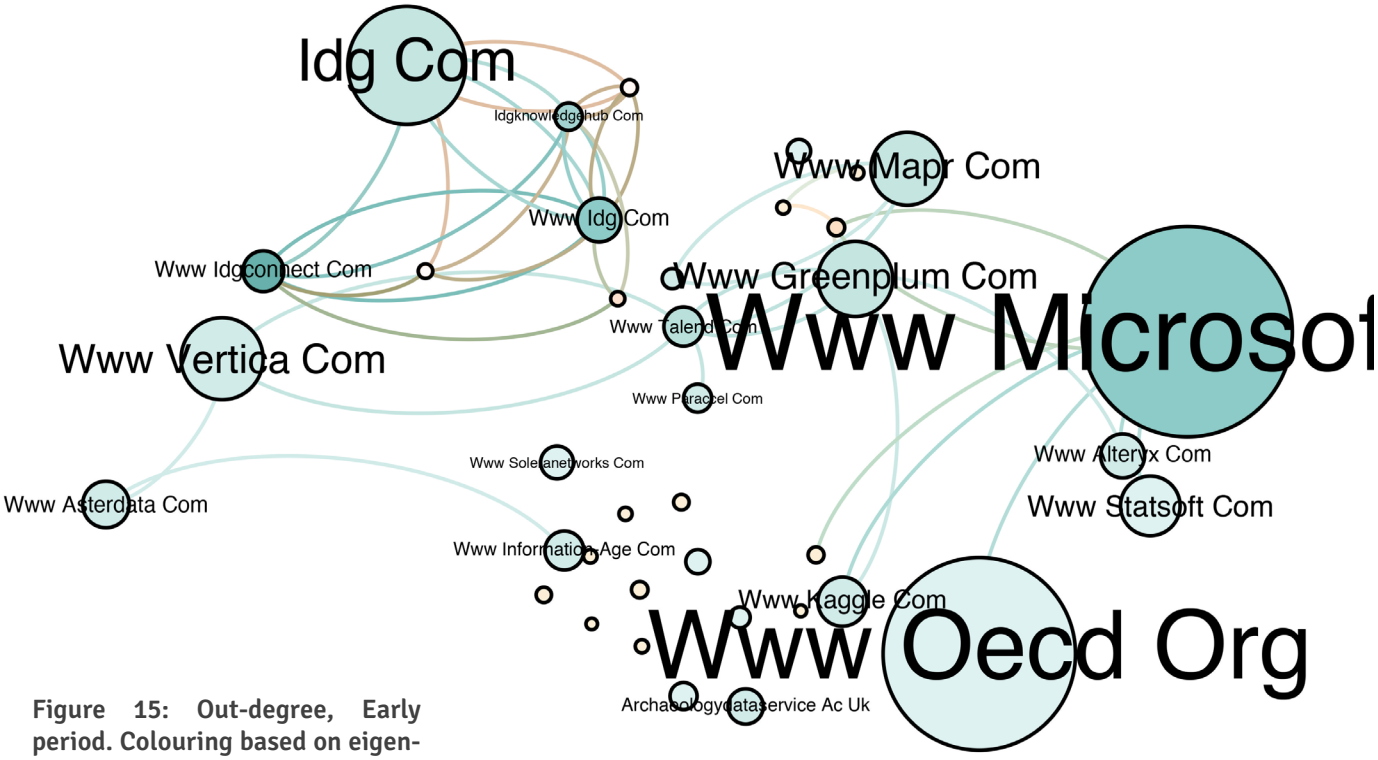


Figure 15: Out-degree, Early period. Colouring based on eigen-vector centrality. Visualization filtred by degree.

by which Big Data grows in size: by translating a host of lesser terms and thereby enrolling a swathe of otherwise unconnected actors in its assemblage.

Secondly we visit the cluster around IDG, a global IT corporation. From this visit, the cluster slowly shows itself to be a syndicate of sites, the IDG network, propagating news and knowledge on IT themes (see figure 16).

At first, this led us to consider removing it from the map, since the mutual linking between their own sites would lead to exaggeratedly high ratings (a common way to "cheat" Google for a higher search rank). Filtering out the subset of subsidiary sites however revealed how two sites in the cluster, IDG-connect and Infoworld, also receive a high number of links from sites outside the IDG-network. Diving deeper into these two sites reveals how they hosts a repository of more than 4,000 white papers on data management contributed by large tech firms. As such, this cluster of sites is revealed to play a role as instigator of discussion in a technically and professional commercial setting. This repository also shows an interesting facet of research and development when compared to the academic journals in Chapter 4: not only do commercial actors figure prominently in academic contexts, they also appear to have their own portals for knowledge sharing and collaboration in the form of IDG's white paper repository.

Figure 16: Footer on Infoworld.com revealing the members of the IDG-network.

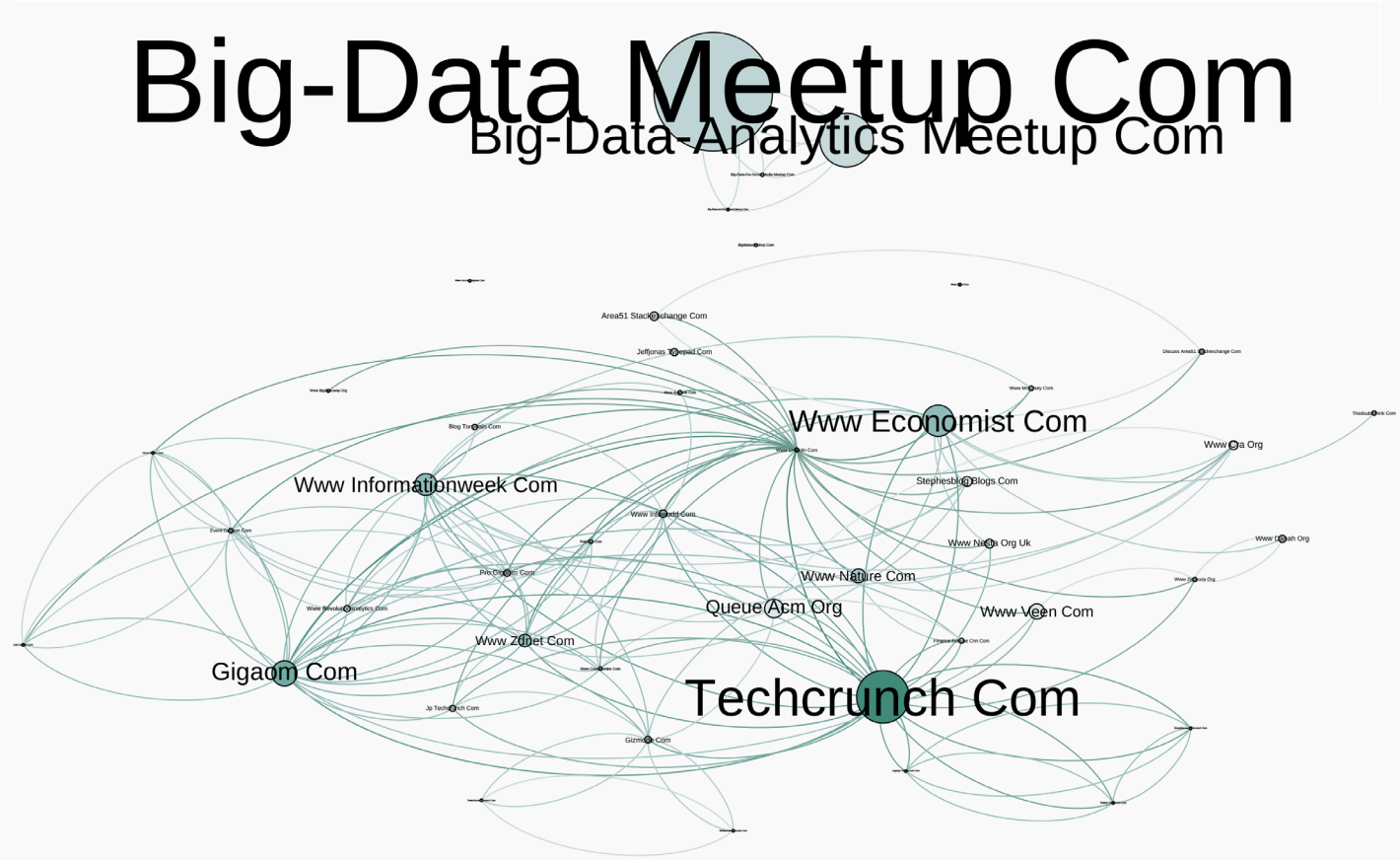


Figure 17: Out-Degree, Middle period. Colouring based on eigenvector centrality. Visualization filtred by degree.

When we subject the data set from the middle period (Figure 17) to the same treatment, we witness a clear shift in the sites depicted. Gone are Microsoft and most IT-firms.

Instead of sites targeting a minority of professionals, the biggest hubs are now all consumer oriented technology news sites such as *Information week, Techcrunch, Gigaom, Gizmodo and ZDNet*. Not only are the consumer-oriented technology sites increasing in size, the IDG-network simultaneously decreases in relevance and nearly disappears from the map in this period. Shifting again to a qualitative level, these tech news sites are also revealed to have a slightly different thematic focus than the IDG-Network. Where IDG-network imitates the role of trade journals providing information on mergers, hiring and other more business like topics for the IT industry, these new sites are more sensationalist outlets for consumers with a general interest in technological matters. They contain reviews of gadgets, reports on new software applications, speculation on what is next for the cell-phone industry and the sort - often in a highly futuristic, techno utopian vein. So Big Data appears in broader and less professionally orientated news channels - and in the process, the expectations and interest vested in it are translated anew; no longer simply a topic for researchers and IT consultants, Big Data emerges as a topic surrounding our common future.

This pattern is also seen in the appearance of T*he Economist*, a major business journal on international politics and economy news, and *nestea.org*, an independent charity organisation. devoted to solving the big social and economic challenges through technological innovation, signifies that Big Data has been translated into a topic of not just technological importance, but also of relevance to the broader topics of business and politics.

Another important observation is the emergence of the highly esteemed scientific magazine *Nature* and *Queue ACM* makes their entrance in this period. While the mapping by itself does not reveal the reason for this, we see, through comparing this inscription with our scientomet-

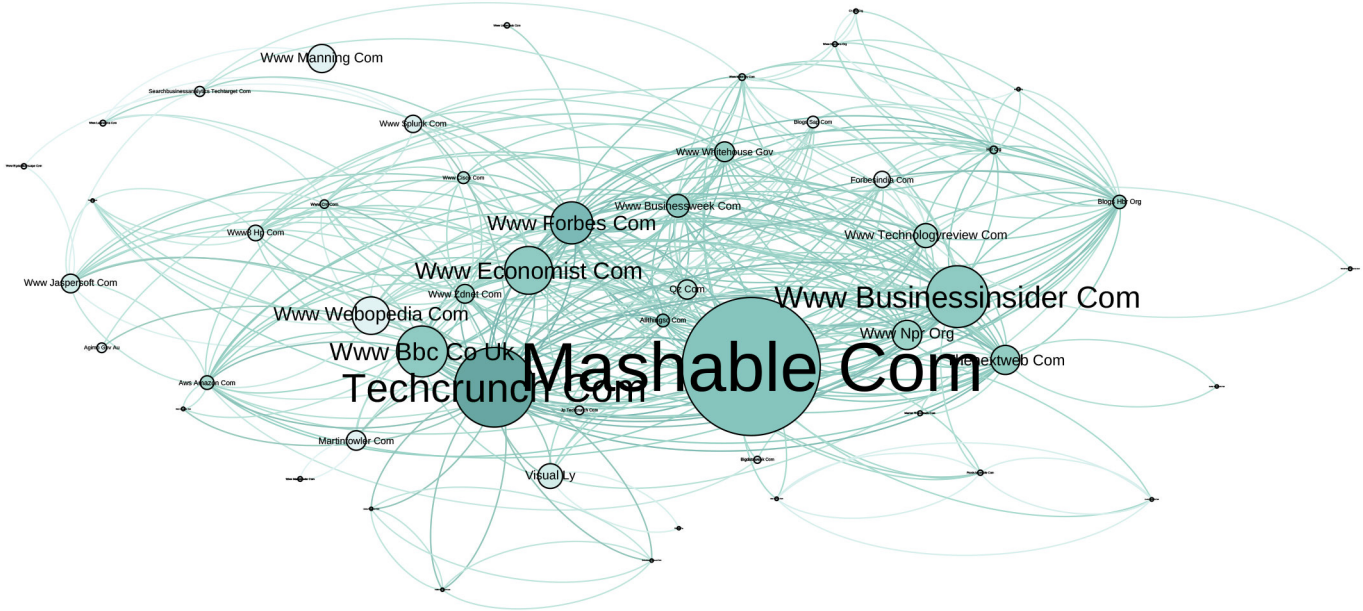
ric analysis, a possible indication: a 2008 special issue of Nature, *Big Data: Science in the petabyte era,* and a highly cited 2009 article in Queue ACM, *The pathologies of Big Data*. Another new science actor is *CRA.org*, an association of more than 220 North American computer science departments and affiliated with an early 2008 article, *Big-Data Computing* that appears to have been influential in the middle period. Also the appearance of *Danah.org*, the private page of Microsoft's Danah Boyd, co-author on *The six provocations on big Data* discussed earlier (see chapter. 4), makes its way into the map. In all these cases the actors appear to have attained their prominent position through few or even just a single referenced publication. This cements our suspicion that the aforementioned individual publications are indeed central in the development of Big Data, and point to the hypothesis that the major development of Big Data are centred around a relatively small number of organisations: a field has to be sparsely populated if a single text can enable such a central position.

Lastly, there is a fairly large and disjoint cluster around the sites for the conferences *Big Data meetup* and *Big Data analytics meetup.* Already in our Autocomplete experiments where "Big Data Conference" appeared as the 4th most related search term (see chapter. 4.3), we developed a suspicion that conferences served as important forums for bringing together researchers, entrepreneurs and government figures in the act of stabilising Big Data claims (or *facts*, Latour 1987). By mobilising a broad group of actors around Big Data, new relations could be stabilised (such as e.g. linking data analytics to Big Data as discussed before) which could enable the enrolment of a larger heterogeneity of actors. While our data material precludes further investigation of this phenomenon (the actual activity on the conferences being distinctively analogue), the centrality of Big Data conferences can be gleaned by the massive interests revealed by both search patterns and linking.

In the last period (figure 18) we encounter a number of the same entities apparent in the middle period. An extensive rearrangement in centrality however reminds us that Big Data as a term is experiencing an explosive growth during the period, as we found in chapter 4.

In the lower half of the map we see the same type of tech publishers that we found in the middle period, albeit different sites, the big hubs now being *TechCrunch*, *TheNextWeb*, *AllThingsD*, *Mashable* and *ZDNet*. Despite the appearance of these new actors, the similarity between these and the ones found in the middle period is close, which we read as the role of the tech publishers being unchanged, while the actors occupying the position might have been rearranged.

Figure 18: Out-Degree, Late period. Colouring based on eigenvector centrality. Visualization filtred by degree.



Another reappearance is IT-firms such *HP*, *Cisco* and *Amazon*. While these consisted of primarily small firms in the early period (with Microsoft as an exception) and nearly disappeared in the middle period, the group of IT firms reappears but now under the lead of the IT giants in the infrastructure market.

Both general news sites, e.g. *BBC* and *NPR*, and dedicated business journals such as *Forbes*, *Business Insider, BusinessWeek* and *The Economist* now occupy the most central positions. We also see *Harvard Business Review* (HBR), a journal for management studies and *McKinsey*, publishing their influential 2011 Big Data report in this period. Overall this impies that the translation of Big Data into a subject for businesses can be seen to have further intensified in this period.

5.2.3 Discussion

To answer the question of how the central actors change over time, this particular experiment enacts a picture of the biggest contributors of information. In the enactment, we see a clear shift over time from niche sites, over more consumer oriented technology news sites (normally billed as tech-writing) before adding mainstream media and business journals to the network. We also see a surge of conferences in the middle period and consultancies and management studies in the late. Parallel to this development we witness how a thicket of primarily smaller IT firms present in the early period make way for a few major infrastructure providers in the late period. Lastly, we witnessed how academia was quite absent apart from a brief stint in the middle period.

So we find a pattern where Big Data assembles IT insiders in the early period in the form of small firms and niche trade journals. Comparing with the growth curves in chapter 4, this is the low interest period before the massive growth in mentions started. We see this low interest mirrored in the actors presented here: what we see is the seed stage, the early development of Big Data.

In the middle period, Big Data starts to enrol a broader range of actors. Conferences abound, and we speculated as to their role in forming association and interessement across investors, researchers and clients. This is also the period where academic interest makes a brief foray, and the media coverage shifts to consumer oriented tech publications. Comparing once more to the growth curves in 4, this is the period of initial interest. We see here the beginning of hype, as Big Data begins to attract interest beyond computer specialists and build a somewhat broader audience.

In the late period we see Big Data gain hold as a business thing, both in the presence of management consultancies but also the media coverage in business journals. That media coverage extends to general newspapers shows that this is the period where the conversation on Big Data becomes ubiquitous. In terms of IT firms, the vast undershrub of smaller firms has made way for a selection of cemented giants. Comparing with the growth curves in chapter 4, this is the period where Big Data grows the fastest, which aligns with the depiction here of Big Data as a widespread phenomenon, associating a very broad range of actors.

But for now the only actors are traditional: persons, organisations and publications. We have yet to assemble the associations between keywords, terms and buzzwords which will be the subject of chapter 5.5.

First we will however change our focus to study *who is referred to,* since links always runs between two points, studying a site's outbound links is only half the story.

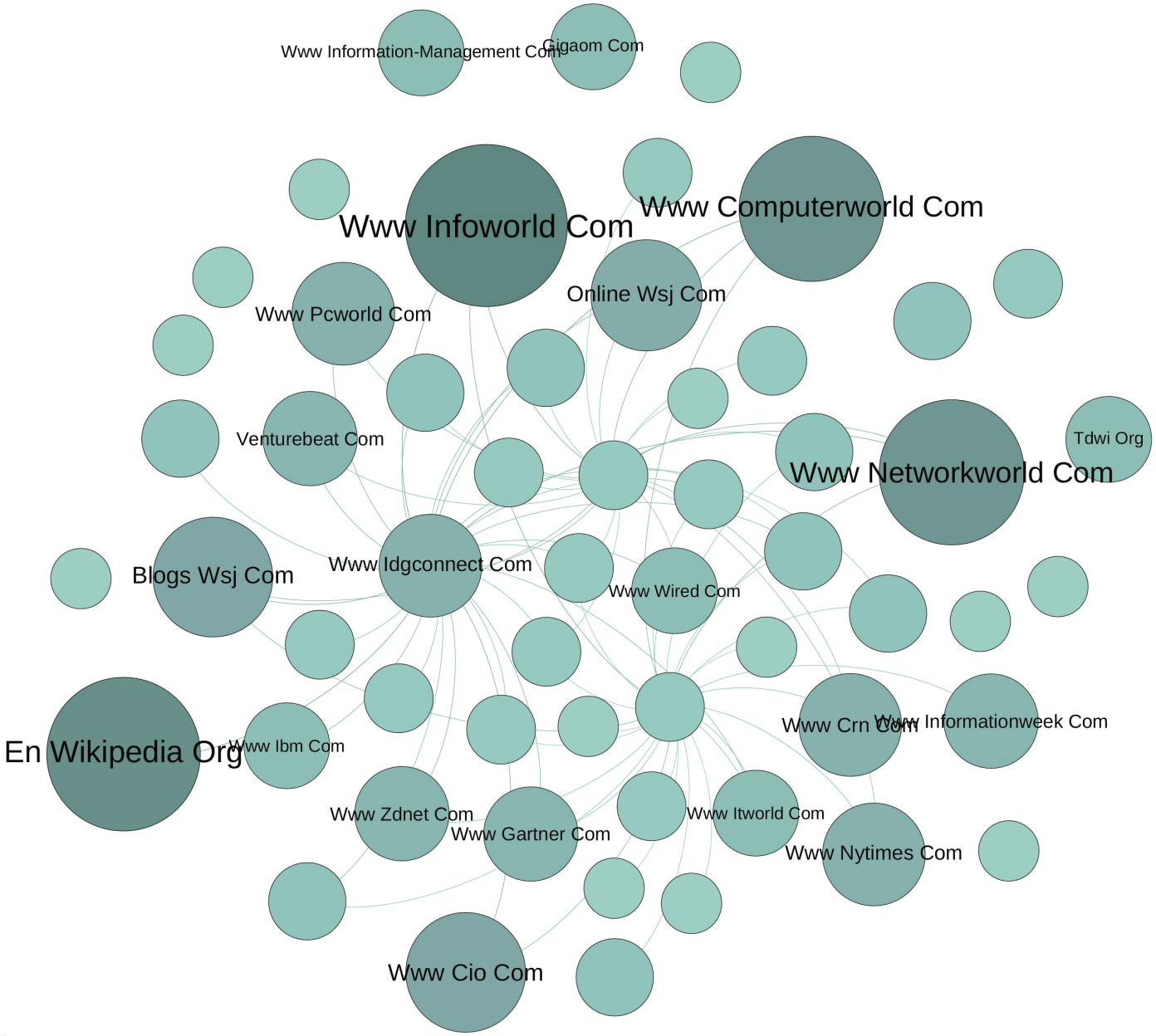
5.3 HYPHER: IN-DEGREE

While the previous analysis gives us some idea of the propagators of Big Data, it does not provide us with any knowledge on which actors are being spoken about: What do people talk about when they talk about Big Data? Who are the propagators referred to as the representatives (Callon 1986) of Big Data?

To gain some idea of the answers to these questions, we switch the rankings of our datasets to ingoing (in degree) links and compare them over time.

Figure 19: In-Degree, Early Period. Colouring based on in-degree. Visualization filtred by degree.

Starting with the early period (figure 19) we notice how the three major nodal points seen in the inscriptions of the previous chapter decreases in centrality. The IDG-cluster is somewhat diminished, while Microsoft and OECD's influence are reduced to a level where their names does not even appear on the map. The central point of reference is instead Wikipedia[4] and a number of general news sites such as *Wall Street Journal* and *New York Times*. The fact that references (links) are primarily directed towards sites with very generalised knowledge implies that the discussion on Big Data is still in a very early stage - the ranking of Wikipedia could stem from articles referencing it as a way of introducing newcomers to the term Big Data.



Another interesting newcomers are *Gartner*, a global IT consultancy behind the earlier discussed *hype cycle*, who has been surprisingly absent in our other experiments. Shifting to a qualitative level, we zoom in on this actor by searching through their publication list and discover that a 2001 Gartner report[5] is the origin of the 3 V's, a ubiquitous dogma of Big Data research ascribing the rise of Big Data to a change in Volume, Velocity and Variety; a quite plausible explanation for the high number of actors referencing Gartner in the early period, when few other reports on Big Data had even been published.

Shifting to the map of the middle period (figure 20) we notice how *Techcrunch* and *Gigaom* as central actors are mirrored from the in-degree map. Beside these two sites, the conclusion is as with the early period map, that the ones who link out and the ones who are linked to are seldom identical.

In comparison to the out degree map we witness a change in focus from dedicated science sites (Nature, Queue ACM etc.) to more popular science coverage. Most importantly *Wired* stands out, most probably due to the special issue on the petabyte age and Chris Anderson's article on *The end of theory,* also significant in our scientometric analysis and a focal point for the theoretical Big Data debate in the social sciences.

Again we observe a much more business-oriented focus than the out-degree maps. Though the economist disappears, business journals such as *Forbes*, *Bloomberg, Wall street Journal*, *Business Week* and *CNN Money* are all referenced. The centrality of these business journals, together with the high ranking of regular news sites like *New York Times, The Register* and *The Guardian* are puzzling. Our intuitive understanding posed that they would be the ones who provided the outbound links referencing experts and analysts'. However, an explanation could be that some translations are able to interest a broader group of actors than others. Though professionals working with Big Data might prefer to cite scientific articles published in Nature, these citations might be drowned in quantity by the majority of internet users - normal folks - who are more interested in an intelligible New York Times article than in the obscure discourse from academic journals.

We also observe the appearance of three providers of IT Infrastructure: *Google*, *Microsoft* and *Amazon*. This stands in sharp contrast to our previous out-degree maps where the IT-firms were completely absent in this period. While Amazon and Microsoft were also visible in the out-degree map we once again see that Google, who only appears in the in-degree maps, is often talked *about*, but do not themselves talk about Big Data, a point we will discuss further in our later analysis. We also note the appearance of *Github*, an online repository for collaborative open source code projects. This indicates that Big Data has gone from something demanding gigantic computational power and primarily limited to big organisations and translated into a practice for which single users in collaboration with others can develop new software.

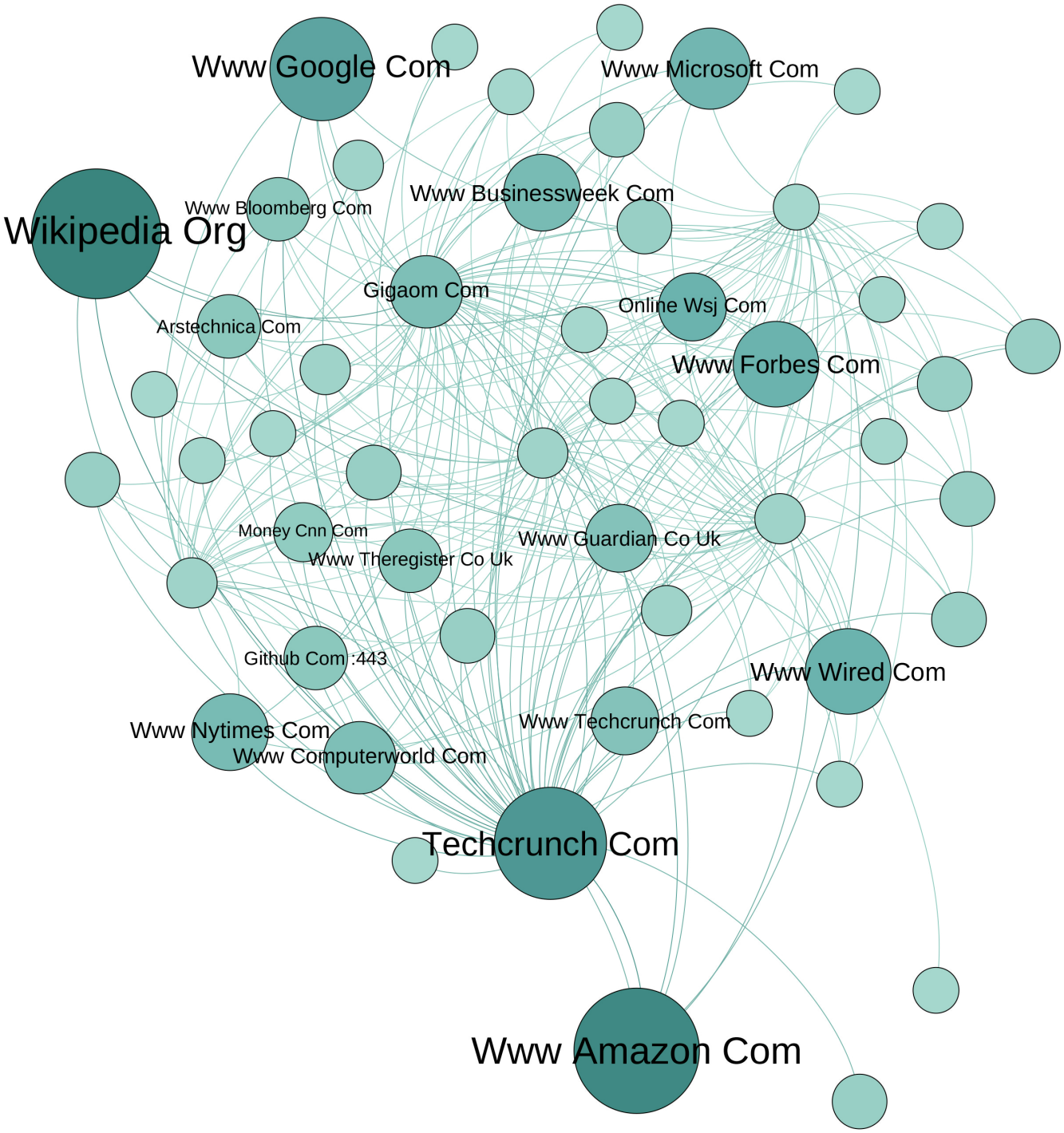


Figure 20: In-Degree, Middle Period. Colouring based on in-degree. Visualization filtred by degree.

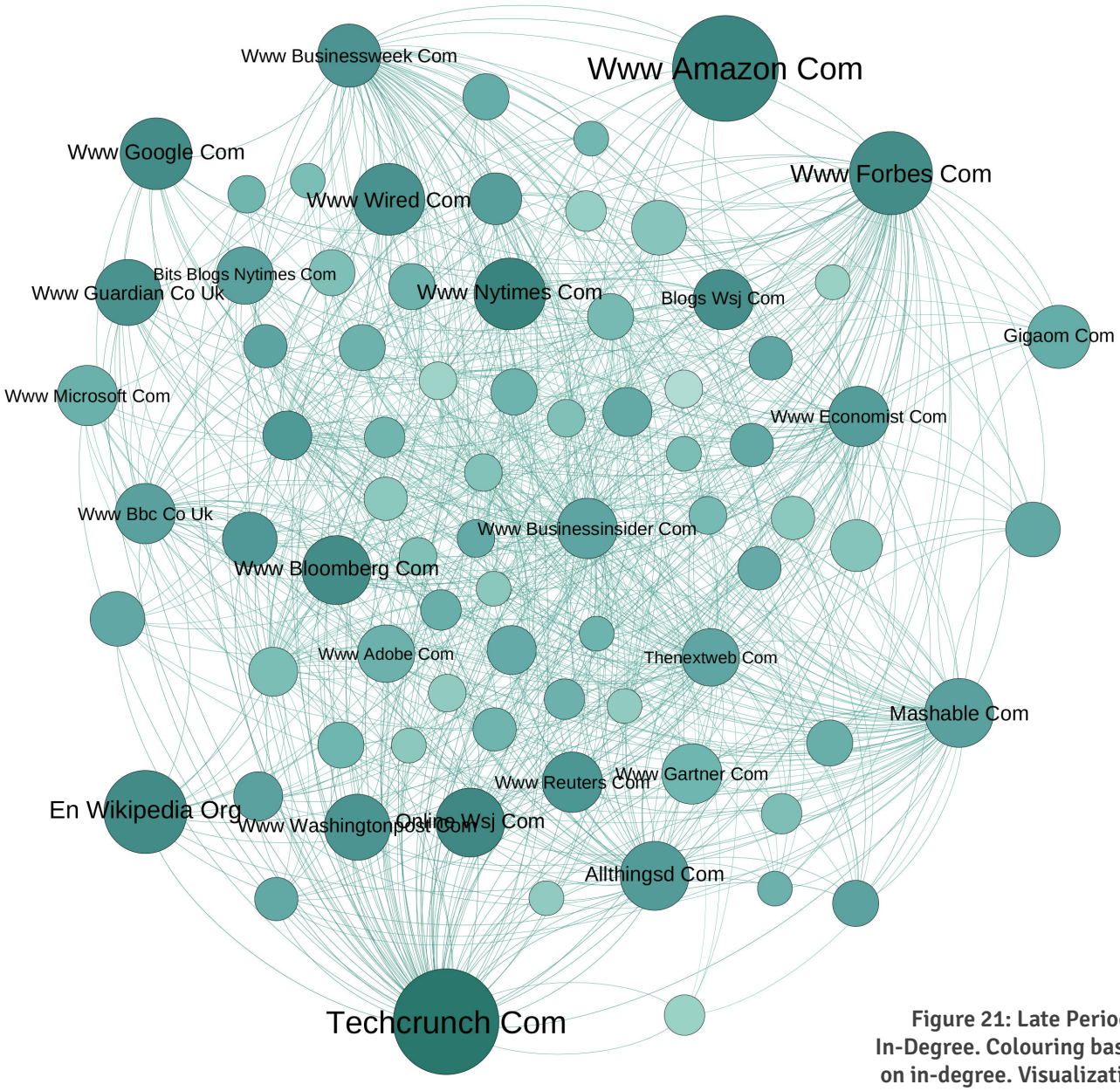


Figure 21: Late Period – In-Degree. Colouring based on in-degree. Visualization filtred by degree.

In the late period (figure 21) we witness an increase in the number of central actors, but very few new types of actors. This implies that the positions and roles might have stabilised, and the colonisation of new areas under the umbrella term of Big Data have decreased in speed, while still being able to attract more attention as seen in Ch. 4.

Again *Google* and *Amazon* are central, while Microsoft, as shown in our out-degree mapping, apparently has left the race. A shift which could be related to Microsoft abandoning their own Big Data product Dryad and shifting their efforts to Hadoop in this period (Jo Foley 2011).

A single new interesting actor is the IT firm *Adobe*, mostly known for its graphical IT tools, who enters the field most probably due to acquisition of the analytic firm Omniture in 2011 and the later launch of their Big Data service *Adobe analytics*. Other relevant changes is that the idea of creating one's own Big Data software (represented with Github) cease to exist and Gartner, absent in our 2008 mapping reappears. Overall are the changes in actors however limited with the map still dominated by business journals, general news sites and a number of tech giants.

Another interesting newcomers are *Gartner*, a global IT consultancy behind the earlier discussed *hype cycle*, who has been surprisingly absent in our other experiments. Shifting to a qualitative level, we zoom in on this actor by searching through their publication list and discover that a 2001 Gartner report[5] is the origin of the 3 V's, a ubiquitous dogma of Big Data research ascribing the rise of Big Data to a change in Volume, Velocity and Variety; a quite plausible explanation for the high number of actors referencing Gartner in the early period, when few other reports on Big Data had even been published.

Shifting to the map of the middle period (figure 20) we notice how *Techcrunch* and *Gigaom* as central actors are mirrored from the in-degree map. Beside these two sites, the conclusion is as with the early period map, that the ones who link out and the ones who are linked to are seldom identical.

In comparison to the out degree map we witness a change in focus from dedicated science sites (Nature, Queue ACM etc.) to more popular science coverage. Most importantly *Wired* stands out, most probably due to the special issue on the petabyte age and Chris Anderson's article on *The end of theory,* also significant in our scientometric analysis and a focal point for the theoretical Big Data debate in the social sciences.

Again we observe a much more business-oriented focus than the out-degree maps. Though the economist disappears, business journals such as *Forbes*, *Bloomberg, Wall street Journal*, *Business Week* and *CNN Money* are all referenced. The centrality of these business journals, together with the high ranking of regular news sites like *New York Times, The Register* and *The Guardian* are puzzling. Our intuitive understanding posed that they would be the ones who provided the outbound links referencing experts and analysts'. However, an explanation could be that some translations are able to interest a broader group of actors than others. Though professionals working with Big Data might prefer to cite scientific articles published in Nature, these citations might be drowned in quantity by the majority of internet users - normal folks - who are more interested in an intelligible New York Times article than in the obscure discourse from academic journals.

We also observe the appearance of three providers of IT Infrastructure: *Google*, *Microsoft* and *Amazon*. This stands in sharp contrast to our previous out-degree maps where the IT-firms were completely absent in this period. While Amazon and Microsoft were also visible in the out-degree map we once again see that Google, who only appears in the in-degree maps, is often talked *about*, but do not themselves talk about Big Data, a point we will discuss further in our later analysis. We also note the appearance of *Github*, an online repository for collaborative open source code projects. This indicates that Big Data has gone from something demanding gigantic computational power and primarily limited to big organisations and translated into a practice for which single users in collaboration with others can develop new software.

In the late period (figure 21) we witness an increase in the number of central actors, but very few new types of actors. This implies that the positions and roles might have stabilised, and the colonisation of new areas under the umbrella term of Big Data have decreased in speed, while still being able to attract more attention as seen in Ch. 4.

Again *Google* and *Amazon* are central, while Microsoft, as shown in our out-degree mapping, apparently has left the race. A shift which could be related to Microsoft abandoning their own Big Data product Dryad and shifting their efforts to Hadoop in this period (Jo Foley 2011).

A single new interesting actor is the IT firm *Adobe*, mostly known for its graphical IT tools, who enters the field most probably due to acquisition of the analytic firm Omniture in 2011 and the later launch of their Big Data service *Adobe analytics*. Other relevant changes is that the idea of creating one's own Big Data software (represented with Github) cease to exist and Gartner, absent in our 2008 mapping reappears. Overall are the changes in actors however limited with the map still dominated by business journals, general news sites and a number of tech giants.

5.3.4 Discussion

On the most general level we found the in-degree maps to be more similar across the different time periods than our previous mapping of out-degree, which could be an indication of the temporal noise we discussed in our introduction. This was especially evident between the middle and the last period, with very little change in the overall composition of the map. Eventual temporal contamination would tend to have a greater effect on in-degree maps since the last degree of the crawl only contains incoming links (see figure 14). While the opposite holds true for the seed URLs (they only contain outgoing links), the effect is much bigger on the final degree because of the exponential growth in numbers of sites for every increase in degree. The temporal separations of the in-degree map should therefore be taken more as an approximation.

This limitation does not mean that our in-degree maps did not produce interesting findings. Especially interesting is the general high ranking of authoritative commentators such as *HBR*, *Wired*, *WSJ*, *Gartner* and *Forrester* in comparison to both regular media sites but especially scientific and technological contributors, which were both markedly reduced in comparison to our outgoing maps. This shows that actors who simply report and convey news are downplayed and actors that provide analyses and commentary overtake their place. This is relevant, to the degree that it exemplifies an inkling that most of the popular discourse on Big Data is not about concrete deployments, but about ideas and opinions. People do not talk about the firm that made agnostic medical sampling techniques; they talk about the visionary Chris Anderson forecasting the end of theory. One explanation for this finding could be the highly technical nature of the subject where most first order observations are far too complex for popular understanding. Instead, Big Data has to be translated into easily accessible forms in order to generate broad attention. Another possible explanation is related to the architecture of news sites, delicately tweaked to give them a them high visibility in the socio-technical mode of seeing that Google's SERP, and by extension we, employ.

To move around this possible skewing of our ranking mechanism and to enact yet another take, we leave our hyperlink data and turn to ANTA and the analysis of text corpora.

5.4 ANTA: ORGANISATIONS

While being talked about and being linked to is obviously different, our previous studies do not provide us with any knowledge on how this difference affects the distribution of actors: are the heavily interlinked actors also the actors who are mentioned in the thousands of Big Data reports and books? And how do these mentions cluster around certain actors? Finally, do the enactments chang when we broaden the scope of actors from primarily organisations, technologies, web sites and includes the now overshadowed terms and keywords?

To answer these questions and widen of our scope to include terminological actors we will in the following conduct a semi-automated text-analysis based on the Actor Network Text Analyser (ANTA) software package. The analysis is based on 500 text pieces scraped from general news, science databases, white papers and reports and business journals (see protocol E). ANTA was then used to identify central terms in the texts and to categorise the types of entities as e.g. field terms, organisations, persons and technologies, divided into our time periods (early, middle and late). The data was then exported to Gephi for visualisation.

In relation to the crawler we employed in the earlier phase, this tool has a much more controlled approach. Primarily we can control its trajectory much more closely, since it does not depart from the text files we give it as input. As

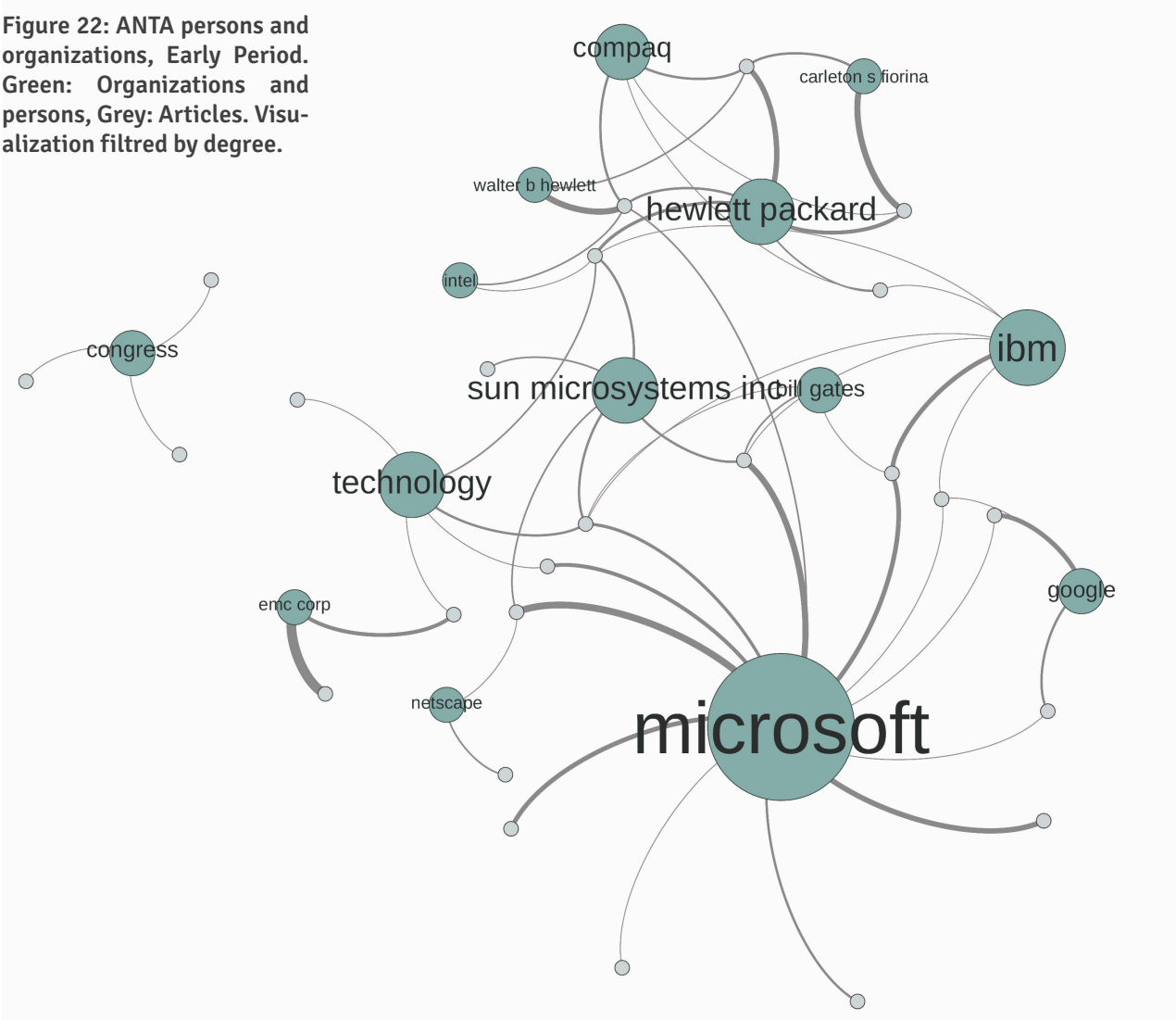


Figure 22: ANTA persons and organizations, Early Period. Green: Organizations and persons, Grey: Articles. Visualization filtred by degree.

such, it does not fall prey to the temporal noise that plagues crawlers as we have just discussed. We can be sure that the list of entities it supplies was relevant in the time period given, since they have to be mentioned in a text that we have confirmed to be from that period. Instead this appraoch however opens up for temporal noise in the selection of relevant articles (see appendix 9.2)

In the first period (figure 22) we find the focus on *Microsoft*, similar to what we saw in the map of out-degree counts in 5.2. But also the other big players in tech - *IBM, HP, Sun, Google* - who were mostly missing in the earlier inscriptions. All of these firms have in common that they are large, heavily invested in research and active developers of Big Data services. So their presence in this map (and absence in others) tells us that while they may not have produced sufficient novelty value to garner new links, other actors nonetheless mentioned them. As such, we find that the assemblage drawn by this method prioritises actors who contribute directly or indirectly to Big Data rather than the earlier inscriptions that put news sites and propagators of information at the front.

In the middle period (figure 23) *Google* rises to prominence and becomes the biggest actor, while Microsoft slinks away. In relation to what we saw in 4, this is probably due to the publication of the articles on Mapreduce and Hadoop and Microsoft shortly after abandoning their own

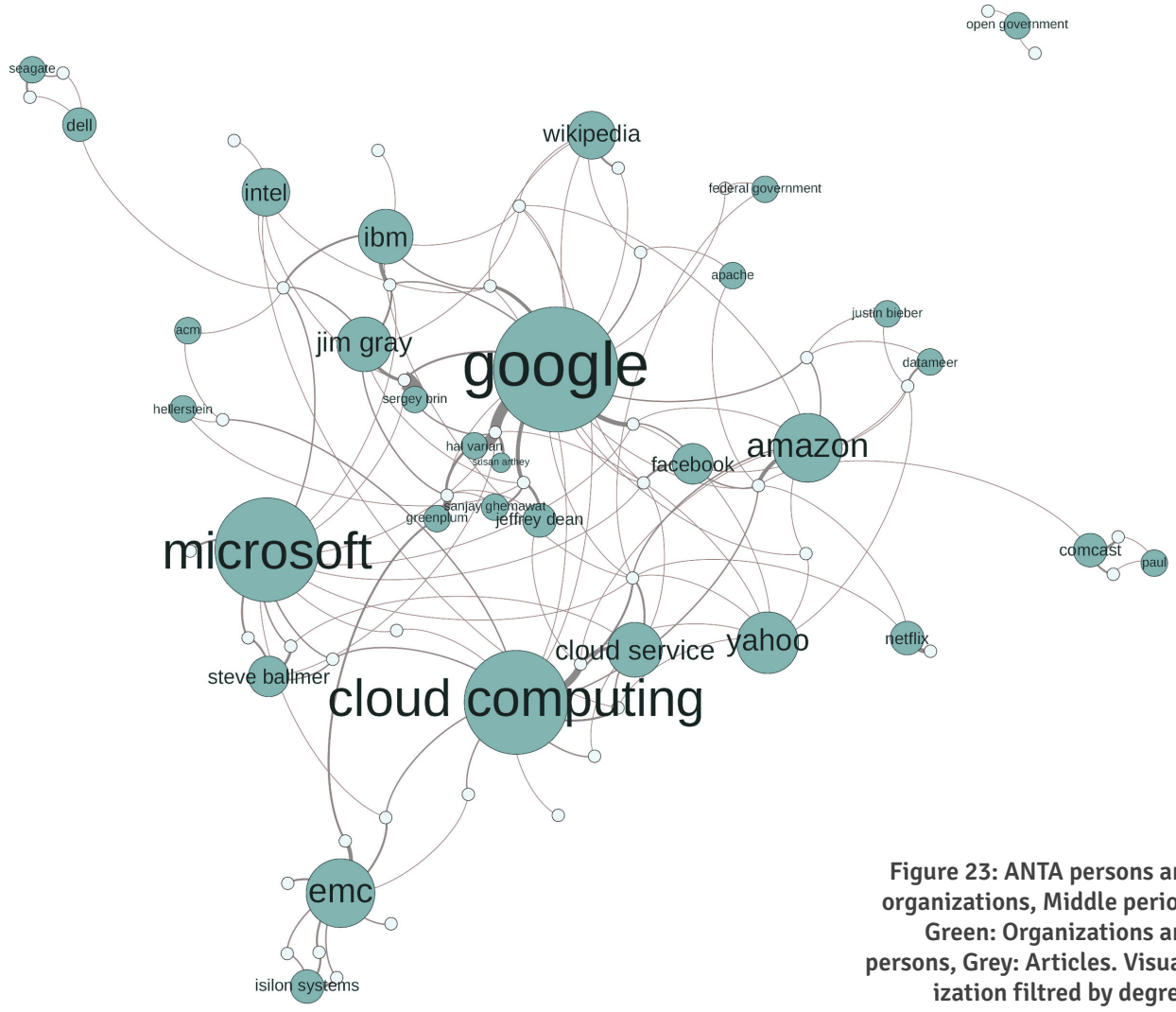


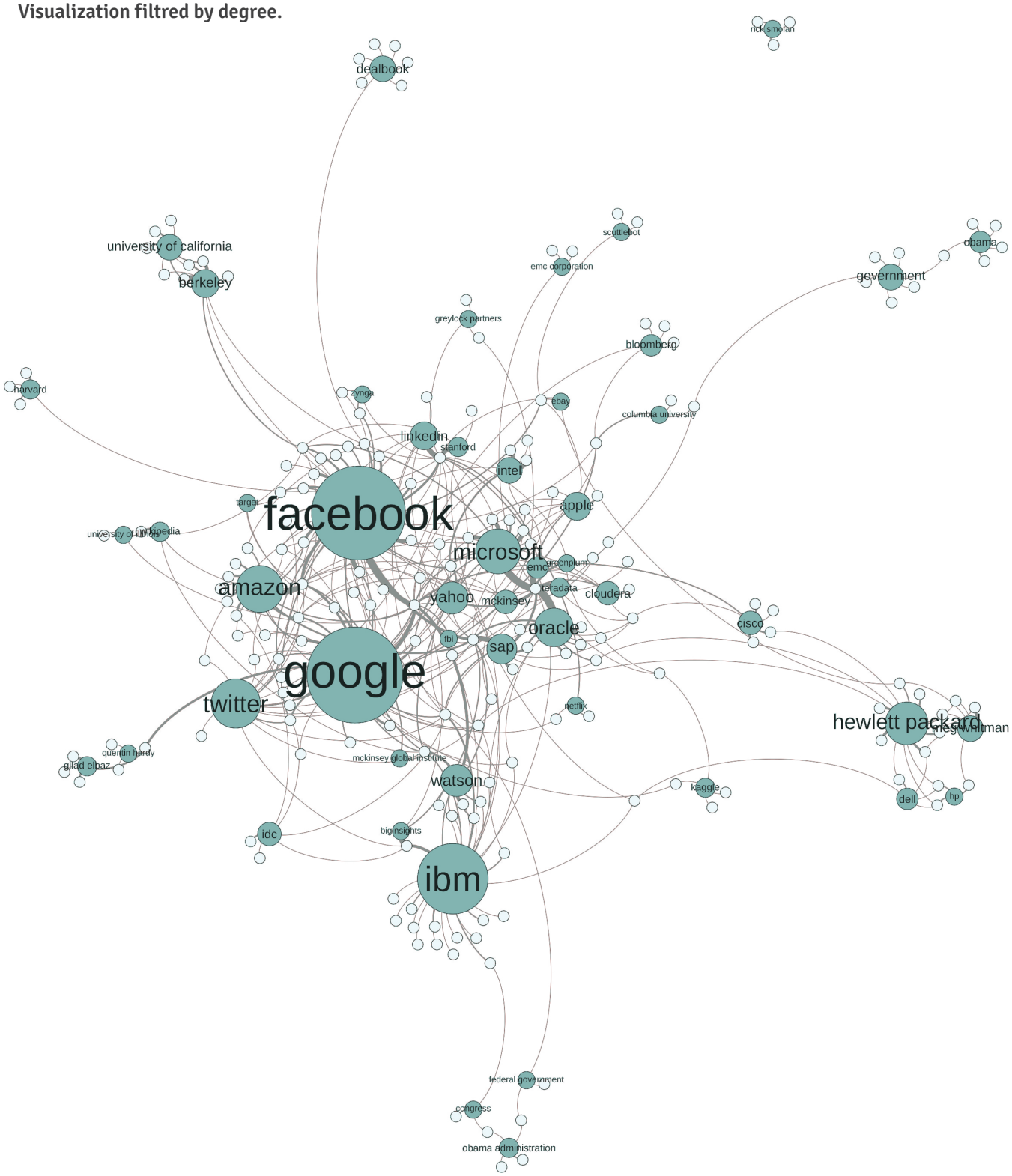
Figure 23: ANTA persons and organizations, Middle period. Green: Organizations and persons, Grey: Articles. Visualization filtred by degree.

Big Data product Dryad (see chapter 5.2). Interestingly, this is the first time that we have seen Google as an organisation gain such a central position; especially compared to their stark absence in chapter 5.2. Why they do not appear in the other inscriptions will be explored later.

Other organisations also include *Amazon* and *Netflix*. Netflix are not purveyors of Big Data, they do not sell a service as such. So their appearance here could imply that they are highlighted as an example of what Big Data can do, they are translated into representatives for others agendas in order to further it. Amazon's figuration is naturally based on their role as one of the first offers of Big Data infrastructure (under the name of AWS). Zooming to a qualitative level however reveals how the recommendation engine underlying the e-commerce has been enrolled as Big Data due to its reliance on large amounts of purchasing and browsing behavioural-data from customers in order to attempt to show them relevant products. Both Amazon and Netflix hereby represent stories of successful applications of Big Data. Their high ranking in mind, this could indicate that inscribing other organisations as examples of the potential success of Big Data is a common way for actors to strengthen their networks.

In the late period (figure 24) we see a huge growth in the mentions of social media giants: Facebook, Twitter and LinkedIn. Originally, we filtered these sites from our Hypher maps, since we assumed their centrality was due to links to content *on* the social networks, not the social networks themselves. This does however not seem to be the case, since this current inscription relies on a method that pays no heed to such measures when assembling its bundle of associations, yet still sees them emerge as central entities. Instead this might be explained by the enormous amount of traffic these services drive and is a key source for the sort of behavioural data that Big Data is computed on, and especially in regard to the more commercial side of Big Data, for which the behavioural data is key to unlocking consumer preference and thereby improving their targeted ads. Their appearance may thus stem from a translation of these social media sites into sources of raw input for Big Data. But as we saw in the map before, the power of examples and representation can be a powerful tool used in the forging of alliances and associations between actors. In this context, the social networks could also be held up as an exemplification for a more general datafication of society. Whether they figure purely as examples or as a source of actual data is beyond our view, but the late periods enrolment of social media in the discussion of Big Data are not.

Figure 24: ANTA persons and organisations, Late period. Green: Organizations and persons, Grey: Articles. Visualization filtred by degree.



5.4.1 Discussion

Comparing the development over time, we see a shift in rankings from firms who develop Big Data infrastructure and analytical tools in the first period (HP, Microsoft, Sun, IBM), to firms for whom data was the backbone of their service in the second[6] (Google, Amazon, Netflix), before finally in the last period those who scour their users' data for profitable insights and give a service in return (social networks). This general shift mirrors the development we found in chapter 5.2; the growth in mentions and interest in Big Data is accompanied by a translation of the term that changes the associated actors from highly technological niche actors into gradually more mainstream entities.

Compared to our scientometric analysis, the IT giants talked about in the early period were quite similar to the ones we found were funding the scientific articles: Microsoft, HP and IBM. While we took it as an indicator of the intermingling of commercial and academic actors then, in this context (where the entities are representing which actors are talked about in general) it makes a slightly different point: that the early debates were likely centred around basic research into the capabilities of Big Data. This is seen in the inclusion of the previously omitted Google, who was behind the altogether most cited article, and had been using the technique for many years in a fully operationalized deployment, but who did not publish until 2008.

That we did not see them as particularly central in any of our earlier mappings, neither scientometrics nor out-degree Hypher crawls, shows that they themselves do not associate with Big Data. Their slight appearance from 2008 in both In-degree maps, and to an even higher degree in our keywords mappings, shows that others nonetheless refer to them as central actors in Big Data.

So even though they were actually deploying Big Data on a massive scale at a time when others mostly dreamed about it, they were not enrolled in the semantic assemblage before the publication of their 2008 article, and appear only sparsely afterwards. More so, they only appear through others associating them, not themselves.

In order to enquire into this seeming discrepancy, we shift once again to a qualitative mode. Here we find that nary a mention of the term Big Data is found on neither Google's homepage nor their corporate communication material (Regalado 2013). Further searching uncovered a story, where a Google PR representative "*was hesitant to participate in a story tied to the term "Big Data." They'd prefer, they said, not to be associated with it. Why? I asked. "It's too Big Brother-ish," came the answer*" (ibid.).

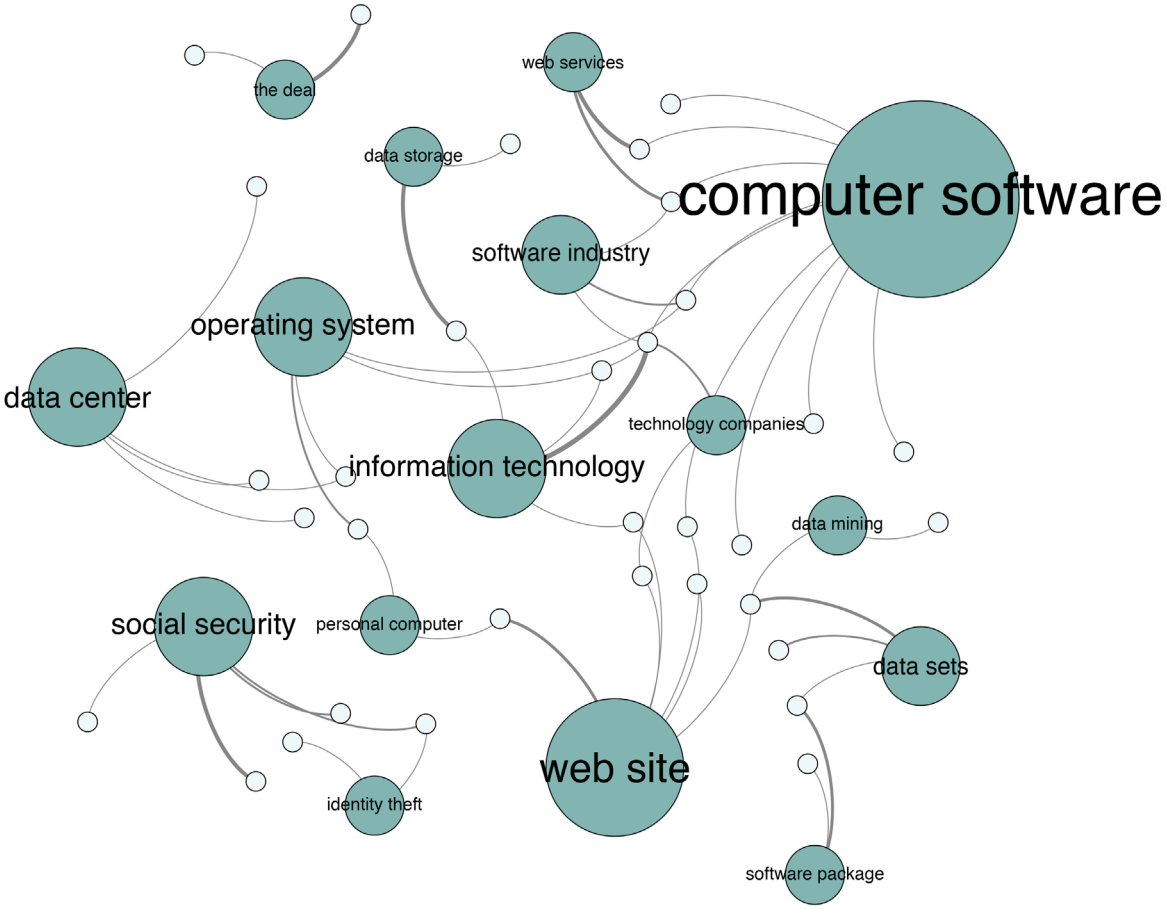
So a central actor chooses to distance itself from the very same phenomenon the rest of the actors are trying to associate themselves with. This anti hype might be due to whom Google are trying to interest: the customers who must willingly supply their personal data. We find it interesting that the semantic connotations of the term Big Data here might serve to scare away, rather than mobilise their allies, and they therefore choose not to use the term.

This leads us to consider the semantic construction of Big Data. To see which terms were translated and enrolled in Big Data over time, we turn to a feature in ANTA that discovers "field terminology".

5.5 ANTA: KEYWORDS

Well arrived in the early period (figure 25) we see how most of the terms listed are very broad, general IT terms, e.g. *web site*, *technology companies*, *software package* and *computer software*, which indicates that Big Data has yet to mobilise its own assemblage of keywords, and merely uses already existing ones. A few of the terms that later gains a central position, such as *data mining* or *data storage*, are however present, indicating the first steps towards establishing a vocabulary.

Figure 25: ANTA Keywords, Early period. Green: Field termiology, Grey: Articles. Visualization filtred by degree.



In the middle period (figure 26) we see a slight maturation of such a vocabulary. Data storage has gone from an outlier to now occupying the centre of the map, terms more specifically related to Big Data, such as *Data management* and *Cloud computing* are also present, and some of the more general terms have faded away.

Overall, we witness how the technological keywords have gained a certain specificity. An example of this movement is the early periods mention on *web services*, which in the middle period have been taken over by *cloud service* and *cloud computing*, similar but more specific and matured terms. That these are in fact related is visible by their spatial location next to web services indicating that the terms occur in the same documents.

Another example of such specification is the almost overarching broad term *operating system* from the early period, linked to the term *personal computer*. In middle period we see this term embedded in a cluster of highly technical and quite uncommon terms: *distributed computing*, *distributed systems*, *parallel processing*, *Mapreduce*. This clustering implies that a general discussion on the performance of single computer operating systems has been translated into a discussion on ways of distributing the operation across multiple computers. Just like cloud computing, these terms represent more concrete technologies to handle Big Data. The growth in specificity that we identify over time is thus centred around technologies or ways to practice or handle Big Data sizes.

The late period (figure 27) continues this process of specifying the discussion on Big Data. In contrast to the specification we witnessed in the middle period, the late period is marked by the arrival of a host of keywords for possible applications of Big Data under titles such as *public health*, *security solutions*, *financial services*, *social security*, *energy efficiency*, *mobile commerce*, *business intelligence* and *presidential election*. The discussions on Big Data is thereby translated into a format where scenarios of usage are being crafted - a picture further strengthened by the advent of a range of keywords pertaining to the economic potential of Big Data, e.g. *venture capital* and *stock markets*, indicating a translation of Big Data into an investment object.

Figure 26: ANTA Keywords, Middle period. Green: Field termiology, Grey: Articles. Visualization filtred by degree.

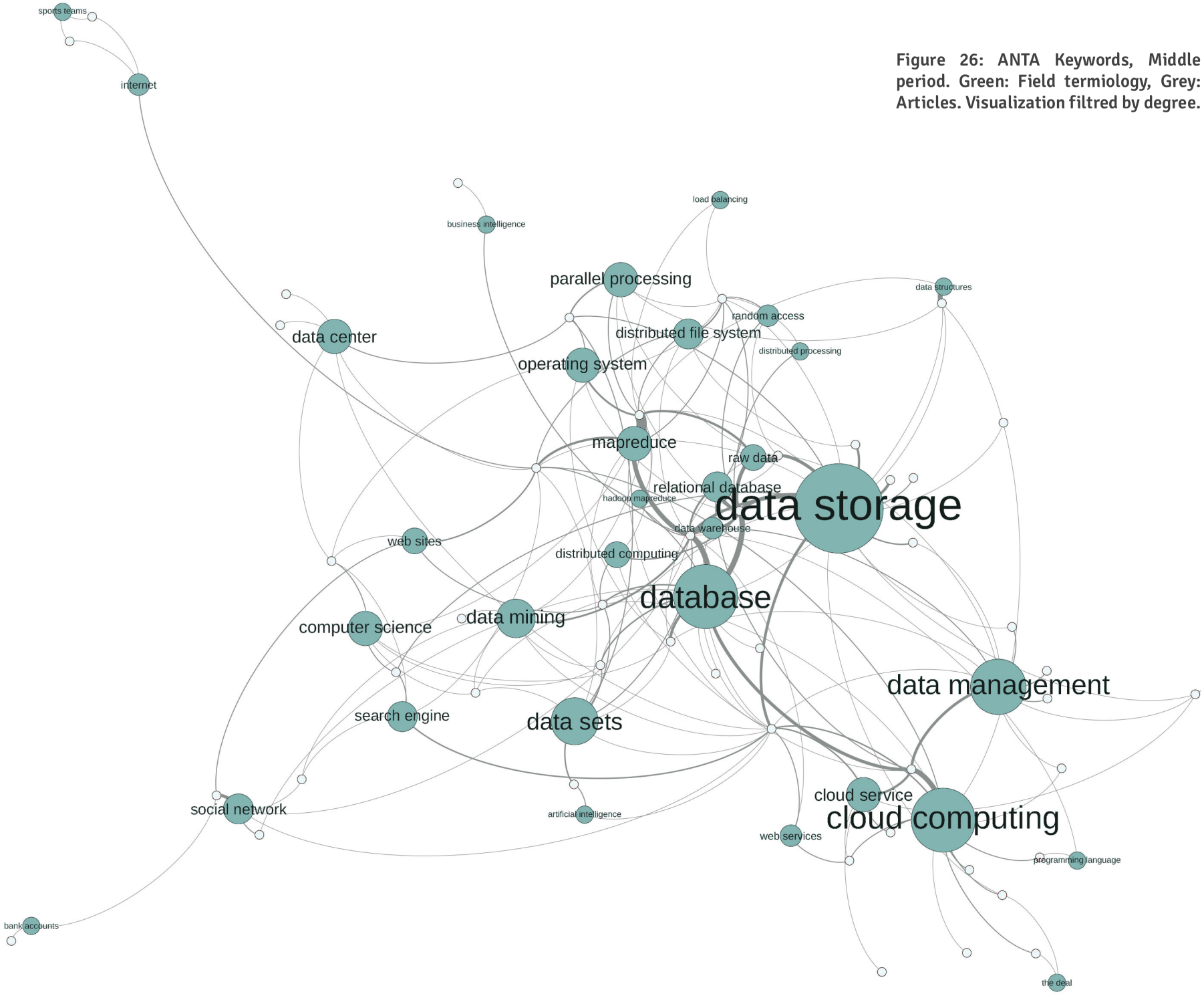
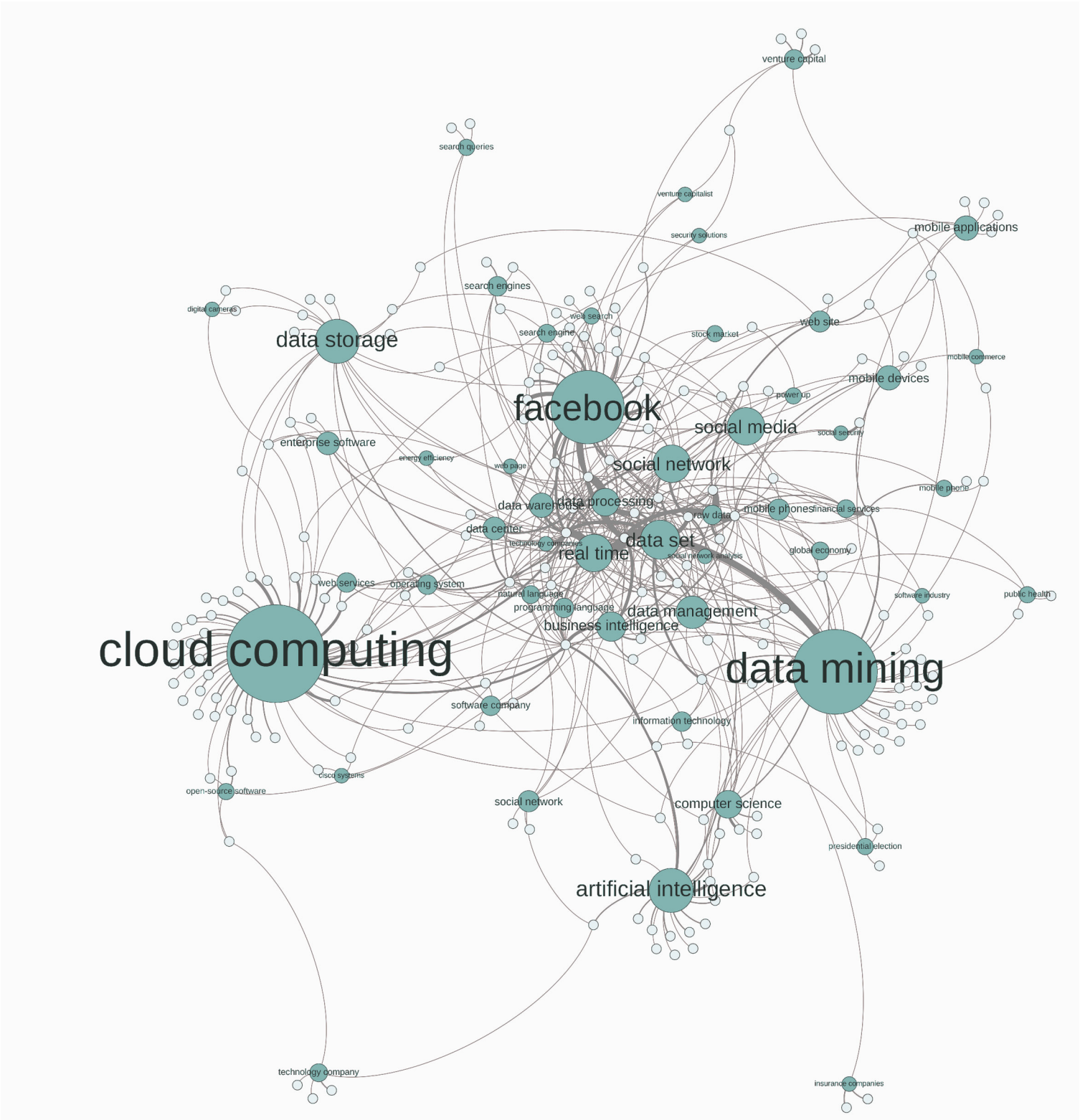


Figure 27: ANTA Keywords, Late period.
Green: Field termiology, Grey: Articles.
Visualization filtred by degree.



5.5.1 Summary

In the previous analysis we explored how the terms (or actors) associated with Big Data changed over time. We witnessed how a vocabulary for Big Data was established distinct from general IT terms, assembling its own distinct assemblage of terms. This is slightly ajar to our preconceptions of how the chain of translations would evolve; we expected Big Data to start out as something quite specific and then gradually gain a more general meaning in order to enrol more actors. But as we see, Big Data translates into more areas by developing specific meanings and terminologies for each, not a singular all-encompassing form.

Tracing the textual network over time has provided us with a picture of the change of Big Data from an un-established terminology into a broad and more stable vocabulary – though still undergoing a constant translation. These translations are naturally not only the result of differences over time, but are equally to be ascribed to differences between clusters of actors each trying to redefine Big Data in their own image.

To explore these clusters' attempts to translate Big Data in a certain way, we will in our last digital experiment trace how the development of keywords varies based on the source of data. Before we reach this point, we will however take a slight intermezzo to coalesce the mass of actors into a slight ordering.

5.6 ANTA: KEYWORDS OVER SPHERES

In all of our earlier experiments, we found clusters of actors with a certain likeness: clusters of interlinked websites, who when zoomed in on reveal a qualitative similitude and create thematic clusters in our mapping. Before we continue, we will create a number of thematic divisions to trace how the enactment of Big Data changes based on the actors propagating it, similar to how we divided our data into three temporal periods.

Concretely, we divide our actors into three groups based on our preliminary mappings (see appendix 9.2):

- General news *(New York Times)*
- Science *(Scopus)*
- Business *(HBR + Whitepaper)*

Without further ado[7], we denote the different types of actors as *spheres.* By sphere we hereby refer to a certain class of actors who all function by the same operating principle[8]. For each sphere we have also identified a data source to mine based on 1) our preliminary findings and 2) a pragmatic evaluation of data availability by e.g. chossing New York Times since they are currently the only global news media with an open API.

In order to gain a view of the relative importance and influence of these spheres over time, we start by mapping out their numbers of publications year by year (see figure 28)

From this simple count of publications in the different spheres we notice especially two aspects on the table. Firstly, we reconfirm our findings in chapter 4 of the development of Big Data growing slowly in the first two periods before exploding in the last period. This late arrival of the majority of publications also supports our early thesis of a relatively small number of articles defining and paving the way for the rapid development of Big Data in the late period (see chapter 5.2).

Secondly, the late but intense appearance of HBR articles recreates the depiction in our previous analysis of an increasing interest for possible real life application of Big Data and a growing interest for Big Data as an object of investment - both themes common to business journals.

But how do the spheres talk about Big Data?

To further explore this, we want to investegate how the enactment of Big Data differs based on the propagator. To do this we map the occurrence of key terms, not over time, but across the different spheres.

While we do see central terms entirely disappearing from some spheres, not surprisingly many of the same keywords appear across the spheres since it indeed still is the same term, Big Data, we filter their output by. What should instead be our focus is how the rankings vary quite distinctly from sphere to sphere, which demonstrates how Big Data ontologically changes based on the spheres enacting it.

Figure 28: Publication based on source and period

Sphere	Source	2001-2007	2008-2010	2011-2013
General news	NYTimes	37	38	321
Sceince	Scopus	48	75	849
Business	HBO	0	1	350
	Whitepapers	0	6	8

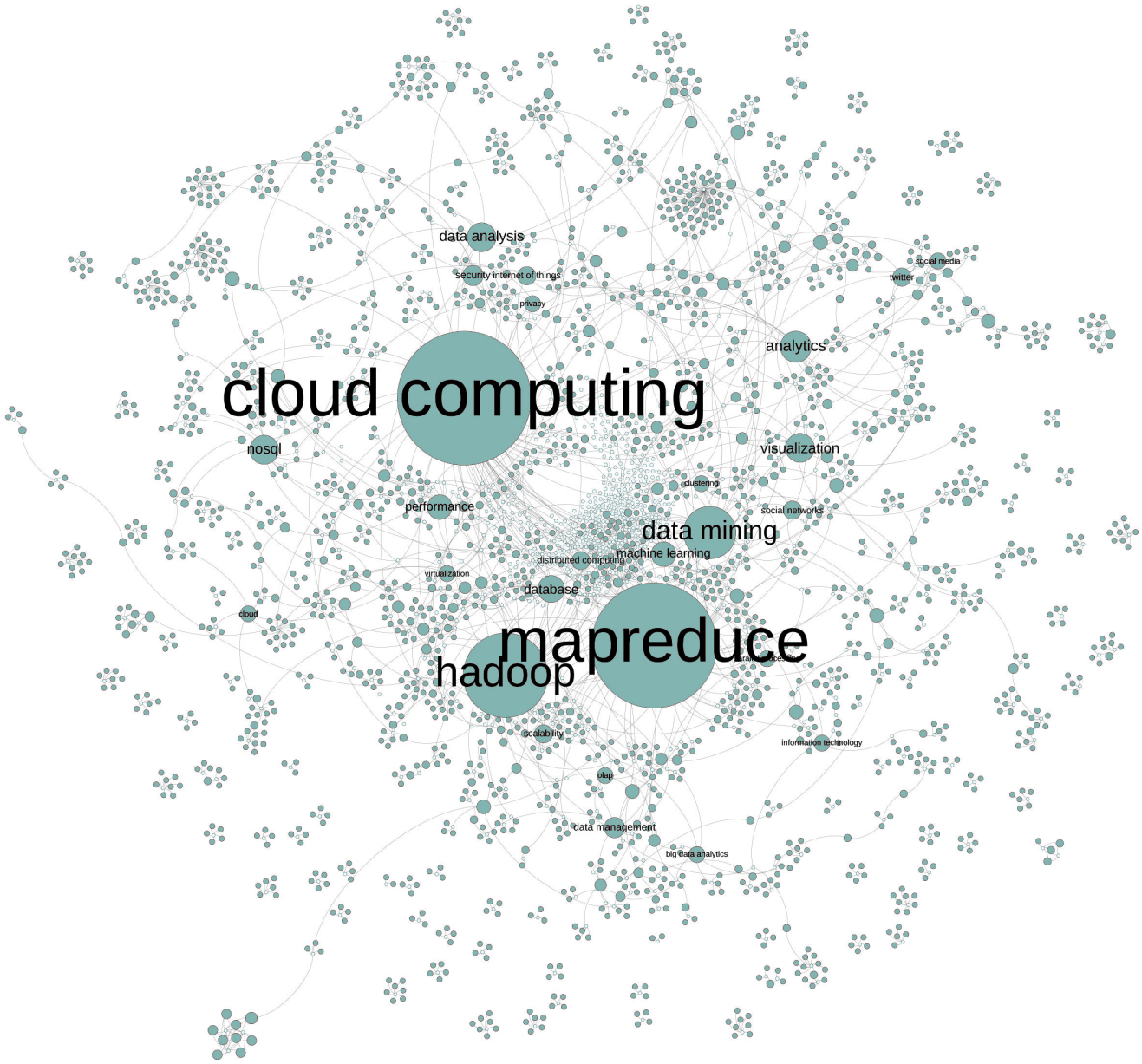


Figure 29: Keywords Scientometrics. Green: Field termiology, Grey: Articles. Visualization filtred by degree.

In the science map[9] (figure 29) we find a predominance of highly specific technical terms: *Hadoop, Mapreduce and cloud computing, NoSQL, machine learning, scalability, Olap*. From these terms we already gain a quite accessible overview of a strongly technical Big Data, underlined by the high ranking of both *Mapreduce* and *Hadoop* (discussed earlier).

By looking into the meaning of some of the keywords we gain a general understanding of the operating principles of this sphere. First and foremost is the tight coupling to *cloud computing*[10], a technology that emerged prior to the widespread interest in Big Data, but serves as a necessary condition for Big Data: the massive computations of Big Data are generally only possible when distributed across servers offered in the cloud. *NoSQL* is a highly scalable database technology that emphasises looser couplings in order to allow for higher variety and volume of data, and *scalability* emerges as a key concern for Big Data: constructing software that can handle exponential growth in data without buckling under stress or drowning out in noise. One way of addressing this is by the widespread use of *machine learning* and *artificial intelligence* where machines learn to identify correlation on their own through iteratively crawling the dataset[11].

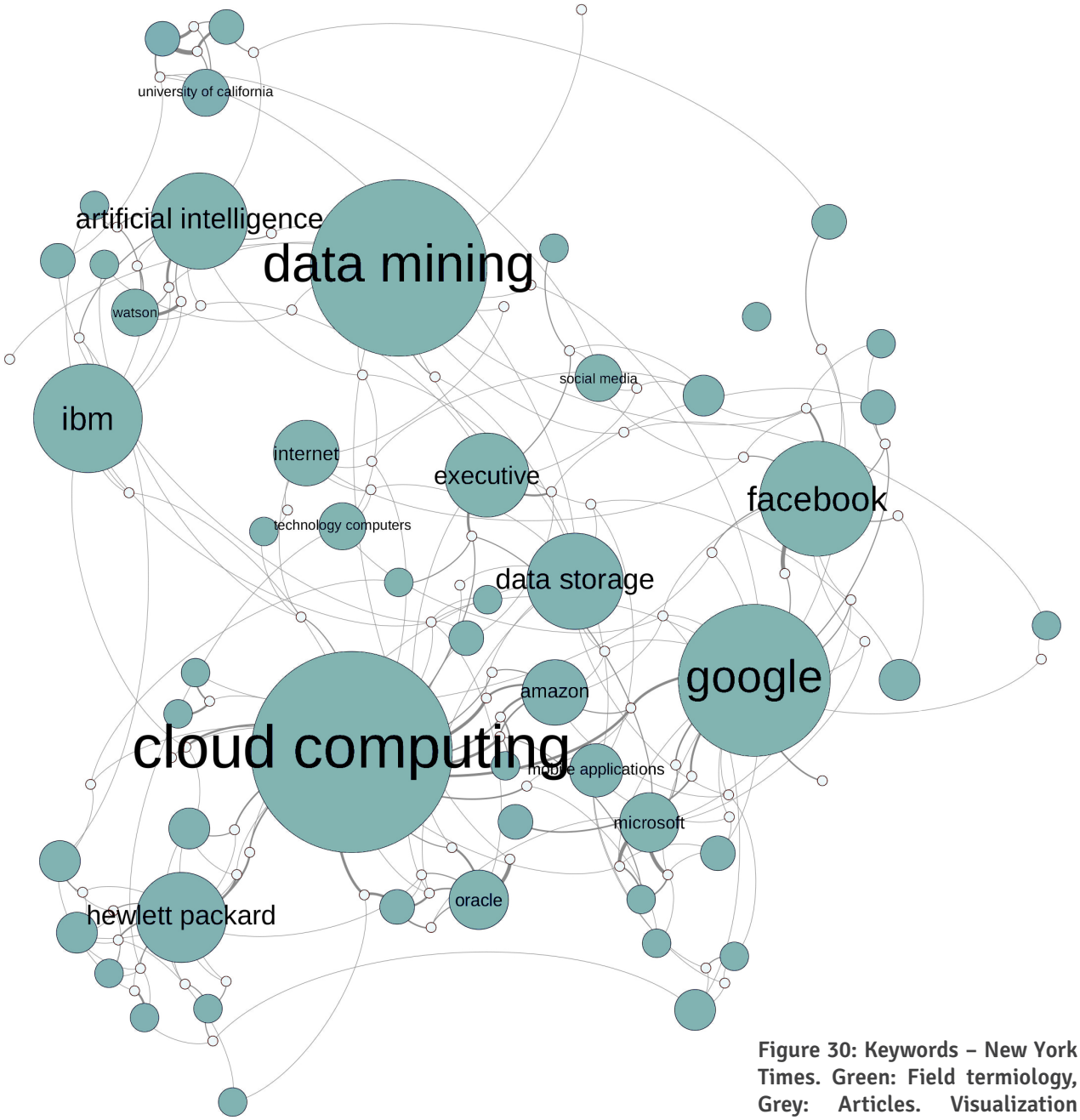


Figure 30: Keywords – New York Times. Green: Field termiology, Grey: Articles. Visualization filtred by degree.

Bundling the different concepts together we observe how the science sphere enacts Big Data as: a mathematical construction of loosely coupled databases hosted on global servers, that allows for exponential increases in data volumes in order to facilitate the machine-learning of predictive capabilities thus generating new insight.

In the map of the general news (figure 30) we find a very different ranking. Although some technical terms like *cloud computing* and *artificial intelligence* figure, the more technically specific field terms (e.g. *NoSQL*, *Mapreduce* and *Hadoop*) have disappeared. Instead the remaining terms are joined by a host of references to organisations such as *Google, HP, IBM, Oracle, Amazon* and *Facebook*. These firms have already been linked to Big Data in various ways in earlier experiments (see e.g. chapter 4 and 5.3), and their re-occurrence here is indicative of the type of text found on the site: they are all news articles, many of them reporting on organisations. The names of the organisations thus figure prominently together with the term *executive*, the person who was probably quoted in the article. This news orientation is also evident by the high ranking of *Watson*, IBM's artificially intelligent supercomputer that defeated the reigning jeopardy champions in 2011 - quite a sensationalist news topic.

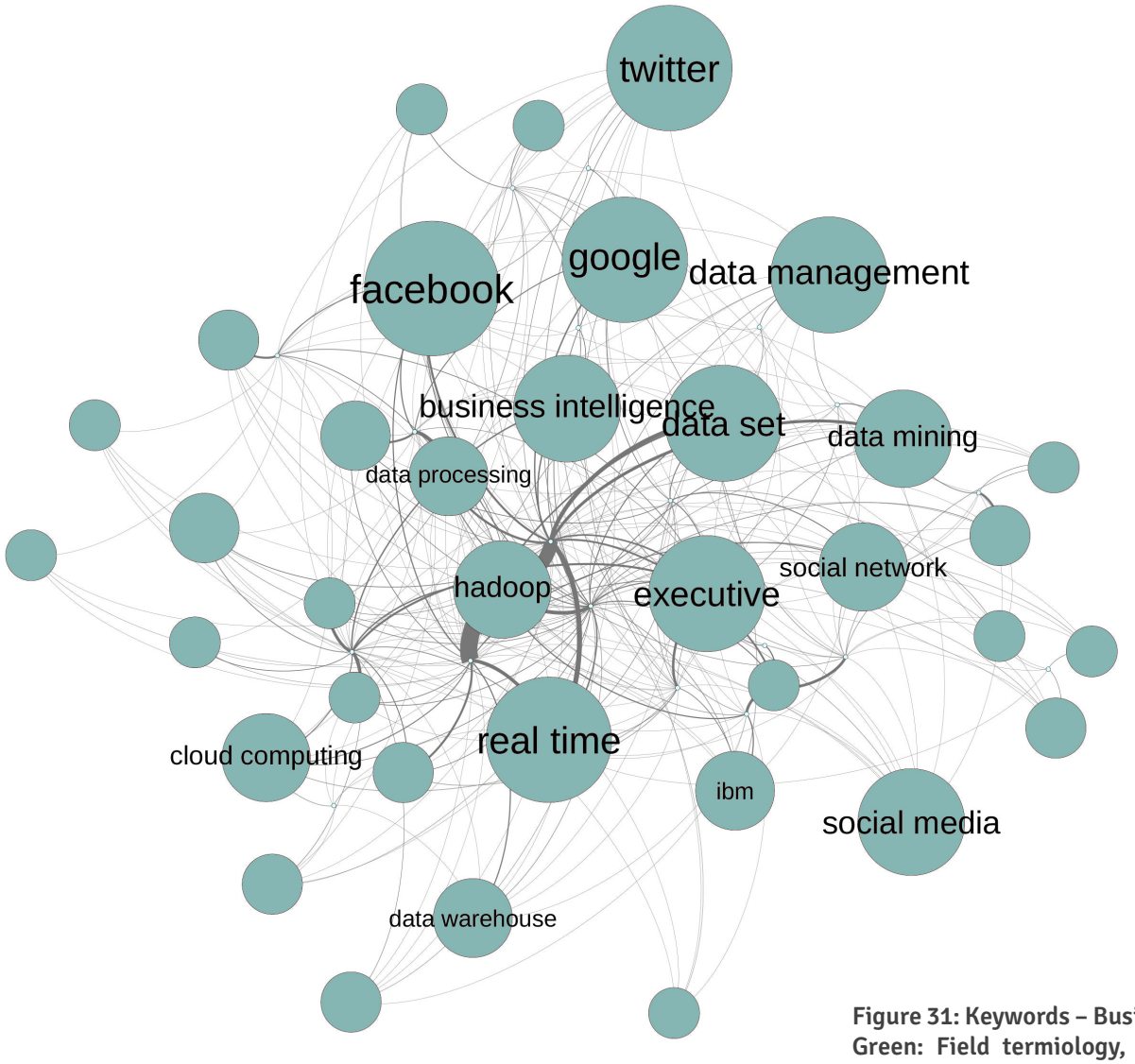


Figure 31: Keywords – Business. Green: Field termiology, Grey: Articles. Visualization filtred by degree.

In the sphere of business (figure 31) we witness again some of the bigger IT organisations and technical words, though fewer and with lower ranking than in the New York Times. *Executive* is ranked even higher positing Big Data even stronger as a venue for business. *Data Management* is another term linked to enterprises, as is the infatuation with *real time* computation - a newly trending topic in the Big Data discussion articulated as a crucial concern if firms are to profit from the knowledge generated by Big Data. We also find a high ranking of the term *data mining* a somewhat crude metaphor for the extraction of insights and correlations through machine learning discussed earlier. One such application is the discipline of *business intelligence,* the application of analytics to further a firm's competitive advantage. This term take a large position in the white papers and consulting reports, unsurprisingly since it is often this very service they sell.

In summary, we find that Big Data is enacted as quite different ontologies across these spheres, an example of the multiplicity we discussed in chapter 2.2. While one enactment is composed of elements (computing architectures and algorithms), others assemble another range of entities (executives, IPO's and competitive advantages). The technical enactment of the science sphere would e.g. only be expected to be able to enrol a narrow range of actors. To enrol more diverse actors and produce a broader interessement outside its own sphere, this enactment would need further translation. A such interessement are witnessed in the general news sphere. Big Data is here still a strongly technological concept, but what has changed from the field of science is the fact that the front figures of Big Data are now the tech giants, their managers and a set of trending technologies that are somewhat familiar to the everyday life of most people. Broadening the concept thereby increases the ability of Big Data to mobilise actors and create interessement. The network that makes up Big Data is thus figurated differently simultaneously in different contexts which will be the foundation of our third and last analytical phase.

5.7 INTERIM CONCLUSION

In the preceding chapter we attempted to map how Big Data translated and enrolled different actors over time. As has become evident from our experiments, our different methods assemble Big Data in a multiplicity of ways. They did that because each of them made certain things absent, and others present in a particular way (Law 2004), a theme we explored by analytically assessing the mode of inscription they each employ. Hereby, we produced not only our own multiplicity of enactments of Big Data, but also inferred how Big Data was enacted differently across time and spheres through the enrolment of different actors in the many forms: websites, organisations, scientific articles, keywords and field terminology. We will in the following summarise and contrast these different enactments to answer which actors were enrolled and how this enrolment was conceived.

In our scientometric analysis (chapter 5.1) we found Google's article on Mapreduce in the centre of the network of academic citations, apparently playing a major role in the development of Big Data; a conclusion which only became greater through discovering how Google's proprietary technology, Mapreduce, was also articulated as the technical origin of the widely used open source equivalent Hadoop. We also found that non-scientific articles, among these Anderson's wired article and Mckinsey's Big Data report received a large number of citations, pointing to a hybridised development (Callon 1987) of Big Data, intermingling science with actors from commercial and tech news. Furthermore, by comparing affiliation counts, we found a large discrepancy between the funding of articles and citation centrality, as well as further evidence of the hybrid nature of Big Data research through the prominence of private companies.

In 5.2 we looked at the main propagators on information on Big Data to explore the role of media in constructing hypes. Here we found a development over time where the central actors started as highly specialised trade press outlets, the middle period saw a prominent rise of consumer oriented tech news with a futuristic bend and Big Data conferences, while the late period ushered in major international news outlets, particularly those focused on business. In short, we found that Big Data was translated as a topic of specific technological interest into one of more general relevance, particularly for business.

In 5.3 we first elaborated on some of the methodological difficulties of investigating historical data with digital methods. But we also looked at how the central actors change when centrality is measured by ingoing rather than outgoing links. Here we found that commentators, and not developers, received the majority of references, indicating that some translation of the technical features of Big Data had to occur for it to associate widely. In general, we note that although thousands of actors are associated to Big Data, it seems to us that the majority of translation work is done by a few central actors, as evidenced by the high degree of concentration of references to central nodal points.

In our first ANTA analysis (chapter 5.4) we left the hyperlinks and proceeded into the semantic universe of text analysis. Here we found a movement from firms who contributed to research over actors who built their business models on data to the last period, where we witnessed the meteoric rise of social media as a new source of behavioural data. Thus the interest in Big Data was translated from an initial focus on researching technologies to the raw data sources and the promises they give. The prominence of Google, who had been suspiciously missing from our out-degree maps, leads us to a slight detour into the question of Google's affiliations with the term Big Data, revealing that they actively distance themselves from the usage out of worry for totalitarian connotations, an understandable choice considering the salience of privacy issues in their engagement with the public. In our context, it is interesting to see a sort of anti hype - the most central actor in the technological development seeking to cut associations to the ascending term.

Furthermore, we looked at the change of keywords over time in chapter 5.5. Overall, we found that the keywords tell a story about Big Data as a phenomenon that has been enacted in multiple ways. We found that the early period used very broad and general technical terms, pointing to the fact that Big Data was yet to mobilise its own assemblage of terms and rather borrowed from the existing fields of IT. In the middle period, the terms gained a certain specificity and a vocabulary of Big Data emerged. In the late period these terms were supplanted by terms denoting areas of application and the potentiality of Big Data as an object of investment. So we found two parallel developments: on one hand Big Data went from a specific niche interest and into a topic of general interest. But simultaneously, it went from a generalised semantic construction into developing its own specific vocabulary. This paves the way for enrolling more actors.

As an introduction to 5.6 we coalesced our preliminary identified actors into a number of spheres, classes of actors representative of particular characteristics found in our earlier experiments. For each of these we found a data source representing the sphere from which we could mine full text articles. We then mapped the occurrence of these articles for each sphere over our temporal periods.

By mapping and comparing our three spheres, we were able to get insight into how different clusters of actors enacted Big Data in quite different manners. For science we found a range of highly technical concepts, in the general news we found keywords indicative of Big Data as a newsworthy subject, as a sensational and exciting new endeavour as seen in the victory of artificial intelligence, Watson and the prominence of organisations and their executives reported on. With the consultancies we saw Big Data as a subject of business, with a focus on executive titles, business intelligence and data management.

In order to delve further into this multiplicity of enactments, we will venture further into the Sociology of Expectations in the next chapter, in order to ascertain the different visions of Big Data.

This will also contribute with a major piece of our puzzle, because while we have accounted for who and when actors were enrolled in the preceding analysis, we have yet to account for how. Why where different actors interested? How were they enrolled? Earlier we discussed the performative functionality of expectations to the future, and in the next analytical phase we will take a look at these expectations as visions, coherent depictions of possible worlds (or possibilities of Big Data in our case) to see how they offer a position to different actors, how they function as interessement devices.

Notes

1. To gain an overview the titles are shortened. The original title is *Mapreduce: simplified data processing on large clusters, Dean et al. 2008.*

2. E.g. “A comparison of joint algorithms for log processing in Mapreduce” (2010), “Hadoop: efficient iterative data processing on large clusters” (2010) and “The Google file system” (2003).

3. Despite the different abbreviation IDG and IDC were found to be subsidiaries of the same firm.

4. To make sure that this is not just a result of self-referencing, e.g. every site on Wikipedia having thousands of links to itself, we ran a filter removing such self references from the map.

5. The report *3D Data Management: Controlling data volume, velocity and variety* is originally published by META group. Gartner acquired this firm shortly after its publication.

6. As discussed earlier, both Amazon and Google contribute to the development of Big Data infrastructure, but only secondarily to their primary services.

7. The basis for this division process and their categorisation is described in appendix, chapter 9.2.

8. We are well aware of the problems and discussions on incorporating the concept of spheres in ANT, often theoretically rejected (Latour 1997:2) but apparently also empirically indispensable (e.g. Latour 2011; Venturini 2010:10). We however do not find it valuable to enter into this broad discussion and will simply use the term in its most general manner, as a way of evaluating different data and data sources against each other (DMI 2007).

9. In the science sphere we have methodically diverted slightly because of technical difficulties of getting access to the full text articles. Instead we used the keywords of the article extracted from the academic databases.

10. An earlier much hyped information technology, where computations and data storage are moved from local computers to global servers, accessible by high-speed Internet connections.Interestingly, this hype actually transitioned into general use quite unproblematically: our readers are probably familiar with services such as Dropbox or Google Docs.

11. This places a high focus on the predictability rather than the veracity of algorithms, and is exemplified by prediction markets such as kaggle.com, where prizes are offered in a competition for who can construct an algorithm that can predict behavioural optimisation with relative accuracy in e.g. aviation or medicine.

6 ANALYSIS III: VISIONS

IN THIS CHAPTER WE WILL TRY TO LOOK AT SOME OF THE WAYS ACTORS ARE ENROLLED THROUGH LOOKING AT HOW DIFFERENT EXPECTATIONS CREATE INTERESSEMENT FOR DIFFERENT ACTORS. WE WILL ANALYSE THESE EXPECTATIONS OR FUTURES AS VISIONS; INTERNALLY COHERENT DEPICTIONS OF ALTERNATIVE FUTURE WORLDS.

In the preceding chapters we have studied first how the interest in Big Data grew over time in chapter 4, and subsequently which different actors were enrolled at different points in the analyses of chapter 5. In this chapter we will try to look at some of the ways of *how* actors were enrolled through looking at how different visions create interessement for different actors.

As stated earlier, one of the key points of the sociology of expectations is the performativity of statements concerning the future. An articulated future, whether in the shape of a scenario, vision or prophesy is not merely a prediction, but an enactment that in itself acts as potentiator of the promise that is proposed (Brown & Michael 2003). Since they thus re-configure any arrangement in which they are inscribed, they fulfil ANT's definition of agency as "*making some difference to a state of affairs"* (Latour 2005:52). So we view expectations as actors, and they will be the focal point of our exploration on this final chapter.

We will analyse these expectations or futures as *visions;* internally coherent depictions of alternative future worlds (Eames et al. 2006).

To do this we draw upon a number of texts and moments identified in the prior analyses, establishing a core text corpus of 30+ articles totalling around 600 pages. Reading through them, we look for any articulations that draw on a future temporal orientation: what Big Data *will be,* what it *will change*, what the world *is becoming* (or *has become),* what Big Data *requires*, what Big Data *offers.* Our focus is statements on impending or on going change and development, statements of potentials and possibilities and statements contrasting past states with current or future states (*we used to x, but now we y*) - in short, statements that touch on possible futures for Big Data. These statements we call *expectations*. By drawing on our previously established depictions of Big Data, we then bundle these expectations into 4 different *visions* of Big Data, where each vision is an assemblage of multiple expectations

Thus, we will study how a vision of a particular future serves to stabilise a given network in a particular arrangement, partially parallel to Callon's (1987) conception of actor-worlds we introduced in chapter 3.2.

Based on these visions we will discuss their possible effects as interessement devices in order to look at *how* actors are enrolled. One of the ways this has been articulated is by Van Lente (1993 & Brown 2000), who studies how expectations of the future are translated from promises of future potentials into requirements. A similar mechanic can undoubtedly be found in the case of Big Data; where a host of actors from tech firms to national governments are initially enthralled by the promises of Big Data, only to find that the entrenchment of this promise becomes so strong that they are required to take heed of it in the formulation of their strategies even though the sheen might have worn off.

6.1 WELCOME TO THE PETABYTE AGE.

First and foremost, we see a tendency to formulate the future of Big Data, not as the future of a singular technology, but as a future of society per se. Big Data becomes not just the name of a given computational method or a business potential, but also a diagnosis of change in society at large - in its staunchest proponents equal in magnitude to the industrial revolution.

> *"We're now entering what I call the 'Industrial Revolution of Data' where the majority of data will be stamped out by machines (...) These machines generate data a lot faster than people can, and their production rates will grow exponentially with Moore's Law."* (Hellerstein 2008)

In this vision, the future of society is predicted as a *revolution* in data across sectors, spanning from housing and healthcare, science and finance, education and business, making it "*possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime and so on. Managed well, the data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account" (Cukier 2010).* In its strongest form, Big Data is pictured as a pervasive explosion of data, the application of which has untold possibilities- a utopia in its clearest form: *"Big Data is a tagline for a process that has the potential to transform everything"* (Kleinberg in Lohr 2012).

Big Data is however not only described as new opportunities; it is also described as a fundamental break with the past. As Anderson states it in wired: *"In the era of Big Data, more isn't just more. More is different"* (Wired 2008). How exactly more is different for Anderson will be further touched upon in 6.3. This vision claims the change to be so pervasive as to affect our fundamental understanding: *"More subtly, it will affect how people think about the world and their place in it."* (Mayer-Schonberger & Cukier in Kelly 2013).

The vision is not just an expectation to the future - it is also slowly but evidently translated into *requirement* the actors are required to take heed of (Van Lente 1993). As exemplified by Berners-Lee:

> *"The information about spending, agriculture, health and education that lies behind locked databases could be used to dramatically improve people's lives. (...) Imagine how quickly impacts such as these would multiply if governments were to openly publish this data, not just about the cost of medicine, but also about student attendance rates or crop productivity compared to use of pesticides"* (Berners-Lee 2012).

As we see in the quote above, this expectation is not just passive encouragement; for this vision to come into play, it requires governments to make their data open and accessible. In this way, the vision functions as an interessement device by not only promising potential benefits to the actors it seeks to enrol, it also formulates the roles they need to play in the network in order to realise it. Thus, the vision translates not only Big Data into a particular form, it also seeks to translate the entities in the network (Callon 1986).

We summarise this vision thus:

Welcome to the petabyte age: the society of tomorrow will produce ever more data on the world around us and the actions of human beings. Nested in this data we can glean insight into phenomena that previously eluded us, and with the right tools potentially transform every field we know. For the revolution to begin government and companies, however, have to change current data practices.

However, the actors thus translated are not limited to humans - a primary role is ascribed to the computers and programs that will handle the data, which leads to a redistribution of agency across the assemblage: *"we could create a world in which it would be programs -- not just people -- that would enjoy the data"* (Berners-Lee 2012).

This idea is the foundation of the next vision - the technical vision of Big Data as a whole new way of computing.

6.2 UNSTRUCTURED DATA: A NEW WAY OF COMPUTING.

This vision articulates a range of expectations to an on-going development in not just computational methods, but also infrastructure and usage patterns. In contrast with the preceding vision, which veered closer to sensationalist and overarching utopianism, this vision is tempered by its reliance on technical metrics. As such, its future orientation is more often expressed in numerical terms of prognoses and projection: *"As of 2012, about 2.5 exabytes of data are created each day, and that number is doubling every 40 months or so. More data cross the Internet every second than were stored in the entire internet just 20 years ago"* (McAfee and Brynjolfsson 2012:62).

The vision is however about more than exabytes of data and data growth per se, but equally on a fundamental change in the sources of data: *"Companies churn out a burgeoning volume of transactional data, capturing trillions of bytes of information about their customers, suppliers, and operations. Millions of networked sensors are being embedded in the physical world"* (Mckinsey 2011:4). The envisioned emergence of these *"millions of networked sensors"* associates Big Data to a far-reaching assemblage of actors that are non-human, yet still reside in the physical world: *"you do not have to breathe oxygen to generate V3 data. Traffic systems, bridges, engines on airplanes, your satellite receiver, weather sensors, your work ID card, and a whole lot more, all generate data"* (IBM 2012:xxvi).

This line-up of non-human actors has important implications for the heterogeneity of the network assembling Big Data, which we will discuss below. But it also changes the formatting of information that gets progressively more unstructured when more data sources are assembled: *"Unstructured information is growing at 15 times the rate of structured information"* (IBM 2012:xv). This vision's depiction of rapid change from structured to unstructured data is similar to the growing attention to 'raw data' and more generally 'data storage' that appeared in the middle period when we mapped field terms in 5.5.

The witnessed massive growth in data and data sources is posited as a condition that requires a wholly new approach to computing:

> *"Used to be that if you wanted to wrest usable information from a big mess of data, you needed two things: First, a meticulously maintained database, tagged and sorted and categorized. And second, a giant computer to sift through that data using a detailed query. But when data sets get to the petabyte scale, the old way simply isn't feasible. Maintenance — tag, sort, categorize, repeat — would gobble up all your time. And a single computer, no matter how large, can't crunch that many numbers"* (Di Justo 2008).

At the petabyte scale our traditional ways of sorting and computing loses its applicability. The "new way" is therefore not to get bigger computers, but to enable more efficient networks in the form of cluster computer systems:

> *"A new form of computer systems, consisting of thousands of "nodes," each having several processors and disks, connected by high-speed local-area networks, has become the chosen hardware configuration for data-intensive computing systems"* (Bryant et al. 2008:3).

The importance of these new techniques in establishing the vision is underlined by the many related terms as *Hadoop*, *MapReduce*, *Distributed computing*, *Distributed processing*, *Parallel processing* etc. that we found in chapter 5.6.

These new data streams not only require new forms of computers, they also demand new ways of analysing: *"The enormous volumes of data require automated or semi-automated analysis – techniques to detect patterns, identify anomalies, and extract knowledge"* (Bryant et al. 2008:3). These automated analytical techniques emerged as actors in chapter 5 under names such as 'artificial intelligence'[1], 'machine learning' and 'evolutionary algorithms'. The important thing with regard to this vision is the expectation that computation will no longer be confined to singular local computers, and analysis will no longer be the sole domain of human beings.

This also moves computation from local mainframes to the cloud *"in which data and software are situated in huge, off-site centres that users can access on demand"* (Marx 2013:257). Cloud computing, identified as a prominent actor in chapter 5, relocates local data sets to a global storage which allows the data to be accessible from anywhere in the world. This relocation also reverses the relationship between user and dataset: the data no longer comes to the user, but rather the user comes to the data: "*There's no reason to move data outside the cloud. You can do analysis right there"* (Sundquist in Marx 2013:258). Once again the data instead of the analyst is placed centrally in the discourse, an observation that will be the focal point in the next vision.

What becomes clear from the previous outline is that the expectations to Big Data as a computational method do not rest on a singular technological development, but rather a host of developments drawing together a range of fields into a more or less stable arrangement. The reach of this arrangement is further extended by Big Data's association to sensors outside of the purely digital, making it possible for Big Data to speak on behalf of not just digital traces from online behaviour, but physical objects and natural phenomena such as bridges, epidemics and tornados. In this perspective Big Data should not be understood as a single technology, but as a *socio-technical system* (Callon et al. 2002), not an object, but a *thing* (Latour 2004:233).

The technical vision of Big Data is thus the vision of a host of parallel developments drawn together in a new assemblage: sensors, digital traces, server farms, cloud computing and machine learning algorithms.

We summarise this vision thus:

Data Ubiquitous: the data of tomorrow comes not from clumsy human data entry, but is born out of an ever-growing array of sensors and the digital traces left by digital activity. Data this big can neither be stored nor processed by a single computer, but must be distributed around magnanimous clusters of computers. Although globally accessible from the cloud, only automated algorithms will ever be able to find meaningful patterns in it.

This vision presents a future for Big Data embedded in machine to machine interaction, a hastily sprawling assemblage of non-human actors that do not award humans any privileged position - sometimes quite the contrary, as in this quote where humans are reduced to little more than mediators: "*Thanks to sensors and the emerging Internet of Things, the digital realm can bypass the pathetic layer of intermediaries"* (Rao 2012:2). As we shall see in the next vision, this anti-anthropocentrism extends to the conception of knowledge that undergoes a translation with serious consequences.

6.3 TOO BIG TO KNOW: BIG DATA AS A NEW WAY OF KNOWING.

The previous visions have hinted at a shift from human cognition to machine learning, *"allowing decisions to be based increasingly on data and analysis rather than intuition and experience"* (Lohr 2012).

This vision deals with expectations to the future of knowledge and Big Data's role in it. It is thus closely tied to discussions on research, and goes as far as to paint a picture of research and science being irrevocably changed due to the impact of Big Data: "*The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world"* (Anderson 2008). Like in the previous visions, we notice how the coming of Big Data is introduced as something radically new and different, and how the access to enormous databases are articulated as key to the new possibilities, enacting what Callon has called an obligatory passage point (1986), a node in the network that others must pass through to stabilise the network around them.

The vision posits this growth in data points to lead not only to new methods of scientific enquiry, as well as the possibility to explore new venues, but to altogether different conception of knowledge: *"As science has gotten too big to know, we've adopted different ideas about what it means to know at all"* (Weinberger 2012:126). In this vision, data is expected to grow at such rates that *"it becomes incomprehensible by a single person, so we have to turn to other means of analysis: people working together, or computers, or both"* (Wattenberg in Horowitz 2008). In other words, the boundaries of human cognition is supplanted by the computing power of machines, whereby computer assistance is slowly translated as an obligatory passage point for the practice of science:

> *"humans cannot understand systems even as complex as that of a simple cell. It's not that we're awaiting some elegant theory that will snap all the details into place. The theory is well established already: Cellular systems consist of a set of detailed interactions that can be thought of as signals and responses. But those interactions surpass in quantity and complexity the human brain's ability to comprehend them. The science of such systems requires computers to store all the details and to see how they interact"* (Weinberger 2012:127)2.

In place of human cognition, unable to deal with the complexity, semi-autonomous machine algorithms will do the work, mirroring Noortje Marres' (2012) redistribution of work between human and non-human actors as discussed in chapter 2.4.5. The implications inferred here, though, are much less tempered:

> *"We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."* (Anderson 2008).

Traditional scientific methods are made obsolete by the predictive power of computer algorithms that are heralded to be able to find answers that we not only cannot predict, but which we might not even be able to understand in the first place.

This has important implications for the role of theory. As we have mentioned already, a cornerstone of the discussion is Anderson's polemic against the role of theory, a sentiment that plays a central role in constructing expectations in this vision:

> *"Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity"* (Anderson 2008).

This (naïve) belief in the predictive power of computer algorithms is pervasive, and leads to a tendency to black box the operations of the algorithms to the extent that they appear almost magical: *"In a world of Big Data the correlations surface almost by themselves"* (Cukier 2010), or as Anderson puts it: *"With enough data, the numbers speak for themselves"* (2008).

The expectations of Big Data's role in both the conception and production of data is both hard-line positivist and radically empiricist. As we shall see later in 6.5, this claim is far from contested. A scepticism we strongly share based on our own experience with the related field of digital methods[3]. The descriptions of answers magically surfacing are clearly at odds with how we have shown the manipulation and visualisation of data to be crucial in the construction of meaning, and we find the idea that machines can find serendipitous results without human engagement glosses over the technical difficulties of programming intelligent algorithms.

These understandings are however imminently well suited as translations of technical concepts into mainstream audiences in a way that gathers inordinately positive expectations by clouding the inner workings and vulnerabilities of the technology. While claiming Big Data will enable access to truths that obsoletes theory seem like superfluous superstitions, and its adoration of data and the predictive power of machine learning appears almost religious, this does not devaluate the visions performativity effects as interessement device, which is beautifully exemplified by the following quote:

> *"The day will come fairly soon where the default view will be to learn from data and temper our individual observation with what we can see from aggregating lots of information. Yes, there may be a small minority who resist this — just as there are people who believe the Earth is flat since from their individual observation, since that's what it looks like. But society advances"* (Mayer-Schonberger and Cukier in Kelly 2013).

The quote shows how Big Data, as new way of knowing, privileges empirical evidence over both intuition and theory, despite itself being based more in the latter than the first. One place where we can observe how this focus on empirical data has attracted specific actors is in regard to the emigration (or *colonisation*, Scott 2011) of empirical natural scientists into the social sciences, where they have gained immediate and significant attention under the paroles of a data driven social science (see e.g. Freeman 2011).

With that said, we define this vision thus:

Too big to know: The future of knowledge will be completely changed by an increase in empirical data and mounting complexity to the degree that will be unfathomable by the human mind. Instead, knowledge will be produced by software algorithms whose operations escape our understanding. Machines will provide us with answers to problems that we cannot understand nor explain, but nonetheless verify through empirical testing. This will mean the demise of theory, intuition and traditional understanding.

While the focal point of the vision of Big Data as a new way of knowing has been focused around the scientific or pseudo-scientific disciplines, the focus on data as a new driving force transcends both the sphere of science and the vision of knowing moving into the field of business: *"Data are becoming the new raw material of business: an economic input almost on a par with capital and labour"* (Cukier 2010).

This is the trace we will follow in our fourth vision, Big Data as a business possibility.

6.4 THE NEXT FRONTIER: BIG DATA AS BUSINESS POSSIBILITY

This vision enacts a future of data as a resource for businesses, and a potential source of big increases in profit: *"In the private sector, we estimate, for example, that a retailer using Big Data to the full has the potential to increase its operating margin by more than 60 percent"* (Mckinsey 2011:2). We note that the language is deliberately cautious, this is no more than a *potential*, but still open ended: the increase is *more than 60 percent,* but how much? This expectation is formulated in a way that allows for practically infinite optimism, but without promising anything, a formulation that is central to the development of wildly positive expectations to Big Data.

Data is thus translated from a computational category, an ancillary service function in organisations, to a highly strategic asset. Central to this translation of Big Data into imminent possibility is the expectation that Big Data will become the core asset for competitive advantage, and thus a strategic differentiator for businesses:

> *"Big Data will confer enhanced competitive advantage over the long term and is therefore well worth the investment to create this capability. But the converse is also true. In a Big Data world, a competitor that fails to sufficiently develop its capabilities will be left behind"* (Mckinsey 2011:6).

So Big Data is at one and the same time both a long-term investment, in which one safely can place funds, and a game where if you wait too long you will be left behind. Expectations are thereby directed towards the future, while action is required now. In this way Big Data is heralded to bring about imminent and tremendous changes:

> *"Indeed, our research suggests that we are on the cusp of a tremendous wave of innovation, productivity, and growth, as well as new modes of competition and value capture - all driven by Big Data"* (Mckinsey 2011:2).

This formulation of being on *the cusp* of the change suggest that if acting fast, an organisation can still be a part of this new gold rush in time. In both quotes we observe an urgency: wait too long (or fail to develop) and your competitors might get too far ahead. This urgency might increase interessement by spurring actors to immediate action.

As for the concrete benefit Big Data will bring, it is formulated as giving managers an unprecedented degree of visibility of the organisation: "*because of Big Data, managers can measure, and hence know, radically more about their businesses, and directly translate that knowledge into improved decision making and performance"* (McAfee and Brynjolfsson 2012:62). By equating more information with better decisions, Big Data is translated into a source of certainty, rather than addressing questions of information overload and the like. This translation is contested, as we will show in the final vision.

Another benefit is tied to the promise of Big Data allowing for real time insights - thus translating the insights provided from post-hoc rationalisations, to input, to immediate action:

> *"Imagine if every company—from retailers to banks to healthcare outlets—was able to conjure, in real-time, what mattered most to users or consumers" (Foster 2012).*

Here we find again the same sense of immediacy: Big Data is something that allows you to act now. In combination with the conception of Big Data's answers being too complex to understand in chapter 6.3, you should not even ask questions later. The promise is to trust the machine, act quickly and get ahead of the competition: *"Real-time or nearly real-time information makes it possible for a company to be much more agile than its competitors"* (McAfee and Brynjolfsson 2012:63). As discussed earlier the promises contained in future visions are often translated into requirements. In this case, the future of Big Data business requires not a reorientation of organisations towards data literacy: "*Companies must develop a "data culture" where executives, employees and strategic partners are active participants in managing a meaningful data lifecycle. Tomorrow's successful companies will be equipped to harness new sources of information and take responsibility over accurate data creation and maintenance."* (Avanade 2010:4).

To assist this process a new set of skills is required:

> *"a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/ artist to extract the nuggets of gold hidden under mountains of data. Hal Varian, Google's chief economist, predicts that the job of statistician will become the 'sexiest' around"* (Cukier 2010).

We note that the inclusion of storyteller/artist in the job description aligns itself with our earlier assessments that the vision of Big Data business increasingly relies on a translation of Big Data into a semi-mystical endeavour. This profession is not only expected to become *the sexiest,* but also increasingly scarce as Big Data grows: "*A shortage of the analytical and managerial talent necessary to make the most of Big Data is a significant and pressing challenge"* (Mckinsey 2011:11). So these expectations construct a vision of an assemblage that hinges on the competences of the very same people who articulate it, inserting themselves as obligatory passages, which others must pass through to stabilise the network around them. Thus the vision functions as an interessement device for data experts to become invaluable in realising it and the wild promises of Big Data coming true.

In summary, we define the vision of Big Data as the future of business thus:

The next frontier: data is the new gold, and for the business of the future it will be the main source of competitive advantage. Mastering data will make the organisation and its surroundings measurable with a precision never seen before, and allow for real-time insight to fuel strategic decision making. But it requires the help of an increasingly scarce species - the data scientist - so act fast, and don't ask how: data doesn't wait.

6.5 BEWARE OF THE BUZZ: BIG DATA AS HYPE

The four preceding visions all enacted highly positive expectations to Big Data. However, as we mentioned, this optimism is not spared from criticism. We found numerous contestations, critiques and pointing out pitfalls, privacy concerns and even describing Big Data as an empty hype. In the following section we will describe these.

First and foremost, the articulation of Big Data as a hype is not a transcendent perspective we impose as researchers. On the contrary, we found it expressed repeatedly, even at the Strata conference that might very well be the biggest gathering of Big Data enthusiasts': *"Is it just me, or is a major meme at #strataconf 'be wary of the religion of data'?"* (Rob Meyer in Brockmeier 2012). Others directly proclaiming it as a hype: "*I missed out on all the hype and irrational exuberance for Big Data"*(Brockmeier 2012), and urged others *"to go beyond the hype of Big Data"* (Mayer-Schonberger in Cook 2013), or denounced it as a *buzzword* (Foster 2012).

One commentator put the semantic mechanism behind the success of Big Data quite succinctly:

> *"Nobody seems quite sure exactly what the phrase means, beyond a general impression of the storage and analysis of unfathomable amounts of information, but we are assured, over and over, that it's going to be big"* (Marcus 2013).

So, as we mentioned in chapter 2.4.6, the reflexivity about hype dynamics is to some extent prevalent among the actors we study. But that does not necessarily mean they avoid it. On the contrary, even though he *scoffed at it initially,* the CMO for SAS declared that SAS *had to hop on the bandwagon* (Lohr 2012). This mirrors Van Lente's (1993) argument that expectations may turn into a requirement for businesses if they are to be seen as progressive, even though they may not share the expectation.

Regarding the contributions of Big Data to business, similar scepticism is expressed: *"The majority of respondents believe information will fundamentally change their business. And yet today, only a minority views their company data as a strategic differentiator. Most, instead, see it as a consequence of doing business."* (Avanade 2010:1).

Contrarily, increasing data is seen as an obstruction to decision making: *"The onslaught of data is making it difficult for executives to make decisions (...) Yet, they are still asking for more and they want it faster."* (Avanade 2010:2). This quote points to a view on the expectations to Big Data as highly irrational: why get more data if you are already unable to handle what you've got? The same sentiment is expressed against the infatuation with real time data we found both in the vision above, but also in 5.6: "*If you don't have the ability to act on real-time data, then don't try to gather real-time data"* (Brockmeier 2012). Both of the quotes point to a concern that business might be eager to gather more data than they can handle - a likely consequence of the incitations we found in 6.4.

The opacity of the claims of Big Data is even held up as a possibility for dubious business practices:

> *"selling Big Data is a great gig for charlatans, because they never have to admit to being wrong.If their system fails to provide predictive insight, it's not their models, it's an issue with your data"* (Marcus 2013).

Overall, we see the visions for Big Data in business to be assaulted on a number of fronts: it has yet to deliver the promised value, it might not contribute to better decisions - and companies might ask for more of it than they can capitalise on, choosing to store data for future benefit though they might not know what to do with it.

Regarding the visions for knowledge, and in particular the role of humans, a warning is addressed at the brazen claims of machine autonomy found in (Anderson 2008) and chapter 6.3:

> *"In the years to come, scientists and engineers will develop a clearer picture of the circumstances in which Big Data can and can't make a big difference; for now, hype needs to be tempered with caution and a sensitivity to when humans should and should not remain in the loop."* (ibid.)

As we stated ourselves, our own experiments and familiarity with digital methods showed us that agency was indeed redistributed to machines to some degree, significant effort still goes into pointing the machines in the right direction, and in the maps we drew in 5 we saw how markedly different assemblages were enacted from giving the machines different starting points and instructions. Others stress that it is exactly the insight and experience of humans that can transform data into insight:

"Getting the most from the data requires interpreting them in light of all the relevant prior knowledge" (Marx 2013:257), and contests the notion that data is inherently superior to other forms of knowing: *"If we see the world only as data, then we run the risk of* fetishizing *the data, of imbuing it with reason and meaning that it does not have. We need to be vigilant that we are not beguiled by data or lured by the false charms of quantifying every problem."* (Mayer-Schonberger and Cukier in Kelly 2013)

This critique of overly relying on massive datasets is also formulated in more technical term:

> *"We're more fooled by noise than ever before, and it's because of a nasty phenomenon called 'Big Data'. Modernity provides too many variables, but too little data per variable. So the spurious relationships grow much, much faster than real information. In other words: Big Data may mean more information, but it also means more false information"* (Taleb 2013).

Here, Taleb claims that the unexpected correlations that Big Data operations comes up with might not be that miraculous, but rather simply false positives. By contesting the translation of Big Data as problem solving panacea, (a notion especially prevalent in 6.4) and instead associating it with calculative devices, Taleb shows another side of the equation.

Lastly, and this is a critique that will regrettably be underexposed, many concerns are raised about what the storing of personal data and the quantification of self will do for citizen privacy: "*With this previously unimaginable growth in the volume and availability of information come some serious questions about privacy and ownership. How much of ourselves do we relinquish with our health apps?"* (The Economist 2010)

In summary, the epistemological claims of Big Data adherents is contested from a variety of positions, all expounding on the naivety of placing too much trust in these opaque mechanisms.

These alarms are not the sole domain of detractors and critics, but are acknowledged by the very same actors who are driving the optimism. Just like we saw with Google in chapter 5, this shows that not only are commercial actors very aware of the strategic impact of hype dynamics, they are also able to play them to their advantage on both sides of the court: *"How many times have you heard, "This changes everything," only for history to show that, in fact, nothing much changed at all?"* (IBM 2012:35).

We define the vision of Big Data as hype thus:

Beware of the Buzz: The vision of Big Data as hype is about being sceptical, critical and contesting Big Data. While the points of criticism range from lack of data integrity and privacy concerns to the risk of quantifying everything every problem, throwing light over the hype seem like the needed cure for the madness surrounding Big Data. This being said is Big Data as hype also a strongly reflexive vision, where actors not only are well aware of the hype surrounding Big Data, but for most parts also are highly attentive of the hype's performativity effects.

6.6 INTERIM CONCLUSION

In our vision of Big Data as a diagnosis of society at large we saw Big Data translated into an omnipresent revolution of data, affecting every part of life and with the potential to unlock new sources of economy, provides new insights into science and gains new ways of governing. We then discussed how these future expectations were translated into requirements to fulfil and how these requirements sought to translate the actors in a network.

From this general vision we ventured into the technical vision of Big Data as a new way of computing. We encountered a vision of Big Data primarily carried by numerical prognoses of data growth and increases in data sources. We witnessed a movement from structured to unstructured data collected through millions of non-human sensors and computed in the cloud by means of artificial intelligence. In this vision computation is no longer confined to local individual computers and analysis no longer the sole domain of humans. Finally we discussed how the inclusion of physical sensors allowed Big Data to speak on behalf of physical objects and how the wide range of technologies and objects assembled under the umbrella of Big Data, makes Big Data better understood as a socio-technical system than as a technology.

Thirdly we traced the vision of Big Data as a new way of knowing. In this vision, an increasing complexity that transcends the capacity of the human mind is enacted as a premise that makes traditional scientific techniques such as the use of theory, models and hypotheses obsolete. Through this Big Data is translated into an obligatory passage points for practice of research in the future.

In the end we discussed these arguments against our own limited experience with Big Data, contesting their claims while highlighting their performative effects in privileging empirical evidence over theory.

In the vision of Big Data as a business possibility we saw Big Data translated into a resource and an asset for businesses. Temporally, this opportunity was marked by a sense of imminence, urging organisations to embrace it before they were left behind. We discussed how this enactment fuelled hype-like expectation dynamics, as did the black boxing of Big Data ascribing it pseudo-mystical properties. Lastly, we will note that the competences of data literate experts are presented as a requisite for benefitting from Big Data; the very same people who present this vision insert themselves as an obligatory actor if the wild promises are to come true.

Thus the vision functions as an interessement device for data experts to become invaluable in realising it.

Finally, we explored how the visions stated above were contested and criticised. As we discussed, the understanding of Big Data as hype is not a reflexivity we find lacking in the other actors' accounts, nor our personal critique, which we reveal from a privileged vantage point, but should be understood as widely expressed expectation. In this vision we most notably observed how commercial actors are very aware of the strategic impact of hype dynamics and that they able to play them to their advantage on both sides of the court

Notes

1. Although strictly not the same – artificial intelligence is a broader related term.

2. While such a notion could easily be contested - did not the written word or printing press similarly augment human cognition? - the purpose here is not to mount a criticism, but rather explore the particularities of how future scenarios for Big Data are enacted.

3. Our methods, though not in the same quantitative scale as Big Data, still share methodological concerns to the degree we find it possible to extrapolate from our own findings.

7 CONTRIBUTION & DISCUSSION

IN THIS CHAPTER WE WILL DISCUSS OUR FINDINGS, PRESENTING THEM AS 3 CONTRIBUTIONS THAT ADDRESSES OUR 3 SUB-QUESTIONS.

In this chapter we will discuss our findings, presenting them as 3 contributions, addressing our three research questions:

Insight into Big Data as an empirical phenomenon,

Insight into what large scale mappings can do for the sociology of expectations and our understanding on how ideas spread through the web

Insight into the potentials, shortcomings and further development of digital methods.

7.1 WHAT IS BIG DATA?

In our study, we have enacted Big Data in many different forms: as an idea, as a technology, as a word, as an industry and as a diagnosis of contemporary society. We have both looked at Big Data as a vision of the future (a purely semantic construction) and as a technological assemblage (as a material construction). Along the way we have encountered many discussions of what qualifies as Big Data, and many examples that stretch the limits of these definitions. We have denied claims that the expectations vested in the technology was extraneous to any "actual" technological potential, instead focusing on how the expectations of potential and the materialisation of actuality are enmeshed and intertwined.

So what is Big Data? To respond to this question we turn to the concept of socio-technical arrangement (Callon et al 2002; Callon 2004). With this concept, Callon implores to understand technologies in the context of associations they are inscribed in. Thus a car is not a car by itself[1], but exists in a network of roads, petrol delivery, taxes, safety measures, agreements on how to drive, status values and even cultural patterns of etiquette in dating. Remove one of these parts, and what a car means, is and does changes.

Similarly, we find it moot to reduce Big Data to one of its components, e.g. distributed computing architectures. Instead, our visions exemplify how Big Data is the assemblage of myriads of highly heterogeneous elements, each of which is crucial to the function of the network. Without data from sensors and online traces, without architectures for handling and storing huge datasets, without automated algorithms to make sense of them, without business propositions,

research programs and conferences, Big Data would not be what it is today. Remove a single part and the network changes accordingly. Only by drawing associations this far and wide was Big Data allowed to come into existence.

So Big Data is at once technology and meaning, material and semiotic and the meshing of these heterogeneous elements. It is also simultaneously something for business and for science, taking on different forms that allow it to interest divergent sets of actors simultaneously as we saw in e.g. our scientometric analysis.

As Star and Griesemer explains (1989:389), when interessement is not just a one-way process, but all actors simultaneously try to interest each other, the coherence of the network depends on the degree to which multiple ontologies are allowed to co-exist.

This point is illustrated through their concept of *boundary objects,* which are *"weakly structured in common use, and become strongly structured in individual-site use. They may be abstract or concrete. They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation."* (Star and Griesemer 1989:393). This is quite similar to the description of Big Data above, where we argued Big Data was simultaneously meaning and material, and took on a multiplicity of forms to be able to engage different actors while still maintaining a degree of cohesion. We also notice how the distinction between common and individual use mirrors the two parallel processes we saw in our semantic analyses, where the vocabulary surrounding Big Data where at one and the same time ascribing new specific terms to Big Data, and generalised, gradually broadening the application of Big Data to cover more fields.

This multiplicity of meaning in boundary objects can be contrasted with a conception of a parallel process in the work of Laclau (Laclau and Mouffe 2002). Laclau's claim relates to the establishment of hegemonies, during which the equating of meanings under specific terms ultimately results in terms being emptied of meaning[2].

In contrast to this, we find in our exploration of the contestations and critique of Big Data, that the idea of Big Data as empty was articulated in a critique from actors describing it as a superficial hype. Rather than proclaiming the emptiness of the concept from a theoretical meta-level, we find it empirically as simply one enactment among a multiplicity of others.

This is worth repeating: we do not find the emptying of Big Data as a theoretical and analytical achievement, but rather as a particular viewpoint in our data - one articulation out of a multiplicity of others.

We hereby argue against Laclau's notion that the assemblage of many different meanings results in an emptying of a term, a notion that has been criticised as being hardly empirical (Andersen 2003). Instead we propose an understanding that terms achieve this breadth not by de-specification, but by re-specification: by translation into a host of specific meanings, each of which allows it to enrol different actors - as a boundary object that accommodates heterogeneity by spawning a multiplicity of forms.

7.2 THE ROLE OF EXPECTATIONS

In the course of this study we have tried to trace the story of how Big Data got to be such a big idea. To account for this we have employed the sociology of expectations to look at how enacted futures have an effect at shaping the actions and associations of actors in the present. One key term in this approach is the notion of hype.

An easy conclusion - and one often jumped to in hype studies - is to view expectation dynamics as irrational herd behaviour, and many of our experiments did in fact encounter a degree of exuberant enthusiasm that seems to be at times quite illogical. When looking at local interactions at least this description of hype as the result of unreflective actors does in fact seem palpable.

But as we saw in the contestations of Big Data in 6, they are not; most actors are indeed strongly aware of the hype around Big Data. Instead the critique comes out as yet another example of contemporary sociology's tendency of too easily pointing fingers (Latour 2004), which from a privileged analytically upstream position (Gorm Hansen 2011) decries other actors as unreflective. Instead we want to take a post critical position to see why hypes and expectations might actually be quite logical, when considering decision making under information constraints (Simon 1991): Not only is the answer to this question less preordained, the enactments produced by digital quali-quantitative experiments also offers a unprecedented overview of the emergence of expectations.

As we have seen throughout the study, expectations do in fact drive development and may be self-serving to some actors. But as we discussed in 7.1, the openness and ambiguity, the multiplicity and tacitness of definitions may actually also serve a broader purpose of enrolling actors.

Disregarding certain buzzwords is easy with the knowledge of hindsight after a situation has been stabilised, after the matters of fact and concern have been settled. But prior to that such discerns are not only impossible to predict, they might also be counterproductive. During the fluid period, how would we handle the descriptions of phenomena that have yet to take a final shape if not with these open, ambiguous and exuberant concepts? They must be open, since we have yet to know what form they will eventually take, and they must be optimistic to garner the support needed for them to develop.

If we momentarily change the object of investigation to the hype surrounding sustainability in the late 00's[3], it is easy to agree that though the term itself was spread so thin for the concept to barely carry any meaning, it did birth a range of more specific usage concepts (e.g. carbon quotes, organic certifications, ISO energy standards etc.), that have survived the hype and progressed as meaningful *and* influential actors. That these terms would be able to stabilise without the drive from the sustainability hype, we see as highly unlikely. Similarly, we argue that Big Data births a range of terms and technologies (e.g. recommendation engines, prediction markets, in-silico drug testing etc.) that can live on their own without the enthusiastic backing of the Big Data hype, and that these terms and technologies could never come to fruition without it.

This argument on specification is in line with what we have discussed in 7.1: the assembling of actors into Big Data is not just emptying the term of meaning (Laclau and Mouffe 2002), but also found in 5.2.5 to generate new meaning, in the movement from singular but quite imprecise terms to a multiplicity of more precisely defined terms.

In conclusion, we find that the terminology of things that are not yet stabilised has to be optimistic to interest, and it has to be open to be able to interest so many actors. Thus hype, visions and expectations do in this light appear as rational ways of approaching emergent phenomena - they might even be necessary for these phenomena to come into being.

In this perspective hype and expectation dynamics seem to be part of a strategy adapted to understand what has still not been completely formalised; less like impervious lemmings jumping off a cliff and more like blind fungus spreadingn random directions to find nourishment.

7.3 DEALING WITH MESS: HOW TO UNDERSTAND DIGITAL METHODS

As a finishing comment we wish to turn back to our introductory concepts of representation and co-creation in the landscape of digital methods, asking once more about its relation to reality and the prediction power of its conclusions.

As we have pointed out repeatedly, Digital Methods are not a shortcut to achieving objectivity in social sciences, even though it is sometimes hailed as such by e.g. the social physics camp (Pentland 2008; Newman, Barabási, and Watts 2006), and indirectly suggested by Latour and Venturini's 2nd order objectivity (Venturini and Latour 2010; Venturini 2012), and Richard Rogers' claims of grounding insights in data (2009:20).

On the contrary, through unpacking the underlying mechanisms of our experiments, we have shown that rather than singular depictions the tools of digital methods are wont to produce a multiplicity of enactments. We also showed how these enactments are highly contingent on the tools, data sources and especially the choices made in visualising them. Regarding visualisations, we found the techniques used to translate data into these visual inscription to be paramount to making them humanly legible, and thereby allowing them to be interpreted, vested with meaning and made to represent the entities they are drawn from. As these visualisations also hinge on contingent choices, the mechanisms and manipulations done on them producing vastly different depictions, they might be called creative calculations. We therefore found the visualisation to be of such importance that we ascribed this work to the analytical phase rather than merely a description of data.

As such, these visualisations, maps and charts should not be seen as holding meaning in themselves. Contrary to Anderson's claims (2008), the numbers do not talk for themselves. They are inscriptions on par with Geiger-counters and meteorological readings (Latour 1987), only achieving meaning by the strenuous work of drawing associations that inscribe them into wider networks of other inscriptions, texts and theories.

So, how should the knowledge we have produced, the claims we have made and the representational power of our maps be seen?

One could doubtlessly make a number of rightful objections to our story of Big Data and its multitudinous origins. One could object how the blurry and at times chaotic story leaves the reader without a clear grasp, claiming we failed to simplify the story adequately. On the other hand, one could also contend the claims between the data we represent and the interpretations we draw from them: How can one ascribe the emergence of a business oriented vision of Big Data by measuring increased interest for 'data management'?

The short answer is that we cannot know this. The longer answer is that no methods, digital or analogue, are able to account for the sum of communication in society, nor the totality of the associations drawn. The social world exceeds any method in complexity, and every depiction must therefore reduce it to a size than can be handled[4]. The map is not the territory - but without it, how can we navigate?

We did however find that the quali-quantitative digital methods applied in this thesis afforded us a glimpse of the endless translations constructing hypes, expectations and visions of the future - a glimpse we followed as far and as wide as it would take us.

Along the way, we chose distance over an ordered itinerary, briefly noting curiosities rather than dwelling at landmarks.

Despite this, we think most scholars cannot help but feel a bit repelled by the messiness, the ambiguity and the equivocality of our conclusions. For every step we took into the world of Big Data, the complexity increased. While we on one hand witnessed the emergence of a more a business orientated vision of Big Data, we did not ignore a parallel increase in interest for Big Data as a new way of knowing. Neither could we ignore how the technical discussion of Big Data did not cease to exist, but continued to increase in scale and multitude. We chose not to reduce our findings to any one of these, not to boil it down to a clear-cut narrative.

From this mess we have tried to extract patterns, retelling a multidimensional story of Big Data where most interpretations were unfixed. The data's heterogeneity was simply too great for one-dimensional and decisive conclusions. We might have stepped into a world that is, with Weinberger's words, too big to know (2012).

But what is the alternative?

Though we all quickly underline the constructed nature of reality when we encounter facts-based science we disagree with, as Latour has argued, the same critical attitude disappears whenever we move into our own field of science (Latour 2004a:240). We appear to have become so accustomed to tracing public discourses or ideas through reading a selective number of articles, that we have forgotten that a collection of 100 articles is but a drop in the ocean of communication produced in society. But a discourse (if we accept such a thing for the sake of argument) is huge beyond comprehension.

Yet, struck by the beautiful neatness of the argument, we accept without hesitation that 100 articles or even 1,000 can be the empirical foundation for unravelling the changes in public discourse on partnerships through decades (Andersen 2006)[5], as long as the author acknowledges that his view is a construct and subjects himself to a strict conditions for analysis (Ratner 2009).

What we observe is a preference for order over mess, to the degree that we accept narrow empirical latitude as long as the scientist can reach a simple and clear-cut conclusion, boiled down through precise composition.

Our point is not to criticise system theory, knowledge archaeology or any other constructivist approaches to which we both owe our academic upbringing and which have often shown themselves to offer highly valuable ways of generating knowledge. What we wish to stress instead, is how order (in contrast to mess) and one dimensional causality based on narrow empirical data appears to hold a prevalence over more chaotic and messy arguments, and how this ideal of tightly composed analysis contributes to an understanding of communication as highly ordered, rather than an uncontrollable and never stabilised mess.

But the world is messy, and meaning and communication is especially messy. And the digital methods we have employed here, while sacrificing order and precision, allow us to handle unprecedented amounts of this mess. Our ANTA analysis of 500 books, our Hypher analysis of several hundred thousand unique pages and our scientometric analysis of 15.000 academic citations would have taken an immeasurable time just a few years ago, but it might not be long before even these data sizes are dismissed as laughable. Along the way, as more people test out the sort of experiments we have done here and digital methods acquires experience, the methods will undoubtedly gain a higher degree of polish. As the discipline proceeds to stabilise, and as consensus is reached on the procedures and their significance, they might get to the point where their results are more uniform and unambiguous. But we hope it will not succumb to prematurely trying to settle the complex chaos of the world.

Rather than trying to reduce it, we have tried to develop tools to encompass more of it. This is accepting mess in all its glory.

Notes

1. and the car itself is of course also an arrangement of parts: motor, wheels, gauges, nuts and bolts.

2. Due to a conception of meaning as a differential logic. Had the scope of this project allowed it, this could have been held up against the relational ontology of ANT.

3. Yes, this is a broad and unsupported claim. Consider it illustrative.

4. Even Latour's *irreductions,* as beautiful and inspiring as we find the idea, is still only an ambition

5. Do not get us wrong; we love this book. We also deeply respect Niels.

8 CONCLUSION

HOW CAN DIGITAL METHODS GENERATE KNOWLEDGE OF THE HYPES, EXPECTATIONS AND VISIONS THAT SURROUNDS BIG DATA? IN THE FOLLOWING WE WILL CONCLUDE ON OUR RESEARCH QUESTION AND PRESENT OUR FINDINGS

In the preceding chapters we have tried to answer how Digital Methods can generate knowledge of the hypes, expectations and visions that surrounds Big Data. We also addressed our three sub-questions: how Big Data emerged, how Digital Methods can contribute to the Sociology of Expectations and how Digital Methods produce knowledge.

By positioning ourselves in the intersection of Digital Methods and the Sociology of Expectations, we have tried to construct a hybrid approach for studying digital expectations and how ideas spread across the Internet. We argued that the granularity and zoomability of digital traces allowed for a quali-quantitative approach that could encompass the translation of expectations across large temporal, social and geographical spans, thereby accounting for how expectations and visions of the future played a role in development of Big Data. While only allowing us to see digital enactments of expectations to Big Data, we propose that this method of digital expectations studies contains the possibility of empirically operationalizing the theories of the Sociology of Expectations – addressing thereby our second sub-question.

We have employed this approach in tracing the emergence of Big Data through large scale, longitudinal mappings of actors, their associations and the translation of expectations to Big Data. Here we constructed a myriad of enactments of Big Data and how it changed over time and space, and by juxtapositioning them accounted for the translation of Big Data from niche phenomena to widespread hype. We found a development where Big Data started as highly specific technical discussion on computing architectures for handling increasing data sizes and gradually changed into a utopian diagnosis for society where the proliferation of data and increased capabilities in wringing insight from them was presented as a promise of a panacea for solving and improving a range of issues across science, administration and especially business.

The movement was however not singular – we found lots of detours and variations, contestations and dead ends. Most significantly, we did not find the translation to be equivocally from specific to general, but rather that concurrently with the change from a specific technical matter to generalized commoditization, Big Data developed its own specific vocabulary of terms rather than borrowing broader terms from general

computer science. At the same time, the general promises of e.g. insight into consumer behaviour were supplanted by specific examples of actualized deployments such as Amazon's recommendation engine. Through these findings on Big Data as empirical phenomena we addressed our first sub-question: how the idea of Big Data emerged.

Along the way, we found not only patterns and nodal points in the translation of Big Data over time; we also identified limitations in both the existing literature and our own approach, and formulated our own method to try to address these limitations. In our work with digital tools, we tried to unpack their black boxed operations and discuss their representational and explanatory power in regard to the effect of contingent ranking and sorting mechanisms and especially the role of visualization in making them decodable. These discussions touched on the limitations, pitfalls and future potential of Digital Methods and thus addressed our third sub-question: how Digital Methods produce knowledge.

In the following we will reiterate our main findings, conclude on our explorations and relate them to our research questions.

8.1 TRACES OF DIGITAL EXPECTATIONS

The journey started with the outline of our methodological inspirations. We began this outline by introducing Actor-network theory as a conceptual vocabulary for our exploration of both digital methods and the sociology of expectations. The core concept of actors and networks were then extended with the theoretical concepts of translation, interessement, enrolment and inscriptions, representing a foundation for exploring actors' on-going attempts at stabilising actor-networks.

With this conceptual vocabulary we positioned our method in the Post-ANT movement of messy methods, ascribing to a co-fabricated and multiple reality where every engagement enacted yet another ontological reality. We also introduced Law's concept of method assemblages as a way of navigating such multiple realities, directing our attention towards the enactment of presence and absence in our analysis.

From this more traditional ANT encounter, we jumped into the less known territories of digital methods and their special ontological, epistemological, and methodological conditions. As an outset we discussed the relation between the digital and non-digital world, arguing for an understanding of the digital and the analogue as heterogeneous extensions of each other. We then introduced emergent digital traces, a native form of digital data marked by its non-hierarchical ordering, its sheer numbers and its intrinsic traceability, and discussed how the organising principles of the digital world such as Google Search Engine, offered both strengths and weaknesses when evaluating relevance.

Having described the major ontological actors surrounding our digital methods we discussed some of the epistemological consequences related to digital representation. We ascribed to the understanding of digital maps as inscriptions stabilising actor portrayals as viewed through the eyes of the co-fabricators. Based on these ideas we discussed how this shared fabrication and redistribution of agency towards especially the digital tools, straddles the divide between the empirical and analytical phases. We also introduced and discussed the quali-quantitative methods as a way of exploring digital traces by zooming between levels of aggregation.

In our last methodological chapter we disassembled the actor group of digital tools (software agents, crawlers, scrapers, APIs and visualisation tools), while we continuously discussed their different ways of enacting reality. The aforementioned discussion provided the foundation for addressing our third research question.

We then positioned us in the sociology of expectations by conducting a literature review. Through a preliminary scientometric analysis we gained an initial overview, delineating the field around a shared interest for looking at how the future was enacted rather than trying to predict it. On this basis two subfields emerge from our reading: Hype studies, quantitative studies of expectation dynamics and qualitative Visions studies of how expectations were merged into shared visions of the future. Based on our literature review we identified a need for more multi-sourced and integrated methodological approaches (such as the quali-quantitative method) as well as a need for exploring the interaction of hypes, expectations and visions.

Based on our methodological considerations, we began the construction of our method assemblage. We started out by declaring three preliminary requirements for our research design: semantic-, spatial- and temporal-multiplicity, summarising some of the limitations identified in the sociology of expectations. With these in mind we outlined our research design for a large scale, longitudinal quali-quantitative study build around three analytical phases. These findings is the foundation for answering our second sub-question.

8.2 TRACING BIG DATA

In our first analytical phase we studied the macro dynamics of the emergence of Big Data as a hype. We confirmed an explosive growth in both the supply and demand of information on Big Data characteristic of hype dynamics. We also found indications that the early development stage of Big Data was founded on mostly technical interest, while more commercial interest contributed to the explosive activity in later stages, providing the first clue to our first sub-question.

Methodologically we discussed how the current incarnation of hype studies often appeared to fail to distinguish between either: 1) supply and demand for information, 2) salience and sentiment, equalling media coverage with optimistic expectations and 3) singular and multiple representations, aggregating everything into a singular representation of total interest. Based on these discussions we concluded that prior approaches provided an incomplete operationalization of the phenomena they try to depict and that they are inadequate in accounting for a more comprehensive range of ways in which expectations serves to propel the translation of trends in technological development. The first two limitations we tried to address in this analytical phase. Regarding the third limitation, we proposed a way in which digital methods could address the problem of aggregation and over-singularization, in line with our second sub-question.

This proposal led us to proceed into our second analysis of expectations and how actors were enrolled in Big Data. Through a range of experiments we examined how the different tools and methods enacted Big Data in a multiplicity of ways by making certain things absent and others present. Through juxtapositioning these different enactments we drew a picture of Big Data's development, addressing our first sub-question.

Our first experiment assembled a network of citations between scientific articles on Big Data through a scientometric analysis. By virtue of this analysis we identified Google and its programming model Mapreduce as central actors. We also traced a number of highly ranked non-scientific articles, among them Anderson's renowned Wired article and Mckinsey's omnipresent report on Big Data, as well as a high number of private sector sponsored articles, all indicating a hybridised development of Big Data intermeshing scientific and commercial actors.

In our second experiment we then broadened our scope to the propagators of Big Data on the web, using the crawler Hypher. Here we identified a temporal development from highly specialised trade journals in the early period, over consumer oriented and futuristic tech news in the middle period, before a number of major international (business) news outlets were ushered in during the later period. In short, this enactment confirmed our thesis from the first analytical phase of Big Data being translated from a topic of specific technological interest into a more general and business orientated topic.

Shifting our focus from the propagators (out-degree) to the propagated (in-degree), we found attention to gather more around commentators than technical practitioners, indicating that some translation of the technical featured Big Data had to occur for it to associate widely.

Following a discussion on the technical limitations of hyperlink crawlers in reconstructing timelines, we proceeded into the semantic universe with the text analyzer ANTA. By mapping organisations and people mentioned in a selection of the Big Data literature, we identified how the organisations initializing Big Data research was replaced with data-driven organisations in the middle period, who was again supplanted by the meteoric rise of social media in the late period. A slight detour tracing Google's unwillingness to affiliate with Big Data, showed us how hype was managed strategically by actors, not only by latching on, but also by distancing themselves.

Shifting to keywords we learned how the early history of Big Data had been told with broad and general technical terms borrowed from the existing fields of IT, pointing out how Big Data had yet to mobilise its own assemblage of terms. We then traced how the terms gained a certain technical specificity in the middle period, and a dedicated vocabulary of Big Data slowly emerged. This vocabulary was then supplemented in the late period by terms denoting areas of application concurrent to a translation of Big Data into a potential object of investment. Through these enactments we were able to identify parallel developments: on one hand, Big Data was translated from a specific niche and into a topic of general interest. On the other hand, the early very generalised semantic constructions were replaced over time with Big Data's own specific vocabulary, which paved the way for a broader enrolment of actors.

Finally we categorised and visually divided our keywords into three empirically defined spheres (science, general news and Business), to gain insight into how different data sources enacted Big Data in quite different manners.

In our third and last analytical phase we left our digital tools behind and continued "by foot"; to point out how posited futures of Big Data function as interessement devices. Based on statements articulating expectations to the future of Big Data in the form of impending changes and potentials, we retold the story of Big Data as five coherent visions.

In our vision of Big Data as a diagnosis of society, Big Data was retold as an omnipresent revolution of data, affecting every part of life unlocking new economic, scientific governmental potentials. These general expectations of the future were then discussed as ways of stabilising Big Data, through their translation into future requirements. In our second vision, Big Data as a new way of computing, we encountered Big Data envisioned as data growth and new unstructured data sources and sensors situated, computed in the cloud by artificial rather than human intelligence. Based on the myriad of technological actors present, we initiated a discussion of Big Data as a socio-technical system rather than a technology. Thirdly we traced Big Data as a new way of knowing, a vision where increasing complexity has overburdened the capacity of the human mind demanding innovative ways of navigating the world. Fourthly we followed onto the vision of Big Data as a business possibility, a vision marked by a sense of imminence, urging organisation to embrace it before they were left behind. In this vision we also discovered how actors shaping the vision inserted themselves in it, thus turning the vision into an interessement device assigning these actors as obligatory passage points for the realisation of the future it promised. Finally, we explored how the visions stated above where contested, criticised and characterized as a hype, not as a reflexivity lacking in the other actors' accounts revealed from a privileged vantage point, but as just another widely expressed expectation.

8.3 CONTRIBUTING

Finally we discussed our findings, presenting them as 3 contributions. We first discussed how the many and varied enactments of Big Data, gave birth to Big Data as a socio-technical arrangement; an arrangement where meaning is derived from its associations, and where the removal of just one actor changes the entire composition, and thus Big Data itself. Following onto this we discussed how our different enactments of the socio-technical arrangement of Big Data mirrored Star's and Griesemer's concept of Boundary Concept, carrying different articulated meanings in different situations while being structured weakly enough to travel across these situations. We then discussed the boundary concept in opposition to Laclau's theory of hegemony, showing how our analytical findings contradicted the idea of gradual emptying a term. In our enactment, we found this idea of emptying not as a transcendent theoretical insight, but simply one vision of Big Data among others. Instead we propose an understanding that Big Data achieved its breadth, not by de-specification, but by re-specification: by translation into a range of specific meanings, each allowing it to enrol different actors. This discussion is central in our answer to our first sub-question: how Big Data emerged.

From this discussion we delved into a post-critical discussion of the performative aspects of hype. Rejecting the idea of hype as irrational herd behaviour, we elaborated on hype as a way of supporting technologies that had yet to take on a final shape, by providing the necessary drive to stabilise them as networks. Thus, hype and expectation dynamics seem to be part of a strategy adapted to understand what has still not been completely formalised; less like impervious lemmings jumping off a cliff and more like blind fungus spreading in random directions to find nourishment. We argued that it was precisely the scope that Digital Methods afforded that allowed us this perspective on the Sociology of Expectations, thus answering our second research question.

Finally we returned to our initial considerations of messy method and digital methods as a way of dealing with or navigating such messiness. We here positioned the digital methods (and our own digital experiments) in opposition to discourse analysis with a preference for order over mess, arguing for an acceptance of the messiness of social science. We argued that digital methods, and especially in the figuration we have demonstrated in this thesis, are imminently suited to account for the full scale of mess inherent in emergence of new phenomena. By forging a connection between Laws proposal of ANT after method and Digital Methods, we proposed a potential direction of future development in digital methods, thus answering our third sub-question.

9 APPENDIX

FOLLOWING OUR METHODOLOGICAL CONSIDERATIONS AND RESEARCH DESIGN, WE WILL IN THE FOLLOWING PROTOCOLS MOVE ONE STEP CLOSER TO THE TECHNIQUES USED DURING OUR ANALYSIS, ZOOMING IN ON THE CONCRETE PROCEDURES OF OUR DIGITAL EXPERIMENTS.

9.1 PROTOCOLS

Through the following protocols we wish to outline more concretely the practice followed when conducting our experiments. We will do so through a number of protocols, describing primarily our usage of the different digital tools and some of their central limitations. Since our study is relying on a high number of different tools and since some of the experiments are referred to across the different analytical phases, we will introduce the section with a table listing all out applied tools and relating them to our different analytical phases (see table 1).

Table 1: Tools, their phases and their objects of investegation.

Phase	Tool (protocol)	Object of Investegation
Time	1. Google Search results and Google Trends (A) 2. Google Autocomplete (B)	- Search results and Search volume - Related terms
Actors	1) Scientometrics (C) 2) Hypher (D) 3) ANTA (E)	- Citation network - Hyperlink network - Keyword + Actor network
Visions	*No digital tools applied. See chapter 3.2.3.*	

9.1.1 Protocol A - Search and trends

This digital perspective will focus on generating quantitative data for the attention (both supply and demand) to Big Data through the Google Search interface and Google Trend.

Google search is an interface most of us use everyday to find specific resources on Internet, though it is also an important source for evaluating relevance in digital methods (Marres 2012:161). In our investigation we make use of Google result estimation as an indicator of supply. By doing a simple search on “Big Data” and filtering the results based on our temporal periods we were able to make a rough estimate of the overall amount of web pages dealing with the topic.

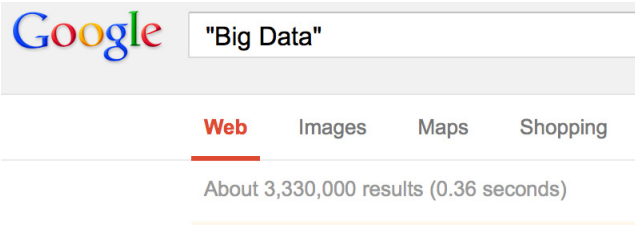
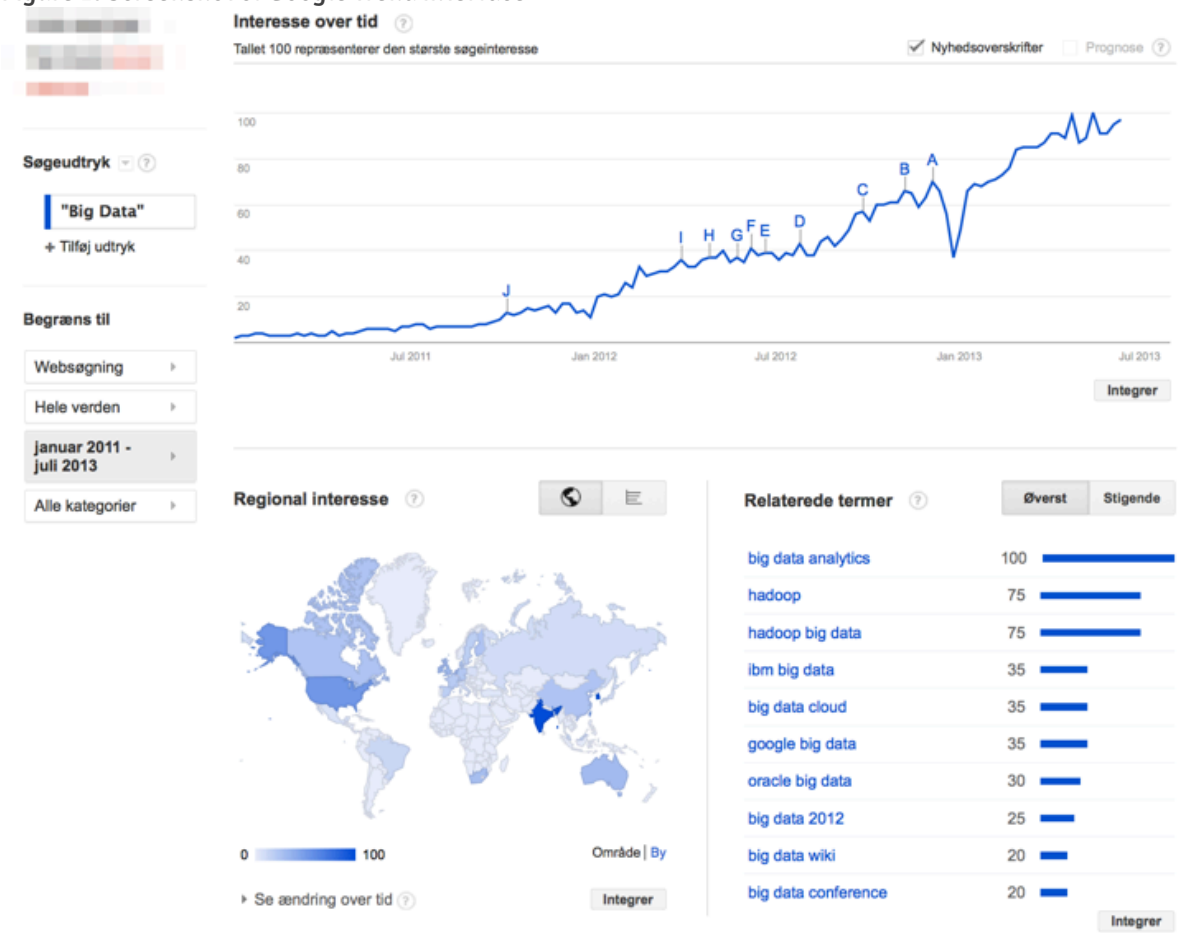


Figure 1: Example of Google Search. The number in the last line shows the number of results found by Google.

Google Trend is another interface by Google, which offers access to search trends and search volumes made on Google. It is in other words related to the Google Search interface, but focuses on the demand of specific search words. Besides extracting the search volumes, to uncover possible correlated developments we also used the interface to extract temporal and the geographical usage of the term “Big Data”, as well as the usage of related search terms.

Figure 2: Screenshot of Google Trend interface



9.1.2 Limitation

It has to be emphasised that the numbers presented by Google search interface alone is a rough and truncated estimate (Google 2013). Also, not all pages underlying the estimation are actually relevant pages, but are merely pages where the keyword appears. To take this into account, we will not focus excessively on the absolute numbers but instead shift attention towards relative growth over time.

Another important limitation derives from language barriers and understanding regional interest. In our analysis we identify the US, UK and India as central hubs for the Big Data interests. While it is not unlikely that these countries are in fact central hubs for the emergence of Big Data, a quick run through of all central countries reveals a predominance of English speaking countries, which could be an indication of local translation of 'Big Data'.

9.1.3 Protocol B - Google Autocomplete

Google Autocomplete[1] is a simple tool developed by the Digital Method Initiative, which makes it possible to gain insight into the related search terms of a given keyword. The tool accesses the Google Search API and retrieves a list of additional keywords that other people searching for the specified term also used. The tool hereby makes it possible to zero in on the interests of people searching for Big Data.

As starting point we used "Big Data" (capitalized and in quotation as with all our experiments). Based on preliminary results from our Google Trend investigation (Protocol A) we chose US, India and UK as the spatial origin for our search, producing a list of 30 related searches as well as the number of queries made on the specific search.

Based on this data we used Microsoft Excel to identify the most used related search in the three countries (figure 3). We also produced an overall ranking of the related terms with which we turned back to Google autocomplete to gain further depth by changing the search term from e.g. "Big Data" to "Big Data University". By repeating this process a number of times we gain increasingly specific insight into the interests of the people searching for Big Data.

9.1.4 Limitation

This remediation of the autocomplete function, with its simple interface and quick response, shouldn't interpreted too generally due to certain limitations, but do offer a quick entrance point to a given term. Most importantly, Google does not archive old recommendations, which makes it impossible for us to progress back in time. Additionally appears the suggestions also to be based on quite recent data, indicated by the "Big Data Week", a global event held only 2 months earlier, topping the related search results implying that recent data is heavily prioritized.

Figure 3: Screenshot of a data list produced by Google Autocomplete.

word	country	language	suggestion	num queries
Big+Data	United States	English	big data	1590000000
Big+Data	United States	English	big data analytics	94200000
Big+Data	United States	English	big data companies	781000000
Big+Data	United States	English	big data university	594000000
Big+Data	United States	English	big data jobs	1150000000

9.1.5 Protocol C - Scientometrics

Scientometrics (or bibliometrics) is more a set of techniques and methods than any specific tools, with a history that took its beginning long before the first computers[2]. Nonetheless the techniques have increased in popularity with digitalization, which have made the practice both easier and faster than earlier manual counting of citations.

Our mapping extracts its data from the online archives Scopus[3] and Web of Science (WOS)[4]. Besides extracting the cross references, we use Scopus' own analytical tools to gain an aggregated overview of the scientific field, publication date and sponsoring firms behind the articles identified.

The downloaded data were then converted from spreadsheets into networks through the converter *Table2Net*[5]. The data is converted into a citation network (articles/references) and a keyword network (articles/keywords). Both networks are also mapped with the publication dates of the articles, to allow a temporal exploration of developments in the three periods of our analysis (see chapter 4.2).

9.1.6 Limitations

Even though most books and articles today are published in some digital formats, the number of databases indexing citations is still incomplete, and most databases do not allow extracting the citations from their databases. In the moment of writing only two major databases, Scopus and Web of Science, fulfil both these requirements[6]. Though they both cover over 40 million articles, this is still far from offering a complete picture. Also problematic is the more prominently appearance of natural and life sciences in the databases.

9.1.7 Protocol D – Hypher

Hypher[7] (*Hypertext corpus initiative*) uses crawlers to uncover clusters of sites on the web through following how they are associated by hyperlinks (Girad 2011). Through lucky coincidence we were able to be the first researchers outside Medialab to experiment with the still not official released tool. The crawl is initiated by inputting a list of URLs of relevant websites. By harvesting all links on these websites, the crawlers creates a new list of websites from where all links is once again harvested. Associations are drawn based on how these sites link to each other through hyperlinks. The crawlers hereby slowly spins a web of interrelated websites, clustering websites together in groups based on their neighbours.

To gain a list of possible starting points we used Google to provide us with the 50 highest ranked articles when searching for Big Data. To make sure that the search results were unaffected of any prior knowledge Google might have had on us, we put up a 'Research Browser' (DMI 2012), which secured that the search result was not personalised. The number of 50 articles was chosen based on technical constraints, with every trial run with higher numbers quickly overburdened our server.

Through the tool *link harvester*[8] these search results were converted into a list of URLs. In the process any direct links to PDFs, movies or pictures, not viable to link crawling, were manually removed. The movies and pictures were simply deleted while the PDFs were stored for us to use in our last analysis (see protocol E below). The entire process was then repeated three times, in turn limiting the Google search to one of the three time periods identified in chapter 4. The departure point of each period is thus the most relevant pages for that period, which we practically equalled to a representation of the central actors.

Finally a 2nd degree crawl from the starting point was then initiated. Due to the high data amounts, the server where then left to itself for a number of days, before the three networks (one for each period) were exported to Gephi. In Gephi the identified pages were filtered based on the number of incoming links (in-degree). Also all social media sites and URL shorteners were removed. Lastly each page was given size based on an Eigenvector calculation (see section 2.4.2).

9.1.8 Limitations

Since Hypher is not officially released to the public, it is still lacking both functionality and stability, which has led to some limitation in our exploration. E.g. some sites are simply not (yet) "crawlable" by Hypher due to their technical structure. To limit the impact of these missing sites we have increased the number of starting points. Hypher also had some problems with some of the big online newspapers such as New York Times, and Financial Times. Our guess is that this was due to crawlers increasingly being blocked on news sites to protect original articles from being "stolen" by news aggregators such as the Huffington post (see e.g. Sullivan 2011; Huffington 2011).

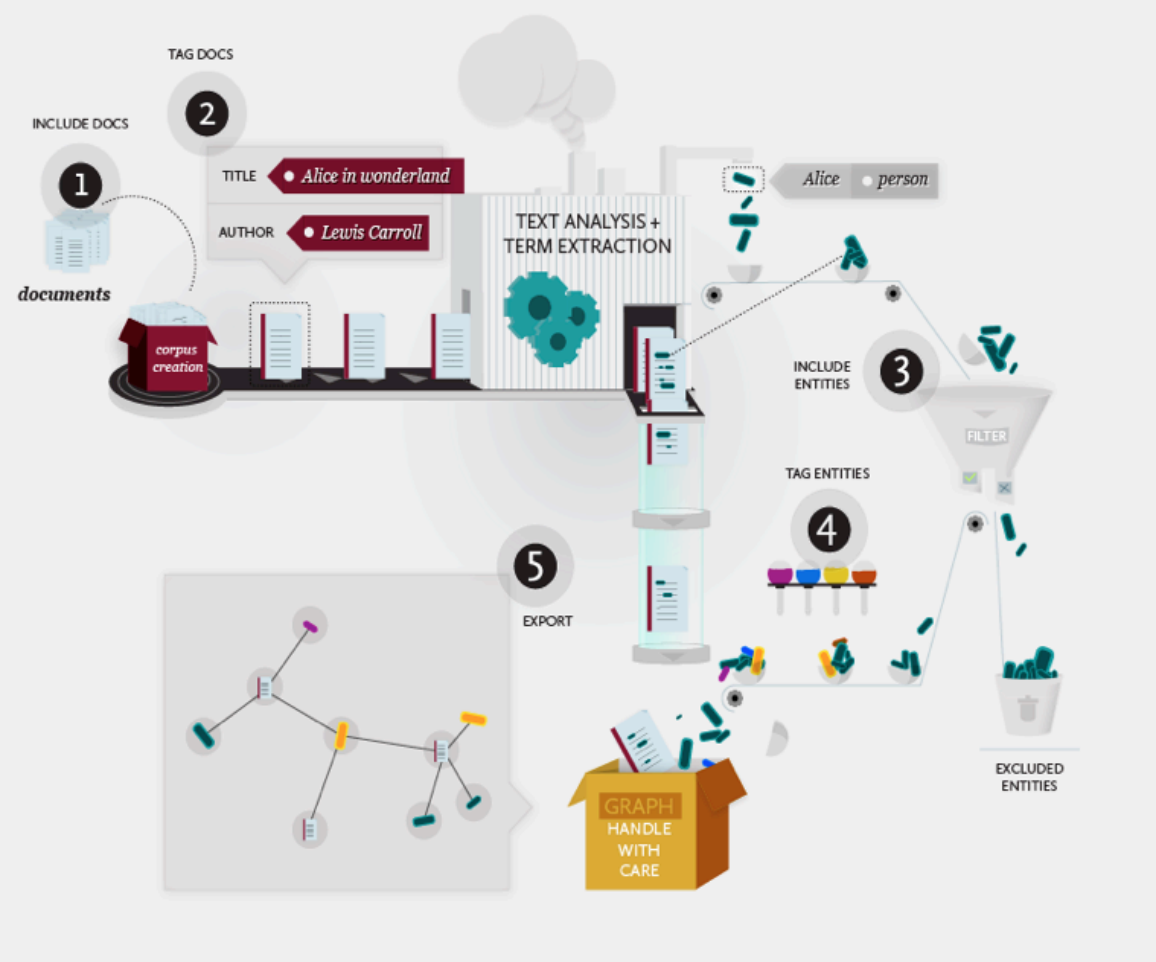
The last prominent limitation has been our limited ability to control the crawlers' temporal directions after they have been unleashed. We have been able to define the starting points in accordance with the three time periods of interests, but what the crawlers finds afterwards can potentially be sources from outside of the period, especially when the degree of the crawl (the number of iterations of harvesting and following links) rise. To counteract this we limited the degree to 2, and increased the number of starting URLs to balance this decrease in data points (see also chapter 5 for a in-depth discussion of this limitation).

9.1.9 Protocol E - ANTA

Actor Network Text Analysis or ANTA is a tool by Médialab developed to analyse text corpora. Through extracting central expressions and terms[9] in a set of texts and drawing a network based on their occurrence, ANTA offers aggregated insight on a delimited textual discourse[10].

Since the corpora studied is always predefined (in contrast to e.g. Hypher), special attention should be given to the selection of text sources. In our study we made use of the fact that our

Figure 4: Overview of the ANTA process, screenshot from the program.



ANTA analysis where the last of our analytical phases, basing our selection of sources on the nodal points emerging in the previous studies. Concretely we pulled 400+ texts from three different sources that we identify as central actors representative of particular spheres: General news (New York Times), Science (Scopus) and Consultancy & IT-firms (HBR + Whitepaper). During the selection, attention was given to secure a somewhat balanced representation[11] of the different sources before the different texts were categorized into our three different time periods. The selection of these spheres and sources are discussed further in chapter 9.2.

The 400+ texts were then analysed by ANTA extracting 20.000+ expressions of special interests and categorizing them based on their type (e.g. *firm, person, field term, country etc.*). We then reduced this mass of expressions to a more manageable amount (~1000) by filtering out expressions based on *document frequency* (number of documents the expression appears in) and *text frequency* (number of times the expression appears in the same document). Though ANTA is a gigantic step forward for automated text analysis, the process of identifying synonymous terms, i.e. Big Data and Big Datum, is still time consuming and tedious. As a final step before exporting the data from ANTA synonymous expressions were merged.

9.1.10 Limitations

Aggregated text analysis is not without problems since semantic value is always dependent on a context that the system might not be sensitive to. In our analysis it became apparent when "Justin Bieber" appeared with high frequency. We first excluded the term from our map, identifying it first as an erroneous anomaly due to ads or such, but a quick search on "Big Data" + "Justin Bieber" made us discover that the singers extreme popularity had made him an important and often studied entity in data analysis of Twitter.

Another important issue was deciding a level of aggregation in the process of merging data. The merging of 'data infrastructure' and 'data infrastructures' is obvious, but including the term 'large data infrastructure' is more doubtful. Should the term be merged with the previous two or is its largeness an important articulation of a new type of data infrastructure? Considerations such as these address the risk of reproducing already dominant narratives (Latour 2005, Bourdieu 1998).

9.2 DIVISION OF SPHERES

Based on our preliminary mappings we divided our actors into three groups:

General News

Source: New York Times

This site appears mainly in our two Hypher experiments, but is here a recurring central actor. Seem in recent years additionally to play an increasing role as commentator on Big Data. It is characterized by being a well-known and global newspaper, with a tendency towards the economic and business oriented news.

As source we selected New York Times based on 1) its status as the first world wide news channel with a well developed API providing easy access to 13 million articles, 2) its special and big *Bits* section focusing on "the business of technology", hereby representing the tendency of the general news actors of our mappings to tilt towards exactly business and technology, 3) its publication of a special issues and prodigious number of articles on big data.

Science

Source: Scopus

Our preliminary findings suggest that science discourse is partly disconnected from the general big data discussion. Wanting to secure a broad and diverse representation of big data we decided on Science as an individual sphere. The actors of this sphere are articles published in journals or academic conferences. Some exported in full length and other only as abstracts.

As source we chose Scopus based on 1) Its status as the biggest online article database supporting export of cross-references and documents, 2) Its availability to CBS students.

Business

Sources: Hayward business Review + Whitepapers

Business appeared as central actors in every experiment. As source we chose articles brought in Hayward Business Review (HBR) based on 1) its centrality in management literature 2) its appearance in our Hypher mapping. Through our investigation we discovered the many freely available whitepapers and reports and their centrality to the discourse. To incorporate these articulations in our view we added to the HBR articles reports and whitepapers that occupied central positions such as e.g. IBM's *Understanding Big Data*.

Notes

1. https://tools.digitalmethods.net/beta/scrapeGoogle/autocomplete.php

2. See (Pritchard 1969) for an historical overview of the term.

3. http://www.scopus.com/

4. http://www.webofknowledge.com/

5. http://tools.medialab.sciences-po.fr/

6. Through hacks it is also partly possible to use Google Scholar through tools such as ScholarScape (https://github.com/medialab/scholarScape). It is however a technically challenging (and partly illegal) act why we after a few quick attempts choose to give up the idea.

7. https://github.com/medialab/Hypertext-Corpus-Initiative

8. https://tools.digitalmethods.net/beta/harvestUrls/

9. Different software packages vary in their use of denominator: *term*, *expression* or *entity*. For clarity's sake, we will stick to *term.*

10. See (Venturini and Guido 2013) for further introduction to the tool.

11. A difficulty by obtaining a balance in representation is the different lengths of the text with e.g. books of 500 pages, which without moderation book would gain a hundred times more salience than a 5 page academic article.

10 BIBLIOGRAPHY

Actor Network Text Analyser. (n.d.). Retrieved from http://jiminy.medialab.sciences-po.fr/anta_dev/

Andersen, N. Å. (2003). Discursive analytical strategies. Bristol: The Policy Press. Retrieved from http://books.google.dk/books?id=solH3hCun70C&source=gbs_navlinks_s

Andersen, N. Å. (2006). Partnerskabelse. Retrieved from http://bog.nu/titler/partnerskabelse-niels-aakerstroem-andersen

Anderson, C. (2008, July 16). The end of Theory: The data Deluge makes the scientific method obsolete. Wired. Retrieved from http://www.uvm.edu/~cmplxsys/newsevents/pdfs/2008/anderson2008.pdf

ANTA. (n.d.). Retrieved January 7, 2013, from http://jiminy.medialab.sciences-po.fr/anta_dev/

Arbesman, S. (2013). Stop Hyping Big Data and Start Paying Attention to “Long Data” | Wired Opinion | Wired.com. Wired Opinion. Retrieved July 4, 2013, from http://www.wired.com/opinion/2013/01/forget-big-data-think-long-data/

Avanade. (2010). Global Survey: The business Impact of Big Data. Retrieved from http://www.avanade.com/Documents/Research%20and%20Insights/Big%20Data%20Executive%20Summary%20FINAL%20SEOv.pdf

Bakker, S. (2010). The car industry and the blow-out of the hydrogen hype. Energy Policy, 38(11), 6540–6544. doi:10.1016/j.enpol.2010.07.019

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In International AAAI conference on weblogs and social media (Vol. 2). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewPDFInterstitial/154Forum/1009

Berners-Lee, T. (2012). Sir Tim Berners-Lee: Raw data, now! Wired UK. Retrieved July 2, 2013, from http://www.wired.co.uk/news/archive/2012-11/09/raw-data

Big data: The next frontier for innovation, competition, and productivity. (n.d.).

Blok, A., Kyllingsbæk, S., Dreyer Lassen, D., & Axel Pedersen, M. (2011, December 4). CCCSS Workshop. CCCSS Workshop. Retrieved from http://sunelehmann.com/2011/12/04/cccss-workshop/

Borch, C. (2012). The politics of crowds : an alternative history of sociology. Cambridge; New York: Cambridge University Press.

Borup, M., Brown, N., Konrad, K., & Van Lente, H. (2006). The sociology of expectations in science and technology. Technology Analysis & Strategic Management, 18(3-4), 285–298. doi:10.1080/09537320600777002

Bourdieu, P. (1993). The field of cultural production: Essays on art and literature. Columbia University Press.

boyd, danah, & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA. Information, Communication & Society, 15(5), 662–679. doi:10.1080/1369118X.2012.678878

Boyd, danah, & Crawford, K. (2011). Six Provocations for Big Data. SSRN Electronic Journal. doi:10.2139/ssrn.1926431

Brockmeier, J. (2012). Strata Conference 2012: The End of Big Data Hype? ReadWrite. Retrieved July 4, 2013, from http://readwrite.com/2012/02/29/strata-conference-2012-the-end

Brown, N. (2000). Contested futures: a sociology of prospective techno-science. Aldershot, England; Burlington, VT: Ashgate.

Brown, N. (2003). Hope Against Hype - Accountability in Biopasts, Presents and Futures. SCIENCE STUDIES AN INTERDISCIPLINARY JOURNAL FOR SCIENCE AND TECHNOLOGY STUDIE, 16(2), 3–21.

Brown, N., & Michael, M. (2003). A Sociology of Expectations: Retrospecting Prospects and Prospecting Retrospects. Technology Analysis and Strategic Management, 15(1), 3–18.

Bryant, R., Katz, R., & Lazowska, E. (2008). Big-Data Computing: creating revolutionary breakthroughs in commerce, science and society.

Callon, M. (1986). Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. In J. Law (Ed.), Power, action, and belief : a new sociology of knowledge? London; Boston: Routledge & Kegan Paul.

Callon, M. (1987). The sociology of an actor-network: The case of the electric vehicle. Mapping the dynamics of science and technology, 23. Retrieved from http://epl.scu.edu:16080/~stsvalues/readings/Callon.pdf

Callon, M. (Ed.). (1998). The laws of the markets. Oxford ; Malden, MA: Blackwell Publishers/Sociological Review.

Callon, M. (2004). The role of hybrid communities and socio-technical arrangements in the participatory design. Journal of the center for information studies, 5(3), 3–10.

Callon, M., Barthe, Y., & Lascoumes, P. (2011). Acting in an uncertain world: an essay on technical democracy. Cambridge: Mit Press.

Callon, M., Méadel, C., & Rabeharisoa, V. (2002). The economy of qualities. Economy and Society, 31(2), 194–217. doi:10.1080/03085140220123126

Carrington, P. J., & Scott, J. (Eds.). (2011). Saga Handbook on Social Network Analysis (1st ed.). Thousand Oaks, CA: SAGE Publications.

Castells, M. (2003). The Internet galaxy: reflections on the Internet, business, and society. Oxford [u.a.]: Oxford Univ. Pr.

Cook, J. (2013). How big data will transform politics, education and just about everything else - GeekWire. GeekWire. Retrieved July 4, 2013, from http://www.geekwire.com/2013/big-data-transform-politics-education/

Crawford, K. (2013, April 1). The Hidden Biases in Big Data. Harvard Business Review. Retrieved from http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html?goback=%2Egde_112700_member_235129730

Cukier, K. (2010, February 25). Data, data everywhere. The Economist. Retrieved from http://www.economist.com/node/15557443

Di Justo, P. (2008). Sorting the World: Google Invents New Way to Manage Data. WIRED. Retrieved July 2, 2013, from http://www.wired.com/science/discoveries/magazine/16-07/pb_sorting

DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. American Sociological Review, 48(2), 147. doi:10.2307/2095101

dlopezgo-WatsonG.ris. (n.d.).

DMI. (2007, July 15). The spheres. Digital method initiative. Retrieved from https://www.digitalmethods.net/Digitalmethods/TheSpheres

DMI. (2012, November 18). DMI Tools firefox extension. Digital method initiative. Retrieved June 4, 2013, from https://wiki.digitalmethods.net/Dmi/FirefoxToolBar

Eames, M., Mcdowall, W., Hodson, M., & Marvin, S. (2006). Negotiating contested visions and place-specific expectations of the hydrogen economy. Technology Analysis & Strategic Management, 18(3-4), 361–374. doi:10.1080/09537320600777127

Elkjær, B. (2012, November 28). Stavefejl sender Krause-Kjær på vildspor. Journalisten.dk. Retrieved from http://www.journalisten.dk/stavefejl-sender-krause-kj-r-p-vildspor

Esmark, A., Bagge Laustsen, C., & Åkerstrøm Andersen, N. (2005). Socialkonstruktivistiske analysestrategier. Frederiksberg: Roskilde Universitetsforlag.

Flyverbom, M. (2011). The Power of Networks: Organizing the Global Politics of the Internet. Edward Elgar Pub.

Foster, I. A. (2012). The Algorithmic Magic of Trendspotting - Datanami. Retrieved July 4, 2013, from http://www.datanami.com/datanami/2012-09-20/the_algorithmic_magic_of_trendspotting.html

Freeman, L. (2011). The development of social network analysis - with an emphasis on recent events. In The Sage handbook of social network analysis (1st ed., pp. 26–39). Thousand Oaks, CA: SAGE Publications.

Fuchs, C. (2003). The Internet as a Self-Organizing Socio-Technological System (SSRN Scholarly Paper No. ID 458680). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=458680

Gad, C., & Jensen, C. B. (2009). On the Consequences of Post-ANT. Science, Technology & Human Values, 35(1), 55–80. doi:10.1177/0162243908329567

Geels, F. W., & smit, W. (2000). Lessons from Failed Technology Futures: Potholes in the Road to the Future. Contested Futures: A Sociology of Prospective Techno-Science, 129.

Giddens, Anthony. (1998). Risk Society: the Context of British Politics. In The politics of risk society. Cambridge ; Malden, Mass: Polity Press.

Girard, P. (2011). HyperText Corpus Initiative: how to help researchers sieving the web? In Proposal for the “Using Web Archives” panel, Out of the Box Conference (Vol. 9). Retrieved from http://www.medialab.sciences-po.fr/publications/Girard-HCI.pdf

Google. (2013, May 27). Google search result count. Retrieved from http://support.google.com/webmasters/bin/answer.py?hl=en&answer=70920

Gorm Hansen, B. (2011). Adapting in the Knowledge Economy: Lateral Strategies for Scientists and Those Who Study Them. Copenhagen Business SchoolCopenhagen Business School, Institut for Ledelse, Politik og FilosofiDepartment of Management, Politics and Philosophy. Retrieved from http://openarchive.cbs.dk/handle/10398/8346

Graham, M. (2012). Geography/Internet: Ethereal Alternate Dimensions of Cyberspace or Grounded Augmented Realities? (SSRN Scholarly Paper No. ID 2166874). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2166874

Gyldendal (Ed.). (2009, February 17). Opslag: falsifikation. In Den Store Danske. Gyldendal. Retrieved from http://www.denstoredanske.dk/Sprog,_religion_og_filosofi/Filosofi/Filosofiske_begreber_og_fagudtryk/falsifikation?highlight=falsifikation

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspectives. Feminist studies, 14(3), 575–599.

Hedgecoe, A., & Martin, P. (2003). The Drugs Don’t Work: Expectations and the Shaping of Pharmacogenetics. Social Studies of Science, 33(3), 327–364. doi:10.1177/03063127030333002

Hellerstein, J. (2008, November 9). Parallel Programming in the Age of Big Data. GigaOM. Retrieved July 4, 2013, from http://gigaom.com/2008/11/09/mapreduce-leads-the-way-for-parallel-programming/

Holder, A. A., Wootton, J. C., Baron, A. J., Chambers, G. K., & Fincham, J. R. (1975). The amino acid sequence of Neurospora NADP-specific glutamate dehydrogenase. Peptic and chymotryptic peptides and the complete sequence. The Biochemical journal, 149(3), 757–773.

Horowitz, M. (2008). Visualizing Big Data: Bar Charts for Words. WIRED. Retrieved July 3, 2013, from http://www.wired.com/science/discoveries/magazine/16-07/pb_visualizing

Huffington, A. (2011, March 10). The Leaky New York Times Paywall & How “Google Limits” Led To “Search Engine Limits.” Huff Pst Media. Retrieved from http://www.huffingtonpost.com/arianna-huffington/bill-keller-accuses-me-of_b_834289.html

Hypertext-Corpus-Initiative. (n.d.). Retrieved January 25, 2013, from https://github.com/medialab/Hypertext-Corpus-Initiative

IBM. (2012). Understanding big data : analytics for enterprise class Hadoop and streaming data. (P. Zikopoulos, Ed.). New York: McGraw-Hill.

Issuemapping.net. (n.d.). Tools for issue mapping. Retrieved January 7, 2013, from http://issuemapping.net/Main/Tools

Jacomy, M., Heymann, S., Venturini, T., & Bastian, M. (2011). ForceAtlas2, a graph layout algorithm for handy network visualization. Paris http://www. medialab. sciences-po. fr/fr/publications-fr.

Järvenpää, H. M., & Mäkinen, S. J. (2008). An empirical study of the existence of the Hype Cycle: A case of DVD technology (pp. 1–5). IEEE. doi:10.1109/IEMCE.2008.4617999

Jo Foley, M. (2011, November 16). Microsoft drops Dryad; puts its big-data bets on Hadoop. All About Microsoft - ZDNet. Retrieved from http://www.zdnet.com/blog/microsoft/microsoft-drops-dryad-puts-its-big-data-bets-on-hadoop/11226

Jun, S.-P. (2011). An empirical study of users' hype cycle based on search traffic: the case study on hybrid cars. Scientometrics, 91(1), 81–99. doi:10.1007/s11192-011-0550-3

Jun, S.-P. (2012). A comparative study of hype cycles among actors within the socio-technical system: With a focus on the case study of hybrid cars. Technological Forecasting and Social Change, 79(8), 1413–1430. doi:10.1016/j.techfore.2012.04.019

jurgenson, nathan. (2011, February 24). Digital Dualism versus Augmented Reality. Cyborgology. Retrieved July 23, 2013, from http://thesocietypages.org/cyborgology/2011/02/24/digital-dualism-versus-augmented-reality/

Justesen, L. N., & Mouritsen, J. (2008). The Triple Visual : Translations between Photographs, 3-D Visualizations and Calculations. Retrieved July 23, 2013, from http://research.cbs.dk/portal/en/publications/the-triple-visual(e47e6250-c5ea-11dd-ab34-000ea68e967b).html

Kelly, M. (2013). The Big Data Revolution. Medium. Retrieved July 3, 2013, from https://medium.com/how-to-use-the-internet/554ecb1eca73

Kemp, J. (2007). Gartner Research's Hype Cycle diagram. Retrieved from http://en.wikipedia.org/wiki/File:Gartner_Hype_Cycle.svg

Konrad, K. (2006). The social dynamics of expectations: The interaction of collective and actor-specific expectations on electronic commerce and interactive television. Technology Analysis & Strategic Management, 18(3-4), 429–444. doi:10.1080/09537320600777192

Laclau, E., & Mouffe, C. (2002). Det radikale demokrati : diskursteoriens politiske perspektiv. Frederiksberg: Roskilde Universitetsforlag.

Latour, B. (1986). Visualization and cognition: Drawing Things Together. Knowledge and society, 6, 1–40.

Latour, B. (1987). Science in action : how to follow scientists and engineers through society. Cambridge, Mass.: Harvard University Press.

Latour, B. (1988). Politics of Explanation. In Knowledge and reflexivity: new frontiers in the sociology of knowledge.

Latour, B. (1993). The pasteurization of France. Cambridge, Mass.: Harvard University Press.

Latour, B. (1994). On Technical Mediation -- Philosophy, Sociology, Genealogy. Common Knowledge, 3(2), 29–64.

Latour, B. (1996). On actor-network theory. A few clarifications plus more than a few complications. Soziale Welt, 47, 369–381.

Latour, B. (1998). On Actor Network Theory: A few clarifications. Nettime. CSI-Paris. Retrieved December 5, 2012, from http://www.nettime.org/Lists-Archives/nettime-l-9801/msg00019.html

Latour, B. (2004a). How to Talk About the Body? the Normative Dimension of Science Studies. Body & Society, 10(2-3), 205–229. doi:10.1177/1357034X04042943

Latour, B. (2004b). Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern. Critical Inquiry, 30(2), 225–248. doi:10.1086/421123

Latour, B. (2005). Reassembling the social : an introduction to actor-network-theory. Oxford; New York: Oxford University Press.

Latour, B. (2007). Beware your imagination leaves digital traces. Times Higher Literary Supplement. Retrieved from http://en.wikipedia.org/wiki/Digital_traces#CITEREFKieronTuffieldShadbolt2009

Latour, B. (2011). Networks, Societies, Spheres: Reflections of an Actor-Network Theorist. In International Journal of Communication (Vol. 5, pp. 796–810). Presented at the Network Multidimensionality in the Digital Age. Retrieved from http://ijoc.org/ojs/index.php/ijoc/article/view/1094/558

Latour, B., Jensen, P., Venturini, T., Sébastian, G., & Boullier, D. (2012). The Whole is Always Smaller Than Its Parts' A Digital Test of Gabriel Tarde's Monads. British Journal of Sociology, Forthcoming. Retrieved from http://www.bruno-latour.fr/node/330

Latour, B., & Woolgar, S. (1979). Laboratory life : the construction of scientific facts. Princeton, N.J.: Princeton University Press.

Law, J. (2004). After method: mess in social science research. London ; New York: Routledge.

Law, J. (2007). Actor Network Theory and Material Semiotics version of 25th April 2007. Retrieved from http://www.heterogeneities.net/publications/Law2007ANTandMaterialSemiotics.pdf

Law, J., & Hassard, J. (1999). Actor Network Theory and After. Wiley.

Law, J., & Urry, J. (2004). Enacting the social. Economy and Society, 33(3), 390–410. doi:10.1080/0308514042000225716

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., … Van Alstyne, M. (2009). SOCIAL SCIENCE: Computational Social Science. Science, 323(5915), 721–723. doi:10.1126/science.1167742

Lehmann, S. (2012). Ultra-detaljerede komplekse netværk. In Villum og Velux Fondens årsskrift (pp. 64–67).

Lohr, S. (2012, August 11). How Big Data Became So Big - Unboxed. The New York Times. Retrieved from http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html

Lösch, A. (2006). Anticipating the futures of nanotechnology: Visionary images as means of communication. Technology Analysis & Strategic Management, 18(3-4), 393–409. doi:10.1080/09537320600777168

MacKenzie, D. A. (1990). Inventing accuracy: a historical sociology of nuclear missile guidance. Cambridge, Mass.: MIT Press.

MacKenzie, D. A. (2006). An engine, not a camera: how financial models shape markets. Cambridge, Mass: MIT Press.

Madsen, A. K. (2012). Web-Visions as Controversy-Lenses. Interdisciplinary Science Reviews, 37(1), 51–68. doi:10.1179/0308018812Z.0000000004

Madsen, A. K. (2013). Web-Visions Repurposing digital traces to organize social attention (PhD Dissertation). Copenhagen Business School.

Marcus, G. (2013). Steamrolled by Big Data. Retrieved July 4, 2013, from http://www.newyorker.com/online/blogs/elements/2013/04/steamrolled-by-big-data.html?mobify=0

Marres, Noorthe, & Weltevrede, E. (2012). Scraping the Social? Issues in real-time social research. Journal of Cultural Economy, 1–52.

Marres, Noortje. (2012). The redistribution of methods: on intervention in digital social research, broadly conceived. The Sociological Review, 60, 139–165. doi:10.1111/j.1467-954X.2012.02121.x

Marres, Noortje, & Rogers, R. (2008). Subsuming the ground: how local realities of the Fergana Valley, the Narmada Dams and the BTC pipeline are put to use on the Web. Economy and Society, 37(2), 251–281. doi:10.1080/03085140801933314

Marx, V. (2013). Biology: The big challenges of big data. Nature, 498(7453), 255–260. doi:10.1038/498255a

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Business Review. Retrieved July 2, 2013, from http://hbr.org/2012/10/big-data-the-management-revolution/ar/1

McCombs, M. E., & Shaw, D. L. (1972). The Agenda-Setting Function of Mass Media. Public Opinion Quarterly, 36(2), 176–187. doi:10.1086/267990

Mckinsey. (2011). Big data: the next frontier for innovation, competition, and productivity. Lexington, KY: McKinsey.

medialab/Hypertext-Corpus-Initiative · GitHub. (n.d.). Retrieved from https://github.com/medialab/Hypertext-Corpus-Initiative

Mol, A. (2002). The body multiple: ontology in medical practice. Durham: Duke University Press.

Newman, M. E. J., Barabási, A.-L., & Watts, D. J. (2006). The structure and dynamics of networks. Princeton, N.J.; Oxford: Princeton University Press. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=399094

Nwana, H. S. (1996). Software agents: An overview. Knowledge engineering review, 11(3), 205–244.

Pariser, E. (2011). The filter bubble: what the Internet is hiding from you. New York: Penguin Press.

[PDF] from sciences-po.fr. (n.d.-a). Retrieved from http://www.medialab.sciences-po.fr/publications/Girard-HCI.pdf

[PDF] from sciences-po.fr. (n.d.-b). Retrieved from http://medialab.sciences-po.fr/publications/Venturini_HCI_DossierINA.pdf

Pentland, A. (2010). Honest signals: how they shape our world. Cambridge, Mass.; London: MIT Press.

Peters, B. (2012, September 6). Do you really want to get aboard the Big Data train? Forbes. Retrieved from http://www.forbes.com/sites/bradpeters/2012/09/06/133/

Pollock, N., & Williams, R. (2010). The business of expectations: How promissory organizations shape technology and innovation. Social Studies of Science, 40(4), 525–548. doi:10.1177/0306312710362275

Pritchard, A. (1969). Statistical bibliography or bibliometrics? Journal of Documentation, 4(25), 348–349.

Rao, V. (2012). Data, Data Everywhere, Not a Byte to Drink. Forbes. Retrieved July 3, 2013, from http://www.forbes.com/sites/venkateshrao/2012/02/24/data-data-everywhere-not-a-byte-to-

drink/

Ratner, H. (2009). License to Kill: The word is not Enough. Hvorfor, 3, 97–107.

Ratner, H. (2012). Promises of Reflexivity: Managing and Researching Inclusive Schools. Frederiksberg.

Regalado, A. (2013). Why Google Hates the Term Big Data | MIT Technology Review. MIT Technology Review. Retrieved July 11, 2013, from http://www.technologyreview.com/view/515941/just-dont-call-it-big-data/

Ritzer, G. (2005). Encyclopedia of social theory Vol. 1, [A - M]. Thousand Oaks, Calif. [u.a.]: SAGE.

Robinson, J. V., & James, A. L. (1975a). Some observations on the effects produced in white mice following the injection of certain suspensions of corroding bacilli. British journal of experimental pathology, 56(1), 14–16.

Robinson, J. V., & James, A. L. (1975b). Some observations on the effects produced in white mice following the injection of certain suspensions of corroding bacilli. British journal of experimental pathology, 56(1), 14–16.

Rogers, R. (2009). End of the virtual digital methods: inaugural lecture delivered on the appointment to the Chair of New Media & Digital Culture at the University of Amsterdam on 8 May 2009. Retrieved May 13, 2013, from http://site.ebrary.com/id/10363484

Rogers, R. (2013). Digital methods. Cambridge, Massachusetts; London: The MIT Press.

Ruef, A., & Markard, J. (2010). What happens after a hype? How changing expectations affected innovation activities in the case of stationary fuel cells. Technology Analysis & Strategic Management, 22(3), 317–338. doi:10.1080/09537321003647354

Schroeder, R., & Meyer, E. T. (2012). The Scientific Styles of e-Research. Retrieved from http://citation.allacademic.com/meta/p_mla_apa_research_citation/5/7/8/7/6/p578762_index.html

Scott, J. (2011). Social physics and social networks. In The Sage handbook of social network analysis (1st ed., pp. 55–66). Thousand Oaks, CA: SAGE Publications.

Serres, M. (1995). Angels: A modern myth. Flammarion.

Simon, H. (1991). Bounded Rationality and Organizational Learning. Retrieved July 11, 2013, from http://orgsci.journal.informs.org/content/2/1/125

Smarty, A. (N/A). The Number of Google Results Found: What It Really Means - SEO Chat. SEO Chat. Retrieved from http://www.seochat.com/c/a/search-engine-spiders-help/the-number-of-google-results-found-what-it-really-means/

Sommerlund, J., & Jespersen, A. P. (2008). Fashion, Mediations & Method Assemblages.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social studies of science, 19(3), 387–420.

Steinert, M., & Leifer, L. (2010). Scruntinizing Gartner's hype cycle approach 2010. In Proceedings of PICMET 10 (pp. 1–13). Presented at the International Conference on Management of Engineering & Technology, IEEE.

Stengers, I. (1997). Power and invention: situating science. Minneapolis: University of Minnesota Press.

Sullivan, D. (2011, March 22). The Leaky New York Times Paywall & How “Google Limits” Led To “Search Engine Limits.” Search Engine Land. Retrieved from http://searchengineland.com/leaky-new-york-times-paywall-google-limits-69302

Taleb, N. (2013). Beware the Big Errors of “Big Data” | Wired Opinion | Wired.com. Retrieved July 4, 2013, from http://www.wired.com/opinion/2013/02/big-data-means-big-errors-people/

Tarde, G. (1903). The laws of imitation. (E. W. C. Parsons & F. H. Giddings, Eds.). Charleston: Bibliolife.

The Economist. (2012, December 10). Let the data flow—and live longer and better... The Economist. Retrieved from http://www.economist.com/blogs/theworldin2013/2012/12/data-data-everywhere

Turner, B. S., & Wiley InterScience (Online service). (2009). The new Blackwell companion to social theory. Chichester, West Sussex, United Kingdom; Malden, MA, USA: Wiley-Blackwell. Retrieved from http://dx.doi.org/10.1002/9781444304992

Urry, J. (2004). Small Worlds and the New “Social Physics.” Global Networks, 4(2), 109–130. doi:10.1111/j.1471-0374.2004.00083.x

Van Couvering, E. (2007). Is relevance relevant? Market, science, and war: Discourses of search engine quality. Journal of Computer-Mediated Communication, 12(3), 866–887.

Van Lente, H. (1993). Promising technology: the dynamics of expectations in technological developments. Eburon, Delft.

Van Lente, H., & Rip, A. (1998). The Rise of Membrane Technology: From Rhetorics to Social Reality. Social Studies of Science, 28(2), 221–254.

Van Lente, H., Spitters, C., & Peine, A. (2013). Comparing technological hype cycles: Towards a theory. Technological Forecasting and Social Change. doi:10.1016/j.techfore.2012.12.004

Van Merkerk, R. O., & Douglas K. R. Robinson. (2006). Characterizing the emergence of a technological field: Expectations, agendas and networks in Lab-on-a-chip technologies. Technology Analysis & Strategic Management, 18(3-4), 411–428. doi:10.1080/09537320600777184

Van Merkerk, R. O., & van Lente, H. (2005). Tracing emerging irreversibilities in emerging technologies: The case of nanotubes. Technological Forecasting and Social Change, 72(9), 1094–1111. doi:10.1016/j.techfore.2004.10.003

Venturini, T. (2010). Building on faults: How to represent controversies with digital methods. Public Understanding of Science, 21(7), 796–812. doi:10.1177/0963662510387558

Venturini, Tommaso. (2009). Diving in magma: how to explore controversies with actor-network theory. Public Understanding of Science, 19(3), 258–273. doi:10.1177/0963662509102694

Venturini, Tommaso. (2011). Hypertext Corpus Initiative. Observa Working papers. Retrieved from http://medialab.sciences-po.fr/publications/Venturini_HCI_DossierINA.pdf

Venturini, Tommaso. (2012). What is second-degree objectivity and how could it be represented. draft. Retrieved from http://www.medialab.sciences-po.fr/publications/Venturini-Second_Degree_Objectivity_draft1.pdf

Venturini, Tommaso, & Guido, D. (2013). Once Upon a Text: An ANT Tale in Text Analysis. Sociologica. Retrieved from http://www.medialab.sciences-po.fr/publications/Venturini_Guido-Once_Upon_A_Text.pdf

Venturini, Tommaso, & Latour, B. (2010). The Social Fabric: Digital Traces and Quali-quantitative Methods. In Future En Seine 2009. Presented at the The digital future of the city, Paris: Cap Digital.

Vimeo: Tutorial ANTA. (n.d.). Retrieved January 7, 2013, from http://vimeo.com/45433706

Watson, G. (2003). Actor Network Theory, After-ANT & Enactment: Implications for method. gavan.ca. Retrieved from http://www.gavan.ca/wp-content/uploads/2007/01/ANT_comp.pdf

Weick, K. E. (1995). Sensemaking in organizations. Thousand Oaks: Sage Publications.

Weinberger, D. (2012, January 3). To Know, but Not Understand: David Weinberger on Science and Big Data. The Atlantic. Retrieved July 3, 2013, from http://www.theatlantic.com/technology/archive/2012/01/to-know-but-not-understand-david-weinberger-on-science-and-big-data/250820/

Whatmore, S. (2003). Generating materials. In M. Pryke & G. Rose (Eds.), Using social theory : thinking through research. London; Thousand Oaks, Calif.: SAGE in Association with the Open University.

Wired. (2008). The Petabyte Age: Because More Isn't Just More — More Is Different. WIRED. Retrieved July 2, 2013, from http://www.wired.com/science/discoveries/magazine/16-07/pb_intro